bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

NEBULA: a fast negative binomial mixed model for differential expression and co-expression analyses of large-scale multi-subject single-cell data

Liang He[1*], Alexander M. Kulminski[1*]

[1] Biodemography of Aging Research Unit, Social Science Research Institute, Duke University, Durham, NC, USA

* Corresponding authors: Liang He, Alexander Kulminski
Email: lh235@duke.edu; alexander.kulminski@duke.edu

Abstract

The growing availability of large-scale single-cell data revolutionizes our understanding of biological mechanisms at a finer resolution. In differential expression and co-expression analyses of multi-subject single-cell data, it is important to take into account both subject-level and cell-level overdispersions through negative binomial mixed models (NBMMs). However, the application of NBMMs to large-scale single-cell data is computationally demanding. In this work, we propose an efficient NEgative Binomial mixed model Using a Large-sample Approximation (NEBULA)), which analytically solves the high-dimensional integral in the marginal likelihood instead of using the Laplace approximation. Our benchmarks show that NEBULA dramatically reduces the running time by orders of magnitude compared to existing tools. We showed that NEBULA controlled false positives in identifying marker genes, while a simple negative binomial model produced spurious associations. Leveraging NEBULA, we decomposed between-subject and within-subject overdispersions of an snRNA-seq data set in the frontal cortex comprising ~80,000 cells from a cohort of 48 individuals for Alzheimer's diseases (AD). We observed that subpopulations and known subject-level covariates contributed substantially to the overdispersions. We carried out cell-type-specific transcriptome-wide within-subject co-expression analysis of *APOE*. The results revealed that *APOE* was most co-expressed with multiple AD-related genes, including *CLU* and *CST3* in astrocytes, *TREM2* and C1q genes in microglia, and *ITM2B*, an inhibitor of the amyloid-beta peptide aggregation, in both cell types. We found that the co-expression patterns were different in $APOE2^+$ and $APOE4^+$ cells in microglia, which suggest an isoform-dependent regulatory role in the immune system through the complement system in microglia. NEBULA opens up a new avenue for the broad application of NBMMs in the analysis of large-scale multi-subject single-cell data.

Word count: 276

Keywords: single-cell RNA-seq, negative binomial mixed model, differential expression analysis, Alzheimer's disease, *APOE*

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Introduction

Single-cell profiling technology has revolutionized our understanding of biology to an unprecedented resolution. With the recent advances in high-throughput single-cell profiling techniques (Hashimshony et al., 2016; Klein et al., 2015, 2015; Picelli et al., 2014), large-scale multi-subject single-cell RNA-seq (scRNA-seq) and ATAC-seq (scATAC-seq) data become increasingly accessible. Notably, through droplet-based technology, tens of thousands of cells from multiple subjects can be sequenced in parallel within a single batch (Klein et al., 2015; Macosko et al., 2015). As an example, a recent large-scale population single-nucleus RNA-seq (snRNA-seq) data in the human frontal cortex involving ~80,000 nuclei from a cohort comprising 24 Alzheimer's disease (AD) patients and 24 healthy controls (Mathys et al., 2019) provide biological information in much finer granularity compared to bulk RNA-seq data. Projects profiling more than one million single cells from hundreds of subjects are underway.

The drastically increasing magnitude of sample size, however, poses a significant computational challenge for efficient analysis of differential expression, expression quantitive trait loci (eQTLs), and co-expression using large-scale scRNA-seq data. This is in contrast to previous statistical models for analyzing bulk RNA-seq data, in which more emphasis is placed upon building a robust estimate under a small sample size (e.g., regularization of standard errors and robust estimation of overdispersion parameters (Law et al., 2014; Love et al., 2014; McCarthy et al., 2012)). Unlike bulk RNA-seq data, multi-subject scRNA-seq data, particularly using the droplet-based technology, often comprise many cells but are featured by high sparsity and a hierarchical structure. Due to the hierarchical nature of multi-subject scRNA-seq data, it is preferable to decompose the overdispersion into the subject and cellular levels. One of the standard approaches to handle hierarchical count responses is a negative binomial mixed-effects model (NBMM) that introduces independent random effects to take into account the overdispersion at both levels. In this paper, we focus on the NBMM rather than a zero-inflated model because multiple recent studies show that a zero-inflated model might be redundant for unique molecular identifiers (UMI)-based single-cell data (Chen et al., 2018; Choi et al., 2020; Hafemeister and Satija, 2019). This is also consistent with the observations for the vast majority of genes in our analysis of real data. The NBMM can increase the statistical power by a more accurate specification of the overdispersion structure, and also helps in eliminating spurious associations to detect marker genes of a cell cluster, as we show in our simulation.

A plethora of estimation methods have been proposed for such a model (Breslow and Clayton, 1993; Lindstrom and Bates, 1990; Ormerod and Wand, 2012; Rue et al., 2009; Tierney and Kadane, 1986; Zhang et al., 2017), and it is the invention of these fast algorithms that popularizes its practical use. However, the computational efficiency of existing tools (Bates et al., 2014; Brooks et al., 2017; Rue et al., 2009) still cannot suffice for the broad application in large-scale multi-subject scRNA-seq data. Due to the intractable marginal likelihood in the NBMM, current estimation algorithms generally rely on a two-layer iteration procedure, which often converges slowly. To address the computational burden, we propose a **NE**gative **B**inomial

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

mixed model **U**sing **L**arge-sample **A**pproximation (NEBULA), a novel fast algorithm for association analysis of scRNA-seq data using an NBMM. As the core idea in NEBULA, we developed an analytical approximation of the high-dimensional integral for the marginal likelihood of the NBMM, by leveraging a large number of cells per subject (CPS), which is a common feature in droplet-based scRNA-seq data. The improvement of computational efficiency is achieved by avoiding high-dimensional optimization and converging within fewer steps.

We demonstrated the efficiency and accuracy of NEBULA through an extensive simulation study. Because NEBULA uses approximation based on large numbers, we explored the questions of how many cells and subjects are required to estimate subject-level and cell-level overdispersions accurately. We also investigated the performance of NEBULA in dealing with low-expressed genes, and controlling the false positives for testing fixed-effects predictors. It should be noted that a predictor of interest can be classified as a subject-level or a cell-level variable. The first category often includes disease or treatment groups, genetic variants, and subject-level covariates such as age, sex, etc, while stages of the cell cycles, subpopulations, and cell-level expression are common variables in the second category. For testing a cell-level variable, it is known that ignoring the subject-level random effects would lead to inflated p-values (Milanzi et al., 2012). In contrast, there is much less literature on the performance of testing subject-level predictors in cell-level data. It is well known that the subject-level random effects need to be included in the model when the predictor of interest is also a subject-level variable. Otherwise, the type I error rate would be inflated due to an underestimated standard error of the effect size. A detailed derivation of such underestimation is given in (Landeghem et al., 2005; Moerbeek, 2004). In this study, we found that testing a subject-level variable was highly sensitive to the estimation of the subject-level variance component. Thus, testing predictors at the two levels might require different estimation strategies.

To date, few studies have investigated the relative contributions of subject-level and cell-level overdispersions in cell-type-specific gene expression. Overdispersion often results from failing to include important explanatory variables or a hierarchical structure in the data (Hilbe, 2011). Using NEBULA, we decomposed gene-specific overdispersion of a large-scale snRNA-seq data in the human frontal cortex from an AD cohort (Mathys et al., 2019). We further explored what factors might contribute to the overdispersions. We showed that NEBULA reduced false positives in selecting cell-type marker genes, especially when the numbers of cells across subjects were unbalanced. As a direct application of NEBULA, we performed a cell-level transcriptomic co-expression analysis of *APOE* to investigate its cell-type-specific regulatory mechanisms in microglia and astrocytes. We further carried out an isoform-specific co-expression analysis for different *APOE* genotypes separately.

Results

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Overview of NEBULA

The variation of raw counts in scRNA-seq data comes from two sources, the Poisson sampling process and overdispersion originating from biological heterogeneity and technical or experimental noise. Cells in large-scale single-cell data are often collected from multiple subjects or batches, leading to an apparent hierarchical structure. To model this strucutre, NEBULA decomposes the total overdispersion into subject-level (i.e., between-subject) and cell-level (i.e., within-subject) components using a random-effects term parametrized by σ^2 and the overdispersion parameter ϕ in the negative binomial distribution. The subject-level overdispersion captures the contributions from e.g., eQTLs, technical artifacts, batch effects, and other subject-level covariates. The cell-level overdispersion reflects the variation of gene expression due to, e.g., cell cycle, sub-cell-type composition, cell-level covariates such as ribosome RNA fraction. The estimation of an NBMM, in general, is computationally intensive, especially for large sample size because the marginal likelihood is intractable due to high-dimensional integrals of the random effects. A standard estimation method for generalized linear mixed models (GLMMs) using penalized quasi-likelihood (PQL) (Breslow and Clayton, 1993) involves a two-step iteration procedure where the variance of the random effects is estimated first, and then it is used to estimate the coefficients of the variables of interest. In genome-wide association studies (GWAS), the variance component is often estimated only once under the null model. In contrast, such a two-step procedure has to be carried out for each of tens of thousands of genes in scRNA-seq data or peak regions in scATAC-seq data, which is considerably more challenging. NEBULA achieves its speed gain by using an approximate analytical marginal likelihood, which was derived leveraging the fact that the data for each subject often include a large number of cells. This approximation, referred to as NEBULA-LN, eliminates the computationally-demanding two-layer high-dimensional optimization procedure, and therefore substantially reduces the computational time for estimating the overdispersions. For situations in which NEBULA-LN fails to achieve accurate estimates of the subject-level overdispersion, we estimate the cell-level overdispersion using NEBULA-LN and implemented a fast estimation procedure, referred to as NEBULA-HL, based on a hierarchical likelihood (h-likelihood) to estimate the subject-level overdispersion. Because only one overdispersion is needed to be optimized in the h-likelihood, NEBULA-HL converges in much fewer steps.

NEBULA is computationally efficient

We compared NEBULA with the existing statistical tools for estimating NBMMs to evaluate its computational efficiency. In addition to NEBULA-LN, we assessed a combination of NEBULA-LN and NEBULA-HL (denoted by NEBULA-LN+HL), in which we first used NEBULA-LN to estimate the cell-level overdispersion, and then fixed this estimate in NEBULA-HL to obtain the subject-level overdispersion. We chose the L-BFGS algorithm from (Ypma, 2014) for optimization in NEBULA-LN. We found that a Newton-Raphson (NR) algorithm (Dennis and Schnabel, 1996) was generally faster in this case but less robust in terms of convergence. For comparison, we included the *glmer.nb* function from the *lme4* R package (Bates et al., 2014), the *glmmTMB* function from the *glmmTMB* R package (Brooks et al., 2017), and the *inla* function from the *INLA* R package (Rue et al., 2009). The *glmmTMB* R package utilizes the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Template Model Builder automatic differentiation engine, and the *INLA* R package is an efficient Bayesian framework based on integrated nested Laplace approximations. In *glmer.nb*, we assessed the default setting (nAGQ=1), which is based on Laplace approximation (LA) (Tierney and Kadane, 1986), and a faster but less accurate setting (nAGQ=0), which is based on a penalized likelihood (Lindstrom and Bates, 1990). The difference is that LA estimates the fixed effects in the marginal likelihood with the variance components, while the latter estimates the fixed effects with the random effects in the penalized likelihood. We did not include methods using adaptive Gaussian quadrature (AGQ) (*glmer.nb* with nAGQ>1) (Pinheiro and Bates, 1995; Pinheiro and Chao, 2006) or Bayesian methods using Markov chain Monte Carlo (MCMC) (Vestal et al., 2020) because of their computational intensity. In addition, the Gaussian variational approximation method (Ormerod and Wand, 2012) is promising, but has not been implemented for the NBMM.

Given a simulated data set comprising 50 subjects and CPS=200 (a total of ~10,000 cells), it took ~400s for NEBULA-LN and ~1000s for NEBULA-LN+HL to analyze 10,000 genes (Fig. 1A). These benchmarks were on average >100-fold and ~50-fold faster than *glmer.nb* with nAGQ=0, respectively. It took much longer for *glmer.nb* with the default setting and *glmmTMB*. We failed to obtain the full benchmarks for *inla* because it took an extremely long time to finish in some scenarios. Generally, the running time of *inla* was between *glmmTMB* and the default *glmer.nb*. The computational time of all methods increased almost linearly with respect to CPS, which was expected because the random effects are assumed to be independent. We also observed that the computational time of all methods scaled in a similar trend with the increasing number of subjects (Fig. 1B). In all settings, NEBULA-LN+HL was ~2-3x slower than NEBULA-LN but was still much faster than *glmer.nb*. In our following analysis of ~34,000 excitatory neurons from 48 individuals in the snRNA-seq data adopted from (Mathys et al., 2019), NEBULA accomplished an association analysis of ~16,000 genes for identifying marker genes in ~40 minutes, compared to ~67 hours using *glmer.nb* with nAGQ=0. When the number of fixed-effects predictors increased from two to ten, the computational time of both NEBULA and *glmer.nb* with nAGQ=0 grew modestly, while the default *glmer.nb* increased by ~10x (Fig. S1). This substantial increase of the computational time in the default *glmer.nb* is expected because it estimates the fixed effects with the overdispersions in the outer layer of the iterative procedure using a derivative-free method. The number of iterations of such an optimization method rises drastically when the dimension of the parameter space increases.

Decomposing overdispersions using NEBULA

We show in Supplementary materials A.5 that the approximation proposed in NEBULA-LN still provides asymptotically consistent estimates. Nevertheless, it is also crucial to assess its practical performance under a finite sample size. As the above benchmarks indicated that NEBULA-LN had a huge speed advantage over NEBULA-HL, the basic strategy in NEBULA is to first fit the data using NEBULA-LN and apply NEBULA-HL only in situations where NEBULA-LN performs poorly. We, therefore, conducted a comprehensive simulation study to investigate under which situations NEBULA-LN could produce sufficiently accurate estimates. The estimation accuracy of NEBULA-LN is primarily affected by CPS, the magnitude of the cell-level

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

dispersion, and the coefficient of variation (CV) of the scaling factor. We thus focus on determining a threshold in terms of these parameters. Because NEBULA-HL was based on a standard h-likelihood method, we used NEBULA-HL as a reference to assess the accuracy of NEBULA-LN in terms of mean square error (MSE).

We first evaluated the performance of NEBULA-LN in a well-designed study where the numbers of cells were highly homogeneous across the subjects and were assumed to follow a Poisson distribution. We considered the CPS value varying from 100 to 800, and the number of subjects ranging from 30 to 100, which are common settings in real droplet-based scRNA-seq data. We simulated counts based on an NBMM with the subject-level overdispersion σ^2 ranging from 0.01 to 1 and the cell-level overdispersion ϕ^{-1} ranging from 0.01 to 100 (A larger ϕ^{-1} means higher overdispersion). These ranges were selected because they were close to our estimates observed in the real snRNA-seq dataset in (Mathys et al., 2019). We simulated genes of different expression level by tuning β_0 from -4 to 2 (When the subject-level overdispersion σ^2 is small, $\exp(\beta_0)$ approximately equals the counts per cell (CPC) defined by the total counts of a gene divided by the total number of cells). We considered the following two situations, (i) all cells share the same scaling factor and (ii) the scaling factor across cells varied substantially with the coefficient of variation equal to 1.

Fig. 2A shows that the MSE of NEBULA-LN for estimating the cell-level overdispersion ϕ^{-1} decreased uniformly with increasing CPS. The convergence of the estimates to their true values supported our theoretical conclusion that NEBULA-LN is asymptotically consistent for estimating ϕ. The MSEs of both methods increased when the mean expression of a gene (measured by CPC) dropped. Both methods struggled to accurately estimate ϕ for low-expressed genes when CPC≈0.018, and the bias of NEBULA-LN became more noticeable. This suggests that genes with a low CPC value could be filtered during the quality control (QC) because it is difficult to obtain accurate estimates of the overdispersions for these genes due to the high sparsity of the data unless the number of cells is very large. In most scenarios, NEBULA-LN showed higher MSEs and was asymptotically less efficient than NEBULA-HL. Nevertheless, the MSEs between the two methods were comparable when CPS was >400. When ϕ was very large (=100), NEBULA-LN produced a more biased estimate if CPS is small. In addition, Fig. S3 shows that increasing the number of subjects alone improved the standard error, but not the bias of $\hat{\phi}$, suggesting that the bias of $\hat{\phi}$ depended mainly on CPS. A larger CV of the scaling factor showed little impact on the MSE of NEBULA-LN (Fig. S2A).

On the other hand, NEBULA-LN had similar or lower MSEs than NEBULA-HL for estimating the subject-level overdispersion when σ^2 is large (≥0.5) (Fig. 2B). However, when σ^2 was small, NEBULA-LN began to produce CPC-dependent biased estimates when ϕ turned larger (Fig. 2B). Increasing CPS alleviated the bias rapidly (Fig. 2D). When the CV of the scaling factor increased to one, an approximately 2-fold increase of $\mathrm{CPS}\cdot\phi$ was required to achieve similar MSEs by comparing Figs. S2B, S2D with Figs. 2B, 2D. Combining these observations suggested that the accuracy of NEBULA-LN for estimating σ^2 approximately depends on $\kappa = \mathrm{CPS}\cdot\phi/(1+c^2)$ (c is the CV of the scaling factor), which is derived in the Methods section. These empirical results suggested that κ determined the lower bound of σ^2 for which an

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

accurate estimate could be achieved. NEBULA-LN was able to obtain good estimates for σ^2 as low as 0.02 when $\kappa = 200$ (corresponding to $\phi^{-1} = 2$ in Fig. 2B and Fig. 2C), which was sufficient to reliably test cell-level predictors, as we will show in the next section. Both NEBULA-LN and NEBULA-HL underestimated σ^2 particularly when the number of subjects was small (Fig. 2C). We found that the underestimate was not specific to NEBULA, but also present in other tools based on PQL (e.g., the *lme4* R package with nAGQ=0). Increasing the number of subjects alleviated this bias, and the influence of such bias on testing subject-level preditors will be discussed in the next section.

We further evaluated the performance of NEBULA in unbalanced samples, where the numbers of cells varied substantially across the individuals. An unbalanced sample is common when analyzing a cell subpopulation obtained from clustering. For example, in the snRNA-seq data in (Mathys et al., 2019), several individuals had only ~20 inhibitory neurons, although the CPS value among the 48 individuals was ~200. Overall, the results from the unbalanced (Fig. S17) were comparable to those from the balanced sample (Fig. 2), suggesting the distribution of the number of cells had little impact on estimating σ^2 and ϕ.

NEBULA controls type I error rate

As shown above, either NEBULA-LN or NEBULA-HL might yield somewhat biased estimates of the overdispersions in specific scenarios. We then assessed the impact of such biases on controlling the type I error rate by testing the fixed-effects predictors under a wide range of settings. As mentioned above, predictors can be classified as either a cell-level or subject-level variable. For a cell-level predictor, Fig. 3A and Fig. 3B show that both methods controlled the type I error rate well in almost all scenarios. The type I error rate was only slightly inflated when both overdispersions were very large, and the CPS value was as small as 100 (i.e., the bottom-right panels in Fig. 3A and Fig. 3B). Although ϕ^{-1} was underestimated in NEBULA-LN more for low-expressed genes than high-expressed genes (Fig. 2A), no inflation of the type I error rate was observed for low-expressed genes (Fig. S7). Besides, biased estimates of the subject-level overdispersion in NEBULA-LN had little effects on testing a cell-level predictor.

In contrast, the type I error rate of testing a subject-level predictor was highly sensitive to a biased estimate of the subject-level overdispersion σ^2. The type I error rate in both methods was inflated in the case of 30 subjects and the inflation diminished with the increasing number of subjects (Fig. 3C, Fig. 3D). This is because both methods underestimated the subject-level overdispersion when the number of individuals was small (Fig. 2C). The comparison between Fig. 2C and Fig. 3C suggested that even a small underestimated subject-level overdispersion would lead to an inflated type I error rate of testing a subject-level predictor when the number of subjects is small. Combining Fig. 2B and Fig. S6 suggested that NEBULA-LN had inflated or deflated type I error rate of testing a subject-level predictor whenever σ^2 was overestimated or underestimated, respectively, even for σ^2 being as low as 0.01. Underestimates of the subject-level overdispersion were not specific to NEBULA but also observed in *glmer.nb* with a penalized likelihood (nAGQ=0). These results suggested that NEBULA or other estimation

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

methods based on a penalized likelihood should be used with caution when testing a subject-level predictor, particularly when the number of subjects is small (<100).

We also evaluated the performance of a Poisson mixed model for testing a subject-level predictor. If we ignore the cell-level overdispersion ϕ^{-1} in NEBULA, we end up with a Poisson gamma mixed model (PGMM) (Sutradhar and Qu, 1998), the likelihood function of which has a simple analytical form. Therefore, it does not suffer from the problem of underestimating σ^2 as in NEBULA-HL and is computationally much faster than the Poisson mixed model estimated by *glmer.nb* with nAGQ=10 adopted in (Mathys et al., 2019). Because the cell-level overdispersion is ignored in the PGMM, it definitely cannot be used to test a cell-level predictor. However, it is still interesting to check its type I error rate when used for testing a subject-level predictor. Fig. S15 shows that overall the PGMM controlled the type I error rate well in scenarios where the number of subjects was 100, except for situations where both the cell-level and subject-level overdispersions were large ($\phi^{-1} \geq 100$ and $\sigma^2 \geq 0.5$). But such situations account for a very small fraction of genes (usually low-expressed genes) in real data which are often excluded during quality control. For instance, we observed <0.1% genes with CPC>0.1% having $\phi^{-1} > 100$ in the major cell types in the snRNA-seq data in (Mathys et al., 2019), and the percentage became almost zero when we filtered out low-expressed genes (CPC<1%) during quality control. Therefore, the PGMM might be considered as a fast tool for testing subject-level predictors followed by a second-round sensitivity analysis for top signals if the computational intensity is a major concern.

Based on these simulation results, the strategy that we recommend for testing a cell-level predictor, as summarized in Table S3, is to fit the data first using NEBULA-LN to obtain estimates $\hat{\phi}$ and $\hat{\sigma}^2$. Both overdispersion parameters are re-estimated using NEBULA-HL if $\hat{\phi} < 20/\mathrm{CPS}$ because biased estimates of both overdispersions might occur in this case. Otherwise, we fix $\hat{\phi}$ and only use NEBULA-HL to estimate σ^2 if $\hat{\phi} < \kappa(1 + c^2)/\mathrm{CPS}$, where we chose $\kappa = 200$ based on the simulation results. We report the estimates directly from NEBULA-LN if $\hat{\phi} > \kappa(1 + c^2)/\mathrm{CPS}$ because in this case, NEBULA-LN can produce a reliable estimate for a sufficiently small σ^2. We used this strategy in our following analysis of the snRNA-seq data in the human frontal cortex (Mathys et al., 2019), and adopted the cell library size (the total number of reads) as the scaling factor. Under this strategy, ~90% and ~80% of genes with CPC>0.1% were analyzed using NEBULA-LN alone in the excitatory neurons (CPS=~700) and oligodendrocytes (CPS=~380). We found that c^2 was ~0.8 for the neurons and ~0.25 for all other cell types. The larger CV of the cell library size in the neurons was probably because the neurons consisted of highly heterogeneous subpopulations (Fig. S8). On the other hand, for testing a subject-level predictor, we recommend using $\kappa = 1000$ or even larger.

Subpopulation heterogeneity and known covariates contribute to overdispersions

The subject-level and cell-level overdispersion parameters likely reflect some intrinsic biological features of a gene. A gene with a larger subject-level overdispersion might indicate that the

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

gene is strongly regulated by subject-level factors such as age, sex, disease status, exposure, and genetic variation. We then dissected the overdispersions using NEBULA for genes in the snRNA-seq data in (Mathys et al., 2019). We carried out the analysis in each of the six major cell types, including excitatory neurons, inhibitory neurons, oligodendrocytes, astrocytes, oligodendrocyte progenitor cells (OPCs), and microglia.

We observed that the mean cell-level overdispersion grew steadily with lower expression of a gene in the excitatory neurons (Fig. 4A) and the other cell types (Fig. S11). This mean-overdispersion trend is reminiscent of that observed in bulk RNA-seq data, in which low-expressed genes often show higher variance (Law et al., 2014). Less than 0.1% of genes in the excitatory neurons had an estimated cell-level overdispersion $\hat{\phi}^{-1}$ larger than 100, and all of them were low-expressed genes. In contrast, the subject-level overdispersion exhibited a similar mean-overdispersion trend, except that the overdispersion leveled off for abundantly expressed genes in the case of CPC>1 in the excitatory neurons (Fig. 4B). Similar patterns were observed in the other cell types (Fig. S12). We observed a higher variation for both estimated overdispersion in the lower-expressed genes due to the larger standard error for those genes, as shown in the simulation (Fig. 2). There was no clear correlation between the subject-level overdispersion and the cell-level overdispersion (Fig. 4C).

We then explored which factors might contribute to both overdispersions of a gene in each of the cell types. First, we found that the median cell-level overdispersion dropped markedly in the subpopulations (Fig. S13A), suggesting that sub-cell-type differential expression was the major source for the cell-level overdispersion. On the other hand, the median subject-level overdispersion dropped to a much smaller extent (Fig. S13B). We then investigated subject-level covariates, including age, sex, AD diagnosis, and race, and cell-level covariates, including the number of total features of a cell and percentage of ribosomal protein genes by adding them to the regression model. We found that the overall subject-level overdispersion dropped substantially in most of the cell types after adjustment for these covariates, but the cell-level overdispersion almost remained at the same level (Fig. 4D, Fig. 4E). These results suggested that the subject-level overdispersion was largely attributed to these factors, while the cell-level overdispersion resulted from subpopulation heterogeneity and other unknown factors, e.g., drop-out rates.

To compare different NBMMs, we applied *glmer.nb* in the *lme4* R package to the same data. The estimated cell-level overdispersions were highly comparable between both methods (Fig. S14), except for an outlier in *glmer.nb*. The estimated subject-level overdispersions were also highly correlated for those genes with a small to moderate subject-level overdispersion (Fig. S9), although it is not straightforward to compare the absolute values because the random effects in these two NBMMs are assumed to follow slightly different distributions (See the Methods section). The larger deviation between NEBULA and *glmer.nb* observed for higher subject-level overdispersion was expected, as previously shown that the two models are almost the same when the subject-level overdispersion is small (Sutradhar and Qu, 1998). Estimated fixed effects were also consistent between NEBULA and *glmer.nb* (Fig. S4).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Identifying sub-cell-type marker genes using NEBULA

Using a negative binomial regression without the subject-level random effects could result in substantially inflated false positives (Milanzi et al., 2012). Disproportionate cell frequency among individuals may exacerbate the inflation of spurious associations, similar to Simpson's Paradox (Simpson, 1951). In this case, individuals can be a confounding factor between gene expression and cell clusters. On the other hand, adding individuals as fixed effects would lead to too many parameters in the model, especially when the number of individuals is large. Therefore, an NBMM is an ideal approach for the identification of marker genes for subpopulations in multi-subject scRNA-seq data.

For the demonstration, we used the 3386 astrocytes from the snRNA-seq data in the frontal cortex, which were classified into four subclusters (AST0 - AST3). Some of the subclusters were distributed across the subjects in a highly unbalanced pattern. For example, the vast majority of the cells in the AST3 cluster came from only four subjects (Fig. S10). We simulated the counts of gene expression using an NBMM under the null hypothesis (i.e., not associated with any of the subclusters) with both subject-level and cell-level overdispersions. We assessed the false-positive rates (FPR) under four scenarios with the variance component of the subject-level overdispersion ranging from 0 to 0.1. As we can see in Fig. 5A, even with the subject-level overdispersion σ^2 as small as 0.05, the FPR started to show inflation. The inflation rose rapidly with the increased subject-level overdispersion. In contrast, NEBULA controlled the FPR very well.

To evaluate how this inflation of FPR might affect the selection of marker genes, we carried out association analysis between the gene expression and the four subclusters in the astrocytes in the snRNA-seq data (Mathys et al., 2019) using a negative binomial regression and NEBULA. Among the top 30 genes identified by a negative binomial regression, most of the p-values were consistent between the two methods in the subclusters AST2 and AST3. However, some of the top genes in the subclusters AST0 and AST1, such as *RASGEF1B* and *LINGO1*, were much less significant in NEBULA than in the negative binomial regression (Fig. 5B). Both genes showed uniformly strong subject-level overdispersion in the astrocytes (Fig. 5B) and the other five cell types, indicating that the highly significant p-values in the negative binomial regression could result from failing to account for the subject-level overdispersion. Furthermore, an independent cell-type-specific bulk RNA-seq data (Zhang et al., 2016) showed that the expression of *RASGEF1B* and *LINGO1* was much lower in astrocytes than that in microglia and oligodendrocytes, respectively. Given the evidence, it should be cautious about classifying these genes as subcell-type marker genes based on the results from the negative binomial regression without the random effects. Besides, *GJA1* was identified by NEBULA as a marker gene of AST1 but not AST0, while a negative binomial regression identified *GJA1* as significant in both subclusters (Fig. 5B).

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Cell-level co-expression analysis of *APOE*

Estimating co-expression between genes is crucial for building interaction or regulatory networks and detecting biological pathways. However, co-expression estimated from bulk RNA-seq data is heavily affected by many hidden confounders such as cell type composition and experimental noise. In contrast, scRNA-seq data opened up new possibilities for investigating co-expression at the cellular level. We interpret the cell-level co-expression as a co-occurrence of two genes in a cell. Because cells with larger sequencing depth or lower ribosomal protein gene expression had higher co-occurrence rate, we included the library size as the normalizing factor, and the number of features and percentage of ribosomal protein genes as the covariates. As *APOE* was abundantly expressed in astrocytes and microglia in the frontal cortex, we identified genes co-expressed with *APOE* in these two cell types by utilizing NEBULA. We normalized the raw *APOE* UMI counts by the library size of each cell and subtracted the mean *APOE* expression of each individual so that the centered normalized *APOE* expression did not correlate with subject-level covariates.

We observed that *APOE* expression in astrocytes was most correlated with *CLU*, one of the top genes from AD GWAS (Harold et al., 2009), on top of other AD-related genes such as *CST3* (Deng et al., 2001) (Table 1). In microglia, *APOE* was co-expressed with multiple immune-related genes, particularly *TREM2* and those involved in the complement system including *C1QA*, *C1QB*, and *C1QC* (Table 1). *ITM2B*, an inhibitor of the amyloid-beta peptide aggregation, was strongly co-expressed with *APOE* in both cell types. It should be noted that the identified co-expression was based on the within-subject expression, thus ruled out effects from subject-level confounders such as AD status. We then carried out an isoform-specific co-expression analysis by dividing cells into *APOE e3e3*, *APOE e4*$^{+}$, and *APOE e2e3*. The co-expression patterns showed substantial differences among the isoform groups in microglia, but not in astrocytes (Table S1). More specifically, strong co-expression was observed in the *APOE e2e3* cell population for *CST3*, and the three genes (*C1QA*, *C1QB*, *C1QC*) in the complement system, but not for *TREM2* (Table S2), which might suggest different biological mechanisms in microglia behind these two *APOE* isoforms.

Discussion

In this work, we propose a fast NBMM for the association analysis of large-scale multi-subject single-cell data. We evaluated and demonstrated the performance of NEBULA using a comprehensive simulation study and a droplet-based snRNA-seq data set in the human frontal cortex. Overall, our results show that combining methods based on the approximated marginal likelihood and the h-likelihood, NEBULA managed to achieve significant speed gain and also practically preserve estimation accuracy for analyzing scRNA-seq data. We expect that NEBULA can also be appropriate for other single-cell data such as scATAC-seq data.

The asymptotic efficiency and estimation accuracy of NEBULA-LN are primarily determined by CPS and the cell-level overdispersion. Specifically, our results suggest that $\phi \cdot \mathrm{CPS} > 20$ was

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

sufficient to produce a highly accurate estimate of the cell-level overdispersion. This means that genes with a cell-level overdispersion <10 in a single-cell data set with ~200 cells per subject can benefit from NEBULA-LN. This criterion covers the vast majority of genes in many large-scale real data. On the other hand, a larger CPS value enables a more accurate estimate for a smaller subject-level overdispersion. Accurate estimation requires a larger $\phi \cdot \mathrm{CPS}$ for the subject-level overdispersion than the cell-level overdispersion and also depends on the CV of the scaling factor. Compared to a standard h-likelihood method, we see that NEBULA-LN sacrifices some asymptotic efficiency in estimating the cell-level overdispersion, which means that more cells and counts per cell are needed to achieve the same MSEs. This is the price to pay for the gain of computational efficiency.

NEBULA is solid in controlling the FPR for testing cell-level variables such as subpopulation marker genes. The simulation study suggested that although the estimated overdispersions for low-expressed genes were often less accurate, there was little influence on the FPR. It is possible to treat subjects as batch effects and regress them out in identifying marker genes, but this approach would compromise statistical power if the number of subjects is large. NEBULA is a powerful tool for robustly identifying marker genes. On the other hand, our simulation study suggests that testing a subject-level predictor is sensitive to the estimate of the subject-level overdispersion. The downward bias of NEBULA-HL in estimating the subject-level overdispersion is in line with theoretical results for binary data (Breslow and Lin, 1995) and count responses (Lin, 2007). This bias is more severe in genes with very low expression, probably because the LA works poorly when the response is sparse. Therefore, it should be cautious about using a penalized likelihood method or a first-order LA in an NBMM for investigating a subject-level predictor, particularly when the number of subjects is small. An empirical simulation study indicates that LA may work better for count data than binary data, and the bias decreases as the cluster size increases (Joe, 2008). AGQ can alleviate this issue by producing less biased estimates, but it is computationally more intensive (Joe, 2008; Rabe-Hesketh et al., 2002). If the computational efficiency is a major concern, we found that using the PGMM followed by a sensitivity analysis for top findings could be a fast and viable option for testing a subject-level predictor. Another strategy for testing a subject-level variable is to pool all cells of each subject and treat the data as a bulk RNA-seq dataset. One potential disadvantage of this strategy is that it cannot adjust for cell-level covariates such as mitochondrial and ribosomal gene concentration to increase the statistical power.

The trend of higher average cell-level overdispersion in low-expressed genes is in line with those reported in bulk RNA-seq data (Law et al., 2014; Love et al., 2014). A justification for this observation unique to scRNA-seq data can be that low-expressed genes might have a higher drop-out rate or have larger variation across subpopulations. In bulk RNA-seq data, a large proportion of overdispersion results from heterogeneous cell type composition across the individuals. While this heterogeneity is much better controlled in the scRNA-seq data, subpopulation composition still accounts for most of the cell-level overdispersion. The observed smaller cell-level overdispersion in abundant genes might be attributed to lower drop-out rate and stable expression across subpopulations. In contrast, the subject-level overdispersion is primarily explained by known covariates. With increased gene expression, the cell-level

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

overdispersion drops while the subject-level overdispersion stabilizes. More work is needed to examine to what extent the subject-level overdispersion is explained by the effects of *cis*-eQTLs and whether eGenes show a larger subject-level overdispersion.

The co-expression of *APOE* in astrocytes and microglia is intriguing. Among the top co-expressed genes, multiple SNPs in *CLU* and *TREM2* are AD GWAS loci (Guerreiro et al., 2013; Harold et al., 2009; Jonsson et al., 2013; Lambert et al., 2009). *CLU, CST3*, and *ITM2B* are involved in regulating beta-amyloid production or its fibril formation (Bell et al., 2007; Kaeser et al., 2007; Kim et al., 2008; Matsubara et al., 1995; Matsuda et al., 2005). *CST3* and *ITM2B* are culprits implicated in hereditary cerebral amyloid angiopathy (Revesz et al., 2009). The co-expression results in astrocytes suggest that *APOE* might be involved in the same biological pathway of these genes regulating beta-amyloid production. On the other hand, the co-expression results in microglia support a recent finding (Yin et al., 2019) that *APOE* is involved in the immune system through the complement system, particularly C1q. We further found that this co-expression pattern is *APOE2*$^+$ and *APOE4*$^+$ dependent in microglia, which might provide more insights into the different role of *APOE* isoforms in AD. These results suggest that the effect of *APOE* on AD is related to regulating both beta-amyloid deposition and the immune system. It should be noted that, because of the data sparsity in the scRNA-seq data, a drawback of this co-expression analysis is that there is little statistical power to identify low-expressed genes because the gene counts have very small variation. Therefore, no observed correlation does not necessarily mean that the two genes are not co-expressed. Interpretation of the co-expression results for low-expressed genes should be cautious.

Methods

The model specification in NEBULA

Consider a raw count matrix of n cells from m individuals in a single-cell data set. We choose to use a negative binomial model to take into account the uncertainty originating from the Poisson sampling process, biological effects, and technical noise. Denote by y_{ij} the raw count of a gene and by $\boldsymbol{x}_{ij} = (x_{ij0}, \dots, x_{ijk}) \in \mathbb{R}^{1\times(k+1)}$ a row vector of k fixed-effects predictors and the intercept ($x_{ij0} = 1$) of cell j ($j \in \{1, \dots, n_i\}$ and $\sum_i n_i = n$) in individual i ($i \in \{1, \dots, m\}$). We model the raw count using the following NBMM

$$y_{ij} \sim NB(\mu_{ij} = \pi_{ij}\exp({\boldsymbol{x}_{ij}}^T\boldsymbol{\beta} + \log(\omega_i)), \phi), \quad (1)$$

where NB is a negative binomial distribution parametrized by its mean μ and cell-level dispersion parameter ϕ such that the variance is $\mu + \mu^2/\varphi$, $\boldsymbol{\beta} \in \mathbb{R}^{(k+\mathbf{1})\times\mathbf{1}}$ is a vector of the coefficients of the predictors, ω_i are subject-level random effects capturing the between-subject overdispersion, and π_{ij} is a known cell-specific scaling factor reflecting the sequencing depth, which can be the library size or some global normalizing factor (e.g., (Robinson and Oshlack, 2010)). Note that π_{ij} can also be gene-specific as recent studies show that a single normalizing factor might be insufficent to normalize scRNA-seq data (Bacher et al., 2017). The negative

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

binomial log-linear mixed model (NBLMM) proposed in (Booth et al., 2003) assumes a log-normal distribution for the random effects ω_i, i.e.,

$$\omega_i \sim lognormal(0, \sigma^2) \text{ or } \ln(\omega_i) \sim N(0, \sigma^2), \quad (2)$$

where σ^2 is the variance component of the between-subject random effects.

The estimation of the NBLMM poses a computational challenge because the marginal likelihood is intractable after integrating out the random effects ω_i. An estimation method based on an Expectation-Maximization (EM) algorithm is proposed in (Booth et al., 2003). As the NBLMM falls in the framework of the GLMM, conventional estimation methods for the GLMM can be applied, including penalized likelihood-based methods (Breslow and Clayton, 1993; Lindstrom and Bates, 1990), Laplace approximation (Tierney and Kadane, 1986), AGQ (Pinheiro and Bates, 1995; Pinheiro and Chao, 2006). These methods, implemented in the *lme4* R package (Bates et al., 2014) through the argument *nAGQ*, differ in how to approximate the high-dimensional integral in the marginal likelihood, among which the penalized likelihood-based method is the fastest but the least accurate while AGQ is more accurate and much slower. We refer to (Tuerlinckx et al., 2006) for more extensive reviews of these methods. More recently, (Zhang et al., 2017) proposes an estimation procedure for the NBLMM by transforming the problem into repeatedly fitting a linear mixed model using iteratively reweighted least square. All these methods involve at least a two-layer optimization procedure. Although the time complexity is $\mathcal{O}(n)$ under the assumption of the independent random effects, the two-layer procedure often requires hundreds of iterations to converge, which becomes the major computational bottleneck. On the other hand, Bayesian methods such as MCMC (Vestal et al., 2020) offer better accuracy, but are even slower than the above methods.

The rationale behind NEBULA is to skip the time-consuming two-layer optimization algorithm for as many genes as possible and to significantly reduce the number of iterations if such an algorithm has to be applied. The key idea is to derive an approximate marginal likelihood in a closed-form so that the model can be estimated using a simple NR or quasi-Newton method. Inspired by (Sutradhar and Qu, 1998), instead of the log-normal distribution in eq. (2), we assume that the between-subject random effects follow the following gamma distribution

$$\omega_i \sim Gamma(\alpha, \lambda), \quad (3)$$

$$\alpha = \frac{1}{\exp(\sigma^2) - 1},$$

$$\lambda = \frac{1}{\exp(\sigma^2/2)(\exp(\sigma^2) - 1)},$$

where α and λ are the shape and rate parameters, respectively. One advantage of this parametrization is that it matches the first two moments to the NBLMM, so that even in a situation where eq. (2) is the true distribution of the random effects, eq. (3) can still provide an accurate estimate of σ^2 when it is not large (Sutradhar and Qu, 1998). Previous studies (Milanzi et al., 2012; Neuhaus and McCulloch, 2011; Neuhaus et al., 2013) indicate that it has practically little influence on the estimation of the fixed-effects predictors to assume a lognormal, gamma, or even misspecified distributions for the random effects, which is also confirmed in our real data analysis (Fig. S4). We call the NBMM with the random effects in eq. (3) as a negative binomial gamma mixed model (NBGMM). Under the NBGMM, the marginal mean and variance of y_{ij} are given by

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

$$E(y_{ij}) = \pi_{ij}\exp\left(\boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + \frac{\sigma^2}{2}\right),$$
$$Var(y_{ij}) = E(y_{ij}) + E^2(y_{ij})(\exp(\sigma^2) - 1) + E^2(y_{ij})\frac{\exp(\sigma^2)}{\phi}, \quad (4)$$

from which we see the first overdispersion term in eq. (4) is controlled by the between-subject variance component σ^2 alone, and the other includes both σ^2 and ϕ (The derivation can be found in Supplementary materials A.1). In general, the marginal likelihood of the NBGMM after integrating out ω_i cannot be obtained analytically. However, we show in the next section how to achieve an approximated closed form of the marginal likelihood when n_i is large. We first focus on the derivation in an asymptotic situation, and then investigate its performance in a finite sample.

Approximation of marginal likelihood in NEBULA-LN

We switch the order of the integral in the marginal likelihood to first integrate out ω_i by rewriting the NBGMM in terms of a mixture of Poisson distributions as follows

$$y_{ij} \sim Poisson\left(\omega_i v_{ij}\exp\left(\boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + \log(\pi_{ij})\right)\right), \quad (5)$$
$$\omega_i \sim Gamma(\alpha, \lambda),$$
$$v_{ij} \sim Gamma(\phi, \phi),$$

where v_{ij} has the mixing gamma distribution. The PGMM proposed in (Sutradhar and Qu, 1998), which has a closed-form likelihood, is a special case of the NBGMM if we assume $v_{ij} = 1$ in eq. (5). As ω_i is a conjugate random effect for the Poisson model (Molenberghs et al., 2010), a partial marginal likelihood after integrating out ω_i for cells in individual i can be expressed explicitly as

$$L_i(\boldsymbol{\beta}, \sigma^2, \phi) = \iint_0^{+\infty}\prod_{j=1}^{n_i} f\left(y_{ij}\middle|\omega_i, \boldsymbol{v}_i\right) f(\omega_i) f(\boldsymbol{v}_i) d\omega_i d\boldsymbol{v}_i$$
$$= \frac{\lambda^\alpha}{\prod_j y_{ij}}\frac{\Gamma(\sum_j y_{ij}+\alpha)}{\Gamma(\alpha)}\exp\left(\sum_{j=1}^{n_i} y_{ij}\left(\boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + \log(\pi_{ij})\right)\right)\int_0^{+\infty}\left(\prod_j v_{ij}{}^{y_{ij}}\right)\underbrace{\left(\lambda + \sum_j\left(v_{ij}\exp\left(\boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + \log(\pi_{ij})\right)\right)\right)^{-(\sum_j y_{ij}+\alpha)}}_{(\Theta)}\prod_j f(v_{ij})\, d\boldsymbol{v}_i, \quad (6)$$

where $\boldsymbol{\Gamma}(\cdot)$ is the gamma function, $\boldsymbol{v}_i = (v_{i1}, \ldots, v_{in_i})^T$, and $f(v_{ij}) = \frac{\phi^\phi}{\Gamma(\phi)} v_{ij}{}^{\phi-1}\exp(-\phi v_{ij})$ is the gamma density function defined in the NBGMM. Generally, it is not straightforward to solve the high-dimensional integral in eq. (6) because of the entanglement of v_{ij} in Θ, but when $n_i \to \infty$, we expect to obtain a simple approximation of the summation in Θ asymptotically by the law of large numbers (LLN). Keeping this notion in mind, we rewrite Θ in eq. (6) as

$$\Theta = \exp\left(-\left(\sum_j y_{ij} + \alpha\right)\log\left(\left(\lambda + \sum_j \mu_{ij}^*\right)\left(1 + \frac{\sum_j\left((v_{ij}-1)\mu_{ij}^*\right)}{\lambda + \sum_j \mu_{ij}^*}\right)\right)\right), \quad (7)$$

where $\mu_{ij}^* = \exp\left(\boldsymbol{x}_{ij}{}^T\boldsymbol{\beta} + log(\pi_{ij})\right)$. Now consider a Monte Carlo method to approximate the integral in eq. (6) by sampling a large number of $\boldsymbol{v}_i$ from $f(\boldsymbol{v}_i)$ and summing up the integrand evaluated at these $\boldsymbol{v}_i$. We show in Supplementary materials A.2 that under very mild conditions,

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

given any $0 < \varepsilon \ll 1$, we have $\left|\frac{\sum_j\left((v_{ij}-1)\mu_{ij}^*\right)}{\lambda+\sum_j \mu_{ij}^*}\right| < \varepsilon$ for almost all Monte Carlo samples when $n_i \to \infty$, that is, $\frac{\sum_j\left((v_{ij}-1)\mu_{ij}^*\right)}{\lambda+\sum_j \mu_{ij}^*} \to 0$ in probability. Given this observation, applying a first-order Taylor expansion to the logarithm in eq. (7), we can approximate the integral in eq. (6) (See Supplementary materials A.3 for the derivation) by

$$\left(\lambda+\sum_j \mu_{ij}^*\right)^{-\left(\sum_j y_{ij}+\alpha\right)} \prod_{j=1}^{n_i} \int_0^1 {v_{ij}}^{y_{ij}} \exp\left(-\breve{\omega}_i\left(v_{ij}-1\right)\mu_{ij}^*\right) \frac{\phi^\phi}{\Gamma(\phi)} {v_{ij}}^{\phi-1} \exp\left(-\phi v_{ij}\right) dv_{ij}, \qquad (8)$$

where $\breve{\omega}_i = \frac{\sum_j y_{ij}+\alpha}{\lambda+\sum_j \mu_{ij}^*}$. The integrand in eq. (8) is now recognized as a kernel of a gamma distribution with respective to v_{ij}. Calculating this integral explicitly and substituting it into eq. (6) gives an approximated marginal log-likelihood

$$l_i(\boldsymbol{\beta}, \sigma^2, \phi) \approx \tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$$
$$= \alpha\log\lambda + \log\frac{\Gamma\left(\sum_j y_{ij}+\alpha\right)}{\Gamma(\alpha)} + \sum_{j=1}^{n_i} y_{ij}\log\left(\mu_{ij}^*\right) - \left(\sum_j y_{ij}+\alpha\right)\log\left(\lambda+\sum_j \mu_{ij}^*\right) + \sum_j\Bigg(\phi\log\phi +$$
$$\log\frac{\Gamma(y_{ij}+\phi)}{\Gamma(\phi)} - \left(y_{ij}+\phi\right)\log\left(\phi+\breve{\omega}_i\mu_{ij}^*\right) + \breve{\omega}_i\mu_{ij}^*\Bigg). \qquad (9)$$

Thus, we may estimate $(\boldsymbol{\beta}, \sigma^2, \phi)$ by optimizing $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$, leading to an M-estimator (Huber, 1964, 2004; Serfling, 2009). Compared to the algorithms based on the LA, the optimization procedure of $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$, described in a later section, becomes straightforward because the analytical form is available. As zero counts are prevalent in scRNA-seq data, the evaluation of the log-gamma function $\log\frac{\Gamma(y_{ij}+\phi)}{\Gamma(\phi)}$ in $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$ can be saved for most cells. We obtain estimating equations by taking the first derivative of $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$ (See Supplementary materials A.4), through which we show in Supplementary materials A.5 that $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$ provides an asymptotically consistent estimate of $(\boldsymbol{\beta}, \sigma^2, \phi)$ when $n_i \to \infty$ and $m \to \infty$. The consistency is also supported by the results in our simulation described in the Results section.

As $n_i \to \infty$ is never achieved in real data, we then investigate the performance of using $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$ in single-cell data with finite n_i. The major question is under which conditions the method works well and how large n_i is required to produce an accurate estimate of σ^2 and ϕ. We see from eq. (7) that the accuracy of $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$ depends on how close $\frac{\sum_j\left((v_{ij}-1)\mu_{ij}^*\right)}{\lambda+\sum_j \mu_{ij}^*}$ is to zero, and it becomes an exact maximum likelihood estimate (MLE) if this term equals zero. By the Lindeberg central limit theorem (CLT), under mild conditions and the null hypothesis of the fixed effects, $\frac{\sum_j\left((v_{ij}-1)\mu_{ij}^*\right)}{\lambda+\sum_j \mu_{ij}^*}$ follows a zero-mean normal distribution with a variance at the rate of

$$\left(1+c^2\right) \Bigg/ \left(n_i\phi + \frac{2\phi\lambda}{E\left(\mu_{ij}^*\right)} + \mathcal{O}\left(\frac{1}{n_i}\right)\right)$$

(See Supplementary materials A.6), where c is the CV of the scaling factor π_{ij}. Intuitively, the estimation accuracy of NEBULA-LN should depend on the dominant term $(1+c^2)/n_i\phi$ when n_i is large, and this relationship was supported by the simulation study presented in the Results section by comparing the estimates between NEBULA-LN and NEBULA-HL which uses an h-

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

likelihood method (described in the next section). Interestingly, we find that a good estimate of ϕ in terms of MSE can be achieved even if $\bar{n}_i\phi$ is as small as 20, where $\bar{n}_i$ is the CPS value. However, the estimate of σ^2 requires a larger $\bar{n}_i\phi$ and $(1+c^2)/\bar{n}_i\phi$ determines the resolution of detecting small σ^2. The simulation study in the Results section shows that $\kappa = \frac{\bar{n}_i\phi}{1+c^2} > 200$ can empirically provide a good estimate of σ^2 as small as ~0.02. As shown in the analysis of the droplet-based scRNA-seq data (Mathys et al., 2019), only ~1% of genes, all of which are very low-expressed genes, have very large cell-level overdispersion ($\phi^{-1} > 10$) in the exitatory neurons. Consequently, NEBULA-LN alone is sufficient for most genes when CPS>200. The standard error (SE) of $(\boldsymbol{\beta}, \sigma^2, \phi)$ can be obtained by the sandwich covariance estimator estimator (Huber, 1967). We find that the SE based on the Hessian matrix of $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$ is practically accurate when $\bar{n}_i\phi$ is large. Instead, we choose to use the h-likelihood (described in the next section) for computing its SE.

Estimation using h-likelihood in NEBULA-HL

As aforementioned, NEBULA-LN is first used to estimate σ^2 and ϕ, and if $\bar{n}_i\hat{\phi}$ or $(1+c^2)/\bar{n}_i\hat{\phi}$ is smaller than a predefined threshold, we then use NEBULA-HL to obtain a more accurate estimate. NEBULA-HL is based on an h-likelihood method (Lee et al., 2006) requiring a two-layer iterative procedure. To derive the h-likelihood, we switch the integral order in $L_i(\boldsymbol{\beta}, \sigma^2, \phi)$ and first integrate out υ_{ij}, which leads to

$$L_i(\boldsymbol{\beta}, \sigma^2, \phi) \propto \int_0^{+\infty} \prod_{\boldsymbol{j}} \left(\boldsymbol{\Gamma}(y_{ij} + \phi) \exp\left(y_{ij} \left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + \log(\pi_{ij}) + \log(\omega_i) \right) \right) \left(\phi + \exp\left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + \log(\pi_{ij}) + \log(\omega_i) \right) \right)^{-(y_{ij}+\phi)} \right) f(\omega_i) d\omega_i. \quad (10)$$

The h-likelihood method first optimizes the integrand in eq. (9) with respect to $(\boldsymbol{\beta}, \omega_i)$ in the inner loop. Note that ω_i cannot be optimized with $\boldsymbol{\beta}$ directly because they are not in the canonical scale (Lee et al., 2006). We thus convert the integrand in eq. (10) to an h-likelihood by re-parametrizing the integral in eq. (9) using $\eta_i = \log(\omega_i)$, which leads to the log-h-likelihood for individual i

$$hl_i(\boldsymbol{\beta}, \eta_i | \sigma^2, \phi) = \sum_j y_{ij} \left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + \log(\pi_{ij}) + \eta_i \right) - \left(y_{ij} + \phi \right) \log\left(\phi + exp\left(\boldsymbol{x}_{ij}{}^T \boldsymbol{\beta} + log(\pi_{ij}) + \eta_i \right) \right) + \alpha\eta_i - \lambda \exp(\eta_i). \quad (11)$$

We optimize $\sum_i hl_i(\boldsymbol{\beta}, \eta_i | \sigma^2, \phi)$ by calculating its first and second derivatives of eq. (11) (See Supplementary materials A.7 for more detail). Because the log-h-likelihood is concave (Both the negative binomial term and the penalizing term are concave), we find that an NR algorithm converges fast and properly. The bottom-right corner of the Hessian matrix corresponding to η_i is diagonal, so the computation for solving the inverse Hessian matrix takes $\mathcal{O}(m)$, which is ignorable. We also use this step to compute the SE of $\boldsymbol{\beta}$ in NEBULA-LN by plugging in $(\hat{\sigma}^2, \hat{\phi})$ estimated from $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$.

Given the current estimate $(\boldsymbol{\beta}^*, \eta_i{}^*)$, we then optimize (σ^2, ϕ) using the marginal likelihood

$$\sum_i L_i(\sigma^2, \phi | \boldsymbol{\beta}^*, \eta_i{}^*) = \sum_i hl_i(\boldsymbol{\beta}^*, \eta_i{}^* | \sigma^2, \phi) - \frac{1}{2} \log\left(\left| \frac{\partial^2 hl_i(\boldsymbol{\beta}, \eta_i | \sigma^2, \phi)}{\partial \eta_i{}^2} \right| \right), \quad (12)$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

where $|\frac{\partial^2 hl_i(\boldsymbol{\beta},\eta_i|\sigma^2,\phi)}{\partial {\eta_i}^2}|$ is the logarithm of the absolute value of the determinant of the second derivative with respect to η_i. Replacing $\frac{\partial^2 hl_i(\boldsymbol{\beta},\eta_i|\sigma^2,\phi)}{\partial {\eta_i}^2}$ with $\frac{\partial^2 hl_i(\boldsymbol{\beta},\eta_i|\sigma^2,\phi)}{\partial (\eta_i,\boldsymbol{\beta})^2}$ in eq. (12) gives a REML estimate of (σ^2, ϕ). Similar to *lme4* (Bates et al., 2014)and *coxmeg* (He and Kulminski, 2020), we use the derivative-free algorithm bobyqa (Powell, 2009) implemented in the *nloptr* R package (Ypma, 2014) for the optimization of $\sum_i L_i(\sigma^2, \phi|\boldsymbol{\beta}^*, {\eta_i}^*)$. The number of iterations in this step grows rapidly with the dimension of the parameters. To reduce the iterations, we plug in $\hat{\phi}$ in $\sum_i L_i(\sigma^2, \phi|\boldsymbol{\beta}^*, {\eta_i}^*)$ as we find that $\hat{\phi}$ from NEBULA-LN is very accurate under mild conditions (e.g., $\bar{n}_i\phi > 20$). Thus, for most genes estimated by NEBULA-HL, we only perform the optimization for σ^2, which substantially reduces the number of iterations. In most cases, the algorithm can reach convergence within ~10 iterations.

Computational implementation of NEBULA-LN

We assessed several algorithms for optimizing $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$ in eq. (8), including (i) the L-BFGS algorithm (Byrd et al., 1995) with box constraints implemented in *nloptr* (Ypma, 2014), (ii) the NR algorithm implemented in the *nlm* R function (Dennis and Schnabel, 1996) and (iii) a trust region algorithm (Fletcher, 1987) implemented in the *trust* R package (https://cran.r-project.org/web/packages/trust/index.html). The NR and the trust region algorithms require the Hessian matrix of $\tilde{l}_i(\boldsymbol{\beta}, \sigma^2, \phi)$. Inspired by (Gilmour et al., 1995), we derive an approximate Hessian matrix (See Supplementary materials A.8) by taking the expectation with respect to y_{ij} for some elements to simplify the computation. As σ^2, ϕ are restricted to be positive values, we reparameterize say what, I guess Hessian matrix using logarithm for the unconstrained optimization. In contrast, the L-BFGS algorithm requires only the estimating equations. We find that the NR algorithm using *nlm* is generally the fastest, and often converges within 10 iterations. However, the NR algorithm may not converge well for low-expressed genes with a small CPC value, and occasionally the convergence depended on the initial values. In contrast, the L-BFGS and trust region algorithms are often slower, but much more robust, and both algorithms produce highly consistent estimates. The trust region algorithm takes fewer iterations, but can be slower than the L-BFGS algorithm when the dimension of $\boldsymbol{\beta}$ becomes larger. This is because the computation of the Hessian matrix is more computationally expensive with the increasing number of the parameters. As most genes in single-cell data have low expression or zero counts, the evaluation of most log-gamma, and digamma functions in the marginal likelihood and the estimating equations can be saved or replaced using a recursive formula.

Comparison with existing tools

We compared the computational performance of NEBULA with three commonly used R packages (*lme4* (Bates et al., 2014), *glmmTMB* (Brooks et al., 2017), and *INLA* (Rue et al., 2009)) for estimating the NBMM. We downloaded the *lme4* and *glmmTMB* R packages via https://cran.r-project.org/ and installed the *INLA* R package via http://www.r-inla.org/download.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

In *lme4*, we included two methods by setting the argument nAGQ=1 (for LA), and nAGQ=0 (for PQL). In *glmmTMB*, we set family=nbinom2. In *INLA*, we set family = "nbinomial", and control.predictor = list(compute = FALSE). We kept the other arguments as default. All benchmarks were run in R 3.5 on Windows 10 and Linux, separately.

Preprocessing of snRNA-seq data

The snRNA-seq UMI count matrix for 17,926 genes and 70,634 cells after QC filtering in 48 human frontal cortex from the ROSMAP project was downloaded from Synapse (https://www.synapse.org/#!Synapse:syn21261143). The details of filtering the cells in the QC step are described in (Mathys et al., 2019). We adopted the clustering results provided by (Mathys et al., 2019), which included 34,976 excitatory neurons, 9196 inhibitory neurons, 18,235 oligodendrocytes, 3392 astrocytes, 1920 microglia, and 2627 OPCs. We did not include pericytes and endothelial cells due to their small sample sizes. We performed a further QC step to remove outlier cells whose total counts or total features were lower than 3 median absolute deviations away from the median using scater (McCarthy et al., 2017), which resulted in 69,458 cells. We also removed genes with CPC<0.1% in the analysis of each cell type (CPC for a gene was calculated by its total UMI counts divided by the number of cells in the cell type). The sub-cell-type clustering annotation provided in (Mathys et al., 2019) was used in the analysis of marker genes for sub-clusters. There were 11, 12, 5, 4, 4, and 3 subclusters in the excitatory neurons, inhibitory neurons, oligodendrocytes, astrocytes, microglia, and OPCs, respectively. The percentage of ribosomal genes (RPS and RPL genes) was computed using the *PercentageFeatureSet* function in the *Seurat* R package (Butler et al., 2018). Subject-level covariates, including age, sex, AD status, race, and *APOE* genotype in the ROSMAP project (Bennett et al., 2012a, 2012b), were downloaded from Synapse (https://www.synapse.org/#!Synapse:syn3157322).

In the co-expression analysis of *APOE*, we first obtained normalized *APOE* expression by dividing the raw count of *APOE* in each cell by its library size. We then subtracted from it the mean value of the normalized *APOE* expression across all cells of each individual. In the isoform-specific analysis, we separated all cells into three categories (e2e3, e3e3, or e4$^+$) based on the *APOE* genotypes. We then applied the same normalizing and centering procedure as in the co-expression analysis of *APOE*.

Simulation study

We generated the number of CPS n_i using a Poisson distribution for the balanced design or a negative binomial distribution with size=3 for the unbalanced design. We evaluated scenarios with the number of subjects $m = 30, 50, 100$ and CPS value $\bar{n}_i = 100, 200, 400, 800$. The count data was generated based on the following generative model

$$y_{ij} \sim Poisson\left(\pi_{ij} \exp\left(\beta_0 + X_1\beta_1 + X_2\beta_2 + \log\left(v_{ij}\right) + \log\left(\omega_i\right)\right)\right),$$
$$\omega_i \sim Gamma\left(\frac{1}{\exp(\sigma^2)-1}, \frac{1}{\exp(\sigma^2/2)(\exp(\sigma^2)-1)}\right),$$
$$v_{ij} \sim Gamma(\phi, \phi),$$

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

where we considered β_0 ranging from -5 to 2, ϕ ranging from 0.01 to 100, and σ^2 ranging from 0.01 to 1. Under the scenarios of a constant scaling factor, we set $\pi_{ij} = 1$, and otherwise we sampled π_{ij} from $Gamma(1,1)$. We simulated a cell-level variable X_1 and a subject-level variable X_2 using a standard normal distribution. Under each of the settings, we generated 500 data sets to calculate the summary statistics. The MSE of an estimate (e.g., $\hat{\phi}$) was computed by $\frac{\sum (\phi-\hat{\phi})^2}{500}$ across the 500 data sets.

Data and code availability

This manuscript was prepared using limited access datasets obtained through Synapse (https://www.synapse.org/#!Synapse:syn3219045, https://www.synapse.org/#!Synapse:syn18485175). The code of this R package can be downloaded and installed via https://github.com/lhe17/nebula.

Acknowledgments

This research was supported by Grants from the National Institute on Aging (R01 AG061853, R01 AG065477, and R01 AG070488). The funders had no role in study design, data collection and analysis, decision to publish, or manuscript preparation. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of Interest

The authors declare that they have no conflict of interest.

Author contributions

LH conceived, developed, and implemented the algorithms. LH designed the simulation study and analyzed the simulated and real data. AK contributed to acquiring the real data and discussion of the final results. LH and AK contributed to the writing of the manuscript.

References

Bacher, R., Chu, L.-F., Leng, N., Gasch, A.P., Thomson, J.A., Stewart, R.M., Newton, M., and Kendziorski, C. (2017). SCnorm: robust normalization of single-cell RNA-seq data. Nat. Methods *14*, 584–586.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2014). Fitting linear mixed-effects models using lme4. ArXiv14065823 Stat.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Bell, R.D., Sagare, A., Friedman, A.E., Bedi, G., Holtzman, D.M., Deane, R., and Zlokovic, B.V. (2007). Transport pathways for clearance of human Alzheimer's amyloid β-peptide and apolipoproteins E and J in the mouse central nervous system. J. Cereb. Blood Flow Metab. Off. J. Int. Soc. Cereb. Blood Flow Metab. *27*, 909–918.

Bennett, D.A., Schneider, J.A., Arvanitakis, Z., and Wilson, R.S. (2012a). Overview and findings from the religious orders study. Curr. Alzheimer Res. *9*, 628–645.

Bennett, D.A., Schneider, J.A., Buchman, A.S., Barnes, L.L., Boyle, P.A., and Wilson, R.S. (2012b). Overview and findings from the rush Memory and Aging Project. Curr. Alzheimer Res. *9*, 646–663.

Booth, J.G., Casella, G., Friedl, H., and Hobert, J.P. (2003). Negative binomial loglinear mixed models. Stat. Model. *3*, 179–191.

Breslow, N.E., and Clayton, D.G. (1993). Approximate inference in generalized linear mixed models. J. Am. Stat. Assoc. *88*, 9–25.

Breslow, N.E., and Lin, X. (1995). Bias Correction in Generalised Linear Mixed Models with a Single Component of Dispersion. Biometrika *82*, 81–91.

Brooks, M.E., Kristensen, K., van Benthem, K.J., Magnusson, A., Berg, C.W., Nielsen, A., Skaug, H.J., Machler, M., and Bolker, B.M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. R J. *9*, 378–400.

Butler, A., Hoffman, P., Smibert, P., Papalexi, E., and Satija, R. (2018). Integrating single-cell transcriptomic data across different conditions, technologies, and species. Nat. Biotechnol. *36*, 411–420.

Byrd, R., Lu, P., Nocedal, J., and Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. SIAM J. Sci. Comput. *16*, 1190–1208.

Chen, W., Li, Y., Easton, J., Finkelstein, D., Wu, G., and Chen, X. (2018). UMI-count modeling and differential expression analysis for single-cell RNA sequencing. Genome Biol. *19*, 70.

Choi, K., Chen, Y., Skelly, D.A., and Churchill, G.A. (2020). Bayesian model selection reveals biological origins of zero inflation in single-cell transcriptomics. Genome Biol. *21*, 183.

Deng, A., Irizarry, M.C., Nitsch, R.M., Growdon, J.H., and Rebeck, G.W. (2001). Elevation of Cystatin C in Susceptible Neurons in Alzheimer's Disease. Am. J. Pathol. *159*, 1061–1068.

Dennis, J.E., and Schnabel, R.B. (1996). Numerical Methods for Unconstrained Optimization and Nonlinear Equations (Society for Industrial and Applied Mathematics).

Fletcher, R. (1987). Practical Methods of Optimization (Wiley).

Gilmour, A.R., Thompson, R., and Cullis, B.R. (1995). Average Information REML: An Efficient Algorithm for Variance Parameter Estimation in Linear Mixed Models. Biometrics *51*, 1440–1450.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Guerreiro, R., Wojtas, A., Bras, J., Carrasquillo, M., Rogaeva, E., Majounie, E., Cruchaga, C., Sassi, C., Kauwe, J.S.K., Younkin, S., et al. (2013). TREM2 variants in Alzheimer's disease. N. Engl. J. Med. *368*, 117–127.

Hafemeister, C., and Satija, R. (2019). Normalization and variance stabilization of single-cell RNA-seq data using regularized negative binomial regression. Genome Biol. *20*, 296.

Harold, D., Abraham, R., Hollingworth, P., Sims, R., Gerrish, A., Hamshere, M.L., Pahwa, J.S., Moskvina, V., Dowzell, K., Williams, A., et al. (2009). Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. Nat. Genet. *41*, 1088–1093.

Hashimshony, T., Senderovich, N., Avital, G., Klochendler, A., de Leeuw, Y., Anavy, L., Gennert, D., Li, S., Livak, K.J., Rozenblatt-Rosen, O., et al. (2016). CEL-Seq2: sensitive highly-multiplexed single-cell RNA-Seq. Genome Biol. *17*, 77.

He, L., and Kulminski, A.M. (2020). Fast Algorithms for Conducting Large-Scale GWAS of Age-at-Onset Traits Using Cox Mixed-Effects Models. Genetics.

Hilbe, J.M. (2011). Negative Binomial Regression (Cambridge University Press).

Huber, P.J. (1964). Robust Estimation of a Location Parameter. Ann. Math. Stat. *35*, 73–101.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. p.

Huber, P.J. (2004). Robust Statistics (John Wiley & Sons).

Joe, H. (2008). Accuracy of Laplace approximation for discrete response mixed models. Comput. Stat. Data Anal. *52*, 5066–5074.

Jonsson, T., Stefansson, H., Steinberg, S., Jonsdottir, I., Jonsson, P.V., Snaedal, J., Bjornsson, S., Huttenlocher, J., Levey, A.I., Lah, J.J., et al. (2013). Variant of TREM2 associated with the risk of Alzheimer's disease. N. Engl. J. Med. *368*, 107–116.

Kaeser, S.A., Herzig, M.C., Coomaraswamy, J., Kilger, E., Selenica, M.-L., Winkler, D.T., Staufenbiel, M., Levy, E., Grubb, A., and Jucker, M. (2007). Cystatin C modulates cerebral beta-amyloidosis. Nat. Genet. *39*, 1437–1439.

Kim, J., Miller, V.M., Levites, Y., West, K.J., Zwizinski, C.W., Moore, B.D., Troendle, F.J., Bann, M., Verbeeck, C., Price, R.W., et al. (2008). BRI2 (ITM2b) Inhibits Aβ Deposition In Vivo. J. Neurosci. *28*, 6030–6036.

Klein, A.M., Mazutis, L., Akartuna, I., Tallapragada, N., Veres, A., Li, V., Peshkin, L., Weitz, D.A., and Kirschner, M.W. (2015). Droplet barcoding for single cell transcriptomics applied to embryonic stem cells. Cell *161*, 1187–1201.

Lambert, J.-C., Heath, S., Even, G., Campion, D., Sleegers, K., Hiltunen, M., Combarros, O., Zelenika, D., Bullido, M.J., Tavernier, B., et al. (2009). Genome-wide association study identifies variants at CLU and CR1 associated with Alzheimer's disease. Nat. Genet. *41*, 1094–1099.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Landeghem, G.V., Fraine, B.D., and Damme, J.V. (2005). The Consequence of Ignoring a Level of Nesting in Multilevel Analysis: A Comment. Multivar. Behav. Res. *40*, 423–434.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Lee, Y., Nelder, J.A., and Pawitan, Y. (2006). Generalized linear models with random effects: unified analysis via H-likelihood (Chapman and Hall/CRC).

Lin, X. (2007). Estimation using penalized quasilikelihood and quasi-pseudo-likelihood in Poisson mixed models. Lifetime Data Anal. *13*, 533–544.

Lindstrom, M.J., and Bates, D.M. (1990). Nonlinear Mixed Effects Models for Repeated Measures Data. Biometrics *46*, 673–687.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Macosko, E.Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., Tirosh, I., Bialas, A.R., Kamitaki, N., Martersteck, E.M., et al. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. Cell *161*, 1202–1214.

Mathys, H., Davila-Velderrain, J., Peng, Z., Gao, F., Mohammadi, S., Young, J.Z., Menon, M., He, L., Abdurrob, F., Jiang, X., et al. (2019). Single-cell transcriptomic analysis of Alzheimer's disease. Nature *570*, 332–337.

Matsubara, E., Frangione, B., and Ghiso, J. (1995). Characterization of Apolipoprotein J-Alzheimer's Aβ Interaction. J. Biol. Chem. *270*, 7563–7567.

Matsuda, S., Giliberto, L., Matsuda, Y., Davies, P., McGowan, E., Pickford, F., Ghiso, J., Frangione, B., and D'Adamio, L. (2005). The familial dementia BRI2 gene binds the Alzheimer gene amyloid-beta precursor protein and inhibits amyloid-beta production. J. Biol. Chem. *280*, 28912–28916.

McCarthy, D.J., Chen, Y., and Smyth, G.K. (2012). Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. Nucleic Acids Res. *40*, 4288–4297.

McCarthy, D.J., Campbell, K.R., Lun, A.T.L., and Wills, Q.F. (2017). Scater: pre-processing, quality control, normalization and visualization of single-cell RNA-seq data in R. Bioinforma. Oxf. Engl. *33*, 1179–1186.

Milanzi, E., Alonso, A., and Molenberghs, G. (2012). Ignoring overdispersion in hierarchical loglinear models: Possible problems and solutions. Stat. Med. *31*, 1475–1482.

Moerbeek, M. (2004). The Consequence of Ignoring a Level of Nesting in Multilevel Analysis. Multivar. Behav. Res. *39*, 129–149.

Molenberghs, G., Verbeke, G., Demétrio, C.G.B., and Vieira, A.M.C. (2010). A Family of Generalized Linear Models for Repeated Measures with Normal and Conjugate Random Effects. Stat. Sci. *25*, 325–347.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Neuhaus, J.M., and McCulloch, C.E. (2011). Estimation of covariate effects in generalized linear mixed models with informative cluster sizes. Biometrika *98*, 147–162.

Neuhaus, J.M., McCulloch, C.E., and Boylan, R. (2013). Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. Stat. Med. *32*, 2419–2429.

Ormerod, J.T., and Wand, M.P. (2012). Gaussian Variational Approximate Inference for Generalized Linear Mixed Models. J. Comput. Graph. Stat. *21*, 2–17.

Picelli, S., Faridani, O.R., Björklund, Å.K., Winberg, G., Sagasser, S., and Sandberg, R. (2014). Full-length RNA-seq from single cells using Smart-seq2. Nat. Protoc. *9*, 171–181.

Pinheiro, J.C., and Bates, D.M. (1995). Approximations to the Log-Likelihood Function in the Nonlinear Mixed-Effects Model. J. Comput. Graph. Stat. *4*, 12–35.

Pinheiro, J.C., and Chao, E.C. (2006). Efficient Laplacian and Adaptive Gaussian Quadrature Algorithms for Multilevel Generalized Linear Mixed Models. J. Comput. Graph. Stat. *15*, 58–81.

Powell, M.J. (2009). The BOBYQA algorithm for bound constrained optimization without derivatives. Camb. NA Rep. NA200906 Univ. Camb. Camb. 26–46.

Rabe-Hesketh, S., Skrondal, A., and Pickles, A. (2002). Reliable Estimation of Generalized Linear Mixed Models using Adaptive Quadrature. Stata J. *2*, 1–21.

Revesz, T., Holton, J.L., Lashley, T., Plant, G., Frangione, B., Rostagno, A., and Ghiso, J. (2009). Genetics and molecular pathogenesis of sporadic and hereditary cerebral amyloid angiopathies. Acta Neuropathol. (Berl.) *118*, 115–130.

Robinson, M.D., and Oshlack, A. (2010). A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol. *11*, R25.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J. R. Stat. Soc. Ser. B Stat. Methodol. *71*, 319–392.

Serfling, R.J. (2009). Approximation Theorems of Mathematical Statistics (John Wiley & Sons).

Simpson, E.H. (1951). The Interpretation of Interaction in Contingency Tables. J. R. Stat. Soc. Ser. B Methodol. *13*, 238–241.

Sutradhar, B.C., and Qu, Z. (1998). On approximate likelihood inference in a poisson mixed model. Can. J. Stat. *26*, 169–186.

Tierney, L., and Kadane, J.B. (1986). Accurate Approximations for Posterior Moments and Marginal Densities. J. Am. Stat. Assoc. *81*, 82–86.

Tuerlinckx, F., Rijmen, F., Verbeke, G., and De Boeck, P. (2006). Statistical inference in generalized linear mixed models: a review. Br. J. Math. Stat. Psychol. *59*, 225–255.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Vestal, B.E., Moore, C.M., Wynn, E., Saba, L., Fingerlin, T., and Kechris, K. (2020). MCMSeq: Bayesian hierarchical modeling of clustered and repeated measures RNA sequencing experiments. BMC Bioinformatics *21*, 375.

Yin, C., Ackermann, S., Ma, Z., Mohanta, S.K., Zhang, C., Li, Y., Nietzsche, S., Westermann, M., Peng, L., Hu, D., et al. (2019). ApoE attenuates unresolvable inflammation by complex formation with activated C1q. Nat. Med. *25*, 496–506.

Ypma, J. (2014). Introduction to nloptr: an R interface to NLopt (Tech. rep).

Zhang, X., Mallick, H., Tang, Z., Zhang, L., Cui, X., Benson, A.K., and Yi, N. (2017). Negative binomial mixed models for analyzing microbiome count data. BMC Bioinformatics *18*, 4.

Zhang, Y., Sloan, S.A., Clarke, L.E., Caneda, C., Plaza, C.A., Blumenthal, P.D., Vogel, H., Steinberg, G.K., Edwards, M.S.B., Li, G., et al. (2016). Purification and characterization of progenitor and mature human astrocytes reveals transcriptional and functional differences with mouse. Neuron *89*, 37–53.

Figures

Figure 1. (A) The computational time (measured in $\log_{10}$(seconds)) of fitting an NBMM for 10,000 genes with respect to CPS. The number of subjects was set at 50. (B) The computational time (measured in $\log_{10}$(seconds)) of fitting an NBMM for 10,000 genes with respect to the number of subjects. The CPS value was set at 200. Two fixed-effects predictors were included in the NBMM. The average benchmarks were summarized from scenarios of varying subject-level and cell-level overdispersions and the CPC value of a gene ranging from exp(-4) to 1. CPS: cells per subject. CPC: counts per cell.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

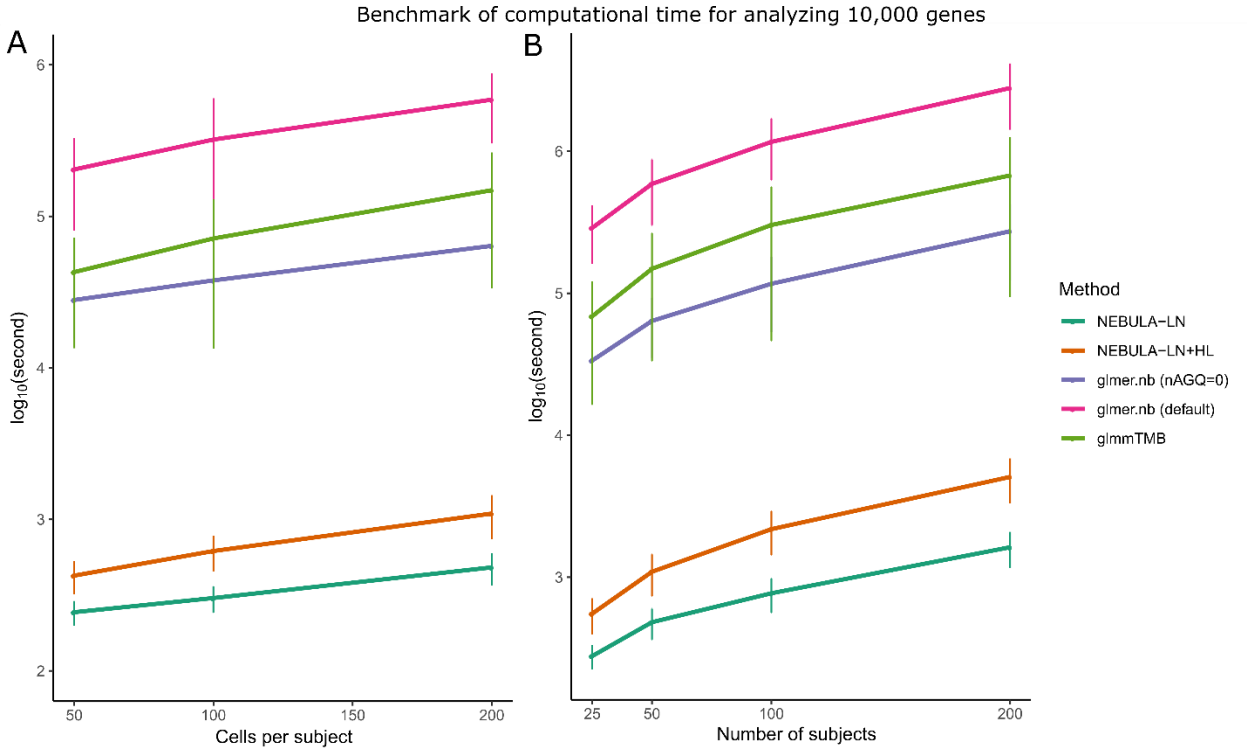


Figure 2. Comparison of estimated cell-level and subject-level overdispersions between NEBULA-LN and NEBULA-HL. The summary statistics were calculated from five hundred simulated replicates in each of the scenarios. (A) The cell-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of CPS, β_0 (a lower β_0 corresponding to a lower CPC value), and ϕ. The number of subjects was set at 50. (B) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of β_0, σ^2, and ϕ. The number of subjects was set at 50. The CPS value was set at 400. (C) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of the number of subjects, σ^2, and ϕ. The CPS value was set at 400, and β_0 was set at 0.05. (D) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of CPS, σ^2, and ϕ. The number of individuals was set at 50, and β_0 was set at 1. CPS: cells per subject. CPC: counts per cell.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

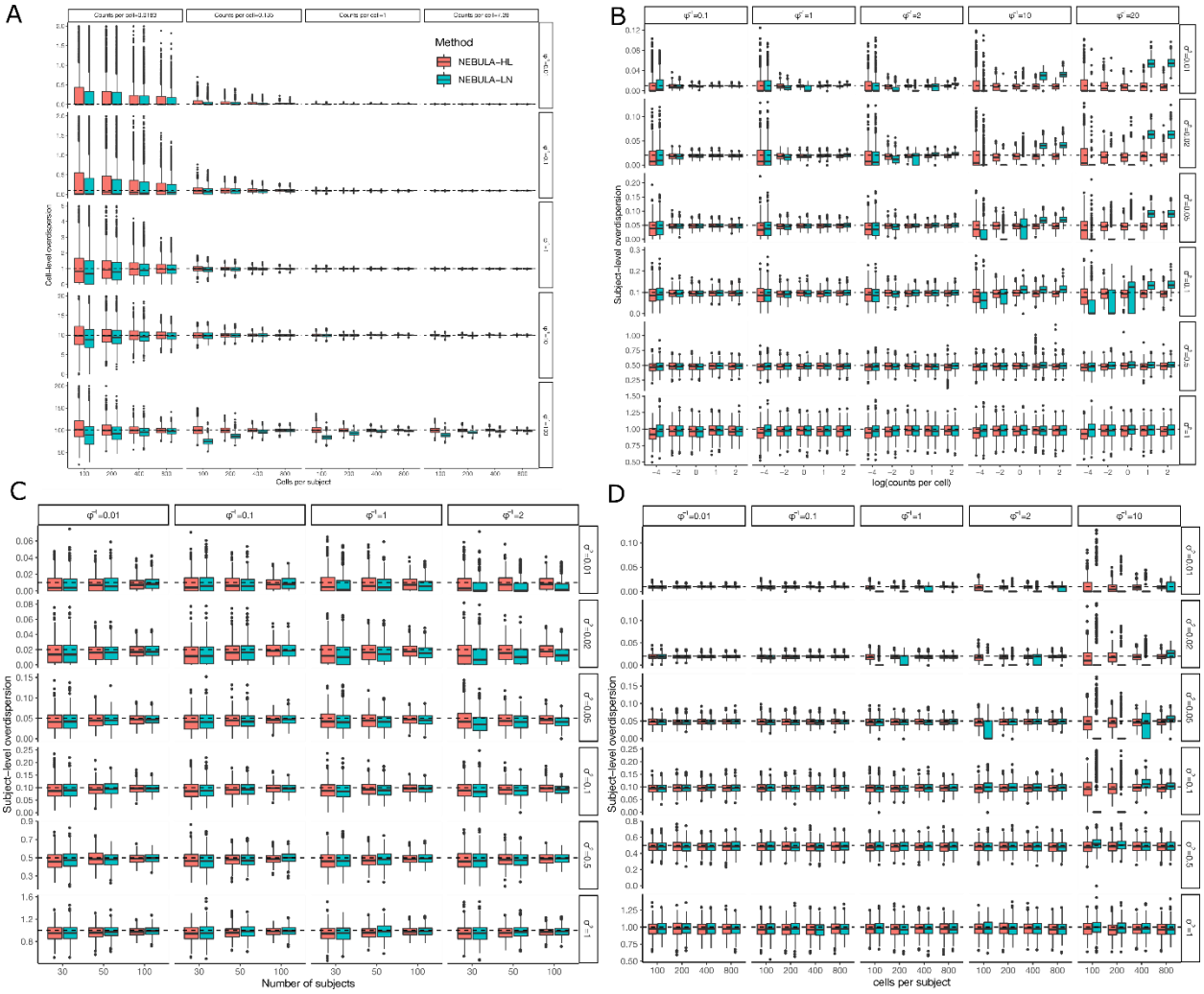


Figure 3. Empirical type I error rate of testing cell-level and subject-level predictors using NEBULA-LN and NEBULA-HL. The type I error rate was calculated from five hundred simulated replicates in each of the scenarios and was evaluated at the significance level of 0.05 (the dashed lines). (A) The empirical type I error rate of NEBULA-LN under different combinations of CPS, σ^2, and ϕ. The number of individuals was set at 50. (B) The empirical type I error rate of NEBULA-HL under different combinations of CPS, σ^2, and ϕ. The number of subjects was set at 50. (C) The empirical type I error rate of NEBULA-LN under different combinations of the number of subjects, σ^2, and ϕ. The CPS value was set at 400. (D) The empirical type I error rate of NEBULA-HL under different combinations of the number of subjects, σ^2, and ϕ. The CPS value was set at 400. CPS: cells per subject. CPC: counts per cell.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

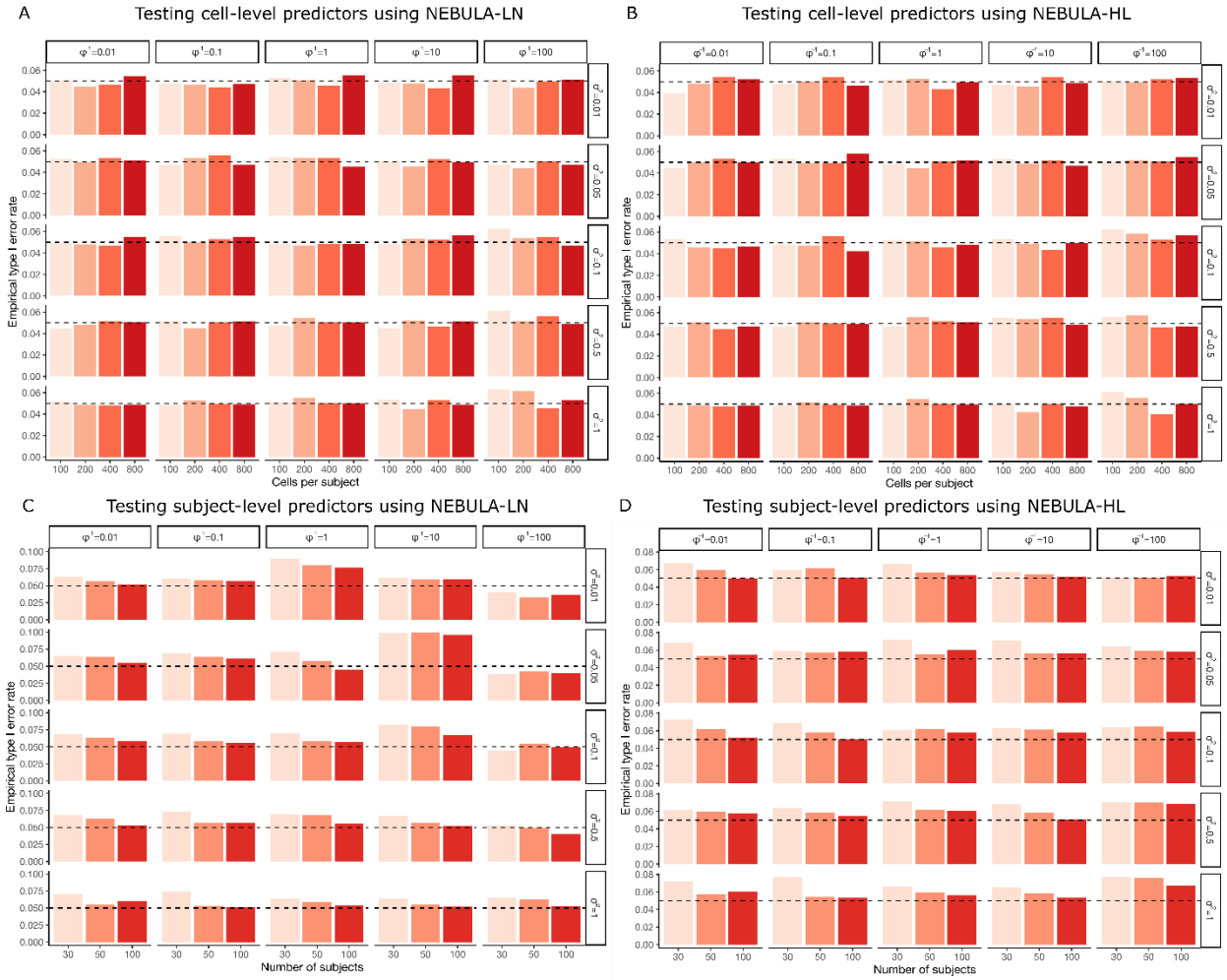


Figure 4. (A) The cell-level overdispersions of 16,207 genes with CPC>0.1% in the excitatory neurons estimated by NEBULA with respect to the CPC of the gene. No covariates other than the intercept were included in the model. The density is the two-dimensional kernel density computed using the *kde2d* R function with n=100. (B) The subject-level overdispersions of 16,207 genes with CPC>0.1% in the excitatory neurons estimated by NEBULA with respect to the CPC of the gene. (C) The cell-level overdispersions of 16,207 genes with CPC>0.1% in the excitatory neurons estimated by NEBULA versus their subject-level overdispersions. (D) The estimated cell-level overdispersions of all genes with CPC>0.1% in each of the six brain cell types before and after adjusting for age, sex, race, AD status, the total number of features of the cell, and the percentage of ribosomal genes. (E) The estimated subject-level overdispersions of all genes with CPC>0.1% in each of the six brain cell types before and after adjusting for age, sex, race, AD status, the total number of features of the cell, and the percentage of ribosomal genes.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

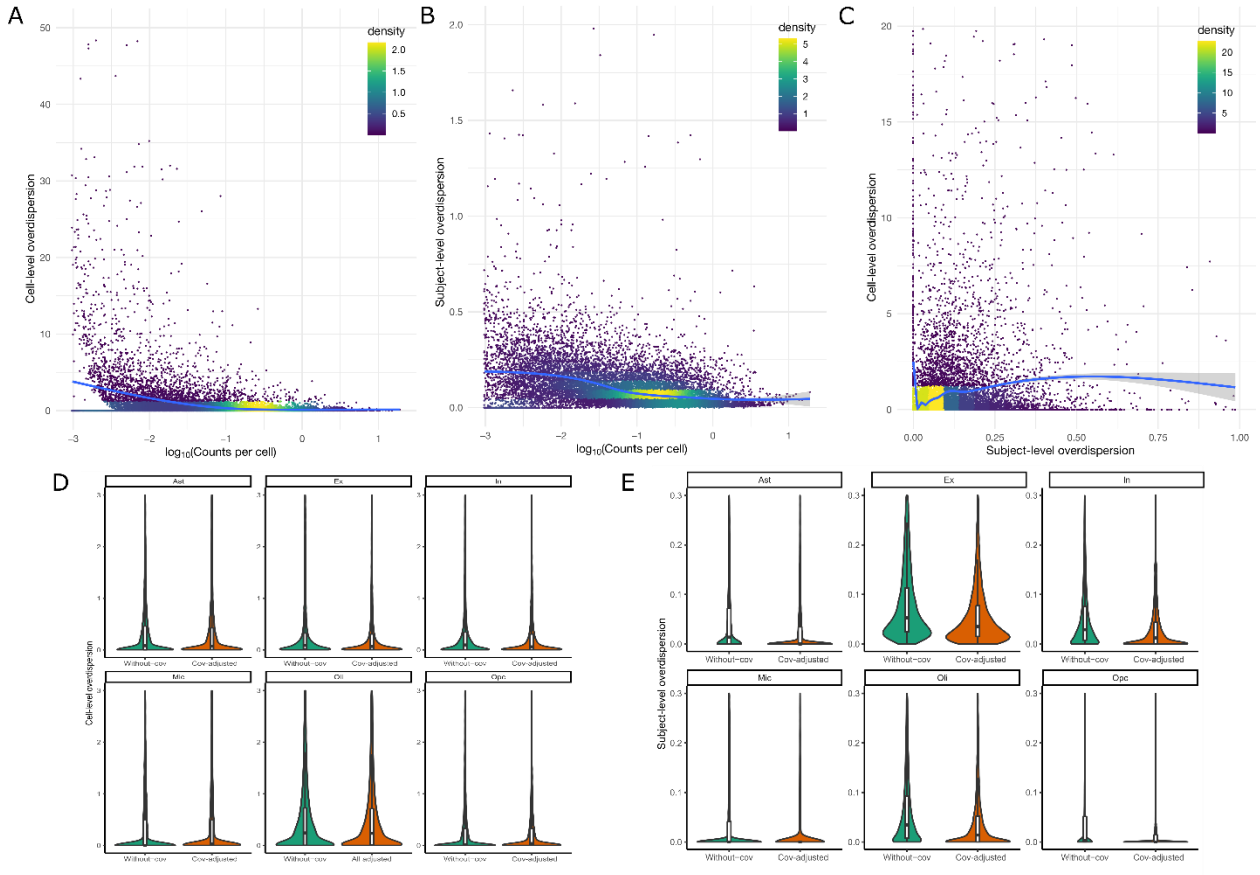
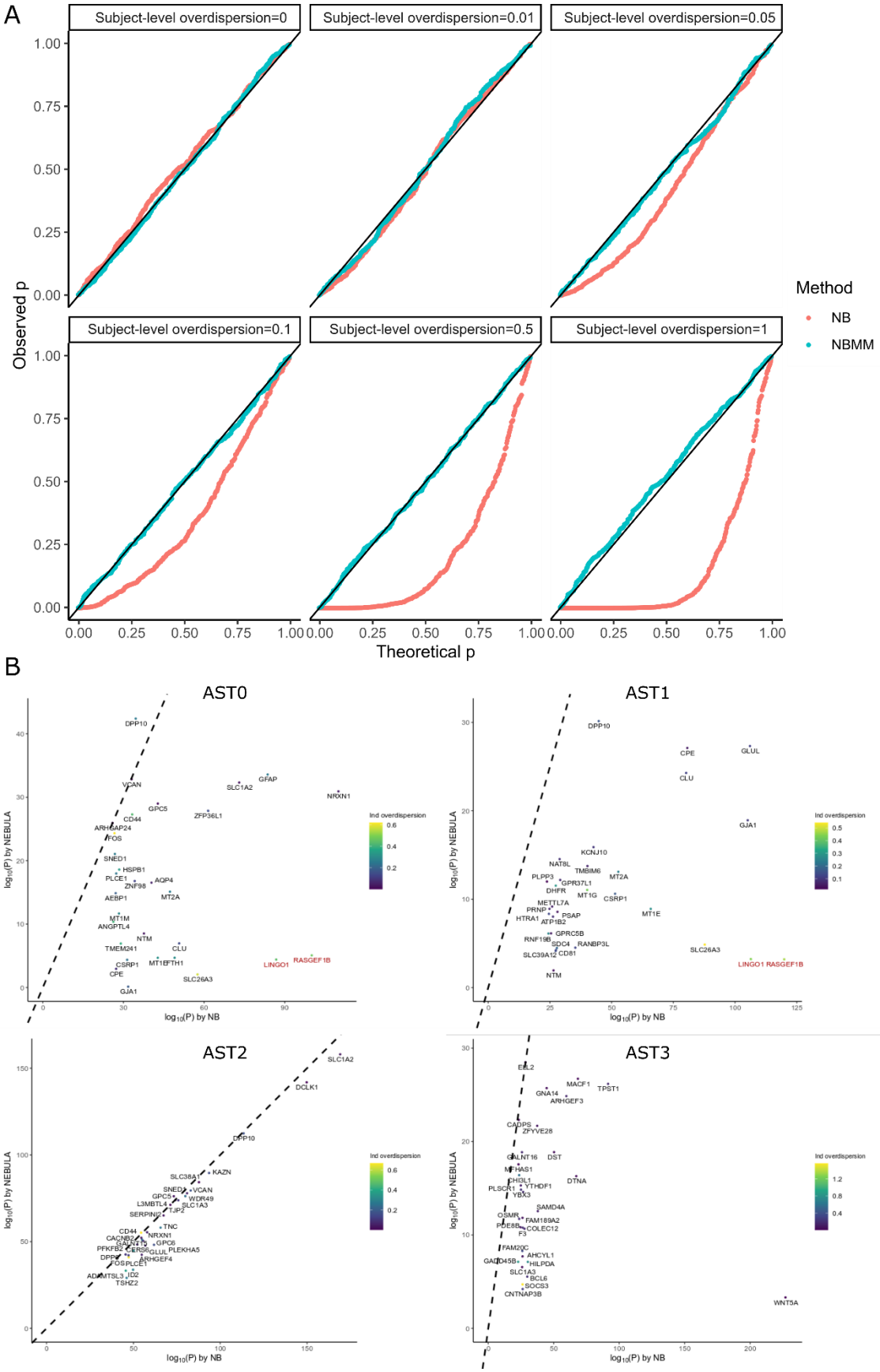


Figure 5. (A) Q-Q plots of p-values for testing the association between simulated counts and cell clusters using NEBULA and a negative binomial regression. The counts were simulated under the null model (i.e., no association between the counts and cell clusters) based on an NBMM with the subject-level overdispersion (σ^2) ranging from 0 to 1 and a cell-level overdispersion fixed at 1. (B) Comparison of the p-values of the top 30 genes identified by a negative binomial regression associated with each of the four astrocyte subpopulations (AST0-AST3) in the snRNA-seq data (Mathys et al., 2019). The labels of the four astrocyte subpopulations are corresponding to those presented in (Mathys et al., 2019). The color of each point gives the subject-level overdispersion of this gene estimated by NEBULA. Two genes (*RASGEF1B* and *LINGO1*) highlighted in red showed highly reduced significance in NEBULA compared to the p-values in the negative binomial regression.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Tables

Table 1: Top 30 genes co-expressed with *APOE* at the single-cell level in astrocytes and microglia in the human frontal cortex. logFC: The log(fold change) of the gene expression if the *APOE* expression is increased. logFC>0 means a positive correlation, and logFC<0 means a negative correlation with the expression of *APOE*. CPC: The UMI counts per cell of this gene in the specific cell type.

Astrocytes				**Microglia**			
logFC	p	Gene	CPC	logFC	p	Gene	CPC
1.48E-01	9.97E-79	*CLU*	3.13	9.57E-02	1.13E-16	*CD74*	0.88
1.56E-01	4.04E-68	*CST3*	2.57	1.09E-01	4.65E-12	*C1QB*	0.37
9.42E-02	3.62E-32	*CPE*	2.36	-8.53E-01	6.61E-12	*FYN*	0.14
1.21E-01	3.20E-31	*TTYH1*	1.03	1.64E-01	2.28E-11	*APOC1*	0.10
1.20E-01	8.97E-31	*GJA1*	1.68	1.08E-01	4.81E-10	*C1QA*	0.26
9.70E-02	3.45E-29	*GLUL*	2.31	5.30E-02	4.06E-09	*DOCK4*	2.79
1.38E-01	7.59E-29	*ITM2C*	0.93	9.54E-02	6.34E-09	*C1QC*	0.30
8.79E-02	1.98E-27	*PTGDS*	2.13	1.49E-01	6.87E-09	*TREM2*	0.10
9.63E-02	6.26E-24	*CKB*	1.57	1.29E-01	1.42E-08	*CD14*	0.27
7.69E-02	7.19E-23	*SPARCL1*	2.87	-8.32E-01	8.97E-08	*SKAP1*	0.09
-1.95E-01	2.19E-22	*DPP10*	3.42	1.09E-01	1.05E-07	*GPX1*	0.22
1.38E-01	2.51E-22	*METTL7A*	0.58	1.48E-01	1.08E-07	*TSPO*	0.05
1.24E-01	4.57E-22	*SERPINE2*	0.72	-1.05E+00	2.02E-07	*CD247*	0.10
8.06E-02	2.11E-21	*ATP1A2*	2.14	8.28E-02	3.75E-07	*ITM2B*	0.38
1.30E-01	7.98E-21	*ITM2B*	0.57	-9.77E-01	9.32E-07	*F2R*	0.01
5.66E-02	1.14E-19	*SLC1A2*	8.09	8.58E-02	9.82E-07	*CD81*	0.18
1.05E-01	1.23E-17	*ATP1B2*	1.02	-9.77E-01	1.04E-06	*PRKCQ*	0.05
1.33E-01	1.44E-17	*TUBB2B*	0.52	6.66E-02	1.20E-06	*PSAP*	0.48
1.03E-01	2.04E-17	*PSAP*	0.93	-6.96E-01	2.24E-06	*CCDC88C*	0.04
9.37E-02	8.32E-17	*FTL*	1.58	6.19E-02	2.44E-06	*FTL*	1.16
1.61E-01	9.20E-16	*VIM*	0.30	-2.26E-01	2.80E-06	*PITPNC1*	0.19
1.17E-01	1.93E-15	*AGT*	0.62	-2.48E-01	4.47E-06	*CBLB*	0.16
1.73E-01	6.32E-15	*CHRDL1*	0.19	3.02E-01	4.93E-06	*PKNOX2*	0.00
8.44E-02	2.22E-14	*PLPP3*	1.40	-4.94E-01	5.36E-06	*SLCO3A1*	0.04
1.35E-01	2.87E-14	*SCG3*	0.38	1.34E-01	7.59E-06	*MX1*	0.05
1.55E-01	4.88E-14	*HLA-A*	0.24	2.61E-01	7.83E-06	*FRAT2*	0.01
9.52E-02	8.15E-14	*PTN*	0.98	-4.75E-01	1.14E-05	*VAV3*	0.05
1.30E-01	8.16E-14	*CLDN10*	0.41	9.21E-02	1.42E-05	*HLA-DRB1*	0.24
1.29E-01	9.36E-14	*SARAF*	0.38	2.22E-01	1.60E-05	*MFSD5*	0.01
1.54E-01	1.01E-13	*WIF1*	0.54	1.94E-01	1.92E-05	*ZNF22*	0.02

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Supplementary Text

Please see the supplementary file.

Supplementary Tables

Table S1. A list of top 200 genes co-expressed with the expression of the *APOE* variants at the single-cell level in astrocytes in the human frontal cortex. logFC: The log(fold change) of the gene expression if the *APOE* expression is increased. logFC>0 means a positive correlation, and logFC<0 means a negative correlation with the expression of *APOE*. APOE2: *APOE* e2e3. APOE3: *APOE* e3e3. APOE4: *APOE* e3e4 and *APOE* e4e4.

Table S2. A list of top 200 genes co-expressed with the expression of the *APOE* variants at the single-cell level in microglia in the human frontal cortex. logFC: The log(fold change) of the gene expression if the *APOE* expression is increased. logFC>0 means a positive correlation, and logFC<0 means a negative correlation with the expression of *APOE*. APOE2: *APOE* e2e3. APOE3: *APOE* e3e3. APOE4: *APOE* e3e4 and *APOE* e4e4.

Table S3. The proposed strategy for testing a cell-level predictor used in the analysis of the real data. This table summarizes the condition under which an estimation algorithm (NEBULA-LN or NEBULA-HL) is used to estimate each of overdispersion parameters.

Supplementary Figures

Figure S1. The computational time (measured in $\log_{10}$(seconds)) of fitting an NBMM for 10,000 genes with respect to the different numbers of fixed-effects predictors in the NBMM. The number of subjects was set at 50, and CPS was set at 100. The average benchmarks are summarized from scenarios of varying subject-level and cell-level overdispersions and the mean count per cell ranging from exp(-4) to 1.

Figure S2. Comparison of estimated cell-level and subject-level overdispersions between NEBULA-LN and NEBULA-HL under a situation in which the CV of the scaling factor (π_{ij}) is one. Five hundred simulated data sets were generated in each of the scenarios to calculate the summary statistics. (A) The cell-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of CPS, β_0 (a lower β_0 corresponding to a lower CPC value), and ϕ. The number of subjects was set at 50. (B) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of β_0, σ^2, and ϕ. The number of subjects was set at 50. The CPS value was set at 400. (C) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of the number of subjects, σ^2, and ϕ. The CPS value was set at 400, and β_0 was set at 0.05. (D) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of CPS, σ^2, and ϕ. The number of subjects was set at 50, and β_0 was set at 1.

Figure S3. Comparison of estimated cell-level overdispersions between NEBULA-LN and NEBULA-HL under different combinations of the number of subjects, β_0, and ϕ. The CPS value

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

was set at 400. The summary statistics were calculated from five hundred simulated replicates in each of the scenarios.

Figure S4. The log(FC) of a subject-level predictor estimated by NEBULA versus those estimated by *glmer.nb* with nAGQ=0 for 16,207 genes with CPC>0.1% in the excitatory neurons.

Figure S5. Comparison of estimated cell-level and subject-level overdispersions between NEBULA-LN and NEBULA-HL under an unbalanced design in which the cell counts across the subjects were sampled from a negative binomial distribution with size=3. The summary statistics were calculated from five hundred simulated replicates in each of the scenarios. (A) The cell-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of CPS, β_0 (a lower β_0 corresponding to a lower CPC value), and ϕ. The number of subjects was set at 50. (B) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of β_0, σ^2, and ϕ. The number of subjects was set at 50. The CPS value was set at 400. (C) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of the number of subjects, σ^2, and ϕ. The CPS value was set at 400, and β_0 was set at 0.05. (D) The subject-level overdispersions estimated by NEBULA-LN and NEBULA-HL under different combinations of CPS, σ^2, and ϕ. The number of subjects was set at 50, and β_0 was set at 1.

Figure S6. Empirical type I error rate of testing a subject-level predictor using NEBULA-LN under different combinations of CPC, σ^2, and ϕ. The CPS value was set at 400 and the number of subjects was set at 50. The empirical type I error rate was calculated from five hundred simulated replicates in each of the scenarios and was evaluated at the significance level of 0.05 (the dashed lines).

Figure S7. Empirical type I error rate of testing a cell-level predictor using NEBULA-LN under different combinations of CPC, σ^2, and ϕ. The number of subjects was set at 50 and the CPS value was set at 400. The empirical type I error rate was calculated from five hundred simulated replicates in each of the scenarios and was evaluated at the significance level of 0.05 (the dashed lines).

Figure S8. Distribution of the library size (the total number of reads) of each cell grouped by the seven major cell types in the snRNA-seq data in the human frontal cortex.

Figure S9. The subject-level overdispersions estimated by NEBULA versus those estimated by *glmer.nb* with nAGQ=0 for 16,207 genes with CPC>0.1% in the excitatory neurons.

Figure S10. The cell frequency of the four subclusters in astrocytes across the 48 individuals. The 3386 astrocytes are from the ROSMAP snRNA-seq data in the human frontal cortex.

Figure S11. The cell-level overdispersions of genes with CPC>0.1% in five major cell types in the frontal cortex estimated by NEBULA with respect to the CPC of the gene. No covariates other than the intercept were included in the model.

Figure S12. The subject-level overdispersions of genes with CPC>0.1% in five major cell types in the frontal cortex estimated by NEBULA with respect to the CPC of the gene. No covariates other than the intercept were included in the model.

bioRxiv preprint doi: https://doi.org/10.1101/2020.09.24.311662; this version posted September 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under a CC-BY-NC-ND 4.0 International license.

Figure S13. The distribution of (A) the estimated cell-level overdispersions and (B) subject-level overdispersions in the excitatory neurons (Ex) and the 11 subpopulations in the excitatory neurons (Ex0-Ex12). All genes with CPC>0.1% in each of the cell populations were included.

Figure S14. The cell-level overdispersions estimated by NEBULA versus those estimated by *glmer.nb* with nAGQ=0 for 16,207 genes with CPC>0.1% in the excitatory neurons.

Figure S15. Empirical type I error rate of testing a subject-level predictor using PGMM under different combinations of the number of individuals, σ^2, and ϕ. The CPS value was set at 200. The empirical type I error rate was calculated from five hundred simulated replicates in each of the scenarios and was evaluated at the significance level of 0.05 (the dashed lines).