Transnational University Limburg is a unique collaboration of two universities in two countries: the University of Hasselt and Maastricht University.

UHASSELT
KNOWLEDGE IN ACTION

www.uhasselt.be
Universiteit Hasselt
Campus Hasselt:
Martelarenlaan 42 | 3500 Hasselt
Campus Diepenbeek:
Agoralaan Gebouw D | 3590 Diepenbeek

UHASSELT
KNOWLEDGE IN ACTION

Maastricht University

Faculty of Sciences
School for Information Technology
Master of Statistics and Data Science

Master's thesis

Optimizing Cell Counts and Sample Sizes for Detecting Cell Type Abundance Changes in Compositional Single-Cell Experiments.

Lulu Asmi Lathifah Sundayana

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :
Prof. dr. Olivier THAS

SUPERVISOR :
Alemu Takele ASSEFA
Bie VERBIST

2024
2025

Maastricht University

Faculty of Sciences

School for Information Technology

Master of Statistics and Data Science

Master's thesis

Optimizing Cell Counts and Sample Sizes for Detecting Cell Type Abundance Changes in Compositional Single-Cell Experiments.

Lulu Asmi Lathifah Sundayana

Thesis presented in fulfillment of the requirements for the degree of Master of Statistics and Data Science, specialization Biostatistics

SUPERVISOR :

Prof. dr. Olivier THAS

SUPERVISOR :

Alemu Takele ASSEFA

Bie VERBIST

Acknowledgement

First and foremost, I praise Allah SWT for His mercy and guidance that enabled me to complete this thesis.

I would like to sincerely thank my thesis supervisor, Prof. Olivier Thas, for his guidance, support, and encouragement throughout this project. I am also very grateful to my external supervisors, Alemu Takele Assefa and Bie Verbist, for their insightful discussions, support, and valuable feedback.

I am deeply thankful to my parents for their unconditional love and unwavering support throughout my journey. I wouldn't have made it this far without you. I would also like to thank my family and friends for their continuous encouragement during this time.

Finally, I am grateful to VLIR-UOS for supporting my master's studies. Their support made this journey possible.

Abstract

Single-cell experiment enables researchers to uncover differences in cell composition between donors by profiling thousands of cells per sample. This approach is widely used, for example, to investigate disease mechanisms, such as immune response and tumor progression. However, designing an effective single-cell experiment remains challenging due to the compositional nature of the data, technical variability, and biological variability between donors. These factors complicate both the estimation of cell type abundance and the detection of differences between groups.

This thesis aims to support the design of single-cell experiments by investigating how the number of cells and the number of biological samples (donors) influence (1) the accuracy of cell type abundance estimation and (2) the ability to detect changes in abundance across groups. A simulation framework based on the Dirichlet-multinomial distribution was developed to generate single-cell count data under various experimental settings. Estimation accuracy was assessed using metrics based on relative error, while differential abundance testing was performed using *voomCLR* method.

The results indicate that cell count and sample size contribute differently to estimation accuracy and statistical power. Abundance estimation accuracy improved primarily with higher number of cells, while the number of donors had little additional effect. In contrast, statistical power for detecting changes in cell type abundance increased mainly with the number of donors, while increasing cell count had limited impact. These findings suggest that the number of cells and the number of donors contribute to different aspects of analysis performance and may need to be considered separately. Optimal study design should therefore align with the specific study objective, whether to estimate cell type proportions or detecting differences between groups.

Keywords: single-cell analysis, cell type abundance, Dirichlet-multinomial, compositional data analysis

Contents

1 Introduction

1.1 Background

Single-cell analysis refers to the study of individual cells to uncover the diversity and differences within cell populations. By understanding its characteristics, single-cell analysis can give information about which cell types exist in a sample and whether there are differences in cell type composition between individuals (Choi and Kim, 2019). This approach is valuable in disease research because it helps reveal important differences between cells. In cancer studies, for example, single-cell analysis has been used to better understand tumor composition, track how cancer changes over time, and support the development of more targeted treatments (Lei et al., 2021).

In practice, single-cell analysis is typically carried out through experiments in which thousands of cells are profiled per sample and classified into types, allowing for comparisons of their relative abundance across individuals (Lähnemann et al., 2020). Comparing these proportions across samples or conditions can reveal important biological differences. For example, changes in immune cell composition may signal the presence of infection or inflammation, while a decrease in specific cell types could signal disease progression. Single-cell analysis allows researchers to detect such changes in cell composition across individuals, which provide insights into underlying biological processes and disease mechanisms (Jovic et al., 2022). However, this analysis is complicated by the compositional nature of the data, as all cell type proportions must sum to one; an increase in one type automatically leads to decreases in other types, which introduce spurious correlations and inflate false discoveries if not properly accounted for (Quinn et al., 2018).

In addition to compositional constraints, single-cell data are also affected by multiple sources of variation. This includes technical variation due to random sampling of cells, as illustrated in the SCOPIT framework (Davis et al., 2019), as well as biological variation, which has been shown to reduce the true positive rate in differential abundance testing (Assefa et al., 2024). These factors can make it difficult to determine whether observed differences in cell type frequencies reflect true biological changes or random variation, especially when studying rare cell types. For example, the SCOPIT framework shows that the number of cells required depends on the expected frequency of each cell type, and a large number may be needed to ensure sufficient representation of

rare cell types (Davis et al., 2019). While SCOPIT provides guidance for planning the number of cells to sample in a single group, it does not address how accurately those proportions can be estimated or how donor-to-donor variation affects the estimation.

When the goal is to detect differences in cell type composition between groups, the number of biological samples plays a key role. Simulation results from Assefa et al. (2024) show that higher between-sample variability reduces the true positive rate in differential abundance testing, but this effect can be reduced by increasing sample size (Assefa et al., 2024). Methods such as LinDA and voomCLR have been developed to improve robustness in differential abundance testing by adjusting for compositional effects and variability across biological samples (Zhou et al., 2022; Assefa et al., 2024). However, they do not provide practical guidance on how to choose the number of cells or samples in designing the experiment. In contrast, Sensei proposed in Liang et al. (2022) directly addresses power and sample size calculation for detecting differences in cell type proportions across groups. It models the variability in cell type proportions across individuals using a beta-binomial model, but does not account for the compositional structure of the data or dependencies between cell types (Liang et al., 2022). As a result, there is a need for a clearer understanding of how cell count and sample size work together to affect estimation accuracy and testing performance in single-cell analysis.

1.2 Objectives

The objective of this study is to support the design of single-cell experiments aimed at cell type abundance estimation and comparison between groups of biological samples. Specifically, the first objective is to evaluate how the number of cells and the sample size influence the accuracy of cell type abundance estimation. The second objective is to assess how the number of cells and the sample size affect the ability to detect changes in cell type abundance between experimental groups. While the effects of cell count and sample size are often considered separately, this study takes an exploratory approach to investigate whether cell count and sample size can be unified into a common framework for guiding both estimation and testing in single-cell analysis.

2 | Methodology

This chapter describes the methodological framework used to evaluate how experimental design parameters influence estimation accuracy and the ability to detect differences in single-cell analysis. The approach is based on parametric simulation of single-cell count data using the Dirichlet-multinomial model, followed by downstream analysis to assess abundance estimation and differential abundance detection under varying design conditions.

2.1 The Dirichlet-multinomial distribution

In single-cell experiments, cell type composition data often take the form of counts, representing how many cells of each type are observed in a given donor. Because the total number of cells per donor is fixed, these counts are compositional, where they represent proportions constrained to sum to one. Moreover, biological and technical variability causes overdispersion, where there is more variation across donors than expected under a simple multinomial model. The Dirichlet-multinomial distribution is commonly used to model single-cell and other count data to account for both compositionality and overdispersion (Assefa et al., 2024; Fordyce et al., 2011; Harrison et al., 2020).

The Dirichlet-multinomial distribution is a compound distribution of a multinomial distribution and a Dirichlet distribution that models cell counts across multiple categories (cell types), where the underlying proportions themselves are random variables drawn from a Dirichlet distribution (Ng and Tian, 2011). Formally, for donor $i = 1, \ldots, n$ and cell type $j = 1, \ldots, P$, the model is defined as follows:

$$\boldsymbol{\pi_i} \sim \text{Dirichlet}(\boldsymbol{\theta_i}), \mathbf{Y_i} \mid \boldsymbol{\pi_i} \sim \text{Multinomial}(N_i, \boldsymbol{\pi_i}) \tag{2.1}$$

where:
i indexes donors ($i = 1, \ldots, n$),
j indexes cell types ($j = 1, \ldots, P$),
$\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{iP})$ is the vector of true cell types frequency,
$\boldsymbol{\theta_i} = (\theta_{i1}, \theta_{i2}, \ldots, \theta_{iP})$ are the parameters of the Dirichlet distribution,
$\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iP})$ is the observed count vector for donor i,
$N_i = \sum_{j=1}^{P} Y_{ij}$ is the total number of cells in the donor i.

The parameter vector $\boldsymbol{\theta}$ controls both the mean proportions and the amount of variation in those proportions across donors. The expected value of the proportion of cell type j

in donor i is given by

$$\mathbb{E}[\pi_{ij}] = \frac{\theta_{ij}}{\theta_0}, \quad \text{where } \theta_0 = \sum_{j=1}^{P} \theta_{ij} \tag{2.2}$$

The variance of the proportion of cell type j in donor i and the covariance between proportions of two cell types j and k are given by

$$\text{Var}[\pi_{ij}] = \frac{\theta_{ij}(\theta_0 - \theta_{ij})}{\theta_0^2(\theta_0 + 1)}, \qquad \text{Cov}[\pi_{ij}, \pi_{ik}] = -\frac{\theta_{ij}\theta_{ik}}{\theta_0^2(\theta_0 + 1)} \tag{2.3}$$

This variance decreases as the total concentration θ_0 increases, indicating that the Dirichlet distribution becomes more concentrated around the mean. The negative covariance arises because of the compositional constraint, where all proportions must sum to one, so an increase in one cell type's frequency implies a decrease in others.

The total concentration parameter θ_0 governs how much variability is present in the proportions, where a large θ_0 leads to low variance in the proportions across donors, meaning that the donors have similar compositions, while a small θ_0 results in more variability across donors.

To introduce heterogeneity across cell types and across experimental groups, a log-linear formulation can be applied to the Dirichlet parameters. For example, Assefa et al. (2024) used the following formulation:

$$\theta_{ij} = \gamma \cdot \exp(\beta_{0j} + X_i \beta_{1j}) \tag{2.4}$$

where $X_i = 0$ if donor i is from group 1 and $X_i = 1$ if donor i from group 2. The parameter γ controls the overall level of between-sample variability, while β_{0j} captures the baseline abundance of cell type j, and β_{1j} controls the magnitude of differential abundance between the two groups of donors.

This parameterization provides a flexible framework for generating realistic variability in cell type composition, and it serves as the basis for the simulation setup described in the next section.

2.2 Abundance estimation

This section discusses how accurately cell type abundances can be estimated under varying design parameters. Two estimation settings were established for the abundance estimation: the single-sample case and the multiple-sample case. The single-sample case focuses on understanding how many cells need to be sequenced from a single donor. It supports a prospective approach to maintain a reasonable accuracy level or provides a guideline to filter data in the retrospective approach. In the multiple-sample case, the goal is to estimate the population-level cell type proportions across donors accurately. It can support a prospective approach, such as determining how many cell counts and donors are required to achieve a desired estimation accuracy. It also applies to retrospective evaluation, where the adequacy of the existing sample size is assessed to determine

whether population-level estimates can be trusted, particularly for low-abundance cell types.

2.2.1 Simulation framework

To investigate the accuracy of cell type abundance estimation, cell count data were simulated using the Dirichlet-multinomial model described in Section 2.1. For this estimation analysis, a single-group setting was used, where all donors were assumed to come from the same group. A log-linear formulation was used for the Dirichlet parameters as defined in Equation (2.4), with the group indicator fixed as $X_i = 0$ since only one group was simulated. Under this setting, the Dirichlet parameters for each donor i were defined as

$$\theta_{ij} = \gamma \cdot \exp(\beta_{0j}) \tag{2.5}$$

where $\beta_{0j} \sim \mathcal{N}(\mu_0, \tau_0)$ represents cell type–specific abundance variation, and γ controls the amount of variability across donors. In the single-sample setting, γ was fixed at 1, as only one donor was generated per simulation replicate. The mean of β_{0j} was fixed at 0, while the τ_0 was fixed at 0.25.

To ensure consistency across simulation replicates, the values of β_{0j} and the resulting Dirichlet parameters were generated once at the beginning of each simulation setting and reused across all replicates. This reflects the assumption that donors within a group share the same underlying distribution of cell type proportions, with variation across donors arising from sampling variation under the Dirichlet-multinomial model.

Simulations were performed over a range of settings to reflect various experimental conditions:

- Number of cells to sequence: $N \in \{1000, 5000, 10000, 30000, 50000, 100000\}$
- Number of sample sizes: $n \in \{1, 5, 10, 20\}$
- Number of cell types: $P \in \{5, 10, 20, 30\}$
- Dirichlet scale: $\gamma \in \{1.5, 1, 0.25\}$ (for multiple-sample case; $n > 1$)

For each simulation setting defined by N, n, P, and γ, $K = 500$ replicates were generated. The following steps were performed in each simulation replicate:

1. For each donor i, draw the true cell type frequencies: $\boldsymbol{\pi}_i \sim \text{Dirichlet}(\boldsymbol{\theta})$.
2. Given $\boldsymbol{\pi}_i$, draw observed counts: $\mathbf{Y}_i \mid \boldsymbol{\pi_i} \sim \text{Multinomial}(N, \boldsymbol{\pi}_i)$.
3. Compute estimated frequencies $P_{ij} = \frac{Y_{ij}}{N}$.

To improve the stability of evaluation metrics, cell type filtering was applied at each simulation run to exclude extremely low-abundance cell types from analysis. Specifically, cell types were excluded in a given simulation replicate if their estimated frequency P_{ij} fell below a predefined threshold, i.e., $P_{ij} < \delta$, where $\delta \in \{0\%, 0.1\%, 1\%, 5\%\}$, to explore how the choice of the exclusion threshold affect the results.

2.2.2 Estimation accuracy and decision criteria

The accuracy of cell type abundance estimation was assessed using the relative absolute error (RAE). This approach is based on the same idea described in Thompson (1987), which is to control how far estimated frequencies are from the true values when choosing a sample size. For each donor i and cell type j, it was defined as:

$$RAE_{ij} = \frac{|P_{ij} - \pi_{ij}|}{\pi_{ij}} \tag{2.6}$$

where $P_{ij} = \frac{Y_{ij}}{N}$ is the estimated frequency based on observed counts, and π_{ij} is the true frequency drawn from the Dirichlet distribution.

In the single-sample setting ($n = 1$), the RAE values were directly used to assess estimation accuracy. In the multiple-sample setting ($n > 1$), two approaches were considered.

1. Mean RAE across donors: RAE was first computed for each donor and cell type, then averaged across donors to summarize accuracy per cell type:

$$mRAE_j = \frac{1}{n}\sum_{i=1}^{n} RAE_{ij} \tag{2.7}$$

This summarizes estimation accuracy across donors for each cell type and reflects average error at the donor level.

2. Population-level RAE: A population-level estimate of RAE was computed by averaging the estimated frequencies across donors, and then comparing this average to the expected true frequency:

$$pRAE_j = \frac{|\bar{P}_j - \bar{\pi}_j|}{\bar{\pi}_j}, \tag{2.8}$$

where $\bar{P}_j = \frac{1}{n}\sum_{i=1}^{n} P_{ij}$ is the average estimated frequency, and $\bar{\pi}_j = \mathbb{E}[\pi_j] = \frac{\theta_j}{\theta_0}$ is the expected true frequency under the Dirichlet distribution (Equation 2.2). This captures the bias in the group-level estimate.

To summarize estimation performance across simulation replicates, three decision criteria were used to define a simulation as successful based on its error values. For each criterion, the success probability Π was calculated as the proportion of simulation replicates that meet the specified threshold conditions.

1. Strict criterion. A simulation replicate was considered successful only if all cell types had estimation errors below a fixed threshold r. This was applied using RAE_j in the single-sample case, while in the multiple-sample case it is either $mRAE_j$ or $pRAE_j$. Different error thresholds $r \in \{5\%, 10\%, 15\%, 20\%\}$ were explored. This strict criterion ensures that all cell types are estimated with given accuracy, but may be

overly conservative, particularly when low-abundant cell types are present. The success probability was defined as:

$$\Pi_{\text{strict}} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}\left(\bigcap_{j=1}^{P}\{\text{Error}_{jk} \leq r\}\right) \tag{2.9}$$

where K is the number of simulation replicates, r is the error threshold, and Error_{jk} is either RAE_j, $mRAE_j$, or $pRAE_j$ in replicate k, depending on the setting.

2. Adaptive criterion. This criterion was only applied in the single-sample setting ($n = 1$), where per-replicate observed frequencies P_{ij} are available. Cell types were stratified into three abundance groups based on their observed frequencies P_{ij}: low ($P_{ij} < 1\%$), medium ($1\% \leq P_{ij} < 10\%$), and high ($P_{ij} \geq 10\%$). Each abundance group was assigned a distinct threshold to reflect its relative estimation difficulty:

$$r_j = \begin{cases} 0.5, & \text{if } P_{ij} < 1\% \text{ (low)} \\ 0.1, & \text{if } 1\% \leq P_{ij} < 10\% \text{ (medium)} \\ 0.05, & \text{if } P_{ij} \geq 10\% \text{ (high)} \end{cases}$$

A replicate was considered successful if all cell types satisfied their respective thresholds:

$$\Pi_{\text{adaptive}} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}\left(\bigcap_{j=1}^{P}\{RAE_{jk} \leq r_j\}\right) \tag{2.10}$$

where K is the number of simulation replicates, and r_j is the error threshold of cell type j.

3. Relaxed criterion. A replicate was considered successful if at least a fraction ψ of cell types satisfied the error threshold r (e.g., $\psi = 0.8$), defined as:

$$\Pi_{\text{relaxed}} = \frac{1}{K}\sum_{k=1}^{K}\mathbb{I}\left(\frac{1}{P}\sum_{j=1}^{P}\mathbb{I}\left(\text{Error}_{jk} \leq r\right) \geq \psi\right) \tag{2.11}$$

where K is the number of simulation replicates, r is the error threshold, and Error_{jk} is either RAE_j, $mRAE_j$, or $pRAE_j$ in replicate k, depending on the setting. In this study, ψ was set to 0.8.

2.2.3 Determining number of cells and sample size

The simulation results were used to assess how different values of the number of cells (N) and the number of donors (n) influence the accuracy of abundance estimation. For each simulation setting, success probabilities Π were computed under a range of decision criteria based on relative error thresholds. Thresholds were then applied to identify the simulation settings that yielded sufficiently accurate estimates. Specifically, a setting was considered acceptable if the success probability exceeded a predefined threshold (e.g., 95%). These results were used to identify minimum values of N and n required to achieve reliable estimation under different settings of the number of cell types (P), and between-sample variability (γ).

2.3 Detecting changes in cell type abundance

This section describes the simulation framework used to evaluate how the number of cells and donors affects the ability to detect changes in cell type abundance between groups. The goal is to evaluate under which design settings the differentially abundant (DA) cell types can be reliably identified with sufficient statistical power. A simulation-based approach was implemented by combining Dirichlet-multinomial modeling of count data with the voomCLR method for differential abundance testing.

2.3.1 Simulation framework

To investigate the ability to detect differential abundance between groups, data were simulated under a two-group setting using the Dirichlet-multinomial model described in Section 2.1. The group-specific Dirichlet parameters were defined using the log-linear model in Equation 2.4.

The set of cell types was divided into two subsets: $\mathcal{V}_0$, representing non-differentially abundant (non-DA) cell types for which the null hypothesis holds, and $\mathcal{V}_1$, representing truly differentially abundant (DA) cell types for which the alternative hypothesis holds. These satisfy $|\mathcal{V}_0| + |\mathcal{V}_1| = P$ and $\mathcal{V}_0 \cap \mathcal{V}_1 = \emptyset$. The proportion of truly DA cell types was set to be 20% of P, rounded up to the nearest integer.

For each DA cell type $j \in \mathcal{V}_1$, the parameter β_{1j} controls the magnitude of the group difference, while γ controls the level of variability among donors. The baseline abundance parameters β_{0j} were generated using the same setting as in the abundance estimation analysis, i.e., $\beta_{0j} \sim \mathcal{N}(0, 0.25)$. The differential abundance effects β_{1j} were sampled from $\mathcal{N}(0, 2)$, following the Lupus case study configuration in Assefa et al. (2024).

Simulations were conducted under various settings to evaluate how design parameters affect statistical power. The following parameter settings were explored:

- Number of cells per sample: $N \in \{1000, 5000, 10000, 30000, 50000, 100000\}$
- Number of donors per group: $n = n_1 = n_2 \in \{5, 10, 20\}$
- Number of cell types: $P \in \{5, 10, 20, 30\}$
- Dirichlet scale: $\gamma \in \{1.5, 1, 0.25\}$

For each simulation setting defined by N, n, P, and γ, a set of DA cell types was selected. The corresponding log-scale parameters $\boldsymbol{\beta_0}$ and $\boldsymbol{\beta_1}$, along with the group-specific Dirichlet parameters $\boldsymbol{\theta^{(1)}}$ and $\boldsymbol{\theta^{(2)}}$ were generated once and held fixed across all $K = 100$ replicates within that setting.

Within each simulation replicate, the following steps were performed:

1. For each donor i in group $g \in \{1,\ 2\}$, draw the true cell type frequencies: $\boldsymbol{\pi}_i \sim \text{Dirichlet}(\boldsymbol{\theta}^{(g)})$.

2. Given $\boldsymbol{\pi}_i$, draw the observed counts: $\mathbf{Y}_i \mid \boldsymbol{\pi}_i \sim \text{Multinomial}(N,\ \boldsymbol{\pi}_i)$.

2.3.2 Differential abundance testing

To evaluate statistical power under each simulation setting, differential abundance (DA) testing was performed using a linear modeling framework based on the *voomCLR* method (Assefa et al., 2024). This approach was designed to improve statistical inference for compositional single-cell data by combining log-ratio transformation and variance modeling. It addresses key challenges such as mean–variance dependence and limited power due to small sample sizes, while also accounting for the relative nature of cell type proportions. This method begins with a centered log-ratio (CLR) transformation of compositional count data, followed by mean-variance modeling to compute observation-level weights. The weighted CLR-transformed data are then analyzed using linear models with empirical Bayes shrinkage. Bias correction is applied prior to hypothesis testing to account for compositional effects.

CLR transformation

Observed count matrices were transformed using the centered log-ratio (CLR) transformation:

$$Z_{ij} = \log \frac{Y_{ij}}{\bar{Y}_i}, \quad \text{with } \bar{Y}_i = \left(\prod_{j=1}^{P} Y_{ij} \right)^{1/P} \tag{2.12}$$

where Y_{ij} denotes the observed count of cell type j in donor i, and P denotes the total number of cell types. This transformation allows standard statistical tools to be applied more appropriately by working with log-ratios rather than raw counts.

The mean-variance modeling and weights

CLR-transformed data often exhibit heteroscedasticity, where the variance depends on the mean. To account for this, the mean-variance trend across cell types was estimated using loess smoothing. When the number of cell types P was small, the variance was approximated analytically using a Poisson or negative binomial assumption. For each cell type j, weight was defined as $w_j = \frac{1}{\hat{\sigma}_j^2}$, where $\hat{\sigma}_j^2$ is the estimated variance of the CLR-transformed abundance. These weights were included in the linear model to reduce the influence of cell types with higher variability.

These two steps on CLR transformation and mean-variance modeling were performed using *voomCLR* function from *voomCLR* package. This function outputs the CLR-transformed values along with observation-level weights, which are then used in the linear modeling step.

Linear modeling of CLR-transformed abundances

A weighted linear model was fitted separately for each cell type j using the CLR-transformed data. The model can be written as:

$$z_{ij} = \alpha_{0j} + \alpha_{1j}\ \text{group}_i + \epsilon_{ij} \tag{2.13}$$

where z_{ij} is the CLR-transformed abundance for cell type j in donor i, group_i is a binary indicator variable equal to 0 for group 1 and 1 for group 2, α_{0j} is the intercept for cell type j, α_{1j} is the group effect representing the log-fold change in CLR-transformed abundance between the two groups, and ϵ_{ij} is the residual error term.

The model was fitted using weighted least squares, where the weights are derived from the estimated mean–variance relationship (described in the previous section). The *lmFit* function from *limma* package was used to fit the model. To improve the stability of the variance estimates, especially in settings with a small number of donors, empirical Bayes shrinkage was applied using the *eBayes* function. This approach results in moderated t-statistics, which borrow information across cell types to produce more stable and reliable hypothesis tests.

Bias correction and hypothesis testing

To address the compositional bias in effect size estimates caused by the CLR transformation, a bias correction method based on LinDA (Zhou et al., 2022) was applied using *topTableBC* function from *voomCLR* package. This approach adjusts the estimated group effects by subtracting the mode of all regression coefficients across cell types, under the assumption that most cell types are not differentially abundant.

Let $\alpha_{1j}^{\text{corrected}}$ denote the bias-corrected effect size for cell type j. Differential abundance was then assessed by testing the null hypothesis: $H_0 : \alpha_{1j}^{\text{corrected}} = 0$. p-values from the moderated t-tests were adjusted using the Benjamini–Hochberg (BH) procedure to control the false discovery rate (FDR) at 5%.

2.3.3 Power and error rate estimation

After hypothesis testing, performance was evaluated by calculating both power and false discovery rate (FDR) under each simulation setting.

For each truly differentially abundant (DA) cell type, per-cell-type power was estimated as the proportion of simulation replicates in which the cell type was detected as significant (the null hypothesis was rejected). Average power was then calculated as the mean of these per-cell-type power estimates across all truly DA cell types.

To quantify the proportion of false positives among all discoveries, the false discovery rate (FDR) was computed within each simulation replicate as

$$\text{FDR} = \frac{\text{number of non-DA cell types detected as significant}}{\text{total number of significant discoveries}} \tag{2.14}$$

The average FDR across replicates was then reported for each simulation setting.

2.3.4 Determining number of cells and sample size

The simulation framework was used to evaluate how different values of the number of cells (N) and the number of donors (n) affected the ability to detect differential abundance. For each simulation setting, power and FDR estimates were summarized across replicates. Thresholds were then applied to identify settings that achieved sufficient statistical performance, defined as achieving average power greater than 80% while controlling FDR below 5%. These results were used to identify minimum values of N and n required to meet performance criteria under different settings of the number of cell types (P), and biological variability (γ).

3 | Results

A simulation framework based on the Dirichlet-Multinomial distribution was implemented, capturing both technical sampling variability and biological heterogeneity across donors. Estimation accuracy and statistical power were evaluated across a range of design parameters. The results are presented in two sections. Section 3.1 focuses on the accuracy of cell type abundance estimation, examining both single-sample and multiple-sample cases. Section 3.2 evaluates the ability to detect differential abundance between groups, highlighting how the number of cells and the number of donors influences statistical power and false discovery rate.

3.1 Abundance estimation

3.1.1 Single-sample case

The accuracy of cell type abundance estimation was first evaluated in the single-sample setting, where the goal is to determine how many cells need to be sequenced from a single donor to achieve acceptable estimation accuracy. The success probability Π was computed under varying total cell counts (N), number of cell types (P), and error thresholds (r) for each simulation setting. Three decision criteria were considered: a strict criterion applied to all cell types, an adaptive criterion that adjusts r based on cell type abundance, and a relaxed criterion that allows a specified proportion of cell types to exceed the threshold.

Figure 3.1 shows the estimated success probability Π as a function of the total number of cells (N), across different numbers of cell types (P). The lines correspond to different decision criteria: strict (all cell types must satisfy the threshold), adaptive (thresholds vary based on abundance), and relaxed (at least 80% of cell types must satisfy the threshold). An error threshold of 5% was used for all cell types under the strict and relaxed rules. The horizontal dashed line indicates the target success probability threshold ($\lambda = 0.95$).

Overall, the success probability Π increases with larger cell counts (N). In all different settings of the number of cell types (P), the strict criterion has the lowest success probability. This reflects the conservative behavior of this criterion, where all cell types' estimated frequencies need to have a lower error than 5%. The adaptive and relaxed criteria result in notably higher success probabilities and show similar performance across

most conditions.

As the number of cell types increases, the gap between the strict and the more flexible criteria widens. For example, at $P = 20$ and $P = 30$, the success probability under the strict criterion remains very low, even at $N = 100,000$, while both adaptive and relaxed criteria approach or exceed the 95% target threshold. These trends suggest that the strict criterion becomes increasingly challenging to satisfy in settings with a high number of cell types.

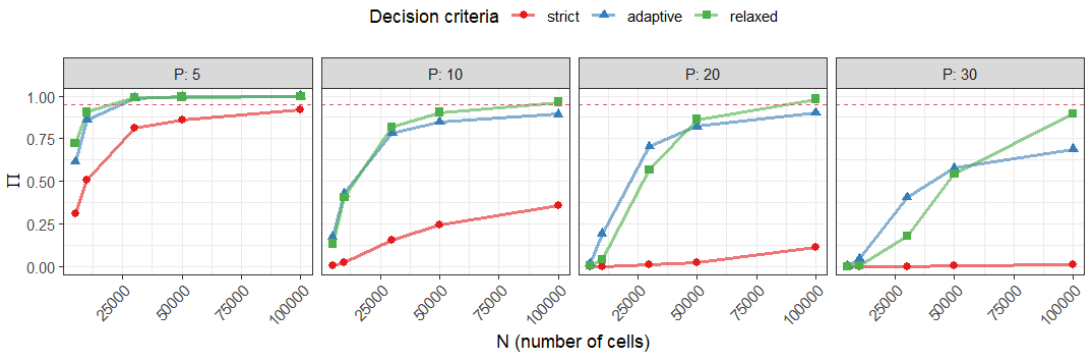

Figure 3.1: Success probability Π, plotted against the number of cells (N) in various number of cell types P and decision criteria, for single-sample case.

In practice, researchers may choose to exclude extremely low-abundance cell types from evaluation, either to avoid unstable estimates or to focus on biologically meaningful cell types. Figure 3.2 illustrates how applying different exclusion thresholds (i.e., 0.1%, 1%, 5%) affects the success probability Π and the number of cells required. Each row in the figure corresponds to a different exclusion threshold, while columns vary the number of cell types (P). The same three decision criteria were applied as in the previous analysis.

Overall, higher exclusion thresholds result in increased success probabilities. This effect is most notable in settings with larger numbers of cell types, where exclusion reduces the challenge of meeting accuracy thresholds across all evaluated cell types. In contrast, when the number of cell types is small (e.g., $P = 5$), exclusion had little to no effect across all three criteria, likely because few cell types fall below the exclusion threshold. As P increased, exclusion led to clearer improvements under the strict criterion, but had a limited effect on the adaptive and relaxed criteria. This is likely because these two criteria already account for the difficulty of estimating low-abundance cell types, either by adjusting the threshold based on abundance (adaptive) or by allowing some cell types to exceed it (relaxed). This highlights how excluding low-abundance cell types can substantially reduce the number of cells needed to meet a desired accuracy level.

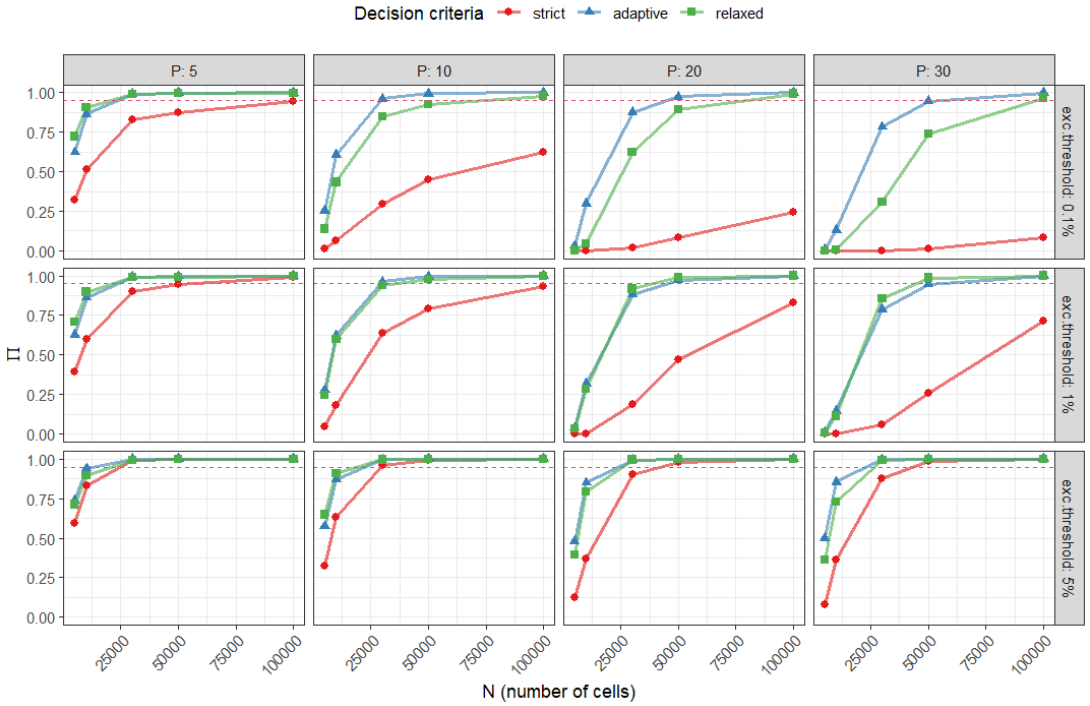

Figure 3.2: Success probability Π, plotted against the number of cells (N) in various number of cell types P and decision criteria, with some cell types exclusion threshold, for single-sample case.

Several error thresholds $r \in \{5\%, 10\%, 15\%, 20\%\}$ for the strict criterion were also explored, and the results can be seen in Appendix A.3 (Figure A.2). As expected, higher error tolerance led to higher success probabilities across all settings. The improvement was especially noticeable for larger values of P, where stricter criteria were harder to satisfy.

3.1.2 Multiple-sample case

Figure 3.3 presents the estimated success probability Π in the multiple-sample setting based on $mRAE$ approach and strict criterion, as a function of the total number of cells (N), across different numbers of cell types (P), under medium between-sample variability. Results are shown for the 5% error threshold. The horizontal dashed line indicates the target success probability threshold ($\lambda = 0.95$). Results for low and high variability are provided in Appendix A.3 (Figure A.3).

Overall, the success probability increased with higher numbers of cell counts (N) and decreased with higher numbers of cell types (P), consistent with trends observed in the single-sample case. In general, success probability also declined with higher levels of between-sample variability. Under high between-sample variability, success probabilities remained close to zero across all conditions, indicating that more heterogeneous populations require sequencing more cells to achieve the same level of accuracy.

At $P = 5$, all donor sizes under low between-sample variability level reached approximately 95% success probability at $N = 30,000$. Under medium variability, the 90% success probability was reached at around $N = 100,000$. For $P = 10$, success probability increases with N, but remains below the 95% threshold across all variability levels even at the maximum cell count. This trend continues for $P = 20$ and $P = 30$, where success probabilities remain low regardless of the number of donors or cells sequenced. An unexpected trend is observed at $P = 10$ under medium between-sample variability, where smaller donor sizes appeared to yield higher success probabilities than larger ones. This will be revisited in the Discussion.

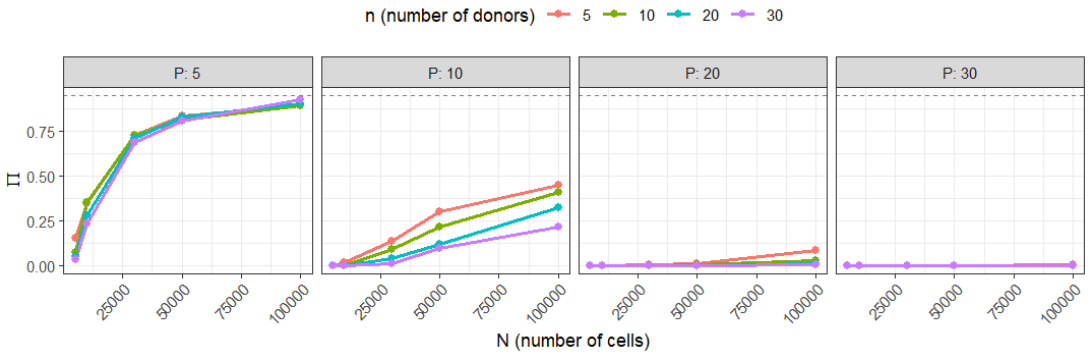

Figure 3.3: Success probability Π based on $mRAE$ approach and strict criterion, evaluated for varying numbers of donors n under medium between-sample variability, plotted against the number of cells N in various number of cell types P, for multiple-sample case.

The effect of applying different exclusion thresholds on success probability in the multiple-sample setting under medium between-sample variability is shown in Figure 3.4. Each row corresponds to a different minimum abundance threshold (0.1%, 1%, and 5%), while columns vary the number of cell types (P). Results for low and high variability are provided in Appendix A.3 (Figure A.4 and A.5).

As observed in the single-sample setting, increasing the exclusion thresholds improves success probability. This is particularly noticeable under high between-sample variability, where the success probabilities increase significantly compared to the case without exclusion. For $P = 5$, excluding cell types with a minimum observed frequency of 0.1% has little impact, likely because few cell types fall below this threshold. In contrast, the effect of exclusion becomes more pronounced for higher values of P. Notably, after exclusion, the relationship between the number of donors and success probability becomes more intuitive, where higher n values consistently yield higher success probabilities. For example, at $P = 10$, increasing $n = 10$ to $n = 30$ reach approximately the 95% success probability at around $N = 50,000$, while $n = 5$ reaches it at $N = 100,000$. These results suggest that excluding low-abundance cell types not only improves overall accuracy but also clarifies the effect of increasing sample size.

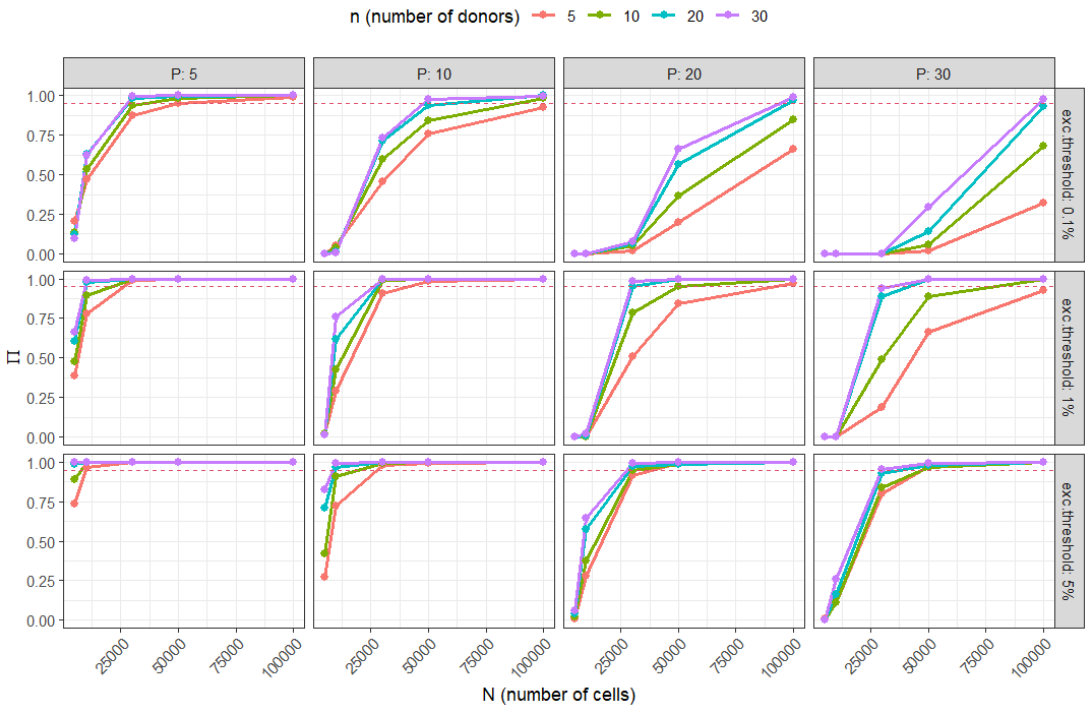

Figure 3.4: Success probability Π based on $mRAE$ approach and strict criterion, evaluated for varying number of donors n under medium between-sample variability, plotted against the number of cells N in various number of cell types P, with some cell types exclusion threshold, for multiple-sample case.

To complement the donor-level analysis, results based on the population-level RAE (pRAE) are presented in Appendix A.3 (Figure A.8). Unlike the mRAE results, the pRAE results indicate that improvements in the success probability are driven by the number of donors. Across all settings, success probability remained nearly constant across different values of N, which suggests that increasing the number of cells has little to no effect on the accuracy. This suggests that population-level accuracy is primarily governed by the number of donors.

3.2 Detecting changes in cell type abundance

This section presents simulation results evaluating the power to detect differential abundance (DA) in a comparison of two groups of donors under varying study designs. Power and false discovery rate (FDR) were estimated across combinations of number of cells (N), sample size (n), number of cell types (P), and between-sample variability.

Because the differential abundance effect sizes ($\boldsymbol{\beta_1}$) were held constant within each P, performance comparisons across values of N, n, and γ within the same P setting reflect the influence of design parameters rather than differences in effect magnitude. This allows a more controlled evaluation of how experimental design affects statistical power and FDR. The exact values of $\boldsymbol{\beta_1}$ used are provided in Appendix A.2 (Table A.3).

3.2.1 Effect of cell count and sample size on power

Figure 3.5 shows the average power to detect DA cell types across varying values of number of cells (N) and number of donors (n). Each panel corresponds to a combination of the number of cell types (P) and between-sample variability level (γ). This allows comparison within each panel to reflect the effect of design parameters. The dashed line indicates the 80% power threshold.

Within each panel, average power increases with the number of donors, while the effect of the number of cells is minimal. This suggests that increasing the number of donors is generally more effective than increasing the number of cells for improving power, especially when each donor already contains a moderate number of cells. For example, in the case of $P = 10$ and medium variability, increasing n from 5 to 20 enables power to exceed 80%, while differences across N remain small.

While power appears to vary across different P and γ settings, comparisons between panels should be interpreted with caution, as the underlying effect size configuration ($\boldsymbol{\beta_1}$) differs across P. For this reason, the results are best interpreted by comparing performance under the same P.

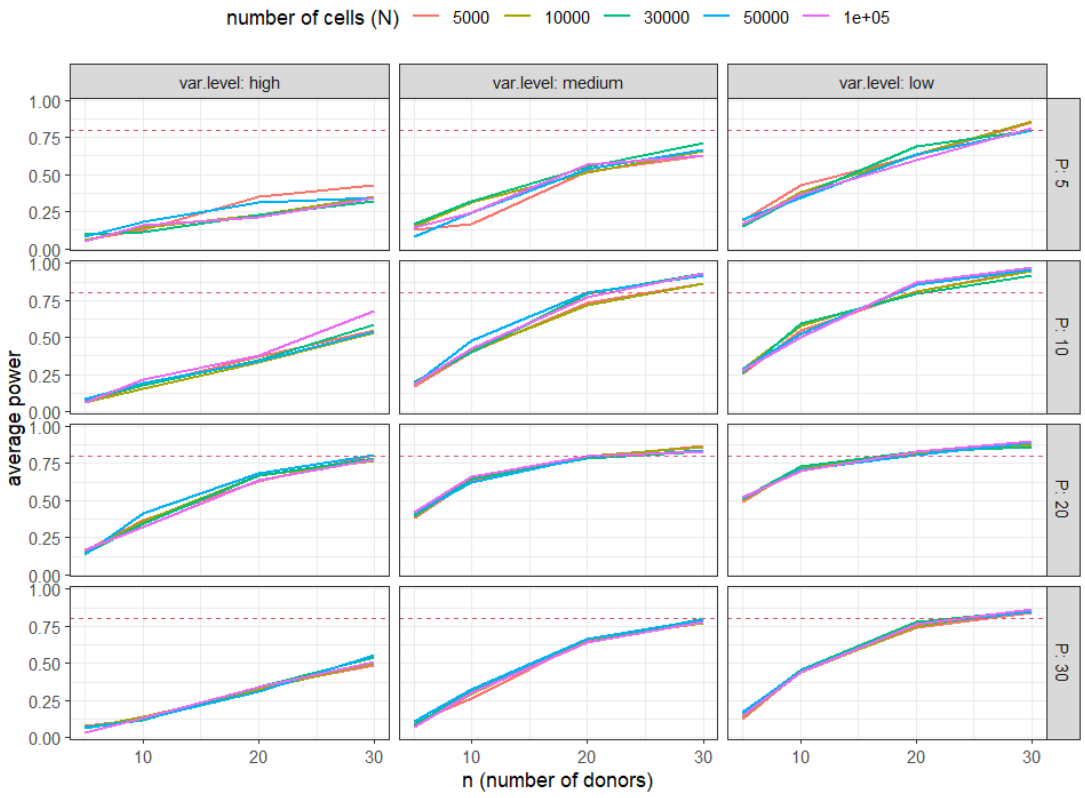

Figure 3.5: Average power to detect differentially abundant cell types across combinations of number of donors n, number of cells N, between-sample variability, and number of cell types P.

3.2.2 Effect of cell count and sample size on FDR

Figure 3.6 shows the average false discovery rate (FDR) across different values of number of cells (N) and sample size (n). Each panel corresponds to a combination of the number of cell types (P) and between-sample variability level (γ). The dashed line marks the 5% FDR threshold.

FDR was highest in settings with high between-sample variability and low number of donors, particularly for lower values of P. Increasing the number of donors tended to reduce FDR in many settings, but the patterns were not strictly decreasing and showed some fluctuation. This is likely due to variation in the number of discoveries across replicates, which can cause FDR estimates unstable, especially when few cell types are detected.

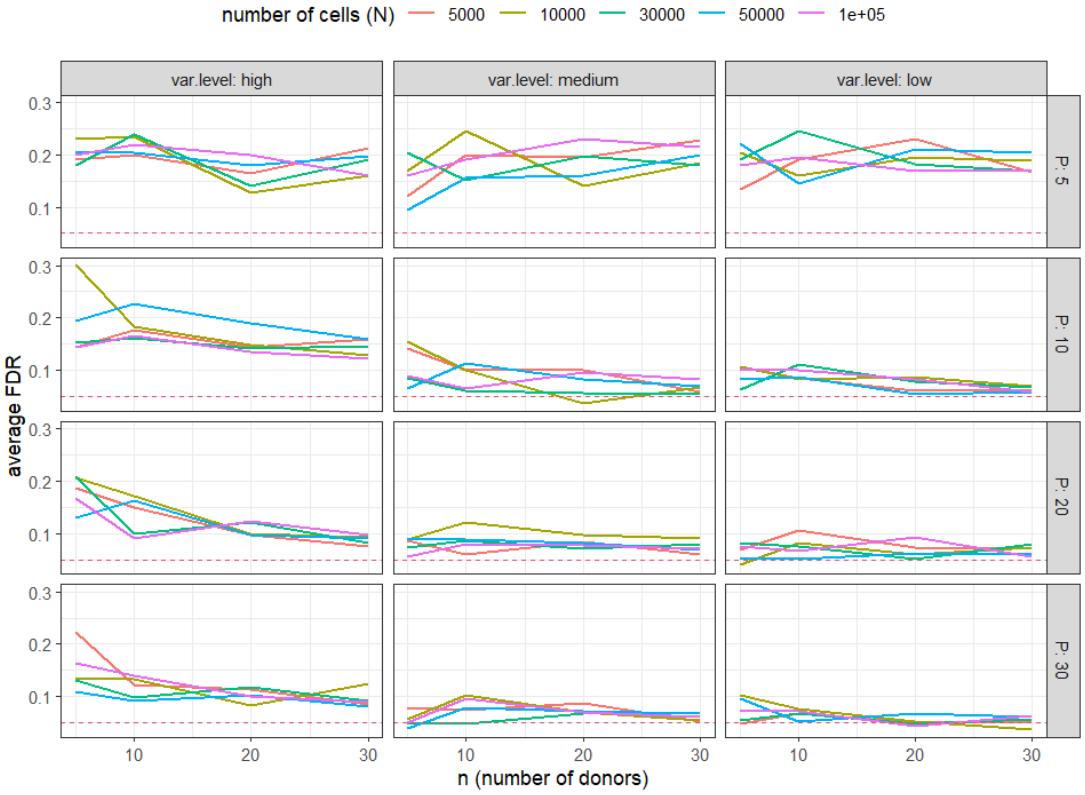

Figure 3.6: Average FDR to detect differentially abundant cell types across combinations of number of donors n, number of cells N, between-sample variability, and number of cell types P.

4 | Discussion and Conclusion

This thesis investigated how different cell counts and sample size settings affect the accuracy of cell type abundance estimation and the ability to detect changes in cell type abundance in single-cell experiments. A simulation framework based on the Dirichlet-multinomial model was employed to generate single-cell count data across a range of settings, including different numbers of cells, numbers of donors, numbers of cell types, and levels of between-sample variability.

The first research question focused on evaluating the accuracy of cell type abundance estimation across different simulation settings. From a prospective perspective, the objective is to determine the required number of cells and donors to meet a desired accuracy level, based on the number of cell types and the degree of between-sample variability. Retrospectively, given a fixed study design setting (e.g., 10 donors with 10000 cells per donor to analyze 20 cell types), researchers can evaluate the accuracy of the resulting cell type frequency estimates and potentially consider excluding unreliable cell types if accuracy is insufficient for downstream analysis.

Estimation accuracy was evaluated using the relative absolute error (RAE), which compares estimated cell type frequencies to the true frequencies. However, since it compares relative frequencies, it tends to produce larger error values for low-abundance cell types. To summarize performance across simulation replicates, the success probability Π was computed as the proportion of replicates that are successful, i.e., satisfied a predefined threshold. It provides an empirical estimate of how a study design yields accurate estimates under a given setting of cell counts, donors, number of cell types, and variability between samples.

The definition of a successful replicate was based on one of three criteria. In the strict criterion, a replicate is deemed successful only if all cell types have an error metric below the threshold. The main advantage of this approach is that it ensures accurate estimation across the entire cell type composition, which is essential when all populations are equally important or required for downstream analyses. However, it may be too stringent when many low-abundance cell types are present, as they typically yield higher RAE values due to their low frequencies (see Figure A.1 in Appendix A.3). This often results in low success probabilities and higher number of donors requirements. The adaptive criterion adjusts the error threshold for each cell type based on its observed

frequency, allowing more tolerance for low-abundance cell types. This helps mitigate the compositional bias of the RAE metric for low-abundance cell types. However, the main drawback is that the chosen thresholds can be somewhat arbitrary, and the application is limited to single-sample settings in this study's framework. For the relaxed criterion, a replicate is considered successful if a specified proportion (e.g., 80%) of cell types meet the error threshold. While this approach improves flexibility, it offers no control over which cell types fail or how large the errors are.

The use of cell type exclusion was also explored, where extremely low-abundance cell types were excluded from error evaluation. This approach improved success probabilities but introduced variability across replicates due to differences in which cell types were excluded. The number of excluded cell types and the type of excluded cells may differ in each replicate. This complicates the interpretation and limits generalizability. Moreover, excessive exclusion could also result in misleading success probabilities or underrepresentation of biologically important cell types.

In the multiple-sample setting, estimation error was summarized using the mean RAE (mRAE), computed by averaging the per-donor RAEs for each cell type. This metric attempts to reflect aggregate accuracy across donors. However, it may not fully capture how both cell count and sample size contribute to estimation performance. Simulations revealed that increasing number of cells reduced the average of mRAE, while increasing number of donors reduced the variability in mRAE but had a limited effect on its average (see Figure A.9 in Appendix A.3). This suggests that increasing the number of cells per donor plays a greater role in improving accuracy, whereas increasing the number of donors improves the stability of estimates. An unexpected trend was observed under certain conditions (e.g., $P = 10$), where a smaller number of donors n yielded higher success probabilities than a larger n. Further exploration showed that this was linked to low-abundance cell types and insufficient cell counts (see Figure A.11 in Appendix A.3). In low N settings, estimates of a low-abundance cell type had high variability. Increasing n alone did not resolve this unless N was also increased. These findings suggest that increasing number of cells may be more impactful than an increasing number of donors, particularly for low-abundance cell types.

An alternative summary metric was also explored to better capture the effects of both the number of cells and the number of donors on the accuracy, that is using population level RAE (pRAE), averaging the estimated cell type proportions across donors and computing RAE on this group-level average, then comparing it to the expected true frequencies derived from the Dirichlet distribution. However, this approach primarily reflected only the effect of the number of donors, and not the number of cells (see Figure A.10 in Appendix A.3). Thus, the choice of summary metric can influence interpretation, and further work may be needed to develop more comprehensive metrics that reflect both cell count and sample size.

The second research question evaluated how the number of cells and the number of

donors influence the ability to detect differential abundance (DA) between groups. Statistical power was the main metric used to evaluate detection performance, with false discovery rate (FDR) also examined to ensure that improved detection of true differences did not come at the cost of more false positives. Using a simulation-based framework and the voomCLR method, the number of cells, the number of donors, the number of cell types, and degree of between-sample variability were varied to explore their effects.

It was found that increasing the number of donors had a stronger impact on power than increasing the number of cells per donor. This is likely because voomCLR summarizes cell type proportions within each donor and performs group comparisons across these donor-level summaries, so each donor is treated as an independent observation in the statistical test. Increasing the number of cells per donor improves the accuracy of the estimates for each donor, but does not provide additional independent information for distinguishing between groups. In contrast, increasing the number of donors adds more biological replicates, reduces the influence of between-donor variation, and makes it easier to detect group-level differences.

Power was also strongly influenced by between-sample variability. In settings with high variability, power was consistently lower across all values of the number of cells and donors. In these settings, increasing the number of donors stabilized group-level estimates and improved power. In contrast, increasing the number of cells had little effect, because improving estimates within donors does not reduce the variability that occurs between donors. This suggests the importance of recruiting more donors rather than solely increasing the number of cells, especially when high between-sample variability is expected.

False discovery rate (FDR) was highest in settings with high between-sample variability and a low number of donors, especially when the number of cell types was small. Increasing the number of donors generally reduced the FDR, but this trend was not strictly monotonic, and some fluctuation was observed. This variability may reflect differences in the number of significant discoveries between replicates. When few cell types are detected as significant, the FDR estimate becomes unstable. Overall, voomCLR maintained reasonable FDR control in most settings, except in settings with high variability, low sample size, or few detected discoveries.

The combined results from both parts of this thesis show that different aspects of study design play important roles depending on the goal of the analysis. When the focus is on accurately estimating cell type proportions, sequencing more cells per donor is important, especially for low-abundance cell types. In contrast, increasing the number of donors is the main factor influencing power and error control in differential abundance testing. This emphasizes the importance of tailoring experimental design to the specific research question.

While this study offers insights into how cell count and sample size influence estimation

accuracy and the ability to detect differential abundance, several limitations should be considered. First, the simulation framework assumes a specific structure of variability based on the Dirichlet-multinomial model, which may not capture all forms of biological and technical variability present in real-world data. In particular, the Dirichlet-multinomial model only allows negative correlations between cell type proportions, since the total must always add up to one. This means it cannot represent situations where certain cell types tend to increase or decrease together, as might occur with biologically related cell types. Second, in the abundance estimation, RAE is sensitive to low-abundance cell types due to the compositional nature of relative frequencies. Although the adaptive criterion was introduced to account for this, it depends on threshold definitions that are somewhat arbitrary and currently limited to single-sample settings. Third, the use of mRAE to summarize performance across donors may mask potential improvements from increasing the number of donors. Fourth, the strategy of excluding low-abundance cell types improved success probability, but the variability in which cell types were excluded across replicates may limit the consistency and generalizability of these results. Lastly, in the differential abundance detection, the power depends on assumptions about effect size, so results may be higher or lower than what would be observed in actual experiments.

Future work could extend this study in several ways. For abundance estimation, new error metrics that are robust to compositional effects and low-abundance cell types should be explored, along with summary measures that better capture the effects of both cell count and sample size. For differential abundance testing, alternative DA detection methods could be evaluated, and it would be valuable to validate the simulation-based findings using real single-cell datasets.

4.1 Ethical thinking, societal relevance, and stakeholder awareness

This study uses simulation-based methods to assess how different numbers of cells and sample sizes affect estimation accuracy and the ability to detect changes in cell type abundance. Although it does not involve real human data, it still relates to several ethical and practical considerations.

The goal of this study is to help researchers make better decisions about how many cells and donors to collect in their studies. A well-planned experiment can produce more reliable results, while a poorly designed experiment can lead to misleading results or a waste of resources. By identifying how many cells and donors are needed to reach a certain level of accuracy or power, this study helps reduce unnecessary sample collection from human or animal donors. This can lower costs and reduce the use of limited donor material.

The methods and findings in this thesis are relevant for researchers, data analysts, and

professionals who are involved in designing and conducting single-cell studies. These stakeholders rely on optimal study designs to obtain meaningful insights from single-cell data while balancing accuracy, cost, and feasibility. Although this thesis does not produce a single unified rule for determining optimal design parameters, it highlights important trade-offs between cell count, sample size, number of cell types, and donor variability.

A good experimental design can also lead to clearer research findings and faster progress in understanding disease or developing treatments in areas like cancer research and drug development. This way, improving methods in study design can have an indirect benefit for society by supporting more effective research.

4.2 Conclusion

This thesis examined how the number of cells and donors impact the accuracy of cell type abundance estimation and the power to detect differential abundance in single-cell analysis. Using simulations based on the Dirichlet-multinomial model, the first part of the study demonstrated that estimation accuracy was primarily driven by the number of cells, while increasing number of donors had little effect. Meanwhile, the second part of the study showed that power increased with number of donors, whereas the effect of number of cells was limited.

This difference arises because abundance estimation is most sensitive to within-sample sampling variability, which decreases as more cells are collected per donor. In contrast, differential abundance testing involves comparing distributions across groups, which benefits more from having multiple donors to reduce uncertainty in group-level estimates. Although the initial aim was to derive a unified sample size framework, the results show that the number of cells and the number of donors serve different purposes. Therefore, this study suggests that they cannot be used interchangeably and should be tailored to the study objective.

Bibliography

A. T. Assefa, B. Verbist, and K. Van den Berge. Assessing differential cell composition in single-cell studies using voomclr. *bioRxiv*, 2024. doi: 10.1101/2024.09.12.612645.

Y. H. Choi and J. K. Kim. Dissecting cellular heterogeneity using single-cell rna sequencing. *Molecules and Cells*, 42:189–199, 2019. ISSN 02191032. doi: 10.14348/molcells.2019.2446.

A. Davis, R. Gao, and N. E. Navin. Scopit: sample size calculations for single-cell sequencing experiments. *BMC Bioinformatics*, 20, 11 2019. ISSN 14712105. doi: 10.1186/s12859-019-3167-9.

J. A. Fordyce, Z. Gompert, M. L. Forister, and C. C. Nice. A hierarchical bayesian approach to ecological count data: A flexible tool for ecologists. *PLoS ONE*, 6, 11 2011. ISSN 19326203. doi: 10.1371/journal.pone.0026785.

J. G. Harrison, W. J. Calder, V. Shastry, and C. A. Buerkle. Dirichlet-multinomial modelling outperforms alternatives for analysis of microbiome and other ecological count data. *Molecular Ecology Resources*, 20:481–497, 3 2020. ISSN 17550998. doi: 10.1111/1755-0998.13128.

D. Jovic, X. Liang, H. Zeng, L. Lin, F. Xu, and Y. Luo. Single-cell rna sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, 12, 3 2022. ISSN 2001-1326. doi: 10.1002/ctm2.694.

Y. Lei, R. Tang, J. Xu, W. Wang, B. Zhang, J. Liu, X. Yu, and S. Shi. Applications of single-cell sequencing in cancer research: progress and perspectives. *Journal of Hematology and Oncology*, 14, 12 2021. ISSN 17568722. doi: 10.1186/s13045-021-01105-2.

S. Liang, J. Willis, J. Dou, V. Mohanty, Y. Huang, E. Vilar, and K. Chen. Sensei: how many samples to tell a change in cell type abundance? *BMC Bioinformatics*, 23, 12 2022. ISSN 14712105. doi: 10.1186/s12859-021-04526-5.

D. Lähnemann, J. Köster, E. Szczurek, D. J. McCarthy, S. C. Hicks, M. D. Robinson, C. A. Vallejos, K. R. Campbell, N. Beerenwinkel, A. Mahfouz, L. Pinello, P. Skums, A. Stamatakis, C. S. O. Attolini, S. Aparicio, J. Baaijens, M. Balvert, B. de Barbanson, A. Cappuccio, G. Corleone, B. E. Dutilh, M. Florescu, V. Guryev, R. Holmer, K. Jahn, T. J. Lobo, E. M. Keizer, I. Khatri, S. M. Kielbasa, J. O. Korbel, A. M. Kozlov, T. H. Kuo, B. P. Lelieveldt, I. I. Mandoiu, J. C. Marioni, T. Marschall,

F. Mölder, A. Niknejad, L. Raczkowski, M. Reinders, J. de Ridder, A. E. Saliba, A. Somarakis, O. Stegle, F. J. Theis, H. Yang, A. Zelikovsky, A. C. McHardy, B. J. Raphael, S. P. Shah, and A. Schönhuth. Eleven grand challenges in single-cell data science. *Genome Biology*, 21, 2 2020. ISSN 1474760X. doi: 10.1186/s13059-020-1926-6.

K. W. Ng and G.-L. Tian. *Dirichlet and Related Distributions: Theory, Methods and Applications*. John Wiley & Sons, Hoboken, NJ, 2011. ISBN 9780470749783.

T. P. Quinn, I. Erb, M. F. Richardson, and T. M. Crowley. Understanding sequencing data as compositions: An outlook and review. *Bioinformatics*, 34:2870–2878, 8 2018. ISSN 14602059. doi: 10.1093/bioinformatics/bty175.

S. K. Thompson. Sample size for estimating multinomial proportions. *The American Statistician*, 41:42–46, 2 1987.

H. Zhou, K. He, J. Chen, and X. Zhang. Linda: linear models for differential abundance analysis of microbiome compositional data. *Genome Biology*, 23(1):95, 2022. doi: 10.1186/s13059-022-02655-5.

A | Appendix

A.1 R code

The R codes used in this thesis are available in `https://github.com/luluasmils/master-thesis`

A.2 Simulation parameters

Table A.1: Generated Dirichlet parameters (θ_j) and expected value of the cell type proportion (μ_j) for each cell type j in different settings of number of cell types P, for single-sample case.

P	cell type j	θ_j	μ_j
5	ppl_1	0.9189	0.1344
5	ppl_2	1.4220	0.2080
5	ppl_3	0.8869	0.1297
5	ppl_4	1.4701	0.2150
5	ppl_5	2.1396	0.3129
10	ppl_1	1.2750	0.1342
10	ppl_2	0.4594	0.0484
10	ppl_3	1.1911	0.1254
10	ppl_4	1.8128	0.1909
10	ppl_5	0.3279	0.0345
10	ppl_6	0.9467	0.0997
10	ppl_7	1.3818	0.1455
10	ppl_8	0.9100	0.0958
10	ppl_9	0.5682	0.0598
10	ppl_10	0.6242	0.0657
20	ppl_1	0.7610	0.0350
20	ppl_2	1.1393	0.0524
20	ppl_3	0.5492	0.0253
20	ppl_4	1.1460	0.0527
20	ppl_5	0.8234	0.0379
20	ppl_6	1.5463	0.0711
20	ppl_7	0.7038	0.0324
20	ppl_8	0.7827	0.0360
20	ppl_9	0.4427	0.0204
20	ppl_10	1.3928	0.0641
20	ppl_11	1.6347	0.0752
20	ppl_12	1.0263	0.0472
20	ppl_13	0.6944	0.0319
20	ppl_14	1.4111	0.0649
20	ppl_15	2.5950	0.1194
20	ppl_16	0.6232	0.0287
20	ppl_17	0.8260	0.0380
20	ppl_18	1.0851	0.0499
20	ppl_19	1.4380	0.0662
20	ppl_20	1.1131	0.0512
30	ppl_1	1.2414	0.0402
30	ppl_2	0.6436	0.0208
30	ppl_3	1.1943	0.0387
30	ppl_4	0.9125	0.0295
30	ppl_5	1.4369	0.0465
30	ppl_6	1.6782	0.0543
30	ppl_7	1.2205	0.0395
30	ppl_8	0.5527	0.0179
30	ppl_9	0.9337	0.0302
30	ppl_10	0.8708	0.0282
30	ppl_11	0.4214	0.0136
30	ppl_12	1.2382	0.0401
30	ppl_13	0.4782	0.0155
30	ppl_14	1.3067	0.0423
30	ppl_15	0.5497	0.0178
30	ppl_16	0.6119	0.0198
30	ppl_17	1.3692	0.0443
30	ppl_18	1.4062	0.0455
30	ppl_19	0.7717	0.0250
30	ppl_20	1.8013	0.0583
30	ppl_21	0.7312	0.0237
30	ppl_22	0.3413	0.0110
30	ppl_23	2.4647	0.0798
30	ppl_24	1.4841	0.0480
30	ppl_25	0.7376	0.0239
30	ppl_26	0.6750	0.0219
30	ppl_27	0.7013	0.0227
30	ppl_28	1.5525	0.0503
30	ppl_29	0.7937	0.0257
30	ppl_30	0.7713	0.0250

Table A.2: Expected value of the cell type proportion (μ_j) for each cell type j in different settings of number of cell types P, for multiple-sample case.

P	cell type j	μ_j
5	ppl_1	0.0908
5	ppl_2	0.1604
5	ppl_3	0.1574
5	ppl_4	0.4356
5	ppl_5	0.1557
10	ppl_1	0.1093
10	ppl_2	0.1633
10	ppl_3	0.0694
10	ppl_4	0.0871
10	ppl_5	0.0702
10	ppl_6	0.1049
10	ppl_7	0.1271
10	ppl_8	0.0349
10	ppl_9	0.0629
10	ppl_10	0.1709
20	ppl_1	0.0424
20	ppl_2	0.0448
20	ppl_3	0.0598
20	ppl_4	0.0631
20	ppl_5	0.0471
20	ppl_6	0.0751
20	ppl_7	0.0139
20	ppl_8	0.0263
20	ppl_9	0.1132
20	ppl_10	0.0308
20	ppl_11	0.0275
20	ppl_12	0.0642
20	ppl_13	0.0221
20	ppl_14	0.0627
20	ppl_15	0.0816
20	ppl_16	0.0219
20	ppl_17	0.0486
20	ppl_18	0.0633

P	cellType	μ_j
20	ppl_19	0.0635
20	ppl_20	0.0278
30	ppl_1	0.0379
30	ppl_2	0.0170
30	ppl_3	0.0207
30	ppl_4	0.0629
30	ppl_5	0.0288
30	ppl_6	0.0438
30	ppl_7	0.0338
30	ppl_8	0.0337
30	ppl_9	0.0203
30	ppl_10	0.0129
30	ppl_11	0.0290
30	ppl_12	0.0534
30	ppl_13	0.0286
30	ppl_14	0.0209
30	ppl_15	0.0332
30	ppl_16	0.0240
30	ppl_17	0.0130
30	ppl_18	0.0551
30	ppl_19	0.0654
30	ppl_20	0.0154
30	ppl_21	0.0287
30	ppl_22	0.0242
30	ppl_23	0.0489
30	ppl_24	0.0294
30	ppl_25	0.0624
30	ppl_26	0.0254
30	ppl_27	0.0219
30	ppl_28	0.0630
30	ppl_29	0.0231
30	ppl_30	0.0234

Table A.3: Generated β_{1j} for each cell type j in different settings of number of cell types P.

P	cell type j	β_{1j}
5	ppl_3	-0.5616
10	ppl_4	-0.8416
10	ppl_6	-0.7588
20	ppl_20	-2.1626
20	ppl_5	-1.0160
20	ppl_6	-0.5051
20	ppl_1	1.3393
30	ppl_12	-1.1714
30	ppl_14	-0.6949
30	ppl_27	-0.5342
30	ppl_10	0.4860
30	ppl_15	0.8643
30	ppl_28	1.0047

A.3 Additional figures

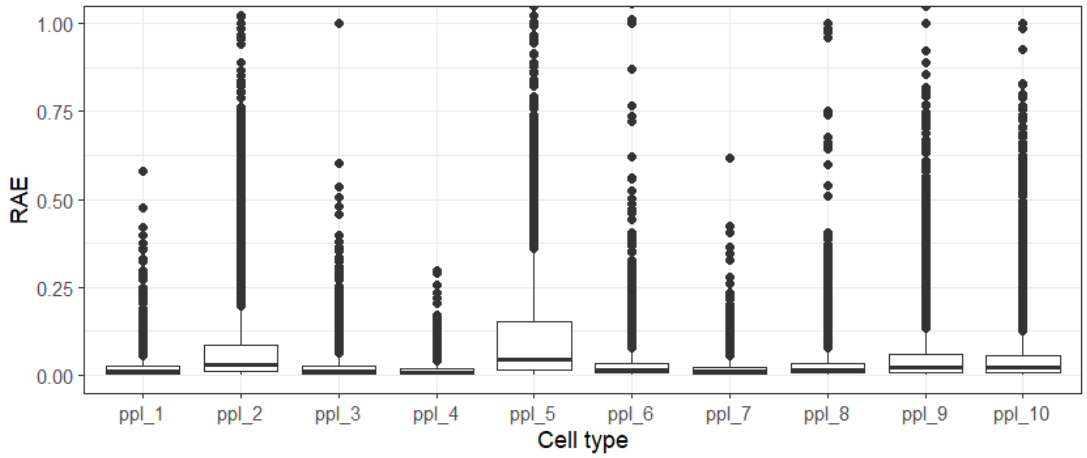

Figure A.1: RAE distribution for single-sample setting with $P = 10$. Y-axis limited to RAE ≤ 1.

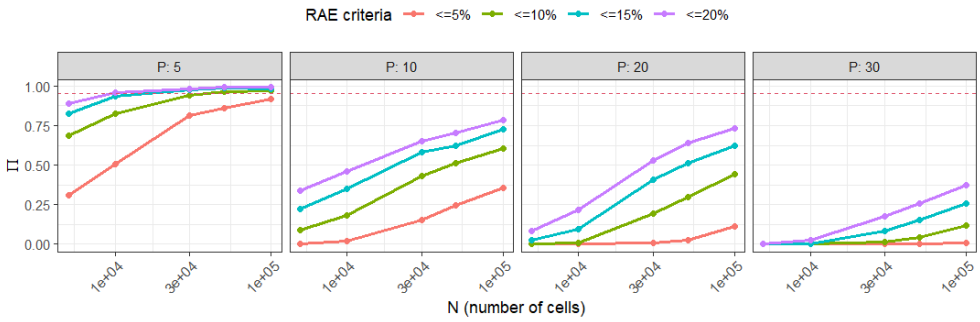

Figure A.2: Success probability Π of different error thresholds r (RAE criteria), plotted against the number of cells N in various number of cell types P, for single-sample case.

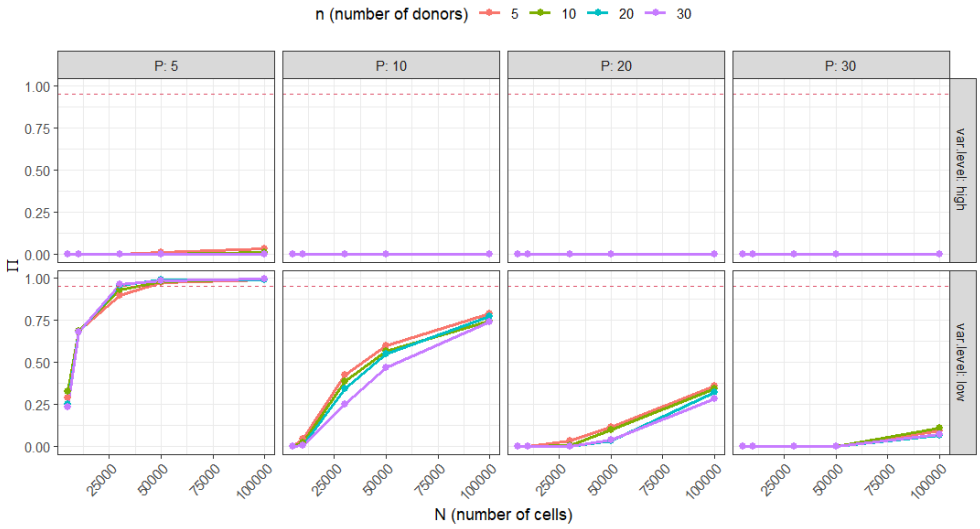

Figure A.3: Success probability Π based on $mRAE$ approach and strict criterion, evaluated for varying number of donors n with high and low between-sample variability, plotted against the number of cells N in various number of cell types P, for multiple-sample case.

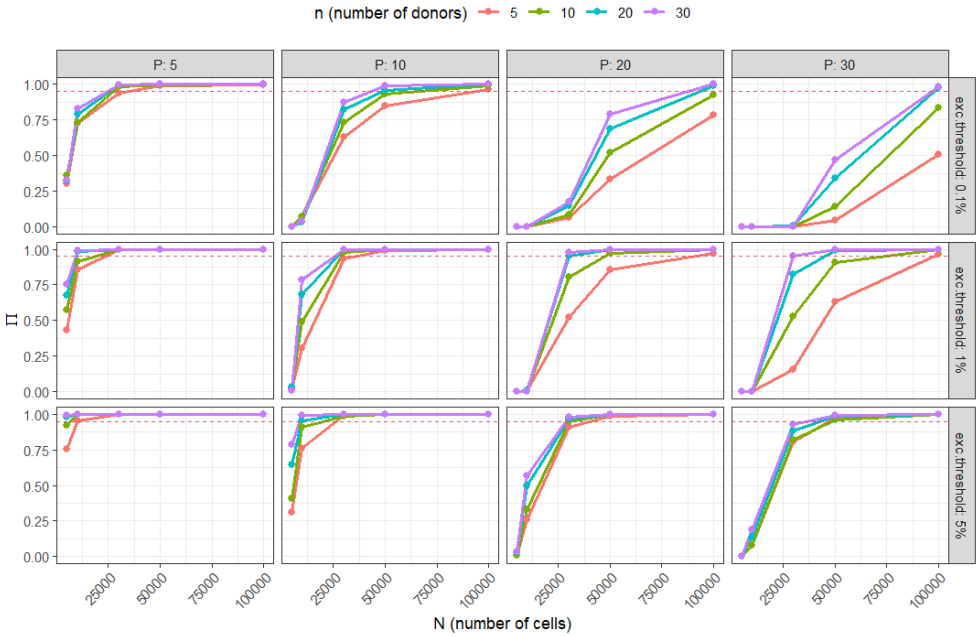

Figure A.4: Success probability Π based on $mRAE$ approach and strict criterion, evaluated for varying number of donors n with low between-sample variability, plotted against the number of cells N in various number of cell types P, with some cell types exclusion threshold, for multiple-sample case.

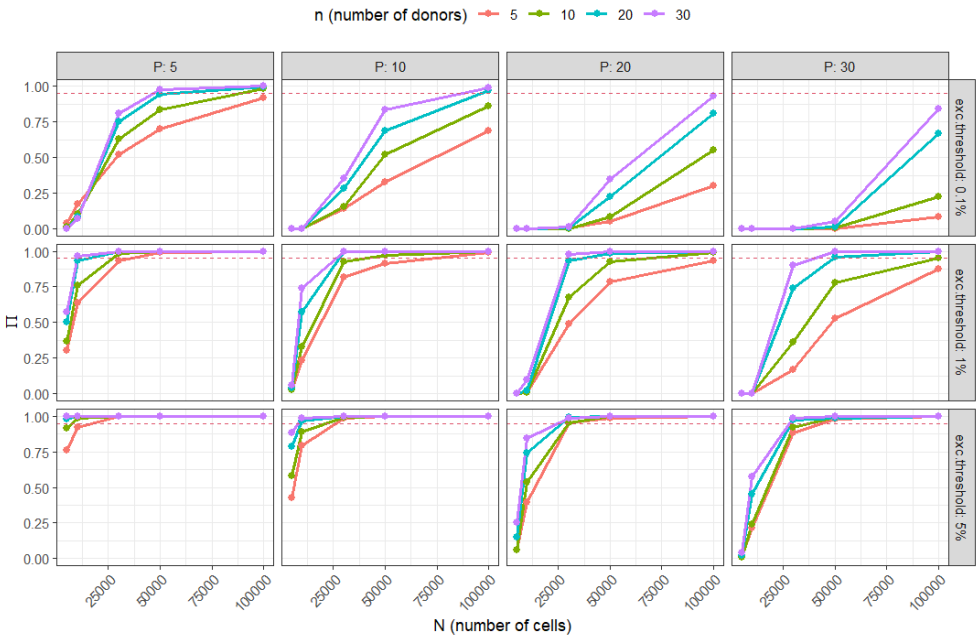

Figure A.5: Success probability Π based on $mRAE$ approach and strict criterion, evaluated for varying number of donors n with high between-sample variability, plotted against the number of cells N in various number of cell types P, with some cell types exclusion threshold, for multiple-sample case.

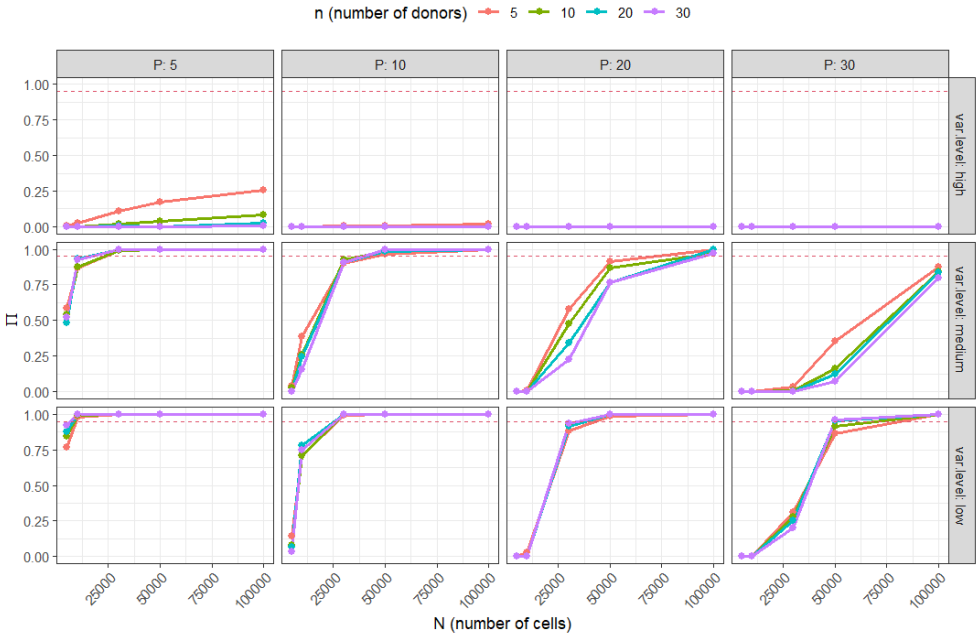

Figure A.6: Success probability Π based on *mRAE* approach and relaxed criterion, evaluated for varying number of donors n, plotted against the number of cells N in various number of cell types P and between-sample variability levels, for multiple-sample case.

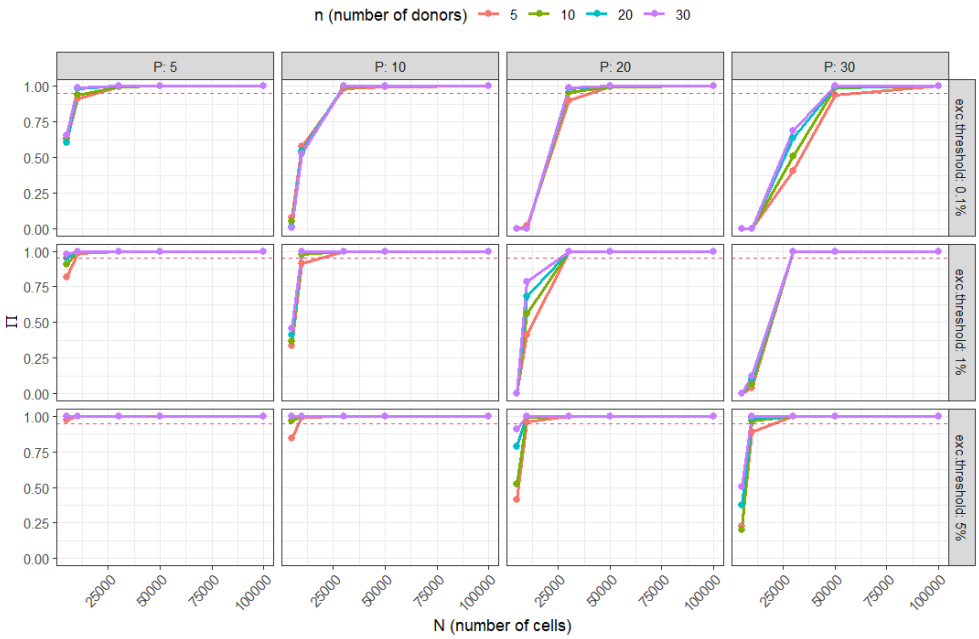

Figure A.7: Success probability Π based on *mRAE* approach and relaxed criterion, evaluated for varying number of donors n under medium between-sample variability, plotted against the number of cells N in various number of cell types P, with some cell types exclusion threshold, for multiple-sample case.

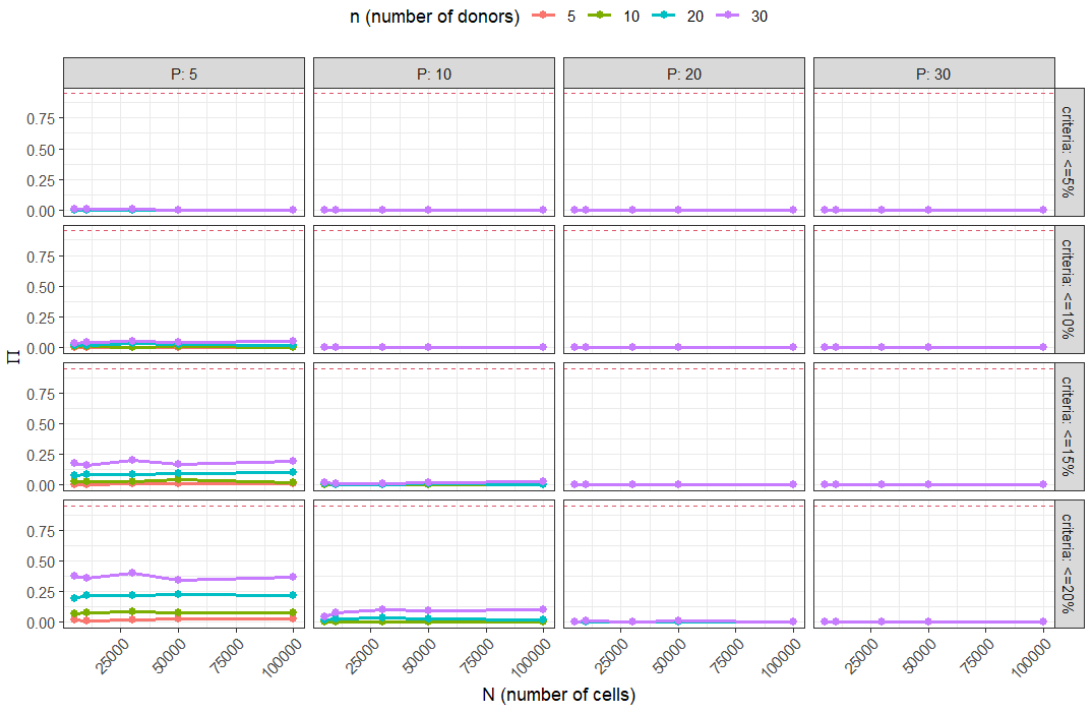

Figure A.8: Success probability Π based on *pRAE* approach and strict criterion, evaluated for varying numbers of donors n under medium between-sample variability, plotted against the number of cells N in various number of cell types P, with some error threshold, for multiple-sample case.

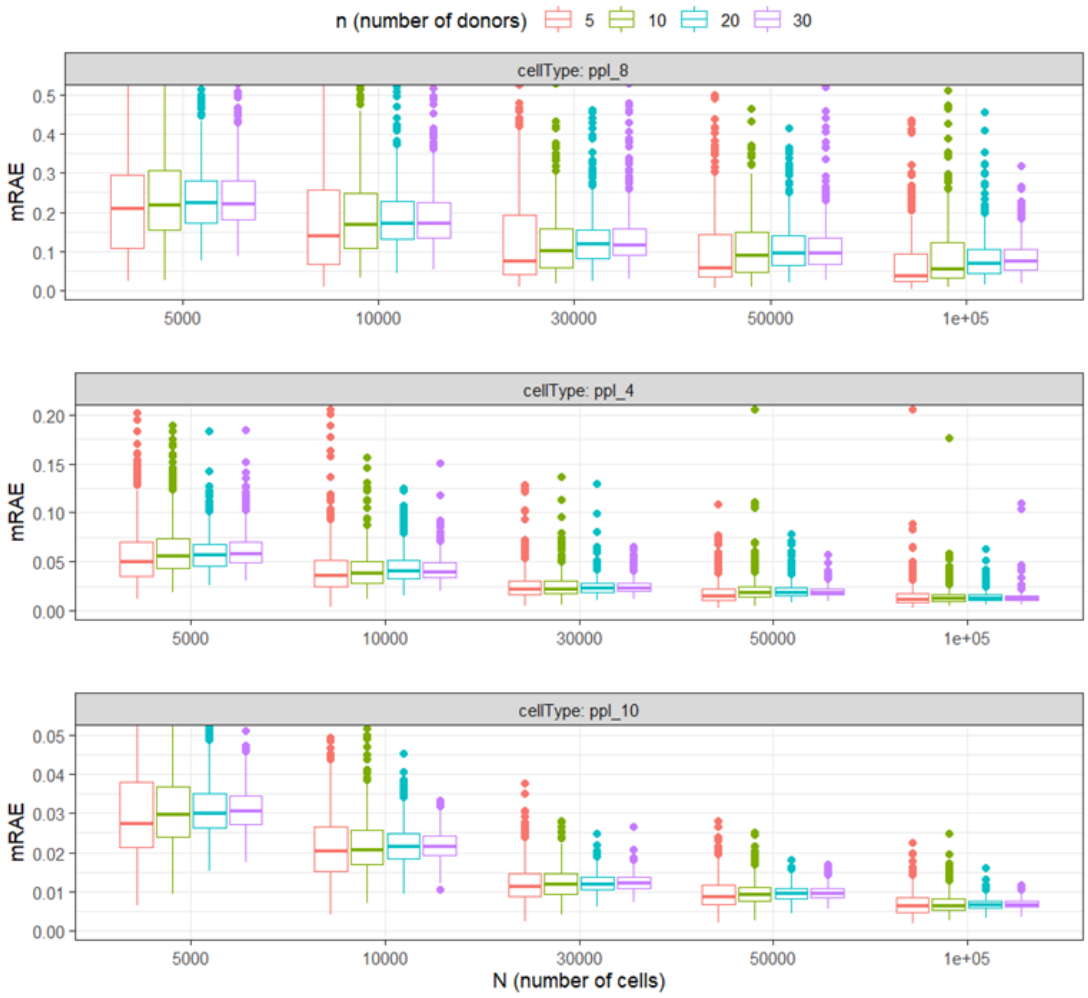

Figure A.9: Distribution of mRAE across different numbers of cells (N) and donors (n), under medium between-sample variability and $P = 10$, for three representative cell types: ppl_8 (lowest frequency), ppl_4 (medium frequency), and ppl_10 (highest frequency).

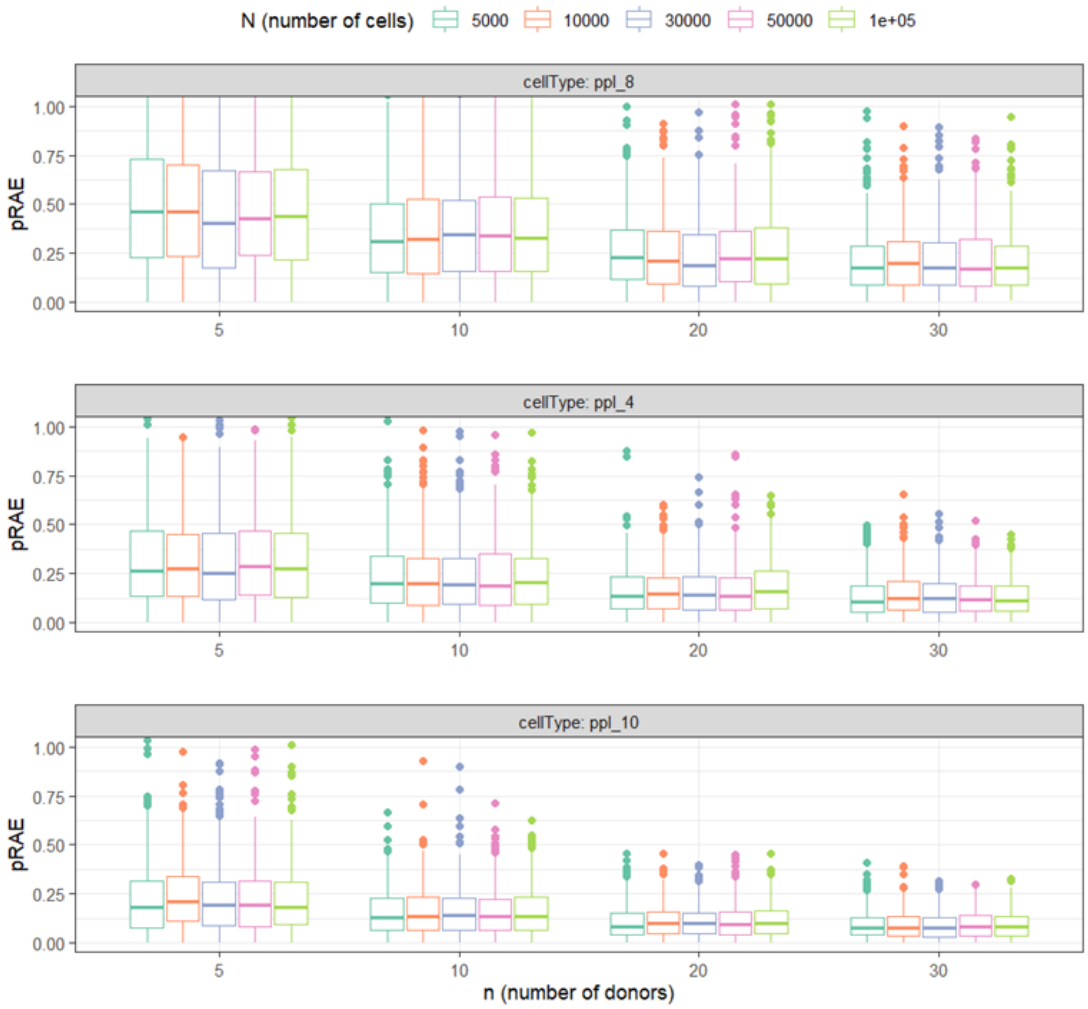

Figure A.10: Distribution of pRAE across different numbers of cells (N) and donors (n), under medium between-sample variability and $P = 10$, for three representative cell types: ppl_8 (lowest frequency), ppl_4 (medium frequency), and ppl_10 (highest frequency).

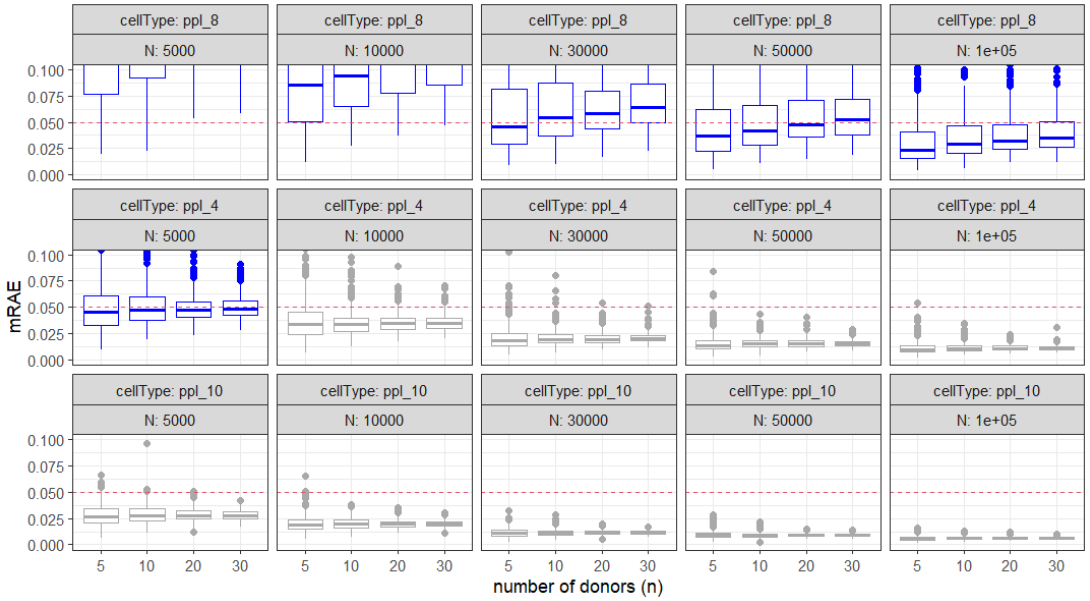

Figure A.11: Distribution of mRAE across different numbers of cells (N) and donors (n), under medium between-sample variability and $P = 10$, for three representative cell types: ppl_8 (lowest frequency), ppl_4 (medium frequency), and ppl_10 (highest frequency). Blue boxes indicate settings where lower n yields a higher proportion of acceptable mRAE values (below 5%), while grey boxes indicate the opposite. The red dashed line marks the 5% error threshold.