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

Valentina Nale[1]*, Noemi Di Nanni[1]*, Alice Chiodi[1], Ingrid Cifola[1], Marco Moscatelli[1], Cinzia Cocola[1], Matteo Gnocchi[1], Eleonora Piscitelli[1], Rolland Reinbold[1], Ileana Zucchi[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

*Equal contribution

+Corresponding author

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 multiple and complementary criteria applied on SC gene expression data.

RESULTS: scMuffin provides a series of functions to calculate several different qualitative and quantitative scores, such as: expression of marker sets for normal and tumor conditions, pathway activity, cell state trajectories, CNVs, chromatin state 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. As a proof-of-concept, we applied scMuffin to a public SC expression dataset of human high-grade gliomas, where we found that some chromosomal amplifications might underlie the invasive tumor phenotype and identified rare quiescent cells that may deserve further investigations as candidate cancer stem cells.

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 with quantitative (e.g. gene set scores) as well as categorical (e.g. mutation states, proliferation states) features; copy number variation (CNV) analysis; chromatin state analysis; proliferation rate quantification; and marker-based two-sample comparisons **(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.

2. Implementation

scMuffin is implemented in R and provides a series of functions that allows the user to perform various tasks, which can be combined to obtain various data analysis pipelines. The package includes a vignette that describes the use of the tool, and every function is documented. The results from the various analyses (e.g., gene set scores at SC level and cell chromatin state) can be organized in dedicated (simple) objects in order to enable subsequent analyses (e.g., assessment of associations between features and cell clusters) that jointly consider multiple cell features and various ways of cell clustering. Computationally intensive tasks (in particular, gene set scoring and CNV inference) are parallelized. In this section, we describe the algorithms used to perform the several tasks offered by scMuffin.

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

The quantification of gene set expression scores follows the approach described in (8,9), in which a gene set is scored on the basis of its average deviation from an empirical null. In scMuffin the implementation of gene set scoring is parallel and the calculation can be tuned acting on a series of parameters, such as: the number of bins, the number of minimum genes that must have non negative values in a cell ("nmark_min"), the minimum number of cells in which at least nmark_min have to be found ("ncell_min"), the number of permutations (k), the minimum number of permutations required (k'). This tuning helps to address the issue of missing values - typical of SC datasets - and therefore maximizes the number of gene sets for which it is actually possible to obtain an expression score supported by an empirical null distribution. Briefly, given a gene set S:

1. the genes occurring in the genes-by-cells matrix are grouped into a number 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}^*;
5. the average value $\bar{Y}_c$, calculated over the Y_c of a given cluster, is used as the representative score of S in that cluster.

2.2 CNV estimation by adjacent gene windows approach

CNV inference in scMuffin is based on the "adjacent gene windows" approach, which has been validated using both single nucleotide polymorphism arrays (10) and whole-exome sequencing (8) technologies. The approach is implemented in parallel and offers various parameter tuning and data filtering possibilities, which allows the investigator to optimize the analysis on the characteristics of its dataset. The CNV profile of each cell is calculated as a moving average of scaled gene expression levels ordered by genomic location, with the possibility of subtracting a normal reference profile to identify sample-specific CNVs. 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}, 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).

2.3 Chromatin state and proliferation rate

The chromatin state R_c of a cell is inferred on the basis of the number of expressed genes over the number of total mapped reads:

$$R_c = \frac{\#\{_{ic} \geq \ \alpha\}}{\sum_i {}_{ic}}$$

where α is a threshold over the gene count $_{ic}$, which defines the gene *i* as "expressed". High values of R_c indicate cells that are expressing many genes in relation to the number of mapped reads.

The proliferation rate P_c of a cell is quantified as the maximum value 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, respectively, characterize the G1/S and G2/M cell cycle phases:

$$P_c = \max\big(Y_c(S_1), Y_c(S_2)\big)$$

where S_1 and S_2 are defined as in Tirosh *et al.* (8).

2.4 Cluster enrichment analysis for quantitative and categorical features

The assessment of cluster enrichment in high values of quantitative features is computed using a procedure that we name "cell set enrichment analysis" (CSEA), because it is analogous to the gene set enrichment analysis (GSEA) (7), but operates on different input types. In particular, instead of a ranked list of genes, the CSEA considers a list of cells ranked by a feature of interest, and instead of testing a gene set, CSEA tests a cell set (i.e., a cluster of cells). Therefore, CSEA tests whether the cells assigned to a cluster are located at the top (or bottom) of the ranked list of cells. The assessment of a cluster enrichment in a particular value of a categorical feature is computed using the over-representation analysis (ORA) approach (11), which is based on the hypergeometric test.

Both CSEA and ORA are implemented in parallel in scMuffin. This is particularly important for CSEA, which uses permutations to build an empirical null distribution. Nonetheless, it is also effective for ORAs that are run over a large number of gene sets.

2.5 Cell clustering

Cell clustering is based on the approaches implemented in the R package Seurat (3). The results from multiple clustering procedures 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:

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

Meta-clusters are defined as the union of cell clusters that have high o_{ij} and are found by hierarchical clustering of the matrix $O = \{o_{ij}\}$.

2.6 Cluster marker-based two-sample comparison

The cluster marker-based two-sample comparison is based on assessing the expression of cluster markers of every cluster of a sample (A) in every cluster of the other sample (B) and *vice versa*. Given a set of markers S_{a_i}, which represents the cell cluster a_i of sample A, the gene set score $\bar{Y}_{b_j}(S_{a_i})$ quantifies the expression of S_{a_i} in the cell cluster b_j of sample B, while *vice versa* $\bar{Y}_{a_i}\left(S_{b_j}\right)$ quantifies the expression of S_{b_j} in the cell cluster a_i of sample A. The average value $\bar{Y}_{a_i b_j} = \frac{\bar{Y}_{b_j}(S_{a_i}) + \bar{Y}_{a_i}(S_{b_j})}{2}$ quantifies the similarity between cell clusters a_i and b_j on the basis of the expression of their markers. The procedure is repeated for all-pairs of clusters of sample A and sample B.

3. Results and Discussion

In this section, we present the user interface **(Table 1)**, and the results obtained using our scMuffin package on the SC dataset generated by Yuan *et al*. (12) from human high-grade glioma (HGG) samples, and available on the Gene Expression Omnibus (GEO) repository (GSE103224) (13).

Table 1. Main tasks and corresponding functions in scMuffin.

Task	Description	User interface
Gene set expression scoring	Average gene set expression deviation from matched empirical reference; provided gene sets from CellMarker (4), PanglaoDB (5), CancerSEA (6) and MSigDB (7)	• `prepare_gsls` • `calculate_gs_scores` • `calculate_gs_scores_in_clusters`

Copy number variations	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	• `calculate_CNV` • `cluster_by_features` • `apply_CNV_reference` • `CNV_heatmap` • `process_GTEx_gene_reads`
Chromatin state	Number of expressed genes in relation to the total reads	• `exp_rate`
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`
Two-sample comparison	Quantification of the expression similarity between all-pairs of clusters between two samples	• `quantify_samples_similarity`
Assembling cell features and cell partitions	Objects that host cell-level feature values and cell partitions	• `create_features_obj` • `create_partitions_obj`
Visualization	Automated UMAP visualizations for multiple features, heatmaps, box plots and dot plots	• `boxplot_cluster` • `dotplot_cluster` • `quantify_samples_similarity` • `heatmap_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 (4), PanglaoDB (5), CancerSEA (6) and MSigDB (7) databases. Unlike many existing tools that are used to perform marker-based cell annotation (14), 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`, which requires the expression matrix, the gene sets, as well as a series of optional parameters to fine-tune the calculation in relation to the characteristics of the SC dataset under analysis. This tuning is important to address the heterogeneity of size and sparsity that characterizes different SC datasets, attributable to both the biological specimen under analysis and the SC platform used. For instance, the following code shows how to quantify cell-level, and then cluster-level, expression score of the cancer functional states from CancerSEA ("SIG_CancerSEA") (6) in a normalized genes-by-cells ("gbc") expression matrix:

```
gsc <- prepare_gsls(gs_sources = "SIG_CancerSEA", genes = rownames(gbc))
gs_scores_obj    <-    calculate_gs_scores(genes_by_cells    =    gbc,    gs_list    =
gsc$SIG_CancerSEA)
res_sig_cl   <-   calculate_gs_scores_in_clusters(gs_scores_obj   =   gs_scores_obj,
cell_clusters = cell_clusters)
```

The cell-level value of any gene set (and more generally of any feature) can be visualized over the UMAP by means of `plot_umap_colored_features`, while cluster-level values of multiple gene sets (features) can be visualized as a heatmap using `plot_heatmap_features_by_clusters`, which relies on the ComplexHeatmap R package (15). For example, in our case study, the analysis of the CancerSEA functional states in the HGG sample PJ016 showed that the two groups of cell clusters (**Figure 2a,** left and right of the UMAP) reflect distinct functional states **(Figure 2b)**: for example, the expression of the CancerSEA "Invasion" markers was particularly high in cell clusters 0 and 9 as compared to all the other clusters **(Figure 2b-c)**.

3.2 CNV estimation and association with CancerSEA functional states

CNV inference from SC expression data estimates the presence of relevant genomic aberrations (amplifications 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 pattern allows the investigator to hypothesize link between genomic alterations and cell phenotypes.

The function `calculate_CNV` basically retrieves the genomic locations and performs the CNV estimation; `cluster_by_features` executes the cell clustering based on CNV profiles; `apply_CNV_reference` redefines the CNV values on the basis of normal reference cells; the dedicated plotting function `CNV_heatmap` handles the visualization, where the cell cluster that contains the reference is marked in red. Here is an example that illustrates CNV inference using a 100 genes window size and a normal reference profile from The Genotype-Tissue Expression (GTEx) portal (16):

```
cnv_res <- calculate_CNV(gbc, wnd_size = 100, reference =
GTEx_mean)
cnv_clustering <- cluster_by_features(cnv_res, cnv=TRUE)
cnv_res_ref <- apply_CNV_reference(cnv = cnv_res, cnv_clustering =
cnv_clustering, reference="reference")
cnv_res_ref <- CNV_heatmap(cnv = cnv_res, cnv_clustering =
cnv_clustering, reference="reference")
```

To illustrate the use of this workflow, we selected two different HGG samples by Yuan *et al.* (12), that is, PJ030, composed by tumor cells as well as not transformed cells and PJ016, including only transformed cells. We observed that the reference profile (obtained using the average gene expression values of the normal brain samples available in GTEx portal) falls into cluster 3 of PJ030 **(Figure 3a-b)**. With respect to cluster 3 (corresponding to the not transformed cells included in this sample), the clusters 0, 1 and 2 showed large recurrent aneuploidies, some of which are typical of HGG, like the amplification of chromosome 7 **(Figure 3a)**. The CNV pattern here inferred is fundamentally coherent with that reported by Yuan *et al*. (12), even if the authors just quantified a summary value per chromosome, while scMuffin provides multiple CNV estimations per chromosome. For sample PJ016, the CNV inference analysis highlighted two groups of "CNV clusters" that map to two distinct components of the UMAP, while it did not identify a diploid cluster, accordingly to the presence of only transformed cells in this sample **(Figure 3c-d).** Interestingly, clusters 1 and 3 were marked by peculiar amplifications in chromosomes 1p and 19p.

scMuffin enables the comparison of clusters obtained using different procedures. In particular, the overlap among all-pairs of clusters can be quantified using:

```
cl_list <- partitions_to_list(clust_obj)
ov_mat <- overlap_matrix(cl_list)
```

In our case study, the comparison between expression clusters and CNV clusters of sample PJ016 confirmed the presence of two groups of cells: for example, CNV clusters 1 and 3 showed a relevant overlap with expression clusters 0, 6, 8 and 9 **(Figure 3e).**

An example of integrative analysis enabled by scMuffin is the functional assessment of CNV patterns. We quantified the expression scores of the CancerSEA functional states throughout the CNV clusters of sample PJ016. As expected, the two aforementioned groups of CNV clusters (0-2-4 and 1-3) were characterized by different functional states **(Figure 3f)**, like the corresponding expression clusters**.** In particular, CNV cluster 3 – which is mainly located in the top-left region of the UMAP visualization (**Figure 3d**) and has a strong overlap with expression clusters 0 and 9 **(Figure 3e)** – contains cells that highly express the CancerSEA "Invasion" markers **(Figure 3f** and **Figure 2)**. This finding suggests that the peculiar amplifications of chromosomes 1p and 19p found in this cluster might be linked to the invasive phenotype. This hypothesis is supported by the finding of two CancerSEA invasion markers, Y-Box Binding Protein 1 *(YBX1)* and Heterogeneous Nuclear Ribonucleoprotein M *(HNRNPM)*, located within the amplified regions of chromosome 1p and 19p specifically found in CNV clusters 1 and 3 **(Figure 3c)**. *YBX1* is a DNA/RNA-binding protein and transcription factor which plays a central role in coordinating tumor invasion in glioblastoma (17). *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 (18,19). In particular, HNRNPM was identified as an interactor of the DNA/RNA binding protein SON, which drives oncogenic RNA splicing in glioblastoma (20). While it is beyond the scope of this article to further study this hypothesis, these findings clearly highlight the usefulness of the integrative analysis of CNVs and CancerSEA functional states provided by our scMuffin tool.

3.3 Clustering, features and annotation

scMuffin contains functions for assessing the association between cell clusters and quantitative as well as categorical features, by means of CSEA and ORA, respectively. Here is the user interface, where, firstly, the objects containing cell clusters and cell features are set up; then, the enrichment is quantified for all partitions (various ways of clustering cells) and all features:

```
clust_obj <- create_partitions_obj(cell_clusters)
feat_obj <- create_features_obj(feature_values)
cl_enrich <- assess_cluster_enrichment(features = feat_obj,
partitions = clust_obj)
```

The results of CSEA and ORA can be extracted to produce features-by-clusters matrices that contain any score calculated by CSEA or ORA, like, for example, normalized enrichment scores (NES) values and enrichment ratios (er):

```
    cl_enrich_table <-
extract_cluster_enrichment_table(cl_enrich, q_type = "nes", c_type
= "er")
```

The results of enrichment analysis can be visualized as box plots (quantitative features) and dot plots (categorical features):

```
top_feat_lab_CSEA <- boxplot_cluster(features = feat_obj,
partitions = clust_obj, clustering_name = "global",
clust_enrich_res = cl_enrich, criterion = "fdr")
top_feat_lab_ORA <- dotplot_cluster(features = feat_obj,
partitions = clust_obj, clustering_name = "global",
clust_enrich_res = cl_enrich, text_val = "p")
```

These plots show, for each cluster, the distribution of values of the most significant features in the cluster in comparison to all the other clusters, and the related scores (e.g., NES, p-value and FDR). In addition, `boxplot_cluster` and `dotplot_cluster` provide the labels of the most significant features associated with any cluster. These labels can be extracted from the enrichment analysis results also by means of `extract_cluster_enrichment_tags`, according to various criteria (e.g., NES, enrichment ratio, p-value, FDR) that are specific to CSEA or ORA.

To illustrate these functions, we assessed the enrichment of the expression clusters of sample PJ016 in terms of both CancerSEA functional states (quantitative features) and three categorical features, namely: cell clusters obtained analyzing ribosomal gene expression, a gene set included in scMuffin because changes in ribosomal gene expression were associated with specific cancer phenotype and can reveal specific malignant subpopulations (21–23); cell clusters obtained using a glioblastoma signature of 500 genes (24) whose expression can be used to classify glioblastoma subtypes; cell cycle phase. Considering as an example the cluster 0 of sample PJ016, the analysis showed that it was significantly enriched in cells that, in comparison with cells of other clusters, highly express the gene markers of CancerSEA “Invasion” state **(Figure 4a)** and are in S and G2M phases **(Figure 4b).** The labels of the most significant features of any cluster can be used, by means of `plot_umap`, to plot an annotated UMAP **(Figure 4c)** .

3.4 Chromatin state, proliferation rate and cell state trajectories

Chromatin state and proliferation rate carry two relevant pieces of information for the characterization of a cancer cell.

In particular, an open chromatin state is peculiar of stem cells (and cancer stem cells (CSCs)), might indicate de-differentiation processes of tumor progression and might influence cell plasticity, favoring cancer cell adaptability and drug resistance (25,26). In a recent study on glioblastoma, chromatin accessibility was associated to a specific subpopulation of putative tumor-initiating CSCs with invasive phenotype and low survival prediction (27). The global state of the chromatin at SC level can be inferred from SC transcriptomic data and provides a simple and useful score that can be used to distinguish specific cell types, such as CSCs. The chromatin state can be quantified by means of the function `exp_rate` on the genes-by-cells count matrix:

```
res <- exp_rate(gbc, min_counts = 5)
```

where 5 is the required threshold above which a gene is considered expressed.

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 (28,29). In scMuffin, we quantify cell proliferation rate on the basis of the expression of G1/S and G2/M genes:

```
res <- proliferation_analysis(gbc)
```

As a proof-of-concept, we show the joint analyses of chromatin state and proliferation rate in sample PJ016 and visualize the results in the state space of cell differentiation trajectories. In scMuffin, cell state trajectories are inferred using the “diffusion maps” approach available in the destiny R package (30), by means of the wrapper function:

```
        res <- diff_map(gbc)
```

Interestingly, we observed that cells showing high values of chromatin state score - cells that are expressing a relatively high number of genes (i.e., an open chromatin state) - are located at the root of the trajectory state space **(Figure 5a)**, while the cells that show the highest proliferation rates are located at a corner of the state space **(Figure 5b)**. This pattern suggests that the cells with high values of chromatin state could be quiescent cells, which express a large number of genes but are not actively dividing. Therefore, these cells are interesting candidates for further analysis aimed at studying CSCs in HGG. More generally, this proof-of-concept demonstrates the usefulness of the chromatin state score defined here, especially if used in combination with the proliferation rate for the identification of particular cell types or cell states.

3.5 Comparison of samples

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 (31). 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 two samples: the expression of the cluster markers of a sample is assessed in the other sample – and *vice versa* – obtaining the similarities among all pairs of clusters. For example, Nguyen *et al*. (9) 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 similarity between all-pairs of clusters of two samples on the basis of cluster-specific markers:

```
sim_res <- quantify_samples_similarity(gbc_1, gbc_2, clusters_1,
clusters_2, cluster_markers_1, cluster_markers_2)
```

Concerning our case study, the comparison of samples PJ016 and PJ018 showed a series of similarities between their clusters. For instance, the clusters 0 and 9 of sample PJ1016 are composed of cells highly similar to those grouped into clusters 2 and 5 of sample PJ018 **(Figure 6a)**. This analysis revealed a pattern of cluster-cluster similarities that is fundamentally coherent with the results obtained by performing the alternative approach of integrating the two datasets and then clustering the cells **(Figure 6b-c)**. For example, both approaches showed that clusters 0 and 9 of PJ1016 are similar to clusters 2 and 5 of PJ018, and cluster 4 of PJ016 is close to cluster 7 of PJ018. There were also some differences, which, yet again, remark the challenge of this task: for example, cluster 8 of PJ016 is similar to cluster 9 of PJ018 using the marker-based similarity **(Figure 6a)**, while the UMAP obtained by the integrated dataset places cluster 8 of PJ016 close to clusters 6 and 3 of PJ018 **(Figure 6b-c**).

4. Conclusions

Here, we presented scMuffin, an R package that we developed to offer a series of useful functions to perform and integrate multiple types of analyses on SC expression data. As a proof-of-concept, we applied scMuffin on a publicly available SC expression dataset of human HGG. We described two examples of integrative analyses which returned particularly interesting findings that would deserve further investigations. The functional characterization of CNVs highlighted a possible link between amplifications of chromosomes 1p and 19p and invasive tumor phenotype. The joint analysis of chromatin state, proliferation rate and cell state trajectories suggested possible candidates of CSCs in HGG. 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. a-d) CNV heatmaps (a, c) where cells (columns) are grouped into CNV clusters, and UMAP visualizations (b, d) where cells are colored by CNV clusters, for sample PJ030 (a, b) and sample PJ016 (c, d). **e)** Overlap between cell clusters of sample PJ016 obtained by analyzing gene expression (rows, "global_" prefix) and CNV clusters (columns, "cnv_" prefix); *YBX1* and *HNRNPM* are two CancerSEA invasion markers located within the amplified 1p and 19p regions found in CNV clusters 1 and 3. **f)** Expression scores for CancerSEA functional states in CNV clusters of sample PJ016.

Figure 4. Cluster enrichment in HGG sample PJ016. a) The top five most significant (fdr < 0.05) CancerSEA functional states in cluster 0: distribution of expression scores in cluster 0 (red) in comparison with all the other clusters (grey); normalized enrichment score (NES) and false discovery rate (FDR) values. **b)** Distribution of cells by their values (red labels) in cluster 0 (red dots) in comparison with all the others (grey dots) for three categorical variables, namely, the clusters obtained analyzing ribosomal gene expression ("ribosomes"), the clusters obtained analyzing the expression of a Glioblastoma signature ("GB500"), and cell cycle phase ("Phase", obtained with the Seurat package function "CellCycleScoring"); the numbers over each cell group

are ORA p-values. **c)** UMAP visualization with expression clusters annotated with the names of the top two CancerSEA gene sets with the highest enrichment (CSEA) for each cluster.

Figure 5. Chromatin state, proliferation rate and cell state trajectories of HGG sample PJ016. (a-b). Cell state trajectories colored by chromatin state (a) and proliferation rate (b).

Figure 6. Cluster marker-based comparison of HGG samples PJ016 and PJ018. a) Similarity among all-pairs of clusters. **b-c)** UMAP visualizations obtained by integrating the two samples with the "FindIntegrationAnchors" and "IntegrateData" Seurat functions, showing PJ016 cells (b) and PJ018 cells (c) colored by the clusters found by independent analysis of each sample.

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

CSEA: Cell Set Enrichment Analysis

GSEA: Gene Set Enrichment Analysis

HGG: high grade glioma

ORA: over representation analysis

SC: single cell

CSC: cancer stem cells

UMAP: Uniform Manifold Approximation and Projection

GTEx: The Genotype-Tissue Expression project

NES: normalized enrichment score

FDR: False discovery rate

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 (13), under the accession GSE103224; the GTEx Portal (16) 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. NDN drafted the package, carried out the analyses and wrote the article. AC implemented clustering functions, carried out the analyses and wrote the manuscript. IC curated the biological aspects of CNV analysis and revised the manuscript. MM and MG set up the computational infrastructure for data analysis. CC and EP curated the biological aspects of solid tumor data analysis. 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 manuscript. 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. 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

5. Franzén O, Gan L-M, 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

6. 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

7. 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. Proc Natl Acad Sci [Internet]. 2005 Oct 25;102(43):15545–50. Available from: https://pnas.org/doi/full/10.1073/pnas.0506580102

8. 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 (80-) [Internet]. 2016 Apr 8;352(6282):189–96. Available from: https://www.science.org/doi/10.1126/science.aad0501

9. 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

10. 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 (80-) [Internet]. 2014 Jun 20;344(6190):1396–401. Available from: https://www.science.org/doi/10.1126/science.1254257

11. 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

12. 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

13. 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

14. 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

15. 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

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

17. 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

18. 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

19. Golan-Gerstl R, Cohen M, Shilo A, Suh S-S, 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

20. Kim J-H, 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

21. Bastide A, David A. The ribosome, (slow) beating heart of cancer (stem) cell. Oncogenesis [Internet]. 2018 Apr 20;7(4):34. Available from: http://www.nature.com/articles/s41389-018-0044-8

22. Guimaraes JC, Zavolan M. Patterns of ribosomal protein expression specify normal and malignant human cells. Genome Biol [Internet]. 2016 Dec 24;17(1):236. Available from: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1104-z

23. Dolezal JM, Dash AP, Prochownik E V. Diagnostic and prognostic implications of ribosomal protein transcript expression patterns in human cancers. BMC Cancer [Internet]. 2018 Dec 12;18(1):275. Available from: https://bmccancer.biomedcentral.com/articles/10.1186/s12885-018-4178-z

24. Teo W-Y, Sekar K, Seshachalam P, Shen J, Chow W-Y, Lau CC, et al. Relevance of a TCGA-derived Glioblastoma Subtype Gene-Classifier among Patient Populations. Sci Rep [Internet]. 2019 Dec 15;9(1):7442. Available from: http://www.nature.com/articles/s41598-019-43173-y

25. Wainwright EN, Scaffidi P. Epigenetics and Cancer Stem Cells: Unleashing, Hijacking, and Restricting Cellular Plasticity. Trends in Cancer [Internet]. 2017 May;3(5):372–86. Available from: https://linkinghub.elsevier.com/retrieve/pii/S240580331730081X

26. 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

27. 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

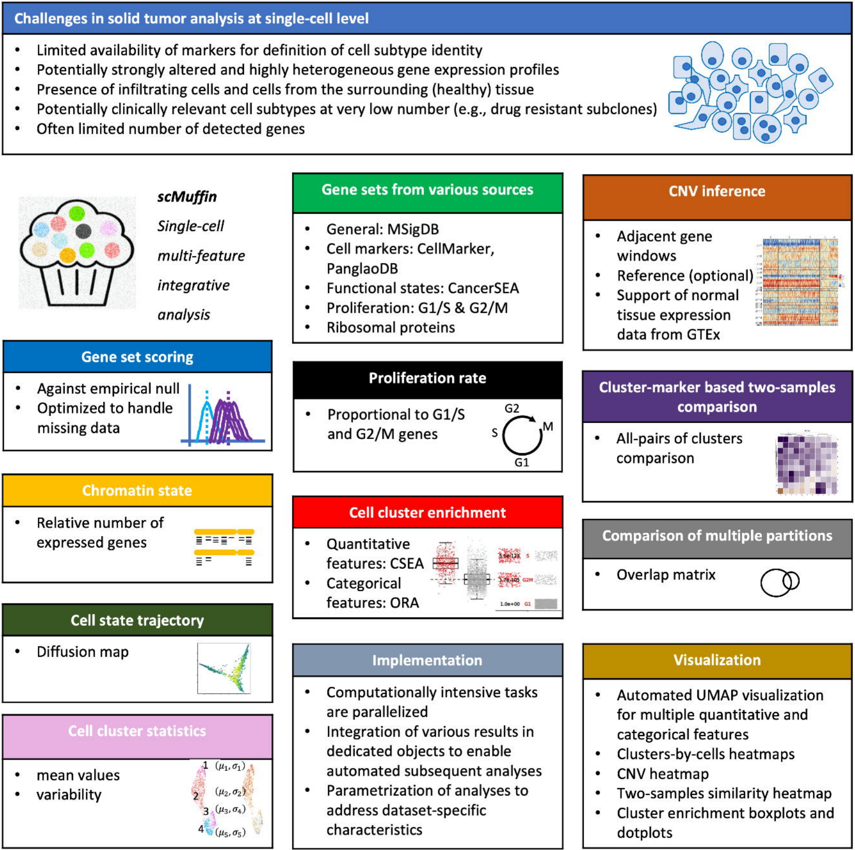
28. Feitelson MA, Arzumanyan A, Kulathinal RJ, Blain SW, Holcombe RF, Mahajna J, et al. Sustained proliferation in cancer: Mechanisms and novel therapeutic targets. Semin Cancer Biol [Internet]. 2015 Dec;35:S25–54. Available from:

https://linkinghub.elsevier.com/retrieve/pii/S1044579X15000140

29. 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

30. 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

31. 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



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

Chromatin state

- Relative number of expressed genes

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

Cell cluster enrichment

- 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 (optional)
- Support of normal tissue expression data from GTEx

Cluster-marker based two-samples comparison

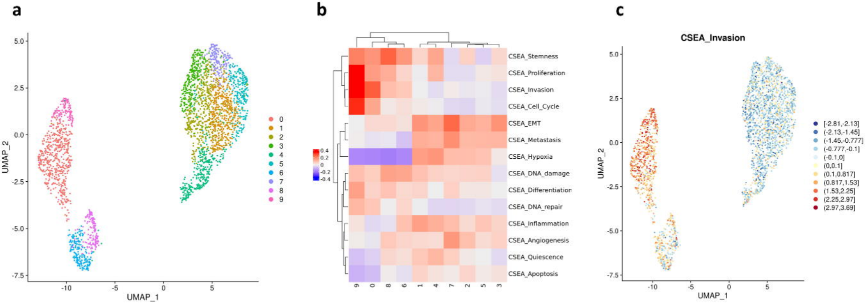
- All-pairs of clusters comparison

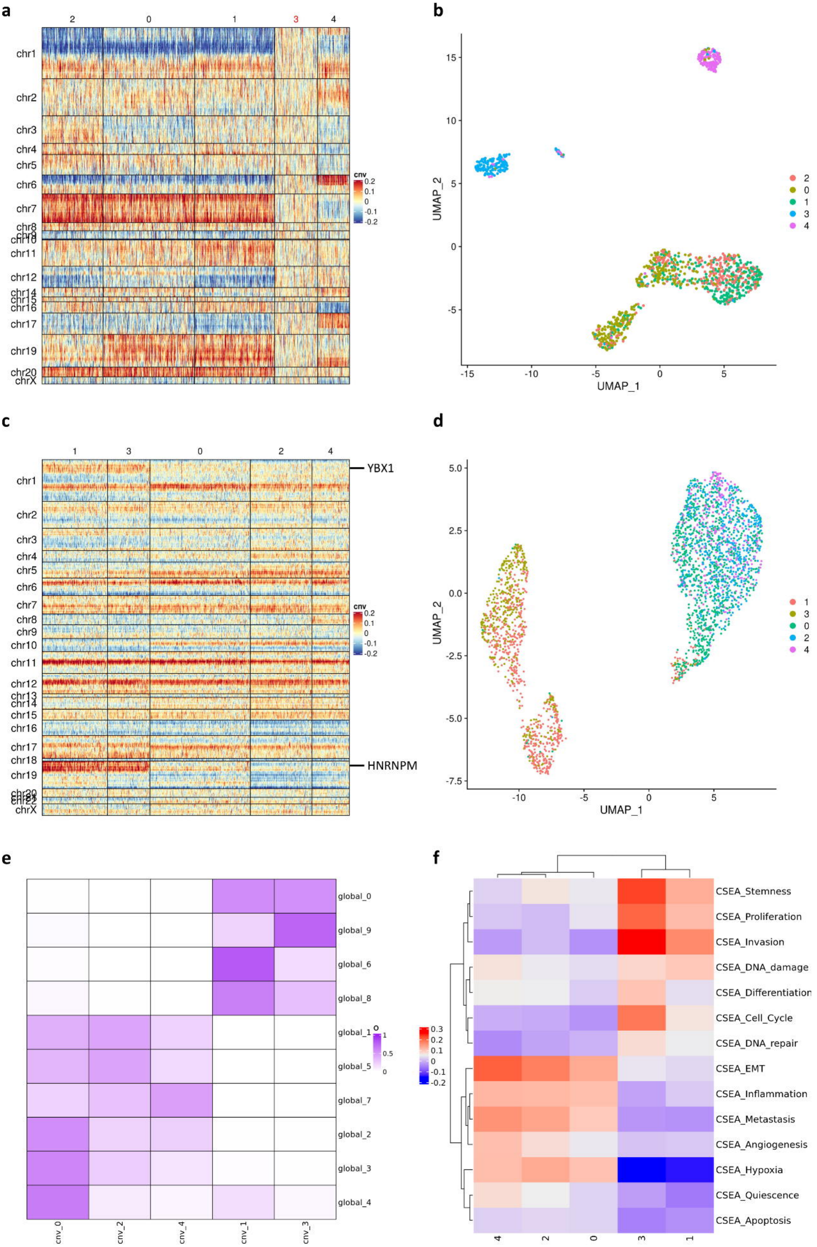
Comparison of multiple partitions

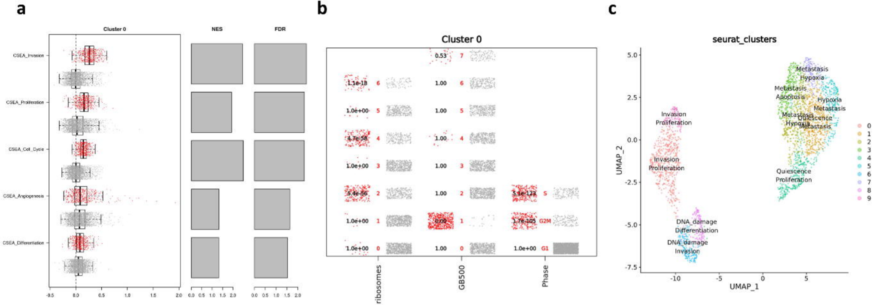
- Overlap matrix

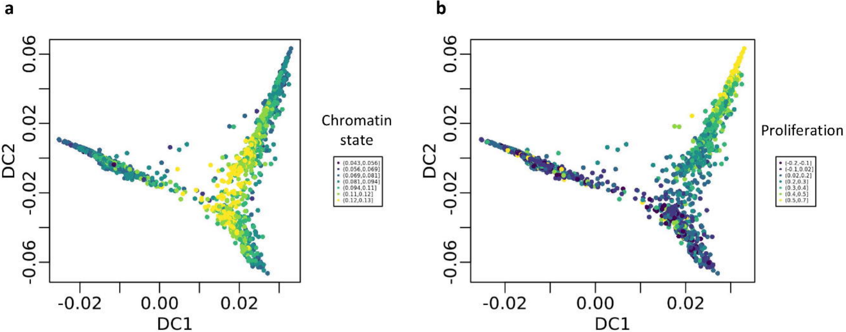
Visualization

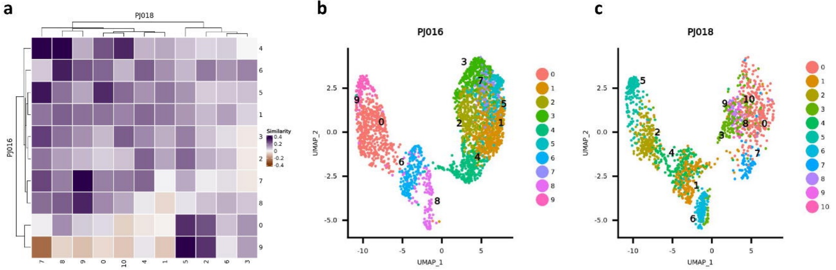
- Automated UMAP visualization for multiple quantitative and categorical features
- Clusters-by-cells heatmaps
- CNV heatmap
- Two-samples similarity heatmap
- Cluster enrichment boxplots and dotplots


a
5.0
2.5
0.0
-2.5
-5.0
-7.5
UMAP_2
-10
-5
0
5
UMAP_1
0
1
2
3
4
5
6
7
8
9
b
0.4
0.2
0
-0.2
-0.4
CSEA_Stemness
CSEA_Proliferation
CSEA_Invasion
CSEA_Cell_Cycle
CSEA_EMT
CSEA_Metastasis
CSEA_Hypoxia
CSEA_DNA_damage
CSEA_Differentiation
CSEA_DNA_repair
CSEA_Inflammation
CSEA_Angiogenesis
CSEA_Quiescence
CSEA_Apoptosis
9
0
8
6
1
4
7
2
5
3
c
CSEA_Invasion
5.0
2.5
0.0
-2.5
-5.0
-7.5
UMAP_2
-10
-5
0
5
UMAP_1
[-2.81,-2.13]
(-2.13,-1.45]
(-1.45,-0.777]
(-0.777,-0.1]
(-0.1,0]
(0,0.1]
(0.1,0.817]
(0.817,1.53]
(1.53,2.25]
(2.25,2.97]
(2.97,3.69]

a
2
0
1
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
UMAP_2
UMAP_1
15
10
5
0
-5
-15
-10
-5
0
5
2
0
1
3
4
c
1
3
0
2
4
YBX1
HNRNPM
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
d
UMAP_2
UMAP_1
5.0
2.5
0.0
-2.5
-5.0
-7.5
-10
-5
0
5
1
3
0
2
4
e
global_0
global_9
global_6
global_8
global_1
global_5
global_7
global_2
global_3
global_4
o
1
0.5
0
cnv_0
cnv_2
cnv_4
cnv_1
cnv_3
f
0.3
0.2
0.1
0
-0.1
-0.2
CSEA_Stemness
CSEA_Proliferation
CSEA_Invasion
CSEA_DNA_damage
CSEA_Differentiation
CSEA_Cell_Cycle
CSEA_DNA_repair
CSEA_EMT
CSEA_Inflammation
CSEA_Metastasis
CSEA_Angiogenesis
CSEA_Hypoxia
CSEA_Quiescence
CSEA_Apoptosis
4
2
0
3
1


a
Cluster 0
NES
FDR
CSEA_Invasion
CSEA_Proliferation
CSEA_Cell_Cycle
CSEA_Angiogenesis
CSEA_Differentiation
-0.5
0.0
0.5
1.0
1.5
2.0
0.0 0.5 1.0 1.5 2.0
0.0 0.5 1.0 1.5 2.0
b
Cluster 0
0.53
7
1.1e-18
6
1.00
6
1.0e+00
5
1.00
5
4.7e-58
4
1.00
4
1.0e+00
3
1.00
3
5.4e-56
2
1.00
2
5.1e-123
S
1.0e+00
1
0.00
1
1.7e-105
G2M
1.0e+00
0
1.00
0
1.0e+00
G1
ribosomes
GB500
Phase
c
seurat_clusters
Metastasis
Hypoxia
Metastasis
Apoptosis
Hypoxia
Metastasis
Metastasis
Quiescence
Hypoxia
Metastasis
Invasion
Proliferation
Invasion
Proliferation
Quiescence
Proliferation
DNA_damage
Differentiation
DNA_damage
Invasion
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

a
b
DC2
DC1
0.06
0.02
-0.02
-0.06
-0.02
0.00
0.02
Chromatin state
(0.043,0.056]
(0.056,0.069]
(0.069,0.081]
(0.081,0.094]
(0.094,0.11]
(0.11,0.12]
(0.12,0.13]
Proliferation
[-0.2,-0.1]
(-0.1,0.02]
(0.02,0.2]
(0.2,0.3]
(0.3,0.4]
(0.4,0.5]
(0.5,0.7]


a
PJ018
PJ016
4
6
5
1
3
2
7
8
0
9
7
8
9
0
10
4
1
5
2
6
3
Similarity
0.4
0.2
0
-0.2
-0.4
b
PJ016
UMAP_2
UMAP_1
2.5
0.0
-2.5
-5.0
-10
-5
0
5
0
1
2
3
4
5
6
7
8
9
c
PJ018
UMAP_2
UMAP_1
2.5
0.0
-2.5
-5.0
-10
-5
0
5
0
1
2
3
4
5
6
7
8
9
10