CHAS, a deconvolution tool, infers cell type-specific signatures in bulk brain histone acetylation studies of brain disorders

Kitty B. Murphy[1,2], Alexi Nott[1,2], Sarah J. Marzi[1,2*]

[1] UK Dementia Research Institute, Imperial College London, London, UK
[2] Department of Brain Sciences, Imperial College London, London, UK

*** Correspondence should be addressed to Sarah J. Marzi (s.marzi@imperial.ac.uk)**

Abstract

Chromatin profiling studies have shown the importance of gene regulation in driving heritability and environmental risk of brain disorders. Acetylation of histone H3 lysine 27 (H3K27ac) has emerged as an informative disease-associated epigenetic mark. However, cell type-specific contributions to epigenetic dysregulation in disease are unclear as studies have often used bulk brain tissue. Therefore, methods for the deconvolution of bulk H3K27ac profiles are critical. Here we developed the Cell type-specific Histone Acetylation Score (CHAS), a computational tool for inferring cell type-specific signatures in bulk brain H3K27ac profiles. CHAS annotates peaks identified in bulk brain studies of H3K27ac to cell type-specific signals in four major brain cell types, and derives cell type-specific histone acetylation scores as a proxy for cell type proportion. Our method was validated in pseudo-bulk samples and applied to three brain disorder epigenome-wide association studies conducted on bulk brain tissue. CHAS exposed shifts in cellular proportions in Alzheimer's disease (AD), in line with neuropathology, and identified disrupted gene regulatory elements in oligodendrocytes in AD and microglia in autism spectrum disorder (ASD). This contrasts with heritability-based enrichment analyses which indicate genetic risk is associated with microglia in AD and neurons in ASD. Our approach identified cell type specific signalling pathways and putative upstream transcription factors associated with these elements. CHAS enables deconvolution of H3K27ac in bulk brain tissue, yielding cell type-specific biological insights into brain disease-associated regulatory variation.

Keywords: Cell types, histone acetylation, epigenetics, neurodegeneration, Alzheimer's disease, deconvolution, brain

Introduction

H3K27ac is a highly cell type-specific epigenetic modification that marks active enhancers and promoters and is thought to be directly involved in regulating gene expression[1]. Brain disorder risk variants predominantly fall into non-coding and regulatory regions[2], such as those marked by H3K27ac[3]. Given the high cell type specificity and direct link to transcriptional regulation, integrating genome-wide profiles of H3K27ac from disease-relevant cell types can be useful for functional interpretation of these risk variants. This was demonstrated by recent efforts in mapping regulatory elements to major cell types in the human cortex and investigating neurological and psychiatric disease-risk associations[4]. Additionally, H3K27ac responds to external stimuli[1], including those associated with disease. Identifying cell type-specific H3K27ac signals in diseased brains can therefore be used to infer dysregulated signalling pathways and transcription factors at cell type resolution.

Studies on post-mortem human brains have identified genome-wide dysregulation of histone acetylation associated with brain disorders, including Alzheimer's disease (AD), Parkinson's disease (PD) and autism spectrum disorder (ASD)[5–8]. Individuals with syndromic and idiopathic ASD were reported to share a significant proportion of their respective histone acetylomes, and epigenetic dysregulation in the ASD brain was associated with genes involved in synaptic transmission, chemokinesis, and immunity[5]. Both Marzi et al 2018 and Nativio et al 2020 reported enrichment of AD-associated H3K27ac for AD risk variants, as well as for functional pathways related to AD neuropathology such as Aβ metabolic process and Wnt receptor signalling pathway. Additionally, Marzi and colleagues observed genes known to be associated with early onset AD in the vicinity of differentially acetylated peaks. Similarly, dysregulated H3K27ac in the PD brain was located near genes previously implicated in the susceptibility and progression of the disease[8]. However, interpretation of these studies is limited by the use of bulk tissue, which does not account for the high cellular heterogeneity in the brain. This can lead to biological findings being driven by differences in cellular abundance rather than disease-associated changes, and limits follow-up studies in the appropriate cell types. Thus far, to control for cellular composition, studies have used approaches such as CETS[9], a metholymic neuronal marker, and by measuring the neuronal fraction using flow cytometry. However, these methods require DNA methylation or flow cytometry profiles for the same samples and have generally only estimated the proportion of neuronal cell types vs non-neuronal cell types.

Although purified cell or nuclei population and single-cell epigenomic profiling is gaining traction, these methods have generated sparse datasets and have not been applied to the human brain. Thus, there is an urgency for the development of cell type deconvolution methods to better interpret bulk brain epigenome profiles. Deconvolution approaches have been developed for bulk DNA methylation profiles[9–

[11], bulk ATAC-seq profiles[12], and bulk tissue transcriptomes[13-14], however, a tool for bulk H3K27ac profiles is lacking.

We have developed CHAS (Cell type-specific Histone Acetylation Score), a novel computational tool for the identification of cell type-specific peaks within bulk brain H3K27ac profiles and generation of cell type-specific histone acetylation scores. These scores can be used to infer cell type proportions of bulk brain H3K27ac samples, and to perform downstream analyses at cell type resolution. We applied CHAS to three brain disorder H3K27ac datasets: Alzheimer's disease (AD)[6], Parkinson's disease (PD)[8], and autism spectrum disorder (ASD)[5], to detect shifts in cellular composition and re-investigate cell type-specific differential histone acetylation between cases and controls. In contrast to cell-typing based SNP heritability, epigenetic regions dysregulated in disease were enriched in oligodendrocytes for AD and microglia for ASD and cell type specific dysregulated pathways were inferred for both diseases. To the best of our knowledge, CHAS is the first publicly available tool for the deconvolution of bulk brain histone acetylation profiles.

Results

The CHAS model

Enhancers and H3K27ac domains are known to be highly cell type specific. CHAS exploits this cell type specificity to annotate peaks identified in bulk brain studies of H3K27ac to their cell type-specific signals in neurons, microglia, oligodendrocytes and astrocytes, as previously identified[4]. CHAS achieves this by overlapping bulk brain H3K27ac peaks with each cell type specific peak set and identifying which of the bulk peaks are specific to a given cell type. For a bulk peak to be defined as cell type-specific two criteria must be met: (i) the bulk peak is annotated only to a single cell type; (ii) the bulk peak overlaps a predefined percentage of that cell type's peak. This step in CHAS outputs the bulk peaks annotated to a single cell type, 'multiple' cell types (the peak is annotated to more than one cell type), and 'other' (the bulk peak is not annotated to any of the cell types).

Analysis of bulk tissue can be difficult due to differences in cell type proportion in response to disease, or resulting from discrepancies in brain region sampling. To overcome this, using each set of cell type-specific H3K27ac peaks, CHAS generates **C**ell type-specific **H**istone **A**cetylation **S**cores. By averaging the normalized signal intensity of a sample across all peaks specific to a given cell type, CHAS derives a proxy of the proportion of that cell type in the given bulk sample. For sample y and cell type x, CHAS sums up the peak-normalised counts ($s_{p,y}$; for peak-normalisation see **Methods**) across all peaks that are specific

to cell type x, divided by the total number of cell type-specific peaks for cell type x. As a constraint from peak-normalisation, the maximum signal intensity for any given peak and sample is 1 and the resulting CHAS will lie in the interval between 0 and 1 for a given sample and cell type.

$$CHAS_{x,y} = \frac{\sum_{p\ in\ P_x} s_{p,y}}{|P_x|}$$

$CHAS_{x,y}$: **C**ell type-specific **H**istone **A**cetylation **S**core for cell type x, sample y

P_x: Set of cell type specific peaks for cell type x

$s_{p,y}$: Peak-normalised signal intensity for peak p, sample y

Y: Set of all samples

$\forall$ x, $\forall$ p $\in$ P_x: $\max_{y \in Y}(s_{p,y}) = 1$

Of note, given the application of CHAS to deconvolute and control for cellular heterogeneity in epigenetic association studies of brain diseases, we must work under the assumption that disease-related differences in histone acetylation are limited to only a subset of cell type-specific peaks, and that cell-type specific epigenetic variation far outweighs variation associated with disease status[15,16]. We can thus use cell type-specific chromatin immunoprecipitation (ChIP)-seq H3K27ac signal intensities to act as a proxy for cell type proportion in bulk tissue data. The cell type annotation and generation of cell type-specific histone acetylation scores is implemented and freely available in our R package CHAS (https://github.com/neurogenomics/CHAS).

Briefly, our cell type deconvolution tool CHAS requires three inputs:

(i) bulk tissue H3K27ac peaks;

(ii) cell sorted H3K27ac reference peaks;

(iii) counts matrix for the bulk H3K27ac peaks.

CHAS then performs two main analytical tasks:

(i) Identification of cell type-specific peaks in bulk tissue H3K27ac profiles using cell sorted H3K27ac data (**Fig. 1**);

(ii) Generation of cell type-specific histone acetylation scores on the basis of genome-wide average ChIP-seq signal intensities (**Fig. 1**).

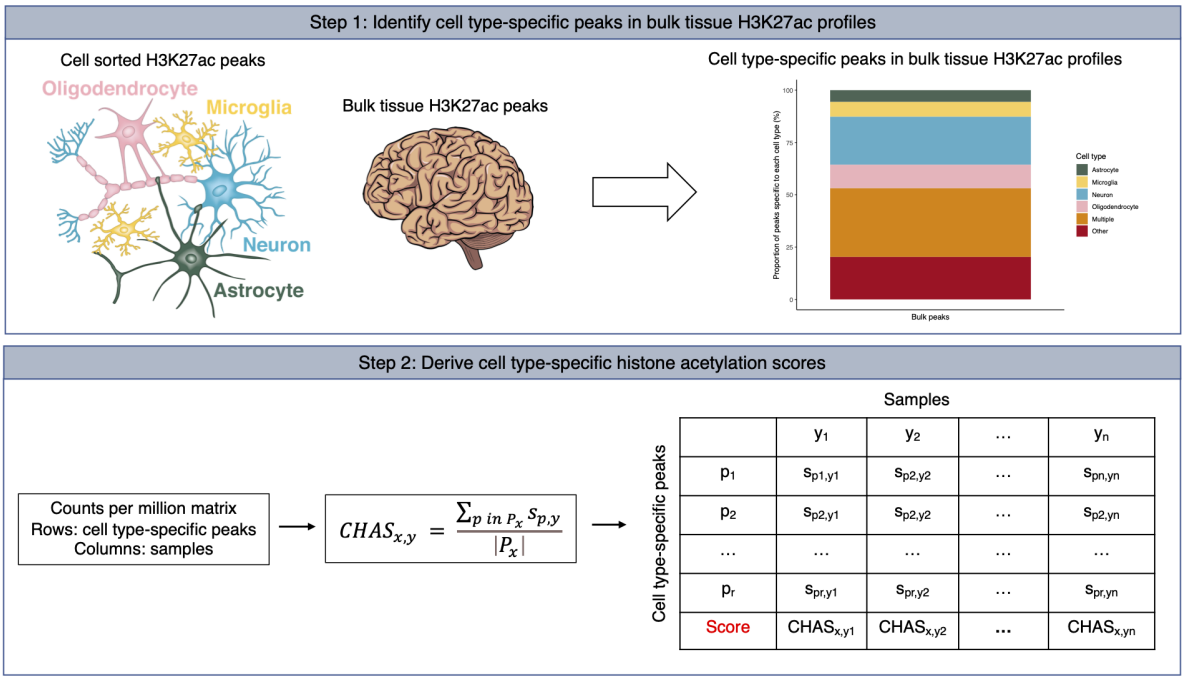


Figure 1. CHAS workflow.

The first step in CHAS is the identification of cell type-specific peaks in bulk tissue H3K27ac profiles, which is achieved using publicly available brain cell sorted H3K27ac data[4]. For a bulk tissue peak to be defined as cell type-specific, it can only be annotated to a single cell type and it must overlap *n*% of that cell type's peak(s). The specific required percentage overlap (*n*) can be specified as input to CHAS. Using the cell type-specific peaks identified in step 1 and a counts per million matrix derived using the bulk tissue H3K27ac data, CHAS generates cell type-specific histone acetylation scores for each sample.

$CHAS_{x,y}$: cell type-specific histone acetylation score for cell type x, sample y; P_x: set of cell type-specific peaks for cell type x; $S_{p,y}$: standardised peak signal intensity for peak p, sample y. A constraint is applied to $S_{p,y}$ whereby for each peak p, the maximum peak signal intensity for any sample equals 1.

CHAS accurately and robustly correlates with known cell type proportions in pseudo-bulk samples

To validate the accuracy of CHAS for predicting cell type proportion in bulk tissue data, we simulated pseudo-bulk H3K27ac profiles of known cell type composition using the cell sorted H3K27ac data previously described[4]. Each pseudo-bulk sample was made up of 30 million randomly sampled reads from astrocytes, microglia, neurons, and oligodendrocytes. The cellular composition was based on experimentally quantified proportions of these cell types from the cortex of 17 individuals with AD and 33 age-matched non-AD subjects[17]. We performed peak calling and read count generation for these pseudo-bulk samples and the resultant peaks and counts were used to run CHAS. For each cell type, we observed

a near-perfect correlation (Pearson's product moment correlation, $R \geq 0.99$, $P < 2.2 \times 10^{-16}$ for each cell type, **Fig. 2**) between the CHAS derived score and the true cell type proportion[17]. To evaluate the robustness of our approach, this process was repeated with varying read depths (20 and 10 million reads) and sample sizes (25 and 10 samples). We again observed a strong correlation between CHAS derived cell type scores and true cell type proportions (Pearson's product moment correlation, $R \geq 0.99$ for each analysis; **Supplementary Figs. 1 and 2**), highlighting that CHAS is robust to read depth and sample size variation.

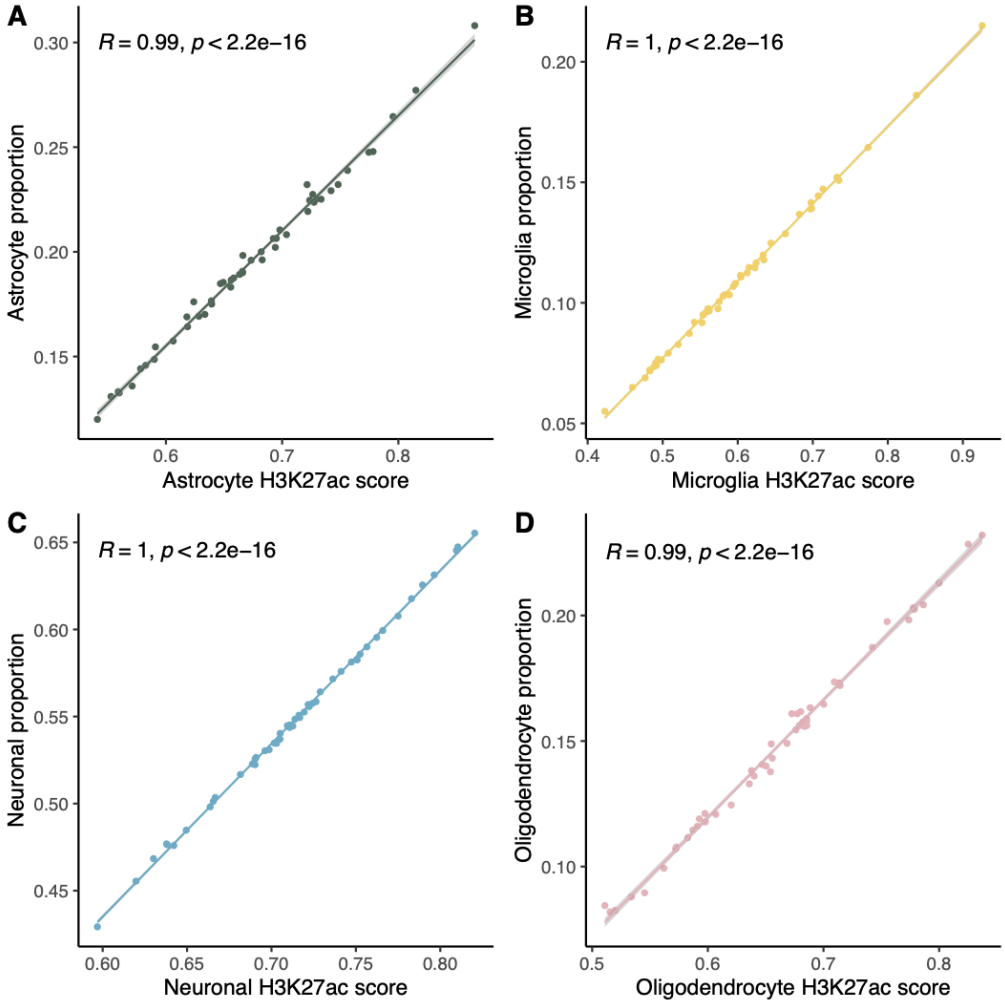


Figure 2. CHAS cell type scores correlate with true cell type proportions.

Scatterplots for the CHAS-derived histone acetylation score vs. the true proportion of the cell type in the pseudo-bulk sample for **A** astrocytes, **B** microglia, **C** neurons, and **D** oligodendrocytes.

Deconvolution of bulk brain H3K27ac in Alzheimer's disease highlights oligodendrocyte-specific epigenetic dysregulation

AD is a complex, highly heritable neurodegenerative disorder. Genetic variants contributing to AD susceptibility predominantly fall into non-coding and regulatory regions[2], such as those marked by H3K27ac. Aberrations in H3K27ac associated with AD have been reported in the human brain[6,7]. However, the brain is a highly heterogeneous tissue and the individual cell types involved in epigenetic dysregulation associated with AD are still unclear. To address this, we used CHAS to deconvolute H3K7ac profiles from the entorhinal cortex of 24 AD cases and 23 controls[6]. Briefly, Marzi and colleagues (2018) found widespread dysregulation of H3K27ac associated with AD, with differentially acetylated regions identified in the vicinity of known early-onset AD risk genes, as well as genomic regions containing variants associated with late-onset AD. Out of 183,353 peaks, 80% (n = 146,144) were annotatable to one or more cell types, with 47% (n = 85,824) specific to a single cell type (**Supplementary Fig. 3A**). In this study, neuronal proportion ($NeuN^+$ fraction) estimates for the same samples had been derived based on matched bulk brain DNA methylation data using the CETS method[6,9], a tool for estimating neuronal proportion. CHAS-derived neuronal scores and CETS-derived neuronal proportion estimates correlated across the 47 samples (Spearman's rank correlation coefficient, $\rho = 0.45$, $P = 0.0016$, **Supplementary Fig. 4**). We next used CHAS to evaluate shifts in cellular composition in the bulk brain data, testing whether these replicate known disease-associated changes. To this end we compared CHAS-derived scores between AD cases and controls for each cell type. As expected, given that AD is primarily associated with neuronal loss, we observed a lower neuronal score in AD brains compared to controls (Welch two-sample t-test, two-sided, $P = 0.028$, average difference in neuronal score = 0.083, 95% CI: 0.0092-0.16, t = 2.27; **Fig. 3A** and **Supplementary Fig. 3B**). The astrocyte score was higher in AD brains when compared to control brains (Welch two-sample t-test, two-sided, $P = 0.049$, average difference in astrocyte score = 0.057, 95% CI: 0.00019-0.11, t = 2.04; **Fig. 3A** and **Supplementary Fig. 3B**), suggesting astrogliosis, a mechanism which has been reported to increase with AD progression[18]. We also report a higher oligodendrocyte score in AD cases (Welch two-sample t-test, two-sided, $P = 0.013$, average difference in oligodendrocyte score = 0.069, 95% CI: 0.015-0.12, t = 2.59; **Fig. 3A** and **Supplementary Fig. 3B**) but no significant difference in microglia score (Welch two-sample t-test, two-sided, $P = 0.072$, average difference in microglia score = 0.053, 95% CI: -0.0049-0.11, t = 1.85; **Fig. 3A** and **Supplementary Fig. 3B**). Overall, these shifts reflect what is known based on disease biology.

We then used the derived cell type scores to re-investigate differential histone acetylation in AD at cell type resolution. Employing the quasi-likelihood *F* test in edgeR[19] we quantified differential acetylation between AD cases and controls, while controlling for cell type proportion (CHAS in neurons, astrocytes, microglia and oligodendrocytes) and age at death. 5,757 peaks were characterised by AD-associated

hyperacetylation and 5,897 were characterised by AD-associated hypoacetylation (false discovery rate (FDR) < 0.05) (**Supplementary Table 1**). To evaluate the likelihood of false-positive associations we repeated the differential histone acetylation analysis using permuted AD case and control labels. Across 100 permuted datasets there was never more than one significant peak at FDR < 0.05, thus making it unlikely that the results of our differential acetylation analysis based on the true AD case and control labels were detected due to chance. Notably, AD-associated hyperacetylated regions were significantly enriched for oligodendrocytes when compared to regions that were not differentially acetylated (two-proportion Z-test, $\chi^2 = 300.57$, $P < 2.2 \times 10^{-16}$, difference in proportion = 0.073; **Fig. 4A**). A similar enrichment of oligodendrocyte-specific peaks was observed for AD-associated hypoacetylated regions (two-proportion Z-test, $\chi^2 = 11.13$, $P = 8.5 \times 10^{-4}$, difference in proportion = 0.014; **Fig. 4A**). In addition, the top AD-associated hyperacetylated peak was also specific to oligodendrocytes and located in the vicinity of both *MVB12B*, a gene implicated in vesicular trafficking, and the transcription factor *PBX3* (**Supplementary Table 1**). *MVB12B* has previously been identified as an AD risk gene[20] and forms part of an oligodendrocyte-enriched gene network in the AD brain[21]. These results are consistent with a previous study in which oligodendrocytes were reported to show the most widespread acetylation differences in the AD brain[22]. The top-ranked AD-hypoacetylated peak was annotated to multiple cell types, and located near *NKAIN3* (**Supplementary Table 1**), an AD GWAS candidate gene that has been shown to be differentially expressed in astrocyte subclusters[23]. Notably, functional enrichment analysis revealed significant association between AD-associated H3K27ac specific to astrocytes and *NKAIN3.* We also observed a significant increase in the microglia proportion amongst AD-associated hypoacetylated regions (two-proportion Z-test, $\chi^2 = 16.61$, $P = 4.60 \times 10^{-5}$, difference in proportion = 0.014; **Fig. 4A**).

Using the Genomic Regions Enrichment of Annotations Tool (GREAT)[24], we were able to match the functional categories associated with AD differentially acetylated bulk peaks to their cell types (**Fig. 4B**). Additionally, we were able to identify cell type-specific dysregulated pathways that were not identified in functional enrichment analyses based on bulk peaks (**Fig. 4B** and **Supplementary Fig. 5**). For instance, oligodendrocyte-specific hyperacetylated peaks were enriched for kinase activity (**Fig. 4B** and **Supplementary Fig. 5A**). Of note, cyclin-dependant kinase 5 was previously implicated in aberrant tau phosphorylation[25,26]. On the other hand, microglia-specific hypoacetylated were enriched for lysine methylation (**Supplementary Fig. 5B**), which has also been associated with tau pathology in the AD brain [27,28]. As reported previously[6], differential H3K27ac was observed in regulatory regions annotated to genes *MAPT*, *APP*, *PSEN1,* and *PSEN2*, known to be associated with early onset AD or directly involved in AD-neuropathology (**Table 1**); with the majority of peaks annotated to *PSEN2* exhibiting oligodendrocyte-specific H3K27ac signatures.

To link CHAS-derived cell type specific differential acetylation with genetic risk for AD, we performed partitioned heritability analysis[29]**.** We quantified enrichment of AD GWAS risk variants within cell type-specific disease-associated H3K27ac, as well as within the background sets of cell type-specific peaks detectable in our bulk brain data. As reported previously, significant enrichment of AD risk loci was found within microglia-specific H3K27ac regions[4] (**Fig. 5A** and **Supplementary Fig. 6**). However, we observed no significant enrichment in any of the other cell type-specific peak sets nor were cell type-specific disease-associated H3K27ac regions enriched for AD heritability, suggesting separate mechanisms between genetic risk and epigenetic dysregulation observed in late-stage disease (**Fig. 5A** and **Supplementary Fig. 6**).

Lastly, using HOMER[30] to assess transcription factor motif enrichment in cell type-specific disease-associated H3K27ac, we report enrichment of binding motifs for both shared and distinct transcription factors across the cell type-specific datasets. For example, only astrocyte-specific AD hyperacetylated regions were enriched for apoptosis associated protein 1 (THAP1) binding motifs ($P < 1 \times 10^{-12}$). Notably, THAP1 has been observed to have a binding site in the specificity protein 1 (Sp1) promoter in humans[31], and Sp1 was reported to be enriched for binding motifs in AD hyperacetylated peaks in Marzi et al (2018). Amongst microglia-specific and oligodendrocyte-specific AD hyperacetylated peaks we observed significant enrichment of binding motifs for Krüppel-like transcription factor 5 (KLF5) (microglia: $P < 1 \times 10^{-16}$, oligodendrocytes: $P < 1 \times 10^{-91}$), a transcription factor implicated in inflammatory responses [32].

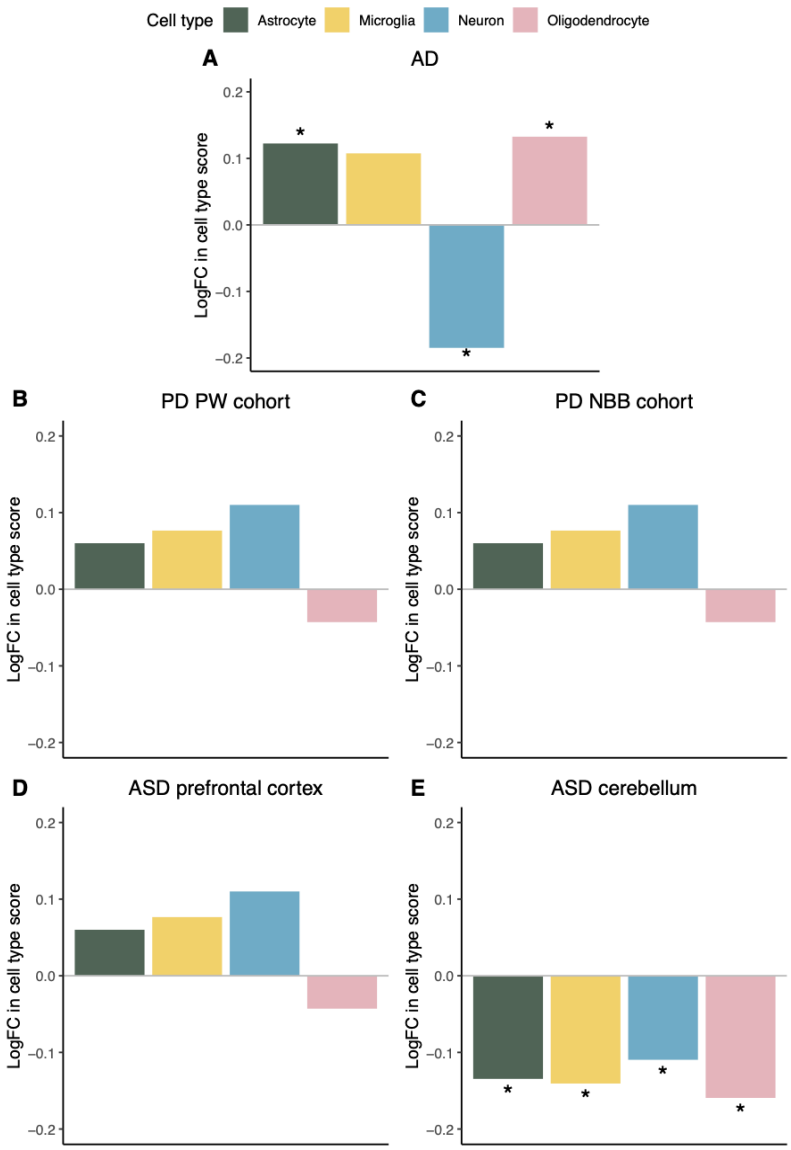


Figure 3. Differences in cell type score between cases and controls recapitulate known cellular alterations in AD.

LogFC in astrocyte, microglia, neuronal, and oligodendrocyte scores in cases compared to controls is shown for multiple histone acetylation studies in brain diseases. Significant differences ($P < 0.05$) are highlighted by an asterix. **A** AD cases show decreased neuronal cell scores in the entorhinal cortex compared to controls ($P = 0.028$), while the astrocyte ($P = 0.049$) and oligodendrocyte scores ($P = 0.013$) are increased in disease[6]**.** No significant differences in cell type scores between PD cases and controls in the prefrontal cortex were observed in the **B** Park West (PW) or **C** Netherlands Brain Bank (NBB) cohorts[8]. **D** ASD cases vs controls showed no cell type score differences in the prefrontal cortex. **E** However, lower cell type scores in ASD cases were found for all four cell types appeared in the cerebellum[5]**.**

Table 1. Top-ranked AD-associated differentially acetylated peaks were identified in close proximity to neuropathology linked and early-onset AD risk genes.

CHR	BP (start-end)	logFC	P value	Cell type	GREAT annotated gene (distance to TSS)
chr1	226824365-226825105	0.49	4.17×10^{-6}	Oligodendrocyte	*PSEN2* (-45828)
chr1	226825314-226826316	0.42	5.07×10^{-6}	Oligodendrocyte	*PSEN2* (-44748)
chr1	226782946-226783977	0.42	1.82×10^{-5}	Oligodendrocyte	*PSEN2* (-87101)
chr1	226798511-226800406	0.48	2.29×10^{-5}	Microglia	*PSEN2* (-71104)
chr1	226753702-226754210	0.88	2.73×10^{-5}	Other	*PSEN2* (-116607)
chr1	226781490-226781832	0.57	3.32×10^{-5}	Oligodendrocyte	*PSEN2* (-88902)
chr1	226784182-226785025	0.34	3.67×10^{-4}	Oligodendrocyte	*PSEN2* (-85959)
chr1	226808312-226810171	0.32	3.85×10^{-4}	Astrocyte	*PSEN2* (-61321)
chr1	226771922-226773890	0.4	5.12×10^{-4}	Other	*PSEN2* (-97657)
chr1	226751565-226752400	0.6	5.33×10^{-4}	Neuron	*PSEN2* (-118580)
chr1	226821015-226822383	0.32	5.63×10^{-4}	Oligodendrocyte	*PSEN2* (-48864)
chr1	226786424-226788137	0.29	0.001	Oligodendrocyte	*PSEN2* (-83282)
chr1	226826797-226828312	0.24	0.001	Astrocyte, microglia, neuron, oligodendrocyte	*PSEN2* (-43008)
chr1	226763768-226764311	0.57	0.003	Neuron	*PSEN2* (-106523)
chr14	73189675-73190186	0.7	7.65×10^{-5}	Other	*PSEN1* (+53484)
chr17	45848349-45850135	0.62	1.61×10^{-9}	Astrocyte, oligodendrocyte	*MAPT* (-45140)
chr17	45870042-45870861	0.42	3.63×10^{-5}	Oligodendrocyte	*MAPT* (-23930)
chr17	45905682-45907558	0.33	9.09×10^{-5}	Neuron, oligodendrocyte	*MAPT* (+12238)
chr17	45907772-45908616	0.42	1.91×10^{-4}	Neuron	*MAPT* (+13812)
chr17	45965650-45966543	0.39	5.04×10^{-4}	Neuron	*MAPT* (+71715)
chr17	45913501-45914698	0.39	5.23×10^{-4}	Neuron, oligodendrocyte	*MAPT* (+19718)
chr17	45942609-45945161	0.19	6.59×10^{-4}	Astrocyte, oligodendrocyte	*MAPT* (+49503)
chr17	45857949-45858752	0.43	8.44×10^{-4}	Oligodendrocyte	*MAPT* (-36031)
chr17	45851477-45851991	0.5	1.47E-03	Oligodendrocyte	*MAPT* (-42648)
chr17	45882681-45884212	0.45	0.002	Other	*MAPT* (-10935)
chr17	45852585-45852866	0.5	0.002	Oligodendrocyte	*MAPT* (-41656)
chr21	26096779-26097326	0.4	0.001	Oligodendrocyte	*APP* (+73601)
chr21	26448550-26449750	0.57	0.002	Neuron, oligodendrocyte	*APP* (-278496)
chr21	25983024-25984204	-0.38	6.45×10^{-5}	Neuron, oligodendrocyte	*APP* (+187040)
chr21	25907479-25908636	-0.37	8.82×10^{-5}	Oligodendrocyte	*APP* (+262596)
chr21	25943915-25945278	-0.33	6.07×10^{-4}	Neuron, oligodendrocyte	*APP* (+226057)
chr21	26115249-26117224	-0.21	0.001	Astrocyte, neuron, oligodendrocyte	*APP* (+54417)
chr21	26006153-26006941	-0.32	0.003	Neuron, oligodendrocyte	*APP* (+164107)

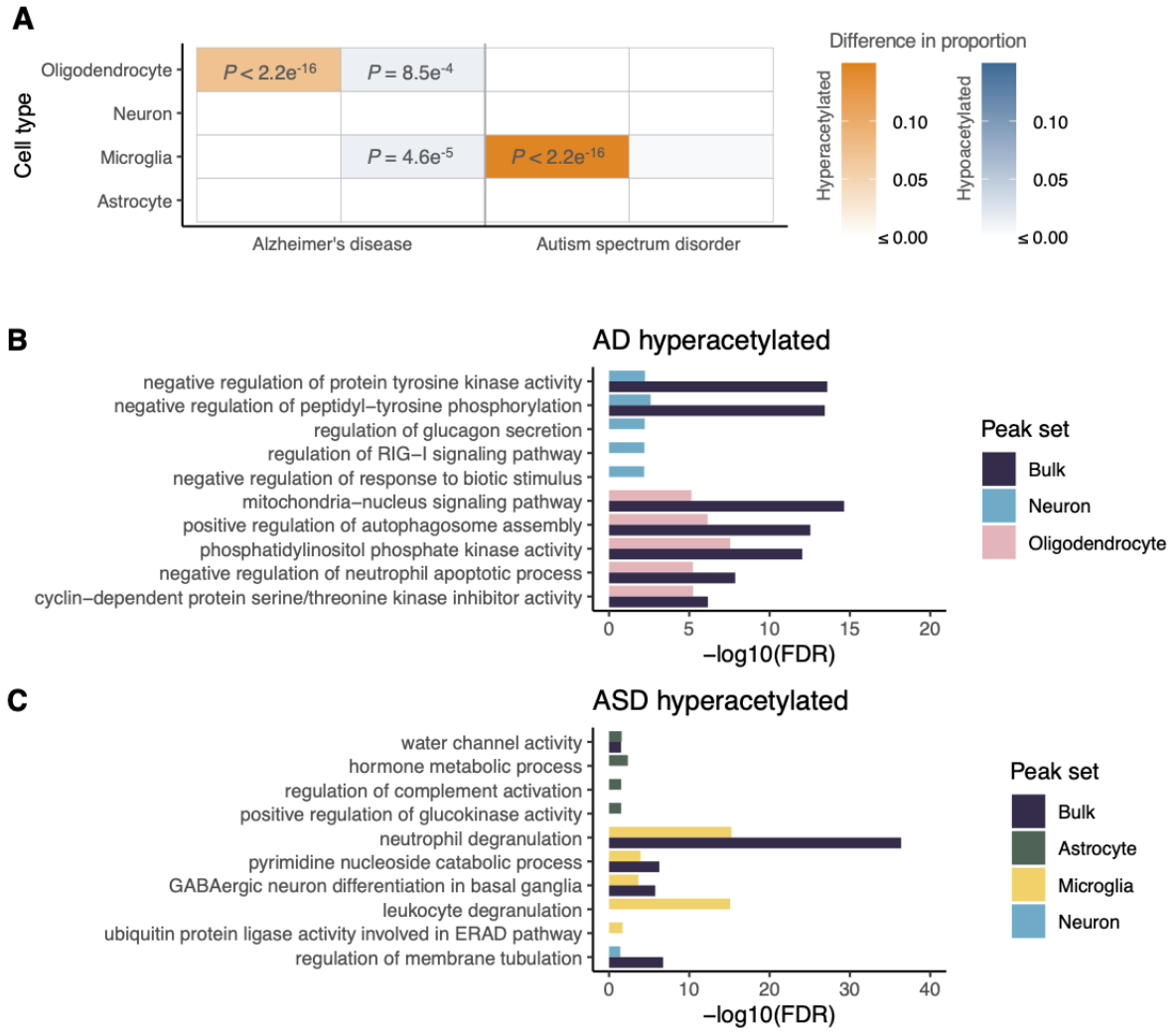


Figure 4. Dysregulated H3K27ac is associated with oligodendrocytes in AD and with microglia in ASD.

A Enrichment of cell types amongst AD- and ASD-differentially acetylated regions vs non-disease associated H3K27ac, calculated using the two proportion Z-test. AD hyper- and hypoacetylated regions were enriched for oligodendrocytes (two-proportion Z-test, $\chi^2 = 300.57$, $P < 2.2 \times 10^{-16}$, difference in proportion = 0.073; two-proportion Z-test, $\chi^2 = 11.13$, $P = 8.5 \times 10^{-4}$, difference in proportion = 0.014). AD hypoacetylated regions were also enriched for microglia (two-proportion Z-test, $\chi^2 = 16.61$, $P = 4.6 \times 10^{-5}$, difference in proportion = 0.014). ASD hyperacetylated regions were enriched for microglia (two-proportion Z-test, $\chi^2 = 1746.5$, $P < 2.2 \times 10^{-16}$, difference in proportion = 0.147). **B** GREAT[24] pathway enrichment analysis using AD entorhinal cortex (bulk) hyperacetylated peaks and cell type-specific hyperacetylated peaks identified using CHAS, highlight cell type specific pathways underlying the bulk signal. **C** Pathway enrichment analysis in ASD prefrontal cortex (bulk) hyperacetylated peaks and cell type-specific hyperacetylated peaks points to activation of microglial immune processes.

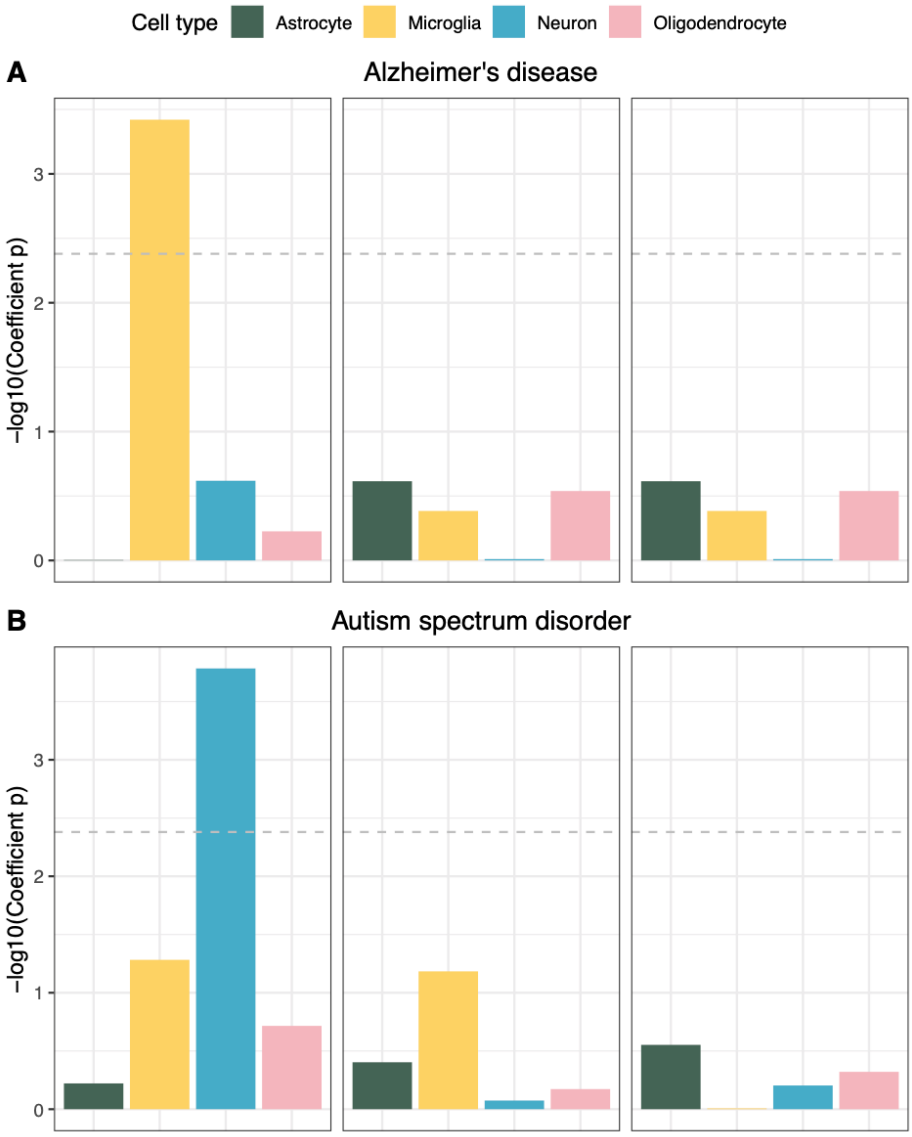


Figure 5. AD risk variants are enriched within microglia-specific H3K27ac peaks, while ASD risk variants are enriched within neuronal-specific H3K27ac domains. Results from partitioned heritability analysis[29] using GWAS for **A** AD[33] and **B** ASD[34] with both disease-associated and background cell type-specific H3K27ac. The y-axis represents the coefficient p value transformed from the coefficient z-score output by LDSC. The grey dashed line at -log10(P) = 2.4 is the cutoff for Bonferroni significance. **A** AD SNP heritability was found to be exclusively enriched in microglial H3K27ac domains. None of the disease-associated cell-type specific peak sets showed any genetic enrichment. **B** ASD genetic risk was found to be enriched in neuron specific H3K27ac peaks. As with AD, disease-associated cell-type specific H3K27ac domains were not enriched for ASD genetic risk.

No significant associations between cortical H3K27ac patterns and Parkinson's disease are found when controlling for cell type scores

Little is known about the etiology of PD, including the role of epigenetics. Although cell type vulnerability in PD is commonly attributed to dopaminergic neurons, genetic risk has also been associated with cholinergic and enteric neurons, as well as oligodendrocytes[35]. However, as our deconvolution of epigenetic signatures of AD highlights, genetic risk and epigenetic dysregulation are not necessarily in agreement in relation to cell types. Therefore, we applied CHAS to a bulk brain H3K27ac study in PD cases and controls. Toker and colleagues (2021) observed genome-wide dysregulation of histone acetylation in prefrontal cortex (PFC) of individuals with PD from two independent cohorts (Park West (PW) and Netherlands Brain Bank (NBB)). The authors reported that PD-associated hyperacetylated regions were annotated to genes implicated in PD pathology, and also report decoupling between promoter H3K27ac and gene expression in the PD brain[8]. Their study included 13 individuals with PD and 10 controls from the PW cohort, in addition to 9 individuals with PD and 9 controls from the NBB cohort. To account for cellular heterogeneity they integrated H3K27ac regions differing between $NeuN^+$ and $NeuN^-$ cell types with brain cell type-specific marker genes, and employed a principal component analysis approach to serve as a proxy for cell type composition across their samples[8]. Using their approach, they found no significant differences in cell type proportion between PD cases and controls in the PW cohort, and they did not test this approach in the NBB cohort.

Using the peaks and read counts generated by Toker and colleagues[8], we filtered out peaks with low read counts before running CHAS to identify cell type-specific peaks in PFC and to generate cell type-specific histone acetylation scores. 66% of bulk peaks were annotatable to at least one cell type in the PW cohort (**Supplementary Fig. 7A**), and 74% of bulk peaks were annotatable to at least one cell type in the NBB cohort (**Supplementary Fig. 7C**). In line with the original study, we found no significant difference in cell type-specific histone acetylation scores between PD cases and controls, in either of the two cohorts (**Figs. 3B and 3C; Supplementary Figs. 7B and 7D**). While this indicates that bulk PFC tissue may be less prone to confounding disease-associated shifts in cellular proportions in PD, it simultaneously does not represent the primarily disease-affected brain region. Differential histone acetylation analysis controlling for cell type proportions revealed no significant differences in acetylation between PD cases and controls in either the PW or NBB cohorts (**Supplementary Tables 2** and **3**).

Cell type deconvolution of autism associated H3K27ac suggest upregulation of microglia in disease

ASD encompasses a group of genetically complex and heterogeneous neurodevelopmental disorders. At the bulk tissue level, dysregulation of H3K27ac in AD brains is associated with genes involved in synaptic transmission and immunity, as well as genes which harbour rare ASD mutations[5]. Taking into consideration the high cell type-specificity of H3K27ac, CHAS provides an opportunity to re-investigate these epigenomic perturbations at cell type resolution. Sun and colleagues[5] performed a histone acetylome-wide association study across 3 different brain regions from ASD cases and age-matched controls. They reported widespread dysregulation of H3K27ac in prefrontal cortex and temporal cortex of ASD cases, with similar changes observed in both brain regions. In contrast, only a small proportion of peaks were differentially acetylated in cerebellum.

Using 80 ChIP-seq samples from prefrontal cortex (40 cases, 40 controls), and 62 samples from cerebellum (31 cases, 31 controls), we called peaks in each brain region using MACS2[36]. After filtering out peaks with low read counts, we defined an optimal peak set for each brain region: 250,614 peaks in prefrontal cortex, and 241,759 peaks in cerebellum. We then used these optimal peak sets to run CHAS to evaluate cell type proportions in each brain region, and to generate cell type-specific scores for each sample in each brain region. In the PFC, 72% of bulk peaks were annotatable to at least one cell type (**Supplementary Fig. 8A**), whereas in the cerebellum less than 50% of peaks could be annotated to a cell type (**Supplementary Fig. 8C**). This is most likely explained by epigenetic differences across brain regions in the reference and test datasets[37,38]: brain region specific differences could exist in the epigenetic state of the same cell type, for instance microglia across multiple brain regions. Similarly, there can be differences in the actual cell types located in different brain regions. For example, the cortex contains highly specialized pyramidal neurons, while only the cerebellum hosts granule cells. Generation of cell type-specific histone acetylation scores using CHAS revealed significantly lower cell type scores for all four cell types in cerebellum of ASD cases when compared to controls (**Fig. 3E** and **Supplementary Fig. 8D**), but no differences in cell type scores in the PFC (**Fig. 3D** and **Supplementary Fig. 8C**). This likely relates to large scale cellular composition differences between cortex and cerebellum, highlighting the need for a brain region appropriate reference. Hence, the observed disease-associated differences in cell type scores in the cerebellum should be interpreted with caution.

Differential histone acetylation analysis controlling for age at death and cell type score using CHAS revealed ASD-associated differentially acetylated peaks nearly exclusively in prefrontal cortex (8,761 differentially acetylated peaks; **Supplementary Table 4**) compared to cerebellum (two differentially acetylated peaks; **Supplementary Table 5**), in line with observations from the original study[5]. ASD-

associated hyperacetylated regions in PFC were significantly enriched for microglia when compared to the background peak set (two-proportion Z-test, χ^2 = 1746.5, $P < 2. \times 10^{-16}$, difference in proportion = 0.147; **Fig. 4A**). Interestingly, altered microglial states in ASD had previously been reported[39,40]. In addition, the top ranking ASD-associated hyperacetylated peak was specific to microglia and located ~14kb upstream of *TRERF1* (**Supplementary Table 4**). Exome sequencing in individuals with intellectual disability previously revealed a missense variant in this gene, suggesting its involvement in intellectual disability pathogenesis[41]. The top ranking ASD-associated hypoacetylated peak peak was not annotatable to any cell type and was located ~21kb downstream of *SVIL* (**Supplementary Table 4**), which has been shown to harbour a single de novo mutation in children with ASD[42].

Performing functional enrichment analysis at cell type resolution allowed us to overlap the categories associated with the differentially acetylated bulk peaks, with the cell types they were specific to, as well as identify distinct functional enrichments across the four cell types that were not identified in the bulk analysis. For example, pathway enrichment analysis using ASD hyperacetylated bulk peaks revealed enrichment for immune related processes, which were predominantly specific to microglia (**Fig. 4C** and **Supplementary Fig. 9**). However, different immune related processes were also found to be enriched when only using astrocyte-specific ASD-hyperacetylated regions in the analysis (**Fig. 4C** and **Supplementary Fig. 9**). This is in line with the growing body of evidence highlighting the role of glial cells and the immune system in ASD[39,43–46]. We quantified enrichment of ASD GWAS risk variants within cell type-specific disease-associated H3K27ac, as well as within the background sets of cell type-specific peaks detectable in our bulk brain data. As reported previously, significant enrichment of ASD risk loci was found within neuron-specific H3K27ac regions[4] (**Fig. 5B** and **Supplementary Fig. 10**).

Discussion

Given that histone acetylation is highly cell type-specific, inferring and deconvolving cell type-specific signatures in bulk tissue H3K27ac profiles is critical to our interpretation of these profiles in neurodegenerative disorders. Here, we demonstrate that we can exploit this specificity to annotate peaks identified in bulk brain studies of H3K27ac to their cell type-specific constituent parts. We furthermore show that cell type-specific ChIP-seq signal intensities can be leveraged to generate cell type-specific scores to act as a proxy for cell type proportion. As a result, we developed CHAS, a novel tool for cell type deconvolution of bulk brain H3K27ac profiles. To the best of our knowledge, CHAS is the only publicly available tool for deconvolution of histone acetylation.

CHAS showed highly convincing performance on pseudo-bulk H3K27ac profiles of known cell type composition yielding near perfect correlations with the true underlying cellular proportions. These correlations remain stable at decreased sample sizes and at lower read depths, highlighting the robustness of our approach. Additionally, comparison of CHAS-generated cell type scores between AD cases and controls[6] reflected known shifts in cellular proportion associated with the disease, as well as significant correlation between neuronal CHAS and DNA methylation derived estimates of neuronal proportion[9].

To illustrate the utility of CHAS for interpretation of bulk tissue H3K27ac profiles, we applied it to three epigenome-wide association studies of brain disorders[5,6,8]. Deconvolution of H3K27ac profiles from the AD brain using CHAS has highlighted that hypoacetylation in late-stage AD is enriched for both oligodendrocyte-specific H3K27ac and microglia-specific H3K27ac, whereas AD-associated hyperacetylation is only enriched for oligodendrocyte-specific H3K27ac. An independent study also found the largest H3K27ac changes in oligodendrocytes in the hippocampus and dorsolateral prefrontal cortex of individuals with AD[22]. Taken together, these data suggest that these oligodendrocyte-specific H3K27ac changes are not limited to a single brain region and warrant further investigation of the role of oligodendrocyte H3K27ac dysregulation in AD. The marked hypoacetylation in microglia could reflect an increase in the activity of histone deacetylases (HDACs). In line with this, a recent study found that genetic ablation of microglial HDAC1 and HDAC2 in an AD mouse model reduced amyloid plaque burden and rescued memory deficits[47]; thus suggesting modulation of HDAC activity in microglia as a potential therapeutic intervention for AD. Whereas epigenetic variation in late-stage AD predominantly points to oligodendrocytes, genetic risk points to microglia, suggesting independent biological mechanisms. Furthermore, our study corroborates the finding that genetic risk for AD is enriched at microglia-specific regulatory elements[4,22,48] and genes[35]. Additionally, we used CHAS to identify disease-associated pathways and transcription factors at cell type resolution. This highlighted a role of kinases in oligodendrocytes, which may be of relevance to neurodegenerative disorders due to their role in CNS myelination[49,50]. In support of this, a recent study reported that pro-myelinating strategies were able to rescue cognitive and physiological deficits in a mouse model of AD[51]. CHAS has broad utility across histone acetylation studies of diverse brain diseases, as illustrated by its application to a study of H3K27ac in ASD and control brains. Our analyses provided evidence for the upregulation of regulatory elements in microglia in ASD brains, and prioritises candidate genes and pathways to study in this cell type. For example, *TRERF1* has been reported to harbour a de novo mutation in children with ASD[52], and we found that the top significant ASD hyperacetylated peak was annotated to this gene and specific to microglia. Taken together, our findings highlight the potential of CHAS in revealing biological insights and aiding prioritisation of relevant cell types and pathways for genetic and epigenetic studies of brain disorders.

There are important caveats to consider in relation to CHAS. The power of CHAS to detect cell type-specific peaks from bulk histone acetylation profiles is inherently linked to the proportion of those cells in the bulk tissue sample. Brain cell types found at low proportions, including for example microglia, will have lower power to be detected compared to high-frequency cell types, such as oligodendrocytes. This may in turn affect the robustness and power of the downstream generation of the cell type specific histone acetylation score and ability to detect differential acetylation in low-frequency cell types. Additionally, the sample sizes of the currently available bulk brain histone modification studies are not ideal for genome-wide epigenomic studies. Here, the inclusion of four additional covariates via the cell type-specific histone acetylation scores potentially contributes to a decrease in power to detect disease-associated differences. In this regard it is interesting to note that, as in the case of AD and ASD, the inclusion of covariates controlling for cell type proportion in moderate and large sized studies appears to improve power to detect acetylation differences, presumably via removing noise from variation caused by cellular heterogeneity. The cell sorted data on which CHAS is based has three limitations: 1) it is only available for four major brain cell types, excluding rarer cell types such as pericytes or endothelial cells. Therefore, whilst we can implicate the role of microglia, neurons, oligodendrocytes, and/or astrocytes in brain disorders, it would be amiss to disregard other cell types for which we do not have adequate data. 2) It is limited with regard to cell subtype diversity. It is well known, for example, that there are numerous types of functionally and regionally distinct neurons, which are all lumped together in the NeuN^{+} population in the current reference. 3) Our reference data does not account for different cell states. For instance, multiple microglial phenotypes have been transcriptionally and functionally characterised[53–55], one example being disease associated microglia[56]. Our current reference is based on a neuropathology-free, pediatric dataset in which such states – if present – are lumped into one category. Future reference atlases focusing on specific sub-cell types and states should enable more detailed deconvolution in this regard. Importantly, the performance of CHAS in bulk cerebellum samples also highlights a need to use brain-region specific reference datasets, to account for regional differences in cell types and states. At present, CHAS is limited to bulk cortex studies of H3K27ac because of reference atlas availability. However, CHAS could easily be extended to other brain cell types, regions or completely distinct tissues, with appropriate reference datasets; and we would expect it to perform in the same manner for these extensions. Most promisingly, we hope that future availability of single cell H3K27ac profiles across brain regions will enable us to adapt CHAS to deconvolute more refined cell subtypes and states.

Going forward epigenetic studies at single cell resolution promise to create a more comprehensive picture of the epigenetic landscape associated with neurological and psychiatric diseases. However, as single cell epigenomic profiling is still in its early stages of application, generates sparse data, and has not yet been achieved in the human brain, CHAS provides a unique opportunity to infer cell type-specific signatures in bulk brain histone acetylation profiles. Importantly, this can yield novel insights into epigenetic changes

contributing to brain disorder risk and progression that are associated with specific cell types. Finally, CHAS is freely and publicly available at https://github.com/neurogenomics/CHAS, adding to the existing repertoire of methods for cell type deconvolution.

Methods

CHAS

Cell type-specific H3K27ac annotation in CHAS is based on H3K27ac profiles from purified populations of astrocytes, microglia, neurons, and oligodendrocytes[4]. For a given human bulk brain H3K27ac dataset, CHAS overlaps peaks detected in bulk with each cell type peak set to identify cell type-specific peaks within the bulk H3K27ac profiles. Overlaps are identified using the GenomicRanges package in R[57]. Of note, no specific threshold of overlap is required for this first cell type annotation step: even if a bulk peak overlaps a purified cell type reference peak by only one base pair, it is annotated to the cell type in this first stage. To derive the cell type-specific histone acetylation score (CHAS), we wanted to ensure that only high-confidence and highly cell type-specific peaks are included and therefore the following two criteria must be met: (i) the bulk peak is annotated only to a single cell type; (ii) the bulk peak overlaps a predefined percentage of the given cell type peak. This predefined percentage can be specified in the *CelltypeSpecificPeaks()* function and is set to 50% by default and for all analyses presented in this manuscript.

For each sample, CHAS generates a cell type-specific histone acetylation score by averaging the normalised signal intensity across all peaks specific to a given cell type. First, read counts across peaks are converted to counts per million (cpm) to account for variation in library size. To further normalise the signal intensity at a given peak, for each peak p, the counts are divided by the highest observed read count for that peak, thereby placing the peak-normalised counts on a scale between 0 and 1. For a sample y and cell type x, the peak-normalised counts per million $s_{p,y}$ are summed up across all peaks p that are specific to cell type x, divided by the total number of cell type specific peaks for cell type x, P_x. As a constraint from peak-normalisation, the maximum signal intensity for any given peak and sample is 1 and the resulting CHAS will lie between 0 and 1 for a given sample and cell type. The cell type annotation and generation of cell type specific histone acetylation scores is implemented and automated in our R package CHAS (https://github.com/neurogenomics/CHAS).

$$CHAS_{x,y} = \frac{\sum_{p\ in\ P_x} s_{p,y}}{|P_x|}$$

$CHAS_{x,y}$: **C**ell type-specific **H**istone **A**cetylation **S**core for cell type x, sample y

P_x: Set of cell type specific peaks for cell type x

$s_{p,y}$: Peak-normalised signal intensity for peak p, sample y

Y: Set of all samples

$\forall$ x, $\forall$ p $\in$ P_x: $\max_{y \in Y}(s_{p,y}) = 1$

Validation of CHAS

We validated CHAS by simulating pseudo-bulk H3K27ac profiles of known cell type composition based on the raw sequencing data from astrocytes, microglia, neurons and oligodendrocytes by Nott and colleagues[4]. The cell type composition of each pseudo-bulk sample was based on proportions of astrocytes, microglia, neurons, oligodendrocytes and endothelial cells in brain tissue from older individuals, which had been quantified in an independent study using immunohistochemistry[17]. Based on the reported cell type proportions of 49 post-mortem brain samples we generated 49 pseudo-bulk samples, pooling a total of 30 million randomly sampled reads per sample from the raw H3K27ac data of the four cell types. As our reference did not include H3K27ac profiles for endothelial cells, we excluded the proportion of this cell type and instead used the relative proportions of the four other cell types. We ran CHAS to generate a cell type-specific histone acetylation score for each pseudo-bulk sample and compared these to the true cell type proportions using Pearson correlation coefficients. To additionally evaluate the robustness of CHAS with respect to sample size and sequencing depth, this process was repeated across the 49 samples with 20 million and 10 million randomly sub-sampled reads, as well as using 30 million reads in random subsets of 25 and 10 samples.

Application of CHAS to bulk brain H3K27ac datasets

H3K27ac in entorhinal cortex from AD cases and controls: Marzi et al. 2018

To demonstrate reproducibility and undertake preprocessing using updated versions of software and the most recent reference genome, raw ChIP-seq data from our previous study was downloaded from SRA under accession number PRJNCA297982[6]. We performed basic quality control using fastQC[58]. Using bowtie2[59] the fastq files were aligned to the most recent human reference genome (GRCh38)[60]. The resulting SAM files were converted to binary (BAM) format using SAMtools[30]. Duplicates, unmapped reads, and reads with a sequence quality score q < 30 were removed from all BAM files and the filtered BAM files were subsequently merged into one grouped file. Next, using MACS2[36] we performed peak calling on the merged file of all samples. The following peak sets were subsequently filtered out: 1) peaks which

overlapped the ENCODE blacklist peaks (https://github.com/Boyle-Lab/Blacklist), 2) peaks which were located in unmapped contigs or mitochondrial DNA, and 3) peaks which did not meet a significance threshold of $P < 10^{-7}$ for peak calling. Read count generation for each sample was performed using featureCounts[61] and read counts were converted to and stored in a cpm matrix, keeping peaks with a minimum of three samples showing ≥ 1 read per million. This resulted in a total of 183,353 peaks to be used in downstream analyses. This optimal peak set and cpm matrix were used as input to run CHAS to identify cell type-specific peaks in the bulk H3K27ac profiles and to generate cell type-specific H3K27ac scores as a proxy for the proportion of each cell type in the bulk peak set. Of note, to annotate cell types to each bulk peak and to calculate the cell type proportions across the bulk peaks, we only required an overlap of at least one base pair between the bulk peak and the cell type peak. However, for peaks included in the cell type specific histone acetylation score we required a more stringent overlap of at least 50% of the cell type peak interval. The CHAS-generated cell type-specific histone acetylation scores were used to detect shifts in cellular composition between AD cases and controls, by comparing the means using a Welch two-sample t-test. Differences in histone acetylation between AD cases and controls were analysed as previously described[6], but including the CHAS derived cell type scores, instead of the neuronal proportion estimator based on CETS[9]. Briefly, the quasi-likelihood F test in the Bioconductor package edgeR[19] was used to test for differences in histone acetylation between AD cases and controls, while controlling for age at death and cell type proportions using the cell type-specific histone acetylation scores for the four brain cell types. All covariates were treated as continuous numeric variables. Peaks were considered differentially acetylated at FDR<0.05 (controlled by Benjamini–Hochberg for n=183,353 tests). To additionally confirm that we had adequately controlled for false-positive associations, we permuted the AD case and control labels 100 times and repeated the differential histone acetylation analysis as described above.

H3K27ac in prefrontal cortex from PD cases and controls: Toker et al. 2021

Peak lists (hg19 reference build) and read count tables for two independent cohorts (Park West (PW) and Netherlands Brain Bank (NBB))[8] were downloaded from https://github.com/ltoker/ChIPseqPD and subsequently used as input to CHAS. Both peak lists were in narrowPeak format and were filtered to include peaks mapping to canonical chromosomes, and to exclude peaks which overlapped those in blacklisted regions https://github.com/Boyle-Lab/Blacklist, as well as those not meeting a significance threshold of $P < 10^{-7}$ for peak calling. For the PW cohort, a total of 171,285 peaks were used for downstream analyses and for the NBB cohort, a total of 132,390 peaks were used for downstream analyses. From the counts tables we excluded the sample outliers identified in Toker et al (2021) and ensured that the peaks in the tables matched those in the filtered peak set and then performed final filtering, keeping peaks with a minimum of three samples showing ≥ 1 read per million for the differential histone acetylation analysis. This left us with 152,823 peaks in the PW cohort, and 111,396 peaks in the

NBB cohort to be used for differential histone acetylation analysis. The counts tables were converted to cpm matrices and along with the filtered peak sets were used as input to CHAS, as previously for the AD dataset, using hg19-based cell type peak sets. The CHAS-generated cell type-specific histone acetylation scores were used to detect shifts in cellular composition between PD cases and controls, by comparing the means using a Welch two-sample t-test. Differences in histone acetylation between PD cases and controls were analysed as described above, controlling for age at death and cell type proportions using the cell type-specific histone acetylation scores.

H3K27ac in prefrontal cortex, temporal cortex, and cerebellum from ASD cases and controls: Sun et al. 2016

ChIP-seq reads mapped to the human reference genome (hg19) using BWA[62] by Sun and colleagues (2016) were downloaded from Synapse under accession number syn4587616. We downloaded 80 libraries from the prefrontal cortex and 62 libraries from the cerebellum. These were the same libraries that were used in the original study for peak calling[5], with exception of one prefrontal cortex sample which was not available on Synapse. Downloaded files were in BAM format and all pre-processing steps were performed as described previously (see H3K27ac in entorhinal cortex from AD cases and controls: Marzi et al. 2018 under **Methods**), using hg19-based cell type peak sets. The optimal peak sets for downstream analyses totaled 250,614 peaks for prefrontal cortex, and 241,759 peaks for cerebellum. These, alongside the derived cpm matrices, were used as input to CHAS. Differences in histone acetylation between ASD cases and controls for each brain region were analysed as described above, controlling for age at death and cell type proportions using the CHAS-generated cell type score, while disease-associated differences in cell type proportions were quantified using a Welch two-sample t-test on the CHAS-derived cell type-specific histone acetylation scores.

Calculating proportions of cell type-specific H3K27ac peaks within disease-associated differentially acetylated regions

To test whether the proportion of cell type-specific peaks in the disease-associated differentially acetylated regions differed significantly from the background set of non-differentially acetylated peaks, we compared the proportions of cell type-specific peaks in the significantly hyper- and hypoacetylated peak sets with the background distribution (based on peaks which were not differentially acetylated) using a two-proportion Z-test.

Genomic annotation and enrichment analysis

Gene annotation and gene ontology analyses were performed as previously described in Marzi et al (2018) for disease-associated cell type-specific hyperacetylated and hypoacetylated peaks, using the full cell type-specific peak sets as the background.

Partitioned heritability analysis

To estimate the proportion of disease SNP-heritability attributable to cell type-specific H3K27ac peaks identified in bulk brain data, we performed partitioned heritability analysis as implemented in LDSC[29]. For each cell type-specific peak set, annotation files were generated and used to compute LD scores. Publicly available GWAS summary statistics for a recent AD GWAS[33] and ASD GWAS[34] were downloaded and converted to the required format for LDSC. Steps for the analysis were followed as instructed here https://github.com/bulik/ldsc/wiki and required files were downloaded from https://alkesgroup.broadinstitute.org/LDSCORE/GRCh38. For each annotation, LDSC was run using the full baseline model[29], thereby computing the proportion of SNP-heritability associated with the annotation of interest, while taking into account all the annotations in the baseline model. LDSC was performed for cell type specific peaks in bulk as well as cell type specific hyper- and hypoacetylated peak sets.

Funding

SJM is funded by the Edmond and Lily Safra Early Career Fellowship Program (https://www.edmondjsafra.org). SJM and AN are funded by the UK Dementia Research Institute (https://ukdri.ac.uk), which receives its funding from UK DRI Ltd, funded by the UK Medical Research Council (https://mrc.ukri.org), Alzheimer's Society (https://www.alzheimers.org.uk) and Alzheimer's Research UK (https://www.alzheimersresearchuk.org). KBM is funded by the UK Medical Research Council Doctoral Training Partnership (https://mrc.ukri.org).

Author contributions

Conceptualization: SJM

Methodology: SJM

Software: KBM

Formal Analysis: KBM

Validation: KBM

Investigation: AN, SJM

Data Curation: KBM, SJM, AN

Writing – original draft: KBM, SJM

Writing – review & editing: KBM, SJM, AN

Visualization: KBM

Supervision: SJM

Project administration: SJM

Acknowledgments

We thank Alan E Murphy and Brian M Schilder at the UK Dementia Institute, Imperial College London for feedback and helpful discussions on the CHAS R package.

References

1. Creyghton, M. P. *et al.* Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 21931–21936 (2010).
2. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
3. Hannon, E., Marzi, S. J., Schalkwyk, L. S. & Mill, J. Genetic risk variants for brain disorders are enriched in cortical H3K27ac domains. *Mol. Brain* **12**, 7 (2019).
4. Nott, A. *et al.* Brain cell type–specific enhancer–promoter interactome maps and disease-risk association. *Science* **366**, 1134–1139 (2019).
5. Sun, W. *et al.* Histone Acetylome-wide Association Study of Autism Spectrum Disorder. *Cell* **167**, 1385–1397.e11 (2016).
6. Marzi, S. J. *et al.* A histone acetylome-wide association study of Alzheimer's disease identifies disease-associated H3K27ac differences in the entorhinal cortex. *Nat. Neurosci.* **21**, 1618–1627 (2018).
7. Nativio, R. *et al.* An integrated multi-omics approach identifies epigenetic alterations associated with Alzheimer's disease. *Nat. Genet.* **52**, 1024–1035 (2020).
8. Toker, L. *et al.* Genome-wide histone acetylation analysis reveals altered transcriptional regulation in the Parkinson's disease brain. *Molecular Neurodegeneration* vol. 16 (2021).
9. Guintivano, J., Aryee, M. J. & Kaminsky, Z. A. A cell epigenotype specific model for the correction of brain cellular heterogeneity bias and its application to age, brain region and major depression. *Epigenetics* **8**, 290–302 (2013).
10. Houseman, E. A. *et al.* DNA methylation arrays as surrogate measures of cell mixture distribution. *BMC Bioinformatics* **13**, 86 (2012).
11. Teschendorff, A. E., Zhu, T., Breeze, C. E. & Beck, S. EPISCORE: cell type deconvolution of bulk

tissue DNA methylomes from single-cell RNA-Seq data. *Genome Biol.* **21**, 221 (2020).

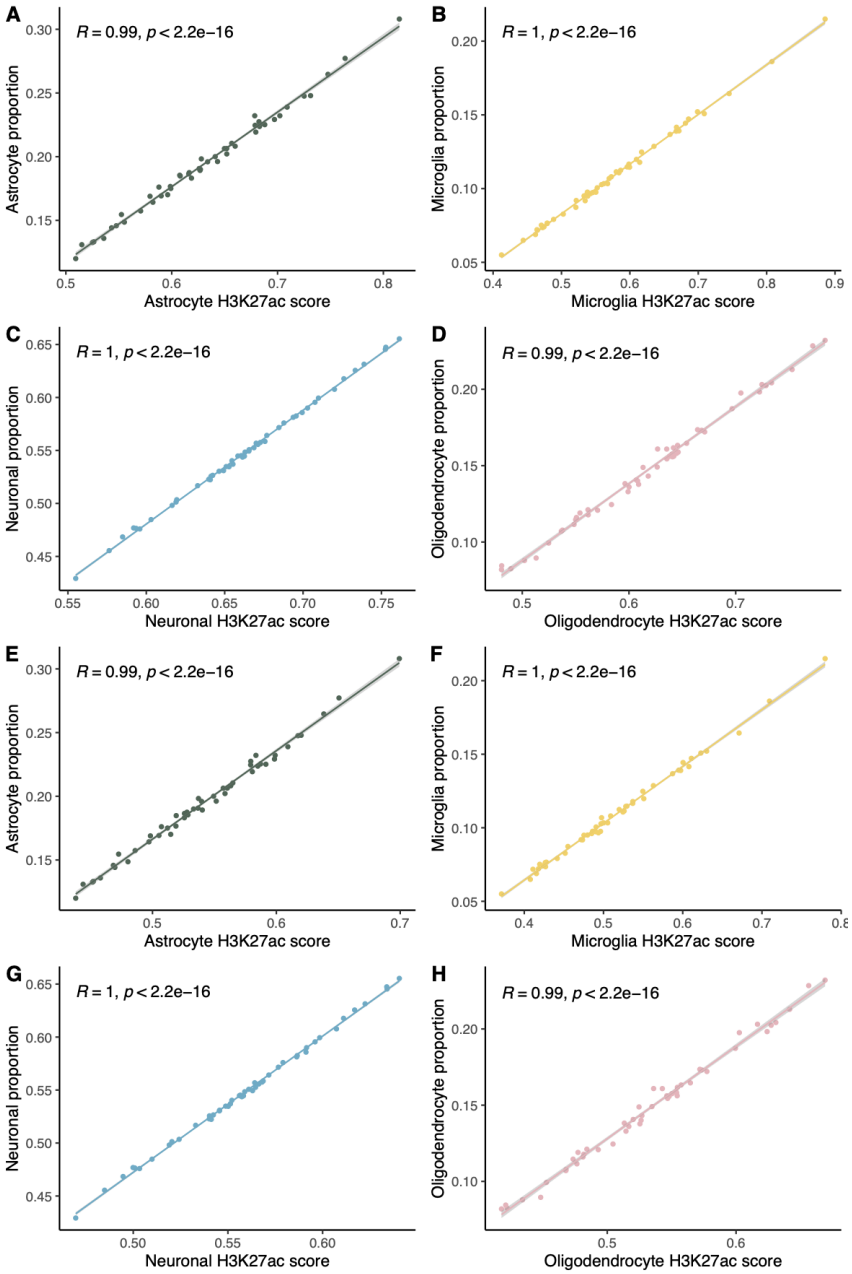
12. Li, H. *et al.* DeconPeaker, a Deconvolution Model to Identify Cell Types Based on Chromatin Accessibility in ATAC-Seq Data of Mixture Samples. *Front. Genet.* **11**, 392 (2020).
13. Gong, T. & Szustakowski, J. D. DeconRNASeq: a statistical framework for deconvolution of heterogeneous tissue samples based on mRNA-Seq data. *Bioinformatics* **29**, 1083–1085 (2013).
14. Dong, M. *et al.* SCDC: bulk gene expression deconvolution by multiple single-cell RNA sequencing references. *Brief. Bioinform.* **22**, 416–427 (2021).
15. Girdhar, K. *et al.* Cell-specific histone modification maps in the human frontal lobe link schizophrenia risk to the neuronal epigenome. *Nat. Neurosci.* **21**, 1126–1136 (2018).
16. Mendizabal, I. *et al.* Cell type-specific epigenetic links to schizophrenia risk in the brain. *Genome Biology* vol. 20 (2019).
17. Patrick, E. *et al.* Deconvolving the contributions of cell-type heterogeneity on cortical gene expression. *PLoS Comput. Biol.* **16**, e1008120 (2020).
18. Kamphuis, W. *et al.* Glial fibrillary acidic protein isoform expression in plaque related astrogliosis in Alzheimer's disease. *Neurobiol. Aging* **35**, 492–510 (2014).
19. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
20. Lambert, J. C. *et al.* Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease. *Nat. Genet.* **45**, 1452–1458 (2013).
21. McKenzie, A. T. *et al.* Multiscale network modeling of oligodendrocytes reveals molecular components of myelin dysregulation in Alzheimer's disease. *Molecular Neurodegeneration* vol. 12 (2017).
22. Ramamurthy, E. *et al.* Cell type-specific histone acetylation profiling of Alzheimer's Disease subjects and integration with genetics. doi:10.1101/2020.03.26.010330.
23. Grubman, A. *et al.* A single-cell atlas of entorhinal cortex from individuals with Alzheimer's disease reveals cell-type-specific gene expression regulation. *Nat. Neurosci.* **22**, 2087–2097 (2019).
24. McLean, C. Y. *et al.* GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* **28**, 495–501 (2010).
25. Noble, W. *et al.* Cdk5 is a key factor in tau aggregation and tangle formation in vivo. *Neuron* **38**, 555–565 (2003).
26. Seo, J. *et al.* A cyclin-dependent kinase 5-derived peptide inhibits Cdk5/p25 activity and improves neurodegenerative phenotypes. *bioRxiv* 2020.05.12.090472 (2020) doi:10.1101/2020.05.12.090472.
27. Thomas, S. N. *et al.* Dual modification of Alzheimer's disease PHF-tau protein by lysine methylation and ubiquitylation: a mass spectrometry approach. *Acta Neuropathol.* **123**, 105–117 (2012).
28. Funk, K. E. *et al.* Lysine methylation is an endogenous post-translational modification of tau protein in human brain and a modulator of aggregation propensity. *Biochemical Journal* vol. 462 77–88

(2014).
29. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
30. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
31. Dosunmu, R. *et al.* Lifespan profiles of Alzheimer's disease-associated genes and products in monkeys and mice. *J. Alzheimers. Dis.* **18**, 211–230 (2009).
32. Yan, W., Wu, J., Song, B., Luo, Q. & Xu, Y. Treatment with a brain-selective prodrug of 17β-estradiol improves cognitive function in Alzheimer's disease mice by regulating klf5-NF-κB pathway. *Naunyn-Schmiedeberg's Archives of Pharmacology* vol. 392 879–886 (2019).
33. Jansen, I. E. *et al.* Genome-wide meta-analysis identifies new loci and functional pathways influencing Alzheimer's disease risk. *Nat. Genet.* **51**, 404–413 (2019).
34. Grove, J. *et al.* Identification of common genetic risk variants for autism spectrum disorder. *Nat. Genet.* **51**, 431–444 (2019).
35. Bryois, J. *et al.* Genetic identification of cell types underlying brain complex traits yields insights into the etiology of Parkinson's disease. *Nature Genetics* vol. 52 482–493 (2020).
36. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
37. Davies, M. N. *et al.* Functional annotation of the human brain methylome identifies tissue-specific epigenetic variation across brain and blood. *Genome Biol.* **13**, R43 (2012).
38. Marzi, S. J. *et al.* Tissue-specific patterns of allelically-skewed DNA methylation. *Epigenetics* **11**, 24–35 (2016).
39. Suzuki, K. *et al.* Microglial activation in young adults with autism spectrum disorder. *JAMA Psychiatry* **70**, 49–58 (2013).
40. Zhan, Y. *et al.* Deficient neuron-microglia signaling results in impaired functional brain connectivity and social behavior. *Nat. Neurosci.* **17**, 400–406 (2014).
41. Anazi, S. *et al.* Clinical genomics expands the morbid genome of intellectual disability and offers a high diagnostic yield. *Mol. Psychiatry* **22**, 615–624 (2017).
42. Neale, B. M. *et al.* Patterns and rates of exonic de novo mutations in autism spectrum disorders. *Nature* **485**, 242–245 (2012).
43. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
44. Morgan, J. T., Barger, N., Amaral, D. G. & Schumann, C. M. Stereological study of amygdala glial populations in adolescents and adults with autism spectrum disorder. *PLoS One* **9**, e110356 (2014).
45. Gupta, S. *et al.* Transcriptome analysis reveals dysregulation of innate immune response genes and neuronal activity-dependent genes in autism. *Nat. Commun.* **5**, 5748 (2014).
46. DiStasio, M. M., Nagakura, I., Nadler, M. J. & Anderson, M. P. T lymphocytes and cytotoxic

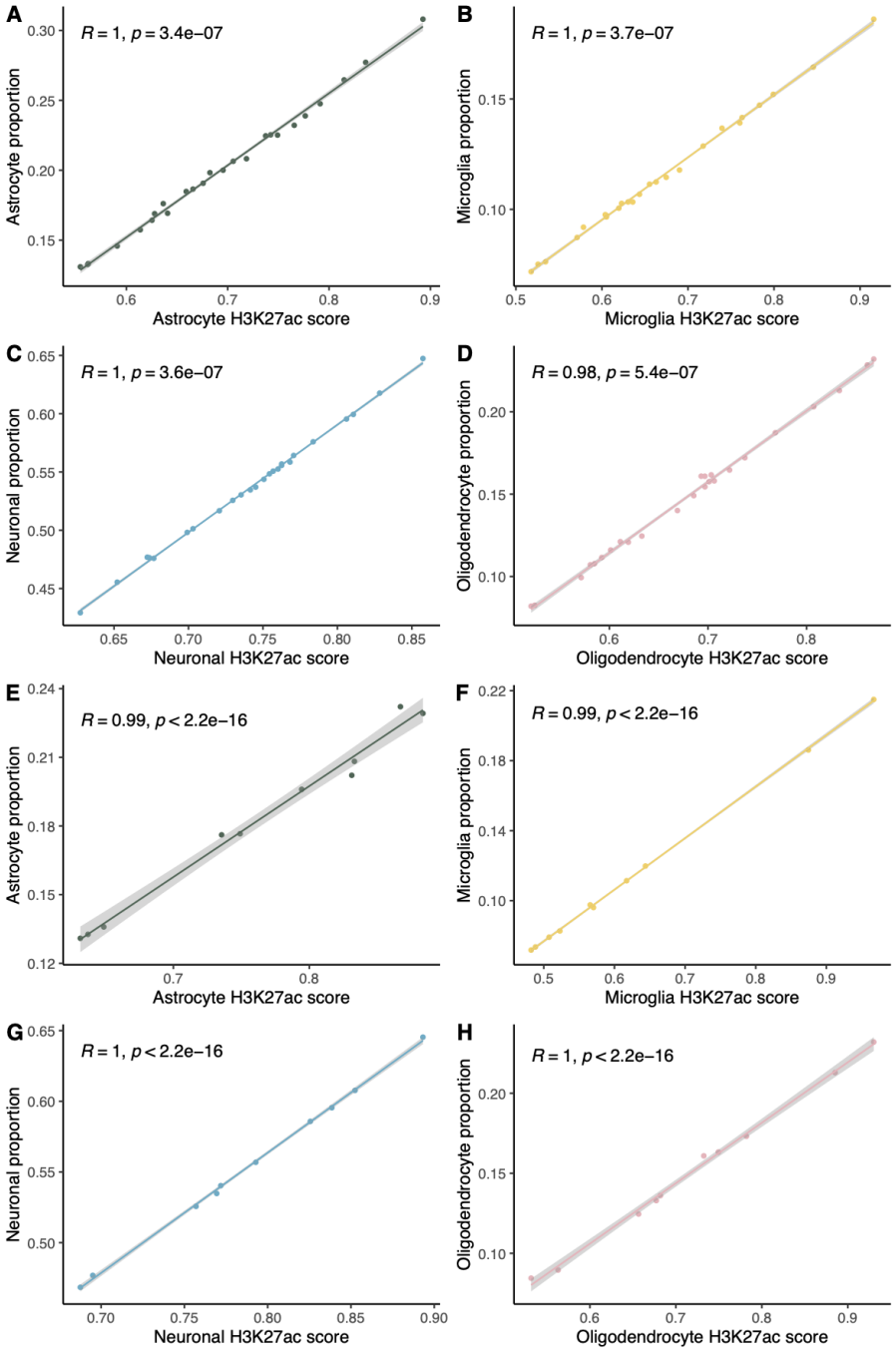
astrocyte blebs correlate across autism brains. *Ann. Neurol.* **86**, 885–898 (2019).

47. Datta, M. *et al.* Histone Deacetylases 1 and 2 Regulate Microglia Function during Development, Homeostasis, and Neurodegeneration in a Context-Dependent Manner. *Immunity* **48**, 514–529.e6 (2018).
48. Novikova, G. *et al.* Integration of Alzheimer's disease genetics and myeloid genomics identifies disease risk regulatory elements and genes. *Nat. Commun.* **12**, 1610 (2021).
49. Gonsalvez, D., Ferner, A. H., Peckham, H., Murray, S. S. & Xiao, J. The roles of extracellular related-kinases 1 and 2 signaling in CNS myelination. *Neuropharmacology* **110**, 586–593 (2016).
50. Ishii, A., Furusho, M., Macklin, W. & Bansal, R. Independent and cooperative roles of the Mek/ERK1/2-MAPK and PI3K/Akt/mTOR pathways during developmental myelination and in adulthood. *Glia* **67**, 1277–1295 (2019).
51. Chen, J.-F. *et al.* Enhancing myelin renewal reverses cognitive dysfunction in a murine model of Alzheimer's disease. *Neuron* **109**, 2292–2307.e5 (2021).
52. Iossifov, I. *et al.* De novo gene disruptions in children on the autistic spectrum. *Neuron* **74**, 285–299 (2012).
53. Mathys, H. *et al.* Temporal Tracking of Microglia Activation in Neurodegeneration at Single-Cell Resolution. *Cell Rep.* **21**, 366–380 (2017).
54. Sankowski, R. *et al.* Mapping microglia states in the human brain through the integration of high-dimensional techniques. *Nat. Neurosci.* **22**, 2098–2110 (2019).
55. Olah, M. *et al.* Single cell RNA sequencing of human microglia uncovers a subset associated with Alzheimer's disease. *Nature Communications* vol. 11 (2020).
56. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* vol. 169 1276–1290.e17 (2017).
57. Lawrence, M. *et al.* Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* **9**, e1003118 (2013).
58. Andrews, S. & Others. FastQC: a quality control tool for high throughput sequence data. (2010).
59. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).
60. Schneider, V. A. *et al.* Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Res.* **27**, 849–864 (2017).
61. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
62. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

Supplementary information

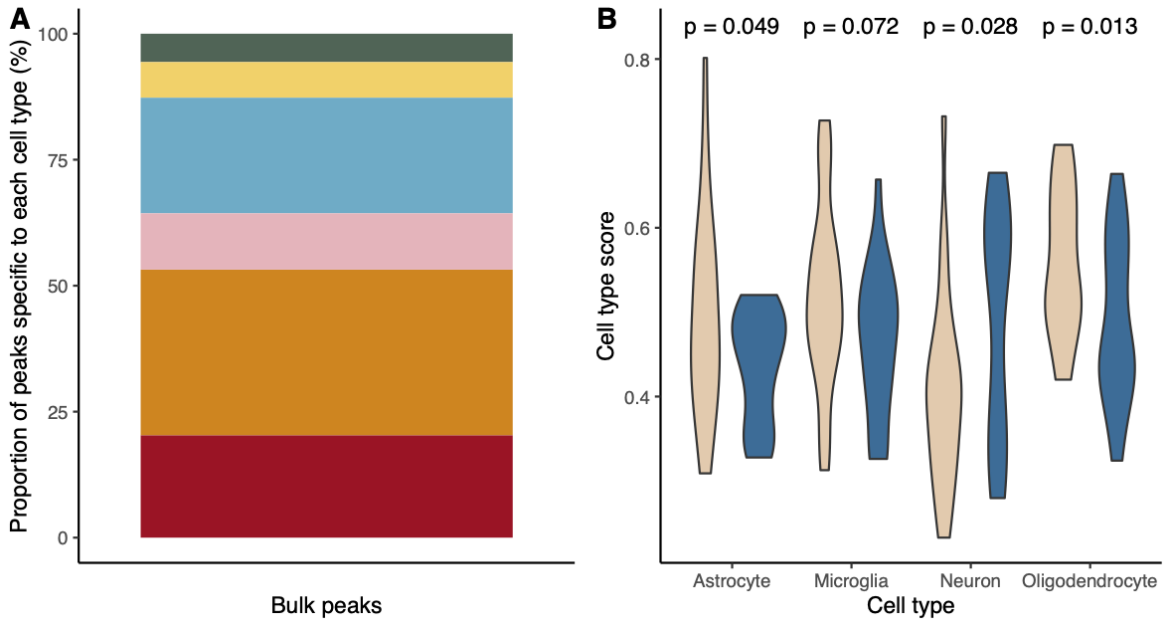


Supplementary Figure 1. CHAS robustly estimates cell type proportions in lower coverage datasets. Validation of CHAS using 49 pseudo-bulk samples made up of 20 million **(A-D)** and 10 million **(E-H)** randomly sampled reads from astrocytes, microglia, neurons, and oligodendrocytes. Shown are scatterplots of the CHAS-derived histone acetylation score for astrocytes (x-axis) vs. the true proportion of the cell type within the pseudo-bulk sample (y-axis) for astrocytes (**A** and **E**), microglia (**B** and **F**), neurons (**C** and **G**) and oligodendrocytes (**D** and **H**).

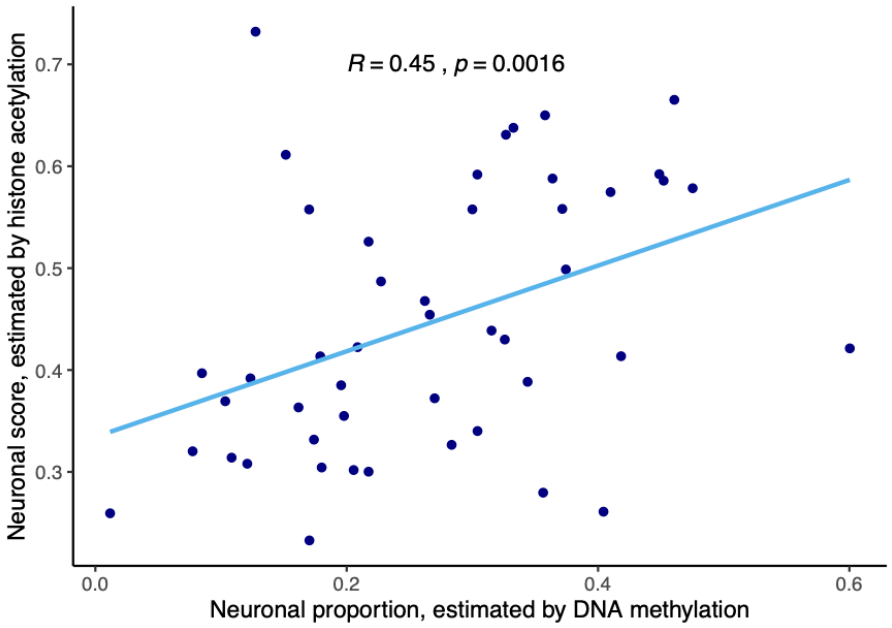


Supplementary Figure 2. CHAS consistently correlates with known cell type proportion even at smaller sample sizes.

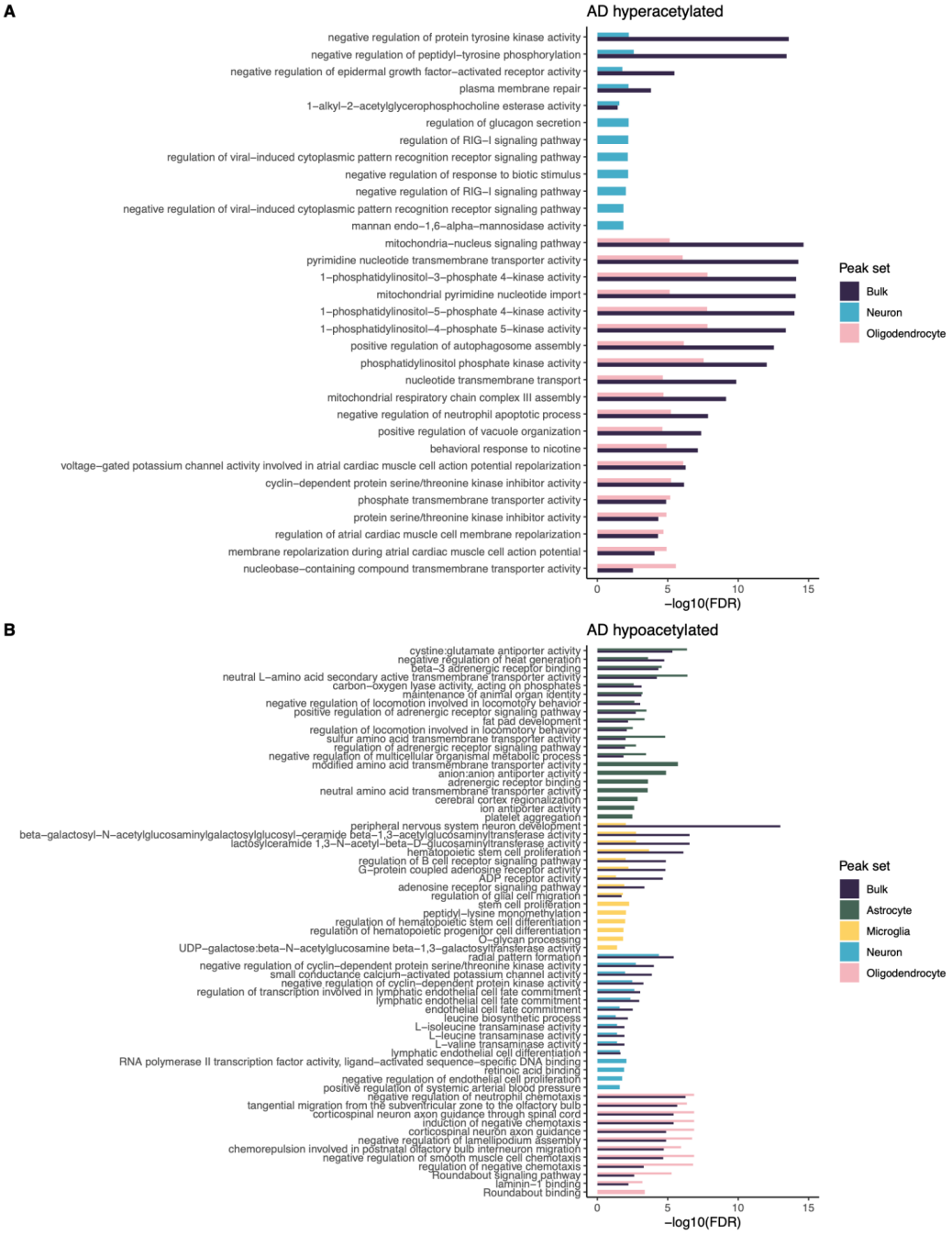
Validation of CHAS using 25 pseudo-bulk samples **(A-D)** and 10 pseudo-bulk samples **(E-H)** made up of 30 million randomly sampled reads from astrocytes, microglia, neurons, and oligodendrocytes. Shown are scatterplots of the CHAS-derived histone acetylation score for astrocytes (x-axis) vs. the true proportion of the cell type within the pseudo-bulk sample (y-axis) for astrocytes (**A** and **E**), microglia (**B** and **F**), neurons (**C** and **G**) and oligodendrocytes (**D** and **H**).



Supplementary Figure 3. Analysis of entorhinal cortex H3K27ac in AD cases and controls using CHAS. Cell type specific peaks within bulk H3K27ac and disease-associated shifts in cell type proportion are shown for[6]. **A)** Bar plot showing the proportion (%) of peaks specific to each cell type, annotated to more than one cell type ('multiple'), or annotated to none of the cell types ('other'). Out of 183,353 peaks, 80% (n = 146,144) were annotatable to one or more cell types, with 47% (n = 85,824) being cell type-specific. 6% (n = 10,303) peaks were specific to astrocytes, 7% (n = 12,993) were specific to microglia, 23% (n = 42,058) were specific to neurons, and 11% (n = 20,470) were specific to oligodendrocytes. **B)** Violin plot comparing the cell type score means between AD cases and controls. P values were calculated using Welch's two-sample t-test.

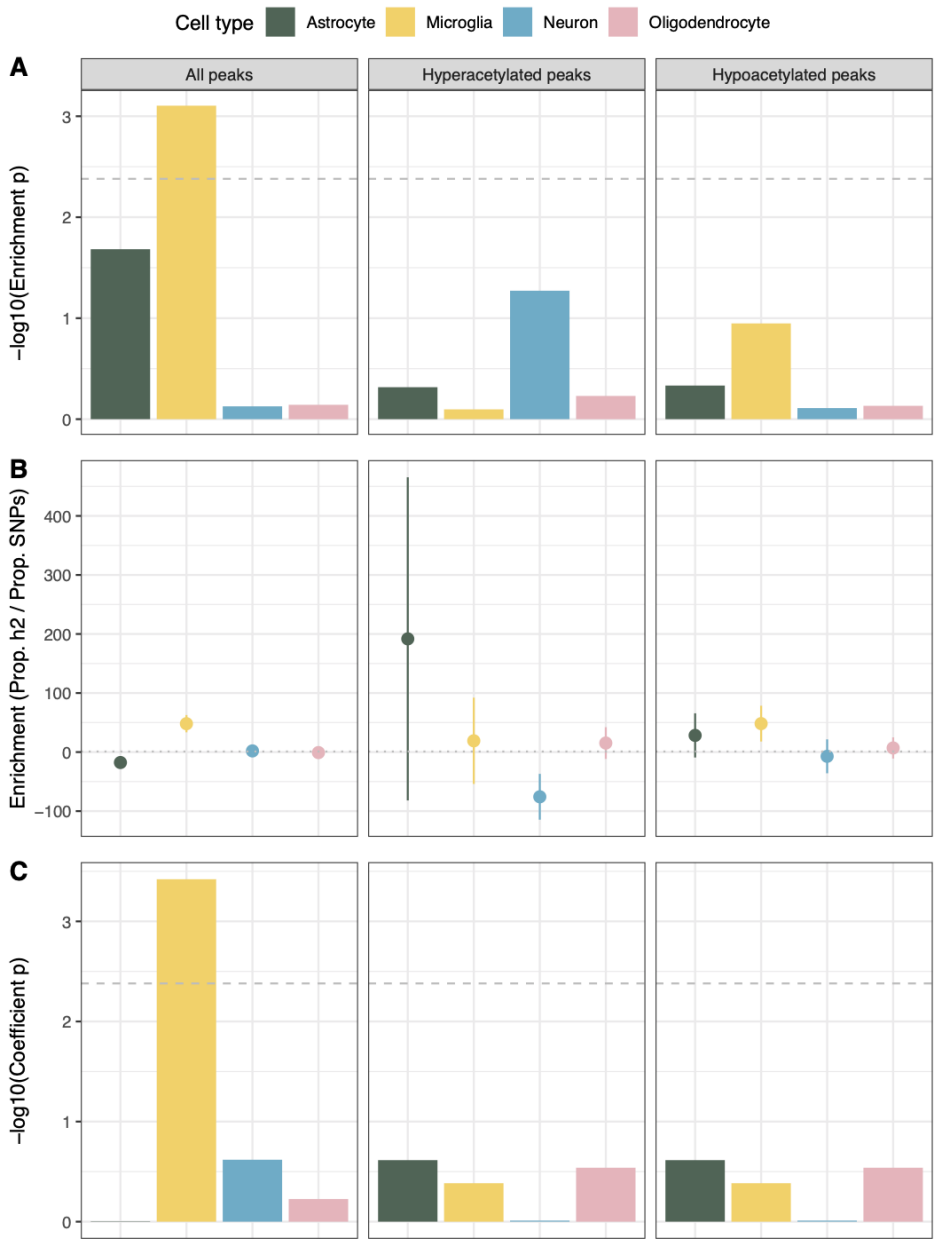


Supplementary Figure 4. CHAS-derived neuronal scores correlate with CETS-derived neuronal proportions. Validation of CHAS using neuronal proportion ($NeuN^+$ fraction) estimates derived using CETS[9] for the AD H3K27ac samples[6] analysed in this study. CHAS-derived neuronal scores and CETS-derived neuronal proportion estimates correlated across the 47 samples.

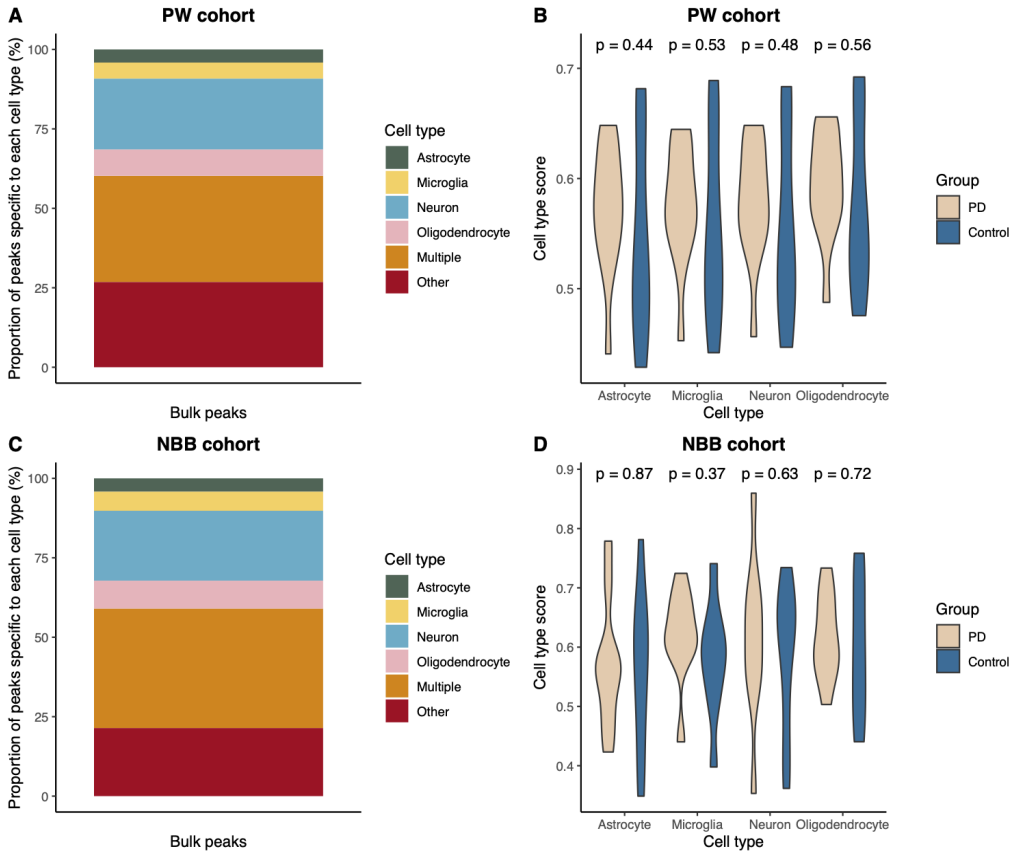


Supplementary Figure 5. **Functional enrichment analysis using AD-associated H3K27ac regions.**

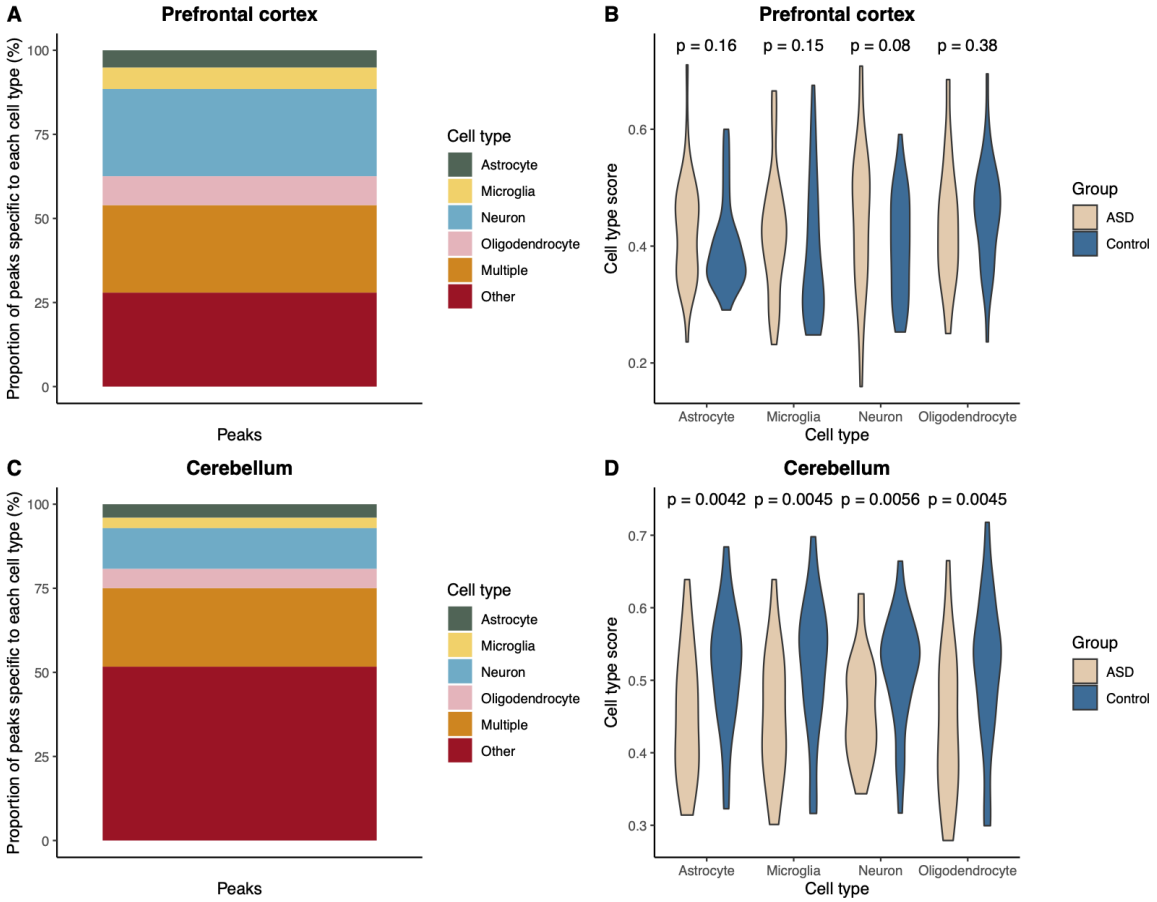
GREAT[24] pathway enrichment analysis using AD-associated bulk and cell type-specific **A** hyperacetylated peaks and **B** hypoacetylated peaks[6]. Shown are GO Biological Processes and GO Molecular Functions which were enriched within bulk and cell type-specific H3K27ac regions, shown are the overlapping pathways between bulk and cell type-specific peak sets, as well as the GO categories enriched only in the cell type-specific H3K27ac regions.



Supplementary Figure 6. AD genetic risk is enriched at microglial peaks detected within bulk brain H3K27ac. Results from partitioned heritability analysis using a GWAS for AD[33] with both disease associated and non-disease associated cell type-specific H3K27ac[6]. a) Enrichment P value from LDSC with Bonferroni significance threshold shown at -log10(P) = 2.4 (grey dashed line). b) Enrichment scores from by LDSC for each cell type-specific H3K27ac peak set in combination with the full baseline mode. The grey dotted line at 1 is the cutoff for enrichment and the error bars represent standard errors around the estimates of enrichment. c) Coefficient P values transformed from the coefficient z-score output by LDSC. The grey dashed line at -log10(P) = 2.4 is the cutoff for Bonferroni significance.

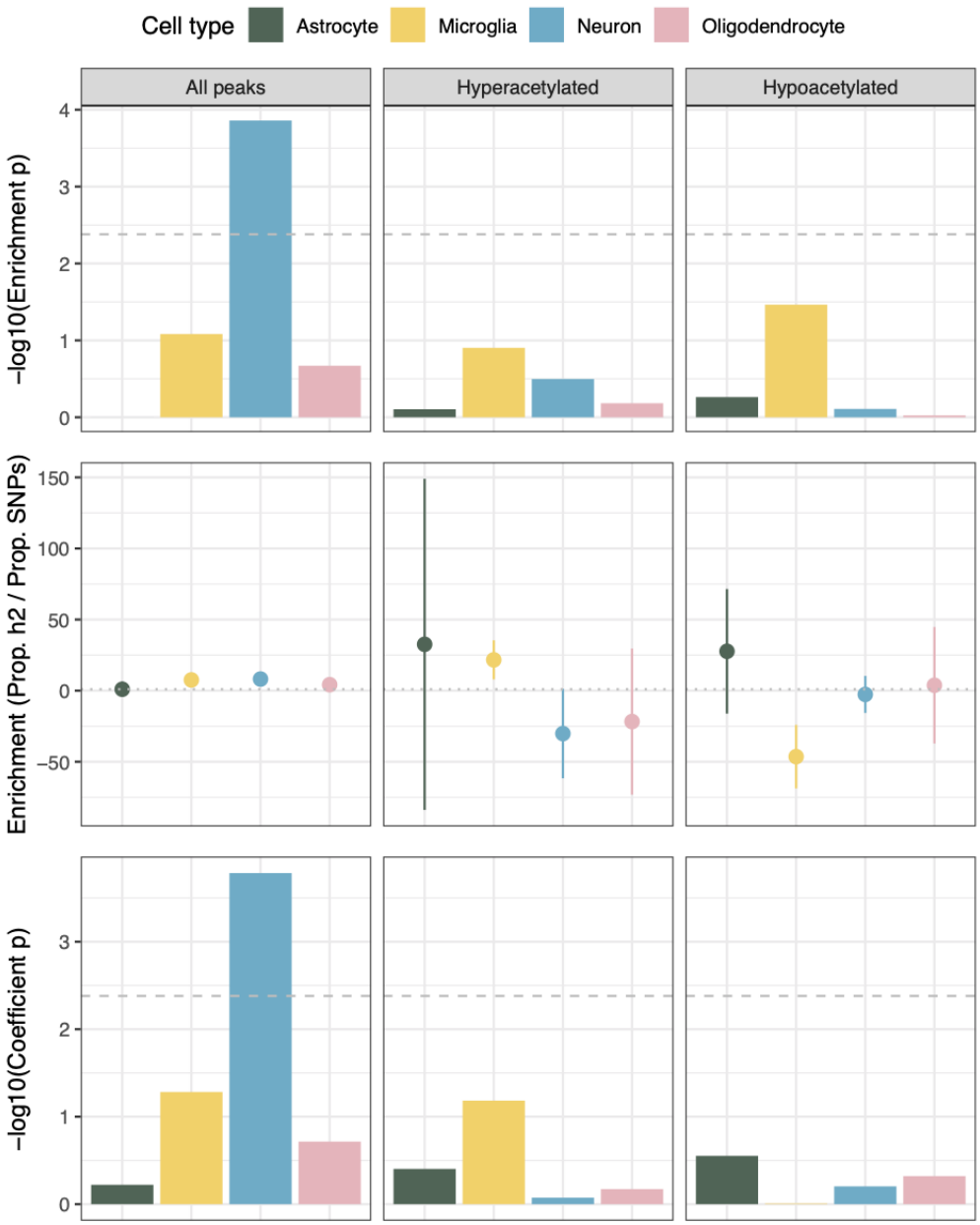


Supplementary Figure 7. Analysis of prefrontal cortex H3K27ac in PD cases and controls using CHAS. Analysis of prefrontal cortex H3K27ac in two independent cohorts of PD cases and controls using CHAS[63] **A** PW cohort: A bar plot showing the proportion (%) of peaks specific to each cell type, annotated to more than one cell type ('multiple'), and annotated to none of the cell types ('other'). In the PW cohort, the percentage of bulk peaks that were specific to a cell type was as follows: 5% were microglia-specific, 5% were astrocyte-specific, 7% were oligodendrocyte-specific, 22% were neuron-specific. The remaining bulk peaks were either annotated to more than one cell type ('multiple', 27%) or didn't overlap any of the cell type peaks and were annotated to 'other' (33%) **B** PW cohort: A violin plot comparing the cell type score means between PD cases and controls, calculated using Welch's two-sample t-test. **C** NBB cohort: A bar plot showing the proportion (%) of peaks specific to each cell type, annotated to more than one cell type ('multiple'), and annotated to none of the cell types ('other'). 4% were astrocyte-specific (n = 4,673), 6% were microglia-specific (n = 6,692), 9% were oligodendrocyte-specific (n = 9,772), 22% were neuron-specific (n = 24,507), 33% were annotated to 'multiple' cell types (n = 41,906), and 26% weren't annotatable ('other', n = 23,846). **D** NBB cohort: A violin plot comparing the cell type score means between PD cases and controls, calculated using Welch's two-sample t-test.



Supplementary Figure 8. Analysis of prefrontal cortex and cerebellar H3K27ac in ASD cases and controls. Analysis of prefrontal cortex, temporal cortex, and cerebellar H3K27ac in ASD cases and controls using CHAS[5]**. A** A bar plot showing the proportion (%) of peaks specific to each cell type, annotated to more than one cell type ('multiple'), and annotated to none of the cell types ('other') in prefrontal cortex. 5% (n = 12,921) of peaks were astrocyte-specific, 6.4% (n = 15,945) were microglia-specific, 26% (n = 65,017) were neuron-specific, and 9% (n = 21,514) were oligodendrocyte-specific. 26% (n = 65,224) of peaks were annotated to multiple cell types, and 28% (n = 69,993) weren't annotatable to any of the cell types. **B** A violin plot comparing the prefrontal cortex cell type score means between ASD cases and controls, calculated using Welch's two-sample t-test. **C** A bar plot showing the proportion (%) of peaks specific to each cell type, annotated to more than one cell type ('multiple'), and annotated to none of the cell types ('other') in cerebellum. 4% (n = 9,740) of peaks were astrocyte-specific, 3% (n = 7,484) were microglia-specific, 12% (n = 29,312) were neuron-specific, and 6% (n = 13,793) were oligodendrocyte-specific. 23% (n = 56,483) of peaks were annotated to multiple cell types, and 52% (n = 124,947) weren't annotatable to any of the cell types. **D** A violin plot comparing the cerebellar cell type score means between ASD cases and controls, calculated using Welch's two-sample t-test.

Supplementary Figure 9. **Functional enrichment analysis using ASD-associated H3K27ac regions.** GREAT [24] pathway enrichment analysis using ASD-associated bulk and cell type-specific **A** hyperacetylated peaks and **B** hypoacetylated peaks. Shown are GO Biological Processes and GO Molecular Functions which were enriched within bulk and cell type-specific H3K27ac regions, shown are the overlapping pathways between bulk and cell type-specific peak sets, as well as the GO categories enriched only in the cell type-specific H3K27ac regions.



Supplementary Figure 10. Enrichment of cell type-specific H3K27ac peaks for ASD risk variants. Results from partitioned heritability analysis using a GWAS for ASD[34] with both disease associated and non-disease associated cell type-specific H3K27ac. a) The y-axis represents the enrichment p value output by LDSC. The grey dashed line at -log10(P) = 2.4 is the cutoff for Bonferroni significance. b) The y-axis represents the enrichment score output by LDSC for each cell type-specific H3K27ac peak set in combination with the full baseline mode. The grey dotted line at 1 is the cutoff for enrichment and the error bars represent standard errors around the estimates of enrichment. c) The y-axis represents the coefficient p value transformed from the coefficient z-score output by LDSC. The grey dashed line at -log10(P) = 2.4 is the cutoff for Bonferroni significance.