1

Text Mining Gene Selection to Understand Pathological Phenotype Using Biological Big Data

Christophe Desterke[1,2] • Hans Kristian Lorenzo[1,3,4] • Jean-Jacques Candelier[1,3]

[1]University Paris-Saclay, UFR Medicine, France; [2]INSERM unit UA9, Hospital P. Brousse, France; [3]INSERM unit 1197, Hospital P.Brousse, bâtiment Lavoisier, 14 avenue P.V.Couturier, 94800 Villejuif, France; [4]Bicêtre Hospital, AP-HP, Department of Nephrology, Le Kremlin-Bicêtre, France

Author for correspondence: Jean-Jacques Candelier, Hospital P.Brousse, bâtiment Lavoisier, 14 avenue P.V.Couturier, 94800 Villejuif, France. Email: jean-jacques.candelier@u-psud.fr

Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021.ch1

Abstract: Whole transcriptome omics experiments allow for the study of gene regulation at the cellular level. During analysis and interpretation of omics data, false discovery can occur. To minimize false discovery and identify true significant cases, multi-test correction has been introduced to bioinformatics algorithms. The scientific literature offers a huge collection of information that can be parsed using a web Application Programming Interface. Gene selection by text mining can rank information according to its importance while taking into account the most recent updates in scientific literature. The integration of text mining selection in biological big data, such as transcriptome experiments including single cell transcriptome, can achieve an important dimensional reduction of the data without any statistical hypothesis. This avoids false discoveries regarding the molecules of interest. Hydatidiform moles and focal segmental glomerulosclerosis (FSGS) nephropathy are the two examples presented in this chapter, which demonstrate the considerable value of these analytical methods to prove the concept. The best FSGS markers expressed can be displayed by building an interactive online web interface as a web resource based on the glomerular cell transcriptome.

In: *Bioinformatics*. Nakaya HI (Editor). Exon Publications, Brisbane, Australia.
ISBN: 978-0-6450017-1-6; Doi: https://doi.org/10.36255/exonpublications.bioinformatics.2021

This chapter shows the value of integrating text mining with omics data analysis to discover specific molecules and determine their locations and functions associated with complex diseases.

Keywords: focal segmental glomerulosclerosis; hydatidiform mole; text mining; transcriptome; web interface

INTRODUCTION

The use of automated literature search tools is often referred to as text or data mining. Text mining was initially defined by Martin Hearst in 1999 as "the discovery by computer of new previously unknown information, by automatically extracting and relating information from different written resources, to reveal otherwise 'hidden' meanings". The automated processing and analysis of text can help researchers evaluate findings in scientific literature. Text mining can be used to answer many research questions, ranging from the discovery of drug targets (1) and biomarkers (2) to drug repositioning (3). Text mining has evolved into a sophisticated and specialized field in the biomedical sciences, where text processing and machine learning techniques are combined with the mining of biological pathways and gene expression databases.

In general, text mining processes are comprised of several steps, such as information retrieval (usually performed by querying databases) followed by named entity recognition and finally information extraction. Text mining algorithms can operate using two distinct methods: co-occurrence-based methods or natural language processing. Co-occurrence based algorithms search for associations between information present in the text. Natural language processing algorithms take account of links between words in a text, for example, by finding a v-structure using the ABC principle. In other words, it can identify a relationship between A and C in the text which in turn will indirectly identify a potential link with B that is not explicitly mentioned in the text (4). The "pubmed.mineR" R library has been developed along these lines, as it analyzes linguistic structure at different levels: sentence tokenization and word tokenization (5).

In R language, some efforts have been made to process text mining transformation of the corpus into a matrix of word citations including tokenization, stemization and atomization processes of linguistic structure. This was previously applied to the "tm" R library, and more recently the "tidytext" R library was developed with text mining functions to enable the structuring of words found in the corpus (6). Focused more specifically on extracting textual information in the scientific literature using the PubMed resource, the "RISMed" R library was developed to extract all annotations of the abstracts loaded into the NIH biomedical database. The "RISMed" R library was recently used to analyze COVID-19-related data to develop a world collaboration map by means of the Biblioshiny application (7).

One important aspect of text mining is the visualization of results after the information has been processed. Two principal result representations have been developed: network visualization of the relationships identified and word cloud representation with the weighted size of words associated in the text

during analysis. Text mining in biomedical context could be integrated transversely with other quantitative methods in the recent field referred to as the "Science of Science". Its goal is to provide a deep, quantified understanding of the relational structure between scientists, institutions, and ideas, because it facilitates identification of the fundamental mechanisms responsible for scientific discovery. Scientific knowledge is made up of concepts and relationships embodied in research papers, books, patents, software, and other scholarly artifacts, organized into scientific disciplines and broader fields. The Science of Science utilizes all the multiple data sources available today such as PubMed, Google Scholar and the US Patent and Trademark Office, among others (8). Text mining processes are essential to enable access to these huge quantities of information located on the web.

BIOINFORMATICS TOOLS FOR THE TEXT MINING OF GENE RANKING

Many text mining applications have been built on the MEDLINE database because it is freely available, features a rich applied programming interface and supplies annotated abstracts containing Medical Subject Heading (MeSH) Terms (9). To improve the efficiency of querying, similar keywords such as synonyms could be used to define concepts with different reformulations such as in the case of the ConQuR-bio algorithm (10).

In the area of biomedical texts, mining is widely used in the context of gene expression annotation, such as understanding large lists of regulated genes during transcriptome experiments. Other biomedical applications for text mining could also be investigated, such as drug-target discovery, which enables the search for new drug targets or candidates; it would permit detailed and automated analysis of scientific literature to discover how genes are related to particular diseases and how they are involved in the effects of medicinal products. For example, text mining analyses could be used to link genes to pathways involved in metabolic adverse events in the transcriptome of immune cells treated with an anti-inflammatory drug (11). Drug repositioning could be also investigated by applying text mining algorithms to discover the identification of hidden connections between drugs, genes, and diseases, such as determining the links between a drug and cell proliferation using gene identifiers as the intermediary support for natural language processing (12).

Numerous methods to prioritize genes are based on the co-occurrence analysis of given keywords and gene names extracted automatically from scientific abstracts. The principal hypothesis underlying this type of analysis is that the more frequently two words co-occur in abstracts, the more likely they are to be functionally linked. However, automatic gene name extraction and normalization methods may wrongly identify a significant proportion of gene mentions in a text, therefore contributing noise and ambiguity to the text mining results (13,14).

In the context of transcriptome analysis, lists of regulated genes are sometimes composed of large numbers of molecules that are difficult to understand during functional annotation with respect to classic biological functions or cellular pathway databases. The co-occurrence of keywords and gene names when searching

in biomedical abstracts could aid in the understanding and discovery of principal relationships between regulated molecules.

For transcriptome analysis, the concept of the "next generation" text mining has emerged by combining gene lists that result from text mining with gene set enrichment analysis (GSEA) (15). This has been applied to enable understanding of the interactions between genes and chemical components. Using text mining, lists of genes were built for each chemical component, and using this custom database the GSEA method was able to retrieve which molecular component was used during experimental stimulation of the transcriptome under investigation (16).

Conventional text mining approaches tend to process PubMed abstracts rather than full text articles and fail to mine information not present in abstracts, but text mining of full text articles has recently gained interest (17,18). Thus, the PubTator central algorithm now enables exploration of the relationships between genes, diseases, chemical components, mutations, cell lines and species in more than six million full text articles uploaded on the PubMed Central website (19). Some website resources have also been developed to extract gene-gene co-occurrence citations based on detecting GeneRIF or AutoRIF references characterizing molecular identities in the corpus; this is the case of the Geneshot application which enables the characterization of gene network relationships in text (20).

To improve detection of biomarkers in pathologies like cancer, and to produce an adapted precision medicine therapy, applications based on deep learning have been developed, such as in the case of the Biomedical Entity Search Tool (BEST) web application (21), which is based on detection of mutation-gene-drug relations in PubMed biomedical corpus (22).

The Genie algorithm originally ranked complete sets of genes in any given organism according to a particular gene function or took advantage of all available orthologous information to expand MEDLINE literature. A biological topic is taken as the input parameter to review the entire MEDLINE database for relevance to that subject, and then evaluated for all genes included in the user's requested organism according to the relevance of their associated MEDLINE records. Genie associates machine learning and text mining processes by creating a train corpus based on 1,000 abstracts to build a naïve linear Bayesian classifier model. Secondly, using text mining, abstracts in which genes occur are compared to the train set by the machine learning model (23).

TEXT MINING GENE SELECTION FOR HYDATIDIFORM MOLES: A CASE STUDY

In some cases, transcriptome experiments corresponding to the biological hypothesis do not exist. A good example of this is the pathophysiology of hydatidiform moles. Hydatidiform moles are a rare complication of pregnancy and are the most common gestational trophoblastic disease. It affects the two layers that make up the placental villi: the trophoblast layer called the cytotrophoblast, and the expanding peripheral syncytial layer, the syncytiotrophoblast, which invades the endometrium and uterine arteries. Hydatidiform moles are characterized by

abnormal trophoblast proliferation giving rise to hydropic villi. It can be either complete or partial. Complete hydatidiform moles are the result of excessive trophoblast proliferation; there is no fetal circulation, and the embryo fails to develop. Partial hydatidiform moles are the result of mild trophoblast proliferation; there is fetal circulation and development of the embryo, but the fetus is abnormal and cannot survive. Hydatidiform moles are subject to severe hypoxia, and the persistent vascular immaturity of the villous stroma can lead to hydropic villi, particularly in the case of complete hydatidiform mole. In about 7-17% of cases, trophoblastic hyperplasia extends to and exceeds the uterine cavity, and is referred to as an invasive mole. An invasive mole can either be premalignant or malignant; malignant mole can transform into a highly aggressive tumor called choriocarcinoma (24).

Because no transcriptome data are publicly available on hydatidiform moles, we performed text mining to search for all genes studied in this context and then looked at their expression in three different transcriptome datasets on normal human placenta, focused on the three biological mechanisms associated with the disease: trophoblast differentiation, trophoblast invasion and hypoxic environment.

The “tm” and “RISMed” R libraries that connect to PubMed were used to perform text mining. “RISMed” library was also used to carry out a meta-analysis on the words used in the scientific literature. Natural Language Processing (NLP) allowed us to communicate information from a variety of biological resources (such as Gene Ontology Terms). We were able to discover semantic relationships with the scientific literature and link them to biological databases. In each case, gene expression matrices were used independently and normalized using our previously developed method (25). Mathematical matrix dimensional reduction was applied to a normalized transcriptome dataset by merging the genes obtained from text mining with identifiers present in gene expression datasets. These transcriptome data were then re-analyzed with the genes retrieved from text mining, using terms such as HM-linked genes studied in cytotrophoblast differentiation, extravillous trophoblast invasion and hypoxia. The expression of genes during mildly or severely invasive trophoblast proliferation (respectively, cyto- and extravillous trophoblasts) was determined using the Significance Analysis for Microarray algorithm with a threshold of false discovery rate fixed under five percent of error. For different oxygen concentrations, a supervised analysis of variance with two factors (culture conditions and oxygen concentration) was performed using Fisher’s test (500 permutations) on hydatidiform mole-linked genes. To validate the text mining approach, we performed a manual search in the PubMed database for the genes identified by text mining. This analysis of the literature enabled validation of the relationship between these genes and hydatidiform mole (26).

In conclusion, by using text mining and associated bioinformatics and mathematics methods we were able to identify 72 unique genes linked to hydatidiform mole (Figure 1). Moreover, our analysis integrated the different aspects of hydatidiform mole pathophysiology and highlighted the importance of trophoblast differentiation in this pathology. We were thus able to demonstrate the importance of some of these genes in chorionic villous invasion and regulation of their expression by oxygen concentrations. Based on this work we were able to build a network of the different relationships between the placenta, placental molecules,

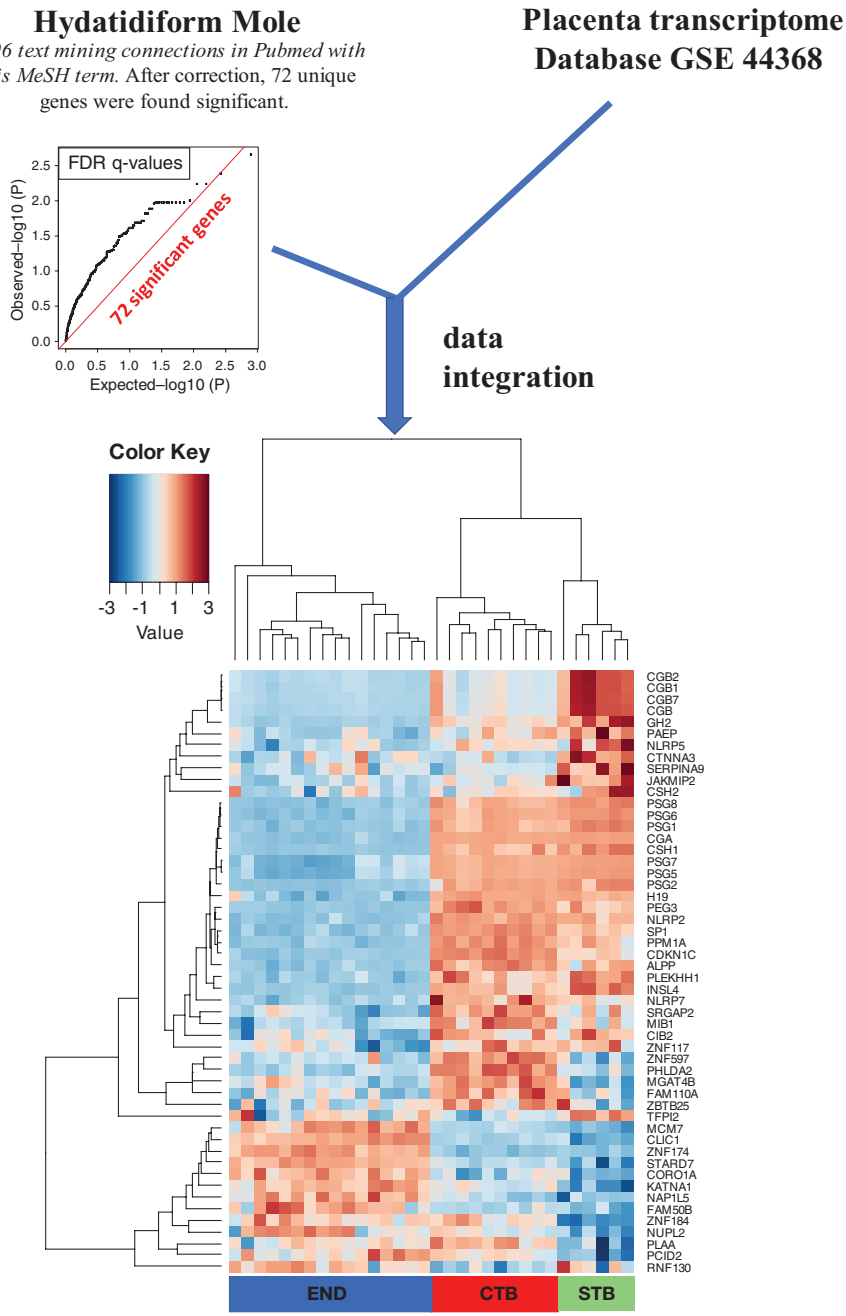


Figure 1. Integration of genes. Obtained from text mining research by MESH term "Hydatidiform Mole" with single cell transcriptome data of Human normal placenta from data set GSE 44368. A Heatmap was performed on the expression of hydatidiform mole related genes which discriminated placenta tissues (unsupervised classification was performed with Pearson correlation distances). End, endothelium; CTB, cytotrophoblast; STB, syncytiotrophoblast.

and their functions. Thus, even without the transcriptome data, it is possible to obtain relevant knowledge that can then be studied by means of appropriate manipulations.

UNDERSTANDING FOCAL SEGMENTAL GLOMERULOSCLEROSIS USING SINGLE CELL TRANSCRIPTOME OF A HEALTHY ADULT DONOR KIDNEY: PROOF OF CONCEPT

Chronic kidney disease (CKD) is a major global public health problem because of its growing epidemic status and the devastating complications it causes at present, such as cardiovascular problems or end-stage renal failure (27). Focal Segmental Glomerulosclerosis (FSGS) is a leading cause of CKD. FSGS describes a histological pattern of kidney damage common in many nephropathies (28). It is characterized by the presence of segmental sclerotic lesions of the glomerulus and an accumulation of focal hyaline deposits. The glomerular filtration structure collapses. As a result, serum proteins are not retained in the blood but lost in the urine (proteinuria). Its progressive nature culminates in end-stage renal disease and loss of renal function. The causes of FSGS are multiple. FSGS may be found in patients with hypertension, diabetic glomerulopathy, reflux nephropathy, drug addiction or HIV infection, as well as in various glomerular protein mutations. There is also an idiopathic form of FSGS, the pathogenesis of which is poorly understood. FSGS is particularly harmful because of its poor response to treatment. Resistance to corticosteroids is common and a relapse of disease after kidney transplantation is frequently observed and leads to graft loss (29). The search for new therapeutic targets has therefore become a major focus for nephrologists for more than 50 years.

The functional links that exist between risk factors (genetic or otherwise) and the phenotype of this disease are not clearly understood. Various strategies have been used to address this challenge, such as study of the genome and the expression profile of microarrays. More recently, single-cell sequencing made it possible to study the expression of the genetic content of an individual cell without the need for prior cell culture. Thanks to this innovative technology, gene expression in thousands of individual cells can be determined in a single experiment (30). These technologies have generated large quantities of bioinformatics data that are often stored non-systematically, thus hindering most researchers without extensive expertise in advanced computing. In the field of nephrology, there are some ongoing initiatives such as Nephroseq (http: //www.nephroseq.org [accessed on 14 January 2021]), the Renal Gene Expression Database (RGED; http://rged.walleva.net/ [accessed on 14 January 2021]), KUPKB: from the Kidney and Urinary Pathways Knowledge Base (31), CKDdb (Chronic Kidney Disease database; www.padb.org/ckdbd/index.html [accessed on 14 January 2021]) as well as others. These “omics” platforms enable the exploration of gene expression associated with kidney diseases. However, they lack multi-omic databases that could unify dispersed “omics” repositories. We performed an FSGS text mining study and integrated it into a gene expression database obtained by single cell RNA-sequencing in glomerular cells from healthy donor kidneys. Based on the result,

we propose a bioinformatics tool that helps visualize the different types of cells associated with a specific gene expression linked to FSGS. Our principal goal was to simply determine which particular renal cells (in glomeruli: podocytes, endothelial and mesangial cells) might be involved in a particular gene expression (and/or function) associated with the pathogenesis of FSGS.

FSGS was investigated by means of a PubMed query with three different gene-disease text mining association algorithms, still operational as of 14 January 2021: ConQuR-bio; http://conqur-bio.lri.fr/ [accessed on 14 January 2021] (10), polysearch2 (32), and Genie; http://cbdm-01.zdv.uni-mainz.de/~jfontain/cms/ [accessed on 14 January 2021] (23). Two of these three algorithms (ConQuR-bio and Polysearch2) enabled human gene ranking by text mining for co-occurrences with the MesH term introduced during the PubMed query, and the last one (Genie) employed a mixed method combining machine learning steps with text mining process (Figure 2).

Preparations for the single cell analysis were made in the Ubuntu 18.04 LTS operating system with R environment version 3.53 and Seurat R-package version 3.1.3 (33). The "CreateSeuratObject" Seurat function was implemented on the GSE140989 dataset (34) after fixing a threshold of 3 for the minimal number of cells and of 200 for the minimal number of features by cell. The resulting Seurat object with metadata on 24 different kidney samples included a total of 19,622 features across 22,268 cells (Figure 3A). The "NormalizeData" Seurat function allowed us to log normalized input sequencing data, and 2,000 variable features were identified before data scaling and dimensional reduction by principal component analysis was performed. The variance of principal component analysis was estimated by performing an Elbow plot on thirty principal components (Figure 3B). Twenty components were found to be important to explaining the heterogeneity of the cells in this dataset. Subsequently, UMAP dimensional reduction was performed with twenty dimensions within the analyses. The sample distribution in the first UMAP analysis revealed a satisfactory distribution between subjects on the totality of cell distribution (Figure 3C), suggesting good integration of the data in this multi-experiment analysis. Downstream clustering analysis with the construction of a KNN graph, based on the Euclidean distance in PCA space, refined the edge weights between any two cells based on the shared overlap in their neighborhoods (Jaccard similarity) in the same way as scRNA-seq data (35) and CyTOF data with the Phenograph algorithm (36). We then applied modularity optimization techniques such as the Louvain algorithm to cluster the cells. Twenty-one cell clusters were found to be representative of this Seurat object (Figure 3D).

ONLINE RESULT VISUALIZATION AND THE DEVELOPMENT OF AN INTERACTIVE WEB INTERFACE

Omics analyses are united by their common feature of describing large numbers of biomolecules relevant to the experimental system under investigation. Various frameworks have been developed to highlight biomolecules that are expected to be the most appropriate for more intensive follow-up study. Transcript

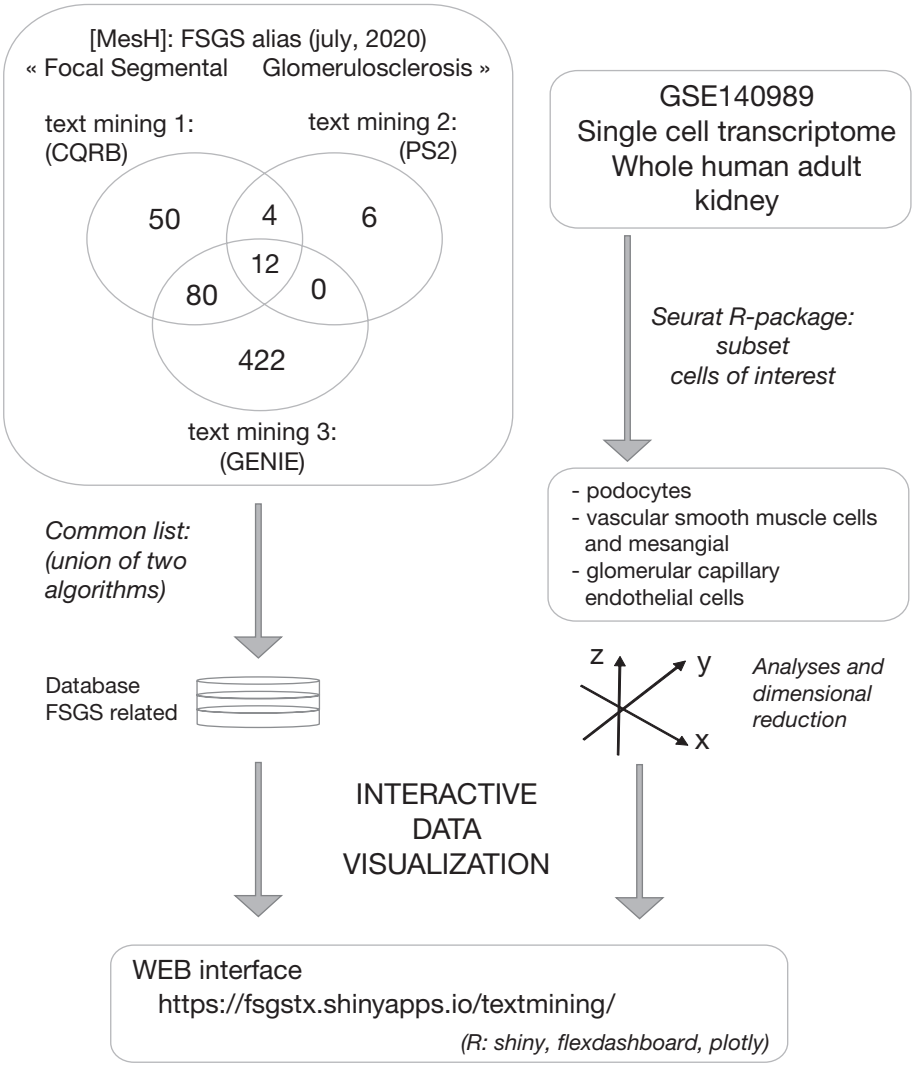


Figure 2. Diagram of the bioinformatics process. This diagram explains the processes that we implemented during this integration of text mining tools to interpret single cell transcriptome data. On the left side of the diagram, we can see the text mining process used to find the genes associated with the term: Focal Segmental Glomerulosclerosis. To make this query, three different text mining algorithms were used: ConQuR-bio (CQRB), polysearch2 (PS2) and GENIE. On the right side of the diagram, we can see Seurat single cell transcriptome process employed to interpret human normal glomerular kidney cells from dataset GSE140989. Finally, at the bottom of the diagram, the intersection of these two analysis processes made it possible to develop a graphical interface with the internet address indicated.

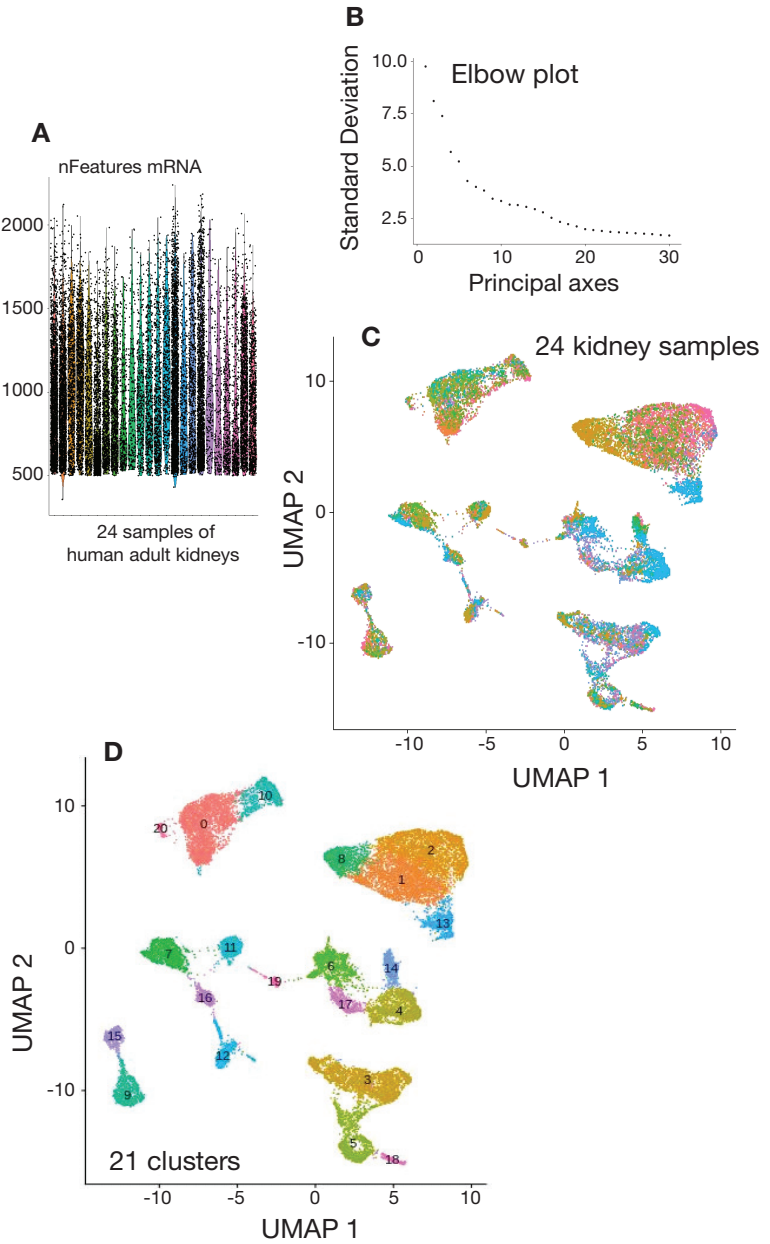


Figure 3. Single cell transcriptome analysis of human adult kidney. A. Violinplot of the numbers of features present in each cell and separated between the 24 merged kidney samples. **B.** Elbow plot presenting the standard deviation of the first 30 principal axes during dimensional reduction by principal component analysis performed on single cell transcriptome of human adult kidney. **C.** UMAP dimensional reduction performed on the single cell transcriptome of the human adult kidney: colors are attributed to the origin of the samples. **D.** UMAP dimensional reduction performed on the single cell transcriptome of the human adult kidney: colors are attributed to the 21 clusters identified during the analyses.

quantifications with omics experiments have been revolutionized during the past 15 years by different technological approaches, such as microarrays, followed by RNA-sequencing and more recently, single cell RNA-sequencing (37). The single cell approach enables us to understand the heterogeneity of cellular expression at the whole transcriptome level thus allowing detailed characterization of the cell subtypes that make up a tissue. Several studies have tried to gain a clearer understanding of tissue cell composition to develop different methods for data visualization. For example, using python language, the Scanpy library (38) was developed to merge processes for single cell visualization already known in different R packages, such as Seurat for clustering (33), monocle for cell trajectories (39) and pagoda for splicing (40).

To facilitate the exploration of FSGS-related biomarkers found by text mining, an interactive web interface was developed and uploaded at the following address: https://fsgstx.shinyapps.io/textmining/ (last accessed on 14 January 2021). Prior to unsupervised analysis, 1,468 glomerular cells were isolated from the GSE140989 normal human kidney dataset (Figure 4). The pre-processed and scaled digital matrix of single cell analysis was restricted to the set of genes identified as being related to FSGS (Table 1). Unsupervised analysis using t-SNE (t-distributed stochastic neighbor embedding) was processed to be displayed in the interface. The website was built with a flexdashboard and shiny application inclusion and with graphical interactivity displayed by R plotly. This data dashboard enabled exploration of the expression of FSGS biomarkers in the three glomerular cell subpopulations: podocytes (POD, n=182), vascular smooth muscle cells and mesangial cells (SMCMG, n=713), and glomerular capillary endothelial cells (GCEC, n=753) (Figure 5). This interactive application enables users to understand the cellular origin of expression for FSGS biomarkers characterized by text mining. Users will need to select a gene ID on the left sidebar and the application will display the expression of this selected marker with interactivity on the t-SNE graph. The number of positive cells for this marker will be displayed in the value box at top right-hand side of the dashboard, and expression by group will be displayed on a violinplot; finally, a statistical summary (mean and standard deviation) will be displayed by group of samples (Figure 5).

Using the *NPHS1* and *NPHS2* genes (congenital nephrotic syndrome of the Finnish types 1 and 2) respectively, nephrin and podocin were confirmed as FSGS-related markers expressed strictly by podocytes associated with glomerular cells. The *MAFB* transcription factor was confirmed as being expressed in podocytes and it was shown that this podocyte transcription factor protected the kidney from developing FSGS (41). The application also confirmed that Wilms' tumor protein (WT1) is strictly expressed by podocytes. *WT1* is known to be a transcription factor and master regulator of podocyte differentiation and homeostasis. It may also be a target repressed by microRNA-193a to induce focal segmental glomerulosclerosis (42). The application further showed that the angiogenic factor *VEGFA* was strictly expressed in podocytes, and indeed this factor is known as a marker of glomerular endothelial cell injury and FSGS lesions in the context of idiopathic membranous nephropathy (43). The application also confirmed a high level of expression of *PLCE1* Phospholipase C Epsilon 1 in podocytes; this gene is known to be mutated and to affect podocytes in familial and genetic forms of FSGS (44). Using the application, *COL4A3* and *COL4A4* were found to be strictly expressed in podocytes at the glomerular level. Mutations in *COL4A3* and

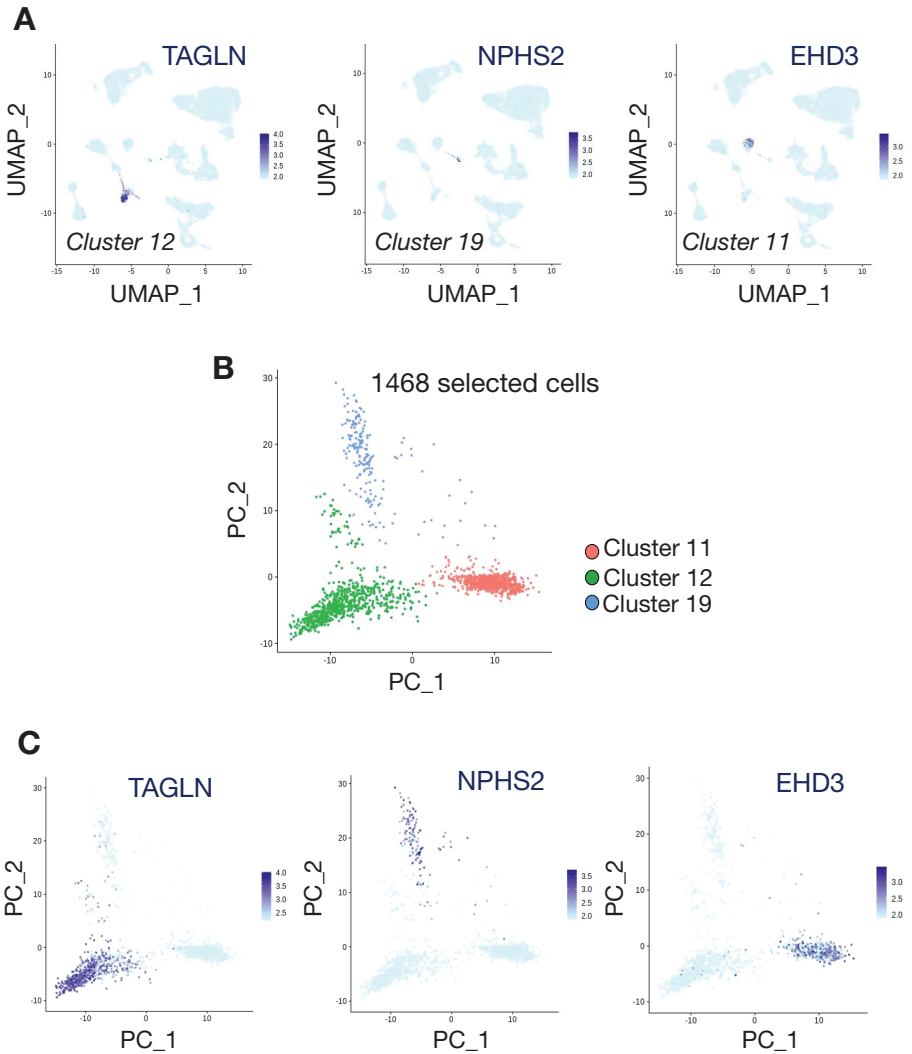


Figure 4. Kidney single cell transcriptome subsetting to select glomerular cells of interest. **A.** UMAP dimensional reduction showing the respective expression of *TAGLN* in cluster 12, *NPHS2* in cluster 19 and *EHD3* in cluster 11. **B.** Principal component analysis after subsetting clusters 11, 12 and 19. **C.** Respective levels of expression of *TAGLN*, *NPHS2* and *EHD3* in principal component analysis after subsetting on clusters 11, 12 and 19.

COL4A4 are known to cause Alport syndrome (AS), a thin basement membrane nephropathy resulting in pathognomonic glomerular basement membrane, and secondary FSGS is known to develop in classic AS at later stages of the disease (45). For *COL4A5*, which may be affected by mutations causing AS with FSGS lesions (46), its expression was found to be shared between different cell subpopulations: podocytes, vascular smooth muscle cells, and mesangial cells. At the

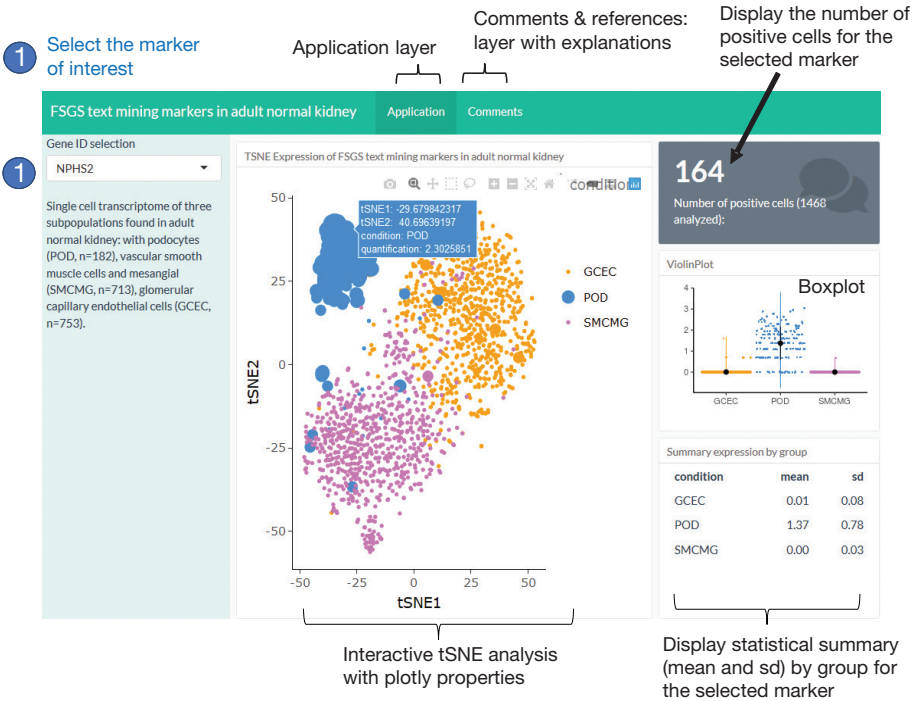
TABLE 1 **List of 96 genes selected using three text-mining algorithms based on a PubMed query for "focal segmental glomerulosclerosis" (2020, July)**

Detection	Number of genes	Genes
Three algorithms	12	*TRPC6, ACTN4, APOL1, INF2, CD2AP, NPHS2, CRB2, NPHS1, COL4A3, PODXL, COL4A4, LMX1B*
Genie and ConQ-R-Bio	80	*MYH9, PAX2, VEGFA, TGFB1, PLAUR, TNF, ACE, CTNNB1, IL6, MYO1E, APOE, ITGB1, ITGB3, IL10, TLR4, WT1, CCL2, MET, PLAU, SMAD3, AGT, YAP1, AGTR1, LCN2, ILK, B2M, FBN1, CD40LG, CYP11B2, SMAD2, COL4A5, PLCE1, CD80, LAMB2, SYNPO, HLA-DRB1, MMP9, ICAM1, ITGA3, SRC, FN1, GREM1, TP53, TNFRSF6B, STAT3, MTOR, NFKB1, MAPK1, CXCR4, RHOA, CAV1, RAC1, VCAM1, NOS3, IGF1, IL17A, ABCB1, ITGAM, SPP1, CST3, FOXP3, KDR, NOTCH1, HMGB1, MAPK3, PDGFRB, LGALS3, PIK3CA, FGF2, FLT1, CASP3, CXCL10, C3, IQGAP1, CFH, EGF, NOX4, ANGPT2, HP, HNF1A*
ConQ-R-Bio and Polysearch2	4	*WWC1, MAFB, CAMK4, CLCF1*

podocyte level, the application highlighted the specific expression of *CRB2* (alias Crumbs Cell Polarity Complex Component 2), a family component of the Crumbs cell polarity complex known to be affected by mutations in FSGS (47).

As for markers that were mainly found to be expressed by the vascular smooth muscle cells and mesangial cell subpopulation, the application focused on the expression of *PDGFRB*, which is known to have an interstitial PDGFR-beta expression that is significantly correlated to monocyte/macrophage infiltration in FSGS. PDGF receptors are prominent in areas of mesangial expansion and intertubular fibrosis (48). In normal kidney cells, *CST3* (or cystatin C) was mainly found to be expressed in vascular smooth muscle cells and mesangial cells. It mapped in the genome near polymorphisms associated with an increased risk of developing end-stage renal disease, possibly followed by FSGS complications (49). The link between the expression of fibronectin 1, the TGF-beta signaling pathway, and chronic progressive kidney disease (50) was confirmed as being restricted to the vascular smooth muscle cells and mesangial cell subpopulation.

Concerning markers which were found to be mainly expressed in the glomerular capillary endothelial cell compartment, *HLA-DRB1* was highly expressed and the rs28366266 polymorphism upstream of the *HLA-DRB1* gene has been characterized as an independent risk allele in steroid-sensitive nephrotic syndrome (51). The application also revealed the marked expression of *NOTCH1* in the glomerular capillary endothelial cell compartment. *NOTCH1* is increased in glomerular epithelial cells in the context of diabetic nephropathy and FSGS (52) and connected to *WT1* deregulation in podocyte (53).



https://fsgstx.shinyapps.io/textmining/

Figure 5. Screenshot of the web interface for users with a description of the interactive visualization process. After gene ID selection (process 1), the single cell expression of text-mining FSGS markers is displayed using in interactive t-SNE analysis in the central panel with the number of positive cells, the violinplot of expression and a statistical summary by group of cells.

CONCLUSION

The public PubMed database is a collection of all scientific publications updated periodically and whose web API can be targeted by crossing information with respect to molecular identifiers. Many bioinformatics tools based on text mining algorithms have been developed to target this type of query. During our work we observed that it was interesting to use these tools in order to interpret omics data by reducing their dimensions to those elicited by these text mining algorithms. This methodology reduces the error of false discovery in these high dimensional biological experiments.

The development of the WEB new generation currently makes it possible to develop online graphical interfaces that can interact with the user. In our case, we were able to develop a web interface allowing the visualization of single cell transcriptome data which were initially selected by text mining tools. By this way the integration of text mining scientific literature in omics experiments, followed by the development of an interactive web visualization application has enabled the rapid establishment of connections between cell deregulations in a pathophysiological context.

Conflict of interest: The authors declare no potential conflicts of interest with respect to the research, authorship and publication of this chapter.

Copyright and permission statement: The authors confirm that the materials included in this chapter do not violate copyright laws. Where relevant, appropriate permissions have been obtained from the original copyright holder(s), and all original sources have been appropriately acknowledged or referenced.

REFERENCES

1. Campillos M, Kuhn M, Gavin A-C, Jensen LJ, Bork P. Drug target identification using side-effect similarity. Science.2008;321(5886):263–6. https://doi.org/10.1126/science.1158140
2. Krallinger M, Valencia A, Hirschman L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol. 2008;9 Suppl 2:S8. https://doi.org/10.1186/gb-2008-9-s2-s8
3. Ashburn TT, Thor KB. Drug repositioning: identifying and developing new uses for existing drugs. Nat Rev Drug Discov.2004;3(8):673–83. https://doi.org/10.1038/nrd1468
4. Swanson DR. Medical literature as a potential source of new knowledge. Bull Med Libr Assoc.1990;78(1):29–37.
5. Rani J, Shah ABR, Ramachandran S. pubmed.mineR: an R package with text-mining algorithms to analyse PubMed abstracts. J Biosci. oct 2015;40(4):671–82. https://doi.org/10.1007/s12038-015-9552-2
6. Robinson JS and D. Text Mining with R [Internet]. [cited 16 sept 2020]. Available on: https://www.tidytextmining.com/
7. Homolak J, Kodvanj I, Virag D. Preliminary analysis of COVID-19 academic information patterns: a call for open science in the times of closed borders. Scientometrics.2020;1–15. https://doi.org/10.20944/preprints202003.0443.v1
8. Fortunato S, Bergstrom CT, Börner K, Evans JA, Helbing D, Milojević S, et al. Science of science. Science.2018;359(6379). https://doi.org/10.1126/science.aao0185

9. Entrez Programming Utilities Help [Internet]. National Center for Biotechnology Information (US); 2010.Available on: https://www.ncbi.nlm.nih.gov/books/NBK25501/
10. Brancotte B, Rance B, Denise A, Cohen-Boulakia S. ConQuR-Bio: Consensus Ranking with Query Reformulation for Biological Data. In: Data Integration in the Life Sciences. Springer; 2014. p. 128–42. https://doi.org/10.1007/978-3-319-08590-6_13
11. Toonen EJM, Fleuren WWM, Nässander U, van Lierop M-JC, Bauerschmidt S, Dokter WHA, et al. Prednisolone-induced changes in gene-expression profiles in healthy volunteers. Pharmacogenomics. 2011;12(7):985–98. https://doi.org/10.2217/pgs.11.34
12. Frijters R, van Vugt M, Smeets R, van Schaik R, de Vlieg J, Alkema W. Literature mining for the discovery of hidden connections between drugs, genes and diseases. PLoS Comput Biol. 2010;6(9). https://doi.org/10.1371/journal.pcbi.1000943
13. Morgan AA, Lu Z, Wang X, Cohen AM, Fluck J, Ruch P, et al. Overview of BioCreative II gene normalization. Genome Biol. 2008;9 Suppl 2:S3. https://doi.org/10.1186/gb-2008-9-s2-s3
14. Wermter J, Tomanek K, Hahn U. High-performance gene name normalization with GeNo. Bioinformatics. 2009;25(6):815–21. https://doi.org/10.1093/bioinformatics/btp071
15. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci U S A. 2005;102(43):15545–50. https://doi.org/10.1073/pnas.0506580102
16. Hettne KM, Boorsma A, van Dartel DAM, Goeman JJ, de Jong E, Piersma AH, et al. Next-generation text-mining mediated generation of chemical response-specific gene sets for interpretation of gene expression data. BMC Med Genomics. 2013;6:2. https://doi.org/10.1186/1755-8794-6-2
17. Westergaard D, Stærfeldt H-H, Tønsberg C, Jensen LJ, Brunak S. A comprehensive and quantitative comparison of text-mining in 15 million full-text articles versus their corresponding abstracts. PLoS Comput Biol. 2018;14(2):e1005962. https://doi.org/10.1371/journal.pcbi.1005962
18. Comeau DC, Wei C-H, Islamaj Doğan R, Lu Z. PMC text mining subset in BioC: about three million full-text articles and growing. Bioinformatics. 2019;35(18):3533–5. https://doi.org/10.1093/bioinformatics/btz070
19. Wei C-H, Allot A, Leaman R, Lu Z. PubTator central: automated concept annotation for biomedical full text articles. Nucleic Acids Res. 2019;47(W1):W587–93. https://doi.org/10.1093/nar/gkz389
20. Lachmann A, Schilder BM, Wojciechowicz ML, Torre D, Kuleshov MV, Keenan AB, et al. Geneshot: search engine for ranking genes from arbitrary text queries. Nucleic Acids Res. 2019;47(W1):W571–7. https://doi.org/10.1093/nar/gkz393
21. Lee S, Kim D, Lee K, Choi J, Kim S, Jeon M, et al. BEST: Next-Generation Biomedical Entity Search Tool for Knowledge Discovery from Biomedical Literature. PloS One. 2016;11(10):e0164680. https://doi.org/10.1371/journal.pone.0164680
22. Lee K, Kim B, Choi Y, Kim S, Shin W, Lee S, et al. Deep learning of mutation-gene-drug relations from the literature. BMC Bioinformatics.2018;19(1):21. https://doi.org/10.1186/s12859-018-2029-1
23. Fontaine J-F, Priller F, Barbosa-Silva A, Andrade-Navarro MA. Génie: literature-based gene prioritization at multi genomic scale. Nucleic Acids Res. july 2011;39(Web Server issue):W455-461. https://doi.org/10.1093/nar/gkr246
24. Candelier J-J. The hydatidiform mole. Cell Adhes Migr. 2016;10(1–2):226–35. https://doi.org/10.1080/19336918.2015.1093275
25. Desterke C, Martinaud C, Ruzehaji N, Le Bousse-Kerdilès M-C. Inflammation as a Keystone of Bone Marrow Stroma Alterations in Primary Myelofibrosis. Mediators Inflamm. 2015;2015:415024. https://doi.org/10.1155/2015/415024
26. Desterke C, Slim R, Candelier J-J. A bioinformatics transcriptome meta-analysis highlights the importance of trophoblast differentiation in the pathology of hydatidiform moles. Placenta. 2018;65:29–36. https://doi.org/10.1016/j.placenta.2018.04.002
27. Tomino Y. Pathogenesis and treatment of chronic kidney disease: a review of our recent basic and clinical data. Kidney Blood Press Res. 2014;39(5):450–89. https://doi.org/10.1159/000368458
28. Candelier J-J, Lorenzo H-K. Idiopathic nephrotic syndrome and serum permeability factors: a molecular jigsaw puzzle. Cell Tissue Res.2020;379(2):231–43. https://doi.org/10.1007/s00441-019-03147-y

29. Beaudreuil S, Lorenzo HK, Elias M, Nnang Obada E, Charpentier B, Durrbach A. Optimal management of primary focal segmental glomerulosclerosis in adults. Int J Nephrol Renov Dis. 2017;10:97–107. https://doi.org/10.2147/IJNRD.S126844
30. Sullivan KM, Susztak K. Unravelling the complex genetics of common kidney diseases: from variants to mechanisms. Nat Rev Nephrol. 2020;16:628–640. https://doi.org/10.1038/s41581-020-0298-1
31. Klein J, Jupp S, Moulos P, Fernandez M, Buffin-Meyer B, Casemayou A, et al. The KUPKB: a novel Web application to access multiomics data on kidney disease. FASEB J. 2012;26(5):2145–53. https://doi.org/10.1096/fj.11-194381
32. Liu Y, Liang Y, Wishart D. PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more. Nucleic Acids Res. 2015;43(W1):W535–542. https://doi.org/10.1093/nar/gkv383
33. Butler A, Hoffman P, Smibert P, Papalexi E, Satija R. Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat Biotechnol.2018;36(5):411–20. https://doi.org/10.1038/nbt.4096
34. Menon R, Otto EA, Hoover P, Eddy S, Mariani L, Godfrey B, et al. Single cell transcriptomics identifies focal segmental glomerulosclerosis remission endothelial biomarker. JCI Insight. 2020;5(6). https://doi.org/10.1172/jci.insight.133267
35. Xu C, Su Z. Identification of cell types from single-cell transcriptomes using a novel clustering method. Bioinformatics.2015;31(12):1974–80. https://doi.org/10.1093/bioinformatics/btv088
36. Levine JH, Simonds EF, Bendall SC, Davis KL, Amir ED, Tadmor MD, et al. Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. Cell. 2015;162(1):184–97. https://doi.org/10.1016/j.cell.2015.05.047
37. Conesa A, Madrigal P, Tarazona S, Gomez-Cabrero D, Cervera A, McPherson A, et al. A survey of best practices for RNA-seq data analysis. Genome Biol.2016;17:13. https://doi.org/10.1186/s13059-016-1047-4
38. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol.2018;19(1):15. https://doi.org/10.1186/s13059-017-1382-0
39. Trapnell C, Cacchiarelli D, Grimsby J, Pokharel P, Li S, Morse M, et al. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. Nat Biotechnol. april 2014;32(4):381–6. https://doi.org/10.1038/nbt.2859
40. Fan J, Salathia N, Liu R, Kaeser GE, Yung YC, Herman JL, et al. Characterizing transcriptional heterogeneity through pathway and gene set overdispersion analysis. Nat Methods.2016;13(3):241–4. https://doi.org/10.1038/nmeth.3734
41. Usui T, Morito N, Shawki HH, Sato Y, Tsukaguchi H, Hamada M, et al. Transcription factor MafB in podocytes protects against the development of focal segmental glomerulosclerosis. Kidney Int. 2020;98(2):391–403. https://doi.org/10.1016/j.kint.2020.02.038
42. Gebeshuber CA, Kornauth C, Dong L, Sierig R, Seibler J, Reiss M, et al. Focal segmental glomerulosclerosis is induced by microRNA-193a and its downregulation of WT1. Nat Med. 2013;19(4):481–7. https://doi.org/10.1038/nm.3142
43. Morita M, Mii A, Shimizu A, Yasuda F, Shoji J, Masuda Y, et al. Glomerular endothelial cell injury and focal segmental glomerulosclerosis lesion in idiopathic membranous nephropathy. PloS One. 2015;10(4):e0116700. https://doi.org/10.1371/journal.pone.0116700
44. Copelovitch L, Guttenberg M, Pollak MR, Kaplan BS. Renin-angiotensin axis blockade reduces proteinuria in presymptomatic patients with familial FSGS. Pediatr Nephrol Berl Ger. 2007;22(10):1779–84. https://doi.org/10.1007/s00467-007-0505-3
45. Malone AF, Phelan PJ, Hall G, Cetincelik U, Homstad A, Alonso AS, et al. Rare hereditary COL4A3/COL4A4 variants may be mistaken for familial focal segmental glomerulosclerosis. Kidney Int. 2014;86(6):1253–9. https://doi.org/10.1038/ki.2014.305
46. Zhang P, Zhuo L, Zou Y, Li G, Peng K. COL4A5 mutation causes Alport syndrome with focal segmental glomerulosclerosis lesion: Case report and literature review. Clin Nephrol. 2019;92(2):98–102. https://doi.org/10.5414/CN109737
47. Fan J, Fu R, Ren F, He J, Wang S, Gou M. A case report of CRB2 mutation identified in a Chinese boy with focal segmental glomerulosclerosis. Medicine (Baltimore). 2018;97(37):e12362. https://doi.org/10.1097/MD.0000000000012362

48. Stein-Oakley AN, Maguire JA, Dowling J, Perry G, Thomsom NM. Altered expression of fibrogenic growth factors in IgA nephropathy and focal and segmental glomerulosclerosis. Kidney Int. 1997;51(1):195–204. https://doi.org/10.1038/ki.1997.24
49. Cyrus C, Chathoth S, Vatte C, Alrubaish N, Almuhanna O, Borgio JF, et al. Novel Haplotype Indicator for End-Stage Renal Disease Progression among Saudi Patients. Int J Nephrol. 2019;2019:1095215. https://doi.org/10.1155/2019/1095215
50. Tamaki K, Okuda S, Ando T, Iwamoto T, Nakayama M, Fujishima M. TGF-beta 1 in glomerulosclerosis and interstitial fibrosis of adriamycin nephropathy. Kidney Int.1994;45(2):525–36. https://doi.org/10.1038/ki.1994.68
51. Debiec H, Dossier C, Letouzé E, Gillies CE, Vivarelli M, Putler RK, et al. Transethnic, Genome-Wide Analysis Reveals Immune-Related Risk Alleles and Phenotypic Correlates in Pediatric Steroid-Sensitive Nephrotic Syndrome. J Am Soc Nephrol. 2018;29(7):2000–13. https://doi.org/10.1681/ASN.2017111185
52. Niranjan T, Bielesz B, Gruenwald A, Ponda MP, Kopp JB, Thomas DB, et al. The Notch pathway in podocytes plays a role in the development of glomerular disease. Nat Med.2008;14(3):290–8. https://doi.org/10.1038/nm1731
53. Asfahani RI, Tahoun MM, Miller-Hodges EV, Bellerby J, Virasami AK, Sampson RD, et al. Activation of podocyte Notch mediates early Wt1 glomerulopathy. Kidney Int. 2018;93(4):903–20. https://doi.org/10.1016/j.kint.2017.11.014