Article

A Topic model analysis of TCGA transcriptomic data

Filippo Valle [1,*], Matteo Osella [1] and Michele Caselle [1]

[1] Physics Department, University of Turin and INFN, via P. Giuria 1, 10125 Turin, Italy

* Correspondence: filippo.valle@unito.it;

Abstract: Topic modelling is a widely used technique to extract relevant information from large arrays of data. The problem of finding a topic structure in a dataset was recently recognized to be analogous to the community detection problem in network theory. Leveraging on this analogy, a new class of topic modelling strategies has been introduced to overcome some of the limitations of classical methods. This paper applies these recent ideas to TCGA transcriptomic data on Breast and Lung cancer. The established cancer subtype organization is well reconstructed in the inferred latent topic structure. Moreover, we identify specific topics that are enriched in genes known to play a role in the corresponding disease and are strongly related to the survival probability of patients. Finally, we show that a simple neural network classifier operating in the low dimensional topic space is able to predict with high accuracy the cancer subtype of a test expression sample.

Keywords: network-based cancer data analysis; topic modelling; gene expression; network theory; stochastic block modelling;

1. Introduction

Thanks to the impressive progress in sequencing technologies and the development of dedicated gene expression databases like *TCGA* [1] it is now possible to study in an unified way the transcriptional status of thousands of cancer samples for several different biological conditions and cancer types. These transcriptomes provide a huge amount of information to link pathological phenotypes to their molecular underpinnings and can be used to identify and classify cancer types and subtypes, find new biomarkers and, as a final goal, elaborate new therapeutic strategies. The early identification of the particular cancer subtype of a given patient may help to customize "ad hoc" therapeutic protocols and may greatly improve the survival probability of patients [2].

To address this issue one must deal with two main steps. First, one must identify the molecular signatures of cancer types and, possibly, of their subtype organization, by suitably clustering gene expression data. Second, one should use these signatures to train a classifier to allow a fast and reliable association of a new sample to the corresponding subtype. As an extra bonus, one may hope in this way to identify specific drivers of the particular cancer subtype under study and reconstruct a map of the altered pathways which promote tumor growth and aggressiveness. In this process the real challenge is the first one. Finding signatures able to distinguish among different cancer subtypes is a highly non trivial task. From a theoretical point of view it is a typical example of a dimensionality reduction process. Starting from the huge dimensional space of gene expression data (with thousands of genes) one aims to find a few (or orders of dozens) of relevant subsets able to summarize the whole information content of the original dataset.

In general a good signature is a collection of genes which are able to capture a large portion of the variation in gene expression across tumors of the same type. These genes can thus be used to define and identify molecular subtypes of that cancer and may explain (and predict) the clinical heterogeneity of the disease (for instance different levels of overall survival). There are several methods for selecting gene signatures. Most of them are based on unsupervised clustering algorithms [3], with a wide variety of different implementations.

Despite the apparent simplicity of this program its actual realization is not so easy. It often happens that lists of genes obtained by different labs and putatively associated to the same cancer subtype

show almost no overlap among them [4]. This is due to the high heterogeneity at the molecular level of tumors [5] and the intrinsic complexity of performing dimensional reduction in the gene expression space.

In the past few years in order to address this issue a new class of clustering algorithm has been proposed based on the so called “topic modeling” approach [6]. This proposal stems form the observation that a similar degree of complexity and heterogeneity is also present in Natural Language Processing. Indeed algorithms which try to identify the “topic” of a given document from the word usage have to face the same type of challenges we are facing here. In this analogy the cancer samples play the role of the documents, the words are the genes, the number of times a particular word is used in a given document is the analogous of the expression level of a particular gene in a given sample and the topics are the gene sets (the "signatures") we use to cluster samples into subtypes. The goal of topic modeling is to identify the “topic" of a given document from the word usage within that document and exactly in the same way our goal is to identify the cancer subtype from the gene expression pattern. The major advantage of topic modeling methods with respect to standard clustering approaches is that they allow a “fuzzy" type of clustering [7]. The output of a typical topic modeling algorithm is a *probability distribution* of membership i.e. the probability of a given document to be composed by a given topic and at the same time the probability of a word to characterize a given topic. In our context this means that we have as output of our analysis a set of values which quantify the probability of a given sample to belong to a particular cancer subtype and the relevance of a given gene in driving this identification.

The most popular tool to perform this kind of analysis is the so called Latent Dirichlet Allocation (LDA) algorithm which assumes a Dirichlet distribution for the gene expression values in the samples. This assumption has no biological motivation, but allows a simple solution of the allocation problem and thus can be used even for databases with a large number of samples and genes. The lack of biological motivation represents the major limitation of LDA in this context.

Another common approach, often used in addressing gene expression data, is the Nonnegative Matrix Factorisation [8]. The major drawback of this approach is that it facilitates the detection of sharp boundaries among subtypes and this could be a limitation in very heterogeneous settings.

In recent years, some important advances in the field have laid the foundations for overcoming the limits of the LDA. First, it has been realized [9,10] that there is a strong connection between topic modeling analysis of complex databases and the community detection problem in bipartite graphs, which is a well know and much studied problem in network theory [11]. Second, a very effective class of community detection algorithms, based on hierarchical stochastic block modeling (hSBM), has been adapted to the topic modeling task [9].

The major advantage of hSBM type algorithms is that they do not require any particular assumption on the probability distribution of genes within samples and can thus adapt to the heterogeneous nature of gene expression data in cancer cell lines. As a consequence they do not require external inputs, like for instance the number of expected cluster (or topics) like it happens for LDA or in general for standard clustering algorithms but they are able to recognize the hierarchical organization of samples (and genes) within the database and thus give in output an optimal choice for the number of topics at different levels of resolution (i.e. the algorithm is able to identify, if it exists, a hierarchical organization of the cancer samples in the database). The major drawback of this class of algorithms is that they require a large amount of computing power and indeed it was only in the last few years, thanks to the improvement in computational power that it became feasible to address large and dense networks, like the ones in which we are interested, with hSBM methods.

In this paper we apply for the first time hSBM based topic modeling to the study of cancer gene expression data. This paper is part of an ongoing effort in our lab to explore the possible applications of advanced, network based, computational tools to the molecular characterization of cancer and, in particular, the identification of cancer drivers [12–15].

In this paper we concentrate in particular on breast and lung cancer, due to their clinical relevance and to the presence of a large number of studies which can be used to benchmark our result, but the methods that we developed could be in principle applied to any type of cancer.

Our main goal is to identify signatures for Breast and Lung cancer subtypes and then use these signatures to construct an efficient classifier to associate samples to their most probable subtype. We shall reach this goal by studying the gene content of the topics associated to the various subtypes. From the analysis of these signatures we shall learn a lot on the peculiar features of gene expression patterns in cancer cell lines.

By comparing our results with those of standard clustering based methods we shall show that hSBM can infer information, like the optimal number of clusters, that cannot be obtained with other approaches.

Finally, we will investigate how topics and clusters are related to the survival probability of patients. This can be accomplished only considering mixtures of topics or latent variables, it is not possible if using hard clustering methods.

2. Results

2.1. A topic modeling analysis of TCGA gene expression data

It has been recently realized [9] that there is a strong connection between topic modeling analysis of complex databases and the community detection problem in bipartite graphs. Among the several different approaches to community detection [11] one of the most powerful is the so called hierarchical Stochastic Block Model (hSBM) [16] which belongs to the class of probabilistic inference approaches to community detection and has the advantage of making the minimal amount of assumptions on the structure of data. A suitable implementation of this algorithm on bipartite networks [9] leads to a particularly effective version of topic modeling which is able at the same time to identify the hierarchical structure of the network and keep track of its bipartite nature. As a consequence it automatically detects the number of topics and hierarchically clusters both sides of the networks (in our case both genes and samples) thus overcoming the typical limitations of standard topic modeling approaches.

These properties are perfectly suited to deal with gene expression data, because they can address the heterogeneous nature of these data, while keeping track of their hierarchical structure (say, the organization in tissues or cell lines).

Moreover they are particularly relevant if one is interested in cancer gene expression data for which this hierarchical structure shows up in the cancer subtypes organization. In the framework of precision medicine it is of utmost importance to be able to identify cancer subtypes in a reliable and reproducible way to improve therapeutic treatment and predict survival probabilities. This is the main goal of this paper.

We address the problem of cancer subtype identification using a hSBM based topic modeling analysis of RNA-Sequencing samples from The Cancer Genome Atlas (TCGA). These data can be represented as a bipartite network relating genes $\{g_i\}$ with samples $\{s_j\}$. Each link of the network has a weight w_{ij} which encodes the expression level of the gene g_i in the sample s_j.

The output of hSBM is a hierarchical, probabilistic, organization of the data in "blocks". Each sample and each gene has in output a certain probability to belong to a given block (see the description of hSBM in the methods section). We shall define as *clusters* the blocks of samples and as *topics* the blocks of genes (see Figure 1). By construction genes are included in topics (and samples in clusters)

with a hard membership: each gene is associated only to a given topic and each sample only to a given cluster[1].

The whole complexity of the database is encoded in the probability distributions $P(\text{topic}|\text{sample})$ and $P(\text{gene}|\text{topic})$ (see the methods section for a precise definition).

- $\mathbf{P(topic|sample)}$ is the probability that a sample (or more precisely its overall gene expression pattern) is driven by the cooperative action of a particular set of genes (topic). These probabilities are not characterized by a hard membership but describe instead a "many to many" interaction (and for this reason they are able to capture the whole complexity of the problem), i.e. a given sample may be controlled by several different sets of genes (topics) and a given set of genes can influence different samples, possibly in different clusters (i.e. belonging to different cancer subtypes).

  To better understand this result it may be useful to address the analogy with the standard topic modeling analysis of linguistic corpora. As mentioned in the introduction in this comparison genes are mapped to words and samples to documents, then $P(\text{topic}|\text{sample})$ gives us the probability that a particular document (sample) deals with a given topic. Similarly in our case this probability tells us the relevance of a set of genes (topic) in describing the sample and, as a consequence, in driving its assignment to a particular phenotype (cluster).
- $\mathbf{P(gene|topic)}$ is instead the probability distribution of the different genes within a topic. We mentioned above that genes are organized within topics with a hard membership partition, however the genes within a topic are not all on the same ground, their assignment to the topic is weighted by the $P(\text{gene}|\text{topic})$ distribution, which allows us to identify within the topic those genes (the ones with a larger $P(\text{gene}|\text{topic})$) which play the most important role and are likely the drivers of the gene set effect on the samples. Indeed we shall see that by ranking the genes within a topic according to their assignment probability we shall identify as top ranking entries well known oncogenes.

  Keeping the above analogy with linguistic, $P(\text{gene}|\text{topic})$ is the analogous of the importance of a given word to define a topic. By selecting the top ranking words we may immediately understand the nature of the topic to which they are associated.

We shall concentrate in the following in particular on two cancer types: Breast and Lung (more precisely on the Non-Small-Cell Lung cancer, which represent the majority of lung cancers). This choice is motivated by their clinical importance but also because they represent two complementary challenges of our approach. While the subtype organization of Breast cancer is based on gene signatures the one of Lung cancer is of clinical nature. While for the Non Small Cell Lung cancer we have only two subtypes, in breast we must identify five different subtypes, which represent a much more complex and non trivial task.

We shall address in the following these two cases separately. We shall first evaluate the ability of hSBM to identify cancer subtypes and then we shall see what we can learn on the pathology (in terms of functional characterization of driving topics, new samples classification and survival prediction) from our analysis.

2.2. Analysis of Breast Cancer samples

Breast Cancer is one of the few examples of a tumor for which there is a widely accepted subtype classification [17,18] based on gene expression which has a relevant therapeutic role and allows for

[1] In principle the algorithm could be extended also to a "fuzzy" version of membership (in which, say, a sample has a certain probability to belong to a given cluster and at the same time a non-zero probability to belong to another cluster). While this option would certainly be interesting in our framework, it adds a further layer of complexity in the analysis and since the results with the hard membership choice turn out to be already interesting we decided to postpone to a forthcoming study the fuzzy membership option.

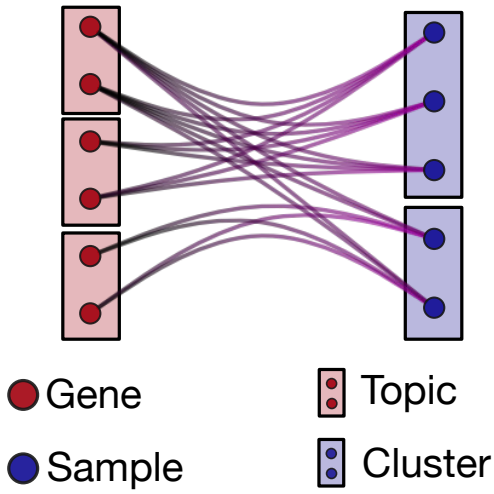

Figure 1. The hSBM partition of samples in clusters and genes in topics. The lines connecting genes and samples encode the weights of the bipartite network (i.e. the gene expression values in the different samples).

improved clinical outcomes (in particular for HER2 subtypes). It is thus a perfect setting for our analysis.

Breast cancer samples are usually divided in 5 different subtypes: *Luminal A*, *Luminal B*, *Triple-Negative/Basal*, *HER2* and *Normal*. On the clinical side this classification is based on the levels of a handful of proteins whose presence in the biopsy is usually detected using immunohistochemistry (IHC) assays. In particular two hormone-receptors (estrogen-receptor (ER) and progesterone-receptor (PR)), HER2[2] and Ki-67 [3].

Here is some information on these subtypes and their clinical outcomes.

- **Luminal A** breast cancer is hormone-receptor positive, HER2 negative, and has low levels of the protein Ki-67. Luminal A cancers are low-grade, tend to grow slowly and have the best prognosis.
- **Luminal B** is very similar to Luminal A for the gene expression point of view, the main difference is that it can be either HER2 positive or HER2 negative and is typically characterized by high levels of Ki-67. As a consequence Luminal B cancers generally grow slightly faster than Luminal A cancers and their prognosis is slightly worse.
- **Triple-negative/Basal** (which we shall simply denote as Basal in the following) are both hormone-receptor negative and HER2 negative.
- **HER2** is hormone-receptor negative and HER2 positive. This class of breast cancers tend to grow faster than the luminal ones and can have a worse prognosis, but they are often successfully treated with targeted therapies aimed at the HER2 protein.
- **Normal** breast cancer is similar to Luminal A: hormone-receptor positive, HER2 negative, and has low levels of the protein Ki-67, however its prognosis is slightly worse than Luminal A cancer's prognosis.

The same classification can be obtained (to a large extent [19]) looking at the expression levels of the well known "Prediction Analysis of Microarray (PAM)50" signature [20]. Given the expression levels of these signature genes, samples are then classified using standard machine learning methods

---

[2] The Human Epidermal growth factor Receptor 2 (HER2) is a growth-promoting protein and plays an important role in several signaling pathways.

[3] Ki-67 is a nuclear antigen expressed by all proliferating cells during late G1 through the M phases of the cell cycle, peaking in the G2-M and with a rapid decline after mitosis and is thus an indicator of cancer cells growth.

(Classification and Regression Trees (CART), Weighted Voting (WV), Support Vector Machine (SVM), Nonnegative Matrix Factorisation (NMF) or k-Nearest Neighbors (k-NN)) or using methods based on the euclidean distance in the signature space like Nearest Template Prediction (NTP) [21] or with more sophisticated network based methods like Hope4genes [15]. The agreement among different classifiers and with the IHC based subtyping is in general reasonably good but far from perfect (see for instance [15] for a recent comparison of the performances of different classifiers in a set of breast cancer classifications tasks). The reason is that the classification task is made particularly difficult by the heterogeneity of cancer tissues (biopsies may contain relevant portions of healthy tissue) and by the intrinsic variability of gene expression patterns in cancer cell lines cells. For instance, the TCGA samples that we shall use for our analysis have been recently reanalyzed in TCGABiolinks [22] leading to a significant relabeling of samples.

To address this particular issue we downloaded both the *PAM50* labels from [23] (which is the most widely used set of annotations) and the more recent and highly curated *SubtypeSelected* annotation provided by the new functionalities of TCGABiolinks [22]. In the following we shall compare the performance of our algorithms against both these annotations.

Our main goal in this framework is not to propose a new signature or a new classifier on top of the existing ones, but to show that it is possible to obtain relevant information on the cancer samples like subtype annotation or the survival probability and a potential list of driver genes and altered pathways **without** resorting to the marker genes mentioned above but looking instead to the **overall** gene expression pattern of the cell. We think this is an important achievement since it allows us to address breast cancer (and in principle any other complex pathology or cancer) without being influenced by the expression levels of few, often wildly fluctuating, marker genes, and opens the possibility to find new driver genes and possibly new subtype structures which may have unexpected therapeutic relevance.

We performed the hSBM analysis on a bipartite network starting from all the 1222 samples of the TCGA-BRCA project on one side and a suitable selection (see methods section) of genes on the other side; the links were weighted by the expression values.

2.2.1. Clustering of Breast Cancer samples

We clustered the 1222 samples of the TCGA-BRCA project using hSBM and a set of other state of art clustering tools: Latent Dirichlet Allocation (LDA) [24], Weighted Gene Correlation Network Analysis (WGCNA) [25] and hierarchical clustering (hierarchical) [26]. We also compared the quality of clustering with the two annotations *PAM50* [23] and *SubtypeSelected* [22]. On the gene side instead of looking to cancer specific markers or differentially expressed genes we selected only breast related genes i.e. genes whose behaviour was different in breast tissues with respect to other tissues (see method section). After this selection we ended up with 978 genes. Among these genes only HER2, among the above discussed markers, was present, but it plays no special role, and in our analysis its expression level is on the same footing of the other 977 genes. We stress again that our goal was to show that it is possible to identify relevant features of cancer samples without assuming previous knowledge on existing markers.

hSBM finds a first layer of clustering in which the samples are divided into 8 clusters and the genes are organized in 6 topics. Then it identifies a further, more refined, level of organization which is composed by 29 clusters and 41 topics [4] .

Results are reported in Figure 2. In Figure 2a and 2b we report the subtype organization in clusters for the first two layers (8 clusters for Figure 2a and 29 clusters for Figure 2b). We see looking at these

[4] The algorithm identifies then two further levels of partition with 149 and 1204 clusters and 147 and 399 topics respectively. These partitions on the cluster side convey little information and their NMI is very low. They correspond to the rightmost points in the figures 2b and c. The lowest partition level on the topic side will instead play an important role in the following when we shall discuss a classifier for breast cancer samples.

figures that hSBM is able to identify rather well Basal, Her2 and Normal samples, while it mixes together Luminal A and B. For this reason we shall treat them as a single subtype "Luminal" in the following.

We used the Normalised Mutual Information (NMI) as a score to evaluate the performance of the various algorithms in identifying cancer subtypes. NMI (see the methods section) is a powerful tool to compare the performances of clustering algorithms in describing labelled partitions and was recently used by Shi et al. [27] to evaluate the performances of different topic modeling algorithms on synthetic corpora.

Results for the NMI are reported in Figure 2d for the comparison among different clustering algorithms and in Figure 2c for the comparison of hSBM results with the two different annotations.

Looking at the figures we see that the highest NMI is reached for the first layer and that hSBM outperforms Weighted Gene Correlation Network Analysis (WGCNA) and hierarchical clustering. Interestingly, at both levels of resolution we find a higher value of NMI for the new, more refined, annotation of samples *SubtypeSelected* than for the old one *PAM50*.

Let us stress again that we obtained these results **without** resorting to the marker genes HER2, KI-67 and hormone receptors, which, except for HER2, were actually excluded from the gene set used in the analysis. Our results show that the expression levels of these genes are crucial to distinguish between the two Luminal types, but they are not mandatory to identify the remaining subtypes and separate them from Luminal A/B.

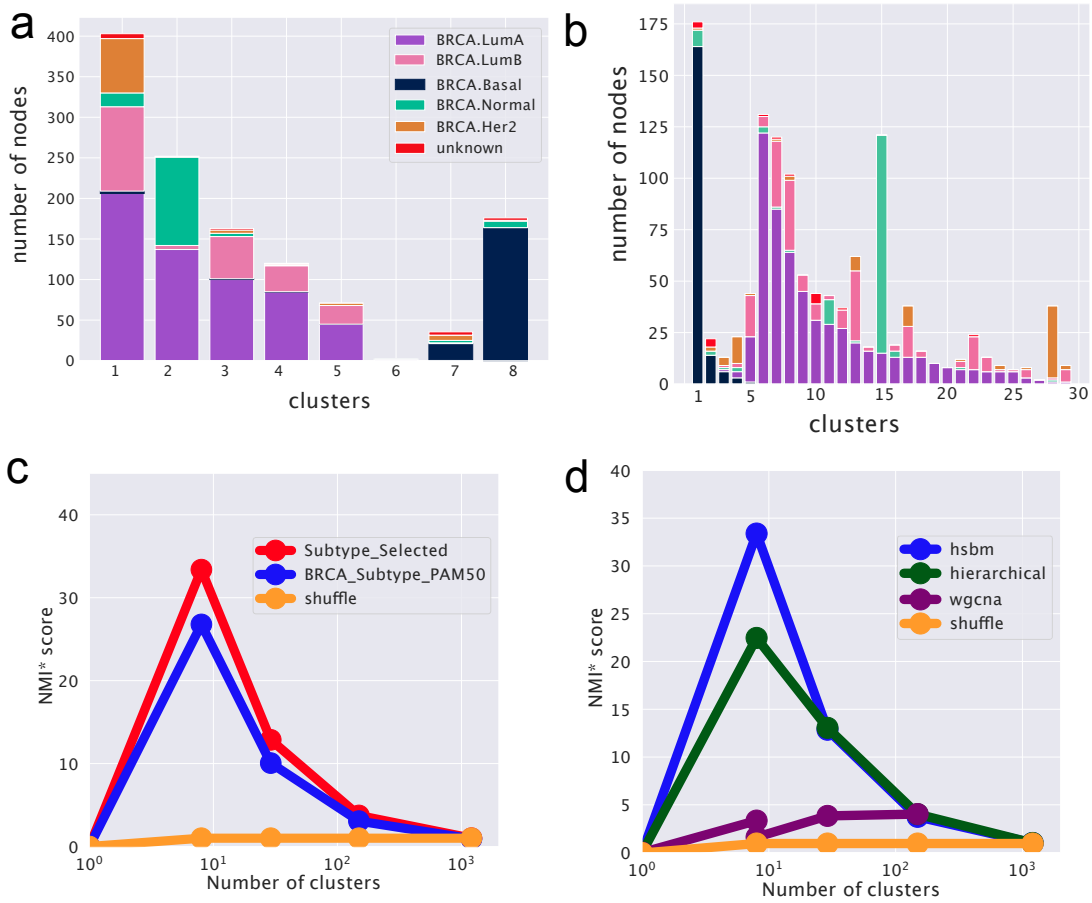

Figure 2. hSBM result for Breast cancer analysis. In (**a**) and (**b**) it is reported the subtype composition of the clusters obtained via hSBM. Each column is a cluster, each color is a *SubtypeSelected* label from TCGABiolinks. The height of each column is proportional to the number of samples within the cluster. In (**a**) we report the results for the first layer of clustering (8 clusters) and in (**b**) those for the second layer (29 clusters). (**c**) Comparison of normalized NMI scores across hierarchy between TCGABiolinks labels [22] and TCGA labels from [23]. (**d**) Comparison of normalized NMI scores for different clustering algorithms.

2.2.2. Gene expression pattern in the topic space and subtype specific topics

One of the advantages of using a topic modeling approach is that, as we mentioned above, each sample is a mixtures of topics which is described by the probability distribution $P(\text{topic}|\text{sample})$. From this distribution it is easy to obtain the probability $P(\text{topic}|\text{subtype})$ by averaging over all samples belonging to the same subtype the $P(\text{topic}|\text{sample})$. By subtracting to $P(\text{topic}|\text{subtype})$ the mean value over the whole dataset we find a new set of quantities, that we define as "centered" distributions $\overline{P}(\text{topic}|\text{subtype})$ (see eq.4) which allow to identify subtype specific topics i.e. topics which are particularly enriched in the samples belonging to a particular subtype and are thus candidates to play a role in driving the specific features of that subtype. We report as an example in Figure 3 the results of this analysis for one particular topic (labelled as "Topic 1") out of the 41 that we obtained at the second hierarchical level of our analysis. This topic turns out to be particularly enriched in the Normal subtype and slightly enriched in the Luminal A subtype. Other examples are reported in Figure 4.

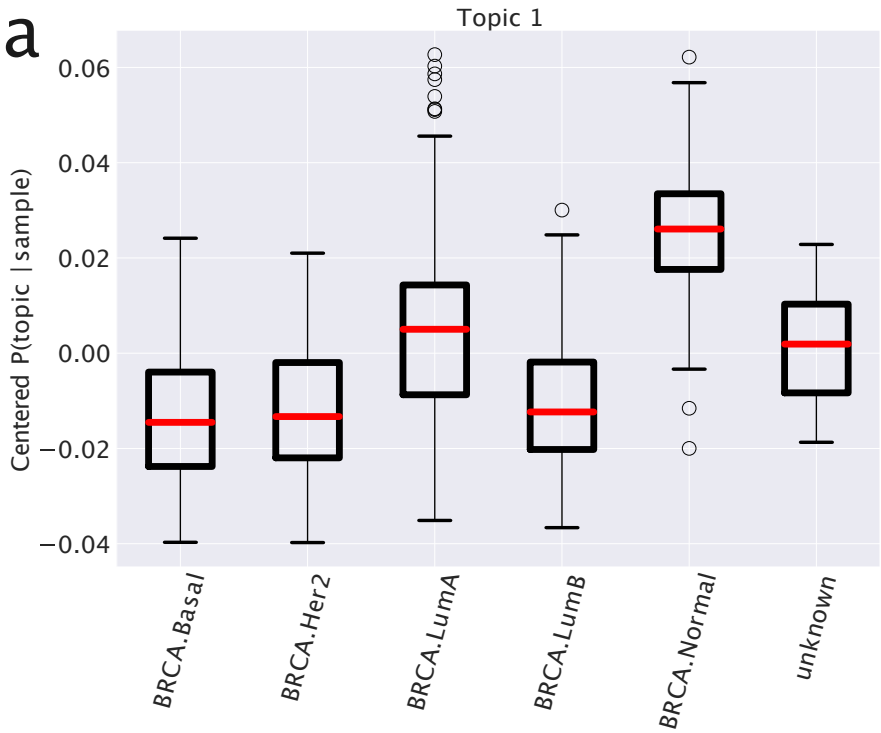

Figure 3. Values of $\overline{P}(\text{topic}|\text{subtype})$ for Topic 1 and the different subtypes (see text for a detailed explanation). This centered version of the probability distribution allows to recognise the differences of Topic expression in different subtypes.

2.2.3. Functional enrichment of the topics.

Topics are nothing but lists of genes, the common way to investigate their properties is by performing enrichment tests using tools like GSEA [28]. Performing this analysis we find enriched functional categories which are in remarkable agreement with the above analysis. We report some of our results in Table 1. For instance, (see the first entries of Table 1) for the Topic 1 mentioned above we find a strong enrichment for two sets of genes (labelled, following the GSEA convention as SMID_BREAST_CANCER_LUMINAL_A_UP and SMID_BREAST_CANCER_NORMAL_LIKE_UP) taken from [29] which fully agree with our subtype annotation in the topic space.

This correspondence between subtype annotation of topics and functional enrichment of their gene content is further confirmed by the other entries of Table 1 and Figure 4. In particular we see, looking for instance at the last entry ("Topic 20") of the table that it works also when the subtype is *depleted* in the topic space (see the last plot of Figure 4 in which the subtype Luminal B is depleted) to which corresponds an enrichment in *downregulated* genes in Luminal B Breast Cancer according to [29]. It is worth mentioning once again that in our analysis we selected only genes which are generically expressed in breast and not differentially expressed in breast cancer. This makes the above results a non trivial consistency check of our procedure and further supports our idea of a role of the whole gene expression pattern of the cell in driving breast cancer subtype phenotypes.

Table 1. Gene ontology results for a few selected topics, those with a number of genes large enough to allow a reliable Gene Set Enrichment Analysis. We report only the entries with the most significant enrichment. In brackets the number of genes in each set (topic). Many topics are enriched for terms related to particular subtypes of breast cancer and this relation is indeed confirmed by the box plots in Figure 3 and 4. In some other cases, like for instance for topic 13, the entries, (like for instance *epithelial mesenchymal transition*) are generic hallmarks of invasiveness or, possibly, of the metastatic nature of the sample and accordingly (see the box plot in Fig.4) they are shared by different subtypes.

Term	FDR q-value
Topic 1 (72)	
SMID_BREAST_CANCER_LUMINAL_A_UP	$3.03E-40$
SMID_BREAST_CANCER_NORMAL_LIKE_UP	$3.23E-15$
Topic 3 (50)	
SMID_BREAST_CANCER_BASAL_DN	$4.89E-9$
VANTVEER_BREAST_CANCER_METASTASIS_UP	$3.46E-8$
Topic 4 (24)	
SMID_BREAST_CANCER_NORMAL_LIKE_UP	$6.16E-11$
SMID_BREAST_CANCER_LUMINAL_B_DN	$1.34E-8$
Topic 5 (21)	
VANTVEER_BREAST_CANCER_ESR1_UP	$3.03E-40$
SMID_BREAST_CANCER_BASAL_DN	$3.23E-15$
SMID_BREAST_CANCER_LUMINAL_B_UP	$1.41E-6$
Topic 13 (129)	
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSINSITION	$7.8E-14$
SCHUETZ_BREAST_CANCER_DUCTAL_INVASIVE_VE_UP	$7.8E-14$
GO_EXTRACELLULAR_MATRIX	$6.9E-12$
Topic 14 (20)	
SMID_BREAST_CANCER_BASAL_DN	$1.47E-8$
SMID_BREAST_CANCER_LUMINAL_B_UP	$3.04E-6$
VANTVEER_BREAST_CANCER_ESR1_UP	$5.66E-4$
Topic 16 (19)	
MODULE_55	$2.24E-5$
Topic 18 (42)	
GO_CONTRACTILE_FIBER	$1.7E-4$
GO_SARCOPLASM	$3.56E-4$
Topic 20 (37)	
SMID_BREAST_CANCER_LUMINAL_B_DN	$1.7E-15$
SMID_BREAST_CANCER_BASAL_UP	$3.3E-10$

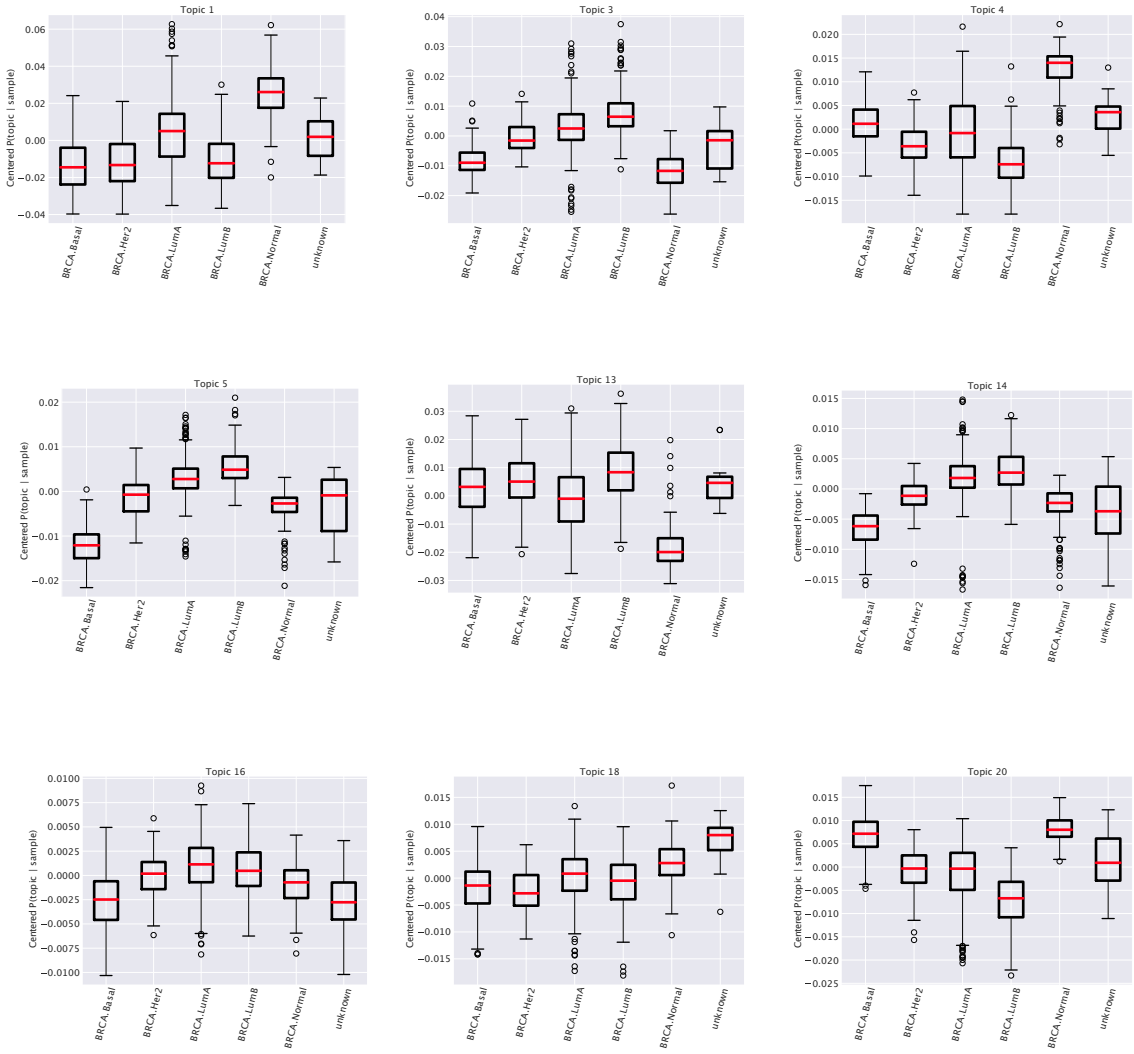

Figure 4. $\overline{P}(\text{topic}|\text{sample})$ for various topics at the second level of resolution of hSBM (the same of Figure 3). Looking at the box plots we recognise several non trivial relationships between topics and subtypes. These relationships are consistent with the functional enrichments listed in Table 1.

2.2.4. Predicting breast subtype annotation of new samples.

One of the advantages of a topic model approach is that it is also a dimensionality reduction process. Topics can be interpreted as new coordinates one can use to visualise and study the data.

We used the topic space as an embedding space to train a neural network model which can then be used as an efficient classifier to associate samples to their specific subtype. Using topics as features and *SubtypeSelected* as labels our task becomes a simple supervised learning classification problem. The use of the topic space simplifies the space to the classifier that therefore could be trained faster and with fewer parameters. Moreover, the topics have a non trivial biological meaning as shown before and this could not be accomplished running just a neural network. We obtained high accuracy results using a 399 dimensional topic space (the lowest level of the hierarchical organization of the topic space) coming from a space with almost 20000 genes. We report in Figure 5 some information on the performances of the classifier.

To evaluate the performances of our predictor we performed the same analysis using K-Nearest Neighbors which a standard and very popular tool in this context. It turns out that the performances of our predictor model (AUC = 0.98) are greater than those, on the same dataset of K-NN (AUC =

0.9). This tells us that the organisation of the samples into the original space is not trivial and that the projection into the topics space improves the ability of the predictor to assign the correct labels to test's samples. We reported the AUC score since it is less influenced by unbalanced classes than, for instance, the accuracy score. We applied K-NN using 5 `n_neighbors` and using *euclidean* `metric` on the $\log_2(FPKM+1)$ data.

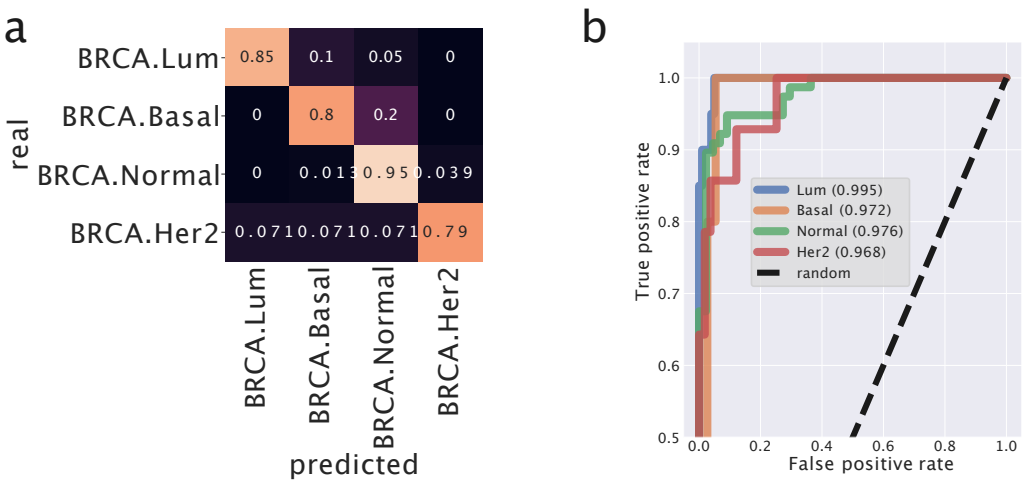

Figure 5. Predictor model for Breast cancer. We built a neural network and trained it on the low-dimensional topic space to classify the different breast subtypes. In (**a**) we report the confusion matrix. In (**b**) the Receiving Operation Characteristic curve and the corresponding Area Under Curve for each subtype estimated using a One-vs-All strategy four times. The diagonal ($TPR = FPR$) corresponding to random guessing is reported for reference. *Luminal* and *basal* subtypes are the ones with the lowest fraction of False Positives. *Normal* is the subtype with the highest fraction of True Positives.

2.3. *Analysis of Non-Small-Cell Lung Cancer samples*

To reveal the potentialities of topic modeling in a different context, we analysed Non-Small-Cell Lung cancer data taken again from TCGA. Lung cancer subtypes are currently defined by their pathological characteristics. The two predominant histological phenotypes of Non-Small-Cell Lung cancer are adenocarcinoma and squamous cell carcinoma [30]. TCGA-LUAD and TCGA-LUSC projects provide transcripts for samples of these two subtypes. In the same way as in the breast analyses we collected FPKM data with GDC's tools. Then 3000 highly variable genes were selected. This setting represents a much easier task for a clustering algorithm and indeed, as we shall see, hSBM is able to correctly separate LUAD from LUSC. The TCGA repository on lung cancer data allows for an unexpected and non-trivial test of clustering algorithms. Indeed, Cline et al. in [31] observed that some samples from TCGA-LUSC have gene expression levels that are more similar to LUAD than LUSC, although their similarity to LUAD is modest. This suggested them that these samples may be near the borderline for subtype classification, for example, tumors that are less differentiated, therefore difficult to classify by pathology. They provide a list of samples labelled as *Discordant LUSC*. In our analyses we searched how hSBM classifies those samples. Finally in the context of Lung cancer, thanks to the recent work of [32], we may perform another non trivial test of our topic modeling approach, combining together healthy and cancer tissues and looking to the ability of hSBM to separate healthy samples from cancer ones.

2.3.1. Classification of Non-Small-Cell Lung cancer subtypes: LUAD and LUSC

Running hSBM on the above data we found three different layers of resolution with 2, 8 and 58 clusters and 5, 12 and 42 topics respectively.

We show the results of our analysis in Figure 6 and 7. We see looking at Figure 6a that hSBM is able to separate well the two subtypes and that indeed most of the *Discordant LUSC* samples are clustered together with the LUAD ones, somehow catching the fact that they are more similar to LUAD than LUSC. It is instructive to follow the hierarchical organization of cluster (see Figure 6b). Already in the first layer many LUAD samples are separated from LUSC, in the next layers the separation is complete.

As we did in the Breast case we can study the distribution of subtype probabilities across topics and their enrichment in LUAD and LUSC samples. We report these results in Figure 8

2.3.2. Classification of LUAD and LUSC samples versus healthy tissue

We also tested the ability of clustering algorithms and in particular hSBM to separate healthy form cancer tissues. We downloaded data with healthy (taken from The Genotype Tissue Expression [33] GTEx project) and cancers samples provided in a unified framework by [32,34]. We selected only samples with valid metadata available from TCGABiolinks or GTEx. Looking at Figure 6c we see that also in this case hSBM identifies LUAD and LUSC subtypes and that both are separated from healthy samples. Moreover, again, the majority of *Discordant LUSC* samples are clustered with LUAD. Even if apparently the task of identifying three categories instead of two is harder, it seems that including healthy tissues the performances of hSBM are greatly enhanced. Looking at Figure 7b we see that in this case the scores are one order of magnitude higher than without healthy tissues.

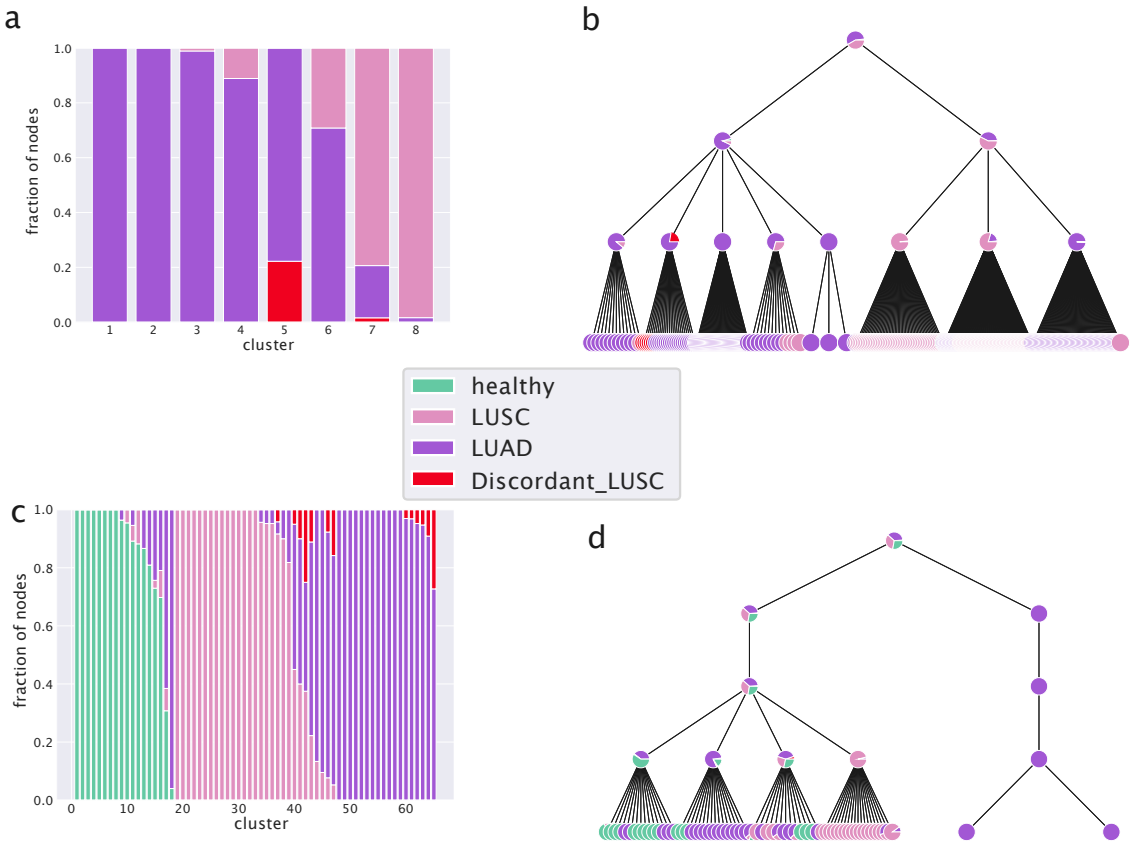

Figure 6. The hSBM classification of Non-Small-Cell Lung cancer subtypes. In (**a**) the columns represent clusters and the colours refer to the sample annotations. Columns are normalized to the total number of samples in the cluster so that the height of different portions of the column are proportional to the fraction of LUAD or LUSC samples in that cluster. (**b**) The hierarchical structure. Already in the first layer many LUAD samples are separated from LUSC, in the next layers the separation is complete. In (**c**) and (**d**) we report the results of hSBM analysis including also healthy samples. In both settings *Discordant LUSC* are classified with LUAD.

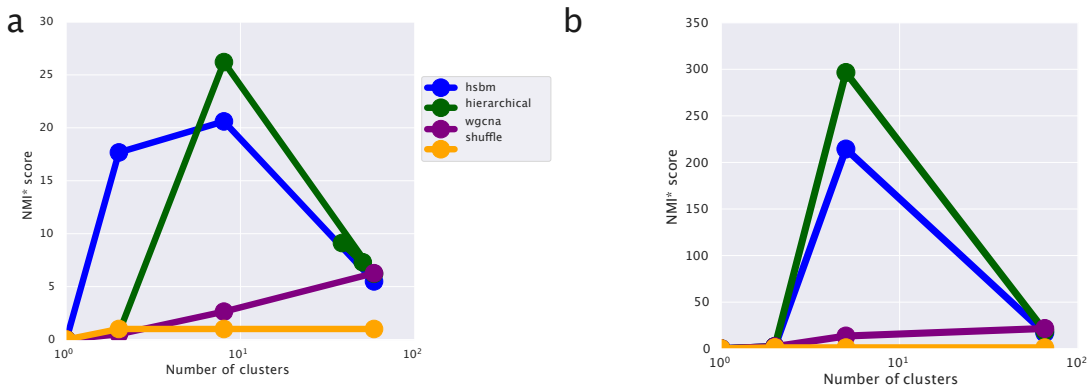

Figure 7. Comparison of different clustering algorithms. In (**a**) we report the scores for the the classification of Non-Small-Cell Lung cancer subtypes without healthy tissues. In (**b**) the scores in presence of healthy tissues. The scores in (b) are one order of magnitude greater than those in (a): adding healthy samples improves the classification.

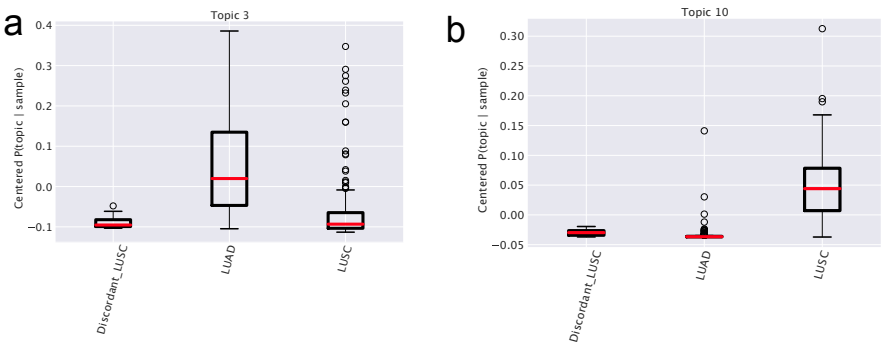

Figure 8. Topic trends in Lung subtypes. In (**a**) and (**b**) there are the values of $\overline{P}(\text{topic}|\text{subtype})$ for two topics grouped by the different subtypes (LUAD, LUSC and Discordant LUSC. The topic enrichment in different subtypes emerges.

2.3.3. Predicting LUAD and LUSC

The topics embedding space can be used to build a predictor similarly to what was done for Breast, in this case the goal is being able to classify LUAD and LUSC.

This predictor is actually a neural network with one hidden layer composed by 20 neurons and an output layer activated by a sigmoid function. We report in Figure 9 the output of this model on the test set. LUAD and LUSC are classified with high accuracy (accuracy: 0.9268, AUC: 0.9493). Also in this case we compared our results with a standard K-NN predictor. In this case K-NN achieves slightly better performances (accuracy: 0.9756, AUC: 0.9733). This is probably due to the fact that this task, which involves only a binary choice, is simpler than the one we studied in the breast cancer case. A major advantage of the topic based predictor with respect to the K-NN one is that since the last layer is a sigmoid function one can extract the probability of being LUAD despite LUSC and this can be helpful in classifying borderline samples. In Figure 10, for example, it is reported the output of the whole predictor Z. Z can be interpreted as the probability that a sample is LUSC (the assignment $LUSC = 1$ and $LUAD = 0$ is completely arbitrary); it is interesting the fact that *Discordant LUSC* samples have a score near $0.3 - 0.4$: they have a probability to be LUSC greater than LUAD samples, but small. This is an example of the use of the predictor to assign labels to ambiguous samples.

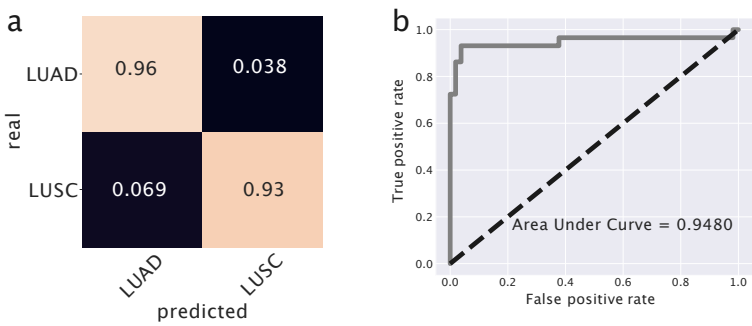

Figure 9. Prediction model for Lung cancer. In (**a**) we used topics as features to train this model and we report the confusion matrix. In (**b**) we report the Receiving Operation Characteristic curve of the test set of this model. The Area Under Curve is reported as a score.

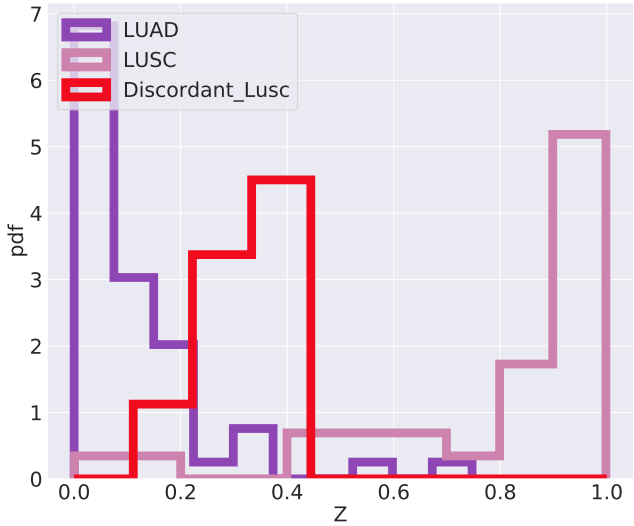

Figure 10. Output of the last layer of the predictor. Z is the output of the sigmoid function on the last layer $Z = \sigma(z)$, namely Z is the probability of being LUSC.

3. Discussion

We saw that a topic modeling analysis is able to extract a lot of important information from cancer gene expression data. This information is encoded in the topic distribution and more precisely in the probability distributions $P(\text{topic}|\text{sample})$ and $P(\text{gene}|\text{topic})$. We have shown that by projecting the data in the topic space it is possible to build efficient predictors which allow to assign samples to the correct subtype. Moreover, by looking at the distribution of genes within the topics it is possible to extract relevant functional information on the subtypes.

We shall see now that if we include in the picture also some additional information on the stage of the tumor and on the follow up of the patients there are some further relevant clinical information that we can extract from the projection of the tumor samples into the topic space.

We shall see two examples, one for Breast and one for Lung.

3.1. Survival analysis for Breast cancer

Looking at Table 1 we see that one of topics: "Topic 13" is characterized by the typical annotations of invasive or already metastatic tumor. Looking at Figure 4 we see that this topic is almost equally distributed across all subtypes. However such a condition should obviously have an important and pervasive effect in the transcriptome and hSBM should be able to detect this effect. Thus we looked at the probability distribution of this topic on the clusters. We constructed the distributions $\overline{P}(\text{topic}|\text{cluster})$ in whole analogy with what we did for $\overline{P}(\text{topic}|\text{subtype})$. We plot them in Figure 11. This time we see a rather pronounced effect with a cluster which is relatively enriched: cluster 5 and one in which the topic is strongly depleted, cluster 2 [5].

At this point we can make a separate survival analysis for each cluster. We report in Figure 12 the results for the two clusters. We evaluated the significance of these results with a standard z-value (obtained by running 1000 times a random reshuffling of samples annotations and keeping the size of cluster unchanged) which is reported in the figure. The same analysis made at 3 years gives even higher z-values, in particular 3.13 for cluster 5 and 4.59 for cluster 2. Only these two clusters had

[5] Cluster 6, which is also depleted, does not allow a statistical analysis since it contains too few samples.

a significant z-value. We see an impressive anticorrelation between the relevance of topic 13 in the cluster and the survival probability of the patients in the cluster. In particular, in cluster 2, a depletion of topic 13 was associated to a significant enhancement of survival probability.

Interestingly, a similar analysis on the subtypes always gave a non-significant z-value in agreement with the above observation that Topic 13 is evenly distributed across subtypes. These analyses are reported in the supplementary materials.

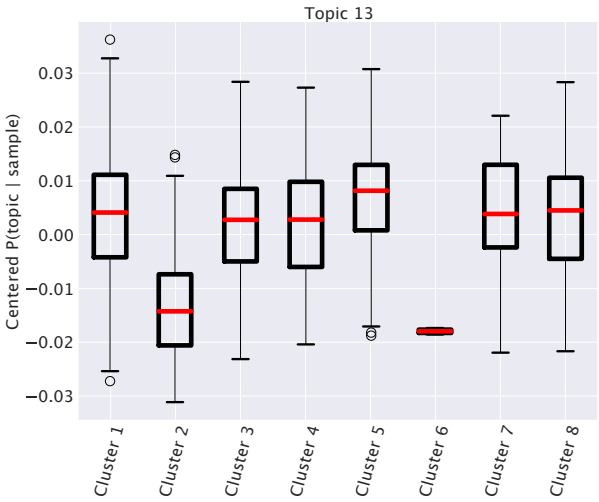

Figure 11. Topic 13's expression in different clusters. $\overline{P}(\text{topic 13}|\text{sample})$ for samples divided into clusters. The mixtures of samples of cluster 2 contains a lower fraction of topic 13 with respect to samples of others clusters.

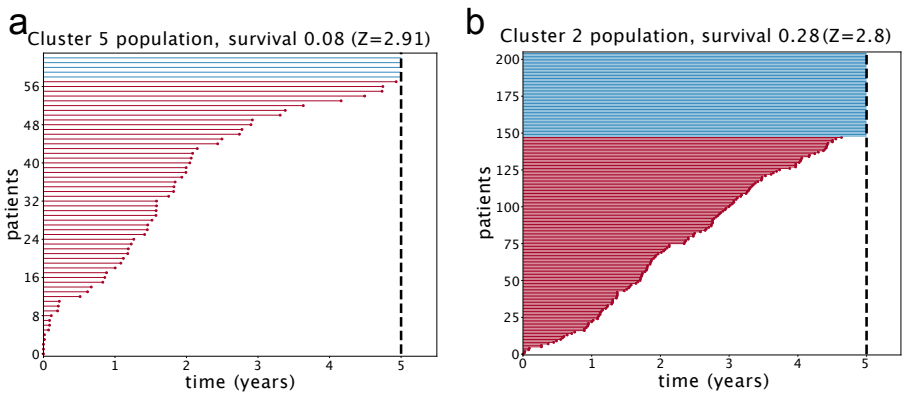

Figure 12. Survival time for two different clusters of patients. Each horizontal line is a patient; in red the ones who died or had the last follow up within 5 years from diagnosis, in blue the others. The survival is estimated as the fraction of patients that are alive after 5 years.

We used the `plot_lifetimes` function of the lifelines package to realise the plot in Figure 12. The *duration* argument was set to the number of days from diagnosis to death or to last follow up; the *event_observed* parameter was set to 1 for dead patients.

3.2. Survival analysis for LUNG cancer

Following a suggestion of Lucchetta et al. [35] we used the information on the tumor stage which is contained in the TCGA database to perform a survival analysis on the topic space.

Recently Lucchetta et al. [35] searched groups of genes effective in separating LUAD and LUSC. Their protocol consisted in extracting the genes that showed a trajectory of up-regulation across stages in one cancer type and down-regulation in the other.

Since topics are effectively group of genes we attempted a similar approach. We searched topics whose pattern across stages in the two different tumor subtypes was different.

In particular we focused on the topic (labelled as "topic 3") in Figure 13, which shows, after an initial common trend, an up-regulation pattern through stages in LUAD and a down-regulation pattern in LUSC. We choose it as the candidate for the upcoming survival analysis. We provide in supplementary Figure S1 the same analyses for all the other topics.

The *lifelines* Python package was used to perform a survival analysis. We performed an estimation of the survival probability $S(t)$, *cox* [36] was used to model the hazard function (the risk of dying at a certain time) and estimate how the survival probability is related to this topic. We defined up-regulated a topic in a sample if the $P(\text{sample}|\text{topic})$ is above a threshold that corresponds to the 50th percentile. We found that patients which had our candidate topic up-regulated had twice the probabilities to die with respect to the other patients. In order to test the significance of this results we performed the same analysis using the *gender* variable instead of the topic up-regulation and found essentially no change in the survival probability between the two classes. As a further test we performed the same analysis for all the other topics at the same hierarchical level (see Fig. S1 in the supplementary material). None of these other topics shew a different trend between LUAD and LUSC and accordingly, for all of them, we found no significant difference in survival in patients in which these topics were up-regulated (see Fig. S1).

At this point we extracted the list of genes contained within topic 3 to see if we could understand the origin of this different survival probability. As we mentioned above, one of the advantages of topic modeling is that groups are not hard-constrained but are actually mixtures weighted by $P(\text{gene}|\text{topic})$: genes with the highest $P(\text{gene}|\text{topic})$ are likely to be those which contribute most to the topic.

We sorted the genes in this topic by $P(\text{gene}|\text{topic})$ and selected the first ones. Looking at the DISEASES database [37] we found that they are all related to lung cancer and lung diseases. Many terms like *cancer*, *bronchitis*, *pneumonia*, *interstitial lung disease* emerged.

Interestingly, the list of the genes contained in the topic has a very small overlap with the one found by Lucchetta et al. [35] even if both sets are able to successfully separate LUAD from LUSC and to predict different survival probabilities between the two. In particular, they analysed 2953 genes in 4 sets divided in up/down and LUAD/LUSC, we filtered 3000 genes; only 633 genes are in common, actually just 20% of genes are shared by the two analyses.

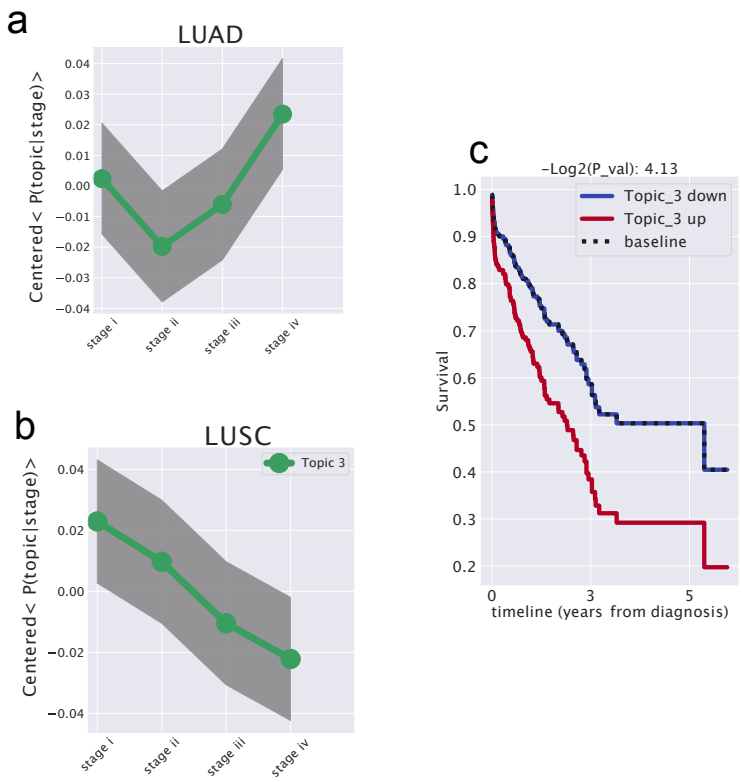

Figure 13. A candidate topic that affects patients' survival probability and shows different trends in LUAD and LUSC. In (**a**) the trend of $P(\text{topic}|\text{stage})$ in LUAD. In (**b**) the trend of $P(\text{topic}|\text{stage})$ in LUSC. The genes belonging to this topic can be considered as a signature which distinguish the evolution across stages of the two subtypes. (**c**) The output of the survival analysis. When the topic is up-regulated the survival probability is halved.

Table 2. Genes in our candidate topic. Genes are sorted by their contribution to the topic. The complete list is available at https://github.com/fvalle1/topicTCGA/blob/master/lung/topic_3_lung.csv

Gene	$P(\textbf{gene}\|\textbf{topic})$
SFTPB	0.24
SFTPA2	0.21
SFTPA1	0.17
RNASE1	0.05
NAPSA	0.04
A2M	0.02
SERPINA1	0.02
...	...

3.3. Some final Remarks

There are three properties of the hSBM topic modeling analysis that we think are at the basis of the effectiveness of this approach and correspondingly there are three lessons that we can learn from our work.

- hSBM imposes the minimal amount of assumptions on the statistical distribution of gene expression values across samples [9]. This is a generic and important feature which one should always require in clustering (or topic modeling) algorithms when addressing complex systems. Complex systems are always characterized by power law probability distributions of entries [38] and different a priori assumptions for these distributions (like for instance those assumed by LDA) can drive the algorithm far from the actual information content of the system [9].

It emerged that even with few assumptions, hSBM outperforms standard algorithms like WGCNA (when performing clustering) or K-NN (when assigning labels to test samples).

- Cell lines phenotype are driven by the expression pattern of the whole set of genes and not by a handful of markers. This is true in general, but it is even more true for complex diseases like cancer. The breast subtype classification is based on the expression levels on a handful of genes but, notwithstanding this, hSBM was able to reproduce the classification to a large extent without resorting to marker genes. Clearly this is telling us that breast cancer subtypes are driven not only by the expression level of one or two genes but by a whole pattern of pathways alterations, which is highlighted by the topic distribution within the different clusters. This has important consequences on the way we approach gene signatures.

  While it is certainly true that gene signatures play an important role in cancer medicine, it is also clear that they are not the end of the story. It is not strange to find different gene signatures aimed to identify the same cancer subtype which have almost no overlap (we have seen a prototypical example in the Lung survival analysis above). This supports the importance of data mining tools able to address the overall behaviour of the transcriptome and not driven by the gene expression pattern of a single gene signature.
- Looking at all our findings we see that the actual information on the complex disease is completely encoded in the topic space and in particular in the probability distributions $P(\text{topic}|\text{sample})$ and $P(\text{gene}|\text{topic})$. This is the probably the most important lesson that we learn from our analysis. The most effective way to address a complex system and in particular a complex disease like cancer is through a stochastic approach in which the relationships between samples and genes are of probabilistic nature. In hSBM these complex associations are mediated by the intermediate layer of topics. This is a very effective solution but non necessarily the only one. Innovative data mining tools should keep into account this observation and allow for fuzzy and complex memberships of samples across topics and genes across clusters.

We think that these lessons could be useful beyond the specific tool (hSBM) that we discussed in this paper. They should drive the conception of a new generation of innovative, network based, data mining approaches, maybe combining the best features of the different tools that we compared in this paper, notably WGCNA and hierarchical clustering. This is a mandatory task if one wants to address the challenging issue of combining the different layers of information which are available in complex databases like TCGA, say, the microRNA expression levels, the mutational content of genes or the epigenetic layer of regulation. A fruitful combination of these different sources of information could greatly enhance our ability to address a personalized therapeutic approach to complex disease like cancer.

4. Materials and Methods

Normalised Mutual Information (NMI)

We used the Normalised Mutual Information *NMI* [39] to evaluate the performance of different protocols to cluster together cancer subtypes. Given a set C of labelled samples and a partition K in clusters of these samples, the NMI is defined as the harmonic average of homogeneity h and completeness $\mathcal{C}$:

$$NMI = 2\frac{h * \mathcal{C}}{h + \mathcal{C}} \ ,$$

where the homogeneity is defined as

$$h = 1 - \frac{H(C|K)}{H(C)}$$

and the completeness as

$$\mathcal{C} = 1 - \frac{H(K|C)}{H(K)} \ .$$

In these definitions $H(C)$ and $H(K)$ are the usual Shannon entropies associated to the partitions C and K; $H(C|K)$ and $H(K|C)$ are defined as:

$$H(C|K) = -\Sigma_{c\in C,k\in K}\frac{n_{ck}}{N}\log\left(\frac{n_{ck}}{n_k}\right)$$

and

$$H(K|C) = -\Sigma_{c\in C,k\in K}\frac{n_{ck}}{N}\log\left(\frac{n_{ck}}{n_c}\right)$$

respectively, where n_c is the number of samples of the cancer subtype c, n_k the number of samples in the cluster k and n_{ck} the number of samples of the subtype c in the cluster k. With this definition it is easy to understand the meaning of homogeneity and completeness. If all the samples in cluster k belong to the cancer subtype c then $n_k = n_{ck}$ and $h = 1$. Similarly the completeness $\mathcal{C}$ equals 1 if all samples belonging to the cancer subtype c are in the same cluster k. A similar approach involving *NMI* was proposed in [27] to evaluate topic modeling performance in reconstructing synthetic corpora.

Due to its definition also randomly shuffled sets may have a nonzero NMI (which steadily increases with the number of clusters). To keep into account this effect we normalized our NMI values dividing them for those obtained with a simple null model which preserves the number of clusters and their sizes, but reshuffles the samples' labels.

TCGA data

The results published here are in part based upon data generated by The Cancer Genome Atlas (TCGA) managed by the NCI and NHGRI. Information about TCGA can be found at http://cancergenome.nih.gov.

We downloaded data from TCGA using tools provided by Genomic Data Commons (GDC) [40]. We downloaded *Gene Expression Quantification* data type in *transcriptome profiling* category. We choose *RNA - Seq* with *HTSeq - FPKM* as workflow type. We downloaded the 1222 samples from TCGA-BRCA project and 1145 from TCGA-LUSC and TCGA-LUAD projects.

During the analyses of Lung cancer we considered only the 408 samples with a subtype annotation available in the TCGABiolinks GUI [41].

TCGABiolinks and metadata

We downloaded metadata for TCGA's samples using the TCGABiolinks GUI at *Version:2.17.1* [41,42].

During the analyses of Breast cancer we downloaded both the *SubtypePam50* classification labels provided by [23], and *SubtypeSelected* obtained via the `PanCancerAtlas_subtypes` function of TCGABiolinks. `PanCancerAtlas_subtypes` gives access to a curated table retrieved from synapse [22]. We discussed the performance of hSBM in classifying the two.

Unified dataset

We downloaded data for lung from a dataset prepared by Wang et al. [32]. They processed data from GTEx and TCGA with the same pipeline and successfully corrected for study-specific biases, enabling comparative analysis. We downloaded the second version of their normalised data from figshare [34].

Only samples with a valid annotation were considered and this left us with 1415 samples from LUAD, LUSC and GTEx. We applied a $\log_2(FPKM + 1)$ transformation, this reduced the number of edges E and let the algorithm to be faster even with a large number of nodes N.

In our repository we provide the code to correctly preprocess the data in order to be loaded by the model.

Gene filtering

As mentioned above we filtered genes with two different strategies in the two settings.

Breast

We searched genes whose behaviour was different in breast and other tissues. We used the following procedure. Firstly we estimated the mean of gene g in tissue T (Breast) $m_g^T = \frac{1}{|T|}\Sigma_{s\in T} n_{gs}$, being s the samples and n_{gs} the expression value, the mean in non-breast is $m_g^{nT} = \frac{1}{R-|T|}\Sigma_{s\notin T} n_{gs}$ being R the total number of samples. Similarly we estimated the variance of breast's samples $v_g^T = \frac{1}{|T|}\Sigma_{s\in T}\left(n_{gs} - m_g^T\right)^2$. We defined the distance between the means of this two distributions $d_g^T = \frac{|m_g^T - m_g^{nT}|}{\sqrt{v_g^T}}$. The genes with highest d_g were selected. Moreover, we got only genes that satisfied filters (e.g. exceeding certain minimum expression levels) that were used during eQTL analyses by the GTEx project whose list was published by [7].

Lung

In the analyses of lung cancer we used a standard approach searching highly variable genes using *scanpy* Python package [43] and selected the 3000 most variable genes.

Hierarchical Stochastic Block Model

We adapted hierarchical Stochastic Block Model (hSBM) to gene expression data. The original code to run hierarchical Stochastic Block Model on a bipartite network was provided by [9] in the repository available at: https://github.com/martingerlach/hSBM_Topicmodel/tree/develop.

Hierarchical stochastic block model is a kind of generative model that tries to maximise the probability that the model θ describe the data $\mathcal{A}$

$$P(\theta|\mathcal{A}) = P(\mathcal{A}|\theta)P(\theta) \tag{1}$$

using a non-parametric approach. In the setting described in this paper $\mathcal{A}$ is the gene expression matrix and the entries $\mathcal{A}_{ij}$ represents the number of FPKM of gene i in sample j. In other words $\mathcal{A}$ is the adjacency matrix of a bipartite network composed by genes and samples, the edges of this network are weighted by the gene expression. The `minimise_nested_blockmodel_dl` function from the graph-tool package [44] minimises the description length $\Sigma = -lnP(\mathcal{A}|\theta) - lnP(\theta)$ of the model. We used the nested version of the model since we expected some sort of hierarchical structure in the data [16,45,46].

We set the algorithm to minimise the description length Σ many times and selected the model that obtained the shortest description length.

As output of the model we find the probability distributions $P(\text{topic}|\text{sample})$ and $P(\text{gene}|\text{topic})$. These probabilities are defined, in terms of entries of the program as follows:

$$P(\text{topic}|\text{sample}) = \frac{\text{number of half-edges on sample coming from topic}}{\text{number of half-edges on sample}} \tag{2}$$

and

$$P(\text{gene}|\text{topic}) = \frac{\text{number of half-edges to topic going to gene}}{\text{number of half-edges to topic}}. \tag{3}$$

The complexity of of hSBM is $O(VLn^2V)$ if the graph is sparse, i.e. if $E \sim O(V)$ [16], where V is the number of vertices (samples and genes) and E the number of edges. For $E >> V$ the complexity increases and the CPU time needed to minimize the description length can become prohibitive. In this case, to reduce the CPU bottleneck, one can apply a log-transformation to the data, which strongly reduces the number of edges E.

In our setting we have $V = O(1\,000)$ vertices, $E = O(100\,000)$ edges and the network is indeed very dense, being $\frac{E}{V} \sim 10^3$. In the Breast case, thanks to the strong reduction in the number of genes, we could face the task of running the program in its original setting. In the Lung case, we kept 3000 genes in input and to analyze them we had to log-transform the data.

Investigate the enrichment of the topics

We centered the $P(\text{topic}|\text{sample})$ obtaining

$$\overline{P}(\text{topic}|\text{sample}) = P(\text{topic}|\text{sample}) - \frac{1}{R}\Sigma_{s\in\text{samples}}P(\text{topic}|\text{s}), \tag{4}$$

being R the total number of samples. This centered $P(\text{topic}|\text{sample})$ can be represented with a box plot, after grouping samples by their subtype.

A topic is nothing but a list of genes, it can be investigated using hypergeometric tests. The results shown in this paper are computed using the GSEA [28] tool.

Survival analysis

We performed survival analysis on lung, we wanted to find how topics are related to the survival probability of a patient.

Our analysis began with the list of the mixtures $P(\text{topic}|\text{sample})$. Samples were annotated due to their subtype (LUAD or LUSC), all samples without a valid cancer type or stage label were dropped, this includes *nan* or *not reported* values. We cleaned up the labels removing any additional letter (e.g *stage ia* became *stage i*) and ended up with four stages *i*, *ii*, *iii* and *iv*.

We averaged each table over stages and obtained $P(\text{topic}|\text{stage})$ for each dataset. $P(\text{topic}|\text{stage}) = \frac{1}{|\text{stage}|}\Sigma_{p\in\text{stage}}P(\text{topic}|p)$ being $|\text{stage}|$ the number of patients p labelled stage. We subtracted the mean to normalise the data and obtained $\overline{P}(\text{topic}|\text{stage}) = P(\text{topic}|\text{stage}) - \frac{1}{4}\Sigma_{\text{s}\in\{\text{i,ii,iii,iv}\}}P(\text{topic}|\text{s})$. The analyses of this $\overline{P}(\text{topic}|\text{stage})$ was used to identify the topics with different behaviour in LUAD and LUSC.

Using GDC tools we downloaded TCGA metadata and in particular: *vital_status*, *days_to_last_follow_up* and *days_to_death*. We estimated the lifetime or the number of days the patient survived after the diagnosis, using *days_to_last_follow_up* if the patient was *Alive* and *days_to_death* for *Dead* patients. A similar approach was recently utilised by [35].

In order to estimate whether a topic is up regulated in a patient, we evaluated the 50th percentile of $P(\text{sample}|\text{topic})$ and considered it as a threshold thr. Then we engineered a feature called up as follows:

$$up(\text{sample}) = \begin{cases} 1 & P(\text{topic}|\text{sample}) > thr \\ 0 & P(\text{topic}|\text{sample}) \leq thr \end{cases} \tag{5}$$

We used these data to fit the hazard with a *cox* model using the `COXPHFitter` module. We used the lifetime, the vital status and the new feature as input for the `fit` function.

The *cox* model quantified how the up regulation of the topic affected the survival probability. *Cox* fits the hazard function conditioned to a variable $h(t|x) = b_0(t) * e^{\Sigma_{i=0}^{n} b_i * (x_i - \bar{x}_i)}$. x is the new feature described above. The hazard is defined as the ratio of the derivative of the survival and the survival itself $h(t) = \frac{-S'(t)}{S(t)}$. $S(t)$ is the probability of being alive at time t, namely the number of patient alive at time t divided by the total number of patients. The package estimated the ratio between the hazard of samples with topic up-regulated and hazard of samples with topic not up-regulated. So we were able to estimate the $\exp(coef)$ or hazard ratio $\exp(coef) = \frac{\text{hazard of samples with topic up-regulated}}{\text{hazard of samples with topic not up-regulated}}$. Note that the *coef* does not depend on time, but it is a sort of weighted average of period-specific hazard ratios. In our test, we obtained an $\exp(coef)$ of 1.8, meaning that patients with topic up regulated have almost twice the chance to die with respect to the remaining patients. The baseline is the case in which

changes of topic does not affect the survival. The P-value reported is the test against this null model without other temporal dependencies. We checked also the *gender* variable to compare and it obtained a lower P-value ($-\log_2(\text{P-value}) = 1.6$).

These analyses were performed using *lifelines* Python package[47].

Predictor

In order to build the predictor for cancer subtypes we prepared the design matrix X with samples on the rows and topics on columns. The element X_{ij} is the $P(\text{topic}_j|\text{sample}_i)$. The model was trained using Stochastic Gradient Descent so we need to apply normalisation to the matrix. We followed the standard procedure and obtained the normalised matrix $\overline{X}$ with entries $\overline{X_{ij}} = \frac{X_{ij} - \langle X_{ij'} \rangle_{j'}}{0.5 * \left(\max_{j'} X_{ij'} - \min_{j'} X_{ij'}\right)}$. We built a one-hot encoded vector of labels corresponding to subtypes. In this analyses we dropped samples with an undefined label. We randomised and split the normalised data into two sets: a training and a test set. Training set was split again to obtain a validation set. In the breast setting the sizes of $[\text{train}, \text{validation}, \text{test}]$ were in proportions of $[0.18, 0.72, 0.1]$ and in lung $[0.48, 0.32, 0.2]$. The models we built in the two settings, breast and lung, are similar with some different parameters.

The model consisted of a neural network with one hidden layer. In order to keep the model as simple as possible, Stochastic Gradient Descent was selected as the optimiser in both cases. We set the loss function to be the *crossentropy*. The model was built using keras [48].

In case of breast the input layer had a dimension of 399, namely the number of topics; the hidden layer was built with 50 neurons activated by *ReLU* and the output layer consisted of 4 (one for each subtype, recall that Luminal A and B were merged into a single subtype) neurons activated by the *softmax* function.

Hyper-parameters were searched maximising the $F1$ score on the validation set. $F1 = 2 * \frac{\text{precision} * \text{recall}}{\text{precision} + \text{recall}}$, being precision $= \frac{\text{true positives}}{\text{predicted positives}}$ and recall $= \frac{\text{true positives}}{\text{positives}}$.

We obtained an accuracy of 0.9008 and an Area Under Curve of 0.9798 on the test set.

In the analysis of lung we prepared the design matrix X with the same process described above. The input layer consisted of 326 neuron, the hidden layer was instead built with 20 neurons activated by *ReLU* functions, and finally, the output layer consisted of a single neuron activated by a *sigmoid* function. The output was a binary array that distinguished LUAD and LUSC.

We obtained an accuracy of 0.9268 and an area under curve of 0.9493 on the test set.

In both cases a confusion matrix and a Receiving Operating Characteristic curve were constructed. The confusion matrix and the ROC curves were estimated using scikit learn [49]. The ROC curve represents the True Positive Rate or sensitivity ($\frac{TP}{TP+FN}$) versus the False Positive Rate or $1 -$ specificity ($\frac{FP}{FP+TN}$) when varying the threshold on the score Z that defines the classes given the outputs of the hidden layer z. In lung Z is the value of the sigmoid function in the last layer $\sigma(z)$. In Breast we used a one-vs-all strategy and considered the softmax (σ_c) values as probabilities for a given class c, $Z = \sigma_c(z)$.

Implementation of WCGNA and hierarchical algorithms

In this work some of the analysis required other clustering methods. Weighted Gene Correlation Network Analysis was performed using the dedicated R package available at https://cran.r-project.org/web/packages/WGCNA/index.html. This was run using default parameters. WGCNA creates *modules* of genes, we considered them as topics. In order to obtain clusters we cut the tree built using modules to estimate distances between samples.

Hierarchical clustering was performed using *sklearn* and we set the model to use `euclidean` metric and `complete` linkage. In this case we cut the tree to match the hSBM number of clusters. In this case it has not been possible extract any information on the genes.

Code availability

The codes, notebooks and data to reproduce this work are available on a GitHub repository at https://github.com/fvalle1/topicTCGA.

Funding: This work was partially supported by the "Departments of Excellence 2018 - 2022" Grant awarded by the Italian Ministry of Education, University and Research (MIUR) (L.232/2016).

Acknowledgments: We would like to acknowledge the Competence Centre for Scientific Computing C^3S which provided us the access to the computing cluster OCCAM. The results shown here are in part based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

hSBM	hierarchical stochastic block model
TP, FP, TN, FN	True Positives, False Positives, True Negatives, False Negatives
FDR	False Discovery Rate
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
GSEA	Gene Set Enrichment Analysis

References

1. Chang, K.; Creighton, C.; Davis, C.; others. The cancer genome atlas pan-cancer analysis project. *Nature Genetics* **2013**, *45*, 1113.
2. Ashley, E.A. Towards precision medicine. *Nature Reviews Genetics* **2016**, *17*, 507–522. doi:10.1038/nrg.2016.86.
3. Eisen, M.; Spellman, P.; Brown, P.; Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **1998**, *95*, 14863–14868. doi:10.1073/pnas.95.25.14863.
4. Ein-Dor, L.; Kela, I.; Getz, G.; Givol, D.; Domany, E. Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics* **2005**, *21*, 171–178. doi:10.1093/bioinformatics/bth469.
5. Andor, N.; Graham, T.A.; Jansen, M.; Xia, L.C.; Aktipis, C.A.; Petritsch, C.; Ji, H.P.; Maley, C.C. Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nature Medicine* **2016**, *22*, 105+. doi:10.1038/nm.3984.
6. Liu, L.; Tang, L.; Dong, W.; Yao, S.; Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *Springerplus* **2016**, *5*. doi:10.1186/s40064-016-3252-8.
7. Dey, K.K.; Hsiao, C.J.; Stephens, M. Visualizing the structure of RNA-seq expression data using grade of membership models. *PLoS Genetics* **2017**, *13*. doi:10.1371/journal.pgen.1006599.
8. Brunet, J.; Tamayo, P.; Golub, T.; Mesirov, J. Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences* **2004**, *101*, 4164–4169. doi:10.1073/pnas.0308531101.
9. Gerlach, M.; Peixoto, T.P.; Altmann, E.G. A network approach to topic models. *Science Advances* **2018**, *4*, eaaq1360. doi:10.1126/sciadv.aaq1360.
10. Lancichinetti, A.; Sirer, M.I.; Wang, J.X.; Acuna, D.; Koerding, K.; Amaral, L.A.N. High-Reproducibility and High-Accuracy Method for Automated Topic Classification. *Physical Review X* **2015**, *5*. doi:10.1103/PhysRevX.5.011007.
11. Fortunato, S.; Hric, D. Community detection in networks: A user guide. *Physics Reports* **2016**, *659*, 1–44. doi:10.1016/j.physrep.2016.09.002.
12. Cantini, L.; Isella, C.; Petti, C.; Picco, G.; Chiola, S.; Ficarra, E.; Caselle, M.; Medico, E. MicroRNA-mRNA interactions underlying colorectal cancer molecular subtypes. *Nature Communications* **2015**, *6*. doi:10.1038/ncomms9878.
13. Cantini, L.; Medico, E.; Fortunato, S.; Caselle, M. Detection of gene communities in multi-networks reveals cancer drivers. *Scientific Reports* **2015**, *5*, 17386.

14. Cantini, L.; Caselle, M.; Forget, A.; Zinovyev, A.; Barillot, E.; Martignetti, L. A review of computational approaches detecting microRNAs involved in cancer. *Frontiers in Bioscience-Landmark* **2017**, *22*, 1774–1791. doi:10.2741/4571.
15. Cantini, L.; Caselle, M. Hope4Genes: a Hopfield-like class prediction algorithm for transcriptomic data. *Scientific Reports* **2019**, *9*. doi:10.1038/s41598-018-36744-y.
16. Peixoto, T.P. Hierarchical Block Structures and High-Resolution Model Selection in Large Networks. *Physical Review X* **2014**, *4*, 011047. doi:10.1103/PhysRevX.4.011047.
17. Perou, C.; Sorlie, T.; Eisen, M.; van de Rijn, M.; Jeffrey, S.; Rees, C.; Pollack, J.; Ross, D.; Johnsen, H.; Akslen, L.; Fluge, O.; Pergamenschikov, A.; Williams, C.; Zhu, S.; Lonning, P.; Borresen-Dale, A.; Brown, P.; Botstein, D. Molecular portraits of human breast tumours. *Nature* **2000**, *406*, 747–752. doi:{10.1038/35021093}.
18. Sorlie, T.; Perou, C.; Tibshirani, R.; Aas, T.; Geisler, S.; Johnsen, H.; Hastie, T.; Eisen, M.; van de Rijn, M.; Jeffrey, S.; Thorsen, T.; Quist, H.; Matese, J.; Brown, P.; Botstein, D.; Lonning, P.; Borresen-Dale, A. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences* **2001**, *98*, 10869–10874. doi:{10.1073/pnas.191367098}.
19. de Ronde, J.J.; Hannemann, J.; Halfwerk, H.; Mulder, L.; Straver, M.E.; Peeters, M.J.T.F.D.V.; Wesseling, J.; van de Vijver, M.; Wessels, L.F.A.; Rodenhuis, S. Concordance of clinical and molecular breast cancer subtyping in the context of preoperative chemotherapy response. *Breast Cancer Research and Treatment* **2010**, *119*, 119–126. doi:10.1007/s10549-009-0499-6.
20. Parker, J.S.; Mullins, M.; Cheang, M.C.U.; Leung, S.; Voduc, D.; Vickery, T.; Davies, S.; Fauron, C.; He, X.; Hu, Z.; Quackenbush, J.F.; Stijleman, I.J.; Palazzo, J.; Marron, J.S.; Nobel, A.B.; Mardis, E.; Nielsen, T.O.; Ellis, M.J.; Perou, C.M.; Bernard, P.S. Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes. *Journal of Clinical Oncology* **2009**, *27*, 1160–1167. doi:{10.1200/JCO.2008.18.1370}.
21. Hoshida, Y. Nearest Template Prediction: A Single-Sample-Based Flexible Class Prediction with Confidence Assessment. *PLoS ONE* **2010**, *5*. doi:{10.1371/journal.pone.0015543}.
22. Mounir, M.; Lucchetta, M.; Silva, T.C.; Olsen, C.; Bontempi, G.; Chen, X.; Noushmehr, H.; Colaprico, A.; Papaleo, E. New functionalities in the TCGAbiolinks package for the study and integration of cancer data from GDC and GTEx. *PLoS Computational Biology* **2019**, *15*, e1006701. doi:10.1371/journal.pcbi.1006701.
23. Koboldt, D.; Fulton, R.; McLellan, M.; others. Comprehensive molecular portraits of human breast tumours. *Nature* **2012**, *490*, 61. doi:10.1038/nature11412.
24. Blei, D.M.; Ng, A.Y.; Jordan, M.I. Latent dirichlet allocation. *Journal of Machine Learning Research* **2003**, *3*, 993–1022. doi:10.5555/944919.944937.
25. Langfelder, P.; Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **2008**, *9*, 559. doi:10.1186/1471-2105-9-559.
26. Ward, J.H.J. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **1963**, *58*, 236–244, [https://amstat.tandfonline.com/doi/pdf/10.1080/01621459.1963.10500845]. doi:10.1080/01621459.1963.10500845.
27. Shi, H.; Gerlach, M.; Diersen, I.; Downey, D.; Amaral, L. A new evaluation framework for topic modeling algorithms based on synthetic corpora. Proceedings of Machine Learning Research; Chaudhuri, K.; Sugiyama, M., Eds. PMLR, 2019, Vol. 89, *Proceedings of Machine Learning Research*, pp. 816–826.
28. Subramanian, A.; Tamayo, P.; Mootha, V.K.; Mukherjee, S.; Ebert, B.L.; Gillette, M.A.; Paulovich, A.; Pomeroy, S.L.; Golub, T.R.; Lander, E.S.; Mesirov, J.P. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences* **2005**, *102*, 15545–15550, [https://www.pnas.org/content/102/43/15545.full.pdf]. doi:10.1073/pnas.0506580102.
29. Smid, M.; Wang, Y.; Zhang, Y.; Sieuwerts, A.M.; Yu, J.; Klijn, J.G.M.; Foekens, J.A.; Martens, J.W.M. Subtypes of breast cancer show preferential site of relapse. *Cancer Research* **2008**, *68*, 3108–3114. doi:{10.1158/0008-5472.CAN-07-5644}.
30. Chen, Z.; Fillmore, C.M.; Hammerman, P.S.; Kim, C.F.; Wong, K.K. Non-small-cell lung cancers: a heterogeneous set of diseases. *Nature Reviews Cancer* **2014**, *14*, 535–546. doi:10.1038/nrc3775.
31. Cline, M.S.; Craft, B.; Swatloski, T.; Goldman, M.; Ma, S.; Haussler, D.; Zhu, J. Exploring TCGA pan-cancer data at the UCSC cancer genomics browser. *Scientific Reports* **2013**, *3*, 2652. doi:10.1038/srep02652.

32. Wang, Q.; Armenia, J.; Zhang, C.; Penson, A.V.; Reznik, E.; Zhang, L.; Minet, T.; Ochoa, A.; Gross, B.E.; Iacobuzio-Donahue, C.A.; others. Unifying cancer and normal RNA sequencing data from different sources. *Scientific Data* **2018**, *5*, 180061. doi:10.1038/sdata.2018.61.
33. Lonsdale, J.; Thomas, J.; Salvatore, M.; Phillips, R.; Lo, E.; Shad, S.; Hasz, R.; Walters, G.; Garcia, F.; Young, N.; others. The genotype-tissue expression (GTEx) project. *Nature Genetics* **2013**, *45*, 580–585. doi:10.1038/ng.2653.
34. Wang, Q.; Gao, J.; Schultz, N. Unified RNA-seq datasets in human cancers and normal tissues - normalized data, 2017. doi:10.6084/m9.figshare.5330593.v2.
35. Lucchetta, M.; da Piedade, I.; Mounir, M.; Vabistsevits, M.; Terkelsen, T.; Papaleo, E. Distinct signatures of lung cancer types: aberrant mucin O-glycosylation and compromised immune response. *BMC Cancer* **2019**, *19*, 824. doi:10.1186/s12885-019-5965-x.
36. Cox, D.R. Regression models and life-tables. *Journal of the Royal Statistical Society* **1972**, *34*, 187–202. doi:10.1111/j.2517-6161.1972.tb00899.x.
37. Pletscher-Frankild, S.; Pallejà, A.; Tsafou, K.; Binder, J.X.; Jensen, L.J. DISEASES: Text mining and data integration of disease–gene associations. *Methods* **2015**, *74*, 83–89. doi:10.1016/j.ymeth.2014.11.020.
38. Mazzolini, A.; Gherardi, M.; Caselle, M.; Lagomarsino, M.C.; Osella, M. Statistics of Shared Components in Complex Component Systems. *Physical Review X*.
39. Rosenberg, A.; Hirschberg, J. V-measure: A conditional entropy-based external cluster evaluation measure. Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning, 2007, pp. 410–420.
40. Grossman, R.L.; Heath, A.P.; Ferretti, V.; Varmus, H.E.; Lowy, D.R.; Kibbe, W.A.; Staudt, L.M. Toward a shared vision for cancer genomic data. *New England Journal of Medicine* **2016**, *375*, 1109–1112. doi:10.1056/NEJMp1607591.
41. Silva, T.C.; Colaprico, A.; Olsen, C.; Malta, T.M.; Bontempi, G.; Ceccarelli, M.; Berman, B.P.; Noushmehr, H. TCGAbiolinksGUI: A graphical user interface to analyze cancer molecular and clinical data. *F1000Research* **2018**, *7*, 439. doi:10.12688/f1000research.14197.1.
42. Colaprico, A.; Silva, T.C.; Olsen, C.; Garofano, L.; Cava, C.; Garolini, D.; Sabedot, T.S.; Malta, T.M.; Pagnotta, S.M.; Castiglioni, I.; others. TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Research* **2016**, *44*, e71–e71. doi:10.1093/nar/gkv1507.
43. Wolf, F.A.; Angerer, P.; Theis, F.J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biology* **2018**, *19*, 15. doi:10.1186/s13059-017-1382-0.
44. Peixoto, T.P. The graph-tool python library. *Figshare* **2014**. doi:m9.figshare.1164194.v14.
45. Peixoto, T.P. Efficient Monte Carlo and greedy heuristic for the inference of stochastic block models. *Physical Review E* **2014**, *89*, 012804. doi:10.1103/PhysRevE.89.012804.
46. Peixoto, T.P. Nonparametric Bayesian inference of the microcanonical stochastic block model. *Physical Review E* **2017**, *95*, 012317. doi:10.1103/PhysRevE.95.012317.
47. Davidson-Pilon, C.; Kalderstam, J.; Jacobson, N.; Zivich, P.; Kuhn, B.; Williamson, M.; Moncada-Torres, A.; Stark, K.; Anton, S.; Noorbakhsh, J.; others. CamDavidsonPilon/lifelines: v0. 24.2. *Context* **2017**, *604*.
48. Chollet, F.; others. Keras. https://keras.io, 2015.
49. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M.; Duchesnay, E. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.