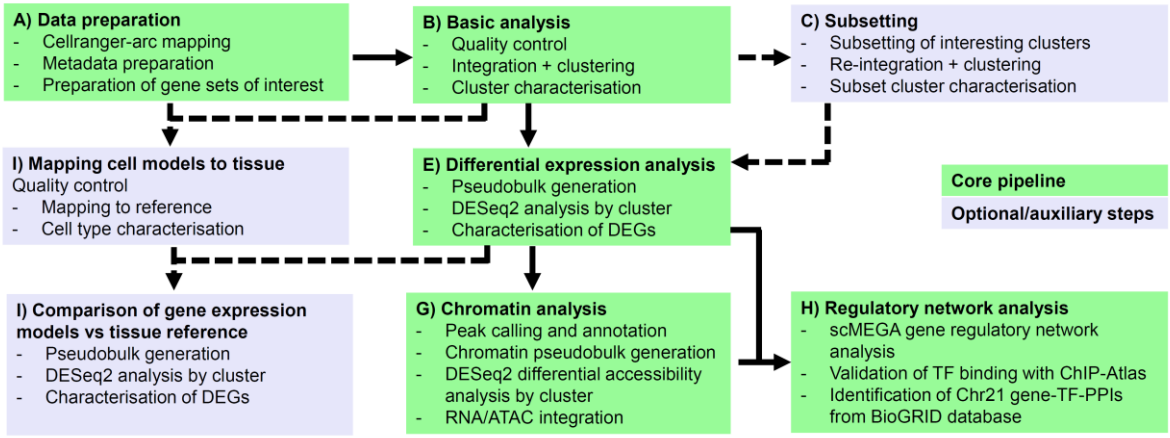
Lattke et al.: Single-Cell Multiomic Atlas of Human Cortical Development in Down Syndrome - Single cell multiome analysis pipeline

This pipeline has been used to analyse our 10X Genomics single cell multiome dataset from human brain tissue of Down syndrome and disomic control foetuses. It includes basic processing, differential gene expression and chromatin accessibility analyses, network analyses to identify upstream regulators of deregulated programmes, and a comparison with human cell models (Lattke et al., bioRxiv, 2024).

Overview



General 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-J (provided here)
- Script outputs will be directed to related output subdirectories B-J, files are labelled with same prefix as script (B01_, B02_, …, C01_, …)

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) was used for compositional analyses. DESeq2 (v1.42.1) was used for differential gene expression/chromatin accessibility analyses. clusterProfiler (v4.10.156) with annotations from DOSE(v 3.28.2), org.Hs.eg.db (v3.18.0) and msigdbr (v7.5.1) 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. HOMER (v5.0.1) with the genome annotation hg38(v6.4) was used for motif enrichment analyses in Fig. 4. 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 R_environment.txt.

Analysis details step by step

A) Data preparation

- **Mapping of 10X genomics single cell multiome data** with cellranger-arc
- **Preparation of main analysis folder** containing input information in "A_input", and script directories B-J (script output)
- **Specification of input datasets** with file path to cellranger output and metadata in **"A_input" folder**
    - Tissue datasets used in main analysis (steps B-H)
        - group_tab_tissue.csv: all generated tissue datasets
        - group_tab_tissue_consolidated.csv: only high-quality datasets retained for the main analysis
    - Cell model datasets (used in steps I and J)
        - group_tab_grafts.csv: datasets from human iPSC-derived neural grafts
        - count_table_neurons_in_vitro_bulk_Peter24.csv: bulk RNA-seq count matrix for bulk RNA-seq data of iPSC-derived neurons in 2D cultures, from related manuscript Peter, Real, et al., bioRxiv (2024).
        - group_tab_neurons_in_vitro_bulk_Peter24.csv: metadata for 2D neuron data above.
- **Preparation of gene sets of interest**
    - 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)

B) Basic analysis

- ***Questions:***
    - *How good is the quality of the data?*
    - *What cell populations are captured?*

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

- **Load cellranger data and extract QC measures**
    - Script: **B01_v030_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:**
    - high quality samples to retain (in **group_tab_tissue_consolidated.csv**)
    - Adapt filtering criteria in following script, if necessary.
    - Stringent filtering criteria used here:
        - keep only samples specified in group_tab_tissue_consolidated.csv
            - excluded datasets: samples with critically low cell number, median UMI/cell or ATAC fragments/cells (each <1000), datasets with better quality after repeated sequencing of sample
        - keep only cells with nCount_ATAC > 500/< 25000, nCount_RNA > 500/< 30000, percent.mt < 2, nucleosome_signal < 2, TSS.enrichment > 1.5

        - remove samples with > 30% removed low quality cells or <1000 remaining cells

- **QC-filter and integrate samples**
    - Script **B02_v030_integrate_samples_RNA_Harmony.R**:
        - Removes low quality samples (retains only samples in group_tab_tissue_consolidated.csv) and cells (cut-off settings see above/script)
        - Integrates samples using Harmony, based on SCTransform-normalised RNA

- **Characterise cell populations**
    - Script **B03_v030_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)
    - **Define final clustering resolution and annotate clusters**
        - adapt clustering resolution in following script
        - define cluster_name, cell_type, cell_class in **B03_cluster_assignment.csv,** order as preferred for plotting
    - Script **B04_v030_charact_clusters.R:**
        - 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)

<u>E: Pseudobulk differential gene expression analysis for whole dataset (after B)</u>

- **Questions:**
    - *What genes are altered in DS?*
    - *Which cell populations are most affected by DS?*
    - *What are the functions of the deregulated genes? What disorders are they associated with? (Gene ontology/human phenotype ontology overlap)*

- **sub-folder:**
    - **E_DESeq_pseudobulk_by_cluster_all/**
    - **E_DESeq_pseudobulk_by_cluster_all_scripts/**

- **Pseudobulk generation**
    - Script **E01_v030_seur_pseudobulk_generation_min_10cells_per_pb.R**
        - create pseudobulks by sample and cluster (only keep pseudobulks with >10 cells)

- **DESeq2 analysis**
    - **E02_v030_pseudobulk_DESeq2_Wald_test_by_cluster_form_only_cluster_group.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 DEGs for each cluster (threshold padj <= 0.05, abs(log2FC) >log2(1.2)), differential TFs and Chr21 genes, quantifies DEG numbers

- **Characterisation of differentially expressed genes**
    - **E03_v030_DEG_characterisation.R**
        - plots number of DEGs (incl Chr21 genes) per cluster
        - performs GO and MSigDB enrichment analysis of combined DEGs
        - plots heatmaps of number of genes up/down for each GO/MSigDB term and cluster
        - calculates relative gene for pseudobulks for(based on vst-normalised expression)
            - calculate mean Z-score per cluster for CON and DS samples separately
            - calculate relative change DS vs CON (DELTA = (mean Z(CON)-mean Z(DS)) for each cluster
        - plot DEG heatmaps by cluster and GO/MSigDB(HPO) term
            - rel. expression in CON (gene x cluster, color by mean Z(CON))
            - expression changes in DS vs CON (gene x cluster, color by mean Z (DELTA))
            - Number of DEGs by GO/MSigDB(HPO) x cluster

C: Subsetting of excitatory lineage PCW11/12 (after B)

- **Questions:**
    - *Which are the earliest changes occurring in the excitatory lineage?*

- **sub-folder:**
    - **C_subsetting_exc_lin_PCW11_12/**
    - **C_subsetting_exc_lin_PCW11_12_scripts/**

- **Delete clusters not to be used for subsetting in copy of B03_cluster_assignment.csv**

- **Subsetting and re-integration**
    - Script: **C01_v030_subsetting_reintegration.R**
        - Subsets dataset
        - Performs (re-)integration as B02

- **Characterise subset cell populations**
    - Script **C02_v030_subset_clustering_tests.R**
        - Performs steps as B03
    - **Define subset clustering resolution and cluster annotations**
        - Adapt following script
        - Define cluster annotations in **C02_subset_cluster_assignment.csv**
    - Script **C03_v030_subcluster_charact_abund_analysis**
        - Performs steps as B04

E: Pseudobulk differential gene expression analysis for excitatory lineage PCW11/12 (after C)

- **Questions:**
    - *What genes are first altered in DS in excitatory neurons?*
    - *What are the functions of the early deregulated genes? What disorders are they associated with? (Gene ontology/human phenotype ontology overlap)*

- **sub-folder:**
    - **E_DESeq_pseudobulk_by_cluster_exc_lin_PCW11_12/**
    - **E_DESeq_pseudobulk_by_cluster_exc_lin_PCW11_12_scripts/**

- **Scripts as for complete dataset (see above):**
    - E01_v030_seur_pseudobulk_generation.R
    - E02_v030_pseudobulk_DESeq2_Wald_test_by_cluster_form_only_cluster_group.R
    - E03_v030_DEG_characterisation.R

G: Chromatin analysis (after E)

- ***Questions:***
    - *What putative cis-regulatory elements are accessible at this stage?*
    - *Does chromatin accessibility change in DS?*
    - *Do chromatin changes underly changes in gene expression?*
    - *What TFs might control changes in gene expression?*

- **Sub-folder:**
    - **G_Chromatin_analysis_by_cluster_exc_lin_PCW11_12/**
    - **G_Chromatin_analysis_by_cluster_exc_lin_PCW11_12_scripts/**

- **Call ATAC peaks by cluster**
    - **G01_v030_seur_call_quant_peaks_by_cluster.R**
        - calls ATAC peaks for each cluster individually (allows to call peaks found only in small subpopulations)

- **Peak annotation and pseudobulk generation**
    - **G02_v030_peak_annot_pseudobulk_generation.R**
        - Annotate peaks with Ensembl annotations: classify in promoter/exon/intron/ intergenic, identify closest promoter as likely target gene
        - Generate pseudobulks as for RNA-seq data (see E)

- **DESeq2 Differential chromatin accessibility analysis**
    - **G03_v030_atac_pseudobulk_DESeq2_Wald_test_by_cluster.R**
        - Differential accessibility analysis with DESeq2 (as for RNA-seq, see E; exclude only peaks with <0.01 counts/cell)
        - save bed files with all increased and all decreased peaks (for HOMER analysis)

- **Run motif enrichment analysis with differential peaks (Homer):**
    - Command: findMotifsGenome.pl $BED hg38 $OUT -size given -mask -p 32

- **Integrate RNA and ATAC data, visualise ATAC data**
    - **G04_v030_peak_RNA_integr_visual.R**
        - Identify differential peaks associated with concordantly differentially expressed target genes
            - plot heatmap of DEG expression vs diff peak accessibility
        - Plot ATAC coverage of differential gene loci with associated differential peaks

H: GRN analyses with scMEGA (Li et al., 2023; after G)

- ***Questions:***
    - *What TFs regulate the observed changes in gene expression programmes?*
    - *What 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-folder:**
    - **H_scMEGA_GRN_analysis_exc_lin_PCW11_12/**
    - **H_scMEGA_GRN_analysis_exc_lin_PCW11_12_scripts/**

- **Preprocess dataset, map TF motifs to peaks, quantify TF activity**
    - **H01_v032_preprocess_get_motifs_run_ArchR.R**
        - Subsample dataset to max 100 cells per sample and cluster (to reduce computational demands; probably not be required)
        - Order cells along pseuodtime using AddTrajectory() and pre-defined order from B03_cluster_assignment_exc_lin.csv
        - 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**
    - **H02_v032_select_network_genes_run_GRN_analysis_select_all_TFs.R**
        - Select TFs for network analysis with SelectTFs(cor.cutoff = -1, p.cutoff = 1)
            - default cut-off removes TFs without change along pseudotime or repressive interactions)
        - Select target genes for network analysis (all DEGs; SCT normalised expression)
        - Calculate gene-TF correlation along pseudotime
        - Extract potential direct interactions (TFs correlating with targets and with motifs in putative regulatory regions of targets)

- **Visualise Network, identify critical regulators**
    - **H03_v036_GRN_analysis_visualisation_filter_DS_rel_int.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 = mean(Z-score DS) - 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 TF targets 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
            - 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**
    - **H04_v036_GRN_interactions_vs_ChIP_atlas_load_peak_sets.R**
        - Check availability of human cell datasets for network TFs in ChIP-Atlas database
        - 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

- **H05_v036_GRN_vs_ChIP_atlas_comp_targets_extract_ChIP_val_network.R**
  - 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 H03)

- **Identify protein-protein-interactions of Chr21 genes with TFs that may affect TF activity using BioGRID database**
  - **H06_v036_PPI_analysis_core_regulators_with_expr_cor.R** (does not require H04/05)
    - 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 H03)
    - Plot network plots (see H03):
      - 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

<u>I: Mapping of grafts to reference dataset (after E)</u>

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

- **sub-folders:**
  - **I_mapping_to_reference_grafts_to_tissue_scripts/**
  - **I_mapping_to_reference_grafts_to_tissue/**

- **Load datasets into Seurat and extract QC measures**
  - **I01_v030_load_merge_datasets_QC.R**
    - *merge and QC datasets (similar to B01)*

- **QC filtering and mapping cells to reference tissue atlas (from B)**
  - **I02_v030_filter_map_to_ref_Harmony.R**
    - filtering*:* nCount_RNA >500 & < 30000, percent.mt < 2
    - mapping to reference dataset with FindTransferAnchors(dims = 1:30, reference.reduction = "pca") and MapQuery(refdata = list(cluster_name = "cluster_name", cell_type = "cell_type", …), reference.reduction = "pca", reduction.model = "umap")

- **Characterise graft cell types**
    - **I03_v030_plot_and_quant_ref_mapping.R**
        - UMAP plots for grafts vs reference
        - quantification of cell proportions (similar to B04 and C03)

J: Gene expression analysis tissue vs grafts and iPSC-derived neurons (after I)

- **Questions:**
    - *Do iPSC derived neurons in vitro (bulk RNA-seq) and in grafts recapitulate gene expression in tissue?*
    - *Do gene expression changes in CON vs DS in grafts recapitulate changes in tissue?*

- **sub-folders:**
    - **J_expression_analysis_mapped_data_vs_reference_with_neurons_in_vitro_bulk_scripts/**
    - **J_expression_analysis_mapped_data_vs_reference_with_neurons_in_vitro_bulk/**

- **Generate pseudobulks**
    - **J01_v030_seur_pseudobulk_generation_min_10cells_per_pb.R**
        - pseudobulk generation for grafts (as for E01)

- **Combine model and reference tissue data and perform DESeq2 analysis**
    - **J02_v030_pseudobulk_DESeq2_Wald_test_combined_ref_mapped_with_in_vitro_bulk.R**
        - merge graft pseudobulks with tissue pseudobulks and bulk RNA-seq from neurons in 2D cultures
        - DESeq2 analysis based on cluster vs group split by sample_type (~comp_group = [cluster_name]_[sample_type]_[group])
            - not for in vitro neurons (not clear which tissue/graft cells to compare with)
        - similar to E02

- **Compare gene expression in models and tissue reference** (incl. changes DS vs CON)
    - **J03_v029_DEG_characterisation.R**
        - plot number of DEGs by cluster tissue vs grafts
        - calculate gene expression Z-scores (mean by cluster for CON, DS pseudobulks)
        - plot expression Z-scores (mean CON and mean DS-CON for each cluster in tissue/grafts) for selected gene sets (Chr21 genes DEGs, DEGs linked to GO terms and HPO terms in tissue analysis (E, whole dataset))