



OPEN Transcriptomics indicate nuclear division and cell adhesion not recapitulated in MCF7 and MCF10A compared to luminal A breast tumours

Jeremy Joon Ho Goh^{1,2}, Corinna Jie Hui Goh¹, Qian Wei Lim², Songjing Zhang², Cheng-Gee Koh² & Keng-Hwee Chiam^{1,2}✉

Breast cancer (BC) cell lines are useful experimental models to understand cancer biology. Yet, their relevance to modelling cancer remains unclear. To better understand the tumour-modelling efficacy of cell lines, we performed RNA-seq analyses on a combined dataset of 2D and 3D cultures of tumourigenic MCF7 and non-tumourigenic MCF10A. To our knowledge, this was the first RNA-seq dataset comprising of 2D and 3D cultures of MCF7 and MCF10A within the same experiment, which facilitates the elucidation of differences between MCF7 and MCF10A across culture types. We compared the genes and gene sets distinguishing MCF7 from MCF10A against separate RNA-seq analyses of clinical luminal A (LumA) and normal samples from the TCGA-BRCA dataset. Among the 1031 cancer-related genes distinguishing LumA from normal samples, only 5.1% and 15.7% of these genes also distinguished MCF7 from MCF10A in 2D and 3D cultures respectively, suggesting that different genes drive cancer-related differences in cell lines compared to clinical BC. Unlike LumA tumours which showed increased nuclear division-related gene expression compared to normal tissue, nuclear division-related gene expression in MCF7 was similar to MCF10A. Moreover, although LumA tumours had similar cell adhesion-related gene expression compared to normal tissues, MCF7 showed reduced cell adhesion-related gene expression compared to MCF10A. These findings suggest that MCF7 and MCF10A cell lines were limited in their ability to model cancer-related processes in clinical LumA tumours.

Abbreviations

2D	Two-dimensional
3D	Three-dimensional
ARCHS4	All RNA-seq and ChIP-seq sample and signature search
BC	Breast cancer
CPM	Counts per million
FDR	False discovery rate
ER	Estrogen receptor
GDC	Genomic data commons
GEO	Gene expression omnibus
GO	Gene Ontology
LFC	Log ₂ -fold change
LumA	Luminal A
ORA	Overrepresentation analysis
PC	Principal component
PCA	Principal component analysis
PCA-UMAP	UMAP projection of principal components

¹Bioinformatics Institute, 30 Biopolis Street, Singapore 138671, Singapore. ²School of Biological Sciences, Nanyang Technological University, Singapore 637551, Singapore. ✉email: chiamkh@bii.a-star.edu.sg

RFC	Random forest classifier
SMOTE	Synthetic Minority Oversampling Technique
TCGA	The Cancer Genome Atlas
TCGA-BRCA	The Cancer Genome Atlas Breast Invasive Carcinoma
TMM	Trimmed mean of M-values
UMAP	Uniform Manifold Approximation and Projection

Breast cancer (BC) is the most common cancer diagnosed in females, representing 11.7% of global cancer diagnoses with over 2.2 million new BC diagnoses and 680,000 BC-related deaths in 2020^{1,2}. Given the concerning global epidemiology of BC, with a projected doubling in incidence to 4.4 million by 2070³, research is needed to better understand and treat BC. However, clinical samples are often challenging to obtain and experimentally manipulate. Cell lines are crucial to facilitating scientific experimentation as a limitless source of biological material⁴. They are key experimental models used in exploring cancer biology⁵ and evaluating drugs^{6,7}.

MCF10A and MCF7 are the most frequently used cell line models for normal breast tissue and BC tumours respectively^{8,9}. MCF10A is a non-tumourigenic human breast epithelial cell line derived from benign proliferative breast tissue, characterised by a lack of estrogen receptor (ER) expression¹⁰. MCF7 is a widely used in vitro model established from pleural effusion samples from a patient with metastatic BC¹¹. It belongs to the Luminal A (LumA) BC subtype^{5,12–14}, an ER-positive subtype with the most optimistic prognoses among the BC subtypes¹⁵.

Research gaps. Past studies on the clinical relevance of cell lines suggest mixed conclusions which require resolution. A compilation of high-throughput data across hundreds of cell lines from the Cancer Cell Line Encyclopaedia found that cell lines had transcriptional profiles which correlated better with that of their corresponding primary tumours than other tumours, concluding that cell lines represented primary tumours well^{16,17}. Within BC, cell lines were found to have expression patterns which correlated with their respective intrinsic subtypes of BC^{5,18,19}. However, other studies suggested poor concordance between transcriptomes of cell lines and primary tumours^{17,20,21}, and where cell lines correlated more with unrelated cell lines from different cancers than their corresponding primary tumours²². These mixed findings call to question the effectiveness of using MCF7 as a model for BC. For MCF10A, two-dimensional (2D) and three-dimensional (3D) cultures of MCF10A were found to express markers which were uncharacteristic of normal human breast tissue⁹, suggesting that MCF10A may be an inappropriate breast tissue model.

We compiled a summary of reported functional differences between cell lines and primary tumours, derived from differential expression analyses between cell line and tumour transcriptomes (Table 1). These studies suggest that cell lines regulated cellular processes differently from primary tumours. Compared to tumours, cell lines displayed upregulation of cell cycle and metabolism, and showed decreased immune, cell adhesion and tissue organisation processes. Moreover, these studies focused on direct differences between cancer cell lines and primary tumours, instead of considering whether the same cancer-related processes could be observed between tumourigenic and non-tumourigenic cell lines, as between primary tumours and non-cancerous tissues. Cancer-invariant gene expression differences between cell lines and primary tumours may be irrelevant to cancer-modelling and should be omitted in the evaluation of cell lines. Additionally, few studies directly validated 3D culture systems relative to primary tumours. 3D cultures are superior to 2D cultures because they retain elements lost in 2D cultures, such as complex cell polarity and lumen formation²³. 3D cultures of MCF7 result in spheroidal microtissues with luminal-like morphology and express more breast-specific biomarkers compared to 2D cultures^{24,25}. Similarly, 3D cultures of MCF10A cells form spheroids with a hollow lumen²⁶, introducing phenotypes that would not have been present in 2D cultures⁹. However, few studies compared gene expression of cell lines in 3D cultures to primary tumours. Given the costs of 3D cultures²⁷, there is a need to assess its benefits in modelling BC over the less expensive 2D culture systems so that experimenters can better justify switching from 3D culture systems.

Aims and overview. In this study, we validated transcriptomic data of MCF7 and MCF10A cell lines in 2D and 3D cultures, augmented with previously published experimental data from all RNA-seq and ChIP-seq sample and signature search (ARCHS4) repository²⁸, against clinical transcriptomic data of LumA BC and normal breast samples in The Cancer Genome Atlas (TCGA).

We aimed to identify the key cancer-related processes that differ between the cell line analyses (MCF7-vs-MCF10A) and analysis of TCGA clinical samples (LumA-vs-normal), to understand the adequacy of cell lines as models for LumA BC. Unlike past methods which directly considered the differences between tumours and cancer cell lines, our method considers the differences-of-differences between cell lines and clinical BC (Fig. 1). That is, we first identified cancer-related differences via separate analyses of 2D MCF7-vs-MCF10A, 3D MCF7-vs-MCF10A and TCGA LumA-vs-normal, by using a random forest classifier (RFC)-based gene selection approach to select cancer-related genes, and identifying overrepresented biological processes in the selected genes. Thereafter, we compared these overrepresented processes across analyses, to determine how well the comparisons of MCF7-vs-MCF10A were able to represent cancer-related processes in the clinical LumA-vs-normal comparison. We found that analyses of MCF7-vs-MCF10A failed to identify cancer-related changes in mitotic nuclear division, and exaggerated cancer-related dysregulation of cell adhesion compared to clinical LumA-vs-normal. We concluded that MCF7 and MCF10A were limited in their ability to model LumA BC.

Processes differentially regulated in cell lines compared to primary tumours	Direction	Scope ^a	References
Cell cycle and proliferation	Upregulated	General tumours, BC	19,20,29,30
Nucleotide metabolism and RNA production	Upregulated	General tumours	29
Glycolysis and energy metabolism	Upregulated	General tumours, BC	19,20
Cell communication	Both directions	General tumours	29
Immune processes and inflammation	Downregulated	General tumours, metastatic BC	18,30
Cell adhesion	Downregulated	General tumours	20,29
Tissue organisation	Downregulated	General tumours	20

Table 1. Brief summary of differentially regulated processes between cell lines and primary tumours. ^aThe scope refers to whether the studies focused on BC data, or looked at cancers in general.

Materials and methods

Data pre-processing and overrepresentation analyses (ORA) were conducted in R version 4.1.2³¹ using RStudio version 2021.9.2.382³². Further processing and feature selection were implemented in Python version 3.9.7 using Jupyter Notebook³³. Where applicable, we specified a random seed of 77 for reproducible analyses. Unless otherwise specified, visualisations were produced with the R package *ggplot2* version 3.3.5³⁴, or the Python packages *seaborn* version 0.11.2³⁵ and *matplotlib* version 3.4.3³⁶.

Our differences-of-differences approach first finds the differences between Luma (or MCF7) and normal (or MCF10A) samples via three separate analyses of the TCGA, 3D culture and 2D culture datasets (Fig. 2a; Steps 1–4), and then compares the significantly overrepresented gene sets across the three analyses (Fig. 2a; Step 5).

Cell lines and monolayer cultures. The ATCC-originated MCF7 cells (HTB-22) were maintained as monolayer in Dulbecco's Modified Eagle's Medium—high glucose (Sigma-Aldrich, USA) supplemented with 10% Fetal Bovine Serum (Gibco, USA), 1% Penicillin/Streptomycin (Gibco, USA). The ATCC-originated MCF10A cells (CRL-10317) were maintained in Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 (D-MEM/F-12) (Gibco, USA) supplemented with 5% horse serum (Gibco, USA), 20 ng/ml Epidermal Growth Factor (PeproTech, USA), 0.5 mg/ml hydrocortisone (Sigma-Aldrich, USA), 10 µg/ml insulin (Sigma-Aldrich, USA), 100 nM (–)-Isoproterenol hydrochloride (Sigma-Aldrich, USA) and 1% Penicillin/Streptomycin (Gibco, USA). All cells were passaged or changed with fresh medium every other day, and grown in 37 °C humidified incubator supplemented with 5% CO₂.

Three-dimensional spheroid formation. The spheroids were generated using the 3D 'on-top' methods as described previously³⁷. Briefly, 8-well chamber slide (Ibidi, Germany) were coated with 50 µl of growth factor-reduced Matrigel (Corning, USA) and left to solidify in the incubator for at least 30 min. MCF7 and MCF10A cells (6500 and 5000 cells respectively) were seeded on top of the solidified Matrigel coating and overlaid with cell medium supplemented with 2% Matrigel (Corning, USA). MCF10A spheroids were cultured in assayed Dulbecco's Modified Eagle Medium: Nutrient Mixture F-12 (D-MEM/F-12) (Gibco, USA) medium supplemented with 2% horse serum (Gibco, USA), 5 ng/ml Epidermal Growth Factor (PeproTech, USA), 0.5 mg/ml hydrocortisone (Sigma-Aldrich, USA), 10 µg/ml insulin (Sigma-Aldrich, USA), 100 nM (–)-Isoproterenol hydrochloride (Sigma-Aldrich, USA) and 1% Penicillin/Streptomycin (Gibco, USA). Cells were refed with fresh assay medium supplemented with 2% Matrigel (Corning, USA) every two days thereafter. MCF7 spheroids were harvested on day 10, while MCF10A spheroids were harvested on day 16.

RNA extraction. Total RNA was extracted from MCF7 and MCF10A cells in both 2D and 3D cultures using RNeasy plus kits (Qiagen, Germany) according to manufacturer's instructions. Briefly, cells grown as monolayer were trypsinized and washed with PBS twice followed by RNA extraction using RNeasy Plus Kit. To collect spheroids, medium was removed from 3D cell culture followed by washing with PBS twice gently. Pre-chilled cell recovery solution (Corning, USA) at the volume of 200 µl were added to 8-well chamber slide (Ibidi, Germany). Matrigel matrix was fully depolymerized, and spheroids were released after 30 min incubation at 4 °C. Suspended spheroids were collected and washed with PBS twice followed by RNA extraction using RNeasy Plus Kit (Qiagen). The integrity and quantity of RNA were determined with NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific, USA) before sending out for RNA sequencing.

RNA-sequencing. RNA-sequencing was done in two batches with the first batch by BGI Group and the second batch by NovogeneAIT Genomics Singapore Pte Ltd. RNA samples were subjected to quality control (QC) after sample submission to the companies. Preliminary quantification of nucleic acid samples was performed using Qubit (Invitrogen, USA). RNA degradation and contamination were ruled out with agarose gel electrophoresis verification. RNA integrity and quantity were measured on 2100 Bioanalyzer (Agilent Technologies, USA). After sample QC, library construction was performed using NEBNext Ultra II RNA Library Prep Kit (Illumina, USA) according to manufacturer's protocols. For the first batch, the libraries were run with BGI group's in-house sequencing platform BGISEQ-500 (20 M single-end reads). For the second batch, libraries were run on NovaSeq PE150 (Illumina, USA) to produce 6 GB of raw data per sample (20 M paired-end reads).

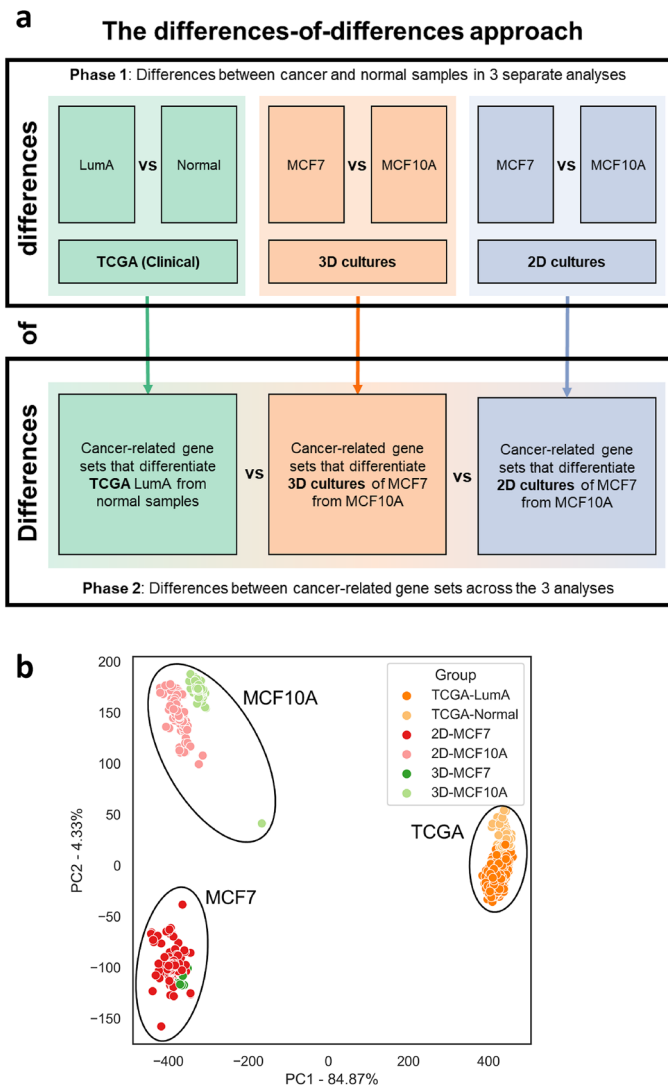


Figure 1. Overview for comparison of cancer progression in cell lines and clinical breast cancer samples. **(a)** Summary of the differences-of-differences approach. **(b)** PCA plot of the first two principal components of the combined dataset of 2D and 3D cultures of MCF7 and MCF10A, and TCGA clinical samples.

Preparation of datasets. Raw RNA-seq data comprised of 9 biological replicates per class (3D-MCF7, 3D-MCF10A, 2D-MCF7, 2D-MCF10A). Data from single-end and paired-end sequencing runs were recorded as separate batches. We harmonised the FASTQ file processing by applying the methods in the ARCHS4 documentation²⁸. Specifically, *Kallisto* version 0.46.0³⁸ with a precalculated human index from the ARCHS4 repository (<https://maayanlab.cloud/archs4/>, downloaded on 14 March 2022) was used to obtain transcript counts, which were then mapped to gene-level counts using *tximport* version 1.22.0³⁹ and the GrCH38.87 human genome annotation⁴⁰.

To improve the generalisability of our findings, we augmented our cell line data with ARCHS4 data. Raw RNA-seq counts from the ARCHS4 repository were downloaded on 21 May 2021²⁸, and a subset of MCF7 and MCF10A samples was manually selected based on metadata from the Gene Expression Omnibus (GEO). Only untransformed samples in control conditions were included for further analysis. This dataset was then combined with our RNA-seq data, and subset into 3D culture and 2D culture datasets.

For clinical RNA-seq data, level 3 unnormalised HTSeq mRNA counts data from the TCGA Breast Invasive Carcinoma (TCGA-BRCA) dataset for “primary tumour” and “solid tissue normal” samples were obtained from the Genomic Data Commons (GDC) on 17 January 2022 using *TCGAbiolinks* version 2.22.3⁴¹. Sample collection and processing protocols are detailed in Ref.⁴². The RNA-seq data had been processed according to GDC’s harmonised pipeline⁴³. We discarded formalin-fixed paraffin-embedded samples, as RNA crosslinking due to the fixation process might confound analyses of the higher quality fresh-frozen samples⁴⁴, and discarded non-female samples as male BC are characterised by distinct molecular profiles from female BC⁴⁵. Metadata from⁴⁶ was used for BC subtype assignment. Given that MCF7 is a LumA cell line¹² and MCF10A is a non-tumourigenic breast

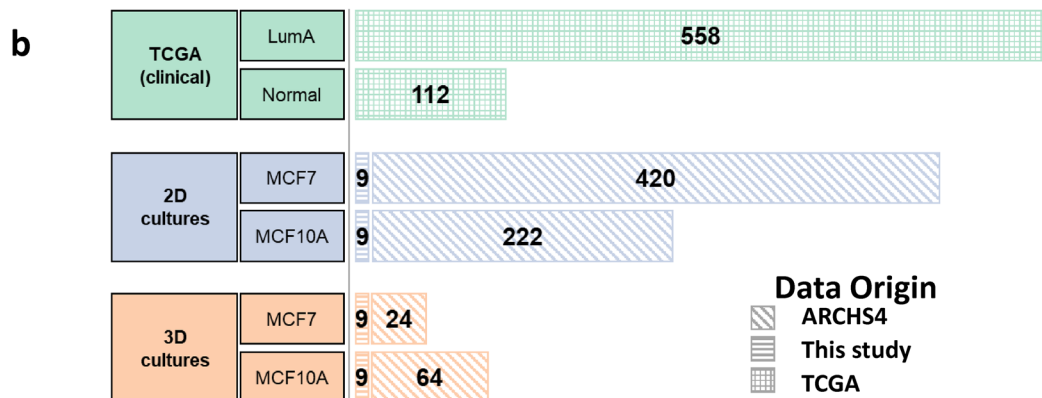
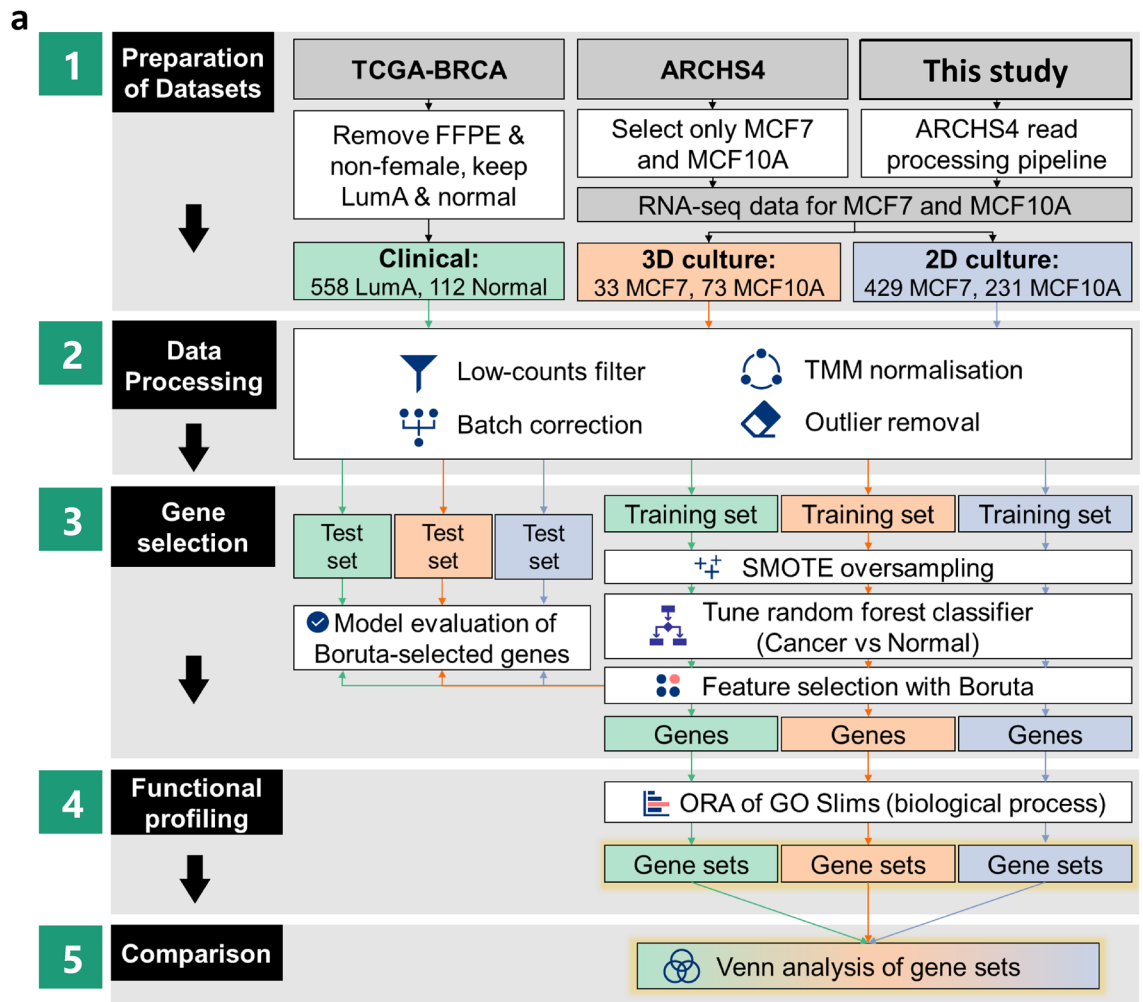


Figure 2. Overall analysis workflow of RNA-seq data. (a) Summary of key steps in the analysis workflow. (Refer to “Preparation of datasets” in “Materials and methods” section). *FFPE* Formalin-fixed paraffin-embedded, *GO* Gene Ontology, *TCGA-BRCA* The Cancer Genome Atlas-Breast Invasive Carcinoma dataset, *ORA* Overrepresentation analysis. (b) Number of samples by class in the TCGA, 2D culture and 3D culture datasets, respectively. The origins of each dataset are indicated by the textures of the bars.

cell line¹⁰, only LumA tumours and normal samples were retained. The number of samples by class in the three datasets are summarized in Fig. 2b.

Data processing: filtering, normalisation, batch correction, and outlier removal. On each dataset, we performed low-counts filtering, normalised counts using the trimmed mean of M-values (TMM) method⁴⁷, and log₂-transformed the counts per million (CPM) values using *EdgeR* version 3.36.0⁴⁸, with default settings.

To visualise data structure, we first performed principal component analysis (PCA) on log-CPM data using *scikit-learn* version 1.0.2⁴⁹. We selected the top principal components (PCs) which cumulatively account for at least 80% of total variation as input into the Uniform Manifold Approximation and Projection (UMAP) algorithm⁵⁰. Selecting only the top PCs accounting for majority of variation helps suppress noise⁵¹, allowing better projection of the relationships between samples. We used the Python package *UMAP* version 0.5.2 with 50 to 200 neighbours, a minimum distance of between 0.05 and 0.3, and random initial embedding positions, to visualise sample relationships. To assess the presence of batch effects and outliers, we generated boxplots by batch along one-dimensional UMAP projections of the PCs (PCA-UMAP) and two-dimensional PCA-UMAP scatterplots. Absolute UMAP values were not provided as they were not intrinsically meaningful. Relative positions of points in UMAP space were more relevant.

One-dimensional PCA-UMAP boxplots of TMM-normalised log-CPM suggested mild batch effects in the TCGA dataset as relative distributions of samples within each class were not homogeneous across plates (Fig. S2a), and substantial batch effects in the 2D and 3D culture datasets because samples tended to cluster by batch without distinct separations between MCF7 and MCF10A samples (Figs. S3a, S4a). Hence, we applied batch correction. Plate IDs (defining separate sequencing runs)⁵² and GEO series IDs (defining separate experiments) were used to specify batch variables for the TCGA and ARCHS4 data respectively. We used the *removeBatchEffect* function in *limma* version 3.50.0⁵³ on log-CPM values, specifying the batch variable as described, and LumA (or MCF7) against normal (or MCF10A) samples as the grouping variable.

Consequent to batch correction, distributions across batches highlighted clear differences between LumA (or MCF7) and normal (or MCF10A) samples (Figs. S2b, S3b, S4b), suggesting that most systematic variation due to technical, cancer-irrelevant differences were removed. However, there were 2 LumA outliers in the TCGA dataset (Fig. S2b) and 12 MCF7 outliers (from GSE91395) in the 2D culture dataset (Fig. S3c) which clustered with non-cancer (or MCF10A) samples. These outliers were removed, and the data processing pipeline described above was rerun on the outlier-removed data.

To understand sample relationships across the three datasets, we combined the outlier-removed unnormalised data and kept genes which were present in all three datasets. We then performed low-counts filtering, normalisation and batch correction as described above. We visualised this combined dataset with a PCA plot instead of using PCA-UMAP because the first two PCs had already accounted for over 80% of variation in this dataset.

Gene selection. To each of the three separate datasets, we applied a random stratified train-test split with a test set size of 20%. Only the training set was used for model fitting.

The number of samples in each class were unequal in all three training datasets (Fig. S1a). This was problematic as class-imbalanced training sets produce classifiers which return biased predictions favouring the majority class⁵⁴. To correct for class imbalance, we applied synthetic minority oversampling technique (SMOTE) on the training sets, implemented in *imbalanced-learn* version 0.9.0⁵⁵, to generate synthetic samples for the minority class based on existing samples⁵⁶. Briefly, SMOTE randomly selected a sample, randomly selected a neighbouring sample out of a specified number (30% of the minority class size) of nearest neighbours in feature space, and generated a random synthetic point between the two samples. This was repeated until the minority class was equivalent in size to the majority class. Two-dimensional PCA-UMAP plots were generated before and after SMOTE to ensure that synthetic datapoints retained data structure.

For each analysis, we tuned a RFC⁵⁷ implemented in *scikit-learn* with 5000 trees using the *GridSearchCV* function, with a pre-specified set of hyperparameters (Supplementary Table 1). In tuning, accuracy was evaluated using a modified tenfold cross-validation, where only the training sets in each fold were oversampled with SMOTE.

To select key genes involved in distinguishing cancer from normal samples, we used *BorutaPy* version 0.1.5, a Python implementation of Boruta⁵⁸, on the SMOTE-oversampled training set and tuned RFC. The implementation entails first generating a set of shadow features by shuffling expression values for each gene. A RFC was fitted on the genes and shadow features to compute Gini importance scores, which quantify how well the given feature classified the samples⁵⁹. Based on whether a gene had a higher importance score than the most important shadow feature, a two-tailed binomial test was performed at each of 1000 successive iterations to determine whether the gene was significantly important (Bonferroni-adjusted p -value < 0.0001) in classifying LumA (or MCF7) and normal (or MCF10A) samples. To ascertain that the Boruta-selected genes were cancer-related, we trained the tuned RFC on the Boruta-selected genes from the oversampled training set and used this to obtain an accuracy score in the test set.

Functional profiling. ORA was performed on the unordered lists of Boruta-selected genes for each of the three analyses using *gprofiler2* version 0.2.1⁶⁰. We used the gene ontology (GO) biological process gene sets from the generic GO slims provided by The Gene Ontology Consortium⁶¹, downloaded on 11 April 2022. For each analysis, the respective list of genes passing low-counts filtering was used as the background gene list. Gene sets with a Benjamini–Hochberg false discovery rate (FDR) below 0.01 were considered significant.

To determine direction of regulation of the significant gene sets, we first used *limma* version 3.50.0⁵³ with *voom* transformation⁶² to compute log₂-fold changes (LFCs) between LumA (or MCF7) and normal (or MCF10A) samples in each dataset, with batch as a covariate. LFCs were input into a pre-ranked gene set enrichment analysis⁶³ implemented in *clusterProfiler* version 4.2.2⁶⁴, using the “DOSE” method with 1000 permutations, without limiting gene set size. We only considered the ORA-significant gene sets. A positive normalised enrichment score meant that the gene set was upregulated, and a negative normalised enrichment score meant that it was downregulated.

Analysis	Number of genes passing low-counts filter	Number of Boruta-selected genes
TCGA LumA-vs-normal	24,657	1031
2D MCF7-vs-MCF10A	15,441	450
3D MCF7-vs-MCF10A	15,001	2007

Table 2. Number of genes (out of those passing the low-counts filter) selected by Boruta.

Differences-of-differences: comparison across the three analyses. We compared the results of each analysis by visualising the number of Boruta-selected genes and significant gene sets identified in each analysis using *VennDiagram* version 1.7.1⁶⁵. We selected the top gene sets (ranked by FDR) occurring in the LumA-vs-normal analysis only and the MCF7-vs-MCF10A analyses only for further discussion. Finally, we listed the top genes (ranked by absolute LFC) associated with the selected gene sets, as well as genes which were associated with common gene sets and selected by Boruta in both cell line and TCGA datasets.

Results

We took a differences-of-differences approach to assess the value of MCF7 and MCF10A cell lines as models for BC (Fig. 1). Contrary to direct comparisons between cell lines and tumours in past studies^{19,30}, the key benefit of this approach is the focus on cancer-related processes which are more relevant for tumour-modelling. Comparing cell lines to tumours does not reveal insights about the cancer-modelling ability of cell lines because cancer-related gene expression differences cannot be divorced from general culture-related but cancer-invariant differences. Our comparisons of tumourigenic MCF7 against non-tumourigenic MCF10A would ensure that the identified processes were cancer-related, hence providing a commentary on whether the cell lines model cancer adequately.

Clinical LumA and normal breast samples had different gene expression profiles from MCF7 and MCF10A. PCA plots of the combined dataset (Fig. 1b) revealed that clinical TCGA samples separated from the cell line samples along the first PC, accounting for 84.87% of variation in the dataset. Hence, majority of variation was driven by differences in gene expression between clinical samples and cell lines, raising doubts about the ability of cell lines to model clinical BC. LumA (or MCF7) and non-cancer (or MCF10A) samples separated along the second PC, suggesting that a minor but substantive amount of variation in the dataset was driven by cancer-related differences. To specifically consider whether these cancer-related differences for cell lines were similar to clinical samples independent of variation driven by broad differences between clinical and cell line samples, we performed separate analyses for TCGA LumA-vs-normal, 3D MCF7-vs-MCF-10A and 2D-MCF7-vs-MCF10A.

Whether in cell cultures or clinical samples, there were clear cancer-related differences in gene expression. Consistent with the second PC in Fig. 1b, PCA-UMAP projections of the training sets (for three separate analyses) demonstrated that LumA (or MCF7) samples clustered separately from normal (or MCF10A) samples in all three datasets both before and after SMOTE (Fig. S1b). This suggests that majority of variation in the three separate datasets was driven by cancer-related differences in gene expression, and that the oversampled data was able to preserve these cancer-related differences between tumour and normal samples. Hence, genes selected downstream based on these datasets would clearly reflect key cancer-related differences.

Different genes were driving cancer-related differences between LumA-vs-normal compared to MCF7-vs-MCF10A. We used an RFC-based feature selection approach (i.e., Boruta) to select key cancer-related genes important for classifying cancer from non-cancer samples. Unlike typical differential expression analysis methods which perform independent gene-wise significance tests, RFC-based gene selection methods can demonstrate greater power in large RNA-seq datasets and account for relationships between genes by selecting genes in a multivariate manner⁶⁶. This is especially important given canonical knowledge that cancers involve interactions between genes and pathways^{67,68}. RFCs were tuned separately for each analysis, and default settings with 5000 trees produced optimal classifiers. A larger number of trees was chosen to ensure a more robust RFC for gene selection. Cross-validation of the optimal RFCs returned a mean accuracy of 0.998 for the TCGA dataset, and perfect accuracy for both the 2D and 3D culture datasets. This suggested that the fitted RFCs successfully learned the gene expression profiles that distinguished LumA (or MCF7) from normal (or MCF10A) samples.

Unlike other minimal-optimal feature selection algorithms which find the smallest feature subset that maximises classifier performance, Boruta's all-relevant approach identifies all genes relevant to the classification of cancer from normal samples⁵⁸, allowing a better understanding of the mechanistic underpinnings of cancer. Table 2 lists the number of genes selected by Boruta in each analysis. We validated the classification accuracy of these genes on the test set. Using only the subset of Boruta-selected genes in the tuned RFC fitted on the oversampled training data, all test sets were predicted with 100% accuracy. Given the perfect accuracy in separating cancer (or MCF7) from normal (or MCF10A) samples, we concluded that the chosen genes in each analysis were important cancer-related genes.

If MCF7 and MCF10A are good models, then they should be distinguished by the same set of cancer-related genes as between LumA BC and normal samples. However, Venn analysis revealed that only 5.1% of genes in the LumA-vs-normal analysis were important in the 2D MCF7-vs-MCF10A analysis (Fig. S5a), and a slightly

larger 15.7% of genes in the LumA-vs-normal analysis were important in the 3D MCF7-vs-MCF10A analysis (Fig. S5b). This demonstrates that the genes distinguishing MCF7 from MCF10A were inconsistent with the genes distinguishing clinical LumA from normal samples.

2D MCF7-vs-MCF10A had no incremental value in modelling cancer-related processes over 3D MCF7-vs-MCF10A.

To understand the biological processes underlying the Boruta-selected genes, we performed ORA within each of the three analyses to identify significant GO slims (FDR < 0.01), and then compared these significant gene sets. Our study focused on gene set-level comparisons because investigations using gene sets are more robust than studies based on genes⁶⁹. We used the generic GO slim, a high-level GO subset intended to give a broad overview of the biological processes involved⁶¹, because our goal was to summarise the key biological processes differentiating LumA (or MCF7) and normal (or MCF10A) samples. Moreover, using the GO slim reduced redundancy in the gene sets, facilitating comparison across the three analyses.

Venn analysis (Fig. 3) of significant gene sets indicated that all cancer-related processes in the 2D MCF7-vs-MCF10A analysis were captured in the 3D MCF7-vs-MCF10A analysis. Both the 3D and 2D MCF7-vs-MCF10A analyses were able to successfully identify 5 concordant gene sets with the LumA-vs-normal analysis, including dysregulation of processes related to anatomical structure development and cell differentiation (Fig. 3c). Beyond 2D MCF7-vs-MCF10A, the 3D MCF7-vs-MCF10A analysis was able to successfully identify dysregulation of 4 additional processes present in the LumA-vs-normal analysis (Fig. 3b). Hence, 3D MCF7-vs-MCF10A showed greater consistency in cancer-related processes with clinical LumA-vs-normal, than 2D MCF7-vs-MCF10A with LumA-vs-normal.

Cell cultures were limited in their ability to model cancer-related processes. We used two criteria to assess the ability of MCF7-vs-MCF10A to model LumA BC at the gene set level. A good model of clinical LumA BC (1) should not miss overrepresentations of cancer-related processes which were found in LumA-vs-normal and (2) should not identify spurious processes which were absent in LumA-vs-normal. Despite similarities in gene sets to the LumA-vs-normal analyses, we found gene sets along these criteria.

MCF7-vs-MCF10A did not model cancer-related processes like mitotic nuclear division which were present in clinical LumA-vs-normal. Among the 10 cancer-related gene sets identified in the LumA-vs-normal analysis, the 3D MCF7-vs-MCF10A analysis failed to represent the upregulation of mitotic nuclear division (Fig. 3a). The 2D MCF7-vs-MCF10A analysis failed to represent 6 processes, including cancer-related dysregulation of mitotic nuclear division and nervous system process (Fig. 3a,b). This finding suggests that cell lines, especially in 2D cultures, missed important cancer-related processes. The cancer-related upregulation of mitotic nuclear division was the top process missed by both 2D and 3D cultures.

MCF7-vs-MCF10A overstated dysregulation of processes like cell adhesion which were absent in clinical LumA-vs-normal. The 2D MCF7-vs-MCF10A analysis showed dysregulation of 2 additional processes, including cell adhesion, which was not present in the LumA-vs-normal analysis (Fig. 3d). However, the 3D MCF7-vs-MCF10A analysis showed dysregulation of 12 additional processes which were absent in the LumA-vs-normal analysis, including cell adhesion and inflammatory response (Fig. 3d,e). This finding suggests that cell lines, particularly from 3D cultures, tended to overstate the dysregulation of numerous processes. Specifically, the downregulation of cell adhesion was the top process overstated by both 2D and 3D cultures.

Finally, we listed the top 10 Boruta-selected genes characterising mitotic nuclear division and cell adhesion (Fig. 4), to explore gene-level differences and propose future experiments such as gene knockdown protocols to improve the modelling ability of MCF7 and MCF10A.

Common cancer-related genes across cell lines and clinical samples present opportunities for future cell line-based experimental study of LumA BC.

Beyond differences in gene sets, our analyses presented opportunities for future experimental research. In our comparison between cell lines and clinical samples, we further identified 37 genes (Fig. 5a) which were important in distinguishing cancer from normal samples (i.e., selected by Boruta) in both 2D and 3D cultures of MCF7-vs-MCF10A, and clinical samples. Among these 37 genes, we selected 20 genes (Fig. 5b) which were involved in overrepresentation of biological processes dysregulated across the three analyses (Fig. 3b,c). We sorted these genes by the average of the LFC values across the three analyses.

Discussion

Our findings suggest that MCF7 and MCF10A were limited in their ability to model LumA BC-related processes, cautioning experimenters about the limited generalisability of these cell line models to real BC. We found low similarity in Boruta-selected genes between cell lines and clinical samples, which suggests that a unique and clinically irrelevant set of genes distinguished MCF7 from MCF10A, compared to the cancer-related genes distinguishing clinical LumA BC from normal breast tissue. Consequently, even when using 3D cultures of MCF7 and MCF10A, researchers may be unable to identify the true set of cancer-relevant genes, and experimental manipulations may end up influencing a set of genes unrelated to clinical BC, leading to wasted time and resources.

Cancer-related dysregulation of structure development and cell differentiation were effectively modelled in MCF7-vs-MCF10A. However, a further investigation of biological processes identified in each analysis revealed a less pessimistic outlook. Despite the limited similarity in Boruta-selected genes

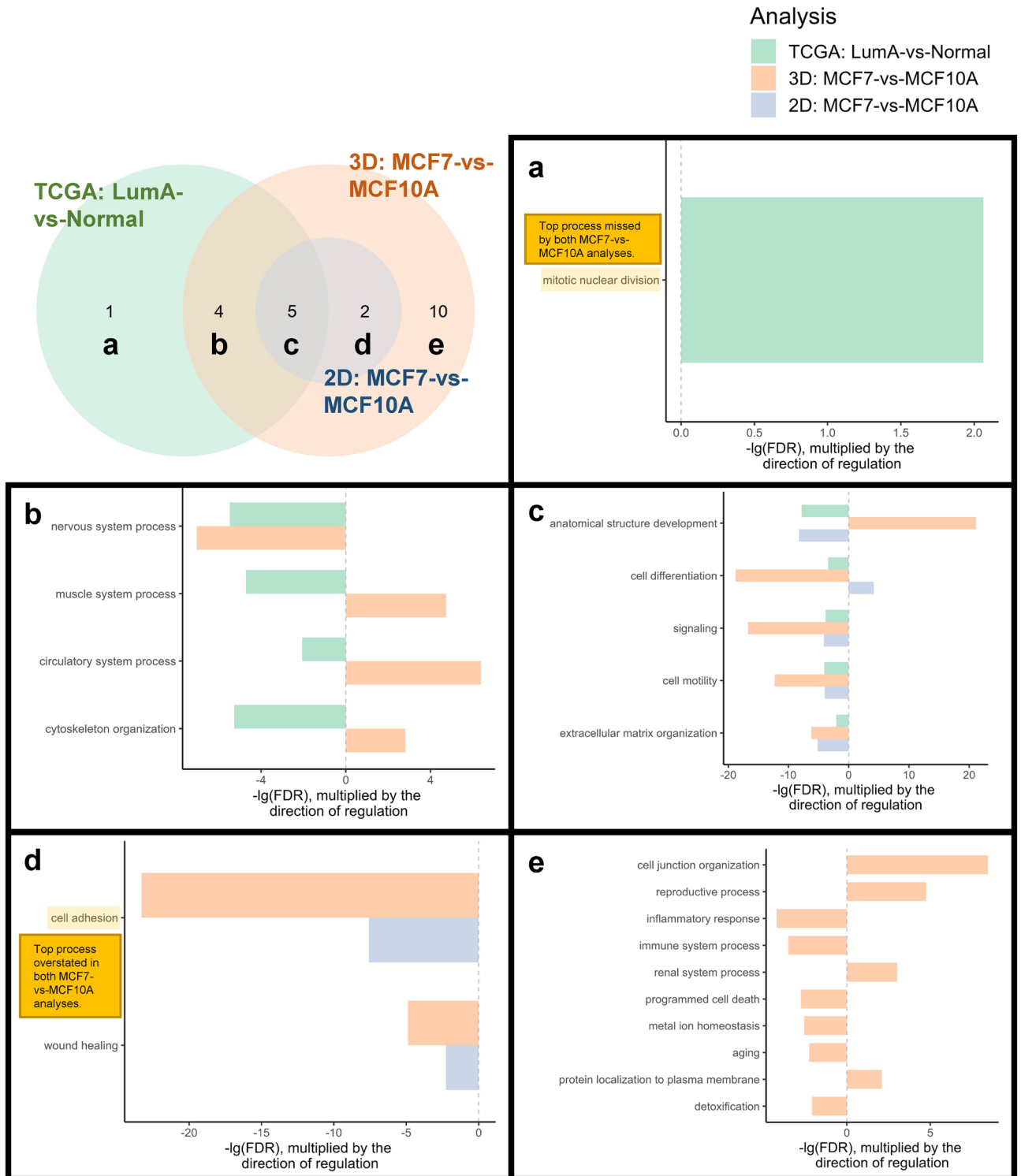


Figure 3. Venn analysis and overrepresented gene sets. (a) Gene sets which were specific to the LumA-vs-normal analysis, (b) specific to both the LumA-vs-normal and 3D MCF7-vs-MCF10A analyses, (c) present in all three analyses, (d) specific to both 3D and 2D cultures of MCF7-vs-MCF10A, and (e) specific to the 3D MCF7-vs-MCF10A analysis only. For further discussion, we highlighted mitotic nuclear division, the top process missed by both 2D and 3D MCF7-vs-MCF10A analyses, and cell adhesion, the top process overstated in both 2D and 3D MCF7-vs-MCF10A analyses compared to LumA-vs-normal.

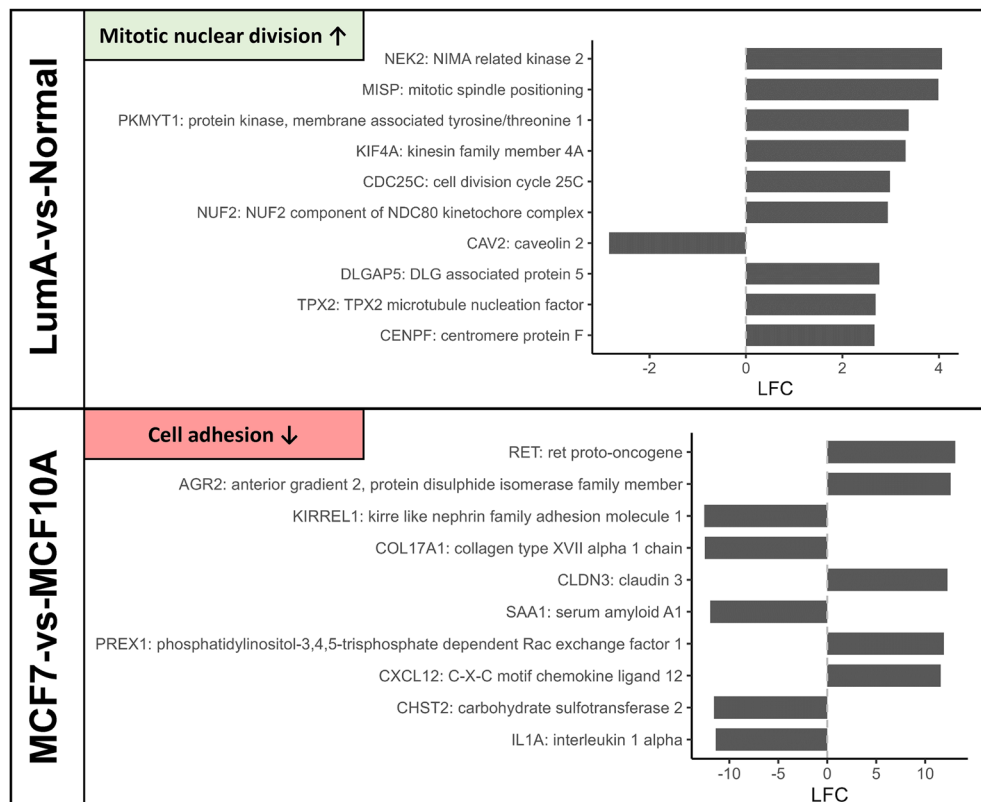


Figure 4. Top Boruta-selected genes which characterise differences in enriched pathways (mitotic nuclear division and cell adhesion) contributing to tumorigenesis in clinical breast tissues and MCF7/MCF10A cell lines respectively. Top panel: Top 10 Boruta-selected genes associated with cancer-related upregulation of mitotic nuclear division in clinical LumA BC compared to normal samples, ranked by absolute LFC (from the TCGA LumA-vs-normal analysis). Bottom panel: Top 10 Boruta-selected genes associated with downregulation of cell adhesion in MCF7 compared to MCF10A, ranked by absolute LFC (from the 3D MCF7-vs-MCF10A analysis).

across analyses, these genes were overrepresented for a similar set of processes. For instance, all analyses showed a dysregulation of anatomical structure development. This is consistent with the known loss of structure organisation in BC⁷⁰, characterised by luminal filling of mammary ducts due to unrestricted cell growth⁷¹, and the findings that developmental pathways are implicated in cancer progression^{72,73}. All analyses also showed dysregulation in cell differentiation, consistent with canonical knowledge that cancer cells undergo epithelial-to-mesenchymal transition, a transient state where cells initially dedifferentiate into a mesenchymal phenotype with invasive potential, and subsequently differentiate into various tumour cell types, creating intratumour heterogeneity which facilitates cancer progression and treatment resistance⁷⁴⁻⁷⁶. As in the LumA-vs-normal analysis, these patterns of dedifferentiation and differentiation in cancer were correctly represented as dysregulation of cell differentiation in the MCF7-vs-MCF10A analyses.

3D cultures of MCF7-vs-MCF10A were superior LumA BC models to 2D cultures, but neither were highly accurate. The 3D MCF7-vs-MCF10A analysis shared more common Boruta-selected genes and enriched gene sets with the LumA-vs-normal analysis compared to the 2D MCF7-vs-MCF10A analysis, suggesting that 2D cultures did not add any value as a model above 3D cultures. The restrictions of monolayer organisation and flat cell morphology in 2D cultures may be responsible for diminished transcriptional patterns^{27,77}, resulting in an attenuated set of biological processes differentiating MCF7 from MCF10A. Our findings support the notion that 3D culture systems are superior models to 2D cultures⁷⁸.

Nonetheless, we found that neither the 3D nor 2D analyses accurately represented LumA BC. The 2D MCF7-vs-MCF10A analysis missed half of the cancer-related processes seen in the LumA-vs-normal analysis, suggesting that it was unable to fully capture the complex processes in LumA BC. While the analysis for 3D cultures may have captured most of the processes in LumA-vs-normal, there were 12 additional significant gene sets being highlighted. This suggests that 3D models may lead experimenters to identify many spurious processes which are rare or non-existent in clinical BC, possibly reducing the validity of studies using 3D cultures. Few other studies have investigated whether 3D cultures introduce false positives. Specifically, 3D cultures of MCF10A was found to express markers which were absent in typical breast tissue⁹, further supporting the observation that 3D culture systems introduced changes to cellular processes which differ from breast tissue in vivo. Future work should be done to confirm and understand the elements of 3D cultures that cause cell lines to present spurious processes.

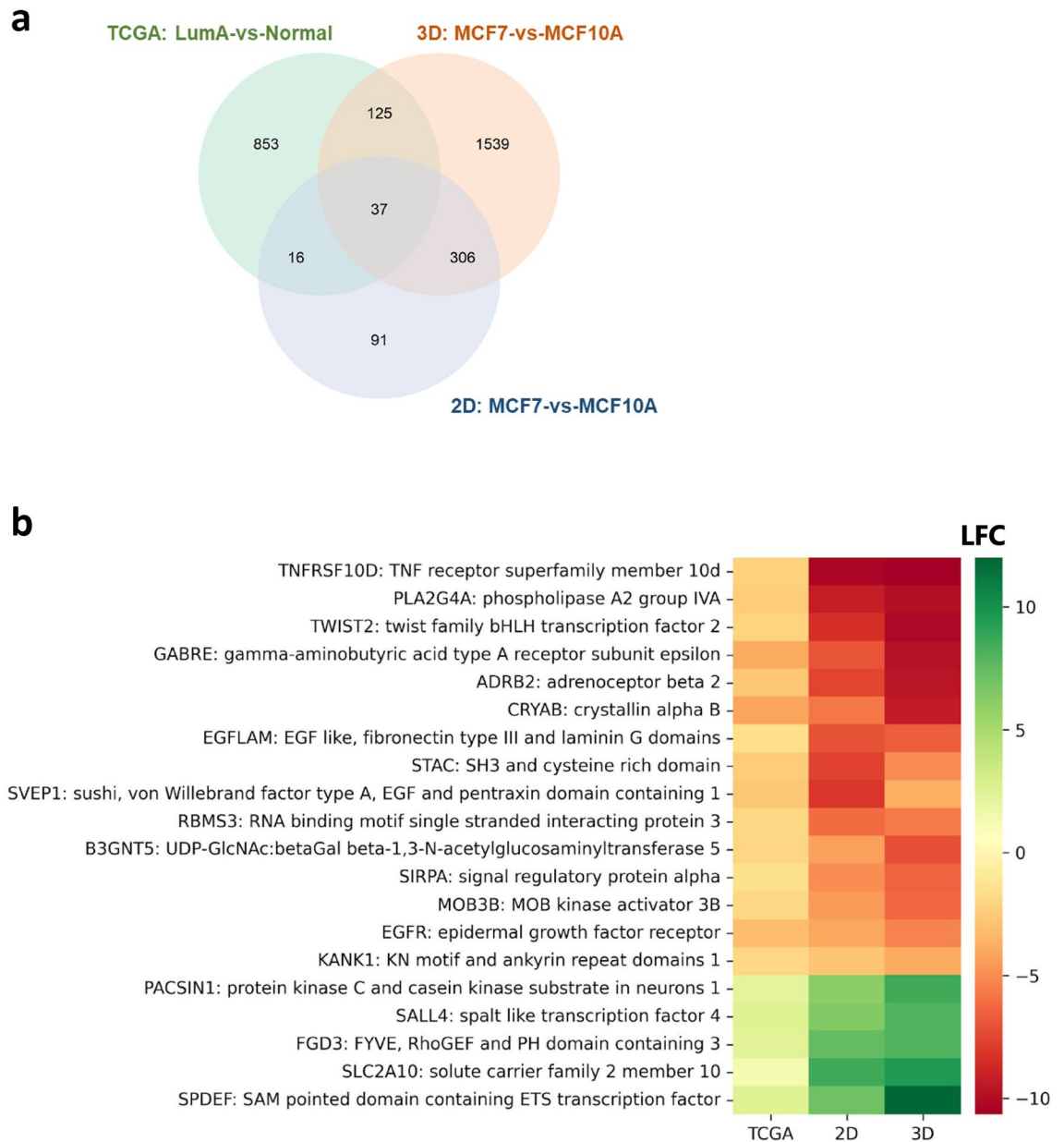


Figure 5. Common cancer-related genes across cell lines and clinical samples. **(a)** Venn diagram showing the number of Boruta-selected genes in each analysis (TCGA: LumA-vs-Normal, 3D: MCF7-vs-MCF10A, 2D: MCF7-vs-MCF10A). 37 genes were selected by Boruta across the three analyses. **(b)** Heatmap of the 20 cancer-related genes which were found to be dysregulated across analyses of 2D and 3D cultures of MCF7-vs-MCF10A cell lines and clinical samples. These genes were sorted by the average LFC across the three analyses. Visualisation was done using the *seaborn* (version 0.11.2) and *matplotlib* (version 3.4.3) Python data visualization libraries.

MCF7-vs-MCF10A comparisons failed to model the cancer-related increase in mitotic nuclear division in LumA-vs-normal. The ability to sustain proliferative signalling is a hallmark of cancer⁷⁹. Therefore, a successful model of cancer should display upregulation in processes related to cell cycle progression. Mitotic nuclear division represents a key step in cell cycle progression⁸⁰. We found that LumA tumours showed greater nuclear division activity compared to normal samples, but MCF7 did not differ in nuclear division compared to MCF10A samples. This suggests that cell lines were unable to accurately model cancer-related cell cycle dysregulation. A possible reason is that both MCF10A and MCF7 were immortalised^{10,11}. Immortalisation entails the impairment of cell cycle checkpoints and oncogenic changes⁸¹, much like cancer cells. Both MCF10A and MCF7 displayed comparable abilities in recovering from cell cycle arrest⁸², suggesting similarities between these immortalised cell lines in cell cycle regulation. Therefore, the cancer-related increase in nuclear division may have been confounded by the immortalisation-related upregulation of nuclear division in MCF10A. Since

MCF10A displayed similar upregulation of nuclear division to tumourigenic MCF7, no cancer-related upregulation of mitotic nuclear division was observed in the MCF7-vs-MCF10A comparison.

Our findings are different but not inconsistent with past studies' findings that cell cycle processes tend to be upregulated in cell lines compared to primary tumours (Table 1). While these studies merely described that cell lines have greater proliferative potential than primary tumours, we found that the immortalisation of cell lines confounded the assessment of cancer-related cell cycle dysregulation, providing a direct commentary on cell line-based cancer-modelling.

MCF7 and MCF10A have been used to understand mechanisms of cell cycle regulation in BC^{83,84}. Given that cell cycle pathways are affected by immortalisation, it becomes difficult to determine whether an experimental manipulation affects cancer-related or immortalisation-related cell cycle pathways, weakening the generalisability of these studies to clinical BC. To address the potential confounding influence of immortalisation, experimenters could perform gene knockdowns of mitotic nuclear division-related genes (Fig. 4; top panel) in cell lines to reverse some of the effects of immortalisation prior to experimental manipulation. For instance, *NEK2*, *MISP* and *PKMYT1* are nuclear division-related genes known to be overexpressed in BC⁸⁵⁻⁸⁷. These genes should not be similarly overexpressed in non-tumourigenic breast models (like MCF10A). Silencing of these genes may help experimenters create experimental models that better represent non-cancerous tissue.

MCF7-vs-MCF10A overstated the downregulation of cell adhesion compared to clinical BC.

Our findings indicated that MCF7 showed downregulated cell adhesion compared to MCF10A samples, but cell adhesion was not dysregulated in LumA tumours compared to normal samples. This finding was consistent with literature reporting decreased cell adhesion in cancer cell lines compared to tumours (Table 1). Given that the MCF7 cell line was derived from a pleural effusion from a late-stage metastatic BC⁸⁸, the observed downregulation of cell adhesion may have been an artefact of using a cell line of metastatic origin. The loss of cell adhesion-related functions allows detachment from the primary tumour, promoting metastasis⁸⁹. In contrast to MCF7's metastatic origin, only 1.4% of the LumA samples in the TCGA dataset had been diagnosed as metastatic. Unlike metastatic cells, many primary tumours may retain normal cell adhesion functions, enabling collective cell migration in tumour invasion and facilitating intercellular interactions with other cancer and stromal cells^{90,91}. Therefore, cell adhesion processes were intact in the mostly non-metastatic TCGA dataset, but downregulated in the MCF7-vs-MCF10A comparison due to MCF7's metastatic lineage. Another possible reason is that the standard of care for luminal tumours involves endocrine therapy⁹². The patients in the TCGA were likely treated with anti-estrogen therapy. ER inhibitors like tamoxifen had been shown to restore cell-cell adhesion, reducing tumour invasion⁹³. Consequently, this may have recovered cell adhesion in TCGA LumA samples.

To address the differences in cell adhesion due to the metastatic lineage of MCF7 compared to LumA tumours, researchers could use techniques like CRISPR/Cas9-mediated gene activation⁹⁴ or RNA interference-based translational silencing⁹⁵. Candidate cell adhesion-related genes (Fig. 4; bottom panel) for activation include *KIRRELI1*, which is known to be overexpressed in BC⁹⁶, and *COL17A1*, whose underexpression is associated with metastatic tumours⁹⁷. Activating these genes may increase resemblance of MCF7's gene expression profile to that of LumA tumours. Candidate cell adhesion-related genes for suppression include *RET*, which has no known prognostic significance in ER-positive BC⁹⁸, and *AGR2*, whose overexpression is associated with metastasis rather than typical LumA BC, and is known to be inhibited by estrogen inhibitor therapy⁹⁹. The absence of estrogen inhibitor-affected Boruta-selected genes like *AGR2* in the LumA-vs-normal analysis further supports the conjecture that endocrine therapies resolved the dysregulation of cell adhesion-related genes in clinical samples. To further address the differences in cell adhesion due to treatment effects in clinical samples, cell lines modelling post-treatment LumA cancers could be cultured in tamoxifen-treated media, better modelling tumour behaviour under endocrine therapy exposure.

Opportunities for future cell line-based experimental study of LumA BC. With the compilation of RNA-seq data from the ARCHS4 and TCGA repositories to augment our RNA-seq dataset, we presented a list of putative biomarkers (Fig. 5b) which exhibited concordance in tumour progression for both clinical breast tissues and cell lines.

We could also use this list of genes to pursue more specific questions of interest. For example, to understand the cancer-related morphogenic changes underlying BC, experimenters may be interested in understanding gene expression differences related to extracellular matrix organisation. Among the genes in Fig. 5b, the top shared gene (by average absolute LFC) involved in extracellular matrix organisation was *SLC2A10*. This gene had been demonstrated to be upregulated in clinical BC¹⁰⁰. *SLC2A10* encodes GLUT10, a facilitative glucose transporter whose role in cancers remains unclear^{101,102}. Downregulation of *SLC2A10* is tied to destabilisation of the extracellular matrix via deficiencies in ascorbic acid processing which are cofactors facilitating collagen and elastin production¹⁰³, and impaired cardiovascular morphogenesis via alterations in respiration and TGFβ signalling¹⁰⁴. However, less is known about the implications of *SLC2A10* overexpression in cancer. The consistent overexpression of *SLC2A10* in cell lines relative to clinical samples suggest that it is possible to further investigate the role of *SLC2A10* overexpression using MCF7 and MCF10A. Future studies could further explore the role of this gene in cancer using experimental cell lines, to better understand its implications on extracellular matrix organisation in cancer cells compared to normal cells.

To further explore cytoskeleton organisation, researchers could analyse *CRYAB*, which was the top cancer-related gene associated with cytoskeleton organisation in the analyses of 2D and 3D cultures of cell lines and clinical samples. *CRYAB* encodes a small heat-shock protein associated with maintaining cytoskeletal integrity under stresses to the cytoskeleton¹⁰⁵, possibly by reducing the aggregation of F-actin¹⁰⁶. Studies on *CRYAB* expression in BC were mixed. While *CRYAB* was overexpressed and associated with poorer prognosis in the

basal BC subtype and metastasis^{107–109}, *CRYAB* was shown to be strongly downregulated in all BC subtypes¹¹⁰. Interestingly, separate clusters of BC samples were identified to be varying in *CRYAB* expression, where *CRYAB* tended to be less expressed in ER-positive BC but more expressed in ER-negative BC¹¹¹, suggesting a subtype-specific gene expression pattern in *CRYAB*. While the tumour-suppressive roles of *CRYAB* in Luma BC are less well-established, *CRYAB* has been shown to reduce progression in nasopharyngeal cancers by associating with membrane-bound β -catenin, preventing the release of β -catenin into the cytoplasm, thereby suppressing the oncogenic abilities of β -catenin by preventing it from interacting with complexes to facilitate transcription of genes involved in tumour progression¹¹². MCF7 and MCF10A cell lines may be useful models to resolve the mixed findings of *CRYAB* expression in cancers and better understand the mechanisms underlying *CRYAB* in BC. Given how its underexpression in Luma BC was consistently reflected in both cell lines and clinical samples, future work could use MCF7 cell lines to examine the roles of *CRYAB* in Luma BC, such as via gene activation protocols. There is a need for such studies to identify downstream effects of *CRYAB* expression, identify *CRYAB*'s tumour-suppressive role in Luma BC, and better determine whether *CRYAB* is an effector in cytoskeletal organisation or disorganisation.

Limitations and future work. Our cell line dataset was a collection of independent experiments with varying culture and sequencing protocols. Specifically, in the ARCHS4 data, many MCF7 samples were obtained from different experiments from the MCF10A samples. Despite our attempts to remove cancer-unrelated variation, batch effects may still exist. Nonetheless, our study aimed to look at the broad cancer-related processes distinguishing cell line models from clinical samples and was able to capture the general differences between analyses of clinical samples and cell lines irrespective of remaining batch effects.

Deriving biological meaning from manually curated GO annotations also provides a limited view of tissue-specific functions. It would be worthwhile to seek an alternative method to impute knowledge for uncovering unannotated biological processes for less studied genes.

BCs exhibit high intra-tumour heterogeneity consisting of different cell subpopulations, genetic heterogeneity and mixed morphologies^{113–115}. Moreover, cell subpopulations have important implications on treatment efficacy. For instance, studies have found that tumour-associated macrophages are related to worse prognoses¹¹⁶ while tumour-infiltrating lymphocytes are associated with good prognoses¹¹⁷. These effects of intratumoral heterogeneity cannot be elucidated directly with bulk RNA-seq data. Future work could collate single cell RNA-seq data to compare specific cell subpopulations between tumours and cell lines. This would allow us to answer more specific questions, such as whether cell lines preferentially model certain cellular subpopulations or whether intratumoral heterogeneity in clinical samples introduces complexity that cannot be modelled in cell lines.

Furthermore, our 3D culture dataset was limited in size compared to the other datasets. There are also various sophisticated 3D culture systems which better model tumour-stromal interactions such as co-cultures with other cell types¹¹⁸ and vascularised systems¹¹⁹, which were not represented in this study. Hence, the list of cancer-related genes and gene sets identified may not generalise well to the full range of 3D culture technology. As 3D culture systems gain popularity, there might be more available data in future which can be used to determine the key cancer-related processes more reliably in 3D MCF7-vs-MCF10A. Future work could amass a larger dataset and further consider the differences between 3D culture techniques, to better assess the modelling capacity of MCF7 and MCF10A.

Finally, our study focused on MCF7 and MCF10A, among the many available BC and non-tumourigenic cell lines. However, our method is scalable and can be easily extended to RNA-seq data mined for other cell lines. Future work can consider expanding implementation to other BC cell lines to obtain a comprehensive idea of the overall representativeness of cell line models to clinical BC data and assist cell line selection in experimental studies by ranking cell lines by similarities in cancer-related processes to relevant breast tissues.

Data availability

RNA-seq data performed in this study for the 2D and 3D cultures of MCF10A and MCF7 were deposited at Gene Expression Omnibus (GEO) repository as GEO Series record GSE208731.

Received: 15 July 2022; Accepted: 16 November 2022

Published online: 03 December 2022

References

1. Lei, S. *et al.* Global patterns of breast cancer incidence and mortality: A population-based cancer registry data analysis from 2000 to 2020. *Cancer Commun.* **41**(11), 1183–1194. <https://doi.org/10.1002/cac2.12207> (2021).
2. Sung, H. *et al.* Global Cancer Statistics 2020: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 Countries. *CA Cancer J. Clin.* **71**(3), 209–249. <https://doi.org/10.3322/caac.21660> (2021).
3. Soerjomataram, I. & Bray, F. Planning for tomorrow: Global cancer incidence and the role of prevention 2020–2070. *Nat. Rev. Clin. Oncol.* **18**(10), 663–672. <https://doi.org/10.1038/s41571-021-00514-z> (2021).
4. Mirabelli, P., Coppola, L. & Salvatore, M. Cancer cell lines are useful model systems for medical research. *Cancers (Basel)* **11**(8), 1098. <https://doi.org/10.3390/cancers11081098> (2019).
5. Jiang, G. *et al.* Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. *BMC Genom.* **17**(7), 525. <https://doi.org/10.1186/s12864-016-2911-z> (2016).
6. Iorio, F. *et al.* A landscape of pharmacogenomic interactions in cancer. *Cell* **166**(3), 740–754. <https://doi.org/10.1016/j.cell.2016.06.017> (2016).
7. Richter, M. *et al.* From donor to the lab: A fascinating journey of primary cell lines. *Front. Cell Dev. Biol.* **9**, 711381. <https://doi.org/10.3389/fcell.2021.711381> (2021).
8. Lee, A. V., Oesterreich, S. & Davidson, N. E. MCF-7 cells—Changing the course of breast cancer research and care for 45 years. *J. Natl. Cancer Inst.* **107**(7), 73. <https://doi.org/10.1093/jnci/djv073> (2015).

9. Qu, Y. *et al.* Evaluation of MCF10A as a reliable model for normal human mammary epithelial cells. *PLoS ONE* **10**(7), e0131285. <https://doi.org/10.1371/journal.pone.0131285> (2015).
10. Soule, H. D. *et al.* Isolation and characterization of a spontaneously immortalized human breast epithelial cell line, MCF-10. *Cancer Res.* **50**(18), 6075–6086 (1990).
11. Soule, H. D., Vazquez, J., Long, A., Albert, S. & Brennan, M. A human cell line from a pleural effusion derived from a breast carcinoma. *JNCI J. Natl. Cancer Inst.* **51**(5), 1409–1416. <https://doi.org/10.1093/jnci/51.5.1409> (1973).
12. Dai, X., Cheng, H., Bai, Z. & Li, J. Breast cancer cell line classification and its relevance with breast tumor subtyping. *J. Cancer* **8**(16), 3131–3141. <https://doi.org/10.7150/jca.18457> (2017).
13. Holliday, D. L. & Speirs, V. Choosing the right cell line for breast cancer research. *Breast Cancer Res.* **13**(4), 215. <https://doi.org/10.1186/bcr2889> (2011).
14. Moon, H. *et al.* Subtype-specific characterization of breast cancer invasion using a microfluidic tumor platform. *PLoS ONE* **15**(6), e0234012. <https://doi.org/10.1371/journal.pone.0234012> (2020).
15. Yersal, O. & Barutca, S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. *World J. Clin. Oncol.* **5**(3), 412–424. <https://doi.org/10.5306/wjco.v5.i3.412> (2014).
16. Barretina, J. *et al.* The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**(7391), 603–607. <https://doi.org/10.1038/nature11003> (2012).
17. Ghandi, M. *et al.* Next-generation characterization of the cancer cell line encyclopedia. *Nature* **569**(7757), 503–508. <https://doi.org/10.1038/s41586-019-1186-3> (2019).
18. Liu, K. *et al.* Evaluating cell lines as models for metastatic breast cancer through integrative analysis of genomic data. *Nat. Commun.* **10**(1), 2138. <https://doi.org/10.1038/s41467-019-10148-6> (2019).
19. Vincent, K. M., Findlay, S. D. & Postovit, L. M. Assessing breast cancer cell lines as tumour models by comparison of mRNA expression profiles. *Breast Cancer Res.* **17**(1), 114. <https://doi.org/10.1186/s13058-015-0613-0> (2015).
20. Sandberg, R. & Ernberg, I. The molecular portrait of in vitro growth by meta-analysis of gene-expression profiles. *Genome Biol.* **6**(8), R65. <https://doi.org/10.1186/gb-2005-6-8-r65> (2005).
21. Tran, V., Kim, R., Maertens, M., Hartung, T. & Maertens, A. Similarities and differences in gene expression networks between the breast cancer cell line Michigan cancer foundation-7 and invasive human breast cancer tissues. *Front. Artif. Intell.* **4**, 674370. <https://doi.org/10.3389/frai.2021.674370> (2021).
22. Gillet, J.-P., Varma, S. & Gottesman, M. M. The clinical relevance of cancer cell lines. *J. Natl. Cancer Inst.* **105**(7), 452–458. <https://doi.org/10.1093/jnci/djt007> (2013).
23. Yamada, K. M. & Cukierman, E. Modeling tissue morphogenesis and cancer in 3D. *Cell* **130**(4), 601–610. <https://doi.org/10.1016/j.cell.2007.08.006> (2007).
24. Amaral, D. B. J., Urabayashi, S. M. & Machado-Santelli, M. G. Cell death and lumen formation in spheroids of MCF-7 cells. *Cell Biol. Int.* **34**(3), 267–274. <https://doi.org/10.1042/CBI20090024> (2010).
25. Vantagoli, M. M., Madnick, S. J., Huse, S. M., Weston, P. & Boekelheide, K. MCF-7 human breast cancer cells form differentiated microtissues in scaffold-free hydrogels. *PLoS ONE* **10**(8), e0135426. <https://doi.org/10.1371/journal.pone.0135426> (2015).
26. Debnath, J., Muthuswamy, S. K. & Brugge, J. S. Morphogenesis and oncogenesis of MCF-10A mammary epithelial acini grown in three-dimensional basement membrane cultures. *Methods* **30**(3), 256–268 (2003).
27. Jensen, C. & Teng, Y. Is it time to start transitioning from 2D to 3D cell culture? *Front. Mol. Biosci.* **7**, 33. <https://doi.org/10.3389/fmolb.2020.00033> (2020).
28. Lachmann, A. *et al.* Massive mining of publicly available RNA-seq data from human and mouse. *Nat. Commun.* **9**(1), 1366. <https://doi.org/10.1038/s41467-018-03751-6> (2018).
29. Ertel, A., Verghese, A., Byers, S. W., Ochs, M. & Tozeren, A. Pathway-specific differences between tumor cell lines and normal and tumor tissue cells. *Mol. Cancer* **5**(1), 55. <https://doi.org/10.1186/1476-4598-5-55> (2006).
30. Yu, K. *et al.* Comprehensive transcriptomic analysis of cell lines as models of primary tumors across 22 tumor types. *Nat. Commun.* **10**(1), 3574. <https://doi.org/10.1038/s41467-019-11415-2> (2019).
31. R. C. Team. *R: A Language and Environment for Statistical Computing*. <https://www.r-project.org/> (R Foundation for Statistical Computing, 2021). Accessed 10 Jan 2022.
32. RStudio Team. *RStudio: Integrated Development Environment for R* (RStudio, 2022).
33. Kluyver, T. *et al.* Jupyter notebooks—A publishing format for reproducible computational workflows. In *20th International Conference on Electronic Publishing (01/01/16)*, 87–90. <https://eprints.soton.ac.uk/403913/> (IOS Press, 2016). Accessed 10 Jan 2022.
34. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2016).
35. Waskom, M. L. Seaborn: Statistical data visualization. *J. Open Source Softw.* **6**(60), 3021 (2021).
36. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**(3), 90–95. <https://doi.org/10.1109/MCSE.2007.55> (2007).
37. Lee, G. Y., Kenny, P. A., Lee, E. H. & Bissell, M. J. Three-dimensional culture models of normal and malignant breast epithelial cells. *Nat. Methods* **4**(4), 359–365. <https://doi.org/10.1038/nmeth1015> (2007).
38. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**(5), 525–527. <https://doi.org/10.1038/nbt.3519> (2016).
39. Soneson, C., Love, M. I. & Robinson, M. D. Differential analyses for RNA-seq: Transcript-level estimates improve gene-level inferences. *F1000Research* **4**, 1521. <https://doi.org/10.12688/f1000research.7563.2> (2015).
40. Aken, B. L. *et al.* The Ensembl gene annotation system. *Database* **2016**, 093. <https://doi.org/10.1093/database/baw093> (2016).
41. Colaprico, A. *et al.* TCGAbiolinks: An R/bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* **44**(8), e71. <https://doi.org/10.1093/nar/gkv1507> (2015).
42. Koboldt, D. C. *et al.* Comprehensive molecular portraits of human breast tumours. *Nature* **490**(7418), 61–70. <https://doi.org/10.1038/nature11412> (2012).
43. Zhang, Z. *et al.* Uniform genomic data analysis in the NCI genomic data commons. *Nat. Commun.* **12**(1), 1226. <https://doi.org/10.1038/s41467-021-21254-9> (2021).
44. Schmeller, J. *et al.* Setting out the frame conditions for feasible use of FFPE derived RNA. *Pathol. Res. Pract.* **215**(2), 381–386. <https://doi.org/10.1016/j.prp.2018.12.027> (2019).
45. Guclalp, A. *et al.* Male breast cancer: A disease distinct from female breast cancer. *Breast Cancer Res. Treat.* **173**(1), 37–48. <https://doi.org/10.1007/s10549-018-4921-9> (2019).
46. Berger, A. C. *et al.* A comprehensive pan-cancer molecular study of gynecologic and breast cancers. *Cancer Cell* **33**(4), 690–705. <https://doi.org/10.1016/j.ccell.2018.03.014> (2018).
47. Robinson, M. D. & Oshlack, A. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**(3), R25. <https://doi.org/10.1186/gb-2010-11-3-r25> (2010).
48. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**(1), 139–140. <https://doi.org/10.1093/bioinformatics/btp616> (2009).
49. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
50. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. Preprint at <http://arXiv.org/1802.03426> (2018).

51. Van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**(11), 2579 (2008).
52. National Cancer Institute. *TCGA Barcode, GDC Documentation*. https://docs.gdc.cancer.gov/Encyclopedia/pages/TCGA_Barcode/. Accessed 10 Jan 2022.
53. Ritchie, M. E. *et al.* Limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**(7), e47. <https://doi.org/10.1093/nar/gkv007> (2015).
54. Blagus, R. & Lusa, L. SMOTE for high-dimensional class-imbalanced data. *BMC Bioinform.* **14**(1), 106. <https://doi.org/10.1186/1471-2105-14-106> (2013).
55. Lemaitre, G., Nogueira, F. & Aridas, C. K. Imbalanced-learn: A python toolbox to tackle the curse of imbalanced datasets in machine learning. *J. Mach. Learn. Res.* **18**(1), 559–563 (2017).
56. Chawla, N. V., Bowyer, K. W., Hall, L. O. & Kegelmeyer, W. P. SMOTE: Synthetic minority over-sampling technique. *J. Artif. Int. Res.* **16**(1), 321–357 (2002).
57. Breiman, L. Random forests. *Mach. Learn.* **45**(1), 5–32. <https://doi.org/10.1023/A:1010933404324> (2001).
58. Kursa, M. B. & Rudnicki, W. R. Feature selection with the boruta package. *J. Stat. Softw.* **36**(11), 1–13. <https://doi.org/10.18637/jss.v036.i11> (2010).
59. Louppe, G., Wehenkel, L., Sutura, A. & Geurts, P. Understanding variable importances in forests of randomized trees. In *Advanced Neural Information Processing System*, Vol. 26 (2013).
60. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2—An R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* **9**, 709. <https://doi.org/10.12688/f1000research.24956.2> (2020).
61. The Gene Ontology Consortium. The gene ontology resource: 20 years and still GOing strong. *Nucleic Acids Res.* **47**(D1), D330–D338. <https://doi.org/10.1093/nar/gky1055> (2018).
62. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**(2), R29. <https://doi.org/10.1186/gb-2014-15-2-r29> (2014).
63. Subramanian, A. *et al.* Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl. Acad. Sci.* **102**(43), 15545–15550. <https://doi.org/10.1073/pnas.0506580102> (2005).
64. Wu, T. *et al.* clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. *Innovation* **2**(3), 100141. <https://doi.org/10.1016/j.xinn.2021.100141> (2021).
65. Chen, H. & Boutros, P. C. VennDiagram: A package for the generation of highly-customizable Venn and Euler diagrams in R. *BMC Bioinform.* **12**(1), 35. <https://doi.org/10.1186/1471-2105-12-35> (2011).
66. Wenric, S. & Shemirani, R. Using supervised learning methods for gene selection in RNA-Seq case-control studies. *Front. Genet.* **9**, 297. <https://doi.org/10.3389/fgene.2018.00297> (2018).
67. Ashworth, A., Lord, C. J. & Reis-Filho, J. S. Genetic interactions in cancer progression and treatment. *Cell* **145**(1), 30–38. <https://doi.org/10.1016/j.cell.2011.03.020> (2011).
68. Iengar, P. Identifying pathways affected by cancer mutations. *Genomics* **110**(5), 318–328. <https://doi.org/10.1016/j.ygeno.2017.12.004> (2018).
69. Maleki, F., Ovens, K., Hogan, D. J. & Kuslik, A. J. Gene set analysis: Challenges, opportunities, and future research. *Front. Genet.* **11**, 654. <https://doi.org/10.3389/fgene.2020.00654> (2020).
70. Burstein, H. J., Polyak, K., Wong, J. S., Lester, S. C. & Kaelin, C. M. Ductal carcinoma in situ of the breast. *N. Engl. J. Med.* **350**(14), 1430–1441. <https://doi.org/10.1056/NEJMra031301> (2004).
71. Pradeep, C. R. *et al.* Modeling ductal carcinoma in situ: A HER2–Notch3 collaboration enables luminal filling. *Oncogene* **31**(7), 907–917. <https://doi.org/10.1038/ncr.2011.279> (2012).
72. Geissler, K. & Zach, O. Pathways involved in Drosophila and human cancer development: The Notch, Hedgehog, Wingless, Runt, and Trithorax pathway. *Ann. Hematol.* **91**(5), 645–669. <https://doi.org/10.1007/s00277-012-1435-0> (2012).
73. Nwabo Kamdje, A. H. *et al.* Developmental pathways associated with cancer metastasis: Notch, Wnt, and Hedgehog. *Cancer Biol. Med.* **14**(2), 109–120. <https://doi.org/10.20892/j.issn.2095-3941.2016.0032> (2017).
74. Battula, V. L. *et al.* Epithelial-mesenchymal transition-derived cells exhibit multilineage differentiation potential similar to mesenchymal stem cells. *Stem Cells* **28**(8), 1435–1445. <https://doi.org/10.1002/stem.467> (2010).
75. Lüönd, F., Tiede, S. & Christofori, G. Breast cancer as an example of tumour heterogeneity and tumour cell plasticity during malignant progression. *Br. J. Cancer* **125**(2), 164–175. <https://doi.org/10.1038/s41416-021-01328-7> (2021).
76. Wang, H. & Unternaehrer, J. J. Epithelial-mesenchymal transition and cancer stem cells: At the crossroads of differentiation and dedifferentiation. *Dev. Dyn.* **248**(1), 10–20. <https://doi.org/10.1002/dvdy.24678> (2019).
77. Birgersdotter, A., Sandberg, R. & Ernberg, I. Gene expression perturbation in vitro—A growing case for three-dimensional (3D) culture systems. *Semin. Cancer Biol.* **15**(5), 405–412. <https://doi.org/10.1016/j.semcancer.2005.06.009> (2005).
78. Kapalczyńska, M. *et al.* 2D and 3D cell cultures—A comparison of different types of cancer cell cultures. *Arch. Med. Sci.* **14**, 910–919. <https://doi.org/10.5114/aoms.2016.63743> (2018).
79. Hanahan, D. & Weinberg, R. A. The hallmarks of cancer. *Cell* **100**(1), 57–70. [https://doi.org/10.1016/S0092-8674\(00\)81683-9](https://doi.org/10.1016/S0092-8674(00)81683-9) (2000).
80. Alberts, B. *et al.* An overview of the cell cycle. In *Molecular Biology of the Cell* 4th edn (eds Alberts, B., Johnson, A., Lewis, J. *et al.*) (Garland Science, 2002).
81. Irfan-Maqsood, M., Matin, M. M., Bahrami, A. R. & Ghasroldasht, M. M. Immortality of cell lines: Challenges and advantages of establishment. *Cell Biol. Int.* **37**(10), 1038–1045. <https://doi.org/10.1002/cbin.10137> (2013).
82. Maya-Mendoza, A. *et al.* Immortalised breast epithelia survive prolonged DNA replication stress and return to cycle from a senescent-like state. *Cell Death Dis.* **5**(7), e1351. <https://doi.org/10.1038/cddis.2014.315> (2014).
83. Eastman, A. Cell cycle checkpoints and their impact on anticancer therapeutic strategies. *J. Cell. Biochem.* **91**(2), 223–231. <https://doi.org/10.1002/jcb.10699> (2004).
84. Foster, J. S., Henley, D. C., Ahamed, S. & Wimalasena, J. Estrogens and cell-cycle regulation in breast cancer. *Trends Endocrinol. Metab.* **12**(7), 320–327. [https://doi.org/10.1016/S1043-2760\(01\)00436-2](https://doi.org/10.1016/S1043-2760(01)00436-2) (2001).
85. Fang, Y. & Zhang, X. Targeting NEK2 as a promising therapeutic approach for cancer treatment. *Cell Cycle* **15**(7), 895–907. <https://doi.org/10.1080/15384101.2016.1152430> (2016).
86. Liu, Y. *et al.* Systematic expression analysis of WEE family kinases reveals the importance of PKMYT1 in breast carcinogenesis. *Cell Prolif.* **53**(2), e12741. <https://doi.org/10.1111/cpr.12741> (2020).
87. Sun, N. *et al.* Screening and identification of key common and specific genes and their prognostic roles in different molecular subtypes of breast cancer. *Front. Mol. Biosci.* **8**, 619110. <https://doi.org/10.3389/fmolb.2021.619110> (2021).
88. Psallidas, I., Kalomenidis, I., Porcel, J. M., Robinson, B. W. & Stathopoulos, G. T. Malignant pleural effusion: From bench to bedside. *Eur. Respir. Rev.* **25**(140), 189–198. <https://doi.org/10.1183/16000617.0019-2016> (2016).
89. Moh, M. C. & Shen, S. The roles of cell adhesion molecules in tumor suppression and cell migration. *Cell Adher. Migr.* **3**(4), 334–336. <https://doi.org/10.4161/cam.3.4.9246> (2009).
90. Cheung, K. J. & Ewald, A. J. Illuminating breast cancer invasion: Diverse roles for cell–cell interactions. *Curr. Opin. Cell Biol.* **30**, 99–111. <https://doi.org/10.1016/j.ceb.2014.07.003> (2014).
91. Janiszewska, M., Primi, M. C. & Izard, T. Cell adhesion in cancer: Beyond the migration of single cells. *J. Biol. Chem.* **295**(8), 2495–2505. <https://doi.org/10.1074/jbc.REV119.007759> (2020).

92. Harbeck, N. *et al.* Breast cancer. *Nat. Rev. Dis. Prim.* **5**(1), 66. <https://doi.org/10.1038/s41572-019-0111-2> (2019).
93. Bracke, M. E. *et al.* Tamoxifen restores the E-cadherin function in human breast cancer MCF-7/6 cells and suppresses their invasive phenotype. *Cancer Res.* **54**(17), 4607–4609 (1994).
94. Liao, H.-K. *et al.* In vivo target gene activation via CRISPR/Cas9-mediated Trans-epigenetic modulation. *Cell* **171**(7), 1495–1507. <https://doi.org/10.1016/j.cell.2017.10.025> (2017).
95. Agrawal, N. *et al.* RNA interference: Biology, mechanism, and applications. *Microbiol. Mol. Biol. Rev.* **67**(4), 657–685. <https://doi.org/10.1128/MMBR.67.4.657-685.2003> (2003).
96. Chen, K. *et al.* Overexpression of kin of IRRE-Like protein 1 (KIRREL) as a prognostic biomarker for breast cancer. *Pathol. Res. Pract.* **216**(7), 153000. <https://doi.org/10.1016/j.prp.2020.153000> (2020).
97. Yodsurang, V. *et al.* Identification of a novel p53 target, COL17A1, that inhibits breast cancer cell migration and invasion. *Oncotarget* **8**(34), 55790–55803. <https://doi.org/10.18632/oncotarget.18433> (2017).
98. Mechera, R. *et al.* Expression of RET is associated with oestrogen receptor expression but lacks prognostic significance in breast cancer. *BMC Cancer* **19**(1), 41. <https://doi.org/10.1186/s12885-018-5262-0> (2019).
99. Salmans, M. L., Zhao, F. & Andersen, B. The estrogen-regulated anterior gradient 2 (AGR2) protein in breast cancer: A potential drug target and biomarker. *Breast Cancer Res.* **15**(2), 204. <https://doi.org/10.1186/bcr3408> (2013).
100. Shi, Z., Liu, J., Wang, F. & Li, Y. Integrated analysis of solute carrier family-2 members reveals SLC2A4 as an independent favorable prognostic biomarker for breast cancer. *Channels* **15**(1), 555–568. <https://doi.org/10.1080/19336950.2021.1973788> (2021).
101. Barron, C. C., Bilan, P. J., Tsakiridis, T. & Tsiani, E. Facilitative glucose transporters: Implications for cancer detection, prognosis and treatment. *Metab. Clin. Exp.* **65**(2), 124–139. <https://doi.org/10.1016/j.metabol.2015.10.007> (2016).
102. Echeverría, C., Nualart, F., Ferrada, L., Smith, G. J. & Godoy, A. S. Hexose transporters in cancer: From multifunctionality to diagnosis and therapy. *Trends Endocrinol. Metab.* **32**(4), 198–211. <https://doi.org/10.1016/j.tem.2020.12.006> (2021).
103. Segade, F. Glucose transporter 10 and arterial tortuosity syndrome: The vitamin C connection. *FEBS Lett.* **584**(14), 2990–2994. <https://doi.org/10.1016/j.febslet.2010.06.011> (2010).
104. Willaert, A. *et al.* GLUT10 is required for the development of the cardiovascular system and the notochord and connects mitochondrial function to TGF β signaling. *Hum. Mol. Genet.* **21**(6), 1248–1259. <https://doi.org/10.1093/hmg/ddr555> (2012).
105. Launay, N., Goudeau, B., Kato, K., Vicart, P. & Lilienbaum, A. Cell signaling pathways to α B-crystallin following stresses of the cytoskeleton. *Exp. Cell Res.* **312**(18), 3570–3584. <https://doi.org/10.1016/j.yexcr.2006.07.025> (2006).
106. Yin, B. *et al.* CRYAB protects cardiomyocytes against heat stress by preventing caspase-mediated apoptosis and reducing F-actin aggregation. *Cell Stress Chaperones* **24**(1), 59–68. <https://doi.org/10.1007/s12192-018-0941-y> (2019).
107. Malin, D., Petrovic, V., Strekalova, E., Sharma, B. & Cryns, V. L. α B-crystallin: Portrait of a malignant chaperone as a cancer therapeutic target. *Pharmacol. Ther.* **160**, 1–10. <https://doi.org/10.1016/j.pharmthera.2016.01.012> (2016).
108. Moyano, J. V. *et al.* α B-Crystallin is a novel oncoprotein that predicts poor clinical outcome in breast cancer. *J. Clin. Investig.* **116**(1), 261–270. <https://doi.org/10.1172/JCI25888> (2006).
109. Sitterding, S. M. *et al.* α B-crystallin: A novel marker of invasive basal-like and metaplastic breast carcinomas. *Ann. Diagn. Pathol.* **12**(1), 33–40. <https://doi.org/10.1016/j.anndiagpath.2007.02.004> (2008).
110. Zoppino, F. C. M., Guerrero-Gimenez, M. E., Castro, G. N. & Ciocca, D. R. Comprehensive transcriptomic analysis of heat shock proteins in the molecular subtypes of human breast cancer. *BMC Cancer* **18**(1), 700. <https://doi.org/10.1186/s12885-018-4621-1> (2018).
111. Quinlan, P. R. *et al.* Cluster analyses of the TCGA and a TMA dataset using the coexpression of HSP27 and CRYAB improves alignment with clinical-pathological parameters of breast cancer and suggests different epichaperome influences for each sHSP. *Cell Stress Chaperones* **27**(2), 177–188. <https://doi.org/10.1007/s12192-022-01258-0> (2022).
112. Huang, Z. *et al.* Tumor suppressor α B-crystallin (CRYAB) associates with the cadherin/catenin adherens junction and impairs NPC progression-associated properties. *Oncogene* **31**(32), 3709–3720. <https://doi.org/10.1038/ncr.2011.529> (2012).
113. Turashvili, G. & Brogi, E. Tumor heterogeneity in breast cancer. *Front. Med.* **4**, 227. <https://doi.org/10.3389/fmed.2017.00227> (2017).
114. Martelotto, L. G., Ng, C. K. Y., Piscuoglio, S., Weigelt, B. & Reis-filho, J. S. Breast cancer intra-tumor heterogeneity. *Breast Cancer Res.* **16**, 210 (2014).
115. Turner, K. M., Yeo, S. K., Holm, T. M., Shaughnessy, E. & Guan, J. L. Heterogeneity within molecular subtypes of breast cancer. *Am. J. Physiol. Cell Physiol.* **321**(2), C343–C354. <https://doi.org/10.1152/ajpcell.00109.2021> (2021).
116. Cassetta, L. *et al.* Human tumor-associated macrophage and monocyte transcriptional landscapes reveal cancer-specific reprogramming, biomarkers, and therapeutic targets. *Cancer Cell* **35**(4), 588–602. <https://doi.org/10.1016/j.ccell.2019.02.009> (2019).
117. Dushyanthen, S. *et al.* Relevance of tumor-infiltrating lymphocytes in breast cancer. *BMC Med.* **13**(1), 1–13. <https://doi.org/10.1186/s12916-015-0431-3> (2015).
118. Yakavets, I. *et al.* Advanced co-culture 3D breast cancer model for investigation of fibrosis induced by external stimuli: Optimization study. *Sci. Rep.* **10**(1), 21273. <https://doi.org/10.1038/s41598-020-78087-7> (2020).
119. Teixeira, F. C., Chaves, S., Torres, A. L., Barrias, C. C. & Bidarra, S. J. Engineering a vascularized 3D hybrid system to model tumor-stroma interactions in breast cancer. *Front. Bioeng. Biotechnol.* **9**, 647031. <https://doi.org/10.3389/fbioe.2021.647031> (2021).

Acknowledgements

This study was funded by MOE Academic Research Fund Tier 1 Grant (Award Number: RG106/20) from Ministry of Education, Singapore, and funding support to CGK from School of Biological Sciences, Nanyang Technological University.

Author contributions

Investigation: J.J.H.G., C.J.H.G., Q.W.L., and S.Z.; Writing—Original Draft: J.J.H.G.; Formal analysis: J.J.H.G., C.J.H.G., Q.W.L., and S.Z.; Data Curation and Visualization: J.J.H.G., C.J.H.G.; Conceptualization and Supervision: C.-G.K, K.-H.C. All authors contributed to reviewing and editing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-022-24511-z>.

Correspondence and requests for materials should be addressed to K.-H.C.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022