International Journal of
Molecular Sciences

MDPI

Article

Identifying Methylation Pattern and Genes Associated with Breast Cancer Subtypes

Lei Chen [1,2,3,†], Tao Zeng [4,†], Xiaoyong Pan [5,6,†], Yu-Hang Zhang [7], Tao Huang [7,*] and Yu-Dong Cai [1,*]

[1] School of Life Sciences, Shanghai University, Shanghai 200444, China
[2] College of Information Engineering, Shanghai Maritime University, Shanghai 201306, China
[3] Shanghai Key Laboratory of PMMP, East China Normal University, Shanghai 200241, China
[4] Key Laboratory of Systems Biology, Institute of Biochemistry and Cell Biology, Chinese Academy of Sciences, Shanghai 200031, China
[5] Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, and Key Laboratory of System Control and Information Processing, Ministry of Education of China, Shanghai 200240, China
[6] IDLab, Department for Electronics and Information Systems, Ghent University, 9000 Ghent, Belgium
[7] Shanghai Institute of Nutrition and Health, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031, China
* Correspondence: tohuangtao@126.com (T.H.); cai_yud@126.com (Y.-D.C.); Tel.: +86-21-5492-3269 (T.H.); +86-21-6613-6132 (Y.-D.C.)
† These authors contributed equally to this work.

Received: 1 August 2019; Accepted: 29 August 2019; Published: 31 August 2019

Abstract: Breast cancer is regarded worldwide as a severe human disease. Various genetic variations, including hereditary and somatic mutations, contribute to the initiation and progression of this disease. The diagnostic parameters of breast cancer are not limited to the conventional protein content and can include newly discovered genetic variants and even genetic modification patterns such as methylation and microRNA. In addition, breast cancer detection extends to detailed breast cancer stratifications to provide subtype-specific indications for further personalized treatment. One genome-wide expression–methylation quantitative trait loci analysis confirmed that different breast cancer subtypes have various methylation patterns. However, recognizing clinically applied (methylation) biomarkers is difficult due to the large number of differentially methylated genes. In this study, we attempted to re-screen a small group of functional biomarkers for the identification and distinction of different breast cancer subtypes with advanced machine learning methods. The findings may contribute to biomarker identification for different breast cancer subtypes and provide a new perspective for differential pathogenesis in breast cancer subtypes.

Keywords: breast cancer; subtype; methylation; pattern; multi-class classification

1. Introduction

According to epidemiological statistics provided by the World Health Organization in 2015, more than 8.8 million deaths, accounting for one in six deaths, are attributed to cancer [1]. Cancer has become one of the major threats to human health. Among different cancer sites from different tissues and organs of the human body, breast cancer is one of the most common in women, with more than 1.7 million new cases recorded in 2012, accounting for one in four new cancer cases among all female cancers worldwide [2,3]. This disease has also been identified as one of the top five common causes of deaths, generating more than 0.57 million deaths in 2015 [1]. Therefore, breast cancer is widely regarded as a severe disease for humans worldwide, especially for women.

Histologically, breast cancer refers to all cancers that develop from breast tissues. Major risks for such disease include active risks (risks that patients can actively avoid) and passive risks (risks that patients can only passively experience) [4]. The major active risks for breast cancer are dietary patterns, obesity, and lack of childbearing [5], and the major passive risks for breast cancer are biological sex (women with higher mobility), genetic background, and age [6,7]. Among these risk factors, genetic background has recently received attention due to the research progression in cancer biology and the development of next-generation sequencing [8]. Various genetic variations, either hereditary or somatic mutations, contribute to the initiation and progression of breast cancers [9]. Genes such as *TP53* [10], *HER2* [11], *BRCA1*, and *BRCA2* [12] are related collectively or independently to breast cancer pathogenesis. Genes such as *BRCA1* and *BRCA2* [12] are even named after particular subtypes of breast cancer, indicating their unequivocal genetic contribution. In general, the major clinical symptom for the initiation and progression of breast cancer is an abnormal region on the breast that feels differently from the rest of the breast tissues [13]. A physical breast exam is the first step in breast cancer diagnosis. The two common diagnostic approaches for further medical testing and verification of breast cancer are mammograms (e.g., low-dose X rays) and lump biopsies [13–15]. Identifying accurate and sensitive biomarkers during cancer detection through blood or biopsy samples is essential. According to histological and biochemical studies, non-specific breast cancer markers such as carcinoembryonic antigen (CA) 15-3, and CA 27.29 have been identified as potential biomarkers for breast cancer at the protein level [16,17]. However, these conventional biomarkers have limited clinical applications because they are also identified in other tumor subtypes and even in healthy people who are under stress. In addition, these biomarkers cannot be distinguished from breast cancer subtypes with different pathogenic mechanisms and corresponding treatments. With the development of liquid biopsy and high-throughput sequencing technologies, the detection of breast cancer biomarkers has been extended to the system level [9,17–20]. Hence, the diagnostic parameters of breast cancer are no longer limited to the protein content and can include newly discovered genetic patterns such as CNV [9], methylation [21], and microRNA [22]. Correspondingly, the task and ability of breast cancer detection extend to the detailed subtyping of breast cancer (e.g., disease or treatment stratifications) to provide subtype-specific indications for further personalized treatment [23,24].

In general cancer studies, cancer epigenetics refers to all the studies on multiple epigenetic modifications to the cancer cell DNA [25]. General cancer epigenetics studies focused on the pathogenic significance of somatic DNA methylation, histone modification and microRNA gene silencing processes. There are three major pathological mechanism for such modification to contribute to tumorigenesis: (1) abnormal gene expression regulation, (2) dysfunctional DNA repair pathways and (3) pathological chromosomal instability. The abnormal epigenetic modification has been identified to be more frequently than other kinds of pathological characteristics in tumors like somatic mutations [25,26]. Therefore, the screening for epigenetic markers of tumors may be one of the major part of basic and clinical study of tumor diagnosis and treatment [27]. Among all such patterns of cancer epigenetics modification, abnormal DNA methylation patterns turn out to be some of the most frequent and significant pathogenesis for various cancer subtypes. Cancer genomes have been shown to be hypo-methylated comparing to adjacent normal cells' genome [26]. The hypo-methylation pattern of cancer genomes is generally triggered by dysfunctional DNA methyl-transferases and may further lead to promoted mitotic recombination and damaged chromosomal structures [26]. Such pathological epigenetic modification may contribute to the tumorigenesis and has been widely identified in multiple cancer subtypes. Apart from such general influences, the abnormal methylation of some specific region on the genome may also be quite important for the trigger of tumorigeiesis. For instance, genes like *BRCA1*, *CDH1*, *RARB2*, *PTEN* have all been reported to have abnormal methylation epigenetically modified DNA fragments in the promoter or exonic regions and such modifications have also been confirmed to participate in the tumorigenesis in breast cancer [28,29]. Therefore, considering that some abnormal methylation patterns/markers are not only specific enough for the identification of tumor cells but also be essential for tumorigenesis, it is quite necessary for the screening of specific epigenetic especially methylation markers in tumors as potential clinical markers that guiding the diagnosis and treatment of specific tumor subtypes.

To date, machine learning-based methods have been widely used for analyzing biological and biomedicine data [30,31]. Model et al. applied feature selection for high-dimensional methylation data to classify different cancers [32], showing selecting the right number of features using feature selection is crucial for cancer classification. Adorjan et al. applied supervised and unsupervised machine learning methods to disseminate tumors using the CpG sites [33]. Chen et al. applied feature selection and supervised classifier to classify samples from different MSI statuses in colorectal cancer using expression data [34]. Shipp et al. applied supervised machine learning models to classify diffuse large B-call lymphoma with expression profiles of 6817 genes [35]. Similarly, Ye et al. trained supervised classifiers to predict hepatitis B virus–positive metastatic hepatocellular carcinomas based on the expression profiles [36]. For methylation profiles, machine learning-based methods can also give useful hints.

One genome-wide expression–methylation quantitative trait loci analysis confirmed that different breast cancer subtypes (i.e., basal, Her2, LumA, and LumB) have varying methylation patterns [21]. According to the statistics provided by the World Health Organization, for different races, the proportion of different subtypes in all breast cancer patients are different. In Asian or Pacific populations, more than 50% of breast cancer cases turn out to be LumA subtype and the basal-like subtype only accounts for 5%. However, in Hispanic populations, basal-like subtype accounts for 11.6% and in African-American populations, this subtype accounts for more than 30%. No matter in which population, LumA subtypes always represent the vast majority, always accounting for more than 40%. As for Her2 and LumB subtypes, in each population, such two subgroups always account for 30% and LumB subtypes accounts for twice comparing to Her2 subtypes [37]. Considering the complicated subgrouping pattern of breast cancer and the imbalanced distribution patterns, it is quite significant to extract key biomarkers for the detailed and accurate subgrouping of breast cancer. Although various potential differentially methylated genes have been identified, the number of potential clinical biomarkers is extremely high. As mentioned above, machine learning-based methods may give considerable help. In this study, we attempted to screen a small group of functional biomarkers for the identification and distinction of different breast cancer subtypes using some advanced machine learning methods. The findings may contribute to the biomarker identification for different breast cancer subtypes and provide a new perspective for the differential pathogenesis in breast cancer subtypes.

2. Results

In this study, we first ranked the input 436,506 methylation features by using the MR score and then selected features with relevance scores larger than 0.2. In total, 9777 highly relevant features were kept for subsequent experiments. The MR scores for individual features are provided in Supplementary Table S1. The obtained 9777 features were analyzed by the MCFS method, producing a feature list provided in Supplementary Table S2.

To further select discriminative features with a supervised classifier, we ran incremental feature selection (IFS) with a multi-class support vector machine (SVM) to classify samples from four breast cancer subtypes. A series of feature subsets was generated with a step interval of 10, and the SVM was trained and evaluated on the training samples consisting of the features from each feature subset by using a 10-fold cross-validation. The result yielded the best Matthews correlation coefficient (MCC) of 0.925 and an overall accuracy of 0.949 when using the top 1890 features (Table 1). These features are termed as optimum features and constitute the optimum feature set. The sensitivity and specificity of each class yielded by the SVM with optimum features are also listed in Table 1. Each of them exceeds 0.9, indicating the good performance of such SVM classifier and the importance of the optimum features. The performance on minority classes (basal, Her2, LumB) is also high, suggesting the utility of Synthetic Minority Over-Sampling Technique (SMOTE). The corresponding confusion matrix yielded by such SVM classifier is illustrated in Figure 1A. Supplementary Table S3 details the information of the top 1890 features, and Supplementary Table S4 reports the performances of the SVMs corresponding to all feature subsets. With these performance measurements, an IFS curve was plotted (Figure 2A) with the performance, including sensitivity of each class, overall accuracy and MCC, as *y*-axis and the number of features as *x*-axis.

Table 1. Performance and optimum number of features of incremental feature selection (IFS) with support vector machine (SVM) and random forest (RF).

Terms	Sensitivity/Specificity	SVM	RF
Number of optimum features	/	1890	840
Matthews correlation coefficient (MCC)	/	0.925	0.860
Overall accuracy	/	0.949	0.906
Basal	Sensitivity	1.000	1.000
	Specificity	1.000	0.991
Her2	Sensitivity	0.973	0.892
	Specificity	0.986	0.982
LumA	Sensitivity	0.942	0.917
	Specificity	0.963	0.918
LumB	Sensitivity	0.921	0.841
	Specificity	0.974	0.963

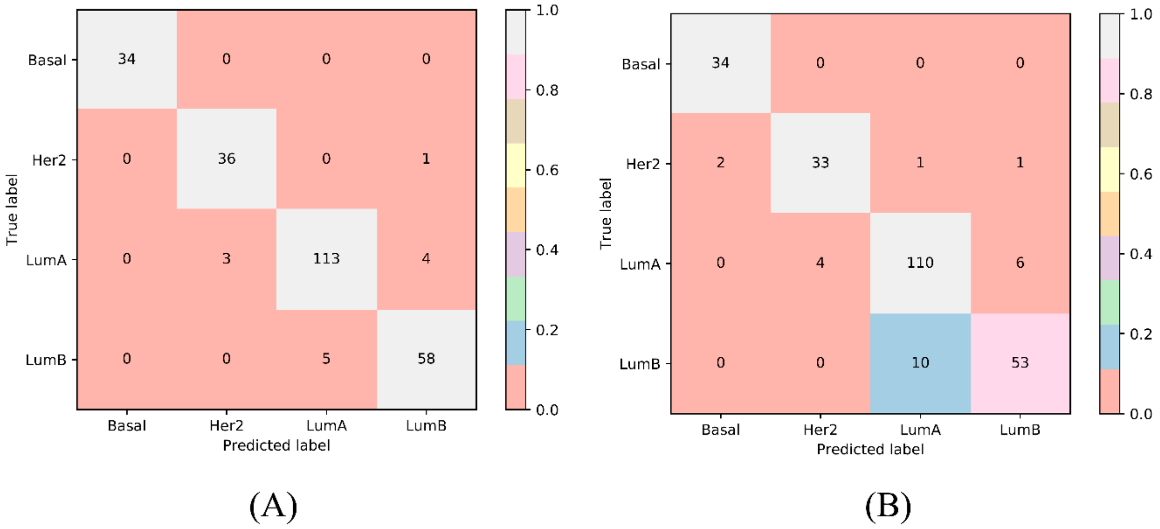


Figure 1. The confusion matrix yielded by the best support vector machine (SVM) and random forest (RF) classifiers. (**A**) The confusion matrix of the best SVM classifier; (**B**) The confusion matrix of the best RF classifier.

Other supervised multi-class classifiers exist, for example random forest (RF) [38–40]. To verify the power of SVMs, we ran IFS with integrated RF for each feature subset in the same way and trained and evaluated RF on samples consisting of features from individual feature subsets by using a 10-fold cross-validation. When the top 840 features were used, RF achieved the best MCC value of 0.860 and an overall accuracy of 0.906 (Table 1). The sensitivity and specificity of each class yielded by such RF classifier are also listed in Table 1. None of them exceed the corresponding measurements yielded by the best SVM classifier. The corresponding confusion matrix is shown in Figure 1B.

Supplementary Table S5 provides the performances of RF for individual feature subsets, and Figure 2B illustrates that the performance of RFs changed with the number of used features. These results reveal that the SVM is a good choice for classifying breast cancer samples from basal, Her2, LumA, and LumB, thereby verifying it as the supervised classifier for IFS in this work.

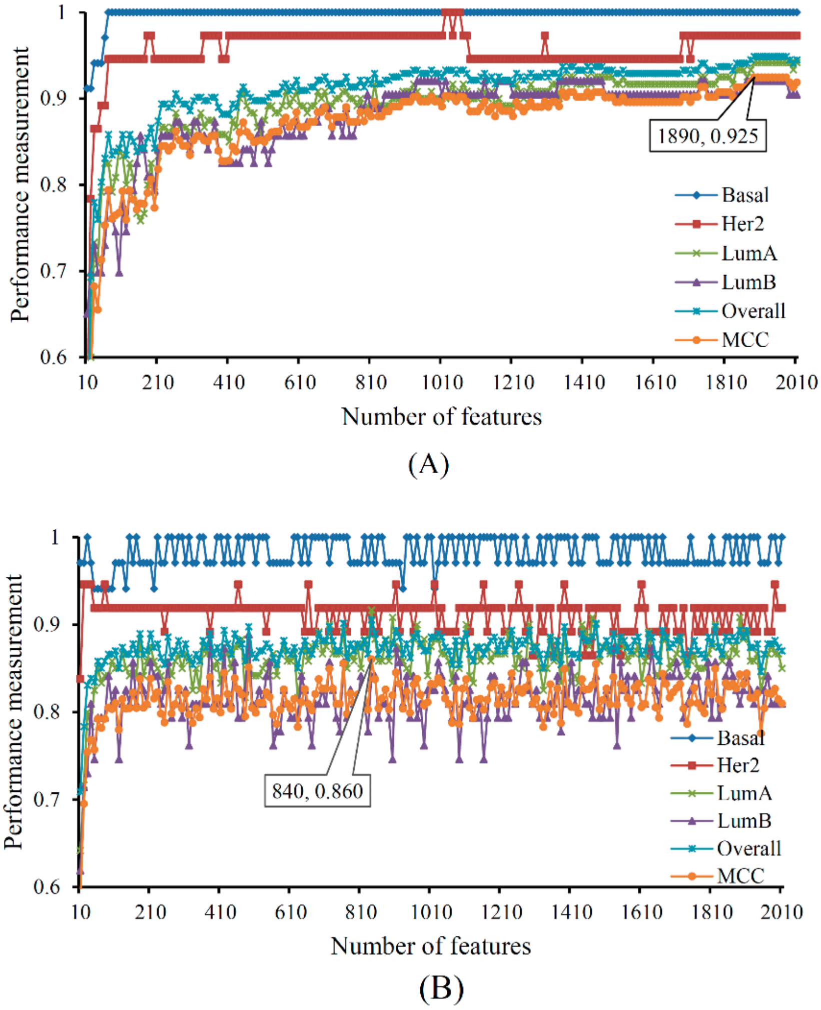


Figure 2. The performance of support vector machine (SVM) and random forest (RF) change with the number of features. (**A**) The performance of SVM; (**B**) The performance of RF.

3. Discussion

3.1. Analysis of Top Ranked Genes

We identified a group of functional genes with different methylation status in various breast cancer subtypes. According to recent publications, the top-10 ranked identified genes with distinctive methylation status have been confirmed, thereby validating the efficacy and accuracy of our prediction. The detailed analysis and discussion of each functional gene are presented below.

In our prediction list, *NTHL1*, which encodes a functional DNA *N*-glycosylase of endonuclease III family, has been predicted to have differential methylation status in various breast cancer subtypes. Although no direct report has confirmed the detailed methylation status of such a gene in breast cancer, a study in 2014 reported that breast cancer with differential breast cancer mutational status can be clustered into various subtypes with different *NTHL1* expression patterns; such expression level distinction of *NTHL1* is probably induced by epigenetic regulation [41]. Therefore, we speculate that in different molecular subtypes of breast cancers, the differential methylation status of *NTHL1* should be associated with its differential expression and could represent a substantial epigenetic modification pattern [42], thereby validating the efficacy and accuracy of our prediction. According to functional annotation and enrichment, the identification of such gene indicated that glycolysis and gluconeogenesis may be alternative pathogenic biological processes for different breast cancer subtypes.

The next predicted gene is *CMBL*, which encodes a specific cysteine hydrolase of the dienelactone hydrolase family and normally has high expression in liver cytosol but not in breast tissues [43,44]. In 2014, an early study on proteasome function in breast cancer confirmed that cysteine hydrolase has an abnormal expression level in certain breast cancer subtypes, and this finding corresponds with our prediction [45]. In 2016, another study confirmed that cysteine hydrolases may be functionally related to proteolytically activated receptors, and the genes encoding these hydrolases such as *CMBL* may have differential expression patterns and biological roles in particular breast cancers subtypes with distinctive epithelial–mesenchymal transition (EMT) tendency [46]. Therefore, *CMBL* may also be a potential biomarker for breast cancer subtyping.

The gene *FLJ43663* is functionally related to an important breast cancer associated gene, *ESR1* [47], but its methylation status has not been directly confirmed in breast cancer subtypes. However, as the clone of *LINC-PINT*, the methylation of this gene has been functionally related to multiple tumor subtypes [48,49]. In 2017, a study on the breast cancer subtyping of Chinese Han women confirmed that the methylation status of the gene *FLJ43663* underlying the expression pattern of *LINC-PINT* is functionally related to the initiation and progression of Luminal A subtype of breast cancer but not of other subtypes [50], thereby validating the efficacy and accuracy of our prediction.

The next predicted gene is *LPP*, which encodes a member of LIM domain protein subfamily contributing to the regulation of cell-cell adhesion and cell motility [51,52]. This gene has pathogenic contribution on breast cancer at methylation level. In 2018, one study on the epigenetic regulation of LIM family genes in different breast cancer subtypes confirmed that the putative promoter region of *LIMD1* and our predicted gene *LPP* may be abnormally methylated in the MDA-MB435 cell line [53]. Further studies on the expression pattern or methylation status of this gene confirmed its abnormal methylation pattern only in breast cancer with high metastatic tendency but not in all samples, thereby indicating that its methylation status may be an effective subtyping biomarker for particular breast cancer patients [54,55].

For the following predicted gene, *ANP32B* is a cell survival factor and participates in cell cycle progression [56,57]. According to a study on *ANP32B* in mouse model, a high expression of this gene may promote the progression of breast cancer [56]. Considering the specific regulatory role of methylation on gene expression pattern, the demethylation of *ANP32B* may contribute to tumorigenesis in breast cancer. Another study in 2016 reported that not all breast cancer subtypes can be induced by the abnormal methylation or expression of our predicted gene *ANP32B* [58], thereby indicating the conditional methylation status of this gene contributes to different breast cancer subtypes.

The next predicted gene *ZCCHC24* encodes a functional zinc finger protein, which is functionally related to platelet biosynthesis and body height [59]. One study on *ZCCHC24* in 2017 confirmed that this gene may participate in tumorigenesis by inhibiting the biological function of BET family proteins [60]. Early in 2016, another study confirmed that different breast cancer subtypes may have various pharmacological reactions against BET inhibitors, thereby reflecting the distinctive contribution of BET signaling pathways in different breast cancer subtypes [61]. Therefore, the expression pattern of *ZCCHC24* in different breast cancer subtypes may be functionally different, and its specific methylation pattern affecting its expression level may also be different and discriminative. These findings validate the efficacy and accuracy of our prediction.

In addition to *ZCCHC24*, another zinc finger protein coding gene named *ZNF282* has also been predicted as a potential distinctive marker for different breast cancer subtypes. As a specific regulator and binder of U5 repressive element, this gene contributes to breast cancer as an estrogen receptor co-activator but not in other pathogenic mechanisms [62]. Hence, the methylation status and expression pattern of *ZNF282* may be different in breast cancers with high and low expression levels of estrogen receptor. The expression of estrogen receptor is one of the clinical diagnostic and subgrouping biomarkers for breast cancer [63–65]. Therefore, our predicted gene *ZNF282* can be regarded as an additional biomarker related to estrogen receptor for breast cancer subtyping.

As the next predicted gene associated with breast cancer subtyping, *SFT2D2* contributes to the fusion of retrograde transport vesicles [66]. Although only a few descriptive studies have been published about this gene, breast cancer *SFT2D2* has been confirmed to contribute to metastatic pathogenic behaviors [67]. The methylation status and expression pattern of this gene vary in breast cancer subtypes such as basal and HER2-like ones, thereby validating its potential distinctive function in molecular subtyping of breast cancer.

BCL9 is a specific gene functionally related to B-cell acute lymphoblastic leukemia and participates in Wnt/β-catenin and GPCR signaling pathways [68]. According to recent publications, *BCL9* is a specific breast cancer associated gene that contributes to the invasion and EMT of breast ductal carcinoma but not of other subtypes, thereby indicating its potential subtyping relevance [69]. Another study on *BCL9* in breast ductal carcinoma confirmed that its methylation status and expression pattern are definitely functionally related to the expression of *ERBB2* and *HER2* in breast cancers [70]. Considering that *ERBB2* and *HER2* are both confirmed molecular subtyping biomarkers for breast cancer [71], the functional links of the two genes with our predicted gene indicate that *BCL9* may also be a potential biomarker at the methylation level, thereby validating the efficacy and accuracy of our prediction. Classifying this gene as potential biomarker also confirms the specific pathogenic role of beta-catenin binding processes in different breast cancer subtypes.

As a specific member of the sorting nexin family, *SNX1b* contributes to the regulation of cell-surface expression of epidermal growth factor receptor [72]. *SNX1b* and its homologues participate in breast cancer tumorigenesis by mediating *BRMS1*-dependent transcriptional repression [73]. Breast cancer subtypes with high metastatic tendency have a low methylation status and a high expression pattern of this gene [74], thereby indicating the potential sub-grouping relevance of *SNX1b* with its pathogenic contribution to different breast cancer subtypes.

In summary, all predicted optimal genes with high rank (the top 10) have been confirmed to participate in breast cancer associated biological processes and have various methylation status and biological roles in different breast cancer subtypes. On the one hand, these identified genes may be potential clinical biomarkers in biopsy samples for breast cancer subtyping. On the other hand, they contribute to different pathogenic mechanisms for various breast cancer subtypes, thus establishing a panorama for breast cancer tumorigenesis at the conditional cellular and molecular levels.

3.2. Case Study on LumB

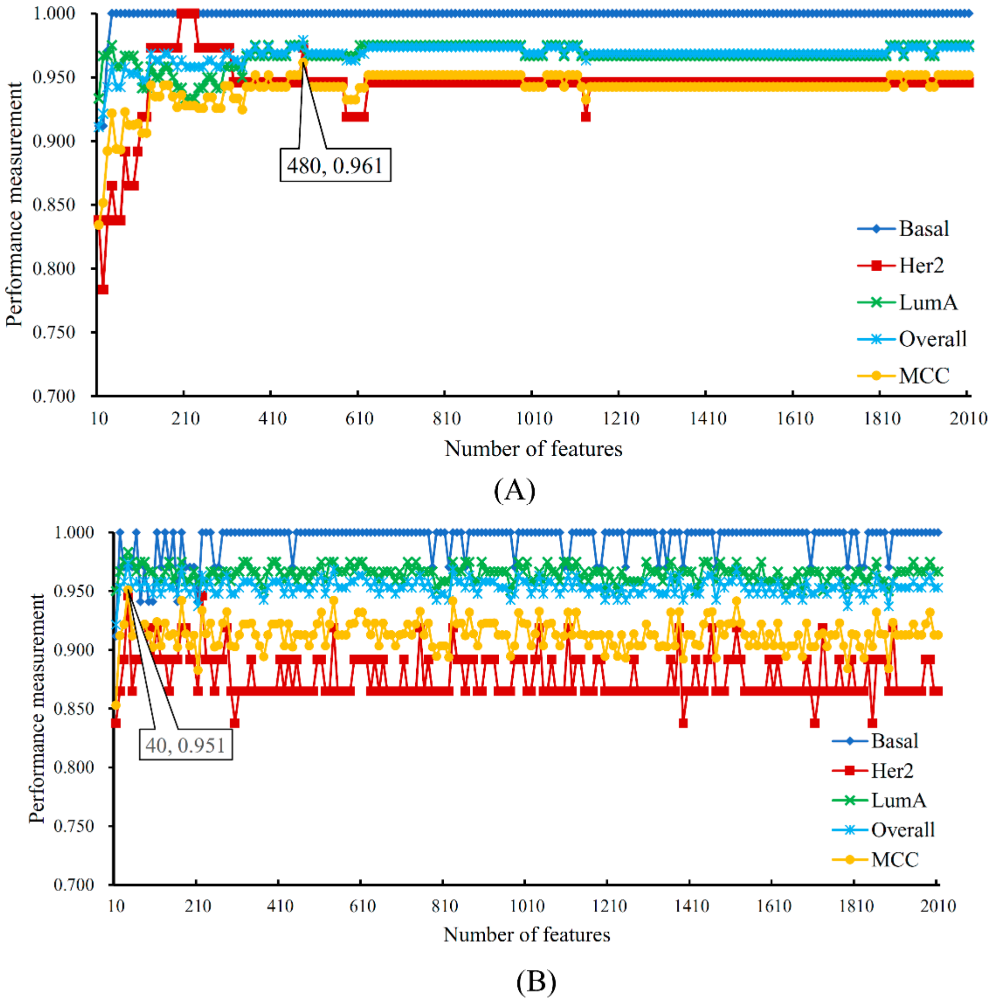
The SVM with optimum features provided good performance as listed in Table 1, from which we can see that the sensitivity on LumB was lowest. It is interesting to investigate whether the SVM can provide better performance if LumB samples are removed. This section gave the results on such study and made further analysis.

Samples in other three breast cancer subtypes: Basal, Her2 and LumA, were represented by 9777 features, which were analyzed by the MCFS method. All 9777 features were sorted in the decreasing order of their RI scores. Obtained feature list is provided in Supplementary Table S6. Similarly, an IFS with SVM was applied on this feature list. The performance of SVM on different feature sets is provided in Supplementary Table S7. The IFS curve was plotted in Figure 3(A). The highest MCC was 0.961 when top 480 features were used. The detailed performance of such SVM is listed in Table 2.

Compared with the results on the whole dataset (Table 1), when LumB samples were excluded, the performance of SVM improved and it can be achieved using much fewer features. These cases also occurred for RF, please see Figure 3B, Table 2 and Supplementary Table S8.

Table 2. Performance and optimum number of features of IFS with SVM and RF on the dataset without LumB samples.

Terms	Sensitivity/Specificity	SVM	RF
Number of optimum features	/	480	40
MCC	/	0.961	0.951
Overall accuracy	/	0.979	0.974
Basal	Sensitivity	1.000	0.971
	Specificity	1.000	0.994
Her2	Sensitivity	0.973	0.946
	Specificity	0.981	0.987
LumA	Sensitivity	0.975	0.983
	Specificity	0.986	0.972



(A)

(B)

Figure 3. The performance of support vector machine (SVM) and random forest (RF) change with the number of features on the dataset without LumB samples. (**A**) The performance of SVM; (**B**) The performance of RF.

As mentioned above, when LumB samples were excluded, the performance of SVM and RF improved and much fewer features were used. It is necessary to investigate the reason why this phenomenon occurred. By checking the results of IFS with SVM on the dataset without Lumb samples (Supplementary Table S7), the MCC arrived at 0.922 when top 40 features were adopted. It was a little lower than that obtained by the SVM with optimum features on the whole dataset, which was 0.925 (Table 1). On the othe hand, the MCC obtained by the SVM with top 40 features on the whole dataset was only 0.655, which was much lower than 0.922. The sensitivities on Basal and Her2 were all high (higher than 0.860), while those on LumA and LumB were much lower (about 0.700). In addition, by checking the confusion matrix, as shown in Figure 4, among 120 LumA samples, 35 samples were misclassified, where 29 samples were classified to LumB; for 63 LumB samples, 19 samples were not correctly classified and 18 samples were assigned to LumA. All these indicated that LumA and LumB samples were quite similar, inducing difficulities to identify LumB samples from LumA samples. LumA and LumB are two subtypes of luminal breast cancer. Luminal breast cancer can be identified by three major parameters named as estrogen receptor (ER), progesterone receptor (PR) and human epidermal growth factor receptor (HER-2). The typical molecular characteristics of luminal breast cancer turn out to be either ER or PR positive together with HER negative. Among the top 40 features (probes), thirty-one probes can be annotated onto the potential functional genes, several of which have been validated to have different methylation patterns or expression levels that possibly controlled by epigenetic factors by recent publications. These features can effectively distinguish luminal breast cancer from other breast cancer subtypes, but cannot perfectly distinguish different luminal breast cancer subtypes. We picked up some features to confirm such a fact.

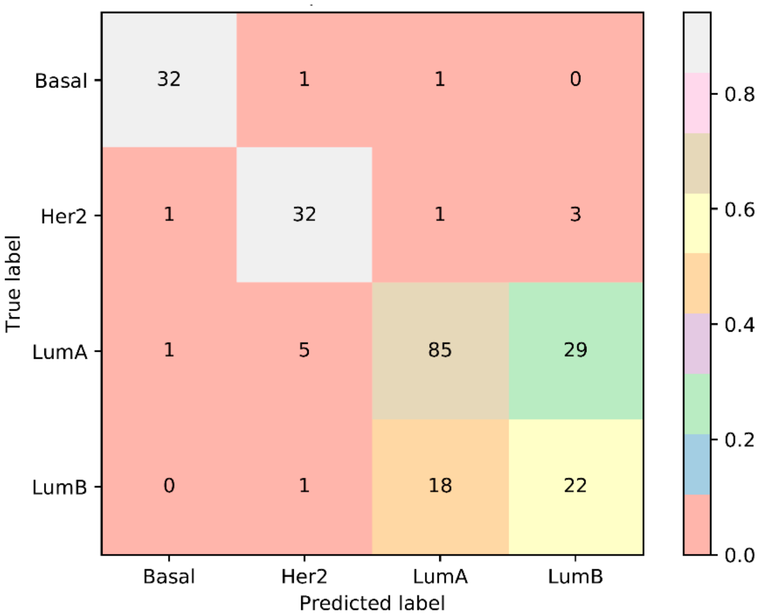


Figure 4. The confusion matrix yielded by the SVM with top 40 features.

A probe named as cg24921140 detected the 3’UTR region of *SSR3*, implying the specific expression alteration of such gene in different subtypes of breast cancer. *SSR3* encodes a glycosylated endoplasmic reticulum membrane receptor and has been reported to contribute to the regulation of calcium metabolism [75,76]. The expression of *SSR3* has been confirmed to be correlated with the biological effects of *HER2*, *Ki-67* and *TP53* [77,78]. As we all know, luminal subtype of breast cancer refers to a subtype of breast cancer with either ER (estrogen receptor) or PR (progesterone receptor) positive together with HER-2 (human epidermal growth factor receptor) negative. Considering the differential expression pattern of *HER2* and *Ki-67* mentioned above, it is reasonable to imply that the expression level of *SSR3* may also contribute to the distinction of luminal breast cancer from other breast cancer subtypes.

ALDH3B1, as another target for the identified probe cg18174530, has been generally reported as a major regulator in the detoxification of aldehydes [79,80]. As for its specific methylation pattern in different breast cancer subtypes, according to recent publications, such a gene has been confirmed to be differentially methylated and expressed in luminal breast cancer subtypes [81]. The next probe, cg13345122, targets another effective gene, named as *STK39*. According to recent publications, this gene has been confirmed to function in the cellular stress response associated processes [82,83]. As for the methylation pattern of such gene in different breast cancer subtypes, it has been confirmed that the expression level of such gene varies in basal and luminal breast cancer samples, indicating the potential distinctive effects of such gene for breast cancer subtyping [84]. Considering that epigenetic silencing like methylation has been reported to be one of the most effective processes for the regulation of *STK39* expression, it is reasonable to speculate that such probe targeting *STK39* may also distinguish luminal subtypes from other breast cancer subtypes [85]. However, no direct reports confirmed the differential methylation pattern of such gene in different subtypes of breast cancer, revealing the inapplicability of such probe on distinguishing LumA and LumB subtypes.

As for the probe targeting effective gene called cg04326566, it targets a member of the homeodomain family of NDA binding protein coding gene, named as *CUX1*. According to recent publications, such gene contributes to the regulation of neuronal differentiation in the brain [86,87]. As the upstream of FGF1/HGF signaling pathway, *CUX1* has been reported to contribute to the progression of luminal and basal breast cancer, help distinguishing Her2 subtype of breast cancer [88]. Apart from that, further study on cell-based assays confirmed that such gene has different expression pattern in basal breast cancer subtype comparing to the luminal subtypes [89].

As for the probe cg12642725, targeting the 5′UTR of *RERG*, was also identified. *RERG* encodes a member of the RAS superfamily participating in the regulation of cell proliferation and tumor formation [90–92]. As for its specific role in different breast cancer subtypes, early in 2011, *RERG* has been screened out as a specific biomarker for ER-positive luminal, like breast cancer subtype [91].

As discussed above, several top features can be validated to distinguish luminal breast cancer from other breast cancer subtypes. However, they cannot further distinguish different luminal breast cancer subtypes (LumA and LumB), inducing difficulties for SVM to identify LumA and LumB samples. In addition, we further investigated the confusion matrices for top 10–100 features, which are available in Supplementary Table S9. The same phenomenon also occurred. Thus, the SVM with small numbers of features cannot provide good performance on LumB because LumA and LumB samples were too similar to identify them.

Besides, for top ten features on the whole dataset, we investigated their distributions on four breast cancer subtypes, which are shown in Figure 5. It can be observed that for amlost all features, their distributions on LumA were most similar to those on LumB, which further confirmed the poor performance of SVM with top features on LumB, as discussed above. The distributions on Her2 followed and those on basal were most different.

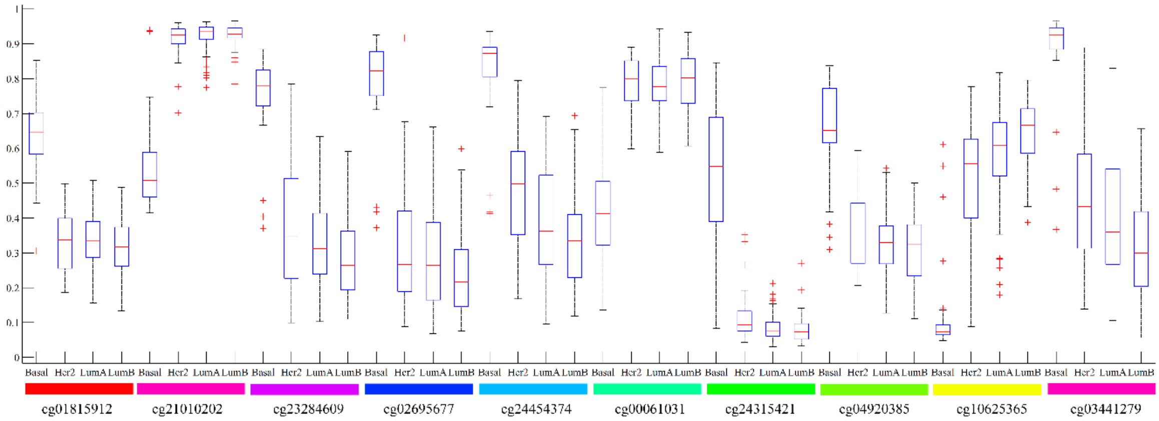


Figure 5. Boxplots to illustrate the distributions of top ten features on four breast cancer subtypes.

4. Materials and Methods

4.1. Datasets

We downloaded the methylation profiles of 34 basal, 37 Her2, 120 LumA, and 63 LumB breast cancer patients from the Gene Expression Omnibus under the accession number of GSE84207 [21]. The methylation profiles of 436,506 probes were measured with Illumina Human Methylation 450 Bead Chip, and the methylation levels were represented with beta values. We investigated the methylation patterns of different breast cancer subtypes and whether or not these subtypes can be rediscovered by using methylation profiles.

4.2. Feature Selection

Figure 6 shows a two-step feature selection strategy, including combining maximum relevance (MR) feature selection [93–102] and Monte Carlo feature selection (MCFS) [9,103–106] to rank input features (e.g., methylation sites). The top-ranked features were further fed into the IFS with SVM to classify samples from four breast cancer subtypes, for example basal, Her2, LumA, and LumB.

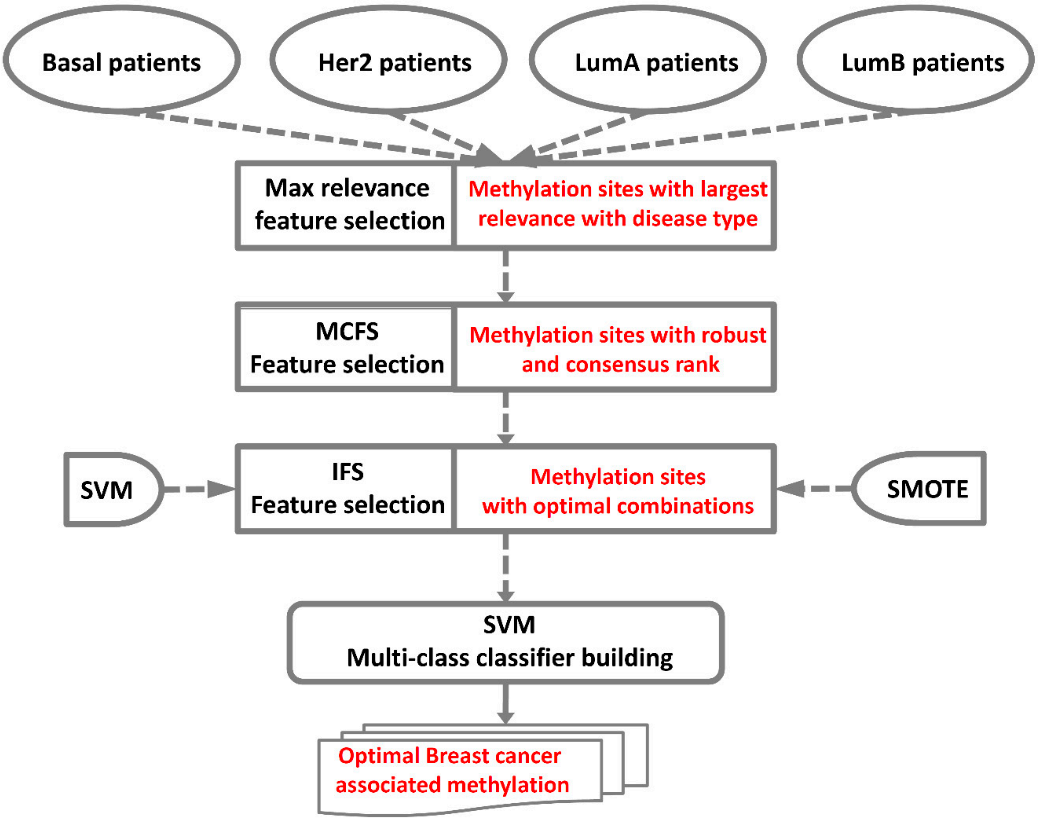

Basal patients
Her2 patients
LumA patients
LumB patients
Max relevance feature selection
Methylation sites with largest relevance with disease type
MCFS Feature selection
Methylation sites with robust and consensus rank
SVM
IFS Feature selection
Methylation sites with optimal combinations
SMOTE
SVM Multi-class classifier building
Optimal Breast cancer associated methylation

Figure 6. Flowchart for classifying samples from four breast cancer subtypes.

4.2.1. Maximum Relevance Score

To select the relevant features to output labels (e.g., breast cancer types), we calculated relevance scores, which are a part of the maximum relevance and minimum redundancy feature selection method [93–101]. The relevance score is defined as the mutual information between label x and feature y:

$$I(x,y) = \iint p(x,y) \log \frac{p(x,y)}{p(x)p(y)} dxdy \tag{1}$$

where $p(x)$ and $p(y)$ are the marginal probability density for x and y, respectively, and $p(x,y)$ is the joint probability density. When this value is high, the relevance of this feature to class label is also high, implying that this feature is important for classification. By setting a threshold for the relevance score, several important features can be obtained.

4.2.2. Monte Carlo Feature Selection

MCFS is a supervised feature selection method based on multiple decision trees and bootstrap sets [9,103–107]. First, p bootstrap sets, which are randomly sampled from the original training set with replacement, are generated, and t feature subsets (each subset includes m features from the original M features and m is much smaller than M) are then produced. One decision tree is grown from one combination of the bootstrap sets and feature subsets; hence, $p \times t$ decision trees should be grown in total.

An important feature should be selected as node feature with high choice for growing decision trees on the basis of which relative importance (RI) score can be calculated for individual features. This score is calculated depending on several factors, including the number of splits where this feature is involved in all nodes of the total $p \times t$ trees, the weight for each split according to its information gain during tree construction, the number of samples corresponding to this split node, and the classification accuracy of the final whole decision tree. The formulation of RI score for feature g is:

$$RI_g = \sum\nolimits_{\tau=1}^{pt} (wAcc)^u \sum\nolimits_{n_g(\tau)} IG(n_g(\tau)) (\frac{\text{no. in } n_g(\tau)}{\text{no. in } \tau})^v \tag{2}$$

where $n_g(\tau)$ represents the nodes in tree τ on which the split is made on feature g, $IG(n_g(\tau))$ denotes the gain information of $n_g(\tau)$, $(\text{no. in } n_g(\tau))$ denotes the number of samples in node $n_g(\tau)$, $(\text{no. in } \tau)$ represents the number of samples in the root of tree τ, $wAcc$ stands for the weighted accuracy of tree τ, and u and v are the two regular factors. Clearly, a feature with a high RI score is important for classification.

In this study, we used MCFS program downloaded from http://www.ipipan.eu/staff/m.draminski/mcfs.html. For convenience, this program was executed with its default parameters. u and v are set to one. After obtaining the RI score for each feature, we sort features in the decreasing order of their RI scores. The obtained feature list is formulated by:

$$F = [f_1, f_2, \ldots, f_n] \tag{3}$$

where f_i denotes a feature and n represents the number of investigated features.

4.2.3. Incremental Feature Selection

Optimum features are required to obtain the best classification performance for distinguishing four breast cancer subtypes, which were further selected by an IFS with an integrated SVM classifier [94,96,103,104,108–111]. With the feature list F, a series of feature subsets is generated with a step interval of 10, in which the first feature subset has the top 10 features, and the second feature subset has the top 20 features and so on. We denote these feature subsets as $F_{10}, F_{20}, F_{30}, \ldots$, where the subscript of F indicates the number of features included in such feature subset. For each feature subset F_i, an SVM classifier can be built on all samples, which are represented by features from this feature subset. 10-fold cross-validation is adopted to evaluate the performance of such classifier. After all SVM classifiers have been evaluated, the classifier with the best cross-validation classification performance is picked up and the corresponding feature subset is selected as the optimum feature set. Features in such set are termed as optimum features (i.e., optimum methylation sites or genes).

4.3. SVM

SVM is a statistics-based supervised model that identifies a hyperplane with a maximum margin between two groups of samples. When a number of samples from two classes are given, a separation line can be used to separate them. However, in most cases, the samples are not linearly separable in a low-dimensional space. Considering that samples will be easily separated in a high-dimensional space, the data are first mapped into a high-dimensional space via kernel trick to make them linearly separable. SVM is a powerful approach to implement such concept and model.

In this study, classifying samples from four breast cancer subtypes is required, which is a multi-class classification problem. Thus, a multi-class SVM was trained by using a one-versus-rest strategy, in which multiple binary SVMs are constructed. Each binary SVM is separately trained on the positive samples from one class (e.g., one breast cancer subtype) and the negative samples from other classes (e.g., remaining breast cancer subtypes). For a query sample, each binary SVM gives the probability of it belonging to the corresponding class. The class with the highest probability is assigned to the sample. To quickly implement the SVM, a tool "SMO" in Weka [112] is employed in this study. John Platt's sequential minimal optimization algorithm [113,114] is used to optimize the training procedures. We selected the polynomial function as the kernel.

4.4. Synthetic Minority Over-Sampling Technique

As indicated in Section 4.1, sample sizes in different breast cancer subtypes are of great difference. The biggest subtype has more than 3.5 times the samples of the smallest subtype. In this case, a perfect classifier is difficult to build because the predicted results are apt to the biggest subtype. To tackle such problem, Synthetic Minority Over-sampling Technique (SMOTE) [115] is employed in this study. In such method, new samples are constructed and added into all classes except the largest class. In detail, let x be a sample in a minority class. Its Euclid distances to all other samples in this class are computed. Accordingly, k nearest neighbors can be found, where k is a pre-defined parameter. Randomly select one neighbor, say y, and construct a new sample z, which is defined as the linear combination of x and y. Because x and y are all in the same class and the new sample z has strong associations with them, z is more likely to be in such class and is added into this class.

In this study, we use the tool "SMOTE" in Weka [112], which implements the above-mentioned SMOTE method. For subtypes: basal, Her and LumB, several new samples are constructed via "SMOTE" and added to each subtype. Finally, each subtype contained 120 samples.

4.5. Performance Measurement

In this study, we used multi-class classifier to discriminate samples from four breast cancer subtypes. We evaluated the trained multi-class classifiers by using a 10-fold cross-validation [39,40,116–118]. For the predicted results yielded by 10-fold cross-validation, the sensitivity and specificity for each class were calculated. In addition, the overall accuracy and Matthews correlation coefficient (MCC) [119–121] were also computed to measure the classification performance. When defining X as the binary matrix of the predicted labels and Y as the binary matrix of the true labels, the MCC is calculated as follows:

$$MCC = \frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{cov}(X,X)\mathrm{cov}(Y,Y)}} = \frac{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{C}(x_{ij}-\overline{x}_j)(y_{ij}-\overline{y}_j)}{\sqrt{\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{C}(x_{ij}-\overline{x}_j)^2\sum\limits_{i=1}^{n}\sum\limits_{j=1}^{C}(y_{ij}-\overline{y}_j)^2}} \tag{4}$$

where cov(X,Y) represents the correlation coefficent of X and Y, $\overline{x}_j$ and $\overline{y}_j$ are the mean values in the j-th columns of X and Y, respectively, n represent the total number of samples and C indicates the number of labels.

4.6. Enrichment Analysis

We mapped the selected methylation probes onto genes based on the probe annotation file of Illumina HumanMethylation450 BeadChip downloaded from https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL13534. These mapped genes were enriched onto GO and KEGG by using the

Hypergeometric test. *p*-value was adjusted as false discovery rate (FDR). The GO terms and KEGG pathways with FDR smaller than 0.05 were considered significantly enriched biological functions.

5. Conclusions

This study investigated the methylation profiles of patients of four breast cancer subtypes with several machine learning algorithms. Some functional genes with different methylation status in different subtypes were identified. These genes can be latent biomarkers or can be used to build an efficient predictor for the identification of different breast cancer subtypes.

Supplementary Materials: The following are available online at http://www.mdpi.com/1422-0067/20/17/4269/s1.

Author Contributions: Conceptualization, T.H. and Y.-D.C.; methodology, L.C. and X.P.; formal analysis, T.Z. and Y.-H.Z.; data curation, T.H.; writing—original draft preparation, L.C. and T.Z.; writing—review and editing, X.P.; supervision, Y.-D.C.

Funding: This research was funded by the Shanghai Municipal Science and Technology Major Project (2017SHZDZX01), National Key R&D Program of China (2018YFC0910403), National Natural Science Foundation of China (31701151), Natural Science Foundation of Shanghai (17ZR1412500), Shanghai Sailing Program (16YF1413800), the Youth Innovation Promotion Association of Chinese Academy of Sciences (CAS) (2016245), the fund of the key Laboratory of Stem Cell Biology of Chinese Academy of Sciences (201703), Science and Technology Commission of Shanghai Municipality (STCSM) (18dz2271000).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Morris, M.; Woods, L.M.; Bhaskaran, K.; Rachet, B. Do pre-diagnosis primary care consultation patterns explain deprivation-specific differences in net survival among women with breast cancer? An examination of individually-linked data from the uk west midlands cancer registry, national screening programme and clinical practice research datalink. *BMC Cancer* **2017**, *17*, 155.
2. Cedolini, C.; Bertozzi, S.; Londero, A.P.; Bernardi, S.; Seriau, L.; Concina, S.; Cattin, F.; Risaliti, A. Type of breast cancer diagnosis, screening, and survival. *Clin. Breast Cancer* **2014**, *14*, 235–240. [CrossRef] [PubMed]
3. Seneviratne, S.; Campbell, I.; Scott, N.; Shirley, R.; Lawrenson, R. Impact of mammographic screening on ethnic and socioeconomic inequities in breast cancer stage at diagnosis and survival in new zealand: A cohort study. *BMC Public. Health* **2015**, *15*, 46. [CrossRef] [PubMed]
4. Hayes, J.; Richardson, A.; Frampton, C. Population attributable risks for modifiable lifestyle factors and breast cancer in new zealand women. *Intern. Med. J.* **2013**, *43*, 1198–1204. [CrossRef] [PubMed]
5. Howell, A.; Anderson, A.S.; Clarke, R.B.; Duffy, S.W.; Evans, D.G.; Garcia-Closas, M.; Gescher, A.J.; Key, T.J.; Saxton, J.M.; Harvie, M.N. Risk determination and prevention of breast cancer. *Breast Cancer Res. Bcr.* **2014**, *16*, 446. [CrossRef] [PubMed]
6. Huang, Z.; Wen, W.; Zheng, Y.; Gao, Y.T.; Wu, C.; Bao, P.; Wang, C.; Gu, K.; Peng, P.; Gong, Y.; et al. Breast cancer incidence and mortality: Trends over 40 years among women in shanghai, china. *Ann. Oncol.* **2016**, *27*, 1129–1134. [CrossRef]
7. Sung, H.; Ren, J.; Li, J.; Pfeiffer, R.M.; Wang, Y.; Guida, J.L.; Fang, Y.; Shi, J.; Zhang, K.; Li, N.; et al. Breast cancer risk factors and mammographic density among high-risk women in urban china. *NPJ Breast Cancer* **2018**, *4*, 3. [CrossRef]
8. Nelson, H.D.; Pappas, M.; Zakher, B.; Mitchell, J.P.; Okinaka-Hu, L.; Fu, R. Risk assessment, genetic counseling, and genetic testing for brca-related cancer in women: A systematic review to update the u.S. Preventive services task force recommendation. *Ann. Intern. Med.* **2014**, *160*, 255–266. [CrossRef]
9. Pan, X.; Hu, X.; Zhang, Y.-H.; Chen, L.; Zhu, L.; Wan, S.; Huang, T.; Cai, Y.-D. Identification of the copy number variant biomarkers for breast cancer subtypes. *Mol. Genet. Genom.* **2019**, *294*, 95–110. [CrossRef]
10. Deb, S.; Wong, S.Q.; Li, J.; Do, H.; Weiss, J.; Byrne, D.; Chakrabarti, A.; Bosma, T.; kConFab, I.; Fellowes, A.; et al. Mutational profiling of familial male breast cancers reveals similarities with luminal a female breast cancer with rare tp53 mutations. *Br. J. Cancer* **2014**, *111*, 2351–2360. [CrossRef]
11. Krishnamurti, U.; Silverman, J.F. Her2 in breast cancer: A review and update. *Adv. Anat. Pathol.* **2014**, *21*, 100–107. [CrossRef] [PubMed]

12. Gangi, A.; Cass, I.; Paik, D.; Barmparas, G.; Karlan, B.; Dang, C.; Li, A.; Walsh, C.; Rimel, B.J.; Amersi, F.F. Breast cancer following ovarian cancer in brca mutation carriers. *JAMA Surg.* **2014**, *149*, 1306–1313. [CrossRef] [PubMed]
13. Waldrep, A.R.; Avery, E.J.; Rose, F.F., Jr.; Midathada, M.V.; Tilford, J.A.; Kolberg, H.C.; Hutchins, M.R. Breast cancer subtype influences the accuracy of predicting pathologic response by imaging and clinical breast exam after neoadjuvant chemotherapy. *Anticancer Res.* **2016**, *36*, 5389–5395. [CrossRef] [PubMed]
14. Buist, D.S.; Bosco, J.L.; Silliman, R.A.; Gold, H.T.; Field, T.; Yood, M.U.; Quinn, V.P.; Prout, M.; Lash, T.L.; Breast Cancer Outcomes in Older Women, I. Long-term surveillance mammography and mortality in older women with a history of early stage invasive breast cancer. *Breast Cancer Res. Treat.* **2013**, *142*, 153–163. [CrossRef] [PubMed]
15. Giannakeas, V.; Lubinski, J.; Gronwald, J.; Moller, P.; Armel, S.; Lynch, H.T.; Foulkes, W.D.; Kim-Sing, C.; Singer, C.; Neuhausen, S.L.; et al. Mammography screening and the risk of breast cancer in brca1 and brca2 mutation carriers: A prospective study. *Breast Cancer Res. Treat.* **2014**, *147*, 113–118. [CrossRef] [PubMed]
16. Sana, M.; Malik, H.J. Current and emerging breast cancer biomarkers. *J. Cancer Res. Ther.* **2015**, *11*, 508–513. [CrossRef] [PubMed]
17. Weigel, M.T.; Dowsett, M. Current and emerging biomarkers in breast cancer: Prognosis and prediction. *Endocr. -Relat. Cancer* **2010**, *17*, R245–R262. [CrossRef]
18. Wang, D.; Li, J.-R.; Zhang, Y.-H.; Chen, L.; Huang, T.; Cai, Y.-D. Identification of differentially expressed genes between original breast cancer and xenograft using machine learning algorithms. *Genes* **2018**, *9*, 155. [CrossRef]
19. Cai, Y.D.; Zhang, Q.; Zhang, Y.H.; Chen, L.; Huang, T. Identification of genes associated with breast cancer metastasis to bone on a protein-protein interaction network with a shortest path algorithm. *J. Proteome Res.* **2017**, *16*, 1027–1038. [CrossRef]
20. Li, X.C.; Liu, C.; Huang, T.; Zhong, Y. The occurrence of genetic alterations during the progression of breast carcinoma. *BioMed Res. Int.* **2016**, *2016*, 5237827. [CrossRef]
21. Fleischer, T.; Tekpli, X.; Mathelier, A.; Wang, S.; Nebdal, D.; Dhakal, H.P.; Sahlberg, K.K.; Schlichting, E.; Oslo Breast Cancer Research, C.; Borresen-Dale, A.L.; et al. DNA methylation at enhancers identifies distinct breast cancer lineages. *Nat. Commun.* **2017**, *8*, 1379. [CrossRef] [PubMed]
22. Bertoli, G.; Cava, C.; Castiglioni, I. Micrornas: New biomarkers for diagnosis, prognosis, therapy prediction and therapeutic tools for breast cancer. *Theranostics* **2015**, *5*, 1122–1143. [CrossRef] [PubMed]
23. Ali, H.R.; Rueda, O.M.; Chin, S.F.; Curtis, C.; Dunning, M.J.; Aparicio, S.A.; Caldas, C. Genome-driven integrated classification of breast cancer validated in over 7500 samples. *Genome Biol.* **2014**, *15*, 431. [CrossRef] [PubMed]
24. Hagemann, I.S. Molecular testing in breast cancer: A guide to current practices. *Arch. Pathol. Lab. Med.* **2016**, *140*, 815–824. [CrossRef] [PubMed]
25. Kanwal, R.; Gupta, K.; Gupta, S. Cancer epigenetics: An introduction. *Methods Mol. Biol.* **2015**, *1238*, 3–25.
26. Herceg, Z.; Ushijima, T. Introduction: Epigenetics and cancer. *Adv. Genet.* **2010**, *70*, 1–23. [PubMed]
27. Zeleznik-Le, N.J. Introduction to progress and promise of epigenetics for diagnosis and therapy in cancer. *Cancer Genet.* **2015**, *208*, 165–166. [CrossRef]
28. Santos, G.C., Jr.; da Silva, A.P.; Feldman, L.; Ventura, G.M.; Vassetzky, Y.; de Moura Gallo, C.V. Epigenetic modifications, chromatin distribution and tp53 transcription in a model of breast cancer progression. *J. Cell. Biochem.* **2015**, *116*, 533–541. [CrossRef]
29. Stefansson, O.A.; Esteller, M. Epigenetic modifications in breast cancer and their role in personalized medicine. *Am. J. Pathol.* **2013**, *183*, 1052–1063. [CrossRef]
30. Ching, T.; Himmelstein, D.S.; Beaulieu-Jones, B.K.; Kalinin, A.A.; Do, B.T.; Way, G.P.; Ferrero, E.; Agapow, P.M.; Zietz, M.; Hoffman, M.M.; et al. Opportunities and obstacles for deep learning in biology and medicine. *J. R. Soc. Interface* **2018**, *15*, 20170387. [CrossRef]
31. Camacho, D.M.; Collins, K.M.; Powers, R.K.; Costello, J.C.; Collins, J.J. Next-generation machine learning for biological networks. *Cell* **2018**, *173*, 1581–1592. [CrossRef] [PubMed]
32. Kerschbaum, H.H.; Kainz, V.; Hermann, A. Sarcoplasmic calcium-binding protein-immunoreactive material in the central nervous system of the snail, helix pomatia. *Brain Res.* **1992**, *597*, 339–342. [CrossRef]

33. Adorjan, P.; Distler, J.; Lipscher, E.; Model, F.; Muller, J.; Pelet, C.; Braun, A.; Florl, A.R.; Gutig, D.; Grabs, G.; et al. Tumour class prediction and discovery by microarray-based DNA methylation analysis. *Nucleic Acids Res.* **2002**, *30*, e21. [CrossRef] [PubMed]
34. Chen, L.; Pan, X.; Hu, X.; Zhang, Y.H.; Wang, S.; Huang, T.; Cai, Y.D. Gene expression differences among different msi statuses in colorectal cancer. *Int. J. Cancer* **2018**, *143*, 1731–1740. [CrossRef] [PubMed]
35. Shipp, M.A.; Ross, K.N.; Tamayo, P.; Weng, A.P.; Kutok, J.L.; Aguiar, R.C.; Gaasenbeek, M.; Angelo, M.; Reich, M.; Pinkus, G.S.; et al. Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nat. Med.* **2002**, *8*, 68–74. [CrossRef] [PubMed]
36. Ye, Q.H.; Qin, L.X.; Forgues, M.; He, P.; Kim, J.W.; Peng, A.C.; Simon, R.; Li, Y.; Robles, A.I.; Chen, Y.; et al. Predicting hepatitis b virus-positive metastatic hepatocellular carcinomas using gene expression profiling and supervised machine learning. *Nat. Med.* **2003**, *9*, 416–423. [CrossRef] [PubMed]
37. Sweeney, C.; Bernard, P.S.; Factor, R.E.; Kwan, M.L.; Habel, L.A.; Quesenberry, C.P., Jr.; Shakespear, K.; Weltzien, E.K.; Stijleman, I.J.; Davis, C.A.; et al. Intrinsic subtypes from pam50 gene expression assay in a population-based breast cancer cohort: Differences by age, race, and tumor characteristics. *Cancer Epidemiol. Biomark. Prev.* **2014**, *23*, 714–724. [CrossRef] [PubMed]
38. Breiman, L. Random forests. *Mach. Learn.* **2001**, *45*, 5–32. [CrossRef]
39. Zhao, X.; Chen, L.; Guo, Z.-H.; Liu, T. Predicting drug side effects with compact integration of heterogeneous networks. *Curr. Bioinform.* **2019**. [CrossRef]
40. Zhao, X.; Chen, L.; Lu, J. A similarity-based method for prediction of drug side effects with heterogeneous information. *Math. Biosci.* **2018**, *306*, 136–144. [CrossRef]
41. De Summa, S.; Pinto, R.; Pilato, B.; Sambiasi, D.; Porcelli, L.; Guida, G.; Mattioli, E.; Paradiso, A.; Merla, G.; Micale, L.; et al. Expression of base excision repair key factors and mir17 in familial and sporadic breast cancer. *Cell Death Dis.* **2014**, *5*, e1076. [CrossRef] [PubMed]
42. Lee, C.J.; Evans, J.; Kim, K.; Chae, H.; Kim, S. Determining the effect of DNA methylation on gene expression in cancer cells. *Methods Mol. Biol.* **2014**, *1101*, 161–178. [PubMed]
43. Ishizuka, T.; Rozehnal, V.; Fischer, T.; Kato, A.; Endo, S.; Yoshigae, Y.; Kurihara, A.; Izumi, T. Interindividual variability of carboxymethylenebutenolidase homolog, a novel olmesartan medoxomil hydrolase, in the human liver and intestine. *Drug Metab. Dispos.* **2013**, *41*, 1156–1162. [CrossRef] [PubMed]
44. Xu, H.; Lam, S.H.; Shen, Y.; Gong, Z. Genome-wide identification of molecular pathways and biomarkers in response to arsenic exposure in zebrafish liver. *PLoS ONE* **2013**, *8*, e68737. [CrossRef] [PubMed]
45. Shashova, E.E.; Lyupina, Y.V.; Glushchenko, S.A.; Slonimskaya, E.M.; Savenkova, O.V.; Kulikov, A.M.; Gornostaev, N.G.; Kondakova, I.V.; Sharova, N.P. Proteasome functioning in breast cancer: Connection with clinical-pathological factors. *PLoS ONE* **2014**, *9*, e109933. [CrossRef] [PubMed]
46. Andrade, S.S.; Gouvea, I.E.; Silva, M.C.; Castro, E.D.; de Paula, C.A.; Okamoto, D.; Oliveira, L.; Peres, G.B.; Ottaiano, T.; Facina, G.; et al. Cathepsin k induces platelet dysfunction and affects cell signaling in breast cancer—Molecularly distinct behavior of cathepsin k in breast cancer. *BMC Cancer* **2016**, *16*, 173. [CrossRef] [PubMed]
47. Xia, P.; Jin, T.; Geng, T.; Sun, T.; Li, X.; Dang, C.; Kang, L.; Chen, C.; Sun, J. Polymorphisms in esr1 and flj43663 are associated with breast cancer risk in the han population. *Tumour Biol.* **2014**, *35*, 2187–2190. [CrossRef]
48. Li, L.; Zhang, G.Q.; Chen, H.; Zhao, Z.J.; Chen, H.Z.; Liu, H.; Wang, G.; Jia, Y.H.; Pan, S.H.; Kong, R.; et al. Plasma and tumor levels of linc-pint are diagnostic and prognostic biomarkers for pancreatic cancer. *Oncotarget* **2016**, *7*, 71773–71781. [CrossRef]
49. Garitano-Trojaola, A.; Jose-Eneriz, E.S.; Ezponda, T.; Unfried, J.P.; Carrasco-Leon, A.; Razquin, N.; Barriocanal, M.; Vilas-Zornoza, A.; Sangro, B.; Segura, V.; et al. Deregulation of linc-pint in acute lymphoblastic leukemia is implicated in abnormal proliferation of leukemic cells. *Oncotarget* **2018**, *9*, 12842–12852. [CrossRef]
50. Xu, Y.; Chen, M.; Liu, C.; Zhang, X.; Li, W.; Cheng, H.; Zhu, J.; Zhang, M.; Chen, Z.; Zhang, B. Association study confirmed three breast cancer-specific molecular subtype-associated susceptibility loci in chinese han women. *Oncologist* **2017**, *22*, 890–894. [CrossRef]
51. Van Itallie, C.M.; Tietgens, A.J.; Aponte, A.; Fredriksson, K.; Fanning, A.S.; Gucek, M.; Anderson, J.M. Biotin ligase tagging identifies proteins proximal to e-cadherin, including lipoma preferred partner, a regulator of epithelial cell-cell and cell-substrate adhesion. *J. Cell Sci.* **2014**, *127*, 885–895. [CrossRef] [PubMed]

52. Gregory Call, S.; Brereton, D.; Bullard, J.T.; Chung, J.Y.; Meacham, K.L.; Morrell, D.J.; Reeder, D.J.; Schuler, J.T.; Slade, A.D.; Hansen, M.D. A zyxin-nectin interaction facilitates zyxin localization to cell-cell adhesions. *Biochem. Biophys. Res. Commun.* **2011**, *415*, 485–489. [CrossRef] [PubMed]
53. Huggins, C.J.; Andrulis, I.L. Cell cycle regulated phosphorylation of limd1 in cell lines and expression in human breast cancers. *Cancer Lett.* **2008**, *267*, 55–66. [CrossRef] [PubMed]
54. Ngan, E.; Northey, J.J.; Brown, C.M.; Ursini-Siegel, J.; Siegel, P.M. A complex containing lpp and alpha-actinin mediates TGF β-induced migration and invasion of ERBB2-expressing breast cancer cells. *J. Cell Sci.* **2013**, *126*, 1981–1991. [CrossRef] [PubMed]
55. Ngan, E.; Stoletov, K.; Smith, H.W.; Common, J.; Muller, W.J.; Lewis, J.D.; Siegel, P.M. Lpp is a src substrate required for invadopodia formation and efficient breast cancer lung metastasis. *Nat. Commun.* **2017**, *8*, 15059. [CrossRef]
56. Yang, S.; Zhou, L.; Reilly, P.T.; Shen, S.M.; He, P.; Zhu, X.N.; Li, C.X.; Wang, L.S.; Mak, T.W.; Chen, G.Q.; et al. Anp32b deficiency impairs proliferation and suppresses tumor progression by regulating akt phosphorylation. *Cell Death Dis.* **2016**, *7*, e2082. [CrossRef]
57. Shen, S.M.; Yu, Y.; Wu, Y.L.; Cheng, J.K.; Wang, L.S.; Chen, G.Q. Downregulation of anp32b, a novel substrate of caspase-3, enhances caspase-3 activation and apoptosis induction in myeloid leukemic cells. *Carcinogenesis* **2010**, *31*, 419–426. [CrossRef]
58. Leo, V.I.; Bunte, R.M.; Reilly, P.T. Balb/c-congenic anp32b-deficient mice reveal a modifying locus that determines viability. *Exp. Anim.* **2016**, *65*, 53–62. [CrossRef]
59. Cieply, B.; Park, J.W.; Nakauka-Ddamba, A.; Bebee, T.W.; Guo, Y.; Shang, X.; Lengner, C.J.; Xing, Y.; Carstens, R.P. Multiphasic and dynamic changes in alternative splicing during induction of pluripotency are coordinated by numerous rna-binding proteins. *Cell Rep.* **2016**, *15*, 247–255. [CrossRef]
60. Lin, X.; Huang, X.; Uziel, T.; Hessler, P.; Albert, D.H.; Roberts-Rapp, L.A.; McDaniel, K.F.; Kati, W.M.; Shen, Y. Hexim1 as a robust pharmacodynamic marker for monitoring target engagement of bet family bromodomain inhibitors in tumors and surrogate tissues. *Mol. Cancer Ther.* **2017**, *16*, 388–396. [CrossRef]
61. Zeng, H.; Qu, J.; Jin, N.; Xu, J.; Lin, C.; Chen, Y.; Yang, X.; He, X.; Tang, S.; Lan, X.; et al. Feedback activation of leukemia inhibitory factor receptor limits response to histone deacetylase inhibitors in breast cancer. *Cancer Cell* **2016**, *30*, 459–473. [CrossRef] [PubMed]
62. Yeo, S.Y.; Ha, S.Y.; Yu, E.J.; Lee, K.W.; Kim, J.H.; Kim, S.H. Znf282 (zinc finger protein 282), a novel e2f1 co-activator, promotes esophageal squamous cell carcinoma. *Oncotarget* **2014**, *5*, 12260–12272. [CrossRef] [PubMed]
63. Rakha, E.A.; Lee, A.H.; Roberts, J.; Villena Salinas, N.M.; Hodi, Z.; Ellis, I.O.; Reis-Filho, J.S. Low-estrogen receptor-positive breast cancer: The impact of tissue sampling, choice of antibody, and molecular subtyping. *J. Clin. Oncol.* **2012**, *30*, 2929–2930. [CrossRef] [PubMed]
64. Balleine, R.L.; Wilcken, N.R. High-risk estrogen-receptor-positive breast cancer: Identification and implications for therapy. *Mol. Diagn. Ther.* **2012**, *16*, 235–240. [CrossRef] [PubMed]
65. Nielsen, T.O.; Parker, J.S.; Leung, S.; Voduc, D.; Ebbert, M.; Vickery, T.; Davies, S.R.; Snider, J.; Stijleman, I.J.; Reed, J.; et al. A comparison of pam50 intrinsic subtyping with immunohistochemistry and clinical prognostic factors in tamoxifen-treated estrogen receptor-positive breast cancer. *Clin. Cancer Res.* **2010**, *16*, 5222–5232. [CrossRef] [PubMed]
66. Breusegem, S.Y.; Seaman, M.N.J. Genome-wide rnai screen reveals a role for multipass membrane proteins in endosome-to-golgi retrieval. *Cell Rep.* **2014**, *9*, 1931–1945. [CrossRef] [PubMed]
67. Savci-Heijink, C.D.; Halfwerk, H.; Koster, J.; van de Vijver, M.J. A novel gene expression signature for bone metastasis in breast carcinomas. *Breast Cancer Res. Treat.* **2016**, *156*, 249–259. [CrossRef] [PubMed]
68. Takada, K.; Zhu, D.; Bird, G.H.; Sukhdeo, K.; Zhao, J.J.; Mani, M.; Lemieux, M.; Carrasco, D.E.; Ryan, J.; Horst, D.; et al. Targeted disruption of the bcl9/beta-catenin complex inhibits oncogenic wnt signaling. *Sci. Transl. Med.* **2012**, *4*, 148ra117. [CrossRef]
69. Elsarraj, H.S.; Hong, Y.; Valdez, K.E.; Michaels, W.; Hook, M.; Smith, W.P.; Chien, J.; Herschkowitz, J.I.; Troester, M.A.; Beck, M.; et al. Expression profiling of in vivo ductal carcinoma in situ progression models identified b cell lymphoma-9 as a molecular driver of breast cancer invasion. *Breast Cancer Res.* **2015**, *17*, 128. [CrossRef] [PubMed]

70. Toya, H.; Oyama, T.; Ohwada, S.; Togo, N.; Sakamoto, I.; Horiguchi, J.; Koibuchi, Y.; Adachi, S.; Jigami, T.; Nakajima, T.; et al. Immunohistochemical expression of the beta-catenin-interacting protein b9l is associated with histological high nuclear grade and immunohistochemical ERBB2/HER-2 expression in breast cancers. *Cancer Sci.* **2007**, *98*, 484–490. [CrossRef]
71. Bastien, R.R.; Rodriguez-Lescure, A.; Ebbert, M.T.; Prat, A.; Munarriz, B.; Rowe, L.; Miller, P.; Ruiz-Borrego, M.; Anderson, D.; Lyons, B.; et al. Pam50 breast cancer subtyping by RT-qPCR and concordance with standard clinical molecular markers. *BMC Med. Genom.* **2012**, *5*, 44. [CrossRef] [PubMed]
72. Ogi, S.; Fujita, H.; Kashihara, M.; Yamamoto, C.; Sonoda, K.; Okamoto, I.; Nakagawa, K.; Ohdo, S.; Tanaka, Y.; Kuwano, M.; et al. Sorting nexin 2-mediated membrane trafficking of c-met contributes to sensitivity of molecular-targeted drugs. *Cancer Sci.* **2013**, *104*, 573–583. [CrossRef] [PubMed]
73. Rivera, J.; Megias, D.; Bravo, J. Sorting nexin 6 interacts with breast cancer metastasis suppressor-1 and promotes transcriptional repression. *J. Cell. Biochem.* **2010**, *111*, 1464–1472. [CrossRef] [PubMed]
74. Bendris, N.; Williams, K.C.; Reis, C.R.; Welf, E.S.; Chen, P.H.; Lemmers, B.; Hahne, M.; Leong, H.S.; Schmid, S.L. Snx9 promotes metastasis by enhancing cancer cell invasion via differential regulation of rhogtpases. *Mol. Biol. Cell* **2016**. [CrossRef] [PubMed]
75. Ng, B.G.; Lourenco, C.M.; Losfeld, M.E.; Buckingham, K.J.; Kircher, M.; Nickerson, D.A.; Shendure, J.; Bamshad, M.J.; University of Washington Center for Mendelian, G.; Freeze, H.H. Mutations in the translocon-associated protein complex subunit ssr3 cause a novel congenital disorder of glycosylation. *J. Inherit. Metab. Dis.* **2019**. [CrossRef] [PubMed]
76. Bano-Polo, M.; Martinez-Garay, C.A.; Grau, B.; Martinez-Gil, L.; Mingarro, I. Membrane insertion and topology of the translocon-associated protein (TRAP) gamma subunit. *Biochim. Biophys. Acta Biomembr.* **2017**, *1859*, 903–909. [CrossRef]
77. Hadad, S.M.; Coates, P.; Jordan, L.B.; Dowling, R.J.; Chang, M.C.; Done, S.J.; Purdie, C.A.; Goodwin, P.J.; Stambolic, V.; Moulder-Thompson, S.; et al. Evidence for biological effects of metformin in operable breast cancer: Biomarker analysis in a pre-operative window of opportunity randomized trial. *Breast Cancer Res. Treat.* **2015**, *150*, 149–155. [CrossRef] [PubMed]
78. Hadad, S.; Iwamoto, T.; Jordan, L.; Purdie, C.; Bray, S.; Baker, L.; Jellema, G.; Deharo, S.; Hardie, D.G.; Pusztai, L.; et al. Evidence for biological effects of metformin in operable breast cancer: A pre-operative, window-of-opportunity, randomized trial. *Breast Cancer Res. Treat.* **2011**, *128*, 783–794. [CrossRef]
79. Marchitti, S.A.; Brocker, C.; Orlicky, D.J.; Vasiliou, V. Molecular characterization, expression analysis, and role of aldh3b1 in the cellular protection against oxidative stress. *Free Radic. Biol. Med.* **2010**, *49*, 1432–1443. [CrossRef]
80. Marchitti, S.A.; Orlicky, D.J.; Vasiliou, V. Expression and initial characterization of human aldh3b1. *Biochem. Biophys. Res. Commun.* **2007**, *356*, 792–798. [CrossRef]
81. Sladek, N.E. Transient induction of increased aldehyde dehydrogenase 3a1 levels in cultured human breast (adeno)carcinoma cell lines via 5′-upstream xenobiotic, and electrophile, responsive elements is, respectively, estrogen receptor-dependent and -independent. *Chem. Biol. Interact.* **2003**, *143*, 63–74. [CrossRef]
82. Zhao, Q.; Zhu, Y.; Liu, L.; Wang, H.; Jiang, S.; Hu, X.; Guo, J. Stk39 blockage by rna interference inhibits the proliferation and induces the apoptosis of renal cell carcinoma. *Onco Targets Ther.* **2018**, *11*, 1511–1519. [CrossRef] [PubMed]
83. Donner, K.M.; Hiltunen, T.P.; Hannila-Handelberg, T.; Suonsyrja, T.; Kontula, K. Stk39 variation predicts the ambulatory blood pressure response to losartan in hypertensive men. *Hypertens. Res.* **2012**, *35*, 107–114. [CrossRef] [PubMed]
84. Astolfi, A.; Landuzzi, L.; Nicoletti, G.; De Giovanni, C.; Croci, S.; Palladini, A.; Ferrini, S.; Iezzi, M.; Musiani, P.; Cavallo, F.; et al. Gene expression analysis of immune-mediated arrest of tumorigenesis in a transgenic mouse model of her-2/neu-positive basal-like mammary carcinoma. *Am. J. Pathol.* **2005**, *166*, 1205–1216. [CrossRef]
85. Balatoni, C.E.; Dawson, D.W.; Suh, J.; Sherman, M.H.; Sanders, G.; Hong, J.S.; Frank, M.J.; Malone, C.S.; Said, J.W.; Teitell, M.A. Epigenetic silencing of STK39 in b-cell lymphoma inhibits apoptosis from genotoxic stress. *Am. J. Pathol.* **2009**, *175*, 1653–1661. [CrossRef] [PubMed]
86. Malek, N.P. Cux1 mediates tumour cell survival: Implications for future therapies? *Gut* **2010**, *59*, 1014–1015. [CrossRef] [PubMed]

87. Cubelos, B.; Sebastian-Serrano, A.; Beccari, L.; Calcagnotto, M.E.; Cisneros, E.; Kim, S.; Dopazo, A.; Alvarez-Dolado, M.; Redondo, J.M.; Bovolenta, P.; et al. Cux1 and cux2 regulate dendritic branching, spine morphology, and synapses of the upper layer neurons of the cortex. *Neuron* **2010**, *66*, 523–535. [CrossRef]
88. Chen, J.; Zhou, Z.; Yao, Y.; Dai, J.; Zhou, D.; Wang, L.; Zhang, Q.Q. Dipalmitoylphosphatidic acid inhibits breast cancer growth by suppressing angiogenesis via inhibition of the CUX1/FGF1/HGF signalling pathway. *J. Cell. Mol. Med.* **2018**, *22*, 4760–4770. [CrossRef]
89. Hulea, L.; Nepveu, A. Cux1 transcription factors: From biochemical activities and cell-based assays to mouse models and human diseases. *Gene* **2012**, *497*, 18–26. [CrossRef]
90. Zhang, Y.; Ren, S.; Yuan, F.; Zhang, K.; Fan, Y.; Zheng, S.; Gao, Z.; Zhao, J.; Mu, T.; Zhao, S.; et al. Mir-135 promotes proliferation and stemness of oesophageal squamous cell carcinoma by targeting rerg. *Artif. Cells Nanomed. Biotechnol.* **2018**, *46*, 1210–1219. [CrossRef]
91. Habashy, H.O.; Powe, D.G.; Glaab, E.; Ball, G.; Spiteri, I.; Krasnogor, N.; Garibaldi, J.M.; Rakha, E.A.; Green, A.R.; Caldas, C.; et al. Rerg (ras-like, oestrogen-regulated, growth-inhibitor) expression in breast cancer: A marker of er-positive luminal-like subtype. *Breast Cancer Res. Treat.* **2011**, *128*, 315–326. [CrossRef] [PubMed]
92. Finlin, B.S.; Gau, C.L.; Murphy, G.A.; Shao, H.; Kimel, T.; Seitz, R.S.; Chiu, Y.F.; Botstein, D.; Brown, P.O.; Der, C.J.; et al. Rerg is a novel ras-related, estrogen-regulated and growth-inhibitory gene in breast cancer. *J. Biol. Chem.* **2001**, *276*, 42259–42267. [CrossRef] [PubMed]
93. Li, J.; Lu, L.; Zhang, Y.H.; Liu, M.; Chen, L.; Huang, T.; Cai, Y.D. Identification of synthetic lethality based on a functional network by using machine learning algorithms. *J. Cell. Biochem.* **2019**, *120*, 405–416. [CrossRef] [PubMed]
94. Chen, L.; Pan, X.; Zhang, Y.-H.; Liu, M.; Huang, T.; Cai, Y.-D. Classification of widely and rarely expressed genes with recurrent neural network. *Comput. Struct. Biotechnol. J.* **2019**, *17*, 49–60. [CrossRef] [PubMed]
95. Zhang, T.M.; Huang, T.; Wang, R.F. Cross talk of chromosome instability, cpg island methylator phenotype and mismatch repair in colorectal cancer. *Oncol. Lett.* **2018**, *16*, 1736–1746. [CrossRef] [PubMed]
96. Li, J.; Huang, T. Predicting and analyzing early wake-up associated gene expressions by integrating gwas and eqtl studies. *Biochim. Et Biophys. Acta Mol. Basis Dis.* **2018**, *1864*, 2241–2246. [CrossRef]
97. 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]
98. Chen, L.; Wang, S.; Zhang, Y.H.; Wei, L.; Xu, X.; Huang, T.; Cai, Y.D. Prediction of nitrated tyrosine residues in protein sequences by extreme learning machine and feature selection methods. *Comb Chem High Throughput Screen* **2018**, *21*, 393–402. [CrossRef]
99. Cai, L.; Huang, T.; Su, J.; Zhang, X.; Chen, W.; Zhang, F.; He, L.; Chou, K.-C. Implications of newly identified brain EQTL genes and their interactors in schizophrenia. *Mol. Ther.-Nucleic Acids* **2018**, *12*, 433–442. [CrossRef]
100. Wang, S.-B.; Huang, T.J.M.B.R. The early detection of asthma based on blood gene expression. *Mol. Biol. Rep.* **2019**, *46*, 217–223. [CrossRef]
101. Chen, L.; Zhang, S.; Pan, X.; Hu, X.; Zhang, Y.H.; Yuan, F.; Huang, T.; Cai, Y.D. Hiv infection alters the human epigenetic landscape. *Gene Ther.* **2019**, *26*, 29–39. [CrossRef] [PubMed]
102. Li, C.; Wang, X.; Dong, W.; Yan, J.; Liu, Q.; Zha, H. Joint active learning with feature selection via cur matrix decomposition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 1382–1396. [CrossRef] [PubMed]
103. Chen, L.; Pan, X.; Zhang, Y.-H.; Kong, X.; Huang, T.; Cai, Y.-D. Tissue differences revealed by gene expression profiles of various cell lines. *J. Cell. Biochem.* **2019**, *120*, 7068–7081. [CrossRef] [PubMed]
104. Pan, X.; Hu, X.; Zhang, Y.H.; Feng, K.; Wang, S.P.; Chen, L.; Huang, T.; Cai, Y.D. Identifying patients with atrioventricular septal defect in down syndrome populations by using self-normalizing neural networks and feature selection. *Genes* **2018**, *9*, 208. [CrossRef] [PubMed]
105. 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.* **2018**, *119*, 3394–3403. [CrossRef] [PubMed]
106. Pan, X.; Chen, L.; Feng, K.-Y.; Hu, X.-H.; Zhang, Y.-H.; Kong, X.-Y.; Huang, T.; Cai, Y.-D. Analysis of expression pattern of snornas in different cancer types with machine learning algorithms. *Int. J. Mol. Sci.* **2019**, *20*, 2185. [CrossRef]

107. Chen, L.; Pan, X.; Zhang, Y.-H.; Hu, X.; Feng, K.; Huang, T.; Cai, Y.-D. Primary tumor site specificity is preserved in patient-derived tumor xenograft models. *Front. Genet.* **2019**. [CrossRef]
108. Chen, L.; Pan, X.; Zhang, Y.-H.; Huang, T.; Cai, Y.-D. Analysis of gene expression differences between different pancreatic cells. *ACS Omega* **2019**, *4*, 6421–6435. [CrossRef]
109. Li, J.; Lan, C.-N.; Kong, Y.; Feng, S.-S.; Huang, T. Identification and analysis of blood gene expression signature for osteoarthritis with advanced feature selection methods. *Front. Genet.* **2018**, *9*, 246. [CrossRef]
110. Li, J.; Chen, L.; Zhang, Y.-H.; Kong, X.; Huang, T.; Cai, Y.-D. A computational method for classifying different human tissues with quantitatively tissue-specific expressed genes. *Genes* **2018**, *9*, 449. [CrossRef]
111. Cui, H.; Chen, L. A binary classifier for the prediction of ec numbers of enzymes. *Curr. Proteom.* **2019**, *16*, 381–389. [CrossRef]
112. Witten, I.H.; Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*; Morgan Kaufmann: San Francisco, CA, USA, 2005.
113. Platt, J. *Fast Training of Support Vector Machines Using Sequential Minimal Optimization*; MIT Press: Cambridge, UK, 1998.
114. Keerthi, S.S.; Shevade, S.K.; Bhattacharyya, C.; Murthy, K.R.K. Improvements to platt's smo algorithm for svm classifier design. *Neural. Comput.* **2001**, *13*, 637–649. [CrossRef]
115. Chawla, N.V.; Bowyer, K.W.; Hall, L.O.; Kegelmeyer, W.P. Smote: Synthetic minority over-sampling technique. *J. Artif. Intell. Res.* **2002**, *16*, 321–357. [CrossRef]
116. Kohavi, R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *International Joint Conference on Artificial Intelligence*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA; Lawrence Erlbaum Associates: Mahwah, NJ, USA, 1995; pp. 1137–1145.
117. 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]
118. Che, J.; Chen, L.; Guo, Z.-H.; Wang, S.; Aorigele. Drug target group prediction with multiple drug networks. *Comb. Chem. High Throughput Screen.* **2019**. [CrossRef] [PubMed]
119. Matthews, B. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochim. Et Biophys. Acta (BBA)-Protein Struct.* **1975**, *405*, 442–451. [CrossRef]
120. Chen, L.; Chu, C.; Zhang, Y.-H.; Zheng, M.-Y.; Zhu, L.; Kong, X.; Huang, T. Identification of drug-drug interactions using chemical interactions. *Curr. Bioinform.* **2017**, *12*, 526–534. [CrossRef]
121. Gorodkin, J. Comparing two k-category assignments by a k-category correlation coefficient. *Comput. Biol. Chem.* **2004**, *28*, 367–374. [CrossRef]