# Identification of Tissue Independent Cancer Driver Genes

Alexandros Manolakos, Idoia Ochoa, Kartik Venkat
Supervisor: Olivier Gevaert

***Abstract*—Identification of genomic patterns in tumors is an important problem, which would enable the community to understand and extend effective therapies across the current (tissue-based) tumor boundaries. In this work, we make two specific contributions. First, we develop a robust system to discover cancer driver genes, via an unsupervised clustering of similarly expressed genes across cancer patients. This is the 'module network identification' phase. Second, we develop a methodology to compare the quality, homogeneity and similarity of the generated clusters (or module networks). This step enables us to discover and quantify the tissue-independent genomic similarity across tumors.**

## I. Introduction

Traditionally, medical science has converged upon cancer treatment strategies unique to the tumor type (organized by the affected tissue), such as breast cancer, lung cancer, etc. Recently, however, there has been a significant push by the research community in discovering shared molecular and genomic similarities across different tumors. For example, recent studies [1] have shown that basal-like breast cancer has more similarities, genomically speaking, to high-grade serous ovarian cancer than to other subtypes of breast cancer.

The advantages of discovering such connections at various molecular and genetic levels are quite apparent. The common inter-tumor molecular patterns can suggest a unified clinical treatment strategy to combat multiple tumors. Thus, the statistical evidence for molecular, proteomic and epigenetic similarities across various tumors is fundamentally interesting, both from the perspective of scientific discovery and the future of personalized medicine.

Until now, research efforts have mainly focused on studying and analyzing tissue dependent genomic patterns. The Cancer Genome Atlas (TCGA) Research Network [2] has collected and analyzed a large amount of data from different human tumors (cancers), to discover molecular aberrations at the DNA, RNA, protein and epigenetic levels. Recently, the Pan-Cancer initiative has been created to compare the first 12 tumor types profiled by the TCGA. In this project, we will use the Pan-Cancer data to help uncover underlying genomic patterns across several different tumors.

Central to our discussion, is the knowledge that a small number of important genes, known as 'regulatory' or 'driver' genes, play a crucial role at the molecular pathway level and directly influence the expression of several other genes. This network of genes, where these driver genes are connected with other downstream targets, is known as the module network [7]. It seems natural that some of these regulatory genes should be able to explain the variability of gene expression in genes that appear downstream in these biological pathways. Thus, researchers are attempting to identify the module network structure based on gene expression data in cancer patients, using machine learning techniques. For example, in [4], the authors identify the module network structure in ovarian cancer.

Motivated by the aforementioned reasons, we are interested in understanding inter-tumor genomic similarities. We divide this problem into two parts. First, we develop an independent 'module network identification' system, which discovers the latent gene clustering[1] structure for a given dataset. After identifying the module network for individual tumors, we now wish to see which of these networks (corresponding to different tumors), are structurally similar and homogeneous. For this, we develop a method to score the similarity of two different clusterings of the data. This will enable us to find groups of tumors that display a similar module network structure, which is the goal of our project.

The rest of this report is organized as follows. In Section II we discuss the format and type of data. In Section III we develop and propose the module network identification system, which in turn is comprised of two different algorithms. In Section IV, we introduce a procedure to collate the results from the 'module network identification' stage, to quantitatively identify similar groups of tumors. In Section V we describe the evaluation criteria and discuss the findings of our study. Finally, we present concluding remarks in Section VI.

## II. Data

**The Pan-Cancer data [2]**: It consists of the expression value of $n = 19451$ genes for $m = 3452$ patients spanning a total of 12 tumor types; i.e., we have a matrix with dimensions $19451 \times 3452$, organized as *genes* $\times$ *patients*.

We pre-process this data in two phases. Firstly, we normalize each gene expression vector to have zero mean and standard deviation one. Secondly, we apply a variance filter to retain only the $50\%$ most varying genes; i.e., genes that seem to contain useful information across patients. We apply this procedure to both the training and the test data.

**Regulatory genes**[2]: This is a subset of genes which are identified via certain biological regulatory mechanisms - and are known to drive other genes. This set has been created based on transcription factor data extracted from the HPRD database [3]. Our data-set consists of $p = 3609$ regulatory genes. However, only those that appear in the Pan-Cancer data after pre-processing are relevant.

## III. Module Network Identification

### A. System Model

We start by introducing the notation. Denote by $n$ the number of genes, by $p$ the number of regulatory genes ($p \ll n$), and by $m$ the number of patients. We refer to the vector gene $\mathbf{g}_i, i \in [1:n]$, as its expression across all patients, i.e., $\mathbf{g}_i \in \mathbb{R}^m$. Recall that some of these genes form part of the regulatory genes, that we denote by $\mathbf{g}^{(r)}$. Without loss of generality, we assume that the regulatory genes correspond to the first $p$ genes, i.e., $\mathbf{g}_i^{(r)} = \mathbf{g}_i$, for $i \in [1:p]$.

We now formulate the module network identification problem, as an unsupervised clustering problem in the gene space. In other words, we seek to perform an unsupervised clustering of the genes (which are represented as vectors in $\mathbb{R}^m$), in a manner that each cluster is sparsely representable in the regulatory-gene

[1] It will become clear later, why we refer to the module network as an unsupervised clustering of the genes.

[2] Also referred to (interchangeably) as driver genes throughout the report

basis. In the literature, this is referred to as a module-network based approach towards genomic profiling of tumors.

More formally, given a cluster (module) $C$, let us denote by $I$ the set of indices of genes that belong to it ($|I| < n$), and by $\boldsymbol{\alpha} \in \mathbb{R}^p$ the sparse coefficients in the regulatory basis. Note that the regulatory gene $\mathbf{g}_i^{(r)}$ is a driver gene of this cluster iff $\alpha_i \neq 0$. We are looking for clusters whose centroid is well approximated by the sparse linear combination of the regulatory genes; i.e.,

$$\boldsymbol{\mu}^{(r)} \doteq \sum_{j=1}^{p} \alpha_j \mathbf{g}_j^{(r)} \approx \boldsymbol{\mu} \doteq \frac{1}{|I|} \sum_{i \in I} \mathbf{g}_i. \tag{1}$$

The intuition is that if $\boldsymbol{\mu}^{(r)}$ and $\boldsymbol{\mu}$ are similar, then the gene expression of the genes belonging to the cluster is well explained by a small linear combination of the expression of the regulatory genes.

### *B. Proposed Algorithms*

We now outline two methods that we use as a baseline towards solving this clustering problem across patients of different tumors. These methods comprise our module network identification system.

**Method 1**: Our primary starting point is the recent work by Gevaert et al. [4] in which they propose the AMARETTO algorithm, an iterative clustering algorithm, and apply it to understand the genomic profile of ovarian cancer. Here is a brief summary of the algorithm:
i) Firstly, the genes are clustered into groups using standard k-means with 100 clusters (modules).
ii) Then, the centroid of each cluster is expressed in terms of the driver genes using linear regression with L1 and L2 regularization (c.f. elastic net regularization [5]). Spasm Toolbox [6], a Matlab toolbox for sparse statistical learning, is employed for this purpose. The L1 regularization weight is chosen based on 10-fold cross-validation, whereas the L2 weight is fixed. In the end of this step, each module (cluster) contains a set of genes whose average expression is described using a small number of cancer driver genes.
iii) Finally, the correlation of each gene with the sparse representation of all the centroids is calculated. Each gene is re-assigned to the cluster whose centroid it is most positively correlated with.
iv) The algorithm repeats steps 2 and 3 until the gene re-assignment process converges (to less than $1\%$ of the total genes being reassigned).

**Method 2**: The second approach is inspired by AMARETTO, but inverts the order of elastic-net and clustering operations. This method is summarized below.
i) A sparse representation of every gene over the driver-gene basis is performed using elastic-net regression. A sparsity level of 30 regulatory genes is imposed. Observe that currently we do not perform any cross validation in order to find the best L1 weight and we just express each gene as a linear combination of 30 regulatory genes. L2 weight is chosen to be 10. In the end of this step, each gene is represented as a sparse vector in the $\mathbb{R}^p$ space with 30 non-zero elements.
ii) We perform a standard k-means clustering of the sparse vectors using Euclidean distance. The best value of $k$ is optimized using the elbow method, i.e., we perform the k-means clustering for $k = \{2, 10, 20, \cdots, 100\}$ and we use the $k$ that leads to explained energy that is $85\%$ of the explained energy for $k = 100$.
iii) The centroids of each cluster are expressed in terms of the regulatory genes using the elastic net method with $10-$fold cross validation for identifying the best L1 coefficient. We choose not to express each centroid with less than 7 regulatory genes and The L2 coefficient of the elastic-net is fixed to 10.

The sparsity level of each gene, the elbow parameter, the sparsity level of the centroids used in steps i),ii) and iii), respectively, and the L2 weight used throughout Method 2, have been chosen after performing the following parameter optimization procedure. We fixed all the parameters to a default value and optimized separately each parameter. Then, we used the performance criteria described in Subsection III-C to independently choose their value.

The basic differences between the above two algorithms are the following:
$-$ The second approach is a non-iterative procedure. Each gene is transformed to a sparse vector and then k-means is performed in this sparse representation to identify the clusters. This leads to a very fast implementation that scales well with the number of patients. For example, using a $70-30$ cross validation in the pan-cancer data, the second approach needs for both training and testing approximately 20 minutes, whereas the first approach (due to its iterative nature), needs several hours.
$-$ The first approach (AMARETTO) clusters points in $\mathbb{R}^m$ where $m$ is the number of patients. This cannot scale well when the number of patients increase. In the second approach, the classification is performed using very sparse points in the $\mathbb{R}^p$ space, e.g., points with only 10 non-zero values out of the $\sim 2000$ elements.
$-$ In AMARETTO a gene is re-assigned to the cluster that it is mostly positively correlated with. In the second approach we use Euclidean distance in order to cluster genes together.

### *C. Performance Criteria*

In order to evaluate these clustering algorithms, we use the coefficient of determination, known as $R^2$, to measure the similarity between $\boldsymbol{\mu}^{(r)}$ and $\boldsymbol{\mu}$, which were defined in (1). In our case, $R^2 \doteq 1 - \frac{S_{res}}{S_{tot}}$, where $S_{res}^j = \|\boldsymbol{\mu}^{(r)} - \boldsymbol{\mu}\|_F^2$, $S_{tot}^j = \|\frac{1}{|I|}\sum_{i \in I} \mathbf{g}_i - \boldsymbol{\mu}\|_F^2$. High $R^2$ means that the residual energy that is not explained by the assigned regulatory genes is relatively small. In order to adjust for the number of regulators relative to the number of patients, we compute the adjusted $R^2$ for each cluster, defined as $\bar{R}^2 \doteq R^2 - (1 - R^2)\frac{r}{m-r-1}$, where $r$ is the number of regulators that are present in the cluster.

We are looking for clusters with a high $R^2$ value. However, a high $R^2$ value is not enough for a cluster to be considered "good". It is also important to see how many genes are explained in the cluster. For example, it is more meaningful to explain 200 genes with 6 regulators, than 5 genes with 6 regulators. With this in mind, we only consider clusters with high $R^2$ ($> 0.10$), and with number of genes between 10 and 500. We refer to such clusters as "good" clusters. Throughout our discussion, we will only consider the "good" clusters for downstream applications.

These two methods have proven to work well, in the sense that they are able to find "good" clusters given a dataset composed of the expression of genes across patients (i.e., a matrix of genes $\times$ patients). Therefore, we can apply them directly on the Pan-Cancer dataset to find "good" clusters of genes across all the different tumors that are well explained by a sparse representation of the regulatory genes.

However, applying the method to the whole Pan-Cancer[3] dataset will fail in finding clusters that are good only among some tumors, rather than all tumors. Recall that we are trying to find clusters of genes across patients of different tumors (but not necessarily all tumors at the same time, which could be very restrictive). A trivial approach would be to apply both methods to all possible subsets of tumor-types, but this is computationally very expensive, and does not scale well with increasing patient data. This leads us to the next section, where we propose two approaches for this task, which use as a baseline the aforementioned module network identification methods.

## IV. Comparing Clusters Across Tumors

Having identified the module network structure for individual tumors using the methods outlined in Section III, we now wish to investigate which tumors are similar. In other words, which groups of tumors exhibit a similar gene expression clustering? This boils down to quantifying the similarity of two clusterings of a given dataset. Once we do that, the natural next step would be to identify the driver genes for these 'similar' tumors taken together. We now propose two distinct approaches for the same.

**Approach 1:** In this approach, for each of the methods we do the following:
i) We apply the method to each of the individual tumors. As a result, for each tumor type we will have a set of clusters.
ii) From those clusters, we retain only those that are "good", as explained in Section III-C.
iii) Given all the "good" clusters from the individual tumors, we now find similarity across clusters of different tumors. I.e., given a "good" cluster, we compare it with all the other "good" clusters that are not of the same tumor type. In order to compare two clusters, we compute the similarity of the genes and regulators that belong to each of the clusters. Specifically, we compute the Jaccard index and a modified Jaccard index that we introduce next. The Jaccard index between two sets $A$ and $B$ is given by the ratio of cardinalities of the intersection and the union; i.e., $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ In our case, the two sets will consist of the genes (or regulators) that belong to two different clusters. When analyzing the Jaccard index, we found that two different situations could lead to a similar Jaccard index. Consider the case where the union is large and one set is contained in the other (e.g., $|A \cap B| = \min(|A|, |B|)$), and the case where the intersection between the two sets is very small. It is easy to see that in both cases the Jaccard index will be very small, even though in the first case one cluster is contained in the other, and in the second case the clusters are very different. To avoid this situation, we also consider a modified Jaccard index, given by $J'(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$.
iv) We use the Jaccard index and the modified Jaccard index across clusters of different tumors to find tumors of different type that are related to each other. With this information we create combination of tumors that we believe are correlated (in the sense of similar gene expression), and select the best $50\%$ combinations (i.e, those with high indexes).
v) Finally, we apply the method again in those combinations of tumors that were detected in the previous step. As a result, we will now have a set of clusters for each combination of tumors. Among those, we select the "good" ones, as before.

**Approach 2:** In this approach the steps are the following:
i) We apply both methods to each of the individual tumors.
ii) Among all the clusters found by both methods, we retain only the "good" ones.
iii) Given the set of "good" clusters of both methods, we compute the modified Jaccard index across clusters of different tumors (independently of the method that found the specific cluster). In a sense, we use Method 1 and 2 as experts that find clusters, and make no difference about which method found a specific cluster.
iv) We discard all the combinations of clusters that have a modified Jaccard index below $0.5$, for either the genes or the regulators.
v) We then select sets of clusters from different tumors that are potentially similar (e.g., that have a high modified Jaccard index between them). Given a set of clusters, we create four new clusters composed of the intersection/union of the genes and the intersection/union of the regulators of the clusters of the set. Denote by $I_g$ the index of the genes that belong to a new cluster, and by $I_r$ the index of the regulators. Then, we use the normal equations to solve for the vector $\boldsymbol{\alpha}$ that minimizes

$$\underset{\alpha}{\operatorname{argmin}} \frac{1}{2} || \sum_{j \in I_r} \alpha_j \mathbf{g}_j^{(r)} - \frac{1}{|I_g|} \sum_{i \in I_g} \mathbf{g}_i ||^2. \tag{2}$$

Note that $\alpha_j \neq 0$ iff the regulator $\mathbf{g}_j^{(r)}$ is part of the new cluster. Therefore, since the set of regulators of the new cluster is already small, the vector $\boldsymbol{\alpha}$ is sparse by construction, and this is why we can use the normal equations directly. Among the four combinations, we retain only the best one, as long as it satisfies that it is "good" and that the $R^2$ computed on the tumors involved in the cluster is similar to the $R^2$ computed on each of the tumors separately. By doing so, we filter out the cases where the $R^2$ is very high for some tumors but very small for the others, and still give a high average $R^2$. We believe a good cluster should have a similar performance across the involved tumors.

### A. Brief comment on the two approaches

Here we will briefly highlight the differences in the two approaches for integrating the clusters and discovering similar tumors. Approach 1 discovers similar clusters to gauge which tumors appear similar. We then use the architecture developed in 'module network identification' to discern the module network structure for this group of tumors. Thus, we use Approach 1 to help us select which tumor groups could potentially have common driver genes, and then use the methods of Section III to find them. On the other hand, Approach 2 is more direct. We directly compute the resulting driver genes that emerge from analyzing the module networks discovered in the identification phase. I.e., we directly look for driver genes that explain the good clusters. Further, the computation is quite simple, since we reduce it to a least squares problem.

## V. Results

In all the simulations we created a training set containing $70\%$ of the data, and a testing set with the remaining $30\%$. We then use the training set to find the clusters, and then analyze them on the testing set.

### A. Module Network Identification: Pan-Cancer dataset

We applied both algorithms from Section III-B on the Pan-Cancer dataset. To analyze the clusters found by each method, we first filter the "good" clusters, and for each of them we show the number of regulators, the number of genes and the $R^2$ value

[3]All the 12 tumors

on the same plot, for ease of visualization. Specifically, the x-axis represents the number of regulators, the y-axis the number of genes, and for each "good" cluster we create a bubble proportional to the size of the $R^2$ value.

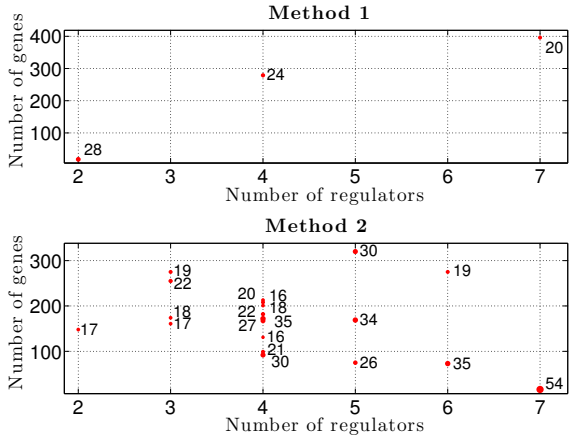


Fig. 1: "Good" clusters found by each of the methods on the Pan-Cancer data. We show the clusters as a function of the number of regulators and the number of genes. The size of the bubbles is proportional to the value $100 \times R^2$.

Fig. 1 shows the results for both methods when applied to the Pan-Cancer data. Most of the clusters disappear after the filtering. For example, in Method 1 only three clusters remain. One of the clusters of Method 1 can explain around 300 genes with 4 regulators, with an $R^2$ on the test data of 0.24. Method 2 gets more good clusters than Method 1 for this dataset, and some of them with higher $R^2$ value. For example, there is a cluster that explains approximately 180 genes, also with 4 regulators, but with an $R^2$ value of 0.35. Combining both methods, we get in total 13 clusters with an $R^2$ above 0.20.

### *B. Comparing Clusters Across Tumors*

**Approach 1:** As explained in Section IV, we first apply each of the methods in the individual tumors, and then we select the "good" clusters. For each method, we compute the Jaccard distance (for both the genes and the regulators) among clusters that belong to different tumors. With this information, we detect tumors that are related. Specifically, given two different tumors, we look at the maximum Jaccard index of both the genes and the regulators among all the combinations of clusters belonging to the two specific tumors. If the minimum of these two quantities is bigger than 0.2, we assume the tumors are related. Fig. 2 shows the best connections for each pair of tumors when using Method 2. Looking at the graph, we can infer that LUSC and LUAD seem to be highly correlated, for example, whereas LAML does not seem to be very related to other tumors regarding the gene expression. We also see that the tumors with more connections are BLCA, LUAD, UCEC, BRCA, KIRC and HNSC, in that order. The results we obtain when applying Method 1 are not exactly the same, as expected, but agree in the best two connections, which are again given by LUAD-KIRC and LUAD-LUSC.

From these connections we estimate the best combinations of tumors to which apply the method (1 or 2). For this, we use the k-clique algorithm [8], which detects overlapping communities in a network, and select the best 50%. After doing so, we retain only the "good" clusters for each of the combinations. We were able

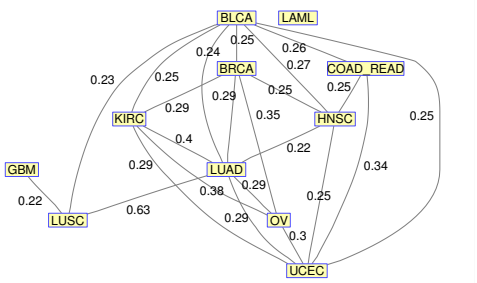


Fig. 2: Relations between pair of tumors computed using the Jaccard index between clusters that belong to them. The results correspond to Approach 1 applied with Method 2.

to find several good clusters using this approach, for both of the methods. Next we show some of the results we obtained.

Fig. 3 shows the clusters obtained by Method 1 on the combination LUAD-LUSC. We show the histogram of the number of regulators and number of genes for the clusters found by running Method 1 in each of the tumors separately and in the combined dataset. As it can be observed, less clusters are obtained in the combination, as expected. The $\bar{R}^2$ value remains similar for both the individual and combine tumors, although it is not very high. We observed similar results when developing Approach 1 with Method 1: i) less clusters than in the individual tumors, and ii) not very high $\bar{R}^2$ values. Finally, we compare the clusters with some natural pathways. These natural pathways establish relations across genes, not necessarily from a specific tumor and not necessarily in the gene expression driven by regulatory genes. Therefore, we can not use them as a criterion to establish if a cluster is good or not, instead, we are just interested to see if some of the natural pathways are present in the clusters we found. For the same combination of tumors, LUAD-LUSC, Method 2 identified a total of 6 "good" clusters, four of which present an $\bar{R}^2$ close to 0.5 and one of them around 0.6.

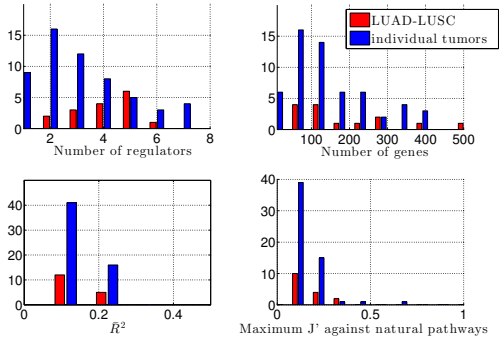


Fig. 3: Some results of the clusters found by Approach 1 with Method 1 for the individual tumors LUAD and LUSC and its combination.

Fig. 4 shows similar graphs but for the combination KIRC-LUAD when Method 2 is applied. As in the previous example, less clusters are found in the combination of tumors than on the individual tumors. Regarding the $\bar{R}^2$ values, in this case we obtain some cluster with $\bar{R}^2$ close to 0.5. In summary, for the combination KIRC-LUAD Approach 1 with Method 2 was able to identify 6 clusters with $\bar{R}^2$ between 0.1 and 0.5. With Method 1 we identified 5 clusters, $\bar{R}^2$ below 0.25. Regarding the comparison with the natural pathways, we observe that two of the clusters of the combined tumors have a modified Jaccard index close to 0.5.

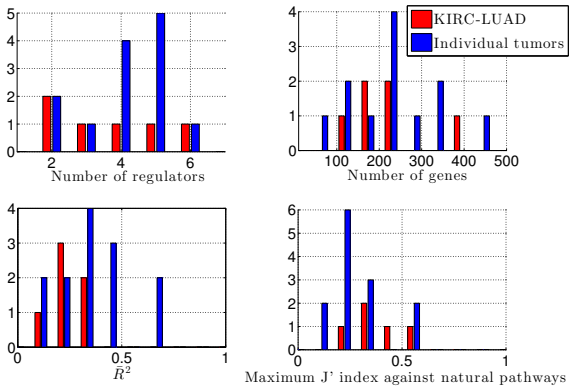


Fig. 4: Some results of the clusters found by Approach 1 with Method 2 for the individual tumors KIRC and LUAD and its combination.

Finally, we compare the $\bar{R}^2$ obtained by this approach when applied with both of the methods. Specifically, for each method, we show the histogram of the $\bar{R}^2$ corresponding to the clusters found on the individual tumors and the clusters found on all the combinations of tumors that were found to be related. As we can observe on Fig. 5, both methods present an histogram on the combined tumors similar to the one obtained on the individual tumor. Another observation is that Method 2 obtains on this data values of $\bar{R}^2$ higher than those obtained by Method 1.

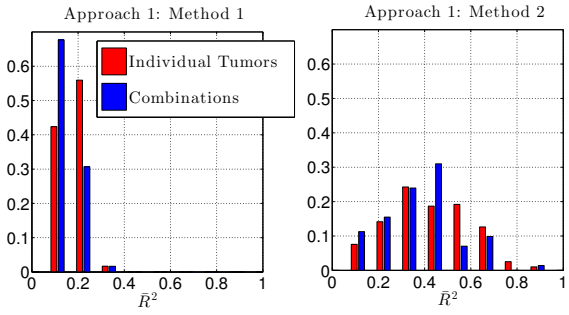


Fig. 5: Histogram of the $\bar{R}^2$ obtained with the Approach 1 when applying Method 1 and 2 on the individual and combined tumors.

**Approach 2:** We applied Method 1 and Method 2 in each of the tumors separately, yielding 442 "good" clusters in total. We then computed the modified Jaccard index of both the genes and the regulators for all the possible combinations of clusters among different tumors, and retained only those with an index of at least 0.5. Fig. 6 shows the values obtained for both the genes and the regulators. Note that the x-axis and the y-axis represent the same clusters, and thus the resulting plot is symmetric. As it can be observed, there are some clusters that intersect with several clusters from other tumors (vertical and/or horizontal lines of circles). This suggests that some clusters may overlap, meaning that we can create new clusters that can potentially explain a subset of genes with a subset of regulators across different tumors. Following the steps of Approach 2 we were able to find several "good" clusters across tumors of different types. Some of them are shown in Table I.

## VI. Conclusions

In this work we have studied the problem of finding set of genes (clusters), expressed across patients of different tumors, that are well explained by a small subset of regulatory genes.

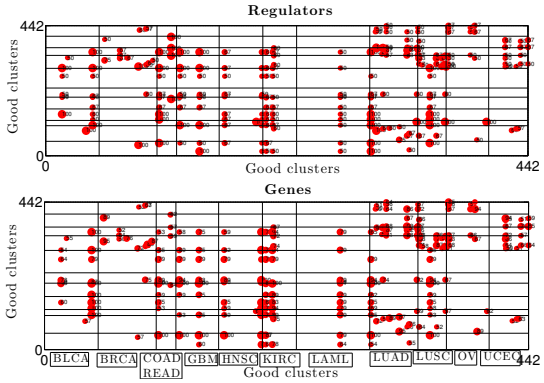


Fig. 6: Modified Jaccard index for clusters from different tumors, for both the genes and the regulators. The black lines separate clusters from different tumors.

| Type of tumors | # Genes | # Regulators | $R^2$ |
|---|---|---|---|
| BLCA, LUAD | 52 | 3 | 0.71 |
| BLCA, BRCA, LUAD | 53 | 1 | 0.56 |
| BRCA, LUAD, LUSC | 345 | 3 | 0.74 |
| KIRC, LUAD, LUSC | 155 | 3 | 0.63 |
| BRCA, KIRC, LUAD, LUSC | 435 | 2 | 0.63 |
| LUAD, LUSC, OV, UCEC | 104 | 2 | 0.62 |
| BLCA, HNSC, KIRC, LAML, LUAD | 56 | 1 | 0.68 |
| BLCA, COAD, GBM, HNSC, READ | 59 | 2 | 0.57 |

TABLE I: Some of the clusters found by Approach 2.

To find such clusters, we have first proposed two methods to perform unsupervised clustering of genes whose expression can be explained by a small number of regulatory (driver) genes. We have also developed two approaches that use the aforementioned methods as a baseline for finding clusters across patients of different tumors.

We have applied our methods on the Pan-Cancer dataset, which is composed of the gene expression of patients of 12 different tumors. We have shown that the methods proposed for the unsupervised clustering are able to find good clusters when applied to the Pan-Cancer dataset. Furthermore, the two approaches to find clusters across patients of different tumors have proven to work well also. We were able to find several clusters and to establish relations between tumors.

Whereas researchers have focused in the past on finding clusters across patients of the same tumor, recently there has been a significant interest in discovering shared molecular and genomic similarities across different tumors. We believe the algorithms and methods proposed in this paper, together with the found clusters, will provide some insight in towards this direction.

## References

[1] The Cancer Genome Atlas: `cancergenome.nih.gov/`.
[2] Weinstein, John N., et al. "The cancer genome atlas Pan-Cancer analysis project", *Nature genetics*, 45.10 (2013): 1113-1120.
[3] Vaquerizas, J. M., et al., "A census of human transcription factors: function, expression and evolution", *Nature Reviews Genetics*, 10(4), 252-263, 2009.
[4] Olivier Gevaert, et al., "Identification of ovarian cancer driver genes by using module network integration of multi-omics data", *Interface Focus*, 2013.
[5] Zou, Hui, et al., "Regularization and variable selection via the elastic net", Journal of the Royal Statistical Society, 67.2 (2005): 301-320.
[6] Sjstrand, Karl, et al. "Spasm: A matlab toolbox for sparse statistical modeling", Journal of Statistical Software. Accepted for publication (2012).
[7] Segal, Eran, et al. "Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data", Nature genetics 34.2 (2003): 166-176.
[8] Palla, Gergely, et al. "Uncovering the overlapping community structure of complex networks in nature and society", Nature 435.7043 (2005): 814-818.