Identification of deregulated transcription factors involved in subtypes of cancers

Magali Champion, Julien Chiquet, Pierre Neuvial, Mohamed Elati, Etienne E. Birmelé

Identification of deregulated transcription factors involved in subtypes of cancers

Magali Champion[1], Julien Chiquet[2], Pierre Neuvial[3], Mohamed Elati[4] and Etienne Birmelé[1]

[1] Laboratoire MAP5, Université Paris Descartes, Sorbonne Paris Cité, France

[2] MIA Paris, AgroParisTech/INRA, France

[3] Institut de Mathématiques de Toulouse, Université Paul Sabatier/CNRS, France

[4] Institute of Systems and Synthetic Biology (ISSB), CNRS, Université d'Evry, France

Abstract

We propose a methodology for the identification of transcription factors involved in the deregulation of genes in tumoral cells. This strategy is based on the inference of a reference gene regulatory network that connects transcription factors to their downstream targets using gene expression data. The behavior of genes in tumor samples is then carefully compared to this network of reference to detect deregulated target genes. A linear model is finally used to measure the ability of each transcription factor to explain those deregulations.

We assess the performance of our method by numerical experiments on a breast cancer data set. We show that the information about deregulation is complementary to the expression data as the combination of the two improves the supervised classification performance of samples into cancer subtypes.

Keywords: Deregulation Gene regulatory networks Cancer systems biology

Background

Today, after decades of intensive research, cancer is still one of the most deadly diseases worldwide, killing millions of people every year. Cancer is mainly caused by mutations triggered by environmental factors (e.g. obesity, smoking, alcohol, lifestyle,...) often promoted by certain genetic configurations. To cure it, past research mostly focused on the effect of these environmental factors [1, 2] but, more recently, also on internal factors (e.g. inherited mutations, copy number alterations, hypo/hypermethylation, over/under-expression,...) [3, 4, 5]. In the last two decades, large-scale projects, such that the Cancer Genome Atlas project (TCGA), which has produced massive amounts of multi-omics data, have launched to improve our understanding of cancers [6, 7]. In this context, developing statistical algorithms able to interpret these large data sets and identify the genes that are the origin of diseases and their causal pathways still remains an important challenge.

Genes are commonly affected by genomic changes and deregulated in the pathogenesis of human cancer. Cancer is moreover an heterogeneous disease, with affected gene sets which may be highly different depending on specific subtypes, and thus different treatment of patients [8]. Specific analysis of subtypes have revealed significant differences between

breast cancer subgroups [9, 10] but also pancancer similarities between breast and bladder cancer subgroups [11].

Using transcriptional data allows to look beyond DNA, that is to study abnormalities in terms of gene expression. A common approach is to perform differential expression analysis, for which statistical procedures have been intensively explored, and to consider as deregulated the genes that are differentially expressed [12, 13]. This approach points to relevant genes but does not take into account the relationships (activation and inhibition) between genes, which we consider as crucial in the notion of deregulation.

Another approach, on which we will focus in the present paper, is to take into account the regulation structure between genes, and in particular the transcription factors (TFs), which have been the focus of many studies [14, 15]. Indeed, TFs play a preponderant role in the regulation of gene expression: by binding the promoter region of their target genes, TFs can activate or inhibit their expression, which make them an attractive target for cancer therapy [16]. Regulation processes between TFs and their targets are usually represented by Gene Regulatory Networks (GRNs), which give an overview of the mechanisms of cancer. In the last few years, many different approaches have been proposed to solve the GRN inference problem from collections of gene expression data. In a discrete framework, gene expressions are discretized depending on their status (under/over-expressed or normal) and truth tables provide the regulation structure [17]. In the continuous case, regression methods, including the popular Lasso [18] and its derivatives, have provided powerful results [19, 20, 21]. The notion of deregulated genes then corresponds to genes whose expression does not correspond to the expected level, given its regulator expressions. To unravel deregulated genes, a first possibility is to infer one network per condition and to compare them. Statistical difficulties due to the noisy nature of transcriptomic data and the large number of features compared to the sample size can be taken into account by inferring the networks jointly and penalizing the differences between them [22, 23]. A second possible approach is to assess the adequacy of gene expression in tumoral cell to a reference GRN, in order to exhibit the most striking discrepancies, i.e. the regulations which are not fulfilled by the data [24, 25, 26]. Such methods however focus on checking the validity of the network rather than highlighting genes with an abnormal behavior. Finally, analyses may be conducted at the pathway level rather than the gene elevel, using the SPIA [27] or PARADIGM [28] method. They are however not network-wide in the sense that each gene has a deregulation score by pathway it belongs to and pathways are treated independently. Moreover, as the pathways are extracted from curated databases, the regulations taken into account are not tissue-specific.

The main goal of this paper is to propose a statistical deregulation model that integrates gene expression data to identify deregulated TFs involved in specific subtypes of cancer. This article is structured as follows: in the first section, we present the 3-step method we developed and the method we used to validate it. Then, we illustrate its interest on a breast cancer data set and show that it improves the prediction of patients subgroup based on gene expression only. We finally discuss the obtained results.

Methods

Overview

Our approach for the identification of deregulated transcription factors (TFs) involved in specific subtypes of cancers is based on a 3-step strategy that (i) creates a gene regulatory network (GRN) of reference, which represents regulations between groups of co-regulated TFs and target genes based on gene expression data, (ii) computes a deregulation score for each target gene in each tumor sample by carefully comparing their behavior with the GRN of reference, (iii) identifies the most significant TFs involved in the deregulation of target genes in specific subtypes

of cancers. These steps are described in detail in the next paragraphs.

Step 1: Inferring a regulatory network

The first step of the algorithm consists in inferring a gene regulatory network that connects TFs to their downstream targets. Among the large number of existing methods, we chose hLICORN, available in the `CoRegNet` R-package [29]. This algorithm is based on a hybrid version of the LICORN model [30], in which groups of co-regulated TFs act together to regulate the expression of their targets. More precisely, LICORN uses heuristic techniques to identify co-activator and co-inhibitor sets from discretized gene expression matrices and locally associates each target gene to pairs of co-activators and co-inhibitors that significantly explain its discretized expression. The hybrid variation of LICORN then ranks these local candidate networks according to how well they predict the target gene expression, through a linear regression, and selects the GRN that minimizes the prediction error.

In this work, we slightly enriched the LICORN model by creating a copy of each TF in the target layer, such allowing to infer the regulation structure controlling a given TF. This view is clearly a simplification of reality as the regulator and regulated roles of a TF are considered independent. However, all GRN models are simplifications of the real biological mechanisms and may however allow to point out relevant TFs through well-chosen measures, as for instance the *influence* measure introduced in [29].

To construct a specific GRN, note that one may prefer using another inference method, such as the cooperative lasso [31], or some pre-existing regulatory network, which can be loaded from the RegNetwork database [32] for example. Here, we focused on hLICORN since the induced model was particularly suitable for the rest of our analysis. In addition, it was shown to provide powerful results for cooperative regulation detection, especially on cancer data set [29, 30].

Step 2: Computing a deregulation score

The second step of the algorithm aims at identifying deregulated target genes by carefully comparing their expression across all tumor samples with the GRN of reference inferred in Step 1. For this purpose, we used the method of [33], which assumes that all genes from a hLICORN model are allowed to be deregulated, i.e. not to respond to their regulators as expected.

More precisely, a binary deregulation variable, assumed to be non-zero with probability E, is introduced to compare the true status (under/over-expressed or normal) of each target gene in each sample with its expected value, resulting from the truth Table 1 [17] and the inferred GRN. To avoid discretization of the data, the status of all genes are considered as hidden variables. As the likelihood of the model is intractable due to the large number of hidden variables, the unknown model parameters (including the deregulation score E) are estimated using an EM-algorithm. The model is described in Figure 1.

Table 1: LICORN truth table, which gives the expected status of a target gene according to the collective status of its co-activators and co-inhibitors. The collective status are set by default to 0 except if and only if all of its elements share the same status. This table was established by biological considerations [30].

	Activator collective status		
Inhibitor collective status	-	0	+
-	0	+	+
0	-	0	+
+	-	-	-

Note that the deregulation score E does not capture information about differentially expressed genes but genes whose expression does not correspond to

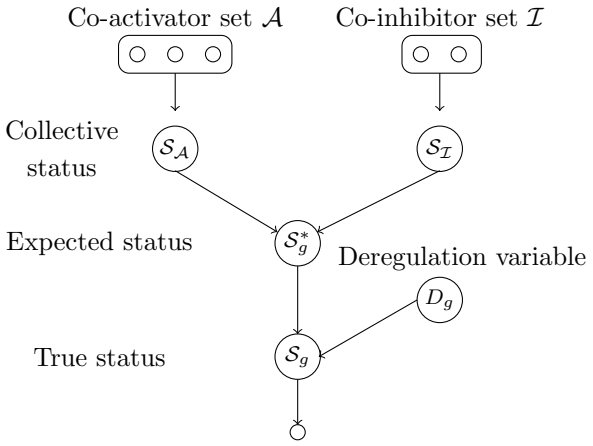

Figure 1: The model introduced in [33] and used to compute a gene deregulation score for each target gene in each sample: each gene g is associated to a hidden status $\mathcal{S}_g$ (under, over-expressed or normal). Target genes are allowed to be deregulated, i.e. not follow their co-regulator rules (see Table 1). The binary variable D_g indicates whether the corresponding target gene g is deregulated ($D_g = 1$) or not ($D_g = 0$). The deregulation score of gene g in sample j is then the probability, given the observation, that $D_g = 1$ in sample j.

the expected level given its regulator expression. Consider for instance a gene g regulated by a single TF a activating g. If both g and a have a fold change of 10, g is differentially expressed but is not deregulated as the regulation relationship is conserved. Conversely, if the respective fold-changes of g and a are of 1 and 10, g is not differentially expressed but is deregulated.

Step 3: Identifying TFs involved in the deregulation of the target genes

The main goal of the third step consists in identifying the TFs that significantly cause the deregulation of the target genes in cancers. Our approach is based on linear regression models, in which we try to explain the deregulation score of all target genes in one sample (Step 2) using their co-regulator TFs as explanatory variables (Step 1).

Assume that the total number of genes p is split into q TFs and $p-q$ target genes. Denote by Y_{ij} the deregulation score of target gene i ($1 \leq i \leq p-q$) in sample j ($1 \leq j \leq n$). In addition, denote by $G := (G_{i\ell})_{1\leq i\leq p-q, 1\leq \ell\leq q}$ the adjacency matrix of the GRN, whose non-zero elements encode the structure (edges) of the graph. We then cast our model as follows:

$$\forall j \in [\![1,n]\!], \forall i \in [\![1,p-q]\!], Y_{ij} = G_{i\ell} \cdot B_{\ell j} + \varepsilon_{ij}, \quad (1)$$

or, in a matrix form, $Y = G \cdot B + \varepsilon$, where each element $B_{\ell j}$ of matrix B, to estimate, measures the deregulation importance of TF ℓ in sample j and ε stands for the presence of noise in the model.

Solving the B-estimation problem (1) can be viewed as a classical multi-task linear learning problem, which is known to be particularly critical in the high-dimensional setting. Note however that we are far from such a case, the total number of observations, which corresponds to the number of target genes $p-q$, being extremely large compared to n (number of linear tasks) and of the same order as q (number of variables).

To estimate B, we used a constrained least squares estimation procedure. As we only expected to find TFs positively causing the deregulation of their targets in each sample, we considered the induced constrained optimization problem:

$$\forall j \in [\![1,n]\!], \quad \hat{B}_{\cdot j} := \underset{\beta\in\mathbb{R}^q}{\operatorname{argmin}} \|Y_{\cdot j} - G\beta\|_2^2, \quad (2)$$
$$s.t \quad \begin{cases} \forall \ell \in [\![1,q]\!], \beta_\ell \geq 0 \\ \sum_{\ell=1}^{q} \beta_q = 1 \end{cases}$$

where $\|.\|_2^2$ stands for the euclidian norm. The first constraint makes all coefficients of $\hat{B}$ positive, whereas the second constraint allows us to interpret $(\hat{B}_{\cdot j})_{1\leq j\leq n}$ as an influence deregulation score of TFs in each sample j. The closer to 1 $\hat{B}_{\ell j}$, the more important the role played by TF ℓ in the deregulation of its targets in sample j. To solve Equation (2), we used the lsei(.) function of the R-package `limSolve`.

Explaining molecular subtypes

Step 3 of the algorithm provides a new deregulation score $\hat{B}$, measuring how important each TF is in each sample to explain the deregulation of its targets.

To assess the information contained in $\hat{B}$, two approaches were considered. The first one quantifies its potential to predict the cancer subtype of a given sample, whichever TFs are used to do so. The second approach consists in looking for small sets of TFs whose deregulations characterize a given cancer subtype.

Classification methods for predicting subtypes

We first used a classification framework, considering a partition of the cancer samples into K known subtypes. More precisely, based on $\hat{B}$, we predicted the classification of the n samples into the K groups in a ten-fold stratified cross validation scheme. As a benchmark for classification methods, we chose the following ones:

- k-nearest neighbors [34]. The knn-algorithm is one of the simplest and fastest algorithm, which can be used both in a regression or classification setting, for a binary or multi-label classification [35] problem. It efficiently assigns a class label to the input pattern based on the class labels represented by its k closest neighbors.

- linear discriminant analysis [36]. LDA is a discriminant analysis and a dimensionality reduction technique, which aims at finding linear combination of features that separates multiple classes [37, 38]. It can be viewed as a supervised version of the Principal Component Analysis (PCA), the principal axes being defined in such a way to maximize the separation between classes.

- random forest [39]. RF is one of the most popular and powerful machine learning algorithms. It mainly consists in aggregating a multitude of decision trees: since decision trees are very sensitive to the specific data on which they are trained, Bootstrap Aggregation (or Bagging) [40] is used to reduce over-fitting by combining multiple predictions, which produces the so-called forest.

- support vector machine [41, 42]. The SVM algorithm is a powerful technique for classification. It looks for the optimal separating hyperplane between two or more classes by maximizing the margins between the classes' closest points.

All these algorithms are implemented in the R package `CMA` that we used for our analysis, which is dedicated to high-dimensional class prediction problems. All the hyperparameters of each method were tuned internally using a second ten-fold stratified cross validation loop and experiments were repeated 100 times each to evaluate accuracy.

Predictions of each test set of the 10-fold stratified cross-validation scheme were aggregated in a unique prediction vector, associating each sample j with label z_j to a unique class $\hat{z}_j$. Performance was then measured by computing the error of classification E_{cl}, defined as the proportion of missclassified samples:

$$E_{cl} := \frac{1}{n} \sum_{j=1}^{n} 1_{\hat{z}_j \neq z_j}.$$

We also compared the obtained classification error to that obtained by a baseline (dummy) classification rule consisting in randomly assigning each sample to a class, while conserving the class frequencies.

Sparse logistic regression for differentiating subtypes

The second approach aims at identifying the TFs whose deregulations are particularly suitable to explain a given subtype of cancer. To this end, we used logistic regression with lasso regularization to build a sparse model for the considered subtype based on $\hat{B}$. For a given subtype $k \in [\![1, K]\!]$, let us denote by $Z^{(k)} := (z_j^{(k)})_{1 \leq j \leq n}$ the vector of length n whose

components are defined as:

$$\forall 1 \leq j \leq n,\ z_j^{(k)} = \begin{cases} 1 & \text{if sample } j \text{ belongs to class } k, \\ 0 & \text{otherwise.} \end{cases}$$

The sparse logistic model we cast is then obtained by minimizing the penalized negative binomial log-likelihood:

$$\min_{(\beta_0,\beta)\in\mathbb{R}^{q+1}} -\frac{1}{n}\sum_{j=1}^{n}\left\{z_j^{(k)}\cdot\left(\beta_0+\hat{B}_{.j}^T\beta\right) - \log\left(1+e^{\left(\beta_0+\hat{B}_{.j}^T\beta\right)}\right)\right\} + \lambda\|\beta\|_1, \quad (3)$$

where, for any $\beta \in \mathbb{R}^q$, we denote by $\|\beta\|_1 := \sum_{\ell=1}^{q}|\beta_\ell|$ the ℓ_1-norm and λ the penalization parameter, which controls the amount of sparsity in β. The scalar value β_0 corresponds to the intercept, i.e. the constant term in the logistic model (3), whereas the vector β captures subsets of TFs that are sufficient to explain their joint effect on the type of developed cancer.

The prediction performance was again computed using the missclassfication error rate in a 10-fold internal cross-validation loop. To identify the most important TFs, we also ranked them based on the number of times they were used in the logistic model (3) for each subtype, experiments being repeated 100 times.

The breast cancer data set

We applied the method we developed on real data, which were produced in the framework of the Cancer Genome Atlas (TCGA) project and available at the Genomic Data Commons Data Portal [43]. These data include a set of 804 breast cancer samples with gene expression data for a total number of 16,020 genes, split into 1,607 TFs and 14,413 targets. Gene expression data were produced using RNA-sequencing on breast cancer tissues. Preprocessing was done by log-transformation and quantile-normalization of the arrays. TCGA samples were analyzed in batches and significant batch effects were observed based on a one-way analysis of variance in most data modes. We applied Combat to adjust for these effects [44].

Previous multi-omics analysis from the TCGA data portal [45] led to the identification of five main breast cancer classes:

- cluster I ("luminal A"), the most common breast cancer subtype, enriched in hormone-receptor positive tumors with negative HER2 and low Ki67 (proliferating cell nuclear antigen) and is associated with good prognosis,
- cluster II ("luminal B"), similar to "luminal A" but with high levels of Ki67, a more aggressive phenotype and a slightly worse prognosis,
- cluster III ("basal-like"), also referred to as triple-negative, corresponding to negative hormone-receptor and HER2 negative,
- cluster IV ("HER2-positive"), characterized by high expression of HER2 and other genes associated with the HER2 pathway, high proliferation and more aggressive biological and clinical behavior.
- cluster V ("normal-like"), a particular luminal A breast cancer subtype, whose biological functions are not clearly established but which presents similar expression patterns to normal breast tissue.

The classification we used was performed using PAM50 [46], a 50 gene expression assay based on microarray and quantitative real time that was developed by analyzing a set of 189 breast tumor samples to separate them into these five molecular subtypes.

For a brief summary of the whole data set we used, refer to Tables 2a and 2b. Note that among the 16,020 available genes, we only kept the top 75% varying genes.

Table 2: Overview of the gene expression data set we used for our analysis, provided by the Cancer Genome Atlas project from breast tumor tissus.
(a) Number of samples and genes (target genes and TFs)

Samples	TFs	Target genes
804	1,607	14,413

(b) Molecular subtypes distribution of the 804 tumor samples.

Luminal A	Luminal B	Basal	HER2	Normal
406	175	133	65	25

Results and discussion

Description of the results

To validate our method, we had to provide a tissue-specific reference gene regulatory network (Step 1). We had thus to choose a cluster of samples on which no deregulation score can be computed. In many cancers, the pure normal tissue of origin is not available. Even if this choice is arbitrary from a statistical point of view, we used as a reference the "normal-like" breast tumor subtype, whose expression is the closest to normal tissue.

The inferred regulatory network will thus reflect averaged relationships between genes for normal-like patients, and deregulations will point deviations between other sample cluster and the "normal-like" cluster.

After calibrating the internal parameters of the hLicorn algorithm, the co-regulatory network we inferred is made of a total number of 74,557 edges connecting 1,182 TFs to 7,780 of their targets. Among these 7,780 targets, 839 are TFs and also regulate other genes. This network is relatively sparse, each of the target genes being associated with an averaged number of 7.3 activators and 4.9 inhibitors. The remaining genes, which are not connected to any other genes, were removed from the rest of the analysis. As it was learnt in a high-dimensional setting, this network is probably still quite noisy but, as demonstrated by [29], it may capture true biological information.

We then ran the EM procedure (Step 2) on the remaining gene expression data matrix to compute a deregulation score of each of the 7,780 target genes in each of the 779 samples. From now on, all samples were thus treated individually, the results reflecting how genes behaved in each sample. Analyzing the score deregulation matrix, we found five genes, BANK1, ESCO2, FAT, ZNF488 and OST4, deregulated with a probability larger than 0.5 in more than the 10% of the samples. These genes are frequently deregulated when compared to the distribution of the deregulation scores, as can be seen in Figure 2.

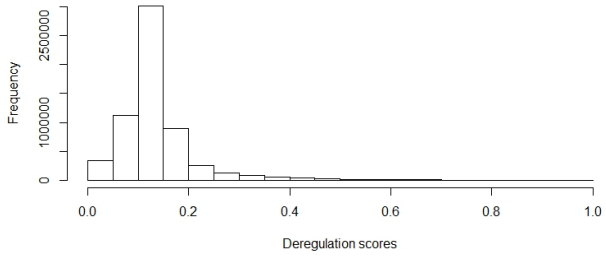

Figure 2: Histogram representing the deregulation scores distribution across all samples and all target genes.

We finally applied the last part of our procedure to identify TFs involved in the deregulation of the target genes, that is having a non-null coefficient in the $\hat{B}$ matrix.

For each of the 779 tumor samples, an average of 19.1 TFs is used to explain the deregulation score of the target genes. Conversely, a TF explains the

deregulation scores of an averaged number of 12.6 samples. Interestingly, among the 1,182 TFs, 51 play a significant deregulation role in more than the 10% of the samples. The first column of Table 3 shows the top 10 list of those TFs. In addition, Figure 3 proposes a visualization in a heatmap form of a submatrix of $\hat{B}$ corresponding to those TFs. It illustrates that $\hat{B}$ is sparse and that the TF deregulation score determination needs no regularization to control sparsity.

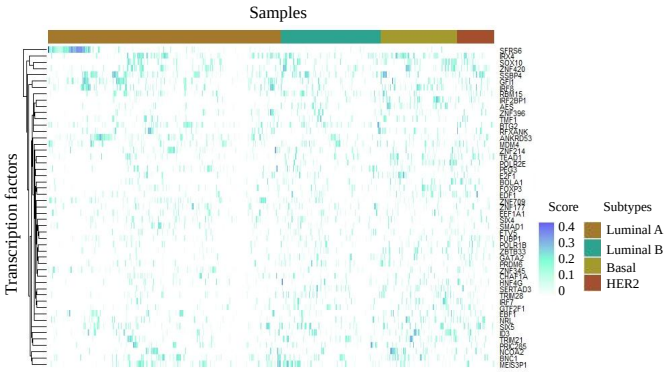

Figure 3: Heatmap representing the involvment of the 51 most important transcription factors in the deregulation of their target genes in the four subtypes.

Among the TFs having a non-null deregulation score in numerous samples (Table 3), we found IRX4, which is implicated in the pathogenesis of prostate cancer, and has been shown to act as a tumor suppressor gene for it [47]. Interestingly, we also found BNC1, which is frequently methylated and silenced in breast brain metastasis (BBM), a spread and advanced breast cancer type [48].

We similarly ranked the TFs according to the number of times they were involved in the deregulation of their targets in each subtype separately. The results are shown in Table 3 (columns 2 to 5). Some of the highly represented TFs are common to all subtypes (IRX4, SSBP4, ZNF420, ANKRD53, BNC1 and MEIS3P1). These TFs may be involved in all of them. However, the rankings are different between types, and some TFs appear only in some subtypes, suggesting that some of them can be characteristic of given subtypes. For both luminal subtypes, we found for instance BTG2, an antiproliferative gene associated with survival in breast cancer [49] and highly sensitive to estrogene response, a marker of luminal breast cancer types [50]. In the basal-like subtype, we retrieved ID3, a transcriptional inhibitor of differentiation, which has been shown to mediate lung metastatic invasion in this particular subgroup of breast tumors [51].

Classification results

To validate further the information contained in the inferred score matrix $\hat{B}$, we checked how well it predicted the four TCGA subtypes for four different classification methods. Prediction results can be found in Table 4.

The missclassification rates are comparable for all tested methods, at the order of 0.4, which means that around 470 of the 779 samples are classified into the subtype they belong. They have to be compared to the 0.64 averaged error rate, obtained by performing a million of random stratified classifications. This highlights the fact that the matrix $\hat{B}$ contains some information about the subtypes of the considered samples.

Table 4: Averaged missclassification error for the prediction of the four breast cancer subtypes in a ten-fold cross validation scheme for four different methods (k-nearest neighbors - knn -, Linear Discriminant Analysis - LDA -, Random Forests - RF - and Support Vector Machines - SVM). Experiments were repeated 100 times each.

Method	knn	LDA	RF	SVM
Error	0.409	0.376	0.396	0.388

However, gene expression is known to contain such

Table 3: List of the most important TFs for explaining the deregulation scores of their downstream targets in each subtype and percentage of samples of each subtype in which they are involved (only the ten top genes are shown).

Subtypes									
All		Luminal A		Luminal B		Basal		HER2	
TF	%	TF	%	TF	%	TF	%	TF	%
IRX4	32	SSBP4	33	BTG2	31	IRX4	44	ZNF420	40
SSBP4	30	ANKRD53	32	SSBP4	30	AES	38	FUBP1	35
ZNF420	26	IRX4	30	ZNF420	29	ID3	37	IRX4	33
ANKRD53	23	GFI1	27	IRX4	28	POLR2E	36	SSBP4	31
BNC1	23	BTG2	27	RBM15	28	SOX10	35	TEAD1	28
GFI1	23	SFRS6	27	MEIS3P1	27	BNC1	32	MLL3	26
BTG2	22	MEIS3P1	24	GTF2F1	24	ZNF420	29	SIX5	25
MEIS3P1	22	MDM4	23	TRIM21	23	TRIM21	29	BOLA1	25
SOX10	22	BNC1	23	MDM4	22	IRF8	29	POLR1B	25
MDM4	22	ZNF709	22	SOX10	22	IRF2BP1	29	SOX10	23

information. The question is thus to determine if the information contained in $\hat{B}$ is redundant with direct gene expression comparison or if the notions of deregulations and differential expression are complementary. To do so, we compared the subtypes prediction performance obtained using the inferred score matrix $\hat{B}$ with the one obtained using the gene expression data matrix only and the one obtained using both types of data. The methods and procedure used were the same in the three cases.

Results, shown in Figure 4, are clearly better for expression data, which is not surprising as the PAM50 classification was built on expression and PCR data [46]. However, the combination of expression data and deregulation scores leads to an improved classification, showing that the deregulation scores capture some information that is missing in expression data.

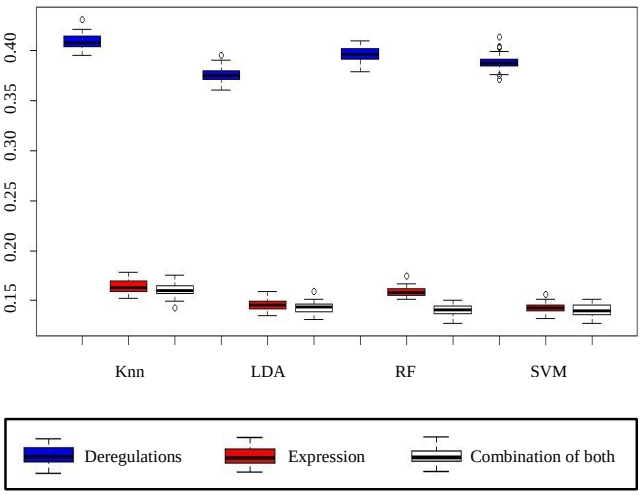

Figure 4: Comparison of the prediction results for the four considered classification methods (k-nearest neighbors - knn -, Linear Discriminant Analysis - LDA -, Random Forests - RF - and Support Vector Machines - SVM) using the $\hat{B}$ TFs deregulation score matrix only, the gene expression data matrix only and a combination of both data.

Subtype characterization

The last validation of $\hat{B}$ consists in determining if small sets of TFs can be used to characterize a given

subtype, by applying the sparse logistic regression procedure. Contrary to the former section, the aim is not to solve the classification procedure into four subtypes but, for a given subtype, to find a small number of TFs characterizing it.

As shown in Table 5, all subtypes are accurately predicted with a missclassification error rate of at most 0.267. All the classification results are again better than a million of random clasifications. Note that HER2-positive subgroup is almost perfectly predicted with less than 10% of missclassified samples, and using a small number of TFs in average. The other subtypes are more difficult to characterize through TF deregulation scores, and the number of selected TFs for the predictions is larger, especially for Luminal A. This suggests a greater heterogeneity of those subtypes in terms of deregulation profiles. Table 6 lists, for each subtype, the TFs that were selected in more than 20% of the folds.

Table 5: Missclassification error rate and averaged number of TFs used for predicting each subtype separately based on a lasso regularized logistic model. The last column indicates the averaged missclassification error rate obtained by performing a million of random stratified classifications.

Subtype	Error	Number of TFs	Random error
Luminal A	0.267	46.02	0.49
Luminal B	0.225	2.54	0.35
Basal	0.141	27.95	0.28
HER2	0.083	1.61	0.15

Discussion

With the aim of understanding the deregulation processes in tumoral cells, we developed a three-step strategy that measures the influence of each transcription factor in the deregulation of genes in each tumor sample. While hardly but significantly predicting existing molecular subgroups of cancer, it can be used to accurately retrieve these subgroups when combined with gene expression data. One has to note that this is not specific to breast cancer, as can be seen in Figure 5, which was produced by performing the same experiment on a TCGA bladder cancer data set [7] with the same conclusion.

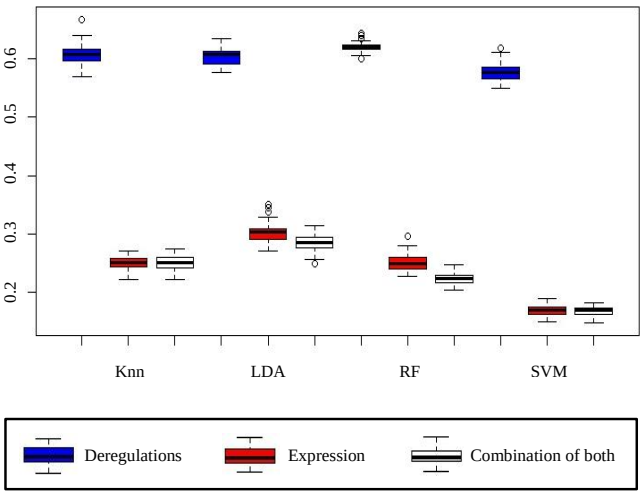

Figure 5: Bladder cancer analysis: comparison of the prediction results for the four considered classification methods (k-nearest neighbors - knn -, Linear Discriminant Analysis - LDA -, Random Forests - RF - and Support Vector Machines - SVM) using the $\hat{B}$ TFs deregulation score matrix only, the gene expression data matrix only and a combination of both data.

Lists of transcription factors characterizing a given subtype can moreover be established. A validation step of such lists has to be done in future work.

An open question which has also to be tackled is to determine to which extent the information carried by the deregulation is different or redundant with Copy Number Varation, methylation or mutation data, all of which being phenomena that may imply deregulation.

To go further, the question of predicting clinical data, such as survival, grade or stage of cancers, would be of particular interest to understand better the evolution of such a complex disease and to

Table 6: List of TFs ranked according to the number of times they were used to predict each subtype (indicated in %) based on a lasso regularized logistic model in a 10-fold cross validation loop (experiments were repeated 100 times). Only TFs that are selected more than the 40% of the times are shown.

Subtype	TFs (%)				
Luminal A	AES (100%)	ANKRD53 (100%)	ETV2 (100%)	IRF2BP1 (100%)	POLR2E (100%)
	SERTAD3 (100%)	SFRS6 (100%)	TBX21 (100%)	ZNF395 (100%)	EDF1 (99.8%)
	IRF7 (99.8%)	ZZZ3 (99.8%)	BOLA1 (99.6%)	MYST2 (99.6%)	GATA2 (99.4%)
	ZNF709 (99.0%)	NRL (98.8%)	TEAD1 (98.8%)	POLR1B (97.6%)	ZBTB33 (97.0%)
	AFF4 (93.0%)	TAF13 (98.8%)	GTF2F1 (90.2%)	IRF3 (88.6%)	SMARCC1 (88.0%)
	GTF3C4 (83.6%)	E2F1 (80.2%)	ZKSCAN1 (79.8%)	ZNF217 (75.4%)	ZNF567 (75.0%)
	MLLT10 (72.8%)	HDAC5 (71.8%)	ZNF3 (67.8%)	ZNF358 (67.2%)	SIRT1 (66.6%)
	SMAD7 (64.0%)	MNF46 (59.2%)	ZFP36 (53.6%)	POU2AF1 (50.0%)	ZNF32 (48.8%)
	SOX11 (44.6%)	CENPB (42.2%)			
Luminal B	DIP2C (59.2%)	SFRS6 (53.0%)			
Basal	AES (64.4%)	AFF4 (64.4%)	BTG2 (64.4%)	ETV2 (64.4%)	IRF2BP1 (64.4%)
	NCOA2 (64.4%)	POLR2E (64.4%)	ZBTB33 (64.4%)	ZMYND11 (64.4%)	ID3 (64.2%)
	NRL (64.0%)	ZZ3 (63.8%)	ALS2CR8 (63.6%)	IRX4 (63.6%)	SMARCC1 (62.8%)
	ZNF192 (60.0%)	NHLH1 (59.6%)	KLF11 (59.2%)	SOX10 (59.0%)	FOXP3 (57.4%)
	BOLA1 (56.0%)	ZNF358 (54.8%)	TFDP1 (53.8%)	IRF7 (51.0%)	TAX1BP3 (50.4%)
	ETS1 (50.2%)	SERTAD3 (49.2%)	EDF1 (47.8%)	E2F1 (47.4%)	CTNNB1 (46.0%)
	SMAD4 (46.0%)	IVNS1ABP (42.2%)	TRIM21 (42.8%)	ZNF446 (41.6%)	ZNF641 (40.8%)
	NFE2 (40.0%)				
HER2	FUBP1 (43.4%)				

use the deregulation in machine learning methods for personalized medicine [52]. We collected cancer stages for the 779 breast cancer samples of our analysis and obtained significant prediction results but that were not as good as for the subtype prediction. Indeed, the error rates were around 0.45, for all tested methods (knn, LDA, SVM and RF) and all data set (deregulation score matrix $\hat{B}$, gene expression and combination of both). This task is however known to be difficult due to the heterogeneous nature and quality of data, which may lead to limited accuracy [53] and unstable biomarker selection. Current investigations thus focus more on predictive models, which aim at predicting the benefit in outcome of a treatment on an individual, rather than prognostic models, which aim at directly predicting an outcome of a disease on an untreated individual.

References

[1] Tuyns, A.J.: Epidemiology of alcohol and cancer. Cancer Research **39**, 2840–3 (1979)

[2] Doll, R., Peto, R.: The causes of cancer: quantitative estimates of avoidable risks of cancer in the united states today. Journal of National Cancer Institute **66**(6), 1191–308 (1981)

[3] Perou, C.M.: Molecular portraits of human breast tumors. Nature **406**(6797), 747–752 (2000)

[4] Shlien, A., Malkin, D.: Copy number variations and cancer. Genome Medicine **1**(6), 62 (2009)

[5] Kulis, M., Esteller, M.: Dna methylation and cancer. Advances in Genetics **70**, 27–56 (2010)

[6] The Cancer Genome Atlas: Comprehensive molecular profiling of lung adenocarcinoma. Nature **511**(7511), 543–550 (2014)

[7] The Cancer Genome Atlas: Comprehensive molecular characterization of urothelial bladder carcinoma. Nature **507**(7492), 315–322 (2014)

[8] Liu, Z., Zhang, X.S., Zhang, S.: Breast tumor subroups reveal diverse clinical prognostic power. Scientific Reports **4**(4002), 10–103804002 (2014)

[9] Lehman, B.D., Bauer, J.A., Chen, X., Sanders, M.E., Chakravarthy, A.B., Shyr, Y., Pietenpol, J.A.: Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. The Journal of Clinical Investigation **121**(7), 2750–2767 (2011)

[10] Chen, X., Li, J., Gray, W.H., Lehman, B.D., Bauer, J.A., Shyr, Y., Pietenpol, J.A.: Tnbctype: A subtyping tool for triple-negative breast cancer. Cancer Informatics **11**, 147–156 (2012)

[11] Damrauer, J.S., Hoadley, K.A., Chism, D.D., Fan, C., Tiganelli, C.J., Wobker, S.E., Yeh, J.J., Milowsky, M.I., Iyer, G., Parker, J.S., Kim, W.Y.: Intrinsic subtypes of high-grade bladder cancer reflect the hallmarks of breast cancer biology. Proccedings of the National Academy of Sciences **111**(8), 3110–3115 (2014)

[12] Hu, M.L., Yeh, K.T., Lin, P.M., Hsu, C.M., Hsiao, H.H., Liu, Y.C., Lin, H.Y., Lin, S.F., Yang, M.Y.: Deregulated expression of circadian clock genes in gastric cancer. BMC Gastroenterology, 14–67 (2014)

[13] Kaczkowski, B., Tanaka, Y., Kawaji, H., Sandelin, A., Andersson, R., Itoh, M., Lassmann, T., Hayashizaki, Y., Carninci, P., Forrest, A.R.R., the FANTOM5 consortium: Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan-cancer biomarkers. Cancer Research **76**(2), 216–226 (2016)

[14] Nebert, D.W.: Transcription factors and cancer: an overview. Toxicology **181-182**, 131–141 (2002)

[15] Bhagwat, A.S., Vakoc, C.R.: Targeting transciption factors in cancer. Trends in Cancer **1**(1), 53–65 (2015)

[16] Yeh, J.E., Toniolo, P.A., Frank, D.A.: Targeting transcription factors: promising new strategies for cancer therapy. Current Opinion in Oncology **25**(6), 652–658 (2013)

[17] Elati, M., Rouveirol, C.: Unsepervised Learning for Gene Regulation Network Inference from Expression Data: a Review. John Wiley and Sons, Inc., Hoboken, NJ (2011)

[18] Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B **58**(1), 267–288 (1996)

[19] Liu, B., de la Fuente, A., Hoeschele, I.: Gene network inference via structural equation modeling in genetical genomics experiments. Genetics **178**(3), 1763–1776 (2008)

[20] Vignes, M., Vandel, J., Allouche, D., Ramadan-Alban, N., Cierco-Ayrolles, C., Schiex, T., Mangin, B., de Givry, S.: Gene regulatory network reconstruction using bayesian networks, the dantzig selector, the lasso and their meta-analysis. PloS one **6**(12), 29165 (2011)

[21] Haury, A.C., Mordelet, F., Vera-Licona, P., Vert, J.P.: Tigress: Trustful inference of gene regulation using stability selection. BMC Systems Biology **6**(1), 145 (2012)

[22] Kojima, K., Imoto, S., Yamaguchi, R., Fujita, A., Yamauchi, M., Gotoh, N., Miyano, S.: Identifying regulational alterations in gene

regulatory networks by state space representation of vector autoregressive models and variational annealing. BMC Genomics **13**(Suppl 1), 6 (2012)

[23] Chiquet, J., Grandvalet, Y., Ambroise, C.: Inferring multiple graphical structures. Statistics and Computing **21**(4), 537–553 (2011)

[24] Karlebach, G., Shamir, R.: Constructing logical models of gene regulatory networks by integrating transcription factor-dna interactions with expression data: An entropy-based approach. Journal of Computational Biology **19**(1), 30–41 (2012)

[25] Guziolowski, C., Bourde, A., Moreews, F., Siegel, A.: Bioquali cytoscape plugin: analysing the global consistency of regulatory networks. BMC Genomics **10**(1), 244 (2009)

[26] Samaga, R., Klamt, S.: Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. Cell Communication and Signaling **11**(1), 43 (2013)

[27] Tarca, A.L., Draghici, S., Khatri, P., Hassan, S.S., Mittal, P., Kim, J.S., Kim, C.J., Kusanovic, J.P., Romero, R.: A novel signaling pathway impact analysis. Bioinformatics **25**(1), 75–82 (2009)

[28] Vaske, C.J., Benz, S.C., Sanborn, J.Z., Earl, D., Szeto, C., Zhu, J., Haussler, D., Stuart, J.M.: Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using paradigm. Bioinformatics **26**, 237–245 (2010)

[29] Nicolle, R., Radvanyi, F., Elati, M.: Coregnet: reconstruction and integrated analysis of co-regulatory networks. Bioinformatics **31**(18), 3066–3068 (2015)

[30] Elati, M., Neuvial, P., Bolotin-Fukuhara, M., Barillot, E., Radvanyi, F., Rouveirol, C.: Licorn: learning cooperative regulation networks from gene expression data. Systems Biology **23**(18), 2407–2414 (2007)

[31] Chiquet, J., Grandvalet, Y., Charbonnier, C.: Sparsity in sign-coherent groups of variables via the cooperative-lasso. The Annals of Applied Statistics **6**, 795–830 (2012)

[32] Liu, Z., Canglin, W., Miao, H., Wu, H.: Regnetwork: an integrating database of transcriptional and post-transcriptional regulatory networks in human and mouse. Database: The Journal of Biological Databases and Curation. **2015**, 095 (2015)

[33] Picchetti, T., Chiquet, J., Elati, M., Neuvial, P., Nicolle, R., Birmelé, E.: A model for gene deregulation detection using expression data. BMC Systems Biology **9**, 6 (2015)

[34] Ripley, B.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge, MA (1996)

[35] Zhang, M.L., Zhou, Z.H.: A k-nearest neighbor based algorithm for multi-label classification. In: IEEE International Conference on Granular Computing, vol. 2, pp. 718–721 (2005)

[36] McLachlan, G.J.: Discriminant Analysis and Statistical Pattern Recognition. Wiley, Newark, NJ (2005)

[37] Fisher, R.A.: The use of multiple measurements in taxonomic problems. Annals of Human Genetics **7**(2), 179–188 (1936)

[38] Rao, C.R.: The utilization of multiple measurements in problems of biological classification. Journal of the Royal Statistical Society. Series B **10**(2), 159–203 (1948)

[39] Breiman, L.: Random forests. Machine Learning **45**(1), 5–32 (2001)

[40] Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth Statistics/Probability Series, p. 358. Wadsworth Advanced Books and Software, Belmont, CA (1984)

[41] Cortes, C., Vapnik, V.: Support-vector network. Machine Learning **20**, 1–25 (1995)

[42] Scholkopf, B., Smola, A., Williamson, R.C., Barlette, P.: New support vector algorithms. Neural Computation **12**, 1207–1245 (2000)

[43] The Genomic Data Commons Data Portal. https://portal.gdc.cancer.gov/

[44] Johnson, W.E., Li, C., Rabinovic, A.: Adjusting batch effects in microarray expression data using empirical bayes methods. Biostatistics **8**, 118–127 (2007)

[45] The Cancer Genome Atlas: Comprehensive molecular portraits of human breast tumours. Nature **490**, 61–70 (2012)

[46] Parker, J.S. and Mullins, M. and Cheang, M.C.U. and Leung, S. and Voduc, D. and Vickery, T. and Davies, S. and Fauron, C. and He, X. and Hu, Z. and Quackenbush, J.F. and Stijleman, I.J. and Palazzo, J. and Marron, J.S. and Nobel, A.B. and Mardis, E. and Nielsen, T.O. and Ellis, M.J. and Perou, C.M. and Bernard, P.S.: Supervised risk predictor of breast cancer based on intrinsic subtypes. Journal of Clinical Oncology **27**(8), 1160–1167 (2009)

[47] Nguyen, H.H., Takata, R., Akamatsu, S., Shigemizu, D., Tsunoda, T., Furihata, A., Kubo, M., Kamatani, N., Ogawa, O., Fujioka, T., Nakamura, Y., Nakagawa, H.: Irx4 at 5p15 suppresses prostate cancer growth through the interaction with vitamin d receptor, conferring prostate cancer susceptibility. Human Molecular Genetics **21**(9), 2076–2085 (2012)

[48] Pangeni, R.P., Channathodiyil, P., Huen, D.S., Eagles, L.W., Johal, B.K., Pasha, D., Hadjistephanou, N., Nevell, O., Davies, C.L., Adewumi, A.I., Khanom, H., Samra, I.S., Buzatto, V.C., Chandrasekaran, P., Shinawi, T., Dawson, T.P., Ashton, K.M., Davis, C., Brodbelt, A.R., Jenkinson, M.D., Bieche, I., Darling, J.L., Warr, T.J., Morris, M.M.: The galnt9, bnc1 and ccdc8 genes are frequently epigenitically dysregulated in breast tumours that metastasise to the brain. Clinical Epigenetics **7**, 57 (2015)

[49] van de Vivjer, M.J., He, Y.D., van't Veer, L.J., Dai, H., Hart, A.A.M., Voskuil, D.W., Schreiber, G.J., Peterse, J.L., RObberts, C., Marton, M.J., Parrish, M., Atsma, D., Witteveen, A., Glas, A., Delahaye, L., van der Velde, T., Bartelink, H., Rodenhuis, S., Rutgers, E.T., Friend, S.H., Bernards, R.: A gene-expression signature as a predictor of survival in breast cancer. The New England Journal of Medicine **347**, 1999–2009 (2002)

[50] Sotiriou, C., Neo, S.Y., McShanes, L.M., Korn, E.L., Long, P.M., Jazaeri, A., Martiat, P., Fox, S.B., Harris, A.L., Liu, E.T.: Breast cancer classification and prognosis based on gene expression profiles from a population-based study. Proceedings of the National Academy of Sciecnes of the United States of America **100**(18), 10393–10398 (2002)

[51] Gupta, G.P., Perk, J., Acharyya, S., de Candia, P., Mittal, V., Todorova-Manova, K., Gerald, W.L., Brogi, E., Benezra, R., Massagué, J.: Id genes mediate tumor reinitiation during breast cancer lung metastasis. Proceedings of the National Aademy of Sciences in the United States of America **104**(49), 19506–19511 (2007)

[52] Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V., Fotiadis, D.I.: Machine learning applications in cancer prognosis and prediction. Computational and Structural Biotechnology Journal **13**, 8–17 (2015)

[53] Bilal, E., Dutkowski, J., Guinney, J., Jang, I.S., Logsdon, B.A., Pandey, G., Sauerwine, B.A., Shimoni, Y., Moen Vollan, H.K., Mecham, B.H., Rueda, O.M., Tost, J., Curtis, C., Alvarez, M.J., Kristensen, V.N., Aparicio, S., Borresen-Dale, A.L., Cladas, C., Califano, A., Friend, S.H., Ideker, T., Schadt, E.E., Stolovitzky, G.A., Margolin, A.A.: Improving breast cancer survival analysis through competition-based multidimensional modeling. PLoS Computational Biology **9**(5), 1003047 (2013)