







-  - Data
-  - Results that need action
-  - Result that does not need action
-  - Code step
-  - Action item to be taken
-  - Final result that does not need action

Explanation of each step-

Step 1 - We use the source Azimuth data from – (<https://hubmapconsortium.github.io/asctb-azimuth-data-comparison/>) These are Azimuth references formatted as ASCT+B tables. In this analysis we have used data for Blood PMBC, Bone Marrow, Brain, Kidney, Lung and Pancreas

Step 2 - ASCT+B tables are obtained from – (https://docs.google.com/spreadsheets/d/1tK916JyG5ZSXW_cXfsyZnzXfjyoN-8B2GXLbYD6_vF0/edit#gid=2137043090) Version 1.1.

Step 3 – Pass Azimuth data into a code chunk to check for author Labels that do not have CT/ID.

Step 4 – Pass ASCT+B data into a code chunk to check for author Labels that do not have CT/ID.

Step 5 – This step gives Azimuth CTs with missing CT/ID. Most Azimuth CTs have been mapped to more generalized CTs if author specified CT is not present in Ontology. Hence the count of Azimuth CTs with missing CT/ID is 0 for all organs as of now. This data can be found for each organ present in Final [result folder](#) under the sheet name - `Az_missing_cts`.

Action Item 1- Confirm whether count of Azimuth CTs with missing CT/ID is 0 for all organs. If not contact Azimuth to let them make a note of this.

Step 6 – This step gives ASCT+B CTs with missing CT/ID. This might be because of certain CTs that are currently being added to Ontology. This data can be found for each organ present in Final [result folder](#) under the sheet name - `Asctb_missing_cts`

Action Item 2- Work with curators & authors to add these to Ontology.

Example result for ASCT+B-

CT/ID	CT/LABEL	CT/LABEL.Author
	Pulmonary vein smooth muscle cell	Pulmonary vein smooth muscle cell
	Pulmonary vein fibroblast	Pulmonary vein fibroblast
	ciliated cuboidal epithelium	ciliated cuboidal epithelium
	airway smooth muscle cells	smooth muscle cells
	Fibroblast type 2 (F2)	lung matrix fibroblast 2
	lung matrix fibroblast 2	lung matrix fibroblast 2
	capillary endothelial cell 1	CAP1
	capillary endothelial cell 2	CAP2
	Matrix fibroblast 1/(Lipofibroblast)	matrix fibroblast 1/(Lipofibroblast)
	lung matrix fibroblast 1/(Lipofibroblast?)	matrix fibroblast 1/(Lipofibroblast)
	resident memory CD4 T cell	resident memory CD4 T cell
	resident memory CD8 T cell	resident memory CD8 T cell

(Results of Step 5 and 6)

Organ	5	6
Lung	0	12
Pancreas	0	0
Kidney	0	12
Bone Marrow	0	0
Blood PMBC	0	0

Step 7 - Push the remainder of the Azimuth and ASCT+B CTs that do have an CT/ID further down the pipeline.

Step 8- This code chunk finds perfect math between ASCT+B CTs and Azimuth tables. For example- If the cell CL:000158 is present in Azimuth and the same cell is present in ASCT+B table for a respective organ then we call it a perfect match i.e CTs that are present in both the data sets.

Step 9- The result of the perfect matches is stored in Final [result folder](#) under the sheet name - `Az_Ascbtb_cts_perfect_matches` for each organ. We denote this as **Result 1**.

Example result for Lung reference-

AZ.CT/ID	AZ.CT/LABEL	AZ.CT/LABEL.Author	ASCTB.CT/ID	ASCTB.CT/LABEL	ASCTB.CT/LABEL.Author
CL:0000158	club cell	Club	CL:0000158	club cell of bronchiole	club
CL:0000186	myofibroblast cell	Myofibroblast	CL:0000186	secondary crest myofibroblasts	secondary crest myofibroblasts
CL:0000236	B cell	B	CL:0000236	B cell	B cell
CL:0000623	natural killer cell	Natural Killer	CL:0000623	natural killer	natural killer
CL:0000669	pericyte cell	Pericyte	CL:0000669	pericyte	pericyte
CL:0000784	plasmacytoid dendritic cell	Plasmacytoid Dendritic	CL:0000784	plasmacytoid dendritic cell	plasmacytoid dendritic cell
CL:0000786	plasma cell	Plasma	CL:0000786	plasma cell	plasma cell
CL:0000814	mature NK T cell	Natural Killer T	CL:0000814	NK T cell	NK T cell
CL:0000860	classical monocyte	Classical Monocyte	CL:0000860	Classical Monocyte	classical monocyte
CL:0000860	classical monocyte	OLR1+ Classical Monocyte	CL:0000860	Classical Monocyte	classical monocyte

The first three columns are derived from Azimuth table and the last three are from ASCT+B table.

Step 10- We move the remainder of the Azimuth CTs that do not have a perfect match (Azimuth mismatches) in ASCT+B further down the pipeline.

Step 11- We move the remainder of the ASCT+B CTs that do not have a perfect match in Azimuth (ASCT+B mismatches) further down the pipeline.

Step 12- We pass the Azimuth mismatches and ASCT+B mismatches from step 10 and 11 to a code chunk which checks if the those CTs are present in Ontology or not.

Step 13- Push the Azimuth CT/IDs that are not in the Ontology. This data can be found in the final summary folder under the sheet name- `Az_incorrect_cts`

Action Item 3- Contact Azimuth regarding these set of CTs. These might be due to issues in Azimuth Github repo.

Example in Azimuth Kidney reference-

CT/ID	CT/LABEL	CT/LABEL.Author
[kidney loop of Henle cortical thick ascending limb epithelial cell](http://www.ontology.org/terms/CT/0000158)	kidney loop of Henle cortical thick ascending limb epithelial cell	Cortical Thick Ascending Limb
Outer Medullary Collecting Duct Intercalated Type A	Outer Medullary Collecting Duct Intercalated Type A	Outer Medullary Collecting Duct Intercalated Type A
Outer Medullary Collecting Duct Principal	Outer Medullary Collecting Duct Principal	Outer Medullary Collecting Duct Principal
Papillary Tip Epithelial	Papillary Tip Epithelial	Papillary Tip Epithelial
Parietal Epithelial	Parietal Epithelial	Parietal Epithelial
Peritubular Capillary Endothelial	Peritubular Capillary Endothelial	Peritubular Capillary Endothelial

Step 14- Push the ASCT+B CT/IDs that are not in the Ontology. This data can be found in the final summary folder under the sheet name- `Asctb_incorrect_cts`

Action Item 4- Work with Authors and curators to get these into CL.

Example in ASCT+B Lung –

CT/ID	CT/LABEL	CT/LABEL.Author
CL:1000388	brush cell of epithelium of trachea	brush
LMHA:00142	ciliated cells of terminal ciliated ducts of tracheal submucosal glands	submucosal gland ciliated duct cells
LMHA:00087	basal cells of terminal ciliated ducts of tracheal submucosal glands	submucosal gland basal cells
LMHA:00143	secretory cells of terminal ciliated ducts of tracheal submucosal glands	submucosal gland secretory
LMHA:00693	epithelial cells of collecting ducts of tracheal submucosal glands	submucosal gland collecting duct epithelium
LMHA:00238	mucus cells of tracheal submucosal glands	submucosal gland mucous cells
LMHA:00340	serous cells of tracheal submucosal glands	submucosal gland serous cells
LMHA:00805	Venous endothelial cell	venous endothelial cell

Results of step 13 and 14-

Organ	13	14
Lung	0	22
Pancreas	0	0
Kidney	6	0
Bone Marrow	2	0
Blood PMBC	0	0

Step 15- Take Azimuth CTs that are present in Ontology and push it to the next step where crosswalk will be done on Azimuth CTs.

Step 16- Take ASCT+B CTs that are present in Ontology and push it to the next step where crosswalk will be done on ASCT+B CTs.

Step 17- This code chunk performs crosswalk matching on ASCT+B CTs. Using CL Ontology we walk up in hierarchy for ASCT+B CTs to match it to Azimuth CT/ID.

Step 18- This code chunk performs crosswalk matching on Azimuth CTs. Using CL Ontology we walk up in hierarchy for Azimuth CTs to match it to ASCT+B CT/ID.

Step 19- Results of the crosswalk matching from step 17 is pushed out to generate a final list of ASCT+B CTs that are not present in Azimuth even after cross walk matching. This data can be found for each organ present in Final [result folder](#) under the sheet name - `Asctb_cts_mismatch_final`

Action item 6- These are ASCT+B CTs that are not present in Azimuth. This might be because of AS regions not present in tissue sample run for scRNAseq assay. A different tissue sample including those regions would need to be run in order to include these CTs.

Example in Lung reference-

ASCTB.CT/ID	ASCTB.CT/LABEL	ASCTB.CT/LABEL.Author
CL:0000185	myoepithelial cells of glands	submucosal gland myoepithelium
CL:0000233	platelet	platelet
CL:0000484	connective tissue mast cell	connective tissue mast cell
CL:0000485	mucosal type mast cell	tissue resident mucosal type mast cell
CL:0000556	lung megakaryocyte	lung megakaryocyte
CL:0002075	brush cell of tracheobronchial tree	tuft
CL:0002138	endothelial cell of lymphatic vessel	lymphatic endothelial cell
CL:0002329	basal epithelial cell of tracheobronchial tree	basal
CL:0002619	endothelial progenitor cells	endothelial progenitors
CL:0019002	tracheobronchial chondrocyte	chondrocyte

Step 20- Results of the crosswalk matching from step 18 is pushed out to generate a final list of Azimuth CTs that are not present in ASCT+B even after cross walk matching. This data can be found for each organ present in Final [result folder](#) under the sheet name - `Az_cts_mismatch_final`

Action item 5- These are the CTs that are not present in ASCT+B. Work with ASCT+B curators to add these CTs into ASCT+B tables in future.

Example in Lung reference-

AZ.CT/ID	AZ.CT/LABEL	AZ.CT/LABEL.Author
CL:0000077	mesothelial cell	Mesothelial
CL:0000160	goblet cell	Goblet
CL:0000165	neuroendocrine cell	Neuroendocrine
CL:0000319	mucus secreting cell	Mucous
CL:0000763	myeloid cell	Platelet/Megakaryocyte
CL:0002393	intermediate monocyte	Intermediate Monocyte
CL:0005006	ionocyte	Ionocyte
UBERON:0001473	lymphatic vessel	Lymphatic
UBERON:0001637	artery	Artery
UBERON:0001638	vein	Vein
UBERON:0001982	capillary	Capillary
UBERON:0004225	respiratory system smooth muscle	Airway Smooth Muscle

Results of step 19 and 20-

Organ	19	20
Lung	12	20
Pancreas	2	12
Kidney	8	8
Bone Marrow	8	8

Blood PMBC	6	10
------------	---	----

Step 21- The Azimuth CTs for which match was found in ASCT+B after using crosswalk are pushed out in this step. The result of the Azimuth crosswalk matches is stored in Final [result folder](#) under the sheet name - `Az_match_tree_crosswalk` for each organ. We denote this as **Result 2**.

Example in Lung reference-

AZ.CT/ID	AZ.CT/LABEL	AZ.CT/LABEL.Author	Match Found	ASCTB.CT/ID	ASCTB.CT/LABEL	Hierarchy Length
CL:0000624	CD4-positive, alpha-beta T cell	CD4 T	Yes	CL:0000084	T cell	4 CL:0000624 (CD4-positive, alpha-beta T cell) >> CL:0000791 (mi
CL:0000624	CD4-positive, alpha-beta T cell	CD4+ Memory/Effector T	Yes	CL:0000084	T cell	4 CL:0000624 (CD4-positive, alpha-beta T cell) >> CL:0000791 (mi
CL:0000625	CD8-positive, alpha-beta T cell	CD8 T	Yes	CL:0000084	T cell	4 CL:0000625 (CD8-positive, alpha-beta T cell) >> CL:0000791 (mi
CL:0000625	CD8-positive, alpha-beta T cell	CD8+ Memory/Effector T	Yes	CL:0000084	T cell	4 CL:0000625 (CD8-positive, alpha-beta T cell) >> CL:0000791 (mi
CL:0002057	CD14-positive, CD16-negative classical monocyte	CD14+ Monocyte	Yes	CL:0000860	classical monocyte	2 CL:0002057 (CD14-positive, CD16-negative classical monocyte)
CL:0002396	CD14-low, CD16-positive monocyte	CD16+ Monocyte	Yes	CL:0000875	non-classical monocyte	2 CL:0002396 (CD14-low, CD16-positive monocyte) >> CL:000087

Step 22- The ADCT+B CTs for which match was found in Azimuth after using crosswalk are pushed out in this step. The result of the ASCT+B crosswalk matches is stored in Final [result folder](#) under the sheet name - `Asctb_match_tree_crosswalk` for each organ. We denote this as **Result 3**.

Example in Lung reference-

ASCTB.CT/ID	ASCTB.CT/LABEL	ASCTB.CT/LABEL.Author	Match Found	AZ.CT/ID	AZ.CT/LABEL	Hierarchy Length
CL:0000084	T cell	T cell	Yes	CL:0000542	lymphocyte	2 CL:0000084 (T cell) >> CL:00000
CL:0000094	neutrophil	neutrophil	Yes	CL:0000766	myeloid leukocyte	2 CL:0000094 (neutrophil) >> CL
CL:0000583	Alveolar Macrophage	alveolar macrophage	Yes	CL:0000235	macrophage	3 CL:0000583 (Alveolar Macroph
CL:0000767	Basophil	basophil	Yes	CL:0000766	myeloid leukocyte	3 CL:0000767 (Basophil) >> CL:0
CL:0000815	regulatory T cell	regulatory T cell	Yes	CL:0000542	lymphocyte	4 CL:0000815 (regulatory T cell)
CL:0000904	CD4+ T cell central memory	CD4+ T cell central memory	Yes	CL:0000542	lymphocyte	7 CL:0000904 (CD4+ T cell centra
CL:0000905	CD4+ T cell effector memory	CD4+ T cell effector memory	Yes	CL:0000542	lymphocyte	7 CL:0000905 (CD4+ T cell effect

The column Hierarchy length shown in step 21 and 22 represents the number of steps required in crosswalk to find a corresponding match.

Crosswalk is nothing but traversing UP the CL Ontology tree for a particular Cell and look for a match for the parents CT.

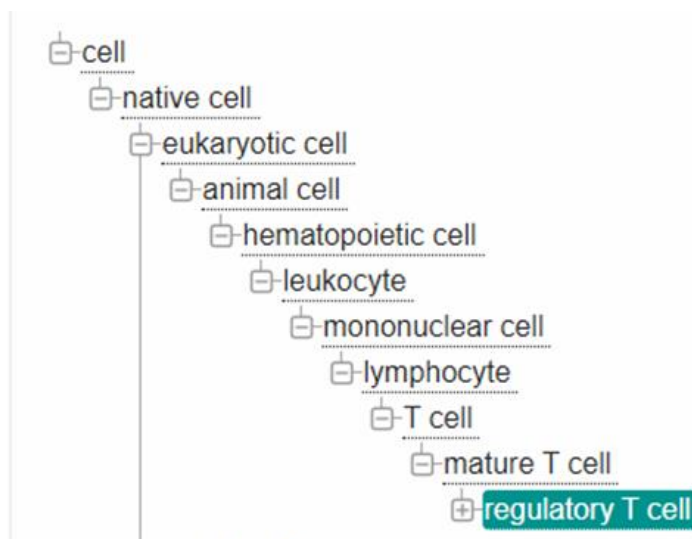
Results of step 21 and 22-

Organ	21	22
Lung	6	30
Pancreas	0	5
Kidney	3	9
Bone Marrow	9	6
Blood PMBC	39	2

Notes-

- 1) Comparison between gold standard (author label) and current results (ontology labels) for kidney
 - 49 mapping records are present in Austin's repo for kidney
 - We are getting 39 mapping records for kidney
 - Our mapping is done based on CT/LABEL (not author label).
 - Austin's mapping is done based on Author Label
 - Our results are found here - <https://github.com/maddy3940/ASCTB-Azimuth-Comparison-22/blob/main/Python/Data/Final/kidney.xlsx> . Look for sheet name is Final_matches in this file for Kidney

Austin's repo - <https://github.com/hubmapconsortium/azimuth-annotate/blob/main/data/kidney.json>
- 2) Currently we are matching on the basis of Ontology CL/IDs which is same as matching with Ontology CL labels as unique label exist for each ID. So 1st and the 2nd steps are essentially the same.
- 3) Ellen will be required for doing matching on basis of Author labels. For each organ I have generated a final mismatch for Azimuth and ASCT+B CTs that are still not matching even after crosswalk.
- 4) Ontology Cell hierarchy can move from more specific cells to generalized cells. Oftentimes we don't want to go to the most generic type of cells in this hierarchy.



- 5) Azimuth and ASCT+B have levels of hierarchy which goes from annotation level 1 to annotation level n i.e., from generalized cells to more specific type of cells.
- 6) Annotation level 1 is the most generalized level in Azimuth. As we increase the annotation levels the idea is that we move towards more specific type of cells. But that is not always true with respect to Azimuth. There are cases where generalized cells are found in annotation level 2 (or level n) i.e. a mix of generalized and specific cells.

- 7) As we go further up in annotation level of Azimuth there are cells that currently do not have Cell IDs in the Ontology. Such cells are mapped to more generalized cells that are found in the Ontology. This makes the ASCT+B and Azimuth mapping more difficult.
- 8) One way to identify the Cell types in Azimuth that have been mapped to a more generalized Cell is to look at the Author assigned labels. If the Ontology label and the Author assigned label are very different then it can mean that cell was mapped to a more generic cell.

Example in Azimuth Kidney reference

AS/3	AS/3/LABEL	AS/3/ID
Descending Thin Limb Type 1	kidney loop of Henle thin descending limb epithelial cell	CL:1001111
Descending Thin Limb Type 2	kidney loop of Henle thin descending limb epithelial cell	CL:1001111
Descending Thin Limb Type 3	kidney loop of Henle thin descending limb epithelial cell	CL:1001111

All the three cells have separate Author Labels but same Ontology Cell ID and Cell Label

- 9) A part where ASCT+B table is different is all such specific cell types are assigned Cell IDs that are present in the Ontology. We do not have such a problem with the ASCT+B tables. ASCT+B effort makes sure all the Cells that the Authors want are present in the ASCT+B tables along with the Ontology ID. Cells that are not present in Ontology are identified and added to Ontology as soon as possible. Cells that are currently being added to Ontology have Author labels in the ASCT+B tables, the Cell IDs and Cell labels are kept blank.

Results-

Final summary table can be found [here](#).

Summary tables for organs can be found [here](#).

Steps to reproduce the work-

- 1) Clone the repo and run the notebook.
- 2) If source needs to be changed then update the [config file](#).