Exploratory Analysis of the Gene Expression Matrix based on Dual Conditional Dimensionality Reduction

Ignacio Díaz, *Member, IEEE,* José M. Enguita, Abel A. Cuadrado, *Member, IEEE*, Diego García, Ana González, Nuria Valdés, María D. Chiara

***Abstract*—One of the major goals in gene expression data analysis is to explore and discover groups of genes and groups of biological conditions with meaningful relationships. While this problem can be addressed by algorithms, their results require an analysis within context, since they may be affected by many side processes —such as tissue differentiation— that could hinder the target goal. Visual analytics-based methods for exploratory analysis of the *gene expression matrix* (GEM) are essential in biomedical research since they allow us to frame the analysis within the user's knowledge domain. In this paper, we present a visual analytics approach to discover relevant connections between genes and samples based on linking a reordered GEM heatmap and dual 2D projections of its rows and columns, which can be recomputed conditioned by subsets of genes and/or samples selected by the user during the analysis. We demonstrate the capability of our approach to discover relevant knowledge in three case studies involving two cancer types plus normal tissue from the TCGA database.**

***Index Terms*—visual analytics, gene expression, exploratory data analysis**

I. Introduction

One of the challenges in gene expression data analysis is to discover patterns involving subsets of genes showing interesting or meaningful behaviors across subsets of biological samples [1]. The first reference to tackling this problem dates back to 1972, when Hartigan proposed an algorithm to cluster cases and variables simultaneously [2]. However, it was not until 2000 that Cheng and Church introduced the concept of *biclustering* as an approach for knowledge discovery from gene expression data, in terms of algorithms to find subsets of genes and subsets of conditions with a high similarity [3]. Since then, this kind of technique has become widely spread in the field of bioinformatics, and many other algorithms have been developed for the discovery of biclusters, as shown in some comprehensive surveys [4], [5].

While the algorithmic approach alone can be very helpful in the analysis and discovery of potentially relevant patterns, the results may lack insight since they reflect correlation, but not necessarily causation. Biomedical research often involves a vast amount of domain knowledge in terms of known biological pathways, with complex cascades of interactions among functionally related genes, external factors and biological conditions, through which the discoveries must be framed to make sense. In this way, approaches such as *visual analytics* (VA) [6], [7], that combine data visualization, interaction and machine learning to take into account the user in the analytic process, have become increasingly more prominent as a powerful alternative to provide insight and perspective into the analysis. Several biclustering analysis tools have been developed with visualization capabilities, such as BiGGEsTS [8], which includes a visualization module capable of rendering GEM heatmaps, dendrograms and expression pattern charts, BicAT [9] or BiVisu [10], which features parallel coordinate visualization of computed biclusters. However, although visualization allows some user supervision of the results, these methods are mainly *one-directional*, with reduced possibilities for the user to reconfigure the analysis based on the observed outcome.

Arguably, mechanisms to provide rich feedback and foster an active role of the user in the analytics discourse are of utmost importance to reconciliate high-dimensional data information with complex domain knowledge involving pathways of gene coregulations and biological conditions related in intricate ways. Some tools go one step further in the VA paradigm, providing richer interaction mechanisms, such as BicOverlapper [11], [12], Bicluster viewer [13] or VisBicluster [14], enabling the user to explore the data by integrating interaction mechanisms linking different views such as parallel coordinates, GEM heatmaps and cluster network visualizations.

In line with this, other methods following a different approach based on projection techniques have also been proposed to address the problem of analysis and discovery of relevant gene-sample patterns, often featuring a stronger visual and interactive component. In [15], for instance, the authors propose *nonnegative matrix factorization* (NMF) to

Ignacio Díaz, José M. Enguita, Abel A. Cuadrado, Diego García and Ana González are with the Dept. of Electrical Engineering, University of Oviedo, Gijón 33204, Spain, e-mail: (see `http://isa.uniovi.es/GSDPI/contacto.html`).

Nuria Valdés is with Dept. of Internal Medicine, Section of Endocrinology and Nutrition, Hospital Universitario de Cabueñes, Gijón, 33204, Spain.

María D. Chiara is with Institute of Sanitary Research of the Principado de Asturias, Hospital Universitario Central de Asturias, Oviedo, 33011, Spain and CIBERONC (Network of Biomedical Research in Cancer), Madrid, 28029, Spain.

decompose the GEM into biologically relevant factors that are embedded in a 2D visualization along with genes and samples. *Dimensionality reduction* (DR) methods such as t-SNE [16] and UMAP [17] have been extensively used to visualize gene expression data in large collections of samples with thousands of genes. These methods are able to map the high dimensional gene expression patterns of the samples to 2D points, so samples with similar gene expression profiles will be projected to close locations, resulting in a visual map where the samples are spatially organized by genetic similarity. This idea has been extensively used as a powerful way to discover and visualize clusters of biologically similar samples in an intuitive way [18], [19] and has led to many data visualization tools for gene expression analysis in recent years [20], [21]. Interestingly, while DR methods have been mostly used to project samples, they can also be used in a *dual* way, projecting genes —instead of the samples— to reveal their expression similarity across samples, such as in the Neuroblast tool, which identifies networks of genes coexpressed within or across neuroanatomic structures [22]. Indeed, the use of DR methods in both the primal and dual ways allows us to discover and explore patterns in both the sample and gene domains, as shown in the BrainScope tool [23].

Despite the fact that the previous approaches tackle the problem from different angles, there remain some shortcomings to be addressed, from which we highlight two:

- First, regrouping the samples (or genes) in a GEM by means of algorithms implicitly assumes finding a permutation operation in the rows (or the columns), so that items with similar expression patterns are placed in close positions in the final arrangement; such an operation is closely related to a 1D dimensionality reduction, that preserves the topological closeness between the sample/gene space and an implicit latent 1D space of scores used for sorting. Obviously, the intrinsic dimension of the input data —in the sample or gene spaces— may not be 1D, potentially leading to cluster overlapping [11], [12].
- Second, biclustering algorithms automatically find subsets of samples and genes to optimize some *agnostic* cost function, involving some priors that might not be optimal for explainability, being able to find correlations, but not necessarily causal connections. User-guided selection of subsets of genes and samples in an exploratory way, followed by algorithms for rearranging the GEM rows and columns according to the similarity of the expressions only in these subsets, poses an alternative way in which the system takes into consideration the user's domain knowledge.

In this paper, we propose a methodology to overcome the above shortcomings. For this purpose, we consider matrix reordering methods by means of 1D DR of the rows and columns of the GEM, which can be conditioned by user-defined subsets of genes and samples, aided by selectable and interactive dual DR projections —a 2D projection of the samples plus a 2D projection of the genes—, in a similar way as proposed in [23]. Meaningful gene/sample subset selections to condition the sample/gene reordering processes respectively, are crucial for the success of the analysis, since they define in what (biological) sense genes or samples are considered "similar" for the GEM reconfiguration. The visualization and selection of groups of genes and samples in 2D mappings overcomes the implicit 1D limitation of algorithmic reordering approaches, being more able to disentangle data and to reveal a richer cluster structure than 1D mappings. At the same time, since this process is supervised by the user with the help of efficient interaction mechanisms, it provides insight and sensemaking.

The remainder of the paper is organized as follows. Section II includes the materials and methods used throughout the paper, including: the data sources used for the experiments; the definitions and notation related to GEM, permutations and selections; reordering methods for heatmap visualization; the dual 2D projections of genes and samples; the possible bidirectional interactions between the GEM and the dual projections; and, finally, we present additional higher level user interaction through conditional reconfiguration of the GEM and the dual views for user-selected subsets of genes and/or samples. Section III explains the analysis workflow allowed by the proposed approach and discusses several case studies using a prototype implementation of the proposed approach to discover relevant patterns in gene expression data involving microRNA (miRNA) and mRNA expression in samples including pheochromocytoma-paraganglioma (PCPG), kidney clear cell carcinoma (KIRC) and normal kidney tissue. Finally, Section V concludes the paper, providing a general discussion including the main contributions of this work and suggesting lines for future research work.

II. Materials and methods

A. Gene expression data sources

RNA sequencing (RNAseq) technologies use massive sequencing to provide gene expression data of large amounts of samples with a large number of expression measurements of transcripts involving mRNA and miRNA. The Cancer Genome Atlas (TCGA) database provides gene expression measurements involving more than 20000 measurements for thousands of biological samples from more than 33 cancer types.

The dataset used in this paper was obtained from the TCGA database (downloaded from the Xenabrowser portal[1]), and curated by removing samples with erroneous or missing data for some of the genes or miRNAs of interest. The resulting dataset contains 157 samples of pheochromocytoma and paraganglioma (PCPG), some carrying mutations in hypoxia related genes such as VHL, SDH and EPAS1, 221 of kidney clear cell carcinoma (KIRC), and 71 of normal renal tissue. For each sample, we selected the expression levels of 129 miRNA and 442 hypoxia-related genes (those available after data curation from the list of 446 described in [24]), including the so-called canonical hypoxia genes, and others with special functions (angiogenesis, extracellular matrix, etc.).

[1] https://xenabrowser.net/datapages/

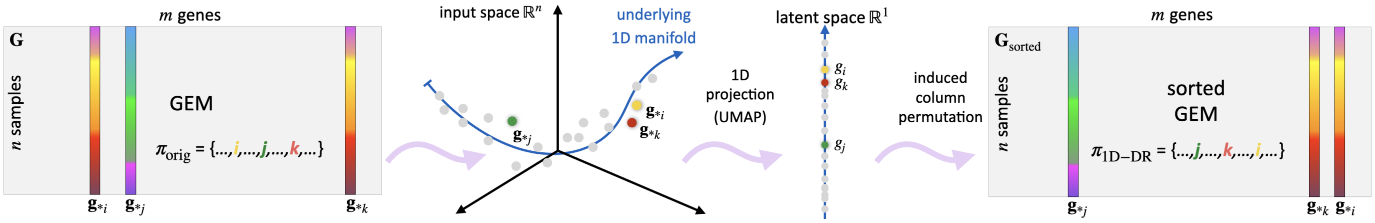

Fig. 1. Schematic representation of DR induced permutation for GEM column sort.

B. Gene expression matrix

Gene expression data are commonly organized in a GEM involving conditions and genes. In this paper we shall use columns for the attributes (genes) and rows for the samples (conditions). A GEM can be defined as an $n \times m$ matrix:

$$\mathbf{G} = (g_{ij}) \tag{1}$$

whose rows $i = 1, \ldots, n$ represent the samples (e.g., tumors or other biological conditions) and whose columns $j = 1, \ldots, m$, represent the attributes (e.g., gene or miRNA expressions) that define each sample. Thus, the scalar element g_{ij} represents the expression level of gene j for sample i. Similarly, the j-th *column vector* $\mathbf{g}_{*j}$ describes the expression behavior of gene j across all the samples in $\mathbf{G}$, and the *row vector* $\mathbf{g}_{i*}$ represents the expression pattern of sample i for all the genes in $\mathbf{G}$.

C. Permutations and selections

A *permutation* is defined as a sequence $\pi = \{i_1, i_2, \ldots, i_n\}$ containing a rearrangement of n indices $\{1, \ldots, n\}$ in a different order. Similarly, a *selection* is defined as a subset $\sigma = \{i_1, \ldots, i_r\}$ of r out of n indices $\{1, \ldots, n\}$.

We shall denote with $\mathbf{G}_{\pi_r *}$ the matrix $\mathbf{G}$ with the rows permuted according to permutation π_r. Similarly, $\mathbf{G}_{*\pi_c}$ represents the columns permuted according to π_c. A simultaneous row and column permutation is denoted as $\mathbf{G}_{\pi_r \pi_c}$. A similar notation will be used for row and column selections, σ_r and σ_c, being $\mathbf{G}_{\sigma_r *}$, $\mathbf{G}_{*\sigma_c}$ and $\mathbf{G}_{\sigma_r \sigma_c}$ submatrices of $\mathbf{G}$ with a subset σ_r of the rows, a subset σ_c of the columns, and subsets σ_r, σ_c of both, respectively.

D. GEM reordering methods

In an unordered GEM, the rows (samples) and the columns (genes) are at arbitrary positions. The typical heatmap visualization in this case will not show patterns revealing genes with similar functions for all the samples or for certain samples in a specific condition, such as a cancer type. Similarly, it will not reveal samples with similar expression patterns for all the genes or for subsets of genes of interest —e.g. a certain gene cluster related to a known pathway.

Matrix reordering methods have been extensively used in the gene expression analysis literature [25], [26]. These methods involve finding proper permutations for the rows and the columns of a given matrix without changing the values of the matrix elements, so the resulting matrix reveals patterns to the user that help them in the analysis.

GEM reordering methods, in general, should follow a *similarity principle* (similar ≈ close), that is, if the expression of genes i and j across all the samples —i.e., vectors $\mathbf{g}_{*i}$ and $\mathbf{g}_{*j}$— are similar according to some distance measure $d(\cdot, \cdot)$, their columns should be placed *close* to each other in the matrix. Conversely, if the behaviors are different, their columns should appear far apart. The same argument is applicable to rows (samples). The idea behind the similarity principle is closely related to a *continuity* or *smoothness* requirement and induces an order in the representation that allows the user to visually identify clusters of genes or samples that behave similarly, thereby making the user aware of the overall behavior of genes and their relationships.

As mentioned before, biclustering algorithms are a special kind of clustering method able to perform simultaneous row-column clustering [4], resulting in a set of submatrices of $\mathbf{G}$ called *biclusters*. A bicluster (I, J) is defined by a subset of the row indices $I \subset \{1, \ldots, n\}$ and a subset of the column indices $J \subset \{1, \ldots, m\}$ for which the gene expressions exhibit a similar behavior [4]. This information can be used to find proper row and column permutations so that the rows and columns of the same bicluster are placed together, thereby turning the bicluster into a visible pattern. For exhaustive, nonoverlapping biclusters (checkerboard type) [2] it is easy to find row and column permutations π_r and π_c, so that all the biclusters are visually revealed in the reordered GEM, $\mathbf{G}^{\text{bicluster}} = \mathbf{G}_{\pi_r \pi_c}$. An example of such reordered GEM can be seen in Fig. 4, to be discussed later in the results section.

Alternatively, DR algorithms [27] allow us to define a mapping $\varphi_d : \mathbb{R}^D \to \mathbb{R}^d$ that transforms a dataset $\mathbf{X}$, with dimensionality D into a new dataset $\mathbf{Y}$ of a much smaller dimensionality d, while retaining the geometry of the data and the mutual similarities among the samples, as much as possible. For convenience we shall denote $\mathbf{Y} = \varphi_d(\mathbf{X})$, assuming that the mapping φ_d is learned from $\mathbf{X}$, and applied to all rows of it, that is, $\mathbf{y}_{i*} = \varphi_d(\mathbf{x}_{i*})$. Particularly, DR mappings on 1D latent spaces can be used to reorder the rows or columns of the GEM. For a row reordering of $\mathbf{G}$ with n samples (rows) and m genes (columns), a 1D dimensionality reduction mapping $\varphi_1 : \mathbb{R}^m \to \mathbb{R}^1$ can be defined, using state-of-the-art methods such as t-SNE [16], [18] or UMAP [17], [19], and then applying the mapping to the rows of $\mathbf{G}$, that is, $\varphi_1(\mathbf{G})$, resulting in n scalars g_i. Under good topology preservation of the mapping it is expected that close scalars g_i and g_j refer to samples i and j with similar gene expression patterns $\mathbf{g}_{i*}$ and $\mathbf{g}_{j*}$. Sorting the scalars, the resulting permutation $\pi_r = \arg\operatorname{sort}(g_1, \ldots, g_n) = \{i_1, \ldots, i_n\}$ induces an

[2] This does not apply to more general types, such as overlapping biclusters; specific visualization methods have been proposed for these —see [11], [12].

ordered sequence in the rows of the gene expression matrix $\mathbf{g}_{i_1*}, \mathbf{g}_{i_2*}, \ldots, \mathbf{g}_{i_n*}$, where consecutive rows can be expected to be similar. Reordering the matrix $\mathbf{G}$ according to this permutation yields a reordered matrix, $\mathbf{G}_{\pi_r*}$, for which an image heatmap representation shows groupings far more easily perceived by the user.

In a similar fashion, it is straightforward to define an ordering in the columns, by first computing a 1D-mapping that transforms the columns $g_j = \varphi_1(\mathbf{G}^T)$, obtaining the permutation of the sorted scalars $\pi_c = \arg\mathrm{sort}(g_1, \ldots, g_m)$, and finally reordering the columns to yield the sorted matrix $\mathbf{G}_{*\pi_c}$. Finally, applying both methods to obtain the matrix simultaneously sorted by rows and columns is also straightforward, as $\mathbf{G}_{\pi_r\pi_c}$. A schematic diagram of the GEM reordering based on DR described in this section can be seen in Fig. 1.

E. Dual 2D projections

In general, matrix reordering methods used to sort rows or columns according to similarities, including the ones described in Section II-D, but also many other standard techniques, such as hierarchical clustering techniques used in heatmap visualizations [28], are inherently a 1D projection problem, since they imply a mapping between row or column vectors on a 1D vector of scores used to define the permutations to be done for rows or columns. This poses an important limitation, since regions in the input space with a larger intrinsic dimensionality may require more than one factor to visually explain how samples are organized.

Therefore, we propose *dual* 2D UMAP projections —similar to the dual t-SNEs proposed by [23]— as a powerful complementary method for clustering (and relating) samples and genes. This is done by defining two 2D dimensionality reduction mappings, one for samples and one for genes, which we call the *sample view* and the *gene view*, respectively. Both mappings, visually represented as scatterplots, can be used to interact with the GEM in many ways and may also be recomputed conditioned by user-selected subsets of samples or genes —described later in II-H—, providing a powerful comprehensive view, better than the GEM or the dual projections alone.

1) The sample view: A 2D mapping learned using UMAP, $\varphi_2 : \mathbb{R}^m \to \mathbb{R}^2$, is applied to all the samples (rows) of the GEM to produce an $n \times 2$ sample projection matrix $\mathbf{P} = \varphi_2(\mathbf{G})$. The rows of $\mathbf{P}$, i.e., $\mathbf{p}_{i*}$, are coordinates of 2D points that are visually displayed in a scatterplot representation, called the *sample view*.

Each point i in the sample view has a color that represents the expression level of the currently selected gene j for that sample. Since close samples in the latent space have similar gene expression profiles, spatial changes in the color result in visually identifiable patterns of expression for gene j that help the user to spot interesting details. Whenever the user changes the current gene j in the pointer selection with a simple mouse move over the GEM view, this color pattern of expressions is immediately and *fluidly* recomputed, allowing the user to *browse* the whole gene collection looking for relevant patterns across the samples.

This scatterplot can be used to interact with the GEM in a bidirectional way. The user can interactively select samples on the sample view using a *lasso* selection. This produces a highlight of the corresponding rows in the GEM, which is updated in real time, as the user draws the lasso, allowing them to obtain immediate feedback on the relationships between both views. On the other hand, a selection of rows in the GEM (the vertical span of a box selection) will highlight the samples in this view. These selections of samples can also be used to define a conditional reordering of the GEM genes (columns) —see section II-G.

2) The gene view: Similarly, a 2D mapping learned with UMAP for the genes, $\varphi_2 : \mathbb{R}^n \to \mathbb{R}^2$, is applied to all m genes (each a column with n expression levels) of the GEM to produce an $m \times 2$ *gene projection matrix* $\mathbf{Q} = \varphi_2(\mathbf{G}^T)$. The rows of $\mathbf{Q}$, i.e., $\mathbf{q}_{i*}$, are also 2D points, that are represented in a scatterplot representation, called the *gene view*.

F. Bidirectional interactions among the views

Linked selection is a powerful interaction mechanism able to highlight relationships among different representations of a multifaceted problem. This mechanism —which implies both selection and connection [29]— is particularly suited for interactive GEM analysis. Indeed, an interface may have at least these three elements sharing indices:

$\mathbf{G} = (g_{ij})$ a gene expression matrix with two dimensions, samples i and genes j that is presented in a heatmap representation. Selections of items both in samples and genes can be performed.

$\mathbf{P} = (\mathbf{p}_{i*})$ a 2D projection of the samples represented in a scatterplot view. This view allows the selection of a subset of the samples.

$\mathbf{Q} = (\mathbf{q}_{j*})$ a 2D projection of the genes also represented in a scatterplot view. This view allows the selection of a subset of the genes.

Four bidirectional linked selection operations can be carried out in a simple way, as shown in Fig. 2: (a) *point selection*, matching the i, j pixel in the GEM with sample i in the sample view and gene j in the gene view; (b) *sample matching*, relating a row of the GEM to a point in the sample view; (c) *gene matching*, relating a column of the GEM to a point in the gene view; and (d) *bicluster matching*, relating an area of the GEM to a subset I of samples and a subset J of genes in the sample and the gene view respectively. For instance, the user can select samples in the 2D sample view or genes in the 2D gene view and see the highlighted rows ($\mathbf{G}_{\sigma_s*}$) or columns ($\mathbf{G}_{*\sigma_g}$) in the heatmap. Conversely, the user can select a box area in the heatmap, and see the highlighted points both in the sample view ($\mathbf{P}_{\sigma_s*}$) and the gene view ($\mathbf{Q}_{\sigma_g*}$).

These operations do not involve algorithmic computations and can be performed at framerate. Lasso or box selection tools available in most interactive visualization libraries can trigger updated events for any incoming or outgoing point in the selection area "on the fly", allowing for *fluid interaction* [30], whereby the user queries and the immediate feedback are synchronized, resulting in a seamless analysis loop, fostering user engagement and boosting the discovery of interesting patterns.

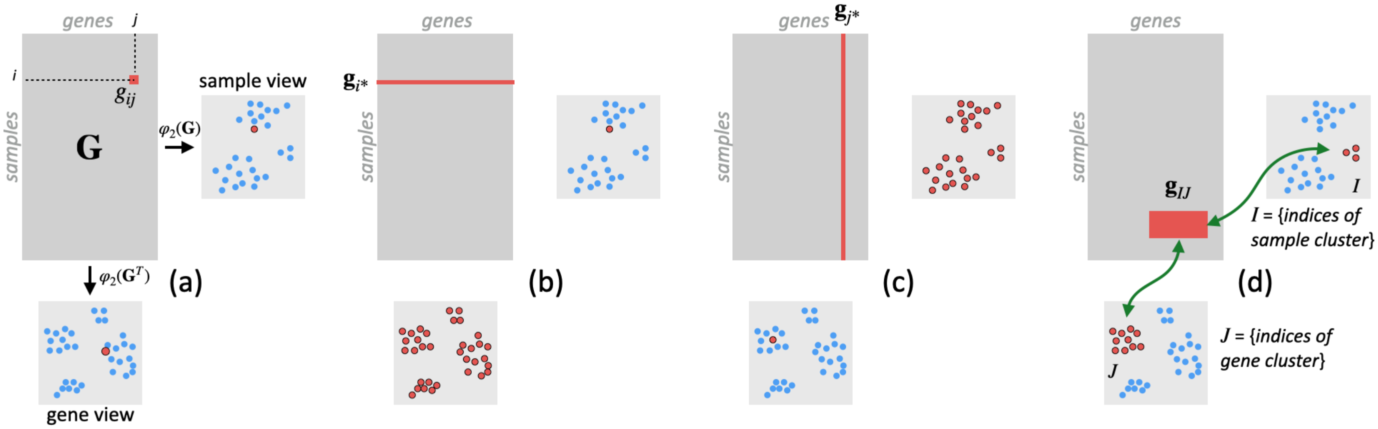

Fig. 2. Bidirectional interactions between the GEM and the dual maps (sample view and gene view): (a) point selection; (b) sample selection; (c) gene selection; (d) bicluster selection.

G. Conditional matrix reordering

While most state-of-the-art tools allow us to reorder the GEM both by genes and samples, it is possible to customize the ordering criteria by selecting the attributes considered. With the approach described in this paper, this can be done through simple interaction pipelines.

Example: conditional sorting of GEM rows (samples). This example describes a procedure for GEM reordering of rows according to similarity in the expression pattern of the samples for a specific set of genes (for instance, known gene clusters, or genes known to have specific functions of interest):

1) The user selects a cluster of related genes in the gene view (e.g. using a *lasso selection* tool) or directly specifying them in a text window. This produces a selection σ that includes the indices of the selected genes.
2) Compute a 1D-DR mapping to project the samples using only the selected genes as attributes, first applying the column selection operator on $\mathbf{G}$ and then projecting the rows (samples) of the resulting matrix, that is, $\varphi_1\left(\mathbf{G}_{*\sigma}\right)$.
3) Sort the resulting scalars g_i and apply the resulting permutation of samples $\mathbf{G}^{\text{sorted}} = \mathbf{G}_{\pi*}$, where $\pi = \arg\text{sort}(g_1, \ldots, g_n)$.

Following similar procedures as the one in the previous example, other conditioning schemes can be carried out. For instance, conditional sorting of the GEM columns (genes) can be performed by $\mathbf{G}^{\text{sorted}} = \mathbf{G}_{*\pi}$ with

$$\pi = \arg\text{sort}(\varphi_1[(\mathbf{G}_{\sigma*})^T])$$

Finally, simultaneous sorting of rows and columns conditioned by subsets of genes and samples, respectively can be achieved by $\mathbf{G}^{\text{sorted}} = \mathbf{G}_{\pi_s\pi_g}$ with

$$\begin{aligned}\pi_s &= \arg\text{sort}(\varphi_1(\mathbf{G}_{*\sigma_g}))\\ \pi_g &= \arg\text{sort}(\varphi_1[(\mathbf{G}_{\sigma_s*})^T])\end{aligned}$$

These procedures, while simple (they indeed can be carried out with a few clicks), provide powerful analytic capabilities, since they allow us to *condition* the matrix reordering patterns to user-specified elements having known common traits. This makes it possible, for instance, to organize the matrix samples according to their similarity in a specific set of genes belonging to a functional cluster or known to take part in a certain pathway. Conversely, the GEM columns (genes) can be organized according to their expression patterns for a certain biological condition (e.g., a cancer type) represented by a selection of the samples.

H. Conditional rearrangement of the 2D views

The sample and gene views, being 2D projections, provide a far more powerful —yet complementary— representation of the similarities between the samples and the genes than the implicit 1D mechanism behind any reordering operation of rows or columns of the GEM.

In addition to conditional GEM reordering, the user can compute a conditional 2D projection of the samples using only a selection σ of genes to define similarity as $(\mathbf{p}_{i*}) = \varphi_2\left(\mathbf{G}_{*\sigma}\right)$. The resulting 2D points $\mathbf{p}_i*$ reflect the samples organized by their similarities in the expression of the selected genes. Similarly, a conditional 2D projection of the genes according to a subset σ of the samples (e.g., a cancer subtype) $(\mathbf{q}_{j*}) = \varphi_2\left[(\mathbf{G}_{\sigma*})^T\right]$.

III. Results

Fig. 3 describes the workflow of the analysis following the knowledge model of [31]. The gene expression visualizations (GEM, dual views) are fed to the user through the perception P, increasing their current knowledge K. This knowledge can condition how information is perceived (P), suggesting new questions, patterns to look and further reconfiguration (exploration E) of the GEM and the dual views through classical zoom, pan, etc., as well as through selection of biologically meaningful groups of samples and genes to condition the GEM ordering and dual map computation (sample view and gene view).

A. Case 1. Similar expression profile of KIRC and VHL-mutated PCPG for a specific set of genes

To start the analysis with a big picture, we decide to visualize a bicluster arrangement of the GEM selecting, for

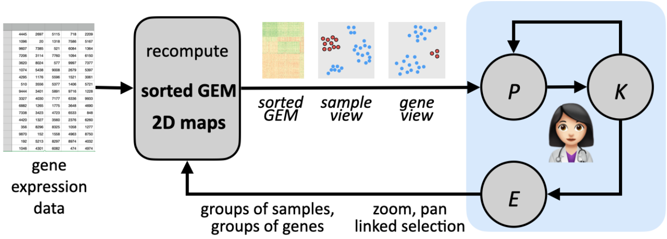

Fig. 3. Workflow for knowledge discovery with the proposed approach.

example, 4 clusters as an initial specification (see Fig. 4). We see that the PCPG, KIRC and normal kidney samples are clearly separated, revealing a remarkably different genetic profile, for the genes considered in the analysis, between both cancer types, as well as between KIRC and normal renal tissue.

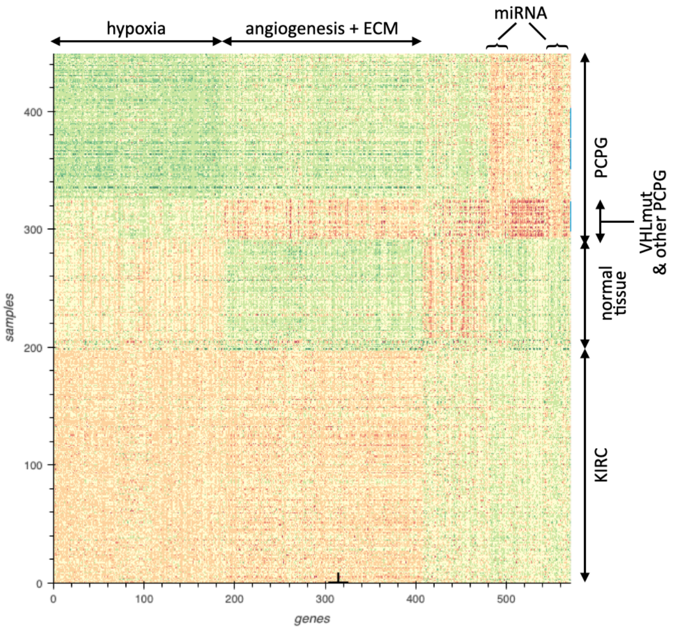

Fig. 4. Bicluster arrangement of the GEM.

Alternatively, we can carry out a 1D-UMAP sorting of the GEM, involving rows and columns, resulting in the GEM arrangement shown on the left of Fig. 5. In this case, a more detailed bicluster organization is observed, where biclusters are not restricted to rectangular areas; however, the qualitative information about subsets of genes and samples is consistent with the previous bicluster representation.

Combining the other elements in the tool with any of the former GEM arrangements, a more detailed analysis can be performed that actually provides richer and more detailed genetic information, highlighting relevant biclusters that involve featured groups of samples, which are related to featured groups of genes. A first bicluster could be identified comprising a subset of the PCPG samples with a gene expression profile more similar to KIRC than to the rest of PCPG samples. An analysis of the genetic features published for these samples [32] revealed that they correspond to all PCPG tumors carrying mutations in VHL genes, a genetic defect also present in most KIRC samples. Complementary to the former, their location in the 2D sample view reveals their relative position with respect to the rest of the samples.

Using basic interaction mechanisms (zoom and pan in the GEM view, as well as in the sample view and the gene view), the user can quickly identify a set of genes commonly upregulated in KIRC and VHL-PCPG (large dotted area of the GEM view in Fig. 5). A visual inspection of this area, with the help of linked selections, that connect the GEM and the 2D views, revealed the presence of three biclusters that were found to correspond to functionally related genes: 1) genes of the hypoxia pathway highly overexpressed in all KIRC samples but moderately overexpressed in VHL-PCPG samples; 2) genes of the extracellular matrix (ECM in Fig. 5) upregulated in both, KIRC and VHL-PCPG; and 3) genes involved in angiogenesis upregulated in VHL-PCPG and most KIRC but not altered in approximately 30% of KIRC (smaller dotted area in the rightmost part and bottom half of the GEM).

Collectively, this analysis provided compelling evidence that VHL-PCPG and KIRC, as expected, share a similar hypoxia-related gene expression profile. Importantly, it also allowed the finding of nonpreviously published data: a bicluster of KIRC not overexpressing angiogenic genes compared with normal kidney tissue and most KIRC.

B. Case 2. Discovering genes with similar behavior for tumor subtypes

As another interesting case of analysis, our approach allows the discovery of genes with similar genetic behavior for a subset of the samples that corresponds to tumors of specific cancer subtypes. In this case —see Fig. 6— the biomedical researcher is interested in the behavior of the `SLC16A3` gene. This gene is essential for cancer cell survival and predicts tumor progression in patients with KIRC [34], [35]. It has been biochemically associated with VHL-regulated signaling pathways [34] such that it is expected to be upregulated in VHL-deficient KIRC and PCPG. The `SLC16A3` gene can be located either interactively using zoom, pan and mouseover with pointer selection in the GEM or by a simple text search. According to our prediction, all KIRC samples and the VHL-mutated PCPG tumors are highlighted in orange color in the resulting sample view indicating that SLC16A3 is overexpressed in the two types of tumors.

These samples can then be selected for conditioned analysis in the resulting 2D sample view, and then the gene view can be updated so that it reflects the similarities of the genes' expressions, *constrained* to the selected samples. It is expected that genes in the neighborhood of `SLC16A3` in the resulting gene view show similar gene expression behavior for KIRC and VHL-mutated samples. A simple inspection reveals that some of the neighboring genes and miRNAs are `ALDOA`, `ENO1`, `FAM57A`, `GYS1`, `RAB40C`, `SLC16A3`, `TPI1` and `hsa-let-7b-5p`. Hovering the mouse pointer in the GEM view over each of these genes, the user can observe their expressions with a color scale across all samples in the sample view. In Fig. 6 the expression patterns in the sample view for two of the neighboring genes (`ENO1` and `FAM57A`) are shown, confirming that they are

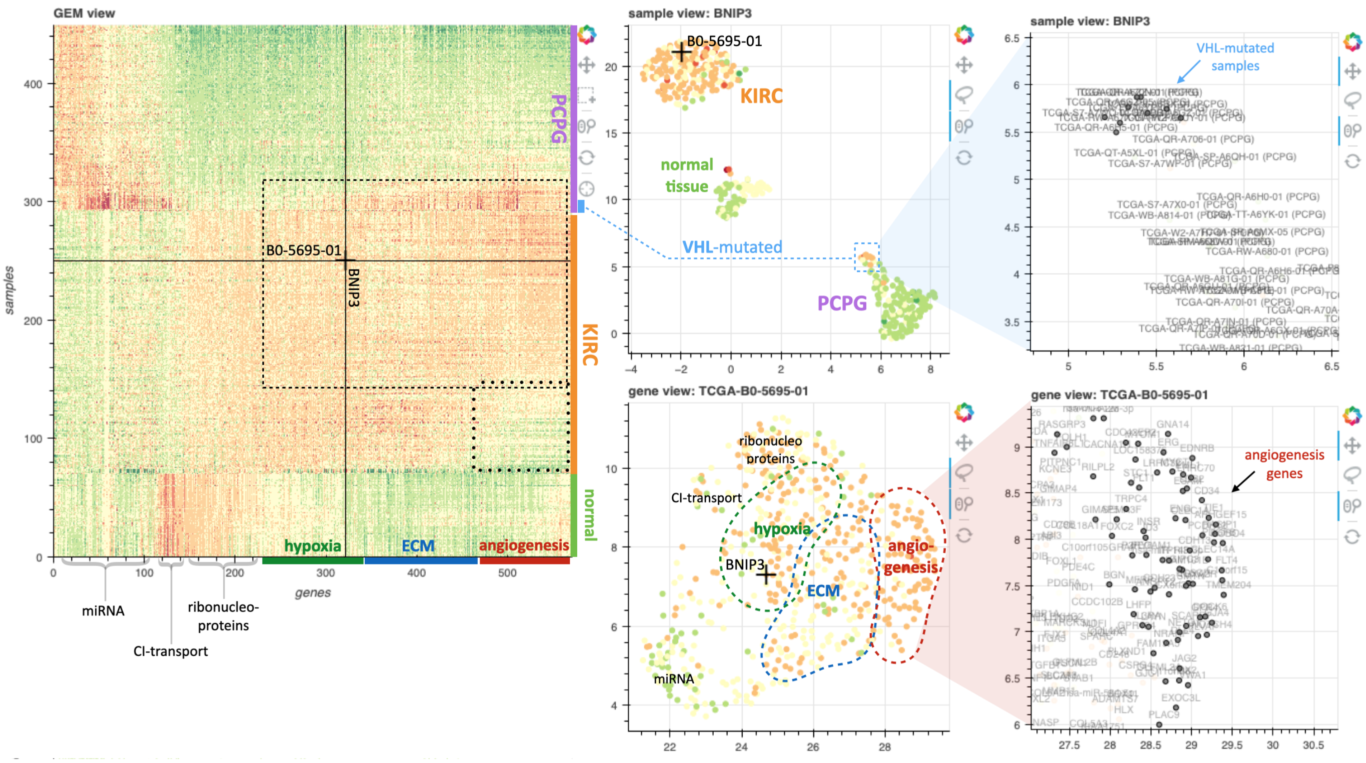

Fig. 5. Case 1. On the left, the GEM with the pointer selection marker, selecting at this moment the sample `TCGA-B0-5695-01` and gene `BNIP3`. The cancer types and subtypes, as well as relevant clusters of genes have been color-highlighted on the right and bottom sides of the GEM. On the right, the gene and sample views with the corresponding sample and gene clusters and zoom views (rightmost subfigures). Note the pointer selection marker also appears on both views showing the current sample and gene positions. The annotated areas reveal a predominance of genes with the corresponding function, according to a gene enrichment analysis with Metascape® tool [33].

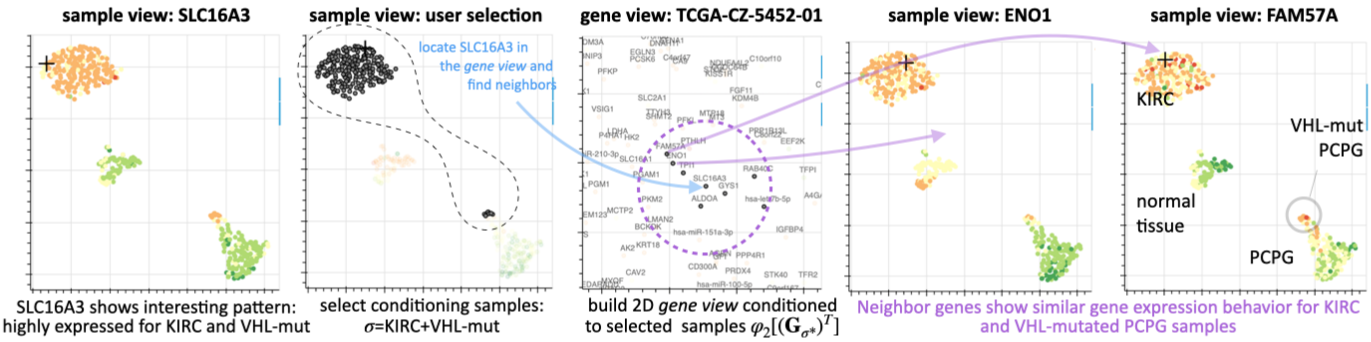

Fig. 6. Workflow for discovering genes with similar behavior to `SLC16A3` on KIRC samples and VHL-mutated PCPG samples.

upregulated in KIRC and VHL-mutated samples, as in the case of `SLC16A3`.

C. Case 3. Conditional GEM sort and 2D maps

Conditional analysis according to angiogenesis genes: Our approach may also help to identify similarities and differences between tumors with different gene mutations. We illustrate in Fig. 7 how, by using the 2D sample view, the different gene expression profiles in PCPGs carrying mutations in the VHL or SDH genes can be easily and interactively studied. By selecting hypoxia-related genes known to be involved in angiogenesis (see supplementary material for a list of genes considered), we can conditionally rearrange the GEM and the sample view to identify similarities between samples according to the angiogenesis gene set. This analysis led to the reconfiguration of the sample view showing increased closeness of PCPG and KIRC samples, indicating that the angiogenic profiles of both types of tumors are similar. More specifically, considering two selections of PCPG samples, one for those with VHL mutations and the other for those with SDH mutations, it also highlights that VHL-PCPG samples cluster together in the vicinity of KIRC. In contrast, PCPGs with mutations in SDH genes are scattered among PCPGs lacking VHL or SDH mutations. Therefore, VHL-PCPG but not SDH-PCPG are closely related to angiogenesis in KIRC. Similar conclusions can be drawn from the conditioned GEM, as shown in Fig. 7. VHL mutated PCPG appear as rows in the KIRC cluster of the GEM (except one that appears in

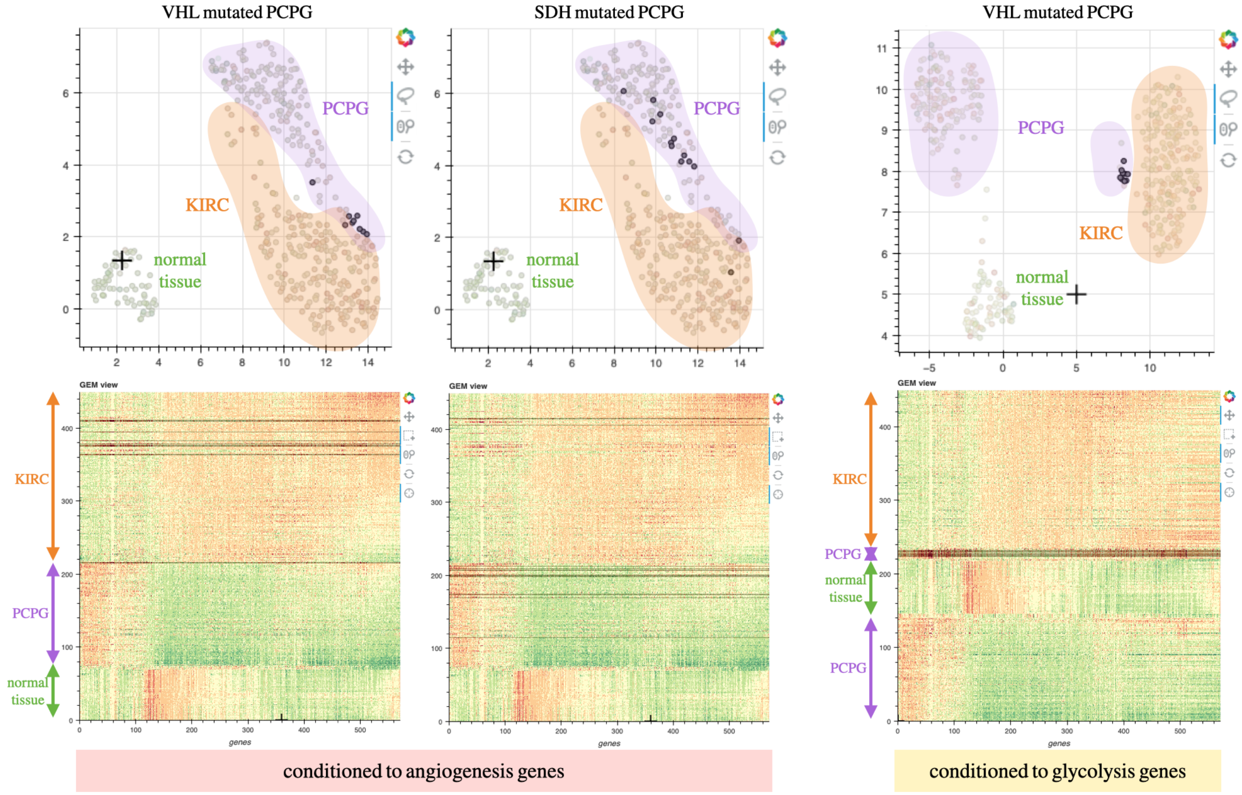

Fig. 7. Conditional analysis using the GEM and the 2D sample view. Left two columns (according to angiogenesis genes): VHL-mutated and SDH-mutated PCPG samples shown as dark points and rows. Right column (according to glycolysis genes): VHL-mutated PCPG samples shown as dark points and rows.

the PCPG cluster, but next to the KIRC cluster), confirming that their angiogenic profile is similar to that of KIRC. Additionally, in accordance with the sample view, most SDH samples appear in the PCPG cluster of the GEM, except two that are in the KIRC cluster.

Conditional analysis according to glycolysis genes: A similar analysis using a different gene set shows relevant novel data. The hypoxia-related genes involved in glycolysis can be selected for reordering the samples in the GEM view and recomputing the sample view. This analysis shows that PCPG and KIRC samples remain apart in differentiated clusters except for VHL-PCPG samples that lay in the vicinity of KIRC samples, thus revealing that activation of glycolysis is a feature of tumors with deleterious mutations or deletions of the VHL gene and that this does not occur in other types of PCPG, including PCPG with SDH mutations. This situation is clearly shown in the GEM, where all the VHL-mutated samples appear together next to the KIRC cluster.

IV. Discussion

The first case study demonstrates the potential of dual representation of samples and genes on similarity maps, where the user can utilize common tools such as pan, zoom and select to explore the data and reveal functionally related genes that display similar expression patterns in different types of cancer (KIRC and VHL-mutated PCPG). The exploratory analysis shows, in an intuitive and rapid manner, a new discovery in the data: a group of KIRC samples that do not overexpress angiogenesis genes in the same way as the rest. This type of finding can lead to new hypotheses.

The other two case studies highlight the potential of interactive conditional reorganization of the visualization based on both sample and gene subgroups. In Case 2, selecting KIRC tissues and VHL-mutated PCPG samples in the sample view allows for rapid identification of genes that behave similarly in these subgroups. In Case 3, reorganizing the GEM based on only selected genes in the gene view shows the similarities and differences in KIRC and PCPG tissues with VHL or SDH mutations with regard to functions such as angiogenesis or glycolysis.

There are several methods in the literature for visualizing the gene expression matrix, including those that incorporate interaction mechanisms and more complex bicluster representations with overlapping clusters [11], [12], [14]. However, the most commonly used methods yield simple checkerboard representations. Our proposed 1D-DR reordering technique presents an alternative that retains the heatmap representation while providing more detailed and specific regions in the GEM. Unlike some biclustering methods, it does not require a predetermined number of clusters to be identified. For instance, in one case, the 1D-DR reordered GEM allowed for the identification of a specific bicluster in KIRC samples that was not visible in the standard checkerboard representation.

Most importantly, using 2D projections, such as sample maps and gene maps, can help to overcome the inherent 1D limitation in GEM reordering. These projections provide an extra degree of freedom for representing samples and genes, allowing for a more detailed representation of the mutual relationships, similarities, and cluster structures among genes and samples.

The proposed method also has some limitations that may be noted. One of them is scalability. The GEM, like any other heatmap representation, has limitations on the amount of data that it can effectively represent. The size of the data also impacts the efficiency of the algorithms used for reordering the GEM and calculating the 2D maps. As a result, the latencies in the workflow can make it less smooth and the recommended GEM size is limited to approximately 1,000 items per dimension.

It must also be considered that the proposed approach is exploratory in nature. Unlike machine learning methods, it is not automated and does not provide precise or quantifiable answers. It requires a user-guided process, and may be prone to subjectivity and potential misinterpretation by the user. Additionally, it is worth mentioning that while our approach allows for the discovery of bicluster structures that are more general than the baseline checkerboard type, it is limited to nonoverlapping constant value biclusters, according to [4]. However, the proposed approach does not lose validity if another GEM sorting algorithm is used that detects more complex types of biclusters, but this remains an area for future work.

It may be challenging to fully understand the value of such an interactive process through a written description alone (it is suggested to view the accompanying video for further understanding). However, the case studies demonstrate the potential of the proposed technique. By combining a GEM heatmap visualization with linked 2D dual representations, the technique offers a more comprehensive and useful representation of the data compared to the traditional checkerboard representation. Additionally, the use of conditional reordering based on user-selected subsets of genes and samples in various views allows for a more in-depth exploratory analysis that is guided by the biomedical researcher's domain knowledge, rather than that of the algorithms used. This is, in our opinion, a key differential factor for success in its utilization.

V. Conclusions

In this paper, we have proposed a visual analytics approach that integrates machine learning, data visualization and user interaction in the analysis pipeline, helping the user to keep insight along the whole discovery process.

To achieve this, we combined two related analysis mechanisms: the visualization of the GEM and dual 2D UMAP projections of genes and samples. Although both techniques —especially the first one— have been used in the literature, our approach bridges them in novel ways by means of tight interaction elements involving linked selections and conditional rearrangement of both the GEM and the dual 2D projections, according to user-specified subsets of genes and/or samples.

We tested the proposed approach on three case studies, showing potentially relevant discoveries through a progressive exploration process where the user keeps insight through all the steps. It should be stressed, however, that while our approach facilitates the exploration of gene expression data, it does not generate medical results or evidence. It allows users to raise hypotheses that could lead to new knowledge, but the quality of the data and the user's judgment are important factors in this process. During the analysis, conclusions should be accompanied by the assumptions and hypotheses used to generate them to avoid bias, and the observations should not be considered conclusive and must be subsequently validated by other means.

We argue that properly designed interaction mechanisms to efficiently connect complementary techniques (visualization and/or machine learning), and having the user as an active agent in the analytics discourse lead to improved knowledge discovery, with respect to the techniques used independently. The framework of analysis presented here opens new avenues for the development of tools that integrate other data visualization techniques and machine learning algorithms for knowledge discovery in transcriptomic data analysis. Our methodology can also be extrapolated to other high dimensional biomedical problems that can be posed in data matrix form, such as single-cell RNAseq analysis, where the samples are individual cells, allowing for a more in-depth understanding of cell diversity and leading to broader types of analyses involving, for instance, cell-wide response to stimuli and conditions.

Finally, while the findings are potentially relevant and provide insight into the role of hypoxia mechanisms in KIRC and PCPG cancers, they are preliminary results aiming mainly to demonstrate the usefulness of the proposed approach for discovering valuable knowledge from genomic data. The insights obtained from these results suggest further work including further downstream analyses to confirm their clinical utility, such as for the development of biomarkers.

VI. Availability and Supplementary Material

A video demo of the approach and its application to the case studies in this paper is available at `https://youtu.be/NcJSxbF9E4w`. The source code of a small demo app used for the results is available at `https://gitlab.com/idiazblanco/gem-i` and can be tried at `https://gsdpi.edv.uniovi.es/matrices_ordenables_interactivas`. Finally, we also include the lists of genes and samples used in the case studies.

Acknowledgments

The results shown here are based upon data generated by the TCGA Research Network: https://www.cancer.gov/tcga.

Funding

This work was supported by the Ministerio de Ciencia e Innovación / Agencia Estatal de Investigación (MCIN/AEI/10.13039/501100011033) grant [PID2020-115401GB-I00]. The authors would also like to thank the financial support provided by the Principado de Asturias government through the predoctoral grant "Severo Ochoa".

References

[1] P. Carmona-Saez, R. D. Pascual-Marqui, F. Tirado, J. M. Carazo, and A. Pascual-Montano, "Biclustering of gene expression data by non-smooth non-negative matrix factorization," *BMC bioinformatics*, vol. 7, no. 1, pp. 1–18, 2006.

[2] J. A. Hartigan, "Direct clustering of a data matrix," *Journal of the american statistical association*, vol. 67, no. 337, pp. 123–129, 1972.

[3] Y. Cheng and G. M. Church, "Biclustering of expression data," in *Proceedings of the International Conference on Intelligent Systems for Molecular Biology*, vol. 8, 2000, pp. 93–103.

[4] S. C. Madeira and A. L. Oliveira, "Biclustering algorithms for biological data analysis: a survey," *IEEE/ACM transactions on computational biology and bioinformatics*, vol. 1, no. 1, pp. 24–45, 2004.

[5] B. Pontes, R. Giráldez, and J. S. Aguilar-Ruiz, "Biclustering on expression data: A review," *Journal of biomedical informatics*, vol. 57, pp. 163–180, 2015.

[6] J. J. Thomas and K. A. Cook, "A visual analytics agenda," *IEEE computer graphics and applications*, vol. 26, no. 1, pp. 10–13, 2006.

[7] A. Endert, W. Ribarsky, C. Turkay, B. W. Wong, I. Nabney, I. D. Blanco, and F. Rossi, "The state of the art in integrating machine learning into visual analytics," *Computer Graphics Forum*, vol. 36, no. 8, pp. 458–486, 2017.

[8] J. P. Gonçalves, S. C. Madeira, and A. L. Oliveira, "Biggests: integrated environment for biclustering analysis of time series gene expression data," *BMC research notes*, vol. 2, no. 1, pp. 1–11, 2009.

[9] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, and E. Zitzler, "Bicat: a biclustering analysis toolbox," *Bioinformatics*, vol. 22, no. 10, pp. 1282–1283, 2006.

[10] K.-O. Cheng, N.-F. Law, W.-C. Siu, and T. Lau, "Bivisu: software tool for bicluster detection and visualization," *Bioinformatics*, vol. 23, no. 17, pp. 2342–2344, 2007.

[11] R. Santamaría, R. Therón, and L. Quintales, "Bicoverlapper: a tool for bicluster visualization," *Bioinformatics*, vol. 24, no. 9, pp. 1212–1213, 2008.

[12] R. Santamaría, R. Therón, and L. Quintales, "Bicoverlapper 2.0: visual analysis for gene expression," *Bioinformatics*, vol. 30, no. 12, pp. 1785–1786, 2014.

[13] J. Heinrich, R. Seifert, M. Burch, and D. Weiskopf, "Bicluster viewer: a visualization tool for analyzing gene expression data," in *International Symposium on Visual Computing*. Springer, 2011, pp. 641–652.

[14] H. Aouabed, R. SantamaríA, and M. Elloumi, "Visbicluster: A matrix-based bicluster visualization of expression data," *Journal of Computational Biology*, vol. 27, no. 9, pp. 1384–1396, 2020.

[15] Y. Wu, P. Tamayo, and K. Zhang, "Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding," *Cell systems*, vol. 7, no. 6, pp. 656–666, 2018.

[16] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne." *Journal of machine learning research*, vol. 9, no. 11, 2008.

[17] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," arXiv, https://arxiv.org/abs/1802.03426, 2018. [Online]. Available: https://arxiv.org/abs/1802.03426

[18] D. Kobak and P. Berens, "The art of using t-sne for single-cell transcriptomics," *Nature communications*, vol. 10, no. 1, pp. 1–14, 2019.

[19] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell, "Dimensionality reduction for visualizing single-cell data using umap," *Nature biotechnology*, vol. 37, no. 1, pp. 38–44, 2019.

[20] F. A. Wolf, P. Angerer, and F. J. Theis, "Scanpy: large-scale single-cell gene expression data analysis," *Genome biology*, vol. 19, no. 1, pp. 1–5, 2018.

[21] R. Hillje, P. G. Pelicci, and L. Luzi, "Cerebro: interactive visualization of scrna-seq data," *Bioinformatics*, vol. 36, no. 7, pp. 2311–2313, 2020.

[22] M. Hawrylycz, L. Ng, D. Page, J. Morris, C. Lau, S. Faber, V. Faber, S. Sunkin, V. Menon, E. Lein *et al.*, "Multi-scale correlation structure of gene expression in the brain," *Neural networks*, vol. 24, no. 9, pp. 933–942, 2011.

[23] S. M. Huisman, B. Van Lew, A. Mahfouz, N. Pezzotti, T. Höllt, L. Michielsen, A. Vilanova, M. J. Reinders, and B. P. Lelieveldt, "Brainscope: interactive visual exploration of the spatial and temporal human brain transcriptome," *Nucleic acids research*, vol. 45, no. 10, pp. e83–e83, 2017.

[24] L. Celada, T. Cubiella, J. San-Juan-Guardado, A. San José Martínez, N. Valdés, P. Jiménez-Fonseca, I. Díaz, J. M. Enguita, A. Astudillo, E. Álvarez-González *et al.*, "Differential hif2α protein expression in human carotid body and adrenal medulla under physiologic and tumorigenic conditions," *Cancers*, vol. 14, no. 12, p. 2986, 2022.

[25] R. Sharan, A. Maron-Katz, and R. Shamir, "Click and expander: a system for clustering and visualizing gene expression data," *Bioinformatics*, vol. 19, no. 14, pp. 1787–1799, 2003.

[26] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14 863–14 868, 1998.

[27] L. Van Der Maaten, E. Postma, J. Van den Herik *et al.*, "Dimensionality reduction: a comparative," *J Mach Learn Res*, vol. 10, no. 66-71, p. 13, 2009.

[28] L. Wilkinson and M. Friendly, "The history of the cluster heat map," *The American Statistician*, vol. 63, no. 2, pp. 179–184, 2009.

[29] J. S. Yi, Y. ah Kang, J. Stasko, and J. A. Jacko, "Toward a deeper understanding of the role of interaction in information visualization," *IEEE transactions on visualization and computer graphics*, vol. 13, no. 6, pp. 1224–1231, 2007.

[30] N. Elmqvist, A. V. Moere, H.-C. Jetter, D. Cernea, H. Reiterer, and T. Jankun-Kelly, "Fluid interaction for information visualization," *Information Visualization*, vol. 10, no. 4, pp. 327–340, 2011.

[31] J. J. Van Wijk, "The value of visualization," in *VIS 05. IEEE Visualization, 2005.* IEEE, 2005, pp. 79–86.

[32] L. Fishbein, I. Leshchiner, V. Walter, L. Danilova, A. G. Robertson, A. R. Johnson, T. M. Lichtenberg, B. A. Murray, H. K. Ghayee, T. Else *et al.*, "Comprehensive molecular characterization of pheochromocytoma and paraganglioma," *Cancer cell*, vol. 31, no. 2, pp. 181–193, 2017.

[33] Y. Zhou, B. Zhou, L. Pache, M. Chang, A. H. Khodabakhshi, O. Tanaseichuk, C. Benner, and S. K. Chanda, "Metascape provides a biologist-oriented resource for the analysis of systems-level datasets," *Nature communications*, vol. 10, no. 1, pp. 1–10, 2019.

[34] M. S. Ullah, A. J. Davies, and A. P. Halestrap, "The plasma membrane lactate transporter mct4, but not mct1, is up-regulated by hypoxia through a hif-1α-dependent mechanism," *Journal of Biological Chemistry*, vol. 281, no. 14, pp. 9030–9037, 2006.

[35] Y. Kim, J.-W. Choi, J.-H. Lee, and Y.-S. Kim, "Expression of lactate/h+ symporters mct1 and mct4 and their chaperone cd147 predicts tumor progression in clear cell renal cell carcinoma: immunohistochemical and the cancer genome atlas data analyses," *Human pathology*, vol. 46, no. 1, pp. 104–112, 2015.