Available Online: 22 July 2026 DOI: 10.54254/3049-5458/2026.35570

Statistical integration of GWAS, eQTL and pQTL data for multi-omics analysis of Alzheimer's disease

Zhuolan Li

Guangzhou Dublin International College of Life Sciences and Technology, South China Agricultural University, Guangzhou, China

anyumi83@163.com

Abstract. Alzheimer's Disease (AD) is a complex neurodegenerative disorder with a strong genetic architecture. Genome-Wide Association Studies (GWAS) have identified numerous susceptibility loci. However, the majority of associated variants reside in non-coding regions, making it difficult to resolve their functional consequences and identify causal genes. To address this limitation, integration of GWAS with expression Quantitative Trait Loci (eQTL) and protein Quantitative Trait Loci (pQTL) data has emerged as a key strategy for linking genetic variation to downstream molecular phenotypes. This review discusses statistical frameworks for multi-omics integration in AD research, with a focus on approaches that enable causal inference and gene prioritization. Major methods include colocalization analysis for detecting shared causal variants, Mendelian Randomization (MR) for assessing putative causal relationships, and Transcriptome-Wide Association Studies (TWAS) for linking genetically predicted gene expression to disease risk. Applications of these frameworks have facilitated the identification of candidate causal genes and proteins, thereby improving the mechanistic interpretation of AD-associated loci. However, challenges remain, including tissue specificity and cell-type specificity, limited ancestral diversity in available datasets, and constraints in causal inference. Emerging single-cell and spatial multi-omics approaches are expected to provide a more detailed characterization of AD-associated molecular mechanisms while supporting therapeutic target discovery.

Keywords: Alzheimer's disease, Genome-Wide Association Studies, expression Quantitative Trait Loci, protein Quantitative Trait Loci, multi-omics integration

1. Introduction

Alzheimer's disease is the leading cause of dementia worldwide and has become a major public health challenge as populations continue to age [1]. Genetic factors play an important role in the development of Alzheimer's Disease (AD), and Genome-Wide Association Studies (GWAS) have identified numerous AD-associated genetic loci.

Despite these advances, most variants identified by GWAS are located in non-coding regions of the genome, making it difficult to determine the genes they regulate or the biological processes through which they contribute to AD pathogenesis [2]. As the number of susceptibility loci continues to grow, research has

gradually shifted from identifying genetic associations to understanding their functional consequences. This has led to increasing interest in molecular quantitative trait loci, particularly expression Quantitative Trait Loci (eQTL) and protein Quantitative Trait Loci (pQTL), which serve as an important bridge between genetic variation and downstream molecular phenotypes [3].

At the same time, the growing availability of transcriptomic and proteomic datasets has promoted the development of statistical approaches that integrate GWAS with molecular QTL data [4]. This review discusses statistical methods for integrating GWAS, eQTL and pQTL data in Alzheimer's disease research. It focuses on how these approaches have been applied to interpret genetic associations and identify candidate causal genes, providing an overview of current progress in statistical multi-omics analysis of AD.

2. Genetic associations underlying Alzheimer's disease

Large-scale Genome-Wide Association Studies have greatly expanded the catalogue of genetic loci associated with AD. Increasing sample sizes and large international meta-analyses have identified additional AD-associated loci, providing a more comprehensive view of the genetic architecture of AD [2]. Rather than implicating only a few high-risk genes, these findings indicate that genetic susceptibility involves multiple biological processes.

AD genetic studies have implicated multiple biological processes, including amyloid processing, innate immunity, lipid metabolism, and endosomal trafficking [1]. Early genetic studies supported the involvement of amyloid-related pathways in AD, with pathogenic variants affecting Amyloid Precursor Protein (APP) processing providing strong evidence for the role of amyloid biology in disease development [2]. However, large-scale GWAS have broadened this perspective by identifying risk loci associated with diverse cellular functions beyond amyloid pathways. In particular, genetic studies have increasingly implicated innate immunity and microglial function in AD susceptibility, with TREM2 emerging as a prominent example linking microglial responses to disease risk [1]. Additional associations involving lipid metabolism and endosomal trafficking, including APOE, BIN1 and PICALM, support the contribution of multiple cellular processes to AD genetic risk [5].

Although GWAS has greatly expanded the catalogue of AD susceptibility loci, genetic associations alone cannot determine which genes within these loci are functionally involved in disease development. Functional interpretation therefore requires molecular evidence that links genetic variants to downstream changes in gene expression and protein abundance.

3. Functional interpretation through eQTL and pQTL analyses

3.1. eQTL analyses linking genetic variants to gene expression

Most AD-associated risk variants are located outside protein-coding regions, making it difficult to determine which genes they influence. eQTL analyses address this question by examining how genetic variants affect gene expression, thereby offering functional evidence for GWAS loci. Large-scale eQTL resources, particularly the GTEx project, have enabled systematic mapping of regulatory variants across human tissues, laying the foundation for linking disease-associated variants to gene expression [4].

In AD, integrating GWAS with brain eQTL data has helped prioritize candidate genes beyond their genomic proximity to associated variants. Using summary-level data integration, Lee et al. identified several genes whose genetically regulated expression was associated with AD susceptibility, including previously unreported candidates [6]. Cross-tissue analyses also identified additional candidate susceptibility genes by

incorporating genetically predicted expression across multiple relevant tissues [7]. However, regulatory signals often vary between tissues, meaning that the interpretation of AD risk variants depends heavily on the availability of disease-relevant brain eQTL datasets [8]. In addition, eQTL associations may differ among specific brain regions because distinct cellular environments can shape patterns of gene regulation. This issue is particularly relevant to AD, where pathological changes involve multiple brain regions and cell populations. Therefore, eQTL-based interpretations require careful selection of biologically appropriate datasets rather than relying solely on publicly available molecular resources.

3.2. pQTL analyses revealing downstream protein regulation

Compared with eQTL analyses, pQTL analyses examine the effects of genetic variation at the protein level rather than the transcript level. Although changes in gene expression provide important clues about regulatory effects, transcript levels do not always correspond to protein abundance because protein production and stability are influenced by additional biological processes, including translation and degradation.

Protein-level regulation is particularly relevant to AD because many disease-associated processes involve changes in extracellular and cellular proteins, including pathways related to amyloid biology and immune responses. Recent advances in Cerebrospinal Fluid (CSF) proteomic profiling have facilitated the application of pQTL analyses in AD research, enabling systematic investigation of the genetic regulation of disease-related proteins [8, 9]. Integration of GWAS with CSF pQTL data has helped prioritize candidate proteins associated with AD susceptibility and provided additional evidence for interpreting genetic risk loci.

The distinct regulatory information captured by eQTL and pQTL analyses highlights the importance of examining multiple molecular layers when interpreting AD-associated variants. Combining transcript-level and protein-level evidence allows genetic associations to be evaluated from complementary perspectives, improving the functional understanding of disease-related loci.

4. Statistical integration frameworks for multi-omics analysis

4.1. Colocalization approaches for identifying shared causal variants

Although GWAS and molecular QTL studies frequently identify association signals within the same genomic region, such overlap does not necessarily indicate that the signals are driven by the same causal variant. Linkage disequilibrium can produce similar association patterns among neighboring variants, making it difficult to determine whether disease-associated loci influence gene expression or protein abundance directly. To address this issue, Bayesian colocalization methods were developed to estimate the probability that two association signals share a common causal variant, providing statistical support for linking GWAS findings with molecular QTLs [10].

In AD research, the value of colocalization extends beyond prioritizing candidate genes. It has strengthened the understanding of GWAS loci, particularly because disease-associated variants often reside in regulatory regions where the nearest gene may not represent the functional target. Rather than relying solely on physical proximity, recent studies have incorporated colocalization together with molecular QTL data to evaluate functional evidence for disease-associated loci. For example, Bellenguez et al. integrated colocalization analyses with eQTL, pQTL, transcriptome-wide association studies, and other functional evidence to prioritize candidate genes across newly identified AD risk loci, thereby improving biological interpretation of GWAS findings [2].

Despite its widespread application, colocalization provides statistical evidence for shared genetic signals rather than direct proof of causal mechanisms. In addition, the reliability of the results depends heavily on the

availability and quality of QTL datasets, particularly for disease-relevant brain tissues and cell populations. These limitations have encouraged the use of colocalization alongside complementary statistical approaches when evaluating candidate genes and potential disease mechanisms.

4.2. Mendelian randomization and transcriptome-wide association studies

Although colocalization analyses can indicate whether GWAS and QTL signals are likely to share a causal variant, they cannot determine whether molecular changes contribute directly to disease development. Even when a genetic variant is associated with both gene expression and AD risk, shared genetic regulation alone does not establish a causal relationship. To evaluate potential causal mechanisms and prioritize candidate genes, Mendelian Randomization (MR) and Transcriptome-Wide Association Studies (TWAS) have become widely adopted in the integration of GWAS with molecular QTL data.

MR uses genetic variants as instrumental variables to evaluate causal relationships between molecular traits and disease risk while reducing confounding from environmental factors. Combined with eQTL or pQTL data, MR helps assess the effects of genetically predicted changes in gene expression or protein abundance on AD risk. Compared with conventional GWAS, this approach provides additional evidence for distinguishing potential causal effects from genetic association alone.

TWAS addresses a related but distinct question by testing whether genetically predicted gene expression is associated with disease risk. Instead of evaluating causality directly, TWAS combines gene expression prediction models derived from eQTL reference datasets with GWAS summary statistics to identify genes whose predicted expression is associated with AD susceptibility [11]. Applications in AD have revealed genes associated with disease risk that cannot be explained by GWAS signals alone, with cross-tissue analyses incorporating genetically predicted expression from multiple disease-relevant tissues [12].

Although MR and TWAS provide complementary evidence for interpreting GWAS findings, both approaches have methodological limitations. MR depends on the validity of instrumental variable assumptions, whereas TWAS may be influenced by linkage disequilibrium and shared regulatory signals.

4.3. Integrative multi-omics applications in Alzheimer's disease research

Integrative multi-omics analysis is widely applied in AD research, particularly for candidate gene prioritization and functional interpretation of genetic associations. Its primary contribution is not simply the incorporation of additional datasets, but the refinement of candidate genes underlying disease-associated variants. Because a single GWAS risk locus often contains multiple nearby genes, genetic association alone is usually insufficient to determine which gene is functionally involved in disease susceptibility [2]. By integrating transcriptomic and proteomic information across different omics layers, multi-omics analysis provides a stronger basis for prioritizing genes that are more likely to participate in AD-related biological processes [7].

Rather than relying on a single analytical method, integrative analyses combine complementary approaches to evaluate candidate genes from multiple perspectives. A candidate gene is typically first identified within a GWAS risk locus. Brain eQTL analyses then examine whether the associated variant influences gene expression, providing functional support for candidate genes [13]. The pQTL analyses extend this interpretation to the protein level by evaluating genetically regulated changes in protein abundance [6]. Colocalization analysis further assesses whether GWAS and molecular QTL signals are likely to arise from the same causal variant, reducing the possibility that the observed association is explained solely by linkage disequilibrium [10]. TWAS identifies genes whose genetically predicted expression is associated with AD risk [11]. MR further evaluates whether genetically predicted molecular traits are likely to contribute to disease

susceptibility. Confidence is greatest when different analytical approaches independently support the same candidate gene or protein.

From a statistical perspective, the value of multi-omics integration lies in evaluating whether independent sources of evidence converge on the same biological interpretation rather than generating additional association signals. Because each approach captures a different aspect of molecular regulation and is subject to its own assumptions, no single method is sufficient to support a robust biological conclusion. Greater confidence is achieved when complementary approaches consistently support the same candidate gene or molecular pathway. This integrative strategy has also facilitated applications beyond gene prioritization, including biomarker discovery through proteomic profiling and the identification of molecular subtypes that reflect biological heterogeneity in AD [8, 14].

Integrative multi-omics approaches have also contributed to therapeutic target evaluation and molecular characterization of AD. Candidates supported by both genetic associations and downstream molecular alterations may represent stronger priorities for experimental validation because they connect inherited risk with functional molecular changes. In addition, multi-omics analyses can reveal biological heterogeneity by identifying distinct molecular patterns among individuals with AD, highlighting biological differences that may underlie variation in disease progression and treatment response.

5. Current challenges and future perspectives

5.1. Challenges in multi-omics integration for Alzheimer's disease research

Although multi-omics statistical integration has substantially improved the functional interpretation of genetic findings in AD research, the reliability of these conclusions does not depend solely on statistical methods. As increasing numbers of omics datasets become available, differences in data sources, sample characteristics, and biological contexts may accumulate, making it difficult for evidence from different molecular layers to always produce consistent biological interpretations. Therefore, multi-omics analysis requires not only the integration of diverse datasets, but also careful consideration of data compatibility and the limitations of statistical inference [7].

One major challenge arises from the heterogeneity among different omics datasets. Most current GWAS, eQTL and pQTL datasets are generated from independent research cohorts, which often differ in sample size, disease stage, experimental platforms, and data processing procedures. These differences may introduce systematic biases during cross-omics integration and affect the stability of downstream conclusions. In addition, AD exhibits substantial tissue and cell-type specificity, whereas many publicly available molecular datasets are still mainly derived from bulk tissues. As molecular signals from different cell populations are averaged, regulatory effects associated with specific disease-relevant cell types may be obscured [15]. Consequently, even when multiple statistical analyses generate consistent findings, the results may not fully capture the underlying biological processes occurring within specific cellular contexts.

Beyond limitations related to data characteristics, the assumptions underlying statistical models also influence the interpretation of integrated results. Colocalization analysis, MR and TWAS each rely on specific statistical assumptions, which may not always be completely satisfied in complex diseases. The presence of multiple causal variants, widespread pleiotropic effects, and complex linkage disequilibrium patterns can affect the reliability of statistical inference. As a result, inconsistent results across different methods do not necessarily indicate analytical failure, but may instead reflect differences in the biological aspects captured by each statistical model. From this perspective, multi-omics statistical integration improves the reliability of

candidate gene assessment and functional interpretation, but does not directly provide a definitive explanation of AD mechanisms.

5.2. Future directions in statistical multi-omics analysis

Future developments in statistical multi-omics analysis will depend not only on the integration of additional data types, but also on improvements in data resolution, statistical inference, and the connection between computational findings and biological validation. With the advancement of single-cell omics technologies, future studies may overcome limitations associated with bulk tissue analysis and characterize AD-related molecular regulation at more refined cellular levels [15]. Spatial omics approaches can complement these analyses by revealing the spatial distribution of molecular alterations within AD-related tissues [16]. These high-resolution datasets can help identify cell-type-specific regulatory effects and clarify relationships between genetic variants, molecular traits, and disease processes.

In addition to improving data resolution, the development of more flexible statistical models will be essential for increasing the reliability of multi-omics integration. Most current analytical methods rely on simplified assumptions, whereas the genetic and molecular regulation of complex diseases often involves multiple causal variants, dynamic molecular changes, and interactions across different biological layers. Future methodological advances should focus on statistical models capable of accommodating complex, heterogeneous, and multi-dimensional biological data. In addition, future analyses may consider broader biological contexts, including gene networks, cellular interactions, and disease progression processes, rather than focusing only on individual genes or proteins when interpreting AD mechanisms [7].

Ultimately, the value of statistical multi-omics analysis lies not in replacing experimental studies, but in providing more reliable hypotheses for biological validation. Future research should establish a continuous workflow connecting genetic association discovery, statistical inference, and functional validation, allowing computational predictions to be further evaluated through cellular models, animal studies, or clinical samples. Through closer integration between statistical approaches and experimental evidence, multi-omics analysis may contribute to a deeper understanding of disease mechanisms and facilitate the identification of potential therapeutic strategies in AD.

6. Conclusion

Recent advances in Alzheimer's disease genetics have gradually shifted the focus from identifying risk-associated loci toward understanding the molecular mechanisms underlying these genetic signals. This review has highlighted how statistical integration of GWAS, eQTL and pQTL data provides a systematic approach for interpreting the functional consequences of genetic variation. Compared with relying solely on genetic associations, multi-omics integration incorporates complementary molecular evidence at the transcript and protein levels, thereby improving the identification of potential functional genes and the interpretation of AD-related biological processes.

Taken together, these analytical approaches have reshaped the interpretation of AD genetic findings. The integration of genetic, transcriptomic, and proteomic evidence enables disease-associated loci to be investigated from multiple molecular perspectives, providing greater confidence in the biological relevance of prioritized genes and pathways.

Statistical integration does not directly resolve all aspects of Alzheimer's disease pathogenesis. The interpretation of integrated results remains influenced by data heterogeneity, the specific biological contexts of different tissues and cell populations, and limitations arising from statistical model assumptions. Conclusions

derived from multi-omics analyses therefore require careful evaluation across different sources of evidence and further biological validation. Overall, the integration of GWAS, eQTL and pQTL data represents an important step toward a more comprehensive understanding of Alzheimer's disease molecular mechanisms. The development of more robust statistical approaches and the increasing availability of molecular datasets will enhance the ability of multi-omics analysis to interpret genetic findings in complex diseases.

References

[1] Reitz, C., Pericak-Vance, M. A., Foroud, T., & Mayeux, R. (2023). A global view of the genetic basis of Alzheimer disease. *Nature Reviews Neurology, 19*(5), 261–277.

[2] Bellenguez, C., Küçükali, F., Jansen, I. E., Kleineidam, L., Moreno-Grau, S., Amin, N., Naj, A. C., Campos-Martin, R., Grenier-Boley, B., Andrade, V., Holmans, P. A., Boland, A., Damotte, V., van der Lee, S. J., Costa, M. R., Kuulasmaa, T., Yang, Q., de Rojas, I., Bis, J. C., … EADB, GR@ACE, & DEGESCO. (2022). New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nature Genetics, 54*(4), 412–436.

[3] Jansen, I. E., Savage, J. E., Watanabe, K., Bryois, J., Williams, D. M., Steinberg, S., Sealock, J., Karlsson, I. K., Hägg, S., Athanasiu, L., Voyle, N., Proitsi, P., Witoelar, A., Stringer, S., Aarsland, D., Almdahl, I. S., Andersen, F., Bergh, S., Bettella, F., … Posthuma, D. (2019). Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nature Genetics, 51*(3), 404–413.

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

[5] Kunkle, B. W., Grenier-Boley, B., Sims, R., Bis, J. C., Damotte, V., Naj, A. C., Boland, A., Vronskaya, M., van der Lee, S. J., Amlie-Wolf, A., Bellenguez, C., Frizatti, A., Chouraki, V., Martin, E. R., Sleegers, K., Badarinarayan, N., Jakobsdottir, J., Hamilton-Nelson, K. L., Moreno-Grau, S., … Goate, A. (2019). Genetic meta-analysis of diagnosed Alzheimer's disease identifies new risk loci and implicates Aβ, tau, immunity and lipid processing. *Nature Genetics, 51*(3), 414–430.

[6] Western, D., Timsina, J., Wang, L., Wang, C., Yang, C., Phillips, B., Wang, Y., Liu, M., Ali, M., Beric, A., Gorijala, P., Kohlfeld, P., Budde, J., Levey, A. I., Morris, J. C., Perrin, R. J., Ruiz, A., Marquié, M., Boada, M., … Cruchaga, C. (2024). Proteogenomic analysis of human cerebrospinal fluid identifies neurologically relevant regulation and implicates causal proteins for Alzheimer's disease. *Nature Genetics, 56*(12), 2672–2684.

[7] Rahimzadeh, N., Srinivasan, S. S., Zhang, J., & Swarup, V. (2024). Gene networks and systems biology in Alzheimer's disease: Insights from multi-omics approaches. *Alzheimer's & Dementia, 20*(9), 3587–3605.

[8] Del Campo, M., Peeters, C. F. W., Johnson, E. C. B., Vermunt, L., Hok-A-Hin, Y. S., van Nee, M., Chen-Plotkin, A., Irwin, D. J., Hu, W. T., Lah, J. J., Seyfried, N. T., Dammer, E. B., Herradon, G., Meeter, L. H., van Swieten, J., Alcolea, D., Lleó, A., Levey, A. I., Lemstra, A. W., … Teunissen, C. E. (2022). CSF proteome profiling across the Alzheimer's disease spectrum reflects the multifactorial nature of the disease and identifies specific biomarker panels. *Nature Aging, 2*(11), 1040–1053.

[9] Luquez, T., Gaur, P., Kosater, I. M., Lam, M., Lee, D. I., Mares, J., Paryani, F., Yadav, A., & Menon, V. (2022). Cell type-specific changes identified by single-cell transcriptomics in Alzheimer's disease. *Genome Medicine, 14*(1), 136.

[10] Giambartolomei, C., Vukcevic, D., Schadt, E. E., Franke, L., Hingorani, A. D., Wallace, C., & Plagnol, V. (2014). Bayesian test for colocalisation between pairs of genetic association studies using summary statistics. *PLoS Genetics, 10*(5), e1004383.

[11] Gusev, A., Ko, A., Shi, H., Bhatia, G., Chung, W., Penninx, B. W. J. H., Jansen, R., de Geus, E. J. C., Boomsma, D. I., Wright, F. A., Sullivan, P. F., Nikkola, E., Alvarez, M., Civelek, M., Lusis, A. J., Lehtimäki,

T., Raitoharju, E., Kähönen, M., Seppälä, I., … Pasaniuc, B. (2016). Integrative approaches for large-scale transcriptome-wide association studies. *Nature Genetics, 48*(3), 245–252.

[12] Sun, Y., Zhu, J., Zhou, D., Canchi, S., Wu, C., Cox, N. J., Rissman, R. A., Gamazon, E. R., & Wu, L. (2021). A transcriptome-wide association study of Alzheimer's disease using prediction models of relevant tissues identifies novel candidate susceptibility genes. *Genome Medicine, 13*(1), 141.

[13] Lee, B., Yao, X., Shen, L., & Alzheimer's Disease Neuroimaging Initiative. (2022). Integrative analysis of summary data from GWAS and eQTL studies implicates genes differentially expressed in Alzheimer's disease. *BMC Genomics, 23*(Suppl 4), 414.

[14] Tijms, B. M., Vromen, E. M., Mjaavatten, O., van der Lee, S. J., Sleegers, K., den Braber, A., Vermunt, L., van Nee, M., Del Campo, M., Teunissen, C. E., Visser, P. J., van der Flier, W. M., & Barkhof, F. (2024). Cerebrospinal fluid proteomics in patients with Alzheimer's disease reveals five molecular subtypes with distinct genetic risk profiles. *Nature Aging, 4*(1), 33–47.

[15] İş, Ö., Min, Y., Wang, X., Oatman, S. R., Daniel, A. A., & Ertekin-Taner, N. (2025). Multi layered omics approaches reveal glia specific alterations in Alzheimer's disease: A systematic review and future prospects. *Glia, 73*(3), 539–573.

[16] Ma, Y., Shi, W., Dong, Y., Sun, Y., & Jin, Q. (2024). Spatial multi-omics in Alzheimer's disease: A multi-dimensional approach to understanding pathology and progression. *Current Issues in Molecular Biology, 46*(5), 4968–4990.