Description of scripts used for single-cell multiomic analysis of hepatoblastomas

This folder contains all the scripts used in the manuscript entitled ' Single-cell multiomics reveals the interplay of clonal evolution and cellular plasticity in hepatoblastoma' by Roehrig *et al.*

Scripts were run in R version 4. The following **tools and packages** are required:
cellranger-arc (v2.0.0)
Seurat (v3)
ArchR (v1.0.1)
Signac (v1.6)
InferCNV (v1.6)
spaceranger (v1.3.1)
screadcounts (v1.1.8)
samtools view (v1.14)
Installing these tools should take ~2-3 hours on a desktop computer.

Input data are the raw sequencing files available on the European Genome Archive (EGA, accession code: EGAS00001006932).

Below is a description of the **functionalities** of each script. Scripts related to each data type (snRNAseq, snATACseq etc.) are grouped in separate folders.

snRNAseq

1. Preprocessing

Quality control of the 8 Multiome samples based on number of UMI counts, detected genes and percentage of mitochondrial reads is performed in 1_Preprocessing.R and visualized in 2_Plot_QC_all_samples.R.

2. Global_visualization

Normalization, dimensionality reduction, clustering and 2D plot projection for 1) each sample individually (in *1_Seurat_SCTransform_per_sample.R*) or 2) merging cells from all 8 samples (in *2_Seurat_SCTransform_merge_all.R*). Normalization is performed with SCTransform for the main analyses and with NormalizeData when focusing on expression levels between cells. Projection of known signatures to identify the cell subtypes is performed in *3_Seurat_signature_analyses.R* and 4_Seurat_H_LP_M_signature_violin_plots.R.

3. Denoising

In-house smoothing of missing values by aggregating the expression values in hundreds of small-size cell clusters. Imputation is performed in *Imputation_in_house_tumor_cells.R*.

4. Pseudo_CNV

Computation of virtual copy-number alteration profiles from cell-level expression data, using normal hepatocytes from one non-tumor sample as reference cells. Input matrices are generated in *1_InferCNV_table_preparation.R* (*Seurat*). *InferCNV* is run on each sample separately or on all cells from the 8 samples in *2_InferCNV_pipeline_individual_sample.R* and

2_InferCNV_pipeline_no_cluster_by_sample.R. Alteration clusters are defined in *3_InferCNV_analyses_no_cluster_by_sample.R*.

5. Tumor_study

Identification of non-tumor, tumor cells and potential doublets based on PCA of all cells, projection of known signatures, presence of chromosomal aberrations and ATAC doublet enrichment is performed in *1_Identification_normal_tumor_on_merged_UMAP.R*. Analysis pipeline is performed again on tumor cells from all 6 samples (*2_Seurat_analyses_tumor_cells_strict.R*) or each sample separately (*4_Seurat_analyses_tumor_cells_individual_strict.R*). scH, scH/LP, scLP and scM Multiome subtypes are identified in *3_Identification_H_LP_M_strict.R* and the correlation between PCA components PC1 and PC2, and known HB markers, is realized in *3_PCA_component_identification_H_LP_M_correlation.R*. Differential expression analysis between subtypes is performed in *5_Differential_expression_subtypes.R*. Signatures related to normal liver development or signatures from other HB single-cell studies are investigated in *Gene_signatures /1_Development_signatures.R* and *Gene_signatures/2_Other_HB_scRNAseq_studies.R*.

6. Clonal_evolution

Definition of genetic subclones from copy-number alteration profiles in *1_Final_definition_CNV_clusters.R*, and their characterization with diverse annotations in *2_Characterization_of_CNV_clusters.R* and *2_Characterization_of_CNV_clusters_CHC2959T_2960T.R*. Somatic mutations are investigated in the « *Somatic mutations* » folder.

BAM files are filtered on reads kept for UMI counting (tag xf=25) in *0_Filter_BAM_on_xf_tag_before_screadcounts.R*.

Lists of somatic mutations from bulk WGS are prepared in *1_Preparation_tables_screadcounts.R*. *scReadCounts* is run on the single-nucleus BAM files of each sample to detect WGS mutations in *2_scReadCounts_command_line.R*. Study of mutations clusters from patient 2959 is performed in *3_scReadCounts_analyses_2959_patient.R*. Study of tumor-specific sets of mutations is performed in *4_scReadCounts_analyses_global_mutations.R*. Study of specific mutations like *CTNNB1* is performed in *4_scReadCounts_analyses_individual_mutations.R*. Investigation of clonal mutation specificity and sensitivity is performed in *5_scReadCounts_clonal_mutations_all_samples.R*. Detection of 11p15 alterations is performed in the folder « *11p15* » in *1_scReadCounts_analyses_11p15_cnLOH.R*.

7. Cancer_stem_cells

Identification of cancer stem cell and liver cancer stem cell markers is performed in *1_Cancer_stem_cells_study_all_csc_markers.R* and *1_Cancer_stem_cells_study_liver_csc_markers.R*.

snATACseq

ATAC_tracks_per_subtype.R enables to visualize chromatin accessibility in selected regions aggregated for each Multiome HB subtype.

Create_promoter_gene_body_table.R enables to create a table containing genomic features for peaks.

1. Preprocessing

Quality control of the 8 Multiome samples based on number of fragments and TSS enrichment, and creation of the corresponding Arrow files, are performed in

1_Preprocessing_quality_control_individual_sample_QC.R. Arrow files are gathered in an *ArchR* project in 2_Create_project_from_individual_QC.R.

2. Global_visualization

Normalization, dimensionality reduction, clustering and 2D plot projection for 1) merging cells from all 8 samples (in 1_Normalization_clustering_UMAP_all_samples.R), 2) merging tumor cells from all 6 tumor samples (in 1_Normalization_clustering_UMAP_tumor_cells.R), or 3) tumor cells from each individual patient (in *1_Normalization_clustering_UMAP_tumor_cells_by_sample.R* and *1_Normalization_clustering_UMAP_tumor_cells_by_sample_2959.R*).

3. Peak_calling

We used *ArchR* for most snATAC-seq analyses and *Signac* for peak calling. Intersection between cells identified as non-empty droplets by *cellranger* and cells retained after quality control in *ArchR* is performed in *1_Preprocess_individual_samples_ArchR_QC.R*. Peak calling is then performed on each HB subtype in each tumor sample in *2_Peak_calling_individual_samples.R*, and the resulting peak sets are merged in *3_Peak_calling_combine_samples.R*. The global peak set obtained with *Signac* is integrated in the *ArchR* pipeline and the corresponding peak count matrix is computed in *4_Integrate_Signac_peakset_in_ArchR_peak_calling.R*. Peaks are annotated with their nearest gene in *5_Compute_nearest_gene_per_peak.R*.

4. Differential_peaks

Differential peak analysis is performed between tumor subtypes and non-tumor cells on the one hand, and between tumor subtypes on the other hand, in *1_Differential_peaks_tables.R*. The resulting differential peaks are annotated by their logFC, FDR, nearest gene, distance to nearest TSS, linked genes and regulating transcription factors in *2_Annotate_differential_peaks.R*.

5. Peaks_study

The number of differential peaks for each comparison is computed in *1_Differential_study_nb_peaks.R*. Differential peaks are characterized by their chromatin state enrichment in *1_Differential_study_chromatin_states.R*.

6. Peak_to_gene_linkage

Linkage between peaks and genes based on correlation between peak accessibility and gene expression is performed in *1_Peak_to_gene_linkage_multiome.R*.

7. Motif_TF

A custom position weith matrix list to include motif information from version 2 of the cisbp database is created in *0_Create_custom_PWM_list_cisbp_v2.R*. Motif enrichment analysis in differential peaks related to each Multiome HB subtype is performed in *1_Motif_enrichment_differential_peaks_cisbp_v1_v2.R*. TF footprints for specific TF are computed in *2_TF_footprints_cisbp_v2.R*. TF motif deviations are computed using cisbp versions 1 and 2 in *2_TF_motif_deviations_cisbp_v1.R* and *2_TF_motif_deviations_cisbp_v2.R*.

Integration_snRNAseq_snATACseq

Identification of key TF regulators based on specific criteria reyling on bulk RNA-seq, snRNA-seq and snATAC-seq results is performed in *1_Identify_TF_regulators.R*. Gene regulatory networks between the defined key TF and their potential target genes are identified in *2_Identify_GRN_networks_H_LP_M.R*. Aggregated single-nucleus expression of each resulting module (TF and target genes) is projected in boxplots in *3_Boxplots_GRN_modules_per_Multiome_subtype.R*. Single-nucleus expression and motif deviation of the corresponding TF, as well as single-nucleus expression of their targets, are projected on a PC2-ordered heatmap in *3_Heatmap_Multiome_TF_targets_expression.R*. Bulk expression of the corresponding TF and their targets is projected on a PC2-ordered heatmap in *3_Heatmap_bulk_RNA_seq_TF_targets_expression.R*. Recapitulating tables of correlation between TF and their targets, as well as mean expression of the TF and targets in each Multiome HB subtype, can be found in *Tables.R*.

Visium

Multiome scH and scLP markers are projected in the Visium data of one tumor from the Multioem cohort (#3133T) in *1_Project_H_LP_snRNAseq_markers_on_visium_3133T.R*.