genes

MDPI

Article

Identification of Differentially Expressed Genes between Original Breast Cancer and Xenograft Using Machine Learning Algorithms

Deling Wang [1,2,†], Jia-Rui Li [3,†], Yu-Hang Zhang [1], Lei Chen [4], Tao Huang [1,*] and Yu-Dong Cai [3,*]

1 Institute of Health Sciences, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China; wangdl@sysucc.org.cn (D.W.); zhangyh825@163.com (Y.-H.Z.)
2 Department of Medical Imaging, Sun Yat-sen University Cancer Center, State Key Laboratory of Oncology in South China; Collaborative Innovation Center for Cancer Medicine, Guangzhou 510060, China
3 School of Life Sciences, Shanghai University, Shanghai 200444, China; jiaruili@shu.edu.cn
4 College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China; chen_lei1@163.com
* Correspondence: tohuangtao@126.com (T.H.); cai_yud@126.com (Y.-D.C.); Tel.: +86-021-6613-6132 (Y.-D.C.)
† These authors contributed equally to this work.

Received: 3 January 2018; Accepted: 6 March 2018; Published: 12 March 2018

Abstract: Breast cancer is one of the most common malignancies in women. Patient-derived tumor xenograft (PDX) model is a cutting-edge approach for drug research on breast cancer. However, PDX still exhibits differences from original human tumors, thereby challenging the molecular understanding of tumorigenesis. In particular, gene expression changes after tissues are transplanted from human to mouse model. In this study, we propose a novel computational method by incorporating several machine learning algorithms, including Monte Carlo feature selection (MCFS), random forest (RF), and rough set-based rule learning, to identify genes with significant expression differences between PDX and original human tumors. First, 831 breast tumors, including 657 PDX and 174 human tumors, were collected. Based on MCFS and RF, 32 genes were then identified to be informative for the prediction of PDX and human tumors and can be used to construct a prediction model. The prediction model exhibits a Matthews coefficient correlation value of 0.777. Seven interpretable interactions within the informative gene were detected based on the rough set-based rule learning. Furthermore, the seven interpretable interactions can be well supported by previous experimental studies. Our study not only presents a method for identifying informative genes with differential expression but also provides insights into the mechanism through which gene expression changes after being transplanted from human tumor into mouse model. This work would be helpful for research and drug development for breast cancer.

Keywords: Monte Carlo feature selection; breast cancer; random forest; patient-derived tumor xenograft

1. Introduction

According to epidemiological data from the World Health Organization (WHO), breast cancer has been confirmed to be the most frequently diagnosed cancer in women from developed and developing countries and is accompanied with rising expectancy of average life span [1,2]. Early in 2011, more than 508,000 women died from breast cancer, accounting for more than 58% in developing countries [1]. However, the incidence of breast cancer is higher in Western Europe (89.7/100,000) is higher than those in Eastern Africa (19.3/100,000) and most developing countries (40/100,000); this finding reflects

the specific regional distribution of breast cancer incidence [3] and suggests the need for development of precise and effective methods for diagnosis and treatments.

To identify a cure for breast cancer, researchers have applied five levels of research, namely, molecular, cellular, histopathological, animal model, and clinical levels [4]. Among these levels, the animal model level not only confirms the conclusion obtained from studies on the other levels as an integrated creature but also lays the foundation for further clinical studies [5,6]. In studies on this level, mice are a typical and widely used species because of their high degree of genetic similarity to human beings and simple cultivation requirements. However, murine tumors considerably differ from human tumors at both genetic and molecular levels [7]. Therefore, a murine tumor model cannot thoroughly reflect the characteristics of human tumors. To solve this limitation, scholars have developed patient-derived tumor xenograft (PDX) model with various modified subtypes. In general, PDX mouse model is produced by directly transplanting cancerous tissues from a unique patient into an immune-deficient mouse model; this model can be used for studies on the biological, pharmacological, or clinical characteristics of such tumor tissues [8,9]. In comparison with conventional xenograft models using cell lines, the PDX model retains the molecular signatures and cancerous heterogeneity, which includes the cancer stem cells of the original tumor [10].

Although the PDX mouse model maintains most of the characteristics of the original tumor, genomic and transcriptional variations induced by murine microenvironment selection have been characterized. In 2014, researchers from University of North Carolina (USA) found that in pancreatic and colorectal cancer, the mutation frequencies of two oncogenic genes, namely, *KRAS* and *PIK3CA*, were greatly altered by the murine microenvironment across passages; hence, at the genomic level, the PDX mouse model cannot completely maintain the characteristics of in situ tumor tissues [11]. At the transcriptomic level, a study on PDX tumor expression patterns of 58 patients with eight different tumors reported that 48 genes were differentially expressed between the PDX mouse model and the tumor tissues of the patients [12].

In this study, we obtained and used the gene expression data from a recent study [13] on PDX mouse models and original breast cancer to identify differentially expressed genes, which are called informative genes. We collected a total of 657 PDX and 174 human tumors. By using Monte Carlo feature selection (MCFS), we evaluated the relevance of the gene expression in distinguishing the original breast cancer and the PDX models. Subsequently, we compared several algorithms for building classifiers including random forest (RF), rough set-based rule learning, support vector machine (SVM) and dagging. We found that building classifier using RF algorithm with the 32 identified informative genes could achieve the best performance, obtaining the highest Matthews coefficient correlation (MCC) value of 0.777. This result demonstrated the predictive power of the 32 identified informative genes in distinguishing PDX tumors from original human tumors. Furthermore, we applied rough set-based rule learning method to detect human-interpretable rules to reveal the interactions between the 32 informative genes. These genes rules, and interactions were found to be supported by previous studies through a literature review.

2. Materials and Methods

Our entire data set contained 831 tumor cells, including 657 PDX and 174 tumor cells. The 831 tumor cells were encoded into feature vectors by the expression levels of 69 genes. The 69 genes we used were derived from the work by Lawson et al. [13]. To investigate metastatic breast cancer, Lawson et al. did a literature survey and collected a small set of genes that were highly possible as the key genes in breast cancer and clearly involved in breast cancer related functions, such as stemness, pluripotency, epithelial-to-mesenchymal transition (EMT), mammary lineage specification, dormancy, cell cycle and proliferation. Based on the curated gene set, they designed a special single-cell array called Biomark 96.96 dynamic array [13] which can measure the expression level of 116 genes. In our study, since some genes were not expressed in many tumor cells and did not play important roles as

indicated by the literature, we only analyzed 69 genes by filtering those with missing expression in more than half of tumor cells.

Our method for identifying differentially expressed genes consisted of three steps: (i) applying MCFS method [14–16] to rank features; (ii) using incremental forward selection (IFS) and RF [17,18] methods to construct prediction model; and (iii) employing rough set-based rule learning [19] to detect interactions between differentially expressed genes.

2.1. Data Preparation and Feature Construction

The expression levels of 116 genes in 1293 tumor cells from three patients with breast cancer and the corresponding PDX models were obtained from a recent publication [13]. These genes are involved in stemness, pluripotency, epithelial-to-mesenchymal transition, mammary lineage specification, dormancy, cell cycle, and proliferation [13]; these processes suggest their crucial roles in tumor development. For the missing data in the original dataset, we filtered out those genes with missing data in more than half of tumor cells and those tumor cells with missing data in more than half of 116 genes. The rest of the missing data retained in the dataset were imputed using the nearest neighbor averaging method [20,21]; this filter process resulted to 69 genes and 831 tumor cells, in which 657 PDX mouse tumor cells and 174 original human tumor cells were found. The expression levels of the 831 tumor cells were used as features to encode each cell in this study.

2.2. Monte Carlo Feature Selection Method

MCFS method is designed to select informative features for supervised classifiers by randomly constructing a large number of tree classifiers from an original training dataset. MCFS assigns higher relative importance (*RI*) to a feature if this feature is selected by more tree classifiers. Suppose the total of $s \times t$ classification trees are constructed by selecting s subsets of m features ($m \ll d$, where d is the total number of features), and constructing t trees for each of the s subsets. Each of the t trees is built and trained by a random selection of training and test sets using the original training set. The *RI* of a feature can be measured by estimating the overall number of splits involving the feature in all nodes of all trees constructed. Particularly, the *RI* of a feature g can be estimated by the following equation

$$RI_g = \sum_{\tau=1}^{st} (wAcc) \sum_{n_g(\tau)}^{u} IG(n_g(\tau)) \left(\frac{\text{no. in } n_g(\tau)}{\text{no. in } \tau}\right)^{v}, \quad (1)$$

where $wAcc$ is the weighted accuracy over all samples, $IG(n_g(\tau))$ is the information gain of the node $n_g(\tau)$, $(\text{no. in } n_g(\tau))$ is the number of samples in node $n_g(\tau)$, $(\text{no. in } \tau)$ is the number of samples in tree τ, and u and v are fixed real numbers.

In this study, we performed MCFS ranking by using the MCFS software package [14]. MCFS program can output a list in which features were ranked according to their *RI* values. The ranking list can be formulated as

$$F = [f_1, f_2, \ldots, f_N], \quad (2)$$

where N is the total number of features (N = 69 in this study).

2.3. Incremental Forward Selection Method and Random Forest Classification

We used MCFS method to evaluate the importance of the 69 features, resulting in a feature list F. However, determining the optimal number of features to be used is difficult. Therefore, IFS [22–24] and RF methods [25,26] were employed to tell how many features should be used.

IFS method first constructed a series of candidate feature sets, denoted as $F_1, F_2, \ldots, F_N$, where $F_i = \{f_1, f_2, \ldots, f_i\}$ by incrementally adding features to precedent feature set according to their orders in the ranking list. For example, F_1 was composed of only the top 1 feature in the list F, F_2 was composed of the top 1 and top 2 features in the list F, and so on. To select the optimal feature set, i.e., determining

the optimal features to be used, RF classification algorithm was then performed on all candidate feature sets. The classification performance of each candidate model was evaluated by cross-validation [27,28] (see Section 2.5 for details). The candidate model achieving the best performance was selected as the optimal prediction model. Candidate feature set that corresponded to the optimal model was determined as the optimal feature set. Features contained in the optimal feature set were actually the informative genes identified to be differentially expressed between PDX and human tumor cells.

2.4. Rough Set-Based Rule Learning

After obtaining the informative genes, the rough set-based rule learning algorithm can be applied to detect interaction between informative genes. The detected interactions were represented as rules. A rule is a terminology to describe a relation between rule conditions (the left-hand-side of the rule) and the rule outcome (the right-hand-side). For example, in our study, a rule can be presented as IF–THEN relationship based on gene expression: IF Gene1 = *high* AND Gene2 = *low* THEN type = *human cell*. Details for creating rough set-based rule model are presented below, including an introduction of rough set theory, and the repeated incremental pruning to produce error reduction (RIPPER) algorithm which was used to generate decision rules based on rough set theory.

Rough set is described briefly as follows. A decision (classification) system $\mathcal{A}$ can be defined as:

$$\mathcal{A} = (U, A \cup \{d\}), \tag{3}$$

where U denotes all the objects, A represents all the condition attributes (features), and d is the decision attribute (class label). An attribute $a \in A$ can be regarded as a function such that $a(x)$ is the value of the attribute a of object x. Generalized decision attribute ∂_A can be defined to replace d, to deal with the situation where the values of all the condition attributes of some objects are the same, but the values of the decision variable d of these objects are different [29]. A function, $discerns(a, x, y, d)$, is defined as the situation in which object x and y are discernable using attribute a in term of the generalized decision variable ∂_A. A symmetric $|U| \times |U|$ matrix, called discernibility matrix M_A^d, is defined as follows

$$M_A^d[i,j] = M_A^d(x_i, x_j) = \left\{ a \in A \middle| discerns(a, x_i, x_j, d) \right\} \; i, j = 1, 2, \ldots, |U|, \tag{4}$$

The entry $M_A^d(x_i, x_j)$, i.e., the (i,j)th entry of the matrix M_A^d, contains all the attributes that can separate object x_i from object x_j. For example, an entry $M_A^d(x_i, x_j) = \{a_1, a_2, a_5\}$ means that x_i and x_j can be distinguished by a_1, a_2 or a_5.

A Boolean product-of-sums (POS) function, called discernibility function $f_A^d(x)$ that is relative to object x, can be computed from the row x of M_A^d as follows

$$f_A^d(x) = \prod_{y \in U} \left\{ \sum a^* \middle| a \in M_A^d(x,y) \text{ and } M_A^d(x,y) \neq \varnothing \right\}, \tag{5}$$

where a^* corresponds to the membership of attribute a. Function $f_A^d(x)$ is composed of some conjunctions and each conjunction contains some terms corresponding to a non-empty entry of the row x of M_A^d. For example, $f_A^d(x) = (a_1^* \vee a_3^* \vee a_5^*) \wedge (a_1^* \vee a_2^*)$ contains two conjunctions indicating that two entries of the row x of M_A^d are not empty with one set as $\{a_1, a_3, a_5\}$ and another set as $\{a_1, a_2\}$. Finding the set of all the prime implicants of $f_A^d(x)$, all the reducts, $RED(\mathcal{A}, x, d)$, that are relative to an object x can be determined, where a reduct is a minimal set of features that can separate x from other objects equally well with the full set of attributes A. A reduct preserves the boundaries between the approximation regions defined by the rough set [29]. A discernibility function $g_A^d(U)$ can be defined to discern all objects as follows

$$g_A^d(U) = \prod_{x \in U} f_A^d(x), \tag{6}$$

Accordingly, the reducts $RED(\mathcal{A}, d)$ of the whole decision system are the prime implicants of $g_A^d(U)$. The Boolean POS function $f_A^d(x)$ and $g_A^d(U)$ can often be considerably simplified according to multiplicative idempotence and absorption [29]. However, finding all reducts using $f_A^d(x)$ or $g_A^d(U)$ is non-deterministic polynomial (NP)-hard, and for many applications only one reduct is needed. In our work, a heuristic method called Johnson Reducer algorithm [29] is applied to find a single reduct which is generally close to minimal size.

Johnson Reducer [30], which is based on a discernibility function, first initializes the set of the reduct $R(\mathcal{A}, x, d)$ to be empty. Then (1) find the attribute a that appears most frequently in the remaining conjunctions of the discernibility function; (2) add attribute a to $R(\mathcal{A}, x, d)$; (3) remove all the conjunctions that contain a. Repeat (1), (2) and (3) until all conjunctions are removed from the discernibility function.

Based on the reduct, the RIPPER algorithm is applied to generate decision rules. RIPPER, proposed by Cohen [31] in 1995, is a rule learning algorithm which is capable of handling large noisy datasets effectively. RIPPER is the improved version of Incremental Reduced Error Pruning (IREP) [32] which combines both the separate-and-conquer technique used first in the relational learner FOIL [33], a system learning Horn clauses from data expressed as relations, and the reduced error pruning strategy proposed by Brunk and Pazzani [34]. RIPPER algorithm is described briefly in Figure 1.

Initialize a set E to be the training set.

Choose a class C that contains least instances:

Initialize a rule R to have an empty left-hand side that predicts C;

Split E into growing and pruning sets;

While there are positive samples (instances of C) in the growing set, or the description length (DL) is 64 bits greater than the smallest DL found so far, or the error rate is greater than 50%

Until R is perfect (or no more attributes to add):

For each attribute a not included in R, and for each value v:

Consider $a = v$ to add to the left-hand side of R;

Choose the a and v that have the highest Foil's information gain;

Add $a = v$ to R;

Prune R using reduced error pruning.

Remove the instances covered by R from the growing set.

Global optimization strategy is applied to further prune the rule.

Figure 1. The repeated incremental pruning to produce error reduction (RIPPER) algorithm. RIPPER starts from an empty rule and splits the training set into a growing set and a pruning set. Then it repeatedly grows rules that achieve the highest Foil's information gain using the growing set and prunes the rules using the pruning set until certain conditions are met. Finally, a global pruning is applied to prune the rules to gain the final rule set.

In our study, we fulfilled the detection of rules by using the Johnson Reducer algorithm [29], which was also implemented in the MCFS software package [14,35].

2.5. Measurements

Tenfold cross-validation [27,36–46] was used to evaluate the prediction performance. For each fold, prediction performance was measured by the MCC. Final results were summarized as the 10-fold average. MCC considers the sample numbers among classes and is generally regarded as a balance

measurement to indicate prediction accuracy even though sample numbers among classes are of very different sizes. In this study, we adopted MCC for the performance measurement because the numbers of the two classes (human cell and PDX cell) were of great imbalance (174 vs. 657).

Given the involved n samples, denoted by $s_1, s_2, \ldots, s_n$, and N classes, represented by 1, 2, ... , N. True classes of samples were defined as matrix $Y = (y_{ij})_{n \times N}$, where $y_{ij} = 1$ if the s_i belongs to class j; otherwise, this variable is set to 0. Predicted classes of samples were defined as matrix $X = (x_{ij})_{n \times N}$ $X = (x_{ij})_{n \times N}$, where $x_{ij} = 1$, if s_i is predicted to be class j; otherwise, $x_{ij} = 0$. Thus, MCC is defined as

$$MCC = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathrm{cov}(X, X)\mathrm{cov}(Y, Y)}}, \tag{7}$$

where cov(X,Y) was covariance function of X and Y, which can be computed by

$$\mathrm{cov}(X, Y) = \frac{1}{N}\sum_{k=1}^{N} \mathrm{cov}(x_k, y_k) = \frac{1}{N}\sum_{i=1}^{n}\sum_{k=1}^{N} (x_{ik} - \overline{x}_k)(y_{ik} - \overline{y}_k), \tag{8}$$

where x_k and y_k denote the kth column of X and Y, respectively; $\overline{x}_k$ and $\overline{y}_k$ denote the mean values of numbers in x_k and y_k, respectively.

The range of MCC is between -1 and 1. With high MCC value, the classifier yields a good performance (1 means the given classifier yields a perfect classification, 0 indicates a classification no better than random prediction, and -1 represents a total misclassification).

3. Results

In this study, we collected the expression levels of 69 genes in 831 tumor cells, which included 657 PDX and 167 human tumor cells. By using MCFS, we ranked the 69 genes according to their *RI* on the tumor cell classification. Then, by applying IFS and RF methods, we drew the IFS-curve, which illustrated the relationship between the number of features used for building the model and the corresponding performance measured by the MCC value (see Figure 2). As shown in Figure 2, when using the top 32 genes to build the model, the prediction performance can reach the highest MCC value of 0.777. Details of the IFS curve can be found in Supplementary Material I.

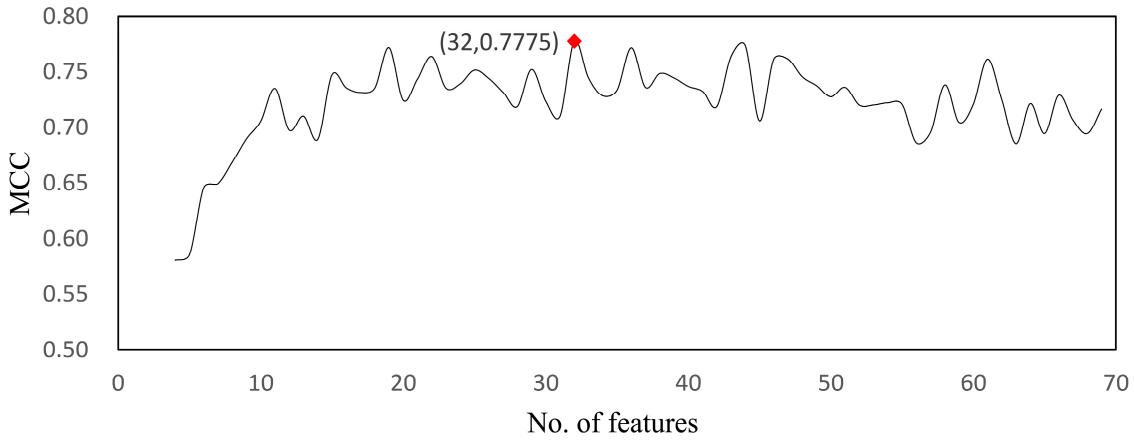

Figure 2. Incremental forward selection (IFS) curve illustrating the relationship of the prediction performance and the number of features incorporated in building the prediction engine. MCC: Matthews coefficient correlation.

Furthermore, we compared our method to other similar methods, including the rough-set based method as mentioned above, the SVM [47] and the dagging [48]. The comparison result was shown in Table 1. As can be seen in Table 1, using RF with the top 32 features achieved the highest MCC, the

32 genes were identified as informative genes for predicting PDX and human tumor cells. Also, we performed enrichment analysis (multiple testing corrections) on the 32 identified genes. The significant Kyoto encyclopedia of genes and genomes (KEGG) and gene ontology (GO) enrichment results of 32 genes with false discovery rate (FDR) < 0.05 can be found in Supplementary Material II.

Table 1. Comparison on performance of different models. The RF model with top 32 features achieved the highest MCC (0.777). RF: random forest; SVM: supporter vector machine; MCFS: Monte Carlo feature selection.

Model	Feature No.	MCC	Sensitivity	Specificity	Accuracy
RF	32	**0.777**	0.996	0.672	0.929
RF	57 (MCFS cutoff)	0.695	0.995	0.598	0.905
Rough Set	57 (MCFS cutoff)	0.665	0.950	0.680	0.893
SVM	41	0.695	0.995	0.563	0.904
Dagging	58	0.599	0.996	0.436	0.878

To somewhat open the RF model, we applied the rough set-based rule learning method to detect human-interpretable IF–THEN rules to reveal the interactions between the 32 informative genes. Finally, seven IF–THEN rules were detected (see Table 2). Details of the seven IF–THEN rules were provided in Supplementary Material III, including support values, coverage, ranking of the rules, etc. Moreover, in the Section 4, comprehensive interaction analysis as well as visualization of the seven IF–THEN rules were provided.

Table 2. Seven rules quantitatively defining the criteria for classification. These rules were produced by the MCFS software package. If a tumor cell satisfies the first six criteria, it would be classified into Human tumor; otherwise, it would be classified into PDX tumor.

Rules	Criteria	Classification
Rule 1	KRT19 $\geq$ 1.939224	Human tumor
	KRT5 $\leq$ 0.148786	
	CDH3 $\leq$ 0.868794	
Rule 2	EMP1 $\geq$ 4.237572	Human tumor
	CAV2 $\leq$ 1.610886	
Rule 3	TP53 $\leq$ 0.291193	Human tumor
	CXCR4 $\geq$ 4.367387	
	TGFBR2 $\leq$ 1.868461	
Rule 4	CXCR4 ≤ -2.474571	Human tumor
	CD44 $\geq$ 0.086944	
	PTEN $\geq$ 0.143515	
	VIM $\leq$ 0.647694	
Rule 5	PARP2 $\geq$ 3.111536	Human tumor
Rule 6	PLCB4 $\geq$ 3.744729	Human tumor
	AKT1 ≤ -0.070679	
Rule 7	Other conditions	PDX tumor

PDX: Patient-derived tumor xenograft.

4. Discussion

Based on the detailed expression profile of PDX mouse tumor tissue and the corresponding clinical tissue, we applied our newly presented computational method for the identification of core differentially expressed genes between PDX model and clinical sample. The 32 important genes used for constructing RF prediction model are listed in Table 3. All these genes have been identified to be differentially expressed in clinical tumor samples and PDX mouse model by recent publications, which validated the efficacy and accuracy of our predicted genes. Furthermore, most of these differentially

expressed genes contribute to tumorigenesis; this finding implies the potential difference of oncogenic mechanisms in PDX mouse model from in situ. This study at the transcriptomic level may not only deepen our understanding on the tumorigenesis in PDX tumor mouse model but also further reveal the potential restriction of PDX tumor mouse model due to the differential expression profile between PDX tumor tissue and clinical samples. The detailed analysis of optimal genes can be seen below.

Table 3. Thirty-two differentially expressed genes identified as optimal features.

HUGO Symbol	HUGO Name	*RI*
EMP1	epithelial membrane protein 1	0.16895404
PARP2	poly(ADP-ribose) polymerase 2	0.15058246
KRT19	keratin 19	0.12158414
MUC1	mucin 1, cell surface associated	0.11115772
CXCR4	C-X-C motif chemokine receptor 4	0.07917199
PROM1	prominin 1	0.06480689
ERBB2	erb-b2 receptor tyrosine kinase 2	0.048957534
ERBB3	erb-b2 receptor tyrosine kinase 3	0.04209958
KRT5	keratin 5	0.037512265
ID4	inhibitor of DNA binding 4, HLH protein	0.03389286
PTEN	phosphatase and tensin homolog	0.029668033
NTRK2	neurotrophic receptor tyrosine kinase 2	0.022596486
PGR	progesterone receptor	0.020494139
TP53	tumor protein p53	0.019557578
CDH3	cadherin 3	0.01846532
BMI1	BMI1 proto-oncogene, polycomb ring finger	0.013900218
TGFBR2	transforming growth factor beta receptor 2	0.013375987
CCNB1	cyclin B1	0.013296658
PLCB4	phospholipase C beta 4	0.013219586
CLDN4	claudin 4	0.013182897
CXCL12	C-X-C motif chemokine ligand 12	0.010324035
EGFR	epidermal growth factor receptor	0.010273729
CD44	CD44 molecule (Indian blood group)	0.009676576
LGR5	leucine rich repeat containing G protein-coupled receptor 5	0.008659011
NOTCH4	notch 4	0.007799821
BCL2	BCL2, apoptosis regulator	0.007518955
CAV2	caveolin 2	0.007474113
VEGFC	vascular endothelial growth factor C	0.006789302
TGFBR1	transforming growth factor beta receptor 1	0.006149265
VIM	vimentin	0.005953075
TGFB2	transforming growth factor beta 2	0.005226418
KRT8	keratin 8	0.00506866

HUGO: Human gene nomenclature; *RI*: relative importance.

4.1. Differentially Expressed Genes

Among our predicted differentially expressed gene, *EMP1* is at the top rank. In encoding a tumor-associated membrane protein, this gene has been confirmed to be downregulated during the passage and transplantation of breast cancer tumor cells [12,13].

The second gene *PARP2* has also been predicted to have differential expression patterns in our study. According to recent publications, this gene has been confirmed to contribute to DNA repair regulation in tumor tissues; this finding is quite common and significant in tumor malignant tissues [49,50]. Studies based on PDX model further confirmed that this gene mediates the chemotherapy resistance of breast cancer and has a differential expression pattern between breast cancer PDX model and corresponding tumor tissue in situ [50,51]. Previous findings demonstrated that during the transplantation processes, the biological process of DNA repair is inhibited, which indicates that *PARP2* might be lowly expressed in PDX model [52].

The third ranked gene *KRT19* has been widely reported to contribute to breast cancer tumorigenesis by its involvement in the organization of myofibers [53,54]. This gene is differentially expressed in PDX model compared with the original tumors not only in breast cancer but also in pancreatic cancer and hepatocellular carcinoma [55–57]. The upregulation of this gene helps maintain the structural integrity of epithelial cells [58]. During the transplantation of primary tumor into the mouse model, epithelial cell integrity became unstable in the PDX mouse model; this finding indicates that this gene can be downregulated in PDX model [58]. Interestingly, the homolog of *KRT19* and *KRT5* is also in the optimal feature set. The products of these two genes may form a specific dimer and contribute to specific biological processes as a whole [59]; this observation suggests that this gene can also be differentially expressed in the PDX mouse model.

In encoding a specific membrane glycoprotein, *MUC1* is usually expressed in the epithelial female cancer subtypes (e.g., ovarian cancer and breast cancer) and regarded as the biomarker of cancer cell stemness [60–62]. A recent study on the stemness of cancer cells confirmed that during the passage of PDX mouse model, the stemness of cancer cells is gradually shaped by the murine microenvironment [62]; this finding indicates that the expression of *MUC1* can be altered in this progress. Similarly, another gene involved in the stemness maintenance of cancer cells, *PROM1* [55,63], can also have differential expression patterns.

CXCR4 encodes a specific G protein coupled receptor, which has been widely reported to be expressed on the immature $CD34^+$ hematopoietic stem cells [64,65]. Recent studies on breast cancer found that the expression profiles of *CXCR4* are very diverse among the primary tumor of breast cancer, the metastatic tumor tissues, and the PDX tumors filtered by specific murine microenvironment [66,67].

In our results, two specific Erb-B2 Receptor Tyrosine Kinase (ERBB)-pathway-associated genes, i.e., *ERBB2* and *ERBB3*, have been predicted as potential distinctive biomarkers. ERBB family signaling pathway has been widely reported to contribute to transcriptional regulation during the tumorigenesis of breast cancer [68]. Further studies on the detailed expression pattern of these two genes in clinical tumor tissues and PDX models confirmed that after the selection of murine tumor microenvironment and PDX model passage, the expression patterns of *ERBB2* and *ERBB3* have been systematically regulated for further adaption to the murine mouse microenvironment [55,69].

Another gene, *ID4*, has also been identified as a breast-cancer-associated gene, which participates in the regulation of basic helix-loop-helix transcription factors [70,71]. Although no direct evidence is found on the differential expression pattern in paired PDX mouse model and tumor tissues in situ, the expression of creatine kinase is altered due to the microenvironment of PDX mouse model [72,73]. As the negative regulator of creatine kinase, the expression pattern of *ID4* may also be changed in a different microenvironment.

A recent study on single-cell transcriptome of PDX mouse model reported a series of genes with differential expression patterns in the PDX mouse model compared with the original tumors [13]; this series included *PTEN*, *NTRK2*, *PGR*, and *TP53*, which are identified in this study as the differentially expressed genes.

4.2. Rules of Quantitative Expression Level Requirements

As we have mentioned above, we detected seven specific rules to distinguish primary tumor tissues and PDX tumor tissues based on the expression levels of 14 genes. In the seven rules, the first six provided criteria for classifying a tumor cell into the original human tumor, which also represents the primary tumor-specific gene expression patterns (see Table 2). Among the 14 genes, seven genes, i.e., *KRT19*, *EMP1*, *CXCR4*, *CD44*, *PTEN*, *PARP2*, and *PLCB4*, were required to have high expression levels in the original tumor cells, whereas others should have low expression levels. As we discussed above, some of these genes, such as *KRT19*, *EMP1*, and *PARP2*, have evidence from previous studies, which indicate their higher expression levels in the original tumors compared with PDX tumor cells [12,13,52,58].

Further, to facilitate our understanding of the seven rules, we visualized the seven rules by using Ciruvis, a web-based tool for rule networks and interaction detection using rule-based classifiers, as suggested in the literature [74]. The result is shown in Figure 3, from which we found that, strong interactive relationships involved genes of *KRT19*, *KRT5* and *CDH3* (interactions of dark red color). Thus, we provided full analysis on the interactions among *KRT19*, *KRT5* and *CDH3*.

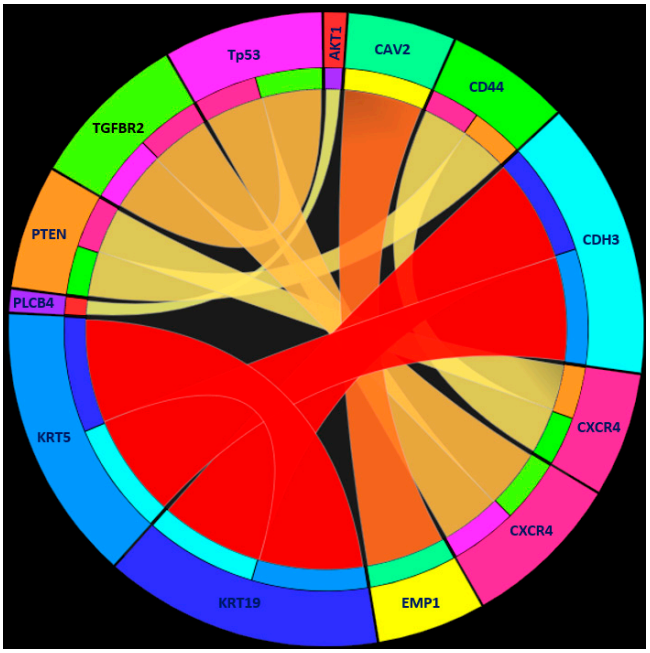

Figure 3. Rule networks for the seven detected rules generated by Ciruvis [74]. For the purpose of visualization, Ciruvis assigned specific colors for genes and placed genes into a circle. Meanwhile, Ciruvis indicated the strength of interaction among genes by using the degree of red color on the interactions. The darker the red color is, the stronger the interaction among genes.

First, the dimerizing of Krt5 with Krt14 has been confirmed to be involved in the regulation of EMT in tumorigenesis [75–77]. Further studies on the protein–protein interactions confirmed that this heterodimer could also interact with another component of cytokeratins, Krt6b [78]. As a functional mediator, Krt6b connects Krt5 with Krt19 through protein–protein interaction [79]. Krt6b shows negatively correlated expression pattern with Krt19 during tumorigenesis [80]. Therefore, the strong interaction between *KRT19* and *KRT5* has been indicated by previous experiments.

Second, the Ciruvis figure also show strong correlations between *KRT5* and *CDH3*, which encodes a member of the cadherin superfamily. A previous study in 2010 [81] reported that *BRCA1* repressed the expression of various cytokeratins including Krt5. It is also known that BRCA1 product manipulate the expression pattern of both *KRT5* and *CDH3* [81], suggesting the correlation and co-regulation of the expressions of these two genes.

Third, the correlation of *KRT19* and *CDH3* is suggested by the previous studies which reported the participation of these two genes in the essential proliferation and metastasis associated pathway MAPK/ERK cascade in tumorigenesis [82,83]. Furthermore, *KRT19* shows co-expression pattern with various *CDH3* associated genes like *CDH1* [84], *EGFR* [85] and *CTNNB1* [54] in certain pathological conditions, indicating that there may be potential biological relevance between *KRT19* and *CDH3*.

In summary, these findings strongly support the correlations of both functions and expression patterns among the three genes *KRT19*, *KRT5* and *CDH3*, which are all related to EMT associated pathological processes in human tumors, but not in PDX mouse model.

5. Conclusions

Through a new computational method, we identified the differentially expressed genes and quantitative rules that can classify the PDX tumor and the original tumor with high accuracies. Multiple well-known oncogenes and tumor suppressors are differentially expressed; this finding indicates the risk of using those genes as therapeutic targets in breast cancer through PDX mouse model. Therefore, our study not only provides a functional tool to distinguish the primary tumor from PDX tumor by gene expression levels but also reveals the detailed gene expression alterations and shaping characteristics in murine microenvironment. In this work, we employed SVM, RF, dagging and rough set classifiers due to the small samples problem. Some of the latest classifiers, such as ensemble classifier [86], will be tested in future work.

Supplementary Materials: The following are available online at www.mdpi.com/2073-4425/9/3/155/s1. Supplementary Material I: Incremental forward selection by using Random Forest, Supplementary Material II: The significant KEGG and GO enrichment results of 32 identified genes with FDR < 0.05, Supplementary Material III: Details of rough-set based model.

Acknowledgments: This work was supported by the National Natural Science Foundation of China (31371335, 31701151), the Natural Science Foundation of Shanghai (17ZR1412500), the Shanghai Sailing Program, the Youth Innovation Promotion Association of the Chinese Academy of Sciences (CAS) (2016245), the fund of the key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703).

Author Contributions: T.H. and Y.D.C. conceived and designed the experiments; D.W. and L.C. performed the experiments; D.W., J.R.L. and Y.H.Z. analyzed the data; D.W., J.R.L. and Y.H.Z. contributed reagents/materials/analysis tools; D.W. wrote the paper.

Conflicts of Interest: The founding sponsors had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, and in the decision to publish the results.

References

1. Sopik, V.; Sun, P.; Narod, S.A. The prognostic effect of estrogen receptor status differs for younger versus older breast cancer patients. *Breast Cancer Res. Treat.* **2017**, *165*, 391–402. [CrossRef] [PubMed]
2. Boerman, L.M.; Maass, S.; van der Meer, P.; Gietema, J.A.; Maduro, J.H.; Hummel, Y.M.; Berger, M.Y.; de Bock, G.H.; Berendsen, A.J. Long-term outcome of cardiac function in a population-based cohort of breast cancer survivors: A cross-sectional study. *Eur. J. Cancer* **2017**, *81*, 56–65. [CrossRef] [PubMed]
3. Lundberg, F.E.; Iliadou, A.N.; Rodriguez-Wallberg, K.; Bergh, C.; Gemzell-Danielsson, K.; Johansson, A.L.V. Ovarian stimulation and risk of breast cancer in Swedish women. *Fertil. Steril.* **2017**, *108*, 137–144. [CrossRef] [PubMed]
4. Kawaguchi, T.; Foster, B.A.; Young, J.; Takabe, K. Current update of patient-derived xenograft model for translational breast cancer research. *J. Mammary Gland Biol. Neoplasia* **2017**, *22*, 131–139. [CrossRef] [PubMed]
5. De la Cruz, F.S.; Diolaiti, D.; Turk, A.T.; Rainey, A.R.; Ambesi-Impiombato, A.; Andrews, S.J.; Mansukhani, M.M.; Nagy, P.L.; Alvarez, M.J.; Califano, A.; et al. A case study of an integrative genomic and experimental therapeutic approach for rare tumors: Identification of vulnerabilities in a pediatric poorly differentiated carcinoma. *Genome Med.* **2016**, *8*, 116. [CrossRef] [PubMed]
6. Furuyama, T.; Tanaka, S.; Shimada, S.; Akiyama, Y.; Matsumura, S.; Mitsunori, Y.; Aihara, A.; Ban, D.; Ochiai, T.; Kudo, A.; et al. Proteasome activity is required for the initiation of precancerous pancreatic lesions. *Sci. Rep.* **2016**, *6*, 27044. [CrossRef] [PubMed]
7. Zhan, B.; Wen, S.; Lu, J.; Shen, G.; Lin, X.; Feng, J.; Huang, H. Identification and causes of metabonomic difference between orthotopic and subcutaneous xenograft of pancreatic cancer. *Oncotarget* **2017**, *8*, 61264–61281. [CrossRef] [PubMed]
8. Chijiwa, T.; Kawai, K.; Noguchi, A.; Sato, H.; Hayashi, A.; Cho, H.; Shiozawa, M.; Kishida, T.; Morinaga, S.; Yokose, T.; et al. Establishment of patient-derived cancer xenografts in immunodeficient NOG mice. *Int. J. Oncol.* **2015**, *47*, 61–70. [CrossRef] [PubMed]
9. Unno, K.; Ono, M.; Winder, A.D.; Maniar, K.P.; Paintal, A.S.; Yu, Y.; Wei, J.J.; Lurain, J.R.; Kim, J.J. Establishment of human patient-derived endometrial cancer xenografts in NOD scid gamma mice for the study of invasion and metastasis. *PLoS ONE* **2014**, *9*, e116064. [CrossRef] [PubMed]

10. Bertotti, A.; Migliardi, G.; Galimi, F.; Sassi, F.; Torti, D.; Isella, C.; Cora, D.; Di Nicolantonio, F.; Buscarino, M.; Petti, C.; et al. A molecularly annotated platform of patient-derived xenografts ("xenopatients") identifies HER2 as an effective therapeutic target in cetuximab-resistant colorectal cancer. *Cancer Discov.* **2011**, *1*, 508–523. [CrossRef] [PubMed]
11. Tignanelli, C.J.; Herrera Loeza, S.G.; Yeh, J.J. KRAS and PIK3CA mutation frequencies in patient-derived xenograft models of pancreatic and colorectal cancer are reflective of patient tumors and stable across passages. *Am. Surg.* **2014**, *80*, 873–877. [PubMed]
12. Li, H.; Zhu, Y.J.; Tang, X.Y.; Li, J.Y.; Li, Y.Y.; Zhong, Z.M.; Ding, G.H.; Li, Y.X. Integrated analysis of transcriptome in cancer patient-derived xenografts. *PLoS ONE* **2015**, *10*, e0124780. [CrossRef] [PubMed]
13. Lawson, D.A.; Bhakta, N.R.; Kessenbrock, K.; Prummel, K.D.; Yu, Y.; Takai, K.; Zhou, A.; Eyob, H.; Balakrishnan, S.; Wang, C.Y.; et al. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature* **2015**, *526*, 131–135. [CrossRef] [PubMed]
14. Draminski, M.; Rada-Iglesias, A.; Enroth, S.; Wadelius, C.; Koronacki, J.; Komorowski, J. Monte Carlo feature selection for supervised classification. *Bioinformatics* **2008**, *24*, 110–117. [CrossRef] [PubMed]
15. Dramiński, M.; Kierczak, M.; Nowak-Brzezińska, A.; Koronecki, J.; Komorowski, J. The Monte Carlo feature selection and interdependency discovery is unbiased. *Control Cybern.* **2011**, *40*, 199–211.
16. Chen, L.; Li, J.; Zhang, Y.-H.; Feng, K.; Wang, S.; Zhang, Y.; Huang, T.; Kong, X.; Cai, Y.-D. Identification of gene expression signatures across different types of neural stem cells with the Monte-Carlo feature selection method. *J. Cell. Biochem.* **2017**, *119*, 3394–3403. [CrossRef] [PubMed]
17. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
18. Ho, T.K. Random decision forests. In Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC, Canada, 14–16 August 1995; pp. 278–282.
19. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*; Kluwer Academic Publisher: Dordrecht, The Netherlands, 1991.
20. Hastie, T.; Tibshirani, R.; Sherlock, G.; Eisen, M.; Brown, P.; Botstein, D. *Imputing Missing Data for Gene Expression Arrays*; Technical Report; Stanford University Statistics Department: Stanford, CA, USA, 1999.
21. Troyanskaya, O.; Cantor, M.; Sherlock, G.; Brown, P.; Hastie, T.; Tibshirani, R.; Botstein, D.; Altman, R.B. Missing value estimation methods for DNA microarrays. *Bioinformatics* **2001**, *17*, 520–525. [CrossRef] [PubMed]
22. Wang, S.; Zhang, Y.H.; Lu, J.; Cui, W.; Hu, J.; Cai, Y.D. Analysis and identification of aptamer-compound interactions with a maximum relevance minimum redundancy and nearest neighbor algorithm. *Biomed. Res. Int.* **2016**, *2016*, 8351204. [CrossRef] [PubMed]
23. Wang, S.; Zhang, Q.; Lu, J.; Cai, Y.-D. Analysis and prediction of nitrated tyrosine sites with the mRMR method and support vector machine algorithm. *Curr. Bioinform.* **2018**, *13*, 3–13. [CrossRef]
24. Wang, S.; Zhang, Y.-H.; Huang, G.; Chen, L.; Cai, Y.-D. Analysis and prediction of myristoylation sites using the mRMR method, the IFS method and an extreme learning machine algorithm. *Comb. Chem. High Throughput Screen.* **2017**, *20*, 96–106. [CrossRef] [PubMed]
25. Pugalenthi, G.; Kandaswamy, K.; Chou, K.-C.; Vivekanandan, S.; Kolatkar, P. RSARF: Prediction of residue solvent accessibility from protein sequence using random forest method. *Protein Pept. Lett.* **2011**, *19*, 50–56. [CrossRef]
26. Zhao, X.; Zou, Q.; Liu, B.; Liu, X. Exploratory predicting protein folding model with random forest and hybrid features. *Curr. Proteom.* **2014**, *11*, 289–299. [CrossRef]
27. Kohavi, R. *A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection; International Joint Conference on Artificial Intelligence*; Lawrence Erlbaum Associates Ltd.: Montreal, QC, Canada, 1995; pp. 1137–1145.
28. Geisser, S. *Predictive Inference*; CRC Press: Boca Raton, FL, USA, 1993; Volume 55.
29. Øhrn, A. *Discernibility and Rough sets in Medicine: Tools and Applications*; Norwegian University of Science and Technology (NTNU): Trondheim, Norway, 1999.
30. Johnson, D.S. Approximation algorithms for combinatorial problems. *J. Comput. Syst. Sci.* **1974**, *9*, 256–278. [CrossRef]
31. Cohen, W.W. Fast effective rule induction. In Proceedings of the Twelfth International Conference on Machine Learning, Tahoe City, CA, USA, 9–12 July 1995; pp. 115–123.

32. Furnkranz, J.; Widmer, G. Incremental reduced error pruning. In Proceedings of the Machine Learning: Proceedings of the Eleventh Annual Conference, Rutgers University, New Brunswick, NJ, USA, 10–13 July 1994.
33. Quinlan, J.R. Learning logical definitions from relations. *Mach. Learn.* **1990**, *266*, 239–266. [CrossRef]
34. Brunk, C.A.; Pazzani, M.J. An investigation of noise-tolerant relational concept learning algorithms. In Proceedings of the 8th International Workshop on Machine Learning, Evanston, IL, USA, June 1991; pp. 389–393.
35. Dramiński, M.; Dąbrowski, M.J.; Diamanti, K.; Koronacki, J.; Komorowski, J. Discovering networks of interdependent features in high-dimensional problems. In *Big Data Analysis: New Algorithms for a New Society*; Springer International Publishing: Cham, Switzerland, 2016; pp. 285–304.
36. Chen, L.; Zhang, Y.-H.; Lu, G.; Huang, T.; Cai, Y.-D. Analysis of cancer-related lncRNAs using gene ontology and KEGG pathways. *Artif. Intell. Med.* **2017**, *76*, 27–36. [CrossRef] [PubMed]
37. Li, B.-Q.; Zhang, Y.-H.; Jin, M.-L.; Huang, T.; Cai, Y.-D. Prediction of protein-peptide interactions with a nearest neighbor algorithm. *Curr. Bioinform.* **2018**, *13*, 14–24. [CrossRef]
38. Zhang, Q.; Sun, X.; Feng, K.; Wang, S.; Zhang, Y.H.; Wang, S.; Lu, L.; Cai, Y.D. Predicting citrullination sites in protein sequences using mRMR method and random forest algorithm. *Comb. Chem. High Throughput Screen.* **2017**, *20*, 164–173. [CrossRef] [PubMed]
39. Chen, L.; Zhang, Y.H.; Huang, T.; Cai, Y.D. Gene expression profiling gut microbiota in different races of humans. *Sci. Rep.* **2016**, *6*, 23075. [CrossRef] [PubMed]
40. Chen, L.; Zhang, Y.-H.; Zheng, M.; Huang, T.; Cai, Y.-D. Identification of compound–protein interactions through the analysis of gene ontology, KEGG enrichment for proteins and molecular fragments of compounds. *Mol. Genet. Genom.* **2016**, *291*, 2065–2079. [CrossRef] [PubMed]
41. Ni, Q.; Chen, L. A feature and algorithm selection method for improving the prediction of protein structural classes. *Comb. Chem. High Throughput Screen.* **2017**, *20*, 612–621. [CrossRef] [PubMed]
42. Fang, Y.; Chen, L. A binary classifier for prediction of the types of metabolic pathway of chemicals. *Comb. Chem. High Throughput Screen.* **2017**, *20*, 140–146. [CrossRef] [PubMed]
43. Chen, L.; Wang, S.; Zhang, Y.-H.; Li, J.; Xing, Z.-H.; Yang, J.; Huang, T.; Cai, Y.-D. Identify key sequence features to improve CRISPR sgRNA efficacy. *IEEE Access* **2017**, *5*, 26582–26590. [CrossRef]
44. Chen, L.; Zhang, Y.-H.; Huang, G.; Pan, X.; Wang, S.; Huang, T.; Cai, Y.-D. Discriminating cirRNAs from other lncRNAs using a hierarchical extreme learning machine (H-ELM) algorithm with feature selection. *Mol. Genet. Genom.* **2018**, *293*, 137–149. [CrossRef] [PubMed]
45. Zhang, Y.H.; Xing, Z.H.; Liu, C.L.; Wang, S.P.; Huang, T.; Cai, Y.D.; Kong, X.Y. Identification of the core regulators of the HLA I-peptide binding process. *Sci. Rep.* **2017**, *7*, 42768. [CrossRef] [PubMed]
46. Huang, G.; Chu, C.; Huang, T.; Kong, X.; Zhang, Y.; Zhang, N.; Cai, Y.D. Exploring mouse protein function via multiple approaches. *PLoS ONE* **2016**, *11*, e0166580. [CrossRef] [PubMed]
47. Cherkassky, V. The nature of statistical learning theory. *IEEE Trans. Neural Netw.* **1997**, *8*, 1564. [CrossRef] [PubMed]
48. Ting, K.M.; Witten, I.H. Stacking bagged and dagged models. In Proceedings of the Fourteenth International Conference on Machine Learning, San Francisco, CA, USA, 8–12 July 1997.
49. Almeida, G.S.; Bawn, C.M.; Galler, M.; Wilson, I.; Thomas, H.D.; Kyle, S.; Curtin, N.J.; Newell, D.R.; Maxwell, R.J. PARP inhibitor rucaparib induces changes in NAD levels in cells and liver tissues as assessed by MRS. *NMR Biomed.* **2017**, *30*, e3736. [CrossRef] [PubMed]
50. Ghosh, R.; Roy, S.; Kamyab, J.; Dantzer, F.; Franco, S. Common and unique genetic interactions of the poly(ADP-ribose) polymerases PARP1 and PARP2 with DNA double-strand break repair pathways. *DNA Repair* **2016**, *45*, 56–62. [CrossRef] [PubMed]
51. Pignochino, Y.; Capozzi, F.; D'Ambrosio, L.; Dell'Aglio, C.; Basirico, M.; Canta, M.; Lorenzato, A.; Lutati, F.V.; Aliberti, S.; Palesandro, E.; et al. PARP1 expression drives the synergistic antitumor activity of trabectedin and PARP1 inhibitors in sarcoma preclinical models. *Mol. Cancer* **2017**, *16*, 86. [CrossRef] [PubMed]
52. Johnson, S.F.; Cruz, C.; Greifenberg, A.K.; Dust, S.; Stover, D.G.; Chi, D.; Primack, B.; Cao, S.; Bernhardy, A.J.; Coulson, R.; et al. CDK12 inhibition reverses de novo and acquired PARP inhibitor resistance in BRCA wild-type and mutated models of triple-negative breast cancer. *Cell Rep.* **2016**, *17*, 2367–2381. [CrossRef] [PubMed]

53. Foyle, A.; Sangalang, V.E. Intraskeletal myofiber metastasis of breast-carcinoma. *Hum. Pathol.* **1984**, *15*, 198. [CrossRef]
54. Saha, S.K.; Choi, H.Y.; Kim, B.W.; Dayem, A.A.; Yang, G.M.; Kim, K.S.; Yin, Y.F.; Cho, S.G. KRT19 directly interacts with β-catenin/RAC1 complex to regulate NUMB-dependent NOTCH signaling pathway and breast cancer properties. *Oncogene* **2017**, *36*, 332–349. [CrossRef] [PubMed]
55. Whittle, J.R.; Lewis, M.T.; Lindeman, G.J.; Visvader, J.E. Patient-derived xenograft models of breast cancer and their predictive power. *Breast Cancer Res.* **2015**, *17*, 17. [CrossRef] [PubMed]
56. Jiang, Z.; Jiang, X.; Chen, S.; Lai, Y.; Wei, X.; Li, B.; Lin, S.; Wang, S.; Wu, Q.; Liang, Q.; et al. Anti-GPC3-CAR T cells suppress the growth of tumor cells in patient-derived xenografts of hepatocellular carcinoma. *Front. Immunol.* **2016**, *7*, 690. [CrossRef] [PubMed]
57. Leca, J.; Martinez, S.; Lac, S.; Nigri, J.; Secq, V.; Rubis, M.; Bressy, C.; Serge, A.; Lavaut, M.N.; Dusetti, N.; et al. Cancer-associated fibroblast-derived annexin A6+ extracellular vesicles support pancreatic cancer aggressiveness. *J. Clin. Investig.* **2016**, *126*, 4140–4156. [CrossRef] [PubMed]
58. Katdare, M.R.; Bhonde, R.R.; Parab, P.B. Analysis of morphological and functional maturation of neoislets generated in vitro from pancreatic ductal cells and their suitability for islet banking and transplantation. *J. Endocrinol.* **2004**, *182*, 105–112. [CrossRef] [PubMed]
59. Khanom, R.; Sakamoto, K.; Pal, S.K.; Shimada, Y.; Morita, K.; Omura, K.; Miki, Y.; Yamaguchi, A. Expression of basal cell keratin 15 and keratin 19 in oral squamous neoplasms represents diverse pathophysiologies. *Histol. Histopathol.* **2012**, *27*, 949–959. [PubMed]
60. Deng, J.; Wang, L.; Chen, H.; Li, L.; Ma, Y.; Ni, J.; Li, Y. The role of tumour-associated MUC1 in epithelial ovarian cancer metastasis and progression. *Cancer Metast. Rev.* **2013**, *32*, 535–551. [CrossRef] [PubMed]
61. Jeschke, U.; Wiest, I.; Schumacher, A.L.; Kupka, M.; Rack, B.; Stahn, R.; Karsten, U.; Mayr, D.; Friese, K.; Dian, D. Determination of MUC1 in sera of ovarian cancer patients and in sera of patients with benign changes of the ovaries with CA15–3, CA27.29, and PankoMab. *Anticancer Res.* **2012**, *32*, 2185–2189. [PubMed]
62. Qadir, A.S.; Ceppi, P.; Brockway, S.; Law, C.; Mu, L.; Khodarev, N.N.; Kim, J.; Zhao, J.C.; Putzbach, W.; Murmann, A.E.; et al. CD95/Fas increases stemness in cancer cells by inducing a STAT1-dependent type I interferon response. *Cell Rep.* **2017**, *18*, 2373–2386. [CrossRef] [PubMed]
63. Snowden, E.; Porter, W.; Hahn, F.; Ferguson, M.; Tong, F.; Parker, J.S.; Middlebrook, A.; Ghanekar, S.; Dillmore, W.S.; Blaesius, R. Immunophenotyping and transcriptomic outcomes in PDX-derived TNBC tissue. *Mol. Cancer Res.* **2017**, *15*, 429–438. [CrossRef] [PubMed]
64. Freitas, C.; Wittner, M.; Nguyen, J.; Rondeau, V.; Biajoux, V.; Aknin, M.L.; Gaudin, F.; Beaussant-Cohen, S.; Bertrand, Y.; Bellanne-Chantelot, C.; et al. Lymphoid differentiation of hematopoietic stem cells requires efficient Cxcr4 desensitization. *J. Exp. Med.* **2017**, *214*, 2023–2040. [CrossRef] [PubMed]
65. Lis, R.; Karrasch, C.C.; Poulos, M.G.; Kunar, B.; Redmond, D.; Duran, J.G.B.; Badwe, C.R.; Schachterle, W.; Ginsberg, M.; Xiang, J.; et al. Conversion of adult endothelium to immunocompetent haematopoietic stem cells. *Nature* **2017**, *545*, 439–445. [CrossRef] [PubMed]
66. Lefort, S.; Thuleau, A.; Kieffer, Y.; Sirven, P.; Bieche, I.; Marangoni, E.; Vincent-Salomon, A.; Mechta-Grigoriou, F. CXCR4 inhibitors could benefit to HER2 but not to triple-negative breast cancer patients. *Oncogene* **2017**, *36*, 1211–1222. [CrossRef] [PubMed]
67. Nobutani, K.; Shimono, Y.; Mizutani, K.; Ueda, Y.; Suzuki, T.; Kitayama, M.; Minami, A.; Momose, K.; Miyawaki, K.; Akashi, K.; et al. Downregulation of CXCR4 in metastasized breast cancer cells and implication in their dormancy. *PLoS ONE* **2015**, *10*, e0130032. [CrossRef] [PubMed]
68. Jacobi, N.; Seeboeck, R.; Hofmann, E.; Eger, A. ErbB family signalling: A paradigm for oncogene addiction and personalized oncology. *Cancers* **2017**, *9*, 33. [CrossRef] [PubMed]
69. Manning, H.C.; Buck, J.R.; Cook, R.S. Mouse models of breast cancer: Platforms for discovering precision imaging diagnostics and future cancer medicine. *J. Nucl. Med.* **2016**, *57*, 60–68. [CrossRef] [PubMed]
70. Choy, L.; Hagenbeek, T.J.; Solon, M.; French, D.; Finkle, D.; Shelton, A.; Venook, R.; Brauer, M.J.; Siebel, C.W. Constitutive NOTCH3 signaling promotes the growth of basal breast cancers. *Cancer Res.* **2017**, *77*, 1439–1452. [CrossRef] [PubMed]
71. Baker, L.A.; Holliday, H.; Swarbrick, A. ID4 controls luminal lineage commitment in normal mammary epithelium and inhibits BRCA1 function in basal-like breast cancer. *Endocr.-Relat. Cancer* **2016**, *23*, R381–R392. [CrossRef] [PubMed]

72. Kashi, V.P.; Hatley, M.E.; Galindo, R.L. Probing for a deeper understanding of rhabdomyosarcoma: Insights from complementary model systems. *Nat. Rev. Cancer* **2015**, *15*, 426–439. [CrossRef] [PubMed]
73. Berezovsky, A.D. Sox2 promotes malignancy in glioblastoma by regulating plasticity and astrocytic differentiation. *Neoplasia* **2014**, *16*, 193–206. [CrossRef] [PubMed]
74. Bornelov, S.; Marillet, S.; Komorowski, J. Ciruvis: A web-based tool for rule networks and interaction detection using rule-based classifiers. *BMC Bioinform.* **2014**, *15*, 139. [CrossRef] [PubMed]
75. Mehrazarin, S.; Chen, W.; Oh, J.E.; Liu, Z.X.; Kang, K.L.; Yi, J.K.; Kim, R.H.; Shin, K.H.; Park, N.H.; Kang, M.K. The p63 gene is regulated by grainyhead-like 2 (GRHL2) through reciprocal feedback and determines the epithelial phenotype in human keratinocytes. *J. Biol. Chem.* **2015**, *290*, 19999–20008. [CrossRef] [PubMed]
76. Kiselyov, A.; Bunimovich-Mendrazitsky, S.; Startsev, V. Key signaling pathways in the muscle-invasive bladder carcinoma: Clinical markers for disease modeling and optimized treatment. *J. Int. Cancer* **2016**, *138*, 2562–2569. [CrossRef] [PubMed]
77. Moheimani, F.; Roth, H.M.; Cross, J.; Reid, A.T.; Shaheen, F.; Warner, S.M.; Hirota, J.A.; Kicic, A.; Hallstrand, T.S.; Kahn, M.; et al. Disruption of β-catenin/CBP signaling inhibits human airway epithelial–mesenchymal transition and repair. *Int. J. Biochem. Cell Biol.* **2015**, *68*, 59–69. [CrossRef] [PubMed]
78. Glatter, T.; Wepf, A.; Aebersold, R.; Gstaiger, M. An integrated workflow for charting the human interaction proteome: Insights into the PP2A system. *Mol. Syst. Biol.* **2009**, *5*, 237. [CrossRef] [PubMed]
79. Rual, J.F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G.F.; Gibbons, F.D.; Dreze, M.; Ayivi-Guedehoussou, N.; et al. Towards a proteome-scale map of the human protein-protein interaction network. *Nature* **2005**, *437*, 1173–1178. [CrossRef] [PubMed]
80. Koringa, P.G.; Jakhesara, S.J.; Bhatt, V.D.; Meshram, C.P.; Patel, A.K.; Fefar, D.T.; Joshi, C.G. Comprehensive transcriptome profiling of squamous cell carcinoma of horn in Bos indicus. *Vet. Comp. Oncol.* **2016**, *14*, 122–136. [CrossRef] [PubMed]
81. Gorski, J.J.; James, C.R.; Quinn, J.E.; Stewart, G.E.; Staunton, K.C.; Buckley, N.E.; McDyer, F.A.; Kennedy, R.D.; Wilson, R.H.; Mullan, P.B.; et al. BRCA1 transcriptionally regulates genes associated with the basal-like phenotype in breast cancer. *Breast Cancer Res. Treat.* **2010**, *122*, 721–731. [CrossRef] [PubMed]
82. Li, C.; Ma, H.; Wang, Y.; Cao, Z.; Graves-Deal, R.; Powell, A.E.; Starchenko, A.; Ayers, G.D.; Washington, M.K.; Kamath, V.; et al. Excess PLAC8 promotes an unconventional ERK2-dependent EMT in colon cancer. *J. Clin. Investig.* **2014**, *124*, 2172–2187. [CrossRef] [PubMed]
83. Ju, J.H.; Oh, S.; Lee, K.M.; Yang, W.; Nam, K.S.; Moon, H.G.; Noh, D.Y.; Kim, C.G.; Park, G.; Park, J.B.; et al. Cytokeratin19 induced by HER2/ERK binds and stabilizes HER2 on cell membranes. *Cell Death Differ.* **2015**, *22*, 665–676. [CrossRef] [PubMed]
84. Markou, A.; Lazaridou, M.; Paraskevopoulos, P.; Chen, S.; Swierczewska, M.; Budna, J.; Kuske, A.; Gorges, T.M.; Joosse, S.A.; Kroneis, T.; et al. Multiplex gene expression profiling of in vivo isolated circulating tumor cells in high-risk prostate cancer patients. *Clin. Chem.* **2018**. [CrossRef] [PubMed]
85. Bredemeier, M.; Edimiris, P.; Mach, P.; Kubista, M.; Sjoback, R.; Rohlova, E.; Kolostova, K.; Hauch, S.; Aktas, B.; Tewes, M.; et al. Gene expression signatures in circulating tumor cells correlate with response to therapy in metastatic breast cancer. *Clin. Chem.* **2017**, *63*, 1585–1593. [CrossRef] [PubMed]
86. Lin, C.; Chen, W.; Qiu, C.; Wu, Y.; Krishnan, S.; Zou, Q. LibD3C: Ensemble classifiers with a clustering and dynamic selection strategy. *Neurocomputing* **2014**, *123*, 424–435. [CrossRef]