Evaluating cell lines and organoids as models for metastatic cancer through integrative analysis of open genomic data

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

Affiliations:

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. Bakar Computational Health Sciences Institute, University of California San Francisco, San Francisco, CA, USA

4. Department of Physiology, Michigan State University, East Lansing, MI, USA

*To whom correspondence should be addressed: 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 test drug candidates, yet it is still unknown to what extent do they adequately resemble the disease in patients. The recent accumulation of large-scale genomic data in cell lines, organoids, mouse models, and patient tissue samples provides an unprecedented opportunity to evaluate the suitability of cancer cell lines as models for metastatic cancer research. Using metastatic breast cancer as a case study, we systematically evaluate their suitability as models for metastatic cancer research. The comprehensive comparison of the genomic profiles of 57 breast cancer cell lines with those of metastatic breast cancer samples revealed substantial genomic differences. We also identified cell lines that more closely resemble different subtypes of metastatic breast cancer. Surprisingly, a combined 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. In addition to cell lines, we analyzed the RNA-Seq data of patient-derived organoids and found organoids outperformed cell lines in resembling the transcriptome of metastatic breast cancer samples. Finally, we characterized systematic difference between metastatic breast cancer and the *in vitro* models. Our work both provides a guide of cell line selection in metastasis-related study and sheds light on the large potential of organoids in translational 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 promising preclinical evidence 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 transcriptome profiles of more than 10,000 human tissue samples across 32 tumor types[5]. The Cancer Cell Line Encyclopedia (CCLE) characterized both genetic and transcriptome 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[9].

Cancer metastasis is the most common cause of cancer-related death, thus there is an urgent need of new drugs for treating cancer metastasis[10,11]. Previous cell line evaluation analysis was 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)[12]. 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 focused on breast cancer and comprehensively compared multiple types of molecular features between breast cancer cell lines and metastatic breast cancer samples (Fig 1). Based on our analyses, we identified cell lines that are suitable for modeling metastatic breast cancer samples of different subtypes. In addition, we also evaluated patient-derived organoids and showed their superior potential in pre-clinical trial. Our work provides useful guidance for choosing cell lines in metastasis-related translational research and could be easily extended to other cancer types.

Results

Comparison of genetic profiles between metastatic breast cancer and cell lines

We first compared somatic mutation profiles between MET500 breast cancer samples and CCLE breast cancer cell lines. Whole-exome sequencing was performed for MET500 samples, while hybrid capture sequencing was performed for CCLE cell lines. We thus only focused on the 1630 genes genotyped in both studies. We were 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 and 69 of them were highly mutated (mutation frequency > 0.05, Fig S1a). The five most-altered genes were TP53 (0.67), PIK3CA (0.35), TTN (0.29), OBSCN (0.19), and ESR1 (0.14). We identified 19 differentially-mutated genes between MET500 and TCGA samples (FDR < 0.001) and the five most significant genes were ESR1, TNK2, OBSN, CAMKK2, and CLK1 (Fig S1b). Interestingly, all of these 19 differentially-mutated genes had higher mutation frequency in

MET500 than TCGA, which is consistent with previous study showing that metastatic cancer has increased mutation burden compared to primary cancer[12]. 68% of them were also among the 69 highly mutated genes mentioned above. After merging the two gene lists, 75 unique genes remained (Fig 2a and Table S1). The median mutation frequency of the 75 genes across CCLE breast cancer cell lines was 0.07 and only 9% of them (PRKDC, MAP3K1, TTN, ADGRG4, TP53, FN1, and AKAP9) were mutated in at least 50% of cell lines, suggesting that majority of these gene mutations could be recapitulated by only a few cell lines. Surprisingly, nine of the 75 genes (ESR1, GNAS, PIKFYVE, FFAR2, RNF213, MYBL2, KAT6A, MAP4K4, and FMO4) were not mutated in any cell line. Notably, ESR1 has been identified as a driver gene of cancer metastasis and associated mutations can cause endocrine resistance of metastatic breast cancer cells[13,14], yet none of the cell lines could be used to accurately model it.

We next asked whether there were genes specifically hypermutated in breast cancer cell lines. To address this question, we examined the mutation spectrum of the 32 genes that were mutated in at least 50% of the breast cancer cell lines. Surprisingly, 25 of them had 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 hypermutations were specific to breast cancer cell lines (Fig S1c).

In addition to somatic mutation spectrum, we also compared copy number variation (CNV) profiles between MET500 breast cancer samples and CCLE breast cancer cell lines. We observed a high correlation of median-CNV values across the 1630 commonly genotyped genes (spearman rank correlation = 0.81, Fig 2b). However, we also 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. For the 711 genes showing copy-number-loss in CCLE breast cancer cell lines (median-CNV < 0), their cell line derived median-CNV values were significantly lower than that from MET500 breast cancer samples;

however, no significant difference was detected in the 919 genes with copy-number-gain (Fig S1d). Out of the 57 CCLE breast cancer cell lines, 23 were derived from metastatic sites (Table S2). We further divided the cell lines into two groups (according to whether they were derived from metastatic sites or not) and then compared the CNV profile of each group with MET500 breast cancer samples. We found cell lines derived from metastatic sites more closely resembled the CNV status of the 109 genes with high copy-number-gain (median-CNV >= 0.4) in MET500 breast cancer samples (Fig 2c, 2d, and Fig S1e).

Correlating CCLE breast cancer cell lines with metastatic breast cancer samples using transcriptome data

Transcriptome correlation analysis (TC analysis) is proven to be an effective approach to evaluate the suitability of cell lines for research purpose [7,8,15]. Therefore, we performed TC analysis and ranked all 1019 CCLE cell lines according to their transcriptome-similarly with MET500 breast cancer samples (see Methods). The top 20 cell lines were all breast cancer cell lines, suggesting metastatic breast cancer cells retain the information of their originated tissue and cell lines have the potential to resemble the transcriptome of them (Fig 3a). MDAMB415 and HMC18 were the two breast cancer cell lines that had highest and lowest transcriptome-similarity, respectively.

We next assessed whether cell lines resembling the transcriptome of samples from different metastatic sites were identical. We were only able to consider liver and lymph node (the two sites which have at least nine samples) due to the lack of enough samples from other sites in the MET500 dataset. For each of them, we performed metastatic-site-specific TC analysis (i.e., compute transcriptome-similarity of cell lines with samples derived from a specific metastatic site) and found the results were highly correlated (Fig 3b, spearman rank correlation = 0.97) with MDAMB415 being the most correlated cell line for both sites. In addition, we detected no significant difference of expression correlation (with MDAMB415) between the two sites (Fig S2a).

Given the genomic heterogeneity of breast cancer, we further asked whether cell lines resembling the transcriptome of metastatic breast 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 then applied t-SNE to visualize them (Fig 3c). We found Basal-like samples were clustered together and separate to other subtypes; 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). These results suggested that subtype information were well maintained in metastatic breast cancer samples and additionally confirmed the feasibility of genefu for subtyping metastatic breast cancer though it was initially developed with primary breast cancer data. We further confirmed the subtyping results by performing the same analysis on a combined dataset which contains both MET500 and TCGA breast cancer samples (Fig S2b). Next, we performed subtype-specific TC analysis (i.e., compute transcriptome-similarity of cell lines with samples of a specific subtype) and found high correlation within LuminalA/LuminalB/Her2-enriched subtypes (spearman rank correlation values were 0.96, 0.97, and 0.96 respectively), in contrast to their relatively lower correlation to Basal-like subtype (Fig 3d).

To confirm the robustness of our TC analysis on MET500 dataset, we searched the GEO database and assembled a microarray dataset containing the expression value of another 106 metastatic breast cancer samples, and then repeated the above analysis. Results obtained from two the datasets were highly consistent with each other. First, there was a large overlap of the top-ranked cell lines. Out of the 10 cell lines having highest transcriptome-similarity with the 106 metastatic breast cancer samples, six of them were within the 10 cell lines having highest transcriptome-similarity with MET500 breast cancer samples. Second, both metastatic-site-specific and subtype-specific TC analysis results showed high correlations (Fig S3). Due to such high consistency, it is not surprising that we observed similar correlation trends in metastatic-site-specific (or subtype-specific) TC analysis results (Fig S4, S5). About 24% of the 106 samples in the microarray dataset were derived from bone. Remarkably, the metastatic-site-specific TC analysis result of bone showed lower correlation with other sites (Fig S4). To exclude the possibility that this was caused by tumor purity issues, we applied ESTIMATE[16] on the microarray data and found the tumor purity of bone-derived samples was not significantly lower than that of liver, lymph node, and lung (Fig S6). Our results may not be too surprising given the fact that bone provides a very unique microenvironment including enriched expression of osteolytic genes[17]; however, this result needs to be confirmed in the future as more data becomes available.

Suitable cell lines for metastatic breast cancer research

We attempted to identify suitable cell lines for different subtypes of metastatic breast cancer based on the results of subtype-specific TC analysis. Given a subtype, we noticed that for a random cell line, its transcriptome-similarity with MET500 breast cancer samples of that subtype approximately followed a normal distribution (Fig S7). Therefore, those breast cancer cell lines showing significantly higher transcriptome-similarity were considered as suitable. Driven by this finding, we first fit a normal distribution (which is used as null distribution) with the transcriptome-similarity values of all non-breast-cancer cell lines and then assigned each of the 57 breast cancer cell lines a right-tailed p-value. The most significant cell lines for LuminalA, LuminalB, Her2-enriched, and Basal-like subtypes were MDAMB415 (p-value =3.59e-05), BT483 (p-value=2.22e-07), EFM192A (p-value=0.11e-03) and HCC70 (p-value =0.40e-03), respectively. Using a criteria of FDR <= 0.01, we identified 20, 28 and 24 suitable cell lines for LuminalA, LuminalB, and Her2-enriched subtypes respectively. Notably, most of these significant cell lines were derived from metastatic sites and 18 were shared by the three subtypes. Surprisingly, no cell line passed the criterion for Basal-like subtype. We further examined whether this was due to the limited number of Basal-like MET500 breast cancer samples, but found that the number of LuminalA samples was even less than that of Basal-like samples. After we used a more loosed FDR cutoff of 0.05, we found 22 suitable cell lines for Basal-like subtype. All statistical testing results are listed in Table S3.

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, it was less correlated with MET500 LuminalB breast cancer samples than BT483 (Fig S8a). Following MCF7 in mentions is MDAMB231 (n=2118 PubMed citations); however, it was not a suitable cell line for any subtype. The third most popular cell line T47D (n=204 PubMed citations) was a suitable cell line for LuminalA and Her2-enriched subtypes. It did not show significantly lower correlation with LuminalA MET500 breast cancer samples than MDAMB415 (Fig S8b); however, compared to EFM192A, it was significantly less correlated with Her2-enriched MET500 breast cancer samples (Fig S8c).

While the triple negative cell line MDAMB231 is one of the most frequently used cell lines in metastatic breast cancer research, we found that it might not be the most suitable cell line to model metastasis biology in breast cancer. We ranked all of the 1019 CCLE cell lines according to their transcriptome-similarity with the 15 MET500 Basal-like breast cancer samples and the rank of MDAMB231 was 583 (Fig 4a). Consistent with this, MDAMB231 was significantly less correlated with MET500 Basal-like breast cancer samples than HCC70. We observed similar patterns with CNV data (Fig 4b). We also examined how MDAMB231 recapitulated the somatic mutation spectrum of Basal-like metastatic breast cancer samples and found only three of the 25 highly-mutated genes (mutation frequency >= 0.1 in Basal-like MET500 breast cancer samples) were mutated in MDAMB231 (Fig 4c). Since CCLE data for MDAMB231 was generated *in vitro*, to confirm our finding we obtained another independent microarray dataset which profiled the gene expression of MDAMB231 cell lines derived from lung metastasis *in vivo* [18]. 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 breast cancer cell line which showed highest correlation with lung metastasis breast cancer samples was EFM192A (Fig 4d). Our analysis indicates that although MDAMB231 presents many favorable properties for metastatic breast cancer research, its genomic profile is substantially different from metastatic breast cancer samples.

Recently established patient-derived organoids more closely resemble the transcriptome of metastatic breast cancer samples

Owing to the advancement of 3D culturing technology, more and more tumor patient-derived organoids have been established and widely used in translational research[19,20]. However, their suitability to model metastatic cancer has not been comprehensively evaluated with large-scale genomic data. To fill this gap, we performed additional transcriptome correlation analysis on 26 patient-derived organoids using RNA-Seq data. The aforementioned subtype-specific TC analysis showed that the Basal-like subtype had relatively lower correlation with other subtypes and we also observed similar trend in organoids (Fig 5a). We next asked whether organoids performed better than cell lines in resembling the transcriptome of metastatic breast cancer. For each of the non-Basal-like organoids, we computed its transcriptome-similarity with non-Basal-like MET500 breast cancer samples and found organoids had significantly higher transcriptome-similarity values than CCLE breast cancer cell lines (Fig 5b, left panel). The superiority of organoids was also observed in the TC analysis of Basal-like subtype (Fig 5b, right panel). The previous analysis revealed that MDAMB415, BT483 and EFM192A were the three most suitable cell lines for LuminalA, LuminalB and Her2-enriched subtypes, respectively. Interestingly, for all the three subtypes MMC01031 was the organoid showing highest transcriptome-similarity and had significantly higher correlation with MET500 samples than the corresponding most-correlated cell line. Organoid W1009 had the highest transcriptome-similarity with Basal-like MET500 breast cancer samples and the correlation values were also significantly higher than HCC70, the triple-negative cell line showing highest transcriptome-similarity with Basal-like MET500 breast cancer samples (Fig 5c).

Characterization of systematic difference between metastatic breast cancer samples and *in vitro* models

Our TC analysis has shown that *in vitro* models such as cell lines and organoids could resemble the transcriptome of metastatic breast cancer at some extent. However, they are still different in many aspects. To characterize such differences, we performed differential gene expression analysis among MET500 breast cancer samples, CCLE breast cancer cell lines and organoids (Fig S9). For non-Basal-like subtypes, 2,380 genes (2,179 up-regulated, 201 down-regulated) were identified as differentially expressed in both MET500 vs CCLE and MET500 vs organoids comparisons. For Basal-like subtype, there were 2842 common differential expressed (DE) genes (1,117 up-regulated, 261 down-regulated). After intersecting the above two common DE gene lists, we finally obtained 1,016 subtype-and-model-independent DE genes (948 up-regulated, 68 down-regulated) and then performed GO enrichment analysis. For the 948 up-regulated ones, 30 GO terms were identified as significant (FDR < 0.001) and most of them were immune-related, illustrating the large gap between culture media and tumor micro-environment (Table S4). The two terms "platelet degranulation", and "chemotaxis" were also detected as significant. Besides micro-environment, our results also implicated the difference of intrinsic characteristics between metastatic breast cancer cells and *in vitro* models. For example, the enrichment on "steroid metabolic process" suggested that neither cell lines nor organoids resemble the reprogrammed metabolism of metastatic breast cancer sufficiently. Likewise, the enrichment on "cell adhesion" indicated that the *in vitro* models may not recapitulate epithelial-to-mesenchymal-transition-related process of metastatic breast cancer. Surprisingly, for the 68 down-regulated subtype-and-model-independent DE genes, no GO terms passed the FDR < 0.001 criteria, which could be due to the small gene number. We decreased the FDR cutoff to 0.1 and observed 5 significant terms with cell division being the most significant (FDR = 0.029).

We further compared single sample gene set enrichment analysis (ssGSEA) scores of the 50 MSigDB hallmark gene sets among MET500 breast cancer samples, CCLE breast cancer cell lines and organoids to characterize their differences regarding to specific biological process (Fig 6a, Fig S10). For non-Basal-like MET500 breast cancer samples/CCLE breast cancer cell lines/organoids, we performed gene set differential activity analysis (DA) analysis on ssGSEA scores and identified 35 and 32 significant gene sets in MET500-vs-CCLE and MET500-vs-organoids comparisons, respectively (FDR < 0.001, Table S5). There were 26 differentially activated (DA) gene sets in common and for majority of them (23 of 26), the p-values derived from MET500-vs-CCLE comparison were lower than that derived from MET500-vs-organoid comparison, which may be unsurprising given that organoids more closely resemble the transcriptome of metastatic breast cancer samples (Fig 6b, left panel). We also performed the DA analysis for Basal-like subtype, identifying 19 and 24 significant gene sets in MET500-vs-CCLE and MET500-vs-organoids comparisons, respectively (Fig 6b, right panel and Table S5). For each of the subtypes, we classified the 50 hallmark gene sets into 4 categories according to DA analysis results:

I. Only significant in MET500-vs-organoids comparison (e.g., ANDROGEN RESPONSE).

II. Only significant in MET500-vs-CCLE comparison (e.g., E2F TARGETS).

III. Significant in both MET500-vs-organoids and MET500-vs-CCLE comparisons (e.g, COMPLEMENT).

IV. Not significant in either comparison (e.g, FATTY ACID METABOLISIM).

Interestingly, 27 gene sets could be consensually classified into one specific category, regardless of the subtype. Fig 6c shows the distribution of ssGSEA scores of the representative gene set for each category.

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 and organoids as models for metastatic breast cancer.

The genetic profile comparison showed that breast cancer cell lines poorly recaptured the mutation patterns of metastatic breast cancer samples, while the CNV profiles were more consistent. However, it is also worth noting that cell lines carried many specific genomic alternations, possibly due to culture effects. For example, we identified 25 genes showing cell-line-specific hypermutation and found that copy-number-loss events of some genes appeared to be quite limited in cancer cell lines.

Selecting cancer cell lines representative of tumors is vital for metastatic-cancer-related pre-clinical studies, and many factors are need to be considered. This study focused on two of them: metastatic site and subtype. Although the tumor micro-environment of different metastatic sites has large impact in shaping the genomic profiles of cancer cells, gene differential expression analysis demonstrated that cell lines failed to model such effect. This may explain the high correlation among different metastatic-sites in metastatic-site-specific TC analysis. Bone appears to be an exception, but requires further investigation. Breast cancer is quite heterogeneous and we showed that PAM50 subtypes were maintained in metastatic breast cancer cells. Considering the large genomic difference between Basal-like and other subtypes, it is not surprising to see that in subtype-specific TC analysis Basal-like subtype showed lower correlation with others. Prior to this study, a lot of research has been done to select representative cell lines as models for breast cancer, but the subtype information was not taken into consideration. Our analysis reveals the importance and necessity of subtype-specific cell line selection. In the future as data continues to accumulate, more factors can be considered for appropriate cell line selection and we can start building an ad-hoc mapping algorithm to metastasis samples and cell lines. Inputs into this algorithm would be the characteristics of metastatic cancer samples (subtype, metastatic site, even age, race, stage etc.) as well as the specific scientific question of interest and the output would be a list of appropriate cell lines.

We picked out suitable cell lines according to subtype-specific transcriptome correlation analysis results. Surprisingly, we found MDAMB231, the widely-used triple-negative cell line in metastatic breast cancer research, was dramatically different with Basal-like metastatic breast cancer samples. According to our analysis, HCC70 seems to be a better model, but this does not mean it can be directly employed to study cancer metastasis as many other criteria are needed for the assessment. In addition, although MDAMB231 has poor transcriptome (and CNV) correlation with metastatic breast cancer samples, it could still be very useful in studying some specific processes. Finally, since Basal-like breast cancer is itself highly heterogeneous, it is possible that MDAMB231 represents a rare subtype not delineated in the MET500 dataset.

Organoids are recently established *in vitro* models by 3D culturing and have shown large potential in translational research. Our analysis suggested that compared to cell lines, they have significantly better capacity to resemble the transcriptome of patient samples, which is a very useful characteristic in drug testing. It is also important to note that the cell lines evaluated in our study were established much earlier than organoids. Cell lines could have accumulated additional genomic alternations (during culturing process) which may result substantial transcriptome change and this may explain why recently established organoids are more correlated with patient samples. It is still unknown whether organoids bypass the issue mentioned above. In addition, through gene set differential activity analysis, we showed that some gene sets had organoid-specific high activity. Therefore, we conclude that organoids and cell lines are complementary with each other and further comparative studies are still needed.

In summary, by leveraging publicly available gnomic data, we comprehensively evaluated the suitability of breast cancer cell lines as models for metastatic breast cancer. Our study introduces a simple framework for cell line selection which can be easily extended to other cancer types. 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 to gain biological insights of cancer metastasis. 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.

Methods

Datasets

The raw RNA-Seq data of MET500 samples were downloaded from dbGap (under accession number phs000673.v2.p1) and further processed using RSEM[21,22]. FPKM values were used as gene expression measure. To keep consistent with other RNA-Seq datasets, only the RNA-Seq samples profiled with PolyA protocol were considered. 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 RNA-Seq 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[23,24] and CNV data were downloaded from BROAD GDAC Firehose (https://gdac.broadinstitute.org/). RSEM-processed gene expression data were downloaded from UCSC Xena data portal (https://xena.ucsc.edu/)[25].

The RNA-Seq data of patient-derived organoids was from Biobank[26].

We also searched GEO and manually assembled another microarray dataset containing gene expression value of 106 metastatic breast cancer samples[27,28,29,30]. The GEO accession numbers used were GSE11078, GSE14017, GSE14018, and GSE54323.

The gene expression data of lung-metastasis-derived MDAMB231 were downloaded from GEO under accession number GSE2603.

Detailed statistics of the above datasets are listed in Table S6.

Identification of differentially mutated genes between MET500 and TCGA samples

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

$$P_1 = 1 - \sum_{i=0}^{n} \Pr(i; N, \hat{q})$$

Where Pr 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 (for genes with zero mutation frequency, we used the minimum mutation frequency across all genes). Similarly, we computed left-tailed p-value (1- P_1) to test whether a gene has significantly lower mutation frequency in metastatic breast cancer samples. To control FDR, we applied the Benjamini-Hochberg procedure on left-tailed and right-tailed p-values respectively[31].

Transcriptome correlation analysis with RNA-Seq and microarray data

To perform transcriptome correlation analysis with RNA-Seq data, we first rank-transformed gene RPKM values for each CCLE cell line and then ranked all the genes according to their rank variation across all CCLE cell lines. The 1000 most-varied genes were kept as “marker genes”. Given RNA-Seq profiles of a cell line (or an organoid) and several patient samples, we compute the spearman rank correlation (across the 1000 marker genes) between the cell line and each sample and the median value of computed spearman rank correlations was defined as the transcriptome-similarity of the cell line with the patient samples. For microarray data, a similar procedure was applied and the 1000 most-varied probe sets were used to compute correlation values.

We also extended the above method to compute CNV similarity between a cell line and patient samples. Instead of selecting “marker genes”, all of the 1630 commonly genotyped genes were used.

PAM50 sub-typing and t-SNE visualization

The genefu package was used to determine breast cancer subtype[32,33]. 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[34].

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 and differentially activated gene sets

DESeq2 was used to identify differentially expressed genes (FDR < 0.001 and log2FC > 1) and DAVID bioinformatics sever was used to perform Gene Ontology enrichment analysis[35,36]. To increase statistical power, only protein coding genes were considered. 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[37–40]. In DA analysis, to identify gene sets which have differential activity, Wilcoxon rank test was used to assign p-values.

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[41,42]. The tumor purity was estimated using ESTIMATE[16]. CNTools was used to map the segmented CNV data to genes[43]. If not specified, the Wilcoxon rank test was used to compute p-value in hypothesis testing.

Figures

Fig 1. Overall design of the study. The upper panel lists data sources as well as sample types used in our study and the lower panel is a summary of the evaluations performed. TC analysis: transcriptome correlation analysis; DE analysis: gene differential expression analysis; DA analysis: gene set differential activity analysis.

Fig 2. Comparison of genetic profiles between MET500 breast cancer samples and CCLE breast cancer cell lines. (a) Somatic mutation profile of the 75 genes across MET500 breast cancer samples and CCLE breast cancer cell lines. The top-side color-bar indicates data sources (MET500 or CCLE) and the right-side color-bar indicates mutation frequency. (b) Comparison of CNV profiles (with the 1630 genes that are genotyped in both datasets) between MET500 breast cancer samples and all the 57 CCLE breast cancer cell lines. (c) Comparison of CNV profiles between MET500 breast cancer samples and the 33 primary-site derived CCLE breast cancer cell lines. (d) Comparison of CNV profiles between MET500 breast cancer samples and the 24 metastatic-site derived CCLE breast cancer cell lines. In panel (b), (c) and (d), each dot is a gene, y-axis represents its median CNV value across MET500 breast cancer samples and x-axis represents its median CNV value across cell lines. In panel (c) and (d), genes with high copy-number-gain in MET500 samples were marked as red.

Fig 3. Transcriptome correlation analysis between MET500 breast cancer samples and CCLE breast cancer cell lines. (a) Ranking 1019 CCLE cell lines according to their transcriptome-similarity with MET500 breast cancer samples. Each dot is a CCLE cell line with breast cancer cell lines marked as red. (b) Metastatic-site-specific transcriptome correlation analysis results are highly correlated between liver and lymph node. Each dot is a CCLE breast cancer cell line with x-axis represents its transcriptome-similarity with the 9 lymph node derived MET500 breast cancer samples and y-axis represents its transcriptome-similarity with the 27 liver

derived MET500 breast cancer samples. (c) The t-SNE plot of MET500 breast cancer samples (only the expression values of PAM50 genes were used). Metastatic-sites are labeled by color and subtypes are labeled by shape. (d) Pair-wise comparison of subtype-specific transcriptome correlation analysis results. In the lower-triangle part, each dot is a CCLE breast cancer cell line with the two axis representing transcriptome-similarity of the cell line with MET500 breast cancer samples of the corresponding two subtypes; the upper-triangle part shows the pair-wise spearman rank correlation values of every two subtypes.

Fig 4. MDAMB231 has substantial genomic difference with metastatic breast cancer samples. (a) MDAMB231 does not closely resemble the transcriptome of Basal-like breast cancer. The left panel shows the ranking of all 1019 CCLE cell lines according to their transcriptome-similarity with Basal-like MET500 breast cancer samples. 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 Basal-like MET500 breast cancer samples. The boxplot on the right panel shows the distribution of expression correlation values (with Basal-like MET500 breast cancer samples) for MDAMB231 and HCC70. (b) MDAMB231 does not closely resemble CNV profile of Basal-like breast cancer. The left panel shows the ranking of all 1019 CCLE cell lines according to their CNV similarity with Basal-like MET500 breast cancer samples; the boxplot on the right panel shows the distribution of CNV correlation values (with Basal-like MET500 breast cancer samples) for MDAMB231 and HCC70. (c) Somatic mutation profile of the 25 highly mutated genes across MDAMB231 and Basal-like MET500 breast cancer samples. (d) Boxplot of expression correlation between cell lines (including CCLE breast cancer cell lines, lung-metastasis-derived MDAMB231 which are colored by red) and lung-derived metastatic breast cancer samples.

Fig 5. Comparing CCLE breast cancer cell lines with patient-derived organoids using

gene expression data. (a) Pair-wise comparison of subtype-specific transcriptome correlation analysis results. In the lower-triangle part, each dot is an established organoid with the two axis representing its transcriptome-similarity with MET500 breast cancer samples of the corresponding two subtypes; the upper-triangle part shows the pair-wise spearman rank correlation values of every two subtypes. (b) Boxplot of transcriptome-similarity values (with MET500 breast cancer samples of different subtypes) of CCLE breast cancer cell lines and organoids. (c) For each subtype, the most-correlated organoid has significantly higher expression correlation values with MET500 breast cancer samples (of that subtype) than the most-correlated cell line.

Fig 6. Comparison of ssGSEA scores of the 50 MSigDB hallmark gene sets. (a) Visualization of ssGSEA scores across CCLE breast cancer cell lines, MET500 breast caner samples and organoids (non Basal-like subtype). (b) DA analysis results of different breast cancer subtypes. Each dot is a hallmark gene set with x-axis represent -log10(FDR) derived from MET500 vs organoids analysis and y-axis represent -log10(FDR) derived from MET500 vs CCLE analysis. (c) Boxplot of ssGSEA scores of the four representative gene sets.

Supplementary Figures

Fig S1. (a) Long-tailed gene mutation spectrum in MET500 breast cancer samples. (b) Volcano plot of gene differential mutation analysis. The dashed line corresponds to FDR = 0.001. (c) Visualization of log10-transformed mutation frequency of the 25 genes that are specifically hyper-mutated 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 breast cancer samples. 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.

Fig S2. (a) MET500 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. Data-sources are labeled by color and subtypes are labeled by shape.

Fig S3. Metastatic-site-specific and subtype-specific transcriptome correlation analysis results are highly correlated between MET500 dataset and the microarray dataset. In each plot, a dot is a CCLE breast cancer cell line with x-axis represents transcriptome-similarity derived from MET500 dataset and y-axis represents transcriptome-similarity derived from the assembled microarray dataset.

Fig S4. Pair-wise comparison of metastatic-site-specific transcriptome correlation analysis results among metastatic sites (microarray dataset).

Fig S5. Pair-wise comparison of subtype-specific transcriptome correlation analysis results among subtypes (microarray dataset).

Fig S6. Boxplot of tumor purity (microarray dataset).

Fig S7. Normal qqplot to confirm the transcriptome-similarity value between a random cell line 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.

Fig S8. (a) Compared to BT483, MCF7 shows significantly lower expression correlation with LuminalB MET500 breast cancer samples. (b) Compared to MDAMB415, T47D does not show significantly lower expression correlation with LuminalA MET500 breast cancer samples. (c) Compared to EFM192A, T47D shows significant lower expression correlation with Her2-enriched MET500 breast cancer samples.

Fig S9. Workflow of gene differential expression analysis.

Fig S10. Visualization of ssGSEA scores across CCLE breast cancer cell lines, MET500 breast cancer samples and organoids (Basal-like subtype).

Supplementary Tables

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

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

Table S3. Testing suitability of CCLE breast cancer cell lines for different subtypes.

Table S4. Results of GO enrichment analysis.

Table S5. Results of DA analysis.

Table S6. Detailed statistics of the datasets used in our study.

Acknowledgments

We thank Dr.Nijman of Utrecht Medical Center for providing us organoids RNA-Seq data. 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. Iorio, F. *et al.* A Landscape of Pharmacogenomic Interactions in Cancer. *Cell* **166,** 740–754 (2016).

2. Wilding, J. L. & Bodmer, W. F. Cancer cell lines for drug discovery and development. *Cancer Research* **74,** 2377–2384 (2014).

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

4. Gillet, J.-P. *et al.* Redefining the relevance of established cancer cell lines to the study of mechanisms of clinical anti-cancer drug resistance. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* **108,** 18708–18713 (2011).

5. Weinstein, J. N. *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *NATURE GENETICS* **45,** 1113–1120 (2013).

6. Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity (vol 483, pg 603, 2012). *NATURE* **492,** 290 (2012).

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

8. Chen, B., Sirota, M., Fan-Minogue, H., Hadley, D. & Butte, A. J. Relating hepatocellular carcinoma tumor samples and cell lines using gene expression data in translational research. *BMC MEDICAL GENOMICS* **8,** (2015).

9. Jiang, G. *et al.* Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. *BMC Genomics* **17,** (2016).

10. Lambert, A. W., Pattabiraman, D. R. & Weinberg, R. A. Emerging Biological Principles of Metastasis. *CELL* **168,** 670–691 (2017).

11. Mehlen, P. & Puisieux, A. Metastasis: A question of life or death. *Nature Reviews Cancer* **6,** 449–458 (2006).

12. Robinson, D. R. *et al.* Integrative clinical genomics of metastatic cancer. *Nature* **548,** 297–303 (2017).

13. Lefebvre, C. *et al.* Mutational Profile of Metastatic Breast Cancers: A Retrospective Analysis. *PLOS MEDICINE* **13,** (2016).

14. Bartels, S. *et al.* Estrogen receptor (ESR1) mutation in bone metastases from breast cancer. *Modern Pathology* **31,** 56–61 (2018).

15. Sandberg, R. & Ernberg, I. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* **102,** 2052–2057 (2005).

16. Yoshihara, K. *et al.* Inferring tumour purity and stromal and immune cell admixture from expression data. *NATURE COMMUNICATIONS* **4,** (2013).

17. Kang, Y. B. *et al.* A multigenic program mediating breast cancer metastasis to bone. *CANCER CELL* **3,** 537–549 (2003).

18. Minn, A. J. *et al.* Genes that mediate breast cancer metastasis to lung. *Nature* **436,** 518–524 (2005).

19. Weeber, F., Ooft, S. N., Dijkstra, K. K. & Voest, E. E. Tumor Organoids as a Pre-clinical Cancer Model for Drug Discovery. *Cell Chemical Biology* **24,** 1092–1100 (2017).

20. Drost, J. & Clevers, H. Organoids in cancer research. *Nature Reviews Cancer* **18,** 407–418 (2018).

21. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12,** (2011).

22. Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A. & Dewey, C. N. RNA-Seq gene expression estimation with read mapping uncertainty. *BIOINFORMATICS* **26,** 493–500 (2010).

23. Gao, J. *et al.* Integrative Analysis of Complex Cancer Genomics and Clinical Profiles Using the cBioPortal. *SCIENCE SIGNALING* **6,** (2013).

24. Cerami, E. *et al.* The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data (vol 2, pg 401, 2012). *CANCER DISCOVERY* **2,** 960 (2012).

25. Vivian, J. *et al.* Toil enables reproducible, open source, big biomedical data analyses. *NATURE BIOTECHNOLOGY* **35,** 314–316 (2017).

26. Sachs, N. *et al.* A Living Biobank of Breast Cancer Organoids Captures Disease Heterogeneity. *CELL* **172,** 373+ (2018).

27. Landemaine, T. *et al.* A six-gene signature predicting breast cancer lung metastasis. *CANCER RESEARCH* **68,** 6092–6099 (2008).

28. Xu, J. *et al.* 14-3-3ζ Turns TGF-β's Function from Tumor Suppressor to Metastasis Promoter in Breast Cancer by Contextual Changes of Smad Partners from p53 to Gli2. *Cancer Cell* **27,** 177–192 (2015).

29. Zhang, X. H.-F. *et al.* Latent Bone Metastasis in Breast Cancer Tied to Src-Dependent Survival Signals. *CANCER CELL* **16,** 67–78 (2009).

30. Foukakis, T. *et al.* Gene expression profiling of sequential metastatic biopsies for biomarker discovery in breast cancer. *MOLECULAR ONCOLOGY* **9,** 1384–1391 (2015).

31. BENJAMINI, Y. & HOCHBERG, Y. CONTROLLING THE FALSE DISCOVERY RATE - A PRACTICAL AND POWERFUL APPROACH TO MULTIPLE TESTING. *JOURNAL OF THE ROYAL STATISTICAL SOCIETY SERIES B-METHODOLOGICAL* **57,** 289–300 (1995).

32. Bernard, P. S. *et al.* Supervised risk predictor of breast cancer based on intrinsic subtypes. *Journal of Clinical Oncology* **27,** 1160–1167 (2009).

33. Gendoo, D. M. A. *et al.* Genefu: an R/Bioconductor package for computation of gene expression-based signatures in breast cancer. *BIOINFORMATICS* **32,** 1097–1099 (2016).

34. Van Der Maaten, L. J. P. & Hinton, G. E. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* **9,** 2579–2605 (2008).

35. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *GENOME BIOLOGY* **15,** (2014).

36. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *NATURE PROTOCOLS* **4,** 44–57 (2009).

37. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA* **102,** 15545–15550 (2005).

38. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *BIOINFORMATICS* **27,** 1739–1740 (2011).

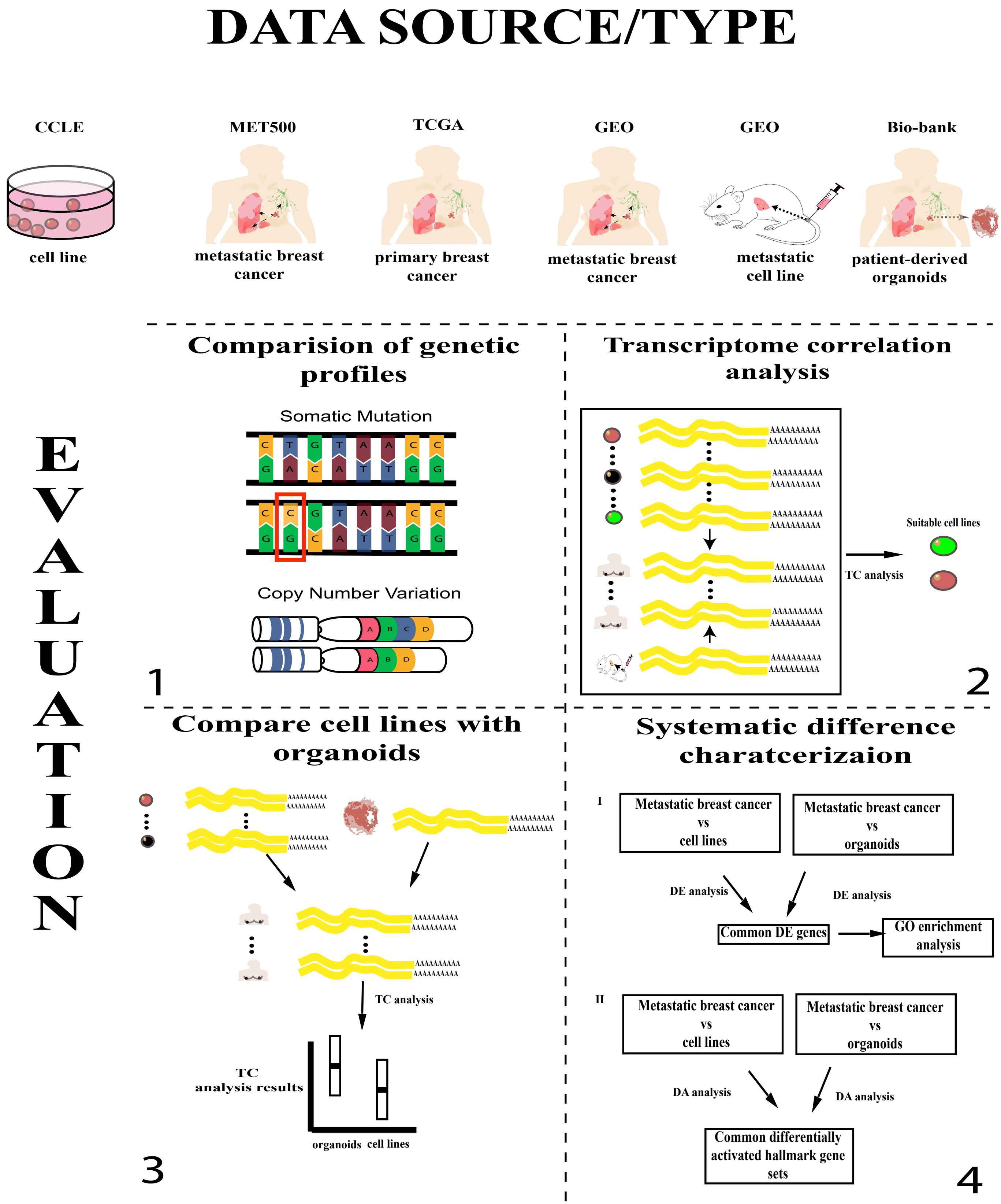
39. Liberzon, A. *et al.* The Molecular Signatures Database Hallmark Gene Set Collection. *CELL SYSTEMS* **1,** 417–425 (2015).

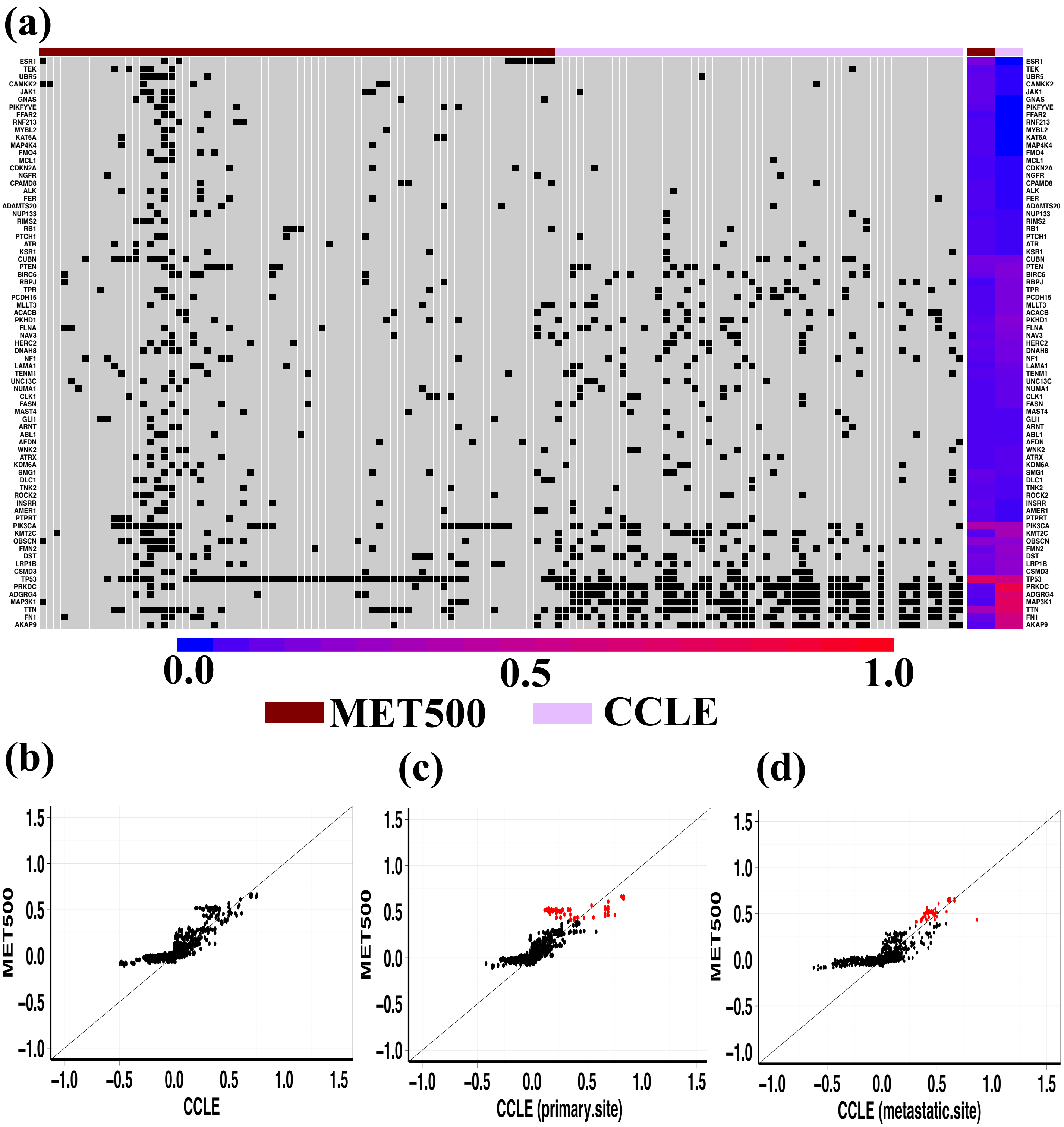
40. Haenzelmann, S., Castelo, R. & Guinney, J. GSVA: gene set variation analysis for microarray and RNA-Seq data. *BMC BIOINFORMATICS* **14,** (2013).

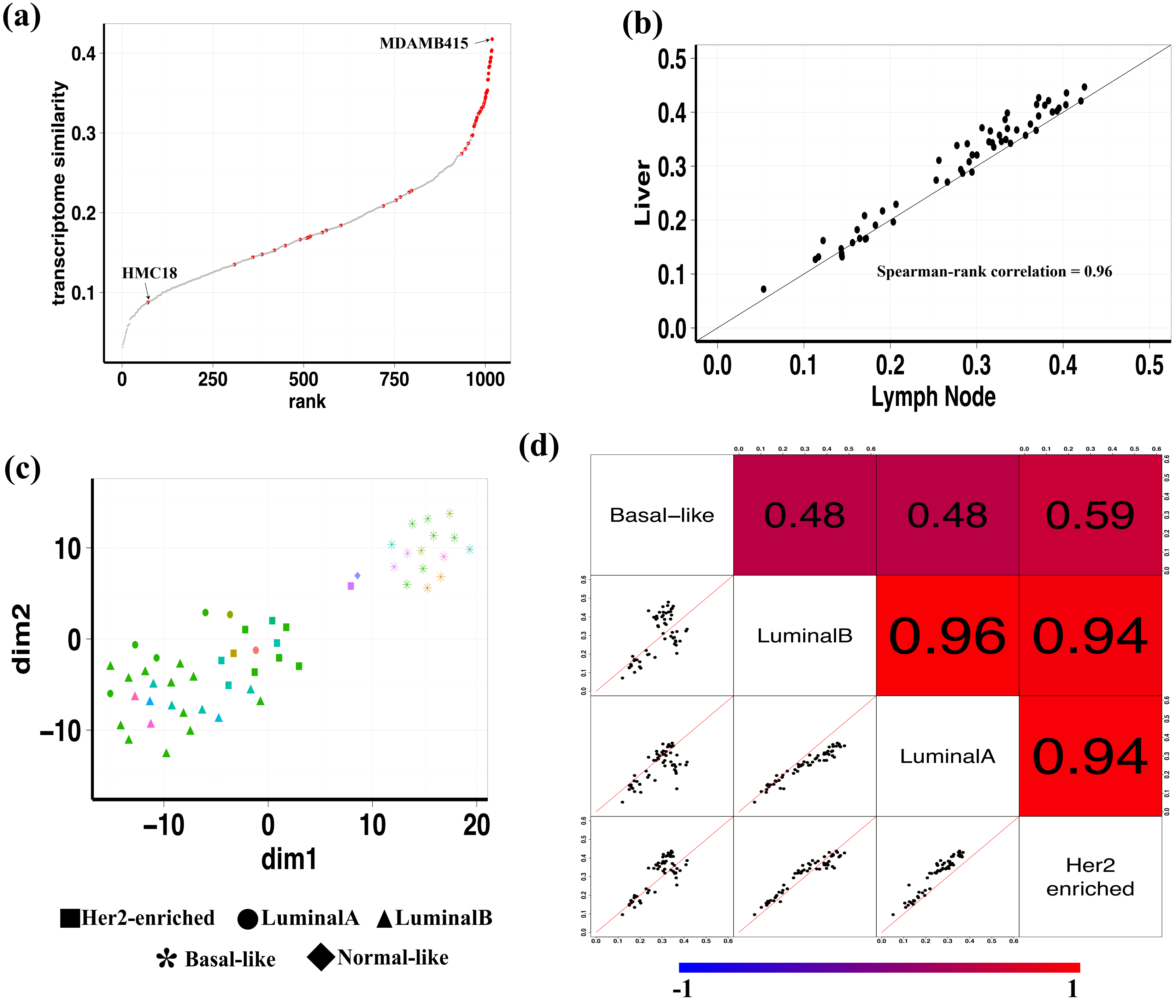
41. The ComplexHeatmap package. Available at: https://bioconductor.org/packages/release/bioc/html/ComplexHeatmap.html.

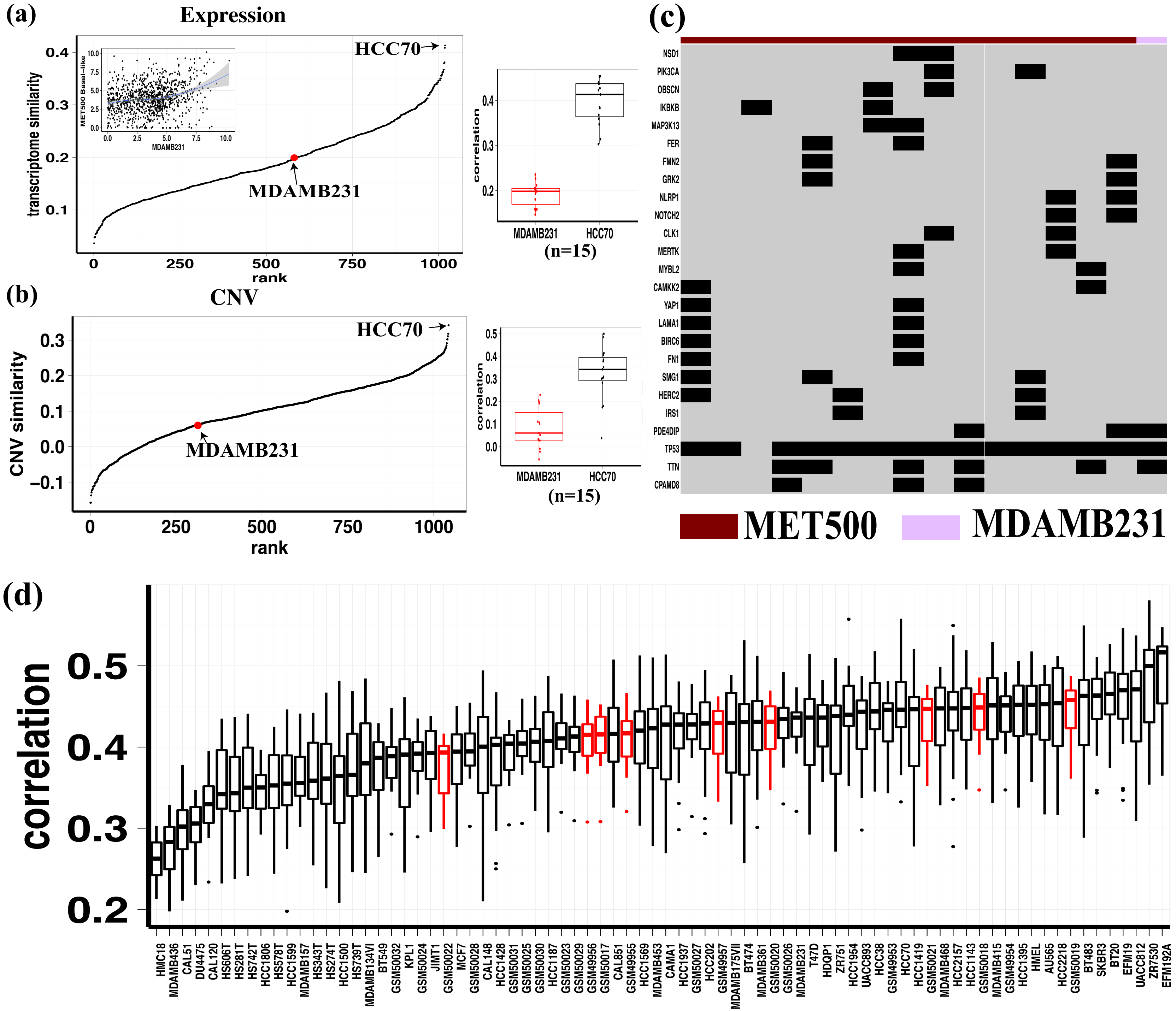
42. The ggplot2 package. Available at: https://cran.r-project.org/web/packages/ggplot2/index.html.

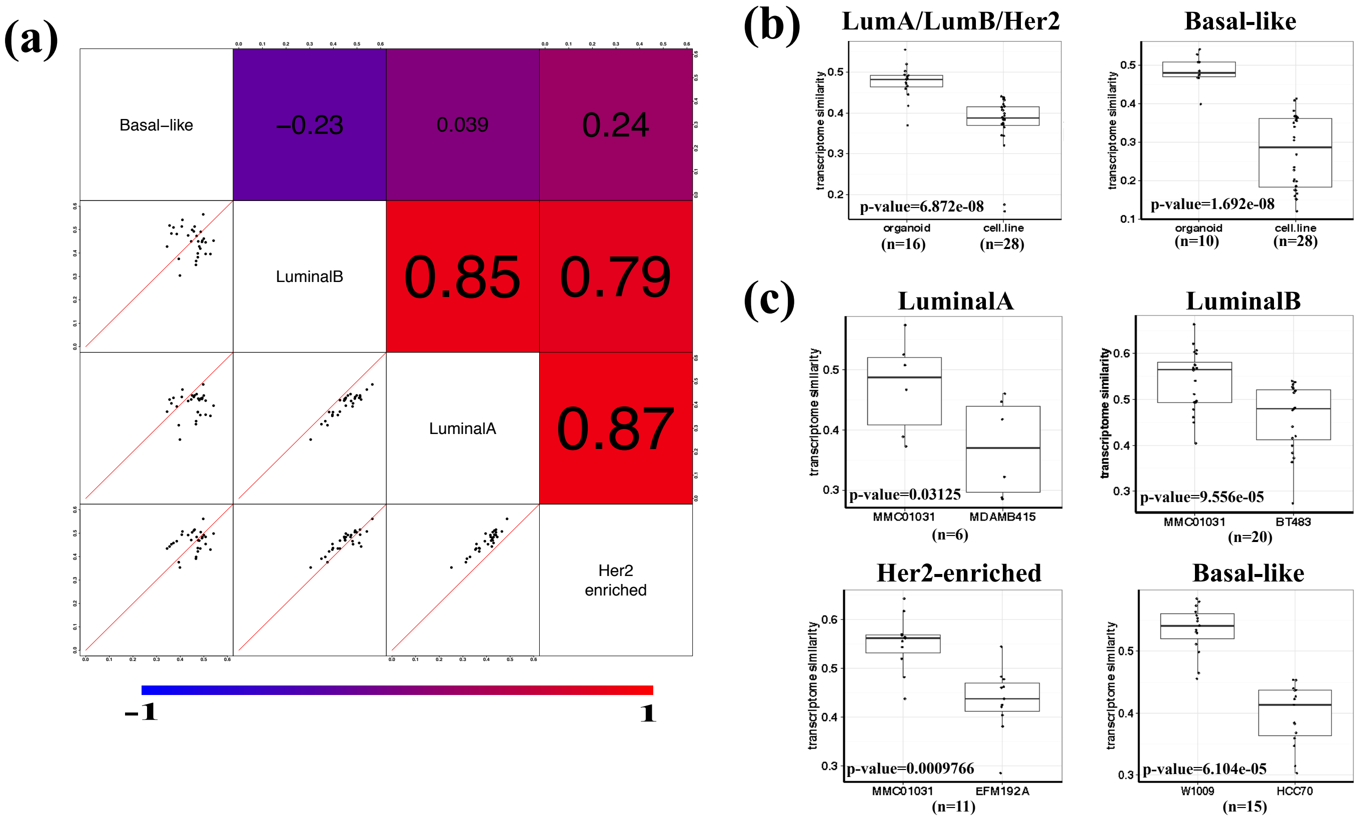
43. Jianhua Zhang. The CNTools package. Available at: https://bioconductor.org/packages/release/bioc/html/CNTools.html.

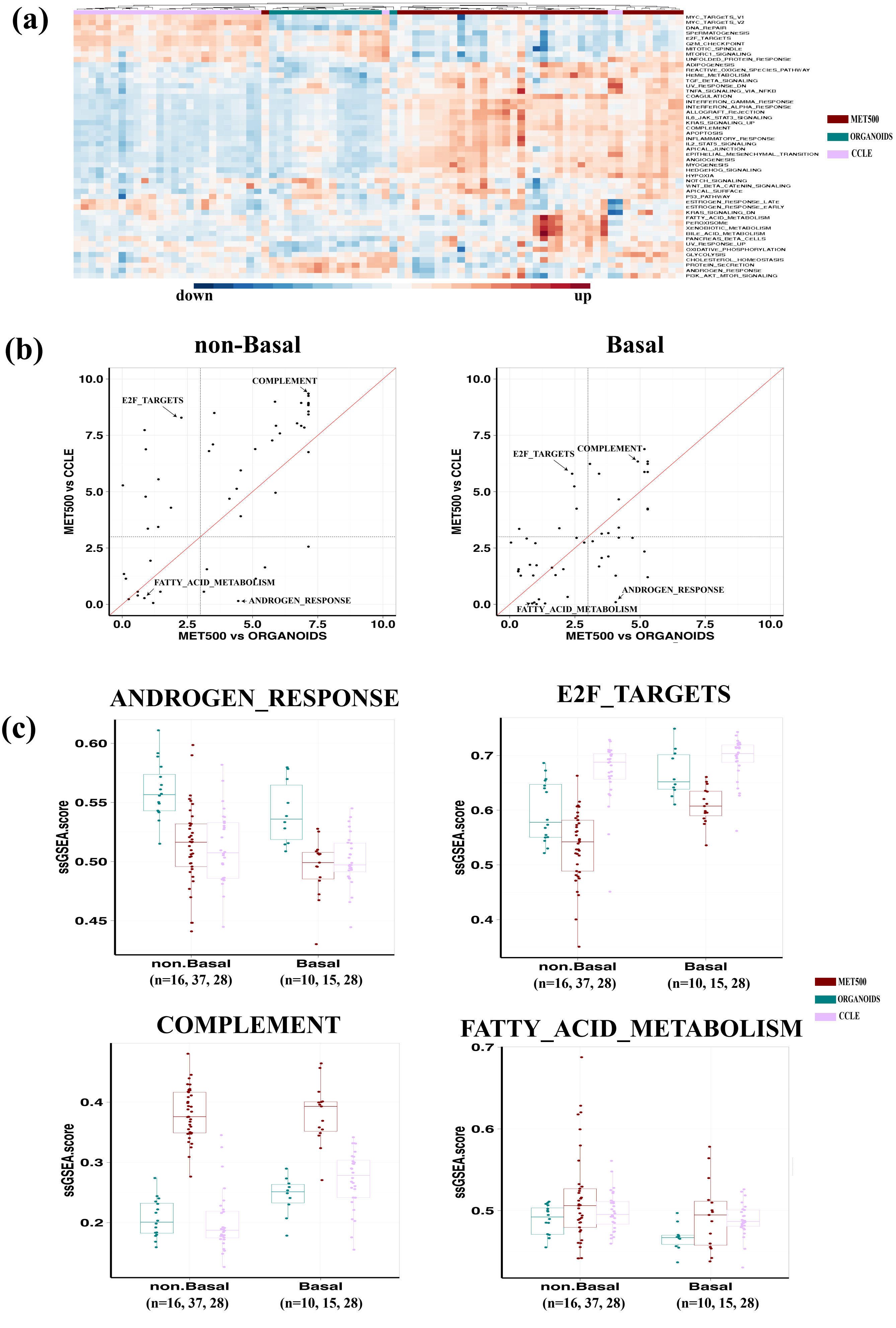

DATA SOURCE/TYPE
CCLE
cell line
MET500
metastatic breast cancer
TCGA
primary breast cancer
GEO
metastatic breast cancer
GEO
metastatic cell line
Bio-bank
patient-derived organoids
EVALUATION
Comparision of genetic profiles
Somatic Mutation
C T G T A A C C
G A C A T T G G
C C G T A A C C
G G C A T T G G
Copy Number Variation
A B C D
A B D
1
Transcriptome correlation analysis
AAAAAAAAAA
Suitable cell lines
TC analysis
2
Compare cell lines with organoids
AAAAAAAAAA
TC analysis
TC analysis results
organoids
cell lines
3
Systematic difference charatcerizaion
I
Metastatic breast cancer vs cell lines
Metastatic breast cancer vs organoids
DE analysis
DE analysis
Common DE genes
GO enrichment analysis
II
Metastatic breast cancer vs cell lines
Metastatic breast cancer vs organoids
DA analysis
DA analysis
Common differentially activated hallmark gene sets
4

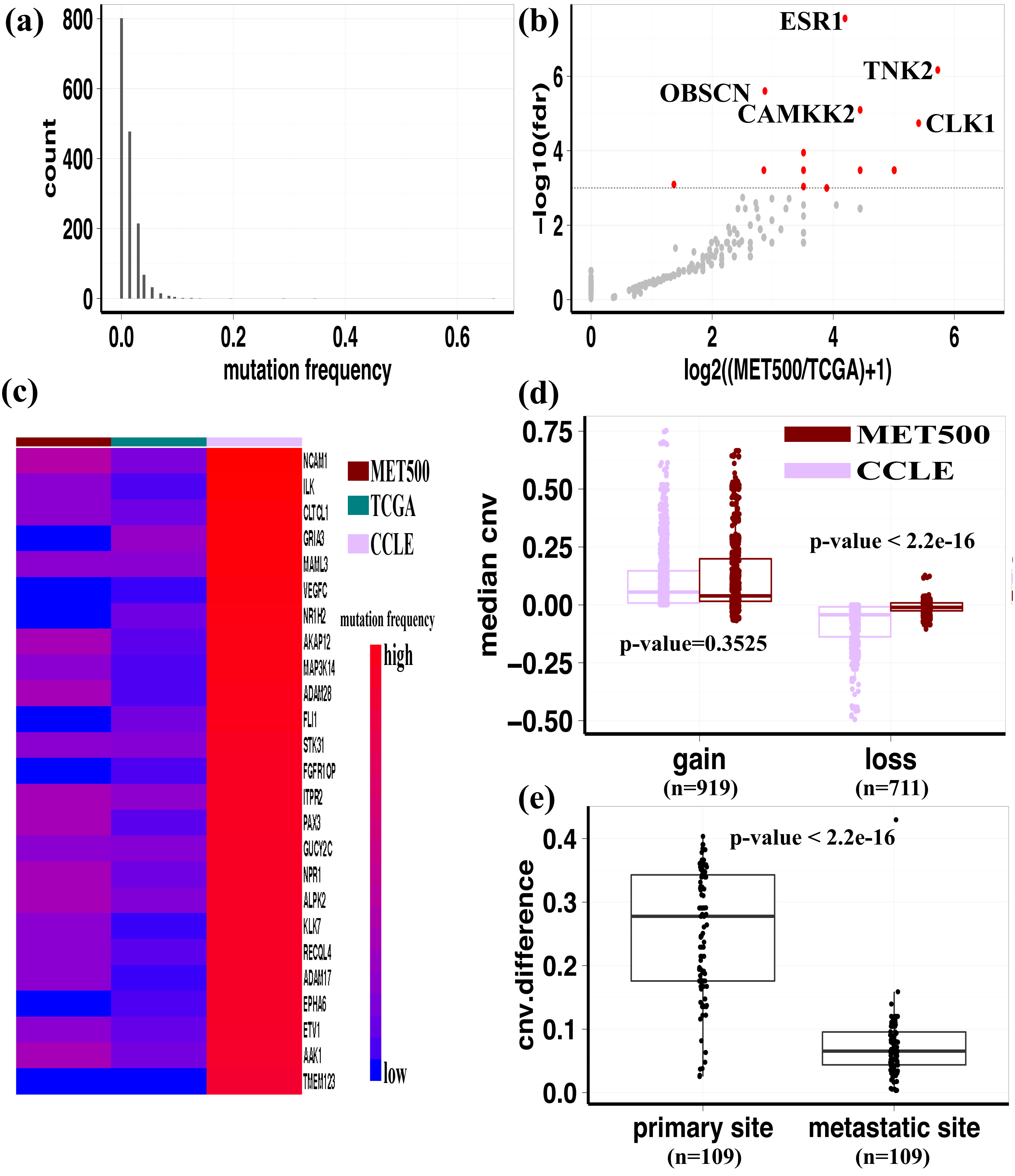

(a)
ESR1
TEK
UBR5
CAMKK2
JAK1
GNAS
PIKFYVE
FFAR2
RNF213
MYBL2
KAT6A
MAP4K4
FMO4
MCL1
CDKN2A
NGFR
CPAMD8
ALK
FER
ADAMTS20
NUP133
RIMS2
RB1
PTCH1
ATR
KSR1
CUBN
PTEN
BIRC6
RBPJ
TPR
PCDH15
MLLT3
ACACB
PKHD1
FLNA
NAV3
HERC2
DNAH8
NF1
LAMA1
TENM1
UNC13C
NUMA1
CLK1
FASN
MAST4
GLI1
ARNT
ABL1
AFDN
WNK2
ATRX
KDM6A
SMG1
DLC1
TNK2
ROCK2
INSRR
AMER1
PTPRT
PIK3CA
KMT2C
OBSCN
FMN2
DST
LRP1B
CSMD3
TP53
PRKDC
ADGRG4
MAP3K1
TTN
FN1
AKAP9
0.0
0.5
1.0
MET500
CCLE
(b)
MET500
CCLE
(c)
MET500
CCLE (primary.site)
(d)
MET500
CCLE (metastatic.site)
1.5
1.0
0.5
0.0
-0.5
-1.0

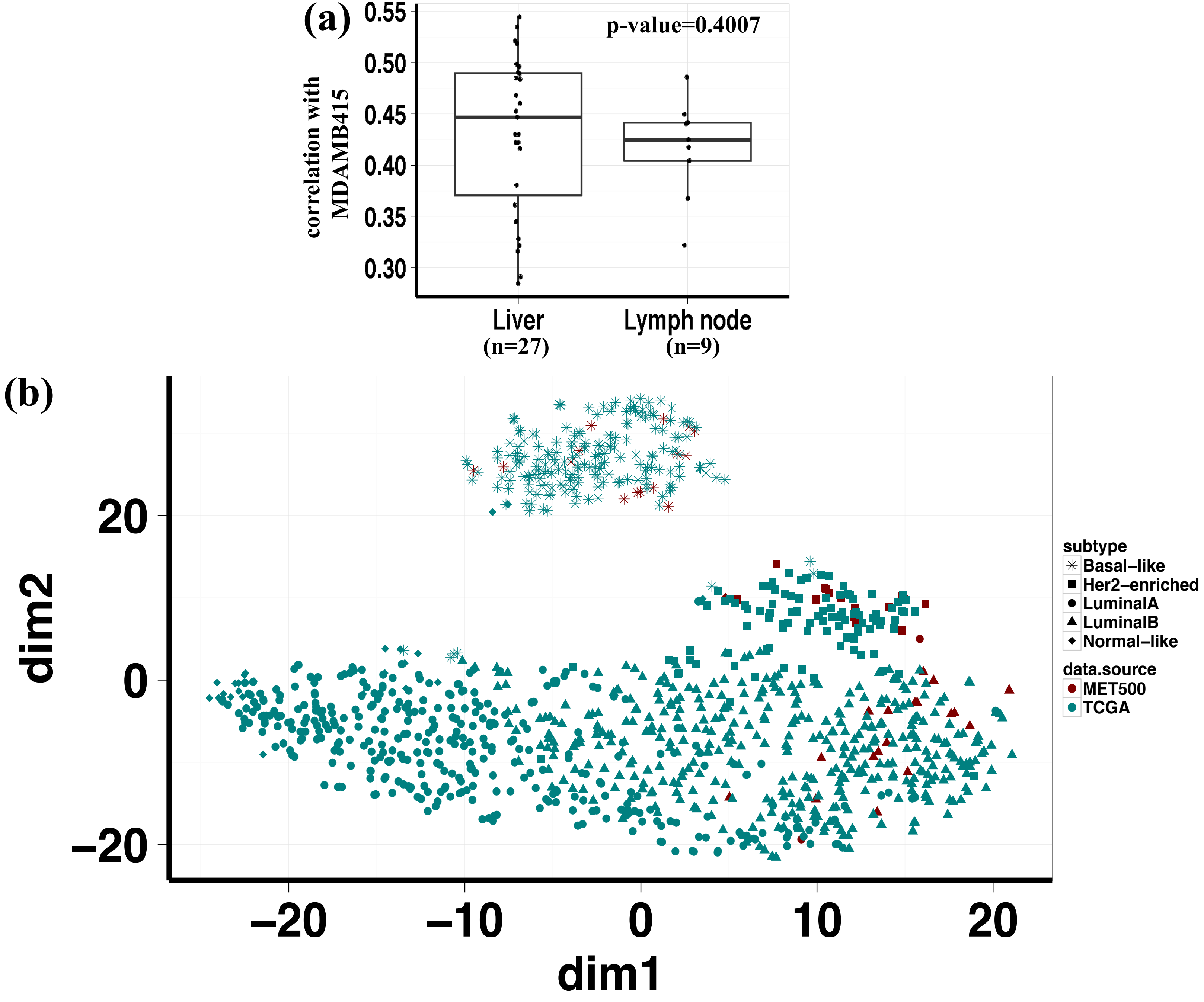

(a)
MDAMB415
HMC18
transcriptome similarity
0.1
0.2
0.3
0.4
0
250
500
750
1000
rank
(b)
Liver
Lymph Node
Spearman-rank correlation = 0.96
0.0
0.1
0.2
0.3
0.4
0.5
(c)
dim2
dim1
−10
0
10
20
■Her2-enriched
●LuminalA
▲LuminalB
✻ Basal-like
◆Normal-like
(d)
Basal−like
LuminalB
LuminalA
Her2 enriched
0.48
0.48
0.59
0.96
0.94
0.94
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-1
1

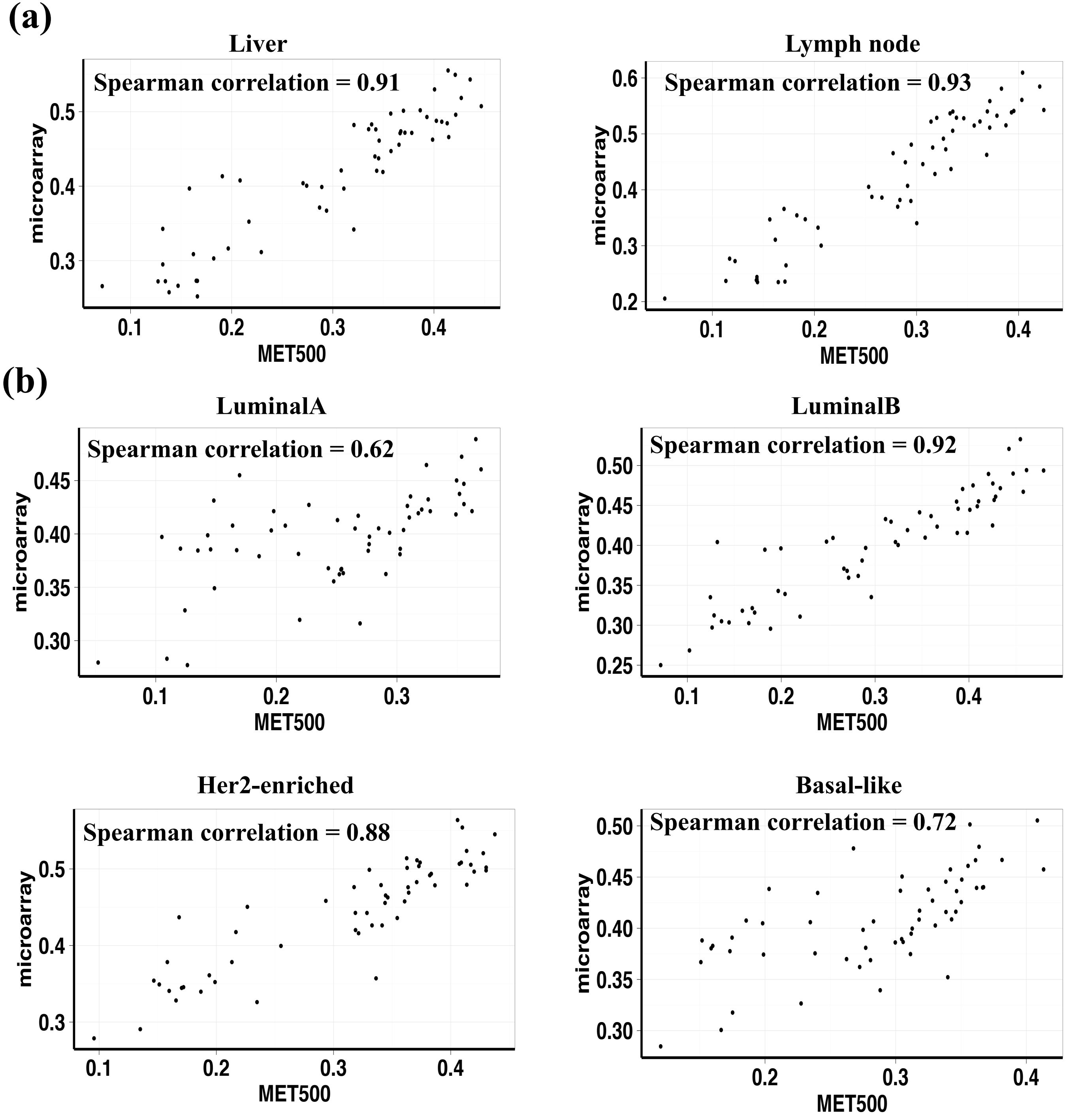

(a)
Expression
transcriptome similarity
0.4
0.3
0.2
0.1
0
250
500
750
1000
rank
HCC70
MDAMB231
MET500 Basal-like
10.0
7.5
5.0
2.5
0.0
0.0
2.5
5.0
7.5
10.0
MDAMB231
correlation
0.4
0.3
0.2
MDAMB231
HCC70
(n=15)
(b)
CNV
CNV similarity
0.3
0.2
0.1
0.0
−0.1
0
250
500
750
1000
rank
HCC70
MDAMB231
correlation
0.5
0.4
0.3
0.2
0.1
0.0
MDAMB231
HCC70
(n=15)
(c)
NSD1
PIK3CA
OBSCN
IKBKB
MAP3K13
FER
FMN2
GRK2
NLRP1
NOTCH2
CLK1
MERTK
MYBL2
CAMKK2
YAP1
LAMA1
BIRC6
FN1
SMG1
HERC2
IRS1
PDE4DIP
TP53
TTN
CPAMD8
MET500
MDAMB231
(d)
correlation
0.5
0.4
0.3
0.2
HMC18
MDAMB436
CAL51
DU4475
CAL120
HS606T
HS281T
HS742T
HCC1806
HS578T
HCC1599
MDAMB157
HS343T
HS274T
HCC1500
HS739T
MDAMB134VI
BT549
GSM50032
KPL1
GSM50024
JIMT1
GSM50022
MCF7
GSM50028
CAL148
HCC1428
GSM50031
GSM50025
GSM50030
HCC1187
GSM50023
GSM50029
GSM49956
GSM50017
CAL851
GSM49955
HCC1569
MDAMB453
CAMA1
HCC1937
GSM50027
HCC202
GSM49957
MDAMB175VII
BT474
MDAMB361
GSM50020
GSM50026
MDAMB231
T47D
HDQP1
ZR751
HCC1954
UACC893
HCC38
GSM49953
HCC70
HCC1419
GSM50021
MDAMB468
HCC2157
HCC1143
GSM50018
MDAMB415
GSM49954
HCC1395
HMEL
AU565
HCC2218
GSM50019
BT483
SKBR3
BT20
EFM19
UACC812
ZR7530
EFM192A

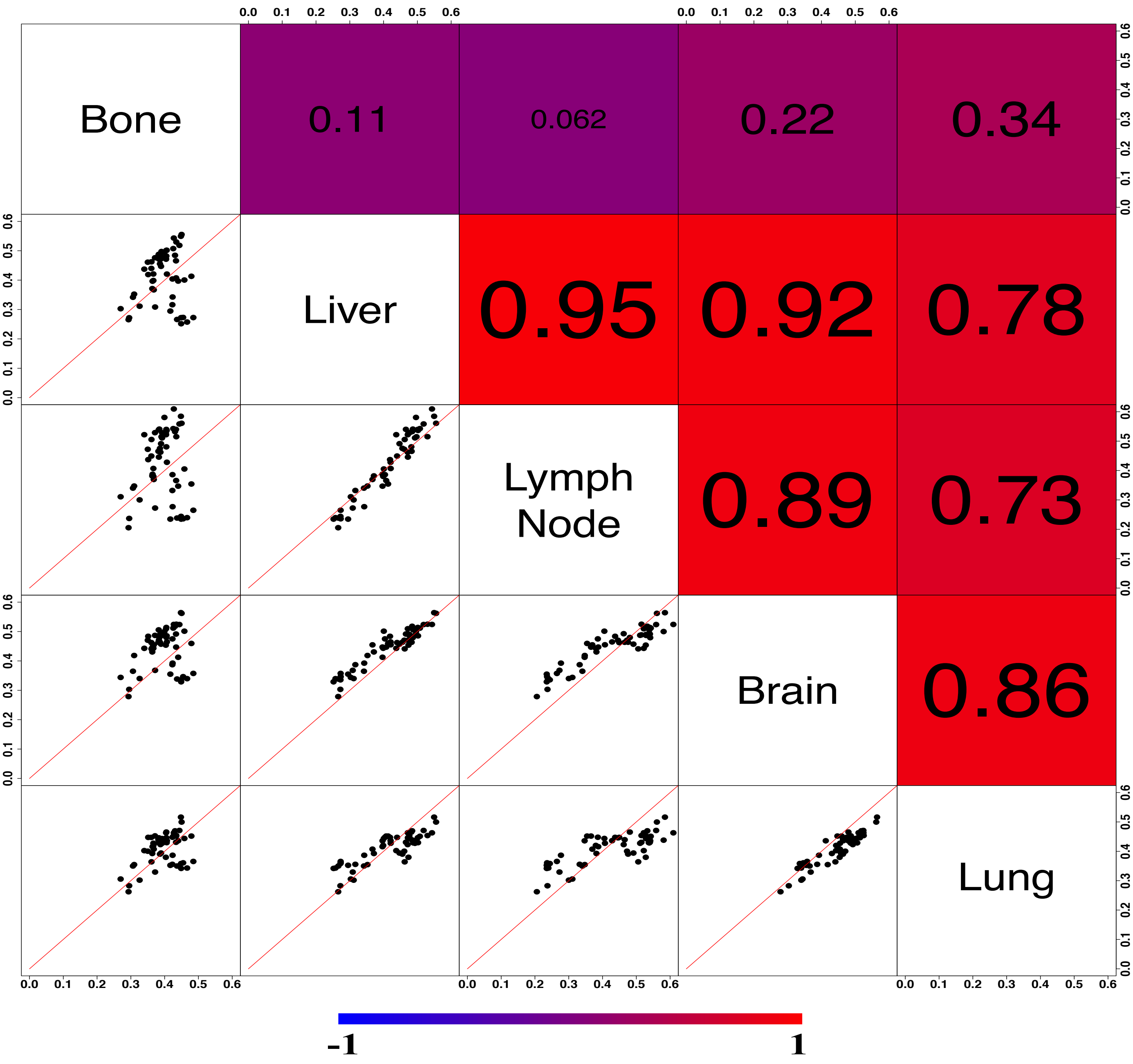

(a)
Basal–like
–0.23
0.039
0.24
LuminalB
0.85
0.79
LuminalA
0.87
Her2
enriched
0.0
0.1
0.2
0.3
0.4
0.5
0.6
–1
1
(b)
LumA/LumB/Her2
transcriptome similarity
0.5
0.4
0.3
0.2
p-value=6.872e-08
organoid
(n=16)
cell.line
(n=28)
Basal-like
transcriptome similarity
0.5
0.4
0.3
0.2
0.1
p-value=1.692e-08
organoid
(n=10)
cell.line
(n=28)
(c)
LuminalA
transcriptome similarity
0.5
0.4
0.3
p-value=0.03125
MMC01031
MDAMB415
(n=6)
LuminalB
transcriptome similarity
0.6
0.5
0.4
0.3
p-value=9.556e-05
MMC01031
BT483
(n=20)
Her2-enriched
transcriptome similarity
0.6
0.5
0.4
0.3
p-value=0.0009766
MMC01031
EFM192A
(n=11)
Basal-like
transcriptome similarity
0.5
0.4
0.3
p-value=6.104e-05
W1009
HCC70
(n=15)


(a)
MYC_TARGETS_V1
MYC_TARGETS_V2
DNA_REPAIR
SPERMATOGENESIS
E2F_TARGETS
G2M_CHECKPOINT
MITOTIC_SPINDLE
MTORC1_SIGNALING
UNFOLDED_PROTEIN_RESPONSE
ADIPOGENESIS
REACTIVE_OXIGEN_SPECIES_PATHWAY
HEME_METABOLISM
TGF_BETA_SIGNALING
UV_RESPONSE_DN
TNFA_SIGNALING_VIA_NFKB
COAGULATION
INTERFERON_GAMMA_RESPONSE
INTERFERON_ALPHA_RESPONSE
ALLOGRAFT_REJECTION
IL6_JAK_STAT3_SIGNALING
KRAS_SIGNALING_UP
COMPLEMENT
APOPTOSIS
INFLAMMATORY_RESPONSE
IL2_STAT5_SIGNALING
APICAL_JUNCTION
EPITHELIAL_MESENCHYMAL_TRANSITION
ANGIOGENESIS
MYOGENESIS
HEDGEHOG_SIGNALING
HYPOXIA
NOTCH_SIGNALING
WNT_BETA_CATENIN_SIGNALING
APICAL_SURFACE
P53_PATHWAY
ESTROGEN_RESPONSE_LATE
ESTROGEN_RESPONSE_EARLY
KRAS_SIGNALING_DN
FATTY_ACID_METABOLISM
PEROXISOME
XENOBIOTIC_METABOLISM
BILE_ACID_METABOLISM
PANCREAS_BETA_CELLS
UV_RESPONSE_UP
OXIDATIVE_PHOSPHORYLATION
GLYCOLYSIS
CHOLESTEROL_HOMEOSTASIS
PROTEIN_SECRETION
ANDROGEN_RESPONSE
PI3K_AKT_MTOR_SIGNALING
MET500
ORGANOIDS
CCLE
down
up
(b)
non-Basal
Basal
10.0
7.5
5.0
2.5
0.0
MET500 vs CCLE
MET500 vs ORGANOIDS
E2F_TARGETS
COMPLEMENT
FATTY_ACID_METABOLISM
ANDROGEN_RESPONSE
(c)
ANDROGEN_RESPONSE
0.60
0.55
0.50
0.45
ssGSEA.score
non.Basal
(n=16, 37, 28)
Basal
(n=10, 15, 28)
E2F_TARGETS
0.7
0.6
0.5
0.4
ssGSEA.score
non.Basal
(n=16, 37, 28)
Basal
(n=10, 15, 28)
MET500
ORGANOIDS
CCLE
COMPLEMENT
0.4
0.3
0.2
ssGSEA.score
non.Basal
(n=16, 37, 28)
Basal
(n=10, 15, 28)
FATTY_ACID_METABOLISM
0.7
0.6
0.5
ssGSEA.score
non.Basal
(n=16, 37, 28)
Basal
(n=10, 15, 28)

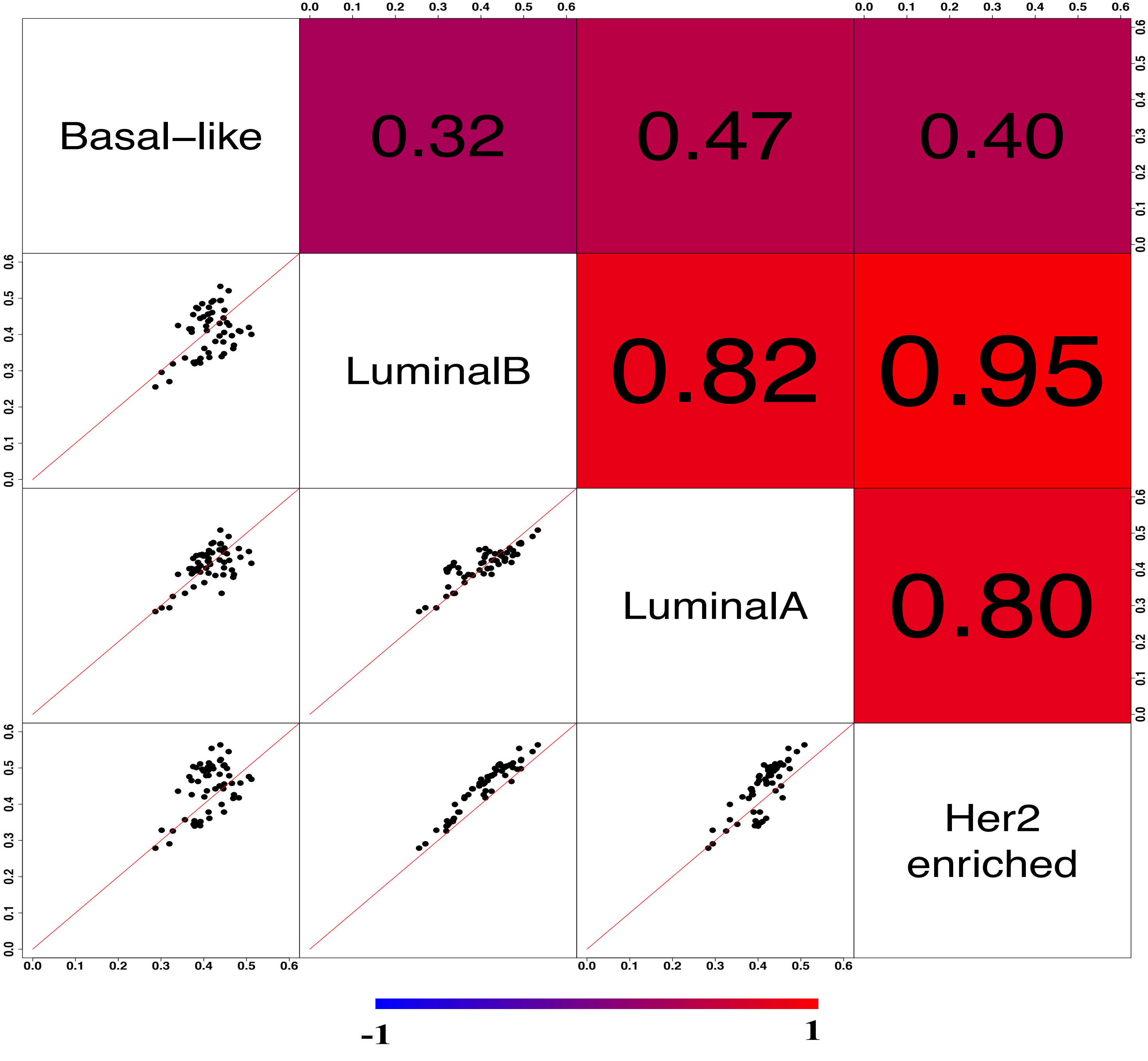

(a)
800
600
400
200
0
count
0.0
0.2
0.4
0.6
mutation frequency
(b)
ESR1
TNK2
OBSCN
CAMKK2
CLK1
6
4
2
0
−log10(fdr)
0
2
4
6
log2((MET500/TCGA)+1)
(c)
NCAM1
ILK
CLTCL1
GRIA3
MAML3
VEGFC
NR1H2
AKAP12
MAP3K14
ADAM28
FLI1
STK31
FGFR1OP
ITPR2
PAX3
GUCY2C
NPR1
ALPK2
KLK7
RECQL4
ADAM17
EPHA6
ETV1
AAK1
TMEM123
MET500
TCGA
CCLE
mutation frequency
high
low
(d)
0.75
0.50
0.25
0.00
−0.25
−0.50
median cnv
MET500
CCLE
p-value < 2.2e-16
p-value=0.3525
gain
(n=919)
loss
(n=711)
(e)
0.4
0.3
0.2
0.1
0.0
cnv.difference
p-value < 2.2e-16
primary site
(n=109)
metastatic site
(n=109)

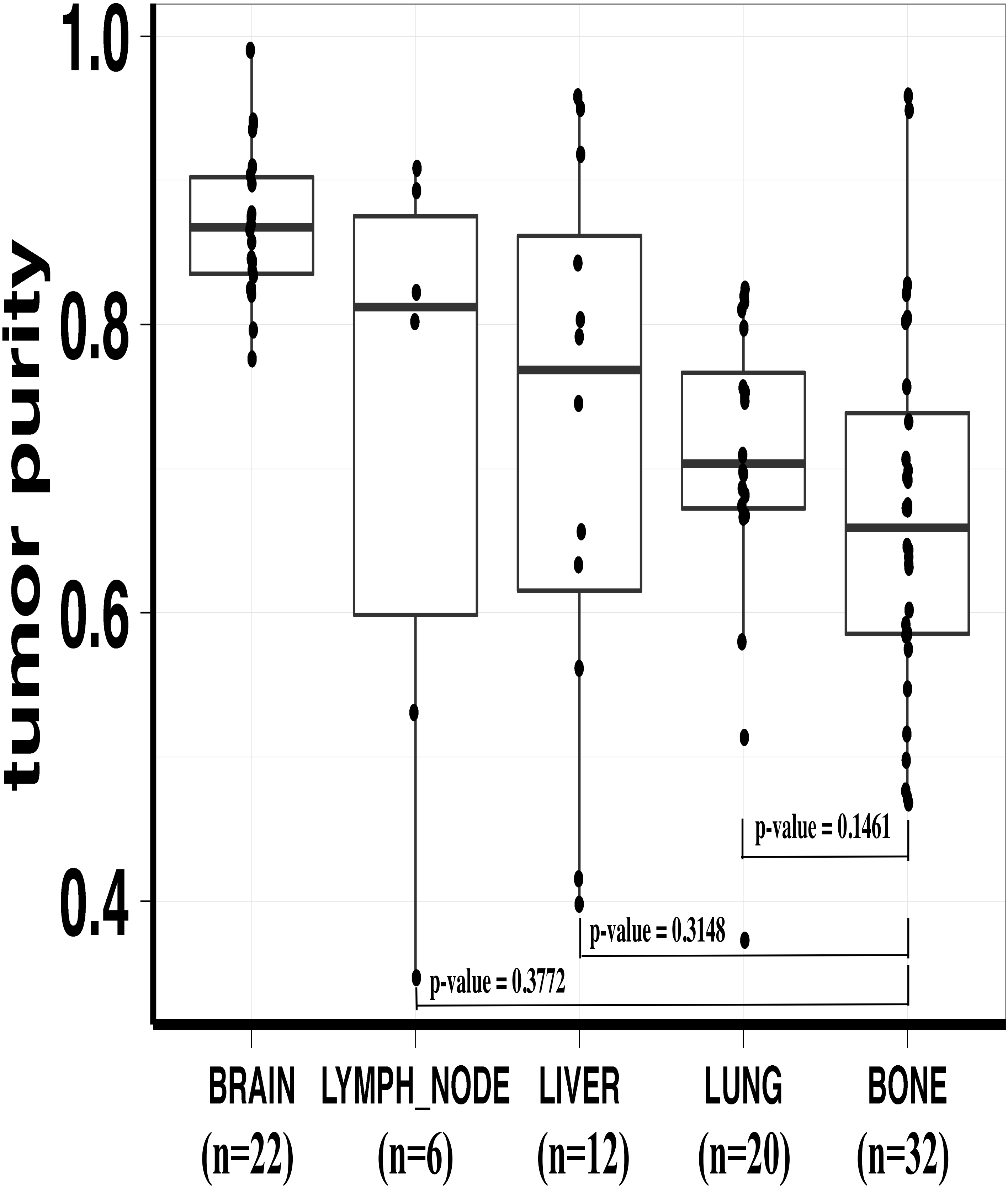
(a)
p-value=0.4007
correlation with
MDAMB415
0.55
0.50
0.45
0.40
0.35
0.30
Liver
(n=27)
Lymph node
(n=9)
(b)
dim2
20
0
−20
−20
−10
0
10
20
dim1
subtype
Basal−like
Her2−enriched
LuminalA
LuminalB
Normal−like
data.source
MET500
TCGA

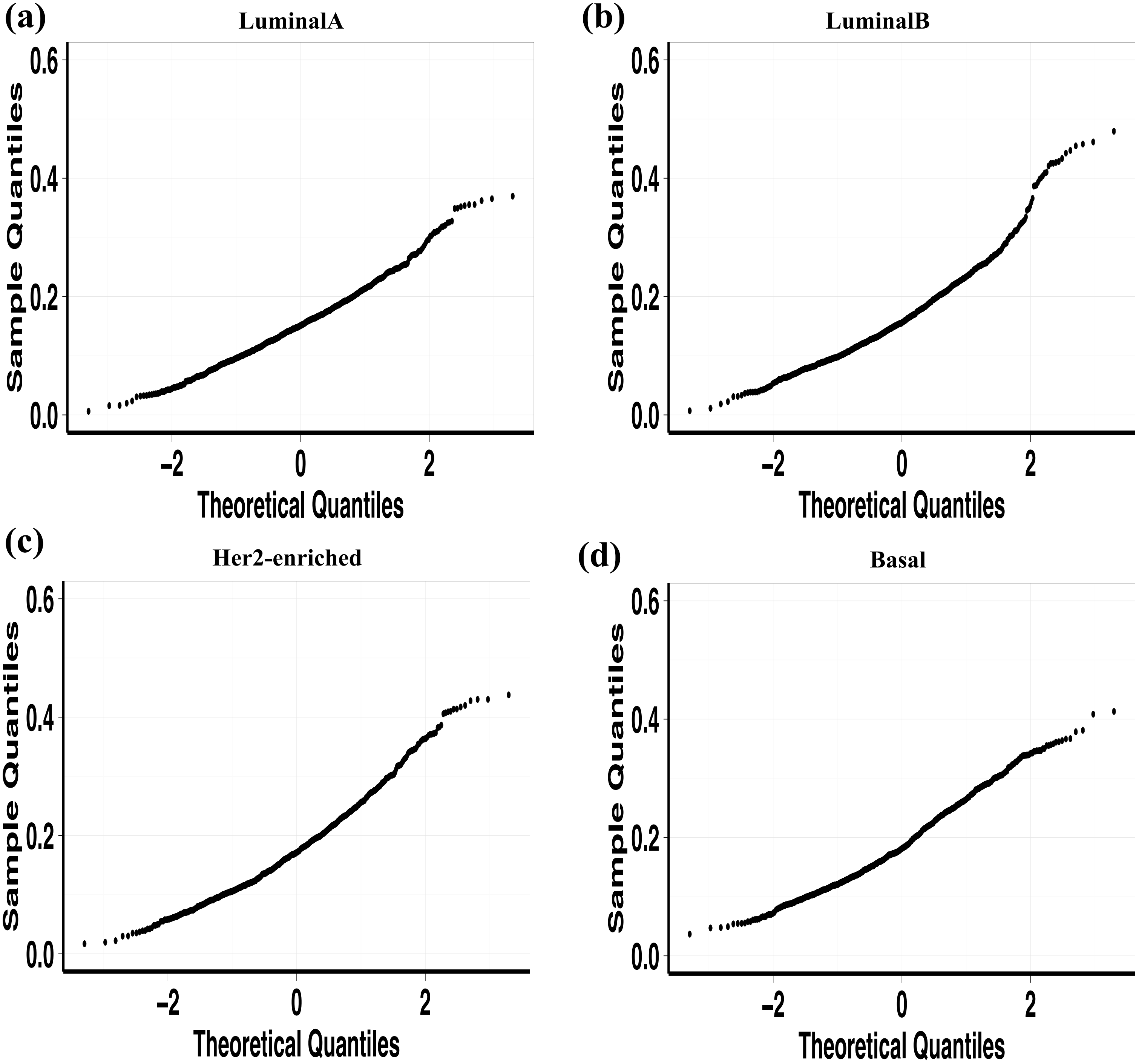

(a)
Liver
Spearman correlation = 0.91
microarray
0.3
0.4
0.5
0.1
0.2
0.3
0.4
MET500
Lymph node
Spearman correlation = 0.93
microarray
0.2
0.3
0.4
0.5
0.6
0.1
0.2
0.3
0.4
MET500
(b)
LuminalA
Spearman correlation = 0.62
microarray
0.30
0.35
0.40
0.45
0.1
0.2
0.3
MET500
LuminalB
Spearman correlation = 0.92
microarray
0.25
0.30
0.35
0.40
0.45
0.50
0.1
0.2
0.3
0.4
MET500
Her2-enriched
Spearman correlation = 0.88
microarray
0.3
0.4
0.5
0.1
0.2
0.3
0.4
MET500
Basal-like
Spearman correlation = 0.72
microarray
0.30
0.35
0.40
0.45
0.50
0.2
0.3
0.4
MET500

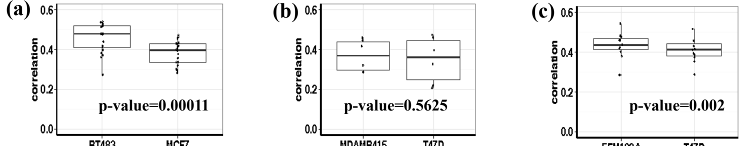


Bone
Liver
Lymph Node
Brain
Lung
0.11
0.062
0.22
0.34
0.95
0.92
0.78
0.89
0.73
0.86
0.0 0.1 0.2 0.3 0.4 0.5 0.6
-1
1


Basal–like
0.32
0.47
0.40
LuminalB
0.82
0.95
LuminalA
0.80
Her2
enriched
0.0
0.1
0.2
0.3
0.4
0.5
0.6
-1
1


tumor purity
1.0
0.8
0.6
0.4
p-value = 0.1461
p-value = 0.3148
p-value = 0.3772
BRAIN (n=22)
LYMPH_NODE (n=6)
LIVER (n=12)
LUNG (n=20)
BONE (n=32)

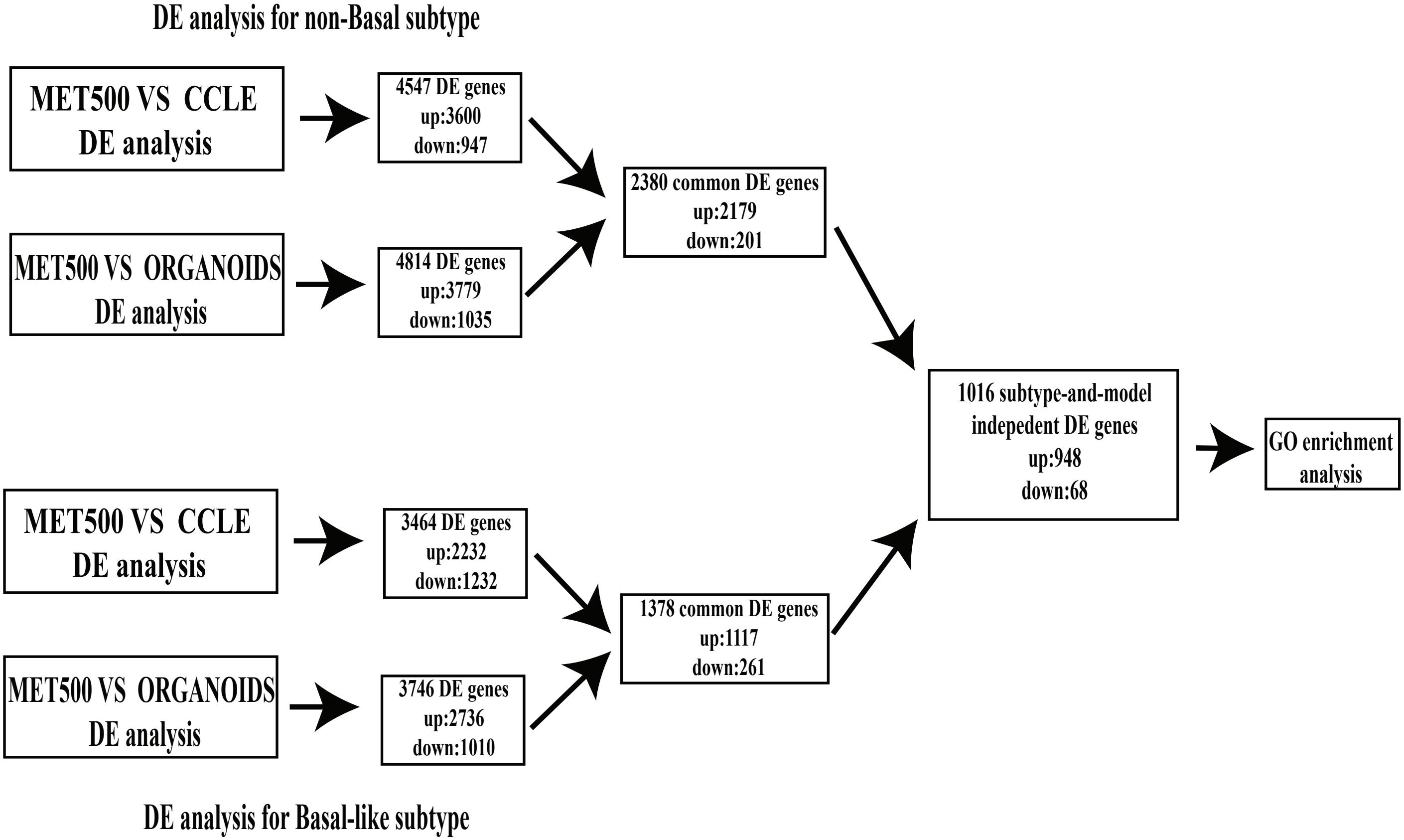

(a)
LuminalA
(b)
LuminalB
(c)
Her2-enriched
(d)
Basal
Sample Quantiles
Theoretical Quantiles
0.0
0.2
0.4
0.6
−2
0
2

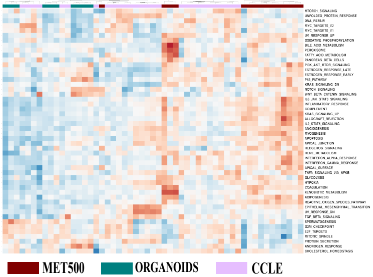


(LuminalB, n=20)

(LuminalA, n=6)

(Her2-enriched, n=11)


DE analysis for non-Basal subtype
MET500 VS CCLE DE analysis
4547 DE genes up:3600 down:947
MET500 VS ORGANOIDS DE analysis
4814 DE genes up:3779 down:1035
2380 common DE genes up:2179 down:201
1016 subtype-and-model indepedent DE genes up:948 down:68
GO enrichment analysis
MET500 VS CCLE DE analysis
3464 DE genes up:2232 down:1232
MET500 VS ORGANOIDS DE analysis
3746 DE genes up:2736 down:1010
1378 common DE genes up:1117 down:261
DE analysis for Basal-like subtype


MTORC1_SIGNALING
UNFOLDED_PROTEIN_RESPONSE
DNA_REPAIR
MYC_TARGETS_V2
MYC_TARGETS_V1
UV_RESPONSE_UP
OXIDATIVE_PHOSPHORYLATION
BILE_ACID_METABOLISM
PEROXISOME
FATTY_ACID_METABOLISM
PANCREAS_BETA_CELLS
PI3K_AKT_MTOR_SIGNALING
ESTROGEN_RESPONSE_LATE
ESTROGEN_RESPONSE_EARLY
P53_PATHWAY
KRAS_SIGNALING_DN
NOTCH_SIGNALING
WNT_BETA_CATENIN_SIGNALING
IL6_JAK_STAT3_SIGNALING
INFLAMMATORY_RESPONSE
COMPLEMENT
KRAS_SIGNALING_UP
ALLOGRAFT_REJECTION
IL2_STAT5_SIGNALING
ANGIOGENESIS
MYOGENESIS
APOPTOSIS
APICAL_JUNCTION
HEDGEHOG_SIGNALING
HEME_METABOLISM
INTERFERON_ALPHA_RESPONSE
INTERFERON_GAMMA_RESPONSE
APICAL_SURFACE
TNFA_SIGNALING_VIA_NFKB
GLYCOLYSIS
HYPOXIA
COAGULATION
XENOBIOTIC_METABOLISM
ADIPOGENESIS
REACTIVE_OXIGEN_SPECIES_PATHWAY
EPITHELIAL_MESENCHYMAL_TRANSITION
UV_RESPONSE_DN
TGF_BETA_SIGNALING
SPERMATOGENESIS
G2M_CHECKPOINT
E2F_TARGETS
MITOTIC_SPINDLE
PROTEIN_SECRETION
ANDROGEN_RESPONSE
CHOLESTEROL_HOMEOSTASIS
MET500
ORGANOIDS
CCLE