Welcome to the DKFZ!

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

Unite Single Cell Course
Day1

Christian Heyer
University of Augsburg & DKFZ Graduate School

Course Organization

- You will need
  - Laptop with internet access (eduroam)
- Denbi Cloud Account
  - If you haven‘t registered, please do this now (email instructions sent 03.04.2024)

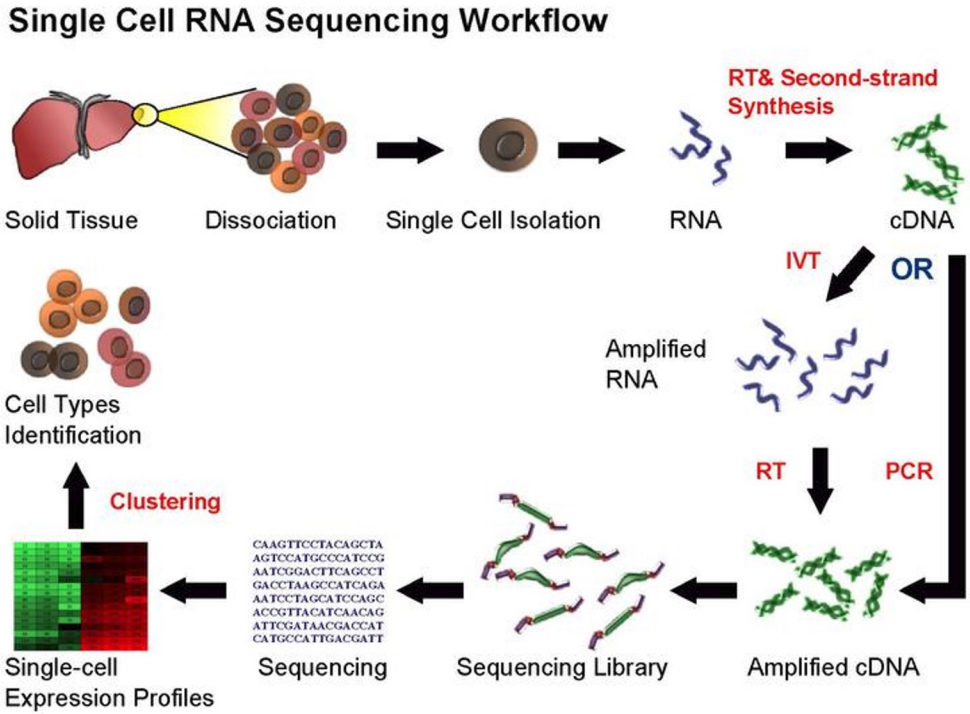
Overview

- What are your expectations?

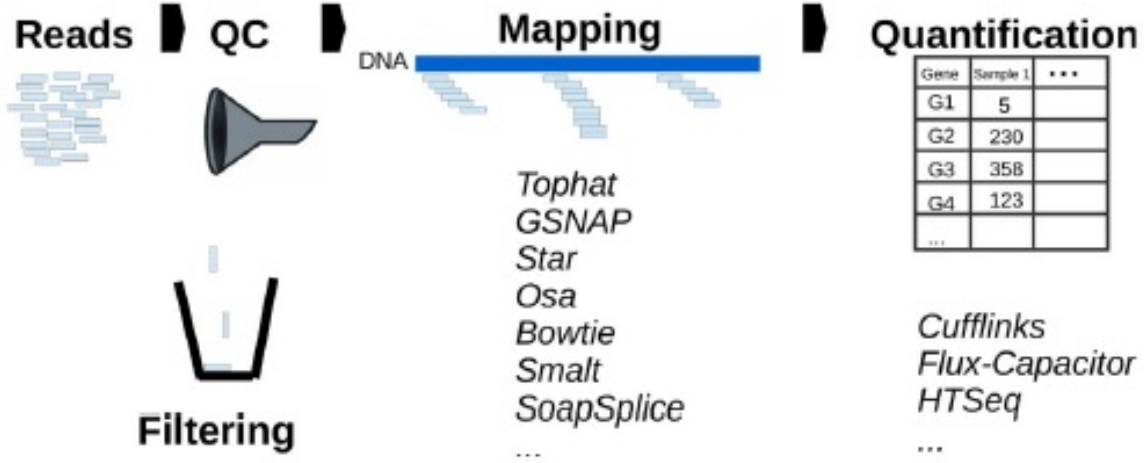
Course Overview

- Day1:
  - Getting 10x Data into R
  - Introduction to seurat Package
  - Quality Control
- Day2:
  - UMAP
  - Clustering
  - Cell Type Identification
- Day3 Multi Sample analysis:
  - Data Integration
  - Differential Expression between conditions.

Workflow Overview



Basic bioinformatics wokflow (cell Ranger)



- QC filtering
- Mapping: Alignment of reads to the genome/transcriptome
- Quantification: count reads per feature

Where do we start?

•	•SRR1039508	•SRR1039509	•SRR1039512	•SRR1039513	•SRR1039516	•SRR1039517	•SRR1039520	•SRR1039521
•ENSG00000000003	•679	•448	•873	•408	•1138	•1047	•770	•572
•ENSG00000000005	•0	•0	•0	•0	•0	•0	•0	•0
•ENSG00000000419	•467	•515	•621	•365	•587	•799	•417	•508
•ENSG00000000457	•260	•211	•263	•164	•245	•331	•233	•229
•ENSG00000000460	•60	•55	•40	•35	•78	•63	•76	•60
•ENSG00000000938	•0	•0	•2	•0	•1	•0	•0	•0
•ENSG00000000971	•3251	•3679	•6177	•4252	•6721	•11027	•5176	•7995
•ENSG00000001036	•1433	•1062	•1733	•881	•1424	•1439	•1359	•1109
•ENSG00000001084	•519	•380	•595	•493	•820	•714	•696	•704
•ENSG00000001167	•394	•236	•464	•175	•658	•584	•360	•269

Rstudio and the Denbi Cloud

- Denbi Cloud provides computational resources at no cost for academic bioinformatics projects
- Rstudio Server -> Server running Rstudio

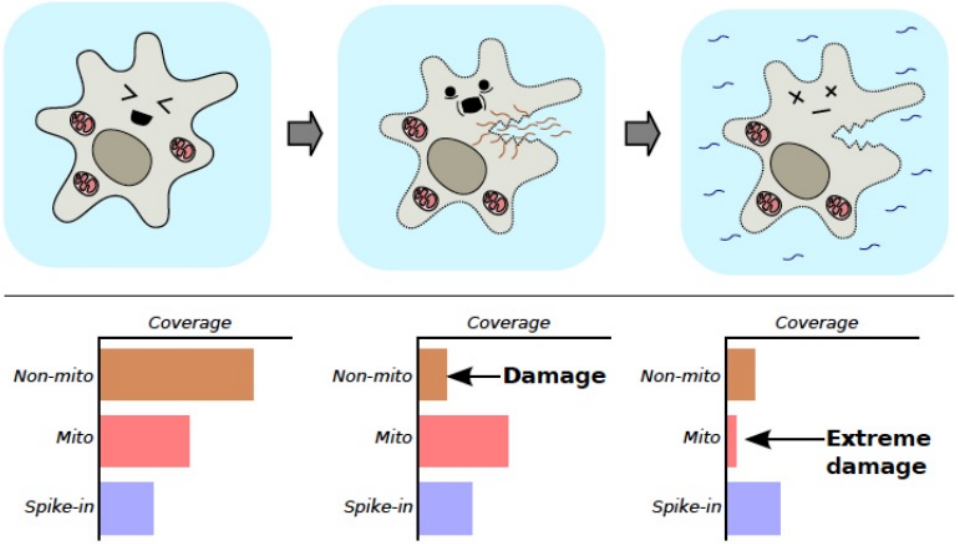
First steps

- Login
- Copy course scripts.
  - sc_course/scripts
  - Click on sc_course.
  - Click the checkmark besides scripts.
  - Click on the Gear (More) and Copy to. Copy to your home directory
- Navigate to the new copied directory and open the Day1 folder

Preprocessing single Cell RNA data

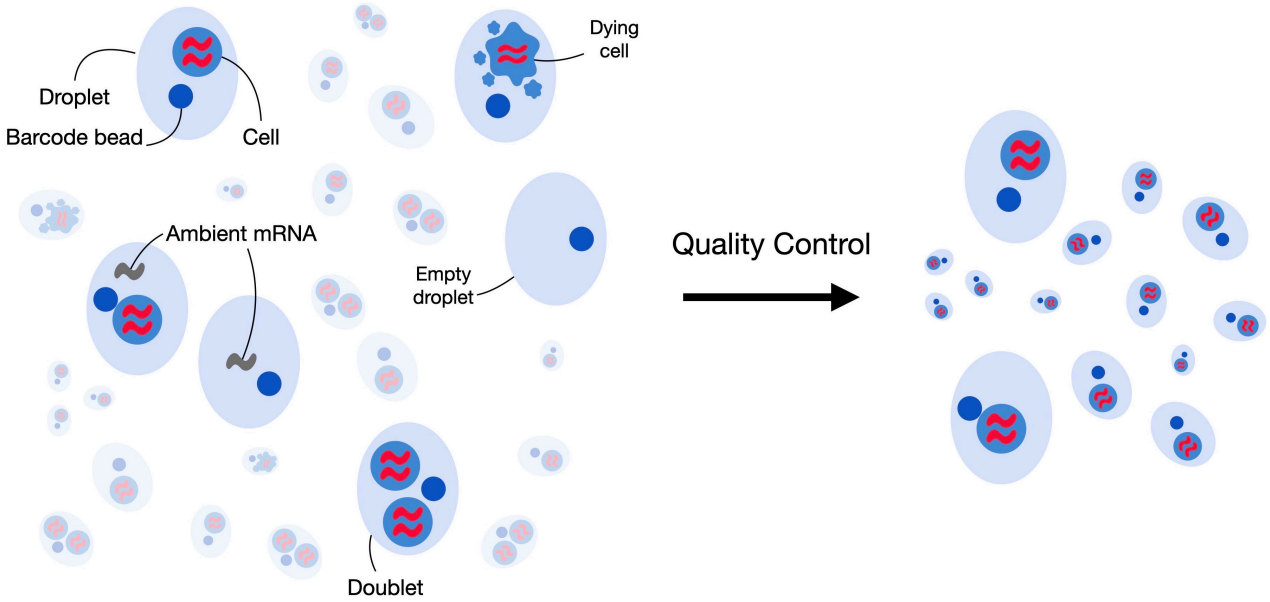
- Quality control
  - cell filtering
  - gene filtering
- normalization
- batch correction
- identify highly variably genes
- visualization through dimension reduction
- cell cycle scoring

Biological cell states to detect and filter out.



- RNA composition changes with cell state
- How can we identify these cells?

Quality control



https://www.sc-best-practices.org/_images/quality_control.jpeg

QC Goals

- **Filter outlier cells**
  - Damaged cells
  - In Droplet seq: Remove Doublet cells
  - Strategy -> Remove extreme outliers.
- **filter out uninformative genes**
  - Remove genes expressed in very few cells
  - What is the smallest cell cluster size of interest?

Fitering strategy

- **iterative strategy**
  - Start with permissive filtering steps
  - Only if issues occur downstream, make more stringent
- **Filter batches separately, if necessary**
  - If QC covariates differ bewteen samples -> separate QC
- **Many datasets are mixture of different cells with different properties**
  - Multivariate distributions
  - Be permissive in QC limits

Low dimensional data representation

- Curse of high dimensionality
- Depending on datasets, we have 10.000-20.000 variables
- How can we deal with this?
- Feature selection
- (Maybe UMAP)

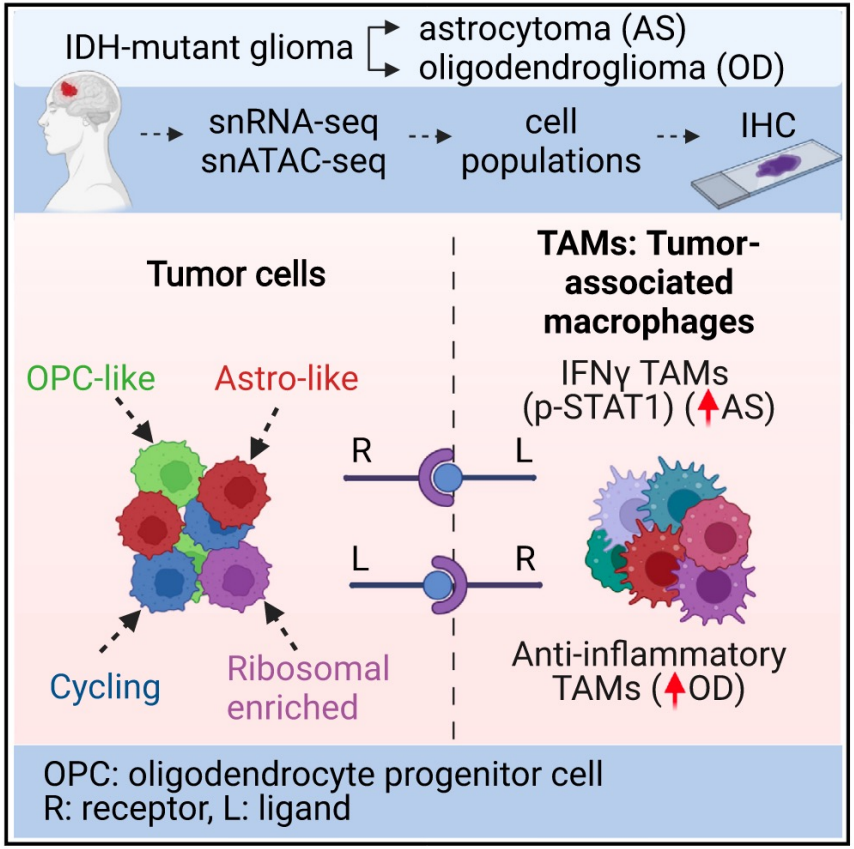
Today

- Get everything setup
  - Login, Copy scripts. Open Day1 getting started.
- Load Data into R and Seurat.
- Try to solve exercises in the Notebook.
- Do QC analysis
- Choose thresholds for your data

- At the End of the day. Save your progress my saving your Seurat object.

Tumor Heterogeneity in IDH mutant Glioma

- Blanco-Carmona, E. *et al.* Tumor heterogeneity and tumor-microglia interactions in primary and recurrent IDH1-mutant gliomas. *Cell Reports Medicine* **4**, 101249 (2023).



https://www.sciencedirect.com/science/article/pii/S2666379123004263
dkfz.

Day2: UMAPs, Cell type annotations and scoring cellular processes

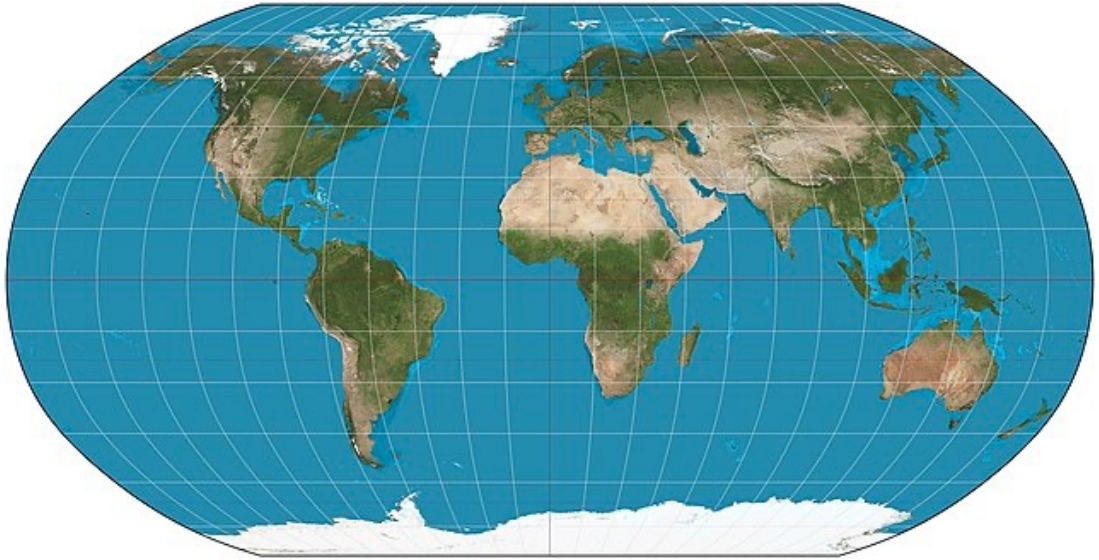
Recap

- How to connect to Rstudio in the Denbi Cloud
- Read 10x data into Seurat
- Seurat Basics
- QC and cell/gene filtering
- Data Normalization and PCA

Low Dimensional Data representation

- PCA
  - Linear data transformation
  - PCs represent axis of highest variances
- Goal: 2D representation
  - Non-linear dimensionality reduction

Non linear dimesionality reduction „warps“space



UMAP (Uniform Manifold Approximation and Projection)

- Step1: Build high dimensional graph representation of the data
- Step2: optimize low dimensional graph to be as simliar as possible

- Advantage:
  - Attempts to balance between local and global structure

UMAP important points

- 1.) Hyperparamters matter
  - # of PCs
  - # of neihbors
- 2.) Cluster sizes mean nothing
- 3.) Distances between clusters may mean nothing
- 4.) Random noise isn‘t always random

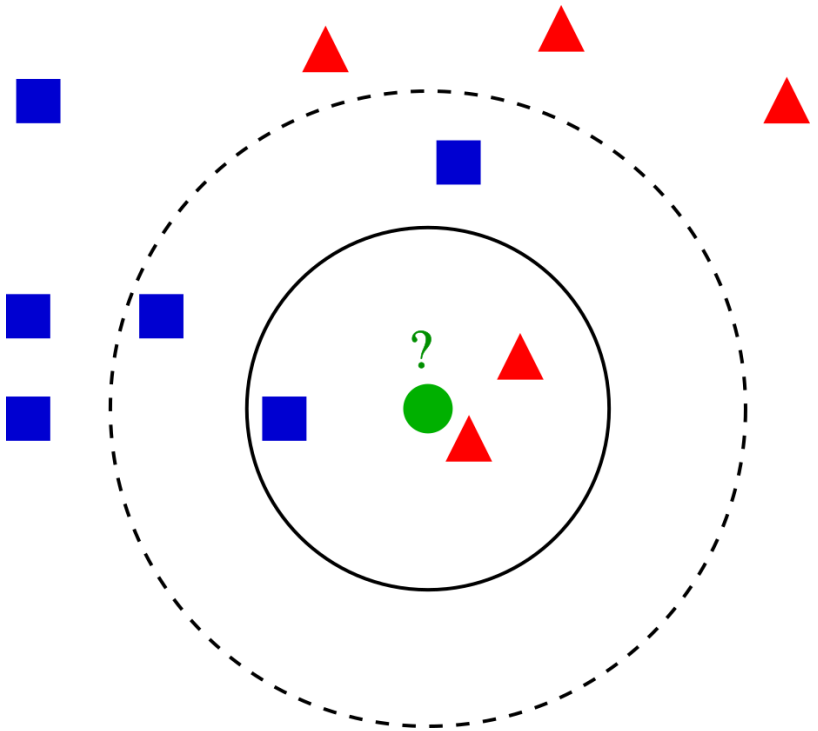
- -> Try multiple settings

https://pair-code.github.io/understanding-umap/

Clustering: Find groups of similar cells

- K Nearest-neighbor graph construction
- 1.) finds K-nearest neihbors in PCA space

- Seurat FindNeighbors
  - creates a shared nearest neighborhood graph between every cell and ist K nearest neighbors



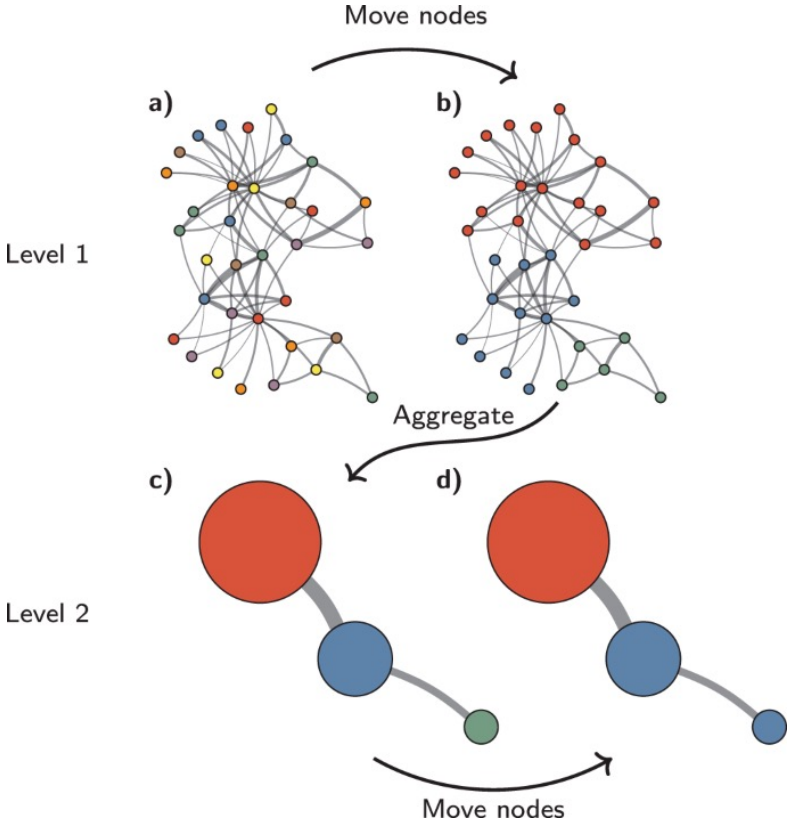
Community Detection

- Community detection methods

- given: (potentially very large) graph
  - twitter social network
  - phone calls between millions of phones
- extract communities: groups of nodes with
  - many nedges between each other
  - few edges to nodes outside of the group

FindClusters

- community search algorithm for graphs called Louvain clustering
- Optimizes „modularity score“ calculated from graph representaiton
- Parameters
  - Resolution (controls „granularity“)
  - K (nearest neighbors)



Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci Rep* **9**, 5233 (2019).

Characterizing clusters

- Now that we have clusters we want to characterize them
  - Finding Marker genes
  - Scoring gene sets

Finding Markers:

- Run differential expression tests between clusters
- Choice of test strong influence on our results.
  - Wilcoxon ranked sum test (non parametric)
- However, almost all methods assume that individual cells are biological replicates
  - High number of replicates -> high p-values -> high false positive rate

Types of tests in Seurat

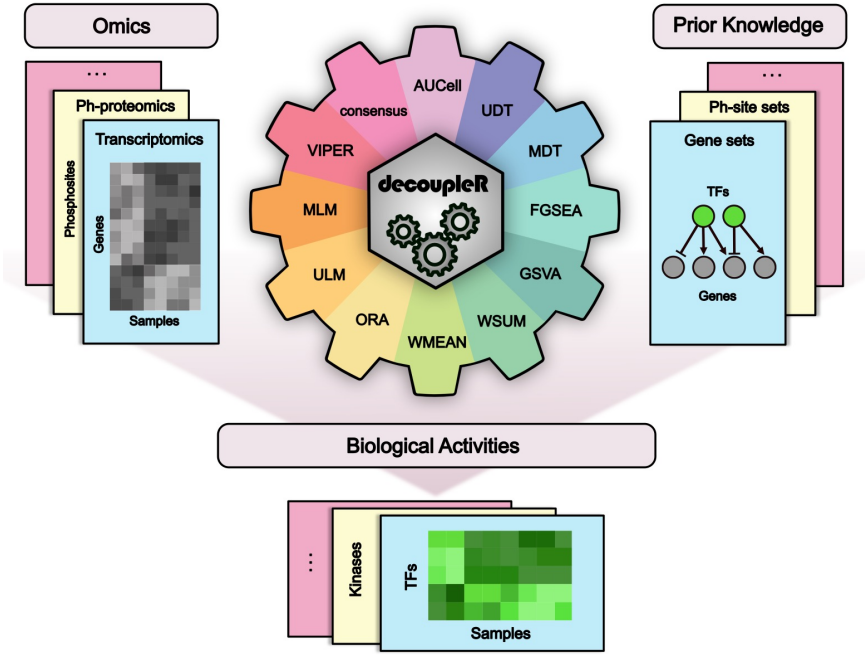
- FindMarkers -> Finds marker genes between ident.1 and ident.2
- FindAllMarkers -> Compares ident.1 to all other cells -> attempt to find genes uniquely changing in ident.1
- FindConservedMarkers -> Finds genes that are conserved in two groups vs all other groups.

How do we classify cell types

- Biological Knowledge
  - Marker genes in FACS or other methods can also work for cell type identification f.e. CD8 and CD4
- Using databases of known cell markers f.e. PanglaoDB
  - https://panglaodb.se/

How do we score gene sets in single cell?

- Many different methods.
- We use decoupleR
  - Implements many methods
  - Works for both bulk and single cell RNAseq
  - Tutorial for R and python
  - https://saezlab.github.io/decoupleR/
- AUCell method



Badia-i-Mompel, P. *et al.* decoupleR: ensemble of computational methods to infer biological activities from omics data. *Bioinformatics Advances* **2**, vbac016 (2022).

Today

- UMAP representation of our Data
- Perform Clustering to Identify clusters in our data
- Find potential Marker genes for each cluster.
  - Panglao
  - Check the marker genes from the original publication
- Score cell type gene sets of Marker genes.

Cell types

- Multiple types of Tumor cells
- Astrocytes
- Endothelial cells
- Microglia
- Neurons
- Oligodendrocytes
- Pericytes
- T-cells

Day 3

Batch effect, integration and between sample comparisons

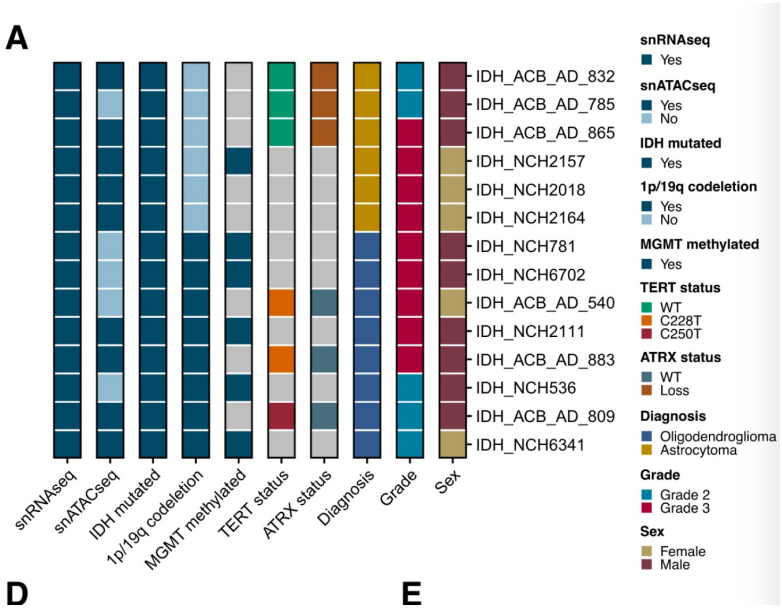
Recap

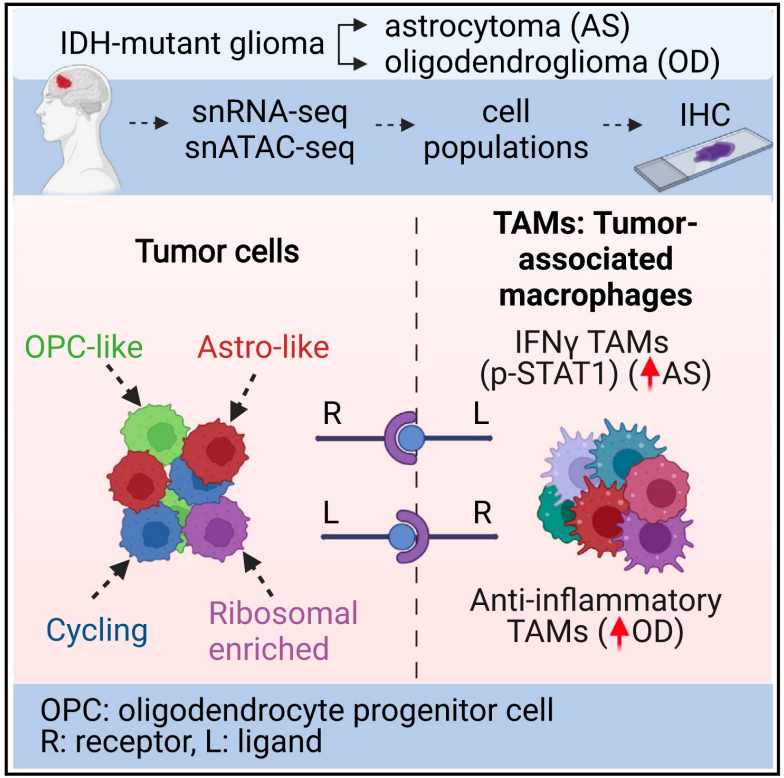
- UMAP
- How to run clustering analysis
- Subsequently, how to characterize clusters using
  - Differential expression tests
  - Enrichment analysis

Today

- Batch effect correction and Integration of multiple samples
- Running Differential expression analysis between conditions across multiple samples.

Our Dataset





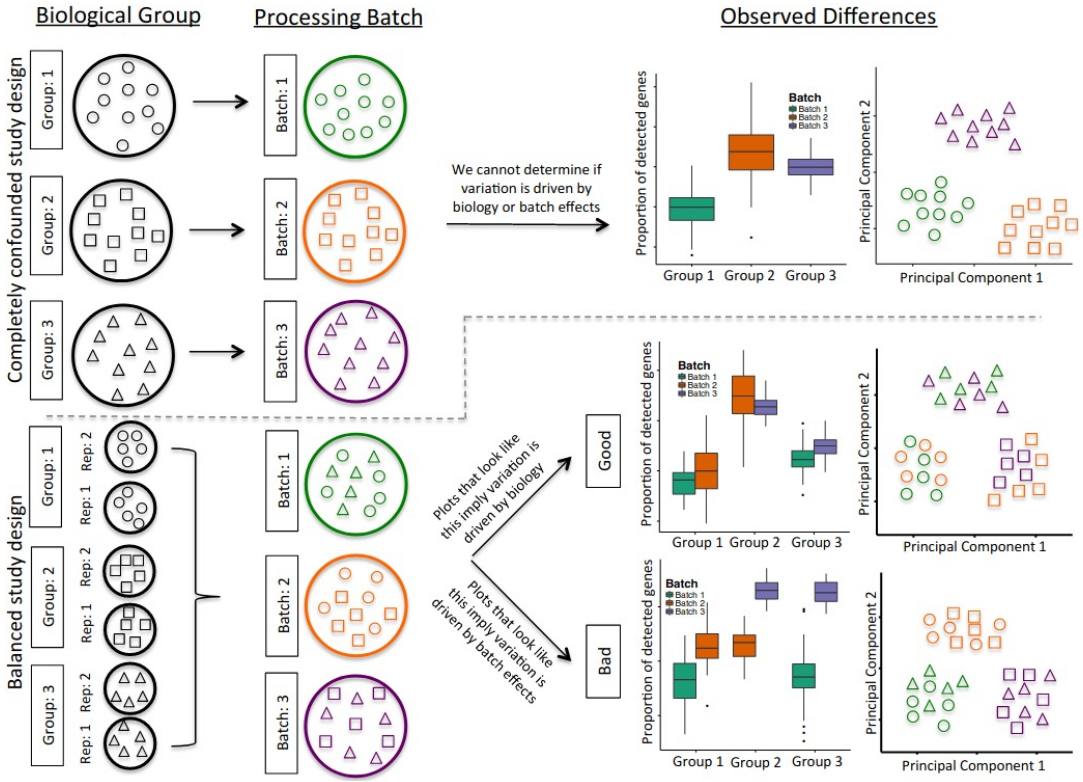
Blanco-Carmona, E. *et al.* Tumor heterogeneity and tumor-microglia interactions in primary and recurrent IDH1-mutant gliomas. *Cell Reports Medicine* 4, 101249 (2023).

Methods for Batch effect correction

- Fast moving field, new methods every few months
- General goal:
  - Identify batch effects and remove that probably carry no biological information and remove it from the dataset.

- In Seurat: Seuratv5 has completely revamped how integration analysis is run (in case you check older tutorials)

Batch effects and experimental design



Workflow

- N samples with G groups
- Each sample is processed separately.
  - Cell and gene filtering, QC, PCA, normalization etc.

- Different Methods influence different steps of the processes.
  - Seurat CCA
  - Harmony

Types of Single cell Batch correction and Integration models

Global models

- Fit regression model with batch effect covariate

Residuals (often using linear regression):

$$\hat{n}_{gc} = f_D(B_c, \dots)$$

$$r_{gc} = n_{gc} - \hat{n}_{gc} = n_{gc} - (\beta_0 + \beta_1 B_c)$$

in linear model case

Example:
sc.tl.regress_out()

Correct for fitted batch effect:

$$n_{gcb} = \alpha_g + X\beta_g + \gamma_{gb} + \delta_{gb}\epsilon_{gcb}$$

bio design matrix

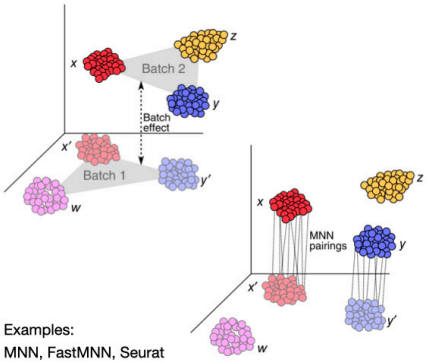
additive batch effect

multiplicative batch effect

Example:
ComBat - ***scanpy.pp.combat()***

Linear embedding models

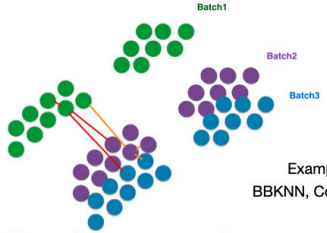
- Project cells into low dimensional embedding
- find most similar cells in other batch e.g., using mutual nearest neighbours (MNNs)
- Use MNNs as anchors to calculate a correction vector



Examples:
MNN, FastMNN, Seurat v3, Scanorama

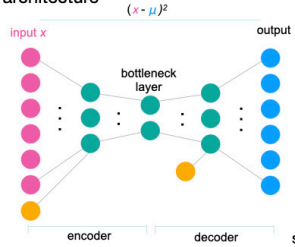
Graph-based methods & Deep learning

Enforce graph connections between different batches



Examples:
BBKNN, Conos

Add condition node into auto-encoder architecture

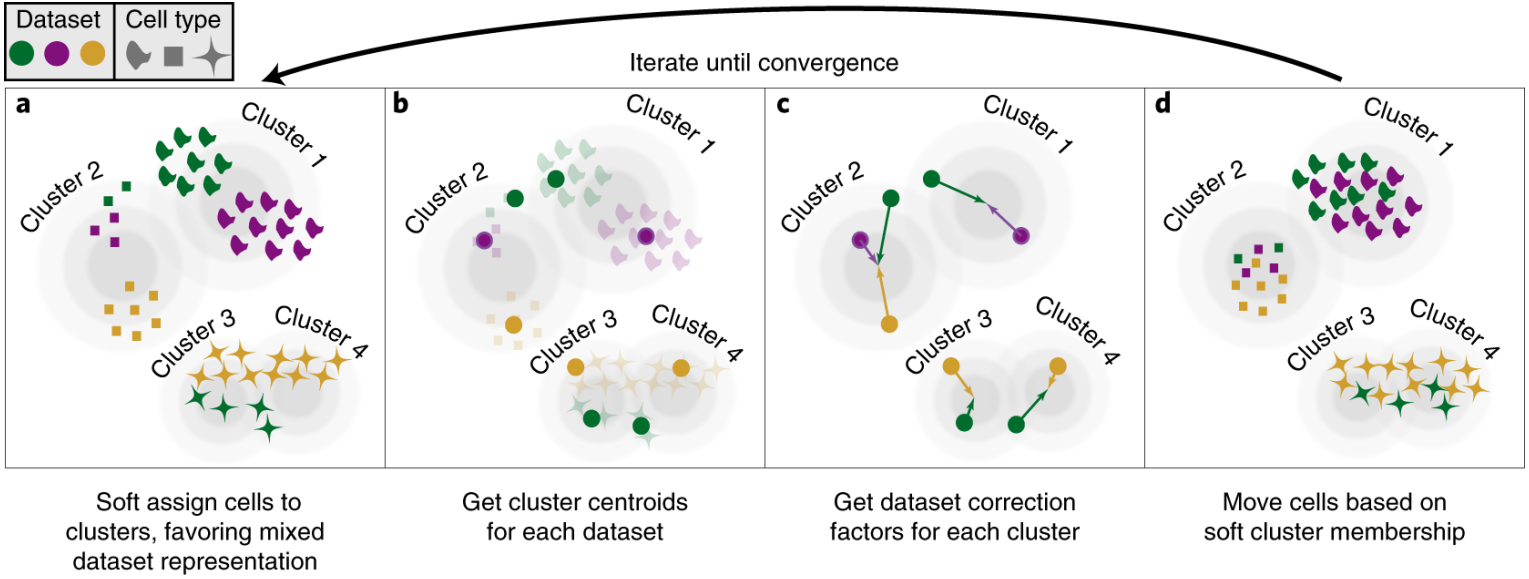


Examples:
scVI, trVAE, SAUCIE

https://www.sc-best-practices.org/cellular_structure/integration.html

dkfz.

Harmony iteratievly finds clusters and corrects the low dimensional representation



New to seurat v5

- One combined Seurat Object an be split with `split`
- Each „split“ object ist normalized separately + PCA
- Run IntegrateLayers
  - Choice of method with `method`
- Use new reduction as low dimensional representation for clustering and UMAP

Today we will:

- We will run Harmony and both Seurat methods on the whole IDH gliomas Dataset and attempt to integrate the data.
- Compare integration methods. Which seems to work best.

Differential Expression analysis

- When we want to compare expression between conditions in the single cell space, we have more to consider
  - We cant just compare all cells from condition A to B since we ignore our multi sample structure
- Options
  - Single cell methods (MAST)
  - Pseudobulk
- Surprisingly, Pseudobulk is still competitive to sc specific methods

Pseudobulk

- We aggregate expression by
  - Cell type
  - Sample
- If we aggregate a sufficient amount of cells, we reduce the sparsity of our data -> bulk RNAseq methods can then be used on the resulting data (DESeq2)
- Run analysis between conditions for each cell type

Today

- We will compare our dataset of glimoas if we can find any expression differences between grade2 and grade3 tumors across the entire dataset
- Pseudobulk using cell type annotations by the authors.
- Run DESeq2 using Seurat
- See if we can find any differences between our samples.

Finally

- This is just the beginng

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer

The German Cancer Research Center (DKFZ) in Heidelberg

Innovative Cancer Research in a Historic City

Career Opportunities at all Levels:

- Professors
- Junior Group Leaders
- Postdocs
- PhD and MSc Students

www.dkfz.de

dkfz.

Thank you for your attention!

Further information on www.dkfz.de

dkfz. GERMAN CANCER RESEARCH CENTER IN THE HELMHOLTZ ASSOCIATION

Research for a Life without Cancer