Efficient differential expression analysis of large-scale single cell transcriptomics data using dreamlet

Gabriel E. Hoffman[1,2,3]*, Donghoon Lee[1,2,3,4], Jaroslav Bendl[1,2,3,4], Prashant Fnu[1,2,3,4], Aram Hong[1,2,3,4], Clara Casey[1,2,3,4], Marcela Alvia[1,2,3,4], Zhiping Shao[1,2,3,4], Stathis Argyriou[1,2,3,4], Karen Therrien[1,2,3,4], Sanan Venkatesh[1,2,3,4], Georgios Voloudakis[1,2,3,4], Vahram Haroutunian[2,5,6], John F. Fullard[1,2,3,4], Panos Roussos[1,2,3,4,5,6]*

[1] Center for Disease Neurogenomics,
[2] Department of Psychiatry,
[3] Department of Genetics and Genomic Sciences,
[4] Friedman Brain Institute,
[5] Department of Neuroscience,
Icahn School of Medicine at Mount Sinai, New York, NY, USA
[6] Mental Illness Research, Education and Clinical Centers,
James J. Peters VA Medical Center, Bronx, New York

*Correspondence to:
G.E.H (gabriel.hoffman@mssm.edu) and P.R. (panagiotis.roussos@mssm.edu)

Abstract

Advances in single-cell and -nucleus transcriptomics have enabled generation of increasingly large-scale datasets from hundreds of subjects and millions of cells. These studies promise to give unprecedented insight into the cell type specific biology of human disease. Yet performing differential expression analyses across subjects remains difficult due to challenges in statistical modeling of these complex studies and scaling analyses to large datasets. Our open-source R package dreamlet (DiseaseNeurogenomics.github.io/dreamlet) uses a pseudobulk approach based on precision-weighted linear mixed models to identify genes differentially expressed with traits across subjects for each cell cluster. Designed for data from large cohorts, dreamlet is substantially faster and uses less memory than existing workflows, while supporting complex statistical models and controlling the false positive rate. We demonstrate computational and statistical performance on published datasets, and a novel dataset of 1.4M single nuclei from postmortem brains of 150 Alzheimer's disease cases and 149 controls.

The human body is composed of hundreds of cell types, each with their own role in the biology of health and disease [1]. The cell type specificity of fundamental biological mechanisms has long been appreciated [2,3], and genome-wide association studies have revealed strong cell type specific enrichments of risk variants that inform our understanding of disease biology [4–6].

Recent advances in single-cell and -nucleus transcriptomics technology have enabled profiling of over a million cells from hundreds of subjects in order to study variation in gene expression associated with disease or other traits [7–11]. Multiplexing samples and assigning cells to subjects following sequencing, using genetic variation [12,13] or hashing with barcoded antibodies [14], has enabled a further increase in the scale of these studies. As the scale of single cell data continues to increase, and studies of cross-subject variation profile more subjects and cells, analytical workflows must keep pace.

Existing methods for performing differential expression analysis across subjects were designed for small to moderate size datasets, and these model gene expression at either the single cell or the pseudobulk level for each cell type cluster. Modeling expression at the single cell level is the most direct approach and can be performed with a range of statistical models [15–18] while allowing cell-level covariates. Yet statistical modeling of cell-level counts is challenging due to the low read depth and pervasive dropout effects [15,18], and it is essential to model the fact that multiple cells are sampled from the same subject [16,17]. Even then, controlling the false positive rate remains challenging [15]. Moreover, as studies assay more cells, the computational cost required to fit thousands of cell-level regression models becomes prohibitive [15].

Alternatively, pseudobulk approaches aggregate reads across cells within a cell cluster and then use methods originally developed for bulk RNA-seq [15]. This works well for moderate sized datasets, but aggregating reads with existing methods across cells in large-scale studies can be very computationally and memory intensive. More importantly, emerging single-cell and -nucleus transcriptome datasets use complex study designs including technical or biological replicates, or use sample multiplexing that can introduce high-dimensional batch effects that existing pseudobulk methods cannot model adequately. Here, we introduce dreamlet, an open-source R package that shows superior computational and statistical performance compared to existing methods, addressing previous challenges related to statistical modeling of data from large cohorts.

Results

Dreamlet workflow for large-scale differential expression analysis

Dreamlet applies a pseudobulk approach and fits a regression model for each gene and cell cluster to test differential expression associated with variation in a trait across subjects. Use of precision-weighted linear mixed models enables accounting for repeated measures study designs, including technical or biological replicates, high dimensional batch effects due to sample multiplexing, and variation in sequencing depth [19] **(Figure 1)**. Dreamlet incorporates precision weights at two levels to account for uncertainty in the observed gene expression measurements. First, dreamlet uses precision weights to account for the fact that measuring

more cells from a biospecimen increases the measurement precision of the underlying expression state, since the observed cells are sampled from a larger population of cells [15]. Second, dreamlet uses a mean-variance trend weighting in order to account for the fact that higher read counts give higher measurement precision [20]. Dreamlet also extends the use of empirical Bayes moderated t-statistics, which borrow information across genes to increase power and control of false positive rate [21], to the case of precision-weighted linear mixed models (**Supplementary Methods**). Finally, the dreamlet workflow is designed for large-scale data and is substantially faster and uses less memory than existing workflows, while supporting complex statistical models and controlling the false positive rate.

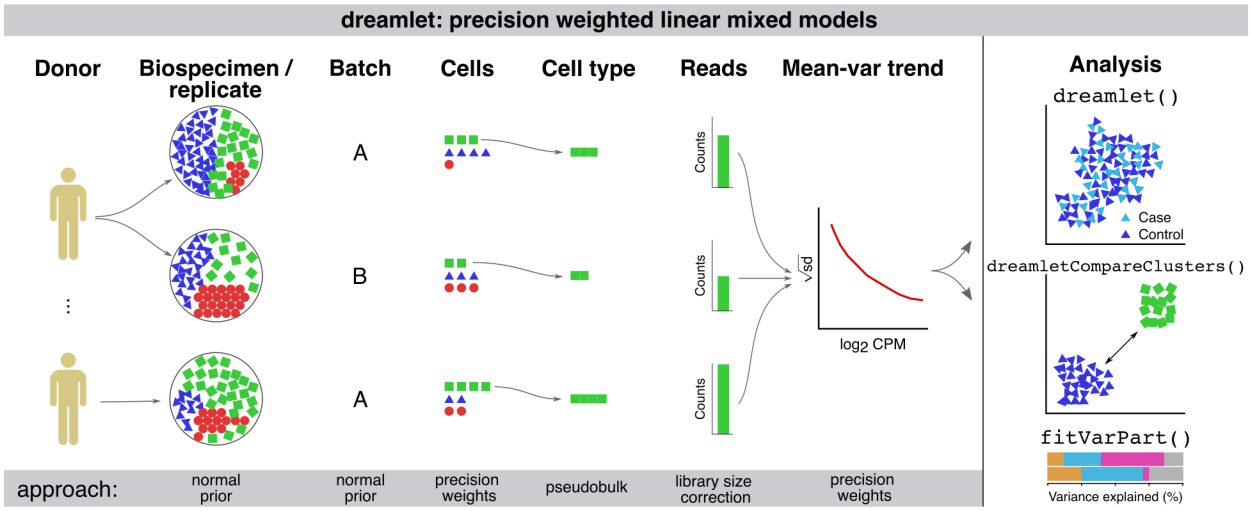

Figure 1. Illustration of dreamlet workflow using precision weighted linear mixed models. Expression variation across multiple biological or technical replicates, and technical batches are modeled using a random effect with a normal prior in a linear mixed model. Variation in the number of cells observed for each sample and cell type are modeled using precision weights. Pseudobulk counts are computed for each cell type and standard library size correction is performed. Precision weights are then used in a second weighting step to model the mean-variance trend. The dreamlet package provides interfaces for differential expression analysis across donors and cell clusters, for variance partitioning analysis, and for downstream analyses and visualization.

Computational performance on large-scale datasets

The dreamlet workflow uses an efficient implementation to compute pseudobulk counts and scales to larger datasets than competing methods. Dreamlet processes pseudobulk for 1000 donors across 12 cell types for 2.2 million cells in 10 minutes using only 20 Gb of memory (**Figure 2A,B**). Performance of pegasus [22] is slightly faster but its memory footprint is substantially larger and increases rapidly with sample size. Other methods are 10-100x slower and are limited by memory constraints even for moderate sample sizes.

Dreamlet fits precision-weighted linear mixed models for each gene parallelized across multiple CPU cores and enables analysis of 12 cell types across 326 subjects in 45 CPU minutes corresponding to a wall time of 15 minutes in this case (**Figure 2C**). This is over an order of

magnitude faster than using generalized linear mixed models (GLMMs) at the single cell level, and competitive with other pseudobulk approaches using a negative binomial model that are not able to model random effects (i.e. edgeR [23], DESeq2 [24]).

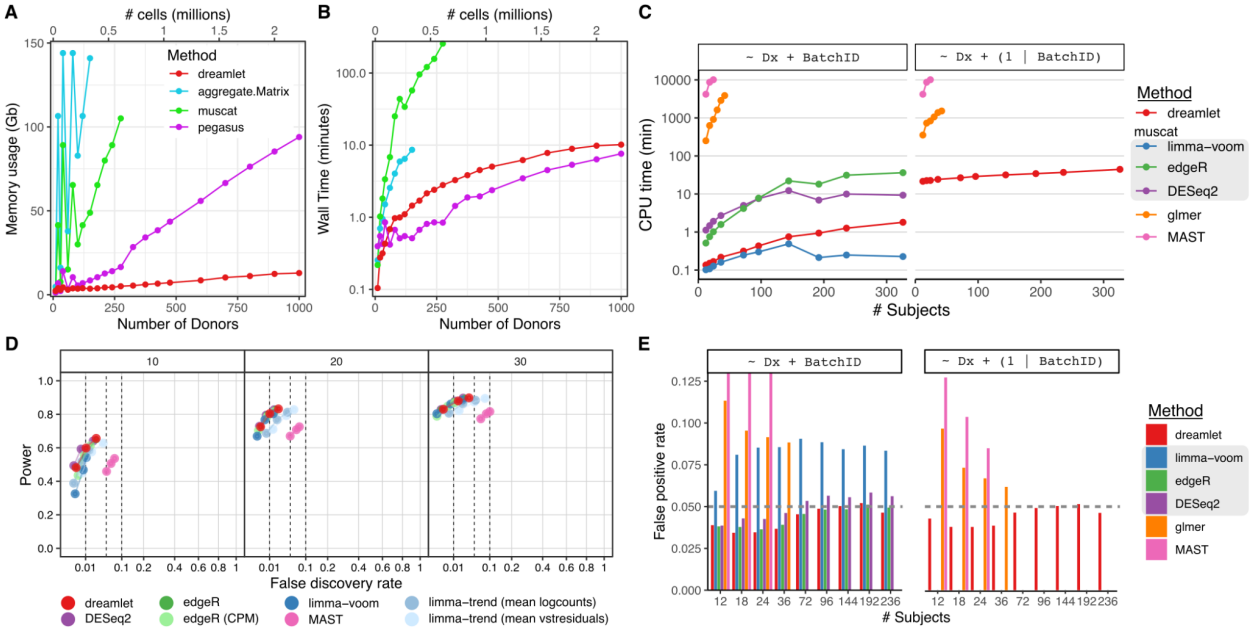

Figure 2: Computational and statistical performance of dreamlet workflow. A,B) Peak memory usage (**A**) and CPU time (**B**) averaged across 10 runs using 1 CPU core on a machine with 144 Gb memory. (**C**) CPU times for differential expression analysis for an increasing number of subjects for 6 differential expression methods. Methods in the gray box use a pseudobulk approach run using muscat software, while glmer and MAST model the data at the single cell level using generalized linear (mixed) models. (**D**) Power plotted against false discovery rate (FDR) for 8 methods for 10, 20 and 30 subjects. Dashed vertical gray lines indicate target FDR values of 1,5, and 10%, and points are shown for these target values. (**E**) False positive rates for an increasing number of subjects using real single nucleus RNA-seq from postmortem human brains and permuting the disease status. Dashed horizontal gray lines indicate the target false positive rate of 5%. Results are shown for testing the effect of diagnosis, while modeling batch as a fixed (left) or random effect (right). Methods in the gray box use a pseudobulk approach run using muscat software, while glmer and MAST model the data at the single cell level.

Statistical performance on large-scale datasets

Using a simulation pipeline and benchmarks developed by an independent group [15], dreamlet using only a fixed effects model shows statistical performance comparable to the best performing current methods while giving the most accurate estimates of the simulated effect sizes (**Figure 2D, Supplementary Figures 1-3**). Yet simulating single cell transcriptomics data that accurately recapitulates the complexity of real datasets is notoriously challenging [15,25]. Instead, we used real data from human postmortem brains and permuted the disease status to estimate the empirical false positive rate across differential expression methods within each of 12 cell types. Of the methods that control the false positive rate across a range of sample sizes,

dreamlet is the only one that can model random effects (**Figure 2E**). Results are similar when examined within each of the 12 cell types. Notably, the limma/voom method applied by muscat [15] shows inflated false positive rates in most conditions. Generalized linear mixed models used by the glmer [26] and MAST [18] methods can model random effects, but have an inflated false positive rate for small and moderate sample sizes. This is consistent with the fact that the null distributions of the coefficient estimates for these methods are normally distributed only in the asymptotic limit of large sample size. Additional methods tested by others do not scale to these large datasets [15–17].

Multiplexing combines multiple samples in parts of the experimental workflow and reduces the cost of generating large-scale datasets. This process creates many small batches, each of 6 to 12 samples, that can share technical artifacts. Statistically, these batches are termed 'high dimensional' when their number is large compared to the sample size. In order to control the false positive rate, a statistical model should account for these high dimensional batch effects. Yet widely used fixed effects models can suffer from a substantial decrease in power, and can even perform worse when including the batch effect compared to omitting it from the model[27]. We show that using a linear mixed model to account for batches with a random effect controls the false positive rate while retaining power (**Supplemental Figure 4**).

Next we apply dreamlet to single cell RNA-seq from T-cells following tuberculosis, and bone metastasis from prostate cancer. We then describe a novel single nucleus RNA-seq data from an Alzheimer's disease cohort, and perform analysis using dreamlet to examine cell type specific biology of the disease.

Dreamlet uncovers upregulation of interferon-gamma pathway in a subpopulation of T-cells in response to tuberculosis infection

Nathan, et al. [9] studied the post-disease steady state of T-cells following tuberculosis (TB) infection and disease resolution, by isolating 500K T-cells from 259 donors, including 128 of which had previously had TB. Yet the original work did not report differential expression analysis between cases and controls. We applied dreamlet to identify genes associated with TB status within each of the 31 data-derived T-cell clusters. The number of expressed genes varied widely across cell types (**Supplementary Figure 5**). Based on differential expression findings downstream, we highlight the cluster of CD4/8$^+$ PD-1$^+$TIGIT$^+$ T-cells. Analysis of pseudobulk counts for this cluster (**Figure 3A**) shows a strong mean-variance trend in the $\log_2$ counts that dreamlet models using precision weights (**Figure 3B**). Variance partitioning analysis for each gene in this cluster estimates the fraction of expression variance attributable to variation across 46 technical batches, 259 donors and 2 disease states (**Figure 3C**). While the median fraction of variation explained by batch is 8.47%, 20.8% of expressed genes have more than 25% of their expression variance attributed to cross-batch variation. Importantly, TB status explains > 1% of variance for just 7.4% of expressed genes, underscoring the small effect of TB infection following disease resolution. Genes can have very different variance partitioning profiles (**Figure 3D**), where *HIST1H1E* shows high variance across technical batches (**Figure 3E**) while *LTB* is among the genes with the largest variation explained by disease status (**Figure 3F**). While differential expression analysis between TB and controls did not identify any

study-wide significant genes at FDR < 5% (i.e. accounting for 64,057 total tests across 31 cell clusters), gene set analysis using the full spectrum of test statistics [28] identified a study-wide significant upregulation in TB cases of 'cellular response to interferon-gamma' in the CD4/8$^+$ PD-1$^+$TIGIT$^+$ cluster, and a handful of other genesets in other T-cell clusters (**Figure 3G**). Examining genes in this geneset shows consistent upregulation in TB cases for genes with sufficient expression to be analyzed (**Figure 3H**). Furthermore, these gene sets show upregulation in other cell clusters, but not passing the study-wide significance cutoff, indicating a subtle residual immune activation even following resolution of TB disease.

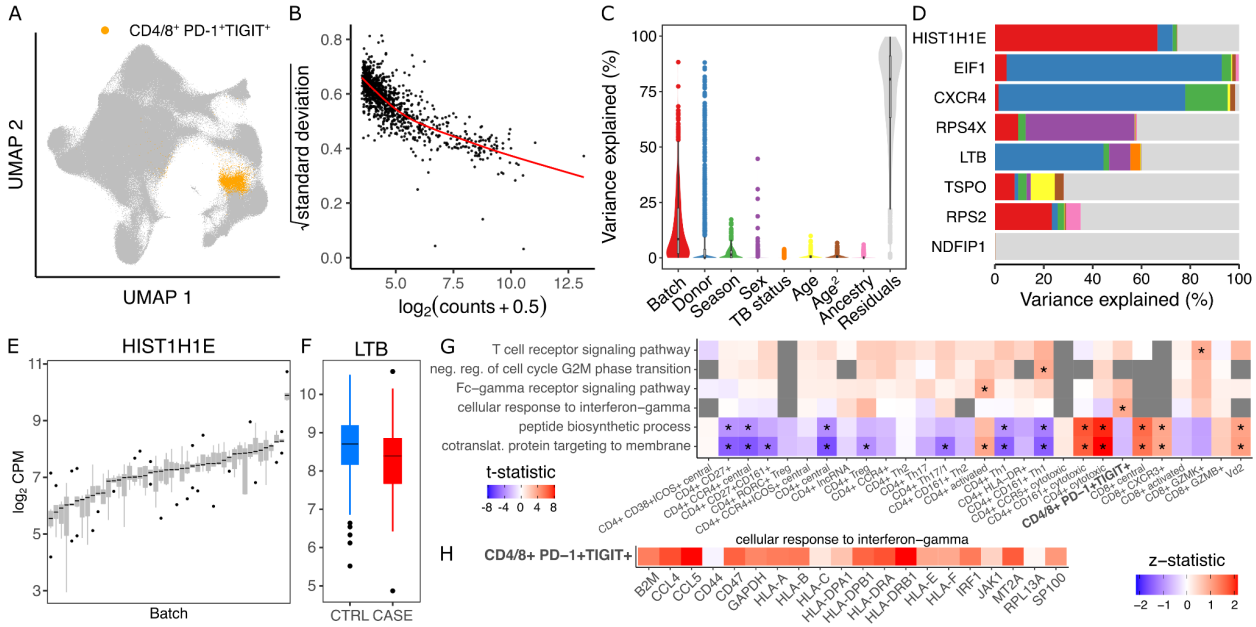

Figure 3: Dreamlet analysis of expression response to tuberculosis infection in T-cell subpopulations in 259 donors. A) Dimensionality reduction highlighting CD4/8$^+$ PD-1$^+$TIGIT$^+$ subpopulation of T-cells. **B)** Plot of mean-variance trend for expression counts in this subpopulation. **C)** Violin plot summarizing variance partitioning analysis separating the fraction of expression variation for each gene into 9 components. **D)** Representative genes with high variance fractions explained by variance components. **E)** *HIST1H1E* shows high expression variation across batches. **F)** *LTB* shows high expression variation across disease status. **G)** Gene set analysis using the full spectrum of test statistics shows specific immune activation in the CD4/8$^+$ PD-1$^+$TIGIT$^+$ subpopulation. Study-wide FDR < 0.05 is indicated by '*'. **H)** Test statistics from dreamlet differential expression analysis in this subpopulation for genes involved in cellular response to interferon-gamma.

Dreamlet identifies robust transcriptional changes in bone metastases from prostate cancer

Prostate cancer can metastasize to the bone and result in very poor patient prognosis. Kfoury et al. [8] performed single cell RNA-seq on solid tumor, involved bone marrow, and distal bone marrow from 9 prostate cancer patients with bone metastases, as well as benign bone marrow from 7 patients without cancer. We applied the dreamlet workflow to identify genes that were differentially expressed based on disease status by modeling the multiple measurements from each cancer patient using a random effect. The number of expressed and differentially expressed genes varied widely across cell types and group comparisons (**Supplementary Figure 6**). Analysis of pseudobulk counts for a monocyte cluster (**Figure 4A**) shows the typical mean-variance trend that dreamlet models (**Figure 4B**). Variance partitioning analysis in this cell type estimates the fraction of expression variance attributable to variation across patients, disease status of each sample (i.e., tumor, involved bone marrow, etc), and disease status of the patient (i.e. cancer vs non-cancer). Expression variation across patients is the strongest, while disease status explains greater than 10% of expression variation for 398 of 1376 (28.9%) genes with sufficient expression to be included in the analysis (**Figure 4C**). This is consistent with cancer being a much stronger driver of gene expression changes than TB, above. Genes show a range of variance partitioning profiles (**Figure 4D**). For example, while *CCNL1* shows high variance across disease status, *VIM* has high variation explained by patient status (**Figure 4E**). Dreamlet analysis identified 134 genes in this monocyte cluster as differentially expressed between tumor and involved bone marrow at a study-wide FDR of 5% (**Figure 4F**). Gene set analysis using the full spectrum of test statistics identified upregulation of MHC class II proteins and protein folding chaperones in the tumor samples across a range of cell types, including multiple monocyte clusters (**Figure 4G**). Examining protein folding chaperones shows upregulation in tumor samples for most genes with sufficient expression (**Figure 4H**).

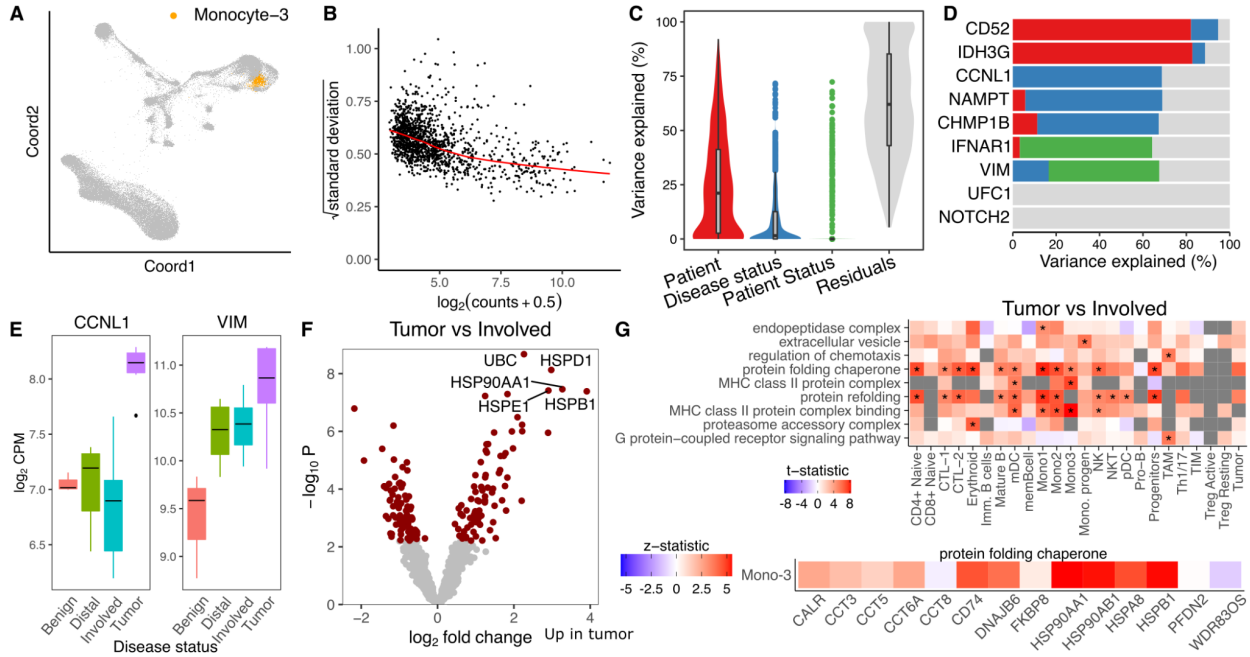

Figure 4: Dreamlet analysis of expression differences in prostate cancer bone metastases. A) Dimensionality reduction highlighting monocyte population. **B)** Plot of mean-variance trend for expression counts in this monocyte population. **C)** Violin plot summarizing variance partitioning analysis separating the fraction of expression variation for each gene into 4 components. **D)** Representative genes with high variance fractions explained by each of the 4 components. **E)** Representative genes with high expression variation across disease state (*CCNL1*) and subject cancer status (*VIM*). **F**) Volcano plot of differential analysis between tumor and involved bone marrow within each subject. **G)** Gene set analysis using the full spectrum of test statistics shows cell type specific signatures of tumor vs involved bone marrow. Study-wide FDR < 0.05 is indicated by '*'.

Modeling multiple sources of expression variation in a large Alzheimer's Disease cohort

We generated single nucleus RNA-seq (snRNA-seq) on tissue samples from the dorsolateral prefrontal cortex (DLPFC) of postmortem brains from donors in the Mount Sinai NIH NeuroBioBank (**Figure 5A**). Samples were multiplexed by pooling 6 donors using the nuclei hashing method [29]. Each pool was processed in duplicate using the 10x Genomics single cell gene expression platform, to produce technical replicates. Following raw data processing and quality control, the dataset comprised 586 samples from 299 donors over 60 years of age and 1.4M nuclei. Of these donors, 150 had Alzheimer's disease (AD) and 149 were age-matched neurotypical controls (**Supplementary Figure 7**). Cell cluster annotation identified 22 cell types including 8 subtypes of excitatory neurons and 6 subtypes of inhibitory neurons (**Figure 5B**). Analysis of pseudobulk counts for microglia (Micro_PVM) shows the strong mean-variance trend (**Figure 5C**) that dreamlet models with precision weights. Variance partitioning analysis for each gene in this cluster estimates the fraction of expression variance attributable to variation across technical replicates from the same subject, as well as variation across age, AD status, sample pool and sex (**Figure 5D**). This identifies genes with different variance profiles where, for example, 31.6% of the variation in *PTPRG* is explained by Alzheimer's status while 98% of variance in NXPE1 is explained by sample pool (**Figure 5E**).

In microglia, the median fraction of variation explained by subject is 50.9%, indicating good reproducibility in expression values across technical replicates. The effect of AD status is modest with only 143 genes having more than 5% variance explained, underscoring the need for large sample sizes in order to characterize expression changes associated with the disease. While the sample pool explains minimal variance for the majority of genes, 668 genes have >5% variance explained across the pools. This indicates a substantial batch effect for these genes. The high dimensionality of the batch effect due to multiplexing 6 samples per pool indicates that modeling this as a random effect using dreamlet can retain high power while avoiding false positive findings. These findings are consistent across other cell types (**Supplementary Figure 8**). Further characterizing the batch effect reveals significant correlations between the variance across sample pools and the GC content of each gene (computed using all exons in the reference genome) in many cell clusters. This is consistent with previous work showing PCR artifacts driving technical variation [30,31] (**Figure 5F**).

While single cell/nucleus assays are uniquely capable of identifying cell type specific effects, the statistical power for each cell cluster in a given dataset varies widely. In particular, increasing the average number of nuclei per subject is directly related to an increase in the technical reproducibility of the gene expression measurements according to the variance explained across technical replicates from the same subject (**Figure 5G**). Moreover, the number of nuclei per subject impacts the read count per subject, the number of genes that pass a minimum expression cutoff, and the number of genes that are found to be differentially expressed between AD and controls (**Figure 5H**).

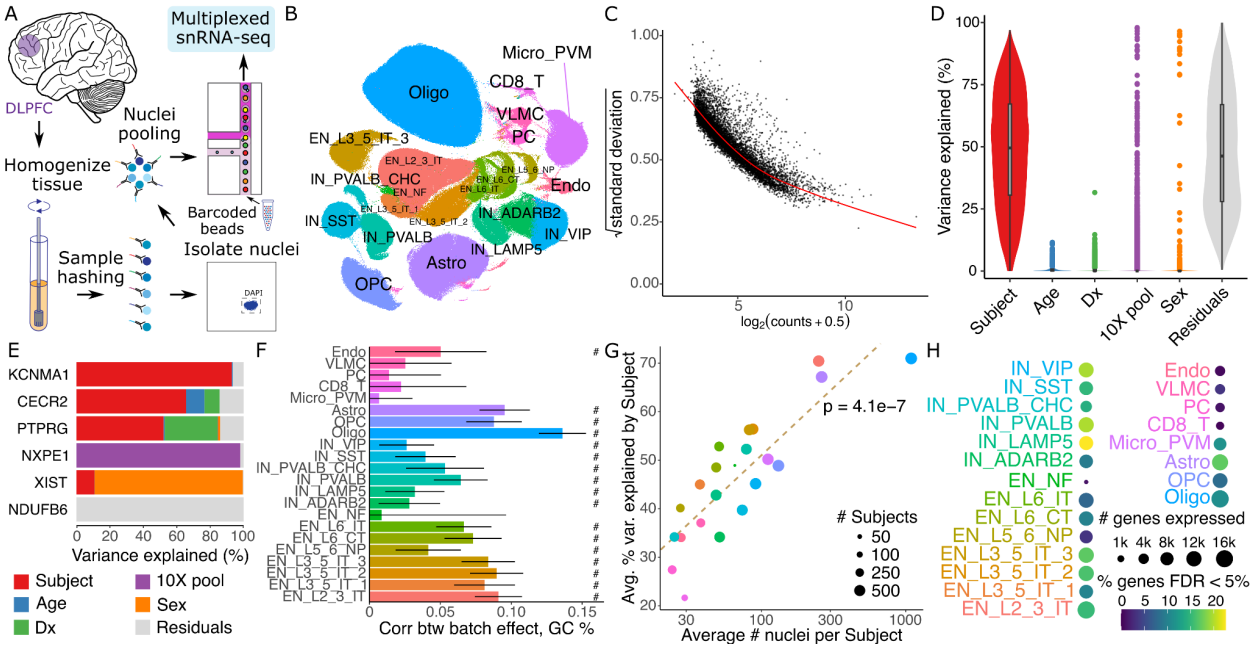

Figure 5: Dreamlet analysis of expression differences in Alzheimer's Disease. A) Multiplexed single nucleus RNA-seq was performed on postmortem brain samples. **B)** UMAP dimensionality reduction with cell cluster annotations. **C)** Plot of mean-variance trend for expression counts in the microglia population. **D)** Violin plot for microglia summarizing variance partitioning analysis separating the fraction of expression variation for each gene into 6

components. **E)** Representative genes with high variance fractions explained by each of the 6 components in microglia. **F)** Spearman correlation of variance explained by sample pool ID with GC content of each gene. Bars indicate 95% confidence interval, '#' indicates < 5% FDR. **G)** Annotated clusters with more nuclei per subject show higher concordance in technical replicates. Circle size indicates the number of subjects with at least 10 nuclei observed for the cluster. P-value from linear regression is shown. **H)** Number of genes passing expression cutoffs and the fraction of genes differentially expressed between AD subjects and controls at 5% FDR.

Large effect up-regulation of *PTPRG* in microglia of Alzheimer's Disease cases

The number of differentially expressed genes between donors with AD and controls varies widely between cell types and increases on the number of nuclei observed per subject (**Supplementary Figure 9**). The role of microglia in AD has received much recent attention due to findings from human genetic studies and mouse models [32–34], but the understanding of disease-associated gene expression changes has been more limited. Here, microglia have 826 differentially expressed genes passing the study-wide 5% FDR cutoff (**Figure 6A**). Even so, *PTPRG* stands out with a $\log_2$ fold change of 1.52 and p-value of 9.28e-28. Genes demonstrate a range of cell type specificity patterns with many differentially expressed genes shared across multiple excitatory and inhibitory neuron subtypes (**Figure 6B**). Notably, the *PTPRG* up-regulation with this large effect size is specific to microglia, with significant but much smaller $\log_2$ fold changes observed in 1 excitatory and 2 inhibitory neuron subtypes (**Figure 6C**). Other genes show effects across a range of cell types, as *PDE10A* is significantly down-regulated in AD in 12 of 14 neuron subtypes (**Figure 6D**). Gene set analysis using the full spectrum of test statistics in each cell cluster identifies up-regulation of synapse assembly and glutamate pathways in subsets of excitatory and inhibitory neurons, up-regulation of synaptic vesicle endocytosis most strongly in EN_L2_3_IT, neuronal action potential most strongly in IN_PVALB (**Figure 6E**). In astrocytes there is a specific down-regulation of cholesterol biosynthesis and up-regulation of ERBB signaling. Oligodendrocyte precursors (OPC) have a specific upregulation of neuropeptide signaling and neural nucleus development. Notably, microglia have a specific up-regulation of the p38MAPK cascade, which is involved in microglial inflammatory response [35]. This pathway includes DUSP10 which is one of the top upregulated genes in microglia (**Figure 6F**).

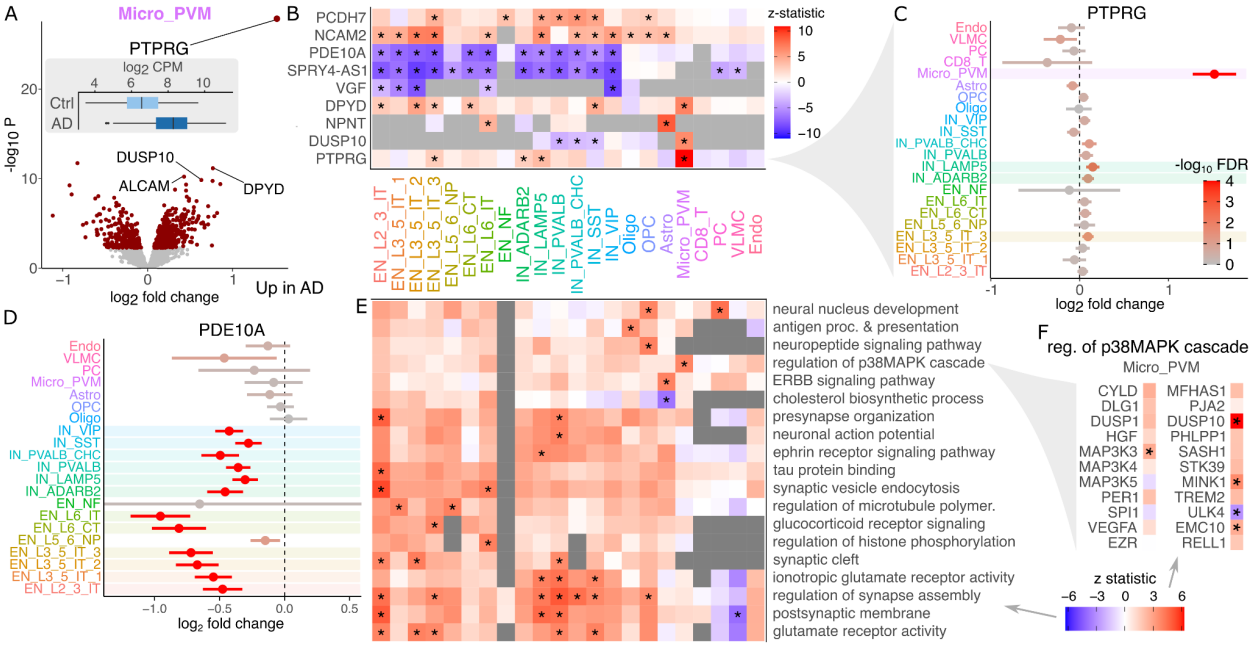

Figure 6: Gene expression signatures of Alzheimer's disease. A) Volcano plot of differential expression between AD cases and controls in microglia. Red points indicate FDR < 5%. Inset shows boxplot of PTPRG stratified by disease status. **B)** Heatmap showing differential expression z-statistic for genes in each cell cluster. '*' indicates study-wide FDR < 5% in all panels. Grey box indicates a gene did not pass expression cutoff in that cell cluster. **C)** Forest plot of $\log_2$ fold change for *PTPRG* in each cell type. Bars indicate 95% confidence interval. Color indicates FDR. **D)** Forest plot for *PDE10A*. **E)** Gene set analysis using the full spectrum of differential expression test statistics. **F)** Differential expression results for genes involved in regulation of p38MAPK cascade.

Discussion

The advent of single cell technology has enabled generation of large-scale high-resolution atlases to characterize cell type specific biology [36–38]. As the scale of datasets continues to increase, there is new potential to study variation across subjects at the cell type level and how gene expression changes relate to a subject's age, sex, disease state, and many other traits [7–11].

We present the dreamlet software, an open-source R package that enables analysis of massive-scale single cell/nucleus transcriptome datasets. Dreamlet addresses both CPU and memory usage limitations by performing preprocessing and statistical analysis in parallel on multicore machines, and can distribute work across multiple nodes on a compute cluster. Dreamlet also uses the H5AD format for on-disk data storage to enable data processing in smaller chunks to dramatically reduce memory usage [39]. The dreamlet workflow easily integrates into the Bioconductor ecosystem [40], and uses the SingleCellExperiment class [41] to facilitate compatibility with other analyses. Fitting precision-weighted linear mixed models [19] enables control of the false positive rate while retaining high power, even in the presence of high dimensional batch effects. Beyond differential expression testing, dreamlet provides seamless integration of downstream analysis including quantifying sources of expression variation [42], gene set analysis using the full spectrum of gene-level t-statistics [28] and visualizing results.

We also introduce a novel dataset of 1.4M single nuclei from postmortem brains from 150 Alzheimer's disease cases and 149 controls. Analysis using the dreamlet software examines the cell type specific biology of AD. Highlighting the role of microglia in the disease, we observed that *PTPRG*, a protein tyrosine phosphatase receptor, is upregulated in AD and stands out substantially from all other genes in terms of effect size and p-value. PTPRG is an inflammatory marker but its role in the molecular etiology of AD is unclear [43,44]. Interestingly, *PTPRG* is among the three genes (the other two being *APOE* and *DYPD*) that are reliably upregulated in two out of three previous human microglia transcriptome studies [45]. Recent genome-wide association studies of AD do not identify risk variants in the region of the gene, suggesting that *PTPRG* upregulation is reactive, and might behave differently depending on the stage and progression of AD.

Yet our work also highlights challenges in single-cell and -nucleus studies. First, technical batch effects can be substantial for some genes, and downstream analysis must account for these effects in order to control the false positive rate. Second, the findings within each cell cluster can be driven by biology, but also by limitations in the precision of gene expression measurements. We observe that increasing the measurement precision by increasing the number of reads and nuclei increases the reproducibility across technical replicates, and increases the number of differentially expressed genes identified. This wide variation in measurement precision across cell clusters yields wide differences in power to detect differential signals even within the same dataset. A finding that a gene is differentially expressed in only one cell cluster may be driven by cell type specific biology of disease, but could also be due to lower power in the other cell clusters. While the prospect of understanding the cell type specific

biology of disease motivates these large cohort-scale studies, we recommend caution in the interpretation of cell type specific findings.

In conclusion, here we introduce dreamlet, an open-source R package (DiseaseNeurogenomics.github.io/dreamlet) that addresses previous challenges to perform efficient differential expression analyses across subjects in large-scale single-cell datasets. Dreamlet is substantially faster and uses less memory, supports complex statistical models and better controls the false positive rate compared to existing workflows, providing an important tool to address the need of expanding single cell/nucleus transcriptome datasets.

Online Methods

Efficient computation of pseudobulk using on-disk memory

The dreamlet package creates pseudobulk data from the raw read counts for each sample and cell cluster stored in a SingleCellExperiment object [41]. SingleCellExperiment is Bioconductor's core data class for single cell data, and data from other formats (i.e. Seurat [46]) can easily be converted to it. Dreamlet handles SingleCellExperiment objects to support storing large single cell datasets, either in-memory or on-disk, in a way that is seamless for the end user.

For large-scale studies, loading the entire dataset into memory can be prohibitive. Storing 18K genes across 2M cells with double precision would require 288 Gb memory. If the dataset is sparse so that 80% of the entries are zero, loading the entire dataset as a `sparseMatrix` object requires 58 Gb of data. Yet, instead of loading the entire dataset into memory, the data can remain on-disk and be accessed using an approach that takes advantage of the H5AD file format built on top of the HDF5 format [39]. The zellkonverter package [47] uses a DelayedArray backend to provide a seamless interface to an on-disk H5AD dataset through the interface of the SingleCellExperiment class. This enables any analysis designed for the SingleCellExperiment class to take advantage of on-disk access to large-scale datasets. This can dramatically reduce memory usage while still retaining high performance.

Creating pseudobulk from a large dataset involves summing reads across a set of cells for each gene. While on-disk access to H5AD avoids loading the entire dataset into memory at the same time, creating pseudobulk requires accessing each entry of the dataset. For large datasets, the amount of time required for this step can vary dramatically depending on implementation details. Following extensive experimentation using R, Rcpp and C++ code, we use the beachmat library [48] to summarize each block containing all genes and a subset of cells into pseudobulk. Using dreamlet to compute pseudobulk from real single cell data reduces both compute time and peak memory usage by more than an order of magnitude. For large datasets of hundreds of donors, only dreamlet is tractable without purchasing an expensive high memory machine.

Performance comparison for computing pseudobulk

H5AD files were created from real single nucleus RNA-seq data generated here using nuclei annotated using 12 cell type clusters. Donors were sampled with replacement to evaluate performance on up to 1,000 donors. We compared `aggregateToPseudoBulk()` in our dreamlet R package, `aggregate.Matrix()` in the Matrix.utils R package, `aggregateData()` in the muscat R package, and `pegasus.pseudobulk()` in the pegasus python library [22]. Each method was run 10 times for each condition and the average memory usage and CPU time are shown. Performance was assessed on a compute cluster where each run had access to 144 Gb memory. For a fair comparison, results are shown using only 1 thread to compute the pseudobulk values. (Dreamlet can use more threads to achieve faster performance at the expense of using more memory.)

We note that peak memory usage of native R code can vary dramatically based on available system memory, and when R chooses to invoke garbage collection. Despite substantial fluctuations, especially for memory usage, these performance results are very robust across multiple runs and changes to internal parameters.

Precision-weighted linear mixed models

In widely cited work on differential expression analysis of RNA-seq data, Law, et al [20] demonstrate feasibility of modeling measurement uncertainty in a count response by weighting by the precision (i.e. reciprocal of observation-level variance). Importantly, they show that approximating the log transformed counts using a weighted linear regression can outperform NB regression that models counts explicitly but suffers from poor hypothesis testing for finite sample sizes (since the null distribution of the test statistics relies on asymptotic theory). Our motivation for using weighted linear regression to model transformed counts follows that of Law, et al [20], but becomes more pressing with repeated measures and complex study designs. In previous work on bulk RNA-seq, we have demonstrated that precision weighted linear mixed models are computationally efficient and hypothesis tests have good finite-sample performance in retaining power while controlling the false positive rate [19]. Using generalized linear mixed models (GLMM) such as a negative binomial mixed model, can be very computationally demanding, suffer from convergence issues on real data, and produce poorly calibrated p-values on finite samples.

While Law, et al [20] consider precision weights to model heteroskedasticity of bulk RNA-seq counts, here we consider two levels of precision weights.

Variation in cell counts

The original biospecimen is often composed of thousands of cells of a particular type, or corresponding to a particular empirically defined cell cluster. In order to study the biology of a given cell type, experimental workflows randomly sample single cells for RNA sequencing. Since this is a stochastic process that samples cells from a much larger population, gene expression measurements aggregated across a larger number of cells more precisely

represents expression in the full population of cells. Consequently, the precision of expression measurements of a gene across samples is directly related to the number of cells sequenced per specimen. Statistically, gene expression measurements vary in their precision and are thus heteroskedastic. A standard approach to accounting for heteroskedasticity is to use a weighted regression model.

Consider gene expression observed for specimen i, cell j and gene g indicated by $y_{i,j,g}$. Aggregating expression across n_i cells for specimen i gives an average expression value $\mu_{i,g} = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{i,j,g}$. Since this value is averaged over n_i cells, the sampling variance of the estimated mean is $var(\mu_{i,g}) \propto 1/n_i$, and the corresponding precision is proportional to n_i. Importantly, weighting a regression model to account for heteroskedasticity requires only *relative* weights. Moreover, the relative precision n_i is the same across all genes for specimen i. We use the number of cells from each specimen and each cluster as precision weights for the regression model in the next step.

Variation in read counts

The widely used voom approach [20] estimates precision weights by fitting a linear model for each gene. Voom then smoothes the relationship between the log2 counts per million and the square root residual standard deviation from the model fit of each gene. We have previously extended the voom approach to enable fitting of linear mixed models for estimating precision weights [19]. Here we further extend this work to allow modeling of heteroskedasticity using initial precision weights from the previous step. The resulting estimated precision weights incorporate measurement uncertainty due to stochastic sampling effects at both the level of cells and reads.

Empirical Bayes shrinkage for precision-weighted linear mixed models

For small sample sizes, parameter estimates can have high sampling variance. In seminal work, Smyth [21] developed an empirical Bayes approach that borrows information across genes to estimate the residual variance. The widely used limma package [49] fits a linear model for each gene, performs the empirical Bayes step, and then computes a moderated t-statistic with a modified null distribution. In the case of a linear model, Smyth's empirical Bayes method uses a conjugate prior on the residual variances and assumes they are drawn from an inverse gamma distribution (i.e. precisions are drawn from a scaled chi-squared distribution) with parameters estimated from the data. A key value in this calculation is the residual degrees of freedom.

In the case of a linear model with n samples and p covariates (including the intercept), the residual degrees of freedom (df_r) is simply $n-p$. However, the case of a linear mixed model used here is more complicated. In this case, we show that the residual variance estimates follow a distribution given by a weighted mixture of n chi-squared random variables, where the weights depend on both the data and the estimated model parameters (**Supplementary Information**). We match the expected value of this mixture distribution using a single chi-square and use its degrees of freedom to approximate the df_r of the linear mixed model.

Importantly, this method is exact in the case of a linear model, is approximate for linear mixed models with any number of random effects, and the approximation improves with the sample size.

Analysis of simulated single cell data

We followed the workflow from Cromwell, et al. [15] to estimate parameters from read single cell data and then simulate data based on this. We added our dreamlet software to their existing workflow and used performance metrics as described in their paper.

Analysis of real single nucleus data with permuted disease labels

We used the single nucleus data generated here and randomly permuted the disease labels of each donor while retaining the same fraction of Alzheimer's disease cases and controls within each sample batch. Each method was then run with default parameters.

Single nucleus RNA-seq data generation

Study cohort

Frozen brain tissue samples derived from DLPFC (Brodmann area 9/46) were obtained from the Mount Sinai Brain Bank (MSBB–Mount Sinai NIH Neurobiobank), which holds over 2,000 human brains. Since we wanted to leverage samples from donors with either no discernable neuropathology or cognitive complaints (controls), or with only AD-associated neuropathology, we narrowed our initial selection of brain donors using a combination of neuropathological and clinical criteria inspired by previous work [50,51]. AD samples needed to be classified by (1) CERAD protocol [52] as “AD possible”, “AD probable” or “AD definite”, (2) Braak AD staging protocol [53] within the stage 3-6, and (3) clinical dementia rating [54] within the rate 0.5-5. Furthermore, our subset of AD donors cannot be diagnosed with Parkinson’s disease or Diffuse Lewy body disease. Conversely, control samples needed to be classified by (1) CERAD protocol as “no AD” or “possible AD” and (2) Braak AD staging protocol within the stage 0-2. Additionally, control samples cannot be diagnosed with any other neurodegenerative, neurological or neuropathological diagnosis. Neuropathological assessments, cognitive, medical status and neurological status were performed according to established procedures [55]. All neuropsychological, diagnostic and autopsy protocols were approved by the Mount Sinai and JJ Peters VA Medical Center Institutional Review Boards.

Isolation and fluorescence-activated nuclear sorting (FANS) of nuclei with hashing.

All buffers were supplemented with RNAse inhibitors (Takara). 6 samples were processed in parallel. 25mg of frozen postmortem human brain tissue from each specimen was homogenized in cold lysis buffer (0.32 M Sucrose, 5 mM $CaCl_2$, 3 mM Magnesium acetate, 0.1 mM, EDTA, 10 mM Tris-HCl, pH8, 1 mM DTT, 0.1% Triton X-100) and filtered through a 40 µm cell strainer. The flow-through was underlaid with sucrose solution (1.8 M Sucrose, 3 mM Magnesium acetate, 1 mM DTT, 10 mM Tris-HCl, pH8) and centrifuged at 107,000 g for 1 hour

at 4 °C. Pellets were resuspended in PBS supplemented with 0.5% bovine serum albumin (BSA). Resuspended nuclei were quantified (Countess II, Life Technologies) and 2M from each sample were pelleted at 500 g for 5 minutes at 4˚C and re-suspended in 100 µl staining buffer (2% BSA, 0.02% Tween-20 in PBS). Each sample incubated with 1 µg of a distinct TotalSeq-A nuclear hashing antibody (Biolegend) for 30 min at 4 °C. Prior to FANS, volumes were brought up to 250 µl with PBS and 7-AAD (Invitrogen) added to facilitate detection of nuclei. 7-AAD positive nuclei were collected in tubes pre-coated with 5% BSA using a FACSAria flow cytometer (BD Biosciences).

snRNA-seq and library preparation.
Following FANS, nuclei were washed twice in staining buffer before being re-suspended in 22 µl PBS and quantified. Nuclei concentrations were normalized and equal amounts from each sample were pooled together. 2 aliquots of 60,000 pooled nuclei (i.e. 10,000 each) were processed in parallel using 3' v3.1 reagents (10x Genomics). At the cDNA amplification step (step 2.2), reactions were supplemented with a hash-tag oligo (HTO) cDNA "additive" primer (GTGACTGGAGTTCAGACGTGTGCTCTTCCGAT*C*T; *Phosphorothioate bond). Following cDNA amplification, supernatants from the 0.6x SPRI (Beckman Coulter) selection step were retained for HTO library generation. Otherwise, cDNA libraries were prepared according to the manufacturer's instructions (10x Genomics). HTO libraries were prepared as described previously [56]. All libraries were sequenced at NYGC using the Novaseq platform (Illumina).

Single nucleus data processing and quality control

Alignment and demultiplexing

Sequencing reads from all pools of multiplexed samples were aligned to the hg38 reference genome using STARsolo [57,58]. To assign the cells from each pool to their respective donors, we applied a genotype-based demultiplexing approach followed by genotype concordance. First, cellSNP [59] was used to pile up the expressed alleles from polymorphic sites overlapping snRNA-seq reads. Then, vireo [13] utilized those pile-ups to split cells into clusters corresponding to six distinct donors per pool. The assignment of identity of each cluster of cells to a particular donor was derived from genotype concordance analysis that compared the clusters of cells against reference SNP-array data using QTLtools-mbv [60]. While the majority of pools contained the cells from the expected sets of donors, we leveraged the genotype concordance results to detect and correct occasional sample swaps and mislabelings.

Quality control, UMAP, cell type annotation

QC. After genome alignment and demultiplexing, we applied rigorous three-step QC to remove ambient RNA and retain viable nuclei for downstream analysis. First, the QC was applied at the cell level. A battery of QC tests was performed to filter low-quality libraries and non-viable cells within each library. Poor-quality cells were detected by thresholding based on UMI counts, gene counts, and mitochondrial contents. We also checked for possible contamination from ambient

RNA, a fraction of reads mapped to non-mRNA like rRNA, sRNA, and pseudogenes, as well as known confounding features such as lncRNA *MALAT1*. Further filtering was carried out by removing doublets using the Scrublet method[61]. Second, the QC is applied at the feature level. We removed features (genes) that are not robustly expressed by at least 0.05% of the cells/nuclei. Lastly, the QC was applied at the donor level. We remove donors with less than 50 cell counts, which can introduce more noise to the downstream analysis.

Batch correction. We have developed a tracking platform to record all technical covariates (such as 10x Genomics lot kit number, dates of different preparations, viable cell counts, etc.) and quality metrics derived from data preprocessing. We assessed the correlation between all pairs of technical variables using Canonical Correlation Analysis and used the Harmony method[62] to regress out the effect of sequencing pools before performing clustering and taxonomy analysis.

Clustering. Highly variable features were selected from mean and variance trends, and we used the k-Nearest-Neighbor (kNN) graph calculated on the basis of harmony-corrected PCA embedding space to cluster cells in the same cell-type using Leiden[63] clustering algorithm. We used UMAP[64] for the visualization of resulting clusters.

Cellular taxonomy. Identified cell types will be annotated based on a combination of expert curation and machine-learning-based algorithms to query known gene marker signatures previously curated by Human Cell Atlas.

Sex check. For each donor the labeled sex was checked to be consistent with expression on *XIST* and *UTY* genes on the X and Y chromosomes, respectively.

Code Availability

The dreamlet R package and documentation is available at DiseaseNeuroGenomics.github.io/dreamlet and has been submitted to Bioconductor. Code for simulations is available at github.com/GabrielHoffman/muscat-comparison_v2. Data analysis code and results are available at https://github.com/GabrielHoffman/dreamlet_analysis. The snRNA-seq data generated here is available at https://www.synapse.org/PsychAD_public.

Software versions:

dreamlet v0.0.55,
muscat v1.11.2,
DESeq2 v1.36.0,
limma 3.52.1,
edgeR 3.38.0,
MAST 1.22.0,
pegasus 1.6.0
STARsolo 2.7.9.
cellSNP 1.2.0
vireo 0.5.6
QTLtools-mbv 1.3

Funding

This work was supported by R01AG067025 (to P.R. and V.H.), R01AG065582 (to P.R. and V.H.) and R01AG050986 (to P.R.), P30AG066514 (to V.H.) from the NIA; R01MH109677 (to P.R.), R01MH125246 (to P.R.), RF1MH128970 (to P.R.) and U01MH116442 (to P.R. and V.H.) from NIMH; 75N95019C00049 (to V.H.) from NIDA; U01NS125580 (to P.R. and V.H.) from NINDS; and supplement 3R01AG067025-03S1 (to P.R.) from the Office of Data Science Strategy.

Author Contributions

G.E.H. developed the dreamlet package and performed analysis. V.H. provided tissue specimens from Alzheimer's disease brains and controls. A.H., C.C., M.A., Z.S. and S.A. generated novel snRNA-seq data from Alzheimer's disease brains and controls under the supervision of J.F.F. D.L., J.B., P.F., K.T., S.V., G.V. performed processing of snRNA-seq data from Alzheimer's disease brains and controls generated here. G.E.H. and P.R. supervised analysis. G.E.H., D.L., J.B., J.F.F. and P.R. wrote the manuscript with input from all authors.

References

1. Zeng, B. *et al.* Multi-ancestry eQTL meta-analysis of human brain identifies candidate causal variants for brain-related traits. *Nat. Genet.* **54**, 161–169 (2022).
2. Zeng, H. What is a cell type and how to define it? *Cell* **185**, 2739–2755 (2022).
3. Tanay, A. & Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature* **541**, 331–338 (2017).
4. Finucane, H. K. *et al.* Partitioning heritability by functional annotation using genome-wide association summary statistics. *Nat. Genet.* **47**, 1228–1235 (2015).
5. Maurano, M. T. *et al.* Systematic localization of common disease-associated variation in regulatory DNA. *Science* **337**, 1190–1195 (2012).
6. Farh, K. K.-H. *et al.* Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
7. Jerber, J. *et al.* Population-scale single-cell RNA-seq profiling across dopaminergic neuron differentiation. *Nat. Genet.* **53**, 304–312 (2021).
8. Kfoury, Y. *et al.* Human prostate cancer bone metastases have an actionable immunosuppressive microenvironment. *Cancer Cell* vol. 39 1464–1478.e8 Preprint at https://doi.org/10.1016/j.ccell.2021.09.005 (2021).
9. Nathan, A. *et al.* Multimodally profiling memory T cells from a tuberculosis cohort identifies cell state associations with demographics, environment and disease. *Nat. Immunol.* **22**, 781–793 (2021).
10. Sikkema, L. *et al.* An integrated cell atlas of the human lung in health and disease. *bioRxiv* 2022.03.10.483747 (2022) doi:10.1101/2022.03.10.483747.
11. Oelen, R. *et al.* Single-cell RNA-sequencing of peripheral blood mononuclear cells reveals widespread, context-specific gene expression regulation upon pathogenic exposure. *Nat.*

Commun. **13**, 3267 (2022).

12. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
13. Huang, Y., McCarthy, D. J. & Stegle, O. Vireo: Bayesian demultiplexing of pooled single-cell RNA-seq data without genotype reference. *Genome Biol.* **20**, 273 (2019).
14. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
15. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).
16. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
17. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
18. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
19. Hoffman, G. E. & Roussos, P. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics* **37**, 192–201 (2021).
20. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
21. Smyth, G. K. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, Article3 (2004).
22. Li, B. *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nat. Methods* **17**, 793–798 (2020).
23. 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).

24. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, (2014).
25. Crowell, H. L., Morillo Leonardo, S. X., Soneson, C. & Robinson, M. D. Built on sand: the shaky foundations of simulating single-cell RNA sequencing data. *bioRxiv* 2021.11.15.468676 (2022) doi:10.1101/2021.11.15.468676.
26. Bates, D., Machler, M., Bolker, B. M. & Walker, S. C. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* http://arxiv.org/abs/1406.5823\nhttp://listengine.tuxfamily.org/lists.tuxfamily.org/eigen/2011/06/pdfKU_S0z6LjT.pdf (2015).
27. Pinheiro, J. & Bates, D. *Mixed-Effects Models in S and S-PLUS*. (Springer Science & Business Media, 2000).
28. Wu, D. & Smyth, G. K. Camera: a competitive gene set test accounting for inter-gene correlation. *Nucleic Acids Res.* **40**, e133 (2012).
29. Gaublomme, J. T. *et al.* Nuclei multiplexing with barcoded antibodies for single-nucleus genomics. *Nat. Commun.* **10**, 2907 (2019).
30. Risso, D., Schwartz, K., Sherlock, G. & Dudoit, S. GC-content normalization for RNA-Seq data. *BMC Bioinformatics* **12**, 480 (2011).
31. Aird, D. *et al.* Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries. *Genome Biol.* **12**, R18 (2011).
32. Keren-Shaul, H. *et al.* A Unique Microglia Type Associated with Restricting Development of Alzheimer's Disease. *Cell* **169**, 1276–1290.e17 (2017).
33. Bellenguez, C. *et al.* New insights into the genetic etiology of Alzheimer's disease and related dementias. *Nat. Genet.* **54**, 412–436 (2022).
34. Leng, F. & Edison, P. Neuroinflammation and microglial activation in Alzheimer disease:

where do we go from here? *Nat. Rev. Neurol.* **17**, 157–172 (2021).

35. Asih, P. R. *et al.* Functions of p38 MAP Kinases in the Central Nervous System. *Front. Mol. Neurosci.* **13**, 570586 (2020).
36. Rood, J. E., Maartens, A., Hupalowska, A., Teichmann, S. A. & Regev, A. Impact of the Human Cell Atlas on medicine. *Nat. Med.* **28**, 2486–2496 (2022).
37. Eraslan, G. *et al.* Single-nucleus cross-tissue molecular reference maps toward understanding disease gene function. *Science* **376**, eabl4290 (2022).
38. Tabula Sapiens Consortium* *et al.* The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science* **376**, eabl4896 (2022).
39. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
40. Huber, W. *et al.* Orchestrating high-throughput genomic analysis with Bioconductor. *Nat. Methods* **12**, 115–121 (2015).
41. Amezquita, R. A. *et al.* Orchestrating single-cell analysis with Bioconductor. *Nat. Methods* **17**, 137–145 (2020).
42. Hoffman, G. E. & Schadt, E. E. variancePartition: Interpreting drivers of variation in complex gene expression studies. *BMC Bioinformatics* **17**, (2016).
43. Boni, C., Laudanna, C. & Sorio, C. A Comprehensive Review of Receptor-Type Tyrosine-Protein Phosphatase Gamma (PTPRG) Role in Health and Non-Neoplastic Disease. *Biomolecules* **12**, (2022).
44. Luo, J. *et al.* PTPRG activates m6A methyltransferase VIRMA to block mitochondrial autophagy mediated neuronal death in Alzheimer's disease. *bioRxiv* (2022) doi:10.1101/2022.03.11.22272061.
45. Schwabe, T., Srinivasan, K. & Rhinn, H. Shifting paradigms: The central role of microglia in Alzheimer's disease. *Neurobiol. Dis.* **143**, 104962 (2020).
46. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902.e21

(2019).

47. Zappia, L. & Lun, A. *zellkonverter: Conversion Between scRNA-seq Objects*. (2022).
48. Lun, A. T. L., Pagès, H. & Smith, M. L. beachmat: A Bioconductor C++ API for accessing high-throughput biological data from a variety of R matrix types. *PLoS Comput. Biol.* **14**, e1006135 (2018).
49. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
50. Beckmann, N. D. *et al.* Multiscale causal networks identify VGF as a key regulator of Alzheimer's disease. *Nat. Commun.* **11**, 3942 (2020).
51. Bendl, J. *et al.* The three-dimensional landscape of cortical chromatin accessibility in Alzheimer's disease. *Nat. Neurosci.* **25**, 1366–1378 (2022).
52. Fillenbaum, G. G. *et al.* Consortium to Establish a Registry for Alzheimer's Disease (CERAD): the first twenty years. *Alzheimers. Dement.* **4**, 96–109 (2008).
53. Braak, H., Alafuzoff, I., Arzberger, T., Kretzschmar, H. & Del Tredici, K. Staging of Alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta Neuropathol.* **112**, 389–404 (2006).
54. Morris, J. C. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology* **43**, 2412–2414 (1993).
55. Haroutunian, V., Katsel, P. & Schmeidler, J. Transcriptional vulnerability of brain regions in Alzheimer's disease and dementia. *Neurobiol. Aging* **30**, 561–573 (2009).
56. Stoeckius, M. *et al.* Cell Hashing with barcoded antibodies enables multiplexing and doublet detection for single cell genomics. *Genome Biol.* **19**, 224 (2018).
57. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
58. Kaminow, B., Yunusov, D. & Dobin, A. STARsolo: accurate, fast and versatile mapping/quantification of single-cell and single-nucleus RNA-seq data. *bioRxiv* 2021.05.05.442755 (2021) doi:10.1101/2021.05.05.442755.

59. Huang, X. & Huang, Y. Cellsnp-lite: an efficient tool for genotyping single cells. *Bioinformatics* (2021) doi:10.1093/bioinformatics/btab358.

60. Fort, A. *et al.* MBV: a method to solve sample mislabeling and detect technical bias in large combined genotype and sequencing assay datasets. *Bioinformatics* **33**, 1895–1897 (2017).

61. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).

62. Korsunsky, I. *et al.* Fast, sensitive and accurate integration of single-cell data with Harmony. *Nat. Methods* **16**, 1289–1296 (2019).

63. Traag, V. A., Waltman, L. & van Eck, N. J. From Louvain to Leiden: guaranteeing well-connected communities. *Sci. Rep.* **9**, 5233 (2019).

64. McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv [stat.ML]* (2018).