[Frontiers In Bioscience, Landmark, 22, 1697-1712, June 1, 2017]

Pathway-based classification of breast cancer subtypes

Alex Graudenzi[1,2], Claudia Cava[1], Gloria Bertoli[1], Bastian Fromm[3], Kjersti Flatmark[3,4,5], Giancarlo Mauri[2,6], Isabella Castiglioni[1]

[1]Institute of Molecular Bioimaging and Physiology of the Italian National Research Council (IBFM-CNR), Milan, Italy, [2]Department of Informatics, Systems and Communication, University of Milan-Bicocca, Milan, Italy, [3]Department of Tumor Biology, Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway, [4]Department of Gastroenterological Surgery, Norwegian Radium Hospital, Oslo University Hospital, Oslo, Norway, [5]Institute of Clinical Medicine, University of Oslo, Oslo, Norway, [6]SYSBIO Centre of Systems Biology (SYSBIO), 20126 Milan, Italy

TABLE OF CONTENTS

1. ABSTRACT

Cancer heterogeneity represents a major hurdle in the development of effective theranostic strategies, as it prevents to devise unique and maximally efficient diagnostic, prognostic and therapeutic procedures even for patients affected by the same tumor type. Computational techniques can nowadays leverage the huge and ever increasing amount of (epi)genomic data to tackle this problem, therefore providing new and valuable instruments for decision support to biologists and pathologists, in the broad sphere of precision medicine. In this context, we here introduce a novel cancer subtype classifier from gene expression data and we apply it to two different Breast Cancer datasets, from TCGA and GEO repositories. The classifier is based on Support Vector Machines and relies on the information about the relevant pathways involved in breast cancer development to reduce the huge variable space. Among the main results, we show that the classifier accuracy is preserved at excellent values even when the variable space is reduced by a 20-fold, hence providing a precious tool for cancer patient profiling even in case of limited experimental resources.

2. INTRODUCTION

Cancer is a complex multistep disease, characterized by high levels of *heterogeneity* at several scales, which, so far, have prevented a thorough understanding of the molecular interplay ruling its emergence and development. Not only different types of cancer display very different (epi)genomic mutational profiles, but also different patients with the same cancer (*sub*)type usually display few common alterations (i.e., *inter-tumor heterogeneity*) and, furthermore, single tumors are often characterized by the coexistence of various cancer clones with different evolutionary histories (i.e., *intra-tumor heterogeneity*) (1).

Cancer heterogeneity represents a key problem from the *theranostic* perspective, as it is not possible to devise a unique and maximally efficient strategy to tackle the disease of different cancer patients (2). Moreover, in cancer tissues where genetic heterogeneity exists, it is possible that the different subpopulations present will have different sensitivity to therapy and show differential responses to treatment. This process could in principle lead to the development of a clonal selection of a group of tumor cells which are resistant to therapeutic treatment, increasing the risk of progression or spreading of the tumor (3).

Accordingly, one of the major challenges of current cancer biology is the development of *personalized* diagnostic and therapeutic strategies, within the broad field of *precision medicine* (4).

In the last decades, the ever-increasing availability of (*big*) *data* regarding the genomic (and epigenetic) profiles of cancer patients has provided a new source of essential information, which however needs efficient theoretical frameworks, instruments and computational tools in order to be exploited.

On one hand, several algorithmic approaches attempt to exploit the massive amount of *Next Generation Sequencing* (NGS) data available in public projects such as e.g. The Cancer Genome Atlas (TCGA) (5), in order to reconstruct cancer progression models that can deliver new experimental hypotheses on the evolutionary trajectories of the various tumor types, which in turn might help in defining new theranostic strategies (see, e.g., (6-9)).

On the other hand, *gene expression* experiments, such as those based on *DNA microarrays*, which detect the simultaneous expression of thousand genes in different experimental conditions, have been widely used to link the expression profiles to cancer phenotypes (10,11), and can now be exploited to face the problem from a complementary perspective. *Machine learning techniques* can, in fact, make use of microarray data to perform *classification tasks*, allowing the detection of key *biomarkers* for diagnosis (12), prognosis (13, 14) and response to treatment of various diseases (15).

Nonetheless, despite several successful applications, the classification problem on microarray data still suffers of the "*many variable – few samples*" problems, as the number of variables (i.e., *genes*) is usually much larger than the number of observations (i.e., *tumor samples*). For instance, in *Breast Cancer* (BC) different gene signatures have been identified, especially focused on grade classification (e.g., (16-22)), yet with poor reproducibility and overlap.

In this regard, *feature selection* techniques are essential in order to achieve a more reliable performance of inference and classification. To this end, some attempts have been recently proposed by combining different features, such as, e.g., gene expression profiles, copy number variations and epigenetic profiles (18). The key idea is to concentrate on those (hopefully few) genes that are indeed relevant in characterizing the molecular progression of the disease, by referring to the complex interplay that rules the various molecular activities, often summarized under the concept of (molecular) *pathways,* which could also be used to combine previously proposed microRNA signatures (23) and eventually might lead to new actionable targets for clinical applications.

The use of explicit knowledge about pathways in feature reduction and disease classification is also fundamental to contrast the idiosyncrasies of single-gene classifier, especially with respect to inter- and intra-tumor heterogeneity and biological and experimental noise (24). Moreover, the interpretation of high-throughput genomic data indeed needs the identification of signaling and metabolic pathways of specific phenotypes. Hence, pathway-based classificatory have shown to be more reproducible and with similar or better performances with respect to classifiers based on independent genes (25-27).

In this context, we here propose a novel classification framework based on *Support Vector Machines* (SVMs) and with a feature selection strategy based on the concept of *pathway activity* (28, 29)[1], and we apply it to the most recent classification of BC subtypes (31).

In particular, we identified and analyzed a list of *enriched pathways* and, accordingly, of differentially expressed genes (DEGs) in four different subtypes, by using a recent Breast Cancer (BC) dataset curated by The Cancer Genome Atlas (TCGA) (32) (see below for details), and we used this information to perform the feature selection in the classifier implementation. The classifier was then trained and tested on a different gene expression BC dataset, from Gene Expression Omnibus (GEO) database (33), with the goal of efficiently classify the samples in the four BC subtypes. We specifically investigated the variation of accuracy and other standard performance measures with respect to distinct parameter settings. Among the most important results, we show that a 20-fold reduction of the variable space does not affect the performance of the classifier in a significant way.

To summarize, we here introduce a new valuable tool for the classification of BC subtypes, which is efficient even when limited experimental resources do not allow collecting information on the

[1] Notice that the complex interaction among pathways ruling cancer development is sometimes referred to as *pathway cloud* (30).

whole gene pool, and which provides significant indications on the key biological features that should be considered in the classification task (e.g., the *relevant* genes to employ). Accordingly, the results might allow identifying groups of genes, pathways and *biomarkers* to opportunely modulate in the development of personalized therapies.

In Section 2 a brief biological overview on BC is presented. In Section 3 the data used in the study are described, as well as the computational methods used in the definition of the classification framework. In Section 4 the results of the performance evaluation of the classifier are presented, whereas in Section 5 we briefly discuss on the possible application and repercussions of the method, with particular regard to the theranostic sphere.

3. BIOLOGICAL BACKGROUND

In the last ten years, several distinct studies have suggested that Breast Cancer (BC) is a deeply heterogeneous disease, which conventional histopathology is not capable of describing. Some attempts of classification have been made by molecular profile analyses, which have lastly classified BC in four major subtypes.

The gene expression profile proposed in (34) demonstrated that BC could be molecularly classified in Luminal A, Luminal B, HER2-like, Basal-like and Normal-like, based on 496 genes. Each subtype was shown to correlate with clinical outcome (10). This intrinsic subtype classification has been proven stable across several platforms and patient cohorts (35). In a microarray analysis a minimal set of 50 genes (PAM50) was able to describe molecularly the intrinsic subtypes (36). Some attempts have been made to combine the intrinsic subtype classification with immunohistochemical markers, such as estrogen receptor (ER), progesterone receptor (PgR) and HER2 (37, 38), but also including proliferation marker Ki67, basal cell markers CK5/6 and EGFR (39-41).

The recently proposed classification of primary breast cancer subtypes according to the 2015 St. Gallen Consensus Conference and recommended by the ESMO Clinical Practice Guidelines (30) suggests that the BC should be divided in four intrinsic subtypes, on the basis of routine histology and immunohistochemistry data:

1. *Luminal A subtype*, that includes BC that are ER-positive, HER2-negative, Ki67 low, PgR high, low-risk molecular signature, if available

2. *Luminal B subtype*, which includes two groups: HER2-negative BC (ER-positive, HER2-negative and either Ki67 high or PgR low, high-risk molecular signature, if available); and HER2-positive BC (ER-positive, HER2-positive, any Ki67, any PgR)

3. *HER2 overexpression subtype*, which includes those BC that shows HER2-positive staining and ER and PgR negative staining

4. *Basal-like subtype*, which includes those BC negative for ER, PgR and HER2. In this case there is a 80% overlap between 'triple negative' and 'basal-like' subtype, but triple negative BC also includes some special histological types (i.e., medullary, adenoid cystic carcinoma with low risks of distant recurrence)

This last classification was considered in the application of our approach, to describe differentially expressed genes (DEG) in each of the four intrinsic BC subtypes vs. normal samples (NS).

4. METHODS

4.1. Data sources

In our study we used two distinct cancer datasets in the different phases of the classifier implementation, in order to avoid any possible cohort-specific bias.

The pathway enrichment phase was performed on a specific BC dataset retrieved from The Cancer Genome Atlas (TCGA) public repository: TCGA-BRCA[2]. In particular, we used the expression level of mRNAs extracted from Illumina HiSeq RNASeqV2 platform with respect to 233 BC luminal A samples, 103 BC luminal B samples, 74 BC Basal samples, 43 BC HER2 samples, and 113 Normal Samples (NS).

In the training and test of the classifier we used a second BC dataset taken from Gene Expression Omnibus (GEO): GSE58212[3], including 121 BC Luminal A, 69 BC Luminal B, 37 BC Basal and 32 BC HER2 samples.

4.2. Multiclass classification of BC subtypes

The goal of a classifier is to assign a (qualitative) category to a sample (e.g. a tumor sample) based on the value of a certain set of variables (e.g., gene expression levels) and prior information. Usually, *supervised* classifiers are trained on samples with established categories (e.g., tumor subtypes classification), and the resulting model can be later applied to predict the category of a new sample, thus we here split the reference dataset in *training* and *test* sets (see below for details).

[2] website: https://gdc-portal.nci.nlm.nih.gov/projects/TCGA-BRCA.
[3] website: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE58212.

As we aim at classifying tumor samples in more than two distinct classes, i.e., distinct tumor subtypes we here speak of *multiclass classifier* (not to be confused with multi-label classifiers, in which each instance can belong to more than one class). Even if there exist algorithms that are naturally developed to deal with multiclass classification tasks, we here make use of an extension of a largely used binary classifier, purposely extended to multiple classes (see below), because of ascertained advantages from the computational perspective.

In general, the performance of a classifier (trained on the training set) is determined by measuring the misclassified samples in the test set (we here do not consider the so-called training errors, i.e. the number of misclassified samples in the training set). More in detail, in the following we will analyze the performance in terms of:

Accuracy: $\frac{tp+tn}{tp+tn+fp+fn}$, Precision: $\frac{tp}{tp+fp}$,

Recall: $\frac{tp}{tp+fp}$,

where true/false positives/negatives (*tp, fp, tn, fn*) are computed by looking at each class distinctly, i.e., by analyzing whether each sample of the test set belonging to a specific class is correctly classified within that class, or to another one. In this way it is possible to compute values accuracy, precision and recall specific for each class. In general, the accuracy value indicates the capability of the method in correctly classifying a random sample, precision indicates the ratio of samples that are assigned to a certain class and that actually belong to it, whereas recall the ratio of samples belonging to a class that are correctly classified. For all the measures, values closer to 1 indicate a better performance.

As specified in the introduction, feature selection is fundamental in reducing the space of variable, mostly considering that the number of samples is usually much lower than the number of genes, in cancer classification tasks. In general, feature selection is aimed at selecting the most informative variables that can discriminate among groups, i.e., cancer subtypes in our case.

Feature selection methods reduce the dimensionality of the original feature space to a lower dimensional space, by selecting a subset of variables. In our case feature selection is performed via enrichment of the relevant pathways that characterize the distinct subtypes, in order to reduce the number of variables (i.e., the differentially expressed genes) to be used by the classifier.

Hence, the classifier implementation is based on a 2-step procedure:

1. Enrichment of relevant pathway for feature selection

2. SVM-based OvO (one vs. one) multiclass classification

4.2.1. Enrichment of relevant pathway for feature selection

Differentially expressed genes (DEGs) between each subtype class of BC samples and the class of N of the TCGA-BRCA dataset were identified by statistical analysis using the function *TCGAanalyze DEA* from the package TCGAbiolinks Bioconductor. The following parameters were used: quantile-adjusted conditional maximum likelihood, *abs*(*log* fold change) > 1, and FDR < 0.0.1 (8). The obtained p-values were adjusted by using the Benjamin-Hochberg procedure for multiple testing correction (42).

Given 1077 pathways derived from the REACTOME (43), BIOCARTA (44) and KEGG (45) databases, a pathway enrichment analysis was applied. We used these three databases since they are currently the most commonly used pathway databases. The enrichment was evaluated using the Fisher's Exact Test between differentially expressed genes and the selected pathways. We considered a pathway to be enriched if p-value was <0.0.1.

We finally obtain lists of differentially expressed genes (DEGs) for each considered subtype, and for each gene in the list we have a fold-change value with respect to the baseline value.

4.2.2. SVM-based OvO classifier

Support vector machines (SVMs) (46, 47) are widely used in binary classification tasks. SVMs transform nonlinearly separable problems into linearly separable ones by projecting the data into a higher dimensional feature space, there searching for the hyperplane that separates two classes of data with the largest possible margin.

Given a training set $T=\{(\vec{x}_i,y_i|(\vec{x}_i,y_i)\in R^m\times R, i=1,2,\ldots,l\}$ with labels $y_i\in\{-1,1\},(i=1,2,\ldots,l)$,

where l is the number of samples in the training set, *xi* is the vector of *m* different attributes for each sample, the SVM searches for a maximum margin hyperplane where *T* can be linearly separated, via the solution of the primal problem

$$\text{min: } P(\vec{w},b,\vec{\zeta})=\frac{1}{2}\vec{w}^T\cdot\vec{w}+C\sum_{i=1}^{l}\zeta_i,$$

$$\text{with: } \begin{cases} y_i(\vec{w}^T\varphi(\vec{x}_i)+b)\geq 1-\zeta_i \\ \zeta_i\geq 0, i=1,2,\ldots,l \end{cases}$$

where $\vec{w}$ is a m-dimensional vector, b is a scalar and ζ_i the slack variables. C is a penalty parameter to adjust the trade-off between the margin maximization and the classification error minimization. SVM maps data $\vec{x}_i$ to a higher-dimensional space via the function $\phi(\cdot)$. The solution of the primal optimization problem is usually done by solving its dual problem of a Lagrangian formulation (see, e.g., (48)):

$$\text{min: } D(\vec{\alpha}) = -\sum_{i=1}^{l}\alpha_i + \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j K(\vec{x}_i, \vec{x}_j),$$

$$\text{subject to: } \begin{cases} \sum_{i-1}^{l} y_i \alpha_i = 0 \\ 0 \le \alpha_i \le C, i = 1,2,\ldots,l \end{cases}$$

where α_i, i = 1,2,…, l is a positive Lagrange multiplier. The kernel function $K(\vec{x}_i, \vec{x}_j)$ computes the dot-product for the data in the high-dimensional feature space. The decision function is finally defined as:

$$g(\vec{x}) = sign\left\{ \sum_{\vec{x}_i \in SV_S} \alpha_i y_i K(\vec{x}_i, \vec{x}) + b \right\}, \qquad \text{where}$$

vector $\vec{x}$ is one of the Support Vectors when $0 < \alpha_i \le C$.

Binary SVMs were extended to account for multi-class classification tasks with several techniques. In particular, we here used the *one vs. one* (OvO) method for the multiclass classification via SVM. A distinct classifier is trained for each pair of classes, resulting in *k(k-1)/2* independent classifiers.

In the prediction/validation step each sample is tested with each classifier and the classification label that results as the most frequent is chosen (i.e., *majority rule*). In case a tie among different labels occurs, the label that is returned by the binary classifiers with the best scores combined (computed via the distance of data points from the separating hyperplane) is chosen. The method is criticized to solve the problem symmetrically and for the overlap in training set due to the pairwise comparison. Yet, it is computationally cheap and the accuracy is proven to be very high with respect to cancer subtypes classification.

4.2.3. Dataset preprocessing

The original GEO: GSE58212 BC microarray dataset, *G* from now on, is processed as follows prior to the classifier training.

Given $G = \{p_{i,j} \mid i=1,2,\ldots,P,\ j=1,2,\ldots M\}$, each element $p_{i,j}$ contains the expression level of the i^{th} probe in the j^{th} sample, where *P* is the total number of probes and *M* the number of samples (for the exact number of samples in our case please refer to Data Sources subsection). The gene expression levels are then normalized via standard *Z-score* procedure, so that the average value for the whole dataset is 0 and the standard deviation is 1, $n_{i,j}=(p_{i,j}-|G|)/\sigma(G),\ i=1,2,\ldots,P,\ j=1,2,\ldots M$.

In particular, for each gene the median value of the expression level of each probe mapped to that gene is considered, $g_{z,j} = median(n_{1,j}, n_{2,j}, \ldots, n_{Pz,j})$, where P_z is the number of probes mapped to the gene *z* in sample *j*. Thus, the pre-processed dataset is defined as $J=\{g_{z,j} \mid z=1,2,\ldots,T,\ j=1,2,\ldots M\}$, where *T* is total number of the unique genes (one gene can be mapped to more than one probe).

As known, learning the parameters of a classifier and testing it on the same dataset is not sufficient to ensure reproducible prediction outcomes. To this end it is common practice to divide the original dataset in two distinct portions, named *training* and *test* set.

In our case, we adopt a *10-fold cross validation* and we repeat the classification training and test 20 times to obtain a measure of the robustness of the method. Given *k=4* different classes (i.e., the four BC subtypes: Luminal A, Luminal B, Her 2, Basal), each run of cross-validation consists in training *k(k-1)/2=6* distinct binary models on the training test and evaluating the predictions on the test set. More in detail, the steps are the followings:

1. The original dataset *J* is split in *w=10* groups, meaning that the samples are randomly divided into 10 different groups of the same size

2. *w-1=9* randomly chosen groups are merged into the training data set

3. the remaining group is the test set.

The whole training/test procedure is repeated 20 times, leading to the instantiation of 20 different classifiers, which will be used to evaluate the accuracy and the robustness of the overall methodology (see Section 4)[4].

4.2.4. Features selection

In order to reduce the huge variable space and to obtain a reduced number of key gene signatures able to discriminate the different cancer subtypes, we used different strategies, especially based on the information on pathway activity. In particular, different subsets of DEGs derived during the pathway enrichment phase are selected as variables for each

[4] As different classifiers might result in different sample-class associations we here do not show the classification results for each single classifier, yet we provide a performance evaluation of the method based on average values of accuracy, precision and recall.

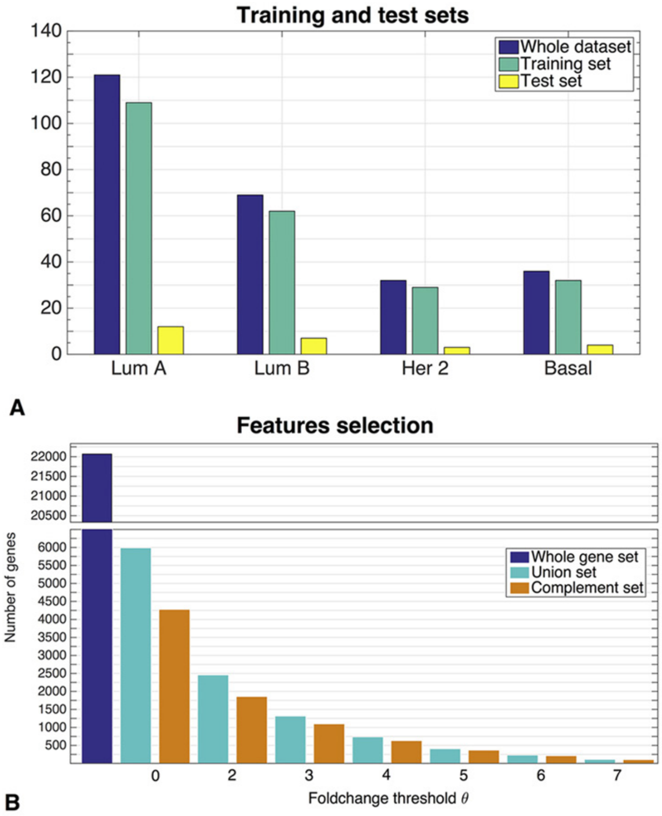

Figure 1. (A) Cardinality of i) the whole sample set, ii) the training set and the iii) test set with respect to the four distinct subtypes, i.e., Luminal A, Luminal B, Her 2 and Basal, as taken from the Gene Expression Omnibus (GEO): GSE58212 original dataset, used in the classification phase. (B) Number of genes (i.e., variables) selected for classifier training and test after features selection, with respect to: i) union or complement set of DEGs, and ii) fold-change threshold in the expression level between the cancer subtype case and the normal case (see Feature selection subsection for details).

specific subtype, in the training and test phases of the classifier, according to the following criteria.

1. Union vs. Complement set of DEGs

   Given that any two-classes classification task involves the two distinct sets of DEGs concerning the two classes, two selection criteria are possible:

   1.1. Complement of the DEG sets, i.e., selecting the genes that belong to either one DEG set or the other, but not the genes belonging to the intersection set. In this case the rationale is that the largest discriminatory capability is hidden in those genes that are indeed differentially expressed in one subtype and not in the other and vice versa.

   1.2. Union of the DEG sets, i.e., selecting all the genes that belong to one DEG set and to the other, i.e., the union of the sets. In this case the justification is that some useful information to be used in the classification task could be retrieved also from those genes that are differentially expressed in both classes.

2. Fold-change threshold

Another criterion to reduce the variable space could be derived from the difference observed in the (average) expression of a certain gene in the wild-type case (NS) versus the cancer case, which is evaluated via a fold-change assessment in the pathway enrichment phase. Accordingly, it is possible to select as relevant variables only those genes that (on average) exceed a certain fold-change threshold in terms of expression level in the distinct cases. In this case, the rational is that the classification benefits most from those genes that are more differentially expressed in the various cases.

In particular, we performed a parameter scan analysis, by assessing the performance of the classifier with different fold-change thresholds: θ={0,2,3,4,5,6,7} which accordingly lead to different cardinalities of the variable sets.

The combination of the union/complement choice and of different fold-change results in distinct modeling choice affecting the classifier performance, which will be evaluated in the following section. In Figure 1 (right) one can see the cardinality of the gene sets selected after feature selection, in case of either union or complement set, and of the distinct values of θ.

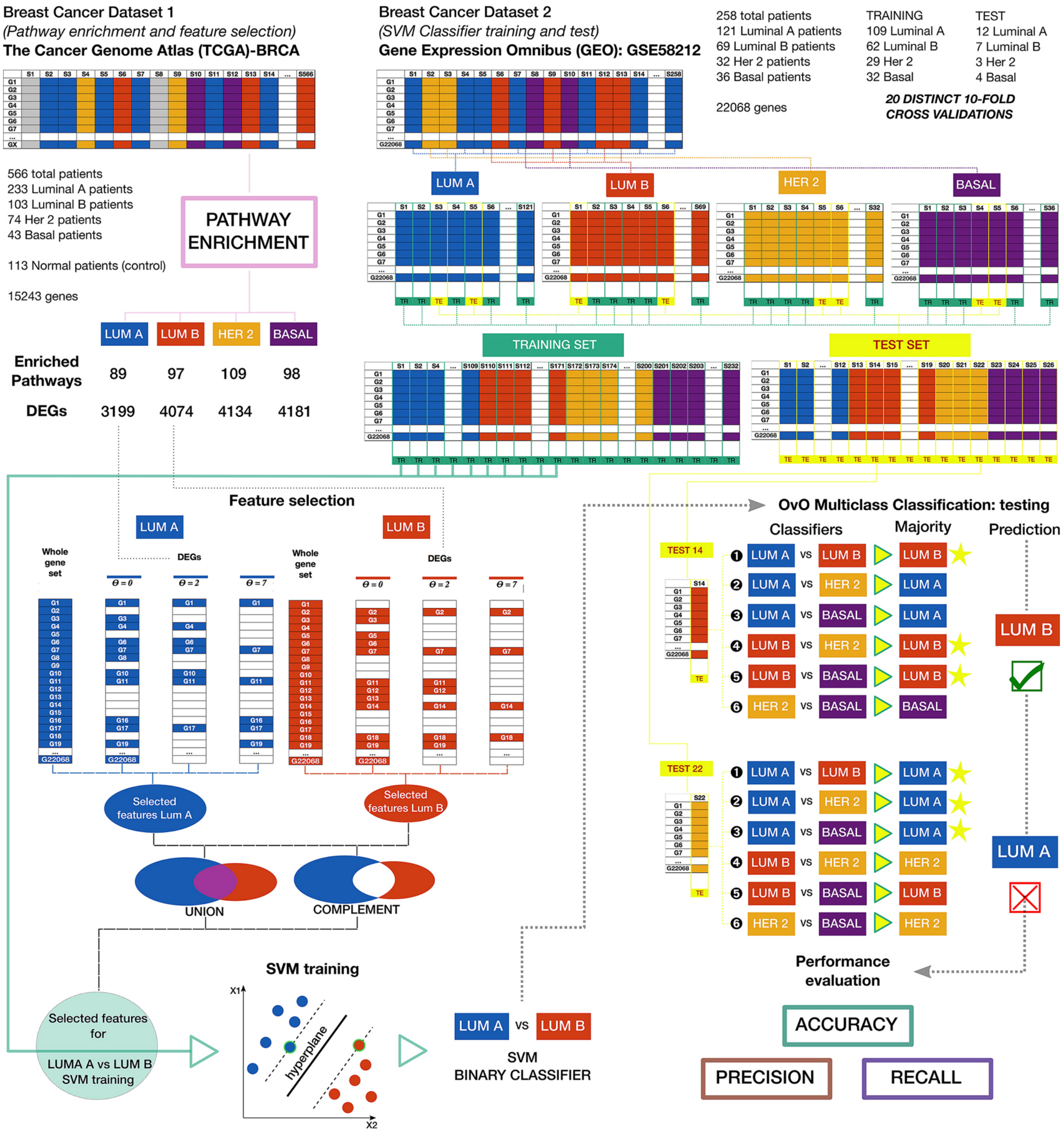

Figure 2. Schematized representation of the classifier implementation and structure. Two distinct BC datasets are used, namely: i) TCGA-BRCA and ii) GEO: GSE58212. The former dataset is used in the pathway enrichment phase and leads to the detection of different enriched pathways and differentially expressed genes with respect to the four distinct BC subtypes. The DEGs are then used in the subsequent feature selection phase, in which either the DEG union or complement set can be employed, in addition to the filtering of DEGs on the base of the expression level fold-change. The second dataset is split in training and test sets, according to 20 distinct 10-fold cross-validations, which allow to test the classifier robustness. The training sets are used to train 6 distinct SVM binary classifiers (i.e., all the possible couples of the four subtypes), on the bases of the features selected in the enrichment phase. Finally, a OvO strategy is used in the prediction phase, hence allowing to the test the classifier performance on the test set in terms of accuracy, precision and recall.

Finally, the SVM binary classifiers are implemented via the *Matlab* functions *svmtrain* and *svmclassify*. In our case we use a *linear* kernel function, as it displays the best performance with respect to the other tested nonlinear kernel functions (results not shown here). In Figure 2 one can see a summarization of the whole procedure.

5. RESULTS

5.1. Relevant pathway enrichment

On the basis of the original TCGA dataset, in the enrichment phase we found: 89 pathways enriched with 3199 DEGs for Luminal A vs. NS, 97 pathways

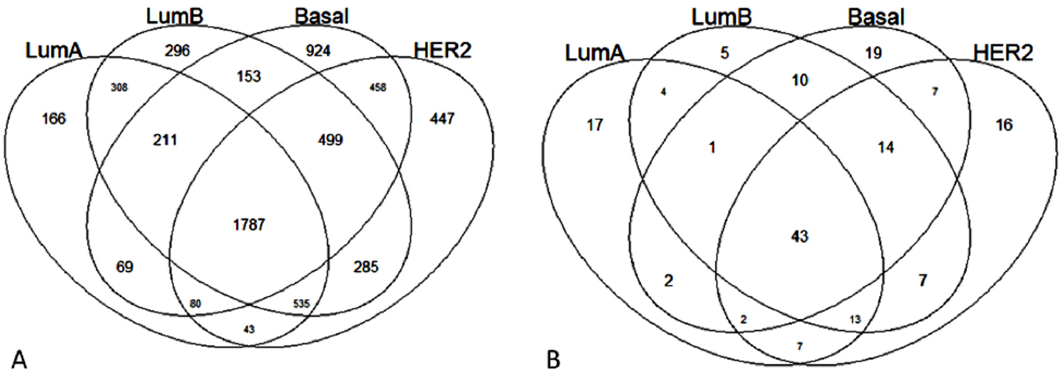

Figure 3. Venn diagrams for: (A) differentially expressed genes (DEGs), (B) pathways enriched of DEGs in the four breast cancer subtypes.

enriched with 4074 DEGs for Luminal B vs. NS, 98 pathways enriched with 4181 DEGs for basal vs. NS and 109 pathways enriched with 4134 DEGs for HER2 vs. NS.

Figure 3 A and B show Venn Diagrams for DEGs and pathways enriched of DEGs in the distinct BC subtypes, respectively.

In particular, we found 1787 DEGs in common among subtypes. Specific genes for each subtype are 166 DEGS for luminal A, 296 for luminal B, 924 for basal and 447 for HER2. Furthermore, we found 43 pathways in common among subtypes. We detected pathways specific for each subtype in this proportion: 17 for luminal A, 5 for luminal B, 19 for basal and 16 for HER2. In Table 1 one can find the list of specific pathways enriched for each subtype.

More in detail, in BC luminal A we identified 17 specific pathways enriched of DEGs, among them BIOCARTA intrinsic pathway and REACTOME ethanol oxidation. Intrinsic Prothrombin Activation Pathway performs an essential role in coagulation, a crucial step for the organization of metastasis also in experimental models of cancer (49). The pathway of ethanol oxidation contains several genes belonging to the family of ALDH genes and that, despite the known role in ethanol detoxification, are also considered biomarkers of cancer stem cells (50).

In BC luminal B we identified 5 specific pathways enriched of DEGs, among them BIOCARTA cell cycle pathway and REACTOME cell junction organization. Cell junctions are structures for cell-cell adhesion machinery related to the differentiation and normal growth of the tissue (51). The development of cancer represents a modification of normal tissue homeostasis and an alteration in cell-cell interaction. In addition, cancer metastasis spreads through the circulatory system caused by cell adhesion (51). Loss of control of cell cycle pathway is considered a hallmark of many cancer and deregulated genes involved in cell cycle regulation are implicated in cancer progression (52).

In BC basal we identified 19 specific pathways enriched of DEGs, among them REACTOME amino acid synthesis and interconversion transamination and REACTOME metabolism of amino acids and derivatives. High protein production in cancer cells increases the overall need for amino acids. Deregulated amino acid metabolism also has a function in immune tolerance in cancer (53).

In BC HER2 we identified 16 specific pathways enriched of DEGs, among them REACTOME p2y receptors and REACTOME axon guidance. P2Y receptors (e.g., P2Y1, P2Y2) have strong direct roles on the tumour by modulating cell development. *In vivo* data confirm *in vitro* evidence that lowering the intratumour adenosine concentration and targeting the P2X7 receptor have a strong anti-tumor outcome (54). Axon guidance pathway includes four families of secreted or membrane bound factors (i.e., netrin 1, semaphorine, ephrins, and Slit, all with their receptors), which have recently studied as central agents in tumour progression. Far from being confined to the developing brain, Axon guidance pathway seems to play an important function in tumour cell migration, tumour cell survival and tumour angiogenesis (55).

5.2. Classification performance evaluation

As specified above, different multiclass classifiers are instantiated with distinct feature selection strategies and different parameter settings, in order to empirically identify an optimal tradeoff between the desired accuracy and the number of variables to analyze. As in our case we deal with gene expression levels or, similarly, with gene sequencing data, it might be desirable to reduce the number of gene to analyze, in terms of experimental costs and times, yet without compromising the reliability of the results.

As described above, we here compare different gene sets, namely: i) DEG union set, ii) DEG complement set, and for each case we reduce the size of the variable set by analyzing seven different values of the DEG fold-change threshold, i.e., θ = 0,2,3,4,5,6,7.

Table 1. Unique pathways enriched of differentially expressed genes for each breast cancer subtype: 17 pathways for luminal A, 5 for luminal B, 19 for basal and 16 for HER2

LUMINAL A	LUMINAL B	BASAL	HER2
REACTOME degradation of the extracellular matrix	KEGG arrhythmogenic right ventricular cardiomyopathy arvc	REACTOME mrna splicing	REACTOME activation of the mrna upon binding of the cap binding complex and eifs and subsequent binding to 43s
BIOCARTA intrinsic pathway	REACTOME crmps in sema3a signaling	REACTOME activation of the pre replicative complex	REACTOME unfolded protein response
REACTOME abc family proteins mediated transport	BIOCARTA cellcycle pathway	KEGG spliceosome	REACTOME developmental biology
REACTOME ethanol oxidation	KEGG ribosome	KEGG apoptosis	KEGG ether lipid metabolism
BIOCARTA ami pathway	REACTOME cell junction organization	REACTOME activation of atr in response to replication stress	REACTOME axon guidance
REACTOME metabolism of carbohydrates		KEGG dna replication	REACTOME nucleotide like purinergic receptors
REACTOME glycerophospholipid biosynthesis		REACTOME interferon alpha beta signaling	REACTOME cgmp effects
REACTOME platelet activation signaling and aggregation		BIOCARTA ranms pathway	REACTOME fgfr ligand binding and activation
KEGG chemokine signaling pathway		KEGG type i diabetes mellitus	REACTOME phospholipase c mediated cascade
BIOCARTA eryth pathway		BIOCARTA g2 pathway	REACTOME glycolysis
BIOCARTA longevity pathway		KEGG bladder cancer	KEGG histidine metabolism
REACTOME triglyceride biosynthesis		REACTOME s phase	KEGG natural killer cell mediated cytotoxicity
REACTOME transmembrane transport of small molecules		REACTOME g1 s transition	REACTOME asparagine n linked glycosylation
BIOCARTA cftr pathway		REACTOME amino acid synthesis and interconversion transamination	REACTOME p2y receptors
REACTOME o linked glycosylation of mucins		REACTOME metabolism of amino acids and derivatives	REACTOME keratan sulfate keratin metabolism
REACTOME transport of glucose and other sugars bile salts and organic acids metal ions and amine compounds		KEGG arginine and proline metabolism	KEGG melanoma
REACTOME factors involved in megakaryocyte development and platelet production		KEGG glycolysis gluconeogenesis	
		BIOCARTA mcm pathway	
		REACTOME extension of telomeres	

As one can notice in Figure 1, the difference in the size of the variable sets with respect to the feature selection choice are remarkable, as the whole datasets include more than 22.0.00 genes, whereas, for instance, by choosing the complement DEG set with *θ = 0* the number of genes is already reduced to around 4200 genes (5-fold difference), while with the union DEG set with *θ = 4* we deal with around 700 genes (30-fold difference) and in our limit case, i.e., the complement set with *θ = 7,* only 104 genes are considered as relevant variables (around 210-fold difference). This aspect must be considered in the choice of the optimal tradeoff between accuracy and number of variables to analyze.

In Figure 4 and 5, one can see the variation of the average values of accuracy, precision and recall, computed on 20 different classifiers, implemented with random training and test sets via a 10-fold cross-validation. This procedure is made to assess the robustness of the method with respect to the selection of the sample sets.

One can firstly notice that no statistically significant differences are observed between the performance of the union and the complement sets, with respect to all the indicators, and this first important result suggests that most of the relevant information to discriminate BC subtypes is contained only in the

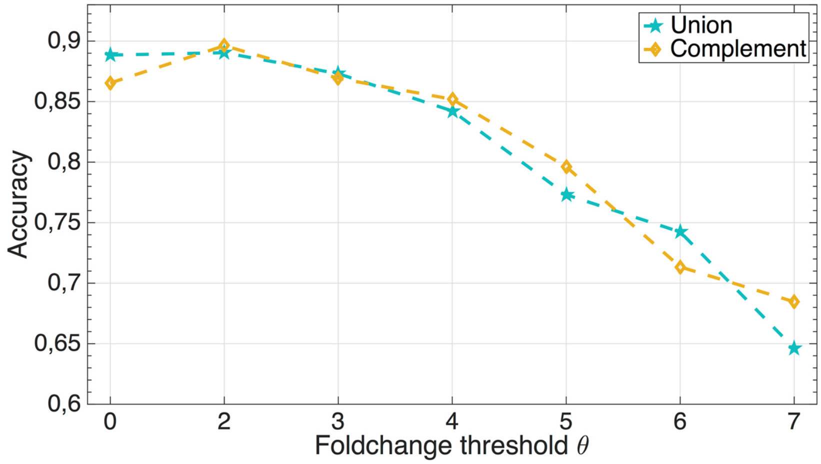

Figure 4. Variation of the average accuracy of the classifier computed on 20 distinct training and test runs, via a 10-fold cross validation of the GEO dataset. In the Figure, the feature selection involving the union/complement DEG set and the different fold-change threshold values are shown.

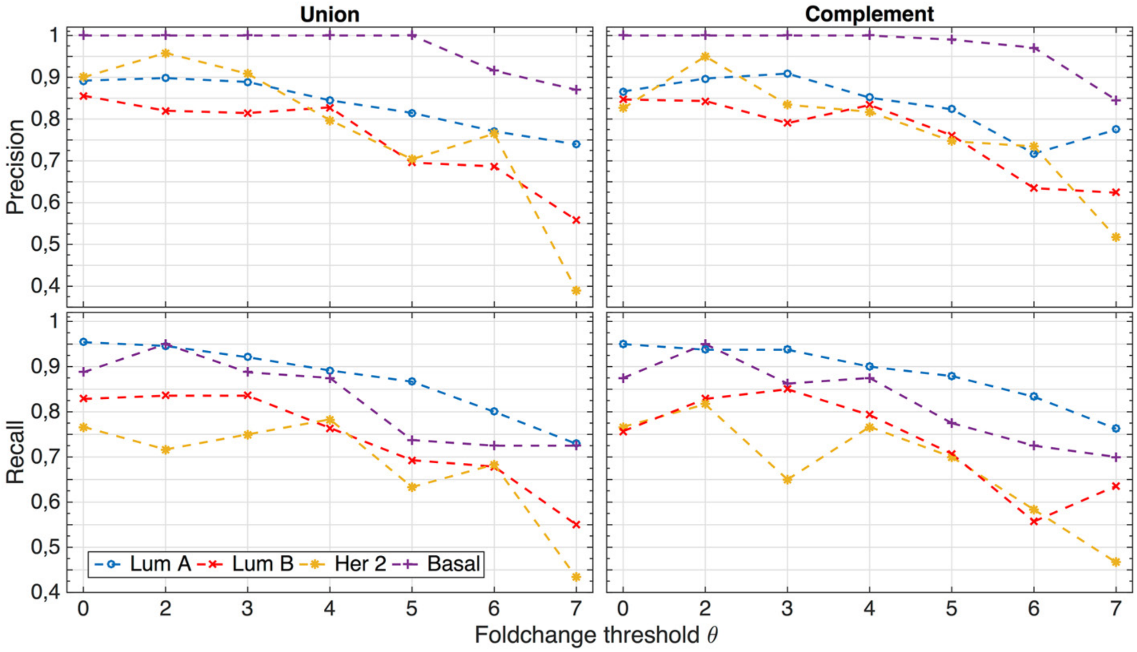

Figure 5. Variation of the average precision and recall of the classifier for each specific BC subtype, computed on 20 distinct training and test runs, via a 10-fold cross validation of the GEO dataset. In the Figure, the feature selection involving the union/complement DEG set and the different fold-change threshold values are shown.

DEGs specific to each subtype, being the intersection DEGs substantially irrelevant. This allows a first important reduction of the feature space, given that, on average, the cardinality of the union set exceeds that of the complement set in a range between the 5% and the 15% (according to the different fold-change threshold).

In terms of overall (average) accuracy, the classifier presents a very good performance, around 0.8.5-0.9., for values of θ up to 4, with the best overall performance observed for the combination of DEG complement set and $\theta = 2$ (accuracy 0.9.), and a slightly worse accuracy for $\theta = 3$ (accuracy 0.8.7). This is an extremely important result as the variable space in these two cases is dramatically reduced: in fact, in the former case (i.e., complement set and $\theta = 2$) 1858 DEGs are selected, indicating a 12-fold change with respect to the whole gene set, whereas in the latter case (i.e., complement set and $\theta = 3$), only 1098 DEGs are used, denoting a 20-fold change. This also suggests that some of the filtered-out genes might even act as confounding factors in the classification task. Furthermore, by combining the information on the enriched pathways and that on the selected DEGs, one could tentatively isolate a list of possible biomarkers to modulate in the design of new prognostic and therapeutic strategies.

For larger values of θ the average accuracy worsens, yet is maintained to acceptable values as compared to a dramatic reduction of the variable space.

By looking at the values of precision and recall specific to each subtype, one can see that all the subtypes present a similar and very good performance, up to certain values of θ (smaller than 5), with the Basal subtypes displaying better and more stable precision values and the Her 2 a slightly worse recall. In general, the precision values are (on average) slightly larger than the corresponding recall values, indicating a high reliability of the classifier in assigning a certain sample to its correct class, y*et al*lowing for some false positives.

5.2.1 Comparison with other techniques

A large number of computational approaches aim at classifying distinct diseases from genomic data, relying on different theoretical frameworks and distinct kinds of source data. For instance, several algorithms make use of SVMs, e.g., (56-58), some others rely on the information on pathways, e.g., (27-29,59-61) or on distinct approaches, e.g., (62).

However, our methodological framework cannot be directly compared with most of the aforementioned techniques, as it aims at categorizing (breast) cancer *subtypes*, which is an intrinsically harder problem because of the similarity of the source mutational profiles, yet displaying a comparable accuracy with the best classifiers, sometimes in spite of a larger variable space.

Three recent methods, i.e., (63-65), instead, focus on a similar problem. In this regard, the performance of our method is definitely comparable to that shown in (63) in which, however, the authors make use also of the additional information on Copy number variations and microRNAs-regulated mRNAs to reduce the number of features (i.e., total accuracy slightly lower than 90%, sensitivity values ranging between 0.5.2 and 0.9.8 according to the different subtypes, and specificity values between 0.8.5 and 0.9.6, see Figure 6 in (63)). With respect to (64) we obtain a considerably better overall accuracy, yet they focus on the classification of metastatic/non metastatic breast tumors, which is a notably different problem (i.e., accuracy values significantly lower than 70% with respect to the two considered datasets and all the tested methods, see Figure 4 in (64)). Finally, we could not directly compare the accuracy of our technique with that of (65), in which they use random forests and methylation data in addition to gene expression data, as they provide a performance evaluation in terms of classification errors on the training set and of leave-one-out bootstrap errors.

We finally remark that, even if further developments to improve the accuracy of our method are ongoing, we here deliver a usable tool that is specifically directed toward the development of personalized treatments.

6. DISCUSSION

Even though the quest for the identification of key biomarkers in cancer research is far from being concluded, the combination of biological knowledge and computational techniques can lead to remarkable results, especially by providing theoretical and practical support to experimentalists and pathologists in the definition of novel and effective diagnostic, prognostic and therapeutic strategies.

In particular, the integrated approach combining the information on genes and pathways at different levels has been recently used in other works of our group, for instance in (7). In this specific case we focused on the use of the pathway activity notion to develop an effective classification tool with respect to cancer subtype.

More in detail, we here introduced a new SVM-based classifier of breast cancer subtypes based on a supervised variable selection related to the differential expression levels of key genes, as detected by a pathway enrichment phase.

The results are significant as the classifier displays an overall accuracy slightly lower than 0.9. when the number of variable is reduced by a 20-fold with respect to the original gene set. In other words, this new instrument allows to efficiently classifying cancer patients into distinct subtypes with an excellent reliability even if limited experimental resources allow to access to information on a limited number of genes only, thus providing an accessible and widely-usable tool to the theranostic community. Furthermore, the list of genes selected by the tool could provide important insights for downstream analyses of previously reported microRNA signatures (23) and eventually indications for the development of targeted therapies, which could be based on the concept of pathway activity.

To conclude, the current efforts will be followed by the implementation of a widely accessible online application via a dedicated web portal, which is currently in developmental phase, and that will allow the user to classify specific cancer expression profiles on the basis of a reduced number of selected genes.

7. ACKNOWLEDGMENTS

This work has been partially supported by grants from the SysBioNet project, a MIUR initiative for the Italian Roadmap of European Strategy Forum on Research Infrastructures (ESFRI). BF and KF are supported by grants from the South-East Norway Health Authority.

8. REFERENCES

1. M Gerlinger, AJ Rowan, S Horswell, J Larkin, D Endesfelder, E Gronroos, P Martinez, N Matthews, A Stewart, P Tarpey, I Varela, B Phillimore, S Begum, NQ McDonald, A Butler, D Jones, K Raine, C Latimer, CR Santos, M Nohadani, AC Eklund, B Spencer-Dene, G Clark, L Pickering, G Stamp, M Gore, Z Szallasi, J Downward, PA Futreal, C Swanton: Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *N Engl J Med*, 366(10):883-92 (2012)
DOI: 10.1056/NEJMoa1113205

2. R Fisher, L Pusztai, C Swanton: Cancer heterogeneity: implications for targeted therapeutics. *Br J Cancer,* 108(3):479-85 (2013)
DOI: 10.1038/bjc.2012.581

3. RA Burrell, C Swanton: Tumour heterogeneity and the evolution of polyclonal drug resistance. *Mol Oncol*, 8(6):1095-111 (2014)
DOI: 10.1016/j.molonc.2014.06.005

4. R Mirnezami, J Nicholson, A Darzi. Preparing for precision medicine. *N Engl J Med*, 366(6):489-91 (2012)
DOI: 10.1056/NEJMp1114866

5. National Cancer Institute; National Genome Research Institute (2015) The Cancer Genome Atlas (Natl Inst Health, Bethesda). Available at https://tcga-data.nci.nih.gov/tcga. Accessed Sept 30, 2016.

6. G Caravagna, A Graudenzi, D Ramazzotti, R Sanz-Pamplona, L De Sano, G Mauri, V Moreno, M Antoniotti, B Mishra: Algorithmic methods to infer the evolutionary trajectories in cancer progression. *Proc Natl Acad Sci U S A*, 113(28):E4025-34 (2016)
DOI: 10.1073/pnas.1520213113

7. C Cava, G Bertoli, I Castiglioni: Integrating genetics and epigenetics in breast cancer: biological insights, experimental, computational methods and therapeutic potential. *BMC Syst Biol*, 1,9:62 (2015)
DOI: 10.1186/s12918-015-0211-x

8. A Colaprico, TC Silva, C Olsen, L Garofano, C Cava, D Garolini, TS Sabedot, TM Malta, SM Pagnotta, I Castiglioni, M Ceccarelli, G Bontempi, H Noushmehr: TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res*,44(8):e71 (2016)
DOI: 10.1093/nar/gkv1507

9. C Cava, I Zoppis, M Gariboldi, I Castiglioni, G Mauri, M Antoniotti: Copy–Number Alterations for Tumor Progression Inference. *Lecture Notes in Computer Science*, 7885:104-109 (2013)
DOI: 10.1007/978-3-642-38326-7_16

10. T Sorlie, CM Perou, R Tibshirani, T Aas, S Geisler, H Johnsen, T Hastie, MB Eisen, M van de Rijn, SS Jeffrey, T Thorsen, H Quist, JC Matese, PO Brown, D Botstein, PE Lonning, AL Borresen-Dale: Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proc Natl Acad Sci U S A*, 98(19):10869–10874 (2001)
DOI: 10.1073/pnas.191367098

11. J Khan, JS Wei, M Ringner, LH Saal, M Ladanyi, F Westermann, F Berthold, M Schwab, CR Antonescu, C Peterson, PS Meltzer: Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med*, 7(6):673–679 (2001)
DOI: 10.1038/89044

12. D Singh, PG Febbo, K Ross, DG Jackson, J Manola, C Ladd, P Tamayo, AA Renshaw, AV D'Amico, JP Richie, ES Lander, M Loda, PW Kantoff, TR Golub, WR Sellers: Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209 (2002)
DOI: 10.1016/S1535-6108(02)00030-2

13. LJ van 't Veer, Dai H, MJ van de Vijver, YD He, AA Hart, M Mao, HL Peterse, K van der Kooy, MJ Marton, AT Witteveen, GJ Schreiber, RM Kerkhoven, C Roberts, PS Linsley, R Bernards, SH Friend: Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536 (2002)
DOI: 10.1038/415530a

14. C Sotiriou, P Wirapati, S Loi, A Harris, S Fox, J Smeds, H Nordgren, P Farmer, V Praz, B Haibe-Kains, C Desmedt, D Larsimont, F Cardoso, H Peterse, D Nuyten, M Buyse, MJ Van de Vijver, J Bergh, M Piccart, M Delorenzi: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*, 98(4): 262–272 (2006)
DOI: 10.1093/jnci/djj052

15. V Popovici, W Chen, BG Gallas, C Hatzis, W Shi, FW Samuelson, Y Nikolsky, M Tsyganova, A Ishkin, T Nikolskaya, KR Hess, V Valero, D Booser, M Delorenzi, GN Hortobagyi, L Shi, WF Symmans, L Pusztai: Effect of training-sample size and classification difficulty on the accuracy of genomic predictors. *Breast Cancer Res*, 12(1):R5 (2010)
DOI: 10.1186/bcr2468

16. AV Ivshina, J George, O Senko, B Mow, TC Putti, J Smeds, T Lindahl, Y Pawitan, P Hall, H Nordgren, JE Wong, ET Liu, J Bergh, VA Kuznetsov, LD Miller: Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer. *Cancer Res*, 1;66(21):10292-301 (2006)
DOI: 10.1158/0008-5472.CAN-05-4414

17. C Sotiriou, P Wirapati, S Loi, A Harris, S Fox, J Smeds, H Nordgren, P Farmer, V Praz, B Haibe-Kains, C Desmedt, D Larsimont, F Cardoso, H Peterse, D Nuyten, M Buyse, MJ Van de Vijver, J Bergh, M Piccart, M Delorenzi: Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *J Natl Cancer Inst*, 98(4):262-72 (2009)
DOI: 10.1093/jnci/djj052

18. C Cava, G Bertoli, M Ripamonti, G Mauri, I Zoppis, PA Della Rosa, MC Gilardi, I Castiglioni: Integration of mRNA expression profile, copy number alterations, and microRNA expression levels in breast cancer to improve grade definition. *PLoS One*, 9(5):e97681 (2014)
DOI: 10.1371/journal.pone.0097681

19. PC Miller, J Clarke, T Koru-Sengul, J Brinkman, D El-Ashry: A novel MAPK-microRNA signature is predictive of hormone-therapy resistance and poor outcome in ER-positive breast cancer. *Clin Cancer Res*,21(2):373-85 (2015)
DOI: 10.1158/1078-0432.CCR-14-2053

20. TM Severson, J Peeters, I Majewski, M Michaut, A Bosma, PC Schouten, SF Chin, B Pereira, MA Goldgraben, T Bismeijer, RJ Kluin, JJ Muris, K Jirström, RM Kerkhoven, L Wessels, C Caldas, R Bernards, IM Simon, S Linn: BRCA1-like signature in triple negative breast cancer: Molecular and clinical characterization reveals subgroups with therapeutic potential. *Mol Oncol*, 9(8):1528-38 (2015)
DOI: 10.1016/j.molonc.2015.04.011

21. SG Zhao, M Shilkrut, C Speers, M Liu, K Wilder-Romans, TS Lawrence, LJ Pierce, FY Feng: Development and validation of a novel platform-independent metastasis signature in human breast cancer. *PLoS One*, 10(5):e0126631 (2015)
DOI: 10.1371/journal.pone.0126631

22. C Cava, I Zoppis, G Mauri, M Ripamonti, F Gallivanone, C Salvatore, MC Gilardi, I Castiglioni: Combination of gene expression and genome copy number alteration has a prognostic value for breast cancer. *Conf Proc IEEE Eng Med Biol Soc*, 2013:608-11 (2013)
DOI: 10.1109/embc.2013.6609573

23. VD Haakensen, V Nygaard, L Greger, MR Aure, B Fromm, IR Bukholm, T Lüders, SF Chin, A Git, C Caldas, VN Kristensen, A Brazma, AL Børresen-Dale, E Hovig, Å Helland: Subtype-specific micro-RNA expression signatures in breast cancer progression. *Int J Cancer*,139(5):1117-28 (2016)
DOI: 10.1002/ijc.30142

24. A Colaprico, C Cava, G Bertoli, G Bontempi, I Castiglioni: Integrative Analysis with Monte Carlo Cross-Validation Reveals miRNAs Regulating Pathways Cross-Talk in Aggressive Breast Cancer. *Biomed Res Int*, 2015:831314 (2015)
DOI: 10.1155/2015/831314

25. J Tomfohr, J Lu, TB Kepler: Pathway level analysis of gene expression using singular value decomposition. *BMC Bioinformatics*, 6:225 (2005)
DOI: 10.1186/1471-2105-6-225

26. F Rapaport, A Zinovyev, M Dutreix, E Barillot, JP Vert: Classification of microarray data using gene networks. *BMC Bioinformatics*, 8:35 (2007)
DOI: 10.1186/1471-2105-8-35

27. J Su, BJ Yoon, ER: Dougherty: Accurate and reliable cancer classification based on

probabilistic inference of pathway activity. *PLoS One*,4(12):e8161 (2009)
DOI: 10.1371/journal.pone.0008161

28. E Lee, HY Chuang, JW Kim, T Ideker, D Lee: Inferring pathway activity toward precise disease classification. *PLoS comput biol*, 4(11), e1000217 (2008)
DOI: 10.1371/journal.pcbi.1000217

29. L Yang, C Ainali, S Tsoka, LG Papageorgiou: Pathway activity inference for multiclass disease classification through a mathematical programming optimisation framework. *BMC Bioinformatics*,15:390 (2014)
DOI: 10.1186/s12859-014-0390-2

30. A Zhavoronkov, AA Buzdin, AV Garazha, NM Borisov, AA Moskalev: Signaling pathway cloud regulation for in silico screening and ranking of the potential geroprotective drugs. *Front Genet,*5:49 (2014)
DOI: 10.3389/fgene.2014.00049

31. E Senkus, S Kyriakides, F Penault-Llorca, P Poortmans, A Thompson, S Zackrisson, F Cardoso: ESMO Guidelines Working Group.. Primary breast cancer: ESMO Clinical Practice Guidelines for diagnosis, treatment and follow-up. *Ann Oncol*, 24 Suppl 6: vi7-23 (2013)
DOI: 10.1093/annonc/mdt284

32. Cancer Genome Atlas Network: Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61-70 (2012)
DOI: 10.1038/nature11412

33. R Edgar, M Domrachev, AE Lash: Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207-10 (2002)
DOI: 10.1093/nar/30.1.207

34. CM Perou, T Sørlie, MB Eisen, M van de Rijn, SS Jeffrey, CA Rees, JR Pollack, DT Ross, H Johnsen, LA Akslen, O Fluge, A Pergamenschikov, C Williams, SX Zhu, PE Lønning, AL Børresen-Dale, PO Brown, D Botstein: Molecular portraits of human breast tumours. *Nature*, 406(6797):747-52 (2000)
DOI: 10.1038/35021093

35. T Sorlie, R Tibshirani, J Parker, T Hastie, JS Marron, A Nobel, S Deng, H Johnsen, R Pesich, S Geisler, J Demeter, CM Perou, PE Lønning, PO Brown, AL Børresen-Dale, D Botstein: Repeated observation of breast tumor subtypes in independent gene expression data sets. *Proc Natl Acad Sci U S A*,100(14):8418-23 (2003)
DOI: 10.1073/pnas.0932692100

36. JS Parker, M Mullins, MC Cheang, S Leung, D Voduc, T Vickery, S Davies, C Fauron, X He, Z Hu, JF Quackenbush, IJ Stijleman, J Palazzo, JS Marron, AB Nobel, E Mardis, TO Nielsen, MJ Ellis, CM Perou, PS Bernard: Supervised risk predictor of breast cancer based on intrinsic subtypes. *J Clin Oncol*, 27(8):1160-7 (2009)
DOI: 10.1200/JCO.2008.18.1370

37. SG Wu, ZY He, Q Li, FY Li, Q Lin, HX Lin, XX Guan: Predictive value of breast cancer molecular subtypes in Chinese patients with four or more positive nodes after postmastectomy radiotherapy. *Breast*, 21(5):657-61 (2012)
DOI: 10.1016/j.breast.2012.07.004

38. M Kyndi, FB Sørensen, H Knudsen, M Overgaard, HM Nielsen, J Overgaard, Danish Breast Cancer Cooperative Group: Estrogen receptor, progesterone receptor,HER-2, and response to postmastectomy radiotherapy in high-risk breast cancer: the Danish Breast Cancer Cooperative Group. *J Clin Oncol*,26(9):1419-26 (2008)
DOI: 10.1200/JCO.2007.14.5565

39. KD Voduc, MC Cheang, S Tyldesley, K Gelmon, TO Nielsen, H Kennecke: Breast cancer subtypes and the risk of local and regional relapse. *J Clin Oncol*, 28(10):1684-91 (2010)
DOI: 10.1200/JCO.2009.24.9284

40. TO Nielsen, FD Hsu, K Jensen, M Cheang, G Karaca, Z Hu, T Hernandez-Boussard, C Livasy, D Cowan, L Dressler, LA Akslen, J Ragaz, AM Gown, CB Gilks, M van de Rijn, CM Perou: Immunohistochemical and clinical characterization of the basal-like subtype of invasive breast carcinoma. *Clin Cancer Res*, 10(16):5367-74 (2004)
DOI: 10.1158/1078-0432.CCR-04-0220

41. MC Cheang, SK Chia, D Voduc, D Gao, S Leung, J Snider, M Watson, S Davies, PS Bernard, JS Parker, CM Perou, MJ Ellis, TO Nielsen: Ki67 index, HER2 status, and prognosis of patients with luminal B breast cancer. *J Natl Cancer Inst*,101(10):736-50 (2009)
DOI: 10.1093/jnci/djp082

42. Y Benjamini, Y Hochberg: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B: Methodological*, 57(1), 289–300 (1995)

43. A Fabregat, K Sidiropoulos, P Garapati, M Gillespie, K Hausmann, R Haw, B Jassal, S Jupe, F Korninger, S McKay, L Matthews, B May, M Milacic, K Rothfels, V Shamovsky, M Webber, J Weiser, M Williams, G Wu, L Stein, H Hermjakob, P D'Eustachio: The Reactome pathway Knowledgebase. *Nucleic Acids Res*,44(D1):D481-7 (2016) DOI: 10.1093/nar/gkv1351

44. D Nishimura: BioCarta. *Biotech Software & Internet Report*, 2(3):117–120 (2001) DOI: 10.1089/152791601750294344

45. M Kanehisa, S Goto: KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28: 27–30 (2000) DOI: 10.1093/nar/28.1.27

46. V Vapnik: The nature of statistical learning theory. *Springer science & business media* (2013)

47. TS Furey, N Cristianini, N Duffy, DW Bednarski, M Schummer, D Haussler: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics*,16(10):906-14 (2000) DOI: 10.1093/bioinformatics/16.10.906

48. XX Niu, CY Suen: A novel hybrid CNN–SVM classifier for recognizing handwritten digits. *Pattern Recognition*, 45(4), 1318-1325 (2012) DOI: 10.1016/j.patcog.2011.09.021

49. A Falanga, MN Levine, R Consonni, G Gritti, F Delaini, E Oldani, JA Julian, T Barbui: The effect of very-low-dose warfarin on markers of hypercoagulation in metastatic breast cancer: results from a randomized trial. *Thromb Haemost*, 79(1):23-7 (1998)

50. P Marcato, CA Dean, CA Giacomantonio, PW Lee: Aldehyde dehydrogenase: its role as a cancer stem cell marker comes down to the specific isoform. *Cell Cycle*, 10(9):1378-84 (2011) DOI: 10.4161/cc.10.9.15486

51. W Shih, S Yamada: N-cadherin-mediated cell-cell adhesion promotes cell migration in a three-dimensional matrix. *J Cell Sci*, 125(Pt15):3661-70 (2012) DOI: 10.1242/jcs.103861

52. R Lamb, S Lehn, L Rogerson, RB Clarke, G Landberg: Cell cycle regulators cyclin D1 and CDK4/6 have estrogen receptor-dependent divergent functions in breast cancer migration and stem cell-like activity. *Cell Cycle*,12(15):2384-94 (2013) DOI: 10.4161/cc.25403

53. A Nagarajan, P Malvi, N Wajapeyee: Oncogene-Directed Alterations in Cancer Cell Metabolism. *Trends in Cancer*, *2*(7), 365-377 (2016) DOI: 10.1016/j.trecan.2016.06.002

54. F Di Virgilio: Purines, purinergic receptors, and cancer. *Cancer Res,*72(21):5441–7 (2012) DOI: 10.1158/0008-5472.CAN-12-1600

55. P Mehlen, C Delloye-Bourgeois, A Chédotal: Novel roles for Slits and netrins: axon guidance cues as anticancer targets? *Nat Rev Cancer*, 11(3):188–97 (2011) DOI: 10.1038/nrc3005

56. S Ramaswamy, P Tamayo, R Rifkin, S Mukherjee, CH Yeang, M Angelo, C Ladd, M Reich, E Latulippe, JP Mesirov, T Poggio, W Gerald, M Loda, ES Lander, TR Golub: Multiclass cancer diagnosis using tumor gene expression signatures. *Proc Natl Acad Sci U S A*, 98(26):15149-54 (2001) DOI: 10.1073/pnas.211566398

57. JC Ang, H Haron, HNA Hamed: Semi-supervised SVM-based feature selection for cancer classification using microarray gene expression data. In *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer International Publishing 468-477 (2015)

58. Z Cai, D Xu, Q Zhang, J Zhang, SM Ngai, J Shao: Classification of lung cancer using ensemble-based feature selection and machine learning methods. *Mol Biosyst*,11(3):791-800 (2015) DOI: 10.1039/C4MB00659C

59. N Bandyopadhyay, T Kahveci, S Goodison, Y Sun, S Ranka: Pathway-Based Feature Selection Algorithm for Cancer Microarray Data. *Adv Bioinformatics*. 2009:532989 (2009) DOI: 10.1155/2009/532989

60. W Engchuan, JH Chan: Pathway activity transformation for multi-class classification of lung cancer datasets. *Neurocomputing* 165: 81-89 (2015)
DOI: 10.1016/j.neucom.2014.08.096

61. W Liu, X Bai, Y Liu, W Wang, J Han, Q Wang, Y Xu, C Zhang, S Zhang, X Li, Z Ren, J Zhang, C Li: Topologically inferring pathway activity toward precise cancer classification via integrating genomic and metabolomic data: prostate cancer as a case. *Sci Rep*,5:13192 (2015)
DOI: 10.1038/srep13192

62. H Wang, H Zhang, Z Dai, MS Chen, Z Yuan: TSG: a new algorithm for binary and multi-class cancer classification and informative genes selection. *BMC Med Genomics*, 6 Suppl 1:S3 (2013)
DOI: 10.1186/1755-8794-6-S1-S3

63. HS Eo, JY Heo, Y Choi, Y Hwang, HS Choi: A pathway-based classification of breast cancer integrating data on differentially expressed genes, copy number variations and microRNA target genes. *Mol Cells*, 34(4):393-8 (2012)
DOI: 10.1007/s10059-012-0177-0

64. S Kim, M Kon, C DeLisi: Pathway-based classification of cancer subtypes. *Biol Direct*, 7:21 (2012)
DOI: 10.1186/1745-6150-7-21

65. M List, AC Hauschild, Q Tan, TA Kruse, J Mollenhauer, J Baumbach, R Batra: Classification of breast cancer subtypes by combining gene expression and DNA methylation data. *J Integr Bioinform,*11(2):236 (2014)
DOI: 10.2390/biecoll-jib-2014-236

Key Words: Cancer Subtypes Classification, Breast Cancer, BC, Pathway Enrichment; Differentially Expressed Genes, DEG, Review

Send correspondence to: Alex Graudenzi, Institute of Molecular Bioimaging and Physiology of the Italian National Research Council (IBFM-CNR), Milan, Italy, Tel: 390221717552, Fax: 390221717558, E-mail: alex.graudenzi@unimib.it