scMuffin: an R package for disentangling solid tumor heterogeneity from single-cell expression data

Valentina Nale[1], Alice Chiodi[1], Noemi Di Nanni[1], Ingrid Cifola[1], Marco Moscatelli[1], Cinzia Cocola[1], Matteo Gnocchi[1], Eleonora Piscitelli[1], Ada Sula[1], Ileana Zucchi[1], Rolland Reinbold[1], Luciano Milanesi[1], Alessandra Mezzelani[1], Paride Pelucchi[1*+], and Ettore Mosca[1*+]

[1]National Research Council, Institute of Biomedical Technologies, Via Fratelli Cervi 93, 20054, Segrate (Milan), Italy

e-mail addresses:

- ettore.mosca@itb.cnr.it

- paride.pelucchi@itb.cnr.it

[+]Corresponding author

*Equal contribution

Abstract

INTRODUCTION: Single-cell (SC) gene expression analysis is crucial to dissect the complex cellular heterogeneity of solid tumors, which is one of the main obstacles for the development of effective cancer treatments. Such tumors typically contain a mixture of cells with aberrant genomic and transcriptomic profiles affecting specific sub-populations that might have a pivotal role in cancer progression, whose identification eludes bulk RNA-sequencing approaches. We present scMuffin, an R package that enables the characterization of cell identity in solid tumors on the basis of a various and complementary analyses on SC gene expression data.

RESULTS: scMuffin provides a series of functions to calculate qualitative and quantitative scores, such as: expression of marker sets for normal and tumor conditions, pathway activity, cell state trajectories, CNVs, transcriptional complexity and proliferation state. Thus, scMuffin facilitates the combination of various evidences that can be used to distinguish normal and tumoral cells, define cell identities, cluster cells in different ways, link genomic aberrations to phenotypes and identify subtle differences between cell subtypes or cell states. We analysed public SC expression datasets of human high-grade gliomas as a proof-of-concept to show the value of scMuffin and illustrate its user interface. Nevertheless, these analyses lead to interesting findings, which suggest that some chromosomal amplifications might underlie the invasive tumor phenotype and the presence of cells that possess tumor initiating cells characteristics.

CONCLUSIONS: The analyses offered by scMuffin and the results achieved in the case study show that our tool helps addressing the main challenges in the bioinformatics analysis of SC expression data from solid tumors.

Keywords: single-cell transcriptomics, cancer, tumor heterogeneity, cell identity.

1. Background

Single-cell (SC) gene expression analysis is crucial to dissect the complex cellular heterogeneity of solid tumors, which is one of the main obstacles for the development of effective cancer treatments (1). A relevant number of software tools has been developed in recent years in the field of SC data analysis (2), a fact that stresses the key opportunities and challenges in this field. A recent study has shown that the development of tools that address common tasks (e.g., clustering of similar cells) and ordering of cells (e.g., definition of cell trajectories) is decreasing, while a greater focus is being paid on data integration and classification (2). These observations reflect the growing availability, scale and complexity of SC datasets (2).

SC datasets of solid tumors are typical examples of complex datasets that present a series of computational challenges and whose analysis demands domain-specific and integrative approaches. In fact, solid tumors typically contain a mixture of cells with aberrant genomic and transcriptomic profiles affecting specific sub-populations that might play a pivotal role in cancer progression, whose identification eludes bulk RNA-sequencing approaches. The use of cell type-specific markers (when available) is limited, and the alterations of gene expression that mark cancer cells makes the use of markers for normal cells not completely adequate. Moreover, the molecular heterogeneity of cancer cells (due to both intra-tumor and inter-individual differences) poses intrinsic limits in the definition of such markers. In addition, solid tumor samples typically comprise cells from the surrounding tissue or infiltrating cells that need to be distinguished from tumor populations for an effective analysis. Another challenge is the identification of clinically relevant cell subtypes that may be rare in the tumor mass, such as cancer stem cells or drug resistant subclones: because of their relatively low number, these cells are typically clustered together with many others. Lastly, an intrinsic problem of many SC datasets is the sequencing depth limit at the SC level. These limitations bound the number of detectable genes to the few

thousands of the highest expressed genes, which implies, for example, that some established markers may not be used for data analysis.

To address these challenges, we developed scMuffin, an R package that implements a series of complementary analyses aimed at shedding light on the complexity of solid tumor SC expression data, including: a fast and customizable gene set scoring system; gene sets from various sources, including pathways, cancer functional states and cell markers; cell cluster association analysis with quantitative (e.g., gene set scores) as well as categorical (e.g. mutation states, proliferation states) features; copy number variation (CNV) analysis; transcriptional complexity analysis; proliferation rate quantification; and gene set based multi-dataset analysis **(Figure 1)**. scMuffin facilitates the integrative analysis of these multiple features, thus allowing the identification of cell subtypes that elude more general clustering and classification approaches. We describe the key aspects of scMuffin implementation and then its user interface by means of a case study on a public SC expression dataset of human high-grade gliomas (HGG).

2. Implementation

scMuffin is implemented as an R package and provides a series of functions that allows the user to quantify various genetic and phenotypic characteristics of single cells, which can be combined to obtain insights on cell identity and function. The functions of scMuffin operate on a common data structure, the “scMuffinList”, by adding, changing or removing its elements. Computationally intensive tasks (in particular, gene set scoring and CNV inference) are parallelized. Package documentation includes R vignettes and function documentation, which are also available as “GitHub Pages” at the URL https://emosca-cnr.github.io/scMuffin. In this section, we describe package inputs and the main algorithms and definitions underlying the analyses offered by scMuffin.

2.1 Input

scMuffin is intended to be used downstream to general purpose tasks like quality control, normalization, cell clustering and dataset integration, for which there are dedicated tools, such as Seurat (3). scMuffin requires the genes-by-cells raw counts matrix, genes-by-cells normalized expression matrix and a partition of cells in clusters. Typically, these matrices are filtered during preprocessing steps to exclude low quality cells and genes that could negatively affect the analyses. However, the characterization of cells that can be achieved with scMuffin offers insights that can be used to (further) filter the dataset and/or to decide on which cells apply particular analyses (e.g. biomarker identification). In general, according to research questions and experimental design one may want to apply strong or mild filters before using scMuffin.

2.2 Quantification of gene set expression scores at cell and cluster levels

The quantification of gene set expression scores follows the approach described in (4,5), in which a gene set is scored on the basis of its average deviation from an empirical null. The approach is implemented enabling the user to tune various parameters (e.g., the minimum number of cells in which a gen set must be expressed, the number of bins of the null model and the possible removal of missing values) to better address the needs of a particular study in relation to the characteristics of the dataset under analysis. In particular, the code is implemented to handle missing values, which are one of the main issues in SC datasets. For example, this is important to quantify the expression of gene sets that are expressed just in a small subset of cells (and have null values elsewhere), or to account for the expression of different members of the same pathway in different cells.

The algorithm for gene set scoring is the following. Given a gene set S:

1. the genes occurring in the normalized genes-by-cells matrix are grouped into a series of bins according to their average expression across cells;

2. a number k of random gene sets S_i^*: are created, of the same size of S, tossing genes from the same bins of S, in order to match the distribution of gene expression of each S_i^* with that of S;
3. the averages m_c and m_{ic}^* are calculated, respectively, over the values of S and S_i^* in every cell c;
4. the expression score Y_c is calculated as the average difference between m_c and m_{ic}^*:

$$Y_c = \frac{\sum_i^k (m_c - m_{ic}^*)}{k}$$

5. the difference between the medians of m_c and m_{ic}^* values in any cluster is used as the cluster-level score of S.

2.3 CNV estimation based on adjacent gene expression

CNV inference in scMuffin is based on the "adjacent gene windows" approach (4,6), which has been validated using both single nucleotide polymorphism arrays (6) and whole-exome sequencing (4). The approach is implemented in parallel and, like gene set scoring, offers various parameter tuning and data filtering possibilities, which allow the investigator to optimize the analysis to specific needs. The CNV profile of each cell is calculated as a moving average of scaled gene expression levels ordered by genomic location. scMuffin offers the possibility of subtracting a "normal" reference profile to highlight relative CNVs. Since a SC dataset derived from a sample of a solid tumor can either contain or not a proper set of physiological cells to be used as reference, this reference can be derived from cells that are part of the dataset under analysis (as described in (4)), or using external dataset (as described in (6)). The main steps are:

1. the reference cells are added to the genes-by-cells matrix (optional);
2. the expression of each gene is scaled subtracting its average (optional);
3. the gene expression matrix is ordered by chromosome and gene location;

4. in each chromosome *h*, the estimated copy number V_{ic} of cell *c* is calculated for all the ordered genes $i \in \left[\frac{w}{2}+1, n_h - \frac{w}{2}\right]$:

$$V_{ic} = \sum_{j=i-\frac{w}{2}}^{i+\frac{w}{2}} \frac{e_{jc}}{w+1}$$

   where *w* is an even number that defines the window size, that is, the number of genes located before and after gene *i* which contribute to the estimation of V_{ic}, n_h is the number of genes in *h*, and e_{jc} is the gene expression value;
5. V_{ic} values are scaled subtracting their average in a cell (optional);
6. cells are clustered by their CNV profile;
7. the average CNV profile of the normal reference cells is subtracted from all the CNV profiles (optional).
8. the summary CNV score of a cell is calculated as:

$$C_c = \sum_i (V_{ic})^2$$

Cells are clustered based on their CNV profile using Seurat (3). A CNV region of a cluster is composed of the adjacent genes i that have absolute median V_{ic} (within the cluster) higher than a given threshold. This threshold is by default defined as the standard deviation of V_{ic} values of all cells.

2.4 Transcriptional complexity

The transcriptional complexity (TC) is calculated using three different approaches. The TC-ratio T_c^R of a cell is inferred based on the number of expressed genes (g_c) over the number of total mapped reads (transcripts) (t_c):

$$T_c^R = \frac{g_c}{t_c} \cdot \frac{\max_c(t_c)}{\max_c(g_c)}$$

where $g_c = \#\{r_{ic} \geq \alpha\}$, α is a threshold over the gene count r_{ic} and defines the gene *i* as "expressed", and $t_c = \sum_i r_{ic}$. Values of T_c^R higher (lower) than 1 indicate a higher (lower) number of expressed genes in relation to the total transcripts of the cell.

The TC-LMR (linear model residual) of a cell corresponds to its residual in the linear regression model between g_c and t_c:

$$\log_{10}(g_c) = \beta_0 + \beta_1 \log_{10}(t_c) + \varepsilon_c$$

$$T_c^{LMR} = \log_{10}(g_c) - \left[\beta_0 + \beta_1 \log_{10}(t_c)\right].$$

Positive (negative) residuals indicate higher (lower) number of expressed genes in relation to total transcripts.

The TC-H (entropy) is quantified as the transcriptional entropy of the cell (7):

$$T_c^H = -\sum_{ic} [p_{ic} \log (p_{ic})]$$

where $p_{ic} = {}^{r_{ic}}/_{t_c}$. The higher the entropy the higher the number of expressed genes.

2.5 Proliferation rate

The proliferation rate P_c of a cell is quantified as the maximum between the two gene set scores $Y_c(S_1)$ and $Y_c(S_2)$, calculated on the gene sets S_1 and S_2 that contain, respectively, genes involved in G1/S and G2/M cell cycle phases:

$$P_c = \max\left(Y_c(S_1), Y_c(S_2)\right)$$

where S_1 and S_2 are defined as in Tirosh *et al*. (4).

2.6 Cluster enrichment analysis for quantitative and categorical features

The association between cell clusters and any cell-level values is performed through Cell Set Enrichment Analysis (CSEA) and Over Representation Analysis (ORA) for, respectively, quantitative

and categorical values. Importantly, these two types of analysis are implemented in parallel, which is particularly important for CSEA that uses permutations to build an empirical null distribution.

CSEA is the application of Gene Set Enrichment Analysis (GSEA) (8) to cells and cell sets in place of, respectively, genes and gene sets: a ranked cell list is used instead of a ranked gene list, and cell sets (cell clusters) are tested instead of gene sets. In addition, the code is implemented to handle missing values. Therefore, scMuffin tests whether the cells assigned to a cluster are located at the top (bottom) of a ranked list of cells.

The assessment of cluster enrichment in a particular value of a categorical feature is computed using the over-representation analysis (ORA) approach (9), which is based on the hypergeometric test. Here, this test assesses whether the occurrence of a particular value between the cells of a cluster in relation to all other clusters is higher than expected in a hypergeometric experiment.

2.7 Overlap between partitions of cells

Cell partitions resulting from different clustering analyses are compared by calculating the overlap coefficients among all-pairs of clusters. Given two partitions *A* and *B*, defined as sets of cell clusters $A = \{a_i\}$ and $B = \{b_j\}$, the similarity between the cell clusters a_i and b_j is calculated as:

$$o_{ij} = \frac{|a_i \cap b_j|}{\min(|a_i|, |b_j|)}$$

3. Results and Discussion

In this section, we present the user interface **(Table 1)**, taking advantage of the results obtained on the SC dataset generated by Yuan *et al.* (10) from human high-grade glioma (HGG) samples (available on the Gene Expression Omnibus (GEO) repository (11) with identifier GSE103224, see Supplementary Methods in Additional File 1).

Table 1. Main tasks and corresponding functions in scMuffin.

Task	Description	User interface
Create the scMuffinList	• Create the scMuffinList from counts and normalized expression data	• `create_scMuffinList()`
Add a partition	• Add a partition of cell (clusters)	• `add_partition()`
Gene set expression scoring	• Average gene set expression deviation from matched empirical reference; • provided gene sets from CellMarker (12), PanglaoDB (13), CancerSEA (14) and MSigDB (8)	• `prepare_gsls()` • `calculate_gs_scores()` • `calculate_gs_scores_in_clusters()`
CNV	• Estimation of CNVs by means of the "moving window" approach, that is, considering the expression of adjacent genes; • calculation of CNV deviation from a normal reference profile; • processing of normal tissue-specific expression data from GTEx; • detection of CNV regions	• `CNV_analysis()` • `process_GTEx_gene_reads()`
Transcriptional complexity	• Number of expressed genes in relation to the total reads, calculated using a ratio, linear model residuals and entropy.	• `transcr_compl()`
Proliferation	• Maximum between G1/S and G2/M gene set scores	• `proliferation_analysis()`
Cell state trajectory	• Diffusion map computation	• `diff_map()`
Cell cluster annotation	• Assessment of cluster enrichment for quantitative and categorical features	• `assess_cluster_enrichment()`
Comparison of datasets	• Assessment of cluster markers or gene set expression across multiple datasets	• `inter_dataset_comparison()`
Visualization	• Automated UMAP visualizations for multiple features; • Heatmaps: gene set scores, CNV, overlap between clusters; • boxplots and barplots of cluster enrichment	• `boxplot_cluster()` • `dotplot_cluster()` • `heatmap_CNV()` • `plot_profile_CNV()` • `plot_umap_colored_features()` • `plot_heatmap_features_by_clusters()`

3.1 Gene set scoring

scMuffin provides functions to set up one or more gene set collections and perform SC-level estimation of gene set expression scores in relation to an empirical null model (see Implementation section). This can be applied to any gene set and can therefore be used to estimate several different cell phenotypes, like pathway activities or marker set expression.

The function `prepare_gsls()` allows the user to collect gene sets of cell types, pathways, cancer functional states, as well as other collections of gene sets (e.g. positional gene sets, hallmarks) from CellMarker (12), PanglaoDB (13), CancerSEA (14) and MSigDB (8) databases.

Unlike many existing tools that are used to perform marker-based cell annotation (15), the availability of these gene sets within scMuffin package spares the user the effort of data collection and harmonization. The function, which also accepts any user-given gene sets, applies a series of criteria (e.g., minimum and maximum number of genes in a gene set) to filter the chosen gene sets.

The cell-level expression scores for these gene sets can be calculated using `calculate_gs_scores()`. For instance, the following code shows how to quantify the activity of "Cancer functional states" from CancerSEA (14) at cell and cell cluster level:

```
gsc <- prepare_gsls(gs_sources = "CancerSEA", scMuffinList =
scMuffinList)
scMuffinList <- calculate_gs_scores(scMuffinList = scMuffinList,
gs_list = gsc$CancerSEA)
scMuffinList <- calculate_gs_scores_in_clusters(scMuffinList =
scMuffinList, partition_id = "global_expr")
```

where "scMuffinList" is the data structure that contains expression data and cell clusters ("global_expr").

Cluster-level gene sets scores can be visualized as a heatmap to obtain a useful summary visualization that can provide insights for cluster annotation. This is accomplished by means of the function `plot_heatmap_features_by_clusters()`, which relies on the powerful ComplexHeatmap R package (16)). For example, the analysis of the CancerSEA functional states in the HGG sample PJ016 showed that the two groups of cell clusters that are spatially separated in the UMAP visualization, that is clusters {0, 6, 8, 9} and {1, 2, 3, 4, 5, 7} (**Figure 2a**), reflect distinct functional states **(Figure 2b)**. Cell-level gene set scores can be visualized over the UMAP by means of `plot_umap_colored_features()` **(Figure 2c)**.

3.2 CNV estimation and association with CancerSEA functional states

CNV inference from SC transcriptomics in cancer provides a means to assess the presence of relevant genomic aberrations (duplications and deletions) based on the expression of adjacent genes. This knowledge offers crucial clues to address the difficult task of distinguishing normal from malignant cells and provides quantitative information to reconstruct the tumor clonal substructure. Moreover, CNV patterns allow the investigator to hypothesize link between genomic alterations and cell phenotypes. As a proof-of-concept, we describe CNV inference in scMuffin with and without a reference expression profile, and the association of CNV clusters with CancerSEA functional states.

CNV inference is performed by means of the function `CNV_analysis()`:

```
scMuffinList <- CNV_analysis(scMuffinList, reference = GTEx_mean,
center_genes = TRUE)
```

where GTEx_mean contains a reference expression profile, which can be obtained by means of the function `process_GTEx_gene_reads()` starting from data available at The Genotype-Tissue Expression (GTEx) portal (17). The resulting scMuffinList contains the matrix with CNV values for each genomic regions in every cell, cell clusters by CNV, detected CNV regions and mapping information between gene locations and CNV regions.

Heatmaps that show CNV profiles of every cell, clustered by similarity, can be generated by means of `plot_heatmap_CNV()`. In the considered dataset (PJ030), we observed that the reference profile (normal brain samples from GTEx) is included into cluster 3, while clusters 0, 1 and 2 show large aneuploidies, some of which are typical of HGG, like the amplification of chromosome 7 as reported by dataset authors (10) **(Table 2, Figure 3a-b)**. The median CNV profile of every cell cluster can be visualized by means of `plot_profile_CNV()` **(Figure 3c)** while

boxplots of the distribution of summary CNV score by clusters can be visualized through `boxplot_points()` **(Figure 3d).**

Table 2. CNV regions detected in chromosomes 1 and 7 of sample PJ030.

Chr	Start (location)	Stop (location)	Cluster	Length
chr7	497259	132784870	0	132287611
chr7	497259	132784870	1	132287611
chr7	497259	132784870	2	132287611
chr19	416582	48872391	0	48455809
chr19	17933014	58551565	1	40618551
chr19	416582	49906825	2	49490243

The analysis of another dataset (PJ016, without reference profile) showed two groups of clusters ({0, 1, 4} and {2, 3}) that map to separated regions of the UMAP visualization **(Figure 4a-b).** Interestingly, clusters 2 and 3 are marked by peculiar amplifications in chromosome arms 1p and 19p. These "CNV clusters" correspond to expression clusters {0, 6, 8, 9} **(Figure 4c)**. The similarity of multiple cell partitions can be quantified using the function `overlap_matrix()`.

An example of integrative analysis enabled by scMuffin is the functional assessment of CNV clusters. We quantified the expression scores of CancerSEA functional states throughout the CNV clusters of sample PJ016. As expected, the two aforementioned groups of CNV clusters {0, 1, 4} and {2, 3} are characterized by different functional states **(Figure 4d).** This finding suggests that the peculiar amplifications of chromosomes 1 and 19 that characterize these clusters might underlie two phenotypes. A suggestive evidence of such hypothesis, as regards the CancerSEA functional state "Invasion", is the location of the two Invasion markers *YBX1* (Y-Box Binding Protein 1) and *HNRNPM* (Heterogeneous Nuclear Ribonucleoprotein M) within the amplified regions of chromosomes 1p and 19p **(Figure 4c)**. *YBX1* is a DNA/RNA-binding protein and transcription factor which plays a central role in coordinating tumor invasion in glioblastoma (18). *HNRNPM* belongs to a family of spliceosome auxiliary factors and is involved in the regulation of splicing; the upregulation of these factors results in tumor-associated aberrant splicing, which

promotes glioma progression and malignancy (19,20). In particular, HNRNPM was identified as an interactor of the DNA/RNA binding protein SON, which drives oncogenic RNA splicing in glioblastoma (21). While it is beyond the scope of this article to validate this hypothesis, these findings clearly highlight the usefulness of the integrative analysis of CNVs and CancerSEA functional states provided by scMuffin.

3.3 Cluster enrichment analysis: association of clusters and features

scMuffin contains functions for assessing the association between cell clusters and quantitative as well as categorical features, by means of CSEA and ORA, respectively. The function `assess_cluster_enrichment()` performs CSEA and ORA accordingly to the feature type under consideration. For example, the following instruction runs CSEA over the gene set scores, considering the partition "global_expr":

```
scMuffinList <- assess_cluster_enrichment(scMuffinList,
feature_name = "gene_set_scoring", partition_id = "global_expr")
```

The results of CSEA and ORA include several quantities, such as the normalized enrichment score (NES), the enrichment ratio (ER) and the false discovery rate (FDR) **(Tables 3-4)**. These results can be visualized as boxplots (quantitative features) and barplots (categorical features). For instance, the cluster 0 of sample PJ016 is particularly enriched – as expected – in "Invasion" markers **(Table 3, Figure 5a)** and cells assigned to G2/M or S phases **(Table 4, Figure 5b).** The function `extract_cluster_enrichment_table()` provides a means to obtain tables of enrichment results across clusters, which can be visualized as heatmaps **(Figure 5c)**, while `extract_cluster_enrichment_tags()` defines the most significant features as cluster labels that are therefore available for annotating UMAPs **(Figure 5d)**.

Table 3. The output of CSEA analysis for a quantitative feature. This result is relative to the CancerSEA Invasion gene set on dataset PJ016. ES = Enrichment Score; p = CSEA p-value; p (BH) = p value after Bonferroni-Hochberg correction; NES = Normalized Enrichment Score; q (FDR) = q value of FDR.

Cluster	ES	p	p (BH)	NES	q (FDR)
0	0.67	0.01	0.02	2.38	0.01
1	0.22	0.99	0.99	0.80	0.97
2	-0.21	0.01	0.02	0.00	1.00
3	-0.15	0.01	0.02	0.00	1.00
4	-0.11	0.01	0.02	0.00	1.00
5	0.25	0.80	0.89	0.92	0.90
6	0.29	0.31	0.39	1.05	0.41
7	0.33	0.11	0.18	1.14	0.20
8	0.32	0.13	0.19	1.13	0.18
9	0.71	0.01	0.02	2.43	0.01

Table 4. The output of an ORA analysis for a categorical feature. This result is relative to the cell cycle phase "G2M" on dataset PJ016. N=number of cells; wb = number G2M cells; bb = N-wb; wbd = number of G2M cells in the cluster; exp = expected number of G2M cells in the cluster; ER = enrichment ratio; p = hypergeometric p; p (BH) = p value after Bonferroni-Hochberg correction.

Cluster	N	wb	bb	bd	wbd	exp	ER	p	p (BH)
0	2828	380	2448	499	239	67.05	3.56	1.67×10^{-105}	8.34×10^{-105}
1	2828	380	2448	475	4	63.83	0.06	1.00	1.00
2	2828	380	2448	378	4	50.79	0.08	1.00	1.00
3	2828	380	2448	309	2	41.52	0.05	1.00	1.00
4	2828	380	2448	296	7	39.77	0.18	1.00	1.00
5	2828	380	2448	257	0	34.53	0.00	1.00	1.00
6	2828	380	2448	185	4	24.86	0.16	1.00	1.00
7	2828	380	2448	172	0	23.11	0.00	1.00	1.00
8	2828	380	2448	142	5	19.08	0.26	1.00	1.00
9	2828	380	2448	115	115	15.45	7.44	2.47×10^{-108}	2.47×10^{-107}

3.4 Transcriptional complexity, proliferation rate and cell state trajectories

The number of expressed genes and proliferation rate are two relevant pieces of information for the characterization of cell type and cell state in cancer.

Undifferentiated stem or progenitor cells show a high number of expressed genes (open chromatin state) compared to differentiated cell types. In cancer context, a high number of expressed genes might suggest tumor initiating cells (TICs), might indicate de-differentiation processes of tumor progression (22,23). For instance, in a recent study on glioblastoma, chromatin accessibility was associated to a specific subpopulation of putative tumor-initiating Cancer Stem

Cells (CSCs) with invasive phenotype and low survival prediction (24). The transcriptional complexity of a cell can be calculated as follows:

```
scMuffinList <- transcr_compl(scMuffinList, min_counts=3, min_cells=10, min_genes=500)
```

where min_counts is the threshold above which a gene is considered expressed, min_cells is the minimum number of cells in which a gene must be expressed and min_genes is the minimum number of genes that a cell must express. In SC datasets, transcriptional complexity quantification is complicated by the presence of a linear relationship between number of expressed genes and total transcripts detected, where biological variability is entangled with sequencing deepness. scMuffin provides three measures of transcriptional complexity, which show three different – but complementary – relations with the number of total transcripts (*t*): the ratio (TC-ratio) prioritizes cells with a relatively low *t* **(Figure 6a)**; the linear model residuals (TC-LMR) are independent from *t* **(Figure 6b)**; lastly, the transcriptional entropy **(TC-H)** (7) highlights genes with relatively high number of *t* **(Figure 6c)**. All three measures may provide useful insights for the identification of particular cell types or states. For example, TC-ratio could be used when one is not interested in cell with a relatively abundant RNA production, like can be the case of quiescent cells with stem-like properties.

The proliferation rate is a relevant indicator for distinguishing cell types in solid tumors and helps to identify cells with potential clinical relevance and interest as candidate therapeutic targets (25,26). Cell proliferation rate is quantified based on the expression of G1/S and G2/M genes:

```
scMuffinList <- proliferation_analysis(scMuffinList).
```

As a proof-of-concept we jointly analysed cell TC and cell proliferation of sample PJ016 and visualized the results on a diffusion map (27) – which can be run by means of the wrapper function

`diff_map()`. This joint analysis shows that cells with the highest values of TC-ratio tend to have lower proliferation scores **(Figure 6d).** These cells are located more closely to the branching point of the diffusion map **(Figure 6e)**, while the cells with the highest proliferation rates are located at a corner of one of the branches **(Figure 6f).** This evidence suggests that the cells with high TC and low proliferation are interesting candidates for further analyses aimed at finding TICs in this dataset of HGG. Interestingly, this particular pattern is captured by TC-ratio, but not by TC-LMR and TC-H **(Supplementary Figure S1, Additional File 1).**

3.5 Comparison of datasets

A SC dataset carries an extensive amount of information. The integration of multiple SC datasets is a challenging task and multiple approaches have been proposed to address it (28). Typically, the integrated datasets are computationally demanding due to their huge size. An alternative possibility lies in cross-checking the expression of cluster markers between datasets: the expression of the cluster markers of every dataset is assessed in all considered datasets. For example, Nguyen *et al.* (5) used this approach to study the occurrence of the characteristic cell types of normal mammary gland across samples collected from different subjects.

scMuffin provides a function to quantify the expression of a series of gene sets across multiple datasets. If these gene sets are defined as cluster markers ("cluster_markers_list" in the code below), then the analysis sheds light over the possible presence of cells with similar expression profiles across datasets:

```
res <- inter_dataset_comparison(seu_obj_list = seu_obj_list, gsl =
cluster_markers_list)
```

We performed this analysis on three datasets (PJ016, PJ017, PJ030) and found cell clusters that are more similar to clusters of other datasets than clusters of the same dataset, suggesting the presence of similar cells **(Figure 8a),** like cluster 7 of PJ030 ("PJ030_C7"), which is close to clusters

352 {0, 6, 8, 9} of PJ016**.** Importantly, the same analysis can be done using gene sets with a functional
353 meaning, to obtain a functional characterization of similar cell types across datasets **(Figure 8b)**.

354 **3.6 scMuffin and scCancer**

355 scCancer (29) is an R package for automated processing of SC expression data in cancer. The
356 scMUffin and scCancer address the same challenge and are complementary. Indeed, each of them
357 offers some analyses that the other does not **(Supplementary Table S1, Additional File 1)**.
358 Examples of this are cluster association analysis (CSEA and ORA), comparison of multiple cell
359 partitions and inter-dataset gene set expression assessment, available in scMuffin, and cell
360 interaction, expression programs and cancer micro-environment cell classification, provided by
361 scCancer. Other types of analyses are common to both packages but differ by the underlying
362 algorithm and/or implementation. This is the case, for instance, of gene set scoring and CNV
363 analysis.

364 4. Conclusions

365 Here, we presented scMuffin, an R package that provides a series of functions to perform and
366 integrate various types of analyses on SC expression data. As a proof-of-concept, to illustrate the
367 user interface and its potentialities, we analysed publicly available SC expression datasets of
368 human HGG. We described two examples of integrative analyses which returned particularly
369 interesting findings that would deserve further investigations: the functional characterization of
370 CNVs highlighted a possible link between amplifications of chromosomes 1p and 19p and invasive
371 tumor phenotype; the joint analysis of transcriptional complexity, proliferation rate and cell state
372 trajectories spotted cells that have some characteristics of TICs. While the inference of CNV from
373 SC expression data is a challenging task, subject to various issues (e.g., missing data, number of
374 aberrancies) still it provides insights useful to clarify cell identity, like the distinction between a

more aberrant cluster and the others or the distinction of tumoral and normal cells. The function of scMuffin can be combined conveniently to address various tasks. For example, different partitions of the same dataset obtained using the same clustering method with different parameters can be characterized by means of their overlap and gene set expression, to gain insights on the appropriate number of clusters. The analyses offered by scMuffin and the results achieved in this case study show that scMuffin helps addressing the main challenges in the bioinformatics analysis of SC datasets from solid tumors.

5. Figure captions

Figure 1. Overview of scMuffin package. scMuffin offers the possibility to perform several different analyses and data integration approaches to address the main challenges of SC gene expression analysis in solid tumors.

Figure 2. Quantification of CancerSEA functional states in the HGG sample PJ016. a) UMAP visualization where cells are coloured by expression clusters. **b)** Cluster-level expression scores of all the CancerSEA functional states. **c)** UMAP visualization where cells are colored by "CSEA_Invasion" gene set score.

Figure 3. CNV analysis (PJ030). a) CNV heatmap where cells (columns) are grouped into "CNV clusters"; the cluster of cells that include the reference profile is highlighted through a red label. **b)** UMAP visualizations where cells are colored by CNV clusters. **c)** Median CNV value of a cluster 0. **d)** Distribution of the summary CNV score in clusters.

Figure 4. CNV analysis and CancerSEA functional states (PJ016). a) CNV heatmap where cells (columns) are grouped into "CNV clusters". **b)** UMAP visualizations where cells are colored by CNV clusters. **c)** Overlap indices between expression clusters (prefix "EXPR_") and CNV clusters (prefix "CNV_"). **d)** Heatmap of CancerSEA functional state expression score in CNV clusters.

Figure 5. Cluster enrichment analysis (PJ016). a) Output of `boxplot_cluster()` function for a quantitative feature – CancerSEA functional states – in cluster 0: distribution of gene set scores values (red) and corresponfing null models (gray, suffix "_ref") (top); Normalized Enrichment Score (NES) and FDR *q*-value (bottom). **b)** Output of `barplot_cluster()` function for a categorical feature – cell cycle phase – in cluster 0: expected ("exp") and observed ("obs") occurrences of every possible category (top); enrichment ratio (ER) and hypergeometric *p*-value (bottom). **c)** Heatmap of NES values for CancerSEA functional states in expression clusters; asterisks indicate FDR q-value < 0.01. **d)** UMAP visualization where clusters are labelled with the top three most significant CancerSEA functional states.

Figure 6. Transcriptional complexity, proliferation rate and cell state trajectories (PJ016). a). Transcriptional Complexity Ratio (TC-R) of cells. **b)** Transcriptional Complexity Linear Model Residuals (TC-LMR) of cells. **c)** Transcriptional Complexity Entropy (TC-H) of cells**. d)** Distribution of TC-R values split by cell proliferation score. **e-f)** Distribution of cells according to the first two "diffusion components" (DC).

Figure 7. Inter-dataset comparison of gene set expression. a) Heatmap of cluster marker (columns, "_Mx" suffix, where "x" indicates cluster number) expression scores across clusters (rows, "_Cx" suffix, where "x" indicates cluster number) of datasets PJ016, PJ017 and PJ030. **b)** Heatmap of CancerSEA functional states scores across clusters of datasets as in panel (a).

6. Availability and requirements

Project name: scMuffin

Project home page: https://github.com/emosca-cnr/scMuffin

Operating system: Platform independent

Programming language: R (>= 4.0.0)

Other requirements: The R Project for Statistical Computing.

License: GPL-3

Any restrictions to use by non-academics: According to GPL-3

7. List of abbreviations

CNV: Copy Number Variation

CSC: cancer stem cells

CSEA: Cell Set Enrichment Analysis

FDR: False discovery rate

ER: Enrichment Ratio

GSEA: Gene Set Enrichment Analysis

GTEx: The Genotype-Tissue Expression project

HGG: high grade glioma

NES: normalized enrichment score

ORA: over representation analysis

SC: single cell

TC-R: Transcriptional Complexity Ratio

TC-LMR: Transcriptional Complexity Linear Model Residuals

TC-H: Transcriptional Complexity Entropy

TICs: Tumor Initiating Cells

UMAP: Uniform Manifold Approximation and Projection

8. Declarations

Ethics approval and consent to participate. Not applicable.

Consent for publication. Not applicable.

Availability of data and materials. The data used for the analyses described in this manuscript were obtained from: The Gene Expression Omnibus (11), under the accession GSE103224; the GTEx Portal (17) on 04/08/2020.

Competing interests. The authors declare that they have no competing interests.

Funding. The work was supported by: the Italian Ministry of Education, University and Research (MIUR) [INTEROMICS PB05].

Author's contributions. VN implemented the CNV functions, carried out the analyses, interpreted the results and wrote the article. AC implemented clustering functions, carried out the analyses, wrote the article, tested the package. NDN drafted the package, carried out the analyses and wrote the article. IC curated the biological aspects of CNV analysis and revised the article. MM and MG set up the computational infrastructure for data analysis. CC and EP curated the biological aspects of solid tumor data analysis. AS tested the package. RR, IZ, LM and AM contributed to the design of the study. PP contributed to the software design, case study definition, interpretation of the results and wrote the article. EM designed the study, implemented the software, performed the analysis, interpreted the results, and wrote the article. All authors read and approved the final manuscript.

9. References

1. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nat Rev Cancer [Internet]. 2017 Sep 24;17(9):557–69. Available from: https://www.nature.com/articles/nrc.2017.58
2. Zappia L, Theis FJ. Over 1000 tools reveal trends in the single-cell RNA-seq analysis landscape. Genome Biol. 2021 Dec;22(1):301.
3. Hao Y, Hao S, Andersen-Nissen E, Mauck WM, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell [Internet]. 2021 Jun;184(13):3573-3587.e29. Available from: https://linkinghub.elsevier.com/retrieve/pii/S0092867421005833
4. Tirosh I, Izar B, Prakadan SM, Wadsworth MH, Treacy D, Trombetta JJ, et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. Science (1979)

[Internet]. 2016 Apr 8;352(6282):189–96. Available from: https://www.science.org/doi/10.1126/science.aad0501

5. Nguyen QH, Pervolarakis N, Blake K, Ma D, Davis RT, James N, et al. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. Nat Commun [Internet]. 2018 Dec 23;9(1):2028. Available from: http://www.nature.com/articles/s41467-018-04334-1

6. Patel AP, Tirosh I, Trombetta JJ, Shalek AK, Gillespie SM, Wakimoto H, et al. Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. Science (1979) [Internet]. 2014 Jun 20;344(6190):1396–401. Available from: https://www.science.org/doi/10.1126/science.1254257

7. Gandrillon O, Gaillard M, Espinasse T, Garnier NB, Dussiau C, Kosmider O, et al. Entropy as a measure of variability and stemness in single-cell transcriptomics. Curr Opin Syst Biol. 2021;27.

8. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences [Internet]. 2005 Oct 25;102(43):15545–50. Available from: https://pnas.org/doi/full/10.1073/pnas.0506580102

9. Khatri P, Sirota M, Butte AJ. Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. Ouzounis CA, editor. PLoS Comput Biol [Internet]. 2012 Feb 23;8(2):e1002375. Available from: https://dx.plos.org/10.1371/journal.pcbi.1002375

10. Yuan J, Levitin HM, Frattini V, Bush EC, Boyett DM, Samanamud J, et al. Single-cell transcriptome analysis of lineage diversity in high-grade glioma. Genome Med [Internet]. 2018 Dec 24;10(1):57. Available from: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-018-0567-9

11. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, et al. NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res [Internet]. 2012 Nov 26;41(D1):D991–5. Available from: http://academic.oup.com/nar/article/41/D1/D991/1067995/NCBI-GEO-archive-for-functional-genomics-data

12. Zhang X, Lan Y, Xu J, Quan F, Zhao E, Deng C, et al. CellMarker: a manually curated resource of cell markers in human and mouse. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D721–8. Available from: https://academic.oup.com/nar/article/47/D1/D721/5115823

13. Franzén O, Gan LM, Björkegren JLM. PanglaoDB: a web server for exploration of mouse and human single-cell RNA sequencing data. Database [Internet]. 2019 Jan 1;2019. Available from: https://academic.oup.com/database/article/doi/10.1093/database/baz046/5427041

14. Yuan H, Yan M, Zhang G, Liu W, Deng C, Liao G, et al. CancerSEA: a cancer single-cell state atlas. Nucleic Acids Res [Internet]. 2019 Jan 8;47(D1):D900–8. Available from: https://academic.oup.com/nar/article/47/D1/D900/5133662

15. Clarke ZA, Andrews TS, Atif J, Pouyabahar D, Innes BT, MacParland SA, et al. Tutorial: guidelines for annotating single-cell transcriptomic maps using automated and manual methods. Nat Protoc [Internet]. 2021 Jun 24;16(6):2749–64. Available from: http://www.nature.com/articles/s41596-021-00534-0

16. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. Bioinformatics [Internet]. 2016 Sep 15;32(18):2847–9. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btw313

17. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Available from: https://www.gtexportal.org/home/index.html

18. Gupta MK, Polisetty RV, Sharma R, Ganesh RA, Gowda H, Purohit AK, et al. Altered transcriptional regulatory proteins in glioblastoma and YBX1 as a potential regulator of tumor invasion. Sci Rep [Internet]. 2019 Dec 29;9(1):10986. Available from: http://www.nature.com/articles/s41598-019-47360-9
19. LeFave C V, Squatrito M, Vorlova S, Rocco GL, Brennan CW, Holland EC, et al. Splicing factor hnRNPH drives an oncogenic splicing switch in gliomas. EMBO J [Internet]. 2011 Oct 5;30(19):4084–97. Available from: http://emboj.embopress.org/cgi/doi/10.1038/emboj.2011.259
20. Golan-Gerstl R, Cohen M, Shilo A, Suh SS, Bakàcs A, Coppola L, et al. Splicing Factor hnRNP A2/B1 Regulates Tumor Suppressor Gene Splicing and Is an Oncogenic Driver in Glioblastoma. Cancer Res [Internet]. 2011 Jul 1;71(13):4464–72. Available from: http://cancerres.aacrjournals.org/lookup/doi/10.1158/0008-5472.CAN-10-4410
21. Kim JH, Jeong K, Li J, Murphy JM, Vukadin L, Stone JK, et al. SON drives oncogenic RNA splicing in glioblastoma by regulating PTBP1/PTBP2 switching and RBFOX2 activity. Nat Commun [Internet]. 2021 Dec 21;12(1):5551. Available from: https://www.nature.com/articles/s41467-021-25892-x
22. Wainwright EN, Scaffidi P. Epigenetics and Cancer Stem Cells: Unleashing, Hijacking, and Restricting Cellular Plasticity. Trends Cancer [Internet]. 2017 May;3(5):372–86. Available from: https://linkinghub.elsevier.com/retrieve/pii/S240580331730081X
23. Gaspar-Maia A, Alajem A, Meshorer E, Ramalho-Santos M. Open chromatin in pluripotency and reprogramming. Nat Rev Mol Cell Biol [Internet]. 2011 Jan 22;12(1):36–47. Available from: http://www.nature.com/articles/nrm3036
24. Guilhamon P, Chesnelong C, Kushida MM, Nikolic A, Singhal D, MacLeod G, et al. Single-cell chromatin accessibility profiling of glioblastoma identifies an invasive cancer stem cell population associated with lower survival. Elife [Internet]. 2021 Jan 11;10. Available from: https://elifesciences.org/articles/64090
25. Feitelson MA, Arzumanyan A, Kulathinal RJ, Blain SW, Holcombe RF, Mahajna J, et al. Sustained proliferation in cancer: Mechanisms and novel therapeutic targets. Semin Cancer Biol [Internet]. 2015 Dec;35:S25–54. Available from: https://linkinghub.elsevier.com/retrieve/pii/S1044579X15000140
26. Al-Hajj M, Clarke MF. Self-renewal and solid tumor stem cells. Oncogene [Internet]. 2004 Sep 20;23(43):7274–82. Available from: https://www.nature.com/articles/1207947
27. Angerer P, Haghverdi L, Büttner M, Theis FJ, Marr C, Buettner F. destiny□: diffusion maps for large-scale single-cell data in R. Bioinformatics [Internet]. 2016 Apr 15;32(8):1241–3. Available from: https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv715
28. Forcato M, Romano O, Bicciato S. Computational methods for the integrative analysis of single-cell data. Brief Bioinform [Internet]. 2021 May 20;22(3). Available from: https://academic.oup.com/bib/article/doi/10.1093/bib/bbaa042/5828125
29. Guo W, Wang D, Wang S, Shan Y, Liu C, Gu J. scCancer: a package for automated processing of single-cell RNA-seq data in cancer. Brief Bioinform. 2021 May 20;22(3).

10. Additional file

Additional file 1

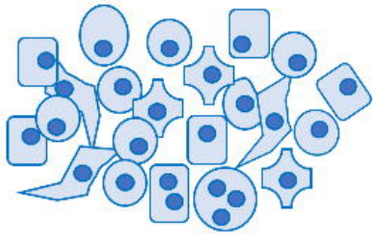
Format: PDF

Title: Supplementary material

Description: supplementary methods, table and figure

Challenges in solid tumor analysis at single-cell level

- Limited availability of markers for definition of cell subtype identity
- Potentially strongly altered and highly heterogeneous gene expression profiles
- Presence of infiltrating cells and cells from the surrounding (healthy) tissue
- Potentially clinically relevant cell subtypes at very low number (e.g., drug resistant subclones)
- Often limited number of detected genes



scMuffin

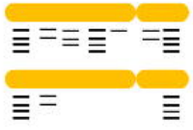
Single-cell multi-feature integrative analysis

Gene set scoring

- Against empirical null
- Optimized to handle missing data

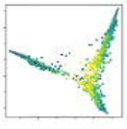
Transcriptional Complexity

- Ratio
- Linear Model Residuals
- Entropy



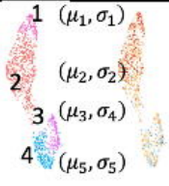
Cell state trajectory

- Diffusion map



Cell cluster statistics

- mean values
- variability

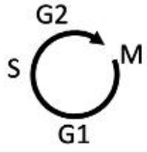


Gene sets from various sources

- General: MSigDB
- Cell markers: CellMarker, PanglaoDB
- Functional states: CancerSEA
- Proliferation: G1/S & G2/M
- Ribosomal proteins

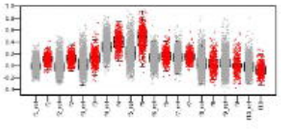
Proliferation rate

- Proportional to G1/S and G2/M genes



Cluster enrichment analysis

- Quantitative features: CSEA
- Categorical features: ORA

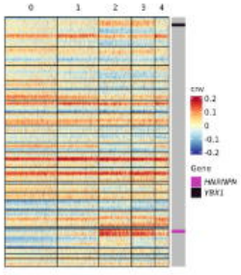


Implementation

- Computationally intensive tasks are parallelized
- Integration of various results in dedicated objects to enable automated subsequent analyses
- Parametrization of analyses to address dataset-specific characteristics

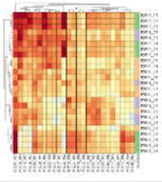
CNV inference

- Adjacent gene windows
- Reference profiles (optional)
- CNV clusters
- CNV regions detection



Gene set-based inter-dataset comparison

- Cluster markers or gene sets expression across datasets

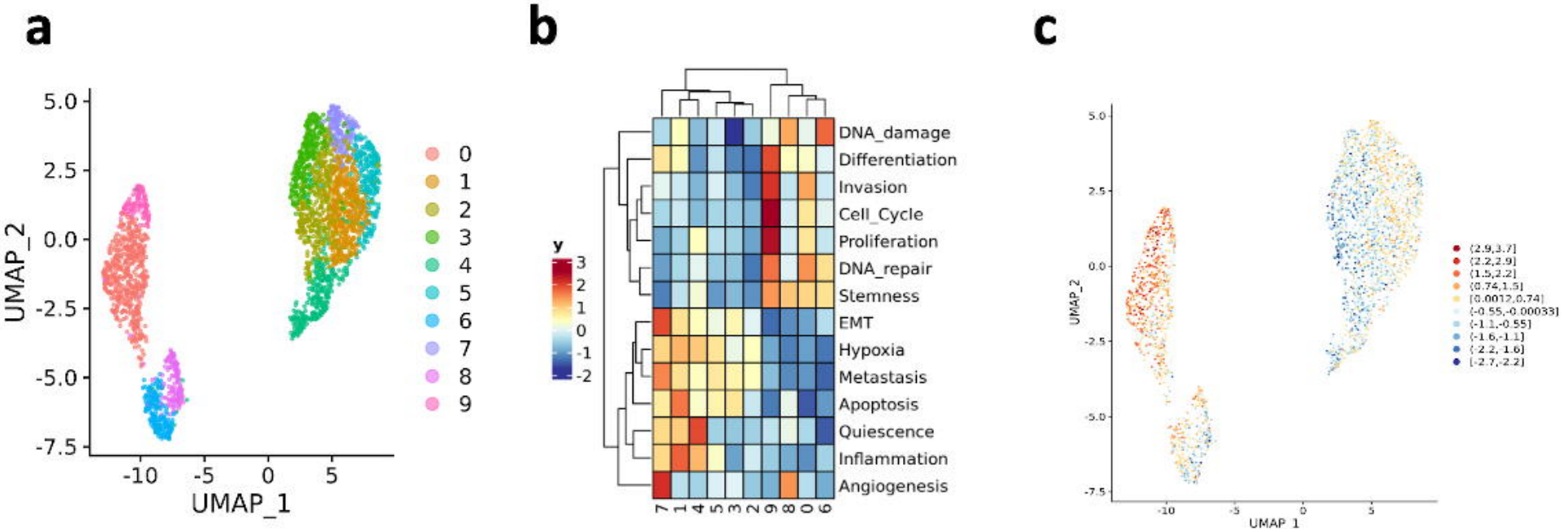


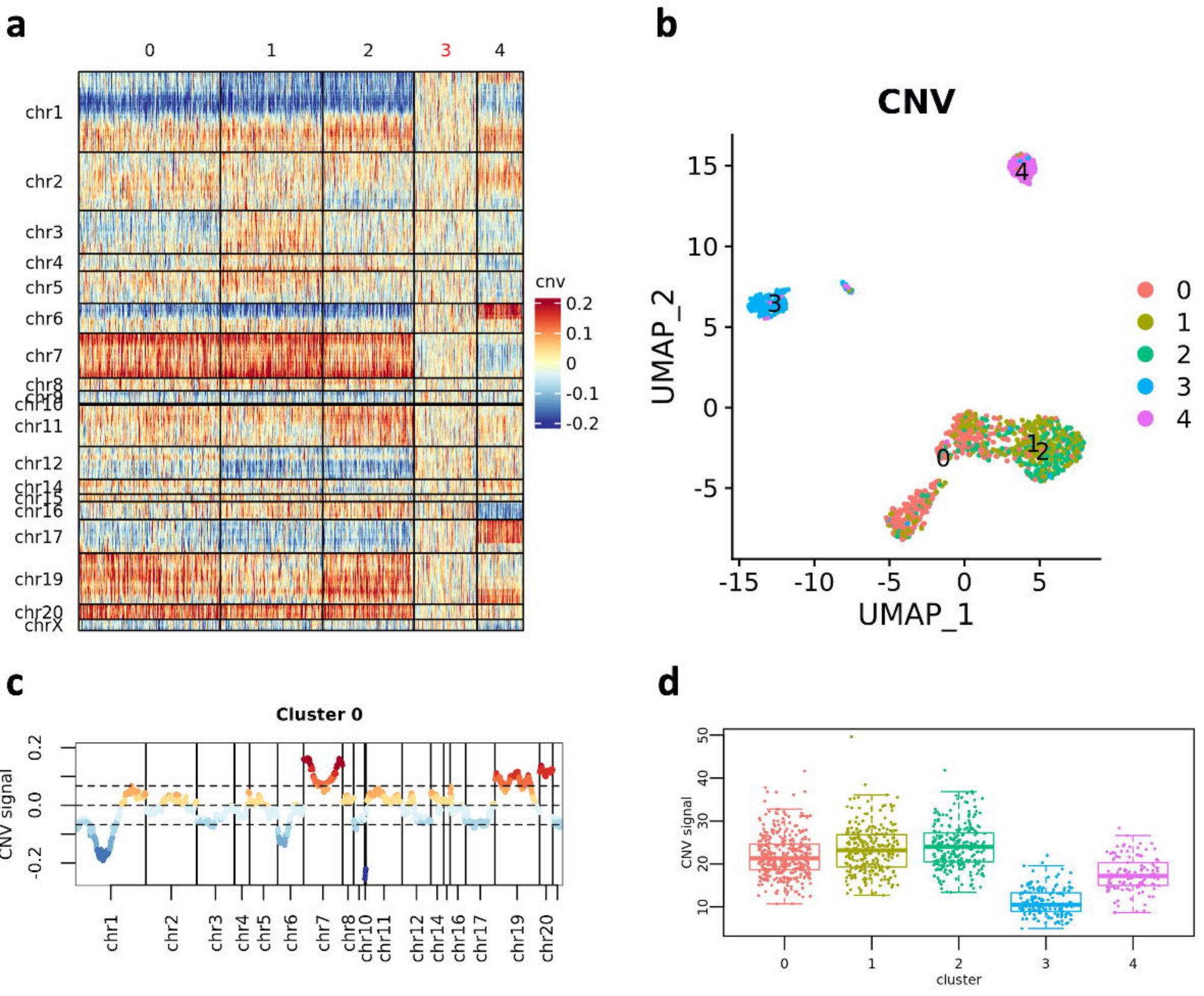
Comparison of multiple partitions

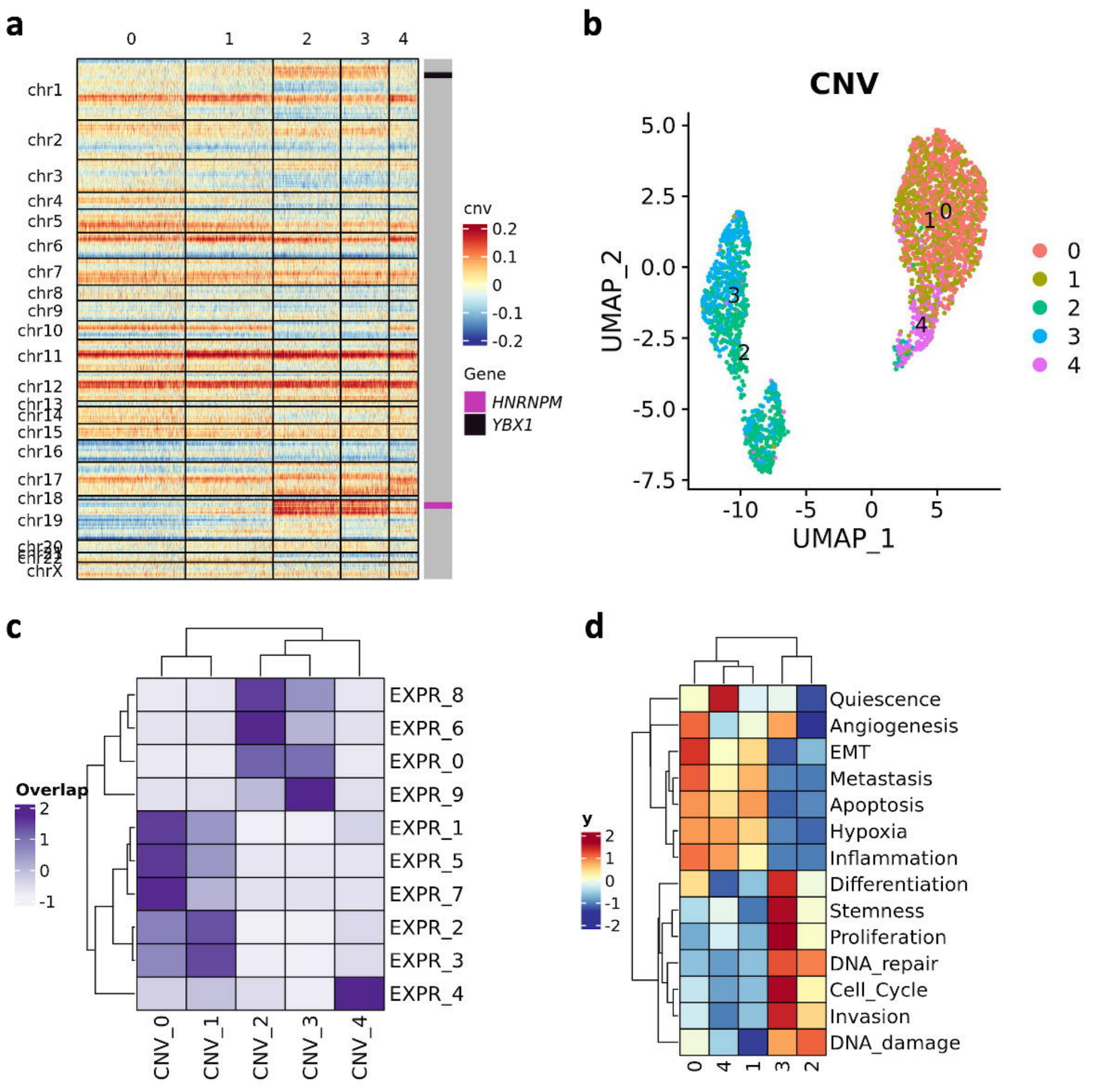
- Overlap matrix

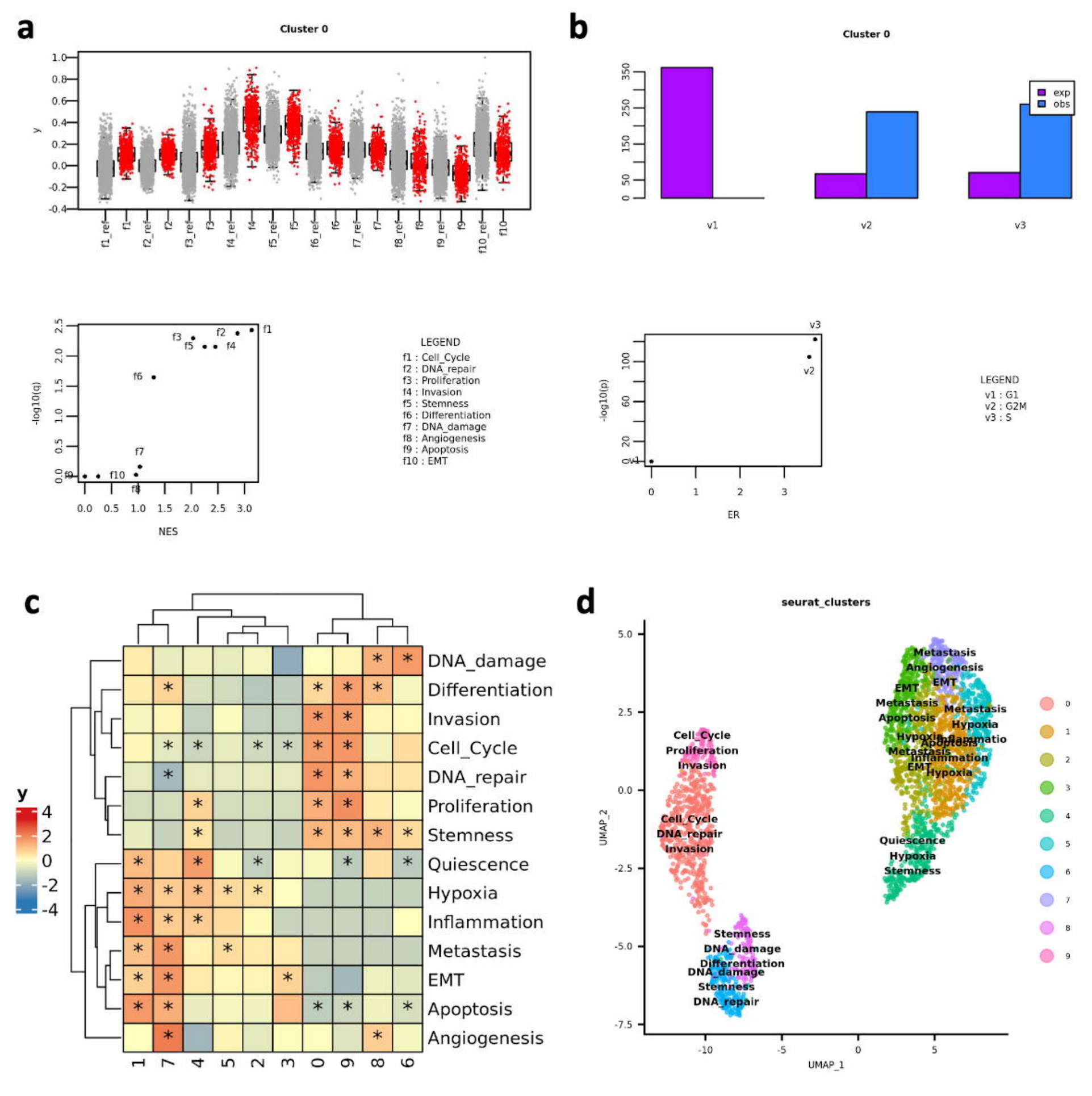
Visualization

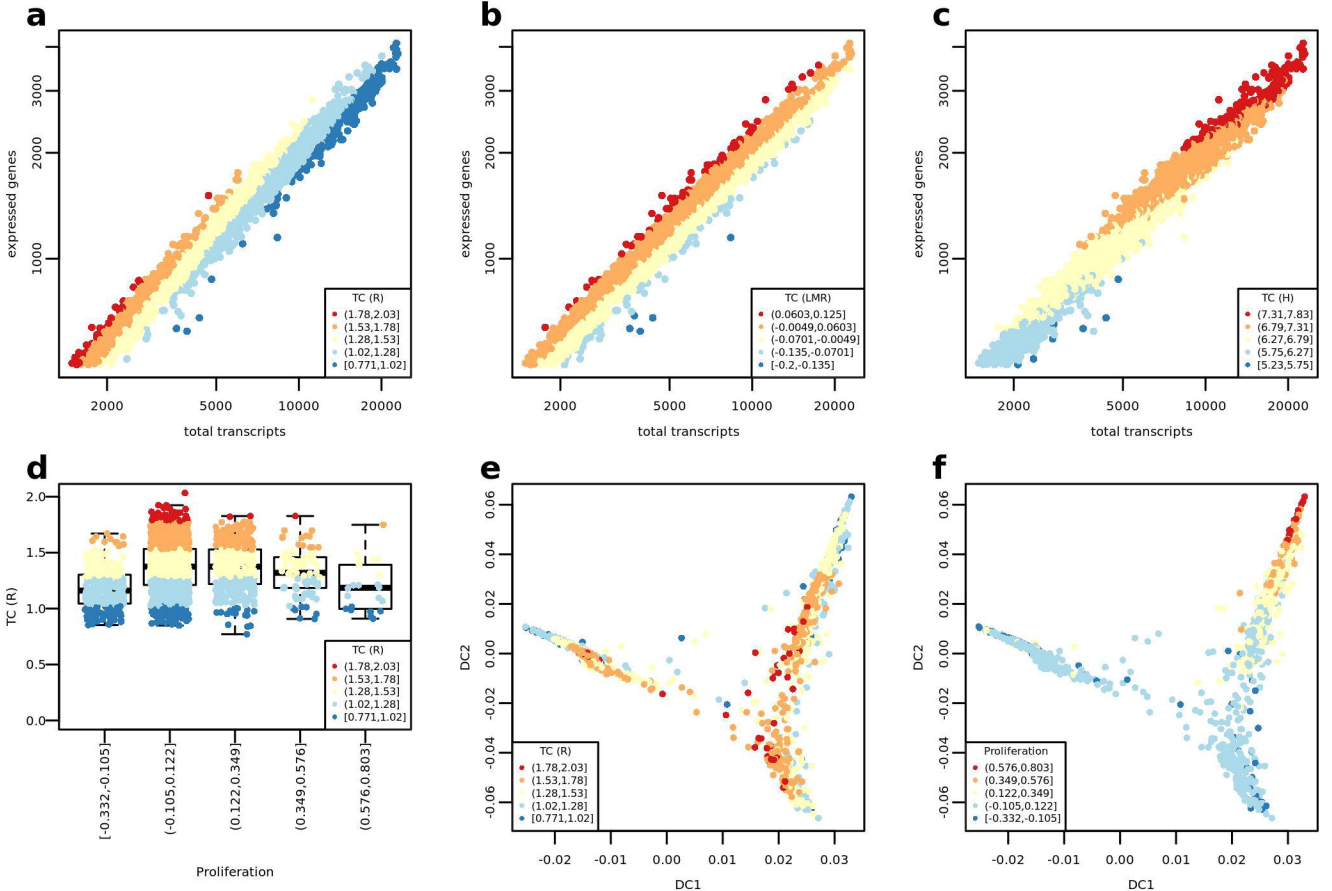
- UMAP visualization: automated for multiple quantitative and categorical features
- Heatmaps: clusters-by-cells, CNV, gene set expression across datasets
- Boxplots and barplots: cluster enrichment in quantitative or categorical values

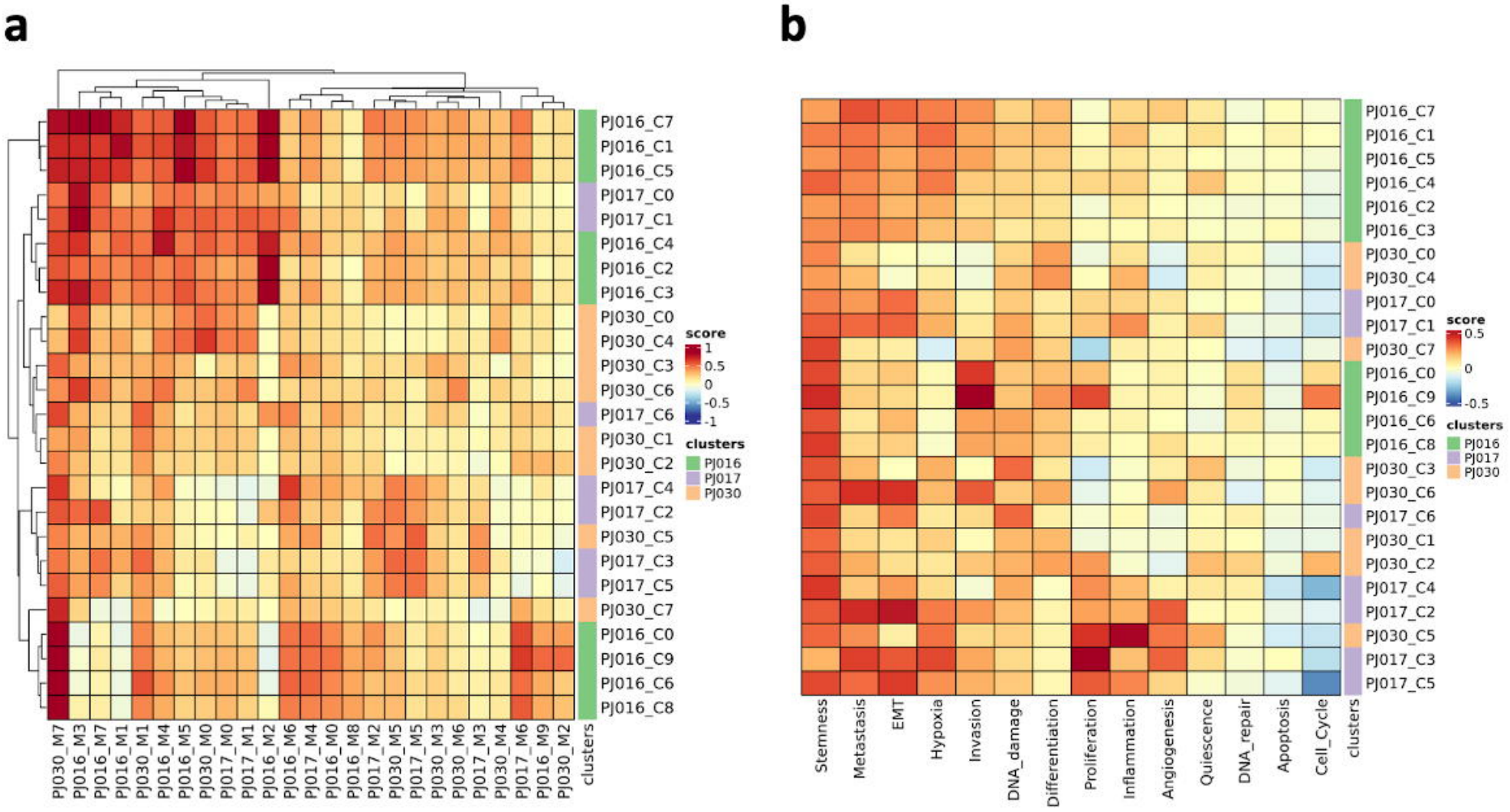

a
UMAP_2
UMAP_1
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10
-5
0
5
0
1
2
3
4
5
6
7
8
9
b
y
3
2
1
0
-1
-2
DNA_damage
Differentiation
Invasion
Cell_Cycle
Proliferation
DNA_repair
Stemness
EMT
Hypoxia
Metastasis
Apoptosis
Quiescence
Inflammation
Angiogenesis
7 1 4 5 3 2 9 8 0 6
c
UMAP_2
UMAP_1
(2.9,3.7]
(2.2,2.9]
(1.5,2.2]
(0.74,1.5]
(0.0012,0.74]
(-0.55,-0.00033]
(-1.1,-0.55]
(-1.6,-1.1]
(-2.2, 1.6]
(-2.7,-2.2]


a
0
1
2
3
4
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr14
chr15
chr16
chr17
chr19
chr20
chrX
cnv
0.2
0.1
0
-0.1
-0.2
b
CNV
UMAP_2
UMAP_1
15
10
5
0
-5
-15
-10
-5
0
5
0
1
2
3
4
c
Cluster 0
CNV signal
0.2
0.0
-0.2
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr10
chr11
chr12
chr14
chr16
chr17
chr19
chr20
d
CNV signal
50
40
30
20
10
0
1
2
3
4
cluster


a
0
1
2
3
4
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
chrX
cnv
0.2
0.1
0
-0.1
-0.2
Gene
HNRNPM
YBX1
b
CNV
UMAP_2
UMAP_1
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10
-5
0
5
0
1
2
3
4
c
Overlap
2
1
0
-1
EXPR_8
EXPR_6
EXPR_0
EXPR_9
EXPR_1
EXPR_5
EXPR_7
EXPR_2
EXPR_3
EXPR_4
CNV_0
CNV_1
CNV_2
CNV_3
CNV_4
d
y
2
1
0
-1
-2
Quiescence
Angiogenesis
EMT
Metastasis
Apoptosis
Hypoxia
Inflammation
Differentiation
Stemness
Proliferation
DNA_repair
Cell_Cycle
Invasion
DNA_damage
0
4
1
3
2


a
Cluster 0
y
f1_ref
f1
f2_ref
f2
f3_ref
f3
f4_ref
f4
f5_ref
f5
f6_ref
f6
f7_ref
f7
f8_ref
f8
f9_ref
f9
f10_ref
f10
-log10(q)
NES
LEGEND
f1 : Cell_Cycle
f2 : DNA_repair
f3 : Proliferation
f4 : Invasion
f5 : Stemness
f6 : Differentiation
f7 : DNA_damage
f8 : Angiogenesis
f9 : Apoptosis
f10 : EMT
b
Cluster 0
exp
obs
v1
v2
v3
-log10(p)
ER
LEGEND
v1 : G1
v2 : G2M
v3 : S
c
y
DNA_damage
Differentiation
Invasion
Cell_Cycle
DNA_repair
Proliferation
Stemness
Quiescence
Hypoxia
Inflammation
Metastasis
EMT
Apoptosis
Angiogenesis
1
7
4
5
2
3
0
9
8
6
d
seurat_clusters
UMAP_2
UMAP_1
Cell_Cycle
Proliferation
Invasion
Cell_Cycle
DNA_repair
Invasion
Stemness
DNA_damage
Differentiation
DNA_damage
Stemness
DNA_repair
Metastasis
Angiogenesis
EMT
EMT
Metastasis
Metastasis
Apoptosis
Hypoxia
Inflammation
EMT
Hypoxia
Quiescence
Hypoxia
Stemness


a
expressed genes
total transcripts
TC (R)
(1.78,2.03]
(1.53,1.78]
(1.28,1.53]
(1.02,1.28]
[0.771,1.02]
b
expressed genes
total transcripts
TC (LMR)
(0.0603,0.125]
(-0.0049,0.0603]
(-0.0701,-0.0049]
(-0.135,-0.0701]
[-0.2,-0.135]
c
expressed genes
total transcripts
TC (H)
(7.31,7.83]
(6.79,7.31]
(6.27,6.79]
(5.75,6.27]
[5.23,5.75]
d
TC (R)
Proliferation
[-0.332,-0.105)
(-0.105,0.122]
(0.122,0.349]
(0.349,0.576]
(0.576,0.803]
e
DC1
DC2
f
DC1
DC2
Proliferation
(0.576,0.803]
(0.349,0.576]
(0.122,0.349]
(-0.105,0.122]
[-0.332,-0.105]


a
PJ016_C7
PJ016_C1
PJ016_C5
PJ017_C0
PJ017_C1
PJ016_C4
PJ016_C2
PJ016_C3
PJ030_C0
PJ030_C4
PJ030_C3
PJ030_C6
PJ017_C6
PJ030_C1
PJ030_C2
PJ017_C4
PJ017_C2
PJ030_C5
PJ017_C3
PJ017_C5
PJ030_C7
PJ016_C0
PJ016_C9
PJ016_C6
PJ016_C8
PJ030_M7
PJ016_M3
PJ016_M7
PJ016_M1
PJ030_M1
PJ016_M4
PJ016_M5
PJ030_M0
PJ017_M0
PJ017_M1
PJ016_M2
PJ016_M6
PJ017_M4
PJ016_M0
PJ016_M8
PJ017_M2
PJ030_M5
PJ017_M5
PJ030_M3
PJ030_M6
PJ017_M3
PJ030_M4
PJ017_M6
PJ016_M9
PJ030_M2
clusters
score
1
0.5
0
-0.5
-1
clusters
PJ016
PJ017
PJ030
b
PJ016_C7
PJ016_C1
PJ016_C5
PJ016_C4
PJ016_C2
PJ016_C3
PJ030_C0
PJ030_C4
PJ017_C0
PJ017_C1
PJ030_C7
PJ016_C0
PJ016_C9
PJ016_C6
PJ016_C8
PJ030_C3
PJ030_C6
PJ017_C6
PJ030_C1
PJ030_C2
PJ017_C4
PJ017_C2
PJ030_C5
PJ017_C3
PJ017_C5
Stemness
Metastasis
EMT
Hypoxia
Invasion
DNA_damage
Differentiation
Proliferation
Inflammation
Angiogenesis
Quiescence
DNA_repair
Apoptosis
Cell_Cycle
clusters
score
0.5
0
-0.5
clusters
PJ016
PJ017
PJ030