Pan-Tissue Cellular Deconvolution Using Single-Cell RNA-Seq References

by

Tianyuzhou Liang

Submitted to the Graduate Faculty of

the School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2024

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Tianyuzhou Liang

It was defended on

April 10th, 2024

and approved by

Jiebiao Wang, PhD, Assistant Professor, Department of Biostatistics

Jeanine M Buchanich, PhD, Associate Professor, Department of Biostatistics

Christopher McKennan, PhD, Assistant Professor, Department of Statistics

Pan-Tissue Cellular Deconvolution Using Single-Cell RNA-Seq References

Tianyuzhou Liang, M.S.

University of Pittsburgh, 2024

Critical questions in biomedical research, such as disease mechanisms and biological processing, require an understanding of cell type proportions in heterogeneous tissues. Due to the complexity of measuring cellular fractions with traditional experimental methods, computational cellular deconvolution methods have been developed to estimate these fractions based on gene expression data. Previously, EnsDeconv, an R package that implements ensemble deconvolution by leveraging multiple deconvolution methods and scenarios, was developed and has been proven to provide a more accurate and robust method to deconvolve bulk gene expression data and estimate cellular fractions. To optimize the package's utility and create a comprehensive cellular deconvolution atlas for the entire human body, we aim to incorporate single-cell RNA sequencing (scRNA-seq) references to deconvolve bulk expression data spanning 43 tissue types into 192 distinct cell types. Using the EnsDeconv package, cellular fractions of 43 Genotype-Tissue Expression (GTEx) bulk samples were estimated based on the corresponding references curated from multiple large-scale scRNA-seq atlases, spanning over 60 datasets and 1.5 million cells. The usage of the estimated cellular fractions was demonstrated with our identified interesting associations between cellular fractions and covariates.

Table of Contents

List of Tables

List of Figures

Preface

I would like to express my deepest gratitude to my thesis advisor, Dr. Jiebiao Wang, for his invaluable guidance and expert advice throughout the development of this thesis. I also extend my sincere thanks to my academic advisor Dr. Jeanine M Buchanich, who has been supportive during my master's journey. I'm also grateful to my committee member Dr. Christopher McKennan for taking the time listening to my defense. Special thanks are due to Manqi Cai, whose assistance with EnsDeconv package was indispensable.

Last but certainly not least, I owe a special thank you to my cat, Omelette, whose unwavering support and companionship were vital during this challenging process. The emotional comfort and distraction provided by my feline friend were essential to my well-being and success.

1.0 Background

1.1 Bulk RNA-seq Resources

To understand critical biomedical questions, including disease mechanisms, developmental processes, treatment responses and so on, it is useful to investigate the gene expression profile in specific tissues of interest. Historically, bulk tissue RNA sequencing has been the method of choice for analyzing gene expression across entire tissue samples. For example, the GTEx portal has been a useful resource that provides bulk gene expression profiles across multiple human tissues, helping to elucidate the tissue-specific regulation of gene expression and its relationship to genetic variation [7].

However, the inherent cellular heterogeneity of tissues, which varies significantly across different organs, poses a challenge for this approach. Thus, although bulk RNA sequencing provides a composite gene expression profile of tissue samples, it lacks the resolution to offer cell type-specific insights, which is crucial to elucidate the underlying mechanisms of interest.

1.2 Single-cell RNA-seq Atlases

Single-cell RNA (scRNA)-sequencing is a popular technique to observe tissue gene expression at a single-cell level, allowing the identification of distinct cell types and states in the heterogeneous tissue sample. Human scRNA-seq atlases are crucial resources that offer comprehensive cellular maps of human tissues, significantly advancing our understanding of human biology. These atlases, rich in cellular and molecular details, are useful for a broad spectrum of biomedical research—ranging from elucidating fundamental biological processes and disease mechanisms to enhancing precision medicine and facilitating drug discovery.

Multiple efforts are underway to create an extensive map of the entire human body, while other projects are focusing on specific organs. A notable example is the Human BioMolecular Atlas Program (HuBMAP), which has integrated over 2300 datasets covering 30 organ types

to uncover intricate relationships among cells, tissues, and organ functions, both in health and disease states [10]. Another example, Allen Brain Cell Atlas, has a specialized focus on the human brain. It offers a detailed map of gene expression at the single-cell level and aids researchers in unraveling the intricacies of brain structure, functionality, and developmental processes.

The high-resolution cellular information and specific gene expression patterns within individual cells provide researchers with ample information to study the sample. However, such techniques are usually high in costs and labor [16].

1.3 Cellular Deconvolution

Current popular methods to quantify cell types include fluorescence-activated cell sorting (FACS) and scRNA-seq techniques. However, each technique has its drawbacks. Isolating cell types through FACS is generally based on markers selected before experiments, which limits the number of cell types and thus sorting might not be detailed [16]. ScRNA-seq, on the other hand, usually enriches specific cell types, thus might be biased as the fraction of cells might not represent the whole cell population in the tissue sample [19].

To overcome the limitations of the cell type analysis of heterogeneous compositions in bulk tissue, cellular deconvolution techniques have been developed. These methods aim to estimate the proportions of specific cell types within the heterogeneous tissue sample, offering a detailed and unbiased understanding of tissue composition.

Cellular deconvolution methods can be broadly categorized into reference-based and reference-free approaches. Reference-based deconvolution methods, as the name suggests, require a reference gene expression profile of specific cell types to deconvolve the bulk gene expression data. Reference-free deconvolution, on the other hand, does not require such reference. Instead, algorithms are developed to infer cell types and their proportion directly from the bulk gene expression data. In this thesis, reference-based deconvolution method is used, thus a detailed introduction is included for reference-based deconvolution methods.

Reference-based cellular deconvolution is essentially framed as a linear mixture model:

$$Y_{G\times S} = B_{G\times K} \times P_{K\times S} + E \tag{1}$$

where Y is the gene expression matrix derived from a bulk sample, encompassing G genes across S samples. B is the average gene expression matrix obtained from the reference gene expression profile for G genes across K distinct cell types. P is the cellular fraction matrix, which denotes the proportion of each of the K cell types within the S samples. E stands for noise or error inherent in the data.

Mathematical models for reference-based cellular deconvolution methods aim to minimize the noise and find the optimal cellular proportion matrix P that, when combined with average gene expression matrix B, approximates the bulk gene expression matrix Y.

The model is also subjected to two biological constraints: 1) non-negativity, as cellular fraction cannot be a negative value; and 2) sums to one, as the proportion of all cell types in one tissue sample should be 1, representing the complete cellular makeup of the sample.

Our effort involves the construction of a tissue atlas based on the deconvolution of bulk gene expression data from 45 diverse tissues obtained from the GTEx portal. We incorporated scRNA-sequencing references from multiple sources to provide the detailed cellular fractions of these organs, offering medical researchers critical insights into the cellular composition of various organs. In addition, we also aim to construct a library of tissue-specific cellular signature matrices, which will allow researchers who have gene expression data from bulk tissues to estimate cellular fraction effortlessly and accurately.

2.0 Data Usage and Methods

2.1 Data Description

Bulk RNA-sequencing data from 45 tissues were obtained from GTEx (Table 1). Each tissue was sequenced for over 50,000 genes. A total of 948 donors have at least one tissue selected for bulk RNA sequencing. Phenotypic data of the donors were also obtained from the GTEx data portal.

scRNA-sequencing data were obtained from multiple public resources, major sources include GTEx [7], HuBMAP (Azimuth) [10], Human Cell Landscape [9], Tabula Sapiens [24], UCSC Cell Browser [22], Broad Institute Single Cell Portal [23], CellxGene [13], and Human Cell Atlas [20](Table 2).

2.2 Ensemble Deconvolution (EnsDeconv)

In this thesis, the EnsDeconv R package is used to deconvolve bulk RNA-sequencing data from the GTEx portal. EnsDeconv is a cellular deconvolution method developed to implement an ensemble over multiple deconvolution results, providing an optimal estimate of cellular fractions from bulk gene expression data [3].

A general step of reference-based deconvolution includes 1) identifying a reference, 2) normalizing bulk and single-cell gene expression data, 3) transforming data,4) selecting marker genes, and 5) applying deconvolution algorithms. In EnsDeconv, each of the 5 steps could be characterized as a parameter. Each unique combination of the five parameters forms a deconvolution scenario. After performing deconvolution in each, the EnsDeconv algorithm performs cell-type-specific (CTS) robust regression to synthesize fraction estimates from all scenarios and outputs an optimal cellular fraction [3].

Table 1: Descriptive statistics of GTEx donors

	Male (N=636)	**Female** (N=312)	**Overall** (N=948)
Donor Cohort			
Organ Donor (OPO)	262 (41.2%)	157 (50.3%)	419 (44.2%)
Postmortem	362 (56.9%)	153 (49.0%)	515 (54.3%)
Surgical	12 (1.9%)	2 (0.6%)	14 (1.5%)
Age (years)			
Mean (SD)	53.0 (12.9)	52.3 (13.1)	52.8 (12.9)
Median [Min, Max]	56 [20, 70]	54 [21, 70]	55 [20, 70]
Race			
Asian	7 (1.1%)	5 (1.6%)	12 (1.3%)
Black or African American	81 (12.7%)	41 (13.1%)	122 (12.9%)
White	539 (84.7%)	265 (84.9%)	804 (84.8%)
American Indian or Alaska Native	2 (0.3%)	0 (0%)	2 (0.2%)
Unknown	7 (1.1%)	1 (0.3%)	8 (0.8%)
Ethnicity			
Not Hispanic or Latino	289 (45.4%)	154 (49.4%)	443 (46.7%)
Hispanic or Latino	16 (2.5%)	3 (1.0%)	19 (2.0%)
Unknown	331 (52.0%)	155 (49.7%)	486 (51.3%)
Height (in)			
Mean (SD)	70.1 (2.85)	64.5 (2.65)	68.2 (3.84)
Median [Min, Max]	70.0 [56.0, 78.0]	64.7 [57.0, 72.0]	69.0 [56.0, 78.0]
Weight (lb)			
Mean (SD)	193 (32.9)	159 (30.0)	182 (35.7)
Median [Min, Max]	197 [86.0, 293]	158 [90.0, 248]	180 [86.0, 293]
BMI			
Mean (SD)	27.6 (4.00)	26.8 (4.28)	27.3 (4.11)
Median [Min, Max]	27.6 [18.6, 35.4]	26.6 [17.0, 35.9]	27.3 [17.0, 35.9]

Table 2: Summary of scRNA-sequencing reference datasets

Organ System	Tissue	Bulk RNA-seq sample size	scRNA-seq reference	# of cell types	# of samples
Cardiovascular	Artery - Aorta	432	Human Cell Landscape	19	9,652
	Artery - Coronary	240	Tabula Sapiens	13	4,867
	Heart - Atrial Appendage	429	GTEx	15	36,574
	Heart - Left Ventricle	432	Human Cell Landscape, Tabula Sapiens	21	6,012
Respiratory	Lung	578	GTEx, Human Cell Landscape, Tabula Sapiens	16	95,017
Nervous	Brain - Amygdala	152	Human Brain Cell Atlas	11	187,225
	Brain - Anterior cingulate cortex (BA24)	176	Human Brain Cell Atlas	17	32,157
	Brain - Caudate (basal ganglia)	246	Human Brain Cell Atlas	11	32,678
	Brain - Cerebellum	456	Human Brain Cell Atlas, Human Cell Landscape	11	7,324
	Brain - Cortex	464	Human Brain Cell Atlas	16	31,065
	Brain - Hippocampus	197	Human Brain Cell Atlas	12	276,997
	Brain - Hypothalamus	202	Human Brain Cell Atlas	13	78,963
	Brain - Nucleus accumbens (basal ganglia)	246	Human Brain Cell Atlas	11	30,132
	Brain - Putamen (basal ganglia)	205	Human Brain Cell Atlas	10	34,416
	Brain - Spinal cord (cervical c-1)	159	Human Brain Cell Atlas	13	24,190
	Brain - Substantia nigra	139	Human Brain Cell Atlas	11	59,505
Skin	Adipose - Subcutaneous	663	Tabula Sapiens	14	9,892
	Adipose - Visceral (Omentum)	541	Human Cell Landscape	18	12,812
	Skin - Not Sun Exposed (Suprapubic)	604	Tabula Sapiens	21	2,007
	Skin - Sun Exposed (Lower leg)	701	GTEx	17	5,327
Musculoskeletal	Muscle - Skeletal	803	GTEx, Tabula Sapiens	17	35,758
Blood	Spleen	241	Human Cell Landscape, Tabula Sapiens	23	49,434
	Whole Blood	755	Human Cell Landscape, Tabula Sapiens	25	67,445
Digestive	Colon - Sigmoid	373	Human Cell Landscape	15	2,813
	Colon - Transverse	406	Human Cell Landscape	20	13,046
	Esophagus - Mucosa	555	GTEx	18	26,060
	Esophagus - Muscularis	515	GTEx	17	34,173
	Liver	226	Human Cell Landscape, Tabula Sapiens	23	15,310
	Minor Salivary Gland	162	Tabula Sapiens	23	27,199
	Small Intestine - Terminal Ileum	187	Human Cell Landscape	19	3,081
	Stomach	359	Human Cell Landscape	26	13,434
Endocrine	Adrenal Gland	258	Human Cell Landscape	18	23,197
	Pancreas	328	Human Cell Landscape, Tabula Sapiens	16	23,033
	Pituitary	283	Yan et al. 2024. Genome Med. [26]	4	5,361
	Thyroid	653	Human Cell Landscape	25	12,587
Urinary	Kidney - Cortex	85	Human Cell Landscape, Tabula Sapiens	30	32,548
Male reproductive	Prostate	245	GTEx, Tabula Sapiens	18	47,436
	Testis	361	Guo et al. 2018. Cell Res. [8]	12	6,500
Female reproductive	Breast - Mammary Tissue	459	GTEx, Tabula Sapiens	14	21,145
	Ovary	180	Fan et al. 2019. Nature. [4]	9	21,000
	Uterus	142	Human Cell Landscape, Tabula Sapiens	19	7,693
	Vagina	156	Li et al. 2021. Nature Comm. [11]	5	81,026

2.2.1 Normalization Methods

Normalization is a fundamental pre-processing step for RNA-seq gene expression analyses. In general, normalization allows accurate biological comparisons across samples. Some key reasons to normalize gene expression data before analysis include correcting for sequencing depth and library size, standardizing data for downstream analyses, and enhancing data quality and interpretability.

In this thesis project, counts per million (CPM) is used as the normalization method for both bulk and single-cell gene expression data. In the CPM normalization method, each raw gene read count is divided by the total read count for the sample and then multiplied by one million.

$$CPM = \frac{\text{Raw Count}}{\text{Total Counts in Sample}} \times 10^6 \tag{2}$$

Sequencing depth, which is the total number of reads in each sample, could vary across samples due to differences in sequencing efficiency, such as those observed between bulk and single-cell sequencing techniques, and variations in library preparation. CPM adjusts this difference and allows reliable comparison between samples.

2.2.2 Transformation Methods

Transformation is another pre-processing step for gene expression data following normalization. Here, either linear or log transformation is performed for the normalized data.

Linear transformation preserves the relationships between variables before the transformation and minimizes the distortion of the distribution of the data, thus the original structure of the data is maintained with linear transformation and is more biologically plausible. However, depending on the original data, the structure could be skewed, and normality and homoscedasticity assumptions might not be met. Log transformation, on the other hand, reduces the skewness and stabilizes the variance.

2.2.3 Marker Gene Selection Methods and Cellular Signature Matrices

Gene expression data is a large dataset in which as many as fifty thousand genes could be sequenced. To reduce the computational burden of such a large dataset, dozens of representative genes, the marker genes, are selected to perform deconvolution. The marker genes are selected through differential expression analysis. The genes that show significant differences in expression between different groups of cells are then selected as marker genes for further analyses.

In this thesis, pairwise Welch-t tests and a method that combines marker statistics from Welch t-tests and Wilcoxon rank sum tests were used for differential expression analysis.

2.2.4 Cellular Deconvolution Methods

In this thesis, four different reference-based deconvolution methods were used, including CIBERSORT, which uses Support Vector Regression (SVR) algorithm that implements e-sensitive loss and L2 regularizer [17]; Estimating the Proportions of Immune and Cancer cells (EPIC) that uses weighted least squares [18]; Digital Cell Quantification (DCQ) that uses quadratic loss and elastic net that combines L1 and L2 penalties [1]; and DeconRNASeq that uses non-negative least squares decomposition algorithms [6].

2.2.5 Ensemble Learning

Ensemble learning is a machine learning method that combines the results from multiple training models that are designed to solve the same problem to achieve an optimal result. The basic idea is to combine several weak learners so that they become a strong learner.

In EnsDeconv, the estimates from each scenario are assumed to mostly resemble true cellular proportions, however subjected to outliers. Thus, robust regression, which is the model commonly used when there are many outliers in the dataset, is incorporated into EnsDeconv to synthesize optimal cellular fraction results.

2.3 Association Tests

The obtained mean cellular abundance was tested for association with phenotypic data from GTEx. The phenotype data from the GTEx portal included demographic data, donor death information, and medical histories.

The association of estimated cellular fractions in each tissue was tested with the phenotypic covariates with either a two-sample t-test (for categorical covariates) or linear regression (for continuous covariates).

3.0 Result

3.1 Diverse Distribution of Mean Cellular Abundances across Different Tissues

Figures 1, 2 and 3 depict the mean cellular abundance across various human tissues, deconvolved from bulk RNA-seq data using distinct scRNA-seq reference datasets. The heatmap visualization demonstrates a diverse landscape of cell types, with marked tissue-specific abundance. For instance, ventricle cardiomyocyte (Figure 2) and cardiac muscle cells (Figure 3), which are essentially the same type of cells but with different names, show a pronounced presence in left ventricular heart tissue, aligning with their known biological function to perform heart contractions and then pump blood into the artery. Similarly, pancreas exocrine cells (Figure 2) and pancreas acinar and ductal cells (Figure 3) are found only in pancreatic tissue, indicating their vital role in pancreas function.

Conversely, certain immune cells, such as myeloid cells and macrophages, display a broad tissue distribution, suggesting a systemic role in immune surveillance. Endothelial cells are also widely distributed in most tissues, reflecting the essential vascularization for organ functionalities.

The clustering patterns of tissues also agree with the biological organ system. For instance, the cellular composition of whole blood and spleen are closely correlated as they both belong to the blood system (Figure 2); the stomach and ileum are also similar as they both belong to the digestive system. Clusters are also found in tissues that are not in the same organ, such as arteries and colon. However, they both exhibit a high abundance of smooth muscle cells, which supports involuntary muscle movements.

Within an organ system, there is still a huge variation in cellular composition in different tissues. Figure 4 shows an example of a complex organ system—the nervous system. Eleven different tissues from the brain are shown here. Vascular cells are found in all brain tissues with similar abundance as they support blood circulation in the brain and provide vital support for brain functions. Astrocytes and oligodendrocytes are commonly found in all brain

tissues as these are the essential components of brain functions, with some tissues enriched in these cells. For example, astrocyte has a higher abundance in the nucleus accumbens compared to other brain regions. Given the nucleus accumbens' critical role in reward, motivation, and addiction, astrocyte function may be critical in understanding the neural basis of these processes as well as being potential therapeutic targets for disorders such as addiction and depression. On the other hand, specialized functional cells like cerebellar inhibitory cells ab Bergmann glial cells are only found with a high abundance in the cerebellum.

3.2 Cellular Fractions Varies with Age and Sex

The association between cellular fraction calculated with GTEx scRNA-seq reference and sex (Figure 5) and age (Figure 6) was also assessed.

In breast mammary tissue, we observed a higher abundance of myoepithelial cells in females, which are integral to the structure of mammary glands, lining the ducts to facilitate milk expulsion (Figure 5). This is consistent with the biological functions of the mammary tissue in females, where active mammary glands are present. Additionally, females exhibited a higher presence of adipocytes in both muscle and skin tissues, aligning with the general observation that females tend to have a higher body fat composition than males. Intriguingly, skin tissue in females showed a lower fraction of sebaceous cells compared to males, potentially correlating with variations in sebum production and skin hydration between the sexes.

Regarding age-related changes, cornification—a specialized form of programmed cell death in epithelial cells leading to the formation of the protective stratum corneum—showed an increased trend in skin tissues with age (Figure 6). This process contributes to the skin's barrier function but may also reflect age-related increases in skin dryness and reduced cellular turnover. Notably, there is a general decline in immune cell abundance in older subjects, aligning with the well-documented trend of diminishing immune function with age. Additionally, a decrease in the fractions of fibroblasts and pericytes in lung tissue could suggest age-related changes in lung structure, potentially impacting lung elasticity and function.

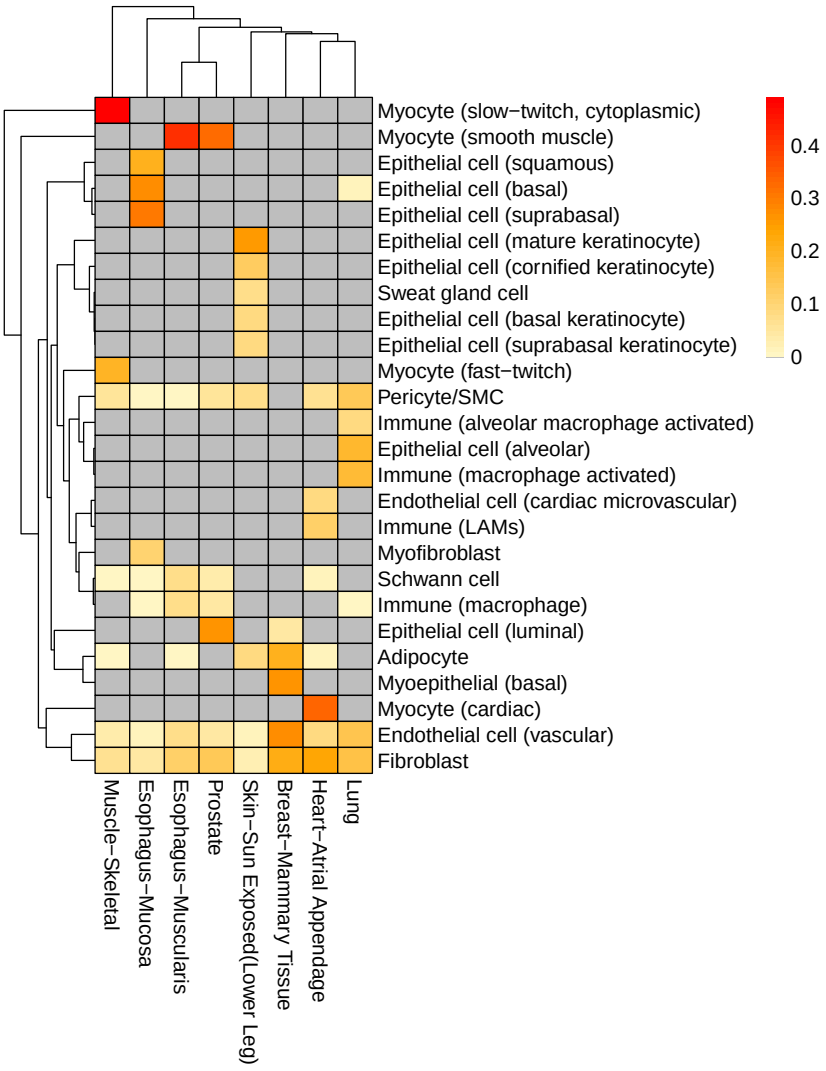

Figure 1: Mean cellular abundance in 8 human tissues using GTEx scRNA-seq reference. Mean cellular abundance deconvolved from GTEx bulk RNA-sequencing data using ensemble deconvolution method via EnsDeconv R package. Single-cell RNA-sequencing reference data was obtained from the GTEx. A total of 26 cell types (with mean abundance greater than 5% in at least one tissue) across 8 tissue types were shown.

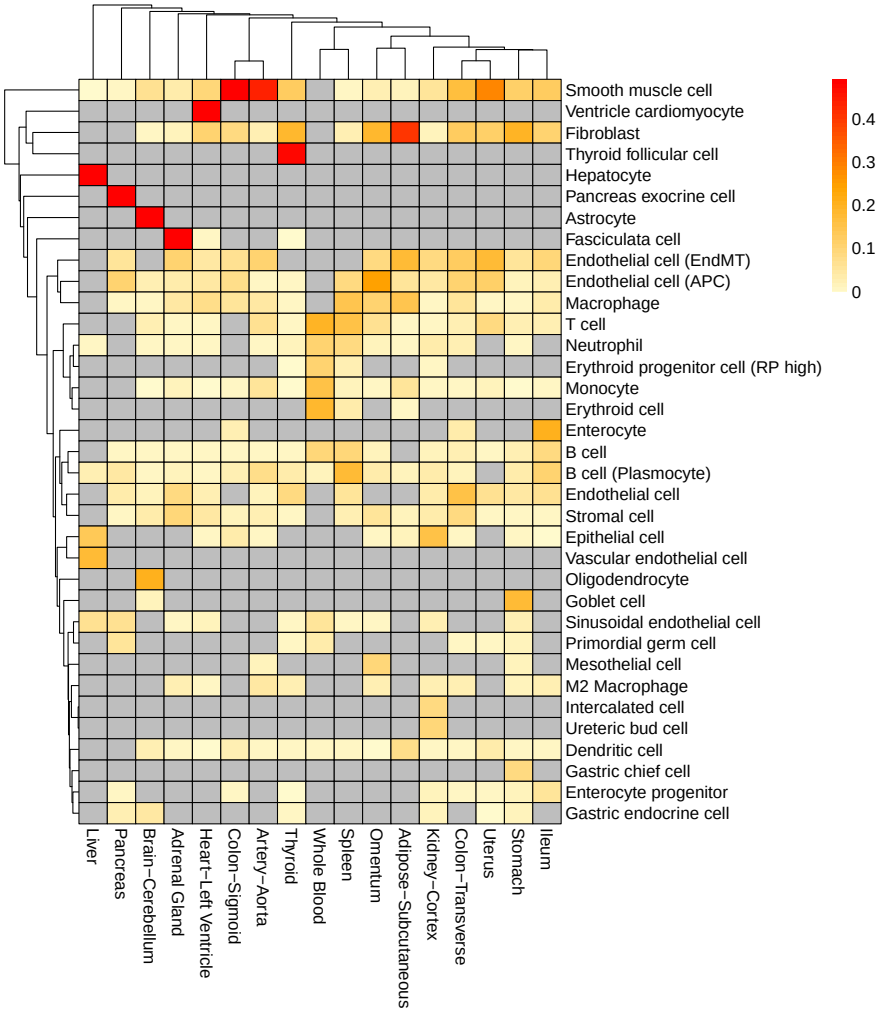

Figure 2: Mean cellular abundance in 17 human tissues using Human Cell Landscape scRNA-seq reference. Mean cellular abundance deconvolved from GTEx bulk RNA-sequencing data using ensemble deconvolution method via EnsDeconv R package. Single-cell RNA-sequencing reference data was obtained from the Human Cell Atlas. A total of 34 cell types (with mean abundance greater than 5% in at least one tissue) across 17 tissue types were shown.

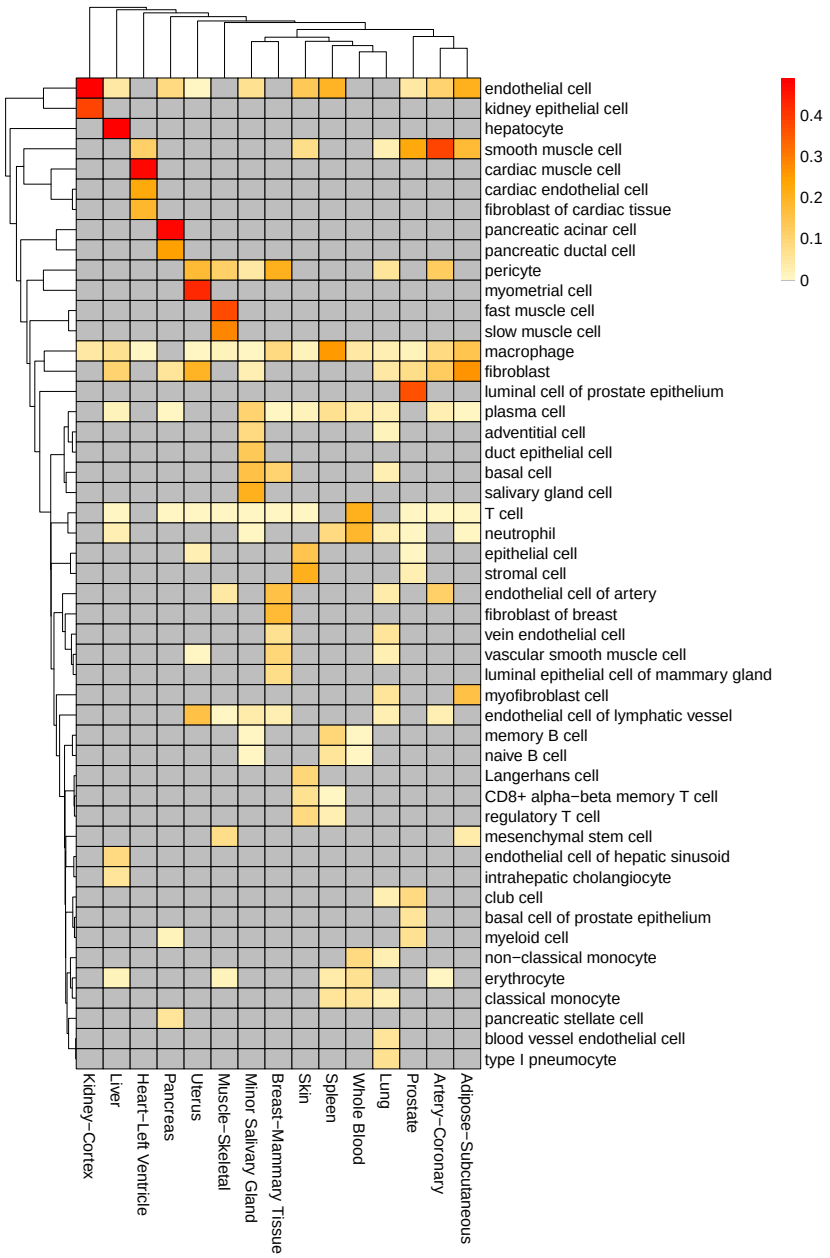

Figure 3: Mean cellular abundance in 15 human tissues using Tabula Sapiens scRNA-seq reference. Mean cellular abundance deconvolved from GTEx bulk RNA-sequencing data using ensemble deconvolution method via EnsDeconv R package. Single-cell RNA-sequencing reference data was obtained from Tabula Sapiens. A total of 49 cell types (with mean abundance greater than 5% in at least one tissue) across 15 tissue types were shown.

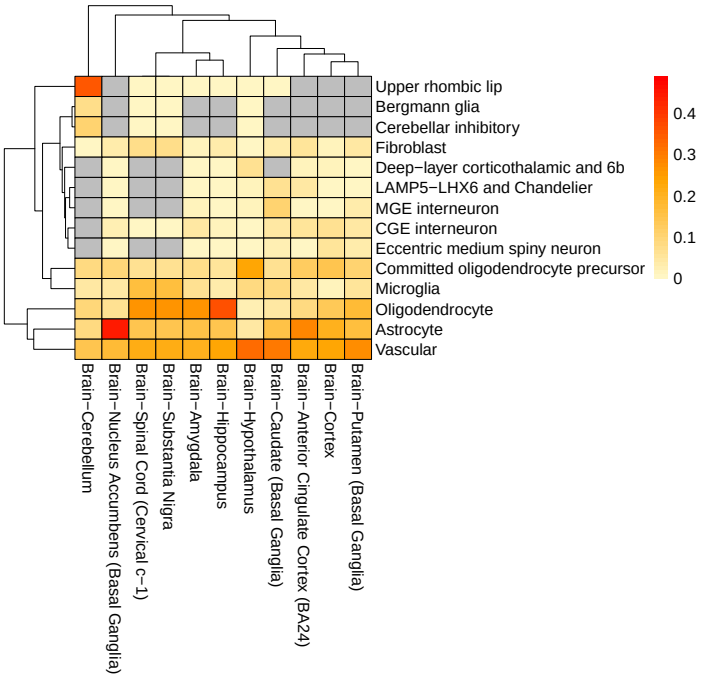

Figure 4: Mean cellular abundance in 11 human brain tissues using Human Brain Cell Atlas scRNA-seq reference. Mean cellular abundance deconvolved from GTEx bulk RNA-sequencing data using ensemble deconvolution method via EnsDeconv R package. Single-cell RNA-sequencing reference data was obtained from the Human Brain Cell Atlas. A total of 14 supercluster types (with mean abundance greater than 5% in at least one cell type) across 11 tissue types were shown.

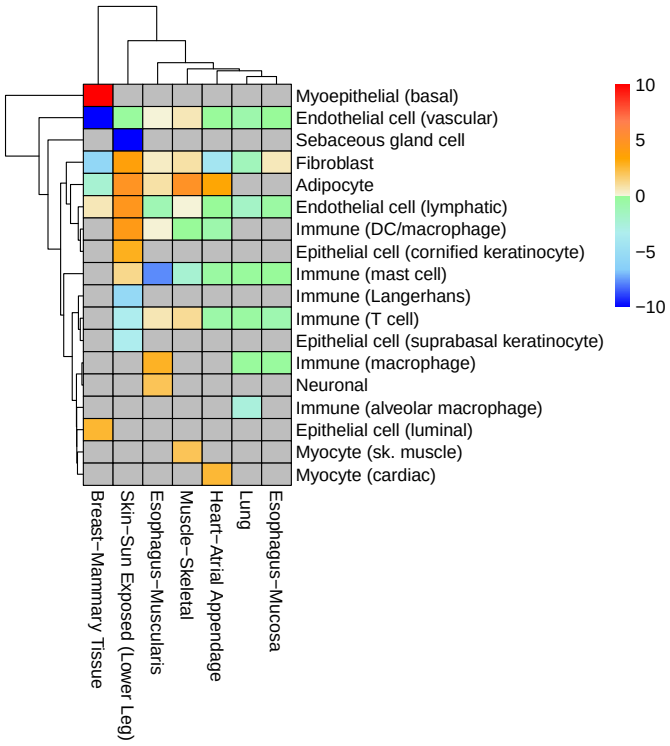

Figure 5: P-value heatmap of association between cellular fractions and sex in 7 tissues. The cellular fraction was estimated by EnsDeconv using scRNA-sequencing reference data obtained from the GTEx portal. Two sample t-tests were performed to assess associations. P-values are log_{10}-transformed with direction added. The male was denoted as 0, and the female was denoted as 1.

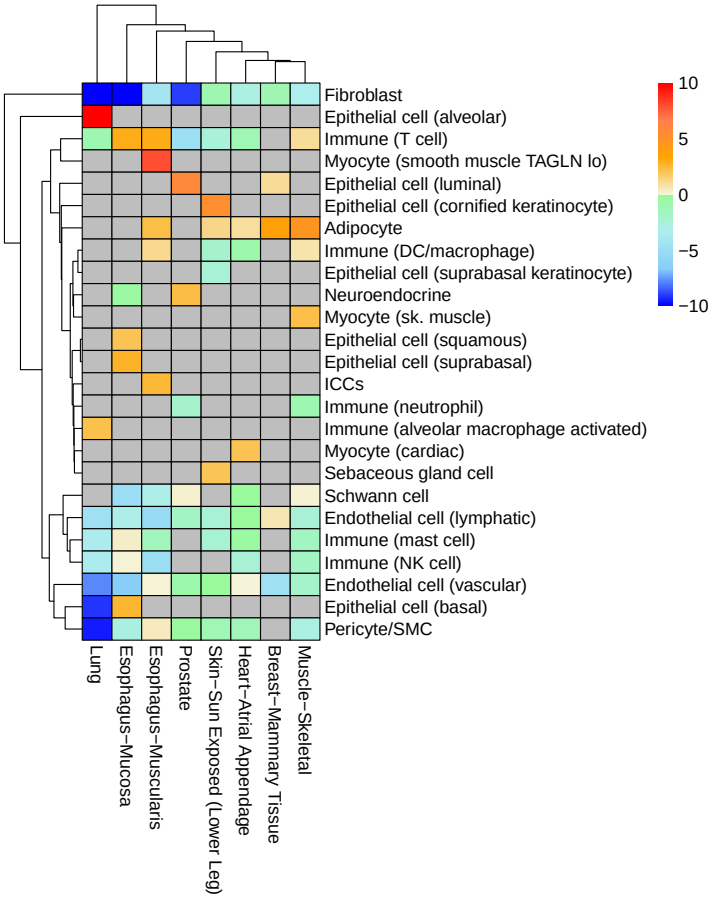

Figure 6: P-value heatmap of association between cellular fractions and age in 8 tissues. The cellular fraction of 8 tissues was estimated by EnsDeconv using scRNA-sequencing reference data obtained from the GTEx portal. Linear regression was performed to assess associations. P-values are log_{10}-transformed with direction added.

4.0 Discussion and Conclusion

This project resulted in the generation of an atlas that includes tissue-specific cellular fractions spanning tissues from the whole human body. The result showed that EnsDeconv can estimate cellular fractions from bulk RNA-seq data using scRNA-seq reference with a biologically explainable result. Our results also showed different regions in one organ can vary quite differently in cellular composition, giving biologically meaningful insights to researchers who would have a special interest in a particular region of an organ.

However, the estimated cellular proportion depends quite heavily on the scRNA-seq reference, as the cell clustering methods vary across different references, and there are still several improvements that could be made to the atlas.

The first challenge is to unify cell type nomenclature across various scRNA-seq reference datasets Due to the diversity of sources, each dataset employs distinct clustering methods, resulting in varying names and levels of specificity for identified cell types. This discrepancy complicates the integration of multiple references within EnsDeconv and the optimization of results. For example, in Figure 2, smooth muscle cells are identified in various tissues, including the uterus. Smooth muscle cells are specialized for involuntary muscle movements, e.g., in the uterus, they support uterine contractions. However, in Figure 3, the term "myometrial cells" is used, which specifically refers to uterine smooth muscle cells. This inconsistency in labeling, which occurs in several tissues, poses significant obstacles to constructing a comprehensive human tissue atlas. There are other ongoing efforts to solve this problem, notably the Human Reference Atlas. Their proposal of the ASCT+B (anatomical structures, cell types, and biomarkers) table might shed light on unifying the nomenclature of cell types and aid in the construction of a comprehensive, and universal map of the whole human body [2]. Allen Institute also proposed a nomenclature format specifically for mammalian brain cell types [14].

Another factor affecting the accuracy of cellular deconvolution is the variability in cell size [12]. Cell sizes can vary widely, ranging from approximately $30\mu m^3$ in sperm cells and $100\mu m^3$ in red blood cells, to massive fat cells at 600,000 μm^3, and even larger oocytes at

$4{,}000{,}000\mu m^3$ [15]. Except for some cell types like adipocytes, most cell types fall within their characteristic size range. Within a single tissue, such as the brain, there is a notable diversity in cell sizes—neurons tend to be larger than both glial and vascular cells, which support the nervous system and blood circulation, respectively [5]. Furthermore, transcriptional activity can differ between cell types. For instance, neurons exhibit higher levels of gene expression compared to glial cells. Cellular deconvolution methods often assume uniform cell size and gene expression activity across all cell types within a tissue, which can lead to inaccuracies in estimating cell proportions. To address these disparities, one proposed method is to implement a cell type-specific scale factor transformation [21]. This approach adjusts the cellular signature matrix to account for the differences in cell size, potentially refining the accuracy of cell proportion estimates in deconvolution analyses.

In the next step, we will address the issues discussed above and extend the framework to cover more tissues beyond GTEx, such as bulk samples saved in recount3 [25] and Gene Expression Omnibus. Our eventual goal is to provide a cell deconvolution atlas that is user-friendly and does not require users to process scRNA-seq data.

Bibliography

[1] Zeev Altboum, Yael Steuerman, Eyal David, Zohar Barnett-Itzhaki, Liran Valadarsky, Hadas Keren-Shaul, Tal Meningher, Ella Mendelson, Michal Mandelboim, Irit Gat-Viks, and Ido Amit. Digital cell quantification identifies global immune cell dynamics during influenza infection. *Molecular Systems Biology*, 10(2), 2014-02-28.

[2] Katy Börner, Sarah A. Teichmann, Ellen M. Quardokus, James C. Gee, Kristen Browne, David Osumi-Sutherland, Bruce W. Herr, Andreas Bueckle, Hrishikesh Paul, Muzlifah Haniffa, Laura Jardine, Amy Bernard, Song-Lin Ding, Jeremy A. Miller, Shin Lin, Marc K. Halushka, Avinash Boppana, Teri A. Longacre, John Hickey, Yiing Lin, M. Todd Valerius, Yongqun He, Gloria Pryhuber, Xin Sun, Marda Jorgensen, Andrea J. Radtke, Clive Wasserfall, Fiona Ginty, Jonhan Ho, Joel Sunshine, Rebecca T. Beuschel, Maigan Brusko, Sujin Lee, Rajeev Malhotra, Sanjay Jain, and Griffin Weber. Anatomical structures, cell types and biomarkers of the human reference atlas. *Nature Cell Biology*, 23(11):1117–1128, 2021.

[3] Manqi Cai, Molin Yue, Tianmeng Chen, Jinling Liu, Erick Forno, Xinghua Lu, Timothy Billiar, Juan Celedón, Chris Mckennan, Wei Chen, and Jiebiao Wang. Robust and accurate estimation of cellular fraction from tissue omics data via ensemble deconvolution. *Bioinformatics*, 38(11):3004–3010, 2022.

[4] X. Fan, M. Bialecka, I. Moustakas, E. Lam, V. Torrens-Juaneda, N. V. Borggreven, L. Trouw, L. A. Louwe, G. S. K. Pilgram, H. Mei, L. Van Der Westerlaken, and S. M. Chuva De Sousa Lopes. Single-cell reconstruction of follicular remodeling in the human adult ovary. *Nature Communications*, 10(1), 2019.

[5] Francisco J Garcia, Na Sun, Hyeseung Lee, Brianna Godlewski, Hansruedi Mathys, Kyriaki Galani, Blake Zhou, Xueqiao Jiang, Ayesha P Ng, Julio Mantero, et al. Single-cell dissection of the human brain vasculature. *Nature*, 603(7903):893–899, 2022.

[6] Ting Gong and Joseph D. Szustakowski. Deconrnaseq: a statistical framework for deconvolution of heterogeneous tissue samples based on mrna-seq data. *Bioinformatics*, 29(8):1083–1085, 2013. OA status: bronze.

[7] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020.

[8] Jingtao Guo, Edward J Grow, Hana Mlcochova, Geoffrey J Maher, Cecilia Lindskog, Xichen Nie, Yixuan Guo, Yodai Takei, Jina Yun, Long Cai, et al. The adult human testis transcriptional cell atlas. *Cell research*, 28(12):1141–1157, 2018.

[9] Xiaoping Han, Ziming Zhou, Lijiang Fei, Huiyu Sun, Renying Wang, Yao Chen, Haide Chen, Jingjing Wang, Huanna Tang, Wenhao Ge, et al. Construction of a human cell landscape at single-cell level. *Nature*, 581(7808):303–309, 2020.

[10] HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature*, 574(7777):187–192, 2019.

[11] Yaqian Li, Qing-Yang Zhang, Bao-Fa Sun, Yidi Ma, Ye Zhang, Min Wang, Congcong Ma, Honghui Shi, Zhijing Sun, Juan Chen, et al. Single-cell transcriptome profiling of the vaginal wall in women with severe anterior vaginal prolapse. *Nature Communications*, 12(1):87, 2021.

[12] Sean K. Maden, Sang Ho Kwon, Louise A. Huuki-Myers, Leonardo Collado-Torres, Stephanie C. Hicks, and Kristen R. Maynard. Challenges and opportunities to computationally deconvolve heterogeneous tissue with varying cell sizes using single-cell rna-sequencing datasets. *Genome Biology*, 24(1), 2023.

[13] Colin Megill, Bruce Martin, Charlotte Weaver, Sidney Bell, Lia Prins, Seve Badajoz, Brian McCandless, Angela Oliveira Pisco, Marcus Kinsella, Fiona Griffin, et al. Cellxgene: a performant, scalable exploration platform for high dimensional sparse matrices. *bioRxiv*, pages 2021–04, 2021.

[14] Jeremy A Miller, Nathan W Gouwens, Bosiljka Tasic, Forrest Collman, Cindy TJ van Velthoven, Trygve E Bakken, Michael J Hawrylycz, Hongkui Zeng, Ed S Lein, and Amy Bernard. Common cell type nomenclature for the mammalian brain. *Elife*, 9:e59928, 2020.

[15] Ron Milo, Paul Jorgensen, Uri Moran, Griffin Weber, and Michael Springer. Bionumbers—the database of key numbers in molecular and cell biology. *Nucleic acids research*, 38(suppl_1):D750–D753, 2010.

[16] Brian B Nadel, Meritxell Oliva, Benjamin L Shou, Keith Mitchell, Feiyang Ma, Dennis J Montoya, Alice Mouton, Sarah Kim-Hellmuth, Barbara E Stranger, Matteo Pellegrini, and Serghei Mangul. Systematic evaluation of transcriptomics-based deconvolution methods and references using thousands of clinical samples. *Briefings in Bioinformatics*, 22(6), 2021.

[17] Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, Ash A Alizadeh, Aaron M Newman, Chih Long Liu, Michael R Green, Andrew J Gentles, Weiguo Feng, Yue Xu, Chuong D Hoang, Maximilian Diehn, and Ash A Alizadeh. Robust enumeration of cell subsets from tissue expression profiles. *Nature Methods 2015 12:5*, 12(5), 2015-03-30.

[18] Julien Racle, Kaat de Jonge, Petra Baumgaertner, Daniel E Speiser, and David Gfeller. Simultaneous enumeration of cancer and immune cell types from bulk tumor gene expression data. *eLife*, 6, 2017.

[19] Xianwen Ren, Boxi Kang, and Zemin Zhang. Understanding tumor ecosystems by single-cell sequencing: promises and limitations. *Genome biology*, 19(1):211, 2018.

[20] Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.

[21] Olukayode A Sosina, Matthew N Tran, Kristen R Maynard, Ran Tao, Margaret A Taub, Keri Martinowich, Stephen A Semick, Bryan C Quach, Daniel R Weinberger, Thomas M Hyde, et al. Strategies for cellular deconvolution in human brain rna sequencing data. *bioRxiv*, pages 2020–01, 2020.

[22] Matthew L Speir, Aparna Bhaduri, Nikolay S Markov, Pablo Moreno, Tomasz J Nowakowski, Irene Papatheodorou, Alex A Pollen, Brian J Raney, Lucas Seninge, W James Kent, et al. Ucsc cell browser: visualize your single-cell data. *Bioinformatics*, 37(23):4578–4580, 2021.

[23] Leyla Tarhan, Jon Bistline, Jean Chang, Bryan Galloway, Emily Hanna, and Eric Weitz. Single cell portal: an interactive home for single-cell genomics data. *bioRxiv*, 2023.

[24] The Tabula Sapiens Consortium. The tabula sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594), 2022.

[25] Christopher Wilks, Shijie C Zheng, Feng Yong Chen, Rone Charles, Brad Solomon, Jonathan P Ling, Eddie Luidy Imada, David Zhang, Lance Joseph, Jeffrey T Leek, et al. recount3: summaries and queries for large-scale rna-seq expression and splicing. *Genome biology*, 22:1–40, 2021.

[26] Nan Yan, Weiyan Xie, Dongfang Wang, Qiuyue Fang, Jing Guo, Yiyuan Chen, Xinqi Li, Lei Gong, Jialin Wang, Wenbo Guo, et al. Single-cell transcriptomic analysis

reveals tumor cell heterogeneity and immune microenvironment features of pituitary neuroendocrine tumors. *Genome Medicine*, 16(1):2, 2024.