Comprehensive analysis of long non-coding RNAs in breast cancer using topic modeling.

Malagoli Gabriele[1,2], Valle Filippo[2], Barillot Emmanuel[1], Caselle Michele[2,*] and Martignetti Loredana[1,*]

[1] Institut Curie / Inserm U900 / Mines ParisTech / PSL Research University, Paris, 75248, France
[2] Physics Department, University of Turin and INFN, Turin, 10125, Italy

* To whom correspondence should be addressed. Tel: +33 01 56 24 69 24; Fax: +33 01 56 24 69 80; Email: loredana.martignetti@curie.fr

Correspondence may also be addressed to: Tel: +39 011 670 72 05 ; Fax: +39 011 670 72 14 ; Email: caselle@to.infn.it

ABSTRACT

Single-cell RNA sequencing is a powerful tool to explore cancer heterogeneity. However, the expression of lncRNAs in single cells is still to be studied extensively and methods to deal with the sparsity of this type of data are lacking. Here, we propose a topic modeling approach to investigate the transcriptional heterogeneity of luminal and triple negative breast cancer cells using patient-derived xenograft models of acquired resistance to chemotherapy and targeted therapy.

We show that using an integrative clustering that combines the information coming from mRNAs and lncRNAs treated as disjoint omic layers greatly improves the accuracy of cell classification. Topics associated with specific breast cancer subpopulations show a clear enrichment for pathways involved in subtyping and progression of breast cancer and to sets of lncRNA encoded in the open chromatin regions of breast cancer cell lines. We identified lncRNAs strongly associated with cell clusters already well known in the literature, such as MALAT1 and NEAT1, and highlighted some others that may be clinically relevant.

INTRODUCTION

The recent development of next-generation sequencing (NGS) technologies has revealed a complex transcriptional landscape in human cells with thousands of long non-coding RNAs (lncRNAs) without protein coding potential (1, 2). In healthy tissues, they have been shown to have lower expression, increased tissue-specificity, and greater expression variability across individuals than protein-coding genes (3, 4). In cancer, lncRNAs are commonly dysregulated and their aberrant expression has been shown to be tumor-type specific (5, 6). However, the expression of lncRNAs in single cells is still to be studied extensively. A single molecule, single cell RNA FISH survey of 61 lncRNAs in three human cell types showed that the low abundance of lncRNAs in bulk population measurements is not due to a small subpopulation of cells expressing lncRNAs at high levels, and overall lncRNA are no different than mRNA in their levels of cell-to-cell heterogeneity (7).

A larger scale study based on Smart-seq-total technology (8) has assayed a broad spectrum of mRNAs and lncRNAs from single cells and showed that the lncRNA content of cells significantly differs across cell types and dynamically changes throughout cellular processes such as cell cycle and cell differentiation. Another large-scale study, combining bulk tissue RNA-seq and scRNA-seq to deeply profile lncRNA expression during neocortical development, found that many lncRNAs are specific to distinct cell types and abundantly expressed in individual cells (9), thus supporting that cell type-specific expression of lncRNAs contributes to the low levels of lncRNAs observed in tissues.

Existing computational methods for analysing single cell RNA-seq (scRNA-seq) data do not consider the differences between protein-coding mRNAs and lncRNAs in terms of data dispersion and sparsity. In particular, unsupervised clustering, a fundamental step of scRNA-seq data analysis, typically suffers from large fractions of observed zeros. In comparison to the analysis of mRNAs, the computational analysis of lncRNAs at single cell is more challenging due to the high sparsity of the data.

To capture the heterogeneity of lncRNA expression at single cell level, we used a new clustering approach called *topic modeling* (10–12), a method originally designed to analyse the words in a corpus of text documents to discover the themes, or topics, that pervade them. Through an analogy between document analysis and single-cell analysis and guided by the fact that similar statistical laws emerge in both of these contexts (13), we applied topic modeling to discover the protein coding genes and the lncRNAs that group thousands of cells into biologically similar clusters. In this sense, the topics are the sets of genes (signatures) associated with the different clusters of cells.

Importantly, the algorithm we used for topic modeling, called hierarchical Stochastic Block Modeling (hSBM) (14, 15), is able to determine the optimal number of clusters at different levels of resolution (i.e., the algorithm is able to reconstruct the hierarchical organization of cells into clusters) without a priori setting the number of expected clusters and provides the features (i.e. protein coding genes and lncRNAs) significantly associated with the identified clusters. A nested SBM algorithm has been shown to outperform state-of-art clustering algorithms on sparse and noisy single cell measurements (16).

In our study, we used hSBM to investigate the transcriptional heterogeneity of luminal and triple negative breast cancer cells using patient-derived xenograft models of acquired resistance to chemotherapy and targeted therapy (17). First, we clustered the cells based on the expression of mRNAs and lncRNAs separately. We then used a multiomic version of hSBM (18) to perform an integrative clustering that combines the information coming from the two RNA families treated as disjoint omic layers. We show that this strategy improves cell classification compared to clustering obtained by investigating mRNAs and lncRNAs separately but also compared to clustering obtained on the expression measures of mRNAs and lncRNAs joined in a single matrix.

Then, we investigated the gene content associated with cell clusters through functional enrichment analysis. To do this, we have developed a procedure to assign the most specific topics to each cluster and query them through gene set functional annotation databases (see Methods). The topics have been interpreted in terms of functionally related gene sets and lncRNA sets associated with upstream and downstream target regulators. The results show a clear enrichment of topics for pathways involved in subtyping and progression of breast cancer and to sets of lncRNA encoded in the open chromatin regions of breast cancer cell lines.

With the topic modeling approach, we have identified some lncRNAs associated with breast cancer cell subpopulations already well known in the literature, such as MALAT1 and NEAT1, and highlighted some others that may be clinically relevant.

RESULTS

1. Clustering of drug-sensitive and resistant breast cancer cells using hSBM

To explore intra-tumor transcriptional heterogeneity in breast cancer, we used a dataset of scRNA-seq profiles of patient-derived xenograft (PDX) models of luminal and triple negative breast cancer with acquired drug resistance (17). After scRNA-seq data pre-processing (see Methods), we obtained two expression matrices containing respectively 1959 mRNAs and 1136 lncRNAs profiled in 3723 cells from drug-sensitive and resistant tumors. We translated each expression matrix into a weighted bipartite network where nodes represent genes (mRNAs or lncRNAs) and cells. Genes are linked to the cells with weighted edges based on their level of expression. We applied the hSBM algorithm to calculate the best partition separately for the two bipartite networks. The output of hSBM is a hierarchical and probabilistic organization of the network in *blocks* of connected genes and cells. For each level of the hierarchy, we obtain three matrices of probability membership, corresponding respectively to the probability of cells to be associated to a given cluster (which is a delta function since we set the model to output “hard” clusters), the probability of association of genes to topics, and the probability of association of a given topic to a given cell. In our context, the output of hSBM informs us about the best partitioning of cells into clusters with similar transcriptional profiles and about which genes are most strongly associated with these clusters.

The partition of the network based on mRNA expression shows three hierarchical levels of cell classification (Fig. 1A,B,C). In the first level, the cells are classified into two clusters containing respectively the luminal and triple negative tumor cells, regardless of their resistant or drug-sensitive phenotype. The next level of clustering resolution separates cells into nine clusters, three of which contain a mixture of sensitive and resistant cells, suggesting that heterogeneous populations of sensitive and resistant cells emerged. The largest of these mixed clusters includes 999 triple negative sensitive and resistant cells, while none of the mixed clusters contain both triple negative and luminal cells. The third level of classification separates cells into 49 clusters, all containing cells from specific tumor types.

Also the partition of the lncRNA-based network exhibits three hierarchical levels of cell classification (Fig. 2A,B,C). Similar to the clustering of the mRNA network, the first level classifies cells in two clusters with luminal and triple negative tumor cells. However, the second and third hierarchical levels show some differences compared to the classification based on mRNAs. The second hierarchical level separates cells into seven clusters, all containing a mixture of resistant and sensitive cells. The largest cluster includes 1333 luminal cells and none of the clusters include both triple negative and luminal cells. The third level of classification separates cells into 249 clusters, many of which include both

resistant and sensitive cells. The high number of clusters containing a limited number of cells that we get for the network with the lncRNAs supports their cell specificity expression.

To evaluate the performance of the hSBM clustering algorithm, we calculated a score based on the normalized mutual information (NMI) between the partition found by the algorithm and the tumor of origin of the cells (see Methods). The NMI/NMI* score decreases as clustering resolution increases, as cells from the same tumor separate into different clusters. At the first and second resolution levels, NMI/NMI* scores are comparable for clustering obtained with lncRNAs and that obtained with mRNAs, while at the third resolution level, NMI/NMI* is an order of magnitude higher for clustering obtained with mRNAs. The low NMI/NMI* score obtained for the third level of resolution of clustering with lncRNAs is consistent with the fact that their expression is highly cell-specific and therefore not very informative of the type of tumor. Instead, the classification of cells obtained at the second level resolution provides information strongly correlated with tumor types. Notably, the mRNA-based clustering classified cells differently than lncRNA-based clustering, suggesting that the expression of these two RNA families carries complementary information about cell heterogeneity.

2. Cells are better classified by analysing the expression of protein coding genes and lncRNAs as separate omics layers

We then incorporate mRNAs and lncRNAs expression data to perform integrative clustering. This data is generated on the same platform in a single experiment, and the standard approach to integrative analysis is to concatenate the expression measurements into a single matrix on which to apply the clustering algorithm. A more sophisticated approach would consider the inherent difference between mRNAs and lncRNAs expression and treat the two RNA families as disjoint omic layers.

To integrate the information coming from mRNAs and lncRNAs, we designed two different experiments. In the first one, we concatenate the mRNAs and lncRNAs expression measurements into a single matrix of 3095 features and apply hSBM algorithm on it. In the second experiment, the mRNAs and lncRNAs matrices are treated as disjoint omic layers and a multiomic version of hSBM is applied, that is called multibranched SBM and is similar to the multilayer SBM algorithm used for text document analysis (8). This extends the network based topic modeling algorithm to a weighted tri-partite network where nodes represent mRNAs, lncRNAs and cells. Similarly to hSBM, mRNAs and lncRNAs are linked to the cells according to their expression level, while no mRNA-lncRNA edges are present, and the best partition of the tri-partite network is computed.

The two experiments produce very different results in terms of cell classification (see Figure 3). In the first one, based on the concatenated expression matrices, we obtain eleven clusters almost exclusively composed of cells specific to a tumor model (Figure 3A). The NMI/NMI* score in this case is lower than the one obtained on the analysis of mRNAs and lncRNAs analysed separately. On the other hand, in the clustering obtained with the multibranched SBM algorithm, the NMI/NMI* score clearly improves compared to all previous experiments, showing that mRNAs and lncRNAs matrices treated as disjoint omic layers promise to deliver more relevant information on the heterogeneity of the tumors under study.

The multibranched SBM algorithm separates the cells into seven clusters, three of which contain a mixture of resistant and sensitive cells (Figure 3B).

3. Functional enrichment analysis of topics

We wanted to assess whether the gene content associated with the breast cancer cell clusters according to tri-partite SBM was enriched for specific biological functions or lncRNA sets associated with upstream and downstream target regulators. To do this, we developed a procedure to select the most specific topics to each cluster, based on the probability matrix of association of a given topic to a given cell (see Methods).

We tested through hypergeometric statistics the enrichment of topics for the gene functional annotations of the MsigDB Molecular Signature Database (19) and for the lncRNA sets associated with upstream regulators and downstream targets, based on the lncSEA database (20). Results of topic enrichment analysis are provided in Figure 4.

The results show a clear enrichment of all but one topic to the pathways involved in subtyping and progression of breast cancer and to sets of lncRNA encoded in the open chromatin regions of breast cancer cell lines. Notably, we found topics specifically associated with luminal or basal cell clusters, with no topics associated with both luminal and triple negative subtypes. As an example, mRNA-Topic 3 shows the keyword SMID BREAST CANCER BASAL UP (21) and is enriched only in basal-related clusters. The same happens for mRNA-Topic 7 which, being also composed of basal cells, shows a keyword related to luminal downregulation. Selected mRNA-Topics or lncRNA-Topics are associated with clusters of a specific subtype and can be used in an unsupervised framework to identify the tumour subtype.

MRNA-Topic 1 is enriched in genes involved in cell migration and specifically associated with a subpopulation of cells from triple negative tumors, supporting a higher migratory capacity of basal cells compared to luminal ones (22) . Gene content in mRNA-Topic 6 is enriched in HALLMARK_TNFA_SIGNALING_VIA_NFKB and specifically associated with the cell cluster_5 containing a mixture of luminal resistant and sensitive cells. This suggests the NFkB regulation by TNF as a transcriptomic hallmark common to a subset of resistant and sensitive luminal cells. Cell cluster_5 is also significantly associated to the lncRNA-Topic 13, which includes MALAT1 and NEAT1 as top ranked non coding genes, two lncRNAs extensively studied in breast cancer progression (23)(24). Another important result shows the association of cluster_6 composed only of therapy-resistant luminal cells, to mRNA-Topic 12 (HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION) and lncRNA-Topic 19 (lncRNAs in chromatin accessible regions of IPC cells). Finally, lncRNA-Topic 15 is specifically associated with cluster_3 and cluster_4 of drug-sensitive luminal cells and has the very highly ranked lncRNA MIR205 host gene (MIR205HG) as top feature.

4. Examples of lncRNAs identified with the topic modeling approach

In Table 1 we report some functional lncRNAs known in the literature that are associated to the identified topics.

SIRLNT

In the lncRNA-Topic 4, which includes a set of 19 lncRNAs, SIRLNT (SIRT1 regulating lncRNA tumor promoter) has the top high probability of association to the topic (P(SIRLNT|lncRNA-Topic 4)=0.6). This lncRNA, also called lncRNA-PRLB (progression-associated lncRNA in breast cancer), is upregulated in human breast cancer tissues and breast cancer cell lines (25). Moreover, it has been described as a modulator of SIRT1 gene both at the mRNA level and protein level. SIRT1 is a member of the sirtuin family of proteins extensively studied in multiple cancers (26), including breast cancer (27, 28).

GATA3-AS1

Also strongly associated with lncRNA-Topic 4 we found GATA3-AS1, a lncRNA recently proposed as predictive biomarker of nonresponsive breast cancer patients to neoadjuvant chemotherapy (29). GATA3-AS1 was described as a functional lncRNA and positive transcriptional regulator of *GATA3* in TH2 lineage of lymphocytes (29).

LncRNAs associated to the Topic 8 are related to ferropoptosis

The lncRNA-Topic 8 is strongly associated with clusters 0 and 2, both of which are composed of triple negative sensitive and resistant cells. LncRNA-Topic 8 presents among its top features some lncRNAs extensively studied in breast cancer, and particularly in triple negative subtype. The top ranked lncRNA in lncRNA-Topic 8 is LINP1, a regulator of DNA repair in triple negative breast cancer (30) whose knockdown increases the sensitivity of the breast cancer cells to radiotherapy (31). The second top ranked lncRNA in lncRNA-Topic 8, LINC02609, is a ferropoptosis lncRNA proposed as prognostic and diagnostic transcript in renal cancer (32). The role of ferropoptosis in dedifferentiation and drug resistance is widely recognized in breast cancer (33–35) and basal-like breast cancers are more susceptible to ferroptosis (36). The third top ranked lncRNA in lncRNA-Topic 8 is LINC00667, which has been reported to be a key mediator of development in breast cancer (37), colorectal cancer (38), hepatocellular carcinoma (39) and other cancer diseases.

The other two top ranked lncRNAs associated with lncRNA-Topic 8 with a P(lncRNA|lncRNA-Topic 8)>0.05 are ROCR (regulator of chondrogenesis RNA) and PART1. The ROCR lncRNA is upregulated during chondrogenic differentiation of mesenchymal stem cells (40) and differentially expressed in clinically relevant morphological structures of breast cancer with varied degrees of MET and stemness (41). The PART1 lncRNA is known to promote proliferation and migration in TNBC (42).

LncRNA Topic 13 includes MALAT1 and NEAT1 as top ranked features

LncRNA-Topic 13 is characterised by two lncRNAs associated with it with a very high probability that are well known in the literature, namely MALAT1 (Metastasis Associated Lung Adenocarcinoma Transcript 1) and NEAT1. MALAT1 is one of the most studied non-coding RNA in cancer, generally

overexpressed in tumor progression and metastasis (23). NEAT1 is also abnormally expressed in many cancers and associated with therapy resistance and poor clinical outcome (24). In our analysis, the lncRNA-Topic 13 is significantly associated with cluster 5 which contains a mixed subpopulation of sensitive and therapy-resistant luminal cells, suggesting that a subset of sensitive cells share this lncRNA signature with resistant cells.

MIR205HG

The host gene for miR-205 (MIR205HG) is the most strongly associated with lncRNA-Topic 15, expressed in luminal sensitive cell cluster 1 and cluster 2. MIR205HG has independent functions as an lncRNA, regulating growth hormone and prolactin production in pituitary gland (43). Remarkably, MIR205HG was previously indicated as predictor of anti-cancer drug sensitivity (43, 44).

LncRNAs associated to the Topic 19 are specifically expressed in luminal drug-resistant cells

The LncRNA-Topic 19 is specifically expressed in cluster 6 of luminal drug-resistant cells. The top ranked lncRNAs associated to LncRNA-Topic 19 are LINC00493 and TP53TG1. LINC00493 and TP53TG1 are enriched and co-released in extracellular vesicles from colorectal cancer cells (45). Previous studies have highlighted the ambivalent oncogenic or tumor suppressor activity of TP53TG1 in luminal breast cancer (46, 47).

MATERIAL AND METHODS

scRNA-seq dataset pre-processing

Single cell expression count matrices were obtained from the NCBI GEO repository (GSE117309). This dataset includes scRNA-seq profiles from 3723 cells of patient-derived xenograft breast cancer models. Cells were collected from four cancer models: a luminal untreated drug-sensitive tumor, a luminal breast tumor with acquired resistance to Tamoxifen, a triple negative tumor initially responsive to Capecitabine and a triple negative tumor with acquired resistance to Capecitabine.
Protein-coding mRNAs and lncRNAs raw expression matrices were obtained based on Ensemble grch37 gene annotations, after removing mithocondrial and ribosomal genes. Data were analyzed with 1.8.2. Library size normalization was applied and highly variable mRNAs and lncRNAs were selected according to the Scanpy procedure, setting different thresholds for the two RNA families, thus selecting 1959 and 1136 features respectively.

hSBM algorithm

We used the original hSBM clustering algorithm available at https://github.com/martingerlach/hSBM_Topicmodel/tree/develop for handling gene expression data.

The algorithm maximises the probability that the model M describe the data A. Formally it maximises the quantity $P(M|A) = P(A|M)P(M)$. In our context, A is the gene expression matrix and the model M is composed of the levels of the hierarchy, each with its own partitions of genes and cells. hSBM minimises the description length of the model, defined as $\Sigma = -\ln P(M|A)$. The most important

hyperparameter in the hSBM implementation of (14) is the number of initializations, i.e. the number of different starting points from which the algorithm begins to estimate the posterior probability $P(M|A)$. We decided to run hSBM with seven initializations for each experiment because we found that seven initializations best balanced the time cost and the ability to reach the minimum description length (see Supp. Fig. 1). Multibranched SBM is the multipartite extension of hBSM: it works with a multipartite graph and find partition on each side of the graph. We also ran Multibranched SBM with seven initializations (18).

Multibranched SBM algorithm

The analytical framework for adding multiple layers of information to a hSBM inference problem was recently introduced for text analysis (48). In the context of scRNA-seq data analysis, the algorithm was adapted to handle a tri-partite network where nodes represent mRNAs, lncRNAs and cells. mRNAs and lncRNAs were connected to cells according to their expression level, therefore there is no edge between the mRNAs and lncRNAs nodes. The statistical inference procedure, as well as the definition of topics and probability distributions, leading to the tri-partite network clustering is an extension of the hSBM algorithm. The n-partite SBM, introduced in (18) extends the hSBM algorithm described in (14). Its implementation is available on GitHub at https://github.com/BioPhys-Turin/nsbm.

The inference procedure is performed using the functions implemented in the graph-tool library (10.6084/m9.figshare.1164194). This algorithm infers the blocks structure on a given network, using a Markov Chain Monte Carlo (MCMC) that moves nodes (i.e. genes, lncRNAs, cells) between blocks in order to minimise the so-called Description Length (DL) Σ (49); this represents, in nat units, the number of bits that the model needs to describe the data. It can be written as the logarithm of the product of the likelihood times the prior $\Sigma = ln(P(A|M)P(M))$ (A is the matrix of the data, M represents all the model parameters) (50).

This description length Σ can be optimised (51) by a MCMC whose moves are accepted with the following probability:

$$P(r \rightarrow s|t) = \frac{e_{ts}+\epsilon}{e_t+\epsilon B}.$$

The probability of moving a node from block (cluster or topic) r to a block s, given that a random neighbour of the considered node belongs to block t is proportional to the ratio between the number of edges that connects group t and block s e_{ts} and the number of edges that go to block t e_t plus a term ϵB that is necessary to have a uniform probability if there are no connections between the blocks considered.

hSBM and the multibranched SBM impose that nodes of a given type (mRNAs, lncRNAs, cells) must not be mixed, so the blocks can contain only homogeneous nodes. This allows us to define clusters (blocks of cells), mRNA-topics (blocks of mRNAs) and lncRNA-topics (blocks of lncRNAs).

Once the model is run, moving all the nodes multiple times and accepting or refusing the moves with the aforementioned probability, it is possible to extract the P(mRNA | mRNA-topic) and P(mRNA-topic | cell) as well as P(lncRNA | lncRNA-topic) and P(lncRNA-topic | cell)

The P(mRNA | topic) is the ratio of the number of edges connected to the mRNA and the number of edges connected to the topic (the total number of edges connected to any gene in the topic). P(topic | cell) is the ratio of the total number of edges connected to the cell and number of edges connecting the cell to the topic. The same definitions apply to P(lncRNA | lncRNA-topic) and P(lncRNA-topic | cell).

Finally, let us highlight that looking at the probability of accepting a move $P(r \rightarrow s|t)$ it is clear the advantage of using multibranched SBM on multi-omics problems: when moving a cell from block r to block s the algorithm samples a neighbour of the cell and it labels its group as t. At this point the model only considers the links between the cell and the nodes in group t, so the normalisations of other branches are irrelevant. In other words, at each move cells are clustered looking at a single omics and this allows to naturally perform inference on multiomics datasets in which each omic has a different normalisation.

Normalized Mutual Information score

To evaluate clustering performance, we computed the Normalized Mutual Information (NMI) as defined in (52) between the partition found by the algorithm and the tumor of origin of the cells. NMI is a widely used measure to compare clustering methods (53). However, the need of adjustment for the so-called selection bias problem is needed, that is NMI steadily increases with the number of clusters. To keep this effect into account, we adjusted the empirical NMI with the NMI$*$ obtained with a null model that preserves the number of clusters and their sizes but reshuffles the labels of samples. Thus, NMI/NMI$*$ represents how much the empirical score is higher than the score obtained with random clusters of the same size and number.

Topic to cluster assignment

We developed a procedure to assign to each cluster its more suitable topics, exploiting the $P(topic|cell)$.

First, we computed the $P_c(topic|cell)$ because it has been shown that is more informative than the $P(topic|cell)$ (11). This is defined as

$$P_{c(topic|cell)} = P(topic|cell) - \frac{1}{N_{cells}} \sum_{c\ in\ all\ cells} P(topic|c)$$

Then, we calculated the $P_c(topic|cluster)$ as mean of $P_c(topic|cell)$ over the cells inside the cluster

$$P_{c(topic|cluster)} = \frac{1}{\dim\ (cluster)} \sum_{cell\ in\ cluster} P_{c(topic|cell)}$$

Finally, we assigned to each cluster all the topics that satisfied the condition

$$P_c(topic|cluster) > \mu + n * \sigma$$

where μ and σ are the mean and the standard deviation of the $P_c(topic|cluster)$ across all the topics for the cluster under examination.

The core of the procedure concerns the choose of n: we started from $n = 3$, then we lowered n by step of 0.05 until two conditions were fulfilled; first we required that at least one topic is assigned to each cluster and that there were no clusters with the same sets of topics, but subsets were allowed. It's worth noting that following this method it's possible to assign to any kind of partitions the most important topics for each subset: for example, we could group cells by subtype and, computing the $P_c(topic|subtype)$, we could identify the most important topics for each label.

Functional enrichment of topics

The gene content of topics was tested through hypergeometric statistics for functional enrichment using annotations from the MsigDB Molecular Signature Database (19) and the lncSEA database (20). We tested all topics assigned to clusters, excluding those not associated with any clusters. After correcting hypergeometric test p-values for multiple testing using Benjamini-Hochberg, we associated each topic to the most enriched functional gene set. When a gene set is associated with multiple topics, the one with the lowest p-value was retained.

DISCUSSION

The topic modeling algorithms proposed here have proved to be a valid approach for unsupervised clustering of lncRNA expression at the single cell level, which allows to compensate for the sparsity of this type of data. The multibranch SBM enables the classification of breast cancer cell populations, showing that mRNAs and lncRNAs datasets provide complementary information about the heterogeneity of the tumors under study, when they are treated as disjoint omics layers.

Combination of several single-cell omics layers represents a new frontier for single-cell genomics and requires appropriate computational methods. Recently, a number of experimental techniques are capable to assay multiple modalities simultaneously from the same single cells, including CITE-seq (54), 10X Genomics ATAC + RNA, SHARE-seq (55), SNARE-seq (56) and many others. The multibranch SBM algorithm can be applied in different contexts of integrative multimodal analysis to learn the information content of each modality in each cell and to define cellular states based on a combination of information derived from multiple modalities.

Differently from state-of-art clustering methods used in single cell data analysis, such as Leiden algorithm and graph-based algorithms (57, 58), multibranched SBM not only provides an integrative clustering of cells based on both mRNAs and lncRNAs expression, but also output the probability distribution of the association of these features to clusters, that is their relative importance in driving cell classification. Clusters of cells and the sets of associated features (topics) are simultaneously identified

with stochastic block modeling procedures. Another great advantage of the algorithm is that it automatically detects the optimal number of clusters and hierarchically clusters both the features and cells. This avoids researchers from fixing the number of expected clusters, which is difficult in real datasets. In addition, we show that clustering performed with multibranched SBM is not affected by batch effect, a common problem affecting the analysis of Large-scale single-cell transcriptomic datasets (Supp. Figure 2). In our study of breast cancer PDX models, the first hierarchical level separated the luminal and basal cells, while the next level showed a finer heterogeneity of the cells present in the tumors. Identified topics associated to the clusters are enriched for genes and pathways clearly involved in breast cancer subtyping, such as cell adhesion and migration and in sets of lncRNAs encoded in the open chromatin regions of breast cancer cell lines. A relevant result of our analysis is the identification of a ferropoptosis-related lncRNA signature (lncRNA topic 8) which is specifically associated with clusters of triple negative drug-resistant cells. Overall, lncRNAs with high probability of association to topics are well known in breast cancer literature, some others are very interesting candidates for functional validation.

AVAILABILITY

Scripts to reproduce all the analyses and figures presented in this paper are provided at GitHub repository https://github.com/sysbio-curie/topic_modeling_lncRNAs.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR online.

ACKNOWLEDGEMENT

We would like to acknowledge the Competence Centre for Scientific Computing C3S which provided us the access to the computing cluster OCCAM (59).

FUNDING

This work was supported by grants from ITMO Cancer AVIESAN (National Alliance for Life Sciences and Health), within the framework of the Plan Cancer 2014–2019 and convention “2018, Non-coding RNA in cancerology: fundamental to translational [18CN039-00]”.
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).

REFERENCES

1. Melé,M., Ferreira,P.G., Reverter,F., DeLuca,D.S., Monlong,J., Sammeth,M., Young,T.R., Goldmann,J.M., Pervouchine,D.D., Sullivan,T.J., *et al.* (2015) The human transcriptome across tissues and individuals. *Science*, **348**, 660–665.

2. Hon,C.-C., Ramilowski,J.A., Harshbarger,J., Bertin,N., Rackham,O.J.L., Gough,J., Denisenko,E., Schmeier,S., Poulsen,T.M., Severin,J., *et al.* (2017) An atlas of human long non-coding RNAs with accurate 5' ends. *Nature*, **543**, 199–204.

3. Cabili,M.N., Trapnell,C., Goff,L., Koziol,M., Tazon-Vega,B., Regev,A. and Rinn,J.L. (2011) Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev*, **25**, 1915–1927.

4. Kornienko,A.E., Dotter,C.P., Guenzl,P.M., Gisslinger,H., Gisslinger,B., Cleary,C., Kralovics,R., Pauler,F.M. and Barlow,D.P. (2016) Long non-coding RNAs display higher natural expression variation than protein-coding genes in healthy humans. *Genome Biol*, **17**, 14.

5. Yan,X., Hu,Z., Feng,Y., Hu,X., Yuan,J., Zhao,S.D., Zhang,Y., Yang,L., Shan,W., He,Q., *et al.* (2015) Comprehensive Genomic Characterization of Long Non-Coding RNAs across Human Cancers. *Cancer Cell*, **28**, 529–540.

6. Chiu,H.-S., Somvanshi,S., Patel,E., Chen,T.-W., Singh,V.P., Zorman,B., Patil,S.L., Pan,Y., Chatterjee,S.S., Sood,A.K., *et al.* (2018) Pan-Cancer Analysis of lncRNA Regulation Supports Their Targeting of Cancer Genes in Each Tumor Context. *Cell Rep*, **23**, 297-312.e12.

7. Cabili,M.N., Dunagin,M.C., McClanahan,P.D., Biaesch,A., Padovan-Merhar,O., Regev,A., Rinn,J.L. and Raj,A. (2015) Localization and abundance analysis of human lncRNAs at single-cell and single-molecule resolution. *Genome Biol*, **16**, 20.

8. Isakova,A., Neff,N. and Quake,S.R. (2021) Single-cell quantification of a broad RNA spectrum reveals unique noncoding patterns associated with cell types and states. *Proc Natl Acad Sci U S A*, **118**, e2113568118.

9. Liu,S.J., Nowakowski,T.J., Pollen,A.A., Lui,J.H., Horlbeck,M.A., Attenello,F.J., He,D., Weissman,J.S., Kriegstein,A.R., Diaz,A.A., *et al.* (2016) Single-cell analysis of long non-coding RNAs in the developing human neocortex. *Genome Biol*, **17**, 67.

10. Liu,L., Tang,L., Dong,W., Yao,S. and Zhou,W. (2016) An overview of topic modeling and its current applications in bioinformatics. *Springerplus*, **5**, 1608.

11. Valle,F., Osella,M. and Caselle,M. (2020) A Topic Modeling Analysis of TCGA Breast and Lung Cancer Transcriptomic Data. *Cancers (Basel)*, **12**, 3799.

12. Ruiz Tejada Segura,M.L., Abou Moussa,E., Garabello,E., Nakahara,T.S., Makhlouf,M., Mathew,L.S., Wang,L., Valle,F., Huang,S.S.Y., Mainland,J.D., *et al.* (2022) A 3D transcriptomics atlas of the mouse nose sheds light on the anatomical logic of smell. *Cell Rep*, **38**, 110547.

13. Lazzardi,S., Valle,F., Mazzolini,A., Scialdone,A., Caselle,M. and Osella,M. (2022) Emergent Statistical Laws in Single-Cell Transcriptomic Data. 10.1101/2021.06.16.448706.

14. Gerlach,M., Peixoto,T.P. and Altmann,E.G. (2018) A network approach to topic models. *Sci Adv*, **4**, eaaq1360.

15. Holland,P.W., Laskey,K.B. and Leinhardt,S. (1983) Stochastic blockmodels: First steps. *Social Networks*, **5**, 109–137.

16. Morelli,L., Giansanti,V. and Cittaro,D. (2021) Nested Stochastic Block Models applied to the analysis of single cell data. *BMC Bioinformatics*, **22**, 576.

17. Grosselin,K., Durand,A., Marsolier,J., Poitou,A., Marangoni,E., Nemati,F., Dahmani,A., Lameiras,S., Reyal,F., Frenoy,O., *et al.* (2019) High-throughput single-cell ChIP-seq identifies heterogeneity of chromatin states in breast cancer. *Nat Genet*, **51**, 1060–1066.

18. Valle,F., Osella,M. and Caselle,M. (2022) Multiomics Topic Modeling for Breast Cancer Classification. *Cancers (Basel)*, **14**, 1150.

19. Liberzon,A., Subramanian,A., Pinchback,R., Thorvaldsdóttir,H., Tamayo,P. and Mesirov,J.P. (2011) Molecular signatures database (MSigDB) 3.0. *Bioinformatics*, **27**, 1739–1740.

20. Chen,J., Zhang,J., Gao,Y., Li,Y., Feng,C., Song,C., Ning,Z., Zhou,X., Zhao,J., Feng,M., *et al.* (2020) LncSEA: a platform for long non-coding RNA related sets and enrichment analysis. *Nucleic Acids Res*, **49**, D969–D980.

21. Smid,M., Wang,Y., Zhang,Y., Sieuwerts,A.M., Yu,J., Klijn,J.G.M., Foekens,J.A. and Martens,J.W.M. (2008) Subtypes of breast cancer show preferential site of relapse. *Cancer Res*, **68**, 3108–3114.

22. Nair,N.U., Das,A., Rogkoti,V.-M., Fokkelman,M., Marcotte,R., de Jong,C.G., Koedoot,E., Lee,J.S., Meilijson,I., Hannenhalli,S., *et al.* (2019) Migration rather than proliferation transcriptomic signatures are strongly associated with breast cancer patient survival. *Sci Rep*, **9**, 10989.

23. Arun,G. and Spector,D.L. (2019) MALAT1 long non-coding RNA and breast cancer. *RNA Biol*, **16**, 860–863.

24. Knutsen,E., Harris,A.L. and Perander,M. (2022) Expression and functions of long non-coding RNA NEAT1 and isoforms in breast cancer. *Br J Cancer*, **126**, 551–561.

25. Liang,Y., Song,X., Li,Y., Sang,Y., Zhang,N., Zhang,H., Liu,Y., Duan,Y., Chen,B., Guo,R., *et al.* (2018) A novel long non-coding RNA-PRLB acts as a tumor promoter through regulating miR-4766-5p/SIRT1 axis in breast cancer. *Cell Death Dis*, **9**, 563.

26. Chalkiadaki,A. and Guarente,L. (2015) The multifaceted functions of sirtuins in cancer. *Nat Rev Cancer*, **15**, 608–624.

27. Shi,L., Tang,X., Qian,M., Liu,Z., Meng,F., Fu,L., Wang,Z., Zhu,W.-G., Huang,J.-D., Zhou,Z., *et al.* (2018) A SIRT1-centered circuitry regulates breast cancer stemness and metastasis. *Oncogene*, **37**, 6299–6315.

28. Jin,X., Wei,Y., Xu,F., Zhao,M., Dai,K., Shen,R., Yang,S. and Zhang,N. (2018) SIRT1 promotes formation of breast cancer through modulating Akt activity. *J Cancer*, **9**, 2012–2023.

29. Contreras-Espinosa,L., Alcaraz,N., De La Rosa-Velázquez,I.A., Díaz-Chávez,J., Cabrera-Galeana,P., Rebollar-Vega,R., Reynoso-Noverón,N., Maldonado-Martínez,H.A., González-Barrios,R., Montiel-Manríquez,R., *et al.* (2021) Transcriptome Analysis Identifies GATA3-AS1 as a Long Noncoding RNA Associated with Resistance to Neoadjuvant Chemotherapy in Locally Advanced Breast Cancer Patients. *J Mol Diagn*, **23**, 1306–1323.

30. Zhang,Y., He,Q., Hu,Z., Feng,Y., Fan,L., Tang,Z., Yuan,J., Shan,W., Li,C., Hu,X., *et al.* (2016) Long noncoding RNA LINP1 regulates repair of DNA double-strand breaks in triple-negative breast cancer. *Nat Struct Mol Biol*, **23**, 522–530.

31. Liang,Y., Li,Y., Song,X., Zhang,N., Sang,Y., Zhang,H., Liu,Y., Chen,B., Zhao,W., Wang,L., *et al.* (2018) Long noncoding RNA LINP1 acts as an oncogene and promotes chemoresistance in breast cancer. *Cancer Biol Ther*, **19**, 120–131.

32. Shu,X., Zhang,Z., Yao,Z.-Y. and Xing,X.-L. (2021) Identification of Five Ferroptosis-Related LncRNAs as Novel Prognosis and Diagnosis Signatures for Renal Cancer. *Front Mol Biosci*, **8**, 763697.

33. Rodriguez,R., Schreiber,S.L. and Conrad,M. (2022) Persister cancer cells: Iron addiction and vulnerability to ferroptosis. *Mol Cell*, **82**, 728–740.

34. Buschhaus,J.M., Rajendran,S., Humphries,B.A., Cutter,A.C., Muñiz,A.J., Ciavattone,N.G., Buschhaus,A.M., Cañeque,T., Nwosu,Z.C., Sahoo,D., *et al.* (2022) Effects of iron modulation on mesenchymal stem cell-induced drug resistance in estrogen receptor-positive breast cancer. *Oncogene*, **41**, 3705–3718.

35. Wu,Z.-H., Tang,Y., Yu,H. and Li,H.-D. (2021) The role of ferroptosis in breast cancer patients: a comprehensive analysis. *Cell Death Discov*, **7**, 93.

36. Z,L., L,C., C,C., Y,Z., D,H., J,Y., Y,C., W,Z., M,M., X,Z., *et al.* (2020) Targeting ferroptosis in breast cancer. *Biomarker research*, **8**.

37. Zhu,M., Lv,Q., Huang,H., Sun,C., Pang,D. and Wu,J. (2020) Identification of a four-long non-coding RNA signature in predicting breast cancer survival. *Oncol Lett*, **19**, 221–228.

38. Yu,J., Wang,F., Zhang,J., Li,J., Chen,X. and Han,G. (2020) LINC00667/miR-449b-5p/YY1 axis promotes cell proliferation and migration in colorectal cancer. *Cancer Cell Int*, **20**, 322.

39. Gu,J., Zhang,X., Miao,R., Ma,X., Xiang,X., Fu,Y., Liu,C., Niu,W. and Qu,K. (2018) A three-long non-coding RNA-expression-based risk score system can better predict both overall and recurrence-free survival in patients with small hepatocellular carcinoma. *Aging (Albany NY)*, **10**, 1627–1639.

40. Barter,M.J., Gomez,R., Hyatt,S., Cheung,K., Skelton,A.J., Xu,Y., Clark,I.M. and Young,D.A. (2017) The long non-coding RNA ROCR contributes to SOX9 expression and chondrogenic differentiation of human mesenchymal stem cells. *Development*, **144**, 4510–4521.

41. Denisov,E.V., Skryabin,N.A., Gerashchenko,T.S., Tashireva,L.A., Wilhelm,J., Buldakov,M.A., Sleptcov,A.A., Lebedev,I.N., Vtorushin,S.V., Zavyalova,M.V., *et al.* (2017) Clinically relevant morphological structures in breast cancer represent transcriptionally distinct tumor cell populations with varied degrees of epithelial-mesenchymal transition and CD44+CD24- stemness. *Oncotarget*, **8**, 61163–61180.

42. Cruickshank,B.M., Wasson,M.-C.D., Brown,J.M., Fernando,W., Venkatesh,J., Walker,O.L., Morales-Quintanilla,F., Dahn,M.L., Vidovic,D., Dean,C.A., *et al.* (2021) LncRNA PART1 Promotes Proliferation and Migration, Is Associated with Cancer Stem Cells, and Alters the miRNA Landscape in Triple-Negative Breast Cancer. *Cancers (Basel)*, **13**, 2644.

43. Du,Q., Hoover,A.R., Dozmorov,I., Raj,P., Khan,S., Molina,E., Chang,T.-C., de la Morena,M.T., Cleaver,O.B., Mendell,J.T., *et al.* (2019) MIR205HG Is a Long Noncoding RNA that Regulates Growth Hormone and Prolactin Production in the Anterior Pituitary. *Dev Cell*, **49**, 618-631.e5.

44. Nath,A., Lau,E.Y.T., Lee,A.M., Geeleher,P., Cho,W.C.S. and Huang,R.S. (2019) Discovering long noncoding RNA predictors of anticancer drug sensitivity beyond protein-coding genes. *Proc Natl Acad Sci U S A*, **116**, 22020–22029.

45. Chen,M., Xu,R., Ji,H., Greening,D.W., Rai,A., Izumikawa,K., Ishikawa,H., Takahashi,N. and Simpson,R.J. (2016) Transcriptome and long noncoding RNA sequencing of three extracellular vesicle subtypes released from the human colon cancer LIM1863 cell line. *Sci Rep*, **6**, 38397.

46. Diaz-Lagares,A., Crujeiras,A.B., Lopez-Serra,P., Soler,M., Setien,F., Goyal,A., Sandoval,J., Hashimoto,Y., Martinez-Cardús,A., Gomez,A., *et al.* (2016) Epigenetic inactivation of the p53-induced long noncoding RNA TP53 target 1 in human cancer. *Proc Natl Acad Sci U S A*, **113**, E7535–E7544.

47. Motalebzadeh,J. and Eskandari,E. (2022) Comprehensive analysis of DRAIC and TP53TG1 in breast cancer luminal subtypes through the construction of lncRNAs regulatory model. *Breast Cancer*, 10.1007/s12282-022-01385-7.

48. Hyland,C.C., Tao,Y., Azizi,L., Gerlach,M., Peixoto,T.P. and Altmann,E.G. (2021) Multilayer networks for text analysis with multiple data types. *EPJ Data Sci.*, **10**, 1–16.

49. Peixoto,T.P. (2012) Entropy of stochastic blockmodel ensembles. *Phys Rev E Stat Nonlin Soft Matter Phys*, **85**, 056122.

50. Peixoto,T.P. (2015) Model Selection and Hypothesis Testing for Large-Scale Network Models with Overlapping Groups. *Phys. Rev. X*, **5**, 011033.

51. Peixoto,T.P. (2017) Nonparametric Bayesian inference of the microcanonical stochastic block model. *Phys Rev E*, **95**, 012317.

52. Hirschberg,J.B. and Rosenberg,A. (2007) V-Measure: A conditional entropy-based external cluster evaluation. 10.7916/D80V8N84.

53. Shi,H., Gerlach,M., Diersen,I., Downey,D. and Amaral,L. (2019) A new evaluation framework for topic modeling algorithms based on synthetic corpora. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*. PMLR, pp. 816–826.

54. Stoeckius,M., Hafemeister,C., Stephenson,W., Houck-Loomis,B., Chattopadhyay,P.K., Swerdlow,H., Satija,R. and Smibert,P. (2017) Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods*, **14**, 865–868.

55. Ma,S., Zhang,B., LaFave,L.M., Earl,A.S., Chiang,Z., Hu,Y., Ding,J., Brack,A., Kartha,V.K., Tay,T., *et al.* (2020) Chromatin Potential Identified by Shared Single-Cell Profiling of RNA and Chromatin. *Cell*, **183**, 1103-1116.e20.

56. Plongthongkum,N., Diep,D., Chen,S., Lake,B.B. and Zhang,K. (2021) Scalable dual-omics profiling with single-nucleus chromatin accessibility and mRNA expression sequencing 2 (SNARE-seq2). *Nat Protoc*, **16**, 4992–5029.

57. Traag,V.A., Waltman,L. and van Eck,N.J. (2019) From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep*, **9**, 5233.

58. Xu,C. and Su,Z. (2015) Identification of cell types from single-cell transcriptomes using a novel clustering method. *Bioinformatics*, **31**, 1974–1980.

59. Aldinucci,M., Bagnasco,S., Lusso,S., Pasteris,P., Rabellino,S. and Vallero,S. (2017) OCCAM: a flexible, multi-purpose and extendable HPC cluster. *J. Phys.: Conf. Ser.*, **898**, 082039.

TABLE AND FIGURES LEGENDS

Table 1. Functional lncRNAs known in the literature that are associated to the identified topics.

lncRNA-Topic 4 = Chromatin in MCF-7		
lncRNA_ID	**Probability of association to topic**	**Ref.**
SIRLNT=CTA-392C11.1=ENSG00000253802	P(SIRLNT|lncRNA-Topic 4)=0.6	(25)
GATA3-AS1 = ENSG00000197308	P(GATA3-AS1|lncRNA-Topic 4)=0.05	(29)
lncRNA-Topic 8 = Chromatin in MCF-10		
lncRNA_ID	**Probability of association to topic**	**Ref.**
LINP1=RP11-554I8.2=ENSG00000223784	P(LINP1|lncRNA-Topic 8)=0.12	(30,31)
LINC02609=RP4-665J23.1=ENSG00000233593	P(LINC02609|lncRNA-Topic 8)=0.01	(32)
LINC00667=ENSG00000263753	P(LINC00667|lncRNA-Topic 8)=0.08	(37-39)
ROCR=AC005152.3=ENSG00000228639	P(ROCR|lncRNA-Topic 8)=0.08	(40)
PART1=ENSG00000152931	P(PART1|lncRNA-Topic 8)=0.08	(42)
lncRNA-Topic 13 =MicroRNA_lncbase_hsa-miR-626		
lncRNA_ID	**Probability of association to topic**	**Ref.**
MALAT1=ENSG00000251562	P(MALAT1|lncRNA-Topic 13)=0.60	(23)
NEAT1=ENSG00000245532	P(NEAT1|lncRNA-Topic 13)=0.40	(24)
lncRNA-Topic 15 = Accessible_Chromatin_CLOSEST_DNaseI_CellLine_KARPAS-422		
lncRNA_ID	**Probability of association to topic**	**Ref.**
MIR205HG=ENSG00000230937	P(MIR205HG|lncRNA-Topic 15)=0.66	(43,44)
lncRNA-Topic 19 = Chromatin in iPSC		
lncRNA_ID	**Probability of association to topic**	**Ref.**
LINC00493	P(LINC00493|lncRNA-Topic 19)=0.32	(45-47)
TP53TG1=LINC00096=ENSG00000182165	P(TP53TG1|lncRNA-Topic 19)=0.15	(45-47)

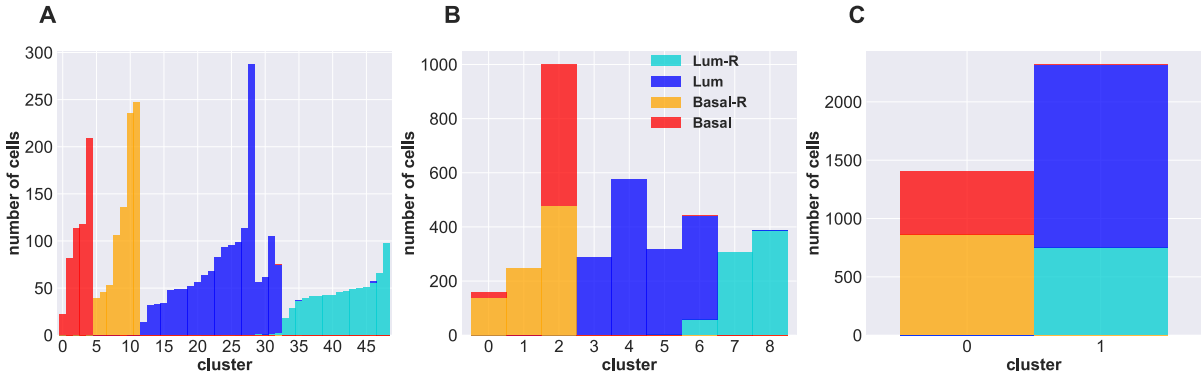

Figure 1. Cells clustering obtained with hSBM applied to the mRNA expression dataset. For each hierarchical level of clustering (A, B, C), the boxplot reports the size of the identified clusters and the tumor subtype of the cells that compose them.

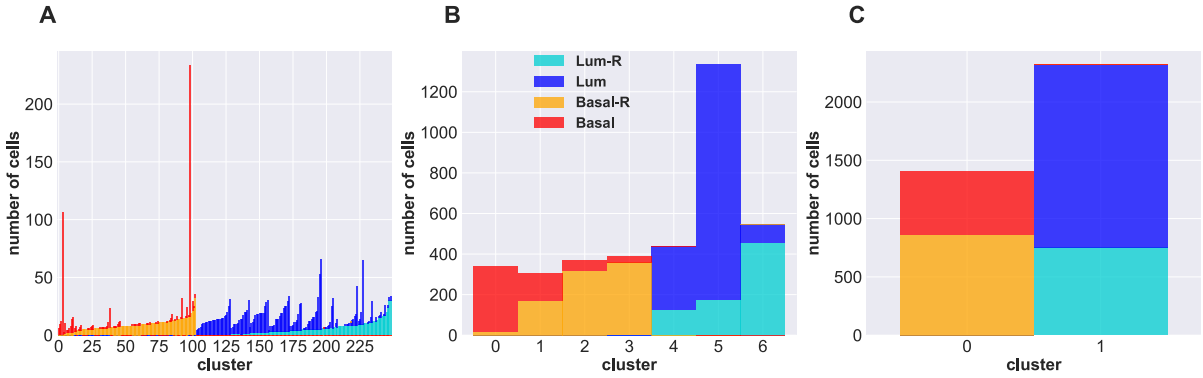

Figure 2. Cells clustering obtained with hSBM applied to the lncRNA expression dataset. For each hierarchical level of clustering (A, B, C), the boxplot reports the size of the identified clusters and the tumor subtype of the cells that compose them.

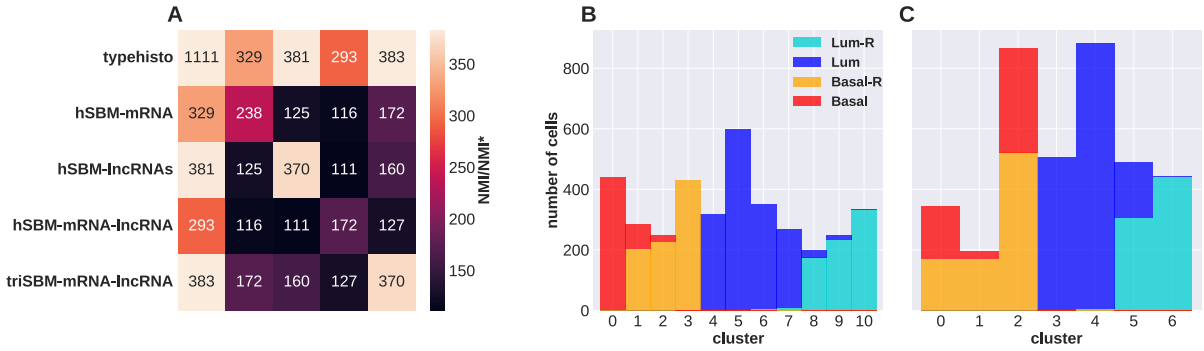

Figure 3. Cells clustering obtained with hSBM applied to the concatenated mRNAs and lncRNAs expression matrices (B) and with the multibranched SBM (C). The boxplot reports the size of the identified clusters and the tumor subtype of the cells that compose them. Heatmap in panel (A) reports the NMI/NMI* scores obtained with different clustering algorithms.

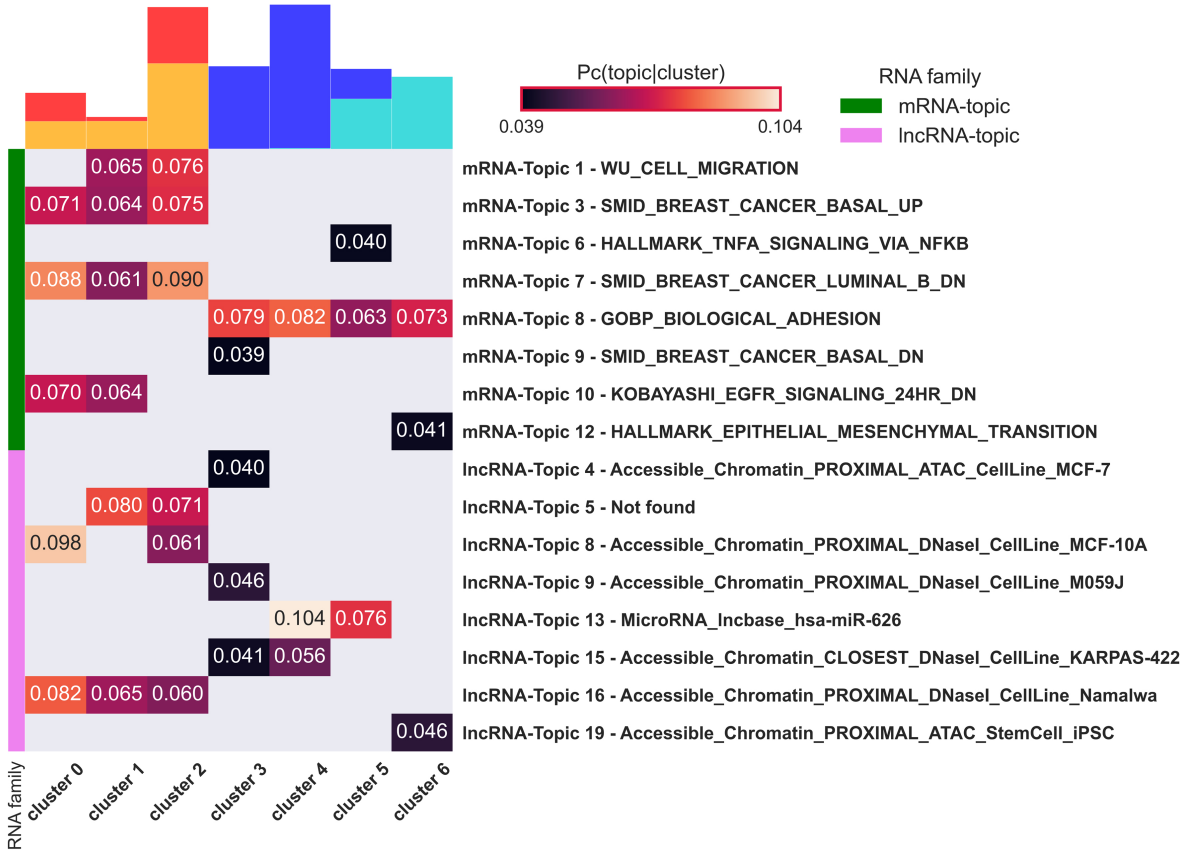

Figure 4. Functional enrichment analysis of topics. The heatmap reports the probability of association of topics to clusters (when statistically significant) and the most enriched functional gene set according to MsigDB and lncSEA databases.