NEURAL REGENERATION RESEARCH
www.nrronline.org

Utilizing Single-cell and Spatial RNA-seq databasE for Alzheimer's Disease (ssREAD) in hypothesis-driven queries

Diana Acosta, Cankun Wang, Qin Ma*, Hongjun Fu*

Alzheimer's disease (AD) is the most common form of dementia. In addition to the lack of effective treatments, there are limitations in diagnostic capabilities. The complexity of AD itself, together with a variety of other diseases often observed in a patient's history in addition to their AD diagnosis, make deciphering the molecular mechanisms that underlie AD, even more important. Large datasets of single-cell RNA sequencing, single-nucleus RNA-sequencing (snRNA-seq), and spatial transcriptomics (ST) have become essential in guiding and supporting new investigations into the cellular and regional susceptibility of AD. However, with unique technology, software, and larger databases emerging; a lack of integration of these data can contribute to ineffective use of valuable knowledge. Importantly, there was no specialized database that concentrates on ST in AD that offers comprehensive differential analyses under various conditions, such as sex-specific, region-specific, and comparisons between AD and control groups until the new Single-cell and Spatial RNA-seq databasE for Alzheimer's Disease (ssREAD) database (Wang et al., 2024) was introduced to meet the scientific community's growing demand for comprehensive, integrated, and accessible data analysis.

Overview of ssREAD database: The ssREAD database includes 381 ST and 277 sc/snRNA-seq AD-related datasets (Jiang et al., 2020; Wang et al., 2024). These sequencing data can be queried by user-friendly modules that allow the user to investigate transcriptomic alterations in AD *vs*. control groups. In the original publication (Wang et al., 2024), ssREAD is compared to existing databases that are relevant to the field, including scREAD, scRNA-seq analysis for AD (SCAD)-Brain, The Alzheimer's Cell Atlas (TACA), Spatial Omics DataBase (SODB), and STOmicsDB. Overall, ssREAD maintains a comprehensive collection of AD-related single-cell RNA sequencing, snRNA-seq, and ST datasets, that stands unrivaled compared to other databases. Compared to the other databases, ssREAD is composed of 277 sc/sn RNAsq samples from 67 studies, as well as 381 ST samples from 18 studies. While other databases cover a fewer range of samples and studies: scREAD (289 sc/sn samples across 67 studies), SCAD-Brain (359 sc/sn samples across 21 studies), TACA (455 sc/sn samples across 26 studies), SODB (155 spatial transcriptomic samples across 4 studies), and STOmicsDB (1 sc/sn samples across 1 study; and 101 spatial transcriptomic samples across 7 studies) (Wang et al., 2024).

To highlight the importance of regional and cellular vulnerability in AD, these transcriptomic alterations can be resolved at the sub-cellular, cellular, and spatial levels. ssREAD includes sc/snRNA-seq data from 144 human samples and 133 datasets from mouse samples. ST datasets included in ssREAD are from 319 mouse datasets, showcasing the difficulty in producing human datasets. From these datasets, ssREAD can be used to query important questions, including the use of ST datasets to analyze subpopulation differences in cortical layer gene expression between disease and control groups. Importantly, the ssREAD publication highlights there are key differences in layer 5, which is also known to have high phosphorylated tau in disease cases (Wang et al., 2024). In addition to these findings, ssREAD was used to unveil sex-specific differences in AD at the cellular level through an integrated analysis of spatial and single-cell transcriptomics (Wang et al., 2024).

Sex-specific differences observed from ssREAD: The integration of datasets included the addition of the Seattle Alzheimer's disease brain cell atlas (Miller et al., 2023), an atlas that includes cells derived from control and AD human middle temporal gyrus. The data revealed there was no batch effect among samples regarding Braak stage, Thal phase, and ethnicity. Obvious sex-oriented differences in cell clusters were observed, which may contribute to the possible pathological sex-bias differences in AD. Additionally, sc/snRNA-seq data from previously published original research that focused on the molecular landscape of over 183k cells in human brain hippocampus vasculature in AD (AD019) was utilized to investigate sex differences across all 16 samples from the original study (Yang et al., 2022). From the original data, a new UMAP was created, and 13 cell types were assigned (i.e., Arterial cell, Astrocyte, Capillary cell, Ependymal cell, Fibroblast, Microglia, Neuron, Oligodendrocyte progenitor cell, Oligodendrocytes, Pericyte, Smooth muscle cell, T cell, and Venous). Sex-specific differences were again observed, including more Oligodendrocytes, Astrocytes, and Microglia in overall female cell types than male, and 44 upregulated and 108 downregulated genes in male microglia *vs*. female. Downregulated genes included previously identified ARM and DAM genes such as ATP1B3 (Keren-Shaul et al., 2017). Upregulated genes included LHFPL2 and RTTN, which are also DAM genes (Keren-Shaul et al., 2017). The findings using ssREAD indicate that both sex and disease status can shape the transcriptomic landscape of cells in disease.

Case Study – Optimizing the use of ssREAD by applying hypothesis-driven queries to identify differentially expressed genes in AD *vs*. CT groups that are unique to the females: The information that can be revealed from the use of ssREAD is a function of the integration of available AD datasets in one repository and of the user's questions. In **Figure 1A** we highlight important features of the available datasets on ssREAD. It is important to note that depending on the user's hypothesis-driven query, one can select to focus on cell-type specific differences in which case the use of sc/snRNA-seq AD-related datasets would be the most appropriate, or layer-specific differences in which case the use of ST datasets would provide the spatial information to best address such investigations. For example, to further understand the differences in differentially expressed genes (DEGs) between two different sexes, one must consider several factors including the region, the cell type, and the type of comparison to be made across four demographic groups (Male AD patients, Male controls, Female AD patients, Female controls), as shown in **Figure 1B**. Insight into the importance of region specificity and cell type should be guided by the user's interest and previous literature suggesting specific regions showcase sex differences. As such, the use of hypothesis-driven queries can optimize the use of databases such as ssREAD, which offers access to a large amount of information and guides downstream analysis and functional studies.

As an example, we use the decision tree to formulate a query in which we wish to compare DEGs across the demographic groups: Male AD patients *vs*. Male controls, and Female AD patients vs. Female controls. The result showcases both unique and shared gene signatures among these groups. By comparing the gene expressions between males and females, we revealed unique sex-specific gene signatures within different cell types (**Figure 1C** and **Additional Table 1**; all data was generated by ssREAD for use in this article and has not been previously published.). For example, within the Excitatory Neuron population, 28 genes overlapped between male and female comparisons, while 10 genes were unique to the female comparison. The eight genes included AC023590.1, ANK2-AS1, KCNMB2-AS1, LINC00513, MAPK10-AS1, RP11-473C19.1, RP11-390N6.1, and RP11-556G22.2. KCNMB2-AS1 has been shown to be overexpressed as an oncogene in cancer cells (Hao et al., 2023). These findings further showcase the importance of hypothesis-driven queries in using ssREAD, to address specific user questions and yield genes of interest that are dependent on the topic of interest. In addition to excitatory neurons, we find 57 genes are unique to female comparisons in oligodendrocytes; 75 genes that are unique to female comparisons in astrocytes; and 30 genes that are unique to female comparisons in microglia (**Additional Table 1**). Further investigation into the function of such genes and how they may be related specifically to AD *vs*. control changes in the female population are warranted.

Conclusions and perspectives: The ssREAD database meets the AD field's need for a database that can consolidate emerging transcriptomic datasets into a user-friendly framework. Importantly, the access to this wealth of data that ssREAD provides can guide users in their ongoing or newly emerging investigations. Furthermore, access to vast data can be under-utilized or ineffective without the proper formulation of hypothesis-driven queries. A hypothesis-driven query such as the case study we highlight, in which we investigate the differentially expressed genes that are unique to female comparisons of AD *vs*. control groups, puts into perspective the literature-driven decisions that must guide this question, including which region to choose, which cell type of interest, and which comparison to delve into. In the above case study, the middle temporal gyrus region was selected due to its prominent role in early AD pathogenesis. The cell types such as excitatory neurons, oligodendrocytes, microglia, and astrocytes, were selected due to the interest in the AD field and their relevance to specific disease functions, including the cell-type vulnerability of Excitatory Neurons (Fu et al., 2019).

The addition of transcriptomic datasets from other neurodegenerative diseases and disease models can support investigations of shared disease mechanisms: The ssREAD database is a perpetual tool for AD investigators. It will continue to grow as new datasets and technology emerge in the field. As a growing resource, it will also be beneficial to include datasets from other neurodegenerative diseases, including frontotemporal lobar degeneration, Parkinson's disease, and amyotrophic lateral sclerosis. The addition of several disease datasets will allow for new queries into the similarities and differences across diseases, which is also of major interest in the field of neurodegeneration due to underlying molecular mechanisms such as deficits in protein folding and clearance that are shared across diseases.

As highlighted by the ssREAD paper (Wang et al., 2024), the transcriptomic datasets are largely comprised of mouse and human datasets, likely due to the availability of good models for AD. However, the addition of transcriptomic datasets from 2D and 3D cell culture systems such as iPSC-derived neurons and human neural organoids can be beneficial. The use of these models can be compared to human and mouse datasets, and their overlaps or deficits can progress the field's use of such *in vitro* models, solidify their interpretations, and highlight their limitations.

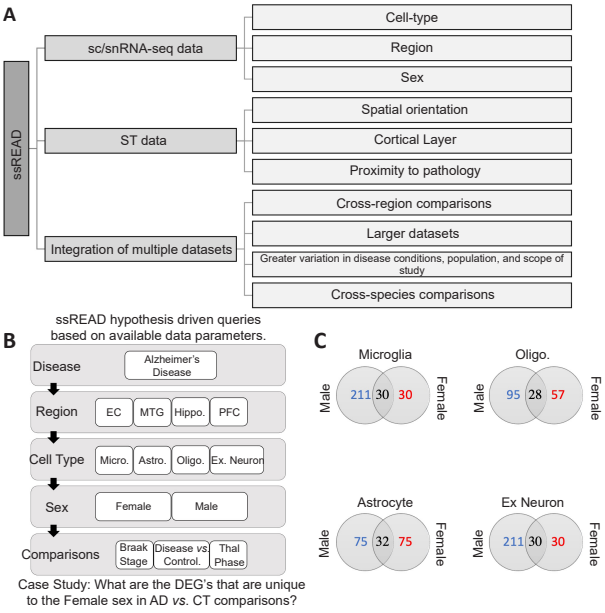

Figure 1 | **Hypothesis-driven queries using ssREAD.**
(A) Overview of available datasets and their unique attributes to consider in building hypothesis-driven queries using ssREAD. (B) Decision tree for the development of hypothesis-driven queries to be used for the ssREAD database. Important components include the region of interest, cell type of interest, and final comparison required. The decision tree is generated to address the case study presented which seeks to investigate the difference in DEGs found in comparisons made for Male controls *vs*. Male AD groups to Female controls *vs*. Female AD groups. (C) Overview of results from a case study investigating the DEGs that are unique to the Females in AD *vs*. CT comparisons separated by cell type (Oligo.: Oligodendrocyte). Blue indicates the number of DEGs that are unique to the Male AD *vs*. CT group. Red indicates the number of DEGs that are unique to the Female AD *vs*. CT group. Black indicates the number of DEGs that are shared by the Male and Female. Created with Microsoft PowerPoint Version 16.77. AD: Alzheimer's disease; Astro: astrocyte; CT: control; DEGs: differentially expressed genes; EC: entorhinal cortex; Ex: excitatory; Hippo: hippocampus; Micro: microglia; MTG: middle temporal gyrus; Oligo: oligodendrocyte; PFC: prefrontal cortex; sc/snRNA-seq: single-cell and single-nucleus RNA-sequencing; ssREAD: Single-cell and Spatial RNAseq databasE for Alzheimer's Disease; ST: spatial transcriptomics.

The addition of OMICS datasets to transcriptomic datasets can complement and guide downstream mechanistic assays: Transcriptomic datasets can guide mechanistic studies that can reveal functional properties of specific proteins of interest as well as their role in pathways that may contribute to disease onset or progression. Incorporating other Omics datasets such as genomics, epigenomics, proteomics, and metabolomics can provide further insight into the changes occurring at multiple biological levels, which may serve as a valuable roadmap for designing and conducting subsequent functional studies. Adding such datasets to ssREAD can be of huge benefit to the field and would transform ssREAD into a toolbox for investigators to plan well-supported experimental designs. Single-cell level omics integration with emerging technologies that can profile thousands of genes to whole transcriptome such as Visium HD also make insight into cellular and regional vulnerability more accessible. However, there are several limitations to the addition of such datasets, including large multimodality datasets integrating technologies and risking the simplicity of ssREAD by adding such datasets. Careful consideration into the user-friendly framework of ssREAD with new datasets is warranted.

Prospective benefits and limitations of applying Artificial Intelligence (AI) to transcriptomic analysis and working databases: To aid in the management of complex and large datasets being consolidated into databases such as ssREAD, AI can be well-equipped to tackle challenges in their maintenance, but the tools are underutilized and are only recently beginning to be applied. For example, scGNN^{+} (single-cell graph neural network) can be used to identify neuron clusters and cell type-specific markers in an AD-based study (Jiang et al., 2024). Other notable uses of AI applied to single-cell RNA sequencing data include scGPT which is used to identify genes for cell states (Cui et al., 2024). However, the use of semi-supervised learning and self-supervised learning in predicting components of data requires preliminary biological knowledge or data into the model to improve predictions as constraints. Additionally, the use of statistical assessments such as *P*-values and z-scores of prediction results can guide non-computational users to make biological inferences. Deep learning models furthermore can be driven by hypothesis-based queries such as feature orders and relationships. However, the use of AI requires several training data from publicly available data that is published or in-house data, computational resources for their training and ultimately requires highly specific computational skills for their development and application. Lastly, these models will require ongoing adaptation due to new data from different species, scalability, and interpretations of the training data and parameters being used to create such models. However, it has been shown that deep learning methods can be applied to a broad range of single-cell studies (Ma and Xu, 2022; Ma et al., 2024). Furthermore, the application of large language models and Conversational AI presents a new approach to data interaction and analysis. Notably, scGNN+ (Jiang et al., 2024) combines the power of GNNs with the capabilities of ChatGPT. This combination enhances reproducibility, code optimization, and visualization, and opens up opportunities for more sophisticated and efficient data analysis in ssREAD. However, it should be noted that Conversational AI models may occasionally provide inaccurate answers and should be employed with care. Additionally, the task of arranging and structuring results within ssREAD to ensure their interpretability by large language models can be a complex and challenging process. BioChatter was a new approach that interfaces with large language models in the biomedical space and uses knowledge graphs and prompt engineering techniques to improve the accuracy of results (Lobentanzer, 2023). By integrating such advancements in conversational AI into ssREAD, it is possible to create an environment where researchers can directly "chat" with data. This development could significantly increase the platform's intuitiveness and accessibility.

Diana Acosta, Cankun Wang, Qin Ma*, Hongjun Fu*
Department of Neuroscience, The Ohio State University, Columbus, OH, USA (Acosta D, Fu H)
Department of Biomedical Informatics, The Ohio State University, Columbus, OH, USA (Wang C, Ma Q)
Chronic Brain Injury Program, The Ohio State University, Columbus, OH, USA (Fu H)
***Correspondence to:** Qin Ma, PhD, qin.ma@osumc.edu; Hongjun Fu, PhD, hongjun.fu@osumc.edu.
https://orcid.org/0000-0002-3264-8392 (Qin Ma)
https://orcid.org/0000-0001-5346-7075 (Hongjun Fu)
Date of submission: October 8, 2024
Date of decision: November 22, 2024
Date of acceptance: December 2, 2024
Date of web publication: January 13, 2025

https://doi.org/10.4103/NRR.NRR-D-24-01201
How to cite this article: *Acosta D, Wang C, Ma Q, Fu H (2026) Utilizing Single-cell and Spatial RNA-seq databasE for Alzheimer's Disease (ssREAD) in hypothesis-driven queries. Neural Regen Res 21(2):677-678.*

Additional file:
Additional Table 1: *Sex-specific gene signatures within different cell types in human AD.*

References

Cui H, Wang C, Maan H, Pang K, Luo F, Duan N, Wang B (2024) scGPT: toward building a foundation model for single-cell multi-omics using generative AI. Nat Methods 21:1470-1480.
Fu H, Possenti A, Freer R, Nakano Y, Hernandez Villegas NC, Tang M, Cauhy PVM, Lassus BA, Chen S, Fowler SL, Figueroa HY, Huey ED, Johnson GVW, Vendruscolo M, Duff KE (2019) A tau homeostasis signature is linked with the cellular and regional vulnerability of excitatory neurons to tau pathology. Nat Neurosci 22:47-56.
Hao L, Zhang J, Liu Z, Zhang Z, Mao T, Guo J (2023) Role of the RNA-binding protein family in gynecologic cancers. Am J Cancer Res 13:3799-3821.
Jiang J, Wang C, Qi R, Fu H, Ma Q (2020) scREAD: a Single-Cell RNA-Seq Database for Alzheimer's Disease. iScience 23:101769.
Jiang Y, Wang S, Feng S, Wang C, Wu W, Huang X, Ma Q, Wang J, Ma A (2024) scGNN+: Adapting ChatGPT for Seamless Tutorial and Code Optimization. bioRxiv doi: https://doi.org/10.1101/2024.09.30.615735 [preprint].
Keren-Shaul H, Spinrad A, Weiner A, Matcovitch-Natan O, Dvir-Szternfeld R, Ulland TK, David E, Baruch K, Lara-Astaiso D, Toth B, Itzkovitz S, Colonna M, Schwartz M, Amit I (2017) A unique microglia type associated with restricting development of Alzheimer's disease. Cell 169:1276-1290.
Lobentanzer S, Feng S, The BioChatter Consortium, Maier A, Wang C, Baumbach J, Krehl N, Ma Q, Saez-Rodriguez J (2023) A platform for the biomedical application of large language models. arXiv doi:10.48550/arXiv.2305.06488.
Ma Q, Xu D (2022) Deep learning shapes single-cell data analysis. Nat Rev Mol Cell Biol 23:303-304.
Ma Q, Jiang Y, Cheng H, Xu D (2024) Harnessing the deep learning power of foundation models in single-cell omics. Nat Rev Mol Cell Biol 25:593-594.
Miller JA, et al. (2023) SEA-AD: scientific analysis and open access resources targeting early changes in Alzheimer's disease. Alzheimers Dement 19:e063478.
Wang C, Acosta D, McNutt M, Bian J, Ma A, Fu H, Ma Q (2024) A single-cell and spatial RNA-seq database for Alzheimer's disease (ssREAD). Nat Commun 15:4710.
Yang AC, et al. (2022) A human brain vascular atlas reveals diverse mediators of Alzheimer's risk. Nature 603:885-892.

C-Editors: Zhao M, Liu WJ, Qiu Y; T-Editor: Jia Y

利用单细胞和空间 RNA-seq 数据库进行阿尔茨海默病研究（ssREAD），用于基于假设的查询
文章特色分析

一、文章重要性

1. 填补 AD 研究中的数据整合空白

- 目前 AD 研究面临数据分散、技术多样、分析工具不统一等问题，缺乏一个专门整合单细胞 RNA 测序与空间转录组数据的数据库。

- ssREAD 的出现填补了这一空白，提供了统一、可访问、多条件比较的数据平台，支持研究者进行跨样本、跨条件、跨细胞类型的差异分析。

2. 推动 AD 机制研究的精细化与空间化

- 文章强调了 AD 中区域特异性与细胞类型特异性的转录组变化，尤其是皮质层、性别差异、病理邻近性等维度，为理解 AD 的异质性提供了新视角。

3. 促进假设驱动的研究范式

- 文章提出并示范了“假设驱动查询”的研究方法，引导用户基于明确的科学问题从数据库中提取有价值的信息，避免了“数据海洋中盲目捕捞”的问题。

二、创新性特色

1. 数据库整合的广度与深度

- ssREAD 整合了来自 381 个空间转录组样本和 277 个单细胞/单核 RNA-seq 样本，覆盖人和小鼠等多种样本类型，是目前 AD 领域最全面的转录组数据库之一。

2. 支持多维度、多条件查询

- 用户可根据性别、脑区、细胞类型、疾病状态等条件进行灵活查询，实现精准的差异表达分析。

3. 结合 AI 与自然语言处理技术

- 文章前瞻性地提出将图神经网络、scGPT、ChatGPT 等 AI 工具应用于数据挖掘与可视化，提升数据分析的自动化与智能化水平。

- 引入 Biochatter 等工具，使研究者能通过“对话”方式与数据库交互，降低使用门槛。

4. 案例研究展示实用性

- 通过具体案例（如性别特异性基因在兴奋性神经元、小胶质细胞等中的表达），展示了如何利用 ssREAD 发现新的生物学见解。

三、对学科的启示

1. 推动神经退行性疾病研究的“数据驱动”转型

- ssREAD 不仅是数据仓库，更是假设生成与验证平台，标志着 AD 研究从“描述性”向“机制性”与“预测性”转变。

2. 促进多组学与空间生物学融合

- 文章呼吁将基因组、表观组、蛋白质组、代谢组等其他组学数据整合进 ssREAD，构建更全面的 AD 分子图谱。

3. 为其他神经退行性疾病提供模板

- ssREAD 的构建理念与方法可推广至帕金森病、肌萎缩侧索硬化等疾病，推动神经科学数据库的标准化与互通性。

4. 强调“可解释 AI”在生物医学中的应用

- 文章指出 AI 模型需结合生物学先验知识，提升结果的可解释性与可靠性，为生物信息学与临床研究的结合指明方向。

总结

这篇文章不仅介绍了一个重要的 AD 研究资源——ssREAD 数据库，更提出了一种以假设驱动、AI 辅助、多组学整合为核心的现代生物医学研究范式。其数据整合的广度、查询设计的灵活性、技术的前瞻性，使其在 AD 乃至整个神经科学领域具有重要的示范与推动作用。