Cureus
Part of SPRINGER NATURE

Open Access Original Article

Actin-Related Protein 2/3 (Arp2/3)-Associated Gene Expression in Repetitive Head Impacts and Chronic Traumatic Encephalopathy: Analysis of Two Human Transcriptomic Datasets

Review began 10/05/2026
Review ended 10/19/2026
Published 10/26/2026

© **Copyright** 2026
Lehrer. This is an open access article distributed under the terms of the Creative Commons Attribution License CC-BY 4.0., which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

DOI: 10.7759/cureus.118619

Steven Lehrer [1]

1. Radiation Oncology, Icahn School of Medicine at Mount Sinai, New York, USA

Corresponding author: Steven Lehrer, stevenlehrer@hotmail.com

Abstract

Background: We evaluated actin-related protein 2/3 (Arp2/3)-associated and SOX9/MEGF10 gene-expression modules in repetitive head impacts (RHI) and chronic traumatic encephalopathy (CTE).

Methods: Public bulk RNA-sequencing data (GSE193407; 136 brains) and single-nucleus RNA-sequencing data (GSE261807; 28 donors) were analyzed. Sensitivity analyses examined neuronal and astrocytic marker adjustment, cell-type selection, and alternative Arp2/3 module definitions.

Results: In bulk tissue, nominal CTE-stage associations with Arp2/3 ($P = 0.0151$) and SOX9/MEGF10 ($P < 0.001$) were attenuated after adjustment for neuronal and astrocytic markers ($P = 0.2751$ and $P = 0.6899$, respectively). In exploratory single-nucleus analyses, the six-cell-type omnibus comparison was nonsignificant ($P = 0.0617$), and no individual cell-type association survived false discovery rate correction. The post hoc three-cell-type composite did not establish a CTE-versus-RHI difference (95% CI, −0.532 to 0.539). Single-nucleus analyses were unadjusted for donor-level covariates; cell-type annotations were not independently validated, and donor overlap between datasets could not be excluded.

Conclusions: The principal sensitivity analyses did not establish robust CTE-specific expression differences. These predominantly null findings are descriptive and hypothesis-generating.

Categories: Neurology, Sports Medicine
Keywords: arp2/3 complex, chronic traumatic encephalopathy (cte), gene expression, neurodegeneration, repetitive head impacts

Introduction

Chronic traumatic encephalopathy (CTE) is a neurodegenerative tauopathy associated with repetitive head impacts (RHI), including those sustained during contact sports and other activities [1,2]. Neuropathologically, CTE is characterized by abnormal perivascular accumulation of hyperphosphorylated tau in neurons at the depths of cortical sulci [1,2]. Neuropathologic staging reflects the distribution of tau pathology but does not itself establish a longitudinal molecular sequence [2-4].

Transcriptomic studies have identified neuronal, glial, and inflammatory differences in postmortem CTE brain. Labadorf et al. examined bulk RNA-sequencing data across CTE stages and reported differences in neuronal and inflammatory gene expression. Butler et al. analyzed single-nucleus RNA-sequencing data from controls, individuals with RHI without CTE, and donors with CTE, identifying changes in multiple cellular populations [5,6]. Additional studies have described altered signaling, microglial responses, and astroglial or oligodendroglial transcriptional characteristics [7-13]. Exposure history and neuropathologic severity are not interchangeable; age at first tackle-football exposure was not associated with greater CTE pathologic severity in one autopsy cohort [10-12,14].

The actin-related protein 2/3 (Arp2/3) complex participates in actin nucleation and has recognized roles in neuronal morphology and development [10]. SOX9 contributes to glial lineage specification, whereas MEGF10 participates in astrocyte-mediated synapse elimination [11,12]. These established functions motivate descriptive examination of their transcript abundance but do not imply that changes in expression demonstrate altered protein activity or developmental program reactivation.

We evaluated associations of neuropathologic CTE stage with a nine-gene Arp2/3 module (ACTR2, ACTR3, ARPC1A, ARPC1B, ARPC2, ARPC3, ARPC4, ARPC5, and ARPC5L) and a two-gene SOX9/MEGF10 module in GSE193407. We also analyzed donor-level expression in GSE261807 across control, RHI, and CTE groups. The primary emphasis is on sensitivity to neuronal and astrocytic marker expression; reconstructed cell-type analyses are supplementary and exploratory. The study was not designed to establish exposure

causality, cell-intrinsic function, or developmental reactivation.

Materials And Methods

Data sources and study design

This secondary, cross-sectional analysis used publicly available, de-identified data from National Center for Biotechnology Information (NCBI) Gene Expression Omnibus (GEO) accessions GSE193407 and GSE261807. GSE193407 comprised 136 postmortem brains: 58 stage 0, 4 stage I, 13 stage II, 31 stage III, and 30 stage IV. GSE261807 comprised 28 donors: 8 controls, 9 with RHI without CTE, and 11 with CTE. The datasets arose from related research programs; independent donor membership could not be verified.

GSE193407 preprocessing and module scores

Bulk counts and metadata were obtained from GEO. Ensembl identifiers were stripped of version suffixes and mapped to gene symbols with org.Hs.eg.db and AnnotationDbi. Gene-level DESeq2 analyses filtered genes with at least 10 counts in at least 10 samples and used the design ~ RIN (RNA integrity number) + AGE + CTE_STAGE. Genome-wide and selected-target Benjamini-Hochberg false-discovery rates (FDRs) were evaluated. The 13 selected transcripts comprised the nine Arp2/3 genes, MYH10, SOX9, MEGF10, and KIF5C.

Module-level analyses used standardized gene-expression values and equally weighted means within each module. The nine-gene Arp2/3 module excluded MYH10; the SOX9/MEGF10 module comprised those two genes. The SOX9/MEGF10-minus-Arp2/3 score was a derived expression contrast, not a validated biological pathway score. The primary age- and RIN-adjusted linear regression coded neuropathologic CTE stage numerically from 0 to 4. Additional analyses excluded the four stage I cases or compared stage 0 with stages III-IV. Spearman correlation was used for descriptive nonparametric comparison.

Bulk marker-adjustment sensitivity analysis

For marker-adjustment sensitivity analyses, bulk counts were mapped to gene symbols, TMM (trimmed mean of M-values)-normalized using edgeR, and converted to log2 counts per million with a prior count of 1. Gene values were standardized across samples. Neuronal marker scores averaged z scores for SLC17A7, CAMK2A, SATB2, CUX2, GAD1, GAD2, SLC6A1, and DLX1; astrocytic scores averaged AQP4, ALDH1L1, SLC1A3, SLC1A2, GJA1, and glial fibrillary acidic protein (GFAP). Regression models included CTE stage, age, and RIN, with or without both marker scores. All 136 samples were included. Marker scores are indirect proxies for cellular transcriptional characteristics, not direct cell-count estimates. The reported module coefficients refer to this sensitivity-analysis normalization and should not be assumed to be numerically interchangeable with all DESeq2 gene-level estimates.

GSE261807 whole-donor analysis

Filtered donor-specific 10x count matrices were processed separately, preserving the donor as the observational unit. Raw counts were summed across nuclei per donor, converted to counts per million using total donor UMIs, transformed as log2(CPM + 1), and standardized gene by gene across 28 donors. Arp2/3 and SOX9/MEGF10 modules were calculated as equally weighted means. An exploratory ordered-group regression coded control = 0, RHI = 1, and CTE = 2. This ordering is a statistical contrast, not a longitudinal disease trajectory. Gene-level FDR was calculated across the selected genes. Count-based donor-level pseudobulk sensitivity analyses used edgeR negative binomial quasi-likelihood models and limma-voom fry gene-set testing.

Covariate availability and dataset overlap

The GSE261807 regression analyses were unadjusted for donor age, sex, postmortem interval (PMI), and RNA integrity number (RIN). A complete, verified mapping of these variables to the analyzed expression matrices was not available for the present analysis; therefore, adjusted or matched analyses were not performed. All GSE261807 P-values are descriptive and exploratory.

Comparison of available donor ages and CTE stages identified 10 of 28 GSE261807 clinical donors with at least one candidate match in GSE193407; several matches were nonunique. Clinical candidate matches could not be linked reliably to individual expression-matrix identifiers. Excluding 10 arbitrary expression profiles would not be a valid donor-overlap sensitivity analysis; therefore, no such exclusion was performed, and independence between datasets remains unverified.

Reconstructed cell-type analyses: supplementary status

GSE261807 cell-type labels were independently reconstructed using marker expression rather than a verified original-author cell annotation object. The classifications were not independently validated against an external reference atlas. Accordingly, all cell-type-specific analyses, three- and six-cell-type composites, leave-one-cell-type-out analyses, and the eight-gene ARPC1B-exclusion analysis are designated as exploratory supplementary analyses. Detailed definitions and classification limitations are provided in the

"Supplementary methods" section, while the estimates are reported in the "Results" section.

Statistical analysis

All analyses were conducted in R (reported source environment version 4.5.2; R Foundation for Statistical Computing, Vienna, Austria), using DESeq2, edgeR, limma, data.table, Matrix, AnnotationDbi, and org.Hs.eg.db. Regression coefficients, 95% confidence intervals, and two-sided nominal P-values are reported. Multiple-testing correction is distinguished from nominal testing; exploratory sensitivity tests were not treated as independent replication.

Results

Bulk RNA-sequencing: initial associations and marker adjustment

In GSE193407 (n=136), the age- and RIN-adjusted Arp2/3 module association with CTE stage was β = −0.0790 (standard error of the mean (SE) = 0.0321, 95% CI, −0.1425 to −0.0155, P = 0.0151). The corresponding SOX9/MEGF10 coefficient was β = 0.1876 (SE = 0.0463, 95% CI, 0.0961 to 0.2791, P = 8.53×10^{-5}). After additional adjustment for neuronal and astrocytic marker scores, the Arp2/3 coefficient was β = 0.0242 (SE = 0.0221, 95% CI, −0.0195 to 0.0680, P = 0.2751), and the SOX9/MEGF10 coefficient was β = 0.0068 (SE = 0.0171, 95% CI, −0.0270 to 0.0407, P = 0.6899). Neither marker-adjusted association remained statistically significant (Table *1*, Figure *1*).

Analysis	β	95% CI	P-value
GSE193407 Arp2/3, all stages	−0.079	−0.143 to −0.016	0.0151
GSE193407 SOX9/MEGF10, all stages	0.188	0.096 to 0.279	8.53×10^{-5}
GSE193407 Arp2/3, stage I excluded	−0.080	−0.141 to −0.019	0.0107
GSE193407 SOX9/MEGF10, stage I excluded	0.188	0.099 to 0.277	5.34×10^{-5}
GSE261807 donor Arp2/3 ordered trend	−0.316	−0.613 to −0.020	0.0375
GSE261807 SOX9/MEGF10 ordered trend	0.043	−0.416 to 0.502	0.849
Neuronal/oligodendroglial ordered trend	−0.314	−0.594 to −0.033	0.0298
Composite: RHI vs control	−0.665	−1.244 to −0.087	0.026
Composite: CTE vs control	−0.662	−1.215 to −0.108	0.021
Composite: CTE vs RHI	0.0034	−0.532 to 0.539	0.9896

TABLE 1: Principal module-level associations in GSE193407 and GSE261807

Regression coefficients (β), 95% confidence intervals (CIs), and two-sided P-values are shown for the primary and sensitivity analyses. GSE193407 models were adjusted for age and RNA integrity number. GSE261807 ordered-trend models encoded control = 0, repetitive head impact (RHI) = 1, and chronic traumatic encephalopathy (CTE) = 2. The neuronal/oligodendroglial composite analysis was exploratory.

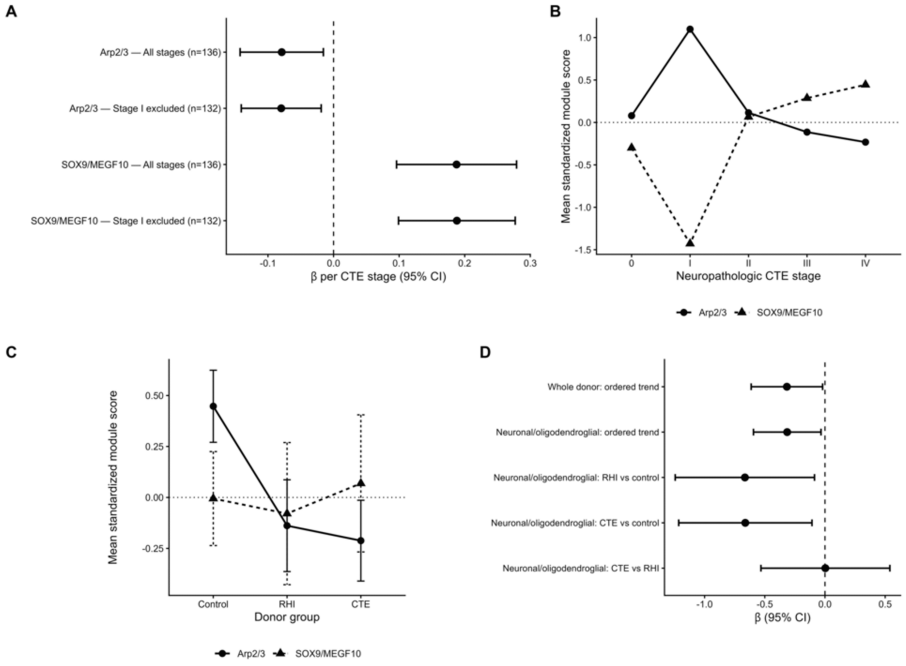


FIGURE 1: Arp2/3 and SOX9/MEGF10 gene-expression associations in two human CTE transcriptomic datasets

(A) Adjusted associations of neuropathologic chronic traumatic encephalopathy (CTE) stage with Arp2/3 and SOX9/MEGF10 module scores in GSE193407 bulk RNA sequencing. Primary analyses included all 136 subjects; sensitivity analyses excluded the four stage I cases (n = 132). Estimates are regression coefficients per one-stage increase in CTE stage with 95% confidence intervals (CIs), adjusted for age and RNA integrity number (RIN).

(B) Mean standardized Arp2/3 and SOX9/MEGF10 module scores across neuropathologic CTE stages in GSE193407. Stage I contained four subjects and should be interpreted cautiously.

(C) Mean standardized Arp2/3 and SOX9/MEGF10 module scores in GSE261807 single-nucleus RNA sequencing across 28 donors: 8 controls, 9 with repetitive head impacts (RHI) without CTE, and 11 with CTE. Error bars represent ±1 standard error of the mean (SE). Donor-level Arp2/3 scores were lower in the RHI and CTE groups than those in controls, whereas SOX9/MEGF10 did not show a significant ordered trend.

(D) Arp2/3 regression effect estimates with 95% CIs for GSE261807, including the whole-donor ordered control-RHI-CTE association and the exploratory neuronal/oligodendroglial composite ordered and pairwise comparisons.

Module scores represent the mean of standardized (z-scored) expression values of their constituent genes. GSE261807 broad cell classes were independently assigned using marker-gene expression because the investigators' final cell-level annotation object was not publicly deposited.

This figure was generated in R version 4.5.2 (R Foundation for Statistical Computing, Vienna, Austria) using ggplot2 version 4.0.3 and assembled using patchwork version 1.3.2.

Bulk individual-gene associations

In age- and RIN-adjusted DESeq2 models, ACTR2 (log2FC −0.0534; genome-wide FDR 0.0373), ARPC2 (−0.0500; FDR 0.00690), ARPC4 (−0.0301; FDR 0.0285), and ARPC5L (−0.0746; FDR 0.00795) had negative associations with CTE stage. MYH10 was negatively associated (−0.1027; FDR 0.00250), and MEGF10 was positively associated (0.1476; FDR 0.00730). SOX9 had a nominal positive coefficient (0.1236; P = 0.0118) but did not meet genome-wide FDR < 0.05 (FDR 0.0518). These gene-level results are descriptive and do not distinguish altered cell composition from within-cell expression differences.

Whole-donor GSE261807 results

All 13 selected genes were detected in the 28 GSE261807 donors. The mean standardized nine-gene Arp2/3 scores were 0.447 in controls, −0.138 in RHI, and −0.212 in CTE. The exploratory ordered control-RHI-CTE coefficient was β = −0.316 (SE = 0.144; 95% CI, −0.613 to −0.020; P = 0.0375). SOX9/MEGF10 showed no ordered association (β = 0.043; 95% CI, −0.416 to 0.502; P = 0.849), and the SOX9/MEGF10-minus-Arp2/3 score was also nonsignificant (β = 0.359; P = 0.265).

Seven of nine Arp2/3 gene coefficients were negative (one-sided exact binomial P = 0.0898). ARPC5 had an ordered coefficient of −0.284 (P = 0.0030; FDR across 13 selected genes = 0.039). In donor-level count-based pseudobulk testing, the nine-gene set yielded P = 0.0374 for the combined RHI/CTE-versus-control contrast, but the result did not survive correction across three tested gene sets (FDR = 0.0628). The CTE-versus-RHI gene-set contrast was nonsignificant (P = 0.8932). No causal exposure interpretation follows from these comparisons.

Exploratory reconstructed-cell-type analyses

The exploratory reconstructed-cell-type findings are summarized below, while the classification methods and their limitations are described in the "Supplementary methods" section. No individual cell-type association survived FDR correction. In the expanded six-cell-type composite, the ordered association was nominal (β = −0.2697, P = 0.0295), but the categorical omnibus test was nonsignificant (P = 0.0617). Excluding ARPC1B did not change the direction (ordered β = −0.2878, P = 0.0239); its six-cell-type omnibus test also remained nonsignificant (P = 0.0517). These unadjusted, unvalidated exploratory analyses do not establish a cell-intrinsic or CTE-specific association.

Discussion

The main finding of this descriptive study is the substantial attenuation of the initial bulk RNA-sequencing associations after adjustment for neuronal and astrocytic marker expression. Neither the Arp2/3 nor the SOX9/MEGF10 module remained significantly associated with CTE stage in the marker-adjusted models. Marker scores may reflect both cellular representation and cell state; their inclusion does not prove that compositional confounding alone explains the initial associations. Nevertheless, the attenuation limits claims that the expression differences are independent of neuronal and astrocytic transcriptional characteristics.

The single-nucleus whole-donor analysis yielded a nominal ordered Arp2/3 association, but the analysis was small, unadjusted for age, sex, PMI, and RIN, and could not be established as independent of the bulk cohort. The supplementary cell-type analyses did not survive FDR correction at the individual cell-type level, and the expanded six-cell-type categorical omnibus test was nonsignificant. The outcome-informed three-cell-type composite cannot be treated as independent validation.

The CTE-versus-RHI comparison in the post hoc three-cell-type composite was β = 0.0034 (95% CI, −0.532 to 0.539; P = 0.9896). The interval encompasses potentially meaningful differences in either direction; a nonsignificant contrast does not establish equivalence. No equivalence margin was prespecified, and no formal equivalence test was performed. The cross-sectional group differences also cannot be attributed specifically to head-impact exposure or CTE progression.

The Arp2/3 complex has established functions in actin regulation, while SOX9 and MEGF10 have established glial functions [10-12]. These functions do not convert associations in transcript abundance into evidence of altered protein activity, cytoskeletal organization, or developmental-program reactivation. Such mechanisms were not tested in this study.

Taken together, the data are best understood as exploratory differences in selected expression measures that depend on analytical choices and the available covariates. The bulk marker-adjustment null results and the nonsignificant expanded cell-type omnibus results should take precedence over nominally significant exploratory contrasts.

Limitations

First, this was a cross-sectional observational analysis. Control, RHI, and CTE are distinct donor groups, not successive observations of the same individuals. Ordinal coding does not establish a temporal sequence.

Second, bulk module associations were attenuated after adjustment for neuronal and astrocytic markers. These marker scores are imperfect proxies for composition and may reflect transcriptional state, so attenuation does not uniquely identify its explanation.

Third, GSE261807 included only 28 donors (8 controls, 9 RHI, and 11 CTE). Analyses were unadjusted for age, sex, PMI, and RIN because a complete verified donor-to-expression crosswalk was unavailable. Small group sizes limit precision and statistical power.

Fourth, cell-type annotations were reconstructed using marker expression and were not independently validated against an external atlas or the original investigators' final labels. Detailed cell-type findings are therefore supplementary. The three-cell-type composite was selected after examining the data, and the eight-gene sensitivity analysis was introduced during revision; neither was prespecified confirmatory evidence.

Fifth, the direct CTE-versus-RHI confidence interval for the post hoc composite (−0.532 to 0.539) does not

establish equivalence. No equivalence test was prespecified.

Sixth, 10 of 28 GSE261807 clinical donors had candidate age/stage matches in GSE193407, but these matches were not linked reliably to expression samples. An exclusion sensitivity analysis could not be performed; independent replication has not been demonstrated.

Seventh, the stage distribution in GSE193407 was uneven, particularly at stage I (n = 4), and stage-specific mean expression was not monotonic. Unmeasured comorbidities, exposure histories, tissue quality, and agonal factors may affect the observations.

Finally, transcript abundance cannot determine Arp2/3 protein abundance, complex activity, actin dynamics, or other cellular functions. No developmental reference transcriptomes or functional experiments were analyzed. Generalizability to other brain regions and exposure populations remains unknown.

Conclusions

In two publicly available human CTE transcriptomic datasets, initially observed associations involving Arp2/3 and SOX9/MEGF10 expression were not consistently supported by sensitivity analyses. Bulk-tissue associations became nonsignificant after neuronal and astrocytic marker adjustment. Single-nucleus analyses were exploratory and limited by small sample size, absent donor-level covariate adjustment, unvalidated reconstructed cell-type annotations, and possible donor overlap. The findings did not establish CTE-specific expression alterations, equivalence between RHI and CTE, or effects attributable to repetitive head-impact exposure. These predominantly null results are descriptive and hypothesis-generating.

Appendices

Code availability

The R source code used for reconstruction of GSE261807 donor-by-cell-type nucleus counts, marker-based cell-type classification, quality-control assessment, and verification of the 20-nucleus pseudobulk inclusion threshold is provided below:

```
# GSE261807 donor-by-cell-type QC reconstruction
# Bracket-free version of Supplementary Appendix 1
# Requires: Matrix, data.table
# Does not reproduce downstream regression models.

suppressPackageStartupMessages({
 library(Matrix)
 library(data.table)
})

root <- "D:/CTE/GSE261807"
input <- file.path(root, "filtered")
out <- file.path(root, "reviewer7_QC")
dir.create(out, recursive = TRUE, showWarnings = FALSE)
stopifnot(dir.exists(input))

markers <- list(
 Excitatory = c("SLC17A7", "SATB2", "CUX1", "CUX2", "RORB"),
 Inhibitory = c("GAD1", "GAD2", "SLC6A1"),
 Astrocyte = c("AQP4", "GFAP", "ALDH1L1", "SLC1A3", "SLC1A2"),
 Oligodendrocyte = c("MBP", "MOG", "PLP1", "MAG", "MOBP"),
 OPC = c("PDGFRA", "CSPG4", "VCAN"),
 Microglia = c("P2RY12", "CX3CR1", "AIF1", "TYROBP", "C1QA"),
 Endothelial = c("CLDN5", "FLT1", "PECAM1", "VWF")
)
classes <- c(names(markers), "Unclassified")

files <- sort(list.files(input, pattern = "_matrix\\.mtx\\.gz$", full.names = TRUE))
files <- Filter(function(f) !grepl("_raw_", basename(f)), files)
if (length(files) != 28L) stop("Expected 28 matrices; found ", length(files))

process_donor <- function(f) {
 prefix <- sub("_matrix\\.mtx\\.gz$", "", basename(f))
 ff <- file.path(input, paste0(prefix, "_features.tsv.gz"))
 if (!file.exists(ff)) stop("Missing features: ", ff)
```

```
 features <- as.data.frame(fread(ff, header = FALSE))
 symbol_column <- if (ncol(features) >= 2L) 2L else 1L
 symbols <- as.character(getElement(features, symbol_column))

 con <- gzfile(f, open = "rt")
 x <- tryCatch(as(Matrix::readMM(con), "dgCMatrix"), finally = close(con))
 if (nrow(x) != length(symbols)) stop("Feature mismatch: ", prefix)

 n_original <- ncol(x)
 original_lib <- Matrix::colSums(x)
 retained <- which(original_lib > 0)
 if (!length(retained)) stop("No nonzero nuclei: ", prefix)

 # Sparse selection matrix keeps only nonzero-UMI columns.
 column_selector <- Matrix::sparseMatrix(
  i = retained,
  j = seq_along(retained),
  x = 1,
  dims = c(n_original, length(retained))
 )
 x <- as(x %*% column_selector, "dgCMatrix")
 lib <- as.numeric(Matrix::colSums(x))

 score <- lapply(markers, function(genes) {
  ix <- which(symbols %in% genes)
  if (!length(ix)) return(rep(0, length(lib)))

  # Sparse row selector replaces matrix indexing.
  row_selector <- Matrix::sparseMatrix(
   i = seq_along(ix),
   j = ix,
   x = 1,
   dims = c(length(ix), nrow(x))
  )
  xx <- as(row_selector %*% x, "dgCMatrix")
  norm <- sweep(xx, 2, lib, "/") * 10000
  norm@x <- log1p(norm@x)
  as.numeric(Matrix::colMeans(norm))
 })

 sm <- do.call(cbind, score)
 colnames(sm) <- names(markers)
 best <- max.col(sm, ties.method = "first")
 celltype <- vapply(best, function(k) getElement(colnames(sm), k), character(1))
 celltype <- replace(celltype, which(apply(sm, 1, max) <= 0), "Unclassified")

 group <- if (grepl("_Control$", prefix)) "Control" else
  if (grepl("_RHI$", prefix)) "RHI" else
   if (grepl("_CTE$", prefix)) "CTE" else NA_character_
 if (is.na(group)) stop("Unknown group: ", prefix)

 tab <- table(factor(celltype, levels = classes))
 ans <- data.frame(
  GSM = sub("_.*$", "", prefix),
  Sample = prefix,
  Group = group,
  CellType = names(tab),
  N_nuclei = as.integer(tab),
  Original_nuclei = as.integer(n_original),
  Nonzero_nuclei = as.integer(length(lib)),
  Zero_UMI_excluded = as.integer(n_original - length(lib)),
  stringsAsFactors = FALSE
 )
 transform(ans, QC20 = N_nuclei >= 20L)
}

qc <- do.call(rbind, lapply(seq_along(files), function(i) {
 message("Processing donor ", i, "/", length(files))
```

```
  process_donor(getElement(files, i))
}))
qc <- as.data.frame(data.table::setorder(as.data.table(qc), Group, GSM, CellType))
if (length(unique(qc$GSM)) != 28L || nrow(qc) != 224L) {
  stop("Donor grid mismatch")
}

assigned <- aggregate(N_nuclei ~ GSM, data = qc, FUN = sum)
nonzero <- aggregate(Nonzero_nuclei ~ GSM, data = qc, FUN = function(v) unique(v))
check <- merge(assigned, nonzero, by = "GSM")
if (any(check$N_nuclei != check$Nonzero_nuclei)) {
  stop("Nucleus count mismatch")
}
fwrite(qc, file.path(out, "donor_celltype_QC_all_strata.csv"))

classified <- subset(qc, CellType != "Unclassified")
groups <- split(classified, interaction(classified$Group, classified$CellType,
                                        drop = TRUE))
summary_ct <- do.call(rbind, lapply(groups, function(d) {
  data.frame(
    Group = unique(d$Group),
    CellType = unique(d$CellType),
    Donors = nrow(d),
    Donors_QC20 = sum(d$QC20),
    Donors_below20 = sum(d$N_nuclei > 0 & !d$QC20),
    Donors_zero = sum(d$N_nuclei == 0),
    Median_nuclei = median(d$N_nuclei),
    Min_nuclei = min(d$N_nuclei),
    Max_nuclei = max(d$N_nuclei)
  )
}))
summary_ct <- as.data.frame(data.table::setorder(
  as.data.table(summary_ct), CellType, Group
))
fwrite(summary_ct, file.path(out, "QC_summary_by_group_celltype.csv"))

excluded <- subset(classified, !QC20)
fwrite(excluded, file.path(out, "excluded_donor_celltype_strata.csv"))

pb_file <- file.path(input, "GSE261807_celltype_target_pseudobulk.csv")
if (file.exists(pb_file)) {
  pb <- fread(pb_file)
  stopifnot(all(c("GSM", "CellType") %in% names(pb)))
  observed <- unique(data.frame(
    GSM = as.character(pb$GSM),
    CellType = as.character(pb$CellType)
  ))
  observed_keys <- paste(observed$GSM, observed$CellType, sep = "||")
  qc <- transform(qc, In_pseudobulk =
    paste(GSM, CellType, sep = "||") %in% observed_keys)
  missing <- subset(qc, CellType != "Unclassified" & QC20 & !In_pseudobulk)
  included <- subset(qc, CellType != "Unclassified" & !QC20 & In_pseudobulk)
  fwrite(qc, file.path(out, "QC_vs_pseudobulk.csv"))
  fwrite(missing, file.path(out, "unexpected_missing_strata.csv"))
  fwrite(included, file.path(out, "unexpected_included_strata.csv"))
  cat("\nEligible strata absent from pseudobulk:\n")
  print(missing)
  cat("\nIneligible strata present in pseudobulk:\n")
  print(included)
} else {
  warning("Pseudobulk file not found; concordance not tested")
}

writeLines(capture.output(sessionInfo()), file.path(out, "sessionInfo.txt"))
cat("\nFINAL QC SUMMARY\n")
print(summary_ct)
cat("\nEXCLUDED STRATA\n")
print(subset(excluded, select = c(GSM, Group, CellType, N_nuclei)))
cat("\nDonors:", length(unique(qc$GSM)),
```

```
  "\nEligible classified strata:", sum(classified$QC20),
  "\nExcluded classified strata:", nrow(excluded), "\n")
cat("Saved to:", out, "\n")
```

Supplementary methods

Cell-Type Annotation and Marker Genes

Because the final annotated cell-level object used by the original investigators of GSE261807 was not publicly deposited, broad cell classes were independently reconstructed using canonical lineage-associated marker genes documented in the investigators' analysis scripts. Marker panels included SLC17A7 and NRGN for excitatory neurons; GAD1 and GAD2 for inhibitory neurons; MBP, PLP1, and MOBP for oligodendrocytes; VCAN for oligodendrocyte precursor cells (OPCs); GFAP, ALDH1L1, and AQP4 for astrocytes; CX3CR1, P2RY12, AIF1, and PTPRC for microglia/myeloid cells; and CLDN5, PECAM1, and FLT1 for endothelial cells. Additional genes appearing in the investigators' cell-characterization panels included CUX2, RORB, FEZF2, PCP4, and other neuronal subtype markers, as well as NOSTRIN and other vascular-associated genes. Thus, the present classifications were informed by the investigators' documented marker framework but represent an independent marker-based reconstruction rather than a reuse of their original cell-level annotations (Table *2*).

Marker genes	
Excitatory neurons	SLC17A7, NRGN
Inhibitory neurons	GAD1, GAD2
Oligodendrocytes	MBP, PLP1, MOBP
OPC	VCAN
Astrocytes	GFAP, ALDH1L1, AQP4
Microglia/myeloid	CX3CR1, P2RY12, AIF1, PTPRC
Endothelial	CLDN5, PECAM1, FLT1

TABLE 2: Cell population and marker genes documented in authors' scripts

OPC: Oligodendrocyte precursor cells.

Additional Information

Author Contributions

All authors have reviewed the final version to be published and agreed to be accountable for all aspects of the work.

Concept and design: Steven Lehrer

Acquisition, analysis, or interpretation of data: Steven Lehrer

Drafting of the manuscript: Steven Lehrer

Critical review of the manuscript for important intellectual content: Steven Lehrer

Supervision: Steven Lehrer

Disclosures

Human subjects: Informed consent for treatment and open access publication was obtained or waived by all participants in this study. **Animal subjects:** All authors have confirmed that this study did not involve animal subjects or tissue. **Conflicts of interest:** In compliance with the ICMJE uniform disclosure form, all authors declare the following: **Payment/services info:** All authors have declared that no financial support was received from any organization for the submitted work. **Financial relationships:** All authors have declared that they have no financial relationships at present or within the previous three years with any organizations that might have an interest in the submitted work. **Other relationships:** All authors have declared that there are no other relationships or activities that could appear to have influenced the

submitted work.

Acknowledgements

ChatGPT (OpenAI; GPT-5.6) was used during manuscript preparation to help develop and debug R code; organize and format statistical results; check consistency across the Abstract, Methods, Results, tables, and figure legends; and revise the manuscript text. The AI tool was not used to independently generate source data or make final scientific conclusions. All statistical outputs, citations, interpretations, and manuscript revisions were reviewed and verified by the authors, who take full responsibility for the content of the manuscript.

References

1. Bieniek KF, Cairns NJ, Crary JF, et al.: The second NINDS/NIBIB consensus meeting to define neuropathological criteria for the diagnosis of chronic traumatic encephalopathy. J Neuropathol Exp Neurol. 2021, 80:210-9. 10.1093/jnen/nlab001
2. McKee AC, Stein TD, Huber BR, et al.: Chronic traumatic encephalopathy (CTE): criteria for neuropathological diagnosis and relationship to repetitive head impacts. Acta Neuropathol. 2023, 145:371-94. 10.1007/s00401-023-02540-w
3. Alosco ML, Cherry JD, Huber BR, et al.: Characterizing tau deposition in chronic traumatic encephalopathy (CTE): utility of the McKee CTE staging scheme. Acta Neuropathol. 2020, 140:495-512. 10.1007/s00401-020-02197-9
4. Stein TD, Alvarez VE, McKee AC: Chronic traumatic encephalopathy: a spectrum of neuropathological changes following repetitive brain trauma in athletes and military personnel. Alzheimers Res Ther. 2014, 6:4. 10.1186/alzrt234
5. Butler ML, Pervaiz N, Breen K, et al.: Repeated head trauma causes neuron loss and inflammation in young athletes. Nature. 2025, 647:228-37. 10.1038/s41586-025-09534-6
6. Labadorf A, Agus F, Aytan N, Cherry J, Mez J, McKee A, Stein TD: Inflammation and neuronal gene expression changes differ in early versus late chronic traumatic encephalopathy brain. BMC Med Genomics. 2023, 16:49. 10.1186/s12920-023-01471-5
7. Seo JS, Lee S, Shin JY, et al.: Transcriptome analyses of chronic traumatic encephalopathy show alterations in protein phosphatase expression associated with tauopathy. Exp Mol Med. 2017, 49:e333. 10.1038/emm.2017.56
8. Cherry JD, Meng G, Daley S, et al.: CCL2 is associated with microglia and macrophage recruitment in chronic traumatic encephalopathy. J Neuroinflammation. 2020, 17:370. 10.1186/s12974-020-02036-4
9. Cherry JD, Tripodis Y, Alvarez VE, et al.: Microglial neuroinflammation contributes to tau accumulation in chronic traumatic encephalopathy. Acta Neuropathol Commun. 2016, 4:112. 10.1186/s40478-016-0382-8
10. Chou FS, Wang PS: The Arp2/3 complex is essential at multiple stages of neural development . Neurogenesis (Austin). 2016, 3:e1261653. 10.1080/23262133.2016.1261653
11. Stolt CC, Lommes P, Sock E, Chaboissier MC, Schedl A, Wegner M: The Sox9 transcription factor determines glial fate choice in the developing spinal cord. Genes Dev. 2003, 17:1677-89. 10.1101/gad.259003
12. Chung WS, Clarke LE, Wang GX, et al.: Astrocytes mediate synapse elimination through MEGF10 and MERTK pathways. Nature. 2013, 504:394-400. 10.1038/nature12776
13. Chancellor KB, Chancellor SE, Duke-Cohan JE, et al.: Altered oligodendroglia and astroglia in chronic traumatic encephalopathy. Acta Neuropathol. 2021, 142:295-321. 10.1007/s00401-021-02322-2
14. Alosco ML, Mez J, Tripodis Y, et al.: Age of first exposure to tackle football and chronic traumatic encephalopathy. Ann Neurol. 2018, 83:886-901. 10.1002/ana.25245