Informatics in Medicine Unlocked 24 (2021) 100629

Contents lists available at ScienceDirect

Informatics in Medicine Unlocked

journal homepage: www.elsevier.com/locate/imu

Integration of meta-analysis and supervised machine learning for pattern recognition in breast cancer using epigenetic data

Reza Panahi [a,*], Esmaeil Ebrahimie [a,b,c], Ali Niazi [a], Alireza Afsharifar [d]

[a] Institute of Biotechnology, Shiraz University, Shiraz, Iran
[b] Genomics Research Platform, School of Life Sciences, College of Science, Health and Engineering, La Trobe University, Melbourne, Victoria 3086, Australia
[c] School of Animal and Veterinary Sciences, The University of Adelaide, South Australia 5371, Australia
[d] Department of Plant Protection, College of Agriculture, Shiraz University, Shiraz, Iran

ARTICLE INFO

Keywords:
Machine learning
Meta-analysis
Systems biology
Breast cancer
ChIP-seq

ABSTRACT

Breast cancer is one of the most widespread diseases with high incidence and mortality rate in females. The accurate biomarker discovery for the early detection of patients prone to breast cancer is crucial in the treatment and diagnosis of breast cancer. The current study employed a comprehensive approach to detect an epigenomic data pattern of breast cancer using meta-analysis and machine learning approaches. Meta-analysis is a precise method that combines the results of multiple experiments. On the other hand, integrating and combining the test results through machine learning algorithms can deal with data complexity and heterogeneity. The main purpose of the current study was to discover the patterns of epigenome changes in the treatment and prognosis of breast cancer. NCBI and EBI databases were searched for ChIP-Seq data regarding the effect of the drugs on breast cancer. There were ten investigations carried out, four of which were appropriate meta-analysis. *NOV*, *JUN* and *ZBTB7A* transcription factors were identified as the biomarkers of breast cancer. Finally, pattern recognition was performed using nine different attribute weighting algorithms. Fourteen genes were selected by the majority of attribute weighting algorithms as the most informative genes including *KIP*, *TCF12*, *ABCC5*, *HDAC11*, *IPP*, *HIST1H2AM*, *ZNF33B*, *PHF2*, *ELAVL3*, *TBC1D9B*, *TMEM217*, *CD34*, *ARHGEF26*, and *CENPL*. The selected genes play vital roles in the occurrence of neoplasms and breast cancer. In this study, using a combination of meta-analysis and data mining, more comprehensive and reliable information were derived compared to the individual studies.

1. Introduction

Breast cancer is one of the most common malignant tumors [1] with the highest prevalence and mortality among women [2–4]. There is a growing concern worldwide associated with rising numbers of patients and their resistance to drugs [5]. Despite the considerable advances made in the early detection and clinical treatments, there are still various constraints including the molecular heterogeneity, resistance to endocrinology, diagnosis of disease progression, and the risk of disease recurrence. These limitations have led many researchers to identify new biomarkers in the disease progression and signaling pathways in order to facilitate the improvement of diagnostic and treatment procedures. A better understanding of cellular and molecular pathways of breast cancer is required to improve the treatment choices, clinical results, and consequently, prevention of the disease [6,7]. The early diagnosis and treatment have a significant importance in order to eliminate the disease before the metastatic stage; therefore, it is highly required to detect it at the early stages [1]. To predict and treat breast cancer in a timely manner, risk prediction models are implemented to identify women at risk of disease. Advanced preventive treatments and screening could also be used to identify eligible individuals and prevent the disease [8]. Moreover, the improvement of clinical outcomes requires discovering the therapeutic and prognostic biomarkers [9].

Epigenetic alterations can change the structure of chromatin through transforming its components, which leads to the transformation of the gene expression pattern. The most important epigenetic mechanisms include chromatin-modifying factors, histone-modifying agents, histone variations, and DNA methylation. These mechanisms are able to regulate the transcription machinery [10]. The functional modification of genes would be achieved through changing the number of regulatory

* Corresponding author.
E-mail address: rezapanahi222@gmail.com (R. Panahi).

https://doi.org/10.1016/j.imu.2021.100629
Received 1 February 2021; Received in revised form 30 May 2021; Accepted 30 May 2021
Available online 3 June 2021

proteins of chromatin with the purpose of reaching the target position, as well as connecting them to these positions [11]. This provides evidence that epigenetic modifications such as DNA methylation and chromatin remodeling at the early stages play a vital role in breast cancer [12].

The chromatin immune-precipitation method is a critical technique applied with the purpose of identifying promoter motifs (binding sites), transcription factors, and regulatory events. Understanding the interaction of protein with DNA and regulating gene expression could also help researchers to recognize the key biological processes [13,14].

Meta-analysis is a statistical analysis applied with the purpose of integrating the data collected from independently conducted studies [15,16]. Moreover, it is a precise process that combines the results of various experiments and derives more accurate and comprehensive conclusions [17].

Data mining is another relatively novel method, which is considered as the most crucial technology for efficient pattern discovery within data [18]. The term data mining refers to the extraction of hidden knowledge, patterns, and relationships in an enormous amount of data [19].

The purpose of current research was to identify the breast cancer-related biomarkers. The results showed an increase in the capacity of epigenetic pattern discovery as a result of combining the two data mining techniques of meta-analysis and machine learning.

2. Material and methods

The current study was carried out based on the following steps: First, ChIP-Seq data regarding the effects of drugs on breast cancer were collected from EBI and NCBI databases. Then, the Human Genome Reference and Human Genotype Annotation sequence leads were retrieved from the Ensembl repository. Using FASTQC software, the quality control was checked in the Linux command line environment. It was found that a wide range of factors could lead to quality control problems. Subsequently, low quality sequencing reads were removed using Trimmomatic, Trimmate software. Quality control was also performed again through FASTQC software to confirm the suitability of the

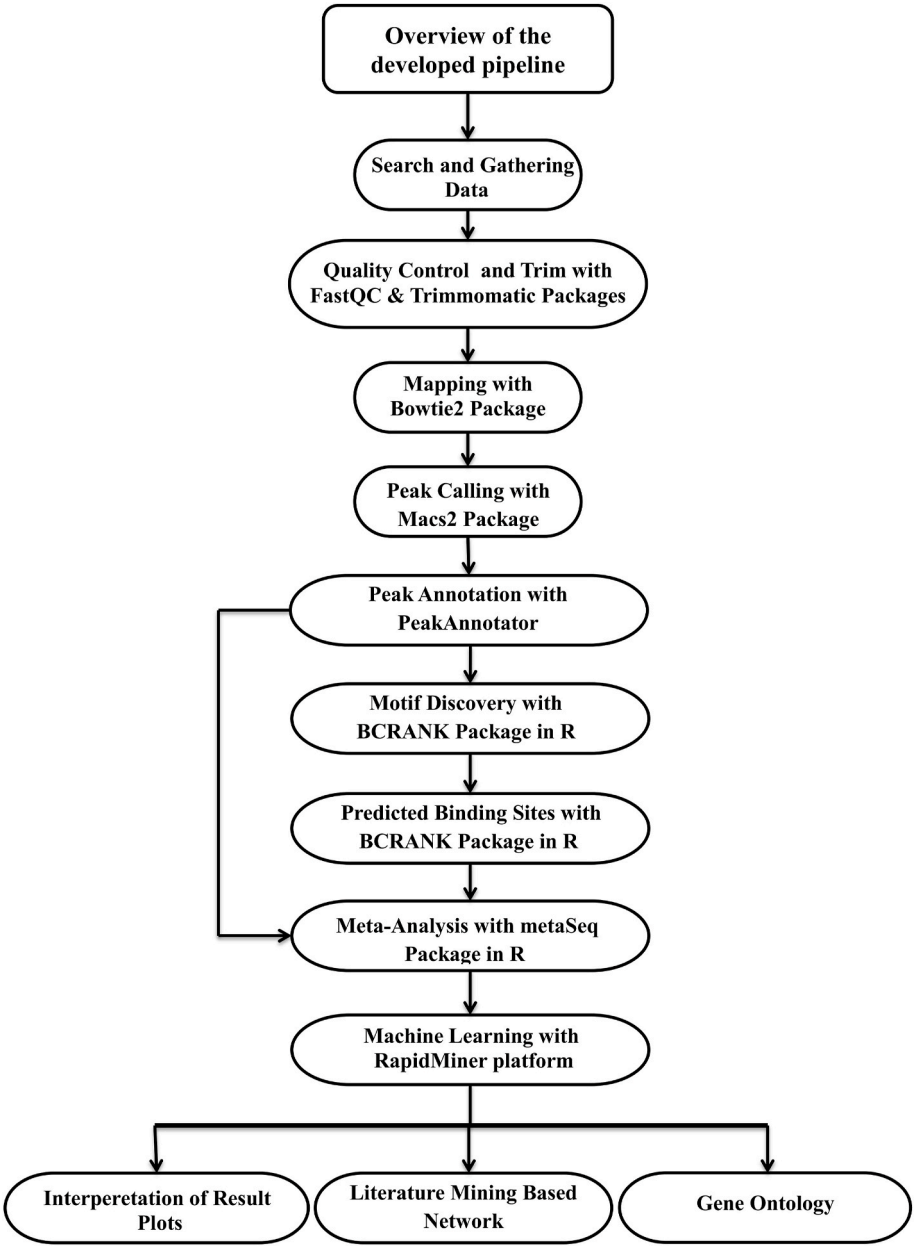

Fig. 1. Overview of the steps of the developed in pipeline.

performed steps. Afterwards, readings were mapped to the human reference genome using Bowtie2 software. Peak calling was applied to recognize those regions of the genome that were enriched by aligned reads in ChIP-Seq experiment. The peak calling was detected using MACS2 software. Then, the peaks achieved from the previous step through applying PeakAnnotator software were added to the Annotation file. The BCRANK package was also used in the R environment in order to predict the binding site consensus from the ranked DNA sequences and the motifs, and to identify downstream genes. Meta-analysis was conducted through metaSeq package in the R environment using the Fisher test on genes output from the peak call and binding site steps. Finally, data mining was carried out using RapidMiner software and nine weighting algorithms in order to determine the biomarkers and key genes (Fig. 1).

2.1. Data collection

ChIP-Seq data, which was related to the effects of drugs on breast cancer, was searched in NCBI and EBI databases. The raw data of experiments were retrieved.

2.2. Data quality control

Data quality control is a significant step in ChIP-Seq testing. The quality control of the collected samples was conducted using FASTQC software [20].

2.3. Data trimming

Based on the data quality control results, the low quality sequence reads were trimmed and corrected data were evaluated using Trimmomatic software to prepare the cleaned data for subsequent analysis [21].

2.4. Generating the referencing index and mapping readings with a reference genome

The human hg19 genome was used in current study (ftp://ftp.ensembl.org/pub/grch37/release-90/fasta/homo_sapiens/dna). To perform the mapping process, the index reference of the genome was first developed through Bowtie 2 software. Then, each sample was mapped using Bowtie 2 with the genome reference index [22].

2.5. Peak calling

At this step, the high density areas (enrichment) of mapped sequence reads (peaks) to the reference genome were identified using MACS2 software [23]. Also, the nearby genes to peaks were identified using PeakAnnotator software [24].

2.6. Motif discovery

To achieve a better understanding of peaks, the processes of finding motifs were carried out [25]. The identification of motifs, which could lead to the prediction of applied binding site, was carried out using BCRANK package in the R environment at this stage [26].

2.7. Predicting the binding site

Using the BCRANK package in R environment [26] and the output of motif discovery, the positions of the binding site were identified.

2.8. Meta-analysis

After performing ChIP-Seq analysis with the purpose of identifying the essential genes that were increased or decreased due to the treatments effects, the outputs of all studies were used for the meta-analysis procedure. The meta-analysis procedure was carried out on the binding site and peak calling outputs through applying metaSeq package in the R environment. The Fisher test was used on peak calling and binding site in the mentioned package [27]. Meta-analysis techniques are widely implemented in order to combine the results of numerous clinical or genomic studies and consequently, increase the statistical power in obtaining accurate conclusions [28]. The Fisher method could implemented for general meta-analysis because this is an effective approach of combining P values derived from independent studies [29,30]. In the randomized trials, especially in studies with small sample sizes, it would be better to implement the Fisher P-value [31] because the Fisher method is very sensitive to the smallest P-value [32].

2.9. Weighing algorithms applied in data mining

To perform the data mining process, the Peak calling output was analyzed through applying RapidMiner software [33] and nine by applying nine different weighing algorithms including the Information Gain, Gini Index, Gain Ratio, Relief, Rule, SVM, Uncertainty, Chi-Squared Statistic, and Deviation to determine the most significant vital genes [21,34–36].

Information Gain (IG) is an entropy-based feature evaluation method that is widely applied in the machine learning and decision tree construction processes. It is defined as the amount of information provided by attribute items to a text group, which is used in the attribute selection. Also, it is calculated through the value of the term that could be implemented for the information classification to measure the importance of related lexical items [3,21,37].

The Gini Index could identify pair patterns with the same entropy measurement. For each of the specific attributes, all states were considered in Pairs [21,38].

The Gain Ratio is implemented to overcome the IG algorithm problems because despite the poor performance, IG selects variables with different values [39].

Relief is considered as one of the most important families in the machine learning algorithms, which implements the nearest neighbors and different classes in order to select the same features or measure the interactions [40,41].

Rule is a data science process that derives rules from datasets or decision trees. Also, it is a part of unsupervised learning processes that identifies the hidden patterns of data in the form of easily recognizable rules [42].

Support vector machines (SVMs) are a collection of related managed learning methods that analyze data and identify patterns in the computational biology, which is applied for the classification and regression analysis [34,41].

Uncertainty can measure the significance of an attribute through evaluating the symmetrical uncertainty with reference to the class. Each attribute is compared to others according to the group in which it is located [41].

Chi-square is a feature selection algorithm that calculates the statistical value of chi-square for each attribute of the input data set to the class property. Chi-square is between each attribute and target variable, which selects the required number of attributes with the best χ^2 scores [43].

Standard Deviation is one of the scatter indicators, which shows the difference of the average data and average value. Low standard deviation indicates that data are close to the average value and have little scatter, while high standard deviation indicates that more data is spaced from the average [42].

The results of meta-analysis and data mining applied with the purpose of identifying the key genes and biomarkers were entered into gene ontology and gene networks. Gene ontology and gene network analysis were conducted through Pathway Studio 2017. Two types of the gene network including Common Target and Common Regulator were also depicted using the above-mentioned software. Common Regulator was

implemented to identify the upstream regulators that could regulate ≥ 2 selected entities. The purpose of applying Common Targets was identification of downstream targets that were set by at least two selected entities [44].

3. Results

Ten related studies were found (see S1 Table); however, only four of them were appropriate enough to be used in the meta-analysis process. Information of the selected studies is provided in Table 1. The Ensembl repository was used as the main reference of the human hg19 genome (Homo_sapiens.GRCh37. dna) and the human genome (http://ftp.ensembl.org/pub/grch37/release-90/fasta/homo_sapiens/dna/).

3.1. Gene ontology and gene regulatory network derived from the output genes of meta-analysis of the binding site's

Output genes derived from the meta-analysis of the binding sites were introduced into the gene ontology (see S3 Table). It was found from Common Regulator analysis that *JUN* transcription factor and *CCNG1*, *NOV*, and *EDN2* genes are biomarkers of neoplasm and breast cancer, neoplasm and cancer, breast cancer, and cancer and neoplasm. Breast cancer could have positive effects on *JUN*, *CCN1*, *NOV*, *USP9Y*, *SMC5*, and *PDE5A* genes, while cancer and neoplasm would have a positive effect on *PDE4D* (Fig. 2-a). It was found from Common Target network analysis that *PDE4D* and *PDE5A* had a negative effect on cancer regulation, while *PDE5A* had a positive effect on neoplasm regulation. *JUN* factor transcription would also play an unknown role in the regulation of cancer, neoplasm, and metastasis. Moreover, *CCVG1* had an unknown effect on cancer and metastasis (Fig. 2-b).

3.2. Gene ontology and network analysis of the output genes of peak calling meta-analysis

The meta-analysis output meta-genes were divided into the following groups: 1) genes with increased the number of peaks after treatment, and 2) genes with reduced peak number.

Meta-genes with increased peaks that were introduced into the gene ontology (see S4 Table). It was found from the Common Regulator analysis that *CCND1* was a biomarker of cancer and neoplasm, while *IER3* was a biomarker of neoplasm. Cancer and neoplasm would have a positive effect on *CCND1* and *RSF1* transcription factors. Also, neoplasm could have a negative effect on *MSH3*, *IER3*, and *MZF1* transcription factors. Cancer had a negative effect on *MSH3* and *IER3*; also, it had an unknown effect on *FRAS1* (Fig. 3-a). The Common Target network analysis showed the negative impact of *MSH3* on neoplasm, carcinogenesis, and metastasis. *MZF1* transcription factor had a negative effect on the neoplasm and apoptosis, a positive effect on metastasis. *CCND1* and *IER3* had unknown effects on regulating metastasis, neoplasm, carcinogenesis, cancer, and apoptosis. *RNF114* had a positive effect on the apoptosis, and an unknown effect on the cancer and neoplasm. *RSF1* transcription factor also had an unknown effect on the regulation of apoptosis, carcinogenesis, and metastasis. Finally, it was found that *AUTS2* had a positive effect on the metastasis (Fig. 3-b).

Meta-analysis output genes with reduced peaks were subjected to the gene ontology (see S5 Table). It was found from Common Regulator analysis that *AIFM1*, *CLU*, *RCHY1*, *GUN* genes, as well as *RFX1*transcription factor were biomarkers of cancer. Cancer had a positive effect on *FOXK1* transcription factor and *POLQ*, *CLU*, and *EFNA1* genes; also, it had an unknown effect on *RFX1* transcription factor and *GUN*, *RAPH1*, *CEP76*, *GSTZ1*, *TP53BP2*, *EIF1AX*, and *AIMP2* genes. Furthermore, *TP53BP2*, *AIFM1*, *CLU*, *DENND2D*, and *PZP* genes were recognized as the biomarkers of neoplasm. Neoplasm had a negative effect on the regulation of *TP53BP2* and *ALFM1*, and a positive effect on *POLQ*, *RCHY1*, *CLU*, *EFNA1*, and *MAT1A*. Moreover, it had an unknown effect on *NR2E1* transcription factor and *SH3GL1*, *EIF1AX*, and *GSTZ1* genes (Fig. 4-a).

Common Target network showed that *BARX2*, *FAM172A*, *DENND2D*, *RCHY1*, *AIMP2*, *MAT1A*, *CLU*, *AIFM1*, *SH3GL1*, *TP53BP2*, and *RAB7A* genes had negative effects, while *SH3GLI* and *NR2E1* had positive effects on neoplasm. Also, *GAN* and *FAM49B*, as well as *FOXK1* transcription factor had unknown effects on the neoplasm. It was found that *MAT1A*, *CLU*, *NR2E1*, *ARHGEF3*, *STRADB* genes and *FOXK1* transcription factor had negative effects and *FAM172A*, *AIMP2*, *RAB7A*, *TP53BP2*, *AIFM1*, *GSPT1* genes, as well as *RFX1* transcription factor had positive effects on the apoptosis. *DENND2D*, *RCHY1*, *CCDC88A*, *GSTA5*, and *PZP* genes had unknown effects on the apoptosis.

The meta-analysis-derived genes were combined with the output of the peak number of those genes in the peak calling stage. Genes were drawn in order to achieve a better result confirmation and heat map visualization [45].

The heat map of essential genes with increased number of peaks with drug treatment in the peak calling stage was represented in current study. The vertical axis showed the genes. In the horizontal axis of studies, the letters S, T, and C respectively represented the study, treatment, and control. It could be found from the above-mentioned heat map that the genes derived from the meta-analysis of treatment had more peak calls than the control factor. This can be seen in red color, which implies an increase in the number of peaks. Furthermore, the green color shows a reduction in the number of peaks (Fig. 5).

The heat map of essential genes with reduced number of peaks with drug treatment on the peak calling stage was also represented. It can be observed in the heat map that the genes derived from the meta-analysis of the control factor had more peak calls than treatment, which could be observed in red color (Fig. 6).

3.3. Gene ontology and network drawing of the output genes of peak calling data mining

Nine different attribute weighting algorithms (AWs) were implemented including Gain, Gini Index, Gain Ratio, Relief, Rule, SVM, Uncertainty, Chi Squares, and Deviation criteria to identify the important genes. It was expected that all weights would be between 0 and 1.0. Values that were closer to 1 indicated that a specific gene was an important attribute. Fourteen genes were detected by the majority of attribute weighting algorithms, which were with > 0.7 weights, as the most informative genes including *KIP*, *TCF12*, *ABCC5*, *HDAC11*, *IPP*, *HIST1H2AM*, *ZNF33B*, *PHF2*, *ELAVL3*, *TBC1D9B*, *TMEM217*, *CD34*,

Table 1
Summary of the ChIP-Seq studies employed for meta-analysis in this study.

Study Number	Accession Number	Title	Number of samples	Cell Line	Progesterone Receptor	Estrogen Receptor	Reference
1	EGEOD605	Drug specific epigenetic reprogramming leads to increased cellular invasion in ERα positive breast cancer via de novo cholesterol biosynthesis	3	MCF7	+	+	[26]
2	EGEOD54027	HoxC11 ChIP-seq of LY2 Breast Cancer Cell Line	3	LY2	+	+	[27]
3	EGEOD28987	SRC-1 targets ADAM22: an ER-independent mechanism of tumor progression in endocrine resistance	5	LY2	+	+	[28]
4	EGEOD26083	Genome-wide maps of Tamoxifen resistance MCF7 cell line	5	MCF7	+	+	[29]

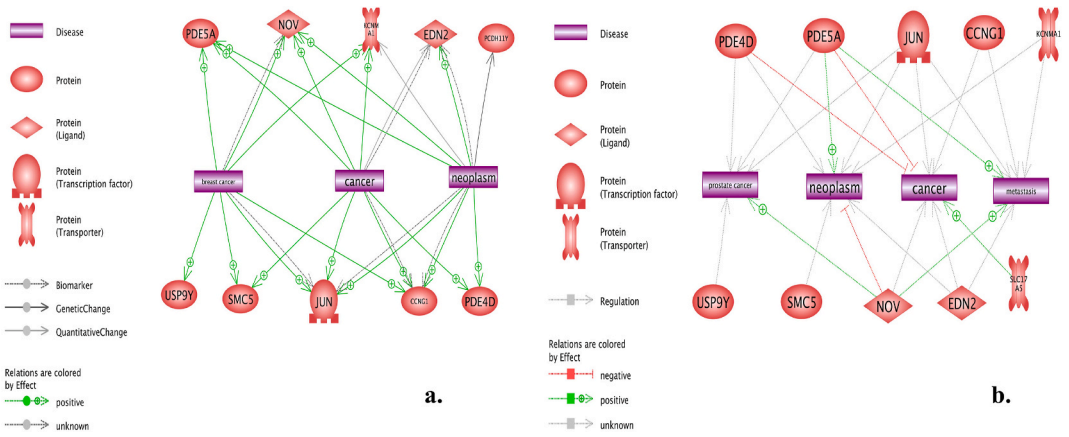

Fig. 2. The meta-analysis output meta-genes were divided into the following two groups: 1) genes with increased the number of peaks after treatment, and 2) genes with reduced peak number. a- Common Regulator network analysis of meta-genes. b- Common Target network meta-analysis output genes.

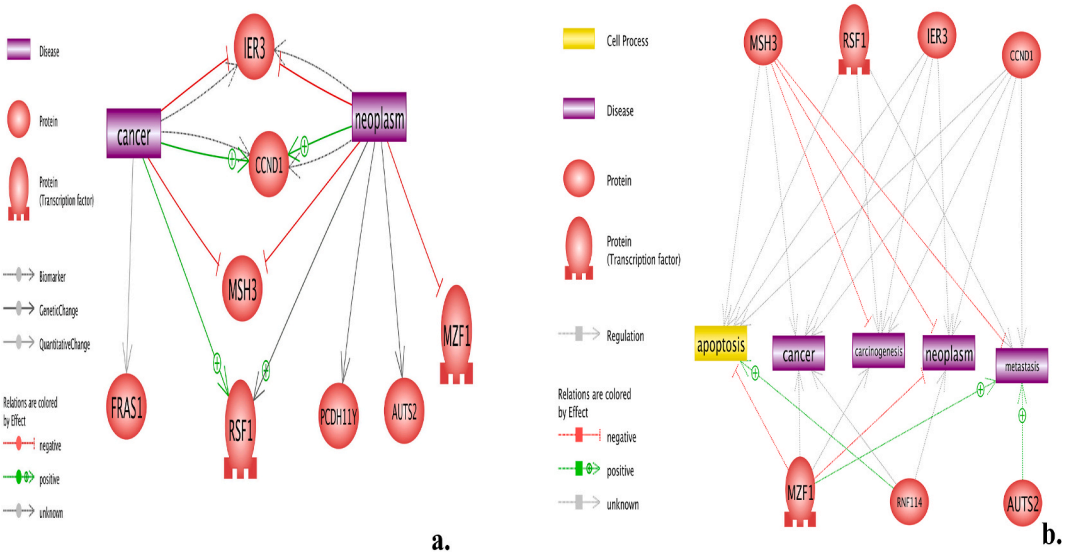

Fig. 3. a-Common Regulatory Network Based on the meta-analysis results and the number of peaks increased by drug treatment. b- Common Target Network Based on the meta-analysis results and the number of peaks increased by drug treatment.

ARHGEF26, and *CENPL* (Table 2). The key genes achieved from this step entered the gene ontology and network drawing.

The key genes were entered from data mining into the gene ontology (see S6 Table). The Common Regulator showed that *ZBTB7A* transcription factor was a biomarker of breast cancer. Breast cancer had a negative effect on *CD34* regulation, and a positive regulatory effect on *EIF3A* and *ZBTB7A* transcription factor. Breast cancer has an unknown effect on the transcription factors of *PLAGL1* and PHF2 (Fig. 7-a).

The Common Target Network showed that *PAPD5* gene and transcription factors of *TCF12, ZBTB7A*, and *PLAGL1* had a negative regulatory effect, while *PHF2* transcription factor had a positive regulatory effect on neoplasm. *CD34* gene and *TOP2B* factor transcription also had an unknown effect on neoplasm. Moreover, *PLAG1* and *ZBTB7A* transcription factors respectively had positive and negative regulatory effects on the apoptosis. *CD34* and *EIF3A* genes, as well as transcription factors of *TOP2B* and *PHF2* had unknown effects on the apoptosis (Fig. 7-b).

The data mining-derived key genes were combined with the output of the peak number of genes in the peak calling stage. Fig. 8 shows the data mining-derived key genes with the most changes in the number of peaks treatment compared to controls. Red color indicates an increase in the number of peaks, while green color shows a decrease in the number of peaks.

4. Discussion

The main purpose of the current study was to discover patterns of epigenome changes in the treatment and prognosis of breast cancer. ChIP-Seq data has a high potential in the prediction and prevention of breast cancer. ChIP-Seq data are important resources to identify the gene regulatory regions, pathways of genes involved in breast cancer, and people prone to cancer. Two important statistical tools including the meta-analysis and machine learning were implemented to identify the biomarkers and key genes from several independent studies. Discovering the proper biomarkers for early detection of patients prone to breast cancer and appropriate identification of high-risk patients are certified ways of the disease treatment and diagnosis [3]. Biomarkers are applied in order to identify the primary molecules and tumors for potential prognostic [46]. Current study aimed to identify the potential biomarkers for the breast cancer treatment. An appropriate biomarker has to be specific to the disease; also, it has to remain constant with unrelated disorders. Moreover, biomarkers have to be reliable and reproducible [21]. We were able to identify *NOV* gene and transcription factors of *JUN* and *ZBTB7A* as the biomarkers of breast cancer. It was

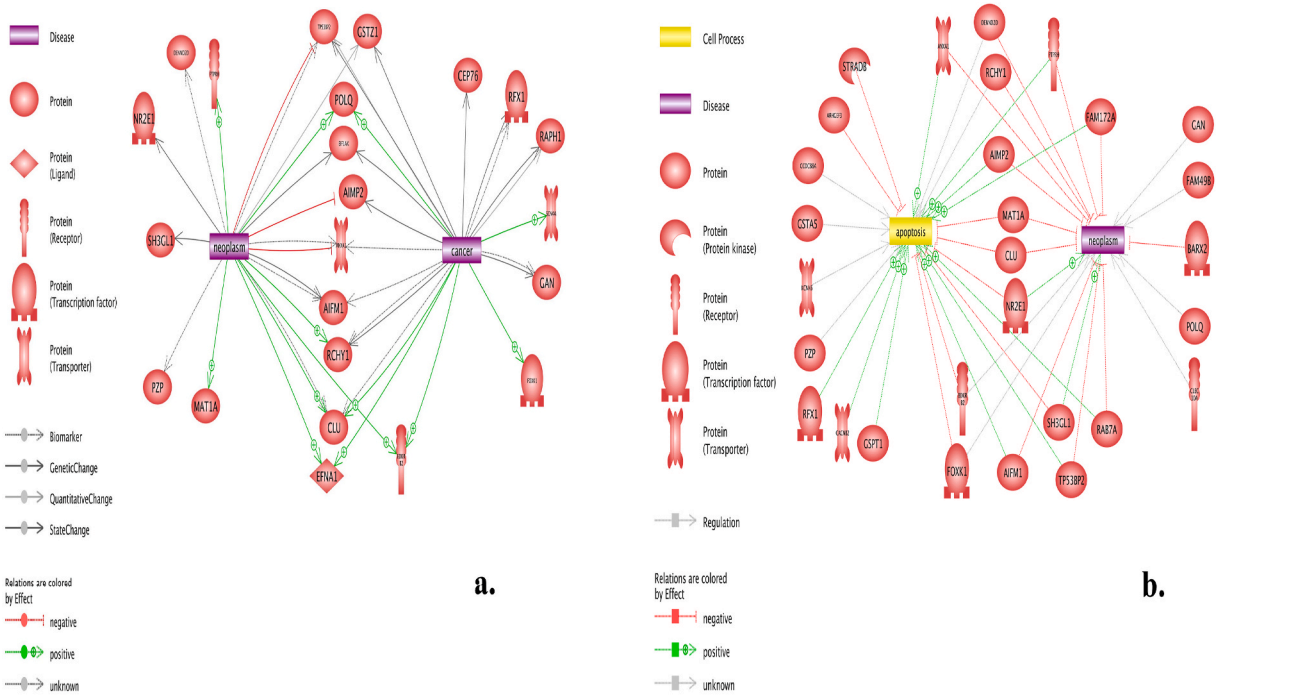

Fig. 4. a- Common Regulatory Network Based on the meta-analysis results, the number of peaks reduced by drug treatment. b- Common Target Network Based on the meta-analysis results, the number of peaks reduced by drug treatment.

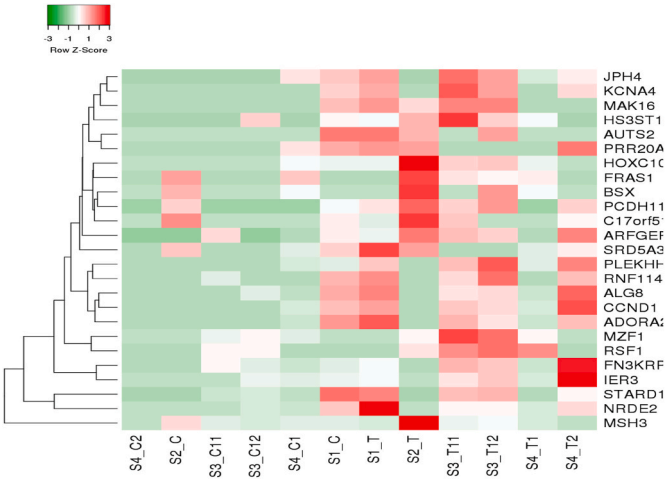

Fig. 5. Heat map of the essential genes that have increased the number of peaks with drug treatment on peak calling stage. Red means an increase in the number of peaks and green indicates a decrease in the number of peaks. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

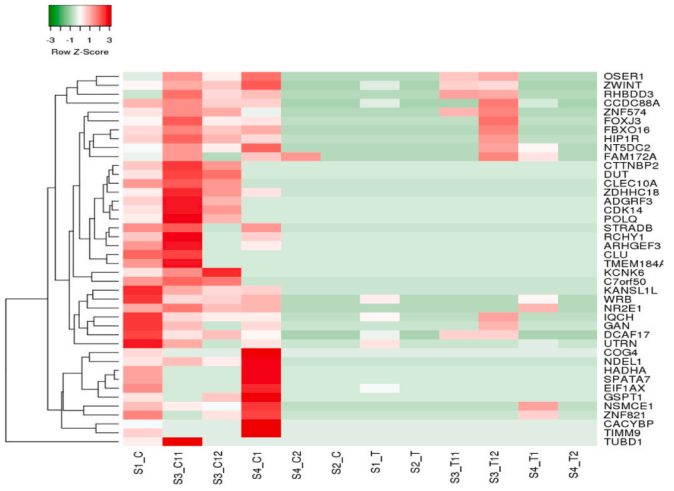

Fig. 6. Heat map of the essential genes that have been treated with medication to reduce the number of peaks on peak calling stage.

found that *NOV* (overexpressed nephroblastoma or *CCN3*) is a member of *CCN* family secreted from matricellular proteins. Moreover, *CCN3* gene played an important role in increasing the metastasis of breast cancer in bone and could be used as a biomarker for prostate cancer [47,48], and *c-Jun* was a protein encoded by *JUN* that plays an important role in carcinogenesis and cancer progression. Also, *c-Jun* overexpression reduced tamoxifen sensitivity in ER + breast cancer cells and could be used as a biomarker in breast cancer [49,50]. *ZBTB7A* transcription factor is involved in breast cancer, apoptosis, and neoplasm. The overexpression of *ZBTB7A* has been observed in numerous tumors including the lung cancer and breast cancer [51]. *ZBTB7A* could directly bind to ERα promoter in ER-positive breast tumors, and act as a tumor suppressor [52,53].

Meta-analysis at the binding sites showed (Fig. 2) that important genes such as *PDE5A*, *PDE4D*, *CCNG1*, *SMC5*, and *EDN2* had the greatest changes in the expression pattern of breast cancer. Catalano et al. (2019) reported that *PDE5A* overexpression was frequently observed in various human cancers such as breast cancer [9]. Results of current study showed that *PDE4D* had a negative effect on cancer. Another study suggested that *PDE4D* is an appropriate target for anti-cancer therapies, and *PDE4D* inhibition might be a means of overcoming tamoxifen resistance in ER-positive models of breast cancer [54]. *SMC5* gene had an increased expression in large intestinal cancers and neuroblastoma [55]. Also, results showed that *CCNG1* and *EDN2* were biomarkers of cancer and neoplasms, respectively. *CCNG1* is involved in the aberrant cell division and tumorigenesis, and its overexpression was also noted in breast and colon cancers [56]. *EDN2* can be served as a potentially

Table 2
Key genes selected by from 9 weighting algorithms (AWs).

Attribute	Weight SVM	Weight Relief	Weight Uncertainty	Weight Gini Index	Weight Chi Squared	Weight Deviation	Weight Rule	Weight Info Gain Ratio	Weight Info Gain
KIP	1	1	1	1	1	0.9	1	1	1
TCF12	1	1	1	1	1	0.9	1	1	1
ABCC5	1	1	1	1	1	0.9	1	1	1
HDAC11	0.7	1	0.7	0.9	0.9	1	0	0.7	0.8
IPP	1	1	1	1	1	0.9	1	1	1
HIST1H2AM	1	1	1	1	1	0.9	1	1	1
ZNF33B	0.7	1	0.7	0.9	0.9	1	0	0.7	0.8
PHF2	1	1	1	1	1	0.9	1	1	1
ELAVL3	1	1	1	1	1	0.9	1	1	1
TBC1D9B	0.7	1	0.7	0.9	0.9	1	0	0.7	0.8
TMEM217	0.7	1	0.7	0.9	0.9	1	0	0.7	0.8
CD34	0.7	1	0.7	0.9	0.9	1	0	0.7	0.8
ARHGEF26	0.7	1	0.7	0.9	0.9	1	0	0.7	0.8
CENPL	0.7	1	0.7	0.9	0.9	1	0	0.7	0.8

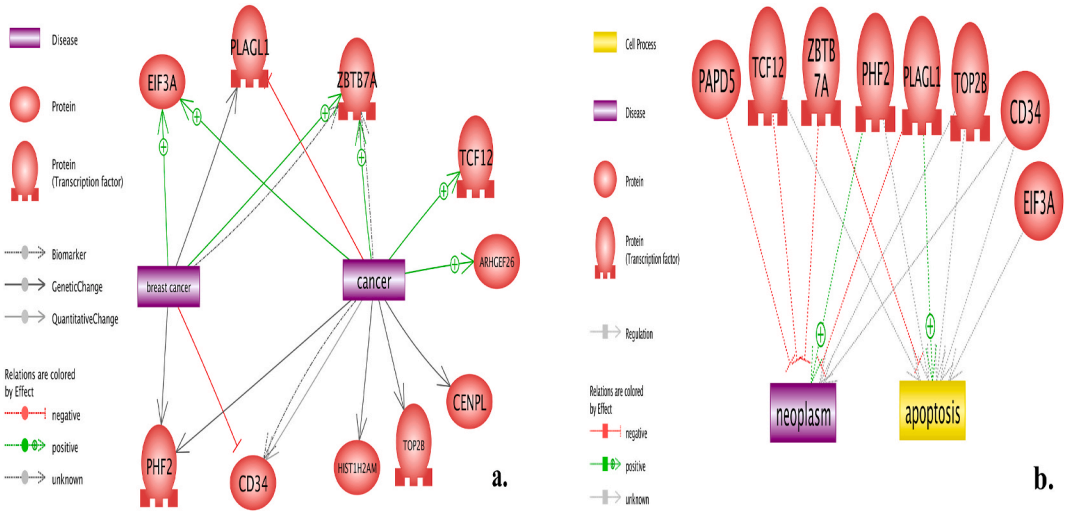

Fig. 7. a- Key genes derived from data mining were subjected to Common Regulatory Network analysis. b- Key genes derived from data mining in the Common Target Network.

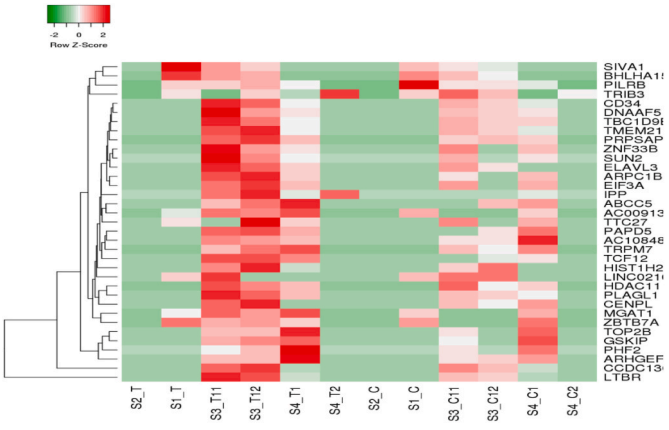

Fig. 8. Heat map of the key output genes of the data mining that has been affected by the drug causing changes in the number of peaks.

effective biomarker in the prognosis of breast cancer and provide a new perspective in order to achieve a better understanding of the molecular network in the breast cancer progression [57].

Results of the meta-analysis of the peak calling stage are presented in Figs. 3 and 4. *RFX1* can be used as a prognostic marker for cancer and breast cancer [58,59]. Results of current study showed that *FOXK1* had a negative regulatory effect on the apoptosis and an unknown effect on the cancer, while previous studies played a vital role in cancer [60,61]. *AIMP2* had an inhibitory effect on neoplasms, which could increase the tumor necrosis-induced signaling apoptosis. *AIMP2* had anti-proliferative activities through specific mechanisms of action and can act as a potent tumor suppressor against various cancers [62]. Results also showed that *NR2E1* played a role in the apoptosis inhibition, which had a positive effect on neoplasms. It was found by another investigation that *NR2E1* could be applied to predict the risk metastasis of breast cancer [63]. Various studies indicated that *CLU* is involved in cancer, inhibiting cell death pathways, and modulating survival signals to enhance the cell growth [64]. Another report also indicated that *CLU* would be up- regulated in breast cancer [65]. *CCN1* could be involved in many cellular biological functions such as mediating cell adhesion, migration, proliferation, apoptosis, and angiogenesis. Moreover, it is commonly expressed in breast cancer [66]. IER3 was involved in the apoptosis and cell cycle arrest [67]. Results of current study showed that cancer and neoplasm had a positive effect on *RSF1* regulation. Another study indicated that interfering with *RSF1* gene expression effectively prevented the proliferation of MCF-7 and SKBR-3 cells and consequently, increased apoptosis. Also, interfering with *RSF1* expression can be served as a new therapeutic target for the breast cancer treatment [68]. Findings also showed that *MZF1* was involved in the development of aggressive breast cancer and metastasis [69].

The machine learning was applied in current study to prioritize the

meta-gens and detect the key differentiating genes in response to breast cancer. The top meta-gens including *KIP*, *TCF12*, *ABCC5*, *HDAC11*, *IPP*, *HIST1H2AM*, *ZNF33B*, *PHF2*, *ELAVL3*, *TBC1D9B*, *TMEM217*, *CD34*, *ARHGEF26*, and *CENPL* are provided in Table 2. The Inhibitory Protein Kinase (*KIP*) family is a mammalian cyclin kinase (CDK) inhibitor involved in the regulation of transcription, apoptosis, and cytoskeleton. CDK abnormal expression would lead to the cancer [70]. *TCF12* may act as a regulator in breast cancer tumors; moreover, it was reported that it could be closely associated with tumor metastasis and invasion [71,72]. *ABCC5* is an ATP-dependent transmitter, overexpressed in skeletal metastasis of breast cancer compared to primary breast tumors [73]. In another study, it was found that *ABCC5* was functionally associated with bone metastases formation in breast cancer [74]. The role of histone acetylation in chromatin organization is completely established and it was found that high levels of histone deacetylase 11 (*HDAC11*) could mediate the breast cancer cell metastasis [75]. In another study, the inhibition of *HDAC11* led to *p53*-dependent cell apoptosis in hepatocellular carcinoma cells [76]. The role of *PHF2* in breast cancer is remained unclear [77]. *PHF2* can act as a tumor suppressor through *p53* epigenetic regulation [78]. Results indicated that *CD34* is a biomarker of cancer, and another study stated that it was a useful angiogenesis marker that could help to identify more aggressive breast tumors [79]. Studies confirmed that *EIF3A* is a proto-oncogene, and many other investigations also reported that it is related to cancer, metastasis, prognosis, therapeutic response [80], and breast cancer [81] (Fig. 7). *PLAGL1* encodes a zinc-finger nuclear transcription factor that can cause apoptosis, and cell cycle arrest [82].

5. Conclusion

Results of current study demonstrated that a combination of machine learning and meta-analysis in the analysis of multiple experiments simultaneously is useful in understanding and identifying key genes in breast cancer progression. The achieved results can be employed in order to both identify the appropriate biomarkers and to predict or find more specific drugs for the breast cancer treatment.

Declaration of competing interest

The authors whose names are listed immediately below certify that they have NO affiliations with or involvement in anyorganization or entity with any financial interest (such as honoraria; educational grants; participation in speakers' bureaus; membership, employment, consultancies, stock ownership, or other equity interest; and expert testimony or patent-licensing arrangements), or non-financial interest (such as personal or professional relationships, affiliations, knowledge or beliefs) in the subject matter or materials discussed in this manuscript.

Acknowledgement

The authors gratefully acknowledge the financial support for this work that was provided by Shiraz University, Shiraz, Iran.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.imu.2021.100629.

References

[1] He Z, Chen Z, Tan M, Elingarami S, Liu Y, Li T, Li W. A review on methods for diagnosis of breast cancer cells and tissues. Cell Prolif 2020;53(7):e12822.
[2] Xu Y, Zhang M, Chen R, Xia X. Genetic alterations of early-stage breast cancers by next-generation sequencing (NGS). Ann Oncol 2018;29:viii67.
[3] Gentile M, Centonza A, Lovero D, Palmirotta R, Porta C, Silvestris F, D'Oronzo S. Application of "omics" sciences to the prediction of bone metastases from breast cancer: state of the art. Journal of Bone Oncology 2020:100337.
[4] Siegel RL, Miller KD, Jemal A. Cancer statistics, 2016. CA A Cancer J Clin 2016;66(1):7–30.
[5] Velaga R, Sugimoto M. Future paradigm of breast cancer resistance and treatment. Resistance to targeted therapies in breast cancer. Cham: Springer; 2017. p. 155–78.
[6] Liedtke C, Mazouni C, Hess KR, Andre F, Tordai A, Mejia JA, et al. Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer. J Clin Oncol 2008;26:1275–81. https://doi.org/10.1200/JCO.2007.14.4147.
[7] von Minckwitz G, Untch M, Ju Blohmer, Costa SD, Eidtmann H, Fasching PA, et al. Definition and impact of pathologic complete response on prognosis after neoadjuvant chemotherapy in various intrinsic breast cancer subtypes. J Clin Oncol 2012;30:1796–804. https://doi.org/10.1200/JCO.2011.38.8595.
[8] Choudhury PP, Brook MN, Hurson AN, Lee A, Mulder CV, Coulson P, Garcia-Closas M. Comparative validation of the BOADICEA and Tyrer-Cuzick breast cancer risk models incorporating classical risk factors and polygenic risk in a population-based prospective cohort of women of European ancestry. Breast Canc Res 2021;23(1):1–5.
[9] Catalano S, Panza S, Augimeri G, Giordano C, Malivindi R, Gelsomino L, Barone I. Phosphodiesterase 5 (PDE5) is highly expressed in cancer-associated fibroblasts and enhances breast tumor progression. Cancers 2019;11(11):1740.
[10] Francastel C, Schübeler D, Martin DI, Groudine M. Nuclear compartmentalization and gene activity. Nat Rev Mol Cell Biol 2000;1(2):137–43.
[11] Gagliano T, Brancolini C. Epigenetic mechanisms beyond tumour–stroma crosstalk. Cancers 2021;13(4):914.
[12] August Pasculli B, Barbano R, Parrella P. Epigenetics of breast cancer: biology and clinical implication in the era of precision medicine. Seminars in cancer biology, vol. 51. : Academic Press; 2018. p. 22–35.
[13] Holmes KA, Brown GD, Carroll JS. Chromatin immunoprecipitation-sequencing (ChIP-seq) for mapping of estrogen receptor-chromatin interactions in breast cancer. Estrogen receptors. New York, NY: Humana Press; 2016. p. 79–98.
[14] Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, Zhang J. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol 2013;9(11):e1003326.
[15] Glass GV. Primary, secondary, and meta-analysis of research 1. Educ Res 1976;5(10):3–8.
[16] Leistico A-MR, Salekin RT, DeCoster J, Rogers R. A large-scale meta-analysis relating the Hare measures of psychopathy to antisocial conduct. Law Hum Behav 2008;32(1):28–45.
[17] Strube MJ, Hartmann DP. Meta-analysis: techniques, applications, and functions. J Consult Clin Psychol 1983;51(1):14.
[18] Ebrahimi M, Mohammadi-Dehcheshmeh M, Ebrahimie E, Petrovski KR. Comprehensive analysis of machine learning models for prediction of sub-clinical mastitis: deep Learning and Gradient-Boosted Trees outperform other models. Comput Biol Med 2019;114:103456.
[19] Fayyad U, Piatetsky-Shapiro G, Smyth P. From data mining to knowledge discovery in databases. AI Mag 1996;17(3):37.
[20] Bioinformatics B. FastQC: a quality control tool for high throughput sequence data. UK: Cambridge; 2011 [Babraham Institute].
[21] Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 2014;30(15):2114–20. , http://www.usadellab.org/cms/index.php?page=trimmomatic.
[22] Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods 2012;9(4):357–9.
[23] Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Li W. Model-based analysis of ChIP-seq (MACS). Genome Biol 2008;9(9):R137.
[24] Salmon-Divon M, Dvinge H, Tammoja K, Bertone P. PeakAnalyzer: genome-wide annotation of chromatin binding and modification loci. BMC Bioinf 2010;11(1):1–12.
[25] Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J. RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. Nucleic Acids Res 2011;40(4). e31-e31.
[26] Ameur A, Ameur MA, Biostrings I, biocViews MotifDiscovery G. Package 'BCRANK'. 2010 ;.
[27] Tsuyuzaki K, Nikaido I. Meta-analysis of RNA-Seq count data in multiple studies. 2013.
[28] Sharifi S, Pakdel A, Ebrahimi M, Reecy JM, Fazeli Farsani S, Ebrahimie E. Integration of machine learning and meta-analysis identifies the transcriptomic bio-signature of mastitis disease in cattle. PloS One 2018;13(2):e0191227.
[29] Fisher RA. Statistical methods for research workers. fourth ed. London: Oliver and Boyd; 1932.
[30] Huo Z, Tang S, Park Y, Tseng G. P-value evaluation, variability index and biomarker categorization for adaptively weighted Fisher's meta-analysis method in omics applications. Bioinformatics 2020;36(2):524–32.
[31] Heard NA, Rubin-Delanchy P. Choosing between methods of combining-values. Biometrika 2018;105(1):239–46.
[32] Bind MA, Rubin DB. When possible, report a Fisher-exact P value and display its underlying null randomization distribution. Proc Natl Acad Sci Unit States Am 2020;117(32):19151–8.
[33] Mierswa I, Klinkenberg R. RapidMiner Studio (9.2. 2019. Data science, machine learning, predictive analytics.
[34] Venkatesh S, Workman JL. Histone exchange, chromatin structure and the regulation of transcription. Nat Rev Mol Cell Biol 2015;16(3):178–89.
[35] Bakhtiarizadeh MR, Moradi-Shahrbabak M, Ebrahimi M, Ebrahimie E. Neural network and SVM classifiers accurately predict lipid binding proteins, irrespective of sequence homology. J Theor Biol 2014;356:213–22. pmid:24819464.

[36] Ebrahimi M, Ebrahimie E, Ebrahimi M. Searching for patterns of thermostability in proteins and defining the main features contributing to enzyme thermostability through screening, clustering, and decision tree algorithms. EXCLI J 2009;8:218–33.
[37] Lei S. March). A feature selection method based on information gain and genetic algorithm. International conference on computer science and electronics engineering, vol. 2. IEEE; 2012. p. 355–8. 2012.
[38] Liu Y, Gastwirth JL. On the capacity of the Gini index to represent income distributions. METRON; 2020. p. 1–9.
[39] Kose U, Alzubi J. Deep learning for cancer diagnosis. : Springer; 2020.
[40] Le TT, Urbanowicz RJ, Moore JH, McKinney BA. Statistical inference Relief (STIR) feature selection. Bioinformatics 2019;35(8):1358–65.
[41] Ebrahimie E, Ebrahimi M, Sarvestani NR, Ebrahimi M. Protein attributes contribute to halo-stability, bioinformatics approach. Saline Syst 2011;7(1):1–14.
[42] Kotu V, Deshpande B. Data science: concepts and practice. Morgan Kaufmann. 2018.
[43] Ul Haq A, Li J, Memon MH, Khan J, Ud Din S. A novel integrated diagnosis method for breast cancer detection. J Intell Fuzzy Syst 2020;38(2):2383–98.
[44] Nikitin A, Egorov S, Daraselia N, Mazo I. Pathway studio—the analysis and navigation of molecular networks. Bioinformatics 2003;19(16):2155–7.
[45] Babicki S, Arndt D, Marcu A, Liang Y, Grant JR, Maciejewski A, Wishart DS. Heatmapper: web-enabled heat mapping for all. Nucleic Acids Res 2016;44(W1):W147–53.
[46] D'Oronzo S, Brown J, Coleman R. The value of biomarkers in bone metastasis. Eur J Canc Care 2017;26(6):e12725.
[47] Dankner M, Ouellet V, Communal L, Schmitt E, Perkins D, Annis MG, Network CPCB. CCN3/Nephroblastoma overexpressed is a functional mediator of prostate cancer bone metastasis that is associated with poor patient prognosis. Am J Pathol 2019;189(7):1451–61.
[48] MacLeod K, Laird BJA, Carragher NO, Hoskin P, Fallon MT, Sande TA. Predicting response to radiotherapy in cancer-induced bone pain: cytokines as a potential biomarker? Clin Oncol 2020;32(10):e203–8.
[49] He H, Sinha I, Fan R, Haldosen LA, Yan F, Zhao C, Dahlman-Wright K. c-Jun/AP-1 overexpression reprograms ERα signaling related to tamoxifen response in ERα-positive breast cancer. Oncogene 2018;37(19):2586–600.
[50] Canzoneri R, Naipauer J, Stedile M, Peña AR, Lacunza E, Gandini NA, Abba MC. Identification of an AP1-ZFP36 regulatory network associated with breast cancer prognosis. J Mammary Gland Biol Neoplasia 2020;25(2):163–72.
[51] Mao A, Chen M, Qin Q, Liang Z, Jiang W, Yang W, Wei C. ZBTB7A promotes migration, invasion and metastasis of human breast cancer cells through NF-κB-induced epithelial–mesenchymal transition in vitro and in vivo. J Biochem 2019;166(6):485–93.
[52] Wang L, Zhang MX, Zhang MF, Tu ZW. ZBTB7A functioned as an oncogene in colorectal cancer. BMC Gastroenterol 2020;20(1):1–7.
[53] Geng R, Zheng Y, Li Q, Li R, Guo X. ZBTB7A, a potential biomarker for prognosis and immune infiltrates, inhibits progression of endometrial cancer based on bioinformatics analysis and experiments. Canc Cell Int 2020;20(1):1–15.
[54] Lai SH, Zervoudakis G, Chou J, Gurney ME, Quesnelle KM. PDE4 subtypes in cancer. Oncogene 2020;39(19):3791–802.
[55] Pryzhkova MV, Jordan PW. Conditional mutation of Smc5 in mouse embryonic stem cells perturbs condensin localization and mitotic progression. J Cell Sci 2016;129(8):1619–34.
[56] Ravicz J, Szeto C, Reddy S, Chawla S, Morse M, Gordon E. Enhanced expression of human cyclin G1 (CCNG1) in tumors, a novel biomarker in development for cancer therapy/gene therapy. 2021 . https://doi.org/10.20944/preprints202103.0213.v1.
[57] Yang H, Mao W, Rodriguez-Aguayo C, Mangala LS, Bartholomeusz G, Iles LR, Bast RC. Paclitaxel sensitivity of ovarian cancer can be enhanced by knocking down pairs of kinases that regulate MAP4 phosphorylation and microtubule stability. Clin Canc Res 2018;24(20):5072–84.
[58] Shibata M, Kanda M, Shimizu D, Tanaka H, Umeda S, Hayashi M, Kikumori T. Expression of regulatory factor X1 can predict the prognosis of breast cancer. Oncology letters 2017;13(6):4334–40.
[59] Ma T, Zhou X, Wei H, Yan S, Hui Y, Liu Y, Mu XX. Long non-coding RNA SNHG17 upregulates RFX1 by sponging miR-3180-3p and promotes cellular function in hepatocellular carcinoma. Front Genet 2020:11.
[60] Liu Y, Ding W, Ge H, Ponnusamy M, Wang Q, Hao X, Wang J. FOXK transcription factors: regulation and critical role in cancer. Canc Lett 2019;458:1–12.
[61] Zheng S, Yang L, Zou Y, Liang JY, Liu P, Gao G, Xie X. Long non-coding RNA HUMT hypomethylation promotes lymphangiogenesis and metastasis via activating FOXK1 transcription in triple-negative breast cancer. J Hematol Oncol 2020;13(1):1–15.
[62] Ku J, Kim R, Kim D, Kim D, Song S, Lee K, Koh Y. Single-cell analysis of AIMP2 splice variants informs on drug sensitivity and prognosis in hematologic cancer. Communications biology 2020;3(1):1–13.
[63] Park SB, Hwang KT, Chung CK, Roy D, Yoo C. Causal Bayesian gene networks associated with bone, brain and lung metastasis of breast cancer. Clin Exp Metastasis 2020;37(6):657–74.
[64] Liu Y, Zhou Y, Ma X, Chen L. Inhibition lysosomal degradation of clusterin by protein kinase D3 promotes triple-negative breast cancer tumor growth. Advanced Science 2021;8(4):2003205.
[65] Merlotti A, Malizia AL, Michea P, Bonte PE, Goudot C, Carregal MS, Sabatte J. Aberrant fucosylation enables breast cancer clusterin to interact with dendritic cell-specific ICAM-grabbing non-integrin (DC-SIGN). OncoImmunology 2019;8(9):e1629257.
[66] Yu S, Yan C, Wu W, He S, Liu M, Liu J, Jia L. RU486 metabolite inhibits CCN1/Cyr61 secretion by MDA-MB-231-endothelial adhesion. Front Pharmacol 2019;10:1296.
[67] Böckers M, Paul NW, Efferth T. Butyl octyl phthalate interacts with estrogen receptor α in MCF-7 breast cancer cells to promote cancer development. World Academy of Sciences Journal 2021;3(2). 1-1.
[68] Liu Y, Gai J, Fu L, Zhang X, Wang E, Li Q. Effects of RSF-1 on proliferation and apoptosis of breast cancer cells. Oncology letters 2018;16(4):4279–84.
[69] Brix DM, Bundgaard Clemmensen KK, Kallunki T. Zinc finger transcription factor MZF1—a specific regulator of cancer invasion. Cells 2020;9(1):223.
[70] Sanaei M, Kavoosi F. Effect of 5-aza-2'-deoxycytidine in comparison to valproic acid and trichostatin A on histone deacetylase 1, DNA methyltransferase 1, and CIP/KIP family (p21, p27, and p57) genes expression, cell growth inhibition, and apoptosis induction in colon cancer SW480 cell line. Adv Biomed Res 2019;8.
[71] Yang J, Zhang L, Jiang Z, Ge C, Zhao F, Jiang J, Li J. TCF12 promotes the tumorigenesis and metastasis of hepatocellular carcinoma via upregulation of CXCR4 expression. Theranostics 2019;9(20):5810.
[72] Gao S, Bian T, Zhang Y, Su M, Liu Y. TCF12 overexpression as a poor prognostic factor in ovarian cancer. Pathol Res Pract 2019;215(9):152527.
[73] Bai F, Yin Y, Chen T, Chen J, Ge M, Lu Y, Liu Y. Development of liposomal pemetrexed for enhanced therapy against multidrug resistance mediated by ABCC5 in breast cancer. Int J Nanomed 2018;13:1327.
[74] Mourskaia AA, Amir E, Dong Z, Tiedemann K, Cory S, Omeroglu A, Siegel PM. ABCC5 supports osteoclast formation and promotes breast cancer metastasis to bone. Breast Canc Res 2012;14(6):1–60.
[75] Bora-Singhal N, Mohankumar D, Saha B, Colin CM, Lee JY, Martin MW, Chellappan S. Novel HDAC11 inhibitors suppress lung adenocarcinoma stem cell self-renewal and overcome drug resistance by suppressing Sox2. Sci Rep 2020;10(1):1–20.
[76] Liu SS, Wu F, Jin YM, Chang WQ, Xu TM. HDAC11: a rising star in epigenetics. Biomed Pharmacother 2020;131:110607.
[77] Zhang L, Hui TL, Wei YX, Cao ZM, Feng F, Ren GS, Li F. The expression and biological function of the PHF2 gene in breast cancer. RSC Adv 2018;8(69):39520–8.
[78] Lee C, Kim B, Song B, Moon KC. Implication of PHF2 expression in clear cell renal cell carcinoma. Journal of pathology and translational medicine 2017;51(4):359.
[79] Hosseini S, Behjati F, Rahimi M, Taheri N, Khorshid HK, Moghaddam FA, Keyhani E. Relationship between PIK3CA amplification and P110α and CD34 tissue expression as angiogenesis markers in iranian women with sporadic breast cancer. Iranian journal of pathology 2018;13(4):447.
[80] Yin JY, Zhang JT, Zhang W, Zhou HH, Liu ZQ. eIF3a: a new anticancer drug target in the eIF family. Canc Lett 2018;412:81–7.
[81] Wang S-q, Liu Y, Yao M-y, Jin J. Eukaryotic translation initiation factor 3a (eIF3a) promotes cell proliferation and motility in pancreatic cancer. J Kor Med Sci 2016;31(10):1586–94.
[82] Kowalczyk AE, Krazinski BE, Godlewski J, Kiewisz J, Kwiatkowski P, Sliwinska-Jewsiewicka A, Kmiec Z. Altered expression of the PLAGL1 (ZAC1/LOT1) gene in colorectal cancer: correlations to the clinicopathological parameters. Int J Oncol 2015;47(3):951–62.