A novel survival prediction signature outperforms PAM50 and artificial intelligence-based feature-selection methods

Reuben Jyong Kiat Foo [1^], Siqi Tian [2,3], Ern Yu Tan [2,4], Wilson Wen Bin Goh [2,3,5*]

1. School of Chemical and Biomedical Engineering, Nanyang Technological University, Singapore
2. Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore
3. School of Biological Sciences, Nanyang Technological University, Singapore
4. Tan Tock Seng Hospital, Singapore
5. Centre for Biomedical Informatics, Nanyang Technological University, Singapore

^ First Author

* Corresponding Author: Wilson Wen Bin Goh, wilsongoh@ntu.edu.sg

Address for correspondence/proofs:

Wilson Wen Bin Goh, PhD

Lee Kong Chian School of Medicine, Nanyang Technological University, 59 Nanyang Drive, Singapore 636921

School of Biological Sciences, Nanyang Technological University, 60 Nanyang Drive, Singapore 637551

Email: wilsongoh@ntu.edu.sg, Tel: +65-65162902

Abstract (203 words)

The robustness of a breast cancer gene signature, the super-proliferation set (SPS), is initially tested and investigated on breast cancer cell lines from the Cancer Cell Line Encyclopaedia (CCLE). Previously, SPS was derived via a meta-analysis of 47 independent breast cancer gene signatures, benchmarked on survival information from clinical data in the NKI dataset. Here, relying on the stability of cell line data and associative prior knowledge, we first demonstrate through Principal Component Analysis (PCA) that SPS prioritizes survival information over secondary subtype information, surpassing both PAM50 and Boruta, an artificial intelligence-based feature-selection algorithm, in this regard. We can also extract higher resolution 'progression' information using SPS, dividing survival outcomes into several clinically relevant stages ('good', 'intermediate', and 'bad) based on different quadrants of the PCA scatterplot. Furthermore, by transferring these 'progression' annotations onto independent clinical datasets, we demonstrate the generalisability of our method on actual patient data. Finally, via the characteristic genetic profiles of each quadrant/stage, we identified efficacious drugs using their gene reversal scores that can shift signatures across quadrants/stages, in a process known as gene signature reversal. This confirms the power of meta-analytical approaches for gene signature inference in breast cancer, as well as the clinical benefit in translating these inferences onto real-world patient data for more targeted therapies.

Keywords

Breast Cancer; Data Science; Machine Learning; Meta-analysis; Principal Component Analysis; Survival Prediction

Abbreviations

- Cancer Cell Line Encyclopaedia (CCLE)
- Database for Annotation, Visualization, and Integrated Discovery (DAVID)
- Oestrogen Receptor (ER)
- Human Epidermal Growth Factor Receptor-2 (HER2)
- Library of Integrated Network-based Cellular Signatures (LINCS)
- Prediction Analysis for Microarrays (PAM50)
- Principal Component (PC)
- Principal Component Analysis (PCA)
- Progesterone Receptor (PR)
- Random Signature Superiority (RSS)
- Reverse Gene Expression Score (RGES)
- Super-Proliferation Set (SPS)
- Top-Left; Top-Right; Bottom-Left; Bottom-Right (TL; TR; BL; BR)
- Tumour, Nodes, and Metastases (TNM)
- Library of Integrated Network-based Cellular Signatures (LINCS)
- The Cancer Genome Atlas (TCGA)

1 Introduction

Cancer is a heterogeneous disease comprising diverse types that harbour varying genetic dependencies. Different cancers rely on distinct sets of genes that can interact with or function together to drive the aberrant activity of genetic pathways essential for their proliferation and survival. Recent advances in whole-genome sequencing have enabled high-throughput biology in the form of various -omics platforms, giving rise to high-dimensional biological big data. Yet, big data is of little clinical value without robust and generalisable analytical approaches.

Breast cancer, in particular, has become a centrepiece in the study of cancer genetics. Gene-expression profiling using DNA microarray data has shown predictiveness of clinical outcome, independently of traditional clinical biomarkers such as tumour size, grade, and oestrogen-receptor (ER) status (1). These multi-gene mRNA biomarkers are also known as gene signatures. However, Venet *et al.* also showed that randomly generated signatures could outperform published gene signatures in terms of outcome prediction, demonstrating that small p-values may not imply domain-relevance, nor generalisability (2). Furthermore, this phenomenon was shown to be deeply confounded with proliferation, making its complete elimination virtually impossible (3).

Nonetheless, through a meta-analysis of published signatures, Goh and Wong showed that isolating common genes among the best performing signatures could still produce a robust gene set that was strongly predictive of patient survival, termed the super-proliferation set (SPS) (4). While SPS has been validated with clinical datasets from both the Netherlands Cancer Institute (NKI) (5) as well as 7 independent datasets from the Gene Expression Omnibus (GEO) (4), there still presents an opportunity to further evaluate the performance of this gene signature on cell line datasets provided by the Cancer Cell Line Encyclopaedia (CCLE) (6). As cell lines remain a mainstay of preclinical cancer research and are experimentally more consistent compared to real-world variances in patient data, we hope to derive deeper insight and clinical value from employing SPS on the CCLE dataset.

We examine the prognostic ability of SPS using gene expression data from CCLE cell lines, instead of clinical data. Through exploratory data analyses and unsupervised learning methods, we should be

able to investigate the mechanisms behind SPS at a more granular level. By identifying individual SPS genes or subsets of SPS that characterise the different stages of risk, it could pave the way for more targeted treatments in breast cancer. Furthermore, we compare SPS with not only random and published signatures, but also signatures that we have derived using a modern feature-selection algorithm, Boruta (7). This would effectively benchmark two distinct approaches towards feature selection in gene expression data, namely the meta-analytical approach employed for SPS, as well as the advanced machine learning approach employed by the Boruta algorithm.

2 Materials and Methods

2.1 Breast cancer subtyping

Breast cancer encompasses a heterogeneity of subtypes, each with distinct biological and morphological features that ultimately govern their response to various treatments, clinical outcomes, and patient survival. As a result, clinical management of breast cancer often relies on a host of prognostic variables to decide on the appropriate course of action.

Traditional prognostic variables include the standard TNM staging system (i.e., tumour size, lymph node metastasis, and extent of tumour spread) (8), histological tumour grade, as well as patient age. Molecular markers such as oestrogen receptor (ER), progesterone receptor (PR), and human epidermal growth factor receptor 2 (HER2) status are also a main consideration in treatment decisions, and have enabled more targeted treatments via endocrine therapy or chemotherapy (9, 10). These molecular markers define several molecular subtypes, of which current literatures tend to use the following classifications, with the bottom rows usually associated with worse prognoses (Table 1).

Yet, these traditional classification systems have its limitations in tailoring treatment strategies to patients, and are unable to fully capture the complex genetic processes underlying cancer development and progression (11).

Table 1: Summary of breast cancer molecular subtypes and their characteristic receptors (12).

Molecular Subtype	Characteristics
Luminal	ER+ and/or PR+, HER2-
Luminal-HER2+	ER+ and/or PR+, HER2+
HER2+	ER-, PR-, HER2+
Basal A	ER-, PR-, HER2- (basal tumours)
Basal B	ER-, PR-, HER2- (claudin-low or metaplastic)

2.2 Gene signatures for breast cancer survival

2.2.1 Random Signature Superiority (RSS)

Venet *et al.* previously identified, through a meta-analysis of 47 breast cancer gene signatures from literature, that most of these signatures are unable to outperform sets of random genes in predicting survival outcomes (2), with 11 of them even performing worse than the median random signature. Furthermore, by taking three control signatures with no obvious connection with breast cancer, namely a signature of the effect of postprandial laughter in diabetes patients (13), that of skin fibroblast spatial positioning (14), and that of social defeat in mice (15), it was shown that they were also significantly associated with breast cancer survival. Manjang *et al.* also demonstrated that a large number of breast cancer signatures are not biologically relevant in explaining breast cancer or drug mechanisms, by systematically eliminating all traces of biological meaning from these signatures. Furthermore, the number of random gene sets that are spuriously significant, is immensely large and hard to avoid (16). Termed as random signature superiority (RSS), Goh *et al.* further examined this phenomenon, determining that RSS was correlated with the signatures' size, overall predictive power, association with confounders, and proportion of confounders (3). These confounders include genes involved in the cell cycle, cell death, and contact-based growth inhibition, which are characteristic of the pathogenetic mechanisms in aggressive cancers. As genes are often deeply correlated in a complex systemic way (17), thousands of genes are likely to be associated with proliferation alone, which explains signatures' entanglement with RSS.

Yet, removing proliferation genes and its correlates entirely would nearly eliminate the outcome association of published and random signatures (2). Besides, the proliferation signal still remains strongly influential in breast cancer prognosis (18, 19). As such, robust feature-selection methods are necessary to ensure that the right genetic signals are extracted from noise.

2.2.2 Super-Proliferation Set (SPS)

The super-proliferation set (Supplementary Table 1) was derived by Goh *et al.* (4), through a meta-analysis of 47 signatures from independent breast cancer studies. Briefly, signatures that were more robust against the confounding effects of proliferation genes were identified, based on the increase in nominal p-value after removing proliferation genes. Then, genes which were supported by at least two or more of these robust signatures were selected as candidate genes for the super-proliferation set.

This feature-selection method is unique in that it does not simply employ the traditional filter, wrapper, or embedded methods (20), but rather a meta-analytical method that resembles a filter technique on independently selected features sets (i.e., 47 different studies) instead of within a single study alone. The proliferation genes in SPS also exhibit an additive power (i.e., prediction performance increases with number of genes used from the gene set), unlike proliferation genes that are not within SPS, demonstrating the effectiveness of this method in filtering out true confounders within the proliferation signal, producing a robust signature that is highly predictive of survival outcomes (4).

2.2.3 Prediction Analysis for Microarrays (PAM50)

The PAM50 gene signature has been widely used in distinguishing between intrinsic molecular subtypes of breast cancer, namely luminal A, luminal B, HER2-enriched, and basal-like (21). It has also been shown to add significant prognostic and predictive information to classical breast cancer parameters. Briefly, these four labels were derived from significant clusters that were generated from hierarchical clustering, with four additional groups that represented intermediary or heterogenous

states. With these labels, gene set reduction was then performed using the top 'N' *t*-test filtering method (22), resulting in 50 remaining genes (Supplementary Table 2).

2.3 Computing environment and functional analyses

All data analyses performed, and graphics produced are written and executed in R 4.1.2, and the relevant packages are detailed in the following sections. Functions were run with the default parameters unless specified otherwise.

To obtain functional gene annotations, official gene symbols were uploaded online to the Database for Annotation, Visualization and Integrated Discovery (DAVID) (https://david.ncifcrf.gov/home.jsp) (23, 24).

2.4 Datasets and data processing

Gene expression data was obtained from the Cancer Cell Line Encyclopaedia (CCLE) from https://data.broadinstitute.org/ccle/. This RNA-Seq data contained 57820 genes over 1019 different cell lines, of which 51 breast cancer cell lines were extracted for the purposes of this study. Metadata containing survival labels for the 51 breast cancer cell lines were obtained from Cell Model Passports (https://cellmodelpassports.sanger.ac.uk/downloads), while metadata containing the breast cancer subtype labels were obtained from the DepMap Portal (https://depmap.org/portal/download/custom/).

As the CCLE RNA-Seq data was provided in transcripts per million (TPM), the data was first log-normalised to transform the gene expressions to be on a similar scale and order of magnitude. The Ensembl gene IDs provided in the microarray annotations were then matched to their corresponding gene symbols via the biomaRt package. 'NA' values for the gene symbol were subsequently removed from the dataset, as well as genes that had no variation across the 51 breast cancer cell lines (i.e., zero variance rows). This left a total of 43651 genes for subsequent analyses.

The SPS gene set (81 genes) used in this study were directly obtained from the supplementary data by Goh *et al.* (4), while the PAM50 gene set (50 genes) was obtained from Parker *et al.* (21). Both

signatures were manually checked to ensure that all genes matched with their corresponding gene symbols in the CCLE dataset, otherwise they were replaced with their alternative gene aliases accordingly.

2.5 Principal component analysis (PCA) and PCA quadrant analysis

We used PCA as our main unsupervised learning technique for dimensionality reduction, while maximizing the variance of the projected data. This was done using R's 'prcomp' function, with both centring and scaling applied to the gene features, and the first two principal components (PC1 and PC2) were plotted using the 'ggplot2' package. The breast cancer cell lines were coloured according to their survival label (Normal, Non-Metastatic, Metastatic), and subtype label (Luminal, Luminal HER2 Amp, HER2 Amp, Basal A, Basal B). Following which, median lines for both PC1 and PC2 were drawn to halve the data along their respective PCs, while also segregating the plot into four distinct quadrants (top-left, top-right, bottom-left, bottom-right: TL, TR, BL, BR). We have previously shown that this two-axis median bisection method can be effective in identifying associations via an enrichment for a particular label (25). In so doing, each cell line was thus assigned and labelled with one of the four distinct quadrants in the PCA scatterplot.

Starting with only the SPS signature (81 genes) in the PCA, enrichment of any label in the quadrants was determined by either the Fisher's exact test or chi-squared test, depending on the sample sizes. To generate a null distribution for any derived numerical quantities, PCA was performed on random selections of 81 genes (excluding SPS genes) over 100000 iterations, to ensure that results were not confounded by the RSS phenomenon. We also plotted the magnitude and direction of primary contributions to each PC, to identify subsets of genes that were most important in segregating the cohort. Finally, the SPS PCA was compared with similar PCA plots arising from PAM50 genes, and all 43651 genes.

To better understand the differential expression of genes between the four PCA quadrants, the Wilcoxon 2-tailed rank-sum test was done between expression values of adjacent quadrants for every

SPS gene, while the Kruskal-Wallis test by ranks was used when comparing more than two quadrants simultaneously. These non-parametric tests were used to avoid any assumptions regarding the distribution of gene expressions. Following which, significant genes ($p < 0.05$) were visualised across the different cell lines and quadrants using heatmaps in the 'pheatmap' R package. Gene expression values were Z-normalised before plotting the heatmaps.

2.6 Feature selection method based on the Boruta machine learning approach

The Boruta algorithm was developed by Kursa *et al.* as a wrapper technique for feature selection and dimensionality reduction, using the random forest classification similar to other wrapper methods (forward selection, backward elimination, stepwise selection (26)). It uniquely employs 'shadow' features, which permutate the original features across samples, so that these features should no longer have any correlation with the response variable. During each iteration of the algorithm, a feature scores a 'hit' only if it has a higher importance (Z-score) than all the shadow features, which effectively constitute the background noise. As the Boruta algorithm uses a stochastic classifier (i.e., random forest), multiple iterations are usually done (~100) to ensure statistical validity of the results, especially when dealing with many features. At each iteration, features with a significantly higher Z-score than the maximum Z-score of the shadow features are selected, while those that are significantly lower are eliminated. Supplementary Figure 1 illustrates the features that are confirmed or rejected based on Z-score, as well as the minimal, average, and maximal performance of the shadow attributes. Tentative features which are not significantly different from the best shadow attribute are not shown.

With respect to gene expression data, Kursa also found that the Boruta algorithm was the most consistent in its selection of important features (27), when compared with other random forest-based methods such as Artificial Contrasts with Ensembles (RF-ACE) (28), recursive feature elimination (RFE), and regularised random forests. Thus, this provides strong justification for further analysis with the candidate genes from Boruta.

Using the 'Boruta' package in R, feature selection was performed on all 43651 genes to predict both survival and subtype. To reduce the pipeline complexity, both labels were converted into binary classification problems; survival was classified into 'Metastatic' and 'Non-Metastatic', while subtype was classified into 'Basal' and 'Non-Basal'. This allows for the classes to be relatively more balanced, while still maintaining sufficient distinction between them. Following which, a min-max normalisation was performed within each sample, ensuring that they all have the same range between 0 and 1. This method of normalisation preserves the distances between samples, which would be beneficial for the algorithm in distinguishing between classes.

To verify that the Boruta algorithm was effective in separating the class labels of our dataset, feature selection was preliminarily performed on all 51 samples, over 200 iterations. These features were then verified using subsequent PCA, importance analysis of genes, and random forest classification. The last of which was also benchmarked against SPS genes as well as all 43651 genes, for a total of 3 different models. For the random forest training, the 'caret' R package was used, and training was done using the in-built 'ranger' function within the package itself. Repeated 5-fold cross validation was done with 100 repeats, to ensure statistical validity of the model accuracy obtained. However, to reduce computational costs when training with all 43651 genes, the 'mtry' parameter was set to the approximate square root of the number of features (i.e., 208), while the other two models retained their default 'caret' parameters for grid search.

For actual testing, an 80:20 train-test split was first performed on the 51 breast cancer cell lines using the 'caret' package, ensuring that class proportions were approximately equal between the train and test groups. As opposed to the previous section, the Boruta algorithm was only performed on the train dataset, leaving the test group unseen by the algorithm. Both confirmed and tentative Boruta features were used for survival, but only confirmed features were used for subtype. This is due to the large difference in number of Boruta features between the two, and hence the Boruta features were selected such that the number of features is similar to SPS's (later elaborated in Table 4).

For the survival label, three models were trained using different subsets of genes (all genes, Boruta genes, and SPS genes), while for the subtype label, five models were trained (all genes, Boruta genes, SPS

genes, PAM50 genes, and a random 81-gene set). Like the preliminary trials, 'mtry' was set to a fixed value when training with all genes, while a grid search was done for the other models based on the default parameters. The entire process described above was performed over 200 iterations, using different train-test splits each iteration. Results were then visualised using ggplot2, and post-hoc analysis was done to explore the Boruta features that were selected over all 200 iterations.

2.7 Translating SPS quadrant-based annotations to new clinical datasets

Using the four quadrant labels derived previously from the CCLE data (i.e., TL, TR, BL, BR), we further annotated three independent breast cancer clinical datasets (GSE81538, GSE202203, TCGA-BRCA), assigning each sample with its quadrant as well. This was done by labelling each clinical sample with a corresponding CCLE cell line of minimum Euclidean distance. Additionally, for the GSE202203 and TCGA–BRCA datasets, survival outcome data was publicly available. Hence, the survival rates of the four quadrants were subsequently analysed as well.

2.8 Gene signature reversal

The methodologies employed by Wagner *et al*. (29) and Chen *et al.* (30) for signature reversal can be closely adapted for our quadrant analysis. Similar to our study, Wagner *et al.* (29) visualises the signature reversal through PCA scatterplots of gene expression spaces in mice models with dyslipidaemia, with the target reversal from the high-fat (high-risk) region to the low-fat (low-risk) region. They found that treatments that more effectively restore gene expressions to their physiological norm, would also be more successful in restoring physiological markers to their baselines (e.g., body weight, white adipose tissue levels, liver health). Correlating this to our PCA quadrants, a targeted movement from the highest-risk quadrant (BR) to lower-risk quadrants (TL, TR, BL), especially in the horizontal (PC1) direction, should lead to better physiological outcomes in breast cancer.

Chen *et al.* (30) also created gene expression signatures based on differentially expressed genes between normal and tumour samples, from a set of 978 'landmark' genes from the Library of Integrated

Network-based Cellular Signatures (LINCS). By observing the change in genetic profiles after distinct drug compounds were used, a Reverse Gene Expression Score (RGES) was computed that measured the compound's potency in reversing the gene expression of tumour samples to that of normal samples. It was also found that RGES was positively correlated with half-maximal inhibitory concentration (IC_{50}) and drug efficacy. Hence, with readily available IC_{50} data from CCLE itself, or compound gene expression profiles from LINCS, drugs that target the differential genes between quadrants while having minimal impact on non-differential genes would be suitable candidates for further investigation.

In our study, we combined the significant gene lists between adjacent SPS quadrants with compounds gene expression change data from LINCS and computed the RGES to find the potential drug treatment that has the potency to reverse significant genes in the adjacent quadrants. This can help to target potential drugs for high-risk breast cancer. LINCS has the gene expression change of 978 landmark genes for 66612 compounds. We only extract the records for the significant genes between adjacent quadrants (BLBR, BLTL, TRTL, TRBR) and calculate the gene reversal score respectively. The compounds with the top gene reversal scores are presented.

3 Results

3.1 SPS predicts survival better than PAM50

In Figure 1a, we observe that SPS has potential in separating both the survival and subtype labels. Using the PC medians (red lines) to separate the plot into equal halves makes these differences even more obvious; ideally, a null effect would result in equally distributed quadrants and label proportions. Yet, for the survival label, the top-left quadrant contains a larger proportion of non-metastatic cell lines (blue and green dots), while the bottom-right quadrant contains a larger proportion of metastatic cell lines (red dots), suggesting that these could represent 'good' and 'bad' prognostic outcomes, respectively. These two quadrants also seem to have fewer cell lines as compared to the remaining two quadrants (top-right and bottom-left), which conversely have almost equal proportions of metastatic and non-

metastatic cell lines, and could serve as 'intermediary' states between the 'good' and 'bad' outcomes. Should these quadrants be sufficiently distinctive in genetic profile (discussed further in next section), a targeted movement from worse quadrants (i.e., bottom-right) to better quadrants (i.e., top-left, top-right, or bottom-left) could signify better outcomes for the patient as well. The 51 breast cancer cell lines and their corresponding quadrants can be found in Supplementary Table 3.

For the subtype label, there is also a clear distinction between two groups, where basal cell lines (Basal A in red, Basal B in brown) are largely contained within the top half, while the non-basal cell lines (Luminal in blue, Luminal HER2 Amp in purple, HER2 Amp in green) are largely within the bottom half. This is to be expected due to the noticeably different genetic profile of the basal-like subtype, being the most reliably identifiable even by single sample predictors (31). However, it is interesting to note that the normal cell line (HMEL) clusters closely with the basal-like subtype, which is usually associated with poorer outcome. This basal-like classification was also derived by Jiang *et al.* (32).

Combining observations from Figure 1a and the top contributing genes from Figure 1b, we posit that PC1 primarily encodes the survival signature, while the subtype signature is mostly encoded within PC2. The subtype signature in PC2 can be observed from the fact that subtype label separation largely occurs across the PC2 median, between basal and non-basal subtypes. Comparing this with the PCA plots derived from using all 43651 genes as well as PAM50 genes only (Supplementary Figure 2), we observe that the subtype separation largely occurs along the horizontal axis (i.e., PC1), particularly between the basal and non-basal cell lines. Interestingly, the PCA for all genes already presents this distinction very clearly, without the need for any feature selection. This is further evidence of the triviality in distinguishing between these two classes, since the basal-like subtype is already easily identifiable (31).

For PC1, the survival signature can be explained not only by the survival label separation as explained earlier, but also by the top genes contributing to PC1, which all show a large positive correlation (0.76 – 0.9) with PC1. Based on DAVID, all 20 genes encode for phosphoproteins, which have been proposed as biomarkers in breast cancer when detected in extracellular vesicles (33). They are also largely associated with cell cycle (17 genes) and cell division (15 genes), as well as proliferation (9 genes), as can be seen in Table 2. As such, cell lines with worse outcomes (in the bottom-right quadrant and

hence above PC1 median) also have a higher expression of these survival-related genes, and the converse is also true for the top-left quadrant.

In terms of proportion of variance explained, it is also interesting to note that PC1 alone explains a large portion (42.36%) of the total variance in the data, further solidifying SPS's ability to extract out and prioritise the survival signal from the gene expression data. Together with PC2 (8.63%) encoding for subtype, the first two PCs in the SPS PCA can explain more than half of the total variance. From Supplementary Figure 3, it is also clear that our findings are not spurious or due to RSS, as the PC1 median for random 81-gene sets sits lowly at 9.54%, with none of the iterations coming anywhere close to the value achieved by SPS. While SPS's PC2 variance is not as high, it still sits far towards the right-tail end of the distribution, with only 7% of iterations exceeding its value.

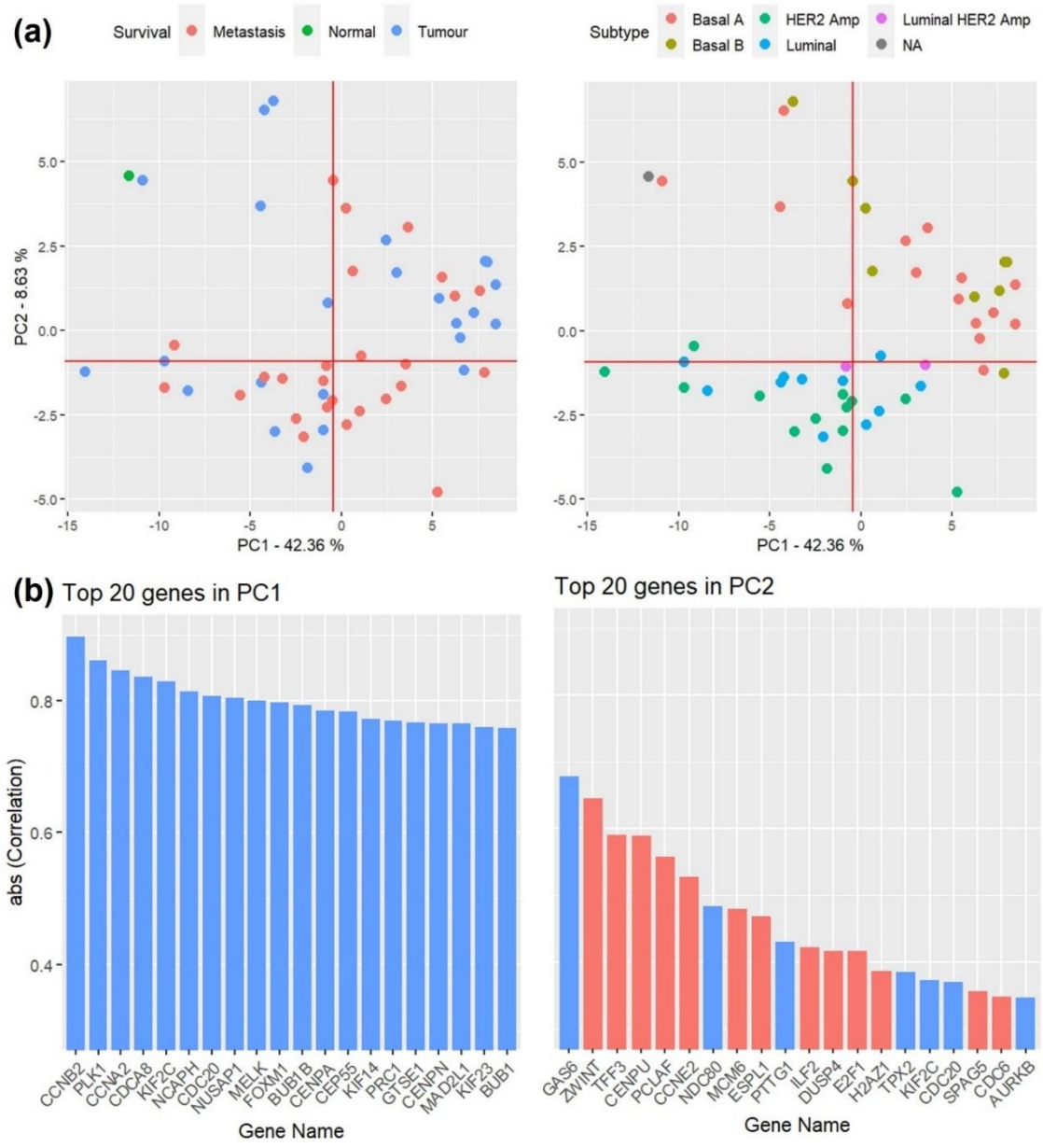

Figure 1: PCA analysis on SPS genes for 51 breast cancer cell lines. (a) PCA plots with survival and subtype labels, respectively. (b) Top 20 genes contributing to PC1 and PC2 and their absolute correlation coefficients. Blue and red bars represent positive and negative coefficients, respectively.

Table 2: Gene symbols for the PC1 top 20 genes in SPS, and their full names.

Gene Symbol	*Full Name*	*Cell Cycle*	*Cell Division*	*Proliferation*
BUB1B	BUB1 mitotic checkpoint serine/threonine kinase B	✓	✓	✓
BUB1	BUB1 mitotic checkpoint serine/threonine kinase	✓	✓	✓
CCNA2	Cyclin A2	✓	✓	-
CCNB2	Cyclin B2	✓	✓	-
CDC20	Cell division cycle 20	✓	✓	✓
CDCA8	Cell division cycle associated 8	✓	✓	-
CENPA	Centromere protein A	✓	✓	-
CENPN	Centromere protein N	-	-	-
CEP55	Centrosomal protein 55	✓	✓	-
FOXM1	Forkhead box M1	✓	-	✓
GTSE1	G2 and S-phase expressed 1	-	-	-
KIF14	Kinesin family member 14	-	-	✓
KIF23	Kinesin family member 23	✓	✓	-
KIF2C	Kinesin family member 2C	✓	✓	✓
MAD2L1	MAD2 mitotic arrest deficient-like 1	✓	✓	-
MELK	Maternal embryonic leucine zipper kinase	✓	-	✓
NCAPH	Non-SMC condensin I complex subunit H	✓	✓	-
NUSAP1	Nucleolar and spindle associated protein 1	✓	✓	-
PLK1	Polo like kinase 1	✓	✓	✓
PRC1	Protein regulator of cytokinesis 1	✓	-	✓

Performing a Fisher's exact test on the counts, we obtained a significant p-value of 0.0218 (Table 3). The post-hoc adjusted residuals (34) calculated also demonstrate that the largest deviations occur in the top-left and bottom-right quadrants, in the same directions as we have previously observed in the PCA plot. Their residuals are also significant since the cut-off Z-value is 1.96 for a two-tailed test at $\alpha = 0.05$. Out of 100,000 random iterations (median = 0.255), SPS also performs better than 95.9% of them (Supplementary Figure 4).

Table 3: Counts and adjusted residuals of survival labels (Non-Metastatic & Metastatic) for the four quadrants. Large residuals are highlighted in red.

		Top Left (TL)	*Top Right (TR)*	*Bottom Left (BL)*	*Bottom Right (BR)*	*Total*
Non-Metastatic	Count	7	10	7	1	25
	Residual	**2.37**	0.69	-0.79	**-2.25**	-
Metastatic	Count	1	8	10	7	26
	Residual	**-2.37**	-0.69	0.79	**2.25**	-

Comparing between adjacent quadrants, the horizontal pairs unsurprisingly had more significant genes (74 genes for TL-TR, 51 genes for BL-BR), as compared to the vertical pairs (25 genes for TL-BL, 15 genes for TR-BR). This is likely due to the much larger influence of PC1 than PC2. Observing the top 25 most significant genes in Supplementary Figure 5, the rightward shift towards worse outcomes was accompanied with the upregulation of most genes, as expected due to the strong proliferation signature associated with PC1. Conversely, only 2 genes were significantly downregulated – both DUSP4 (35-37) and TFF3 (38, 39) have conflicting effects on breast cancer in literature, and may be subtype-dependent as well. A full heatmap with all significant genes can be found in Supplementary Figure 6.

For the vertical pairs of quadrants, this upregulation with worse outcomes (via a downward shift) is not as evident, as we see more genes being downregulated as well. As seen in Supplementary Figure 7, for the TL-BL pair, two out of the 25 significant genes were downregulated, but a majority were nonetheless still upregulated with a downward shift from the TL to BL quadrant. However, for the TR-BR pair, a majority (9 out of 15 genes) are downregulated instead. Furthermore, the vertical pairs have milder differences between quadrants, possibly indicating that the association between gene expression and survival outcomes do not manifest as strongly within these pairs.

3.2 Meta-analysis approach for developing biomarker beats AI-based feature-selection method Boruta

Boruta is an advanced feature-selection algorithm leveraging on the powerful random forest ML model. It is thus useful to know if such new advanced methods can outperform SPS, which is obtained from meta-analysis approaches.

For the preliminary trials, the Boruta algorithm was run on all samples and genes, producing a reduced gene set containing 22 confirmed genes. We further confirmed that the Boruta algorithm works well in separating the survival label between the 'Non-Metastatic' and 'Metastatic' cell lines, with cross-validation accuracies averaging around 0.90, trumping the other two models trained on all genes as well as SPS (Figure 2 a/b). Nonetheless, this is to be expected, as Boruta systematically eliminates irrelevant features while retaining strong features that can discriminate between the two classes easily. What we have also shown here is that Boruta works well on high-dimensional expression data (43651 features), even though it relies on a stochastic random forest classifier to measure the importance of individual genes.

For the actual model, with the Boruta algorithm only training on 80% of the full data, we observed a stark change in the test accuracies for both the Boruta and SPS models. Boruta resulted in the lowest mean test accuracy of 0.597, followed by 0.629 when using all genes, and SPS had the highest mean accuracy of 0.6715 (Figure 2 c/d). Furthermore, SPS performed significantly better than both Boruta ($p = 5.76 \times 10^{-7}$) as well as all genes ($p = 0.00588$). For a more comprehensive evaluation of the performance metrics, the AUC distributions over 200 iterations were plotted (Figure 2e), where class probabilities are also considered. Similar to the test accuracies, the SPS model also had the highest mean AUC of 0.79, followed by Boruta ($p = 2.24 \times 10^{-11}$) and all genes ($p = 2.44 \times 10^{-7}$) at 0.68.

Comparing these back to Figure 2 a/b, while the Boruta algorithm is indeed working effectively in feature selection, it failed to generalise to the unseen data (i.e., the 20% test set in this case). In fact, it would have been more effective to use all genes as compared to the Boruta set of 22 genes ($p = 0.0210$),

likely because the Boruta set was unable to capture enough information from the limited training data that was necessary to distinguish between classes.

For the subtype label, we observed that all models performed reasonably well, with median test accuracies falling between 0.89 and 1 (Figure 2f). This could be due to the already distinguishable features between the basal and non-basal subtypes in the first place. The random model is demonstrative of this, as a set of 81 genes randomly chosen (see Supplementary Table 4) could also have a high median test accuracy of 0.89. Notwithstanding this, its median is still the lowest, as the other models were able to achieve perfect median test accuracies. Interestingly, PAM50 still performs remarkably better than the rest, with an even tighter distribution compared to the other models. This is to be expected as the signature was derived precisely to distinguish between subtypes in clinical data.

A post-hoc analysis of the Boruta features reveals a stark difference between the survival and subtype labels, which could explain the difference in its model performance. The total number of features differ by nearly an order of magnitude, with a mean of 20.5 for the survival label, and 179 for the subtype label (Table 4). To keep the Boruta gene sets similar in size to SPS, the total features were used for survival, while only confirmed features were used for subtype. While many features might be useful in model training, as is the case with subtype, it is also clear that there is a huge dependency for these features, with the top 30 genes appearing in all the 200 iterations (Supplementary Figure 8). In contrast, the top gene for survival only appears in 121 iterations and tapers off very quickly. Thus, even though the Boruta algorithm performed well in distinguishing subtypes compared to survival, each selected gene might not possess as much biological relevance or prognostic value, as they could be easily substitutable for the other high-frequency genes in the signature.

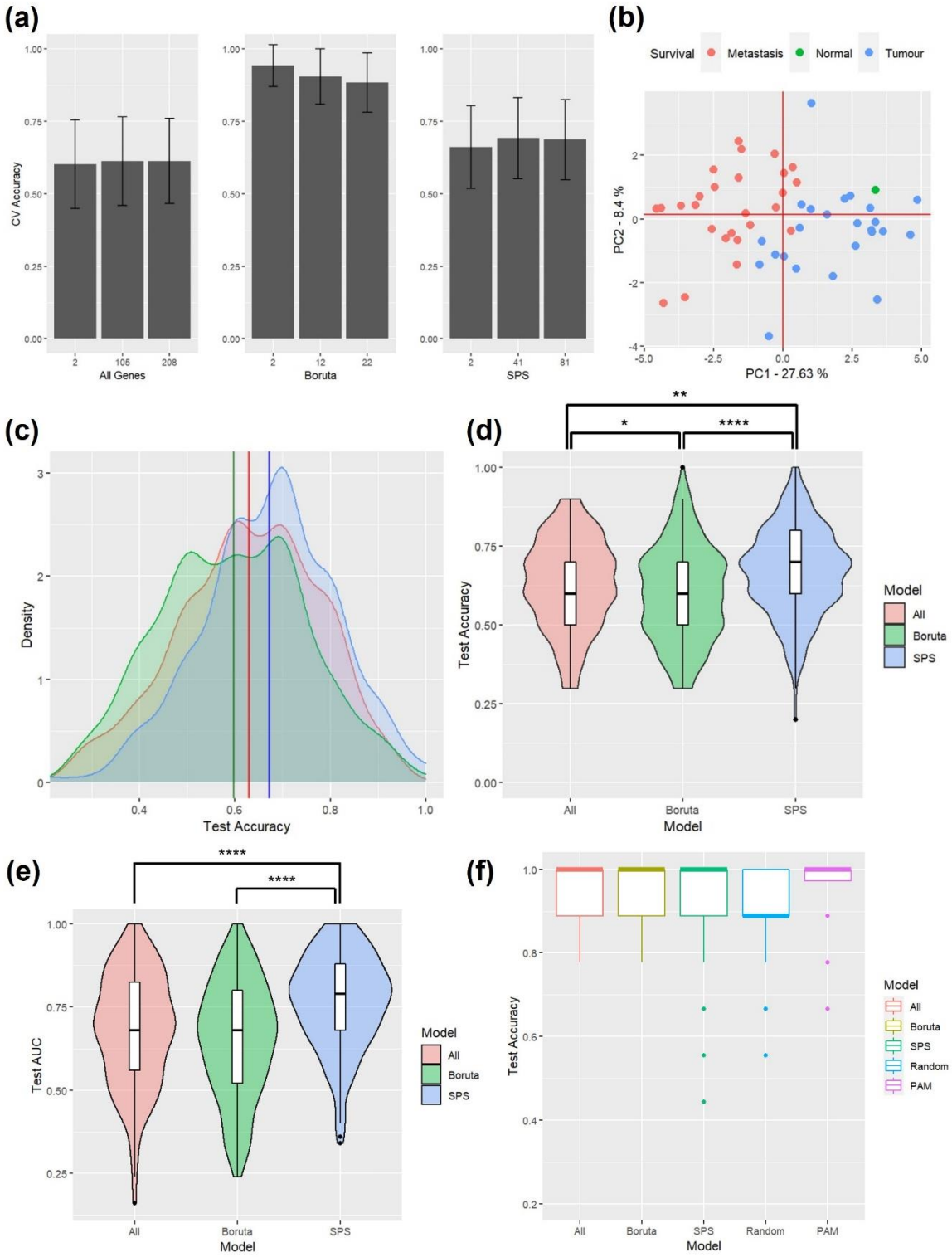

Figure 2: Results from training Boruta algorithm on all 51 breast cancer cell lines using the survival label. (a) Cross-validation accuracies when using all genes, Boruta, and SPS. (b) PCA plot using Boruta features. (c) Distribution of test accuracies in predicting survival for all genes,

Boruta, and SPS, over 200 iterations. Kernel density plots of the test accuracies are overlayed with coloured vertical lines showing the respective means. (d & e) Violin and box plots of the test accuracies and AUC values. (*p < .05, **p<.01, ****p<0.0001) (f) Boxplot of test accuracies in predicting subtype for all genes, Boruta, SPS, random 81-gene set, and PAM50, over 200 iterations.

Table 4: Descriptive statistics for the Boruta features obtained over 200 iterations, for both survival and subtype labels.

		Confirmed	*Tentative*	*Total*
Survival	Median	7	13	**20.5**
	Mean	7.5	13.3	**20.9**
	Max	16	33	**48**
	Min	1	1	**8**
Subtype	Median	**41**	138	179
	Mean	**41.5**	136.7	178.2
	Max	**54**	162	204
	Min	**34**	108	151

3.3 SPS quadrant-based annotations can be generalized towards survival predictions in other datasets

If the SPS-derived survival quadrants are useful (Figure 1), we should be able to extend the quadrant concept to other datasets to predict survival. As described previously, we used the four SPS quadrants to further annotate three independent clinical breast cancer datasets (GSE81538, GSE202203, TCGA-BRCA), and compared these translated annotations to the true prognostic outcomes of the clinical samples.

Figure 3a shows the distribution of the annotated SPS quadrant labels for those 3 independent datasets (GSE81538: 405 breast cancer clinical samples with no survival information from the Swedish National Breast Cancer Registry; GSE202203: 3207 additional clinical samples with survival outcomes from the SCAN-B study; TCGA-BRCA: 943 clinical samples with survival outcomes). We first observe that the label distributions are fairly consistent across the 3 datasets. Interestingly, LumB tends to have a larger proportion in the BR (worst prognosis) compared to Her2 and Basal, despite LumB usually being

associated with milder outcomes. This could indicate the possibility of separating between bad and worse outcomes even within the luminal subtypes.

For all 3 datasets, most Basal samples reside in the top-right quadrant, while most LumA samples localise in the bottom left quadrant; notably, these two quadrants have been associated with intermediary outcomes that fall between the best (TL) and worst (BR) quadrants. This subtype trend is similar to that observed in the cell lines as seen earlier in Figure 1 . Indeed, it appears that PC2 separates samples according to subtype: sorting subtypes by prognosis from worst to best, we observe that better prognostic outcomes (i.e., LumA, LumB) tend to have a smaller proportion of samples projected in the TR quadrant, and a larger proportion projected in the BL quadrant.

For clinical datasets 2 and 3 (GSE202203, TCGA-BRCA), since these come with accompanying survival outcome information, we analysed the survival outcomes for each quadrant based on both the death rate and ratio of dead/alive samples. The pie charts in Figure 3b illustrate the dead/alive ratios in each quadrant for the different subtypes, respectively. For the Basal subtype, samples in the TL quadrant have a higher death rate (42%, death/alive ratio = 0.73) compared to samples in the other quadrants. Thus, for this subtype, TL is likely the most dangerous quadrant; if samples are in this quadrant, there could be a tendency for a poorer outcome. However, there could still be limitations due to the small sample size and inherent differences between cell line and clinical data, which may explain why this finding does not match the supposed 'best' prognosis in the TL quadrant. Furthermore, patterns for other subtypes are not yet inferable. Nonetheless, this result shows that SPS quadrants could potentially capture sensitive features for the Basal subtype in predicting survival outcomes. For future work, we would test the SPS quadrants' performance on more datasets to give more comprehensive features for the quadrants, as well as build robustness in the survival prediction regardless of clinical subtype.

	Basal				Her2				LumB				LumA			
	Alive	Dead	Death rate	Dead/Alive	Alive	Dead	Death rate	Dead/Alive	Alive	Dead	Death rate	Dead/Alive	Alive	Dead	Death rate	Dead/Alive
BR	31	13	30%	0.42	77	23	23%	0.30	306	59	16%	0.19	111	11	9%	0.09
BL	10	3	23%	0.30	41	6	13%	0.15	163	37	19%	0.23	892	111	11%	0.12
TR	173	50	22%	0.29	86	23	21%	0.27	241	48	17%	0.20	76	3	4%	0.04
TL	15	11	42%	0.73	13	9	41%	0.69	31	7	18%	0.23	250	24	9%	0.10

	Basal				Her2				LumB				LumA			
	Alive	Dead	Death rate	Dead/Alive	Alive	Dead	Death rate	Dead/Alive	Alive	Dead	Death rate	Dead/Alive	Alive	Dead	Death rate	Dead/Alive
BR	23	4	15%	0.17	23	5	18%	0.22	70	6	8%	0.09	69	6	8%	0.09
BL	11	2	15%	0.18	12	3	20%	0.25	29	3	9%	0.10	281	40	12%	0.14
TR	109	13	11%	0.12	23	6	21%	0.26	61	15	20%	0.25	21	6	22%	0.29
TL	6	3	33%	0.50	5	1	17%	0.20	5	7	58%	1.4	69	5	7%	0.07

Figure 3: SPS quadrant-based label distribution in the new clinical datasets. (a) Quadrant label distribution for each subtype in the clinical samples. (b) Survival outcomes analysis based on annotated SPS quadrant labels.

3.4 Signature reversal analysis identifies 5-methoxytryptamine as potential drug useful for poor survival outcomes

Based on the significant genes for adjacent quadrants, we calculated the Reverse Gene Expression Score (RGES)(30) - a more negative RGES would indicate a higher likelihood to reverse the gene expression from a pathological state to a normal state. Table 5 lists the top 20 compounds that have the most negative RGES between adjacent quadrants (TRBR, TLBL, TRTL, BLBR). They are hence potential compounds for breast cancer treatment and gene reversal between quadrants. A more negative RGES

would suggest a greater influence of the compound in transitioning between quadrants, however the direction of this transition is not confirmed until further analysis is done.

Notably, there are several compounds that have effects in multiple quadrants. Taking 5-methoxytryptamine (5-MTT) as an example (highlighted red in Table 5), it is enriched in both TRBR and TLBL. Thus, it could be a potential drug for breast cancer treatment from the bottom quadrants to the top quadrants and vice versa. Furthermore, 5-MTT has been proven to exert antitumor, anticachectic, and immunomodulating effects under experimental conditions (40). This suggests that signature reversal analysis on SPS is a feasible strategy in drug discovery or even drug repositioning to find additional treatment options for high-risk breast cancer.

Table 5：RGES for the top 20 compounds on the 4 SPS significant gene lists between adjacent quadrants.

TRBR	sRGES	TLBL	sRGES	TRTL	sRGES	BLBR	sRGES
MW-SHH-250	-0.1109	5-methoxytryptamine	-0.0953	BRD-K76252019	-0.0374	MW-SHH-250	-0.0609
BRD-A84238007	-0.1104	6-hydroxytropinone	-0.0953	BRD-K22384978	-0.0364	BRD-A84238007	-0.0600
anastrozole	-0.1081	AMN-082	-0.0953	BRD-K19882533	-0.0362	BRD-K51644197	-0.0570
BRD-A93507363	-0.1063	BRD-A88878656	-0.0953	BRD-K86631041	-0.0359	BRD-K81164606	-0.0563
RO-15-4513	-0.1063	BRD-K07395346	-0.0953	BRD-K51390937	-0.0356	BRD-K35158088	-0.0563
BRD-K98803880	-0.1054	BRD-K49477212	-0.0953	BRD-K52178187	-0.0355	BRD-K43638827	-0.0563
syringic-acid	-0.1054	BRD-K51731619	-0.0953	BRD-K56301236	-0.0355	BRD-K89288521	-0.0563
BRD-K51644197	-0.1047	BRD-K68020183	-0.0953	BRD-K99597257	-0.0355	anastrozole	-0.0554
BRD-A73594579	-0.1047	hesperidin	-0.0953	BRD-K08138210	-0.0347	BRD-K18067885	-0.0552
AMN-082	-0.1044	BRD-A84238007	-0.0925	BRD-K22496535	-0.0347	SA-419172	-0.0551
BRD-A88878656	-0.1044	BRD-K32602441	-0.0920	BRD-K88581223	-0.0347	BRD-K66383562	-0.0548
BRD-K68020183	-0.1044	BRD-K98803880	-0.0914	BRD-K93062774	-0.0342	BRD-K77634909	-0.0539
BRD-K66383562	-0.1036	BRD-K34079378	-0.0912	alendronic-acid	-0.0340	BRD-K55082668	-0.0535
BRD-K77634909	-0.1032	MW-SHH-250	-0.0898	BRD-K36467636	-0.0335	BRD-K38596298	-0.0533
BRD-K34079378	-0.1030	BRD-A44244100	-0.0895	BRD-K20554239	-0.0334	BRD-K50269042	-0.0533
3,6-dimethoxyflavone	-0.1026	syringic-acid	-0.0895	SD-169	-0.0334	BRD-K67485291	-0.0533
5-methoxytryptamine	-0.1026	3,6-dimethoxyflavone	-0.0893	viomycin	-0.0333	BRD-K94719962	-0.0532
6-hydroxytropinone	-0.1026	AM-630	-0.0893	BRD-K37461197	-0.0331	BRD-K09701578	-0.0529
AM-630	-0.1026	anastrozole	-0.0893	BRD-K68338581	-0.0331	BRD-K47533918	-0.0528
BRD-A17664363	-0.1026	BRD-A17664363	-0.0893	BRD-K06370852	-0.0326	BRD-K67774729	-0.0525

4 Discussions

From the various analyses (PCA, quadrant analysis, machine learning) that were conducted, SPS was able to extract a clear survival signal in all instances. For the PCA analysis, we have shown that SPS prioritises the survival signal in the first principal component with an exceptionally high proportion of

variance explained, while the second principal component accounts for both the subtype and survival signal, giving rise to four quadrants with 'good', 'bad', and 'intermediate' outcomes.

The quadrant analysis via the Fisher's exact test supports this association further, demonstrating that the top-left and bottom-right quadrants in the PCA correspond to relatively 'good' and 'bad' outcomes, respectively. With these SPS-derived survival labels, a heatmap analysis of differential genes between adjacent quadrants provides a basis for signature reversal, which has proven to correlate with drug and therapeutic efficacy (30). Primarily, this reversal should be prioritised along the horizontal axis (i.e., leftwards toward better outcomes) due to the prominence of PC1 as compared to PC2, as well as the strong upregulation of survival/proliferation genes along this axis.

Furthermore, we also showed that SPS was able to outperform the Boruta algorithm in discriminating the survival classes, demonstrating that SPS's meta-analytical derivation from clinical data and published signatures is indeed robust and generalisable to cell line data as well (4). This also shows that a naïve application of powerful machine learning approaches (i.e., Boruta) to biological data might not always be the most appropriate, especially with smaller sample sizes. Furthermore, SPS was also able to discriminate between subtypes to a reasonable extent, as compared to a random 81-gene set. Hence, with respect to CCLE data, the SPS has proven to be a powerful signature that can discriminate between not only survival outcomes as its top priority, but also subtype concurrently as well.

Accordingly, we came up with a methodology to transfer the quadrant annotations (i.e., TL, TR, BL, BR) to unlabelled clinical datasets. The transferred annotations proved to be biologically interpretable, and were indeed able capture survival features for the 'Basal' subtype. This suggests that SPS can provide reproducible features to accurately estimate cancer prognosis and predict survival outcomes on a wide range of breast cancer samples, while the inferred annotations provide the ability to dynamically monitor breast cancer progression. Finally, by calculating the RGES between SPS quadrants, we could refine the treatment process by identifying and providing efficacious drugs that most effectively transfer samples between SPS quadrants, from quadrants with worse prognoses to those with better prognoses.

Future work could build on preclinical applications of the theory that was discussed in this project. For example, with regards to the quadrant analysis, potential drug therapies could be devised that are

able to target the appropriate quadrants via a signature reversal, and their effectiveness would then be evaluated through a time-dependent analysis of the patient or sample's movement through the quadrants. Also, since we have shown that SPS is able to derive survival labels independently (i.e., the quadrant system), it would be interesting to see whether such a method could potentially be applied to unlabelled breast cancer clinical datasets, or even other types of cancer via a similar analytical procedure. Finally, while the Boruta algorithm was shown to be not as effective when applied to the training sets in the CCLE breast cancer data, there could be value in exploring this algorithm with other larger datasets. Through a similar meta-analytical approach to SPS, and looking for consistent features selected across iterations as with Kursa's methodology (27), a powerful alternative signature to SPS could be possible.

5 Conclusions

In breast cancer, we find that SPS is superior in prioritising survival over subtype, when compared against PAM50 or even artificial intelligence-based feature selection based on Boruta. With this higher resolution 'progression' information from SPS, survival outcomes can then be accurately predicted and monitored in clinical datasets, to a greater degree than subtype information alone is able to provide. Furthermore, when coupled with signature reversal analysis, it is possible to identify potential drugs that can most effectively bring patients from high-risk regions to low-risk regions, paving the way for more targeted treatments in clinical settings.

6 Acknowledgements

This research/project is supported by the National Research Foundation, Singapore under its Industry Alignment Fund – Pre-positioning (IAF-PP) Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. WWBG also acknowledges support from a Ministry of Education (MOE), Singapore Tier 1 grant (Grant No. RS08/21).

7 Competing interests

The authors declare no conflicting interests, financial or otherwise.

8 Data availability

All relevant codes and data are available at https://github.com/reubenfoo/SPS_2022.

References

1. Sotiriou C, Pusztai L. Gene-Expression Signatures in Breast Cancer. New England Journal of Medicine. 2009;360(8):790-800.
2. Venet D, Dumont JE, Detours V. Most Random Gene Expression Signatures Are Significantly Associated with Breast Cancer Outcome. PLOS Computational Biology. 2011;7(10):e1002240.
3. Goh WWB, Wong L. Why breast cancer signatures are no better than random signatures explained. Drug discovery today. 2018;23(11):1818-23.
4. Goh WWB, Wong L. Turning straw into gold: building robustness into gene signature inference. Drug discovery today. 2019;24(1):31-6.
5. van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, et al. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002;347(25):1999-2009.
6. Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, et al. The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature. 2012;483(7391):603-7.
7. Kursa MB, Rudnicki WR. Feature Selection with the Boruta Package. Journal of Statistical Software. 2010;36(11):1 - 13.
8. Brierley JD, Gospodarowicz MK, Wittekind C. TNM classification of malignant tumours: John Wiley & Sons; 2017.
9. Early Breast Cancer Trialists' Collaborative Group. Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials. Lancet. 2005;365(9472):1687-717.
10. Piccart-Gebhart MJ, Procter M, Leyland-Jones B, Goldhirsch A, Untch M, Smith I, et al. Trastuzumab after Adjuvant Chemotherapy in HER2-Positive Breast Cancer. New England Journal of Medicine. 2005;353(16):1659-72.
11. Yersal O, Barutca S. Biological subtypes of breast cancer: Prognostic and therapeutic implications. World J Clin Oncol. 2014;5(3):412-24.
12. Dai X, Cheng H, Bai Z, Li J. Breast Cancer Cell Line Classification and Its Relevance with Breast Tumor Subtyping. J Cancer. 2017;8(16):3131-41.
13. Hayashi T, Urayama O, Kawai K, Hayashi K, Iwanaga S, Ohta M, et al. Laughter regulates gene expression in patients with type 2 diabetes. Psychother Psychosom. 2006;75(1):62-5.
14. Rinn JL, Bondre C, Gladstone HB, Brown PO, Chang HY. Anatomic demarcation by positional variation in fibroblast gene expression programs. PLoS Genet. 2006;2(7):e119.

15. Krishnan V, Han MH, Graham DL, Berton O, Renthal W, Russo SJ, et al. Molecular adaptations underlying susceptibility and resistance to social defeat in brain reward regions. Cell. 2007;131(2):391-404.
16. Manjang K, Tripathi S, Yli-Harja O, Dehmer M, Glazko G, Emmert-Streib F. Prognostic gene expression signatures of breast cancer are lacking a sensible biological meaning. Scientific reports. 2021;11(1):156-.
17. Barabási A-L, Gulbahce N, Loscalzo J. Network medicine: a network-based approach to human disease. Nature Reviews Genetics. 2011;12(1):56-68.
18. Karagiannis GS, Goswami S, Jones JG, Oktay MH, Condeelis JS. Signatures of breast cancer metastasis at a glance. J Cell Sci. 2016;129(9):1751-8.
19. Reyal F, van Vliet MH, Armstrong NJ, Horlings HM, de Visser KE, Kok M, et al. A comprehensive analysis of prognostic signatures reveals the high predictive capacity of the proliferation, immune response and RNA splicing modules in breast cancer. Breast Cancer Res. 2008;10(6):R93-R.
20. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507-17.
21. Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, et al. Supervised risk predictor of breast cancer based on intrinsic subtypes. J Clin Oncol. 2009;27(8):1160-7.
22. Storey JD, Tibshirani R. Statistical methods for identifying differentially expressed genes in DNA microarrays. Methods Mol Biol. 2003;224:149-57.
23. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc. 2009;4(1):44-57.
24. Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic Acids Res. 2009;37(1):1-13.
25. Goh WWB, Foo RJK, Wong L. What can scatterplots teach us about doing data science better? International Journal of Data Science and Analytics. 2022.
26. Chowdhury MZI, Turin TC. Variable selection strategies and its importance in clinical prediction modelling. Fam Med Community Health. 2020;8(1):e000262-e.
27. Kursa MB. Robustness of Random Forest-based gene selection methods. BMC Bioinformatics. 2014;15:8-.
28. Tuv E, Borisov A, Runger G, Torkkola K. Feature Selection with Ensembles, Artificial Variables, and Redundancy Elimination. J Mach Learn Res. 2009;10:1341–66.
29. Wagner A, Cohen N, Kelder T, Amit U, Liebman E, Steinberg DM, et al. Drugs that reverse disease transcriptomic signatures are more effective in a mouse model of dyslipidemia. Mol Syst Biol. 2015;11(3):791-.
30. Chen B, Ma L, Paik H, Sirota M, Wei W, Chua M-S, et al. Reversal of cancer gene expression correlates with drug efficacy and reveals therapeutic targets. Nature Communications. 2017;8(1):16022.
31. Weigelt B, Mackay A, A'Hern R, Natrajan R, Tan DSP, Dowsett M, et al. Breast cancer molecular profiling with single sample predictors: a retrospective analysis. The Lancet Oncology. 2010;11(4):339-49.
32. Jiang G, Zhang S, Yazdanparast A, Li M, Pawar AV, Liu Y, et al. Comprehensive comparison of molecular portraits between cell lines and tumors in breast cancer. BMC Genomics. 2016;17(7):525.

33. Chen IH, Xue L, Hsu C-C, Paez Juan Sebastian P, Pan L, Andaluz H, et al. Phosphoproteins in extracellular vesicles as candidate markers for breast cancer. Proceedings of the National Academy of Sciences. 2017;114(12):3175-80.
34. Haberman SJ. The Analysis of Residuals in Cross-Classified Tables. Biometrics. 1973;29(1):205-20.
35. Mazumdar A, Poage GM, Shepherd J, Tsimelzon A, Hartman ZC, Den Hollander P, et al. Analysis of phosphatases in ER-negative breast cancers identifies DUSP4 as a critical regulator of growth and invasion. Breast Cancer Res Treat. 2016;158(3):441-54.
36. Menyhart O, Budczies J, Munkácsy G, Esteva FJ, Szabó A, Miquel TP, et al. DUSP4 is associated with increased resistance against anti-HER2 therapy in breast cancer. Oncotarget. 2017;8(44):77207-18.
37. Balko JM, Cook RS, Vaught DB, Kuba MG, Miller TW, Bhola NE, et al. Profiling of residual breast cancers after neoadjuvant chemotherapy identifies DUSP4 deficiency as a mechanism of drug resistance. Nature Medicine. 2012;18(7):1052-9.
38. Ahmed ARH, Griffiths AB, Tilby MT, Westley BR, May FEB. TFF3 is a normal breast epithelial protein and is associated with differentiated phenotype in early breast cancer but predisposes to invasion and metastasis in advanced disease. Am J Pathol. 2012;180(3):904-16.
39. May FEB, Westley BR. TFF3 is a valuable predictive biomarker of endocrine response in metastatic breast cancer. Endocr Relat Cancer. 2015;22(3):465-79.
40. Lissoni P, Messina G, Rovelli F. Cancer as the main aging factor for humans: the fundamental role of 5-methoxy-tryptamine in reversal of cancer-induced aging processes in metabolic and immune reactions by non-melatonin pineal hormones. Curr Aging Sci. 2012;5(3):231-5.