National University of Singapore

Faculty of Science/University Scholars Program

UIS3921 Independent Study Module

Cell line selection using genomic profiles

Author:
Ng Wee Kiat Jeremy

Supervisor:
Professor Greg-Tucker
Kellogg

Academic Year 2013-2014 Semester 2

Contents

List of Figures

List of Tables

Abstract

Cancer cell lines serve as invaluable *in vitro* models for researchers. Today, there is a large number of available cell lines for the study of different cancer subtypes. Yet, no guidelines exist to aid the selection of cell lines in experimental studies. Using publicly available data from The Cancer Genome Atlas (TCGA) and Cancer Cell Line Encyclopedia (CCLE), Domcke et al proposed a scoring scheme using copy number variation data and somatic mutation data to select a most suitable cell line for the study of ovarian cancer. In this present study, we refine Domcke et al's scoring scheme further by introducing a new scoring scheme that also includes information from gene expression profiles as well as a weighted scoring component for somatic mutations. We then apply our method to identify a suitable cell line model for the study of *TGF-β* sensitive breast cancer.

1 Introduction

Breast cancer is one of the most common cancers worldwide, with an estimated 1,300,000 new cases and 450,000 deaths reported annually. Clinically, breast cancer is classified into three distinct subtypes based on their eostrogen receptor status. Namely, breast cancer is defined along the expression of the *HER*, *ER* and *PR* receptors. This classification has been helpful in devising therapeutic regimes, as different receptor statuses lead to differing response to chemotherapy regimes [1]. However, there are other pathways that are relevant in determining drug response. One such pathway is that of the *TGF-β* pathway, which is particularly important in determining patient response to antihormonal treatment in *ER*-positive breast cancer [2].

TGF-β is a growth factor that has a wide range of effects, including the suppression of tumorigenesis [3]. The canonical pathway of *TGF-β* involves the binding of *TGF-β* to *SMAD* proteins, in which the activated *SMAD* protein then serves as a transcription factor (Figure 1). Despite the fact

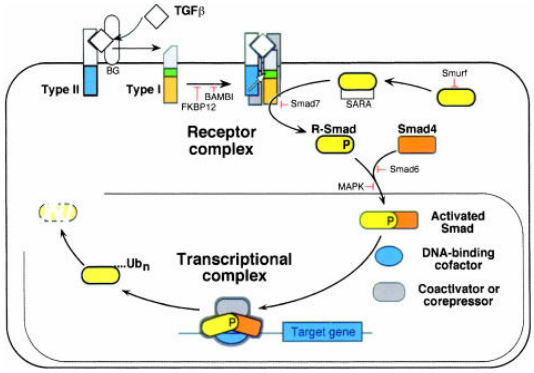

Figure 1: Diagram showing the *TGF-β/SMAD* pathway. Upon activation and dimerization of the *TGF-β* receptor, *SMAD*-4 is then activated, which is then transported into the nucleus to serve as a transcription factor. Taken from Massague and Wotton, 2000.

that *TGF-β* normally functions as a tumor suppressor, it is also recognized that in the later stages of cancer, *TGF-β* instead becomes a promoter of tumor progression, stimulating angiogenesis, inducing extracellular matrix degradation and inhibiting antitumor immune response [4,5]. Likewise, *TGF-β* is known to activate other signaling pathways to module the *TGFβ-SMAD* response [6]. Therefore, the *TGF-β* pathway is an attractive target of study as one can imagine a therapy which aims to reverse the oncogenic effect of *TGF-β*, and hence restoring *TGF-β* as a tumor suppressor.

Cell lines are popular *in vitro* models that have been instrumental in furthering our understanding of cancer biology. Apart from helping us gain insights into the mechanisms of disease progression, cancer cell lines are also

used extensively to screen for potential anti-cancer drugs [7]. Earlier studies have highlighted the many uncertainties that plague cell line selection, such as the risk of cross contamination [8] or cell lines of unknown origin that are mistaken for a wrong cancer type [9]. There is thus great interest in ensuring that the cell line models that are selected are as identical to the actual cancer type in question. One potential criteria for the selection of the most appropriate cell line is the use of genomic data as the basis by which cell lines are selected.

Although the use of genomic data to guide cell line selection is not entirely novel, earlier efforts were concentrated on using only one type of genomic data. For instance, Sandberg and Ernberg (2005) devised the Tissue Similarity Index (TSI) as a measure of how similar a cell line is to a tumor sample, in which they used Singular Value Decomposition to identify similar gene expression profiles [10]. Attempts to integrate multiple data types have been hampered by the lack of comprehensive characterization of tumors and cell line models. Indeed, the first attempt to integrate genomic data from multiple platforms was only made in wake of the extensive characterization of tumor samples by The Cancer Genome Atlas (TCGA), as well as the availability of genomic information of cell lines in the Cancer Cell Line Encyclopedia (CCLE). In their work, Domcke et al proposed a scoring scheme to guide cell line selection. In particular, they proposed cell line suitability to be defined by the formula

$$S = A + B - 2 \times C - \frac{D}{7} \tag{1}$$

,where A is the correlation with the mean CNA of HGSOC tumours, B is 1 for cell lines harbouring a TP53 mutation and 0 otherwise, C is 1 for hypermutated cell lines and 0 otherwise, and D is the number of genes mutated among the seven non-HGSOC genes recurrently altered only in the other ovarian cancer subtypes.

However, one noticeable inadequacy of the method proposed by Domcke et al is that gene expression patterns were not considered. This is particularly significant because it has been shown in other studies that gene expression patterns can be used to define cancer subtypes as well as the therapeutic response of patients [11–13]. Further scrutiny of the scoring scheme proposed by Domcke et al also yields another inadequacy; that is, the penalty for the differences in mutation profile between the cell line and the tumor samples are not weighted. As a consequence, the scheme applies an equal penalty to a cell line that has the absence of a near universal mutation and the absence of a rarer mutation (in the $\frac{D}{7}$ scoring component). In this present study, we attempt to refine the scoring scheme proposed by Domcke et al, addressing

the two inadequacies that we have identified in their earlier scoring scheme. We then apply our method to guide our selection of a suitable cell line for the study of *TGF-β* sensitive breast cancer.

2 Methods and Materials

2.1 Data acquisition

All data obtained are publicly available from the TCGA portal and the CCLE portal. For tumor data, we select only samples which have gene expression data in the form of RNASeq, CNV data and data on somatic mutation. There is no requirement for the samples to be normal-tumor matched. We use TCGA Level 3 data, which is extensively processed and summarized - with the exception of somatic mutations, in which data is provided up to only Level 2. In all, data for 936 samples from TCGA were downloaded. Similarly, processed CCLE data for gene expression, CNV arrays and somatic mutation were downloaded. Data for a total of 1037 cell lines were downloaded from CCLE.

2.2 Softwares

All work was done using the freely available R software (version 3.02). Additional packages were downloaded from either the Comprehensive R Archive Network (CRAN) or the *BioConductor* repository. The codes used for the analysis in this paper are made publicly available on *Github*.

2.3 Validation of batch effects

In particular, because TCGA data is processed across numerous batches, it is thus important to ensure that there are no distinct batch effects. We perform principal component analysis using the R package *FactoMineR*(Francois et al, 2013), and then cluster the samples according to the first and second principle component (PC1 and PC2, respectively). Visualization is done using a contour map that was generated using a *Python* module *matplotlib*.

2.4 Clustering

Clustering was performed using the Ward agglomerative algorithm as implemented in R, using a Euclidean distance matrix. Prior to cluster analysis, feature selection was first performed to reduce the number of gene features

that were used for clustering. Briefly, we first used the *TGF-β* gene set available from Broad Institute to select for genes that are known to be regulated by *TGF-β*. This reduced the total number of gene of interest from about 20,000 genes to 372 genes. Thereafter, all the RNASeq data was $\log_2$-transformed and z-normalized so that all the samples have a mean of 0 and a standard deviation of 1. Because there are genes in which there are no reads, we perform the log-2 transformation as follows

$$N = log_2(M + 1) \tag{2}$$

where N is a matrix of log-transformed reads, and M is the original matrix of reads from RSEM-normalization. Thereafter, we estimate the variability R of each gene g in sample s as follows

$$R = \frac{g_s}{\mu g} \tag{3}$$

Finally, the standard deviation of each gene across all the samples were calculated, and the top 20 most variable genes with the largest standard deviations were selected as features for clustering.

2.5 Analysis of gene expression

Following clustering, we then identified a sub-group of samples that are most representative of *TGF-β* sensitive tumor samples. The gene expression profile of the *TGF-β* was then estimated by taking the median of all the 372 genes known to be regulated by *TGF-β*. Similarity between the tumor model and cell line was measured using the Spearman ρ.

2.6 Analysis of copy number aberrations

The package *BioConductor* package *CNtools* (Zhang, 2013) was used to summarize CNV data from TCGA by genes. Thereafter, we constructed a representative profile by taking the medians of all the CNV counts for each gene across all the samples of interest. Following the lead of Domcke et al, similarity between the tumor model and the cell lines was measured using the Pearson r.

2.7 Analysis of mutations

The curated somatic mutation provided by TCGA is used for downstream analysis. In order to estimate the probability of the sample being from a

TGF-β sensitive origin based on the somatic mutation, we employ the Bayes theorem. Formally, let P(X) denote the probability of the sample being *TGF-β* sensitive, and P(Y) be the probability of a sample having a collection of mutations, we can denote P($X|\,Y$) as

$$P(X \mid Y) = P(Y \mid X) \times \frac{P(X)}{P(Y)} \tag{4}$$

We use the TCGA data to derive the required prior probabilities. In particular,

$$P(X) = \frac{n(TGF\beta \;\; sensitive \;\; samples)}{n(samples)} \tag{5}$$

In order to define P(Y), let the vector ϕ be the vector of mutation frequency in genes *1* to k ; that is,

$$\phi = (F_1, ..., F_k)$$

where F_i is the mutation frequency of gene i observed in the TCGA dataset. We define F_i as

$$F_i = \frac{n(mutations \;\; at \;\; loci \;\; i)}{n(samples)} \tag{6}$$

Assuming that mutations occur randomly, we can derive P(Y) for a given cell line as

$$P(Y) = F_1 \times F_2 \times F_3 \times F_4 \times ... \times F_k \tag{7}$$

The last probability term, P($Y|X$) represents the probability of a particular mutation profile given that the sample is known to be *TGF-β* sensitive. This is also derived from TCGA is defined similar to (7). However, unlike (7) where the mutation frequencies are calculated from all the samples, the mutation frequencies are now calculated only from the *TGF-β* responsive samples. For simplicity, we refer to the conditional probability P($X|\,Y$) as p.

2.8 Scoring and cell line selection

A suitable scoring function is one that will capture the similarities (and differences) in genomic profile between the tumor model and the cell lines. We define the score of a cell line, S_c as

$$S_c = \rho + r + p + 2 \tag{8}$$

, where ρ represents the correlation of gene expression, r represents the correlation of CNV and p represents the probability of a mutation given mutation profile. S_c can take on the range of value from 0 to 5, where a higher S_c suggests a better cell line model for the system of interest.

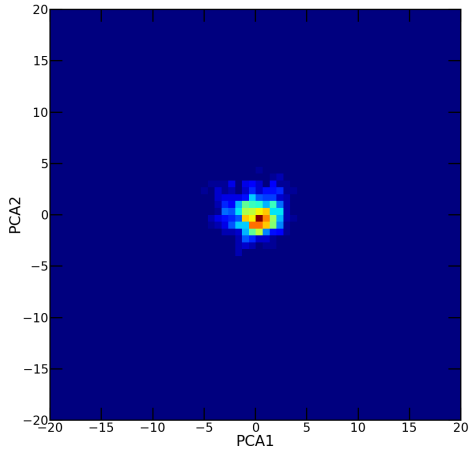

Figure 2: PCA contour plot showing the clustering of the RNASeq data from 936 TCGA samples of multiple batches. There is only one dominant cluster (red), hence indicating the absence of batch effects. Any differences can thus be attributed to biological differences.

3 Results and Discussion

3.1 TCGA data does not demonstrate batch effects

Batch effects serve as a potential confounder in downstream analysis, and must thus be detected and corrected prior to further downstream analysis. Batch effects arise due to technical differences - thus providing a source of variation that is non-biological in origin. This is an especially valid concern for TCGA data, which is processed in multiple batches. To test for batch effect, we perform principal component analysis, and then cluster along the first and second principal components (PC1 and PC2), and represent them in a contour plot (Figure 2). From Figure 2, we observe only 1 centroid, suggesting that all the 936 samples cluster well together and hence indicating an absence of batch effect. Therefore, no correction will be performed on the raw TCGA data prior to downstream analysis.

3.2 Clustering of *TGF-β* sensitive samples occurs along the expression of *SCGB1D2*

Following feature selection and clustering, we visualize the 20 genes using a heatmap (Figure 3).The samples can be clustered into 2 major clusters, indicated by the top color bars in Figure 3.

Interestingly, when we turn to the gene clustering, we notice that the main

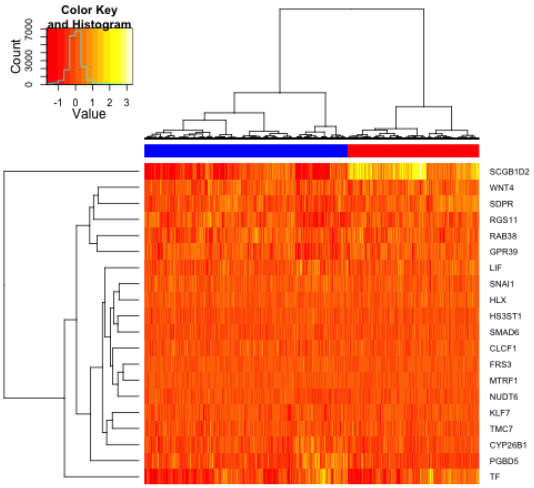

Figure 3: Heatmap showing the clustering and the expression of the 20 most variable genes in *TGF-β* response. The top dendrogram shows the sample clustering, while the side dendrogram shows the gene clustering.

feature that appears to distiguish between both groups is the expression of the gene *SCGB1D2*. Significantly, the expression of *SCGB1D2* is lower in the cluster color-coded blue. This is shown in Figure 4. *SCGB1D2* has

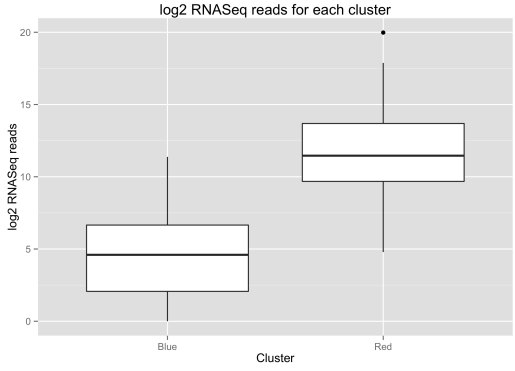

Figure 4: Boxplot showing the gene expression of the gene *SCGB1D2* in the two dominant clusters identified in Figure 3.The blue cluster (left) has a significantly lower ($p < 0.001$) expression of *SCGB1D2* as compared to the red cluster (right).

been shown to be downregulated in 59% of breast cancer cases [14]. This is observed also in the CCLE dataset, where 59 cell lines annotated as being of breast origin showed a significantly lower expression of the *SCGB1D2* gene (Figure 5). However, as noted in Figure 5, the expression of *SCGB1D2* is highly variable, with some breast cell lines having extremely high levels of *SCGB1D2* expression. This observation is consistent with the clustering shown in Figure 3, which shows high variability in expression that allows

us to distinguish between subtypes of breast cancers. In order for us to

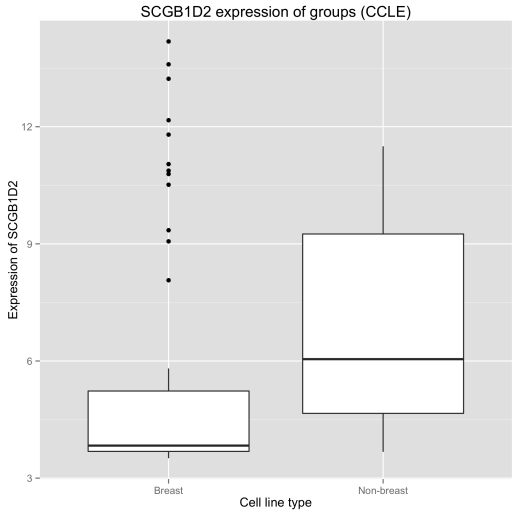

Figure 5: Boxplot showing the expression of *SCGB1D2* gene in cell lines of breast and non-breast origin. Consistent with earlier studies, the *SCGB1D2* gene is expressed at lower levels as compared to cell lines of non-breast origin.

determine which cluster best represents the *TGF-β* sensitive samples, we turn to examine the nature of interaction between *TGF-β* and *SCGB1D2*. From the gene sets provided by Broad Institute, we note that *SCGB1D2* is downregulated by *TGF-β*, and therefore, we selected the cluster with the lowest *SCGB1D2* expression to represent the *TGF-β* sensitive samples. From the clustering presented in Figure 3, we identify a total of 568 samples as being *TGF-β* sensitive.

3.3 Gene expression can be compared using Spearman ρ

We first consider the different platforms used by TCGA and CCLE; in particular, TCGA uses RNASeq to quantify gene expression while CCLE uses a single channel Affymetrix U133-2+ chip. Earlier studies by Guo et al [15] have demonstrated that data from RNASeq and single channeled Affymetrix data is highly concordant, with reported Spearman correlation coefficient of up to 0.9. Likewise, similar to Guo et al, we use the Spearman ρ as a measure to determine the similarity between the tumor model and a cell line. Figure 6 shows that the observed ρ of cell lines of breast origin (mean=0.75) is higher than those of non-breast origin (mean=0.71) ($p < 0.001$).

3.4 CNV analysis using Pearson r shows low correlation between cell lines and tumor samples

Copy number variations is a common hallmark of cancer, with agreement that recurrent aberrations occur at loci containing genes that are important for tumor development. In breast cancer, common aberrations includes the amplification of regions containing *PIK3CA*, *EGFR*, *FOXA1*, *HER2* and deletions of regions containing *MLL3*, *PTEN*, *RB1* and *MAP2K4* (TCGA, 2012). Similar to Domcke et al, we adopted the Pearson r as a measure of similarity to compare the CNV profiles of tumor and cell lines. Strikingly, although those cell lines of breast origins are sighnificantly more correlated (Figure 7), the mean correlation is only 0.27, which is low. Interestingly, when examining the CNV landscape of the tumors which have been identified as being *TGF-β* sensitive - and in particular, the 7 mentioned CNVs that are consistently aberrated, we note that are huge variability (Figure 8), with reported standard deviations between 0.23 and 0.35. Strikingly, the gene *MLL3*, which is reported to be deleted in breast cancer, was noted to have marginal amplifications instead (mean= 0.0266,median=0.0337,σ=0.273). When we compare the CNV landscape of the 7 genes between TCGA and CCLE data, we find that they are largely identical (Figure 9) Given these CNVs have been reported in breast cancer, we expect that the CNV profiles at these 7 loci be largely consistent between both TCGA and CCLE data, as shown in Figure

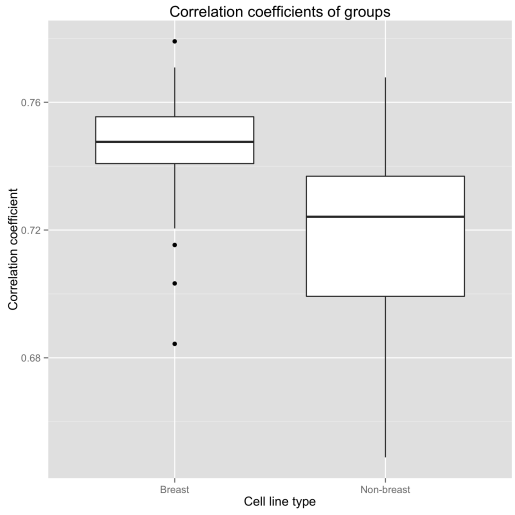

Figure 6: Boxplot showing the Spearman ρ between *TGF-β* sensitive tumor model and cell lines from CCLE. Cell lines from CCLE have been divided into being of breast (left) and non-breast (right) origin. The mean correlation of the samples of breast origin is 0.75, while the mean correlation of samples of non-breast origin is 0.71. The difference is reported to be significant ($p<0.001$, wilcoxon ranked test)

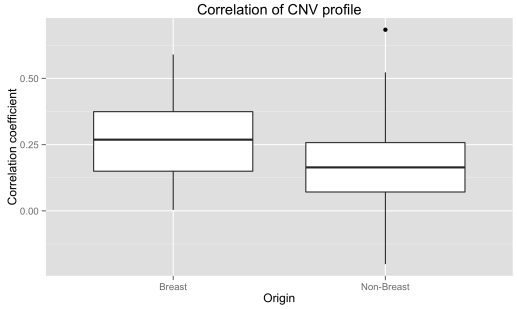

Figure 7: Boxplot showing the Pearson r between *TGF-β* sensitive tumor model and cell lines from CCLE. Cell lines from CCLE have been divided into being of breast (left) and non-breast (right) origin. The difference is reported to be significant (p <0.001, wilcoxon ranked test), although the overall correlation remains low.

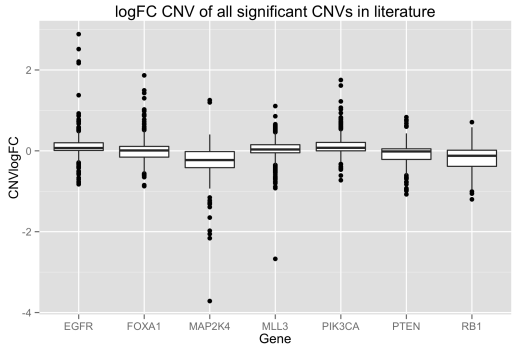

Figure 8: Boxplot showing the CNVs of 7 genes confirmed as being aberrated - either by amplification (*PIK3CA*,*EGFR*,*FOXA1*,*HER2*) or deletion (*PTEN*,*RB1*,*MLL3*,*AMP2K4*). Overall the copy number variations appear to be flat, but with large variations observed among all 568 samples.

9. Thus, the low correlation appears to be due to changes in the global CNV distribution which might be unique to the *TGF-β* sensitive system.

3.5 The mutational landscape of tumor samples is highly heterogeneous

Consistent with earlier reports, the mutation landscape of breast cancer is highly heterogeneous, with few mutations found in >10% of the samples (Figure 10), while table 1 below shows the most common occurring mutations that were observed in the TCGA dataset of 936 breast cancer samples.

That *TP53* is commonly mutated is hardly surprising, as it is estimated to be found in approximately 50% of human cancers. However, studies have shown that in the case of breast cancer, the frequency of *TP53* is about 20%, although this is highly dependent on the type and staging of the breast

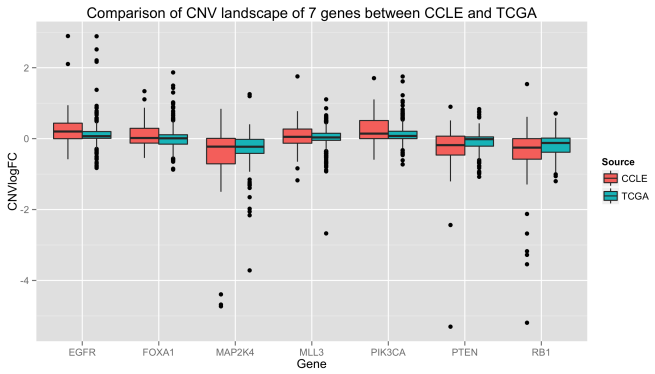

Figure 9: Boxplot showing the CNV landscape of CCLE and TCGA samples at 7 loci identified as being consistently aberrated- either by amplification (*PIK3CA*,*EGFR*,*FOXA1*,*HER2*) or deletion (*PTEN*,*RB1*,*MLL3*,*AMP2K4*). The distribution of CNVs appear to be identical, with relatively flat landscapes

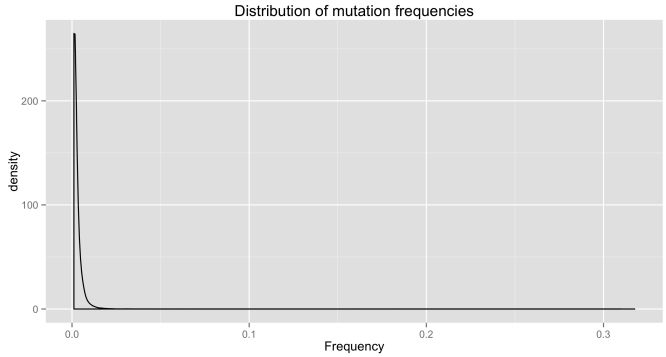

Figure 10: Density showing the frequency of somatic mutations occurring in breast cancer. The frequencies are derived from the TCGA dataset of 936 samples. The heterogeneity of the mutational landscape is shown by the left-skew of the distribution, where each mutation occurs at very small frequencies

cancer ([16]. This value of 20% agrees well with our observation that *TP53* mutations occurring in about 18% of breast cancer cases. Similarly, our observation of *PIK3CA* having the highest mutation is also consistent with literature, and has been estimated to be mutated in 20% - 40% of breast cancer cases [17]. Likewise, our observation of *TTN* being commonly mutated is also confirmed by earlier studies to be a potential oncogene [18].

3.6 Use of conditional probabilities in comparing somatic mutations lends an intuitive interpretation

The use of a conditional probability, $\mathrm{P}(X|\,Y)$ lends itself to intuitive interpretation. The probability, p can be understood to be the probability of a given

Table 1: Table showing the more common mutations in breast cancer, occuring in >10 % of all breast cancer samples

Mutation	Observed Frequency
TP53 Missense Mutation	0.1828877
TTN Missense Mutation	0.1593583
PIK3CA Missense Mutation	0.3176471

sample being *TGF-β* sensitive model given the spectrum of mutation that it possess. The frequency of the mutations occurring - either among breast cancer, or among *TGF-β* sensitive sample - is informed by the available data. A p of 1 - or a sure event - means that given the amount of information we have before hand, we can be sure that the cell line is of a *TGF-β* sensitive origin.

From the available data, we derive the priors and a few of the more important frequencies which we use for the computation of the posterior conditional probability. These values are shown in Table 2.

3.7 Bayesian conditional probabilities suggests that the mutational landscape of cell lines have undergone significant changes

The conditional probability, $\mathrm{P}(X|\,Y)$ (Eq. 4), denotes the probability of a cell line being *TGF-β* sensitive given that it contains a specific set of mutation. Table 3 shows the probability for the top 5 breast cancer cell lines, alongside their correlation to the CNV and gene expression profile of TCGA sample. Interesting, despite the relatively high correlation in terms of gene expression, the mutation scores ($\mathrm{P}(X|\,Y)$) are generally small. There are 2 possible reasons for such an observation. On one hand, it might be due to the fact that only 33 genes were used to derive the mutation scores.

Table 2: Table showing some crucial values involved in calculating the posterior conditional probability $\mathrm{P}(X|\,Y)$

Symbol	Value	Interpretation
$\mathrm{P}(X)$	0.6068	Probability of sample being classified as *TGF-β* sensitive

Table 3: Table showing the conditional probabilities P($X|Y$) of the top 5 BC cell lines, as well as correlations of CNV and gene expression that derive their total scores

Cell line	Correlation(CNV)	Correlation(GE)	P($X|Y$)	Score
HCC2157	0.4470866	0.7326566	6.492890e-02	3.244672
HCC1569	0.4692470	0.7374384	1.265671e-04	3.206812
HCC70	0.3631997	0.7548653	6.492890e-02	3.182994
SKBR3	0.3477014	0.7583189	6.492890e-02	3.170949
DU4475	0.4834161	0.6843618	6.492890e-02	3.167787

Table 4: Table showing the top 10 cell lines that best resembles *TGF-β* sensitive breast cancer

Cell line	Origin	Corr(CNV)	Corr(GE)	P($X|Y$)	Score
HCC2157	Breast	0.4470866	0.7326566	6.492890e-02	3.244
HCC1569	Breast	0.4692470	0.7374384	1.265671e-04	3.206
NCIH1666	Lung	0.4482049	0.7443055	9.492530e-06	3.192
HCC70	Breast	0.3631997	0.7548653	6.492890e-02	3.182
WM88	Skin	0.4581817	0.7234239	9.492530e-06	3.181
COLO704	Ovary	0.5040814	0.6743619	5.695518e-05	3.178
SKBR3	Breast	0.3477014	0.7583189	6.492890e-02	3.170
DU4475	Breast	0.4834161	0.6843618	6.492890e-02	3.167
SNU245	Biliary tract	0.3431889	0.7201723	6.492890e-02	3.128
A2058	Skin	0.3796327	0.7439474	9.492530e-06	3.123

3.8 Scoring scheme suggests HCC2157 as the best cell line model

Table 3 shows the scores of the top 5 breast cancer cell line models that would be *TGF-β* sensitive, and suggests the cell line HCC2157 as the best cell line model. The cell line is shown to have a high correlation in terms of gene expression and copy number aberrations, while also having the highest probability of being *TGF-β* sensitive given its mutational profile. Interestingly, our scoring scheme also suggests a few other cell lines that appear to be closely related to the system of our interest, namely, *TGF-β* sensitive breast cancer. Table 4 shows the top 10 cell line models from our scoring scheme

3.9 Future works

Further improvements to the scoring scheme can be conceived to also include epigenetic information as they become available for cell lines, as CCLE currently does not contain such information. Likewise, as another possible improvement to the scoring scheme involves the refinement of the mutation

scores. On this front, 2 possible improvements can be suggested. Firstly, the global mutation landscape of the cell lines should be considered in our scoring scheme, as this will be more informative in light of the genomic heterogeneity of cancers. Secondly, the somatic mutation landscape of cell lines tend to be different from those from tumor samples because they are grown in artificial environments, and a refined scheme might be cognizant of these differences.

4 Conclusion

In this present study, we have proposed modifications to the initial scoring scheme proposed by Domcke et al in order to use genomic data to guide the selection of cell lines in cancer research. Significantly, we introduced a weighted scoring scheme for somatic mutations, as well as taking into account the gene expression profile of both tumor and cell lines. We then applied our method to identifying cell lines that are most appropriate for the study of $TGF\text{-}\beta$ sensitive breast cancer.

References

[1] Rouzier, R. *et al.* Breast cancer molecular subtypes respond differently to preoperative chemotherapy. *Clinical cancer research* **11**, 5678–85 (2005).

[2] Knabbe, C. *et al.* Evidence that transforming growth factor-β is a hormonally regulated negative growth factor in human breast cancer cells. *Cell* **48**, 417–428 (1987).

[3] Yang, G. & Yang, X. Smad4-mediated TGF-beta signaling in tumorigenesis. *International journal of biological sciences* **6**, 1–8 (2010).

[4] McEarchern, J. a. *et al.* Invasion and metastasis of a mammary tumor involves TGF-beta signaling. *International journal of cancer* **91**, 76–82 (2001).

[5] Bierie, B. & Moses, H. L. Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer. *Nature reviews. Cancer* **6**, 506–20 (2006).

[6] Zhang, Y. E. Non-Smad pathways in TGF-beta signaling. *Cell research* **19**, 128–39 (2009).

[7] Yamori, T. Panel of human cancer cell lines provides valuable database for drug discovery and bioinformatics. *Cancer chemotherapy and pharmacology* **52 Suppl 1**, S74–9 (2003).

[8] Gillet, J.-P., Varma, S. & Gottesman, M. M. The clinical relevance of cancer cell lines. *Journal of the National Cancer Institute* **105**, 452–8 (2013).

[9] Domcke, S., Sinha, R. & Levine, D. Evaluating cell lines as tumour models by comparison of genomic profiles. *Nature* **4** (2013).

[10] Sandberg, R. & Ernberg, I. Assessment of tumor characteristic gene expression in cell lines using a tissue similarity index (TSI). *Proceedings of the National Academy of Sciences of the United States of America* **102**, 2052–7 (2005).

[11] Veer, L. J. V. *et al.* Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530–535 (2002).

[12] Finetti, P. *et al.* Gene Expression Profiling for Molecular Characterization of Inflammatory Breast Franc. *Cancer Research* **1**, 8558–8565 (2004).

[13] Finetti, P. *et al.* Gene Expression Profiling Identifies Molecular Subtypes of Inflammatory Breast Cancer Franc. *Cancer Research* **65**, 2170–2178 (2005).

[14] Zafrakas, M. *et al.* Expression analysis of mammaglobin A (SCGB2A2) and lipophilin B (SCGB1D2) in more than 300 human tumors and matching normal tissues reveals their co-expression in gynecologic malignancies. *BMC cancer* **6** (2006).

[15] Guo, Y. *et al.* Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PloS one* **8**, e71462 (2013).

[16] Pharoah, P. D., Day, N. E. & Caldas, C. Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis. *British journal of cancer* **80**, 1968–73 (1999).

[17] Cizkova, M. *et al.* PIK3CA mutation impact on survival in breast cancer patients and in ERα, PR and ERBB2-based subgroups. *Breast cancer research : BCR* **14**, R28 (2012).

[18] Greenman, C. *et al.* Europe PMC Funders Group Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2009).