Identification of Deregulated Mechanisms Specific to Bladder Cancer Subtypes

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

Identification of deregulation mechanisms specific to cancer subtypes

Magali CHAMPION

Université de Paris, CNRS, MAP5 UMR8145, Paris, France
magali.champion@parisdescartes.fr

Julien CHIQUET

Université Paris Saclay, AgroParisTech, INRAE, UMR MIA-Paris, Paris, France
julien.chiquet@inrae.fr

Pierre NEUVIAL

Institut de Mathématiques de Toulouse, UMR 5219, Université de Toulouse, CNRS, France
pierre.neuvial@math.univ-toulouse.fr

Mohamed ELATI

CANTHER, University of Lille, CNRS UMR 1277, Inserm U9020, 59045 Lille cedex, France
mohamed.elati@univ-lille.fr

Franọis RADVANYI

Institut Curie, PSL Research University, CNRS, UMR144, Paris, France
francois.radvanyi@curie.fr

Etienne BIRMELE

Université de Paris, CNRS, MAP5 UMR8145, Paris, France
Institut de Recherche Mathématique Avancée, UMR 7501 Université de Strasbourg, CNRS, Strasbourg, France
etienne.birmele@parisdescartes.fr

In many cancers, mechanisms of gene regulation can be severely altered. Identification of deregulated genes, which do not follow the regulation processes that exist between transcription factors and their target genes, is of importance to better understand the development of the disease. We propose a methodology to detect deregulation mechanisms with a particular focus on cancer subtypes. This strategy is based on the comparison between tumoral and healthy cells. First, we use gene expression data from healthy cells to infer a reference gene regulatory network. Then, we compare it with gene expression levels in tumor samples to detect deregulated target genes. We finally measure the ability of each transcription factor to explain these deregulations. We apply our method on a public bladder cancer data set derived from The Cancer Genome Atlas project and confirm that it captures hallmarks of cancer subtypes. We also show that it enables the discovery of new potential biomarkers.

Keywords: Cancer systems biology; deregulations; gene regulatory network.

1. Introduction

Cancer is defined by an uncontrolled proliferation of malignant cells, which can invade normal tissues and spread throughout the body. The causes and contribution factors of cancer are multiple and complex but they mostly derive from genetic alterations that accumulate over time. When combined with external factors (from lifestyle factors, *e.g.* nutrition, smoking, alcohol, to environmental exposures, *e.g.* UV-radiations, pesticides), the impact is even more dramatic. After decades of intensive research, a large number of works still focus on understanding the fundamental mechanisms of cancer. In this context, The Cancer Genome Atlas project (TCGA), the International Cancer Genome Consortium (ICGC) and other cancer genome projects have produced massive amounts of multi-omics data.[1] Due to the very high heterogeneity of cancer, a particular effort has also been made on subtyping cancers to improve the treatment of patients, as it has been done to reveal significant differences between breast cancer subtypes.[2]

A large number of genes, implicated in diverse biological processes, are involved in cancer. Many of them are altered by somatic mutations, copy number changes (amplifications/deletions) or DNA methylation.[3,4] At the gene expression level, these abnormalities can explain disregulated expressions (over/under-expression) and can contribute to tumor growth. A common approach for identifying dysfonctional genes in cancer consists in performing differential expression analysis,[5,6] for which statistical procedures have been intensively explored. However, this approach does not take into account regulations between genes. As an alternative, we here focus on deregulated genes,[7] for which expression changes can be observed in the Gene Regulatory Network (GRN) of tumoral cells.

GRNs are usually used to represent activation and inhibition relationships between genes, connecting transcription factors (TFs) with their targets. In the last few years, many different methods have been proposed to infer GRNs from collections of gene expression data. In a discrete framework, gene expression can be discretized depending on the status of the genes (under/over-expressed or normal) and truth tables provide the regulation structure.[8] In the continuous framework, regression methods, including the popular Lasso[9] and its derivatives, have provided powerful results.[10,11] To discover deregulated genes, a first solution consists in inferring GRNs from different tissues and comparing them after penalizing their differences to reduce the noise effect on the data.[12] A second solution we focus on is to evaluate how far gene expression in tumoral cells is with regards to a reference GRN to define a deregulation score for each target gene in each sample.[7] However, the regulation structure between TFs, which are key genes for cancer therapy,[13] and their targets is still not deeply exploited.

In this work, we propose a statistical deregulation model that uses gene expression data to identify deregulation mechanisms involved in specific cancer subtypes. This paper is an extended version of Ref. 14 and is organized as follows: in Section 2, we present the 3-step algorithm we developped and explain how to use it in practice.

In Section 3, we illustrate its performance advantage on a bladder cancer data set from the TCGA. We show that it is complementary to state-of-the-art methods to point to potential biomarkers of cancer subtypes.

2. Methods

2.1. *Deregulation model*

In this work, we aim at identifying deregulation mechanisms specific to cancer subtypes. A deregulated gene is defined as a gene whose expression does not correspond to the expression expected from its regulators. The deregulation model we focus on is derived from Ref. 7 and recalled in the next paragraph.

Regulation processes between TFs and their targets are usually represented by GRNs. Here, we assume that groups of co-regulated TFs act together to regulate (activate or inhibit) the expression of their targets, following the LICORN (*LearnIng COoperative Regulation Networks*) model.[15] Note that we slightly enrich the original LICORN model by creating a copy of each TF in the target layer to allow regulations between TFs. Each gene g is now connected with a set of co-regulated TFs, split into a group of co-activators $\mathcal{A}$ and co-inhibitors $\mathcal{I}$ (see Figure 1 (a)). We denote by $\mathcal{S}_g$ the (unknown) true status (under/over-expressed or normal) of each gene g in each sample. This gene is thus deregulated as soon as:

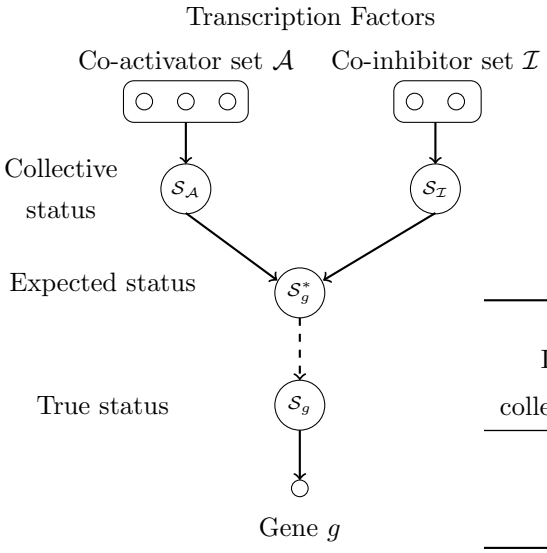
$$\mathcal{S}_g \neq \mathcal{S}_g^*,$$

where $\mathcal{S}_g^*$ is the expected status of g, which results from the deregulation rules presented in Figure 1 (b) and derived from biological experiments (see Ref. 15 for more details).

When considering the deregulation model of Figure 1 (a), one of the main difficulty is to differentiate deregulation mechanisms from other classical alterations that can strongly affect gene expression data and the interpretation results. For example, a Copy Number Alteration (CNA) occuring at the target gene layer influences its expression independently from the regulators expression and makes it wrongly appear as deregulated. This point is discussed in Section 2.3.

2.2. *Identification of deregulation mechanisms*

Based on the deregulation model of Section 2.1, we propose a 3-step strategy for the identification of deregulated mechanisms associated with specific subtypes of cancer. These steps are presented in Figure 2 below and described in detail in the next paragraphs. All codes are available as a `R` package at `https://github.com/magalichampion/LIONS_project`.



(a) Deregulation model.

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

(b) Truth table.

Fig. 1: (a) Deregulation model: each gene g, with unknown status $\mathcal{S}_g$ (under/over-expressed or normal), is activated and inhibited by a group of co-activators $\mathcal{A}$ and co-inhibitors TFs $\mathcal{I}$. A gene is deregulated when its true status $\mathcal{S}_g$ differs from its expected one $\mathcal{S}_g^*$, which is given by the co-regulator rules of the truth table (b).
(b) Truth table: gives the expected status $\mathcal{S}_g^*$ of a gene g (-/+ for under/over-expressed or 0 for normal) according to the collective status of its co-activators and co-inhibitors. The collective status is set by default to 0 except if and only if all of its elements share the same status.

2.2.1. *Step 1: inferring a Gene Regulatory Network*

In Step 1, we infer a GRN of reference that represents regulations between groups of co-expressed TFs and target genes, as presented in Figure 1 (a). To this aim, we use the hLICORN algorithm, available in the `CoRegNet` R-package.[16] By means of heuristic techniques, TFs are gathered into groups of co-expressed genes and each target gene is associated to pairs of co-activators and co-inhibitors that significantly explain its discretized expression. The network is selected among the local candidates based on the best prediction of the target gene expression. Previous work emphasized the ability of hLICORN to detect cooperative regulations on cancer data sets, which motivated its use.[15,16,17] However, when desired, it can be replaced by any other inference methods, such as the cooperative Lasso.[18] Step 1 can also be avoided by loading a pre-existing GRN, for example from the RegNetwork database.[19]

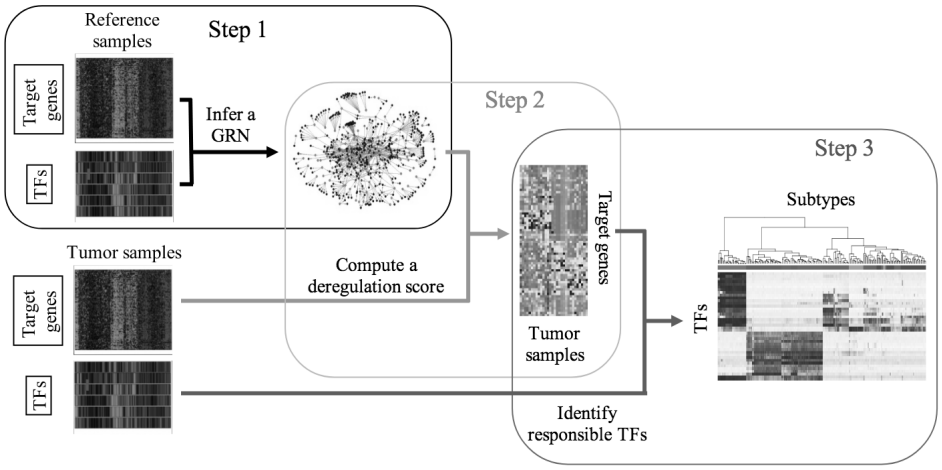


Fig. 2: Workflow of the proposed 3-step algorithm for identifying deregulation mechanisms in specific cancer subtypes.

2.2.2. *Step 2: computing a deregulation score*

In Step 2, we compute a deregulation score for each target gene in each tumor sample, indicating if its expression status $\mathcal{S}_g$ differs from $\mathcal{S}_g^*$ given by the reference GRN. Following the works of Ref. 7, the deregulation score Y of gene g in sample j is defined as the probability of gene g to be deregulated, or, in other words:

$$Y = \mathbb{P}(D_g = 1), \quad \text{where} \begin{cases} \mathcal{S}_g = \mathcal{S}_g^* & \text{if } D_g = 0 \\ \mathcal{S}_g \neq \mathcal{S}_g^* & \text{if } D_g = 1, \end{cases}$$

and D_g is a binary variable indicating whether the corresponding target gene g is deregulated.

To avoid discretization of the data, the status of all genes is considered as a hidden variable. As their number grows exponentially with the number of genes, the likelihood of the model rapidly becomes intractable and the deregulation score is thus estimated using a dedicated EM-algorithm (see Ref. 7 for more details).

2.2.3. *Step 3: identifying TFs involved in the deregulation of target genes*

In Step 3, we go back to the TFs layer to identify TFs that best explain deregulation of the target genes. For this purpose, we introduce a deregulation importance score B, which quantifies the role played by each TF in each sample in the deregulation of its associated targets through the linear regression model:

$$Y = G \cdot B + \varepsilon,$$

where Y is the deregulation score, G is the adjacency matrix of the reference GRN, whose non-zero elements encode the structure (edges) of the graph and ε stands for

the presence of noise in the model. Let n be the total number of samples and p the number of TFs. The deregulation importance score $B_{\cdot j}$ of sample j is estimated by solving the following least-squares problem:

$$\forall j \in [\![1, n]\!], \quad \hat{B}_{\cdot j} := \text{argmin}_{\beta \in \mathbb{R}^p} \quad \|Y_{\cdot j} - G\beta\|_2^2,$$

where $\|.\|_2$ stands for the euclidian norm and $M_{\cdot j}$ is the standard notation for the j-th column of any matrix M. Note that we impose all coefficients of $\hat{B}_{\cdot j}$ to be between 0 and 1, thus focusing on the bounded-variable least-squares problem:

$$\begin{aligned} \forall j \in [\![1, n]\!], \quad \hat{B}_{\cdot j} := \underset{\beta \in \mathbb{R}^p}{\text{argmin}} \quad & \|Y_{\cdot j} - G\beta\|_2^2. \\ s.t \quad & \forall \ell \in [\![1, p]\!], 0 \leq \beta_\ell \leq 1 \end{aligned} \tag{1}$$

The closer $\hat{B}_{ij}$ is to 1, the more important the role of TF i in the deregulation of its targets in sample j. As proved by Ref. 20, the non-negativity imposed on all the coefficients of $\hat{B}_{\cdot j}$ ensures a sparse recovery in the noiseless setting, which can be extended to the noisy setting with a hard-thresholding.

2.3. *Data preprocessing*

To run the procedure described in Section 2.2, gene expression data from two different sample sets have to be provided: reference samples, which ideally come from normal tissues, and tumor-specific samples. In practice, for many different cancer types, the pure normal tissue of origin is not available. However, the whole tumor data set can be split into small subtypes-depending subsets. With the aim of identifying differences between these cancer subtypes in terms of deregulation mechanisms, we thus proceed as presented in Figure 3. Before running our procedure, we extract all samples from one subtype to form the tumor-specific sample set whereas the rest is used as reference samples to infer the reference GRN. Note that the inferred networks will now reflect averaged relationships between genes for patients who are not part of the subtype we focus on. Due to the very-high heterogeneity of cancer, especially of bladder cancer, [21,22] we think that our method still points to relevant deregulations of specific subtypes.

As briefly mentioned in Section 2.1, our deregulation model suffers from the effect of genomic alterations, which can affect gene expression and make some regulations wrongly detected as deregulated. To remove CNAs effects on gene expression, we preprocess the target gene expression data beforehand as proposed in Ref. 23. In this work, gene expression is considered as linearly modified by CNA through the linear regression model:

$$X_{ij} = \alpha_0 + \alpha_1 \text{CNA}_{ij} + \varepsilon_{ij}, \tag{2}$$

where X_{ij} is the expression of gene j in sample i and CNA_{ij} its associated copy number. Denote by $\hat{\alpha}_0$ and $\hat{\alpha}_1$ the estimated solutions of Eq. (2). The corrected expression $\tilde{X}_{ij}$ is then given by:

$$\tilde{X}_{ij} = X_{ij} - \hat{\alpha}_0 - \hat{\alpha}_1 \text{CNA}_{ij}.$$

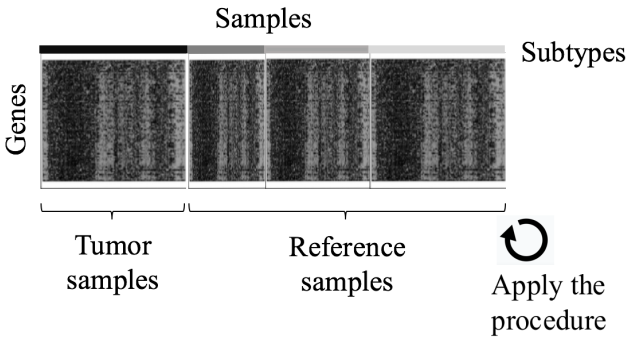


Fig. 3: Construction of the tumor and reference data set: due to the absence of normal tissue data, the whole tumor data set is divided into a tumor-specific one (samples from one subtype) and a reference one (samples from all other subtypes). The procedure is then iteratively applied to discover deregulation mechanisms for each cancer subtype.

Note that other common alterations in cancer may occur, such that hyper/hypomethylations. Even if they have been widely studied,[3] the effect of these epigenetic events on gene expression is however far from well-known and there is no model we can clearly exploit.

3. Application to Real Data

3.1. *The bladder cancer data set*

We apply our method on bladder cancer data, produced in the framework of The Cancer Genome Atlas project (TCGA) and available at the Genomic Data Commons (GDC) data portal (`https://portal.gdc.cancer.gov/`). These data include a set of 399 bladder cancer samples with gene expression and copy number for a total of 15,430 genes, split into 2,020 TFs and 13,410 targets. Gene expression data were produced using RNA-sequencing on bladder cancer tissues. Preprocessing is done by log-transformation and quantile-normalization of the arrays. Missing values are estimated using nearest neighbor averaging.[24] TCGA samples are analyzed in batches and significant batch effects are observed based on a one-way analysis of variance in most data modes. We apply Combat[25] to adjust for these effects. Genes are finally filtered based on their variability: we only keep the 75% most varying genes.

In the last few years, a large number of research groups have proposed molecular subclassifications of bladder cancer.[26] Recently, a consensus classification has been proposed for muscle-invasive bladder cancer,[27] which constitutes most of the tumors of the TCGA. As presented in Table 1, samples are split into six subtypes

with different characteristics not further detailed here (see Ref. 27 if interested): basal-squamous (BaSq), luminal non-specified (LumNS), luminal-papillary (LumP), luminal unstable (LumU), neuroendocrine-like (NE) and stroma-rich (Stroma-rich).

Table 1: Molecular subtypes distribution of the 399 bladder cancer samples.

Subtypes	BaSq	LumNS	LumP	LumU	NE-like	Stroma-rich
Samples	150	21	125	53	6	44

3.2. *Identification of deregulations*

After applying the procedure presented in Section 2.2 on the TCGA bladder cancer data set of Section 3.1, we get a list of deregulation importance scores for each TF in each sample. Table 2 presents the first 25 TFs, ranked according to their number of non-zero coefficients in $\hat{B}$, as given in Eq. (1), across all samples belonging to each specific subtype.

Table 2 includes characteristic genes of bladder cancer. As an example, NOTCH4 is deregulated in 71% and 76% of the BaSq and LumNS samples. This gene is part of the NOTCH pathway, whose inactivation tends to promote bladder cancer progression.[28] SPOCD1, involved in all subtypes except the LumP one, has also been recently shown to be able to distinguish patients with progressive and non-progressive disease.[29] Genes from the SOX family (SOX7 and SOX15), which play a major role in tumorigenesis in multiple cancers[30] including the bladder cancer one, are also highly present. In addition, we retrieve biomarkers of specific bladder cancer subtypes: SNAI2 (79% of the BaSq samples), which discriminates basal from luminal subgroup[31] and TBX2 (70% and 74% of the LumP and LumU samples) an indicator of luminal cancers. [32] Interestingly, TBX2 is also involved in BaSq, NE-like and Stroma-rich subtypes but this may be explained by its frequent overexpression in cancer.[33] We can finally note the presence of PPARGC1B, a coactivator of the well-known PPARG, whose high level of expression is used to describe luminal subtypes,[34] in 74% of the LumU samples.

3.3. *Comparison with other approaches*

The approach we developped captures deregulated genes in the sense that the regulation mechanisms between these genes go wrong. A natural question that arises is then how it behaves towards state-of-the-art methods. In this section, we compare deregulated TFs from our procedure to two classic approaches for biomarkers discovery:

Table 2: List of the 25 TFs that best explain the deregulation scores of their targets and number of non-zero coefficients in $\hat{B}$ (in %) across all samples from each subtype.

Subtypes											
BaSq		LumNS		LumP		LumU		NE-like		Stroma-rich	
TF	$\%\hat{B}$	TF	$\%\hat{B}$	TF	$\%\hat{B}$	TF	$\%\hat{B}$	TF	$\%\hat{B}$	TF	$\%\hat{B}$
SPOCD1	93%	SMARCD3	81%	RARA	91%	HES2	87%	HES2	83%	HES2	91%
IRX3	85%	ANKS1A	81%	PIAS2	86%	SOX7	83%	TBX2	83%	SPOCD1	91%
PEG3	83%	ZNF423	81%	CBFA2T3	81%	ZNF671	79%	MAFG	83%	TBX2	82%
SMARCD3	79%	NOTCH4	76%	SALL2	78%	SPOCD1	79%	PIR	83%	SPIB	77%
SNAI2	79%	HOXC9	76%	LHX6	78%	HOXB3	77%	IRX3	83%	IRX3	75%
ZSCAN12	76%	ZNF563	76%	RARB	77%	TBX2	74%	SETBP1	83%	ZNF713	75%
NAP1L2	76%	SPOCD1	71%	SOX7	76%	NR3C2	74%	RARB	83%	MAFG	73%
ZNF433	75%	HES2	71%	BACH2	76%	BTBD11	74%	SOX15	83%	HTATIP2	73%
JARID2	75%	IKBKB	71%	SMARCD3	74%	PPARGC1B	74%	MITF	83%	SMARCD3	70%
HOXA13	73%	NR3C2	71%	NFIC	74%	TEAD4	72%	HTATIP2	83%	PER3	70%
PRDM8	72%	HOXC6	71%	RUNX1T1	74%	SMARA2	70%	SOX7	67%	ZNF563	70%
TBC1D2B	72%	DLX3	71%	IRF7	74%	MAFG	70%	NR3C2	67%	MYCN	70%
CRY1	72%	FOXD1	71%	ZNF423	73%	EMX2	68%	TEAD4	67%	TEAD4	68%
NOTCH4	71%	HOXA3	67%	PIR	73%	ATF6	68%	EMX2	67%	ZNF420	68%
PHF19	71%	MESP1	67%	CREB3L4	73%	HOXC6	66%	ZNF347	67%	HSF4	68%
NFKB1	71%	HSF4	67%	TLE2	72%	NKD2	66%	ID2	67%	ZNF626	68%
CREM	70%	ZNF626	67%	ZNF71	71%	XBP1	64%	LMO3	67%	ZNF695	68%
PRDM5	70%	BCL3	67%	MXD1	71%	SPIB	64%	ENO1	67%	ZNF214	68%
ZNF215	69%	SOX15	67%	TBX2	70%	ZNF416	64%	NR3C1	67%	NFKBIA	66%
ARNT2	69%	CSDC2	67%	NR1I3	70%	HEY2	64%	PRDM8	67%	NFKBIZ	66%
TBX2	69%	RARB	67%	MAFG	70%	PKNOX2	64%	KLF15	67%	TEAD2	66%
NFIL3	69%	SNAI3	67%	ZNF559	70%	DLX3	62%	BTG2	67%	DLX5	66%
ID3	68%	MCM3	67%	ZNF442	69%	HOXC9	62%	HIVEP3	67%	FOXD4	66%
ETS2	67%	HOXB6	67%	HMGA2	69%	ZNF549	62%	SMAD6	67%	ZNF134	66%
ZNF669	67%	HOXB5	67%	SCML2	68%	PTTG1	62%	ETV7	67%	DAB2	66%

- differential gene expression analysis, which consists in performing statistical tests to discover quantitative changes in terms of expression levels between groups and is frequently used in cancer research to identify genes with important changes between tumor and normal samples, called differentially expressed genes (DEGs),[6]
- copy number alterations (CNAs) analysis, which aims at identifying genes targeted by copy number changes (amplifications/deletions) that critically affect their functions, especially in cancer.[35]

We perform differential expression analysis using the `R`-package `limma`[6] on all genes to check for significant differences between subtypes. In addition, we use the GISTIC algorithm[36] to identify significantly and recurrently deleted/amplified regions. We then verify whether the identified DEGs and CNAs affected genes are different from the deregulated ones (Figure 4). To this aim, we use the following thresholds: a gene is differentially expressed for p-values smaller than 0.01 whereas it is amplified/deleted when detected for more than 75% of the subtype samples. It is finally deregulated for a subtype when more than 50% of the samples have a non-

zero deregulation importance score ($\hat{B} \neq 0$). These thresholds are purely arbitrary but this is not crucial as the results remain almost the same with slight changes.

Obviously, the notion of deregulation should be different from the differential expression since a loss of regulation between a target gene and one of its regulators TFs implies a loss of correlation between them but not necessarily differential expression. This is confirmed by Figure 4, where only a small part of DEGs are deregulated. Except for BaSq and LumP subtypes, the two largest subtypes, more than 50% of the deregulated genes are also not differentially expressed. Similarly, we remove the effects of CNAs on gene expression (see Section 2.3) to ensure that amplified/deleted target genes are not wronlgy identified as deregulated. At the TF layer, a CNA still modifies both TFs and its targets expression level, making it different from deregulation, as can be seen in Figure 4.

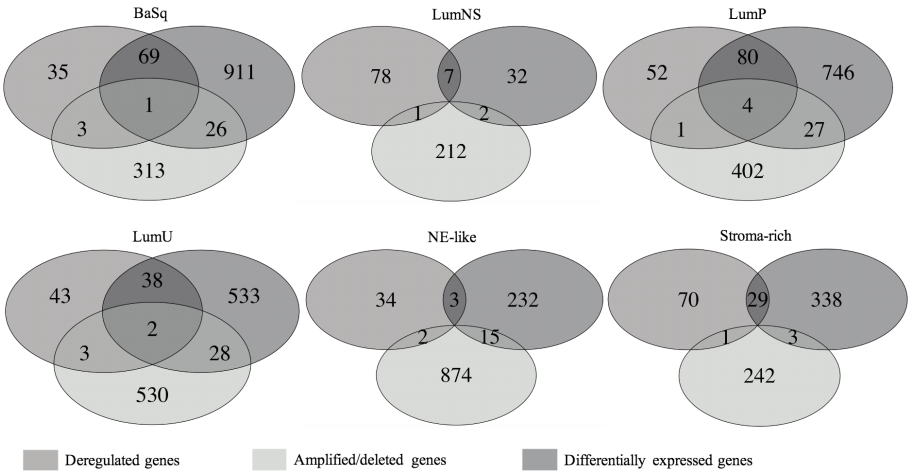


Fig. 4: Venn diagramms representing the number of deregulated TFs identified by our procedure, the number of amplified/deleted genes, the number of differentially expressed genes and their intersections.

4. Discussion

The procedure introduced in Section 2 aims at identifying deregulation mechanisms specific to cancer subtypes from gene expression data measured on two different tissues. By carefully comparing gene expression in tumoral samples with a reference GRN, which models "normal" regulation processes, a list of deregulated TFs characterizing each subtype can be established. The obtained results indicate that it can be used complementary to differential gene expression and CNA analysis to point to potential biomarkers of cancer.

An open question, which has to be tackled, is to determine to which extend the information carried by mutations can explain the deregulations. Mutation data are particularly hard to explore in this context due to various reasons: first of all, mutations do not necessarily affect gene expression. Secondly, mutation data are very sparse: in cancer, many sequencing projects have shown that, besides the most significant mutated genes, genes are mutated in less than 5% of the samples. Particularly, mutations of gene ELF3 have been the focus of a large number of studies, which underlines its tumor suppressive role, especially in bladder cancer.[37,38] Here, we report an association between ELF3 mutations and deregulations (number of non-zero $\hat{B}$ coefficients) in all three luminal subtypes (see Table 3, p-value for chi-test of 10^{-7}): the six deregulated samples are all mutated. However, at this step, supplementary work needs to be done to go further.

Table 3: Confusion matrix indicating the association between mutation and deregulation for TF ELF3 across all 199 luminal samples.

	$\hat{B} \neq 0$	$\hat{B} = 0$
Non mutated	0	165
Mutated	6	28

In Table 2, we also emphasize the presence of SPOCD1, which is deregulated in almost all subtypes, and more particularly in the BaSq one. Going back to the reference GRN, SPOCD1 is connected with 69 target genes. To give a biological meaning to these interactions, we perform gene set enrichment analysis with hypergeometric tests based on the databases GeneSetDB[39] and MSigDB.[40] We then find that the collection of 69 genes is enriched in 15 gene sets, most of which represent metastases-associated pathways. To a little extent, we also find an enrichment in Epidermal Growth Factor Receptor (EGFR)-pathways. EGFR plays a fundamental signalling role in cell growth and is frequently mutated and overexpressed in cancer.[41] It has recently been the subject of further work to explore its behavior in basal samples.[42,43] We can then naturally wonder how SPOCD1, known for inhibiting cell apoptosis,[44] interacts with EGFR.

Next, among the identified deregulated genes, HES2 is highly deregulated in subtypes LumNS, LumU, NE-like and Stroma-rich (Table 2). Interestingly, the subnetworks formed of HES2 and its targets represent genes that are expressed in cancer stem cells. Stem cells are undifferentiate cells that promote tumoral propagation and resistance, being able to differentiate and replicate into multiple cell types. To verify this association, we correlate HES2 expression with CD44 expression, as a marker of stem cells.[45] As can be seen in Figure 5, we observe a significant correlation in BaSq and LumP subtypes, the two subtypes in which HES2 is not deregulated. Literature around HES2 and the family of genes HES it belongs to is

quite limited but this result suggests that there is a close relation between stem cells and HES2 and that this relation is broken by deregulations in specific bladder cancer subtypes.

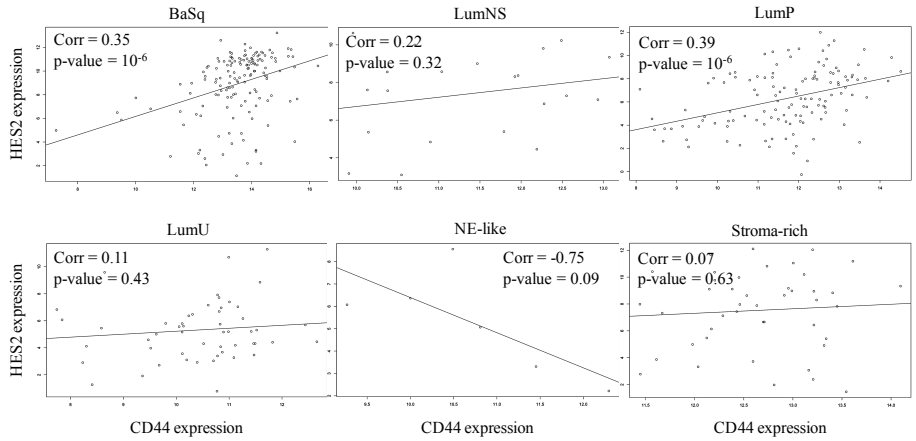


Fig. 5: Correlations between HES2 expression and CD44 expression, a biomarker of stem cells for all bladder cancer subtypes.

In summary, the procedure we developped provides a computational method for the identification of genes which are involved in hallmark cancer pathways. It captures other information than differential expression and copy number variation analysis to point to deregulated mechanisms in specific cancer subtypes. The present study emphasizes the importance of the two transcription factors SPOCD1 and HES2, whose roles have to be investigated deeper.

References

1. The Cancer Genome Atlas, Comprehensive molecular characterization of urothelial bladder carcinoma, *Nature* **507**(7492):315–322, 2014.
2. Lehman BD, Bauer JA, Chen X, Sanders ME, Bapsi Chakravarthy A, Shyr Y, Pietenpol JA, Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies, *J Clin Invest* **121**(7):2750–2767, 2011.
3. Kulis M, Esteller M, DNA methylation and cancer, *Adv Genet* **70**:27–56, 2010.
4. Shlien A, Malkin D, Copy number variations and cancer, *Genome Med* **1**(6):62, 2009.
5. Kaczkowski B, Tanaka Y, Kawaji H, Sandelin A, Andersson R, Itoh M, Lassmann T, Hayashizaki Y, Carninci P, Forrest AR, FANTOM5 Consortium, Transcriptome analysis of recurrently deregulated genes across multiple cancers identifies new pan-cancer biomarkers, *Cancer Res* **76**(2):216–226, 2016.
6. Ritchie et al ME, limma powers differential expression analyses for RNA-sequencing and microarray studies, *Nucleid Acids Res* **43**(7):e47, 2015.
7. 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**:S6, 2015.

8. Elati M, Rouveirol C, *Unsepervised learning for Gene Regulation Network Inference from Expression Data: a review*, John Wiley and Sons, Inc., Hoboken, NJ, 2011.
9. Tibshirani R, Regression shrinkage and selection via the lasso, *J R Statist Soc Ser B* **58**(1):267–288, 1996.
10. 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):e29165, 2011.
11. 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.
12. Chiquet J, Grandvalet Y, Ambroise C, Inferring multiple graphical structures, *Stat Comput* **21**(4):537–553, 2011.
13. Yeh JE, Toniolo PA, Frank DA, Targeting transcription factors: promising new strategies for cancer therapy, *Curr Opin Oncol* **25**(6):652–658, 2013.
14. Champion M, Chiquet J, Neuvial P, Elati M, Radvanyi F, Birmelé E, Identification of deregulated transcription factors involved in specific bladder cancer subtypes, *Proceedings of the 12th International Conference on Bioinformatics and Computational Biology*, EPiC Series in Computing, Vol. 70, pp. 1–10, 2020.
15. Elati M, Neuvial P, Bolotin-Fukuhara M, Barillot E, Radvanyi F, Rouveirol C, LICORN: learning cooperative regulation networks from gene expression data, *Bioinformatics* **23**(18):2407–14, 2007.
16. Nicolle R, Radvanyi F, Elati M, Coregnet: reconstruction and integrated analysis of co-regulatory networks., *Bioinformatics* **31**(18):3066–3068, 2015.
17. Chebil I, Nicolle R, Santini G, Rouveirol C, Elati M, Hybrid method inference for the construction of cooperative regulatory network in human, *IEEE Trans Nanobioscience* **13**(2):97–103, 2014.
18. Chiquet J, Grandvalet Y, Charbonnier C, Sparsity in sign-coherent groups of variables via the cooperative-Lasso, *Ann Appl Stat* **6**:795–830, 2012.
19. Liu Z, Canglin W, Miao H, Wu H, RegNetwork: an integrating database of transcriptional and post-transcriptional regulatory networks in human and mouse, *Database (Oxford)* **2015**:bav095, 2015.
20. Slawski M, Hein M, Sparse recovery by thresholded non-negative least squares, in *Adv Neur In*, eds., Shawe-Taylor J, Zemel RS, Bartlett PL, Pereira F, Weinberger KQ, Curran Associates, Inc., pp. 1926–1934, 2011.
21. Knowles M, Hurst C, Molecular biology of bladder cancer: new insights into pathogenesis and clinical diversity, *Nat Rev Cancer* **15**:25–41, 2014.
22. Togneri FS, Ward DG, Foster JM, J DA, Wojtowicz P, Alyas S, Ramos Vasques F, Oumie A, James ND, Cheng KK, Zeegers MP, Deshmukh N, O'Sullivan B, Taniere P, Spink KG, McMullan DJ, Griffiths M, Bryan T Richard, Genomic complexity of urothelial bladder cancer revealed in urinary cfDNA, *Eur J Human Genet* **24**:1167–1174, 2016.
23. Delatola EI, Lebarbier E, Mary-Huard T, Radvanyi F, Robin S, Wong J, SegCorr: a statistical procedure for the detection of genomic regions of correlated expression, *BMC Bioinformatics* **18**(1):333, 2017.
24. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB, Missing value estimation methods for DNA microarrays, *Bioinformatics* **17**(6):520–525, 2001.
25. Johnson W, Li C, Rabinovic A, Adjusting batch effects in microarray expression data using empirical bayes methods, *Biostatistics* **8**:118–127, 2007.
26. Robertson AG, Kim J, Al-Ahmadie H, Bellmunt J, Guo G, Cherniack AD, Hinoue T,

Laird PW, Hoadley KA, Akbani R, AA CM, Gibb RS Ewan A Kanchi, Gordenin DA, Shukla SA, Sanchez-Vega F, Hansel DE, A CB, Reuter VE, Su X, de Sa Carvahlo B, Chagas VS, Mungall KL, Sadeghi S, Sekhar Pedamallu C, Lu Y, Klimczak LJ, Zhang J, Choo C, Ojesina AI, Bullman S, Leraas KM, Lichtenberg TM, Wu CJ, Schultz N, Getz G, Meyerson M, Mills GB, McConkey DJ, TCGA Research Network, Weinstein N John, Kwiatkowski DJ, Lerner SP, Comprehensive molecular characterization of muscle-invasive bladder cancer, *Cell* **171**(3):540–556, 2017.

27. Kamoun A, de Reyniès A, Allory Y, Sjödahl, Robertson AG, Seiler R, Hoadley KA, Groeneveld CS, Al-Ahmadie H, Choi W, Castro MAA, Fontugne J, Eriksson P, Mo Q, Kardos J, Zlotta A, Hartmann A, Dinney CP, Bellmunt J, Powles T, Malats N, Chan KS, Kim WY, McConkey DJ, Black PC, Dyrskøt L, Höglund M, Lerner SP, Real FX, Radvanyi F, A consensus molecular classification of muscle-invasive bladder cancer, *Eur Urol* **77**(4):420–433, 2016.
28. Maraver A, Fernandez-Marcos PJ, Cash TP, Mendez-Pertuz M, M DP, Martinelli P, Muños-Martin M, Martínez-Fernández M, Cañamero M, Roncador G, Martinez-Torrcuadrada JL, Grivas D, Luis de la Pompa J, Valencia A, Paramio JM, Real FX, Serrano M, NOTCH pathway inactivation promotes bladder cancer progression, *J Clin Invest* **125**(2):824–830, 2015.
29. van der Heijden AG, Mengual L, Lozano JJ, Ingelmo-Torres M, Ribal MJ, Fernández PL, Oosterwijk E, Schalken JA, Alcaraz A, Witjes JA, A five-gene expression signature to predict progression in T1G3 bladder cancer, *Eur J cancer* **64**:127–136, 2016.
30. Thu KL, Becker-Santos DD, Radulovich N, Pikor LA, Lam WL, Tsao M, SOX15 and other SOX family members are important mediators of tumorigenesis in multiple cancer types, *Oncoscience* **1**(5):326–335, 2014.
31. Mistry DS, Chen Y, Wang Y, Sen GL, SNAI2 controls the undifferentiated state of human epidermal progenitor cells, *Stem Cells* **32**(12):3209–3218, 2014.
32. Dhawan D, Paoloni M, Shukradas S, Choudhury DR, Craig BA, Ramos-Vara JA, Hahn N, Bonney PL, Khanna C, Knapp W Deborah, Comparative gene expression analyses identify luminal and basal subtypes of canine invasive urothelial carcinoma that mimic patterns in human invasive bladder cancer, *PLoS One* **10**(9):e0136688, 2015.
33. Abrahams A, Parker M, Prince S, The T-box transcription factor Tbx2: Its role in development and possible implication in cancer, *IUBMB Life* **62**(2):92–102, 2010.
34. Choi W, Porten S, Kim S, Willis D, Plimack ER, Hoffman-Censits J, Roth B, Cheng T, Tran M, I-Ling L, Melquist J, Bondaruk J, Majewski T, Zhang S, Pretzsch S, Baggerly K, Siefker-Radtke A, Czerniak B, Dinney CP, LcConkey DJ, Identification of basal and luminal subtypes of muscle-invasive bladder cancer with different sensitivities to frontline chemotherapy, *Cancer cell* **25**(2):152–165, 2014.
35. Santarius T, Shipley J, Brewer D, Stratton M, Cooper C, A census of amplified and overexpressed human cancer genes, *Nat Rev Cancer* **10**(1):59–64, 2010.
36. Mermel C, Schumacher S, Hill B, Meyerson M, Beroukhim R, Getz G, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers, *Genome Biol* **12**(4):R41, 2011.
37. Luk IY, Reehorst CM, Mariadason JM, ELF3, ELF5, EHF and SPDEF transcription factors in tissue homeostasis and cancer, *Molecules* **23**(9):2191, 2018.
38. Nordentoft I, Lamy P, Birkenkamp-Demtröder K, Shumansky K, Vang S, Hornshøj H, Juul M, Villesen P, Hedegaard J, Roth A, Thorsen K, Høyer S, Borre M, Reinert T, Fristrup N, Dyrskjøt L, Shah S, Pedersen JS, Ørntoft TF, Mutational context and diverse clonal development in early and late bladder cancer, *Cell reports* **7**(5):1649–1663, 2014.

39. Culhane AC, Schwarz l, Sultana R, Picard KC, Picard SC, Lu TH, Franklin KR, French SJ, Papenhausen G, Corell M, Quackenbush J, GeneSigDB, a curated database of gene expression signatures, *Nucleic Acids Res* **38**:D716–D725, 2010.
40. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P, The molecular signatures database (MSigDB) hallmark gene set collection, *Cell systems* **1**(6):417–425, 2015.
41. Mendelsohn J, Baselga J, Status of epidermal growth factor receptor antagonists in the biology and treatment of cancer, *J Clin Oncol* **21**(14):2787–2799, 2003.
42. Zanetti-Domingues LC, Korovesis D, Needham SR, Tynan CJ, Sagawa S, Roberts SK, Kuzmanic A, Ortiz-Zapater E, Jain P, Roovers RC, Lajevardipour A, van Bergen En Henegouwen P, Santis G, Clayton A, Clarke DT, Gervasio FL, Shan Y, Shaw DE, Rolfe DJ, Parker PJ, Martin-Fernandez ML, The architecture of EGFR's basal complexes reveals autoinhibition mechanisms in dimers and oligomers, *Nat commun* **9**(1):4325, 2018.
43. Avci CB, Kaya I, Ozturk A, Ay NP, Sezgin B, Kurt CC, Akyildiz NS, Bozan A, Yaman B, Akalin T, Apaydin F, The role of EGFR overexpression on the recurrence of basal cell carcinomas with positive surgical margins, *Gene* **687**:35–38, 2019.
44. Liang J, Zhao H, Hu J, Liu Y, Li Z, SPOCD1 promotes cell proliferation and inhibits cell apoptosis in human osteosarcoma, *Mol Med Rep* **17**(2):3218–3225, 2018.
45. Thapa R, Wilson GD, The importance of CD44 as a stem cell biomarker and therapeutic target in cancer, *Stem Cells Int* p. 2087204, 2016.