FLASH-MM: fast and scalable single-cell differential expression analysis using linear mixed-effects models

Changjiang Xu* (1), Delaram Pouyabahar* (1,2), Veronique Voisin (1), Hamed Heydari (1,2), Gary D. Bader (1,2,3,4,5,6,7)

1 - The Donnelly Centre, University of Toronto, Toronto, Ontario, Canada
2 - Department of Molecular Genetics, University of Toronto, Toronto, Ontario, Canada
3 - Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
4 - Lunenfeld-Tanenbaum Research Institute, Toronto, Ontario, Canada
5 - Princess Margaret Research Institute, University Health Network, Toronto, Ontario, Canada
6 - CIFAR Multiscale Human Program, CIFAR, Toronto, Ontario, Canada
7 - Corresponding author
*Equal contribution

Abstract

Single-cell RNA sequencing (scRNA-seq) enables detailed comparisons of gene expression across cells and conditions. Single-cell differential expression analysis faces challenges like sample correlation, individual variation, and scalability. We developed a fast and scalable linear mixed-effects model (LMM) estimation algorithm, FLASH-MM, to address these issues. We reformulate aspects of the model estimation procedure to make it faster, by reducing computational complexity and memory use in the case of working with a gene by cell matrix. Simulation studies with scRNA-seq data show that FLASH-MM is accurate, computationally efficient, effectively controls false positive rates, and maintains high statistical power in differential expression analysis. Tests on tuberculosis immune and kidney single cell data demonstrate FLASH-MM's utility in accelerating single-cell differential expression analysis across diverse biological contexts.

Introduction

Differential expression (DE) analysis has long been a cornerstone of transcriptomics research. Single-cell RNA sequencing (scRNA-seq) technology enables researchers to profile the transcriptomes of individual cells, uncovering transcriptional similarities and differences across various biological conditions for specific cell types. Advancements in cost efficiency and throughput have facilitated the generation of large-scale datasets comprising hundreds of subjects and millions of cells, opening new avenues for exploring cellular heterogeneity and dynamics across diverse conditions and samples. Cells from the same individual share common genetic and environmental backgrounds, resulting in a hierarchical structure and statistical dependencies among individual cells in scRNA-seq data[1]. This introduces significant challenges for differential expression analysis in single-cell studies, particularly due to the correlation within cell populations of each subject (intra-subject correlation[1]) and the high variability across cell populations from different subjects (inter-subject variability[2]). Ignoring these correlations and variabilities can inflate false positive rates in statistical tests[1,2]. Furthermore, the large scale of single-cell data, often encompassing hundreds of thousands to millions of cells, adds computational complexity, requiring efficient methods to manage and analyze these vast datasets effectively.

Linear mixed-effects models (LMMs) provide a framework to address the challenges of intra-subject correlation and inter-subject variability in single-cell differential expression analysis by incorporating fixed effects, which capture systematic differences across experimental conditions, and random effects,

which model the correlations within subjects and the variations between subjects[1,3–5]. In multi-subject single-cell studies, cells are nested within subjects, and subjects are often nested within experimental conditions, meaning that cells from the same subject are correlated, and subjects within the same condition share common sources of variation. To account for this hierarchical structure, instead of modeling subjects as fixed effects, mixed models treat them as random effects, efficiently capturing both within-subject correlation and between-subject variability. Many methods and packages have been developed to fit the mixed-effects models[5–10]. The most widely used package is lme4[6], which uses maximum likelihood[11] or restricted maximum likelihood[12] methods for model fitting. However, fitting the LMM is computationally demanding, particularly in large-scale single-cell datasets, where standard implementations struggle with memory usage and runtime constraints[5–10]. As a result, the application of existing tools to large-scale scRNA-seq data has been limited due to the challenges of computational efficiency and memory intensity. As a result, the performance of mixed-effects models for single-cell DE analyses has mostly been examined through simulation studies involving small numbers of subjects and cells or pseudobulk methods[2,13].

To address the challenges of large-scale scRNA-seq data analysis using linear mixed-effects models, we developed FLASH-MM, a fast and scalable LMM estimation algorithm for single-cell differential expression analysis. By leveraging summary statistics, which are precomputed aggregate representations that capture essential information from the data without storing measurements for each individual cell, and by transferring the computation of high-dimension matrices (number of cells) to a lower dimension (numbers of covariates and random effects) in the model estimation step, FLASH-MM achieves both computational efficiency and significantly lower memory usage. Compared to the standard LMM estimation method lmer in the lme4 package[6], the FLASH-MM algorithm requires orders of magnitude less compute time and memory use while maintaining accuracy. We verified the accuracy, efficiency, and DE analysis performance of FLASH-MM using simulation studies. We simulated multi-subject multi-cell-type scRNA-seq datasets using real reference data based on a negative binomial (NB) distribution. We further demonstrate the application of FLASH-MM for case-control comparisons in a tuberculosis immune atlas and for cell-type-specific sex comparisons in healthy kidney data. In summary, FLASH-MM accelerates single-cell differential expression analysis and improves accuracy across diverse biological contexts, supporting the use of mixed models in large-scale, multi-subject single-cell studies.

Results

Overview of FLASH-MM

Single-cell RNA sequencing datasets typically consist of gene expression measurements for thousands of genes (approximately 20,000 in humans) across tens of thousands to millions of cells, often collected from multiple subjects or experimental conditions. We use an LMM to identify differentially expressed genes, where fixed effects are often covariates such as batch, sex, or treatment conditions, and subjects are often modeled as random effects.

We developed FLASH-MM to address the computational challenge of fitting the LMM for large-scale scRNA-seq data by efficiently estimating LMM parameters, using restricted maximum likelihood (REML)[12] with a gradient descent approach (see Supplementary Information). Instead of directly processing large cell-level data matrices, our algorithm computes and operates on summary statistics, which are compact data representations without storing information for each individual cell. Specifically, FLASH-MM operates the matrix computation by transferring the high-dimension n x n matrices (number of cells) to the low-dimension p x p and q x q matrices (numbers of fixed and random effects). This

reformulation substantially reduces computational complexity from $O(mn^3)$ to $O(mn(p^2 + q^2))$, and memory complexity from $O(mn)$ to $O(m*max(p,q))$, where m represents the number of genes, n is the number of cells and p and q denote the number of fixed and random effects, respectively ($n > p$ and $n > q$). By precomputing and directly using the summary statistics as inputs, FLASH-MM further reduces the computational complexity to $O(m(p^3 + q^3))$, which makes LMM estimation independent of the number of cells and achieves both speed and memory efficiency (see Methods and Supplementary Information).

With usual LMM fitting methods, variance components are constrained to be non-negative. As a result, the asymptotic normality of the maximum likelihood estimation of variance components at the null hypothesis is invalid due to the zero variance components being on the boundary of the parameters[14]. Our algorithm allows variance component parameters to take negative values such that the zero variance components are no longer on the boundary of the parameter space. Thus, the usual asymptotic properties of maximum likelihood estimation at the null hypothesis remain valid under regularity conditions, which enables the use of t-statistics or z-statistics for hypothesis testing of both fixed effects and variance components.

The FLASH-MM workflow for single-cell differential expression analysis is illustrated in Figure 1. Figure 1A shows the input structure of scRNA-seq data, represented as a log-transformed gene-by-cell count matrix, where each row corresponds to a gene's expression profile. Fixed effects can include variables such as log-library size, batch effects, biological conditions of interest, and interactions between conditions and cell types. Random effects can capture variations across subjects and correlations within subjects. Figure 1B outlines the linear mixed-effects model framework, constructed for each gene using design matrices of fixed and random effects, informed by prior knowledge of covariates and the biological question. Figure 1C illustrates the model fitting process, comprising LMM parameter estimation and hypothesis testing. Parameter estimation is performed using a gradient descent algorithm applied to summary statistics. Hypothesis testing evaluates fixed effects and their contrasts using t-statistics (see Methods and Supplementary Information).

Simulation studies

We validated FLASH-MM's accuracy and computational efficiency by comparing it to the standard linear mixed model method lmer, implemented in the lme4 package[6], using simulated scRNA-seq data (See Methods and Supplemental Information). We also evaluated the performance of FLASH-MM in single-cell DE analysis through simulations using two key criteria: (1) control of the Type I error rate (false positive rate, FPR) and (2) statistical power (true positive rate, TPR). These metrics were compared against NEBULA[5], a generalized linear mixed-effects model (GLMM) designed for DE analysis.

Simulating scRNA-seq data

Using PBMC 10X droplet-based scRNA-seq data from lupus patients (Kang et al. 2018) as a reference, we simulated six multi-subject multi-cell-type scRNA-seq datasets with 6,000 genes and sample sizes from 20,000 to 120,000 cells, based on a negative binomial (NB) distribution. Genes were randomly selected from the reference dataset, and cells were simulated from 25 subjects across 12 cell types under two treatment conditions. Treatments, cell types, and subjects were assigned randomly with equal probability. A total of 480 differentially expressed genes were designated, specific to a cell type (see Supplementary Information).

FLASH-MM has the same accuracy as lmer, but is much faster

We first validated FLASH-MM's accuracy and computational efficiency using simulated data by comparing it to the standard method, lmer, from the lme4 package[6] with its default settings. The linear mixed model (LMM) was fit to the log-transformed counts using FLASH-MM and lmer. Because lmer in the lme4 package doesn't provide p-values for hypothesis tests of coefficients, we computed p-values in this case by refitting the LMM by the lmer using the lmerTest package[15]. The differences in estimated variance components, coefficients, and p-values between FLASH-MM and lmer are shown in Figure 2A. The model parameters (coefficients and variance components) estimated by FLASH-MM and lmer are identical up to the sixth decimal place, demonstrating the high accuracy of the FLASH-MM implementation. FLASH-MM demonstrates substantially greater computational efficiency compared to lmer with default optimizer settings, achieving a speedup of approximately 50-fold to 140-fold as the sample size increases from 20,000 to 120,000 cells (Figure 2B, Table S1, measured on a 2.8 GHz Quad-Core Intel Core i7 processor with 16 GB DDR3 RAM).

FLASH-MM has a similar statistical performance to NEBULA, but is much faster

We compared the performance of FLASH-MM to NEBULA, a state-of-the-art method for differential expression analysis of multi-subject single-cell data based on Negative Binomial and Poisson mixed models. We ran NEBULA on our simulated data using the arguments method='LN' and model='NBLMM', which specify the negative binomial lognormal mixed model—the same model we used to simulate scRNA-seq data above. We computed type I error using the simulated non-differentially expressed (non-DE) and DE genes and compared the power between the two methods at a sample size of 120,000 cells. FLASH-MM effectively controls Type I error, with p-values remaining within the expected range of a uniform distribution, as shown in the quantile-quantile (QQ) plot of Figure 2C. This expected range is represented by 95% confidence intervals, which indicate the natural variation we would expect if there were no true differences in gene expression. FLASH-MM p-values fall within these confidence intervals, suggesting that the method does not produce an excess of false positives. The plot also shows a slight p-value deflation in NEBULA, meaning its p-values tend to be slightly higher than expected under the null hypothesis. FLASH-MM demonstrates a power comparable to NEBULA, with a Receiver Operating Characteristic (ROC) curve achieving an Area Under the Curve (AUC) of 0.97 for both FLASH-MM and NEBULA (Figure 2D). Additional QQ plots and ROC curves indicate comparable results across various sample sizes (Extended Figures 1 and 2). The t-values and p-values calculated by FLASH-MM and NEBULA demonstrate a strong correlation (Extended Figure 3). However, both NEBULA and lmer required significantly longer runtimes compared to FLASH-MM (319 and 143 times, respectively, at n=120000 cells, Table S1).

FLASH-MM supports DE analysis of diverse biological scRNA-seq data

Kidney scRNA-seq data

We first examined the sex variation within healthy kidney cell types using kidney scRNA-seq data[16] (Figure 3A). The kidney data from 19 subject samples contains 14,175 genes and 27,550 cells consisting of 18 cell types after quality control (Figure 3B). We performed a differential expression analysis using FLASH-MM to identify the DE genes between males and females within each cell type while considering the subjects as a random effect.

Among the various cell populations, connecting tubule (CNT) cells have the highest number of sex-specific differentially expressed genes (200), meeting the criteria of FDR < 0.05 and coefficient < -0.5 or > 0.5 (Figures 3C and 3D). Pathway analysis of these genes highlights distinct enrichments: male

CNT cells show enrichment for pathways related to acid secretion, transporter activity, blood pressure regulation, and ion importation, whereas female CNT cells exhibit enrichment for proton transporter activity and positive regulation of receptor recycling (Figures 3E and 3F).

Tuberculosis (TB) scRNA-seq data

We then applied FLASH-MM to single-cell transcriptomics data from 500K memory T cells from 259 donors in a tuberculosis (TB) progression cohort[17]. After quality control, the large TB dataset contains 11,596 genes and 499,713 cells covering 29 cell states from 46 batches and 259 individual donors. We applied FLASH-MM to identify genes associated with TB status within each cluster. FLASH-MM identified a varying number of differentially expressed genes associated with TB progression across different cell states (Figure 4C; FDR < 0.05 and coefficient > 0.2). The cell types with the highest number of DE genes are activated CD4+ and activated CD8+ populations, with 1266 and 268 DE genes, respectively (Figures 4A and 4B). To further investigate TB-associated signatures, we identified the top TB-enriched genes within these two cell states (Figure 4C). Pathway enrichment analysis of these DE genes identifies cell-cycle pathways in activated CD4+ cells, while activated CD8+ cells show enrichment for pathways related to immune response, TCR-mediated T-cell activation, and chemokine signaling (Figure 4D).

Discussion

Differential expression analysis to detect changes in gene expression across conditions has long been a fundamental aspect of transcriptomics research. However, single-cell RNA sequencing data introduces unique statistical challenges, such as sample correlation, individual variation, and scalability. We developed FLASH-MM as a fast and scalable linear mixed-effects model (LMM) estimation algorithm to address these issues.

Mixed-effects models are powerful tools in single-cell studies due to their ability to model intra-subject correlations and inter-subject variabilities. Classic LMM estimation methods, like lmer in the lme4 package[6], face limitations of speed and memory use in the analysis of large-scale scRNA-seq data. These limitations have encouraged researchers to use traditional bulk RNA-seq differential expression analysis methods with pseudobulk counts by summing reads within each cell type for each subject[3]. While this simplifies the analysis, it sacrifices the resolution inherent in single-cell data and can be problematic for comparisons involving weaker cell state differences or continuous cellular gradients.

To support our simulation studies, we developed a scRNA-seq simulator, implemented in the simuRNAseq function in the FLASH-MM software distribution, to generate multi-subject multi-cell-type scRNA-seq data based on a negative binomial distribution (Supplemental information). simuRNAseq shares similarities with muscat[18] and GLMsim[19]. Both muscat and GLMsim have notable limitations when used for scRNA-seq simulations. Muscat estimates the dispersion of the negative binomial distribution using the edgeR package, but it relies on only a subset of the reference data. As a result, it cannot scale to large scRNA-seq datasets and captures only partial information from the reference dataset. GLMsim estimates the coefficients and dispersion parameters of the NB model for each gene using glm.nb from the MASS package. While effective, this approach is computationally intensive and limited to generating data of a fixed size that matches the reference data. Our scRNA-seq simulator uses the method-of-moments estimate (MME[20]) to compute dispersion parameters for the NB distribution. This approach is faster, more flexible, and uses the full biological reference dataset.

Constructing design matrices for fixed and random effects is an important step in LMM-based DE analysis, requiring identification of the aspects of the data to be modeled, and a balance between reducing residual variance, avoiding overfitting, and managing collinearity among covariates[21]. A design matrix encodes variables such as sample conditions, batch effects, and cell types, specifying how observations are mapped to model parameters. For single cell transcriptomics data, including library size as a fixed effect helps control p-value inflation and should generally be included in the model to help with normalization. Other model design decisions must be defined by the user based on the structure of their data and the biological question under study. If samples are modeled as random effects, their number may be high and the FLASH-MM LMM algorithm may slow down. In such cases, subsampling the data or applying dimensionality reduction techniques, such as PCA, on the random effects can help reduce their number to improve computational efficiency (see Supplemental Information).

Using FLASH-MM, we can comprehensively evaluate the performance of the LMM in large-scale scRNA-seq differential expression analysis. In the future, FLASH-MM's versatile framework can be extended to other data modalities, such as spatial transcriptomics and multiomics. Expanding its application across diverse biological data could provide new opportunities for uncovering novel insights and facilitating integrated analyses in a wide range of research contexts. FLASH-MM's ability to efficiently handle linear mixed models with a large number of samples and a relatively small number of fixed and random effects also makes it broadly applicable to fields beyond biology where similar statistical challenges may arise.

Acknowledgments

This research was supported by the Canadian Institutes for Health Research (grant PJT 469829 to GDB) and by the University of Toronto's Data Sciences Institute Doctoral Student fellowship program.

Author contributions

Conceptualization, D.P., V.V, C.X and G.D.B.; Methodology, C.X, H.H; Formal analysis, D.P., C.X; Writing – Original Draft, D.P., C.X; Writing – Review & Editing, D.P., C.X, V.V and G.D.B.; Supervision, G.D.B.; Funding Acquisition, G.D.B.

Declaration of interests: GDB is on the Scientific Advisory Board of Adela Bio. No other competing interests are declared.

Main Figure Legends

Figure 1. FLASH-MM workflow for single-cell differential expression analysis. A. Data: gene expression matrix Y = log(1 + counts), with each row corresponding to a gene's expression profile and each column corresponding to a cell (gene x cell matrix). Metadata includes various variables such as log-library size, batch effects, biological conditions of interest, and interactions between conditions and cell types, which could be modeled as fixed effects, and individual subjects, which could be modeled as random effects. B. Model: the linear mixed-effects model (LMM) for each gene by design matrices X and Z, which are constructed based on prior knowledge about the covariates and the biological question. C. Model fitting: comprises LMM estimation and tests. LMM estimation is implemented by a gradient descent algorithm over summary statistics. The summary statistics are computed as X^TX, X^TY^T, Z^TX,

Z^TY^T, and Z^TZ. LMM tests perform hypothesis tests on the fixed effects and their contrasts using t-statistics.

Figure 2) Computational and statistical performance of FLASH-MM in differential expression analysis of simulated scRNA-seq data. A) Boxplots of differences: comparison of variance components, coefficients, and p-values between FLASH-MM and lmer fitting across various sample sizes. B) Computation time: comparison of computation time (in minutes) for FLASH-MM, lmer, and NEBULA across different sample sizes. C) QQ-plots of non-DE genes (negative controls) p-values for FLASH-MM and NEBULA. The grey area represents the 95% confidence interval, indicating the expected range under the null hypothesis. D) ROC curves for FLASH-MM and NEBULA.

Figure 3) FLASH-MM identifies sex-specific variations in a healthy human kidney map. A) UMAP projection of the human healthy kidney transcriptomic map shows connecting tubule (CNT) cells in purple, while other populations are shown in lighter tones for contrast. B) Bar plot indicates the fraction of cells from each sex within different cell types. C) The top male-specific differentially expressed genes within the CNT population are identified. Score is calculated as -log(adjusted p-value) x coefficient. Genes are sorted by their score, and dot colors indicate -log(adjusted p-value). D) The top female-specific differentially expressed genes within the CNT population are shown. Pathway enrichment results based on male-specific and female-specific DE genes in the CNT population are shown in panels E and F, respectively.

Figure 4) FLASH-MM identifies TB-enriched signatures within T cell populations while accounting for confounding variables. A) Bar plots indicate the number of differentially expressed genes for each cell type. B) UMAP projection of single-cell RNA-seq data from 500K memory T cells across 259 individuals in a TB progression cohort. CD4+ activated and CD8+ activated T cell populations are highlighted in blue and dark orange, respectively, while other populations are shown in lighter tones for contrast. C) The top TB-associated differentially expressed genes within the $CD4^+$ activated and $CD8^+$ activated T cell populations are identified. Score is calculated as -log(adjusted p-value) x coefficient. Genes are sorted by their score and dot colors indicate -log(adjusted p-value) D) Pathway enrichment results are presented for the TB-enriched differentially expressed genes within the activated $CD8^+$ and $CD4^+$ T cell populations.

Supplementary files:
Extended data: Extended Figures 1-3
Supplementary information: Includes Figures S1, S2 and Table S1

Methods

LMM estimation and hypothesis testing

Consider the linear mixed-effects model (LMM) as expressed below[22]

$$y = Xb + Zb + e$$

where y is a $n \times 1$ vector of observed response (expression for a gene), X is a $n \times p$ design matrix for fixed effects β, Z is a $n \times q$ design matrix for random effects b, and ϵ is a $n \times 1$ vector of residual errors. The term random effects may be a combination of various random-effect components

$$Zb = Z_1 b_1 + \cdots + Z_K b_K,$$

where $Z = [Z_1, \ldots, Z_K]$, $b = [b_1^T, \ldots, b_K^T]^T$, K is the number of the random-effect components, and Z_i is a $n \times q_i$ design matrix for the *i*-th component. The superscript T denotes a transpose of a vector or matrix.

The basic assumptions are as follows: (1) The design matrix X is of full rank, satisfying conditions of estimability for the parameters (needed to support matrix inversion); (2) The random vectors b_i and ϵ are independent and follow a normal distribution, $b_i \sim N(\mathbf{0}, \sigma_i^2 I_{q_i})$ and $\epsilon \sim N(\mathbf{0}, \sigma^2 I_n)$. Here σ_i^2 and σ^2 are unknown parameters, called variance components, $\mathbf{0}$ is a vector or matrix of zero elements, and I_n is an $n \times n$ identity matrix. The random effects reflect variations between groups (subjects) and correlations within groups (subjects). Assumption (1) implies $p < n$. We also assume $q_i < n$. If $q_i > n$, we can use principal component analysis (PCA) to obtain an equivalent LMM with the number of random effects less than *n* (see Supplementary Information).

Maximum likelihood estimation (MLE) and restricted maximum likelihood (REML) are methods for estimating fixed effects and variance components in LMMs. MLE estimates all parameters of fixed effects and variance components together but can produce biased variance component estimates, whereas REML removes fixed effects from variance estimation, resulting in unbiased estimates. Both methods are asymptotically identical. Estimating variance components using either MLE or REML requires numerical methods, among which the iterative gradient-based methods are the most commonly used[22].

Fast and scalable algorithm: The gradient descent methods usually have a computational complexity of $O(n^3)$. We developed the summary statistics-based algorithm, FLASH-MM, to implement the gradient methods for speeding up the LMM estimation and reducing the computer memory usage. Instead of the individual cell-level data: X, Z and y, FLASH-MM uses the summary statistics: $X^T X$, $X^T y$, $Z^T X$, $Z^T y$ and $Z^T Z$ to estimate the LMM parameters. FLASH-MM achieves a computational complexity of $O(n(p^2 + q^2))$, which is fast and linearly scalable with the sample size n. These summary statistics have a low dimension and require less computer memory. By precomputing and directly using the summary statistics as inputs, the algorithm complexity is reduced to $O(p^3 + q^3)$, which makes computations independent of the sample size n (number of cells) and achieves both speed and memory efficiency (see Supplementary Information).

ScRNA-seq data typically consist of gene expression measurements for thousands of genes (approximately 20,000 in human) across thousands to millions of cells. In the single-cell differential expression analysis, the complexity of FLASH-MM is $O(mn(p^2 + q^2))$, m is the number of genes, which is linearly scalable with the number of genes. By using the pre-computed summary statistics as inputs, the algorithm complexity becomes $O(m(p^3 + q^3))$.

Hypothesis testing: The hypothesis testing for fixed effects and variance components can be respectively defined as:

$$H_{0,i}: \beta_i = 0 \text{ versus } H_{1,i}: \beta_i \neq 0,$$

$$H_{0,k}: \sigma_k^2 = 0 \text{ versus } H_{1,k}: \sigma_k^2 > 0.$$

The variance components under null hypothesis, $\sigma_k^2 = 0$, are on the boundary of the parameter space, in which case the MLE asymptotic normality is inappropriate. By reparameterizing the variance components, $\theta_k = \sigma^2 \gamma_k$, the covariance matrix, $V_\theta = \sigma^2 (I + \gamma_1 Z_1 Z_1^T + \cdots + \gamma_K Z_K Z_K^T)$, becomes positive-definite and well-defined when $\gamma_k > -1/\lambda_{max}$, where $\lambda_{max} > 0$, is the largest singular value of ZZ^T. Now the parameters of variance components, θ_k, $k = 1, \cdots, K$, can be negative. If $\theta_k > 0$, $\sigma_k^2 = \theta_k$ is

definable and the mixed model is well-specified. Otherwise, the term of random effects is not needed in the model. Then the hypotheses for the variance components are extended as:

$$H_{0,k}: \theta_k \leq 0 \text{ versus } H_{1,k}: \theta_k > 0,$$

in which the zero components, $\theta_k = 0$, are no longer on the boundary of the parameter space and the MLE asymptotic properties hold. Then we can use z-statistic or t-statistic for hypothesis testing of fixed effects and variance components (see Supplementary Information).

Simulation methods

We generated the multi-subject multi-cell type scRNA-seq dataset by using reference data based on a negative binomial (NB) distribution. The mean of NB distribution is taken as the sample mean for each gene. The dispersion of NB distribution is computed by the method-of-moments estimate (MME)[20]. Compared to the maximum likelihood-based estimates, such as the glm.nb function in the MASS package, the MME is computationally simpler and performs reasonably well. See detailed simulation methods and performance in Supplemental Information.

Biological data preprocessing and model design

Healthy human kidney atlas

The healthy human kidney transcriptomic map[16] was generated from 27,677 cells obtained from 19 living donors (10 female and 9 male). Decontaminated raw data, processed using SoupX[23], was provided upon request by the authors. Cells were filtered based on three criteria: the number of detected features (nFeature), library size, and the number of cells within each cell type (cluster). The number of features per cell, defined as the total number of non-zero genes, was required to meet a minimum threshold of 100. Library size, calculated as the total counts per cell, was restricted to a range of 2^9 to 2^16. Cell types with fewer than 20 cells were excluded, and “Podocyte” cells were removed due to their low sample size (16 cells post filtering). Additionally, genes were filtered based on their expression levels, where the counts per cell ratio had to exceed 0.005 (i.e., the total gene count divided by the number of cells had to be greater than 0.5%). Genes were further filtered to retain those expressed in at least 16 cells, with a minimum of 10 cells in each group, total counts between 2^6 and 2^20, and a counts per cell ratio above the threshold. After these filtering steps, 27,550 cells and a refined set of genes were retained for downstream analyses.

The FLASH-MM model was designed to account for both technical and biological variation. The model formula used was:

~log(library.size) + Cell_Types_Broad + Cell_Types_Broad:sex + (1|sampleID),

where log-transformed library size was included as a fixed effect to normalize for differences in sequencing depth, Cell_Types_Broad captured the cell type-specific effects, and the interaction term Cell_Types_Broad:sex identified sex-specific differences within each cell type. A random effect (1|sampleID) was added to account for inter-sample variability.

Tuberculosis (TB) T cell scRNA-seq data

The Tuberculosis (TB) memory T cell dataset was obtained from Nathan et al., 2021[17], comprising 500,089 cells. The raw count matrix was used for preprocessing. Metadata associated with the cells

includes donor identity, sex, batch, cluster annotations, and TB status. The dataset includes cells from 259 unique donors, spanning 46 batches and 29 cell clusters. In the DE analysis, we modeled the donors as a random effect and ignored the batch effect because the majority of donors were sequenced in a single batch.

Pre-processing steps consisted of removing cells and genes to remove extreme values. Library sizes were assessed using a boxplot, and cells with library sizes outside the lower whisker and above an upper threshold of 2^15 were removed, resulting in the retention of 499,973 cells. Genes were filtered in two steps: first, genes expressed in fewer than 2^9 cells (512 cells) were excluded; second, genes with a counts-per-cell ratio less than 0.005 (i.e., total gene count divided by the number of cells below 0.5%) were removed. These filtering steps reduced the dataset to 11,596 genes and 499,973 cells, which were used for downstream analyses.

The FLASH-MM model was designed to analyze this dataset, accounting for both technical and biological variations. The model formula used was:

~log(library.size) + cluster_name + cluster_name:TB_status + (1|donor),

where log-transformed library size normalized for differences in sequencing depth, cluster_name captured the cell type-specific effects, and the interaction term cluster_name: TB_status identified TB-associated differences within each cell type. A random effect (1|donor) was added to account for inter-donor variability.

Software availability

An R package for FLASH-MM, including example case scenarios, is publicly available at: https://github.com/BaderLab/FLASHMM.
Scripts for data analysis and manuscript figure generation can be found at: https://github.com/BaderLab/FLASH-MM-analysis.

References

1. Zimmerman, K. D., Espeland, M. A. & Langefeld, C. D. A practical solution to pseudoreplication bias in single-cell studies. *Nat. Commun.* **12**, 738 (2021).
2. Squair, J. W. *et al.* Confronting false discoveries in single-cell differential expression. *Nat. Commun.* **12**, 5692 (2021).
3. Hoffman, G. E. & Roussos, P. Dream: powerful differential expression analysis for repeated measures designs. *Bioinformatics* **37**, 192–201 (2021).
4. Gagnon, J. *et al.* Recommendations of scRNA-seq Differential Gene Expression Analysis Based on Comprehensive Benchmarking. *Life (Basel)* **12**, (2022).
5. He, L. *et al.* NEBULA is a fast negative binomial mixed model for differential or co-expression analysis of large-scale multi-subject single-cell data. *Commun. Biol.* **4**, 629 (2021).
6. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
7. Finak, G. *et al.* MAST: a flexible statistical framework for assessing transcriptional changes and characterizing heterogeneity in single-cell RNA sequencing data. *Genome Biol.* **16**, 278 (2015).
8. Brooks, M., E. *et al.* glmmTMB Balances Speed and Flexibility Among Packages for Zero-inflated Generalized Linear Mixed Modeling. *R J.* **9**, 378 (2017).
9. Bürkner, P.-C. brms: an *R* package for bayesian multilevel models using *stan*. *J. Stat. Softw.* **80**,

(2017).
10. Pinheiro, J. C. & Bates, D. M. *Mixed-Effects Models in S and S-PLUS*. (Springer-Verlag, 2000). doi:10.1007/b98882.
11. Hartley, H. O. & Rao, J. N. K. Maximum-likelihood estimation for the mixed analysis of variance model. *Biometrika* **54**, 93–108 (1967).
12. Patterson, H. D. & Thompson, R. Recovery of inter-block information when block sizes are unequal. *Biometrika* **58**, 545–554 (1971).
13. Nguyen, H. C. T., Baik, B., Yoon, S., Park, T. & Nam, D. Benchmarking integration of single-cell differential expression. *Nat. Commun.* **14**, 1570 (2023).
14. Self, S. G. & Liang, K.-Y. Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *J. Am. Stat. Assoc.* **82**, 605 (1987).
15. Kuznetsova, A., Brockhoff, P. B. & Christensen, R. H. B. lmertest package: tests in linear mixed effects models. *J. Stat. Softw.* **82**, 1–26 (2017).
16. McEvoy, C. M. *et al.* Single-cell profiling of healthy human kidney reveals features of sex-based transcriptional programs and tissue-specific immunity. *Nat. Commun.* **13**, 7634 (2022).
17. Nathan, A. *et al.* Multimodally profiling memory T cells from a tuberculosis cohort identifies cell state associations with demographics, environment and disease. *Nat. Immunol.* **22**, 781–793 (2021).
18. Crowell, H. L. *et al.* muscat detects subpopulation-specific state transitions from multi-sample multi-condition single-cell transcriptomics data. *Nat. Commun.* **11**, 6077 (2020).
19. Wang, J., Chen, L., Thijssen, R., Phipson, B. & Speed, T. P. GLMsim: a GLM-based single cell RNA-seq simulator incorporating batch and biological effects. *BioRxiv* (2024) doi:10.1101/2024.03.20.586030.
20. Clark, S. J. & Perry, J. N. Estimation of the Negative Binomial Parameter κ by Maximum Quasi - Likelihood. *Biometrics* **45**, 309 (1989).
21. Gelman, A. & Hill, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. (Cambridge University Press, 2006). doi:10.1017/CBO9780511790942.
22. Searle, S. R., Casella, G. & McCulloch, C. E. *Variance Components*. 536 (Wiley-Interscience, 2006).
23. Young, M. D. & Behjati, S. SoupX removes ambient RNA contamination from droplet-based single-cell RNA sequencing data. *Gigascience* **9**, giaa151. (2020).

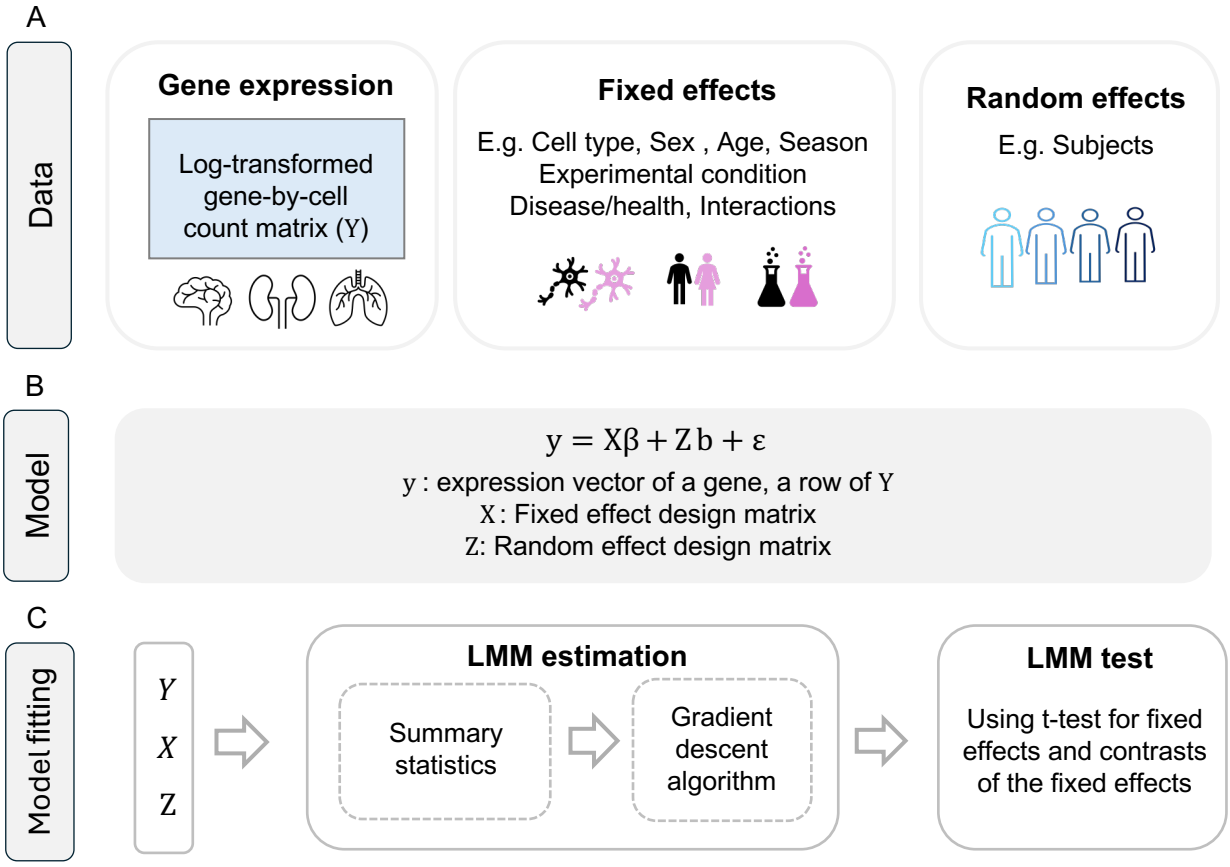

Figure 1) FLASH-MM workflow for single-cell differential expression analysis.

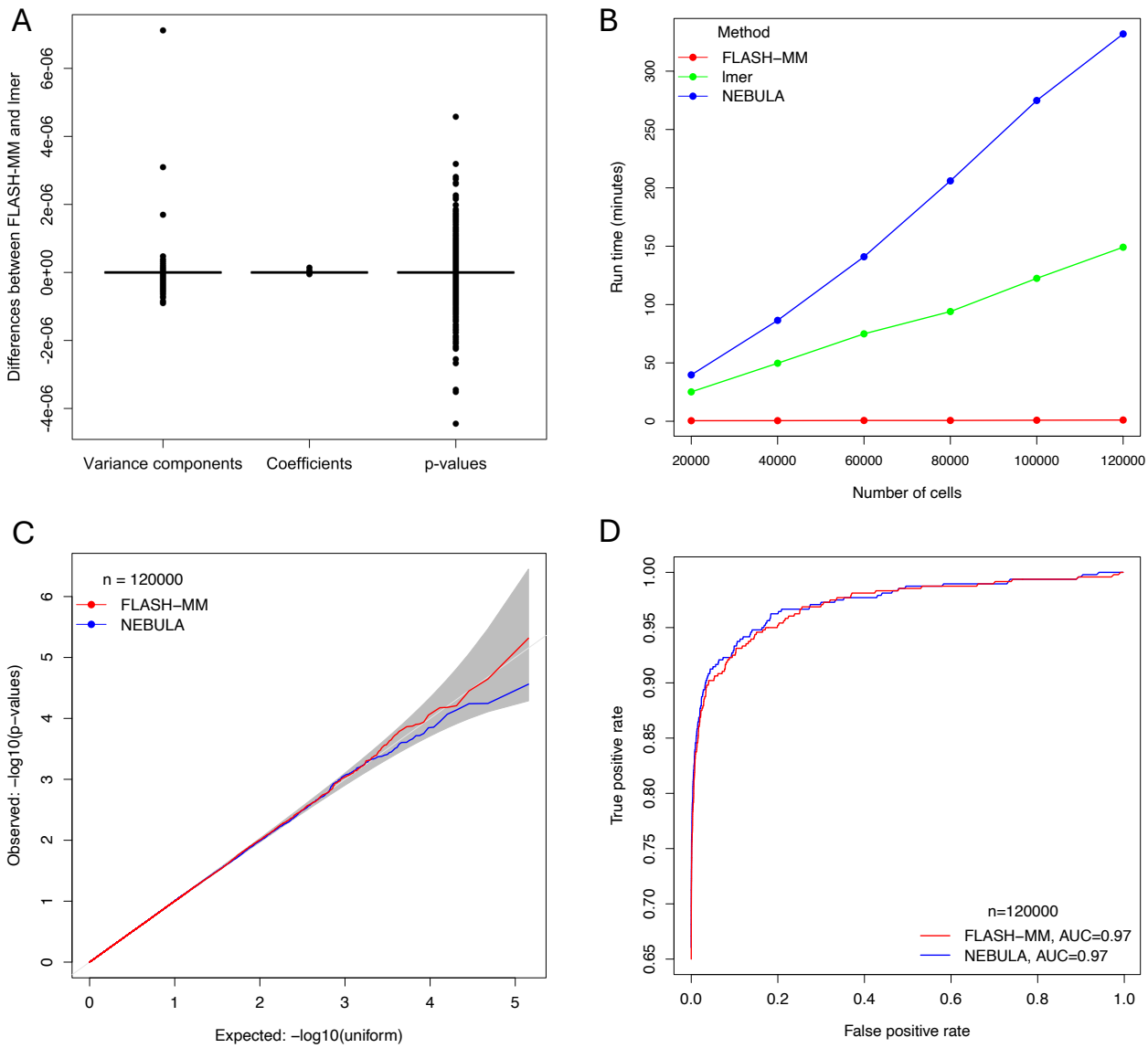

Figure 2) Computational and statistical performance of FLASH-MM in differential expression analysis of simulated scRNA-seq data.

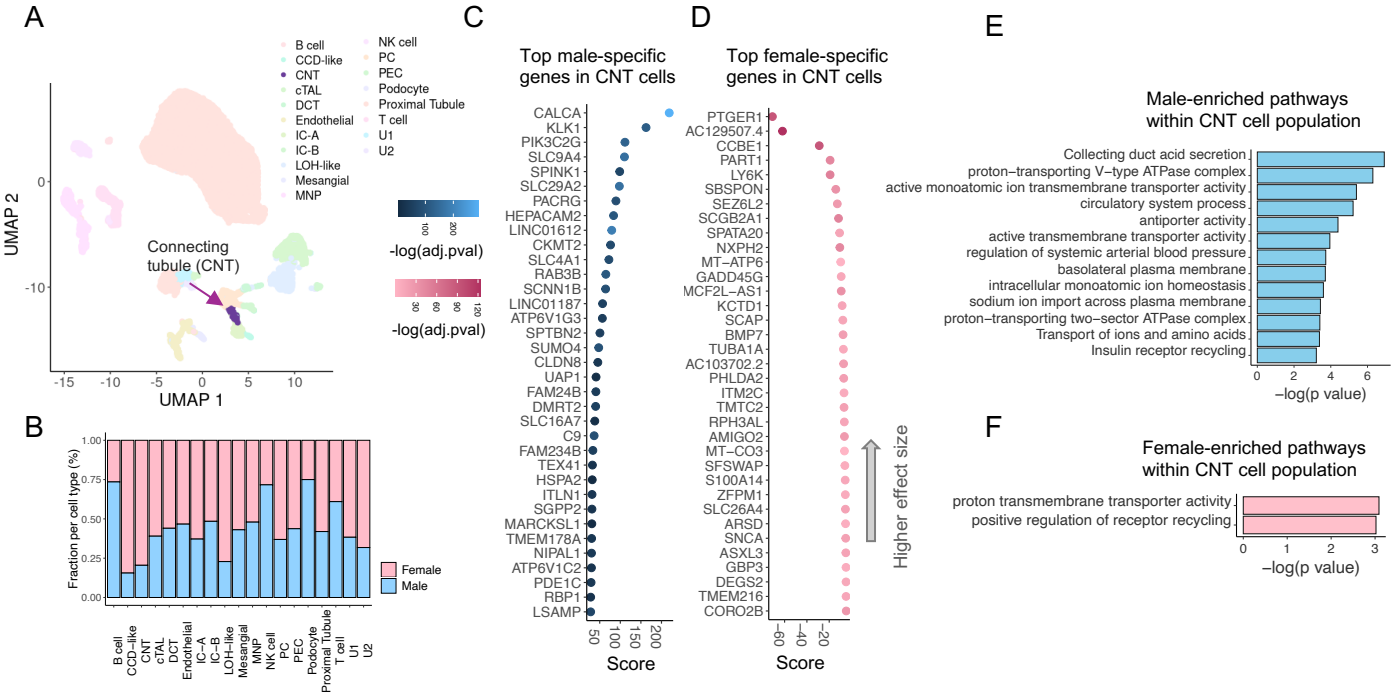

Figure 3) FLASH-MM identifies sex-specific variations in a healthy human kidney map.

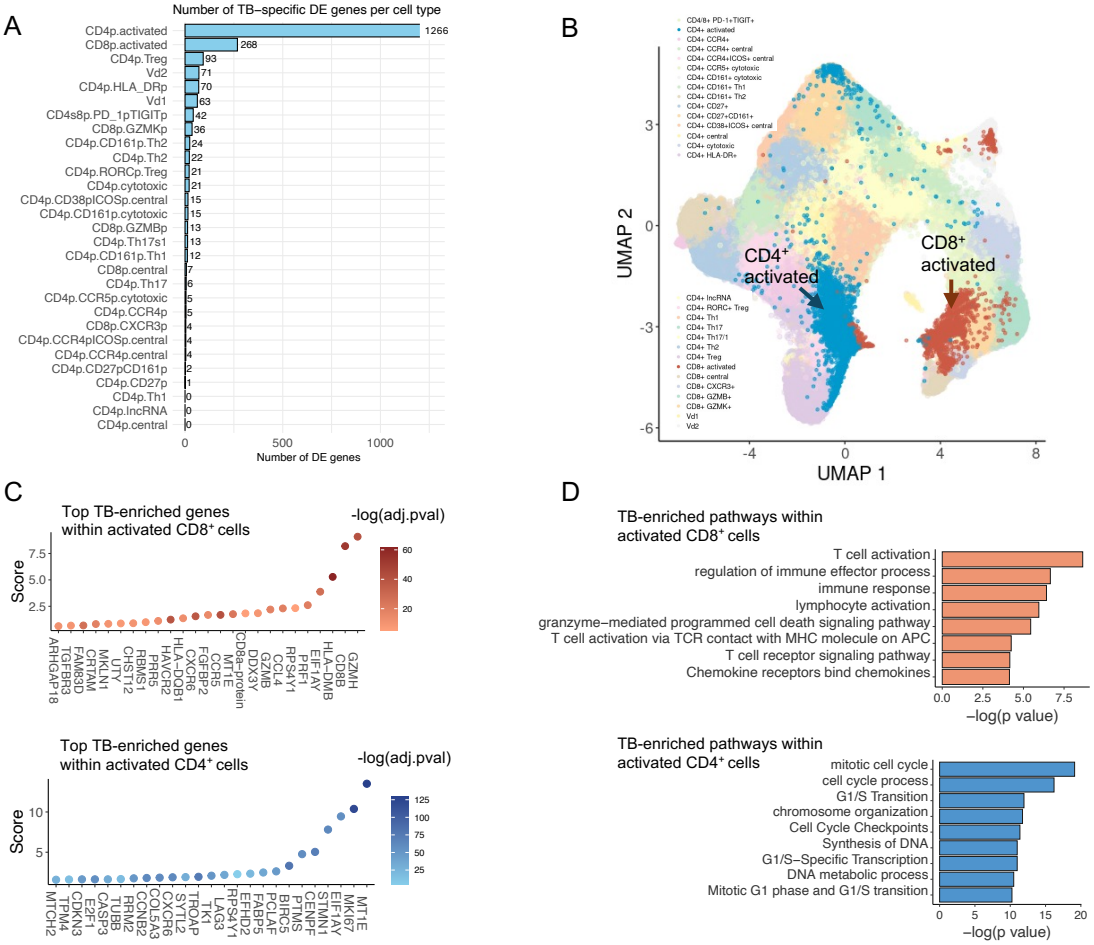

Figure 4) FLASH-MM identifies TB-enriched signatures within T cell populations while accounting for confounding variables.