Cell Genomics Transparent Peer Review Record

Single nucleus multiomics identifies ZEB1 and MAFB as candidate regulators of Alzheimer's disease-specific cis regulatory elements

Ashlyn G. Anderson, Brianne B. Rogers, Jacob M. Loupe, Ivan Rodriguez-Nunez, Sydney C. Roberts, Lauren M. White, J. Nicholas Brazell, William E. Bunney, Blynn G. Bunney, Stanley J. Watson, J. Nicholas Cochran, Richard M. Myers, Lindsay F. Rizzardi

Summary

Initial submission:	Received : October 5th 2022
	Scientific editor: Judith Nicholson
First round of review:	Number of reviewers: 2 Revision invited: November 4th 2022 Revision received : December 6th 2022
Second round of review:	Number of reviewers: 1 Accepted : 12th January 2023
Data freely available:	Yes
Code freely available:	Yes

Referees' reports, first round of review

Reviewer 1

-The work is very well presented and, to my knowledge, there is not other descriptive work of the kind for AD. The text does become too descriptive at times without a lot of emphasis on the biological meaning of the findings. The authors should consider capitalizing (i.e. further the understanding) on both main findings on their understanding of altered gene regulation in AD, namely, the identification of TFs were particularly involved in regulating AD-specific transcriptional programs and the identification of enhancer-like activity for 12 candidate CREs linked to neurodegeneration-related genes.

-The assay used for this work is also unique and this reviewer agrees that the profiling gene expression + chromatin accessibility simultaneously from the nuclei allows for greater confidence in the correlations linking potential cis-regulatory elements to target genes.

-The main limitations have been discussed by the authors.

Reviewer 2

Anderson and collaborators present a single-nucleus multiomic analysis of cortical tissue in Alzheimer's disease (AD) aiming at identifying candidate cis-regulatory elements (CREs) involved in AD-associated transcriptional changes. The authors profiled snRNAseq + snATACseq simultaneously in individual nuclei isolated from cortical tissues of AD (n=7) and unaffected (n=8) donors. The authors report 319,861 significant correlations between gene expression and cell-type specific accessible regions -- 40,831 unique to AD tissues. In vitro experimental validation supports some of the candidate regulators and regulatory links. Using correlation analyses of TFs, CREs, and links; the authors report ZEB1 and MAFB as candidate regulators playing important roles in AD-specific gene regulation in neurons and microglia.

Developing a deeper understanding of potential genomic regulatory mechanisms underlying AD pathogenesis is an important and timely problem in the field of neurodegeneration. The data generated in this study is novel and of good quality, as supported by convincing cell type annotations consistent with previous single-cell analyses of AD. However, I have conceptual and technical concerns regarding

the soundness of the presented analyses, results, and conclusions. The results presented require additional statistical and biological support. As currently presented, the results do not seem to address the original motivation of identifying regulatory mechanisms responsible for the cell type specific molecular alterations observed in AD. More clarity and additional biological interpretation complementing data description would largely benefit the study.

See specific comments below:

Authors should provide a more comprehensive description of subject selection criteria. Were AD subjects defined solely based on the Braak stage? If so, what was the motivation of focusing on the Braak stage only? Do you have any information regarding amyloid (Abeta-plaque load) and/or dementia (cognitive decline) status of the AD subjects? This information is relevant to interpret the molecular changes, for example the potential role of APP dysregulation briefly discussed.

Authors should expand on the biological interpretation of cell type-specific transcriptome changes identified. How do they fit with known AD molecular neuropathology? How do they differ?

The authors identified 189,925 reproducible peaks across cell types (after some filtering) and performed feature linkage analysis using those peaks. The authors include some comparisons with existing data for peaks identified as being linked to genes. Additional analyses validating the biological relevance of the identified peaks and their cell type specificity, prior to analyzing peak-gene linkages, would further strengthen your data. Do existing epigenomic and transcriptomic data support the cell type specificity and physiological relevance of identified peaks?

Because the statistical properties of single-nucleus data are very different from those of bulk data (even more so in snATAC-seq), assumptions of conventional correlation measures (e.g., Pearson or Spearman correlation) are violated and therefore do not work properly. Other than citing cellranger-arc (v2.0), the authors do not discuss any of these technical issues on the methods. Without this information it is difficult to evaluate the validity of the links and/or interpret the reported numbers. How exactly are correlations determined in the cellranger-arc

(v2.0) analysis pipeline? How does it account for data sparseness? Does it account for cell to individual membership? Are positive correlations more likely to be detected than negative? Is the observation that "the majority (76.11%) of linked peaks were positively correlated with gene expression" really supporting an association between open chromatin and transcriptional activation or is it a byproduct of how correlations are computed using sparse data?

Feature linkage analysis to identify peak-gene correlations was performed independently for AD and control cells. What is the rationale behind identifying AD-specific, control-specific, or common correlation-based peak-gene linkages? What would it mean for Alzheimer's disease pathogenesis and progression whether a putative cis-regulatory element (CRE) active in a given cell type correlates with a gene only in AD samples, only in control, or in both? Do you have any example of this "differential regulation" that can be understood on the basis of molecular neuropathological processes already known to be involved in AD? What is the logic of the underlying regulatory mechanisms being analyzed and how does it relate to analyses design?

The authors discuss whether links are unique to one cell-type or shared across cell types. However, it is not clear from the description how cell type specificity is addressed. According to the authors, "cell type specificity of each link was determined by the cell type(s) in which the ATAC peak was identified". Were feature linkages calculated across all cells irrespective of cell type and then annotated by cell type based only where the ATAC peak was detected? Was cell type specific gene expression also considered? If the inference was performed across all cells, how can the authors be sure that correlations are not largely driven by expression and epigenomic differences across cell types? Please clarify.

Related to the previous comment, the authors report a total of 319,905 links. Link estimations were performed independently in AD and control cells. This means that any correlation was estimated based on expression and accessibility patterns across either only 7 (AD) or 8 (control) subjects. Given such small sample sizes, such a large number of "significant" correlations is questionable. I understand that correlations might be computed across all cells, thus ending up with a large sample size of observations. However, the phenotype being analyzed and discussed corresponds to the diagnostic group of the actual individuals. Since the authors present the identification of peak-gene correlations independently in

control and AD datasets as something unique and one of the main contributions of the study, more care should be taken in verifying the veracity of the inferences and their AD-specificity given such a small sample size. Are these correlations supported by the data when accounting for cell-individual membership using, for example, pseudo bulk or hierarchical mixture models?

If correlations are computed across all AD cells, regardless of subject membership, are the resulting peak-gene correlations reproducible across AD subjects and not detectable across cells of individual control subjects?

The study would also benefit from deeper analyses of the relationship between cell type specific peaks and cell type specific genes? Are genes preferentially expressed in a given cell type having links with peaks that are also preferentially active in the same cell type? Is it more likely to identify links for cell type specific genes than for more broadly expressed genes?

Genes with more links (>40 or more linked peaks) were longer and more highly expressed than those with fewer links (Figure S3C), raising the possibility that whether a link is detected or not might be heavily influenced by technical aspects of how the correlations are being computed and the extent to which peak and gene activity can be captured and measured a single-nucleus resolution. Thus, the study would benefit from additional, integrative analyses using existing epigenomic and expression data from brain cells that support the physiological relevance of the identified links for the different cell types. Do links identified as cell type specific involve molecular processes of relevance for the physiology of the cell type in question?

A motivation for this study is to "rigorously interrogate the regulatory mechanisms responsible for these alterations" (i.e., cell type-specific transcriptional differences). However, despite the attempt to provide an example with the gene KANSL1, it is not clear to me how DEGs relate to peak-gene links, and how the two analyses (cell type-specific transcriptome changes and peak-gene correlations) fit together in the study. How do the links explain the differential expression? What are the regulatory mechanisms responsible for the observed alterations?

Similar to the situation with peak-gene links, the authors discuss whether peak-

gene-TF trios are cell type specific or not. However, it is not completely clear how specificity is defined in this case. Is it by peak activity, gene activity, TF activity or a combination? Or is it by computing correlations across only cells of a given type? According to methods, correlations for these analyses were computed on average counts within metacells using Pearson's coefficient. Were cells averaged by individually defining pseudo bulk profiles per cell type? Why were these correlations computed in a different way than those used to define links? What approach is more reliable given the data and goal?

Similar to the question above, what is the rationale behind peak-gene-TF trios and their relevance for disease? Do the authors have a conceptual model for how peak-gene-TF trios might affect disease? What could we learn based on this type of analysis about what might be happening in a specific AD subject?

What does it mean that links were significantly enriched for heritability of AD? How were links annotated to perform (sLDSC) regression? Were both the peak and gene of the link used for genome annotation and therefore the union of all peaks and genes in a link set used as one single annotation across the genome? If so, what would be the difference with performing sLDSC only with cell type specific enhancers (CREs) or only with cell type specific genes as performed in previous studies? Does a link provide any additional information?
What does AD-specific CREs mean? Similarly, please clarify how AD-specific links and their cell type specificity were defined for sLDSC regression analyses.

Authors' response to the first round of review

We appreciate the reviewer's thoughtful consideration of our manuscript (CELL-GENOMICS-D-22-00197) and thank the editors for the opportunity to respond as the manuscript is greatly improved. We have revised our manuscript in accordance with reviewer recommendations and included two new figure panels and several additional analyses in addition to changes to the text. In improving clarity throughout the manuscript, we removed Figure S4B which showed a separate analysis of excitatory-specific trios and instead included them in the neuron-specific trio analysis in Figure 4F as the distinction was confusing and the results were similar. We also added a GO analysis of cell type-specific links as requested by the reviewer as Figure S3E. We performed a permutation analysis to address the validity of our AD- and control-specific links and included this as Figure S3D. The changes to the main text are indicated by track changes in the

revised version of the manuscript and are described in detail below.

Reviewer #1:

-The work is very well presented and, to my knowledge, there is not other descriptive work of the kind for AD. The text does become too descriptive at times without a lot of emphasis on the biological meaning of the findings. The authors should consider capitalizing (i.e. further the understanding) on both main findings on their understanding of altered gene regulation in AD, namely, the identification of TFs were particularly involved in regulating AD-specific transcriptional programs and the identification of enhancer-like activity for 12 candidate CREs linked to neurodegeneration-related genes.

-The assay used for this work is also unique and this reviewer agrees that the profiling gene expression + chromatin accessibility simultaneously from the nuclei allows for greater confidence in the correlations linking potential cisregulatory elements to target genes.

-The main limitations have been discussed by the authors.

We thank Reviewer 1 for their positive comments and address their recommendation to elaborate on our findings both in response to Reviewer 2 (included in the Results) and in the Discussion section. Specifically, we point out additional DEGs that indicate disrupted calcium homeostasis in AD which has been proposed as a potential disease mechanism. We also add more information on the AD-associated target genes for which we validated CREs.

Lines 454-472: “We identified many DEGs associated with calcium homeostasis consistent with the calcium hypothesis of AD which postulates that a synergistic relationship between Aβ accumulation and Ca2+ levels promotes neurodegeneration 74. In AD neurons, we found decreased expression of ryanodine receptor 3 (RYR3) and inositol 1,4,5-trisphosphate receptor type 2 (ITPR2) that both release internal stores of Ca2+ from the endoplasmic reticulum 75. We also measured decreased expression of the Ca2+ sensors calmodulin (CALM1, CALM2, CALM3) and VILIP-1 (VSNL1), the latter which is associated with neuropathologic lesions 49,76. In contrast, two genes encoding calcium channel subunits (CACNA1C and CACNA1B) were upregulated in AD neurons. In addition, astrocytes also demonstrated decreased expression of calneuron 1 (CALN1, a Ca2+ sensor similar to calmodulin), glutamate receptor 2 subunit (GRIA2, limits Ca2+ permeability of AMPA receptors), and glutamate receptor NMDA 2C (GRIN2C, a subunit of the NMDA receptors). Both AD neurons and astrocytes showed decreased expression of the glutamate transporter GLT-1 (SLC1A2). We

identified AD-specific links for all these genes except GRIN2C. Altered expression of these calciumassociated proteins is likely to exhibit complex and cell type-specific effects making the resulting network effect on excitability uncertain. However, one possibility is that this altered expression could lead to increased sensitivity of neurons to glutamate and thus neurotoxicity 75,77. Further study of the candidate regulatory elements we identified for these genes would improve our understanding of how these genes become dysregulated in AD and the emergent resulting effects."

Lines 445-452: "Amyloid precursor protein (APP) is the precursor to the AD hallmark pathology Ab, and, while characteristic of lewy body diseases, a-synuclein (SNCA) aggregates are highly prevalent in AD postmortem brains as well 82 . PHF24 is a modulator of GABAB receptor activity 83 and was recently identified in a study of AD resilience genes 84. ADAMTS1 has been implicated in AD both biochemically 85 and genetically 6 . Our study lays the groundwork for additional functional validation in future studies to confirm these genes as targets of these CREs. Understanding how these genes are regulated and by which TFs could provide new therapeutic targets. In fact, a recent study 86 identified TFs contributing to disruption of gene regulatory networks in AD, demonstrated their ability to predict AD cognitive phenotypes, and used them to prioritize candidate drugs that could be repurposed for AD."

Reviewer #2:

Anderson and collaborators present a single-nucleus multiomic analysis of cortical tissue in Alzheimer's disease (AD) aiming at identifying candidate cis-regulatory elements (CREs) involved in AD-associated transcriptional changes. The authors profiled snRNAseq + snATACseq simultaneously in individual nuclei isolated from cortical tissues of AD (n=7) and unaffected (n=8) donors. The authors report 319,861 significant correlations between gene expression and cell-type specific accessible regions -- 40,831 unique to AD tissues. In vitro experimental validation supports some of the candidate regulators and regulatory links. Using correlation analyses of TFs, CREs, and links; Response to Reviewers the authors report ZEB1 and MAFB as candidate regulators playing important roles in AD-specific gene regulation in neurons and microglia.

Developing a deeper understanding of potential genomic regulatory mechanisms underlying AD pathogenesis is an important and timely problem in the field of neurodegeneration. The data generated in this study is novel and of good quality, as supported by convincing cell type annotations consistent with previous single-cell analyses of AD. However, I have conceptual and technical concerns regarding

the soundness of the presented analyses, results, and conclusions. The results presented require additional statistical and biological support. As currently presented, the results do not seem to address the original motivation of identifying regulatory mechanisms responsible for the cell type specific molecular alterations observed in AD. More clarity and additional biological interpretation complementing data description would largely benefit the study.

See specific comments below:

Authors should provide a more comprehensive description of subject selection criteria. Were AD subjects defined solely based on the Braak stage? If so, what was the motivation of focusing on the Braak stage only? Do you have any information regarding amyloid (Abeta-plaque load) and/or dementia (cognitive decline) status of the AD subjects? This information is relevant to interpret the molecular changes, for example the potential role of APP dysregulation briefly discussed.

By focusing on later Braak stages (IV-VI), we sought to identify general regulatory changes associated with AD. Aside from Braak staging and CERAD criteria, we were not provided with quantitative information on plaque load, though we do note that we are confident in the neuropathological diagnosis given it is from a reliable source, the NIH Neurobiobank. We have included additional information about AD selection criteria in our Methods: Lines 648-651: “AD donors were neuropathologically diagnosed according to CERAD criteria and Braak staging. All AD donors had a clinical diagnosis of AD and evidence of both amyloid beta plaques and neurofibrillary tangles.”

Authors should expand on the biological interpretation of cell type-specific transcriptome changes identified. How do they fit with known AD molecular neuropathology? How do they differ?

We agree that this is a particularly important point as it is always critical to consider where studies exhibit replication. We have highlighted where our results agree with other single cell RNA-seq datasets including comparison to a recent meta-analysis. We have also provided additional information on the DEGs we provide as examples, including MDGA2 which is associated with Braak stage, CERAD score, and cognition (https://agora.adknowledgeportal.org). Further, we elaborate on many DEGs involved in calcium homeostasis and how dysregulation is associated with AD. See section “Cell type-specific transcriptome changes in Alzheimer’s DLPFC” Lines 110-128 Also, see Lines 454-472 included above in response to Reviewer 1.

The authors identified 189,925 reproducible peaks across cell types (after some filtering) and performed feature linkage analysis using those peaks. The authors include some comparisons with existing data for peaks identified as being linked to genes. Additional analyses validating the biological relevance of the identified peaks and their cell type specificity, prior to analyzing peak-gene linkages, would further strengthen your data. Do existing epigenomic and transcriptomic data support the cell type specificity and physiological relevance of identified peaks?

The overlap of all ATAC peaks with H3K27ac from corresponding cell types and ENCODE CREs was similar to that of the linked ATAC peaks. We have included these findings in the results. Lines 149-150: "Nearly half of all peaks overlapped H3K27ac (46%) from the corresponding cell type and 43% overlapped ENCODE distal enhancer-like sequences."

Because the statistical properties of single-nucleus data are very different from those of bulk data (even more so in snATAC-seq), assumptions of conventional correlation measures (e.g., Pearson or Spearman correlation) are violated and therefore do not work properly. Other than citing cellranger-arc (v2.0), the authors do not discuss any of these technical issues on the methods. Without this information it is difficult to evaluate the validity of the links and/or interpret the reported numbers. How exactly are correlations determined in the cellranger-arc (v2.0) analysis pipeline? How does it account for data sparseness? Does it account for cell to individual membership?

The code for the cellranger-arc (v2.0) pipeline is publicly available from 10X Genomics and there is a publication describing the Hotspot algorithm in detail (DeTomaso 2021, PMID: 33951459). Information borrowing across similar cells allows for increased sensitivity to overcome sparsity in the data. These smoothed values are then used in the correlation calculations. We have included a more detailed description of the feature linkage calculation in the methods. Individual is not included for the feature linkage scoring. Lines 788-796: "For feature linkage calculation, ATAC and gene expression counts were normalized independently using depth-adaptive negative binomial normalization. To account for sparsity in the data, the normalized counts were smoothed by taking the weighted sum of the 30 closest neighbors from the KNN graph. The cell weights are determined by using a Gaussian kernel transformation of the euclidean distance. Feature linkage scores were calculated by taking the Pearson correlation between the smoothed counts, while the significance of the correlation was determined using the Hotspot algorithm."

Are positive correlations more likely to be detected than negative? Is the

observation that "the majority (76.11%) of linked peaks were positively correlated with gene expression" really supporting an association between open chromatin and transcriptional activation or is it a byproduct of how correlations are computed using sparse data?

The set of significant feature linkages includes links between a promoter and its gene as well as links to peaks within the gene body. The positive correlation between chromatin accessibility and gene expression in these regions is well supported in the literature (PMIDs: 30795793, 25503965, 28077088, 25103404, 35614386), including from our own previous work (Rizzardi 2019, PMID: 30643296) and added these references to the manuscript. Accessible peaks within the promoter and gene body of the linked-gene(s) account for 38% of the positively correlated links. We require that a gene have at least 200 UMIs across the entire dataset in order to be evaluated in this analysis such that many cell types will have little to no expression of many genes. Therefore, we have not biased our calculations in any way by requiring expression across all cell types or in a large percentage of cells. Lines 207-210: ”The majority (76.11%) of linked peaks were positively correlated with gene expression, as is expected given the association between open chromatin and transcriptional activation 37–41, though negative correlations may be indicative of repressor binding 37,40.”

Feature linkage analysis to identify peak-gene correlations was performed independently for AD and control cells. What is the rationale behind identifying AD-specific, control-specific, or common correlation-based peak-gene linkages? What would it mean for Alzheimer's disease pathogenesis and progression whether a putative cisregulatory element (CRE) active in a given cell type correlates with a gene only in AD samples, only in control, or in both? Do you have any example of this "differential regulation" that can be understood on the basis of molecular neuropathological processes already known to be involved in AD?

We hypothesized that gene regulatory programs would be disrupted in AD as gene expression changes have been readily detected in both bulk and single cell analyses. We performed our linkage analysis in each condition to identify which potential regulatory elements were uniquely utilized in AD and which TFs might be responsible for their activity. One important orthogonally validated example is the BIN1 enhancer identified in microglia from unaffected donor brain tissue in Nott et al 2019 in which deletion of this region reduced BIN1 expression. In primary mouse microglia, loss or reduction of BIN1 expression impaired inflammatory response to LPS (Sudwarts et al 2022, PMID: 35526014). We

identified the same region as a controlspecific microglial link in our dataset and saw reduced BIN1 expression in AD microglia. Together these findings suggest that this regulatory region may no longer be utilized in AD microglia resulting in the decreased expression observed. Additionally, unique peak-gene associations gained or lost in AD could harbor genetic variants that contribute to disease risk such as rs733839 in the microglial BIN1 enhancer. Indeed in microglia, we did find enrichment of AD-associated SNPs in AD-specific links. These results provide several new avenues of inquiry into both specific loci and TFs (through our peak-gene-TF trio analyses) that contribute to altered gene regulation contributing or responding to the AD disease state. Further investigations are needed to determine if any particular SNP drives the AD-specific associations. We have made this point and provided this additional example in the results section: Lines 153-156: "Given the gene expression changes observed in AD, we hypothesized that there would be differential usage of CREs between AD and control samples that would be identified in this analysis as ADor control-specific links." Lines 218-225: "For example, BIN1 expression is significantly reduced in AD microglia compared to controls and this reduced expression hampers proinflammatory microglial responses. We identified six control- specific links and no AD-specific links for BIN1 in microglia. One of these control-specific links was validated as a microglia-specific BIN1 enhancer in Nott et al. and harbors an AD-associated SNP (rs733839). Together, these findings suggest that this CRE may no longer be utilized in AD microglia leading to lower BIN1 expression, though it remains possible that these observations could also result from less sensitive detection with lower expression."

What is the logic of the underlying regulatory mechanisms being analyzed and how does it relate to analyses design?

The underlying logic for this project is to link non-coding, AD-associated variants to their target genes in the affected cell types. By doing so, we can better understand how these genetic variants promote disease onset and/or progression. Further, we identify new regulatory regions for known AD-associated genes that could be amenable to intervention through genome editing or manipulation of TFs that bind the region. For this reason, we identified disease- and control-specific links and the TFs most likely to bind these regions. The common linkages identify CREs that could be important for more general cell type-specific functions of interest for many aspects of neurobiology. While this study is focused on identifying CREs, there are several other non-coding regulatory mechanisms that could also be involved in regulating gene expression

including miRNAs (PMID: 30123182), lncRNAs (PMID: 33353982), transposable elements (PMID:35228718), etc. that this study does not address and we now mention these in the text. Lines 520-522: “While this study is focused on identifying CREs, there are other non-coding regulatory mechanisms that could alter gene expression in AD including miRNAs 90, lncRNAs 92, transposable elements 94, etc. that are not assessed.”

The authors discuss whether links are unique to one cell-type or shared across cell types. However, it is not clear from the description how cell type specificity is addressed. According to the authors, "cell type specificity of each link was determined by the cell type(s) in which the ATAC peak was identified". Were feature linkages calculated across all cells irrespective of cell type and then annotated by cell type based only where the ATAC peak was detected?

We have added text throughout the manuscript to clarify that cell type specificity of links and trios are based on the cell type in which the associated ATAC peak was identified. By calculating feature linkages across all cell types, we increase the dynamic range of both expression and accessibility to better establish peak-gene correlations. We explicitly refer to linked-peaks when we discuss the region of accessibility in the context of links or trios. Lines 139-143:”...identify cell type- and disease-specific CREs and their target genes by correlating gene expression with chromatin accessibility across all nuclei in the dataset. …A feature linkage, or link, is defined as a significant correlation between accessibility of an ATAC peak and the expression of a gene.” Line 153: “We consider the linked-peaks to be candidate CREs.” Lines 265-266: “Cell type specificity was defined based on the cell type in which the linked-peak was identified.“

Was cell type specific gene expression also considered?

All genes with at least 200 UMIs in at least one cell type were considered in the linkage analysis. Lines 798-799: “Links with an absolute correlation score < 0.2 and linked to a gene with < 200 UMIs were removed.”

If the inference was performed across all cells, how can the authors be sure that correlations are not largely driven by expression and epigenomic differences across cell types? Please clarify.

The largest source of variance is indeed across cell types, as has been shown in multiple single cell studies. This analysis is designed to identify cis regulatory elements rather than explain a portion of the expression/accessibility variance by disease status. Therefore, correlations are largely driven by cell type as the variation in expression and accessibility across cell types provides the dynamic

range needed to identify the correlations. The majority of our links are shared between AD and control and are likely important for cell type-specific regulation of many genes with cell type-specific functions. Many of these linked-peaks harbor GWAS SNPs associated with a variety of neurological diseases even though they aren't found in diseasespecific links (Fig5A). We think these are important to identify. Performing our analyses in a single cell type greatly limits our power of detection. This is exemplified by a new analysis (only described in this response) in which we recall links within microglia using AD and control samples together. We identified only 73 links, 35 of which were also identified as microglial links in our original analysis. In addition, 94% of DEGs between AD and control within each cell type also have a link within the same cell type. For genes upregulated in AD, 72% of their positively-correlated links were AD-specific, while for downregulated genes 62% were control-specific. These results support the hypothesis that these linkedpeaks could contribute to the differential expression observed in AD and identification of these regions was a major goal of this study. This information has been added to text both in response to this comment and a comment below. Lines 211-216: "Nearly all (94%) the DEGs identified between AD and control nuclei had a linked peak in the same cell type where the gene was differentially expressed and 85% of these linked peaks overlapped H3K27ac in the same cell type. In addition, we observed that some CREs of differentially expressed genes were uniquely identified in either AD or control datasets. For genes upregulated in AD, 72% of their positively correlated links were AD-specific, while for downregulated genes 62% were control-specific. "

Related to the previous comment, the authors report a total of 319,905 links. Link estimations were performed independently in AD and control cells. This means that any correlation was estimated based on expression and accessibility patterns across either only 7 (AD) or 8 (control) subjects. Given such small sample sizes, such a large number of "significant" correlations is questionable. I understand that correlations might be computed across all cells, thus ending up with a large sample size of observations. However, the phenotype being analyzed and discussed corresponds to the diagnostic group of the actual individuals. Since the authors present the identification of peakgene correlations independently in control and AD datasets as something unique and one of the main contributions of the study, more care should be taken in verifying the veracity of the inferences and their AD-specificity given such a small sample size. Are these correlations supported by the data when accounting for cell-individual membership using, for example, pseudo bulk or hierarchical mixture models? If correlations are

computed across all AD cells, regardless of subject membership, are the resulting peak-gene correlations reproducible across AD subjects and not detectable across cells of individual control subjects?

The reviewer raises an important point that we have addressed through permutation analyses. We performed 100 sample permutations calling links to evaluate the accuracy of AD and control specific links. We chose 100 as these permutation calculations are computationally intensive. For each permutation, 7 or 8 individuals were randomly selected regardless of disease status and links were calculated with the same parameters as the true data. We overlapped permutation pairs to determine the proportion of links that were specific to either group of individuals. The average proportion of group-specific links for permutation pairs was 0.25. This represents the proportion of links that we expect to be specific to any 2 groups given our sample size. In the true data, AD and control-specific links made up 36% of the total links. This proportion was a significant outlier for group-specificity compared to the permutations (Z-test; p-value = 0.027), suggesting that these links are partially driven by phenotype. We have included this distribution as a supplementary figure (Figure S3D) and added a line to the text describing this result We also evaluated the effect of cell-individual membership by calling links on counts that were pseudobulked by individual and cell type. A total of 16,421 links were identified, of which 9,853 (60%) were found in the original links. Additionally, of the pseudobulk links that overlap our original AD-specific links, 63% of them were also called as AD-specific in the pseudobulk analysis, showing that our approach has high concordance with methods that account for cell-individual membership. We find that a pseudobulk link analysis is not well suited for our study given its small sample size, and that permutation testing better addresses the question of a cell-individual membership so we did not include this analysis in the manuscript. Pseudobulking with a small sample size decreases the dynamic range and leads to decreased sensitivity and fewer links called. However, this approach could be more appropriate for studies with a larger sample size. Lines 190-192: “We performed permutation analyses and determined that this fraction of AD/control-specific links (0.36 of total links) was greater than expected by chance (Z-test, p-value=0.027; Figure S3D” Methods section “Permutation Testing” Lines: 801-809

The study would also benefit from deeper analyses of the relationship between cell type specific peaks and cell type specific genes? Are genes preferentially expressed in a given cell type having links with peaks that are also preferentially active in the same cell type?

The majority of cell type specific genes are linked to cell type specific peaks. However, a third of the cell type specific genes are linked to peaks specific to a different cell type and these links are enriched for negative correlations, suggesting there may be active repression at these regions in other cell types.

Is it more likely to identify links for cell type specific genes than for more broadly expressed genes?

Cell type-specific genes do have a higher average number of links. This is expected as there is a larger dynamic range of expression over which correlations can be calculated. In addition, previous work has shown that housekeeping genes (with similar expression levels across cell types) are less dependent on enhancers to regulate their expression (Bergman 2022, PMID: 35594906); thus, having more links in cell type specific genes is consistent with this finding and expected.

Genes with more links (>40 or more linked peaks) were longer and more highly expressed than those with fewer links (Figure S3C), raising the possibility that whether a link is detected or not might be heavily influenced by technical aspects of how the correlations are being computed and the extent to which peak and gene activity can be captured and measured a single-nucleus resolution.

Longer genes most likely have more links because of the larger number of peaks called within the gene body. If we exclude links within a target gene's own gene body, there is not a significant difference in the number of links for longer genes (t-test; p-value = 0.118). Across the entire dataset, 17.8% of linked peaks are present in the promoter or gene body of the target gene. We include these peaks in our analyses as enhancers are often located within the introns of their target genes. We have added this information to the text. Lines 177-183: "This finding is likely due to links being called for peaks within the gene body of longer genes as excluding these peaks abolishes the difference in number of links (t-test, p = 0.12). Across the entire dataset, 17.8% of linked peaks are present in the promoter or gene body of the target gene. While positively correlated links in gene bodies may often be a merely consequence of target gene expression, we retained these peaks in our analyses as enhancers are often located within the introns of their target genes."

Thus, the study would benefit from additional, integrative analyses using existing epigenomic and expression data from brain cells that support the physiological relevance of the identified links for the different cell types. Do links identified as cell type specific involve molecular processes of relevance for the physiology of the cell type in question?

We have integrated our data with previously published functional genomic datasets that include eQTLs, MPRAs, and HiC in brain tissues and/or cell types (Fig 5B). We also intersect our data with H3K27ac data generated from neuronal and glial cell types isolated from human DLPFC (from Nott et al 2019 and Kozlenkov et al 2018) (Fig 3E). We have also performed a new GO analysis of the target genes of cell typespecific links identified in both AD and control samples and added these results as Figure S3E along with a brief description in the results. We find enrichment of myelination pathways in oligodendrocytes, oligodendrocyte differentiation in OPCs, neutrophil-associated processes in microglia, nervous system development in astrocytes, and synaptic and ion channel categories in neurons. Lines 194-195: "Target genes of cell type-specific links identified in both AD and control samples were enriched in expected pathways (Figure S3E)."

A motivation for this study is to "rigorously interrogate the regulatory mechanisms responsible for these alterations" (i.e., cell type-specific transcriptional differences). However, despite the attempt to provide an example with the gene KANSL1, it is not clear to me how DEGs relate to peak-gene links, and how the two analyses (cell type-specific transcriptome changes and peak-gene correlations) fit together in the study. How do the links explain the differential expression? What are the regulatory mechanisms responsible for the observed alterations?

Thank you for pointing out this omission. We have included this additional information in the results section. We hypothesized that CREs most likely to contribute to altered gene expression in a particular cell type would be uniquely identified in that cell type within either AD or control datasets. Lines 213-250: "...Nearly all (94%) the DEGs identified between AD and control nuclei had a linked-peak in the same cell type where the gene was differentially expressed and 85% of these linked-peaks overlapped H3K27ac in the same cell type. In addition, we observed that some links to differentially expressed genes were uniquely identified in either AD or control datasets. For genes upregulated in AD, 72% of their positively correlated links were AD-specific, while for downregulated genes 62% were control-specific. ……These results support the hypothesis that these regions could contribute to the differential expression observed in AD."

Similar to the situation with peak-gene links, the authors discuss whether peak-gene-TF trios are cell type specific or not. However, it is not completely clear how specificity is defined in this case. Is it by peak activity, gene activity, TF activity or a combination?

We have made numerous edits throughout the manuscript to clarify how cell type

specificity is defined for each of our analyses. This definition is based on the cell type that the linked-peak was initially identified in. Whether a link or trio is categorized as AD-specific, control-specific or common is based on which dataset the correlation was found to be significant in. We have clarified this in the results. Lines 261-266: "To identify trios, we performed these additional correlation analyses (linked-peak : TF expression and TF expression : target gene expression) separately using either AD or control data sets to enable identification of TFs whose activities may be associated with disease. AD- or control-specific trios were those uniquely identified in the AD or control dataset, respectively. Cell type specificity was defined based on the cell type in which the linked-peak was identified."

Or is it by computing correlations across only cells of a given type? According to methods, correlations for these analyses were computed on average counts within metacells using Pearson's coefficient. Were cells averaged by individually defining pseudo bulk profiles per cell type?

The pseudobulk profiles were defined by WNN clusters to create metacells. Metacells were not restricted to a cell type and were determined solely on clustering. However, there weren't any metacells that contained cells from multiple cell types. Metacells were used to address the sparsity of the data as an alternative to smoothing and to decrease computation time given the large number of TF motifs to analyze. However, note that the peak-gene links were not recalculated.

Why were these correlations computed in a different way than those used to define links? What approach is more reliable given the data and goal?

Both analyses determined associations using a Pearson correlation. They differed primarily on how significance was assigned. The significance for links using the Hotspot algorithm is more stringent as genepeak links with a -log10(q-value) >= 5 were classified as significant. The significance for TF associations is less stringent because TF inclusion in trios is first defined by motif presence in the linked-peak thus increasing the prior probability of correlation. TF expression had to be significantly and positively correlated with the linked-peak and significantly correlated with the linked-gene. Significant TF correlations with both the linked-peak and linked-gene were defined as those with Pearson correlation -log10(p-value) > 3. As trios assess a three-way regulatory relationship, we therefore accept a lower threshold for TF associations given that all 3 directions must be significant for a trio to be called and the likelihood of spurious correlations for all 3 directions is small. This allows us to identify trios where TF expression may only

be weakly correlated with linked-gene expression.

Similar to the question above, what is the rationale behind peak-gene-TF trios and their relevance for disease? Do the authors have a conceptual model for how peak-gene-TF trios might affect disease? What could we learn based on this type of analysis about what might be happening in a specific AD subject?

Many AD GWAS variants are located in noncoding regions and may disrupt or create TF binding motifs. Identification of trios allows us to link a CRE (that may harbor a GWAS SNP) with a target gene and TF that may regulate target gene expression. Should a SNP disrupt that TFs motif, we then have a testable hypothesis as to why that SNP is associated with disease risk. Further, a recent computational study (PMID:35849618) took a similar conceptual approach using publicly available snRNA-seq and chromatin data to identify gene regulatory networks rewired in AD and the TFs that contributed most to these changes in regulation. They went on to show the predictive value of the top TFs to predict AD cognitive phenotypes and to prioritize cell type candidate drugs that could be repurposed for AD. We now elaborate on this point in the text. Lines 493-498: "Our study lays the groundwork for additional functional validation in future studies to confirm these genes as targets of these CREs. Understanding how these genes are regulated and by which TFs could provide new therapeutic targets. In fact, a recent study identified TFs contributing to disruption of gene regulatory networks in AD, demonstrated their ability to predict AD cognitive phenotypes, and used them to prioritize candidate drugs that could be repurposed for AD."

What does it mean that links were significantly enriched for heritability of AD? How were links annotated to perform (sLDSC) regression? Were both the peak and gene of the link used for genome annotation and therefore the union of all peaks and genes in a link set used as one single annotation across the genome? If so, what would be the difference with performing sLDSC only with cell type specific enhancers (CREs) or only with cell type specific genes as performed in previous studies? Does a link provide any additional information? What does AD-specific CREs mean? Similarly, please clarify how AD-specific links and their cell type specificity were defined for sLDSC regression analyses.

We have clarified link annotations throughout the manuscript and now refer to linked-peaks rather than links in reference to this analysis. We also now define a candidate CRE early in the manuscript, but did replace "AD-specific CREs" with "AD-specific linked-peaks" in this section for clarity. We added a detailed description of the link categories in both the results and methods sections. We note that a peak can have multiple links that fall into multiple categories.

Importantly, we point out that within each cell type, less than a third of the peaks with AD-specific links also have a control-specific link. This emphasizes the specificity of the enrichments we observe for GWAS traits. These results are similar to what has been reported for cell type-specific CREs, but we can further "partition" this signal into disease-associated CREs. Line 153: "We consider the linked-peaks to be candidate CREs." Lines 339-353: "Link categories are defined as "AD" or "Control" if the links were only identified in the analysis of AD or control samples, respectively. "Common" links were identified in both analyses, and "All" is the union of all linked-peaks. While a peak with multiple links can be duplicated across categories, less than a third of peaks with AD-specific links also have a control-specific link emphasizing the specificity of these linked-peaks. Within each link category for each cell type, the union of linked-peaks was used for this analysis. Cell type was assigned based on the cell type(s) in which the linked-peak was identified." Lines 838-843 (Methods): "Each category (all, common, AD, control) corresponds to the analysis in which the peak-gene link was identified. Cell type is assigned based on the cell type(s) in which the linked-peak was identified. Peaks were resized to 1 kb and each set of unique peaks with these categories was tested individually along with the full baseline model (baseline-LD model v2.2.) that included 97 categories capturing a broad set of genomic annotations. Note that a peak can have multiple links that fall in different categories. "

Referees' report, second round of review

Reviewer 2:
The authors have addressed all my concerns and improved the manuscript.

Authors' response to the second round of review

n/a