Evaluating cancer cell lines as models for metastatic breast cancer

Ke Liu[1,2], Patrick Newbury [1,2], Benjamin S. Glicksberg [3], Eran R. Andrechek [4], Bin Chen[1,2,*]

1 Department of Pediatrics and Human Development, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA
2 Department of Pharmacology and Toxicology, College of Human Medicine, Michigan State University, Grand Rapids, MI, USA
3 Institute for Computational Health Sciences, University of California San Francisco, San Francisco, CA, USA
4 Department of Physiology, Michigan State University, East Lansing, MI, USA

*Bin.Chen@hc.msu.edu

Abstract

Metastasis is the most common cause of cancer-related death and, as such, there is an urgent need to discover new therapies to treat metastasized cancers. Cancer cell lines are widely-used models to study cancer biology and evaluate drug candidates. However, it is still unknown whether they adequately recapitulate the disease in patients. The recent accumulation of large-scale genomic data in cell lines, mouse models, and patient tissue samples provides an unprecedented opportunity to evaluate the suitability of cancer cell lines as models for metastatic cancer research. Through comprehensively comparing the genomic profiles of 57 breast cancer cell lines with those of metastatic breast cancer samples, we found their substantial genomic differences. We also identified cell lines that more closely resemble different subtypes of metastatic breast cancer. However, we found none of the currently established Basal-like cell lines sufficiently resemble the samples of Basal-like metastatic breast cancer, a subtype of high interest in therapeutic discovery. Further analysis of mutation, copy number variation and gene expression data suggested that MDAMB231, the most commonly used triple negative cell line, had little genomic similarity with Basal-like metastatic breast cancer samples. Our work demonstrates an urgent need of cell lines that more closely resemble Basal-like metastatic breast cancer samples, and could guide cell line selection in other metastasis-related translational research.

Author summary

Why was this study done?

> Cancer cell lines are commonly used as models to understand cancer metastasis and test drug candidates preclinically, while the degree to which these cell lines accurately reflect metastatic breast cancer *in vivo* is not well established. We leveraged large-scale genomic data to comprehensively evaluate breast cancer cell lines as models for metastatic breast cancer.

What did the researchers do and find?

> The comparison of genetic profiles between breast cancer cell lines and metastatic breast cancer samples revealed that cell lines poorly recapitulated the somatic mutation spectrum of metastatic breast cancer samples, while copy-number variation profiles were considerably consistent.
> Gene expression correlation analysis identified cell lines which closely resembled metastatic breast cancer samples of LuminalA/LuminalB/Her2-enriched subtypes, but none of the currently established cell lines resembles Basal-like metastatic breast cancer samples. Specifically, the most commonly used triple negative cell line MDAMB231 had low genomic similarity with metastatic breast cancer samples from patients.
> Global transcriptome analysis revealed striking differences between breast cancer cell lines and metastatic breast cancer samples.

What do these findings mean?

> These findings indicate that we should keep in mind the large genomic disparity between breast cancer cell lines and metastatic breast cancer samples when using cell lines in translational research, and we are still in urgent need of new cell lines (or other preclinical models) for Basal-like metastatic breast cancer research.

Introduction

Cancer cell lines were initially derived from tumors and cultured in a 2D environment. Because of the merit of cell culture, they have been widely used as models to study cancer biology and test drug candidates [1]. However, the fact that many drugs have great preclinical profiles but fail in the clinic urges the reinvestigation of cell lines as tumor models [2]. The differences between cell lines and tumors have raised the critical question to what extent do cell lines recapitulate the biology of tumor samples [3, 4].

The emergence of large-scale genomic data provides an unprecedented opportunity to quantify their biological differences. The Cancer Genome Atlas (TCGA) project characterized both genetic and transcriptomic profiles of more than 10,000 human tissue samples across 32 tumor types [5]. The Cancer Cell Line Encyclopedia (CCLE) characterized both genetic and transcriptomic profiles for more than 1,000 cell lines [6]. Silvia et al. performed comprehensive comparison of molecular profiles between 47 ovarian cancer cell lines and ovarian tumor samples, and they showed that popular cell line models did not closely resemble high-grade serous ovarian cancers [7]. In addition, they identified several rarely used cell lines that closely resembled the profile of ovarian cancer. We examined the transcriptome similarity between hepatocellular carcinoma (HCC) cell lines and HCC tumor samples and demonstrated that nearly half of the HCC cell lines did not resemble HCC tumors [8]. Jian et al. conducted a comprehensive comparison of molecular portraits between breast cancer cell lines and primary breast cancer samples and found both similar and dissimilar molecular features [10].

Cancer metastasis is the most common cause of cancer-related death, thus there is an urgent need of new drugs for treating cancer metastasis [11, 12]. Previous cell line evaluation analysis were mainly performed in reference to primary tumors. It remains unknown whether cell lines closely resemble metastatic cancer and thus are appropriately used in translational research. Robinson et al. performed whole-exome and transcriptome sequencing on about 500 metastatic cancer samples and recently released their dataset (refer to MET500) [13]. This large-scale genomic profiling combined with existing genomic data allows the evaluation of the suitability of cell lines as models for metastatic cancer. As a case study, in this work we focus on breast cancer, where cell lines are frequently used to study metastasis.

We found breast cancer cell lines poorly recapitulated somatic mutation spectrum while the CNV (copy-number-variation) profiles were highly consistent. In addition, we showed cell lines could resemble the transcriptome of metastatic breast cancer and identified suitable cell lines for LuminalA/LuminalB/Her2-enriched subtypes. Our analysis indicated that none of the cell lines closely resembled Basal-like metastatic breast cancer samples. Specifically, the heavily-used triple negative cell line MDAMB231 shows low genomic similarity with metastatic breast cancer samples, which could be confirmed by using independent *in vivo* and *in vitro* data. Our work reveals the similarity and difference between metastatic breast cancer samples and cell lines and provides guidance for choosing cell lines in metastasis-related translational research.

Materials and methods

Datasets

The raw RNASeq data of MET500 samples were downloaded from dbGap (under accession number phs000673.v2.p1) and further processed using RSEM [14, 15]. The FPKM values were used as gene expression measure. The somatic mutation and copy number variation (CNV) data of MET500 samples were downloaded from MET500 web portal (https://met500.path.med.umich.edu/downloadMet500DataSets). All CCLE data (including gene expression profiled by RNASeq and microarray, somatic mutation call and CNV) were downloaded from the CCLE data portal (https://portals.broadinstitute.org/ccle). Somatic mutation calling results of TCGA breast cancer samples were downloaded from cBioPortal [16, 17] (http://www.cbioportal.org/) and CNV data were downloaded from BROAD GDAC Firehose (https://gdac.broadinstitute.org/). RSEM-processed TCGA gene expression data were downloaded from UCSC Xena data portal (http://xena.ucsc.edu/) [18]. Besides the MET500 dataset, we also searched GEO and manually assembled another microarray dataset which contains gene expression value of 117 metastatic breast cancer samples [19–22]. The GEO accession numbers used were GSE11078, GSE14017, GSE14018, and GSE54323.

Identification of differentially mutated genes between MET500 and TCGA

Given a gene, to test whether it has significantly higher mutation frequency in metastatic breast cancer samples, we computed the right-tailed p-value as follows:

$$p_1 = 1 - \sum_{i=0}^{n} Pr(i; N, \hat{q}) \tag{1}$$

Where Pr(i;N,$\hat{q}$) is the probability mass function of binomial distribution, N is the number of genotyped MET500 breast cancer cohorts, n is the number of MET500 breast cancer cohorts in which the gene is mutated and $\hat{q}$ is the mutation frequency of the gene in TCGA dataset.

Similarly, we computed the left-tailed p-value to test whether a gene has significantly lower mutation frequency in metastatic breast cancer samples:

$$p_2 = 1 - p_1 \tag{2}$$

To control false discovery rate, we applied the Benjamini-Hochberg procedure on the left-tailed and right-tailed p-values respectively [23].

Gene expression correlation analysis

To perform gene expression correlation analysis, we first rank-transformed the gene RPKM values (or probeset intensity values) of each CCLE cell line and then ranked all the genes (or probesets) according to their rank variation across CCLE cell lines. The 1000 most-varied genes (or probesets) were then selected and used to compute the spearman rank correlation between cell lines and metastatic breast cancer samples.

PAM50 sub-typing and t-SNE visualization

The genefu package was used to determine breast cancer subtype [24, 25]. To visualize tumor samples with t-SNE, we first computed the pair-wise distance between every two samples as 1 minus the spearman rank correlation (across PAM50 genes) and then applied the function Rtsne to perform 2D dimensional reduction [26].

Pubmed search

The number of PubMed abstracts or full texts mentioning a CCLE breast cancer cell line was determined using the PubMed Search feature on May 10, 2018 (https://www.ncbi.nlm.nih.gov/pubmed/). For each cell line, we searched with a keyword "[cell line name] metastasis". We repeated this step for the terms "metastatic", "breast cancer", and "metastatic breast". These searches returned highly correlated results, so we used the search terms which returned the most results: "[cell line name] metastasis".

Identification of differentially expressed genes between cell lines and metastatic breast cancer samples

DESeq2 was used to identify differentially expressed genes and DAVID bioinformatics sever was used to perform Gene Ontology (GO) enrichment analysis [27, 28]. The 50 hallmark gene sets were downloaded from MSigDB (http://software.broadinstitute.org/gsea/msigdb/) and the R package GSVA was used to perform ssGSEA analysis [29–32]. To identify gene sets which have different activity between cell lines and metastatic breast cancer samples, we used the Wilcoxon rank test to compute a p-value for each of the 50 gene sets and then applied Benjamini-Hochberg procedure to select significant gene sets (FDR ≤ 0.01).

Software tools and statistical methods

All of the analysis was conducted with R and the code is freely available at https://github.com/Bin-Chen-Lab/MetaBreaCellLine. The ggplot2 and ComplexHeatmap packages were used for data visualization [33, 34]. The tumor purity was estimated using ESTIMATE [35]. CNTools was used to map the segmented CNV data to genes [36]. If not specified, the Wilcoxon rank test was used to compute p-value in hypothesis testing.

Results

Comparison of genetic profiles between metastatic breast cancer and cell lines

We first compared the gene somatic mutation profile between MET500 breast cancer samples and breast cancer cell lines. Whole-exome sequencing was performed for

MET500 samples, while hybrid capture sequencing was performed for cell lines. We thus only focused on the 1630 genes genotyped in both studies. We are particularly interested in two types of genes that may play important roles in breast cancer metastasis: genes that are highly mutated in metastatic breast cancer, and genes that are differentially mutated between metastatic and primary breast cancers.

Consistent with previous research, we identified a long-tailed mutation spectrum of the 1630 genes in MET500 breast cancer samples (Fig S1a). There were 69 highly mutated genes whose mutation frequency is higher than 0.05 and the five most-altered genes were TP53 (0.67), PIK3CA (0.35), TTN (0.29), OBSCN (0.19), and ESR1 (0.14). We applied a statistical method to identify genes which have significantly different mutation frequency between MET500 and TCGA and 19 genes passed the criteria $FDR \leq 0.001$. The top five most significant genes were ESR1, TNK2, OBSN, CAMKK2, and CLK1 (Fig S1b and Table 1). Interestingly, all of these 19 differentially-mutated genes had higher mutation frequency in MET500 breast cancer samples, which is consistent with previous study showing that metastatic cancer has increased mutation burden compared to primary cancer [13]. 68% of them were also among the 69 highly mutated genes mentioned above. After merging the two gene lists, 75 unique genes remained (Fig1 and Table S1). The median mutation frequency of the 75 genes across breast cancer cell lines is 0.07 and only 9% of them (PRKDC, MAP3K1, TTN, ADGRG4, TP53, FN1, and AKAP9) are mutated in at least 50% of cell lines, suggesting that the majority of these gene mutations could be recapitulated by only a few cell lines. In accordance with this finding, the median number of mutated genes of the 57 cell lines is 10, with CAL51, MDAMB453, UACC812, CAL148, and HCC1569 being the five most-mutated cell lines. Additionally, nine out of the 75 genes (ESR1, GNAS, PIKFYVE, FFAR2, RNF213, MYBL2, KAT6A, MAP4K4, and FMO4) are not mutated in any cell lines. Notably, ESR1 has been identified as a driver gene of cancer metastasis and associated mutations could cause endocrine resistance of metastatic cancer cells [37, 38], but none of the cell lines could be used to accurately model it .

We next asked whether there were any genes which were specifically hypermutated in breast cancer cell lines. To address this question, we examined the mutation spectrum of the 32 genes that are mutated in at least 50% of the breast cancer cell lines. Surprisingly, 25 of them (78.1%) have low mutation frequency (< 0.05) in MET500 breast cancer samples. Further analysis of somatic mutation profiles of the 25 genes in TCGA breast cancer samples confirmed their hypermutation was specific to breast cancer cell lines (Fig S1c).

Besides the somatic mutation spectrum, we also compared CNV profiles between MET500 breast cancer samples and breast cancer cell lines. We observed a very strong correlation of median CNV values across the 1630 commonly genotyped genes (spearman rank correlation = 0.81, Fig 1b). Surprisingly, we noticed that the gain-of-copy-number events in cell lines appeared to resemble metastatic breast cancer while loss-of-copy-number events did not. As shown in Fig S1d, for genes that show copy-number-loss in breast cancer cell lines, their median CNV values across breast cancer cell lines are significantly lower than those from MET500; however, no significant difference was detected in genes with copy-number-gain.

Out of the 57 breast cancer cell lines, 24 were derived from metastatic sites (Table S2). We further divided the cell lines into two groups (according to whether derived from metastatic sites or not). Then, we compared the CNV profiles of each group with MET500 breast cancer samples. We found cell lines derived from metastatic sites more closely resembled the CNV status of genes with high copy-number-gain ($CNV \geq 0.4$) in MET500 breast cancer samples (Fig 1c, 1d, and Fig S1e), which is expected and consistent with the results of additional expression analysis (see Section 3 for more details).

Fig 1. Comparison of the genetic profile between metastatic breast cancer samples and breast cancer cell lines. (a) Somatic mutation profile of the 75 genes across MET500 breast cancer samples and breast cancer cell lines. The top-side color-bar indicates data source (MET500 or CCLE) and the right-side color-bar indicates mutation frequency of genes. (b) Comparison of CNV profiles between MET500 breast cancer samples and breast cancer cell lines with 1630 commonly genotyped genes. The x-axis represents the median CNV value of one gene across 57 breast cancer cell lines and y-axis represents the median CNV value of one gene across 53 MET500 breast cancer samples. (c) Comparison of CNV profile between MET500 breast cancer samples and CCLE breast cancer cell lines derived from primary site. The x-axis represents the median CNV value across 33 breast cancer cell lines derived from primary site and y-axis represents the median CNV value across 53 MET500 breast cancer samples. Genes with high CNV value in MET500 breast cancer samples are red. (d) Comparison of CNV profile between MET500 breast cancer samples and breast cancer cell lines derived from metastatic sites. The x-axis represents the median CNV value across 24 breast cancer cell lines derived from metastatic sites and y-axis represents the median CNV value across 53 MET500 breast cancer samples. Genes with high CNV value in MET500 breast cancer samples are red.

Table 1. Differentially mutated genes between MET500 and TCGA

genename	TCGA mutation frequecny	MET500 mutation frequency	FDR
$ESR1$	0.008	0.139	2.76E-08
$TNK2$	0.001	0.069	6.70E-7
$OBSCN$	0.031	0.194	2.46E-06
$CAMKK2$	0.004	0.083	7.93E-06
$CLK1$	0	0.056	1.80E-05
$FN1$	0.009	0.097	1.12E-04
DST	0.020	0.125	3.30E-04
$GNAS$	0.008	0.083	3.30E-04
$MLLT3$	0.003	0.056	3.30E-04
$CDKN2A$	0	0.042	3.30E-04
$NGFR$	0	0.042	3.30E-04
$NUP133$	0	0.042	3.30E-04
$RBPJ$	0	0.042	3.30E-04
$FFAR2$	0	0.042	3.30E-04
$MCL1$	0	0.042	3.30E-04
$TP53$	0.421	0.667	7.98E-04
TEK	0.007	0.069	9.10E-04
FER	0.004	0.056	9.90E-04
$MAP4K4$	0.004	0.056	9.90E-04

Correlating breast cancer cell lines with metastatic breast cancer samples using transcriptomic data

Gene expression correlation analysis is proven to be an effective approach to evaluate the suitability of cell lines for research purpose [7–9]. Therefore, we ranked all 1019 CCLE cell lines according to their median spearman rank correlation with MET500 breast cancer samples. The 20 most-correlated cell lines were all breast cancer cell lines (Fig 2a), illustrating the potential of breast cancer cell lines to resemble the transcriptomic profile of metastatic breast cancer. MDAMB415 and HMC18 are the two cell lines that have highest and lowest correlation respectively.

Since MET500 breast cancer samples were derived from multiple biopsy sites, we asked whether the cell lines resembling the transcriptome of metastatic breast cancer from different biopsy sites were identical. We were only able to consider liver and lymph node due to the paucity of enough samples from other biopsy sites in the MET500 dataset. We performed biopsy-site-specific gene expression correlation analysis (i.e., correlating breast cancer lines with samples derived from a specific biopsy site) and found that the cell line rankings obtained from liver and lymph node were highly correlated (Fig 2b, spearman rank correlation = 0.97), with MDAMB415 being the most correlated cell line for both biopsy sites. In addition, we detected no significant difference of the correlations with MDAMB415 cell line between different biopsy sites (Fig S2a).

Given the genomic heterogeneity of breast cancer, we further asked whether the cell lines resembling the transcriptome of metastatic cancer of different subtypes were identical. To address this question, we first determined the PAM50 subtype of MET500 breast cancer samples with R package genefu. Since genefu was initially developed with primary breast cancer data, we further applied the machine learning method t-SNE on expression data of PAM50 genes and confirmed the PAM50 genes could be used to classify metastatic breast cancer samples. As shown in Fig 2c, Basal-like samples were clustered together and separated with other subtypes, which is in accordance with previous research [39, 42]. Additionally, the majority of LuminalA/LuminalB/Her2-enriched/Normal-like samples were mixed together except two skin-derived samples (HER2-enriched samples seemed to be separated with LuminalA/LuminalB samples but the boundary was not clear). We confirmed the finding by performing the same analysis on a combined dataset which contains both MET500 and TCGA breast cancer samples (Fig S2b). We next performed subtype-specific gene expression correlation analysis (i.e., correlating breast cancer cell lines with samples of a specific subtype) and found the ranking of breast cancer cell lines obtained from LuminalA/LuminalB/Her2-enriched subtypes were highly correlated with each other (spearman rank correlation values were 0.96, 0.97, and 0.96 respectively), but they all showed relatively lower correlation with the Basal-like subtype (Fig 2d).

To confirm the robustness of the results, we searched the GEO database and assembled a microarray dataset containing the expression value of another 117 metastatic breast cancer samples, and repeated the above analysis. As expected, the results obtained from two different platforms were highly consistent with each other. First, there was a large overlap of the top-ranked cell lines. Out of the 10 cell lines that were most correlated with the 117 metastatic breast cancer samples, six of them were within top 10 cell lines that were most correlated with MET500 breast cancer samples. Second, cell line ranking results between liver and lymph node were highly correlated (Fig S3, spearman rank correlation = 0.95). Third, cell line ranking results obtained from Basal-like samples still showed relatively lower correlations with other subtypes (Fig S4).

We also noticed that the expression correlation analysis results derived from bone showed lower correlation with other tissues. To exclude the possibility that this was caused by tumor purity issues, we applied ESTIMATE on the microarray data and found the tumor purity of bone-derived metastatic breast cancer samples was not significantly lower than that of liver, lymph node and lung (Fig S5). Our results may not be too surprising given the fact that bone provides a very unique microenvironment including enriched expression of osteolytic genes [40]; however, this result needs to be confirmed in the future as more data becomes available.

Fig 2. Gene expression correlation analysis between MET500 breast cancer samples and breast cancer cell lines. (a) Ranking 1019 CCLE cell lines according to their median expression correlation with MET500 breast cancer samples. Each dot represents a cell line with breast cancer cell lines marked as red. (b) Cell line ranking results of liver and lymph node are highly correlated. Each dot represents a breast cancer cell line. (c) t-SNE plot of MET500 breast cancer samples. Biopsy-sites are labeled by color. (d) Pair-wise comparison of cell line ranking results among four breast cancer subtypes. The upper-triangle part shows the pair-wise spearman rank correlation. The lower-triangle part shows the pair-wise scatter plot, with each dot representing a breast cancer cell line.

Suitable cell lines for metastatic breast cancer research

We attempted to identify suitable cell line models for metastatic breast cancer based on the results of subtype-specific gene expression correlation analysis. Given a subtype, we noticed that for a random cell line, the median expression correlation (with MET500 breast cancer samples of that subtype) was normally distributed (Fig S6). Based on that, we fit a normal distribution using the median expression correlation values of all non-breast-cancer cell lines and then assigned each of the 57 breast cancer cells lines a right-tailed p-value. We identified 20, 28, and 19 significant cell lines as suitable models for LuminalA, LuminalB, and Her2-enriched subtypes, respectively ($FDR \leq 0.01$, see Table S3). Notably, most of these suitable cell lines were derived from metastatic sites and 18 of them were shared by the three subtypes. Surprisingly, no cell line passed the criteria FDR ≤ 0.01 for the Basal-like subtype. We further examined whether this was due to the limited number of Basal-like samples. However, the number of LuminalA samples was even less than that of Basal-like samples.

We next examined the popularity of the 57 breast cancer cell lines. MCF7 is most commonly used in metastatic breast cancer research (n=2299 Pubmed citations). Although we found it was a suitable cell line for LuminalB subtype, its correlation with MET500 LuminalB samples was lower than that of BT483, the most significant cell line for LuminalB subtype (Fig S7a). Following MCF7 in mentions is MDAMB231 (n=2118 Pubmed citations); however, we found that it was not an suitable cell line to use for every subtype based on our results. The third most popular cell line T47D (n=204 Pubmed citations) was a suitable cell line for both LuminalA and Her2-enriched subtype. T47D did not show significantly lower correlation with LuminalA samples than EFM192A, the most correlated cell line for LuminalA subtype (Fig S7b); however, it was significantly less correlated with Her2-enriched subtype than EFM192A, the most correlated cell line for Her2-enriched subtype (Fig S7c). Additional subtype-specific gene expression correlation analysis in the microarray dataset further confirmed our results (Fig S8).

While the triple negative cell line MDAMB231 is one of the most frequently used cell lines in metastatic breast cancer research, it might not be the most suitable cell line to model metastasis biology. We ranked all of the 1019 CCLE cell lines according to their median expression correlation with MET500 Basal-like breast cancer samples and the rank of MDAMB231 was 583 (Fig 3a). It showed significantly lower correlation with MET500 Basal-like breast cancer samples than HCC70, the most correlated cell line. Similar patterns were observed with CNV data (Fig 3b). We also examined how MDAMB231 recapitulated the somatic mutation spectrum of Basal-like breast cancer samples and found only three of the 25 highly mutated genes (mutation frequency $\geq$0.1 in Basal-like MET500 breast cancer samples) were mutated in MDAMB231. Since CCLE data for MDAMB231 was generated *in vitro*, we obtained another independent dataset which profiled the gene expression of MDAMB231 cell lines derived from lung metastasis *in vivo* [41] in order to confirm our finding. We found, however, that even

these *in vivo* MDAMB231 cell lines did not most closely resemble the transcriptome of lung metastasis breast cancer samples. The cell line which showed highest correlation (with lung metastasis breast cancer samples) is the CCLE cell line EFM192A (Fig 3d). Our analysis indicates that although MDAMB231 presents many favorable properties for metastatic breast cancer research, its genomic profile is substantially different from metastatic tissue samples.

Fig 3. The widely used cell line MDAMB231 may not be the most suitable model for metastatic breast cancer research. (a) MDAMB231 shows poor expression correlation with MET500 breast cancer samples of Basal-like subtype. The left panel shows the ranking of all 1019 CCLE cell lines according to their median expression correlation with MET500 breast cancer samples of Basal-like subtype. The top-left scatter plot shows the expression of the most varied 1000 genes with x-axis represents expression value in MDAMB231 and y-axis represents median expression value across MET500 Basal-like breast cancer samples. The boxplot on the right panel shows the distribution of correlation values (with MET500 breast cancer samples of Basal-like subtype) for MDAMB231 and HCC70. (b) MDAMB231 shows poor CNV correlation with MET500 breast cancer samples of Basal-like subtype. The left panel shows the ranking of all 1019 CCLE cell lines according to their median CNV correlation with MET500 breast cancer samples of Basal-like subtype; the boxplot on the right panel shows the distribution of correlation values (with MET500 breast cancer samples of Basal-like subtype) for MDAMB231 and HCC70. (c) Somatic mutation profile of the 25 highly mutated genes across MDAMB231 and MET500 breast cancer samples of Basal-like subtype. (d) Boxplot of expression correlation between CCLE breast cancer cell lines, lung-metastasis-derived MDAMB231 (colored by red) and lung-derived metastatic breast cancer samples.

Differential gene expression analysis between metastatic breast cancer and cell lines

The gene expression correlation analysis has shown that many cell lines could resemble metastatic breast cancer; however, they are still different in many aspects [3, 4]. To characterize such differences, we compared the gene expression profile of MET500 breast cancer samples with breast cancer cell lines and identified 3044 differentially expressed genes (FDR $\leq$ 0.001, abs(log2FC) $\geq$ 1). We further performed GO enrichment analysis for the up-regulated and down-regulated genes respectively and listed the results in Table S4. The top five most significant enriched GO terms for up-regulated genes are extracellular matrix organization, cell adhesion, type I interferon signaling pathway, interferon-gamma-mediated signaling pathway and immune response; the top five most significant GO terms for down-regulated genes are all related to cell cycle: cell division, mitotic nuclear division, sister chromatid cohesion, DNA replication, and chromosome segregation.

We also compared the ssGSEA score of the 50 MSigDB hallmark gene sets between MET500 breast cancer samples and breast cancer cell lines (Fig 4). In total,37 gene sets were identified as showing differential activity ($FDR \leq 0.01$, Table S6). Out of them, 27 showed significantly higher activity in MET500 breast cancer samples and the remaining 10 showed significantly lower activity. The five gene sets showing largest positive effect size are EPITHELIAL_MESENCHYMAL_TRANSITION, ANGIOGENESIS, COAGULATION, INTERFERON_ALPHA_RESPONSE, and INTERFERON_GAMMA_RESPONSE. The five gene sets showing smallest negative effect size are G2/M check point and DNA repair, E2F_TARGETS, MYC_TARGETS_V2, MYC_TARGETS_V1, and SPERMATOGENESIS. Interestingly,

we noticed that some MET500 breast cancer samples derived from liver (in dashed box of Fig 4) had enriched metabolism-related gene sets (such as XENOBIOTIC_METABOLISM and BILE_ACID_METABOLISM). This suggests that liver-metastasis cancer cells may have their unique metabolic mechanism comparing to primary tumors.

It is worth noting that the ssGSEA results are highly consistent with the gene differential expression analysis. The up-regulated genes (and over-activated gene sets in MET500) reflect the large difference of microenvironment between metastatic breast cancer and cell lines; also, the down-regulated genes (and less-activated gene sets in MET500) suggest that cell lines have more active cell cycles. All of these differences should be kept in mind when using cell lines in translational research.

Fig 4. Comparison of ssGSEA score of the 50 MSigDB hallmark gene sets. The gene sets are re-ordered according to p-value computed by the Wilcoxon rank test. In MET500 dataset, there are 37 LuminalA/LuminalB/Her2-enriched breast cancer samples. For reference, we randomly picked out equal number of TCGA breast cancer samples and included their ssGSEA scores in the figure.

Discussion

In cancer research, cell lines have been traditionally used to test drug candidates and study disease mechanism. Our comprehensive analysis has both raised doubt and shed light on the suitability of breast cancer cell lines as models for metastatic breast cancer research.

Somatic mutation profile analysis indicated that breast cancer cell lines poorly recaptured the mutation patterns of metastatic breast cancer samples. Most of the highly-mutated genes (or differentially-mutated genes between metastatic and primary lesions) were only mutated in a limited number of cell lines. In addition, there were 25 genes showing cell-line-specific hypermutation, which may be due to culture effects. Remarkably, the CNV profiles between breast cancer cell lines and metastatic breast cancer samples were much more consistent. We also performed a gene expression correlation analysis to explore whether breast cancer cell lines could resemble the transcriptome of metastatic breast cancer samples. The results of biopsy-site-specific analysis suggested that for liver and lymph node derived metastatic breast cancer samples, the biospy site did not play a role in determining the cell lines which closely resembled their transcriptome and such conclusion was validated in analysis of two independent datasets. It has been shown that breast cancer is a heterogeneous disease with multiple subtypes. We found that the PAM50 subtype were maintained in metastatic breast cancer samples regardless of the tissues it metastasize to and this corroborates with the results from a recent study [42]. Through a subtype-specific analysis, we found that the cell lines that most closely resembled the transcriptome of LuminalA/LuminalB/Her2-enriched subtypes were highly overlapped. Surprisingly, none of the currently established cell lines adequately resemble Basal-like metastatic breast cancer samples. Moreover, we found that the two most commonly used cell lines, MCF7 and MDAMB231 (together accounting for more than 80% of total PubMed publications mentioning metastatic breast cancer), were not the best choice for metastatic breast cancer research in terms of transcriptomic similarity. Specifically, there is dramatic difference between Basal-like metastatic breast cancer samples and MDAMB231 (the most commonly used triple negative cell line), which was demonstrated by both *in vitro* and *in vivo* data. Note that although some cell lines closely resemble tissue samples of LuminalA/LuminalB/Her2-enriched subtypes, it does

not mean they could be directly employed to study cancer metastasis as many other criteria are needed for the assessment. Nevertheless, this analysis does suggest that we are in urgent need of new Basal-like cell lines which more closely resemble the biology of Basal-like metastatic breast cancer samples.

The results of our gene expression correlation analysis also raises a new question: when picking out cell lines to test drugs targeting breast cancer metastasis, which factors should be taken into consideration? According to our analysis, it appears that for lymph node and liver metastasis, the subtype information is sufficient since the biopsy-site-specific gene expression correlation analysis results were highly concordant with each other. However, we found that the results computed with bone metastases showed low correlation with other tissues. This implies that even for the same subtype, a cell line that is appropriate to model metastasis of other sites may not be appropriate for bone metastasis study.

Even though many breast cancer cell lines resemble the transcriptome of metastatic breast cancer samples, a large number of genes were identified as differentially expressed between them. Some of these genes relate to immune response, possibly reflecting the large difference between tumor microenvironment and the cell culture. In addition, our ssGSEA analysis on the 50 hallmark gene sets suggested that there is systematical difference of important pathway activities.

In summary, by leveraging publicly available gnomic data and machine learning algorithms, we comprehensively evaluate the suitability of breast cancer cell lines as models for metastatic breast cancer. Our study also describes a blueprint which can be easily extended to other cancer types and more advanced model systems, such as organoids [43]. Although there are concerns about data quality and discrepancies between different studies/platforms, our large-scale analysis and cross-platform validation hopefully addresses these concerns and demonstrates the power of leveraging open data and machine learning algorithms to gain biological insights of cancer metastasis. As more data becomes available, we can start building an ad-hoc mapping algorithm linking metastasis samples, cell lines and other models. Inputs into this algorithm would be the characteristics of metastatic cancer samples (subtype, biopsy site, or even age, race, etc) as well as the specific scientific question of interest and the output would be a list of appropriate models. We hope that the recommendations in this study may facilitate improved precision in selecting relevant and suitable cell lines for modeling in metastatic breast cancer research, which may accelerate the translational research.

Supporting information

S1 Fig. (a) Long-tailed gene mutation spectrum in MET500 breast cancer samples. (b) Volcano plot of gene differential mutation frequency analysis. (c) Visualization of log10-transformed mutation frequency of the 25 genes that are specifically hypermutated in CCLE breast cancer cell lines. (d) Boxplot of median CNV of grouped genes (according to whether showing gain or loss of copy number in CCLE breast cancer cell lines) in MET500 breast cancer samples and CCLE breast cancer cell lines. (e) CCLE breast cancer cell lines derived from metastatic sites more closely resemble the CNV status of genes with high copy-number-gain in MET500 dataset. Left: absolute value of median CNV difference between MET500 breast cancer samples and CCLE breast cancer cell lines derived from primary sites; right: absolute value of median CNV difference between MET500 breast cancer samples and CCLE breast cancer cell lines derived from metastatic sites.

S2 Fig. (a) Metastatic breast cancer samples derived from liver and lymph node do not show significantly different expression correlation with MDAMB415. (b) t-SNE plot of TCGA and MET500 breast cancer samples.

S3 Fig. Pair-wise comparison of the results of biopsy-site-specific gene expression correlation analysis (microarray dataset).

S4 Fig. Pair-wise comparison of the results of subtype-specific gene expression correlation analysis (microarray dataset).

S5 Fig. Boxplot of tumor purity of the five biopsy sites (microarray data).

S6 Fig. Normal qqplot to confirm the median expression correlation between cell lines and MET500 breast cancer samples (of a specific subtype) approximately follows normal distribution. (a) LuminalA subtype. (b) LuminalB subtype. (c) Her2-enriched subtype. (d) Basal-like subtype.

S7 Fig. (a) MCF7 cell line shows significant lower correlation with MET500 LuminalB samples than BT483. (b) T47D cell line does not show significant lower correlation with MET500 LuminalA breast cancer samples than MDAMB415. (c) T47D cell line shows significant lower correlation with MET500 Her2-enriched breast cancer samples than EFM192A.

S8 Fig. Subtype-specific gene expression correlation analysis results between MET500 and microarray dataset are highly correlated. (a) LuminalA subtype. (b) LuminalB subtype. (c) Her2-enriched subtype. (d) Basal-like subtype.

S1 Table. Mutation frequency of the 75 highly (or differentially) mutated genes in CCLE, TCGA, and MET500 dataset.

S2 Table. Characteristic of the 57 CCLE breast cancer cell lines.

S3 Table. Suitable CCLE breast cancer cell lines for LuminalA, LuminalB, and Her2-enriched subtypes.

S4 Table. GO enrichment results of differentially expressed genes between CCLE breast cancer cell lines and MET500 breast cancer samples.

S5 Table. Results of differential activity analysis between MET500 breast cancer samples and CCLE breast cancer cell lines (for the 50 MSigDB hallmark gene sets).

Acknowledgments

The research is supported by R21 TR001743 and K01 ES028047 and the MSU Global Impact Initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of sponsors.

References

1. Francesco Iorio,Theo A Knijnenburg,Daniel J Vis et al. A Landscape of Pharmacogenomic Interactions in Cancer. Cell.2016;166(3):740-754.

2. Jennifer L. Wilding and Walter F. Bodmer. Cancer Cell Lines for Drug Discovery and Development. Cancer Res.2014;74(9):2377-84.

3. Adam Ertel1, Arun Verghese, Stephen W Byers, Michael Ochs and Aydin Tozeren. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. Mol Cancer. 2006;5(1):55.

4. Jean-Pierre Gillet,a Anna Maria Calcagno,a Sudhir Varma et al. Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. PNAS.2011;108(46):18708-13.

5. The Cancer Genome Atlas Research Network, John N Weinstein, Eric A Collisson et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet.2013;45(10):1113-20.

6. Jordi Barretina,Giordano Caponigro,Nicolas Stransky et al. The Cancer Cell Line Encyclopedia enables predictive modeling of anticancer drug sensitivity. Nature.2012;483(7391):603-7.

7. Silvia Domcke, Rileen Sinha1, Douglas A. Levine et al. Evaluating cell lines as tumor models by comparison of genomic profiles. Nat Commun.2013;4:2126.

8. Bin Chen,Marina Sirota, Hua Fan-Minogue, Dexter Hadley, Atul J Butte. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. BMC Medical Genomics.2015;8 Suppl 2:S5.

9. Rickard Sandberg and Ingemar Ernberg. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). PNAS.2005;102(6):2052-7

10. Guanglong Jiang,Shijun Zhang,Aida Yazdanparast et al. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics.2016;17;Suppl 7:525.

11. Arthur W. Lambert,Diwakar R. Pattabiraman,and Robert A. Weinberg. Emerging Biological Principles of Metastasis. Cell.2017;168(4):670-691.

12. Patrick Mehlen,Alain Puisieux. Metastasis: a question of life or death. Nat Rev Cancer.2006;6(6):449-58.

13. Robinson DR,Wu YM,Lonigro RJ et al . Integrative clinical genomics of metastatic cancer. Nature.2017;548(7667):297-303.

14. Bo Li,Victor Ruotti,Ron M. Stewart et al. RNA-Seq gene expression estimation with read mapping uncertainty. Bioinformatics.2010;26(4):493-500.

15. Bo Li and Colin N Dewey. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics.2011;12:323.

16. Gao J, Aksoy BA, Dogrusoz U et al. Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. Sci Signal.2013;6(269):pl1.

17. Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz et al. The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data. Cancer Discovery.2012;2(5):401-4.

18. John Vivian, Arjun Arkal Rao, Frank Austin Nothaft et al. Toil enables reproducible, open source, big biomedical data analyses. Nat Biotechnol.2017;35(4):314-316.

19. Landemaine T1, Jackson A, Bellahcène A et al. A six-gene signature predicting breast cancer lung metastasis. Cancer Res.2008;68(15):6092-9.

20. Xu J, Acharya S, Sahin O et al. 14-3-3 ξ turns TGF-$\beta's$ function from tumor suppressor to metastasis promoter in breast cancer by contextual changes of Smad partners from p53 to Gli. Cancer Cell.2015;27(2):177-92.

21. Zhang XH, Wang Q, Gerald W et al. Latent bone metastasis in breast cancer tied to Src-dependent survival signals. Cancer Cell.2009;16(1):67-78.

22. Theodoros Foukakisa,John L€ovrota, Patricia Sandqvistb et al. Gene expression profiling of sequential metastatic biopsies for biomarker discovery in breast cancer. Mol Oncol.2015;9(7):1384-91.

23. Benjamini Yoav, Hochberg, Yosef. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal Statistical Society.1995;57(1):289-300.

24. Joel S. Parker, Michael Mullins, Maggie C.U. Cheang et al. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. J Clin Oncol.2009;27(8):1160-7.

25. Gendoo DM, Ratanasirigulchai N, Schröder MS et al. Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. Bioinformatics. 2016;32(7):1097-9.

26. Laurens van der Maaten. Visualizing Data using t-SNE. Journal of Machine Learning Research

27. Michael I Love, Wolfgang Huber, and Simon Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol.2014;15(12):550.

28. Da Wei Huang, Brad T Sherman, Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc.2009;4(1):44-57.

29. Aravind Subramanian, Pablo Tamayo, Vamsi K. Mootha et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. PNAS.2005;102(43):15545-50.

30. Arthur Liberzon Aravind Subramanian Reid Pinchback et al. Molecular signatures database (MSigDB) 3.0. Bioinformatics.2011;27(12):1739-40.

31. Arthur Liberzon,Chet Birger,Helga Thorvaldsdóttir et al. The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst.2015;1(6):417-425.

32. Sonja Hanzelmann,Robert Castelo1,and Justin Guinney. GSVA: gene set variation analysis for microarray and RNA-Seq data. BMC Bioinformatics.2013 16;14:7.

33. The ggplot2 package is available at https://cran.r-project.org/web/packages/ggplot2/index.html.

34. The ComplexHeatmap package is available at https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html.

35. Kosuke Yoshihara, Maria Shahmoradgoli, Emmanuel Martínez et al. Inferring tumour purity and stromal and immune cell admixture from expression data. Nat Commun.2013;4:2612.

36. Jianhua Zhang. CNTools:Convert segment data into a region by sample matrix to allow for other high level computational analyses.

37. Celine Lefebvre,Thomas Bachelot,Thomas Filleron et al. Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. PLoS Med.2016;13(12):e1002201.

38. Stephan Bartels, Matthias Christgen, Angelina Luft et al. Estrogen receptor (ESR1) mutation in bone metastases from breast cancer. Mod Pathol.2018;31(1):56-61.

39. Clifford A. Hudisa and Luca Gianni. Triple-Negative Breast Cancer: An Unmet Medical Need. Oncologist.2011;16 Suppl 1:1-11.

40. Kang Y, Siegel PM, Shu W et al. A multigenic program mediating breast cancer metastasis to bone. Cancer Cell.2003;3(6):537-49.

41. Andy J. Minn, Gaorav P. Gupta1, Peter M. Siegel1 et al. Genes that mediate breast cancer metastasis to lung. Nature.2005;436(7050):518-24.

42. Juan M. Cejalvo,Eduardo Martínez de Dueñas,Patricia Galván et al. Intrinsic Subtypes and Gene Expression Profiles in Primary and Metastatic Breast Cancer. Cancer Res.2017;77(9):2213-2221.

43. Fleur Weeber,Salo N. Ooft, Krijn K. Dijkstra, Emile E. Voest Tumor Organoids as a Pre-clinical Cancer Model for Drug Discovery. Cell Chem Biol.2017;24(9):1092-1100.