PSC
PITTSBURGH SUPERCOMPUTING CENTER

Carnegie Mellon University

University of Pittsburgh

Integrating and Interpreting Single-Cell Datasets

Jack Lasota
Pittsburgh Supercomputing Center (PSC)
Carnegie Mellon University / University of Pittsburgh

Project Overview & Background

Pittsburgh Supercomputing Center (PSC)

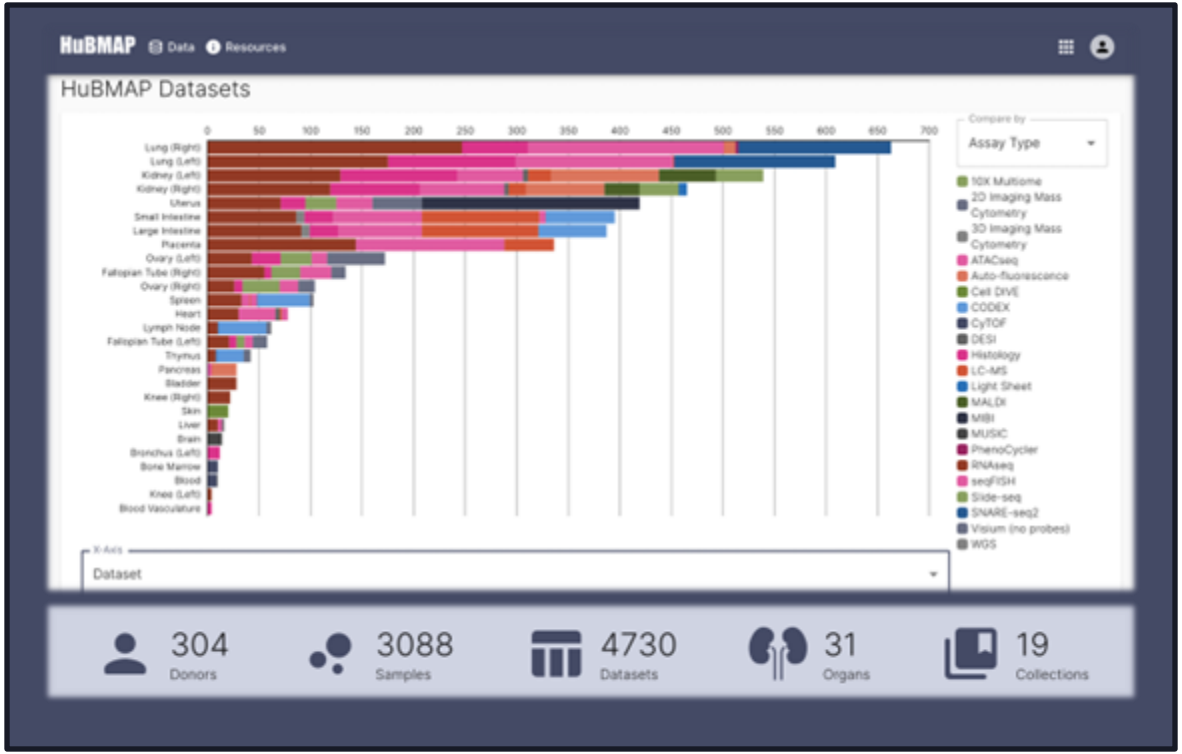
- Jointly operated by Carnegie Mellon University & University of Pittsburgh.
- Provides high-performance computing and networking for university, government, and industry researchers.
- Powers computation-heavy research in:
  - Data analytics, machine learning, AI & deep learning.
  - Biomolecular simulation and scientific discovery.
- Premier supercomputers include Bridges-2 & Anton.

From Raw Data to Integration: Project Goals & Tools

- **Objective:** Utilize pipelines to process and output large-scale biological data, along with providing accessible tools for its interpretation.
- **Standardization:** Uniformly process data so results are comparable across datasets and tissue types.
- **Tools:** Use Python (Jupyter Notebook), along with open-source tools such as Scanpy (Python toolkit), and AnnData (Python data structure).
- **Environment:** All computation is run in high-performance UNIX environments.
- **Products:** Strive for outputs that are clean, structured datasets ready to be explored, visualized, or annotated.

HuBMAP Consortium Data Portal

- **Data Portal:** Centralized access to datasets with tools for analysis and visualization.
- **Discovery:** Search or browse by organ, cell type, or molecule.
- **Visualization:** Interactive tools for spatial and single-cell data.
- **Access:** Download raw and processed datasets.



Carnegie Mellon University
PITTSBURGH SUPERCOMPUTING CENTER
University of Pittsburgh

Sequencing Pipelines and Data Interpretation

Raw Heart Data - scRNAseq

```
heart = ad.read_h5ad("HT_raw.h5ad")
```

Python

```
heart
```

Python

```
AnnData object with n_obs × n_vars = 117782 × 60286
    obs: 'uuid', 'hubmap_id', 'age', 'sex', 'height', 'weight', 'bmi', 'cause_of_death', 'race', 'barcode', 'dataset', 'a
    var: 'hugo_symbol'
    uns: 'annotation_metadata', 'cell_type_counts', 'creation_date_time', 'datasets', 'uuid'
```

```
# value_counts() is a pandas function
heart.obs["predicted_label"].value_counts()
```

predicted_label	
regular ventricular cardiac myocyte	38106
fibroblast	23853
capillary endothelial cell	2479
B cell	1920
myeloid cell	1897
macrophage	1833
pericyte	1806
mesothelial cell	1223
natural killer cell	842
endothelial cell of venule	554
T cell	521
smooth muscle cell	499
endocardial cell	353
endothelial cell of artery	328
mast cell	302
fat cell	207
Schwann cell	204
endothelial cell of lymphatic vessel	186
regular atrial cardiac myocyte	63
endothelial cell	8

Name: count, dtype: int64

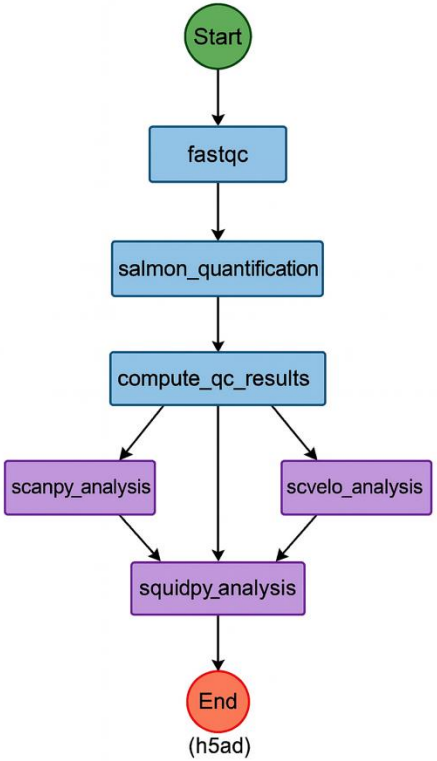
```
heart.var["hugo_symbol"].value_counts()
```

hugo_symbol	
Y_RNA	732
Metazoa_SRP	167
U3	50
U6	29
SNORA70	24
...	
LAP3P1	1
LAP3P2	1
LAPTM4A	1
LAPTM4A-DT	1
snoZ196	1

Name: count, Length: 38618, dtype: int64

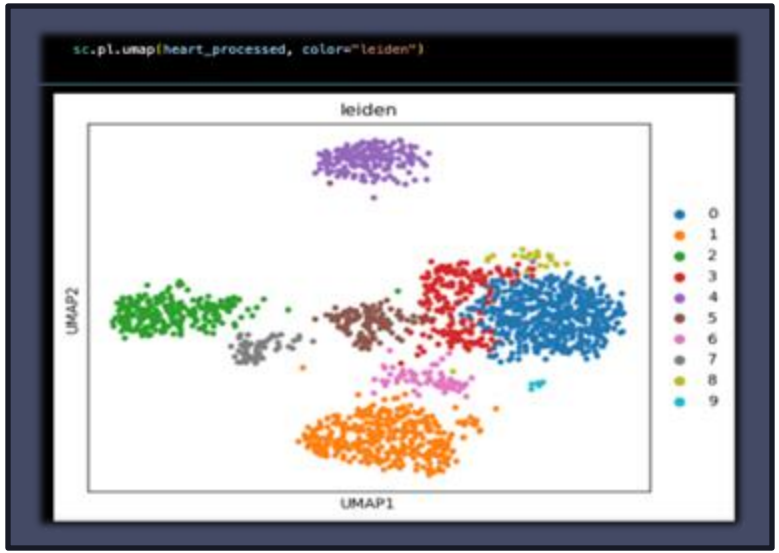
- HuBMAP raw heart data was read using the AnnData package, which provides tools for interpreting and working with complex biological datasets.
- The pandas library enables targeted analysis of specific data points.

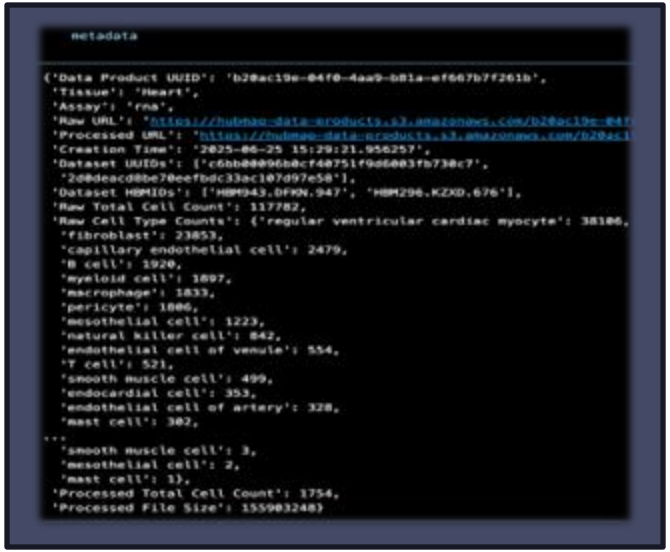
Pipeline Operation



- Datasets were processed from raw FASTQ format using the Salmon-Alevin pipeline, enabling rapid quality control, quantification, initial analysis, and even filtering of low-quality cells and genes.
- This workflow produces clean, high-quality datasets in .h5ad format, ready for integration and further exploration in a Jupyter Notebook.
- Pipelines like this are essential for transforming raw sequencing data into readable, structured formats suitable for interpretation and analysis.

Processed Heart Data – Pipeline Results



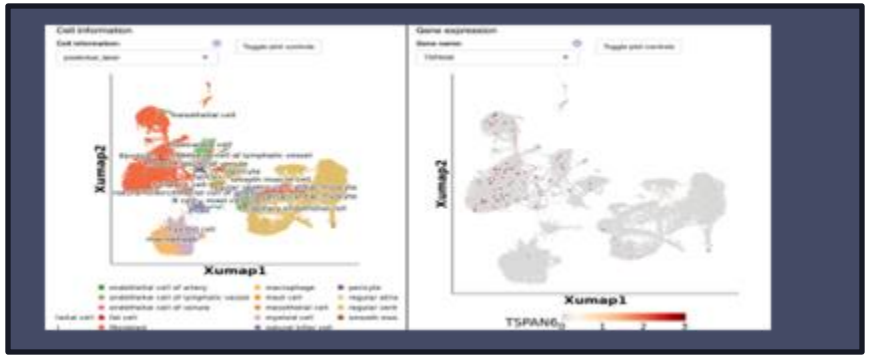


- After filtering low-quality cells and weakly expressed genes, about 99% of cells and 75% of genes were filtered out.
- Pipeline utilized Leiden clustering, viewed with UMAPs plotted by scanpy.
- Metadata gives an overview of the data set.

Carnegie Mellon University PITTSBURGH SUPERCOMPUTING CENTER University of Pittsburgh

User Interface

Tissue Type	Creation Time	Assay Type	Number of Datasets	Total Raw Cells	Number Raw Cell Types	Total Processed Cells	Number Processed Cell Types	Raw Data Product Download	Processed Data Product Download	Shiny App	(All Versions)
Bladder	June 23, 2025, 7:09 a.m.	ma-seq	14	170943	NA	160161	NA	Raw HSAD	Processed HSAD	Shiny App	All versions of Bladder
Fallopian Tube (Left)	June 25, 2025, 6:10 a.m.	multiome-ma-atac	4	32329	NA	17397	NA	Raw h5mu	Processed h5mu	NA	All versions of Fallopian Tube (Left)
Fallopian Tube (Left)	June 23, 2025, 7:09 a.m.	ma-seq	16	34935	NA	14233	NA	Raw HSAD	Processed HSAD	Shiny App	All versions of Fallopian Tube (Left)
Fallopian Tube (Right)	May 28, 2025, 5:49 a.m.	multiome-ma-atac	14	91339	NA	46054	NA	Raw h5mu	Processed h5mu	NA	All versions of Fallopian Tube (Right)
Fallopian Tube (Right)	June 23, 2025, 7:09 a.m.	ma-seq	38	34931	NA	15305	NA	Raw HSAD	Processed HSAD	Shiny App	All versions of Fallopian Tube (Right)
Heart	June 23, 2025, 7:09 a.m.	ma-seq	15	675772	21	22094	19	Raw HSAD	Processed HSAD	Shiny App	All versions of Heart
Kidney (Left)	June 23, 2025, 7:09 a.m.	ma-seq	75	1522897	48	598799	48	Raw HSAD	Processed HSAD	Shiny App	All versions of Kidney (Left)



- Metadata is integrated into the user interface, with tissue type overviews provided. Shiny app enables both gene and cell-level insights/comparisons.

Leveraging Scanpy for Standardized Single-Cell Analysis

Scanpy Pre-processing vs. Integration

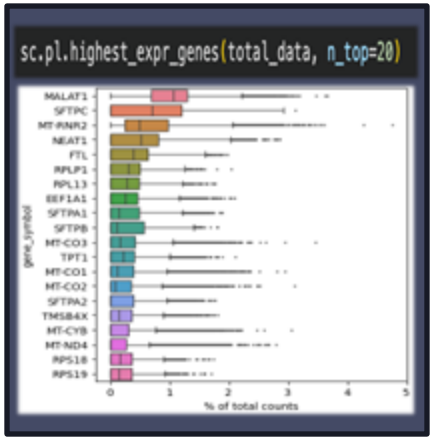
Pre-processing of data:

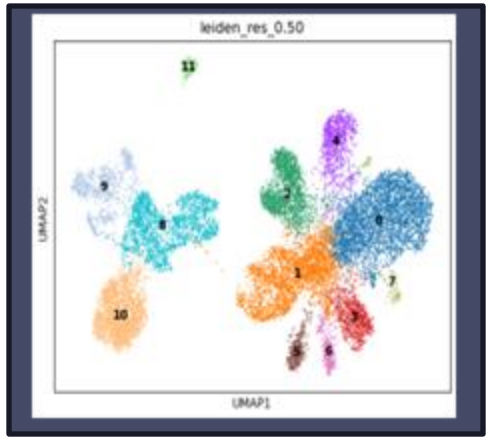
- Filters low-quality cells and genes.
- Normalizes data and visualizes gene/cell counts.
- Identifies highly variable genes.
- Computes Principal Component Analysis (PCA) to capture major variation.

Data integration:

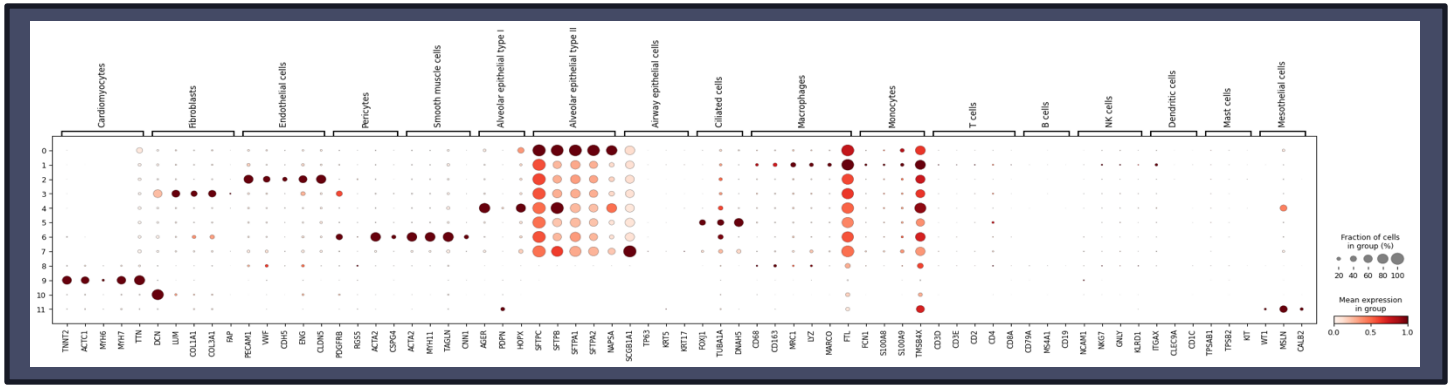
- Aligns multiple datasets in a shared feature space.
- Projects new cells into existing PCA/UMAP embeddings.
- Transfers annotations (e.g., cell types) from reference to new data.

Scanpy pre-processing: Heart and Kidney






```
marker_genes = {
    "Cardiomyocytes": ["TNNT2", "ACTC1", "MYH6", "MYH7", "TTN"],
    "Fibroblasts": ["DCN", "LUM", "COL1A1", "COL3A1", "FAP"],
    "Endothelial cells": ["PECAM1", "VWF", "CDH5", "ENG", "CLDN5"],
    "Pericytes": ["PDGFRB", "RGS5", "ACTA2", "CSPG4"],
    "Smooth muscle cells": ["ACTA2", "MYH11", "TAGLN", "CNN1"],
    "Alveolar epithelial type I": ["AGER", "PDPN", "HOPX"],
    "Alveolar epithelial type II": ["SFTPC", "SFTPB", "SFTPA1", "SFTPA2", "NAPSA"],
    "Airway epithelial cells": ["SCGB1A1", "TP63", "KRT5", "KRT17"],
    "Ciliated cells": ["FOXJ1", "TUBA1A", "DNAH5"],
    "Macrophages": ["CD68", "CD163", "MRC1", "LYZ", "MARCO", "FTL"],
    "Monocytes": ["FCN1", "S100A8", "S100A9", "TMSB4X"],
    "T cells": ["CD3D", "CD3E", "CD2", "CD4", "CD8A"],
    "B cells": ["CD79A", "MS4A1", "CD19"],
    "NK cells": ["NCAM1", "NKG7", "GNLY", "KLRD1"],
    "Dendritic cells": ["ITGAX", "CLEC9A", "CD1C"],
    "Mast cells": ["TPSAB1", "TPSB2", "KIT"],
    "Mesothelial cells": ["WT1", "MSLN", "CALB2"]
}
sc.pl.dotplot(total_data, marker_genes, groupby="leiden_res_0.50", standard_scale="var")
```



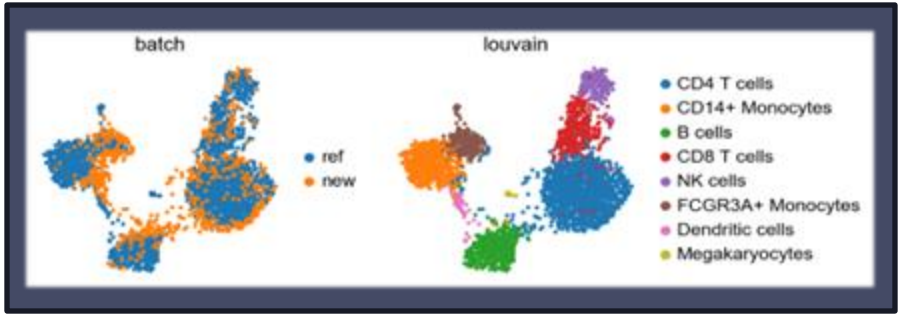
Carnegie Mellon University

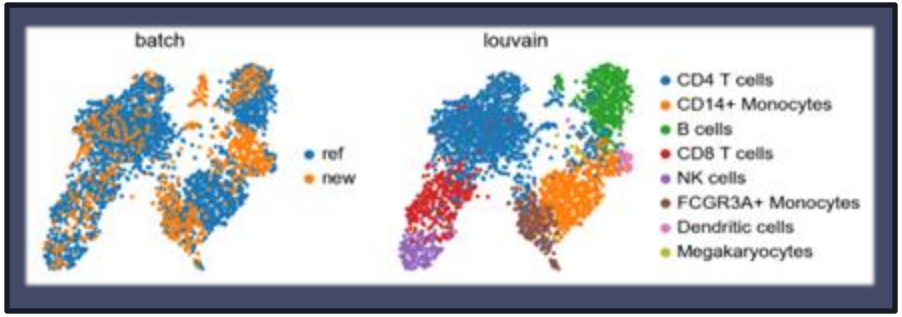
PITTSBURGH SUPERCOMPUTING CENTER

University of Pittsburgh

Scanpy – Data Integration with Ingest + BBKNN

- Purpose: To map new single-cell data (Peripheral Blood Mononuclear Cells) onto a reference dataset while correcting for batch effects across datasets.





Ingest Mapping:

- New cells are projected into the reference UMAP and inherit its clustering and annotations.
- Some separation remains (e.g., in monocytes), but key cell types like T and B cells mix well, indicating partial batch alignment.

BBKNN Mapping:

- BBKNN corrects batch effects by rebuilding the neighbor graph, improving mixing across datasets.
- Monocytes and dendritic cells integrate well, but the Megakaryocyte cluster is lost.

Acknowledgements

- Thank you to the HuBMAP HIVE teams, particularly the CMU Tools Component team and the PSC IEC team for their guidance, technical support, and the tools that enabled my research. They have been instrumental in helping me explore, analyze, and interpret the complex biological data I used throughout this project.

- I want to give a special thanks to:
  - Matt Ruffalo (CMU): Principal Investigator, Systems Scientist
  - Penny Cuda (CMU): Research Programmer, HIVE Tools Component
  - Xiang Li (PSC): Senior Bioinformatics Support Specialist

Sources

- Wolf, Fabian, et al. *Scanpy Tutorials*. Theis Lab, https://scanpy.readthedocs.io/en/stable/tutorials/index.html

- "Tutorials." *AnnData Documentation*, https://anndata.readthedocs.io/en/stable/tutorials/index.html.

- hubmapconsortium. Github.com/hubmapconsortium/salmon-rnaseq. v2.2.8, Zenodo, 17 Sept. 2024, doi:10.5281/zenodo.16421110.

- HuBMAP Consortium. *HuBMAP Data Portal*. Human BioMolecular Atlas Program, https://portal.hubmapconsortium.org/