**Supplementary Tables**

**Single-Cell Data Integration and Cell Type Annotation through Contrastive Adversarial Open Set Domain Adaptation**

Fatemeh Aminzadeh1,2**†**, Jun Wu3**†**, Jingrui He4, Morteza Saberi2, Fatemeh Vafaee1,5,6**\***

1 School of Biotechnology and Biomolecular Sciences, University of New South Wales (UNSW Sydney), Australia

2 School of Computer Science, University of Technology Sydney (UTS), Sydney, Australia

3 Department of Computer Science, University of Illinois Urbana-Champaign, Champaign, IL, USA

4 School of Information Sciences, University of Illinois Urbana-Champaign (UIUC), Champaign, IL, USA

5 UNSW Data Science Hub (uDASH), University of New South Wales (UNSW Sydney), Australia

6 UNSW AI Institute, University of New South Wales (UNSW Sydney), Australia

**†** Equal first Authorship

**\*** Corresponding Author

Correspondence to:

A/Professor Fatemeh Vafaee

School of Biotechnology and Biomolecular Sciences

UNSW SYDNEY NSW 2052 AUSTRALIA

T: +61 (2) 9065 2699

E: [f.vafaee@unsw.edu.au](mailto:f.vafaee@unsw.edu.au)

**Supplementary Table 1.** Summary of Tabula Muris Transcriptomic Mouse Cell Atlas (Page 2)

**Supplementary Table 2**. Summary of Human Pancreatic Islet Datasets (Page 4)

**Supplementary Table 3.** Summary of Simulated Datasets (Page 5)

**Supplementary Table 4.** Summary of Human Lung Atlas (Page 5)

**Supplementary Table 5.** Summary of Cross-species Muscle Datasets (Page 6)

**Supplementary Table 6.** Summary of Human Kidney scATAC seq Datasets (Page 6)

**Supplementary Table 7.** Summary of Human Ovary Multi-Omics Dataset (Page 7)

**Supplementary Table 8.** Summary of Human PBMC Multi-Omics Dataset (Page 7)

**Supplementary Table 9.** Summary of Human PBMC CITE-seq Reference Atlas (Page7)

**Supplementary Table 1.** The scRNA-seq data from eight different tissues in the Tabula Muris cell atlas [1] (GSE132042) was obtained where the gene counts were derived using two techniques: 10X Genomics (10X) and FACS-based cell capture in plates (FACS). The table lists the cell types and the number of cells from both 10X and FACS datasets under two scenarios: (1) 10X is considered as the reference (or source) dataset and FACS as the query (or target) dataset, and (2) the reverse, with FACS as the reference and 10X as target (c.f., ‘Domain’ column).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Organ** | **Domain** | **Platform** | **Cell Types** | **No. of Cells** | **Cell type (No. of Cells)** |
| Bladder | Reference | 10X | 2 | 2109 | bladder cell (923), mesenchymal cell (1186), basal cell of urothelium (0) |
| Target | FACS | 3 | 1287 | bladder cell (532), mesenchymal cell (656), basal cell of urothelium: *unknown* (99) |
| Reference | FACS | 2 | 1188 | bladder cell (532), mesenchymal cell (656), leukocyte(0) |
| Target | 10X | 3 | 2109 | bladder cell (923), mesenchymal cell (1186), leukocyte: *unknown* (57) |
| Heart | Reference | 10X | 6 | 624 | cardiac muscle cell (83), endocardial cell (65), endothelial cell (178), erythrocyte (55), fibroblast (222), smooth muscle cell (21), epicardial adipocyte (0) |
| Target | FACS | 7 | 4029 | cardiac muscle cell (140), endocardial cell (175), endothelial cell (1268), erythrocyte (11), fibroblast (2087), smooth muscle cell (255), epicardial adipocyte: *unknown* (93) |
| Reference | FACS | 5 | 3681 | cardiac muscle cell (140), endocardial cell (175), endothelial cell (1268), erythrocyte (11), fibroblast (2087), smooth muscle cell (0) |
| Target | 10X | 6 | 624 | cardiac muscle cell (83), endocardial cell (65), endothelial cell (178), erythrocyte (55), fibroblast (222), smooth muscle cell: *unknown* (21) |
| Kidney | Reference | 10X | 5 | 2496 | endothelial cell (45), fenestrated cell (345), fibroblast (113), kidney collecting duct cell (72), kidney tubule cell (1921), leukocyte (0) |
| Target | FACS | 6 | 517 | endothelial cell (40), fenestrated cell (69), fibroblast (65), kidney collecting duct cell (44), kidney tubule cell (261), leukocyte: *unknown* (38) |
| Reference | FACS | 5 | 479 | endothelial cell (40), fenestrated cell (69), fibroblast (65), kidney collecting duct cell (44), kidney tubule cell (261), leukocyte (0) |
| Target | 10X | 6 | 2550 | endothelial cell (45), fenestrated cell (345), fibroblast (113), kidney collecting duct cell (72), kidney tubule cell (1921), leukocyte: *unknown* (54) |
| Liver | Reference | 10X | 2 | 1026 | endothelial cell (20), hepatocyte (1006), B cell (0) |
| Target | FACS | 3 | 624 | endothelial cell (196), hepatocyte (399), B cell: *unknown* (29) |
| Mammary | Reference | 10X | 3 | 1341 | basal cell (393), endothelial cell (251), stromal cell (697), luminal epithelial cell of mammary gland (0) |
| Target | FACS | 4 | 2304 | basal cell (1275), endothelial cell (49), stromal cell (428), luminal epithelial cell of mammary gland: *unknown* (552) |
| Reference | FACS | 3 | 1752 | basal cell (1275), endothelial cell (49), stromal cell (428), macrophage (0) |
| Target | 10X | 4 | 1527 | basal cell (393), endothelial cell (251), stromal cell (697), macrophage: *unknown* (186) |
| Marrow | Reference | 10X | 6 | 3288 | B cell (390), Fraction A pre-pro B cell (63), T cell (162), granulocyte (770), hematopoietic stem cell (1381), monocyte (522), natural killer cell (0) |
| Target | FACS | 7 | 4298 | B cell (1848), Fraction A pre-pro B cell (177), T cell (142), granulocyte (380), hematopoietic stem cell (1291), monocyte (324), natural killer cell: *unknown* (136) |
| Reference | FACS | 6 | 4162 | B cell (1848), Fraction A pre-pro B cell (177), T cell (142), granulocyte (380), hematopoietic stem cell (1291), monocyte (324), erythrocyte (0) |
| Target | 10X | 7 | 3430 | B cell (390), Fraction A pre-pro B cell (63), T cell (162), granulocyte (770), hematopoietic stem cell (1381), monocyte (522), erythrocyte: *unknown* (142) |
| Muscle | Reference | 10X | 6 | 3844 | B cell (463), T cell (323), endothelial cell (1290), macrophage (308), mesenchymal stem cell (1111), skeletal muscle satellite cell (349), skeletal muscle satellite stem cell (0) |
| Target | FACS | 7 | 1937 | B cell (151), T cell (35), endothelial cell (206), macrophage (71), mesenchymal stem cell (486), skeletal muscle satellite cell (546), skeletal muscle satellite stem cell: *unknown* (442) |
| Reference | FACS | 6 | 1495 | B cell (151), T cell (35), endothelial cell (206), macrophage (71), mesenchymal stem cell (486), skeletal muscle satellite cell (546), chondroblast (0) |
| Target | 10X | 7 | 4224 | B cell (463), T cell (323), endothelial cell (1290), macrophage (308), mesenchymal stem cell (1111), skeletal muscle satellite cell (349), chondroblast: *unknown* (380) |
| Spleen | Reference | 10X | 2 | 9013 | B cell (6726), T cell (2287), myeloid cell (0) |
| Target | FACS | 3 | 1689 | B cell (1238), T cell (403), myeloid cell: *unknown* (48) |
| Reference | FACS | 2 | 1641 | B cell (1238), T cell (403), dendritic cell (0) |
| Target | 10X | 3 | 9157 | B cell (6726), T cell (2287), dendritic cell: *unknown* (144) |

**Supplementary Table 2.** We obtained human pancreas datasets from distinct studies and platforms to apply open set annotation. We consider Lawlor, Muraro and Segerstolpe as the source domain and annotated Baron dataset as the target domain.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Organ** | **Domain** | **1st Author**  **[Ref]** | **Platform** | **Cell types** | **No. of Cells** | **Accession** | **Cell type (No. of Cells)** |
| Pancreas | Reference | Lawlor  [2] | Fluidigm C1 | 7 | 617 | GSE86469 | Betta (264), Ductal (28), Alpha (239), Acinar (24), Gamma (18), Delta (25), Stellate (19), Macrophage: *unknown* (0) |
| Reference | Muraro  [3] | CelSeq2 | 9 | 2106 | GSE85241 | Betta (446), Ductal (244), Alpha (808), Acinar (218), Gamma (97), Delta (191), Endothelial (21), Epsilon (3), Mesenchymal (78), Macrophage: *unknown* (0) |
| Reference | Segerstolpe  [4] | Smart-Seq2 | 10 | 2122 | E-MTAB-5061 | Betta (270), Ductal (386), Alpha (886), Acinar (185), Gamma (197), Delta (114), Endothelial (16), Epsilon (7), Mast (7), PSC (54), Macrophage: *unknown* (0) |
| Target | Baron  [5] | InDrop | 8 | 8254 | GSE84133 | Betta (2525), Ductal (1077), Alpha (2326), Acinar (958), Gamma (255), Delta (601), Stellate (457), Macrophage: *unknown* (55) |

***Supplementary Table 3.*** *There are two simulated datasets. Simulated dataset1 comprises six batches, designed to replicate an experiment involving multiple samples derived from a single tissue containing seven cell types, each generated using distinct technologies. Simulated dataset 2 is designed to mimic a more complex experimental setup with a nested structure comprising four batches, each containing three sub-batches. This design reflects a multi-center study, where each center processes multiple batches, potentially using different technologies. We integrated different batches in both simulated datasets.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Dataset Name**  **[Ref]** | **Batch** | **Cell Types**  **(Groups)** | **No. of Cells** | **Cell Types (No. of Cells)** |
| Simulated Dataset 1  [6] | Batch 1 | 6 | 2908 | Group 1 (1026), Group 2 (639), Group 3 (440), Group 4 (372), Group 5 (288), Group 6 (143) |
| Batch 2 | 6 | 2422 | Group 1 (601), Group 2 (803), Group 3 (487), Group 4 (240), Group 6 (194), Group 7 (97) |
| Batch 3 | 5 | 2120 | Group 1 (634), Group 2 (638), Group 4 (425), Group 5 (214), Group 6 (209) |
| Batch 4 | 5 | 1929 | Group 1 (288), Group 2 (387), Group 3 (681), Group 4 (285), Group 5 (288) |
| Batch 6 | 4 | 1761 | Group 1 (433), Group 2 (331), Group 3 (144), Group 4 (49) |
| Batch 5 | 7 | 957 | Group 1 (706), Group 2 (87), Group 3 (177), Group 4 (175), Group 5 (299), Group 6 (210), Group 7(107) |
| Simulated Dataset 2  [6] | Batch 1 | 3 | 4806 | Group 1 (1682), Group 2 (1058), Group 3 (1212), Group 4 (854) |
| Batch 2 | 3 | 4905 | Group 1 (1198), Group 2 (1921), Group 3 (1303), Group 4 (483) |
| Batch 3 | 3 | 4809 | Group 1 (1446), Group 2 (726), Group 3 (1673), Group 4: (964) |
| Batch 4 | 4 | 4798 | Group 1 (481), Group 2 (1447), Group 3 (953), Group 4 (1917) |

***Supplementary Table 4.*** *We utilized the single-cell expression data from the human lung atlas, as detailed in the work of [6]. This atlas encompasses samples from three different laboratories, generated using both Drop-seq and 10X Chromium platforms. The Drop-seq data was sourced from GEO under accession code GSE130148. For our experiment, we considered the transplant datasets as the reference domain and the biopsy dataset as the target domain.*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Organ** | **Domain** | **Dataset Name** | **Platform** | **Cell Types** | **No. of Cells** | **Cell type (No. of Cells)** |
| Lung | Reference | Transplant | 10x Chromium | 14 | 22417 | B cell (1302), Ciliated (575), Dendritic cell (1172), Endothelium (678), Fibroblast (682), Lymphatic (339), Macrophage (7468), Mast cell (887), Neutrophil CD14 high (1622), Neutrophils IL1R2 (471), Secretory (718), T/NK cell (1793), Type 1(424), Type 2 (4286), Ionocytes: *unknown* (0) |
| Drop-seq |
| Target | Biopsy | 10x Chromium | 8 | 4998 | B cell (51), Ciliated (2580), Dendritic cell (195), Endothelium (310), Fibroblast (51), Macrophage (24), Secretory (1741), Ionocytes: *unknown* (46) |

***Supplementary Table 5.*** W*e used three datasets from different studies on muscle tissue from mouse[7], bovine [8], and human [9]. We configured the open set annotation settings to map from mouse to human and from bovine to human, running two separate experiments where different cell types were treated as unknown.*

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Organ** | **Domain** | **1st Author**  **[Ref]** | **Species** | **Platform** | **Accession No.** | **Cell Types** | **No. of Cells** | **Cell Types (No. of Cells)** |
| Muscle | Reference | Messmer  [8] | Bovine | 10x | GSE211428 | 11 | 36129 | B/T/NK cells (188), Endothelial (2246), FAPs (21409), MuSCs (10650), Smooth Muscle (913), Myocytes (45), Glial cells (60), Tenocytes (23), Dendritic cells (unknown in Run1) (51), Monocytes (unknown in Run2) (544) |
| Target | Rubenstein  [9] | Human | 10x | GSE130646 | 7 | 2876 | B/T/NK cells (111), Endothelial (1261), FAPs (581), MuSCs (361), Smooth Muscle (370), Myocytes (0), Glial cells (0), Dendritic cells (unknown in Run1) (155), Monocytes (unknown in Run2) (37) |
| Reference | Dell  [7] | Mouse | 10x | GSE126834 | 7 | 4414 | B/T/NK cells (1,249), Endothelial (933), FAPs (1,257), MuSCs (251), Smooth Muscle (240), Dendritic cells (unknown in Run1) (0), Monocytes (unknown in Run2) (0), Tenocytes (337), Macrophages (147) |
| Target | Rubenstein  [9] | Human | 10x | GSE130646 | 7 | 2876 | B/T/NK cells (111), Endothelial (1,261), FAPs (581), MuSCs (361), Smooth Muscle (370), Dendritic cells (unknown in Run1) (155), Monocytes (unknown in Run2) (37) |

***Supplementary Table 6.*** *We utilize two sc-ATAC-seq datasets from studies on the kidney cortex: the Muto dataset [10] and the Wilson dataset [11]. We annotated Wilson as the target dataset according to Muto Dataset which was the source dataset.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Organ** | **Domain** | **1st Author**  **[Ref]** | **Sequencing Method** | **Platform** | **Cell Types** | **No. of Cells** | **Cell type (No. of Cells)** |
| Kidney | Reference | Muto  [10] | ATAC-seq | 10X | 10 | 24835 | Podocyte (133), Renal beta-intercalated cell (620), Epithelial cell of proximal tubule (690), Renal principal cell (1,291), Renal alpha-intercalated cell (613), Parietal epithelial cell (402), Kidney proximal convoluted tubule epithelial cell (6384), Kidney proximal straight tubule epithelial cell (4146), Kidney distal convoluted tubule epithelial cell (2781), Kidney loop of Henle thick ascending limb epithelial cell (7775), Endothelial cell: unknown (0) |
| Target | Wilson  [11] | ATAC-seq | 10X | 11 | 36732 | Podocyte (241), Renal beta-intercalated cell (729), Epithelial cell of proximal tubule (1,734), Renal principal cell (975), Renal alpha-intercalated cell (1112), Parietal epithelial cell (578), Kidney proximal convoluted tubule epithelial cell (12467), Kidney proximal straight tubule epithelial cell (3451), Kidney distal convoluted tubule epithelial cell (5995), Kidney loop of Henle thick ascending limb epithelial cell (8176), Endothelial cell: unknown (1274) |

***Supplementary Table 7.*** *We utilized datasets from the study by [12], which dataset provides an epigenomic perspective, enabling the study of chromatin accessibility alongside gene expression in human ovarian cells and includes transcriptomic (scRNA-seq) and regulatory (scATAC-seq) profiles of the normal postmenopausal ovary and fallopian tube. We focused on the ovary and integrated RNA-seq and ATAC-seq datasets.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Organ** | **Dataset**  **Name** | **Platform** | **Cell Types** | **No. of Cells** | **Cell type (No. of Cells)** |
| Ovary | RNA-seq | Drop-seq | 5 | 26087 | Endothelial cell (291), Smooth muscle cell (150), Stromal cell (21572), Pericyte (3483), Leukocyte (591) |
| ATAC-seq | scATAC-seq | 5 | 18315 | Endothelial cell (156), Smooth muscle cell (165), Stromal cell (15607), Pericyte (2,104), Leukocyte (283) |

***Supplementary Table 8.*** *We utilized a single cell multiome dataset from peripheral blood mononuclear cells (PBMCs)* *[13] obtained from a cryopreserved sample of a healthy female donor. This dataset offers paired ATAC-seq and gene expression profiles, enabling a comprehensive analysis of chromatin accessibility alongside transcriptomic data. We integrated RNA-seq and ATAC-seq datasets.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Organ** | **Dataset Name** | **Platform** | **Cell Types** | **No. of Cells** | **Cell type (Number of Cells)** |
| PBMC | RNA-seq | 10x | 19 | 10412 | CD14 Mono (2,812), CD16 Mono (514), CD4 Naive (1,419), CD4 TCM (1,149), CD4 TEM (298), CD8 Naive (1,410), CD8 TEM\_1 (325), CD8 TEM\_2 (358), cDC (198), gdT (146), HSPC (26), Intermediate B (353), MAIT (137), Memory B (371), Naive B (142), NK (468), pDC (106), Plasma (18), Treg (162) |
| ATAC-seq |

***Supplementary Table 9.*** *This CITE-seq reference consists of a comprehensive dataset of 162000 human peripheral blood mononuclear cells (PBMCs), measured with 228 antibodies [14]. This multimodal dataset was constructed using a technique called Cellular Indexing of Transcriptomes and Epitopes by Sequencing (CITE-seq), which combines RNA sequencing with simultaneous protein measurements through antibody labeling, which were our datasets to be integrated.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Organ** | **Dataset**  **Name** | **Platform** | **Cell Types** | **No. of Cells** | **Cell type (Number of Cells)** |
| PBMC | Protein | CITE-seq | 8 | 161764 | B (13800), CD4 T (41001), CD8 T (25469), DC (3589), Mono (49010), NK (18664), Other (3442), Other T (6789) |
| RNA |

**References**

1. Schaum, N., et al., *Single-cell transcriptomics of 20 mouse organs creates a Tabula Muris: The Tabula Muris Consortium.* Nature, 2018. **562**(7727): p. 367.

2. Lawlor, N., et al., *Single-cell transcriptomes identify human islet cell signatures and reveal cell-type-specific expression changes in type 2 diabetes.* Genome Res, 2017. **27**(2): p. 208-222.

3. Muraro, M.J., et al., *A Single-Cell Transcriptome Atlas of the Human Pancreas.* Cell Syst, 2016. **3**(4): p. 385-394.e3.

4. Segerstolpe, Å., et al., *Single-Cell Transcriptome Profiling of Human Pancreatic Islets in Health and Type 2 Diabetes.* Cell Metab, 2016. **24**(4): p. 593-607.

5. Baron, M., et al., *A Single-Cell Transcriptomic Map of the Human and Mouse Pancreas Reveals Inter- and Intra-cell Population Structure.* Cell Syst, 2016. **3**(4): p. 346-360.e4.

6. Luecken, M.D., et al., *Benchmarking atlas-level data integration in single-cell genomics.* Nat Methods, 2022. **19**(1): p. 41-50.

7. Dell'Orso, S., et al., *Single cell analysis of adult mouse skeletal muscle stem cells in homeostatic and regenerative conditions.* Development, 2019. **146**(12).

8. Messmer, T., et al., *Single-cell analysis of bovine muscle-derived cell types for cultured meat production.* Frontiers in nutrition, 2023. **10**: p. 1212196.

9. Rubenstein, A.B., et al., *Single-cell transcriptional profiles in human skeletal muscle.* Scientific Reports, 2020. **10**(1): p. 229.

10. Muto, Y., et al., *Single cell transcriptional and chromatin accessibility profiling redefine cellular heterogeneity in the adult human kidney.* Nature Communications, 2021. **12**(1): p. 2190.

11. Wilson, P.C., et al., *Multimodal single cell sequencing implicates chromatin accessibility and genetic background in diabetic kidney disease progression.* Nature Communications, 2022. **13**(1): p. 5253.

12. Lengyel, E., et al., *A molecular atlas of the human postmenopausal fallopian tube and ovary from single-cell RNA and ATAC sequencing.* Cell Rep, 2022. **41**(12): p. 111838.

13. Genomics, x., *PBMC from a Healthy Donor - Granulocytes Removed Through Cell Sorting (10k) - Single Cell Multiome ATAC + Gene Expression Dataset*, x. Genomics, Editor. 2020: 10x Genomics.

14. Hao, Y., et al., *Integrated analysis of multimodal single-cell data.* Cell, 2021. **184**(13): p. 3573-3587.e29.