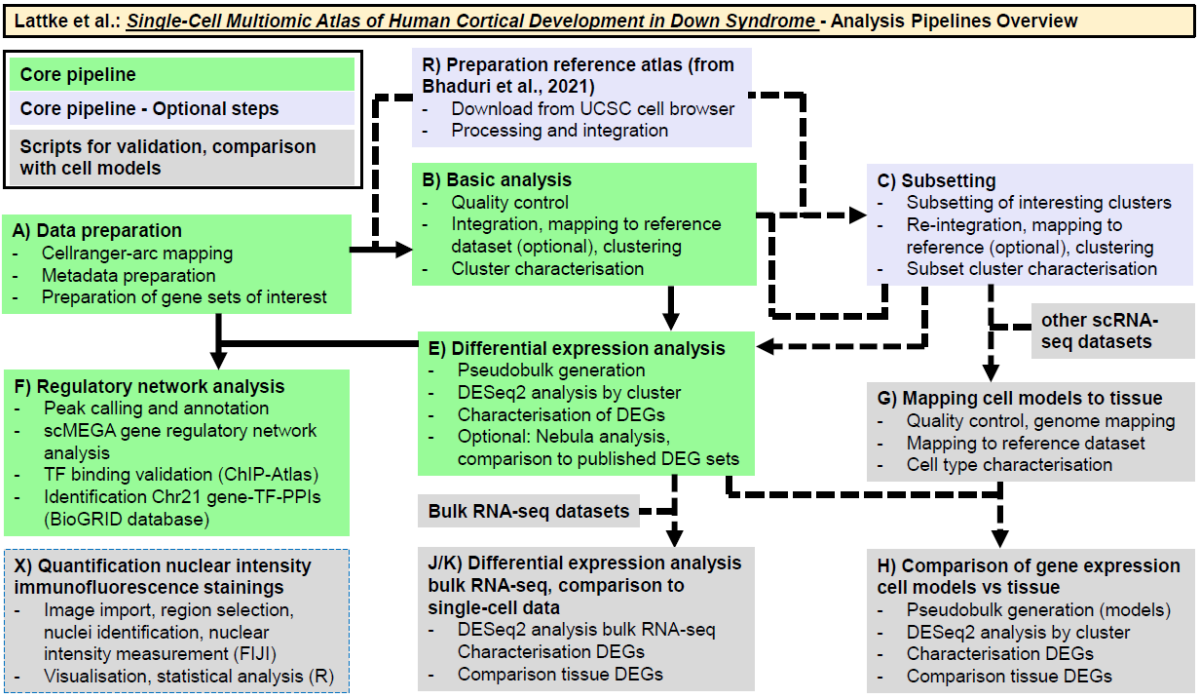
Lattke et al.: Single-Cell Multiomic Atlas of Human Cortical Development in Down Syndrome – Documentation Analysis Pipelines

These pipelines have been used to analyse our 10X Genomics single cell multiome dataset from human brain tissue of Down syndrome and disomic control fetuses (Lattke et al., XYZ). The pipelines include basic processing and clustering, mapping to reference atlases, differential gene expression, and gene regulatory network analyses integrating chromatin accessibility data to identify upstream regulators of deregulated transcriptional programmes. Also included are optional/auxiliary scripts for comparing bulk and scRNA-seq datasets from human cell models with the tissue dataset, and a pipeline for quantifying nuclear fluorescence intensity from immunostainings, used to validate the reduction of a neuronal population expressing the transcription factor FOXP1, and the deregulation of this factor, in DS fetal brain tissue.

Content:

1) Overview



2) General pipeline setup and running information

- Analyses were run in a conda environment on a high-performance computing cluster, either as batch submissions or interactively in R-Studio (most analyses would work on a good laptop with 64GB RAM)
- Analyses are run in one main directory (specified in the script heads), containing a folder with information on datasets and required gene sets ("A_input"), and script directories B-K (provided here)
- Script outputs will be directed to related output subdirectories B-K, files are labelled with same prefix as script (B01_, B02_, …, C01_, …)
- Download of the reference snRNA-seq atlas and immunofluorescence quantification (scripts in folders R, X) were run separately

Key packages/software tools used

Count matrices for snMultiome data were generated using cellranger-arc (v2.0.2; 10X Genomics). Sequencing data were analysed in an R environment (R v4.3.3), using the Seurat single cell analysis package (v5.1.0), with the Signac extension (v1.13.0) for basic analyses. sccomp (v1.7.15) and MiloR (v1.10.0) were used for compositional analyses. ESeq2 v1.42.1 and Nebula (v1.5.3) were used for differential gene expression analyses. limma (v3.58.1) was used for batch correction of expression matrices. clusterProfiler (v4.10.156) with annotations from DOSE(v 3.28.2) and org.Hs.eg.db (v3.18.0) was used for functional enrichment analyses. BSgenome.Hsapiens.UCSC.hg38 (v1.4.5) and EnsDb.Hsapiens.v86 (v2.99.0) were used for ATAC-seq mapping and annotations. scMEGA (v1.0.2) was used for gene regulatory network analyses, using binding motifs from the JASPAR2024 package (v0.99.6). Protein-protein-interactions from the BioGRID database (v4.4.233) were retrieved using the BioGRID API via the R packages jsonlite (v1.8.8) and httr (v1.4.7). tidyverse (v2.0.0) with ggplot (v3.5.1), pheatmap(v1.0.12) and the ggraph package (v2.2.1) were used for general data analyses and visualisations. Full environment specification see Packages_installed_250801.csv.

FIJI/ImageJ v1.53q was used for processing and analysing immunofluorescence images.

3) Core pipeline data preparation details

A) Data preparation

- **Mapping of 10X genomics single cell multiome data** to human genome with cellranger-arc
- **Mapping of Singleron** CeleScope scopeV3.0.1 (kit V2) snRNA-seq data (grafts, for step G, H, J) technology (samples see Extended Data Table 6) to human genome using the CeleScope snRNA-seq mapping software version 2.0.7 (here: mapping and generation of count matrices by Singleron).
- **Mapping of bulk RNA-seq data** (for steps H, J) to human genome (GRCh38) using the pipelines nf-core/rnaseq (v.3.18.0; doi:10.5281/zenodo.1400710) or AccuraCode (v1.2.0; for Singleron data).
- **Preparation of main analysis folder** containing input information in "A_input", and script directories B-K (script output)
- **Specification of input datasets** with file path to cellranger output and metadata and other count matrices in **"A_input" folder** (used files provided as examples, paths need to be adapted to run on different system)
    - Tissue datasets used in main analysis (steps B-J)
        - group_tab_tissue.csv: all generated tissue datasets
        - group_tab_tissue_bulk.csv: tissue samples used for complementary bulk-RNA-seq analysis (step J)
    - Cell model datasets (used in steps G, H, J)
        - group_tab_grafts_DS2U_DS1_SCZ.csv: datasets from human iPSC-derived neural grafts
        - group_tab_in_vitro_bulk_v050.csv: datasets from basic characterisation of human iPSC-derived neural cells in vitro
        - group_tab_in_vitro_bulk_ASO_exp.csv: datasets from transcription factor knockdown in iPSC-derived neural cells in vitro with antisense-oligonucleotides
- **Preparation of gene sets of interest** (in **"A_input" folder; files provided)**
    - Established cell type markers (cell_type_markers_240806_consolidated.csv)
    - Transcription factors from the FANTOM5 database (Transcription Factors hg19 - Fantom5_21-12-21.csv)
    - Protein coding Chr21 genes from Ensembl (HSA21_genes_biomaRt_conversion.csv)
    - Genes with mutations linked to intellectual disability syndromes (Genomics_EnglandPanelApp_Intellectual_disability_v8.243.tsv)
    - Differentially expressed genes in Down syndrome from Rastogi et al., 2024 (DEGs_publ_data/ Rastogi24_TableS2_DEGs_Cor_genes.csv and Rastogi24_TableS2_DEGs_iNeu_genes)

R) Preparation of snRNA-seq reference atlas for human fetal brain from Bhaduri et al., 2021 (not essential for core pipeline)

- ***To map own data to this multi-region dataset to predict brain region, cell type, … by label transfer***

- **sub-folder:**
    - **R_prepare_Bhaduri21_ref/** (scripts)
    - **R_out_Seur5/** (output)

- Download of count matrix from UCSC cell browser (Script: **R01_v020_load_ref_seur_from_UCSC_cell_browser.R**)
- Processing as for main dataset (split in 3 steps):
    - Normalisation of dataset (**R02_v020_ref_QC_SCTransform.R**)
    - Integration of sample datasets (**R03_v020_ref_integrate_Harmony.R**)
    - Re-normalise dataset (**R04_v020_ref_integrate_Harmony_renormalise.R**)

4) Core pipeline analysis details step by step

B) Basic analysis

- ***Questions:***
  - *How good is the quality of the data?*
  - *What cell populations are captured?*
  - *What brain regions and cell types of the reference atlas from Bhaduri et al., 2021 are the samples and cells mapping to?*

- **sub-folder:**
  - **B_basic_analysis/**
  - **B_basic_analysis_scripts/**

- **Load cellranger data and extract QC measures**
  - Script: **B01_v041_load_from_cellranger_arc.R**
    - Loads cellranger output (specified in group_tab_tissue.csv) into Seurat
    - Extracts and plots cellranger and Seurat QC measures

- **Define QC criteria:**
  - Adapt filtering criteria in following script, if necessary.
    - Criteria used here:
      - keep only cells with nCount_ATAC > 100/< 25000, nCount_RNA > 500/< 30000, percent.mt < 2, nucleosome_signal < 2, TSS.enrichment > 1.5
      - remove samples with > 50% removed low quality cells or <500 remaining cells

- **QC-filter and integrate samples**
  - Script **B02_v040_integrate_samples_RNA_Harmony.R**:
    - Removes low quality samples and cells (cut-off settings see above/script)
    - Integrates samples using Harmony, based on SCTransform-normalised RNA

- **Characterise samples and cell populations**
  - Script **B03_v040_integrated_dataset_clustering_tests.R:**
    - UMAP dimension reduction, tests clustering with different resolutions
    - Generates UMAP plots coloured by group, sample, developmental stage, and clusters at different clustering resolutions
    - Generates marker gene expression plots (cell type/subtype markers in cell_type_markers_240806_consolidated.csv as UMAP plot and dotplots for different clustering resolutions)
    - Generates table template for annotating clusters based on marker expression (B03_cluster_assignment.csv)
  - Script **B03a_v042_map_to_ref_Harmony.R (optional)**:
    - Specify reference dataset
    - Map dataset to reference atlas to predict cell types, regional identity of samples (via label transfer => if necessary, adapt labels to transfer in script)
    - Visualisations: reference UMAP split by dataset, sample; transferred labels projected on own dataset UMAP; Fraction of cells per sample mapped to each reference region (here used to identify non-cortical samples)
  - **Define final clustering resolution and annotate clusters**
    - Choose and adapt clustering resolution in following script
    - define cluster_name, cell_type, cell_class in **B03_cluster_assignment.csv,** order as preferred for plotting

        - adapt **B02_gr_tab_filtered.csv** to delete samples with wrong regional identity => **B02_gr_tab_filtered_non_cx_excl.csv** (here provided in output folder)
    - Script **B04_v040_charact_clusters.R (here not used for further analyses, rather proceeded with subsetting (C) to remove non-cortical samples):**
        - Performs clustering at specified resolution
        - Annotates clusters based on B03_cluster_assignment.csv
        - Creates UMAP and expression plots labelled with cluster_name
        - Performs differential abundance analysis with sccomp (Mangolia et al., 2023)

C: Subsetting of cortical samples, excitatory lineage for whole dataset or early/late samples (PCW11-13 or PCW16-20) (after B)

- **Questions:**
    - *Which cellular changes occur in dataset in DS vs CON (excluding non-cortical samples)?*
    - *Which cellular changes occur in the excitatory lineage at different stages?*

- **sub-folder:**
    - **Clean analysis complete dataset after excluding non-cortical samples:**
        - C_subsetting_all_cells_non_cx_excl_scripts/
        - C_subsetting_all_cells_non_cx_excl/
    - **Analysis of excitatory lineage (all stages PCW10-20; from C_subsetting_all_cells_non_cx_excl/**)
        - C_subsetting_exc_lin_from_all_non_cx_excl_scripts/
        - C_subsetting_exc_lin_from_all_non_cx_excl/
    - **Analysis of early excitatory lineage (PCW11-13)**
        - C_subsetting_exc_lin_from_all_non_cx_excl_PCW11_13_scripts/
        - C_subsetting_exc_lin_from_all_non_cx_excl_PCW11_13/
    - **Analysis of late excitatory lineage (PCW16-20)**
        - C_subsetting_exc_lin_from_all_non_cx_excl_PCW16_20_scripts/
        - C_subsetting_exc_lin_from_all_non_cx_excl_PCW16_20/

- **Delete samples not to be used for subsetting in copy of B02_gr_tab_filtered.csv (here non-cortical samples)**
- **For subsetting cell populations, delete clusters to exclude in copy of C_subsetting_all_cells_non_cx_excl/ C02_subset_cluster_assignment.csv)**

- **Subsetting and re-integration**
    - Script: **C01_v040_subsetting_reintegration_from_subset.R**
        - Subsets dataset
        - Performs (re-)integration as B02

- **Characterise subset cell populations**
    - Script **C02_v041_subset_clustering_tests.R**
        - Performs steps as B03
    - Script **C02a_v040_map_to_ref_Harmony.R (optional)**
        - Performs steps as B03a
    - **Define subset clustering resolution and cluster annotations**
        - Adapt following script
        - Define subset cluster annotations in **C02_subset_cluster_assignment.csv**
    - Script **C03_v041_subcluster_charact_abund_analysis.R**
        - Performs steps as B04

- Script **C04_v042_charact_clusters_MiloR_abundance_analysis.R (optional; only performed for complete dataset)**
  - Alternative (cluster-free) abundance analysis with MiloR (Dann et al., 2022) to validate sccomp differential abundance analysis
  - Very computationally intense!
  - Customised visualisations (original ones don't work): significance and fold change by neighborhood on UMAP plot (-log10(FDR), log2FC)

E: Differential gene expression analysis all cells/excitatory lineage (different stages) (after C)

- **Questions:**
  - *What genes are altered in DS?*
  - *Which cell populations are most affected by DS?*
  - *What are the functions of the deregulated genes? (Gene ontology analysis)*
  - *Are the deregulated genes enriched for intellectual disability-associated genes (as functional link to key DS phenotype)?*
  - *How do different differential expression analysis tools compare (DESeq2 pseudobulk analysis vs Nebula single-cell-based mixed-models analysis)?*
  - *Are differentially expressed genes found in published datasets (bulk RNA-seq of adult cortex and iPSC-derived neurons from Rastogi et al., 2024)?*

- **sub-folders:**
  - **Complete dataset with non-cortical samples excluded:** E_DESeq_pseudobulk_by_cluster_all_cells_non_cx_excl_v045_scripts/
  - **Subset excitatory lineage all stages:**
    - E_DESeq_pseudobulk_by_cluster_exc_lin_from_all_non_cx_excl_v045_scripts/
    - E_DESeq_pseudobulk_by_cluster_exc_lin_from_all_non_cx_excl_v045 /
  - **Subsets excitatory lineage stages PCW11-13 or PCW16-20:**
    - E_DESeq_pseudobulk_by_cluster_exc_lin_from_all_non_cx_excl_PCW11_13_v045_scripts/
    - E_DESeq_pseudobulk_by_cluster_exc_lin_from_all_non_cx_excl_PCW16_20_v045_scripts/

- **Pseudobulk generation**
  - Script **E01_v045_seur_pseudobulk_generation_min_30cells_per_pb.R**
    - create pseudobulks by sample and cluster (only keep pseudobulks with >30 cells)

- **DESeq2 analysis**
  - **E02_v045_pseudobulk_DESeq2_Wald_test_by_cluster.R**
    - Keeps only clusters with at least 2 pseudobulks per group
    - DESeq2 pseudobulk analysis (comparison of groups for each cluster with Wald-test; design = ~cluster_group)
    - Extracts vst-normalised expression matrix by pseudobulk (by sample for each cluster cluster) and gene z-score based on vst matrix
    - Extracts DEGs for each cluster (threshold padj <= 0.10, abs(log2FC) >log2(1.2)), differential TFs and Chr21 genes, quantifies DEG numbers
      - plots number of DEGs (incl Chr21 genes) per cluster

- **Characterisation of differentially expressed genes**
  - **E03_v048_DEG_characterisation_GO_combined_heatmaps_w_sign_ID_gene_enr.R**
    - **In script head:** specify cell clusters for plotting gene expression by cluster_sample

- Performs gene ontology analysis for combined DEGs
- Plots heatmaps of number of genes up/down for each GO and cluster (Top20 GO-terms)
- Calculates enrichment of ID-associated genes among DEGs
- Calculates relative gene expression matrices by cell clusters for plotting (gene x cluster; based on vst-normalised expression)
  - mean Z-score per cluster for CON and DS samples
  - relative change DS vs CON (DELTA = difference mean mean Z(DS)) vs mean Z(CON) for each cluster
  - adjusted p-values for each gene and cluster comparison for plotting
- plots DEG heatmaps
  - function based on pheatmap, with option for colored side bars for metadata, and asterisks for adjusted p-value <0.1
  - plots z-scores by cluster and sample for selected clusters (gene x cluster_sample)
  - plots rel. mean expression in CON (gene x cluster, color by mean Z(CON)) and expression changes in DS vs CON by cluster (gene x cluster, color by mean Z (DELTA)) (with asterisk for comparisons with padj <0.1)
  - Gene sets plotted:
    - all DEGs combined
    - cluster DEGs by cluster
    - differentially expressed TFs
    - differentially expressed Chr21 genes
    - enriched DEGs by GO term (Top20 GO-terms)
    - ID-associated genes

- **E04_v048_DEG_characterisation_plot_selected_GO_terms.R (optional; only run for all excitatory lineage cells comparison)**
  - Plots for selected enriched GO terms (specified in E_DESeq_pseudobulk_by_cluster_exc_lin_from_all_non_cx_excl_v045/ E03_GO_results_vs_N_DEGs_by_cluster_selected.csv; provided in output folder)
  - Plots as for E03
- **E01a_v046_Nebula_analysis_by_cluster.R (optional; only run for the all excitatory lineage cells subset):**
  - Nebula analysis as alternative DEG analysis approach for validation, using cell-based mixed-models analysis
  - Runs Nebula for cells of each cluster separately
- **E02a_v047_Nebula_analysis_DEG_characterisation.R (optional; only run for all excitatory lineage cells comparison):**
  - Characterisation of Nebula DEGs as in E03 for DESeq2 DEGs
- **E03a_v048_DEG_overlap_DEGs_other_analyses_datasets_Rastogi24_RNA_volcano_plots.R (optional; only run for all excitatory lineage cells comparison):**
  - Overlap of DEGs with DESeq2 and Nebula
  - Overlap of DEGs from DESeq2 and Nebula with published bulk RNA-seq DEGs (from Rastogi et al.)
  - Plot overlap matrices (Fraction and number of DEGs reproduced by other analysis/dataset)
  - Plot volcano plots for DESeq2 DEGs by cluster

F: GRN analyses with scMEGA (Li et al., 2023; after E)

- ***Questions:***
  - *What putative cis-regulatory elements are dynamically accessible and may contribute to expression changes of DS DEGs?*
  - *What TFs may regulate the observed changes in gene expression programmes?*
  - *Which of these changes might contribute to DS-associated expression changes?*
  - *Which are functionally relevant targets of these TFs?*
  - *How could Chr21 TFs regulate these networks?*
  - *Is there experimental evidence supporting the regulatory predictions?*
  - *How could other Chr21 genes regulate these TFs via PPIs?*

- **Sub-folders:**
  - **F_Chromatin_scMEGA_GRN_analysis_exc_lin_from_all_non_cx_excl_v045_scripts/**
  - **F_Chromatin_scMEGA_GRN_analysis_exc_lin_from_all_non_cx_excl_PCW11_13_v045_scripts/**
  - **F_Chromatin_scMEGA_GRN_analysis_exc_lin_from_all_non_cx_excl_PCW16_20_v045_scripts/**

- **Call ATAC peaks by cluster**
  - **F01_v045_seur_call_quant_peaks_by_cluster.R**
    - calls ATAC peaks for each cluster individually (allows to call peaks found only in small subpopulations)
-
- **Preprocess dataset, map TF motifs to peaks, quantify TF activity**
  - **F02_v045_preprocess_run_ArchR_ChromVar.R**
    - Order cells along pseudotime using AddTrajectory() and pre-defined order from C02_subset_cluster_assignment.csv (remove clusters not part of core excitatory lieage
    - Map motifs to peaks (using JASPAR24 database)
    - Run ChromVar to estimate TF activity from genome-wide accessibility of motifs

- **Select genes for network analysis and extract gene regulatory network**
  - **F03_v046_run_scMEGA_GRN_analysis.R**
    - Manually select TFs for network analysis with approach adapted from SelectTFs() function
      - Default function and cut-off removes TFs with repressive interactions => include repressive interactions
      - function or follow-up analyses cannot deal with TFs without change along pseudotime => manually remove these TFs
    - Select target genes for network analysis (all DEGs; SCT normalised expression)
    - Calculate gene-TF correlation along pseudotime
    - Extract potential direct interactions (for each gene, identify TFs with activity correlating with gene expression and motifs in putative regulatory regions)

- **Visualise Network, identify regulators potentially determining DS-associated changes**
  - **F04_v048_GRN_analysis_visualisation_filter_w_disease_linked_Chr21_TF_targets_sel_GO.R**
    - Extract network interactions for plotting and downstream analyses
    - Calculate mean expression in CON as node size for network plots (vst normalised expression of all control cells from pseudobulk analysis)
    - Calculate relative expression in DS vs CON as node colour for network plots (Z_DELTA = difference mean(Z-score DS) vs mean(Z-score CON) of all cells)
    - Filter only TF-target interactions consistent with role in deregulation in DS:

        - Positive correlation along pseudotime => TF and target both up in DS, or both down
        - Negative correlation along pseudotime => TF up and target down in DS, or vice versa
    - Plot overlap number of predicted TF targets in enriched GO terms by TF vs enriched GO terms (heatmap) => TF functions
    - Generate network plots for all interactions/filtered DS-relevant interactions:
        - all genes
        - all TFs
        - Chr21 TF to direct target gene/direct target TF interactions
        - Chr21 TFs to direct targets linked to ID (and calculate enrichment of ID-linked TF targets over all expressed genes)
        - Plot with CON expression as node size, expression DS vs CON as node colour, strength of correlation as edge thickness (negative: dashed line)

- **Validate TF-regulatory-element-interactions with ChIP-Atlas (optional; only run for all excitatory lineage cells comparison)**
    - **F05_v046_GRN_interactions_vs_ChIP_atlas_TF_binding_validation.R**
        - Checks availability of human cell datasets for network TFs in ChIP-Atlas database (CTCF excluded: huge dataset, probably not critical as TF)
        - Quite large files, take long time to load from web => check whether downloaded version is already available from previous analyses, download only missing TFs (bed file with ChIP peaks from all human cell datasets in ChIP-Atlas database)
        - save for use in analyses
        - For each TF, identify ATAC peaks overlapping with ChIP peaks from ChIP-Atlas
        - Quantify fraction of ATAC peaks predicted to regulate TF targets vs peaks linked to non-target genes, save overlapping peaks as validated interactions
        - Plot fraction of overlapping peaks, and odds ratio and significance of enrichment of targets vs non-targets for each TF
        - Plot regulatory networks only using ChIP-validated interactions (as F04)

- **Identify protein-protein-interactions of Chr21 genes with TFs that may affect TF activity using BioGRID database (optional; only run for all excitatory lineage cells subset)**
    - **F06_v046_PPI_analysis_core_regulators_with_expr_cor.R** (does not require running F05)
        - Trajectory correlation approach as in scMEGA analysis for PPI interactions from BioGRID
            - Calculate Z-score of TF activity DS vs CON
            - Get all Chr21-X-TF interactions from BioGRID
            - Keep interactions where Chr21 expression correlates with TF activity along pseudotime, and consistent with regulatory role in DS vs CON (see F04)
        - Plot network plots (see F04):
            - Chr21-TF interactions (without plotting intermediate interactors) with CON expression as node size, expression DS vs CON as node border colour, TF activity as node fill colour, and strength of correlation as edge thickness (negative: dashed line)
        - Chr21-TF interactions by TF (with intermediate interactors) with CON expression as node size, expression DS vs CON as node border colour, TF activity as node fill colour

5) Scripts for comparison of fetal tissue with cell models

G: Mapping cell models to tissue (xenograft snRNA-seq datasets; after C)

- ***Questions:***
  - *What fetal cell types do iPSC-derived graft cells correspond to?*
  - *Do they recapitulate cell proportion changes between CON and DS in fetal tissue?*

- **sub-folders:**
  - **G_basic_analysis_grafts_map_to_tissue_v05_10X_Singl_DS2U_DS1_scripts/**
  - **G_basic_analysis_grafts_map_to_tissue_v05_10X_Singl_DS2U_DS1_sc/**

- **Load datasets into Seurat and extract QC measures**
  - **G01_v050_load_from_cellranger_arc_Singleron.R**
    - *merge and QC datasets (similar to B01)*
    - can load count data from both 10X-Multiome and Singleron snRNA-seq data (specified in gr_tab$seq_tech as "Multiome_10X" or “snRNA_Singleron”)

- **Map datasets to reference dataset (10X multiome dataset DS fetal cortex)**
  - **G02_v051_map_to_ref_Harmony.R**
    - *Map to reference dataset and transfer cluster labels (similar to B03a/C02a)*

- **Re-map dataset subset to reference dataset subset (excitatory lineage DS fetal cortex)**
  - **G03_v051_map_to_ref_Harmony_remap_to_subset.R**
    - *Map to reference dataset and transfer cluster labels (similar to B03a/C02a)*
    - Keep only cells mapping to reference clusters retained in reference subset
    - Map to re-clustered reference subset clusters

- **Visualise and quantify mapping to reference subset clusters**
  - **G04_v050_charact_clusters_ref_mapping.R**
    - Visualisation of mapping of cells on UMAP of reference dataset (*similar to B03a/C02a)*
    - Quantification of predicted cluster composition in DS vs CON with visualisation and statistical analysis with sccomp (*similar to B04/C03*)

H: Differential gene expression analysis xenograft snRNA-seq and comparison with fetal tissue populations (after F, G, J (basic bulk RNA-seq characterisation in vitro models))

- **Questions:**
  - *How similar are xenograft populations to cells in vitro and fetal tissue reference populations?*
  - *What genes are altered in DS xenografts?*
  - *How similar are changes of relevant gene sets in DS vs CON between fetal tissue and xenografts/cells in vitro?*

- **sub-folder:**
  - **H_expression_analysis_grafts_vs_tissue_v05_10X_Singl_DS2U_DS1_scripts/**

- **Pseudobulk generation**
  - **H01_v050_seur_pseudobulk_generation_min_30cells_per_pb_by_pred_cluster.R**
    - create pseudobulks by sample and predicted cluster (*similar to E01*)

- **DESeq2 analysis**
  - **H02_v050_pseudobulk_DESeq2_LRT_by_pred_cluster_correct_covars.R**
    - *Similar to E02, but using LRT test on individual clusters instead of Wald test, to allow correction for covariates/batch effects (here: sequencing technology)*
    - Removes batch effects from expression matrix and z-scores (vst-normalised) with limma::removeBatchEffect

- **Characterisation of differentially expressed genes**
  - **H03_v050_DEG_characterisation_GO_comb_by_pred_cluster_correct_covars.R**
    - *GO analysis and visualisations similar to E03*
    - Includes PCA plots by cluster to identify covariates/batch effects
    - Heatmaps of selected gene sets with batch-corrected expression values

- **Comparison of gene expression with fetal tissue reference**
  - **H04_v052_DEG_overlap_tissue_grafts_by_pred_cluster_correct_covars.R**
    - *Adapted from to H03*
    - Adapt in head of script:
      - Merge pseudobulk datasets from fetal tissue reference and xenografts, include bulk RNA-seq data from in vitro experiments
      - Select clusters for plotting of expression heatmaps by cluster_sample
    - Re-normalises and corrects expression values (from merged count tables; here: correct for sequencing technology)
    - Plots correlation of vst-norm expression of top variable genes of all fetal and graft clusters/in vitro experiments to identify most similar populations, including most similar of fetal vs graft vs in vitro populations
    - Plots correlation log2FC tissue DEGs between populations to quantify similarity of DS-related changes between fetal tissue and grafts/cells in vitro
    - Plots heatmap of overlaps between tissue DEGs and graft/in vitro DEGs (number and fraction of tissue DEGs by cluster also detected in graft/in vitro comparisons)
    - Heatmaps of expression of selected gene sets with batch-corrected expression values (including graft and tissue populations and bulk expression from in vitro experiments)

J/K: Differential gene expression analysis bulk-RNA-seq data (tissue and in vitro experiments) and comparison with fetal tissue 10X-Multiome dataset (after F)

- **Questions:**
  - *What changes detected in the 10X-Multiome analysis in different populations are also detected in bulk RNA-seq from tissue? Does bulk RNA-seq capture additional DEGs missed by the 10X-Multiome analysis (e.g. non poly-adenylated RNAs)?*
  - *How transcriptionally similar are cells in vitro and fetal tissue reference populations?*
  - *What genes are altered in DS vs CON cells in vitro?*
  - *How do relevant gene sets change in DS vs CON in fetal tissue vs cells in vitro?*
  - *Does antisense-mediated Chr21 TF knockdown revert changes of predicted Chr21 TF targets in DS cells? (K)*

- **sub-folder:**
  - **Comparison 10X-Multiome vs bulk RNA-seq fetal tissue:**
    - J_expression_analysis_tissue_bulk_vs_exc_lin_v050_scripts/
  - **Comparison 10X-Multiome vs basic bulk RNA-seq characterisation cells in vitro:**
    - J_expression_analysis_in_vitro_vs_exc_lin_v050_scripts/
  - **Comparison expression predicted Chr21 targets in DS vs CON and DS + TF knockdown**

    - K_expression_analysis_in_vitro_ASO_exp_v050_scripts/

- **Comparison 10X-Multiome vs bulk RNA-seq fetal tissue**
  (J_expression_analysis_tissue_bulk_vs_exc_lin_v050_scripts/)
  - **J01_v050_DESeq2_by_group_char_overlap_Multiome.R:**
    - DESeq2 analysis of DS vs CON for bulk dataset (Wald test)
    - Plots number of DEGs, PCA plot to identify outliers
    - Volcano plots with selected gene sets highlighted
    - Plots heatmap overlap number/fraction of 10X-Multiome DEGs also detected in bulk
    - Plots expression heatmaps (bulk data by sample): 10X-Multiome DEGs, Bulk DEGs not in Multiome DEGs
  - **J01a_v050_DESeq2_by_group_char_overlap_Multiome_remove_outliers.R:**
    - *as J01, individual outlier sample removed*

- **Comparison 10X-Multiome vs bulk RNA-seq fetal tissue**
  (J_expression_analysis_in_vitro_vs_exc_lin_v050_scripts/)
  - **J01_v050_DESeq2_by_group_by_isog_pair_batch.R:**
    - DESeq2 analysis of DS vs CON for bulk dataset (Wald test; contrasts by individual batch/experiment), Plots number of DEGs
  - **J02_v051_DEG_characterisation_overlap_tissue_TF_targets.R:**
    - Define in script head: fetal tissue clusters to plot in expression heatmaps by sample/cluster_sample for comparison with bulk samples
    - Plots heatmap overlap number/fraction of 10X-Multiome DEGs also detected in vitro
    - Plots correlation of vst-norm expression of top variable genes of all fetal and graft/in vitro clusters to identify most similar fetal and graft populations
    - Plots correlation log2FC tissue DEGs between populations to quantify similarity of DS-related changes between fetal tissue and grafts/cells in vitro
    - Plots heatmaps of expression of selected gene sets with batch-corrected expression values (including tissue populations and bulk expression from in vitro experiments)
      - E.g. all tissue DEGs, Chr21 genes, TFs, tissue DEGs linked to enriched GO terms, predicted TF targets, …)
      - Similar heatmaps with genes with concordant changes in vitro vs fetal only

- **Comparison expression predicted Chr21 targets in DS vs CON and DS + TF knockdown**
  (K_expression_analysis_in_vitro_ASO_exp_v050_scripts)
  - **K01_v055_DESeq2_sel_comps_batch_corr_CON_comb.R:**
    - *Similar to H02*
    - Adapt in script head:
      - Merge original groups for combined comparison (untreated vs non-targeting ASO treated; 2 designs for GABPA ASOs)
      - Define group comparisons (DS vs CON, DS + TF ASO vs DS untreated/non-targeting) and batch variable for batch correction
    - DESeq2 analysis of DS vs CON for bulk dataset (LRT test for each comparison, with correction for batch variable)
    - Extracts comparison results and normalised and corrected expression data
    - Plots number of DEGs
  - **K02_v055_DEG_characterisation_overlap_tissue_TF_targets_batch_corr_CON_comb.R**
    - Plots expression heatmaps for predicted TF targets (by sample and differences of group means DS vs CON and DS+TF-knockdown vs DS); Including plots with predicted targets differentially expressed in DS vs CON only

6) Scripts for quantification of nuclear intensity immunofluorescence stainings

- **Questions:**
  - *Can differences in proportions of cell populations identified in 10X-Multiome analysis (reduced RORB/FOXP1-expressing neurons) be confirmed by immunostaining?*
  - *Can differences in transcription factor expression identified in 10X-Multiome analysis (reduced FOXP1-expression) be confirmed by immunostaining?*

- **sub-folder:**
  - **X_FOXP1_staining_quantification/S00_scripts/**
  - **X_FOXP1_staining_quantification/S00_input/**

- **Save script folder in analysis directory**

- **Import microscopy images into FIJI/ImageJ pipeline**
  - Open all images in FIJI
  - **S01_v032_save_ndpi_greyscale_tif_batch_241003.ijm**
    - Run script, select main analysis directory
    - Saves images as multilayer greyscale TIFF files in analysis subfolder (/S01_greyscale_images_complete/)

- **Select regions of interest (ROIs) in images**
  - **S02_v032_crop_6_layer_bins_241004.ijm**
    - Run script, select main analysis directory
    - Runs for all images in folder from previous step (/S01_greyscale_images_complete/):
      - Allows to select multiple regions of interest in each image (use add to ROI manager or shortcut “t” to add regions to selection
      - Extracts regions of interest into new files and in new folder (/S02_layers_cropped/)
      - Also saves ROI coordinates and images with ROIs drawn into images

- **Identify nuclei (create DAPI masks)**
  - *Critical step for good results*
    - *Masks/segmentation not perfect, but usually acceptable approximation*
    - *Expansion into cytoplasmic space usually allows to also quantify cytoplasmic/perinuclear proteins*
    - *Comparing mask examples with the original images may be advisable to verify acceptable mask generation*
  - **S03_v032_create_mask_DAPI_Huang_tresh_size_filter_batch_241004.ijm**
    - Define in script DAPI channel index (default 1) and size threshold (default 10; check on example images)
    - Runs for all images in folder from previous step (/S02_layers_cropped/)
    - Automatically selects and segments DAPI positive areas
    - Saves DAPI masks into new folder (/S03_DAPI_masks/)

- **Optional step: Convert images for to RGB illustrations**
  - **‘S03a_Greyscale to RGB sel chan adj brightness batch_v022_250609.ijm’**
    - Define in script: output folder, channels to merge, colors in merged image, brightness thresholds
    - Run script => select folder containing images
    - Merges greyscale channels into RBG image for illustration

- **Measure mean intensity for all channels for all DAPI masks (nuclei)**

- **S04_v032_Nuclei_ROI_measure_batch_241004.ijm**
    - Adapt DAPI channel in script (default 1)
    - Measures mean intensity for all images, all channels, all DAPI masks
    - Saves measurements in /S04_FIJI_results_nucl_intens/

- **Import intensity measurements and image metadata into R**
    - **S05_v034_R_batch_input_250528_auto_add_metadata.R**
        - Specify image metadata in /S00_input/group_tab.csv (provided as example)
        - Define channel names in script line 33 (ordered according to channel order in images)
        - Saves dataset in /S05_R_input/

- **Visualise intensity distribution of nuclei for each channel by image/replicate/group**
    - Adapt image analysis metadata (mainly replicate, group) in S05_R_input/image_groups.csv
    - **S06_v035_intens distrib_1Ch_16bit_250805_bin_ROIs_merged.R**
        - Calculates normalised intensities:
            - normalised by mean DAPI intensity in image
            - quantile-normalised by sample: Norm = (Raw – 20% quantile)/(40% quantile - 10% quantile) (rationale: adjust intensity by background intensity offset and spread)
        - Plots intensity histograms (distribution of nuclei) for each channel by image/replicate/group (raw and normalised intensities)
        - Saves output/visualisations in /S06_intens_distrib_1Ch/

- **Test thresholds for positive/negative classification of cells**
    - **S07_v033_threshold_testing_1Ch_250528_FOXP1.R**
        - Define channel and vectors with test thresholds in script (l. 26-32) based on intensity histograms from S06 (for raw and normalised intensities; for selecting raw intensity thresholds also checking on example images can be helpful)
        - Saves and plots quantifications on image and sample level, split by group (into /S07_intens_thresh_analysis_1Ch/)

- **Create optimised plots and run statistics for selected threshold for positive/negative classification of cells**
    - **S08_v034_visualisation_stats_1Ch_250805_FOXP1.R**
        - barplots by group (raw intensity, threshold 10,000 A.U.)
        - group comparison with t-test
        - plots by developmental stage vs group