Add title, author, date

***INSIGHTS OF HuBMAP & RUI.***

Firstly , i would like to thank professor Andreas Buekle and katy borner for giving me this opportunity. It really has been such an exciting and informative week for me. Since then , I have been looking into the recent publication and a few videos of HRA , **HuBMAP**. I would like to share my Insights so far. Here are the following understandings:

HuBMAP , Human BioMolecular Atlas Project. The goal of HuBMAP is to create an atlas that can describe every type of cell in human body & to describe cell types in terms of genomic, proteomic and other Biomarkers.  
  
**ASCT + B** Tables represent Anatomical Structures using terms from an Ontology for vertebrate anatomy, known as Uber Anatomy Ontology ( UBERoN).There are several repo’s that store information about biomedical ontologies. The most comprehensive system known as BioPortal, which has been supported by the NIH and also publicly available.

"Tissue registration and exploration user interfaces in support of a human reference atlas" discusses the development and application of software tools and user interfaces designed to facilitate the spatial and semantic annotation (registration) and exploration of tissue data within the context of a comprehensive, high-resolution, three-dimensional atlas of all cells in the healthy human body. This effort is part of a collaborative initiative involving seventeen international consortia aiming to create a human reference atlas (HRA) by harmonizing tissue data across 25 organs and more than 15 bulk and spatial single-cell assay types. The document highlights the challenges in constructing an HRA, such as harmonizing diverse specimen metadata, collecting biological structure data in a standardized manner, and recording the spatial size and orientation of tissue blocks in relation to reference organs. To address these challenges, the authors present two key software programs: the Registration User Interface (RUI) and the Exploration User Interface (EUI).

***Registration User Interface (RUI)***

The RUI is a tool that allows various groups contributing tissue data to register their data within a 3D spatial reference system and derive anatomical structure annotations. This interface supports the documentation of tissue extraction sites in relation to a 3D reference organ, enabling users to specify the size, position, and rotation of a tissue block. The RUI is instrumental in harmonizing data from multiple sources, including the Human Biomolecular Atlas Program (HuBMAP) and the Genotype Tissue Expression project (GTEx). As of April 22, 2022, the RUI has been used to harmonize and publish data on 5,909 tissue blocks.

***Exploration User Interface (EUI)***

The EUI serves both tissue data providers and users. It allows providers to check and approve tissue size, location, rotation, and semantic annotations for registered tissue blocks. Users can use the EUI to filter, browse, and search for tissue and atlas data using various queries related to cell types, anatomical structures, and spatial data. The EUI supports exploration of anatomical structures and cell types predicted via organ-specific ASCT+B tables and displays registered tissue blocks within anatomically correct 3D reference objects.

The paper highlights several challenges in constructing an HRA, such as harmonizing diverse specimen metadata, collecting biological structure data in a standardized manner, and recording the spatial size and orientation of tissue blocks in relation to reference organs. The authors address these challenges by leveraging a common coordinate framework (CCF), which provides standard terminologies and data structures for describing specimen, biological structure, and spatial data linked to existing ontologies. They enable data harmonization across consortia, donors, organs, and assay types, facilitating the collaborative construction of a consensus HRA. The RUI and EUI are described as user-friendly, requiring minimal training time, and have been adopted by other international consortia mapping the human body.

In conclusion, this paper is a testament to the collaborative effort of seventeen international consortia to create a detailed map of the human body at the cellular level. The RUI and EUI are pivotal in this endeavor, providing the necessary infrastructure to register, harmonize, and explore tissue data within the context of the HRA.

***DATA INSIGHTS (ASCT + B) :***

***Anatomical Structure (AS):*** Each dataset is organized around a hierarchical representation of heart anatomy, starting from the broadest category (e.g., "heart") and moving into more specific regions (e.g., "right atrium," "sinoatrial node"). These are labeled with both a common name and associated unique identifiers, likely to ensure consistency and reference across medical and biological databases.

***Cell Type (CT):*** For each anatomical structure, there are associated cell types identified, providing insights into the cellular composition of each heart region. This includes cells like "pacemaker cells," "endothelial cells," and "cardiomyocytes," crucial for understanding the functional aspects of the heart.

***Biomarkers (BG):*** Biomarkers are indicated for various cell types, pointing to specific genes, proteins, or other markers that are characteristic of or relevant to those cells. This data is essential for biomedical research, offering clues for targeted therapies or diagnostics.

***Evolution Across Versions:*** The progression from version 1.0 to 2.0 suggests an ongoing effort to refine and expand the dataset with new findings, more detailed anatomical distinctions, or improved classification of cell types and biomarkers.Changes in the structure (e.g., additional columns, updated naming conventions) and annotations (e.g., notes or references) across versions indicate an attempt to make the data more comprehensive, accessible, and useful for research purposes.

The datasets represent a significant effort to catalog the human heart's anatomical and cellular composition in unprecedented detail. For a researcher or practitioner in the field of cardiology, regenerative medicine, or cellular biology, this data is a treasure trove that supports a wide range of applications, from fundamental research to the development of novel therapeutic interventions. The evolution of the dataset versions reflects a dynamic field where knowledge is continuously updated and refined, underscoring the importance of adaptable and comprehensive datasets in advancing science and medicine.

Given the extensive nature of the column comparison output, which indicates a variety of changes across versions, here's a focused analysis on the key findings, highlighting new columns added in each version and their potential significance. Then, I'll detail the columns of the final release (v2.0) to provide a clear and deep explanation of what each column represents.

***BRAIN DATASET:***

The final release's structure underscores a comprehensive approach to cataloging brain anatomy, cell types, and biomarkers, with a clear emphasis on genetic aspects (BGene columns). The inclusion of detailed references ensures that the dataset is not only a rich resource for anatomical and cellular information but also well-grounded in scientific research, allowing users to trace back to original studies or findings.

This evolution reflects ongoing efforts to deepen the understanding of the brain's complexity, from macroscopic anatomical structures down to the genetic markers defining cellular functions. The dataset thus serves as a valuable tool for research in neuroanatomy, genetics, and neuroscience, facilitating studies ranging from basic science to applied clinical research.

***VISUALIZATIONS:***

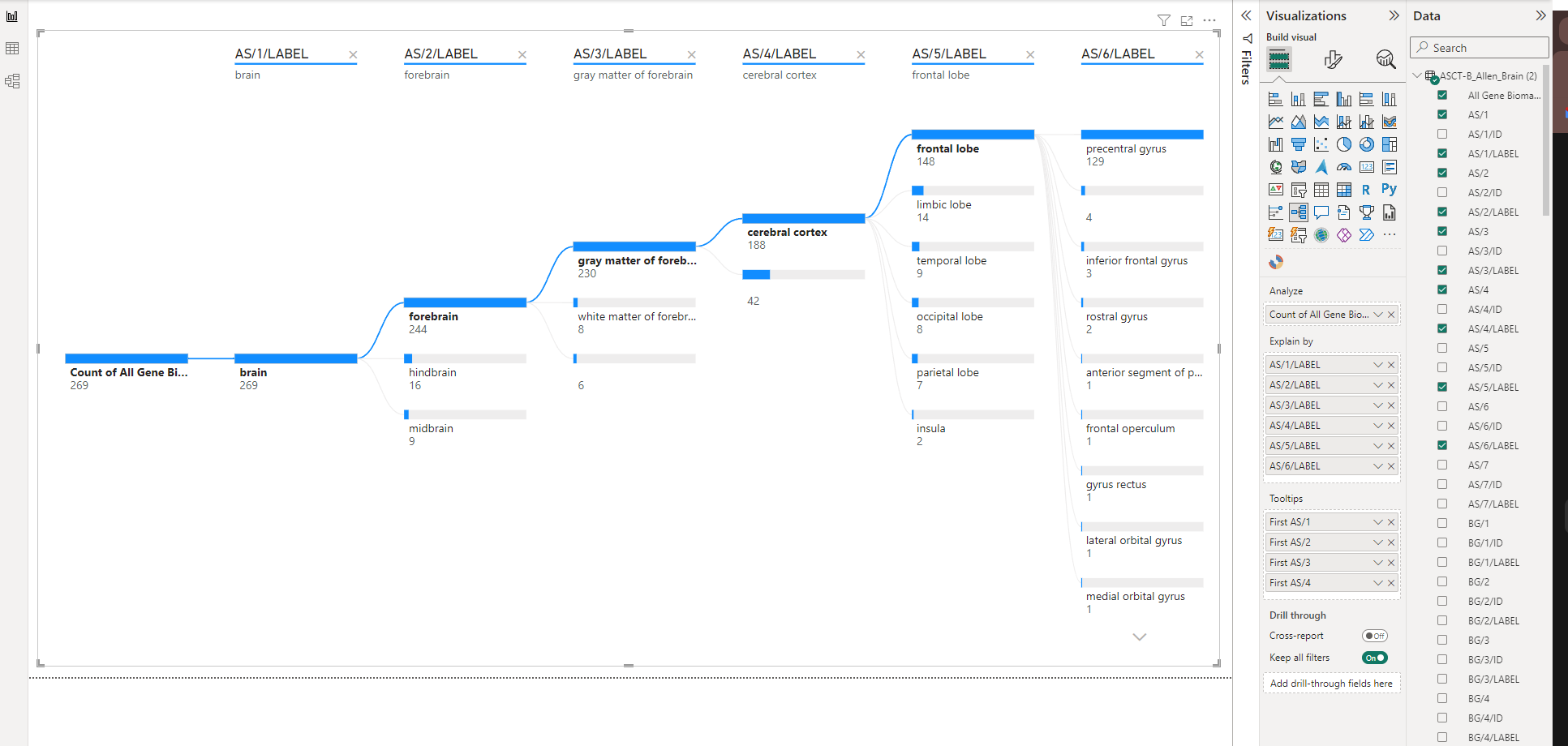
This tree is a hierarchical representation of anatomical structures within the brain, starting from the general category "brain" and breaking down into more specific regions and sub-regions. The visualization helps to clarify the relationships between different parts of the brain and the number of gene biomarkers associated with each part.

we've illustrated a comprehensive overview of the brain's anatomical structure and associated gene biomarkers. This hierarchical approach offers a multi-level breakdown, starting from the broad category of the 'brain' and drilling down into specific regions such as the 'forebrain,' 'hindbrain,' and 'midbrain.'

A key observation from the data is the prominence of the 'forebrain,' which is the most significant contributor among the major divisions of the brain, with a substantial count of 244 gene biomarkers linked to it. This predominance is not surprising, given the forebrain's role in complex behavioral functions and its size relative to other brain regions.

Among the gray matter's components, the 'cerebral cortex' stands out with 188 biomarkers, underlining its importance in higher-order brain functions. It's fascinating to observe the breakdown within the cerebral cortex, where the 'frontal lobe' is distinguished as a key area, with 148 biomarkers, which might be associated with cognitive functions like decision making and problem-solving.

This interpretation of the decomposition tree provides a narrative that captures the hierarchical structure of brain anatomy and its functional implications, as reflected by the distribution of gene biomarkers. It is evident that such data-driven visualizations are powerful in dissecting the complex architecture of the brain and associating it with genetic factors, which are essential for both scientific inquiry and clinical applications.



***2)***

The visualization is Ribbon Chart in Power BI, which is used to visualize data across two or three dimensions, showing the volume of data as it flows between categories. In this case, the chart is visualizing data related to the heart, as indicated by the axis labels and dataset:

**X-Axis (AS/2/LABEL)**: This represents the second level of anatomical structures within the heart (e.g., heart right ventricle, left cardiac atrium). The categories here are specific parts of the heart.

**Y-Axis (Count of AS/1):** This appears to be a count of first-level anatomical structures (AS/1) within each second-level category (AS/2/LABEL). It suggests a quantitative measure, perhaps the number of records, occurrences, or substructures within each part of the heart.

**Legend (CT/1/LABEL):** The colors in the chart correspond to different cell types (CT/1/LABEL) associated with the heart's anatomy. Each color band shows the distribution and prevalence of a particular cell type across various heart regions.

The Ribbon Chart is likely intended to illustrate the complexity and composition of the heart by showing how different cell types are distributed across various anatomical structures of the heart. By observing the thickness of the ribbons, one can compare which cell types are more prevalent in certain parts of the heart. For example, a thicker ribbon indicates a higher count of a particular cell type in a given anatomical structure.

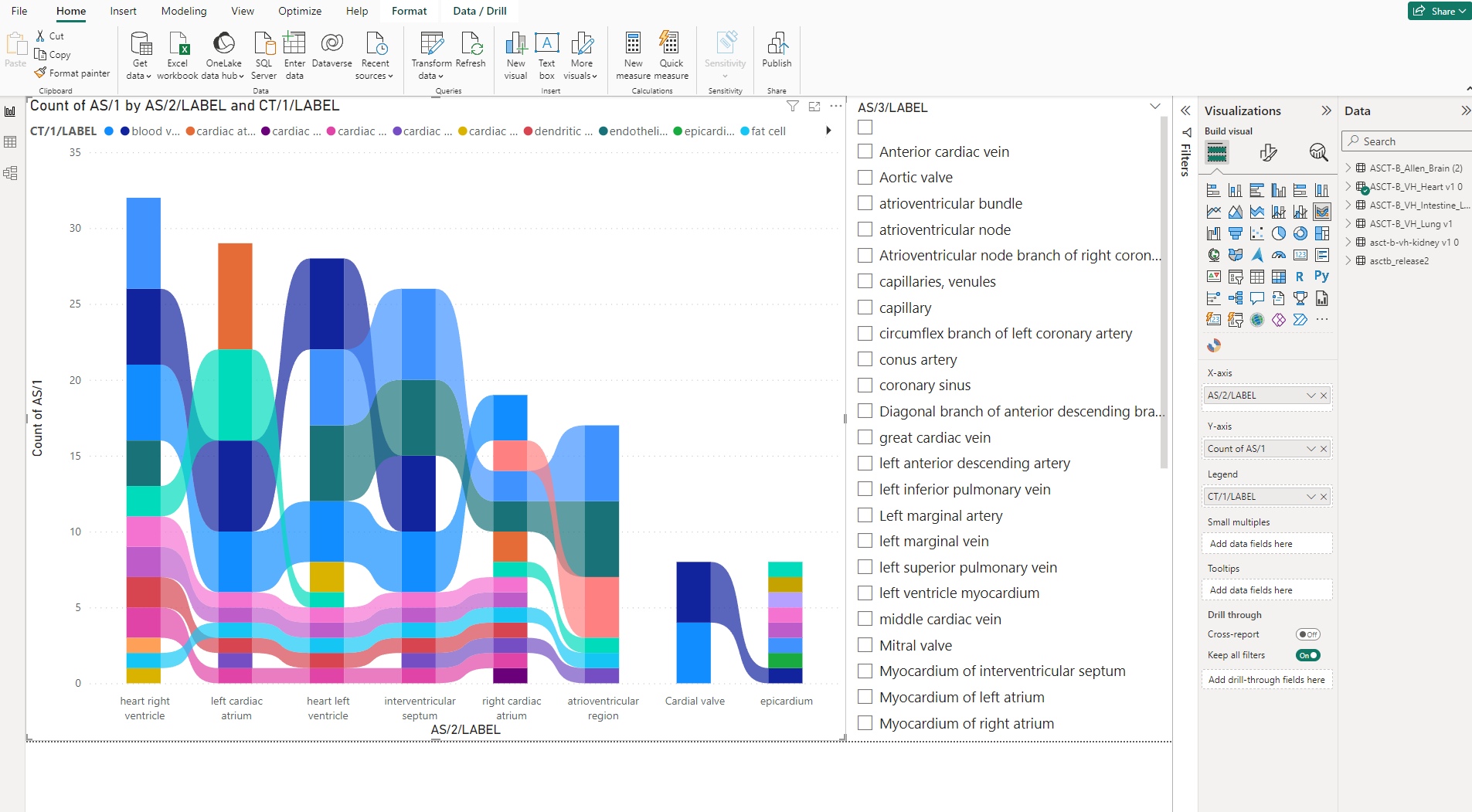
The interconnections of the ribbons can also show relationships between cell types across different anatomical structures, possibly indicating shared functions or commonalities in cell type distribution.

**Interpretation:**

- Areas of the heart with a greater variety of colors (and thus cell types) may be indicative of more complex cellular compositions.

- The chart may also highlight specific anatomical structures that have a high concentration of a certain cell type, which can be useful for medical research and understanding the functional aspects of heart anatomy.

- The flow or movement of ribbon thickness across the heart's anatomy might suggest how different cell types transition or are distributed from one part of the heart to another.

****

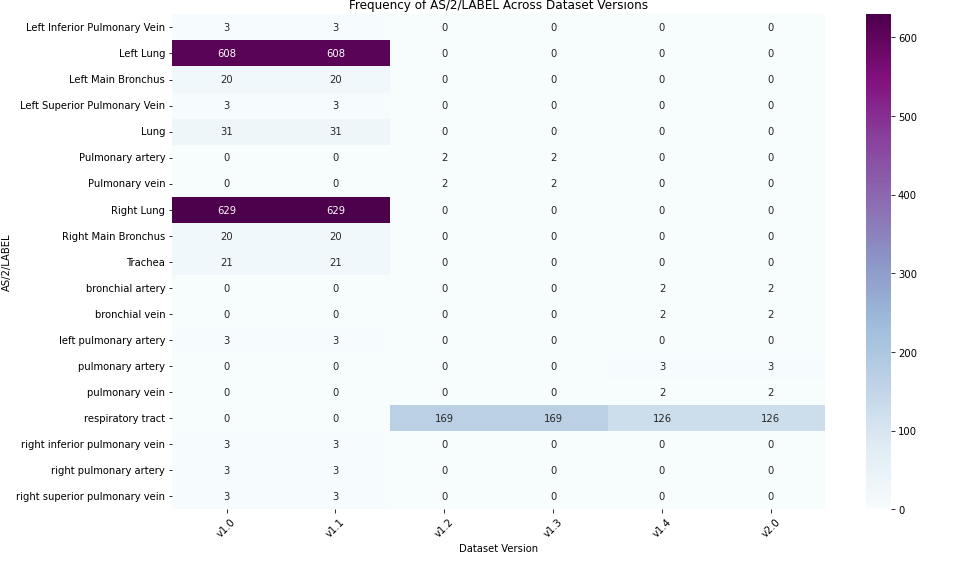
The heatmap's color gradients, which represent the frequency of occurrence for various labels, reveal a consistent emphasis on primary structures such as 'Left Lung' and 'Right Lung'. This stability in recording frequency suggests that these structures are central to the dataset's purpose across its versions.Simultaneously, the heatmap betrays variations, both subtle and pronounced, in the appearances of structures like 'Left Main Bronchus' and 'Trachea'. These fluctuations may be attributable to evolving research foci or refinements in data capture techniques. Notably, the presence or absence of certain structures across different dataset versions highlights an evolutionary trajectory in the dataset—perhaps reflecting updates in scientific understanding or changes in research priorities.

The heatmap also brings attention to the rarer structures and cell types, some of which are absent in certain versions. These gaps could either point to niche areas that have been overlooked or specific focuses that are version-dependent. Moreover, the consistent representation of some labels across different versions speaks to the data collection process's robustness, suggesting that where consistency is expected, the dataset does not disappoint.

However, the heatmap does more than just confirm; it challenges. Anomalies in label frequency might hint at underlying issues such as misclassification or data entry errors, necessitating a closer examination or a potential data cleaning initiative. The data visualization thus acts as both a mirror and a map: reflecting the current state of the dataset and indicating paths for future improvement.

Moreover, the heatmap inadvertently comments on the balance within the dataset. A disproportionate focus on certain labels could lead to a skewed perspective, particularly if this data were to train machine learning models. Such imbalance requires remediation to avoid ingraining biases in predictive algorithms.

Ultimately, the heatmap serves as a strategic tool. It offers a visual summary that can guide decision-making for future data collection and research direction. By observing which areas are well-represented and which are not, researchers can align their data gathering efforts with desired outcomes, whether that be maintaining a focus on established structures or branching out to lesser-charted biological terrains.



**Week - 03/04/24**

**Insights:**

1. The data and visualizations suggest a research environment where certain organs and modalities are more prevalent.

2. There is a noticeable distribution in registration status, which can be critical for understanding the progress and availability of data.

3. The visualizations help identify potential biases or gaps in research, such as disproportionate focus on certain organs or sexes.

4. The distribution of organ labels indicates a variety of focus areas in the research, with some being more central than others.

5. Further investigation into the '30 organs' label is needed to understand its significance.

6. A more in-depth analysis could be conducted to explore the reasons behind the prevalence of certain organs and modalities.

7. The registration status information can be used to manage and prioritize the workflow for data collection and publication.

**Data Analysis:**

-> The data is organized with columns indicating 'Dataset Type', 'Publication Title', 'Organ', 'Registration Status', among others.

-> 'Organ' names are cleaned and standardized to lowercase, and combined organ names are separated.

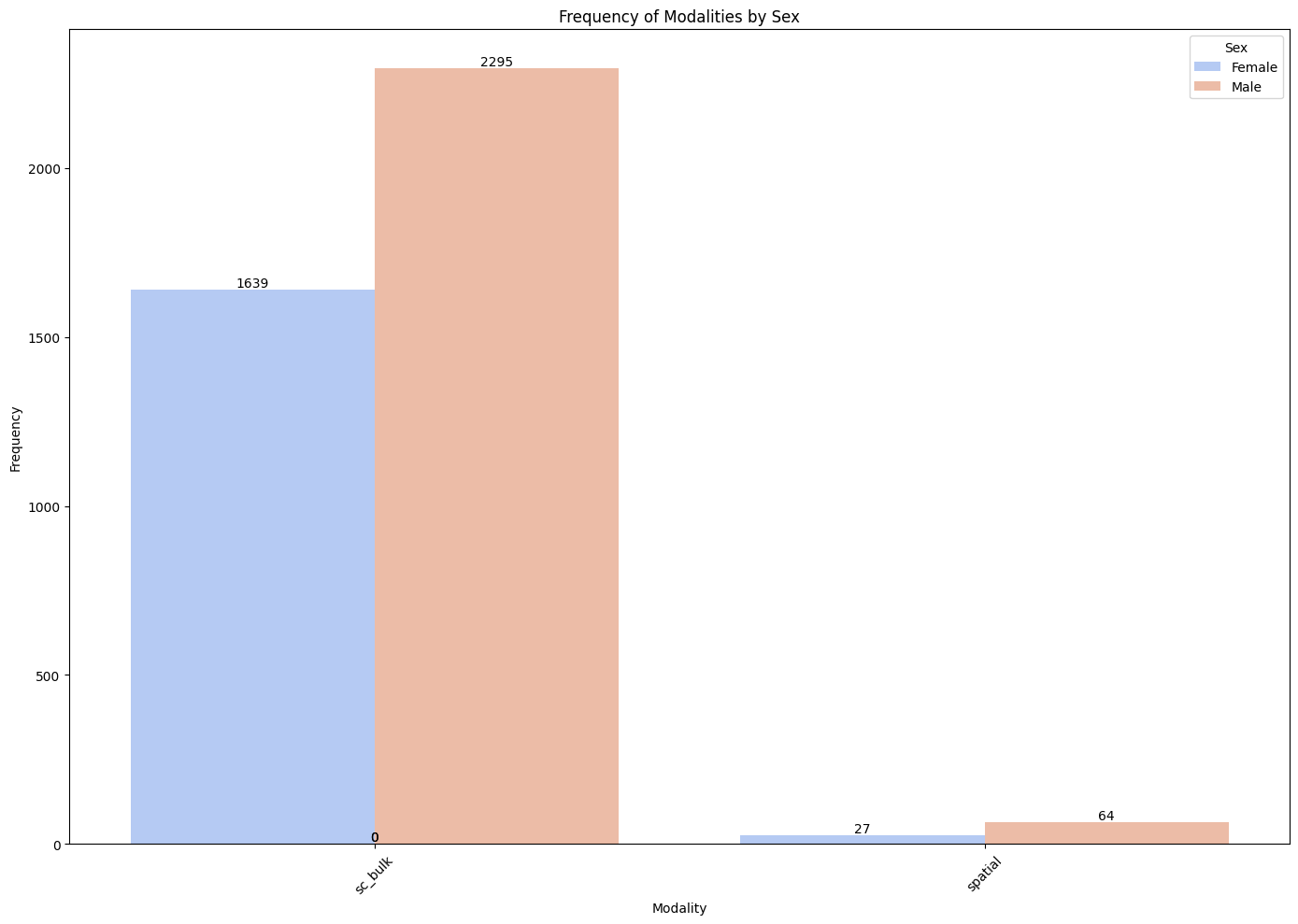
-> 'Registration Status' is also cleaned and standardized to lowercase.

-> There is an 'Organ' labeled as '30 organs', which might be an aggregate label or error; it has been noted once in the 'need author info' registration status.

-> The most frequent registration status is 'need author info' with 202 occurrences, followed by ' no reply from author' with 53 occurrences.

**Visualizations**

1. Frequency of Modalities by Sex (Bar Chart):



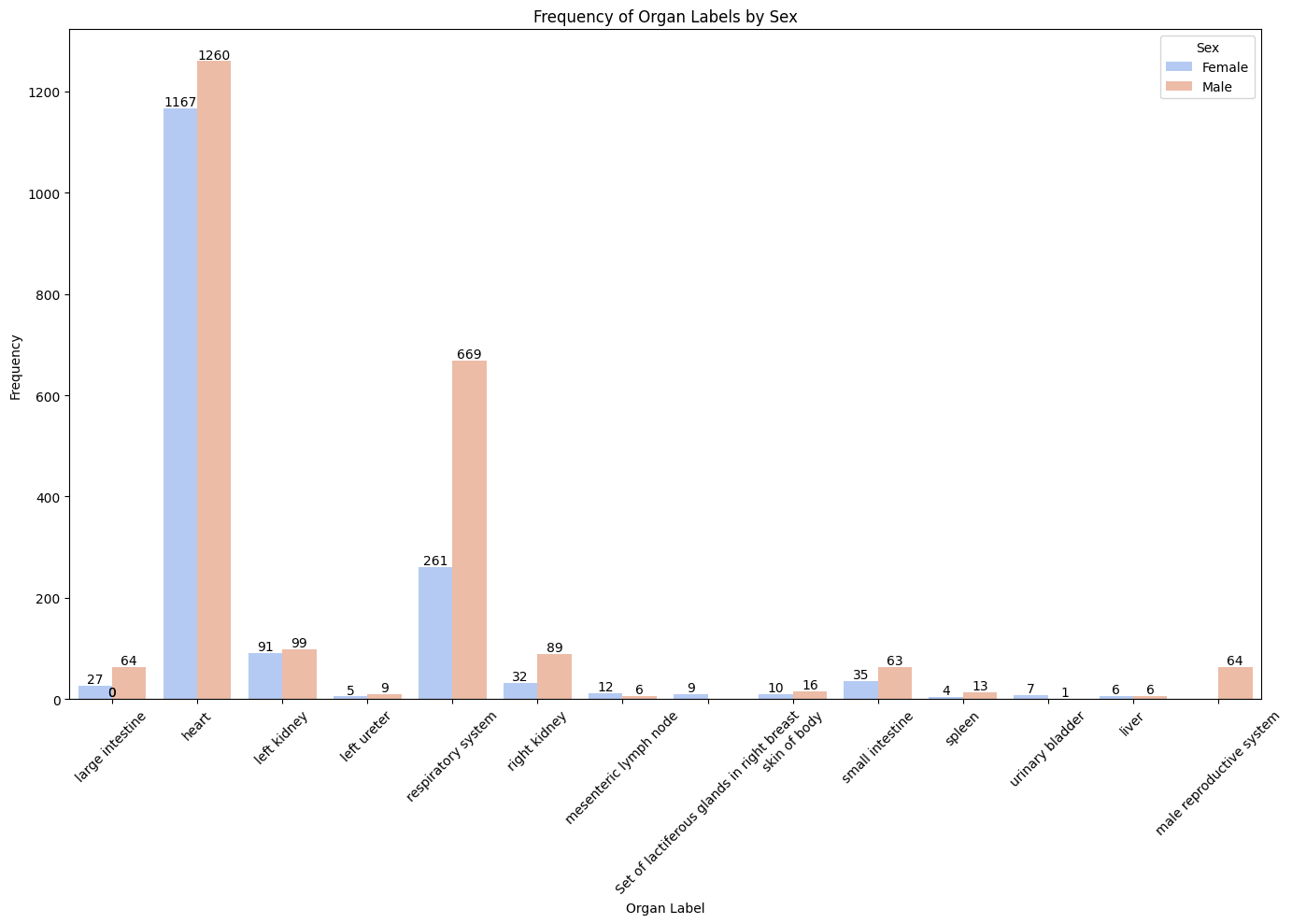
This chart presents a comparison of modalities (likely types of datasets or methods used in the studies) distinguished by sex.

The chart shows a significant disparity between the frequencies for each sex, suggesting potential biases or focus areas in the study design or population.

1. Frequency of Organ Labels by Sex (Bar Chart):

Similar to the first chart, this one compares the frequency of organ labels by sex.

It provides insight into which organs are most frequently studied and how that distribution differs between female and male labels.

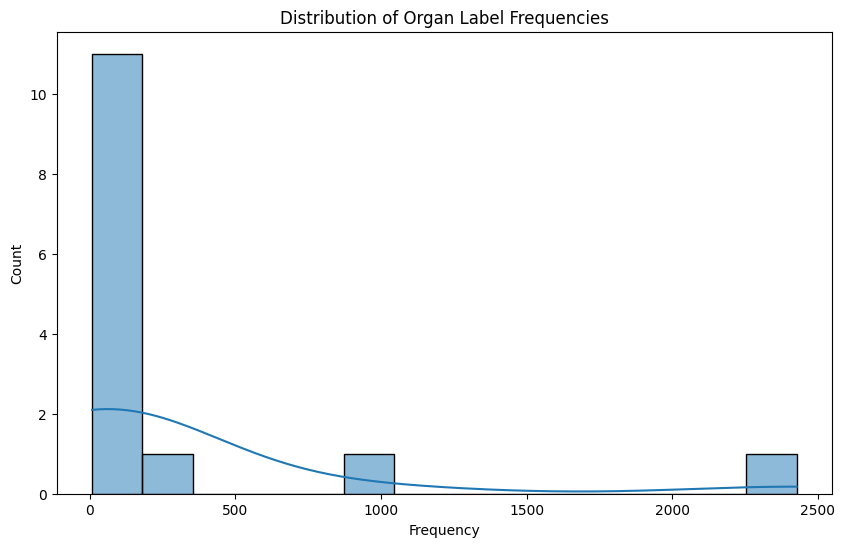


1. Distribution of Organ Label Frequencies (Histogram):

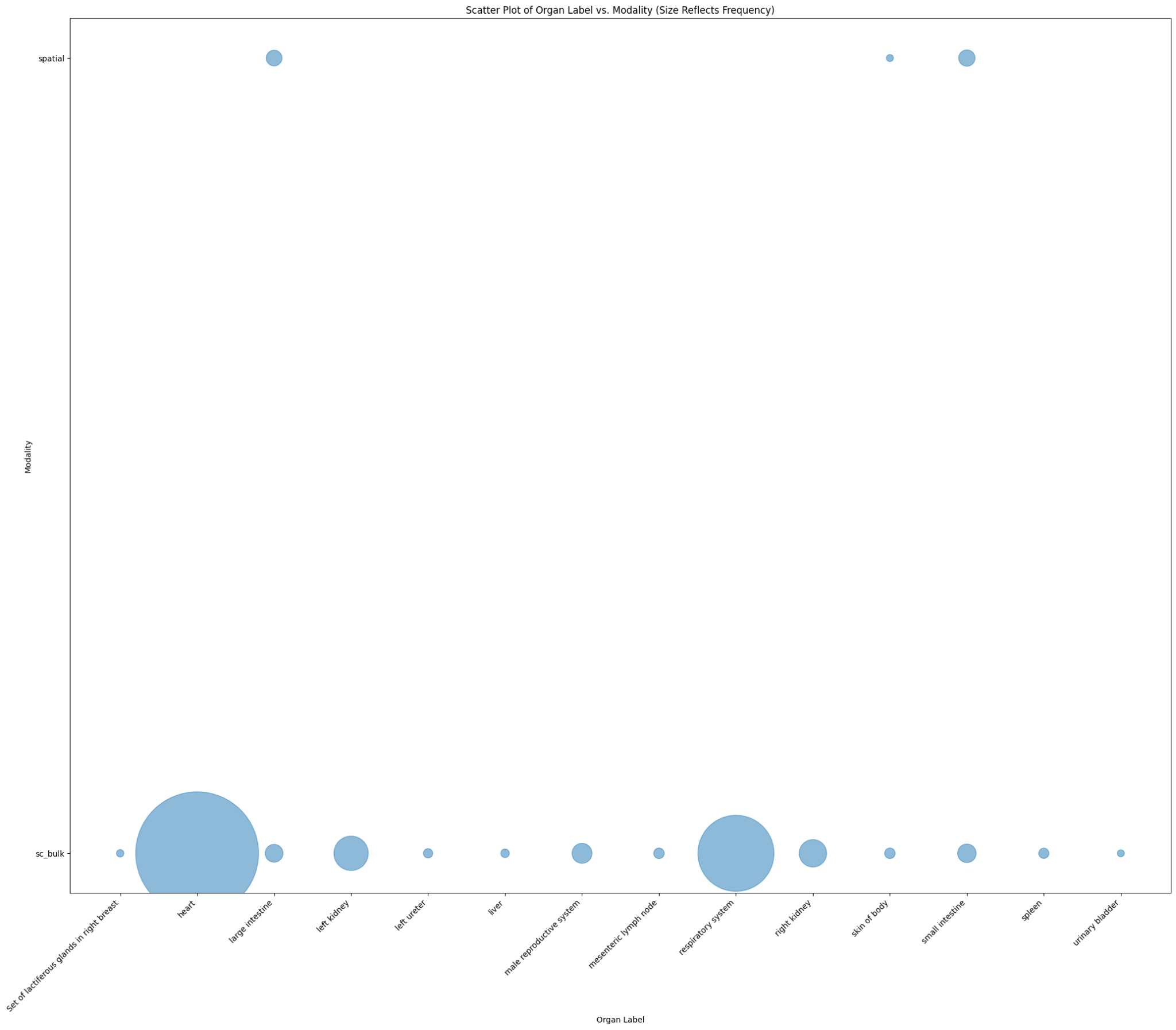
- This histogram displays the distribution of frequencies for different organ labels.

- Most organ labels have low frequencies, indicating a long-tail distribution where few organs are very commonly studied while most are not.

- It is the histogram with a large bar on the left (indicating a high frequency of low occurrence organ labels) and smaller bars to the right (indicating fewer organs with high occurrence). This kind of distribution is typical of a long-tail phenomenon, where many different organ labels are used infrequently, while a small number of organ labels are used very frequently.



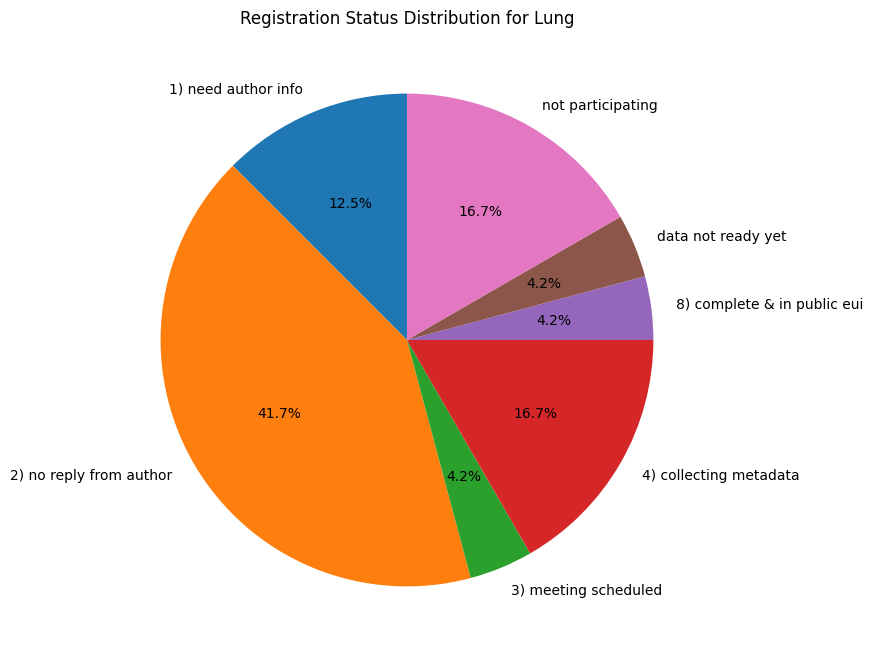
1. Scatter Plot of Organ Label vs. Modality (Size Reflects Frequency):



-The scatter plot illustrates the relationship between organ labels and modalities, with the size of the points reflecting the frequency.

-It helps in identifying which combinations of organ labels and modalities are most common.

1. Registration Status Distribution for Lung (Pie Chart):

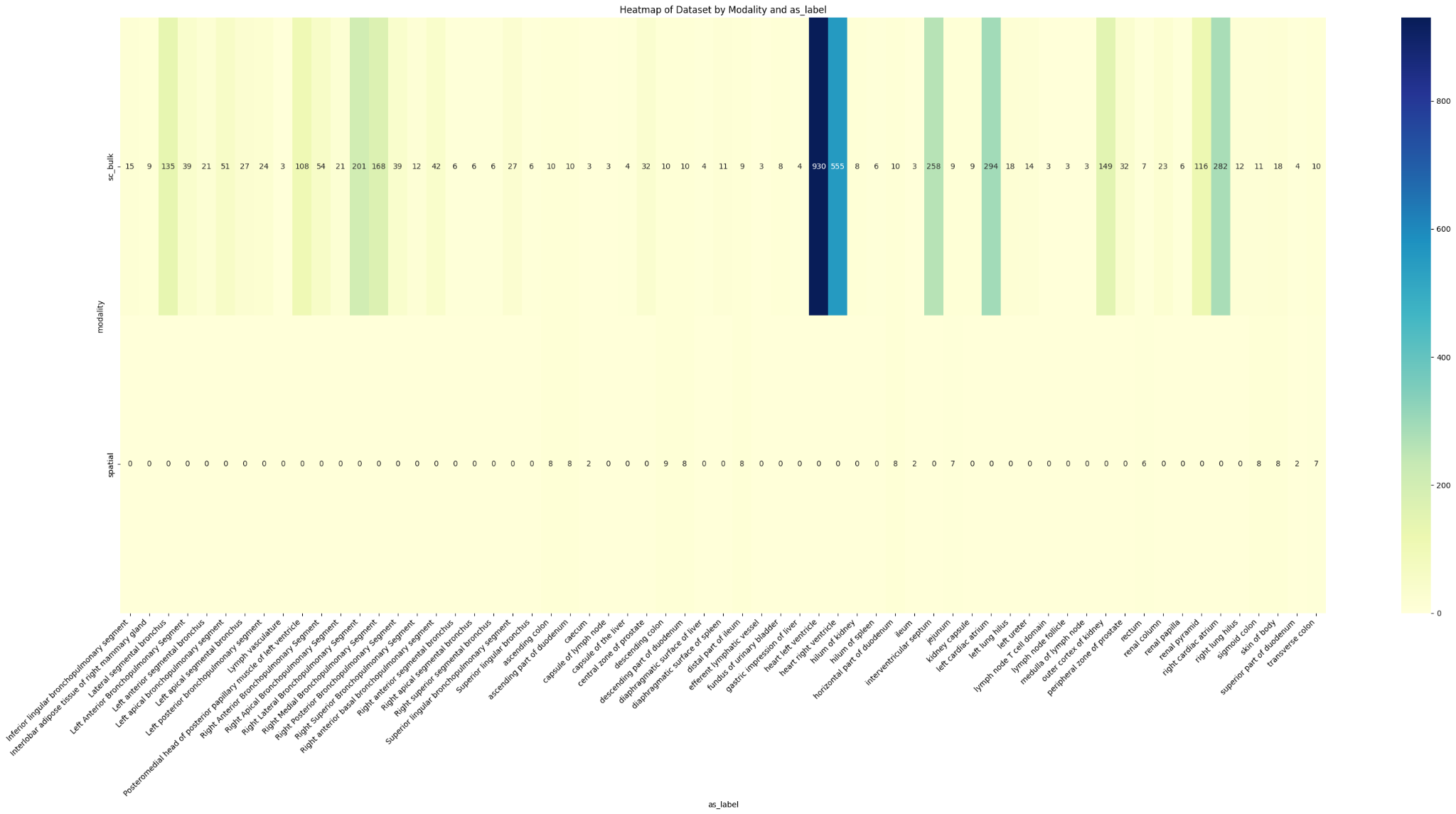


The pie chart shows the distribution of registration status for datasets related to the lung. It highlights the status of the data collection and processing, such as 'not participating', 'data not ready yet', and 'complete & in public eui'.

1. Heatmap of Dataset by Modality and as\_Label:

The heatmap visualizes the relationship between different modalities and labels, with color intensity representing the frequency.

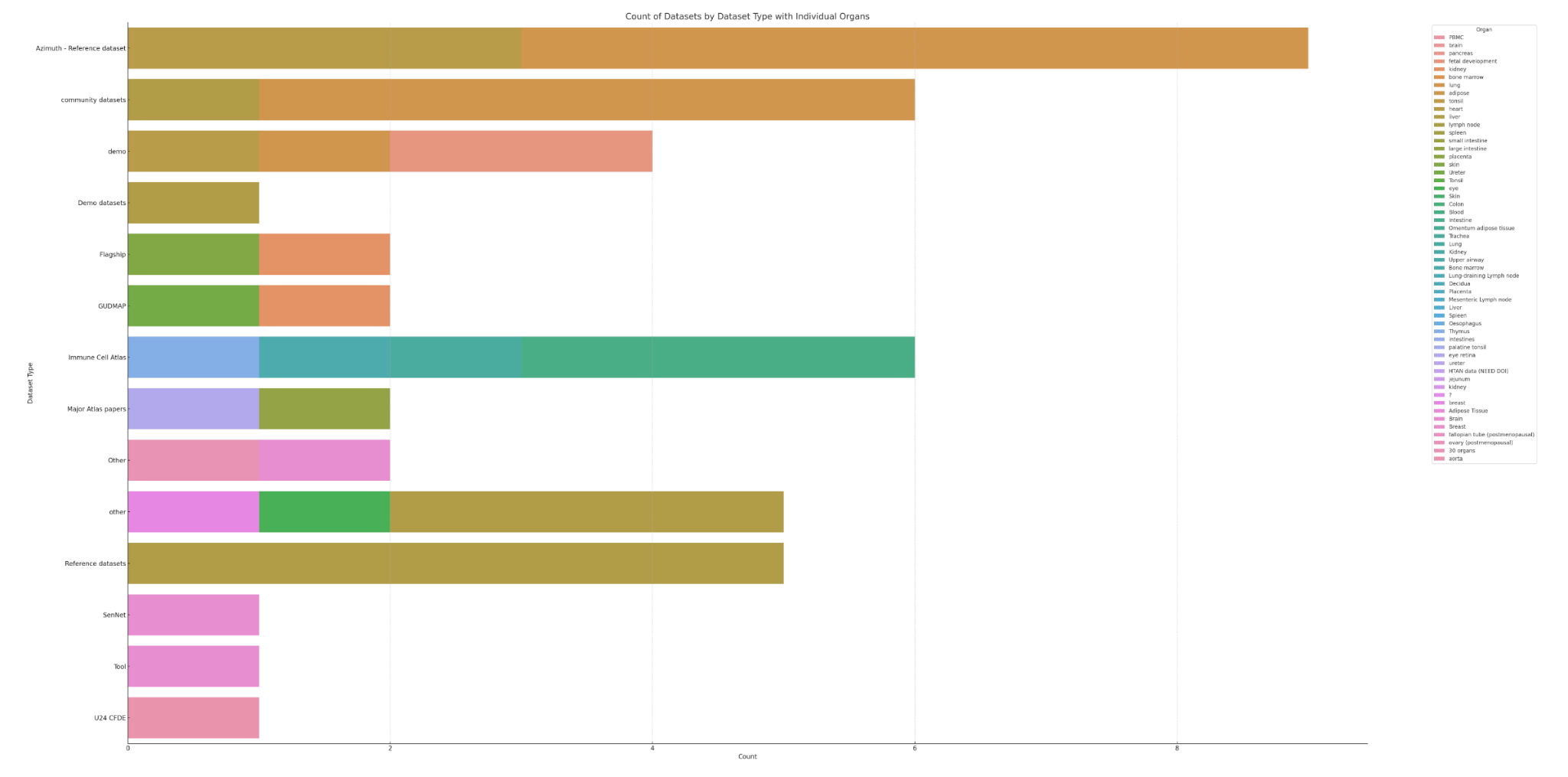
It can identify patterns or correlations between modalities and specific labels.



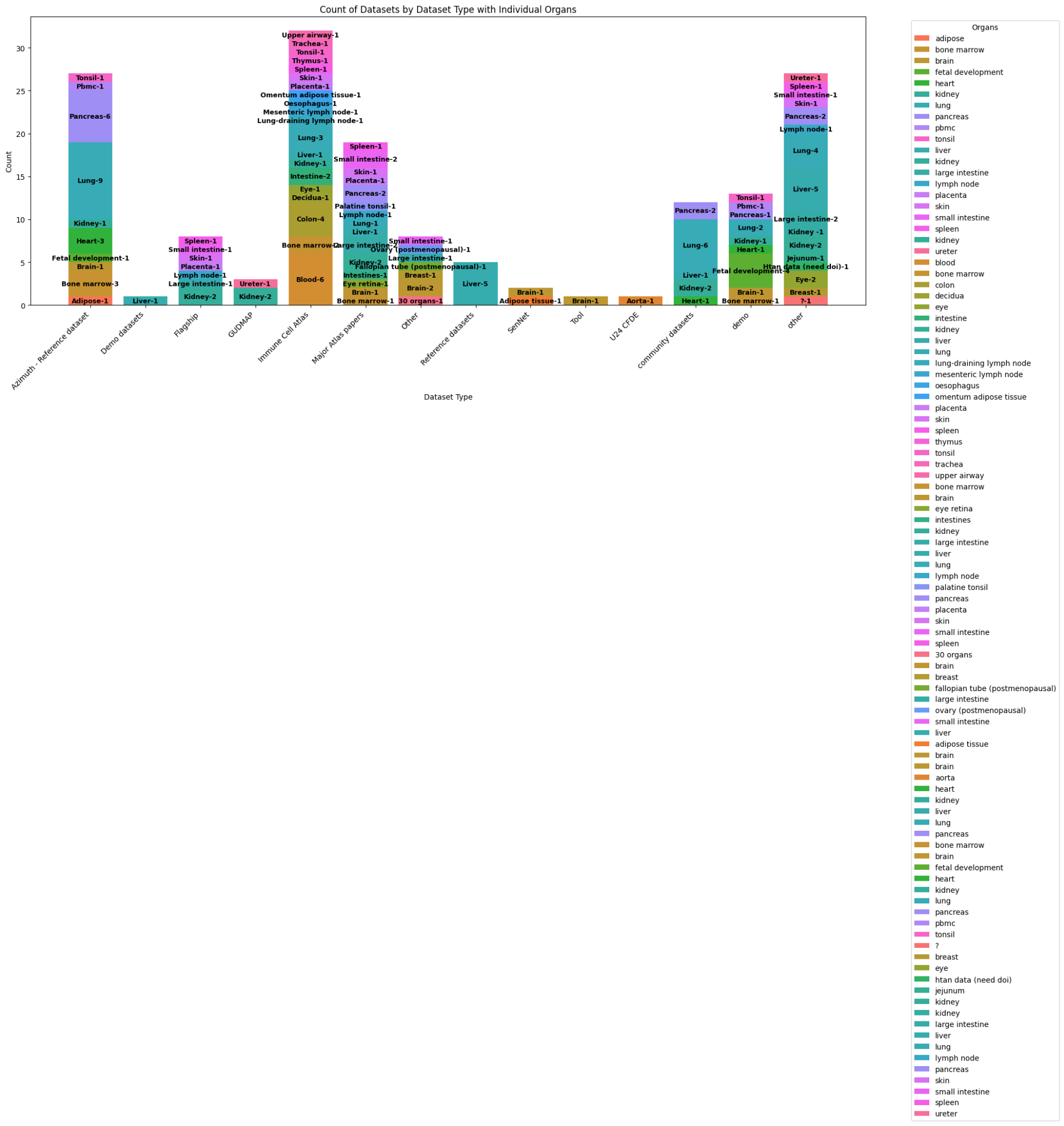
1. Count of Datasets by Dataset Type with Individual Organs (Stacked Bar Chart):

This stacked bar chart provides a count of datasets by dataset type, with colors representing individual organs.

It offers a comprehensive view of the dataset types' composition concerning the organs studied.



8. Bar Chart with Annotations (Unlabeled):

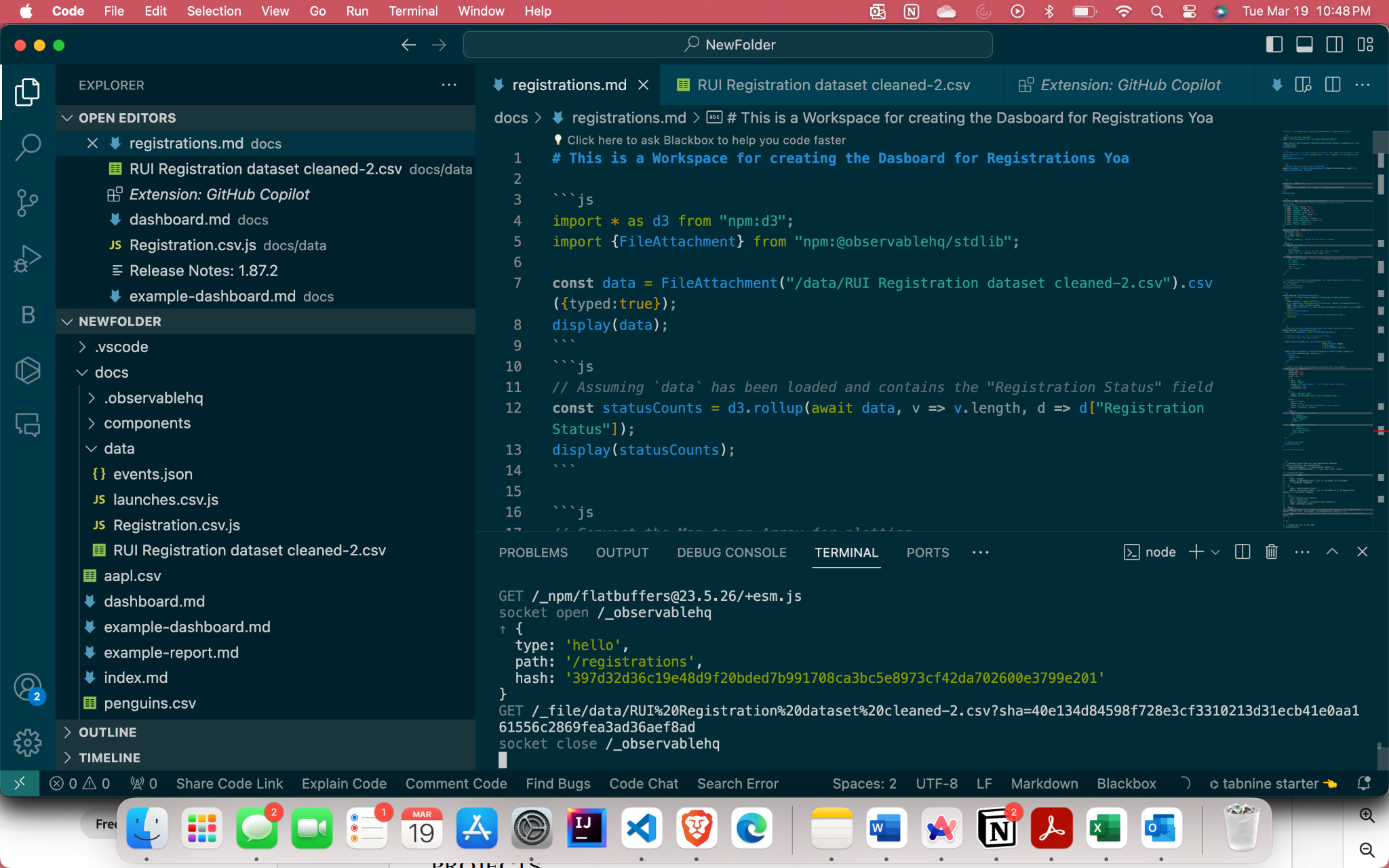


- The bar chart seems to represent the count of datasets by type, with annotations indicating specific counts for individual organs.

- It's similar to the previous chart but focuses on specific count annotations for clarity.

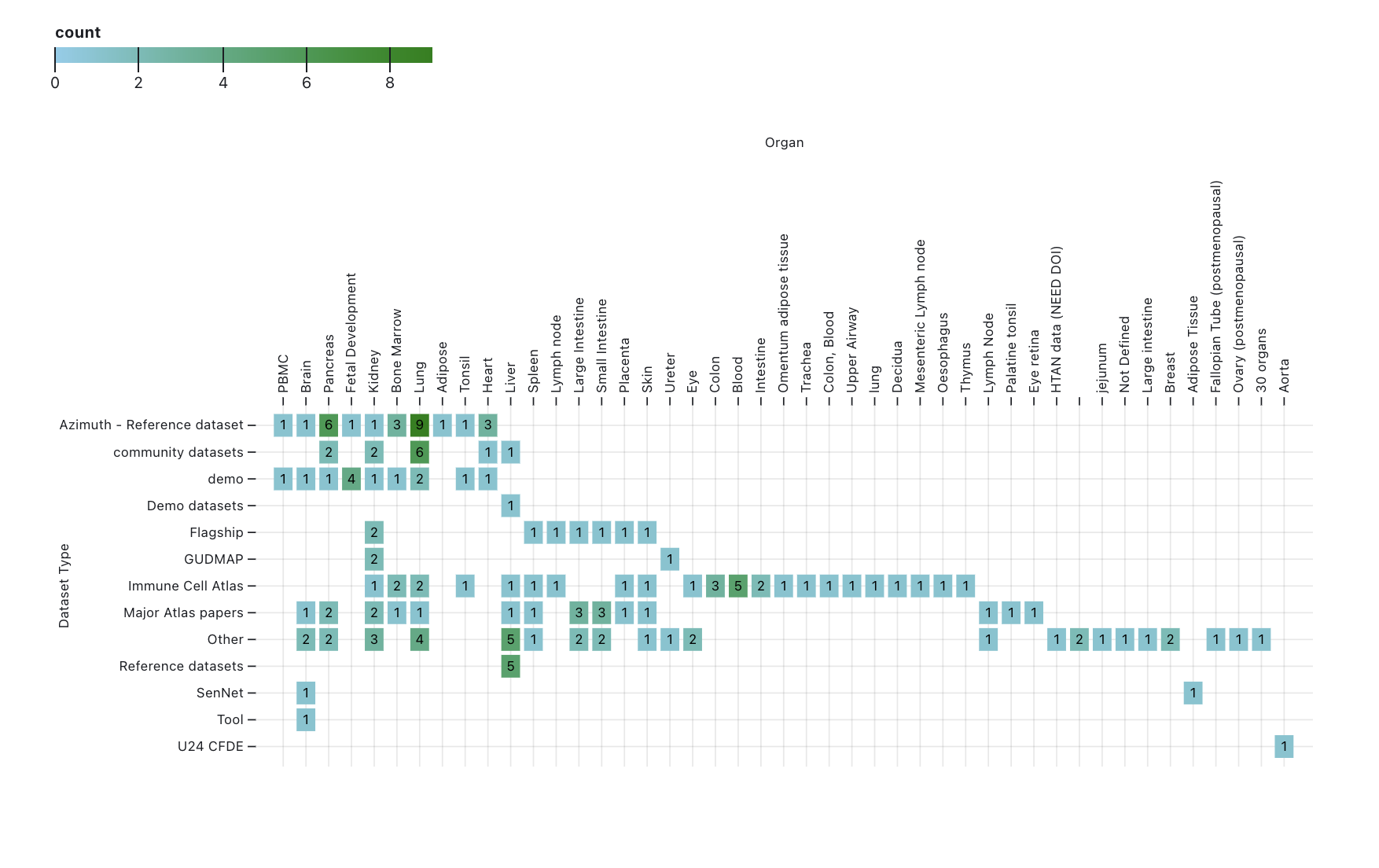
Week 11/03 -

So the past two weeks have been so informative in a learning way. I have been watching tutorials on additional D3 framework and how to come-up with new plots.



Over the past two weeks, considerable progress has been made in the development of the interactive dashboard using the JavaScript D3 framework, as per the specifications and expectations outlined by Dan in our preliminary meetings. The visualizations attached herein represent the latest iteration, demonstrating both the data representation and interactivity aspects. This document serves to provide a comprehensive update on the current status and milestones achieved in the dataset work and visualization development.

I have scripted the code for few visualizations, that didn’t show up quiet well for the first time. After cleaning the data with Dataset Type , Registration Status , Organs. Separating the rows withs combined organs in a cell(excel). It gives the entire organs count across each dataset type. Even though there few organs which are undefined like ‘30 organs’ , ‘not defined’ etc. After a meeting with Dan, organ labels are fetched from an API. the result can be seen below after figure(1).



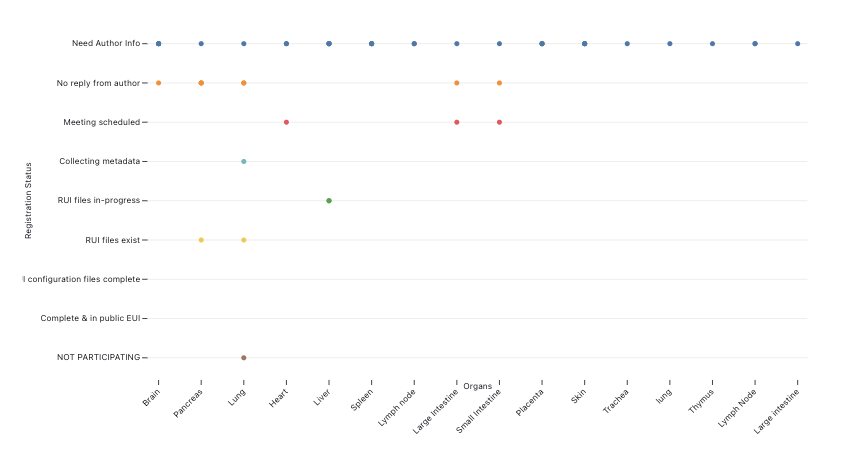
Figure(1)

Week 18/03-



Figure -2

This is new visualization with Dataset Type vs. Organ Distribution  
The first visualization illustrates the distribution of various datasets across different organs. This matrix-like chart enables users to discern patterns and focus areas within the dataset, notably the prominence of certain dataset types such as 'Flagship' and 'Major Atlas papers' in relation to specific organs like the kidney and liver.



Registration Status by Organ

The third visualization provides an overview of the dataset registration process status across various organs. Color-coded markers indicate different stages, from 'Meeting scheduled' to 'Complete & in public EUI.' This helps track progress and identify any bottlenecks.