Distribution Agreement

In presenting this thesis as a partial fulfillment of the requirements for a degree from Emory University, I hereby grant to Emory University and its agents the non-exclusive license to archive, make accessible, and display my thesis in whole or in part in all forms of media, now or hereafter now, including display on the World Wide Web. I understand that I may select some access restrictions as part of the online submission of this thesis. I retain all ownership rights to the copyright of the thesis. I also retain the right to use in future works (such as articles or books) all or part of this thesis.

Boqi Wang April 7, 2023

# Developing enrichment analyzing methods at sub-cell type level to generate novel insights on complex disease pathogenesis

by

Boqi Wang

Zhaohui Qin
Adviser

Biology

Zhaohui Qin

Adviser

Michal Arbilly

Committee Member

David Cutler

Committee Member

2023

# Developing enrichment analyzing methods at sub-cell type level to generate novel insights on complex disease pathogenesis

By

Boqi Wang

Zhaohui Qin

Adviser

An abstract of
a thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2023

Abstract

Developing enrichment analyzing methods at sub-cell type level to generate novel insights on complex disease pathogenesis

By Boqi Wang

The development of biotechnologies and the consequent high throughput experiments have led to an urgent need to utilize such an enormous amount of biomedical data. It is necessary to develop bioinformatics tools that perform gene enrichment analysis at the sub-cell type level in complex diseases and traits for the derivation of disease etiology and the development of new treatment strategies.

In this study, we tackled the problem using newly emerged single-cell gene expression data and developed two approaches to accurately identify affected cell types in specific diseases. The first approach builds logistic regression models using cell type-specific marker genes, and the second one utilizes expression quantitative trait loci (eQTLs) colocalization and target gene read proportions in single nuclei RNA sequencing (snRNA-seq) data. The cell types are ranked based on the significance of cell types' associations with diseases. The central hypothesis is that most disease-associated genes are expressed preferentially in affected cell types. The two methods take advantage of newly emerged single-cell gene expression data from hECA and GEO of NCBI. Other types of biomedical big data like eQTLs from GTEx and disease-associated genes from DisGeNET were utilized as well.

Our approach has presented significantly more accurate results. Various cell type-disease combinations were revealed for 916 diseases and traits while some suggested potential explanations for disease pathogenesis. The results showed great consistency with previous findings. Overall, our methods have shown great potential in uncovering novel pathogenesis mechanisms of complex diseases. In-depth analysis and experimental validation are required to fully understand these discovered tissue-trait associations and their enriched genes.

# Developing enrichment analyzing methods at sub-cell type level to generate novel insights on complex disease pathogenesis

By

Boqi Wang

Zhaohui Qin

Adviser

A thesis submitted to the Faculty of Emory College of Arts and Sciences
of Emory University in partial fulfillment
of the requirements of the degree of
Bachelor of Science with Honors

Biology

2023

Acknowledgements

I would like to acknowledge my deepest thanks to my advisor and committee chair Dr. Zhaohui Qin, who has spent much time guiding me in my research for the past two years. I am also grateful for Dr. Michal Arbilly and Dr. David Cutler for agreeing to be on my committee and providing valuable supports and suggestions throughout the project.

Table of Contents

Lists of Tables and Figures

## 1. Project 1: Loci2path

As biotechnology and analytical tools develop rapidly, investigating biomedical big data has become an increasingly popular topic in life science. Genome-wide association studies (GWAS) have generated such a large quantity of data via high throughput experiments that there exists a consistent need for applications and tools to integrate and analyze the accumulating biomedical datasets [1]. The extensive quantity of data along with the development of biotechnology have also markedly decreased the difficulties of accessing data on public health as we can freely access a large variety of biomedical datasets from projects like Genotype-Tissue Expression (GTEx) and Encyclopedia of DNA Elements (ENCODE) [2-3]. Considering all goals in exploring the big data, identifying target genes critical in the pathogenesis of complex diseases and traits is arguably the most popular focus due to the high demand for treatments for diseases. The development of cutting-edge bioinformatic tools and methods to determine functional genomics could further our understanding of gene functions and what roles they play in disease etiology [4]. However, one major challenge in developing such tools and methods is integrating different types of datasets like function genomics and curated biochemical pathways into one analyzing frame.

Expression quantitative trait loci (eQTLs) have attracted the attention of many researchers among all types of biomedical big data due to their nature of regulating their target gene's expression. Thus, eQTL is a great tool for analyzing associations between genetic variants and complex diseases. Past studies have proven that using eQTLs to connect genomic loci and target genes (eGenes) has significantly higher accuracy than using genomic proximity, corroborating the functionality of eQTLs in determining gene expression [5]. Other studies have used tissue-specific

eQTLs to perform gene enrichment analysis on specific diseases and traits like nephrotic syndrome and fasting glucose [6-7].

Previously, we have extended an R package named loci2path that uses eQTLs localization near genomic variants to locate eGenes and perform enrichment analysis for ten diseases and traits at the gene pathway level [8]. Loci2path identifies eGenes based on tissue-specific eQTLs sets within the input disease-associated genomic intervals and maps them onto gene pathway sets, in which we use Fisher's exact test adjusted by Benjamini & Hochberg correction method to calculate p-values representing significance of the association between the gene pathway and tissue for the input disease [9]. Our results have led to some interesting findings and generated novel hypotheses on pathogenesis of complex human diseases, and one example would be Alzheimer's disease (AD).

Currently, there are three major pathology divisions for AD: protein accumulation, neuron loss, and reactive process [10]. Past studies have shown that the extracellular accumulation and deposition of amyloid-beta (Aβ) protein induce the appearance of senile plaques and create an abnormal neuron environment, which causes cognitive disabilities [11-12]. Such accumulation of Aβ not only enhances the interaction between amyloid-forming protein and neuronal membrane and increases membrane permeability through hypothetical mechanisms like amyloid-forming protein's channel-like conductance, but also contributes to the increase in the reactive oxygen species production and thus the disruption of neuronal membrane integrity [11, 13].

Figure 1A demonstrated the eQTLs enrichment of AD-related genomic intervals in the BioCarta pathway set. There was a distinctly significant enrichment of the D4-GDI pathway in the brain amygdala (Figure 1A). D4-GDI represents the negative regulator of Ras-related Rho GTPases, and its removal is crucial to induce apoptosis since Rho GTPases increase the cytoskeletal and membrane modification related to apoptosis [14]. As an enzyme that cleaves D4-

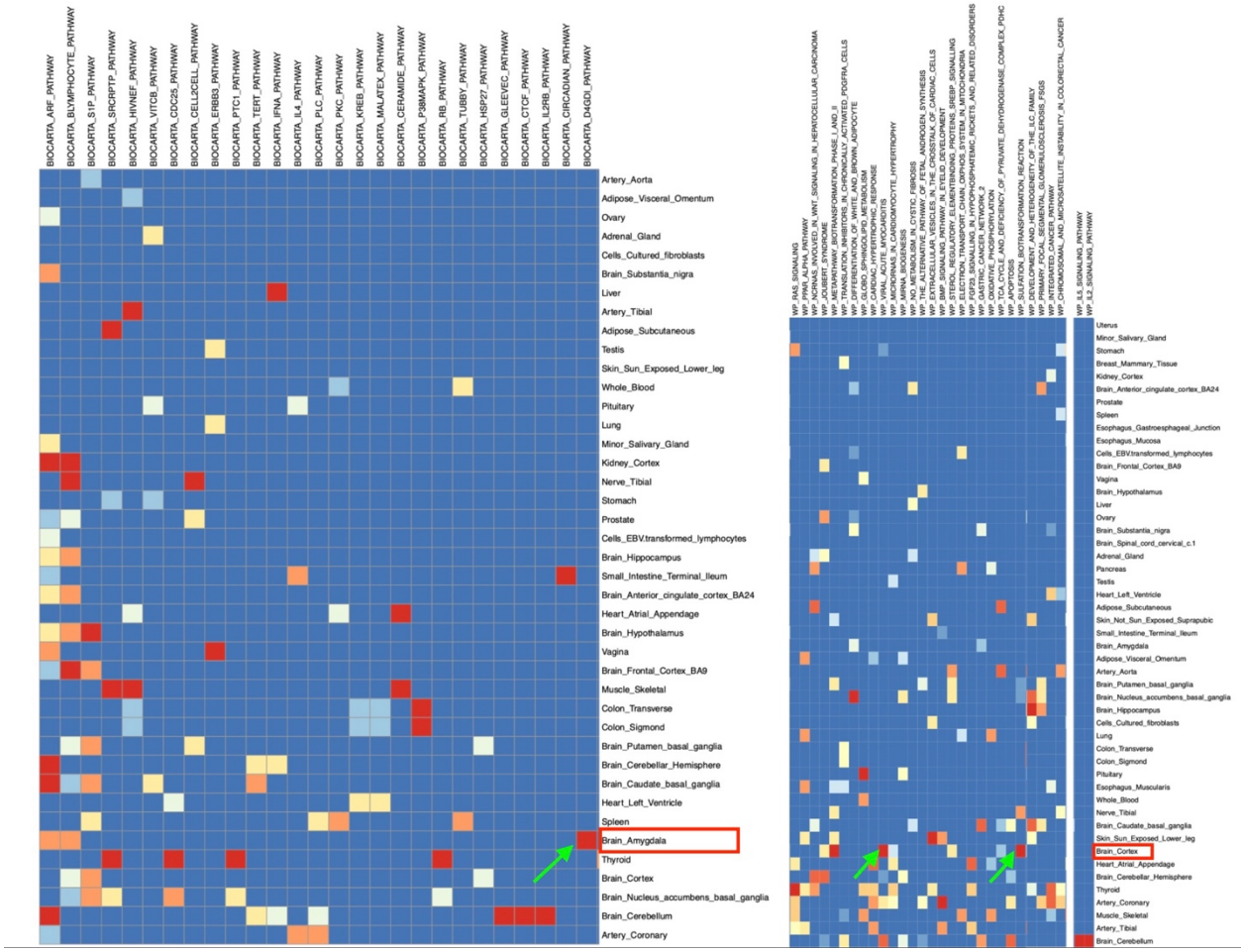


Figure 1. Heatmap of Alzheimer's disease's eQTLs enrichment results in (A) BioCarta and (B) WikiPathways pathway sets, respectively.

GDI, caspase-3 was found to be positively correlated with a mild cognitive deficiency in early AD pathology [15]. Clinical research suggested that Aβ could sequester caspase-3 via direct interaction and induce neuronal apoptosis via caspase-3 activation, thus strengthening AD development [16]. One possible hypothesis was that an increased level of caspase-3 in the amygdala leads to increased apoptosis and neuronal loss and thus contributes to the memory loss symptom of AD.

Similarly, Figure 1B showed significant enrichment of sulfation biotransformation reaction and viral acute myocarditis pathways in brain cortex, IL2 and IL5 signaling pathways in brain cerebellum, and development and heterogeneity of the innate lymphoid cell (ILC) pathway in brain hippocampus for the WikiPathways set (Figure 1B). The significant enrichment of viral acute

myocarditis pathway in the brain cortex suggested that the correlation observed between heart failure and AD was due to not only the majority of patients' age, but also genetic factors (Figure 1B) [17]. Such findings were consistent with a previous study where the viral myocarditis pathway from other pathway sets was identified to be significantly associated with AD [18]. One population study also found a higher than 80% risk of developing AD for patients with heart failures when major confounders like vascular comorbidities were controlled [19]. The significant enrichment in the sulfation biotransformation reaction pathway could also be explained by previous findings (Figure 1B). One research found an increased frequency of reduced metabolism and impaired sulfation of xenobiotics among AD patients [20]. A clinical study showed that sulfated curcumin can bind to copper and iron ions that are enriched in the brain cortex of AD patients and induce Aβ peptide formation, thus indicating that impaired sulfation ability would increase risks of AD [21]. One possible connection between acute viral myocarditis and AD is kynurenine 3-monooxygenase (KMO), which is a key regulatory enzyme in the kynurenine metabolism pathway that converts kynurenine to 3-hydroxykynurenine [22]. Studies have shown that the absence of KMO increased the production of kynurenine pathway metabolite, which lowered the synthesis of chemokine and thus resulted in the decrease of mortality of viral acute myocarditis by encephalomyocarditis virus in mice [22]. Interestingly, another study pointed out that JM6, a KMO inhibitor, was found to be able to prevent memory deficiency and synaptic loss in AD mouse models through the increase of the neuroprotective kynurenine metabolite kynurenic acid [23]. Such interaction may imply a hidden mechanism in AD's pathogenesis that increases KMO production and thus decreases levels of neuroprotective kynurenine metabolite and enhances AD symptoms, which explains AD's connection to acute viral myocarditis.

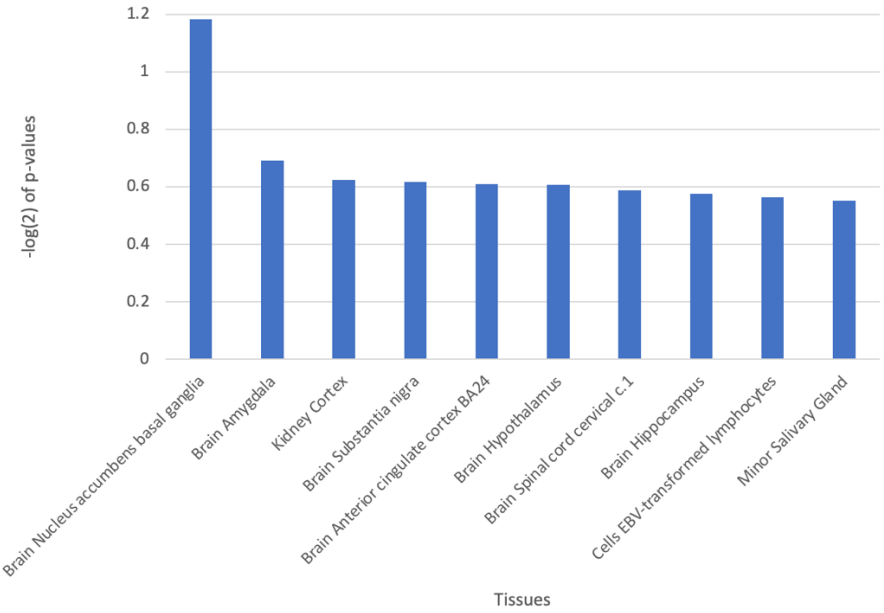


Figure 2. P-values of ten most significantly enriched tissues for Parkinson's disease.

## 2. Project 2: Loci2tissue

While eQTL's power in constructing and analyzing disease connections at the gene pathway level has been validated, it does not work well when dealing with the entire genome. To elaborate on individual genes' contributions to diseases, we developed loci2tissue, a bioinformatic tool that identifies associations between diseases and tissues using eQTL sets and normalized gene expression matrices from GTEx [24].

We have performed loci2tissue over thirteen diseases and analyzed their tissue enrichments. Figure 2 demonstrates tissue enrichments of Parkinson's disease-related genomic intervals. The enrichments of the nucleus accumbens basal ganglia and brain hippocampus are relatively high with significant p-values compared to the rest of the body tissues (Figure 2). Although Parkinson's disease (PD) patients mostly suffer from motor rigidity and bradykinesia, they experience multiple non-motor symptoms as well, including cognitive impairments [25]. The accumulation of alpha-synuclein in brain tissues and the subsequent formation of Lewy bodies and neurites are the central characteristics of PD [26]. At stage 4 of the Lewy pathology, alpha-synuclein has reached the

hippocampal formation through the performant path and axons from the tuberomammillary nucleus to the second sector of Ammon's horn, which initiates the cognitive impairment that completely develops in stages 5 and 6 [27]. Among all types of cognitive impairments exhibited by PD patients, episodic memory impairment is the most common, and subfields CA2 and CA3 of the hippocampus play a key role in episodic recollection [25]. The neurodegeneration of CA3 among PD patients has also been found to influence the pattern recollection and separation, leading to failure in process of recollection [26]. Such a comprehensive correlation between PD and the hippocampus provides sufficient support for its high enrichment index as shown in Figure 2.

The striatal dopamine (DA) depletion and the resulting cognitive dysfunctions in the early stage are considered hallmarks of PD [28]. Past studies have shown that the DA depletion of PD likely progresses from the dorsal to the ventral striatum [29-30]. In addition, Roshan and her colleagues have found that medication of L-DOPA, a precursor to dopamine, deteriorated the probabilistic reversal learning function among PD patients, and it affected reversal-related activity in the nucleus accumbens instead of the dorsal striatum and prefrontal cortex, indicating that nucleus accumbens plays an important role in dopaminergic modulation of reversal learning [31-32]. Nucleus accumbens is also associated with apathy, a frequently occurring symptom of PD. The severity of apathy in PD patients was positively correlated with atrophy in the left nucleus accumbens, and changes in the left nucleus accumbens could be used as a biomarker for dopamine-resistant apathy of PD [33]. Another study also demonstrated a significant decrease in grey matter volume at the left nucleus accumbens, further supporting the statement that the nucleus accumbens and the human reward circuit are closely involved in PD etiology [34].

As shown above, loci2tissue is capable of determining the affected tissues for certain diseases and ranking them based on levels of association, but it generates insignificant results. This

might be contributed by the fact that gene expression data are noisy with pseudogenes and novel transcripts, which increases the difficulty of integrating them into gene enrichment analysis [35]. Hence, gene expression data require an extensive level of filtering and preprocessing to be utilized in bioinformatic tools.

### 3. Project 3: Single-cell approaches

To overcome the challenge, we built a logistic regression-based method named LRDisTissue that performs tissue enrichment analysis using tissue-specific marker genes and disease-associated genes. This method calculates an enrichment score for each gene inside a tissue using the gene expression matrices and extracts the top 200 genes with the highest enrichment scores as the marker genes for this tissue. Compared to loci2tissue, LRDisTissue can more accurately determine the affected tissues with high significance and label enriched genes. However, simply locating the disease-influenced tissues is not specific enough for pathogenesis analysis, even with the help of enriched marker genes. It is necessary to get to the cell type level to develop a more comprehensive understanding of disease etiology.

Another approach to deepen our knowledge in pathogenesis is through the use of single nuclei RNA sequencing (snRNA-seq) datasets. snRNA-seq is a novel technology developed to determine gene expression levels in cells using isolated nuclei, which prevails over whole cells by their easy and rapid isolation process and high production rate [36]. The technology is not only cost-efficient while producing similar gene detection results compared to single-cell RNA sequencing, but also powerful in cases when unimpaired cells are hard to be obtained from the studied tissues [37]. Unlike the common noisy gene expression data, snRNA-seq's high resolution allows it to produce statistically significant results while keeping a relatively high accuracy. Eraslan and colleagues have used snRNA-seq data of skeletal muscles and found that the eQTLs

of type II diabetes are significantly enriched in skeletal muscle adipocytes and lymphatic endothelial cells, which is coherent with the observed tendency of type II diabetes patients in developing vascular diseases [38].

Both methods approach the problem from different angles and utilize different datasets, but they end up coming to the same conclusion that lowering gene enrichment analysis to the cell type and sub-cell type level would significantly increase the accuracy of statistical results and yield more insights about specific mechanisms in disease pathogenesis. The purpose of this research project is to design and develop R-based bioinformatic methods that take advantage of gene expression information at the sub-cell type level to perform enrichment analysis for a set of genomic variants. The program determines sub-cell type enrichment through both the cell type-specific marker genes and the snRNA-seq datasets. It utilizes the eQTLs catalogs of GTEx v8 public data release, disease-associated gene sets from DisGeNET, and snRNA-seq gene expression data from GTEx and Gene Expression Omnibus (GEO) from NCBI.

## 4. Methods & Materials

### 4.1.Overview

To explore the full extent of utilizing single-cell gene expression data, we developed two different methods that used different types of data and strategies to perform enrichment analysis. The first method extracted sub-cell type-specific marker genes and plotted them on disease-associated genes to build logistic regression models, while the second method compares proportions of snRNA-seq gene reads of genes targeted by eQTLs colocalization around genomic variants for diseases. The workflows of both methods are demonstrated in Figure 3.

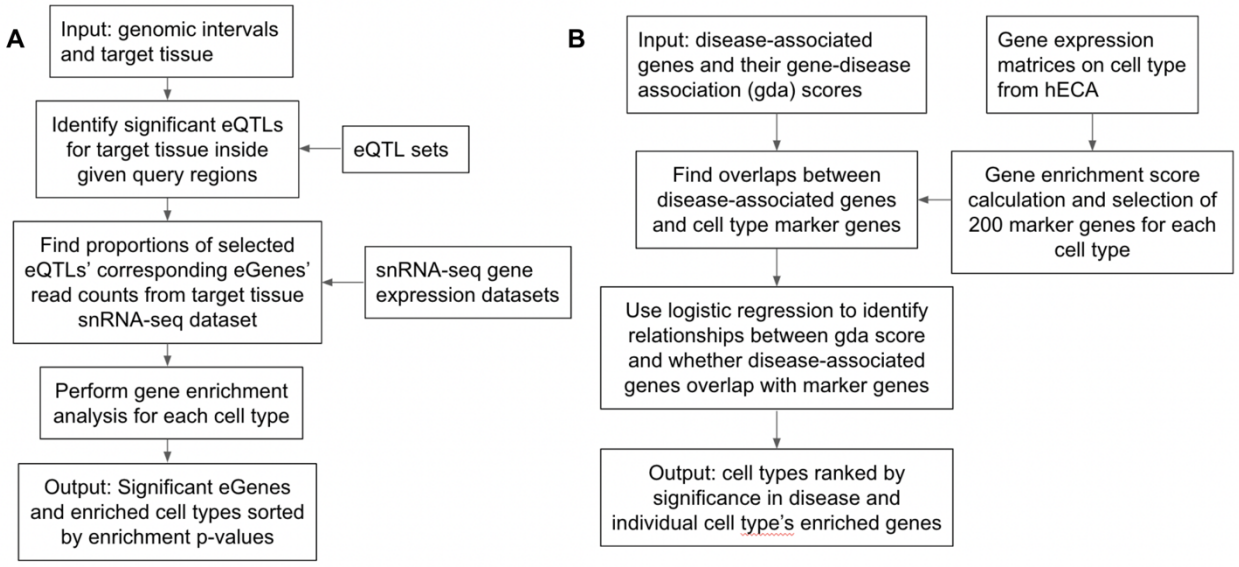


Figure 3. A diagram depicting our study's analysis pipeline, including input data, internal processes, and output results for (A) snRNA-seq approach and (B) cell type marker gene approach.

### 4.2. Cell type-specific marker genes

For the cell type-specific marker gene approach, we used the single-tissue cell-type gene expression matrices from human Ensemble Cell Atlas (hECA) and the disease-associated gene sets accessed via DisGeNET database [39-40]. The gene expression matrices of 105 sub-cell types from 10 various tissues will be treated to extract the top 200 most significant marker genes ranked by enrichment score for each sub-cell type. We will use the marker gene selection method provided in LRCell, in which the enrichment score is calculated by the multiplication of cell type-specific gene expression level and the fraction of samples within the cell type that express the gene [41].

Suppose there are a total of $n$ genes, $m$ cell types, and $s_1, s_2, \ldots s_m$ samples for each cell type that make up a total of $S$ samples. Here we use $x_{ijk}$ to represent the individual gene TPM of *i*th gene in the *k*th sample of the *j*th cell type. Hence, the average gene expression level of *i*th gene within *j*th cell type would be:

$$exp_{ij} = \frac{1}{s_j}\sum_{k=1}^{n} x_{ijk}$$

And the average gene expression level of *i*th gene in all samples would be:

$$exp_i = \frac{1}{S}\sum_{j=1}^{m}\sum_{k=1}^{s_j} x_{ijk}$$

Therefore, the cell type-specific gene expression level for *i*th gene in *j*th cell type is expressed as:

$$texp_{ij} = \frac{exp_{ij}}{exp_i}$$

After the cell type-specific gene expression, the cell type-sample fraction for *i*th gene in *j*th cell type is expressed as:

$$frac_{ij} = \frac{1}{s_j}\sum_{k=1}^{s_j} I_{|E_{ijk}>0|}$$

where *I* is the indicator function of whether *k*th sample in the *j*th cell type expresses the *i*th gene. The enrichment score for the *i*th gene in *j*th cell type is:

$$enrich_{ij} = texp_{ij} \times frac_{ij}$$

For each cell type, genes with the highest enrichment scores may be considered cell type-specific marker genes. For easy and fair comparisons, we use the same number of genes for all 105 cell types. In this study, we tested 200 marker genes for each cell type, which has shown to be able to build valuable models in past experience.

For this method's input, we used disease-associated genes along with their gene-disease association (gda) scores cataloged in the DisGeNET database. The gda scores are calculated based on the number and types of sources and the number of publications supporting the matching associations [40]. A total of 3,261,324 gene-disease associations were used from DisGeNET's SQLite file, which covers a total of 30,710 diseases and traits and 21,666 genes. Only 978 diseases

and traits with greater than or equal to 200 disease-associated genes were selected and used in this study to avoid the bias caused by small gene sizes.

Our logistic regression model proceeds as follows, the disease-associated gene's gda score being the explanatory variable and a binary indicator of whether the disease-associated gene overlaps with cell type-specific marker genes being the response variable. The number of genes used for cell type's logistic regression model and the gene's gda score remain the same for every sub-cell type in one disease or trait. Considering the possibility of cross-tissue progression in pathogenesis, sub-cell types from various tissues were combined and put in the results. The sub-cell types were ranked based on their model's p-values, and the overlapping genes for every sub-cell type were recorded and sorted by gda scores. We considered sub-cell types ranked on top to be more relevant to the disease and the genes sorted upfront to be more involved in the associations between the sub-cell type and the disease. Individual results of diseases and traits were plotted and analyzed to generate novel hypotheses. Additionally, significant tissue-trait pairs were extracted to find potential patterns using a significance threshold of $-ln(\frac{0.05}{105})$.

### 4.3. eQTLs and snRNA-seq

For the eQTLs-snRNA-seq approach, we used the 49 multi-tissue QTL data from GTEx v8 data release, in which each dataset was filtered with a p-value threshold of $10^{-4}$ to extract significant eQTLs. We used the snRNA-seq gene expression data of immune-related cells in the blood, kidney cells, and neurons in the prefrontal cortex from GTEx and GEO in NCBI. The genomic variants of specific diseases and traits were accessed from the Phenotype-Genotype Integrator (PheGenI) website.

The program takes in the genomic variants as input and creates a list of genomic intervals by flanking 50 thousand base pairs on each of the left and right sides of the variant's location,

which spans 100 thousand base pairs. For each tissue, eQTLs within the genomic intervals are pulled out, and the corresponding eGenes' snRNA-seq data are extracted for each sub-cell type. The proportion of reads of eGenes to reads of all genes within the sub-cell type is calculated for each sub-cell type. The overall proportion of eGene reads to all gene reads of all sub-cell types from the input tissue is calculated, and each sub-cell type's eGene read proportion is compared to the overall proportion using a normal distribution test, in which the p-values and negative natural log of them are recorded. We want to find the sub-cell types with the most significant p-values, which may indicate a potential association between the studied disease or trait and the cell cluster from this tissue. Such a relationship could be further analyzed in-depth, and we could deduce a novel hypothesis on the disease or trait's pathogenesis. The genes with the most read counts from the list were annotated for analytical purposes as well.

## 5. Results

### 5.1. Marker gene

We have generated valid results for 916 diseases and traits using the cell type-specific marker gene approach from 978 input diseases. 62 diseases and traits did not produce valid results because their disease-associated genes have unified gda scores, in which the logistic regression models were unable to be built. Using a significance threshold of $-ln(\frac{0.05}{105})$ for the negative-natural-log transformation of p-values, a total of 857 significant cell type-disease association pairs were extracted and recorded into Supplementary Table 1, in which Table 1 shows the top 20 most significant pairs. As demonstrated in Table 1, most of the diseases among the top association pairs are related to cancer.

Table 1. Top 20 Most Significant Disease-Cell Type Associations

| Diseases and traits | Cell types | P-values |
|---|---|---|
| Liver carcinoma | Kidney proliferating cell | 4.02E-23 |
| Neoplasms | Ileum macrophage | 4E-14 |
| Liver carcinoma | Kidney mesenchymal cell | 1.85E-13 |
| Neoplasm Metastasis | Ileum mast cell | 2.37E-13 |
| Neoplasms | Ileum goblet cell | 1.01E-12 |
| Rheumatoid Arthritis | Ileum macrophage | 1.06E-12 |
| Neoplasm Metastasis | Bladder stromal cell | 2.1E-12 |
| Primary malignant neoplasm | Prostate endothelial cell | 2.34E-12 |
| Tumor Cell Invasion | Ileum fibroblast | 6.89E-12 |
| Tumor Cell Invasion | Ileum goblet cell | 7.82E-12 |
| Tumor Cell Invasion | Bladder endothelial cell | 7.98E-12 |
| Primary malignant neoplasm | Adipose adipocyte | 8.03E-12 |
| Neoplasm Metastasis | Bladder mast cell | 9.74E-12 |
| Tumor Cell Invasion | Spinal cord vascular epithelial cell | 1.39E-11 |
| Primary malignant neoplasm | Kidney principle cell | 1.48E-11 |
| Primary malignant neoplasm | Adipose macrophage | 2.11E-11 |
| Neoplasm Metastasis | Ileum dendritic cell | 6.82E-11 |
| Neoplasms | Prostate macrophage | 7.44E-11 |
| Primary malignant neoplasm | Prostate macrophage | 9.15E-11 |
| Dermatologic disorders | Ileum dendritic cell | 1.27E-10 |

During the construction of the results, we combined logistic regression models of cell types from different tissues and ranked them instead of separating cell types by tissues. The built models measure the inclination of one specific type of cells' significant marker genes having associations with certain diseases. Since the gda scores are obtained from DisGeNET for each disease, cell types of different tissues could be directly compared with each other as we are only using marker genes' existence in disease-associated genes and list of gda scores to build the models. This provides a natural advantage of the cell type-specific marker gene approach compared to other methods out there that only test cell type enrichment within the same tissue or have to perform treatments on their data to apply on multiple tissues. The result of obesity was taken out and further analyzed.

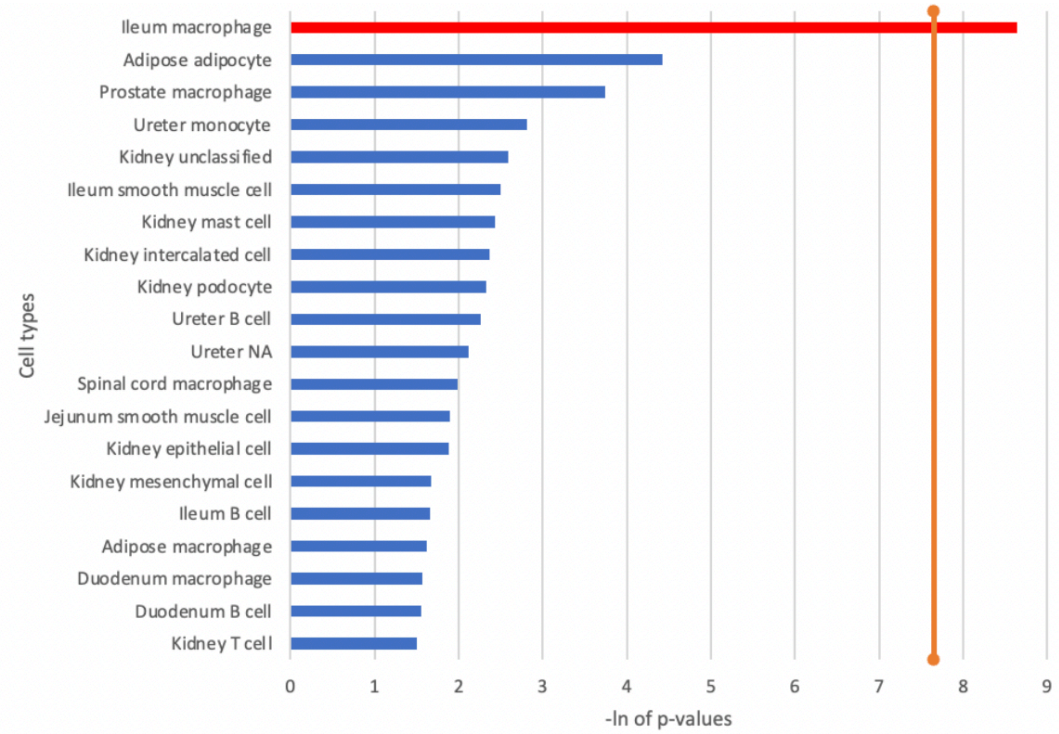


Figure 4. Negative natural log of p-values of twenty most significantly enriched cell types for obesity.

### 5.1.1. Obesity

Figure 4 demonstrated the enrichment of various cell types for obesity. While the adipocytes' relatively high enrichment for obesity is well expected considering that excessive accumulation of body fat is a common trait of obesity, the significant p-values of other lesser-known cell types like ileal and prostate macrophage definitely attract more attention (Figure 4).

Mice treated with a high-fat diet have shown increased expressions of ileal inflammation markers like MCP1 and TLR4 and a decreased expression of the anti-inflammatory cytokine IL-12B in the ileum due to macrophage infiltration, which was likely caused by the more abundant mucosal sulfidogenic bacteria that impairs epithelium integrity in intestines and colons [42]. Hence, the macrophage level in the ileum could serve as an indicator of ileal bacterial invasion due to high-fat diet. Coherently with this finding, another study showed that there was a decrease in anti-inflammatory cytokine IL-10 and an increase in pro-inflammatory cytokine 1L-1B and paracellular

permeability at ileum after one week of high-fat diet treatment to subject rats [43]. A decrease in expression level of ileal antimicrobial peptides like Mmp-7 and ang4 was observed among mice treated with high-fat diet for 30 days, which led to the development of microbiota in the intervillous zone of ileum that was supposed to be bacteria-free [44]. The mice treated with high-fat diet also experienced lower expression levels of cystic fibrosis transmembrane conductance regulator and Na-K-2Cl cotransporter 1 that decreased chloride secretion in ileum and disrupted mucus layer phenotype, which may induce metabolic disorders and hence further strengthen the phenotype of obesity [44-45]. Additionally, Breznik et al. observed a relatively lower expression level of tumor necrosis factor, which also affects intestinal epithelium integrity during homeostasis by ileal macrophage in high-fat diet mice, indicating that the change in ileal macrophage population originates from the same changes that disrupt intestinal homeostasis and eventually induce obesity [46]. Interestingly, most changes mentioned above were reversible in these studies by providing a standard diet regardless of whether the high-fat diet treatment was conducted for one or four weeks, suggesting that the microbiota formation in ileum and the following macrophage invasion could be treated in early stage of obesity by switching to a healthy diet. In general, such a strong association validates ileal macrophage's position as the most significantly enriched cell type in obesity. *IL-1B* was also listed as one of the overlapping genes between obesity-associated genes and ileal macrophage marker genes, which corroborates the effectiveness of our tool in establishing associations between tissue-specific cell types and complex traits (Supplementary Table 1).

The connections between prostate macrophage and obesity take from the aspect of prostate tumor, considering that the prostate gland is surrounded by fat tissues and obesity induces the accumulation of adipocytes. A higher expression level of CCL2, a major recruiter of macrophage

secreted by prostate cancer and stromal cells, was observed in mice under obese conditions [47-48]. Macrophages treated with media of obese sera-conditioned prostate tumor cells exhibited significantly higher expression of IL-10, TGF-β, and arginase-1, which are anti-inflammatory molecules serving as markers of tumor-associated macrophages [48]. This indicates that obese sera could induce normal macrophage's polarization into tumor-associated macrophage at prostate. COX-2 and PGE2 have also shown an increase in prostate cancer epithelial cells under obese conditions compared to those under regular conditions, in which PGE2 may contribute to macrophage polarization as well [48]. Prostatic M2 macrophage responsible for immunosuppression was found to have a significantly higher ratio in prostate cancer mice models treated with a high-fat diet and may participate in the following tumor growth by stimulating the secretion of IL-6 that phosphorylates STAT3 and proliferates myeloid-derived suppressor cells under pro-tumor microenvironment [49]. In one study, Huang et al. demonstrated an association between the high macrophage inhibitory cytokine-1 (MIC1) level in prostate cancer patients and patient's obesity, suggesting that high-fat diet containing palmitic acid could induce MIC1 in prostate cancer cells and hence disrupts metabolic homeostasis [50]. On the other hand, evidence also showed that MIC1 was downregulated among benign prostate hyperplasia patients with obesity as a result of gland destruction by inflammatory infiltrates [51]. Central obesity can lead to a decrease in adiponectin, which may lead to chronic inflammation in prostate since it originally inhibits phagocytic macrophage activity to reduce inflammation [51]. All combined evidence supports prostatic macrophage's third highest enrichment in obesity as indicated in our results.

### 5.2. eQTL-SnRNA-seq

For the eQTL-SnRNA-seq approach, we have performed cell type-enrichment analysis of fourteen diseases and traits using snRNA-seq data of three types of tissues (kidney, frontal cortex,

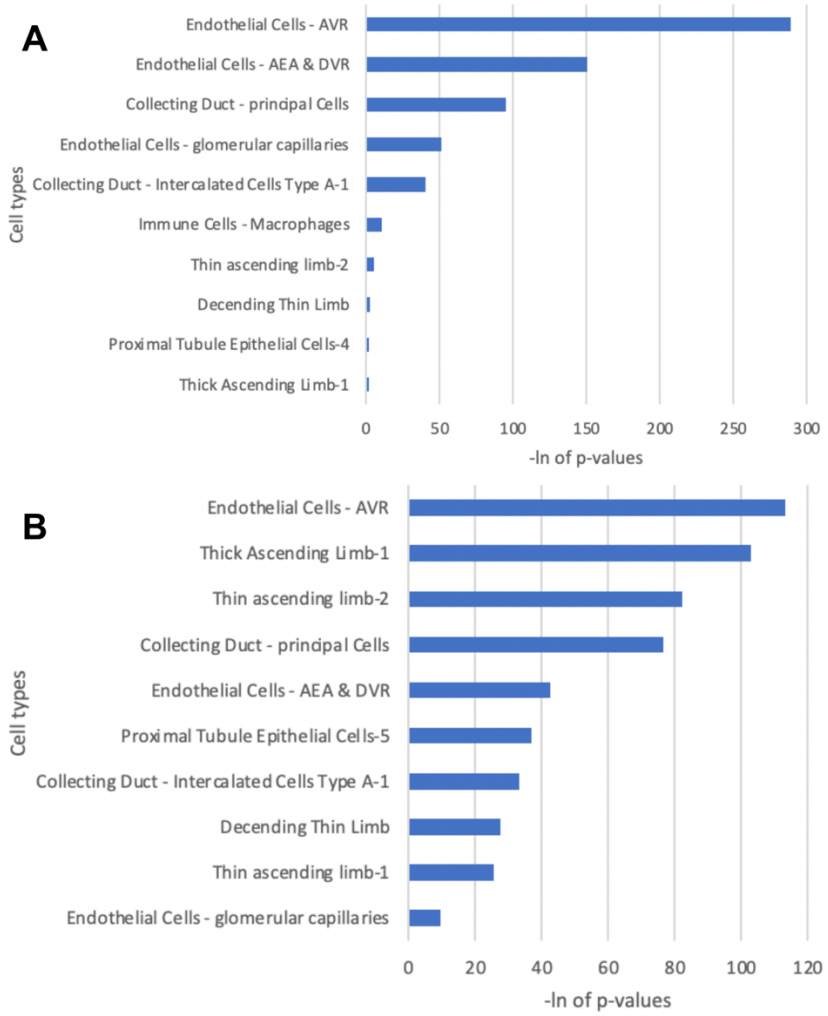


Figure 5. Negative natural log of p-values of twenty most significantly enriched cell types for (A) lupus nephritis and (B) diabetic nephropathy.

blood). Even though we used the same method for the three snRNA-seq tissue data inputs, p-values of cell types from different tissues could not be directly compared with each other since the snRNA-seq data were collected from different studies and their methods to calculate read counts were different. The results generated using snRNA-seq data of kidney tissues for lupus nephritis and diabetic nephropathy were further analyzed for their significance.

### 5.2.1. Lupus Nephritis

For lupus nephritis, the endothelial cells at ascending (AVR) and descending vasa recta (DVR) were shown to exhibit the top two most significant enrichments (Figure 5A). The relatively

higher expression levels of disease-associated genes like *NOTCH4* were observed in AVR and DVR of healthy samples compared to lupus nephritis patients [52]. *NOTCH4* is known to repress *TGF-β* gene expression by degrading phosphorylated SMAD3 protein, and the overexpression of *NOTCH4* could lead to the fibrogenesis in kidney [53-54]. Interestingly, *TGF-β* increases the mRNA expression level of endothelin-1 and endothelin receptors in kidney cells, and the latter two were shown to be upregulated among patients with lupus nephritis [55-56]. Endothelin-1 production in kidney has also been used to measure levels of renal inflammation in lupus nephritis, especially among patients with chronic kidney diseases [57]. Hence, we suspect that a loss-of-function mutation in the *NOTCH4* gene or epigenetic modifications that repress its expression levels in kidney vasa recta endothelial cells could cause an increase in *TGF-β* gene expression that stimulates endothelin-1 and endothelin receptors, which leads to additional signaling cascades and eventually the development of lupus nephritis. This not only supports AVR and DVR endothelial cells' high enrichments, but it also demonstrates our method's potential in coming up with novel hypotheses in complex disease pathogenesis.

The relatively high enrichments of principal cells and type A intercalated cells at the collecting duct reveal interesting biological insights into lupus nephritis. Studies have shown that the proinflammatory IL-18 secreted by intercalated cells of collecting duct is upregulated in lupus nephritis patients compared to healthy individuals, and this phenomenon was observed in animal models as well [58]. Faust et al. also suggested that the upregulation of IL-18 was correlated with the severity of lupus nephritis and was likely due to posttranscriptional processing [59]. The higher IL-18 level leads to more production of IFN-γ induced by IL-12, which causes cytokine imbalance and yields towards T helper 1 immune response [60]. This increase of IL-18 is accompanied by the preservation of vacuolar $H^+$-ATPase in collecting duct intercalated cells among patients with

lupus nephritis [58, 61]. Most $H^+$-ATPase within the collecting duct exists in the apical membrane of type A intercalated cells, in which they contribute to the acid-base homeostasis by transporting protons across apical membrane and thus facilitating movements of other ions like $Cl^-$ and bicarbonate [62]. Hence, the type A intercalated cells' enrichment in lupus nephritis is well supported by past literature findings.

### 5.2.2. Diabetic nephropathy

For diabetic nephropathy, the AVR endothelial cells again rank as the most enriched cell types within the disease followed by thick and thin ascending limbs (Figure 5B). Thick ascending limb reabsorbs 30% of filtered $Na^+$ and $Cl^-$, which in turn regulates urinary concentration, defends artery perfusion, and balances extracellular fluid volume [63]. It also participates in the transportation of ammonium and bicarbonate to maintain acid-base balance [63]. Human serum glucocorticoid-regulated kinase (hSGK) can stimulate the activity of epithelial $Na^+$ channels and $Na^+$, $K^+$, $2Cl^-$ cotransporter BSC-1, and a high transcription level of hSGK was observed in the thick ascending limb of kidney samples from diabetic nephropathy patients [64]. Such stimulations could enhance the $Na^+$ reabsorption at thick ascending limb, which decreases the NaCl delivery to macula densa to increase the glomerular filtration rate that leads to diabetic hyperfiltration, a key phenomenon in early diabetic nephropathy [64]. As a result, hSGK level in thick ascending limbs could conveniently function as a marker of diabetic nephropathy levels. TGF-β has been shown to exhibit regulation over hSGK in human intestinal tissues and suspected to have such function among other tissues, and a remarkedly higher TGF-β expression level was observed in kidney tissues of rats with type II diabetes mellitus, further validating its potential role in the development of diabetic nephropathy [65-66]. On the other hand, a single-cell study on diabetic nephropathy patients showed that there was a decrease in the $Na^+/K^+$-ATPase (NKA) subunits and WNK-1 and

STK39 that regulates the $Na^+$, $K^+$, $2Cl^-$ cotransporter NKCC2 in thick ascending limb cells [67]. Both NKA and NKCC2 perform $Na^+$ and $K^+$ transportation in thick ascending limbs, and their reduction should lower $Na^+$ and $K^+$ reabsorption, which contradicts with the findings of Lang et al. [67]. This inconsistency in evidence suggested a more complicated mechanism of $Na^+$ reabsorption and hyperfiltration within thick ascending limbs for diabetic nephropathy. Additionally, there are not many connections between the thin ascending limb and diabetic nephropathy.

## 6. Discussion

In this study, we have developed two novel bioinformatic methods to uncover interesting associations between complex diseases and specific sub-cell types using single-cell gene expression data. The results generated by our methods have revealed potential mechanisms of disease pathogenesis for obesity, lupus nephritis, and diabetic nephropathy. The proposed mechanism of how the decreased level of *NOTCH4* enhances *TGF-β* mRNA level and stimulates endothelin-1 and its receptor production in kidney vasa recta endothelial cells to promote lupus nephritis progression was well supported by multiple studies' findings. The high enrichments of ileal and prostate macrophages in obesity were consistent with past studies' findings as well.

Our results have generated significant cell type-disease associations that were coherent with past studies' findings in both approaches. The cell type-specific marker gene approach exhibits a key advantage over other existing methods since only the binary indicators of whether marker genes overlap with disease-associated genes change when different cell types were applied, which do not involve with the gene read counts and thus are free from data bias of different tissues. The selection of cell type-specific marker genes and the direct use of disease-associated genes from DisGeNET also avoid potential noises caused by uninvolved genes, which greatly improves the accuracy of this approach. However, the marker gene selection may overlook certain genes not

enriched within a cell type but play a key role in some disease etiology and pathogenesis. This is where our eQTLs-snRNA-seq approach comes in. Our second method takes all genes into consideration and uses eQTLs colocalization to identify target genes of the disease's genomic variants, which provides a broader range for query and hence produces more comprehensive results. Thus, the two approaches can make up for each other's shortcomings and work well when applied to the same diseases and traits together. Overall, both methods have generated novel biological insights into disease etiology by examining diseases' associations with various sub-cell types. Interpretation and analysis of these results could generate hypothesis on specific mechanisms of disease pathogeneses, which can guide genetic laboratories to develop a comprehensive understanding of the disease.

Our study still has room for improvement. While the sub-cell type data brings out more insights into our understanding of disease and trait pathogenesis at the cell level, the marker gene data provided by hECA were relatively limited in terms of tissue types. Only one tissue out of ten from hECA is related to the brain, and it is the spinal cord, which makes evaluating neurodegenerative diseases like bipolar disorder and schizophrenia using the cell type-specific marker gene approach relatively difficult. Such limitation also exists in the eQTLs-snRNA-seq approach since currently we have only been using snRNA-seq data from three tissue types. Additionally, some associations like thin ascending limb with diabetic nephropathy did not make much sense and have little literature covering it, indicating room for improvement in our statistical tests and algorithms. It is also noteworthy that the proposed mechanisms and hypotheses on pathogenesis from both approaches came from statistical tests of existing data and thus should be carefully examined and validated by molecular biology labs.

In future works of our study, we plan to incorporate more single-cell data from hECA and GEO, which would make our methods more accurate and comprehensive. Additional statistical tests and methods to connect eQTLs colocalization genes and snRNA-seq data will be explored to find the optimal approach. We could also add SNPs and RS ID as the input for the snRNA-seq approach to broaden the usage of our methods.

## 7. Supplementary Materials

Supplementary Table 1. Significant disease-cell type associations using cell type-specific marker genes.

## References

[1] J. Luo, M. Wu, D. Gopukumar, and Y. Zhao, "Big Data Application in Biomedical Research and Health Care: A Literature Review," *Biomed Inform Insights*, vol. 8, p. BII.S31559, Jan. 2016, doi: 10.4137/BII.S31559.

[2] GTEx consortium, "The GTEx Consortium atlas of genetic regulatory effects across human tissues," *Science*, vol. 369, no. 6509, pp. 1318–1330, Sep. 2020, doi: 10.1126/science.aaz1776.

[3] ENCODE Project Consortium, "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, Sep. 2012, doi: 10.1038/nature11247.

[4] Z. Qin *et al*., "Statistical challenges in analyzing methylation and long-range chromosomal interaction data," *Stat Biosci*, vol. 8, no. 2, pp. 284–309, Oct. 2016, doi: 10.1007/s12561-016-9145-0.

[5] T. Xu, P. Jin, and Z. S. Qin, "Regulatory annotation of genomic intervals based on tissue-specific expression QTLs," *Bioinformatics*, vol. 36, no. 3, pp. 690–697, Feb. 2020, doi: 10.1093/bioinformatics/btz669.

[6] C. E. Gillies *et al*., "An eQTL Landscape of Kidney Tissue in Human Nephrotic Syndrome," *The American Journal of Human Genetics*, vol. 103, no. 2, pp. 232–244, Aug. 2018, doi: 10.1016/j.ajhg.2018.07.004.

[7] F. Hormozdiari *et al*., "Colocalization of GWAS and eQTL Signals Detects Target Genes," *The American Journal of Human Genetics*, vol. 99, no. 6, pp. 1245–1260, Dec. 2016, doi: 10.1016/j.ajhg.2016.10.003.

[8] B. Wang, J. Yang, S. Qiu, Y. Bai, and Z. S. Qin, "Systematic Exploration in Tissue-Pathway Associations of Complex Traits Using Comprehensive eQTLs Catalog," *Frontiers in Big Data*, vol. 4, 2021, doi: 10.3389/fdata.2021.719737.

[9] Y. Benjamini and Y. Hochberg, "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 57, no. 1, pp. 289–300, 1995.

[10] C. Duyckaerts, B. Delatour, and M. C. Potier, "Classification and basic pathology of Alzheimer disease," *Acta Neuropathol*, vol. 118., no. 1, pp. 5–36, Jul. 2009, doi: 10.1007/s00401-009-0532-1

[11] C. Cheignon, M. Tomas, D. Bonnefont-Rousselot, P. Faller, C. Hureau, and F. Collin, "Oxidative stress and the amyloid beta peptide in Alzheimer's disease," *Redox Biology*, vol. 14, pp. 450–464, Apr. 2018, doi: 10.1016/j.redox.2017.10.014.

[12] S. Sadigh-Eteghad, B. Sabermarouf, A. Majdi, M. Talebi, M. Farhoudi, and J. Mahmoudi, "Amyloid-beta: a crucial factor in Alzheimer's disease," *Med Princ Pract*, vol. 24, no. 1, pp. 1–10. Jan. 2015, doi:10.1159/000369101.

[13] S. M. Butterfield and H. A. Lashuel, "Amyloidogenic Protein–Membrane Interactions: Mechanistic Insight from Model Systems," *Angewandte Chemie International Edition*, vol. 49, no. 33, pp. 5628–5654, Aug. 2010, doi: 10.1002/anie.200906670.

[14] M. L. Coleman and M. F. Olson, "Rho GTPase signalling pathways in the morphological changes associated with apoptosis," *Cell Death & Differentiation*, vol. 9, no. 5, pp. 493–504, May 2002, doi: 10.1038/sj.cdd.4400987.

[15] M. C. Gastard, J. C. Troncoso, and V. E. Koliatsos, “Caspase activation in the limbic cortex of subjects with early Alzheimer’s disease,” *Annals of Neurology*, vol. 54, no. 3, pp. 393–398, Sep. 2003, doi: 10.1002/ana.10680.

[16] Y.-J. Chang, N. H. Linh, Y. H. Shih, H.-M. Yu, M. S. Li, and Y.-R. Chen, “Alzheimer’s Amyloid-β Sequesters Caspase-3 in Vitro via Its C-Terminal Tail,” *ACS Chem. Neurosci.*, vol. 7, no. 8, pp. 1097–1106, Aug. 2016, doi: 10.1021/acschemneuro.6b00049.

[17] D. Li *et al.*, “Mutations of Presenilin Genes in Dilated Cardiomyopathy and Heart Failure,” *The American Journal of Human Genetics*, vol. 79, no. 6, pp. 1030–1039, Dec. 2006, doi: 10.1086/509900.

[18] G. Liu *et al.*, “Cardiovascular disease contributes to Alzheimer’s disease: evidence from large-scale genome-wide association studies,” *Neurobiology of Aging*, vol. 35, no. 4, pp. 786–792, Apr. 2014, doi: 10.1016/j.neurobiolaging.2013.10.084.

[19] C. Qiu, B. Winblad, A. Marengoni, I. Klarin, J. Fastbom, and L. Fratiglioni, “Heart Failure and Risk of Dementia and Alzheimer Disease: A Population-Based Cohort Study,” *Archives of Internal Medicine*, vol. 166, no. 9, pp. 1003–1008, May 2006, doi: 10.1001/archinte.166.9.1003.

[20] S. A. McFadden, “Phenotypic variation in xenobiotic metabolism and adverse environmental response: focus on sulfur-dependent detoxification pathways,” *Toxicology*, vol. 111, no. 1, pp. 43–65, Jul. 1996, doi: 10.1016/0300-483X(96)03392-6.

[21] L. Baum and A. Ng, “Curcumin interaction with copper and iron suggests one possible mechanism of action in Alzheimer’s disease animal models,” *Journal of Alzheimer’s Disease*, vol. 6, no. 4, pp. 367–377, 2004, doi: 10.3233/JAD-2004-6403.

[22] H. Kubo *et al.*, “Absence of kynurenine 3-monooxygenase reduces mortality of acute viral myocarditis in mice,” *Immunology Letters*, vol. 181, pp. 94–100, Jan. 2017, doi: 10.1016/j.imlet.2016.11.012.

[23] D. Zwilling *et al.*, “Kynurenine 3-Monooxygenase Inhibition in Blood Ameliorates Neurodegeneration,” *Cell*, vol. 145, no. 6, pp. 863–874, Jun. 2011, doi: 10.1016/j.cell.2011.05.020.

[24] Wang *et al.*, “Loci2Tissue: Ranking tissues by the e3xpression of disease-associated genes reveals insights of the underlying mechanisms of complex diseases and traits,” in *2022 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Dec. 2022, pp. 2879–2885. doi: 10.1109/BIBM55620.2022.9995232.

[25] T. Das, J. J. Hwang, and K. L. Poston, “Episodic recognition memory and the hippocampus in Parkinson’s disease: A review,” *Cortex*, vol. 113, pp. 191–209, Apr. 2019, doi: 10.1016/j.cortex.2018.11.021.

[26] S. Villar-Conde *et al*., “The Human Hippocampus in Parkinson's Disease: An Integrative Stereological and Proteomic Study,” *Journal of Parkinson's disease*, vol. 11, no. 3, pp. 1345–1365, 2021, doi: 10.3233/JPD-202465.

[27] K. Del Tredici and H. Braak, “Review: Sporadic Parkinson’s disease: development and distribution of α-synuclein pathology,” *Neuropathology and Applied Neurobiology*, vol. 42, no. 1, pp. 33–50, Feb. 2016, doi: 10.1111/nan.12298.

[28] G. U. Höglinger *et al.*, “Dopamine depletion impairs precursor cell proliferation in Parkinson disease,” *Nature Neuroscience*, vol. 7, no. 7, pp. 726–735, Jul. 2004, doi: 10.1038/nn1265.

[29] S. J. Kish, K. Shannak, O. Hornykiewicz, "Uneven Pattern of Dopamine Loss in the Striatum of Patients with Idiopathic Parkinson's Disease," The New England Journal of Medicine, vol. 318, pp. 876-880, 1988, doi: 10.1056/NEJM198804073181402.

[30] P. A. MacDonald and O. Monchi, "Differential Effects of Dopaminergic Therapies on Dorsal and Ventral Striatum in Parkinson's Disease: Implications for Cognitive Function," *Parkinson's Disease*, vol. 2011, p. 572743, Mar. 2011, doi: 10.4061/2011/572743.

[31] R. Cools, R. A. Barker, B. J. Sahakian, and T. W. Robbins, "Enhanced or Impaired Cognitive Function in Parkinson's Disease as a Function of Dopaminergic Medication and Task Demands," *Cerebral Cortex*, vol. 11, no. 12, pp. 1136–1143, Dec. 2001, doi: 10.1093/cercor/11.12.1136.

[32] R. Cools, S. J. G. Lewis, L. Clark, R. A. Barker, and T. W. Robbins, "L-DOPA Disrupts Activity in the Nucleus Accumbens during Reversal Learning in Parkinson's Disease," *Neuropsychopharmacology*, vol. 32, no. 1, pp. 180–189, Jan. 2007, doi: 10.1038/sj.npp.1301153.

[33] N. Carriere *et al.*, "Apathy in Parkinson's disease is associated with nucleus accumbens atrophy: A magnetic resonance imaging shape analysis," *Movement Disorders*, vol. 29, no. 7, pp. 897–903, Jun. 2014, doi: 10.1002/mds.25904.

[34] S. Martinez-Horta *et al.*, "Non-demented Parkinson's disease patients with apathy show decreased grey matter volume in key executive and reward-related nodes," *Brain Imaging and Behavior*, vol. 11, no. 5, pp. 1334–1342, Oct. 2017, doi: 10.1007/s11682-016-9607-5.

[35] J. Oyelade *et al*., "Clustering Algorithms: Their Application to Gene Expression Data," *Bioinform Biol Insights*, vol. 10, p. BBI.S38316, Jan. 2016, doi: 10.4137/BBI.S38316.

[36] R.V. Grindberg et al., “RNA-sequencing from single nuclei,” *Proceedings of the National Academy of Sciences*, vol. 110, no. 49, pp. 19802, Oct. 2013, doi: 10.1073/pnas.1319700110.

[37] H. Wu, Y. Kirita, E. L. Donnelly, and B. D. Humphreys, “Advantages of Single-Nucleus over Single-Cell RNA Sequencing of Adult Kidney: Rare Cell Types and Novel Cell States Revealed in Fibrosis,” *J. Am. Soc. Nephrol.*, vol. 30, no. 1, p. 23, Jan. 2019, doi: 10.1681/ASN.2018090912.

[38] G. Eraslan *et al.*, “Single-nucleus cross-tissue molecular reference maps to decipher disease gene function,” *bioRxiv*, p. 2021.07.19.452954, Jan. 2021, doi: 10.1101/2021.07.19.452954.

[39] S. Chen *et al.*, “hECA: The cell-centric assembly of a cell atlas,” *iScience*, vol. 25, no. 5, p. 104318, May 2022, doi: 10.1016/j.isci.2022.104318.

[40] J. Piñero *et al.*, “The DisGeNET knowledge platform for disease genomics: 2019 update,” *Nucleic Acids Research*, vol. 48, no. D1, pp. D845–D855, Jan. 2020, doi: 10.1093/nar/gkz1021.

[41] W. Ma, S. Sharma, P. Jin, S. L. Gourley, and Z. S. Qin, “LRcell: detecting the source of differential expression at the sub–cell-type level from bulk RNA-seq data,” *Briefings in Bioinformatics*, vol. 23, no. 3, p. bbac063, May 2022, doi: 10.1093/bib/bbac063.

[42] W. Shen *et al.*, “Intestinal and Systemic Inflammatory Responses Are Positively Associated with Sulfidogenic Bacteria Abundance in High-Fat–Fed Male C57BL/6J Mice,” *The Journal of Nutrition*, vol. 144, no. 8, pp. 1181–1187, Aug. 2014, doi: 10.3945/jn.114.194332.

[43] M. K. Hamilton, G. Boudry, D. G. Lemay, and H. E. Raybould, “Changes in intestinal barrier function and gut microbiota in high-fat diet-fed rats are dynamic and region

dependent," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 308, no. 10, pp. G840–G851, May 2015, doi: 10.1152/ajpgi.00029.2015.

[44] J. Tomas *et al.*, "High-fat diet modifies the PPAR-γ pathway leading to disruption of microbial and physiological ecosystem in murine small intestine," *Proceedings of the National Academy of Sciences*, vol. 113, no. 40, pp. E5934–E5943, Oct. 2016, doi: 10.1073/pnas.1612559113.

[45] B. O. Schroeder *et al.*, "Obesity-associated microbiota contributes to mucus layer defects in genetically obese mice," *Journal of Biological Chemistry*, vol. 295, no. 46, pp. 15712–15726, Nov. 2020, doi: 10.1074/jbc.RA120.015771.

[46] J. A. Breznik, J. Jury, E. F. Verdú, D. M. Sloboda, and D. M. E. Bowdish, "Diet-induced obesity alters intestinal monocyte-derived and tissue-resident macrophages and increases intestinal permeability in female mice independent of tumor necrosis factor," *American Journal of Physiology-Gastrointestinal and Liver Physiology*, vol. 324, no. 4, pp. G305–G321, Apr. 2023, doi: 10.1152/ajpgi.00231.2022.

[47] K. Fujita, T. Hayashi, M. Matsushita, M. Uemura, and N. Nonomura, "Obesity, Inflammation, and Prostate Cancer," *Journal of Clinical Medicine*, vol. 8, no. 2, 2019, doi: 10.3390/jcm8020201.

[48] G. C. Galván, C. B. Johnson, R. S. Price, M. A. Liss, C. A. Jolly, and L. A. deGraffenried, "Effects of Obesity on the Regulation of Macrophage Population in the Prostate Tumor Microenvironment," *Nutrition and Cancer*, vol. 69, no. 7, pp. 996–1002, Oct. 2017, doi: 10.1080/01635581.2017.1359320.

[49] T. Hayashi *et al.*, "High-Fat Diet-Induced Inflammation Accelerates Prostate Cancer Growth via IL6 Signaling," *Clinical Cancer Research*, vol. 24, no. 17, pp. 4309–4318, Sep. 2018, doi: 10.1158/1078-0432.CCR-18-0106.

[50] M. Huang *et al.*, "Diet-induced macrophage inhibitory cytokine 1 promotes prostate cancer progression," *Endocrine-Related Cancer*, vol. 21, no. 1, pp. 39–50, Feb. 2014, doi: 10.1530/ERC-13-0227.

[51] D. Parikesit, C. A. Mochtar, R. Umbas, and A. R. A. H. Hamid, "The impact of obesity towards prostate diseases," *Prostate International*, vol. 4, no. 1, pp. 1–6, Mar. 2016, doi: 10.1016/j.prnil.2015.08.001.

[52] E. Der, H. Suryawanshi, J. Buyon, T. Tuschl, and C. Putterman, "Single-cell RNA sequencing for the study of lupus nephritis," *Lupus Sci Med*, vol. 6, no. 1, p. e000329, Jun. 2019, doi: 10.1136/lupus-2019-000329.

[53] B. M. Grabias and K. Konstantopoulos, "Notch4-dependent antagonism of canonical TGF-β1 signaling defines unique temporal fluctuations of SMAD3 activity in sheared proximal tubular epithelial cells," *American Journal of Physiology-Renal Physiology*, vol. 305, no. 1, pp. F123–F133, Jul. 2013, doi: 10.1152/ajprenal.00594.2012.

[54] C. Yuan, L. Ni, C. Zhang, and X. Wu, "The Role of Notch3 Signaling in Kidney Disease," *Oxidative Medicine and Cellular Longevity*, vol. 2020, p. 1809408, Oct. 2020, doi: 10.1155/2020/1809408.

[55] A. Benigni, N. Perico, and G. Remuzzi, "Endothelin Antagonists and Renal Protection," *Journal of Cardiovascular Pharmacology*, vol. 35, 2000, [Online]. Available: https://journals.lww.com/cardiovascularpharm/Fulltext/2000/00002/Endothelin_Antagonists_and_Renal_Protection.17.aspx

[56] T. Nakamura *et al*., "Renal expression of mRNAs for endothelin-1, endothelin-3 and endothelin receptors in NZB/W F1 mice," *Renal physiology and biochemistry*, vol. 16, no. 5, pp. 233-243, Sep. 1993, doi: 10.1159/000173768.

[57] N. Dhaun *et al.*, "Urinary endothelin-1 in chronic kidney disease and as a marker of disease activity in lupus nephritis," *American Journal of Physiology-Renal Physiology*, vol. 296, no. 6, pp. F1477–F1483, Jun. 2009, doi: 10.1152/ajprenal.90713.2008.

[58] S. Gauer *et al.*, "IL-18 is expressed in the intercalated cell of human kidney," *Kidney International*, vol. 72, no. 9, pp. 1081–1087, Nov. 2007, doi: 10.1038/sj.ki.5002473.

[59] J. Faust *et al.*, "Correlation of renal tubular epithelial cell–derived interleukin-18 up-regulation with disease activity in MRL-Faslpr mice with autoimmune lupus nephritis," *Arthritis & Rheumatism*, vol. 46, no. 11, pp. 3083–3095, Nov. 2002, doi: 10.1002/art.10563.

[60] N. Calvani, M. Tucci, H. B. Richards, P. Tartaglia, and F. Silvestris, "Th1 cytokines in the pathogenesis of lupus nephritis: The role of IL-18," *Autoimmunity Reviews*, vol. 4, no. 8, pp. 542–548, Nov. 2005, doi: 10.1016/j.autrev.2005.04.009.

[61] B. Bastani, D. Underhill, N. Chu, R. D. Nelson, L. Haragsim, and S. Gluck, "Preservation of intercalated cell H(+)-ATPase in two patients with lupus nephritis and hyperkalemic distal renal tubular acidosis.," *Journal of the American Society of Nephrology*, vol. 8, no. 7, 1997, [Online]. Available: https://journals.lww.com/jasn/Fulltext/1997/07000/Preservation_of_intercalated_cell_H___ _ATPase_in.8.aspx

[62] A. Roy, M. M. Al-bataineh, and N. M. Pastor-Soler, "Collecting Duct Intercalated Cell Function and Regulation," *Clinical Journal of the American Society of Nephrology*, vol. 10, no. 2, 2015, [Online]. Available:

https://journals.lww.com/cjasn/Fulltext/2015/02000/Collecting_Duct_Intercalated_Cell_Function_and.21.aspx

[63] D. B. Mount, "Thick Ascending Limb of the Loop of Henle," *Clinical Journal of the American Society of Nephrology*, vol. 9, no. 11, 2014, [Online]. Available: https://journals.lww.com/cjasn/Fulltext/2014/11000/Thick_Ascending_Limb_of_the_Loop_of_Henle.22.aspx

[64] F. Lang *et al.*, "Deranged transcriptional regulation of cell-volume-sensitive kinase hSGK in diabetic nephropathy," *Proceedings of the National Academy of Sciences*, vol. 97, no. 14, pp. 8157–8162, Jul. 2000, doi: 10.1073/pnas.97.14.8157.

[65] X.-J. Miao, T.-T. Bi, J.-M. Tang, R. Lv, D.-K. Gui, and X.-F. Yang, "Regulatory mechanism of TGF-β1/SGK1 pathway in tubulointerstitial fibrosis of diabetic nephropathy," *European Review for Medical and Pharmacological Sciences*, vol. 23, no. 23, pp. 10482–10488, 2019, doi: 10.26355/eurrev_201912_19687.

[66] S. Waldegger *et al.*, "h-sgk serine-threonine protein kinase gene as transcriptional target of transforming growth factor β in human intestine," *Gastroenterology*, vol. 116, no. 5, pp. 1081–1088, May 1999, doi: 10.1016/S0016-5085(99)70011-9.

[67] P. C. Wilson *et al.*, "The single-cell transcriptomic landscape of early human diabetic nephropathy," *Proceedings of the National Academy of Sciences*, vol. 116, no. 39, pp. 19619–19625, Sep. 2019, doi: 10.1073/pnas.1908706116.