

# Design and power analysis for multi-sample single cell genomics experiments

**Katharina Schmid**

Helmholtz Zentrum München <https://orcid.org/0000-0001-7082-1099>

**Cristiana Cruceanu**

Max Planck Institute of Psychiatry

**Anika Böttcher**

Helmholtz Zentrum München

**Heiko Lickert**

Helmholtz Zentrum München <https://orcid.org/0000-0002-4597-8825>

**Elisabeth Binder**

Max-Planck Institute of Psychiatry <https://orcid.org/0000-0001-7088-6618>

**Fabian Theis**

Helmholtz Zentrum München <https://orcid.org/0000-0002-2419-1943>

**Matthias Heinig** (✉ [matthias.heinig@helmholtz-muenchen.de](mailto:matthias.heinig@helmholtz-muenchen.de))

Helmholtz Zentrum München <https://orcid.org/0000-0002-5612-1720>

---

## Article

**Keywords:** experimental design, power analysis, single cell genomics, eQTL, differential gene expression

**Posted Date:** March 25th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-331370/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Nature Communications on November 16th, 2021. See the published version at <https://doi.org/10.1038/s41467-021-26779-7>.

# Design and power analysis for multi-sample single cell genomics experiments

Katharina T. Schmid, Cristiana Cruceanu, Anika Böttcher, Heiko Lickert, Elisabeth B. Binder, Fabian J. Theis, Matthias Heinig\*

\* Corresponding author

Katharina T. Schmid  
Institute of Computational Biology, Helmholtz Zentrum München  
Department of Informatics, Technical University Munich  
katharina.schmid@helmholtz-muenchen.de

Cristiana Cruceanu  
Department of Translational Research, Max Planck Institute for Psychiatry, Munich  
cristiana\_cruceanu@psych.mpg.de

Anika Böttcher  
Institute of Diabetes and Regeneration Research, Helmholtz Diabetes Center, Helmholtz Zentrum München  
anika.boettcher@helmholtz-muenchen.de

Heiko Lickert  
Institute of Diabetes and Regeneration Research, Helmholtz Diabetes Center, Helmholtz Zentrum München  
Technical University Munich; German Center for Diabetes Research (DZD)  
heiko.lickert@helmholtz-muenchen.de

Elisabeth B. Binder  
Department of Translational Research, Max Planck Institute for Psychiatry, Munich  
Department of Psychiatry and Behavioral Sciences, Emory University School of Medicine, USA  
binder@psych.mpg.de

Fabian J. Theis  
Institute of Computational Biology, Helmholtz Zentrum München  
Department of Mathematics, Technical University Munich  
fabian.theis@helmholtz-muenchen.de

Matthias Heinig  
Institute of Computational Biology, Helmholtz Zentrum München  
Department of Informatics, Technical University Munich  
matthias.heinig@helmholtz-muenchen.de

## Abstract

Single cell RNA-seq revolutionizes transcriptomics by providing cell type resolution for interindividual differential gene expression and expression quantitative trait loci analyses. However, efficient power analysis methods accounting for the characteristics of single cell data and interindividual comparison are missing.

Here we present a statistical framework for design and power analysis of multi-sample single cell genomics experiments. The model relates sample size, number of cells per individual and sequencing depth to the power of detecting differentially expressed genes within cell types. It enables fast systematic comparison of alternative experimental designs and optimization for a limited budget. We evaluated data driven priors for a range of applications and single cell platforms. In many settings, shallow sequencing of high numbers of cells leads to higher overall power than deep sequencing of fewer cells.

The model including priors is implemented as an R package *scPower* and is accessible as a web tool.

Keywords: experimental design, power analysis, single cell genomics, eQTL, differential gene expression

## Introduction

From the early days of microarrays, one of the main goals of transcriptomic profiling has been to identify changes of gene expression levels (differentially expressed genes; DEGs) between sets of samples, e.g. patients and healthy controls<sup>1-5</sup>. Combining this with genetic information, the analysis of differential expression between genotypes identifies expression quantitative trait loci (eQTLs)<sup>6-9</sup>, which help to understand the molecular basis of trait associated genetic variants<sup>10</sup>.

Single cell RNA-sequencing (scRNA-seq)<sup>11-15</sup> allows for differential gene expression and eQTL analysis on the level of individual cell types. Single cell differential gene expression analysis typically seeks to identify genes whose expression levels are markedly different between different cell types<sup>16-18</sup>. In contrast, multi-sample experiments aim at the identification of DEGs between sets of samples within the same cell type. These sets can be defined by different experimental conditions or genotypes and are each measured at the single cell level. Multi-sample experiments have been identified as one of the grand challenges for single cell data analysis<sup>19</sup>.

Power analysis is an important step to design the most powerful experiments given certain assumptions about the expected effect sizes and constraints on the available resources. Researchers need to decide on parameters such as the sample size, the number of cells per sample and the number of reads. The power is always tightly linked with the statistical testing procedure. Several methods have been established based on the theory of linear regression models<sup>20</sup> and the control of the false discovery rate<sup>21-24</sup> for microarray studies. For bulk RNA-seq studies, power analysis methods based on the theory of negative binomial count regression<sup>25,26</sup>, other parametric models<sup>27-29</sup>, or simulations<sup>30,31</sup> have been proposed and benchmarked<sup>32</sup>.

In principle methods for bulk RNA-seq power analysis could also be applied to compute power or minimally required sample sizes for given effect sizes for single cell experiments, however they fail to take into account specific characteristics of single cell data. In scRNA-seq experiments individual cells are typically not sequenced to saturation, leading to sparse count matrices, where only highly expressed genes are detected with counts greater than zero. In addition, the overall number of transcripts as well as the number of transcripts of individual genes can be highly cell type specific<sup>33</sup>.

Recently, individual aspects of single-cell specific experimental design were addressed (for a detailed comparison see **Supplementary Table S1**). The comparison of sensitivity and accuracy of different technology platforms<sup>34-36</sup> has led to recommendations of sequencing depth. The minimal number of cells sequenced to observe a rare cell type with a given probability can be modelled with a negative binomial distribution<sup>37,38</sup> or multinomial distribution<sup>39</sup>.

While these recommendations also help with the design of multi-sample experiments, there are additional parameters that need to be taken into account here such as the sample size and the effect sizes. For single cell differential expression analyses, several simulation-based methods have been published recently<sup>40-43</sup>. However, only one simulation tool addresses multi-sample comparison with cells from different individuals in each group<sup>44</sup>. An associated benchmarking study, where this simulation tool was applied, demonstrated that the 'pseudobulk' approach in combination with classical differential gene expression methods such as edgeR<sup>45</sup> and limma-voom<sup>46</sup> is outperforming single cell specific methods and mixed models in multi sample DE analysis<sup>41</sup>. The pseudobulk approach approximates cell type specific gene expression levels for each individual as the sum of UMI counts over all cells of

the cell type and was also successfully applied in different single cell eQTL studies<sup>47–49</sup>. Power of the pseudobulk approach can be assessed by simulations<sup>44</sup>.

A big disadvantage of simulation-based studies is that they are not suitable to evaluate a large number of experimental designs in order to identify optimal parameter combinations due to extensive runtimes. Even power analysis for single experiments with large sample size can be very memory and runtime intensive. Additionally, handling more complex designs is not readily accessible for simulation based methods, but could be achieved with analytical power analysis methods.

A first analytic exploration of different experimental designs for single cell eQTL studies showed the importance of the optimizing parameters for a restricted budget<sup>50</sup>, as shallow sequencing of more samples can increase the effective sample size. However, the analysis is based only on the effective sample size and does not quantify the overall power of the experiment, which is crucial for the design of successful experiments. In addition, it does not provide a generalizable tool that can be applied on other data sets. Furthermore, it is not applicable for DEG analysis.

Here we provide a unified resource for experimental design considerations of interindividual comparisons focusing on the power to detect DEGs and eQTLs while also addressing the power to detect rare cell types. Our model was specifically developed for the pseudobulk approach, including a quantification of the probability to detect cell type specific gene expression in scRNA-seq data. In contrast to simulation-based methods, our analytic method enables the optimization of experimental parameters via efficient enumeration of many design options. An accurate power estimation with our model is ensured by the selection of appropriate priors for the cell type specific expression distributions and for the effect size distributions. We derive data driven priors on expression distributions from single cell atlases of three different tissues, two from published studies<sup>51,52</sup> and a newly generated data set. We combine these with cell type specific data driven priors for effect sizes based on DEGs and eQTL genes from bulk RNA-seq experiments on cells sorted by fluorescence activated cell sorting (FACS). The four DE studies<sup>53–56</sup> and one eQTL study<sup>57</sup> cover different biological applications to diseases such as asthma and cancer and to ageing. Our model provides the basis for rationally designing well powered experiments, increasing the number of true biological findings and reducing the number of false negatives.

Efficient calculation including a selection of different possible priors is easily accessible for the user, as we provide our model and parameters as an open source R package *scPower* on github <https://github.com/heiniglab/scPower>. This also comprises a shiny app with a user-friendly graphical user interface, which is additionally available as a web server at <http://scpower.helmholtz-muenchen.de/>. All code to reproduce the figures of the paper is provided in the package vignette.

## Results

### Power analysis framework for scRNA-seq experimental design

Our power analysis framework targets multi-sample genomics experiments analysed with the pseudobulk approach. Each analysis starts with a count matrix of genes times cells. We first focus on counts of unique molecular identifiers (UMI) and later generalize to read counts. Cells are annotated to an individual and a cell type or state. For the sake of simplicity, we will only consider discrete cell types / states. These can be derived by clustering and analysis of marker

genes, potentially considering multiple levels of resolution<sup>58</sup> or using the metacell approach<sup>59</sup>. Individuals are annotated with different experimental covariates, such as disease status. For the discussion we will consider a simple two group comparison, but more complicated experimental designs, which can be analysed with generalized linear models, can be treated analogously.

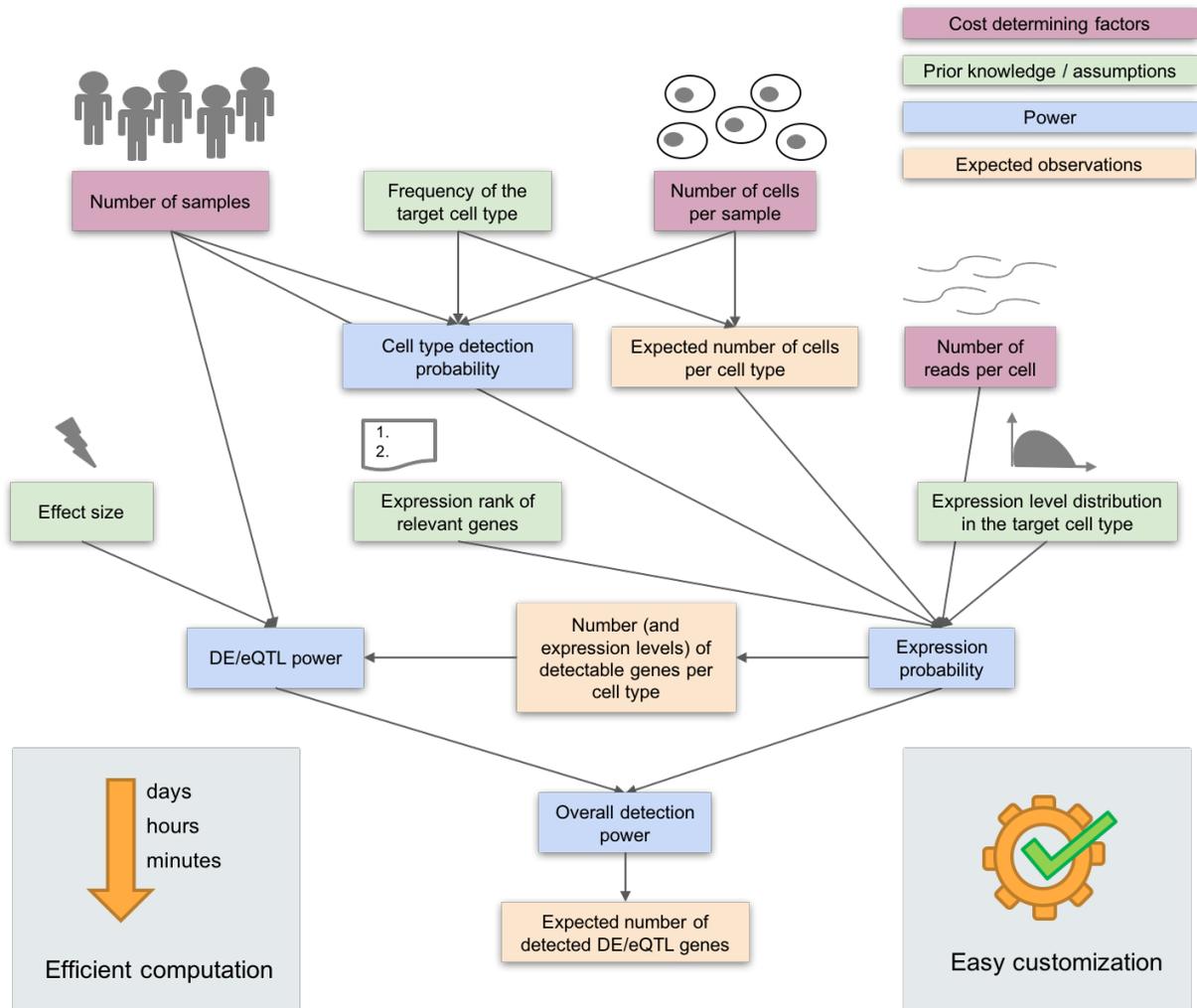
To determine cell type specific differential expression between samples, gene expression estimates for each sample and each cell type are approximated as the sum of UMI counts over all cells of the cell type<sup>47–49</sup>. This pseudobulk approach has been identified as one of the currently best performing approaches for multi sample DE analysis in a recent benchmarking study<sup>41</sup>. It is important to keep in mind that the pseudobulk approach on single cell data is distinct from traditional bulk RNA-seq. In pseudobulk the ability to detect the expression of a gene depends on the number of cells of the cell type and on the expression level of the specific gene. Therefore, we model the general detection power dependent on the number of cells per sample  $n_c$  which is related to the number of cells per cell type. Two additional experimental parameters determine the power in our model and also the cost of a scRNA-seq experiment in general: the number of samples  $n_s$  and the number of reads sequenced per cell  $r$ . In order to compute the power of the experiment, we either need to make explicit assumptions or use prior knowledge about unknown experimental parameters, such as the assumed effect sizes and gene expression levels of eQTLs and DEGs. **Figure 1** shows the dependency between different modifiable experimental parameters, unknown quantities and the expected outcomes. Our framework models the *overall detection power*  $P$  of an experiment across all considered DEGs / eQTL genes  $D$  conditional on the experimental design parameters and the priors. The overall detection power is defined as the mean gene level detection power  $P_i$  conditional on gene specific priors of gene  $i$ :

$$P = \frac{1}{|D|} \sum_{i \in D} P_i$$

In order to identify a gene as an DEG /eQTL gene, it must be both expressed and exceeding the significance cutoff. Therefore, we further decompose the gene level detection power  $P_i$  into the *expression probability*  $P(i \in E)$ , which quantifies the probability to detect gene  $i$  in the set of expressed genes  $E$ , and the *DE/eQTL power*, which we denote as the probability  $P(i \in S)$  that gene  $i$  is in the set of significant differentially expressed genes  $S$ . This quantifies the power (probability to reject  $H_0$  when  $H_1$  is true) of the statistical test for gene  $i$  and depends on the assumed effect sizes  $\theta_p$ , which can be derived from prior data. In addition, both the expression probability and the DE/eQTL power depend on the mean  $\mu$  and dispersion  $\phi$  of expression levels of gene  $i$ . In our model  $\mu$  and  $\phi$  are determined by the experimental design parameters  $(n_c, r)$  and the parameters of cell type specific expression distributions  $\theta_e$ . Conditioning the gene level detection power  $P_i$  on these priors and experimental design parameters, allows for decomposing  $P_i$  as the product of the *expression probability* and the *DE/eQTL power*:

$$\begin{aligned} P_i &= P(i \in E \wedge i \in S | n_s, n_c, r, \theta_e, \theta_p) = \\ &= P(i \in E | n_s, \mu(n_c, r, \theta_e), \phi(n_c, r, \theta_e)) \cdot \\ &\quad P(i \in S | n_s, \mu(n_c, r, \theta_e), \phi(n_c, r, \theta_e), \theta_p) \end{aligned} \tag{1}$$

In the following sections the models for the gene level expression probability and the DE/eQTL power are specified.



**Figure 1: Dependence of experimental design parameters.** The cost determining factors (purple: number of samples, number of cells per sample and number of reads per cell) are related to detection power (blue) and expected number of observations (orange). In addition, power and expected observations also depend on prior knowledge or assumptions (green). Our model enables fast power calculation, independent of the chosen experimental parameters, and easy adaptation to different use-cases through reference priors.

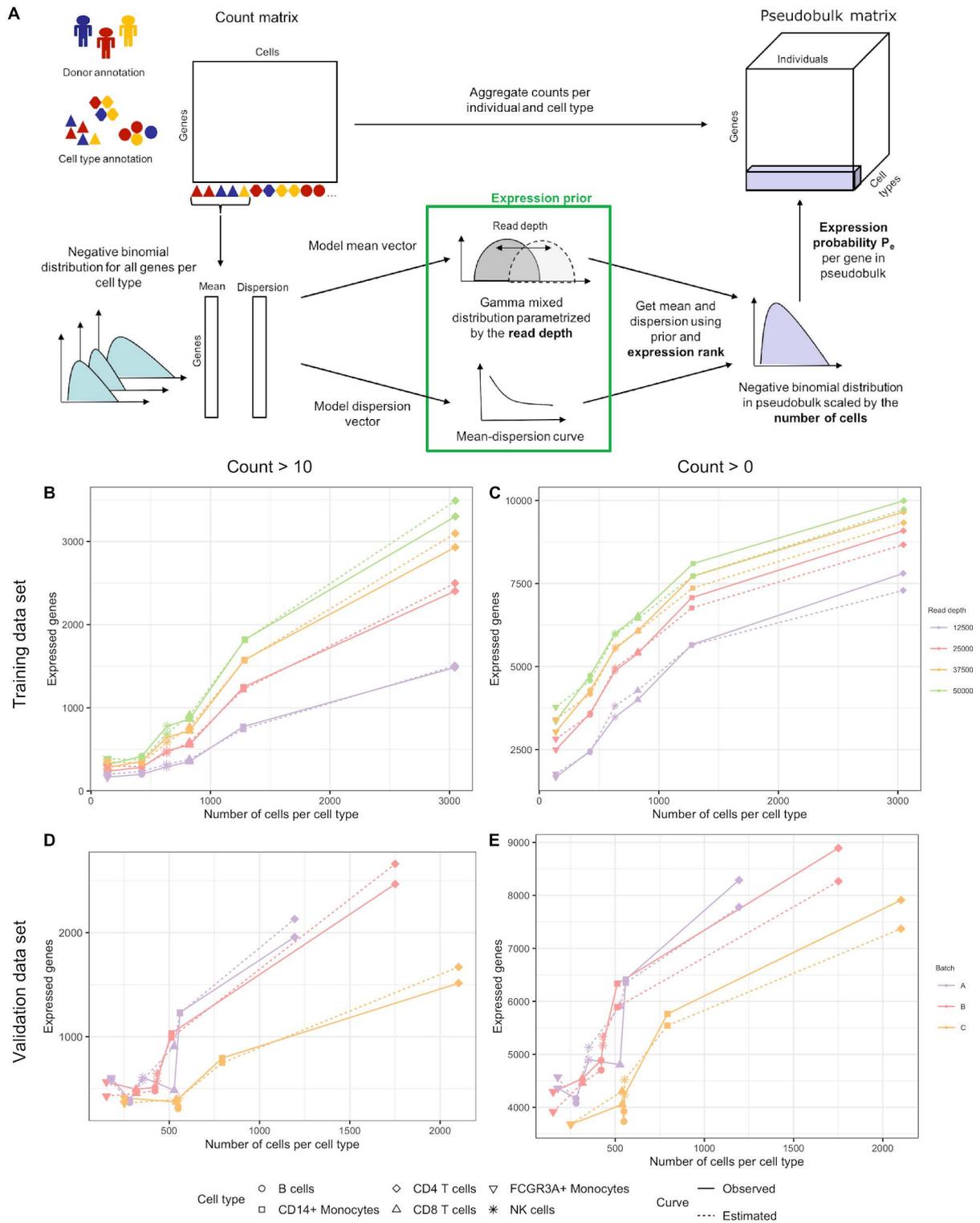
### scPower accurately models the number of detectable genes per cell type

In scRNA-seq experiments typically only highly expressed genes are detected with counts greater than zero<sup>34–36</sup>, therefore a major challenge solved by our single cell specific power framework is the modeling of gene expression levels and the probabilities of detecting expressed genes. We define a gene as detected within a cell type if the pseudobulk count (sum of UMI counts per gene per cell type per individual) is larger than a chosen threshold in a certain fraction of individuals. Therefore, the number of detectable genes per cell type depends on the number of reads sequenced per cell  $r$ , see also<sup>35</sup>, the number of cells of the cell type per individual  $n_{c,s}$  and the number of individuals  $n_s$ . In the following sections, we specify a model parameterized by these three variables.

To model cell type specific expression probabilities a prior is required that describes the general expression distribution in the target cell types. This prior captures cell type specific

differences in overall RNA content and distribution<sup>33</sup>. Exemplarily, we apply our expression probability model to different blood cell types. Therefore, we fit the expression priors per cell type using a scRNA-seq data set of PBMCs from 14 healthy individuals measured with 10X Genomics (**Supplementary Figure S1, Table S2**), in the following called the training data set, and evaluate it on a second independent PBMC data set<sup>47</sup>, the validation data set.

For the cell type specific expression prior, we approximate the single cell count distribution in each cell type with a small number of hyperparameters dependent on the read depth (**Figure 2A**). We model UMI counts per gene  $i$  in a particular cell type  $c$  as independent and identically distributed according to a negative binomial distribution with a mean  $\mu_{i,c}$  and dispersion parameter  $\phi_{i,c}$ . The distribution of means  $\mu_{i,c}$  across all genes is further modeled as a mixture distribution with a zero component and two left censored gamma distributions to cover highly expressed genes (see methods and **Supplementary Figure S2**). Subsampling the read depth of our data shows that the parameters of the mixture distribution are linearly dependent on the average UMI counts (**Supplementary Figure S3**). The dispersion parameter  $\phi_{i,c}$  is modelled dependent on the mean  $\mu_{i,c}$ , using the approach of DEseq<sup>60</sup>. As the initial experimental parameter for our model is the read depth and not directly the UMI counts, average UMI counts are related to the average number of reads mapped confidently to the transcriptome, which are in turn related to the number of reads sequenced per cell (**Supplementary Figure S4**). Taken together, we now have a model of per cell read counts across all genes parameterized by the number of reads sequenced, which was trained on cell type specific expression data. The set of parameters describing the gamma mixture distribution dependent on the UMI counts, the mean-dispersion curves and the read depth-UMI curves is called expression prior in the following. It is required for a correct modelling of the count distribution in unseen data and so the expression probabilities.



**Figure 2: Expression probability model parameterized by UMI counts per cell.** Panel A visualizes how the count distribution in the pseudobulk is modelled using negative binomial distributions with mean and dispersion parameters derived from cell type specific expression priors. Panel B and C show the observed number of expressed genes (solid line) and the number of expressed genes expected under our model (dashed line) on the y-axis and the number of cells per cell type (cell type indicated by the point symbol) on the x-axis for one batch of the trainings PBMC data set (**Supplementary Table S2**). The data is subsampled to different read depths (indicated by the color). Panel D and E show expressed genes for the

three batches of an independent validation PBMC data set <sup>47</sup>. A gene is called expressed with count > 10 (B,D) or count > 0 (C,E) in more than 50% of the individuals.

We quantify the expression distribution of a particular gene in a particular cell type and individual based on its prior expression strength. This prior is represented by the expression rank of the gene compared to all other genes. We determine its mean expression level as the quantile corresponding to this expression rank in the single cell expression prior distribution. This quantity is dependent on the read depth. Next, we derive the pseudobulk count distribution from the single cell expression distributions. This pseudobulk count distribution is again a negative binomial distribution. Its mean and dispersion are scaled by the number of cells per individual and cell type. Based on this gene specific pseudobulk distribution, we compute the probability that the observed counts are greater than the user defined minimal count threshold in at least a given number of individuals. Summing up these gene expression probabilities allows for modelling the expected number of expressed genes (see methods section for detailed formulas).

**Figure 2** shows the number of expressed genes across cell types dependent on the number of cells of the cell type for varying read depth based on subsampling of our data. The observed numbers (solid lines) are closely matched by the expectation under our model (dashed lines) for genes with counts greater ten (**Figure 2B**) and with counts greater than zero (**Figure 2C**). While **Figure 2B-C** shows the results only for one batch of the PBMC data set, the fits of all batches can be found in **Supplementary Figure S5,S6**.

To validate our model, we applied it on a second PBMC data set <sup>47</sup> that was not used during parameter estimation for the expression priors (**Figure 2D-E**). This validation data set was measured at a smaller read depth of 25,000 reads per cell and for a different sample size (batch A and B with 4 individuals and batch C with 8 individuals). The observed numbers are closely matched by the expectation under our model, which demonstrates that it can generalize well between data sets and different experimental parameters. Taken together, we now have a general model for the expected number of expressed genes, which is parameterized by the number of cells per cell type and the number of reads per cell. Of note, gene expression distributions are cell type specific and the model parameters have to be fitted from suitable (pilot) experiments, such as the human cell atlas project <sup>61</sup>.

## scPower models the power to detect differentially expressed genes and expression quantitative trait genes

Building on our expression probability model, we are able to calculate the DE/eQTL power of the expressed genes using analytical power analysis methods. Negative binomial regression is a powerful approach for DEG analysis of both RNA-seq and scRNA-seq <sup>18,62-64</sup>. Well tested tools such as DESeq <sup>5,60</sup> or EdgeR <sup>45</sup> combined with the pseudobulk approach outperform other methods of multi-sample differential expression analysis <sup>41</sup>. We verified that all our training data sets could be modelled by negative binomial distributions after pseudobulk transformation, which showed no evidence of zero inflation (**Supplementary Table S4**). Therefore, it is valid to apply analytical methods for the power analysis of negative binomial regression models <sup>65</sup>. These power calculations are exact when analysing the data with models based on negative binomial regression, but DEG analyses with other tools might lead to different results.

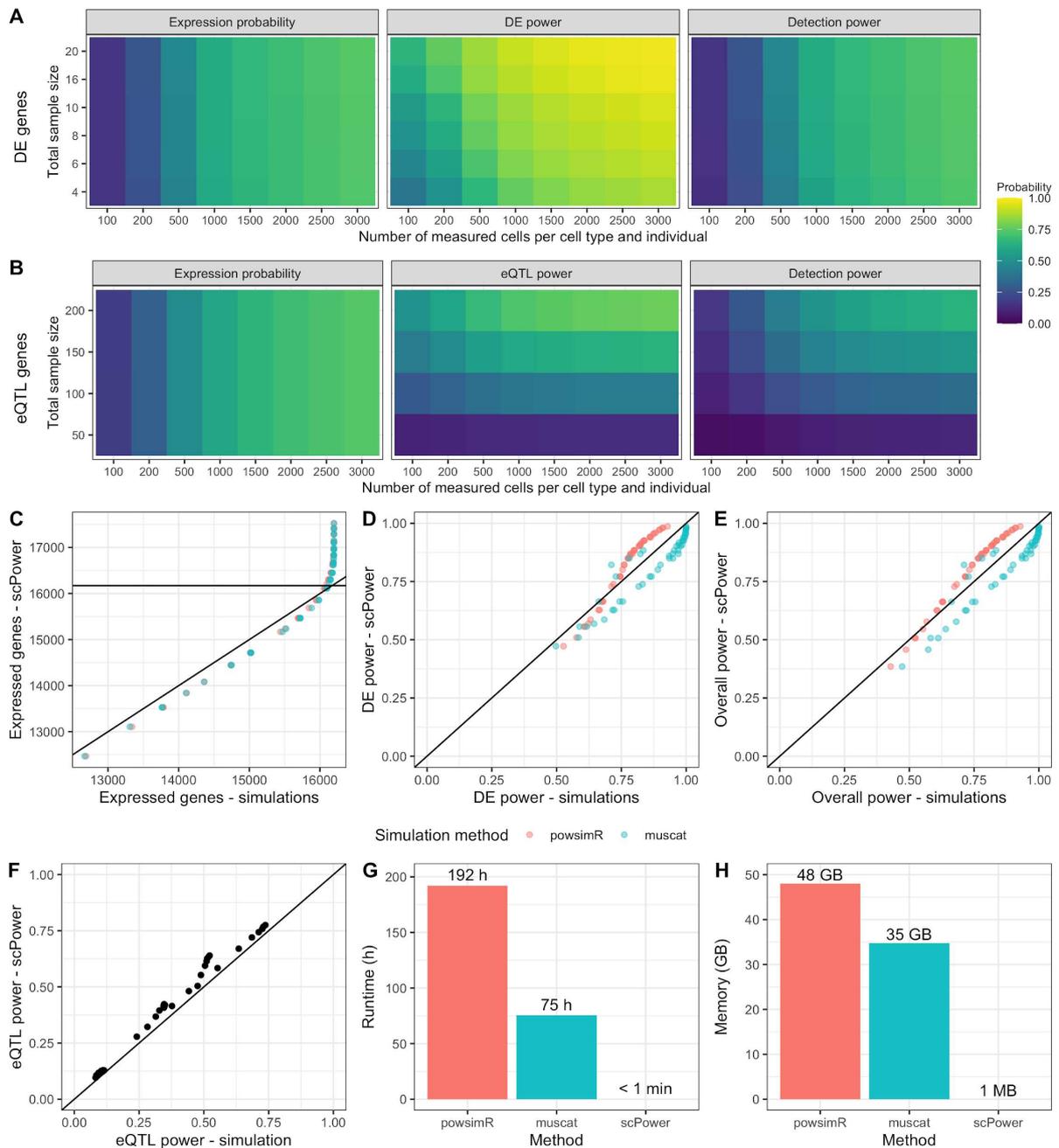
Power to detect an effect of a given effect size (log fold change) depends on the sample size, on the mean expression level and on the significance threshold. The large number of parallel

tests performed in a DEG analysis requires an adjustment of the significance level to avoid large numbers of false positive results. We provide two methods in our framework for that, either controlling the family-wise error rate (FWER) using the Bonferroni method<sup>66</sup> or the false discovery rate (FDR)<sup>22</sup>. In the following analysis, we used the FDR adjustment for DE power and FWER adjustment for eQTL power, as proposed by the GTEx Consortium<sup>67</sup> for a genome-wide cis eQTL analysis.

To obtain a range of typical effect sizes and mean expression distributions in specific immune cell types, we analysed several DEG studies based on FACS sorted bulk RNA-seq<sup>53,54</sup>. Combined with our gene expression model, we can calculate the overall detection power of DE genes averaging over the gene specific expression probability times the power to detect the gene as a DE gene based on fold changes from prior DEG studies.

**Figure 3A** shows that the overall detection power reaches up to 74% for fold changes from a study comparing CLL subtypes iCLL vs mCLL<sup>53</sup> as effect size priors. This power is reached with the experimental parameters of 3,000 cells per cell type and individual, a total balanced sample size of 20, i.e. 10 individuals per group, and FDR adjusted p-values. The original study had a sample size of 6 individuals and detected 84 DEGs with median absolute log fold change of 2.8. For this parameter combination and prior, the DE power would reach even 98% for all DE genes of the study, however, only 74% are likely to be expressed. Overall, the DE power increases with higher number of measured cells and higher sample sizes, while the expression probability is mainly influenced by the number of measured cells.

The influence of the sample size is not so pronounced in this example due to the small sample size of the reference study. Potential weaker effect sizes that would be identified with larger sample sizes could not be considered in the priors, which leads to a low required sample size for the power estimation. For other reference studies the impact of a higher sample size on the power is more visible (**Supplementary Figure S11**). Similar detection ranges are found for the comparison of other CLL subtypes in the same study, while the detection power in a study of systemic sclerosis vs control were much lower with values up to 30% (**Supplementary Figure S11**). Smaller absolute fold changes in this study decrease the DE power and therefore also the overall detection power. The effect of using the FWER adjustment also for the DE power can be seen in **Supplementary Figure S12**.



**Figure 3: Expression probability, DE/eQTL power and overall detection power and their validation in simulation studies.** Power estimation using data driven priors for A. DE genes and B. eQTL genes dependent on the total sample size and the number of measured cells per cell type. The detection power is the product of the expression probability and the power to detect the genes as DE or eQTL genes, respectively. The fold change for DEGs and the  $R^2$  for eQTL genes were taken from published studies, together with the expression rank of the genes. For A, the Blueprint CLL study with comparison iCLL vs mCLL was used, for B, the Blueprint T cell study. The expression profile and expression probabilities in a single cell experiment with a specific number of samples and measured cells was estimated using our expression prior, setting the definition for expressed to  $> 10$  counts in more than 50% of the individuals. Multiple testing correction was performed by using FDR adjusted p-values for DE power and FWER adjusted p-values for eQTL power. The probabilities calculated in A were verified by the simulation-based methods powsimR and muscat (C-E) with each point representing one

parameter combination. Also the eQTL power of B could be replicated with a self-implemented simulation (F). Runtime (G) and memory requirements (H) were drastically higher in the simulations than for our tool scPower during the evaluations of C-E, showing the clear strength of our analytic model.

A different statistical model is used for the power calculation of expression quantitative trait loci (eQTLs). Due to the very large number of statistical tests (~millions), simple linear models are usually applied to transformed read count data <sup>46,68</sup>, as they can be computed very efficiently. Therefore, power calculations here are based on linear models <sup>20</sup>. For large mean values, the power is estimated analytically, for small mean values, this approximation can be imprecise and instead simulations are used that take the discrete nature of scRNA-seq into account. This introduces a dependency between the eQTL power and the expression mean and thus eQTL power is considered conditional on the mean. The mean threshold below which simulations are used, was defined by comparison of simulated and analytic power (**Supplementary Figure S10**).

Overall detection power for eQTL genes (**Figure 3B**) shows a stronger effect of the sample size, which increases the eQTL power. In the depicted use case, the applied priors originate from an eQTL study of T cells from the Blueprint consortium <sup>57</sup>, which had a sample size of 192 and identified 5,132 eQTL genes with a median absolute beta value for the strongest associated SNP of 0.89. Increasing the number of cells per individual increases both the expression probability and the eQTL power by shifting the expression mean of the pseudo bulk counts to higher values. Notably, increasing the number of measured cells per individual and increasing the sample size both result in higher costs. A maximal detection power of 64% was found for a sample size of 200 individuals and 3,000 measured cells per cell type and individual. The Blueprint eQTL data set contains also eQTLs from monocytes where the same trend was shown and a maximal detection power of 65% reached (**Supplementary Figure S11**).

## scPower estimations are supported by simulations

The accuracy of scPower was evaluated by benchmarking against different simulation-based methods (**Figure 3C-F**). In general, simulation-based methods generate and analyze example count matrices. Therefore, they are always approximations and need to be repeated multiple times for accurate results, while we transformatively enable the design of experiments with our analytic model that requires order of magnitude less runtime and memory (**Figure 3G-H**). For single cell DE experiments, we compared our model with powsimR <sup>40</sup> and muscat <sup>41</sup>, which show both well matching power estimations compared to our tool scPower. powsimR is a widely used simulation-based method that is however not designed for multi-sample single cell comparison, i.e. it is only possible to compare groups of single cell measurements and not groups of multiple samples with single cell measurement for each sample. Adaptations of powsimR were necessary to make it comparable to scPower (see Methods for a detailed description of changes). In contrast, muscat is a recent method that incorporates already the pseudobulk approach for multi-sample comparison and can be used directly. Both simulation methods can be combined with different DE analysis methods for the downstream analysis of the simulated counts. We evaluated them in combination with different common DE methods, such as DESeq2 <sup>5</sup>, edgeR <sup>45</sup> and limma <sup>46</sup>.

The simulation based power estimates from the adapted version of powsimR as well as from muscat matched the estimates from scPower very well (**Figure 3C-E**). We compared the

expected number of expressed genes, the DE power of these expressed genes and the overall power for all simulated genes. Running simulations with different DE methods showed that the observed power also depends on analysis choices such as the DE method with scPower estimates being most accurate when using edgeR (**Supplementary Figure S13**). Furthermore, also powsimR and muscat differ slightly, caused by different modelling assumptions. The overall trends when comparing different experimental designs are in good agreement between scPower and all analysis methods applied to the simulated reads. This is true for both FWER adjustment and FDR adjustment as multiple testing correction. Also a comparison over a wide range of experimental design parameters between edgeR applied to simulated data from powsimR and scPower confirms the agreement of power estimates (**Figure 3C-E** and **Supplementary Figure S14**).

Currently, there exists no power estimation method for single cell eQTL, which explicitly accounts for specific effect size priors. Therefore, we compared the analytical eQTL power with our own simulation method, which is also used for power estimation of genes with small mean values. The simulation method applies our underlying expression probability model of scPower for assigning a mean value to each gene. This part of the model is the same for eQTL and DE power and was already shown to be accurate compared to powsimR and muscat. Therefore, we focus on benchmarking the eQTL power, which showed good agreement between the simulated and analytic values (**Figure 3F**).

The analytic calculations of scPower are orders of magnitude faster than the simulation-based approaches: calculations for **Figure 3C-E** took 8 days for powsimR, 3 days for muscat and less than a minute for scPower (**Figure 3G**). Also the memory requirements are much lower, as no count matrices are generated. For the simulation-based methods the memory requirements increase with larger sample size and numbers of cells, leading for example for 20 samples and 3,000 cells per sample to 48GB used memory for powsimR and 35GB used memory for muscat compared to the parameter-independent requirements of scPower of few MB (**Figure 3H**). In addition the installation of scPower is easier due to less dependencies: 8 dependencies of scPower vs. 82 dependencies of powsimR and 28 dependencies of muscat. These advantages of scPower over simulation based approaches enable a systematic evaluation of a large number of design options as described in the next section.

## scPower maximizes detection power for a fixed budget by optimizing experimental parameters

With this model for power estimation in DE and eQTL single cell studies in place, we are now able to optimize the experimental design for a fixed budget. The overall cost function for a 10X Genomics experiment is the sum of the library preparation cost and the sequencing cost (see Methods). The library preparation cost is defined by the number of measured samples and the number of measured cells per sample, while the sequencing cost is defined by the number of sequenced reads, which depends also on the target read depth per cell. **Supplementary Figures S15, S16** show the three parameters maximizing detection power, given a fixed total budget. **Figure 4** shows the optimization with expression priors from our PBMC data set, measured with 10X Genomics.

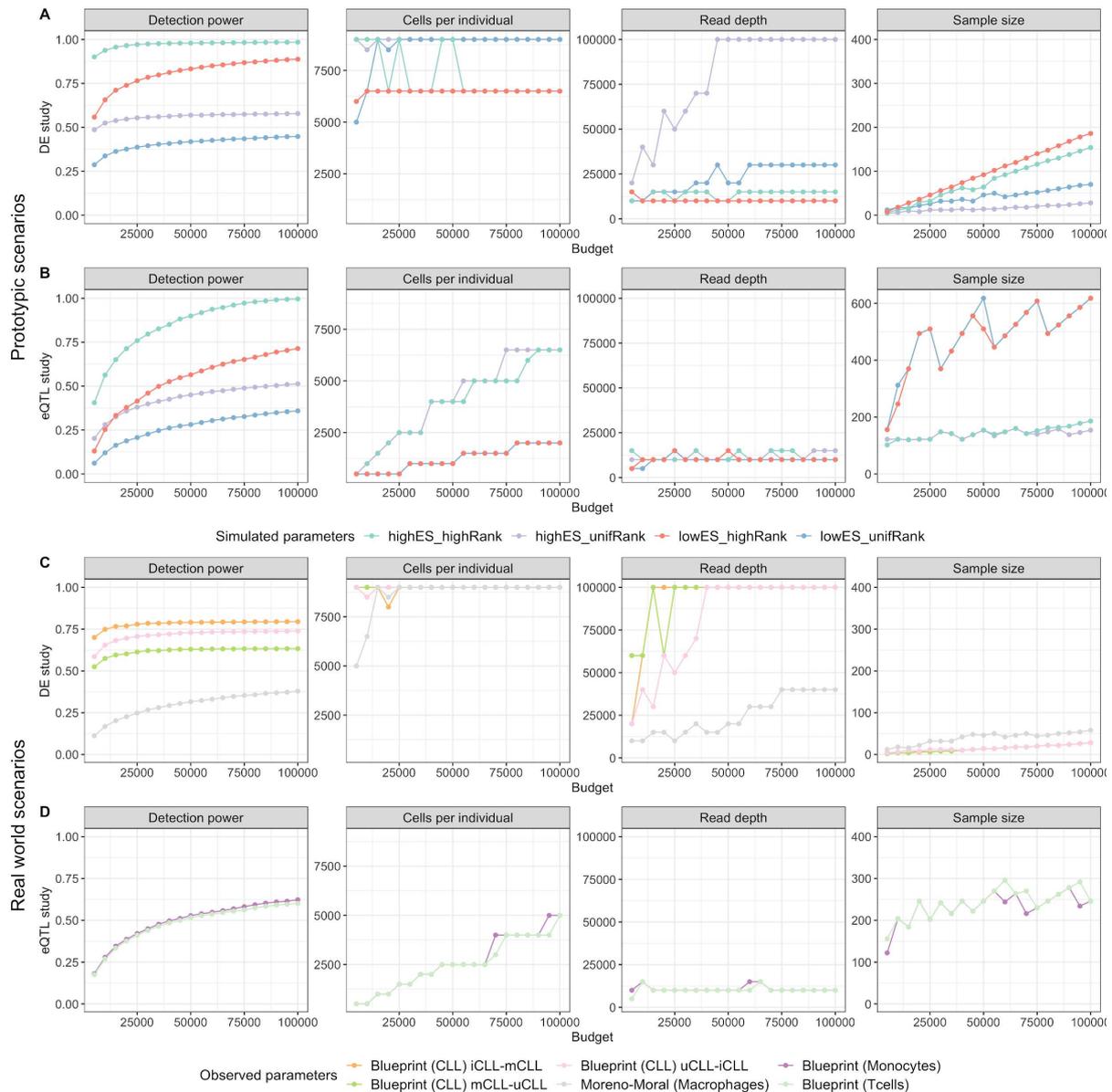
We systematically investigated the evolution of optimal parameters for increasing budgets in four prototypic scenarios for DEG (**Figure 4A**) and eQTL analysis (**Figure 4B**), four scenarios based on prior DEG (**Figure 4C**) and two scenarios on prior eQTL (**Figure 4D**) experiments on FACS sorted cells (for the estimated costs see **Supplementary Table S5**). The prototypic scenarios reflect combinations of effect sizes (high, low) and expression ranks (high, low) of

DEGs and eQTL genes. We observed that the number of cells per individual is the major determinant of power, as this is the variable that is either directly set to maximum values or increased first in the optimization (**Figure 4**). This effect is least pronounced in the prototypic eQTL scenario (**Figure 4B**), where small effect sizes require large sample sizes. For most DEG scenarios, the number of reads per cell is increased before increasing the sample size (**Figure 4A,C**), indicating that strong effects can be detected with relatively few samples, while the detection of expression requires deeper sequencing. For eQTL scenarios increasing first the sample size is more beneficial than increasing the read depth (**Figure 4B,D**), which remains relatively low (10,000 reads per cell).

Figure 4 was generated with FDR adjusted p-values for DE power and FWER adjusted p-values for eQTL power. Using FWER adjustment for DE power changes the observed overall power, but leads to very similar optimal parameter combinations and the same trends overall (see **Supplementary Figure S17**).

In the cost optimization, we also took into account that increasing the number of cells per lane leads to higher numbers of doublets, droplets with two instead of one cell. Doublet detection methods such as Demuxlet<sup>47</sup> and Scrublet<sup>69</sup> enable faithful detection of those to exclude the doublets from the downstream analysis. We validated the doublet detection and donor identification of Demuxlet using our PBMC data set by comparing the expression of sex specific genes with the sex of the assigned donor (**Supplementary Figure S1B**) and found high concordance after doublet removal, also for run 5, which was overloaded with 25,000 cells.

The increase of the doublet rate through overloading was modeled using experimental data<sup>70</sup> to accurately estimate the number of usable cells for the eQTL/DEG analysis. However, we observe in our own data set as well as in published studies<sup>47,71</sup> slightly higher doublet rates than shown in<sup>70</sup>. Therefore, we consider the modeled doublet rate as a lower bound estimation. With a high detection rate of doublets, overloading of lanes is highly beneficial, since larger numbers of cells per individual lead to an increase in detection power, while not causing additional library preparation costs. This supports previous evaluations that demonstrated the benefit of overloading<sup>50</sup>. Although, overloading leads to a decreasing number of usable cells and a decreasing read depth of the singlets, as doublets contain more reads, the overall detection power still rises strongly for both DE and eQTL studies.



**Figure 4: Optimal parameters for varying budgets and 10X Genomics data.** The figure shows the maximal reachable detection power (y-axis, first column) for a given experimental budget (x-axis) and the corresponding optimal parameter combinations for that budget (y-axis, second till fourth column). The colored lines indicate different effect sizes and gene expression rank distributions. Panels A-B visualize different simulated effect sizes and rank distributions (simulation names see text) for DEG studies (A) and eQTL studies (B) with models fitted on 10X PBMC data. Panels C-D visualize effect sizes and rank distributions observed in cell type sorted bulk RNA-seq DEG studies (C) and eQTL studies (D) with model fits analogously to A-B.

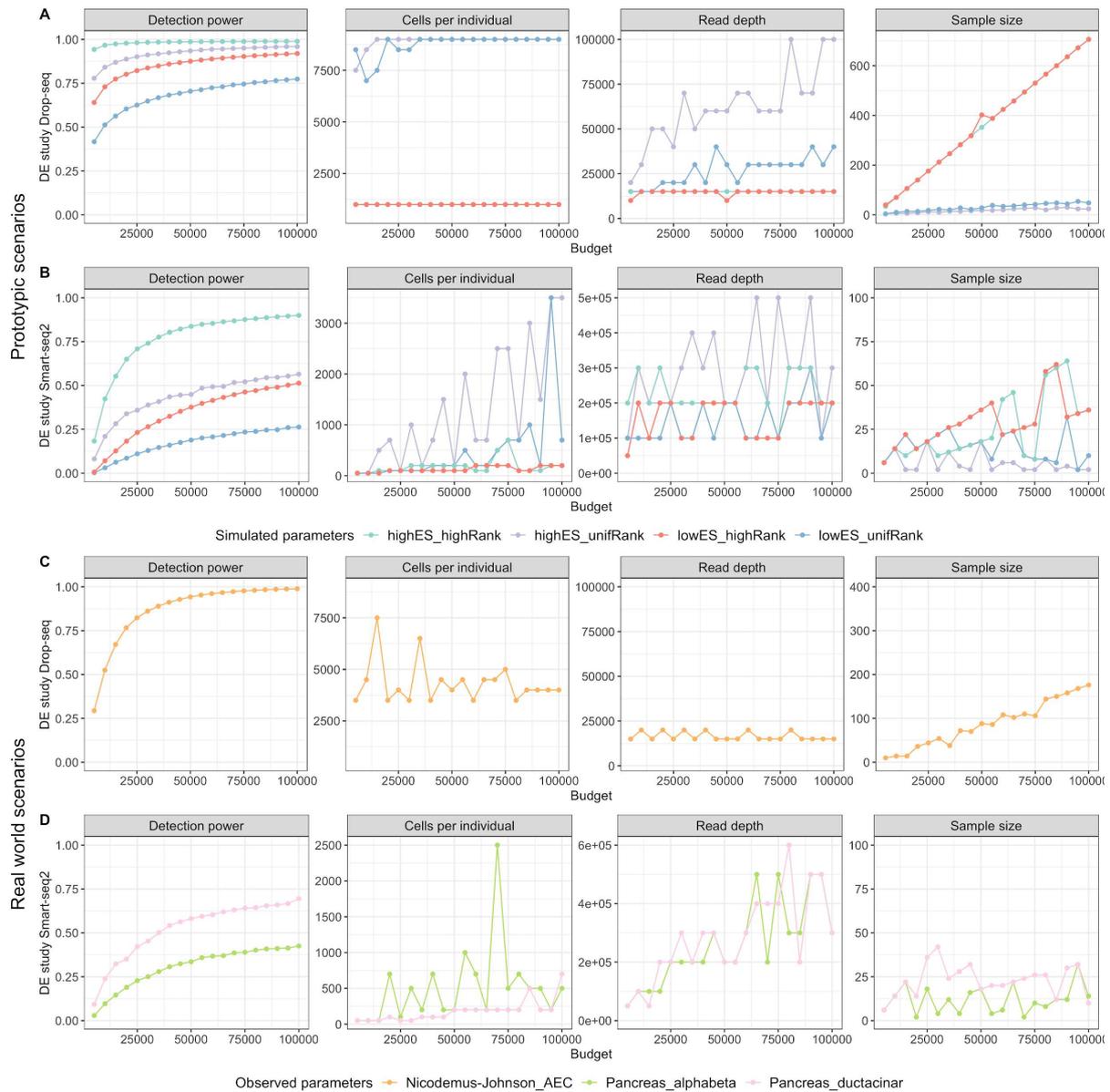
## scPower generalizes across tissues and scRNAseq technologies

Our power analysis framework is applicable on data sets for other tissues besides PBMC and for other single cell technologies besides 10X Genomics. We demonstrate this with a lung cell data set measured by Drop-seq<sup>52</sup> and a pancreas data set measured by Smart-seq2<sup>51</sup>. The model of the expression probability needs to be adapted slightly for other technologies, while

the DE/eQTL power calculation remains the same as for 10X (for the estimated costs see **Supplementary Table S5**).

Smart-seq2 is a plate-based technology, generating read counts from full-length transcripts. Therefore, we express the count threshold for an expressed gene relative to one kilobase of the transcript. We fitted the expression model including the transcript length in the size normalization factor of the count model. Additionally, we modelled the doublet rate as a constant factor. In contrast, Drop-seq is a droplet-based technology similar to 10X Genomics and exactly the same model can be used. However, as we are lacking the experimental data to fit an appropriate model for overloading, we set the doublet rate again constant. With these adaptations our expression probability model (**Supplementary Figure S19**) and the power calculation for other tissues and technologies is in good agreement with simulation based estimates (**Supplementary Figure S20**). However, our framework performs slightly better on droplet-based technologies.

Analogously to **Figure 4**, the evolution of parameters for simulated priors (**Figure 5A,B**) and observed priors (**Figure 5C,D**) was evaluated across the other technologies. Similar trends are observed for the Drop-seq lung data as for the 10X PBMC data set in the prototypic scenarios (**Figure 5A**) as well as for observed priors from cell type sorted bulk studies (**Figure 5C**). In both cases, the number of cells per individual is the major determinant of power. Overall, lower power is observed for the Smart-seq2 pancreas study (**Figure 5B,D**). In contrast to 10X and Drop-seq, the optimal number of reads per cell is much higher and the number of cells per individual and sample size is increased only at higher budgets for both the prototypic and data driven priors.



**Figure 5: Optimal parameters for varying budgets and Drop-seq and Smart-seq2 data.** The figure shows the maximal reachable detection power (y-axis, first column) for a given experimental budget (x-axis) and the corresponding optimal parameter combinations for that budget (y-axis, second till fourth column). The colored lines indicate different effect sizes and gene expression rank distributions. Panels A-B visualize different simulated effect sizes and rank distributions (simulation names see text) for DE studies with models fitted on Drop-seq lung data (A) and Smart-seq2 pancreas data (B). Panels C-D visualize effect sizes and rank distributions observed in cell type sorted bulk RNA-seq DE studies with model fits analogously to A-B.

## Power to detect rare cell types

Additionally to the detection power for DE / eQTL genes, our model also covers the cell type detection probability. This is another important aspect for designing cell type specific multi-sample experiments. In exploratory analyses, the goal is to observe as many cell types as possible. The power to observe rare cells depends on the frequency of this cell type, the number of cells sequenced per individual and the total number of individuals. Following <sup>37</sup>, we

model this problem using the negative binomial distribution (see methods). Here we demonstrate the approach using prior knowledge of cell proportions in PBMCs from the literature to determine the number of cells required for each individual to detect a minimal number of cells of a specific type. The rarest immune cell type we considered are dendritic cells, which occur in PBMCs with a frequency of 1.5%. Consequently, more than 1000 cells per individual are required to observe at least ten dendritic cells in all individuals with probability greater than 95%, while only about 300 cells are required for NK cells, which have a frequency of 7% in PBMCs (**Supplementary Figure S21**). The comparison for varying numbers of individuals shows that the number of cells required for each individual is most strongly affected by the frequency of the cell type and only to a smaller degree by the number of individuals.

## Discussion

We have introduced *scPower*, a method for experimental design and power analysis for interindividual differential gene expression and eQTL analysis with cell type resolution. Our model generalizes across different tissues and scRNAseq technologies and will provide the basis for rationally designing well powered experiments, increasing the number of true biological findings and reducing the number of false negatives.

Previous experimental design methods for multi-sample scRNA-seq<sup>44</sup> are based on simulations. This allows for assessing more complex single cell multi-sample data, including scenarios of cell to cell heterogeneities other than differential gene expression. However, analytical models, such as our framework, are by orders of magnitude faster than comparable simulation-based tools. This transformatively enables the evaluation of many experimental design options in a short time and thus to identify optimal experimental parameters for a limited budget. Additionally, analytical models require only a small amount of memory independent of the assessed experimental parameters, while simulation of data sets with larger sample sizes lead to increasing memory usage. A sample size of 20 with 3,000 cells per sample required already between 35 GB (muscat) and 48GB (powsimR) in our evaluation. Therefore, larger data sets with hundred and more samples, as required for eQTL studies, will be very difficult to simulate.

A previous analytic exploration for single cell eQTL studies<sup>50</sup> had three limitations: first, it was based solely on the effective sample size, ignoring actual effect sizes and expression strength of eQTL genes. Second, it provided no generalizable tool. Third, it is limited to eQTL analysis and does not cover DE studies. In contrast, our approach provides gene level and overall power estimates based on prior data and we provide a generalizable tool for analytic power analysis of single cell DE and eQTL studies. This enables the user the evaluation of his target experiment in order to identify the use-case specific optimal parameter combination. While this optimum is also identified with the approach based in effective sample size, our method also provides an estimate of the overall power. This information is crucial for the design of successful experiments, as researchers would rather attempt an experiment with 80% overall power than one with 20%. The method is implemented in an R package with a user friendly graphical user interface and is freely available on github. Additionally, the graphical interface of our model is also available over this website <http://scpower.helmholtz-muenchen.de/>.

Based on data driven priors on expression distributions from single cell atlases of three different tissues and cell type specific priors for effect sizes based on DEGs and eQTLs from bulk RNA-seq experiments, we show that the number of cells drives power by increasing the sensitivity of gene expression detection. Previous analyses<sup>35</sup> have recommended 1Mio reads

per cell when comprehensive gene expression detection is desired. In line with <sup>50</sup> our analyses suggest that aggregating shallowly sequenced transcriptomes of a large number of cells of the same cell type is a more cost efficient way than increasing read depth to increase the sensitivity for individual level gene expression analysis. Most likely, multiple independent library preparations in individual cells lead to an improved sampling of the transcriptome as compared to fewer independent libraries sequenced more deeply, an effect that has previously been analysed in the context of variant detection <sup>72</sup>. The number of cells to be sequenced has previously been considered with respect to the power of detecting rare cell types <sup>37,38</sup>, however, its effect on gene expression sensitivity is equally important.

Optimal read depths (~10000) are relatively low compared to previous recommendations <sup>34,36,73,74</sup>. In a systematic analysis of ERCC spike-in expression it has been shown that the accuracy of the measurements is not strongly dependent on the sequencing depth and consistently high (~0.9 Pearson correlation for 10X and Drop-seq) for a read depth of 10,000 reads per cell <sup>35</sup>. Hence, we expect accurate individual level gene expression quantification with the optimized experimental design.

The number of cells and sequencing depth also determine the accuracy of the extraction of gene expression programs, which are critical for the annotation of cell types <sup>75</sup>. Shallow sequencing of a higher number of cells has achieved equal accuracy as deeper sequencing of fewer cells <sup>75</sup>. In line with our findings, it has thus been recommended to shallowly sequence more cells <sup>75</sup>. Similar observations have also been made for single cell eQTL studies <sup>50</sup>. To ensure sufficient power for cell type identification, our framework *scPower* can be combined with power analysis tools for cell type identification <sup>38,75</sup>.

As expected, the sample size is mostly dependent on the effect size, with low effect sizes requiring large sample sizes and consequently optimal setting with high sample size typically lead to low sequencing depth and relatively low number of cells.

In general, priors affect the optimal design and should therefore be selected carefully. In the optimal case, priors are known from well matched pilot experiments or knowledge from the literature. In the absence of well-matched pilot experiments it is nevertheless important to make assumptions explicit by either selecting a prior based on a similar biological phenomenon, or by choosing a prototypic case. In our study we have compared the prototypic cases of strong effect sizes and relatively high expression versus intermediate effect sizes and expression levels across the whole range from highly expressed to lowly expressed genes. Both options, processing priors from a selected reference study and simulating proteotypic priors, are possible with *scPower* and described in the package vignette.

The pseudobulk approach presented here leverages well established power analysis methods based on (generalized) linear models. While it represents a baseline method for power analysis, it has some limitations. First, although the (negative binomial) regression model for pseudobulk is currently the most powerful method for assessing individual level differential expression <sup>41</sup>, other more powerful methods might be developed in the future. Our current framework is tightly linked to this approach and cannot easily be extended to arbitrary analysis methods, this is however the case for all analytical power analysis methods. Second, it requires a discrete cell type definition. Therefore, continuous cell annotations such as pseudo time would need to be discretized before the power analysis. Third, we did not address the power for the detection of variance QTLs from scRNAseq data <sup>48</sup> due to the lack of data driven priors for the effect sizes.

Several practical considerations should be addressed when using our approach. First, our data driven priors only allow for reliably assessing the overall power in sample sizes that are smaller or roughly equal to the sample size of the pilot data sets from which the effect sizes

were estimated. Consequently, a larger sample size will identify new significant DEGs with lower effect sizes, which were not identified in the smaller pilot study and thus not included in the computation of the overall detection power. Second, our current modeling of the doublet rate using reference values from 10X Genomics is a lower bound compared to the doublet rates we estimate for our own data and to rates reported by other studies<sup>47,71</sup>. Thus, actual experiments might result in higher doublet rates and lower number of usable cells. Last, the choice of a threshold on the number of reads required for a gene to be called expressed also influences the choice of optimal parameters. Here we used a threshold of >10 and >3 reads, however, some eQTL analyses of bulk RNAseq data advocate using >0 reads<sup>68</sup>, whereas DESeq2 automatically chooses the threshold that optimizes the number of discoveries at a given FDR by applying the independent filtering strategy<sup>5,76</sup>. Best practice guidelines for differential gene expression with RNA-seq recommend cutoffs that remove between 19%-33% of lowly expressed genes, depending on the analysis pipeline<sup>77</sup>. These percentages correspond to 1-10 reads per million sequenced, which translates to 1-5 UMI counts for a median of around 5000 UMI counts per cell in our data set. Our gene expression probability model is cell type specific and has to be fitted based on realistic pilot data. We have shown that our model can be applied to 10X Genomics, Drop-seq and Smart-seq2 and we would expect that it is also applicable to other technology platforms.

Importantly, experimental design recommendations here are optimized for differential expression between individuals. Other applications might result in very different optimal experimental designs. For instance, co-expression analysis requires a high number of quantified genes per cell, especially when one is interested in cell type specific co-expression and comparison of such co-expression relations between individuals. Furthermore, the power to identify new rare cell types by clustering analysis of scRNA-seq data might have different optimal parameters<sup>38</sup>.

The human cell atlas project has outlined a 'skydive' strategy of iteratively sampling human cells with increasing resolution to build a reference map of healthy human cells<sup>61,78</sup>. In combination with the human cell atlas reference transcriptomes and cell type annotations *scPower* will provide the foundation for building a comprehensive resource for the experimental design of interindividual gene expression comparisons with cell type resolution across all organs systems covered in the human cell atlas.

## Methods

### Collection of PBMCs

Blood was collected from psychiatric control individuals according to the clinical trial protocol of the Biological Classification of Mental Disorders study (BeCOME; ClinicalTrials.gov TRN: NCT03984084) at the Max Planck Institute of Psychiatry. All individuals gave informed consent. Perinuclear blood cells (PBMCs) were isolated and cryopreserved in RPMI 1640 medium (Sigma-Aldrich) supplemented with 10% Dimethyl Sulfoxide at a concentration of roughly 1M cells per ml.

## Single cell RNA-sequencing

For single-cell experiments, 14 cell vials from different individuals (7 male and 7 female) were snap-thawed in a 37°C water bath and serially diluted in RPMI 1640 medium (Sigma-Aldrich) supplemented with 10% Fetal Bovine Serum (Sigma-Aldrich) medium. Cells were counted and equal cell numbers per individual were pooled in two pools of 7 individuals each. Cell pools were concentrated and resuspended in PBS supplemented with 0.04 % bovine serum albumin, and loaded separately or as a combined pool with cells of all 14 individuals on the Chromium microfluidic system (10X Genomics) aiming for 8,000 or 25,000 cells per run. Single cell libraries were generated using the Chromium Single Cell 3'library and gel bead kit v2 (PN #120237) from 10X Genomics. The cells were sequenced with a targeted depth of approximately 50,000 reads per cell on the HiSeq4000 (Illumina) with 150 bp paired-end sequencing of read2 (exact numbers for each run in **Supplementary Table S2**).

## Preprocessing of the single cell RNA-seq data

We mapped the single cell RNA-seq reads using CellRanger version 2.0.0 and 2.1.1 <sup>79</sup>. Demuxlet was used to identify doublets and to assign cells to the correct donors <sup>47</sup>. Additionally, Scrublet version 0.1 was run with a doublet threshold of 0.28 to identify also doublets from cells which originate from the same donor <sup>80</sup>. Afterwards, the derived gene count matrices from CellRanger were loaded into Scanpy version 1.4 <sup>81</sup>. Cells identified as doublets or ambivalent by Demuxlet and Scrublet were removed, as well as cells with less than 200 genes or more than 2,500 genes and with more than 10% counts from mitochondrial genes.

## Verification of Demuxlet assignment using sex errors

We validated the donor assignment and doublet detection of Demuxlet by testing if assigned cells express sex specific genes correctly. Xist expression was taken as evidence for a female cell, expression of genes on the Y chromosome as evidence for a male cell.

The male sex error shows the fraction of cells assigned to a male donor among all cells expressing Xist (count > 0). The threshold for the female error was set less strictly, as mismapping of a few reads to the chromosome Y occurs also in female cells. Instead, the female sex error indicates which fraction of cells is assigned to a female donor among all cells having more reads mapped to chromosome Y than the  $q_f$  quantile of all cells, with  $q_f$  being the overall fraction of cells assigned to a female donor among all cells. TPM mapped to chromosome Y is calculated by counting all reads mapped to chromosome Y, excluding reads mapped to the pseudoautosomal regions, times  $10^6$  divided by the total number of read counts per cell.

Both error rates are calculated twice, once with all cells and once without doublets from Demuxlet and Scrublet.

## Cell type identification

We performed the cell type identification according to the Scanpy PBMC tutorial <sup>82</sup>. Genes which occurred in less than 3 cells were removed. Counts were normalized per cell and logarithmized. Afterwards the highly variable genes were identified, the effect of counts and mitochondrial percentage regressed out. We calculated a nearest neighbour graph between the cells, taking the first 40 PCs, and then clustered the cells with a Louvain clustering <sup>83</sup>. Cell types were assigned to the clusters using marker genes (**Supplementary Table S3**).

## Influence of read depths

We used subsampling to estimate the dependence of gene expression probabilities on read depths. The fastq files of all 6 runs were subsampled using fastq-sample from fastq-tools version 0.8<sup>84</sup>. The number of reads was downsampled to 75%, 50% and 25% of the original number of reads. CellRanger was used to generate count matrices from the subsampled reads. Donor, doublet and cell type annotation were always taken from the full runs with all reads.

## Expression probability model

The gene expression distribution of each cell type was modeled separately because there are deviations in RNA content between different cell types<sup>33</sup>. The UMI counts  $x$  per gene across the cells of a cell type are modeled by a negative binomial distribution. We used DESeq<sup>60</sup> to perform the library size normalization as well as the estimation of the negative binomial parameters. The standard library size normalization of DESeq and the variant “poscounts” of DESeq2<sup>5</sup> were both used, depending on the quality of the fit for the specific data set. For the PBMC 10X data set (**Supplementary Table S2**), the standard normalization was taken, for the Drop-seq lung and the Smart-seq2 pancreas datasets the poscount normalization, which is more suitable for sparse data. Only cell types with at least 50 cells were analysed to get a robust estimation of the parameters.

The negative binomial distribution is defined by the probability of success  $p$  and the number of successes  $r$ :

$$f_{NB}(x, r, p) = NB(x, r, p) = \binom{x+r-1}{x} (1-p)^r p^x$$

DESeq uses a parametrization based on mean  $\mu = \frac{p \cdot r}{1-p}$  and dispersion parameter  $\phi = \frac{1}{r}$ .

We formulated the definition of an expressed gene in a flexible way so that users can adapt the thresholds. The definition is based on the pseudobulk approach where the counts  $x_{i,j}$  are summed up per gene  $i$  for all cells  $j$  part of cell type  $c$  and donor  $s$  to a three dimensional matrix  $y_{i,c,s} = \sum_{j \in C \wedge j \in S} x_{i,j}$  with  $C$  the set of all cells of cell type  $c$  and  $S$  the set of all cells of donor  $s$ .

In general, a gene  $i$  is called *expressed* in a cell type  $c$  if the sum of counts  $y_{i,c,s}$  over all cells of the cell type within an individual  $s$  is greater than  $n$  in more than  $k$  percent of the individuals. We assume a negative binomial distribution ( $f_{NB}(x, \mu_{i,c}, \phi_{i,c})$ ) for the counts  $x_{i,j}$  of each gene  $i$  in each cell type  $c$  with  $\mu_{i,c}$  and  $\phi_{i,c}$ . The sum of gene counts  $y_{i,c,s}$  follows a negative binomial distribution where the parameters are altered by the number of cells per cell type and donor  $n_{c,s} = |\{j \in C \wedge j \in S\}|$  to  $\mu'_{i,c,s} = n_{c,s} \cdot \mu_{i,c}$  and  $\phi'_{i,c,s} = \frac{\phi_{i,c}}{n_{c,s}}$ . The probability that the sum of counts  $y$  is greater than  $n$  is

$$p_{i,s} = P(y_{i,c,s} > n) = 1 - F_{NB}(n, \mu'_{i,c,s}, \phi'_{i,c,s})$$

with  $F_{NB}$  as the cumulative negative binomial distribution.

To define a gene as expressed, we require that it can be found in a certain fraction of more than  $k$  percent in all  $n_s$  individuals. The expression probability of a gene  $i$  is obtained from a cumulative binomial distribution  $F_{Bin}$  as

$$P(i \in E) = 1 - F_{Bin}(k \cdot n_s, n_s, p_{i,s})$$

So in total, the expected value of the expected number of expressed genes ( $E$ ) can be defined as

$$\mathbb{E}(E) = \sum_{gene\ i} P(i \in E)$$

To generalize the expression probability model also for unseen data sets, the distribution of the mean values  $\mu_{i,c}$  over all genes in a cell type  $c$  is modelled as a mixture distribution with three components, a zero component  $Z(x)$  and two left-censored gamma distributions  $\Gamma(x, r, s)$

$$f_{\mu_t}(x) = p_1 Z(x) + p_2 \Gamma(x, r_1, s_1) + p_3 \Gamma(x, r_2, s_2)$$

The model is an adaption of the distribution used in the single cell simulation tool Splatter<sup>85</sup>. The largest part of the mean values can be fitted with one gamma distribution, a small fraction with high expressed gene outlier with the second gamma distribution. The genes with zero mean values originate from two sources. Either, the gene is not expressed or the expression level is too low to be captured in the setting. The lower bound for the expression level at which both Gamma distributions are censored depends on the number of cells  $j$  measured for this cell type  $n_c = |\{j \in C\}|$ . The smallest expression level to be captured is  $\frac{1}{n_c}$ .

The density of the gamma distribution is parametrized by rate  $r$  and shape  $s$ :

$$\Gamma(x, r, s) = \frac{s^r x^{r-1} e^{-sx}}{(r-1)!}$$

For modeling of the gamma parameters, also the parameterization by mean  $\mu = \frac{s}{r}$  and standard deviation  $\sigma = \sqrt{\frac{s}{r^2}}$  is used.

The relationship between the mean UMI counts per cell and the gamma parameters (mean and standard deviation of the two gamma distributions) is linear and  $\beta$  values are estimated by linear regression. The mixture proportion of the zero component  $p_1$  is linearly decreasing with the mean UMI counts, also estimated by linear regression. The lower bound of  $p_1$  is set to a small positive number: 0.01. In contrast, the mixture proportion of the second gamma component  $p_3$  is modelled as a constant, independent of the mean UMI counts. We set it to the median value of all fits per cell type. The mixture proportion of the first gamma component is  $p_2 = 1 - p_1 - p_3$  and is linearly increasing with increasing mean UMI counts.

The number of transcriptome mapped reads is linearly related to the logarithm of the mean UMI counts per cell, with an increasing read depth leading to a saturation of UMIs. 10X Genomics describes this also with the sequence saturation parameter. The exact logarithmic saturation curve depends on multiple biological and technical factors, therefore, it needs to be

fitted for each experiment individually. However, *scPower* provides example fits from the different scenarios observed in our analysis.

The dispersion parameter is estimated dependent on the mean value using the dispersion function fitted by DESeq. The parameters of the mean-dispersion curve showed no correlation with the mean UMI counts, therefore the mean of the parameters of the dispersion function across all runs and subsampled runs were taken, resulting in one mean-dispersion function per cell type.

## Power analysis for differential expression

The power to detect differential expression, also denote as the probability  $P(i \in S)$  that gene  $i$  is in the set of significant differentially expressed genes  $S$ , is calculated analytically for the negative binomial model <sup>65</sup>. An implementation of the method can be found in the R package MKmisc. Parameters are sample size, fold change, significance threshold, the mean of the control group and the dispersion parameter (assuming the same dispersion for both groups). Zhu et al. implemented three different methods to estimate the dispersion parameter, we chose method 3 for the power calculation, which was shown to be more accurate in simulation studies in the paper. More complex experimental designs can be addressed using the method of <sup>86</sup>.

## Power analysis for expression quantitative trait loci

Additionally to the DE analyses, the use of scRNA-seq for the detection of expression quantitative trait loci (eQTLs) was evaluated. We distinguish for the eQTL power between genes with high and with low expression levels, where the mean is used to parameterize a simulation. Therefore, the eQTL power is a function of the mean expression level.

For genes with high expression level, the power to detect an eQTL is calculated analytically using an F-test and depends on the sample size  $n_s$ , the coefficient of determination  $R^2$  of the locus and the chosen significance threshold  $\alpha$ .  $R^2$  is calculated for the pilot studies from the regression parameter  $\beta$ , its standard error  $se(\beta)$  and the sample size  $N$  of the pilot study:

$$t = \frac{\beta}{se(\beta)}$$

$$R^2 = \frac{t^2}{N - 2 + t^2}$$

The implementation `pwr.f2.test` of the R package `pwr` is used for the F-test <sup>20</sup>. The degrees of freedom of the numerator are 1 and of the denominator are  $n_s - 2$ , the effect size is  $\frac{R^2}{1-R^2}$ .

This power calculation assumes that the residuals are i.i.d. normally distributed. For large count values, it has been shown that normalized log transformed counts have a constant variance independent of the mean value and can be analysed with linear models <sup>46</sup>. However, for genes with small mean values, i.e. only very few non-zero counts, this normalization might not be effective and the power is overestimated by the analytical power calculation based on the F-test. We performed a simulation study to assess the effect of the mean values on the eQTL power.

To account for the discrete nature of the counts we adopted a simulation scheme similar to a negative binomial regression model and analysed the log transformed counts using linear models <sup>46</sup>. As for the analytical power calculation, the effect size is given by the coefficient of determination  $R^2$  of the locus. To determine the simulation based power for sample size  $n_s$ ,

significance threshold  $\alpha$  and mean count  $\mu_c$  of the allele with lower expression, the following steps are repeated  $B = 100$  times:

- 1) Simulate genotypes. To also account for the discrete nature of the genotypes, we first draw allele frequency  $f_a$  from a uniform distribution between 0.1 and 0.9. A random genotype vector  $g$  with  $g_i \in \{0,1,2\}$  of length  $n_s$  is generated with the expected number of each genotype  $(f_a^2, 2f_a(1-f_a), (1-f_a)^2)$  according to Hardy Weinberg equilibrium.
- 2) Simulate read counts. Using the allele frequency, the beta value  $\beta$  and the standard deviation of the residuals  $\hat{\sigma}$  is calculated:

$$\beta = \sqrt{\frac{R^2}{2 * f_a * (1 - f_a)}}$$

$$\hat{\sigma} = \sqrt{1 - R^2}$$

The associated gene expression count vector  $x$  is sampled from a negative binomial distribution parameterized for each genotype  $g_i$  with mean  $\mu_i = e^{\log(\mu_c) + \beta * g_i}$  and dispersion  $\phi_i$ . In the following, we work with log transformed counts (plus one pseudo count). To match with the effect size  $R^2$ , the dispersion parameter  $\phi_i$  is chosen, such that the variance of the log transformed counts is  $\hat{\sigma}$ . Since the Taylor approximation of the dispersion parameter<sup>46</sup> was not accurate enough, we used instead a numerical optimization. This numerical optimization is precalculated for a range of parameter combinations to speed up calculation for the user.

- 3) Using the linear regression  $\log(x_i + 1) \sim g_i$ , the P-value  $P_i$  for  $H_0: \beta = 0$  is determined.

Finally, the simulation based power is estimated as  $\sum_{i=1}^B P_i < \alpha$

The power of the simulation was compared with the analytic power calculated by scPower to assess at which value of the mean  $\mu_c$  the analytic power starts to overestimate the simulation based empirical power (see **Supplementary Figure S10**) for Bonferroni adjusted significance thresholds used in eQTL analyses. We choose a cut-off of mean count  $< 5$  and estimate the power for genes with smaller mean values based on simulation instead of the F-test to increase accuracy for small count values.

## Overall detection power

The overall detection power for DEGs / eQTLs is the product of the expression probability and power to detect DEGs / eQTLs, as both probabilities are conditionally independent given the expression mean of the gene. Expression probabilities were determined based on the gene expression rank in the observed (pilot) data. The number of considered genes  $G$  was set to 21,000, the number of genes used for fitting of the curves. Ranks  $i$  were transformed to the quantiles  $\frac{i}{G}$  of the gamma mixture model parameterized by the mean UMI counts to obtain the mean  $\mu_c$  of the negative binomial model, which is in turn used to compute the expression probability.

To quantify the overall power of an experimental setup, we compute the expected fraction of detected DEG / eQTL genes with prior expression levels and effect sizes derived from the pilot data. We obtain gene expression ranks of DEGs / eQTLs and their corresponding fold changes to compute overall detection power for each gene.

DE / eQTL power is computed using a significance threshold  $\alpha$  corrected for multiple testing, controlling either the family-wise error rate (FWER) or the false discovery rate (FDR). We used FDR adjustment for the DE power and followed the approach of the GTEx consortium<sup>67</sup> based

on FWER adjustment for the eQTL power. However, our framework allows for any combination of power analysis and multiple testing method. For all analyses shown, adjusted  $\alpha$  was set to 0.05.

The family-wise error rate is defined as the probability of at least one false positive among all tests. Each expressed gene is tested once in the DE analysis, therefore, the adjustment for the family-wise error rate is done by correcting the threshold to  $\frac{\alpha}{(E)}$  for  $(E)$  expected expressed genes. For eQTLs we followed the approach of <sup>67</sup>, which assumes that for each gene on average 10 independent (uncorrelated) SNPs are tested in a genome-wide cis eQTL analysis. Thus, the adjusted P-value threshold is set at  $\frac{0.05}{(E)*10}$ .

Alternatively, for DE analysis the significance threshold can be adjusted for the false discovery rate using the method of Jung <sup>22</sup>. Using his formula, the FDR corrected threshold  $\alpha = FDR(\alpha')$  is estimated depending on the total number of true null hypotheses  $m_0$  and the number of correct rejections of the null hypothesis for true alternative cases at this threshold  $r_1(\alpha')$ :

$$FDR(\alpha') = \frac{m_0 * \alpha'}{m_0 * \alpha' + r_1(\alpha')}$$

$m_0$  is the number of expected expressed genes without the expected expressed DEGs/eQTLs,  $(E) - \left( \frac{E_{DEG}}{eQTL} \right)$ .  $r_1(\alpha')$  is the expected number of significant DEGs/eQTLs, which is the sum of the overall power for a significance threshold of  $\alpha'$  for all true DEGs/eQTLs. Therefore, the unknown adjusted significance threshold  $\alpha'$  can be numerically estimated using the R function *uniroot*.

The overall power of the experimental setup is then the average detection power over all prior DEG/eQTL genes.

## Pilot data sets

Realistic DE and eQTL priors, i.e. effect sizes and expression ranks, were taken from sorted bulk RNA-seq studies of matching tissues (PBMCs, lung and pancreas). For all studies, the significance cut-off of the DE and eQTL genes was set to  $FDR < 0.05$  and the expression levels of the genes were taken from FPKM normalized values. When published, we took directly the effect sizes, otherwise we recalculated the DE analysis with DEseq2.

*Differential gene expression:* To get realistic estimates for effect sizes (fold changes), data sets from FACS sorted bulk RNA-seq studies were taken <sup>53,54</sup>. The data sets were used to rank the expression level of the DEGs among all other genes using the FPKM values. The cell types used in the studies were matched to our annotated cell types in PBMCs for the expression profiles. The expression profile of CD14+ Monocytes was used for the study of Macrophages, the profile of CD4+ T cells for the CLL study.

Lung cell type specific priors were obtained from a DE study of freshly isolated airway epithelial cells of asthma patients and healthy controls <sup>55</sup>. As no effect sizes were reported, the analysis was redone with the given count matrix from GEO using DEseq2.

A DE study analyzing age-dependent gene regulation in human pancreas <sup>56</sup> was used to get pancreas cell type specific priors. We obtained expression ranks and gene length, which is needed for proper normalization of Smart-seq2 expression values.

*eQTLs:* We used eQTL effect and sample sizes from the Blueprint study on bulk RNA-seq of FACS sorted Monocytes and T cells <sup>57</sup>. Neutrophils were excluded as they are not PBMCs. We took the most significant eQTL for each gene, using a significance cutoff of  $10^{-6}$ . We

compared the FPKM normalized expression levels of the eQTL genes among all other genes to get the expression rank for each eQTL gene. Effect sizes were derived from the slope parameters of the linear regression against genotype dosage, its standard error and the sample size of the study.

## Comparison with simulation-based power analysis tools

To validate our model, we compared the DE power estimations of our framework with two simulation-based tools, called *powsimR* and *muscat*<sup>40,41</sup>. For both tools, a few changes needed to be implemented to compare the output exactly with our approach. *powsimR* is not designed for multi-sample comparison and for both methods the option to apply a vector of log-fold changes with matching expression ranks was not available. A detailed explanation of both methods and applied changes can be found below.

The simulation-based methods perform random sampling of their count matrices and therefore the simulation was repeated 25 times for each parameter combination to generate stable results. Both tools allow the power estimation for different DE methods. We evaluated *powsimR* in combination with edgeR-LRT, DESeq and limma-voom, together median-ratio normalization of DESeq ('MR'), and *muscat* in combination with edgeR, DESeq2, limma-voom and limma-trend. No imputation or filtering was applied for any of the methods. In the comparisons with our model *scPower*, the expression probability parameters of *scPower* were set to  $\text{minCounts} > 0$  in at least one individual to match the detected genes of *powsimR* and *muscat*.

Exemplarily, the CD4 T cells of our PBMC data set were used for fitting the simulation models of *powsimR* and *muscat*. We evaluated all DE methods for 4, 8 and 16 samples in combination with 200, 1000 and 3000 cells per person. Additionally, we performed a comparison for a large range of parameter combinations of *powsimR* with edgeR-LRT and *muscat* with edgeR, testing all combinations as evaluated in **Figure 3A**.

In the following, it is important to distinguish the training data set, which is used for model fitting of *powsimR/muscat* and restricts the number of simulated genes, and the simulated data set which is sampled from the trained model.

The three main components of our statistical framework were evaluated in the comparison, the expression probability (by comparing the number of expressed genes), the power (here according to the definition of *powsimR*, i.e. the power of all genes expressed in the simulated data) and the overall detection power.

1. **Expressed genes:** The expected number of expressed genes for *scPower* is compared with the number of expressed genes in *powsimR* and *muscat*, which are all genes with at least one count in the simulated matrix. An important limitation of the simulation based frameworks is here that the number of expressed genes in the simulation tools can never be larger than the number of expressed genes in the training data set, while *scPower* can also approximate expression of unseen genes with smaller mean values and so estimate more expressed genes than seen in the pilot data.
2. **DE power:** The reported power of *powsimR* includes only genes, which are expressed in the simulated data set ( $\text{count} > 0$ ). The same value can also be calculated for *muscat*. To make the DE power of our framework comparable, the mean power for all

expressed DE genes was calculated. An expressed DE gene for scPower is defined by its expression rank, which needs to be smaller than the expected number of expressed genes.

3. **Overall power:** powsimR does not return directly an overall power, which we define as the power over all simulated DE genes (including genes simulated with count > 0 and count = 0). However, the overall detection power of powsimR can simply be calculated by dividing the number of true positives of powsimR by the number of all simulated DE genes. The same was done for muscat.

**powsimR:** uses training data to fit the parameters of the expression distributions for each cell type and gene. Using these parameters, it is randomly generating count matrices for two groups of cells introducing differential gene expression between these two groups for a prespecified number of DE genes. These DE genes are randomly selected and the means of their distributions shifted by a given effect size. In the next step the simulated data is analyzed with different methods and results are compared to the simulated group truth to determine the power.

Adaptations of powsimR are required to simulate a multisample setting and thus make it comparable to scPower: We added an additional step that generates a pseudobulk count matrix for multi-sample comparison. For this, we included an additional parameter for the sample size  $n_s$ , with samples distributed equally across both groups ( $\frac{n_s}{2}$  samples per group). Thus, individual level effect sizes are identical to the cell level effect sizes, as more complex differential distributions are not implemented in powsimR<sup>87</sup>. After simulation of the new count matrix  $C$  with dimensions  $n_C$  (number of cells) times  $n_G$  (number of genes) in powsimR, we changed the algorithm to equally distribute the simulated cells between the samples ( $\frac{n_C}{n_s}$  cells for each sample), while preserving the group structure. Summing up the counts for each sample generates a pseudobulk matrix with dimensions  $n_s$  times  $n_G$ , which can be processed exactly the same way as a single cell matrix in the following steps in powsimR. Furthermore, instead of randomly sampling the position of the DE genes with powsimR, we assigned DE genes based on their expression ranks in the bulk studies, as in scPower.

**muscat:** was specifically designed for multi-sample comparisons, in contrast to powsimR. It fits one negative binomial distribution separately for each sample and subpopulation in the training data set. The subpopulation definition here is equivalent to our cell type definition. We noticed that fitting each sample separately decreases the number of expressed genes quite drastically, if not a sufficient number of cells are available for each sample. This is again due to the fact that only as many genes can be sampled as are detected in the training data set. To get a robust fit of the negative binomial distribution with our training data set, we therefore decided to fit the negative binomial distribution for all samples together, for a very large training data set this is probably not necessary.

Another difference to powsimR is that muscat provides different scenarios for simulating differential expression besides the shift of the mean expression (called DE in muscat). Additionally, they simulate genes with different proportions of low and high expression-state (DP), differential modality (DM) or changes in both proportions and modality (DB). For comparison with powsimR and scPower, we focus on the DE scenario.

Similar to powsimR, we incorporated also here the option to assign genes of a specific expression rank a specific log fold change to simulate same DE genes are in scPower.

As no simulation-based power analysis for eQTLs exists (and also no other method), we benchmarked the eQTL power with our own simulation tool (described in the methods section **Power analysis for expression quantitative trait loci**). Our simulation method uses our expression probability model to estimate the mean parameter, therefore only the power itself is compared (not the expression probability and overall power). We tested again 25 rounds of simulation for all parameter combinations depicted in **Figure 3B**.

## Cost calculation and parameter optimization for a given budget

The overall experimental cost  $C_t$  for a 10X Genomics experiment is the sum of the library preparation cost and the sequencing cost. It can be calculated dependent on the three cost determining parameters sample size  $n_s$ , number of cells per sample  $n_c$  and the read depth  $r$ . The library preparation cost is determined by the number of 10X kits, depending on how many samples are loaded per lane  $n_{s,l}$  and the cost of one kit  $C_k$ . The cost of a flow cell  $C_f$  and the number of reads per flow cell  $r_f$  determine the sequencing cost.

$$C_t = \text{ceiling}\left(\frac{n_s}{6 * n_{s,l}}\right) * C_k + \text{ceiling}\left(\frac{n_s * n_c * r}{r_f}\right) * C_f$$

We optimized the three cost parameters for a fixed budget to maximize the detection power. A grid of values for number of cells per individual and for the read depth was tested, while the sample size is uniquely determined given the other two parameters and the fixed total costs. As an approximation of the sample size, the ceiling functions from the cost formula were removed.

$$n_s = C_t - \left(\frac{C_k}{6 * n_{s,l}} + \frac{n_c * r * C_f}{r_f}\right)$$

The same approach can also be used with a grid of sample size and cells per sample or read depth. In general, two parameters need to be chosen and the third parameter is uniquely determined given the other two and the fixed experimental cost.

Given the three cost parameters, the detection power for a specific cell type and a specific DE or eQTL study can be estimated. However, we also have to account for the appearance of doublets during the experiment. The fraction of doublets depends on the number of cells loaded on the lane. Following the approach of<sup>37</sup>, we model the doublet rate  $d$  linear dependent on the number of recovered cells, using the values from the 10X User guide of  $\sqrt{3}$ <sup>70</sup>. A factor of  $7.67 * 10^{-6}$  was estimated, so that  $d = 7.67 * 10^{-6} * n_c * n_{s,l}$ .

The number of usable cells per individual used for the calculation of detection power is then  $n_u = (1 - d) * n_c$ . We assume that nearly all doublets are detectable using Demuxlet and Scrublet and discarded during the preprocessing of the data set. The expected number of cells for the target cell type with a frequency of  $f_c$  will be  $f_c * (1 - d) * n_c$ .

A second effect of doublets is that the read distribution is shifted, as doublets contain more reads than singlets. Again following the approach of<sup>37</sup>, we assume that doublets contain 80% more reads than singlets. In the following, the ratio of reads in doublets compared to reads in singlets is called doublet factor  $f_d$ , a factor of 1.8 is assumed in the calculations in this manuscript. Therefore, depending on the number of doublets, the read depth of the singlets will be slightly lower than the target read depth.

$$r_s = \frac{r * n_c}{n_u + f_d * (n_c - n_u)}$$

Additionally, the mapping efficiency is taken into account. Assuming a mapping efficiency of 80%,  $r_m = 0.8 * r_s$  mapped read depth remains. In the power calculation, the number of usable cells per cell type will be used instead of the number of cells and the mapped read depth instead of the target read depth.

Instead of defining the number of samples per lane directly, usually the number of cells loaded per lane  $n_{c,l}$  is defined. So, the doublet rate per lane can be directly restricted. We use in our analyses  $n_{c,l} = 20,000$ , which leads to a doublet rate of at most 15.4%. The number of individuals per lane can be derived directly as  $n_{s,l} = \text{floor}\left(\frac{n_{c,l}}{n_c}\right)$ .

## Simulation of effect sizes and gene rank distributions

Model priors, i.e. effect sizes and gene rank distributions, were derived from FACS sorted bulk RNA-seq to get realistic assumptions. Additionally, we simulated different extreme prior distributions to evaluate their influence on the optimal experimental parameters. The log fold changes for the DE studies were modeled as normally distributed. High effect size distributions were simulated with a mean of 2 and a standard deviation of 1, low effect sizes distributions with a mean of 0.5 and standard deviation of 1.

Effect sizes ( $R^2$  values) for the eQTL studies were obtained by sampling normally distributed Z scores and applying the inverse Fisher Z Transformation. Because very small values are not observed due to the significance threshold, the normal distribution is truncated to retain values above the mean. High effect sizes were simulated with a mean of 0.5 and standard deviation of 0.2, low effect sizes with a mean of 0.2 and a standard deviation of 0.2. A similar standard deviation was also observed in the pilot data.

250 DEGs were simulated and 2000 eQTL genes. The ranks were uniformly distributed, either over the first 10,000 genes or the first 20,000 genes. This leads to four simulation scenarios for each, high and low effect sizes (ES) and high or uniformly distributed expression ranks, called in the studies `highES_highRank`, `lowES_highRank`, `highES_unifRank` and `lowES_unifRank`.

## Evaluation of Drop-seq and Smart-seq2 data

We validated our expression probability model for other tissues and single cell RNA-seq technologies. Two data sets of the human cell atlas were used for that, a Drop-seq data set measured in lung tissue<sup>52</sup> and a Smart-seq2 data set measured in pancreas tissue<sup>51</sup>.

The Drop-seq technology is also a droplet-based technique, similar to 10X Genomics. The same model can be used, only adapting the doublet and cost parameter. However, as there was no data available to model the linear increase of the doublet rate during overloading correctly, the doublet rate was modeled instead as a constant factor and the library preparation costs were estimated per cell. *scPower* provides models for both cases and with the necessary prior data, users can also model the overloading for Drop-seq.

Smart-seq2 is a plate-based technique, which produces full length transcripts and read counts instead of UMI counts. To compensate the gene length bias in the counts, the definition of an expressed gene was adapted to at least  $n$  counts per kilobase of transcript, resulting in a gene specific threshold of  $\frac{n * 1000}{l_i}$  with  $l_i$  as gene length for gene  $i$ . The gamma mixed distribution of the mean gene expression levels is modelled using length normalized counts, but the gene length is required as a prior for the dispersion estimation and the power calculation, as DEseq uses counts, which are not normalized for gene length. These priors can be obtained together with the effect sizes and the expression ranks from the pilot bulk studies. In the simulation of non-DE genes, an average mean length of 5,000 bp is assumed. The linear relationship of the parameters of the mixture of gamma distributions is modeled directly based on the mean number of reads per cell. Doublets also appear in Smart-seq2, but as a constant factor, not increasing with a higher number of cells per individual. We observed for the parameter of the DEseq dispersion model a linear relationship with the read depth, which was not visible for Drop-seq and 10X Genomics. So, instead of taking the mean value per cell type, a linear fit is modeled for Smart-seq2.

For both data sets, the cell type frequencies varied greatly among individuals, therefore an estimation of expressed genes in a certain fraction of individuals could not be validated, as this requires similar cell type frequencies for each donor. Instead, the expressed genes were estimated to be above a certain count threshold in all cells of a cell type, independent of the individual.

Both data sets were subsampled to investigate the effect of the read depth. The Drop-seq reads are subsampled using fastq-tools version 0.8<sup>84</sup> and the subsampled UMI count matrix was generated following the pipeline previously described in<sup>88</sup>. The Smart-seq2 read matrix was subsampled directly using the function *downsampleMatrix* of the package *DropletUtils*<sup>89</sup>.

We compared the budget restricted power to our PBMC 10X Genomics results, using the same simulated effect sizes and distribution ranks as well as matched observed priors from FACS sorted bulk studies.

## Frequency of the rarest cell type

The probability to detect at least  $n_{c,s}$  cells of a specific cell type  $c$  in each individual  $s$  depends on the frequency of the cell type  $f_c$ , the number of cells per individual  $n_c$  and the number of individuals  $n_s$ . For one individual, the minimal number of cells can be modeled using a cumulative negative binomial distribution<sup>37</sup> as  $F_{NB}(n_c - n_{c,s}, n_{c,s}, f_c)$  and for all individuals as  $F_{NB}(n_c - n_{c,s}, n_{c,s}, f_c)^{n_s}$ .

The cell type frequencies were obtained by literature research, the frequencies in PBMC are approximately twice as high as in whole blood<sup>90</sup>. All other parameters can be freely chosen (dependent on the expected study design).

## Declarations

### Ethics approval and consent to participate

All investigations have been carried out in accordance with the Declaration of Helsinki, including written informed consent of all participants. Study conduct complies with the recommendations by the ethics committee of the Bavarian Chamber of Physicians, Munich and the ethics committee of the Ludwig-Maximilian University, Munich. Applicable national and EU law, in particular the General Data Protection Regulation (GDPR) (Regulation (EU) 2016/679) has been followed.

Permission for using the data has been obtained from the Biobank of Max Planck Institute of Psychiatry. Consent for secondary use of the existing data has been obtained. In compliance with the consent for secondary use, the data generated in this project will be stored and can be used for future research. All data has been pseudonymized.

### Consent for publication

Written informed consent of all participants allows for publication of data in an access controlled repository.

### Availability of data and materials

The datasets generated and/or analysed during the current study are available in the European Genome Phenome Archive (EGA), accession number pending.

### Competing interests

FJT reports receiving consulting fees from Roche Diagnostics GmbH and Cellarity Inc., and ownership interest in Cellarity, Inc. and Dermagnostix. The other authors declare that they have no competing interests.

### Funding

HL is grateful for support by 'ExNet-0041-Phase2-3 („SyNergy-HMGU“)' through the Initiative and Network Fund of the Helmholtz Association. CC is supported by a Banting Postdoctoral Fellowship. FJT acknowledges support by the BMBF (grant# 01IS18036A and grant# 01IS18053A), by the Helmholtz Association (Incubator grant sparse2big, grant # ZT-I-0007) and by the Chan Zuckerberg Initiative DAF (advised fund of Silicon Valley Community Foundation, 182835). MH acknowledges support by the Chan Zuckerberg Foundation (CZF Grant #: CZF2019-002431).

### Authors' contributions

KTS and MH conceived the power analysis framework and analyzed the data. MH, FJT, EBB and HL designed the scRNA-seq experiment. EBB planned the BeCOME study and recruited the study participants. CC and AB generated scRNA-seq data in PBMCs. KTS and MH wrote the manuscript with input from all authors. All authors approved the final manuscript.

## Acknowledgements

We thank Thomas Walzthoeni for bioinformatics support provided at the Bioinformatics Core Facility, Institute of Computational Biology, Helmholtz Zentrum München. We thank Elisabeth Graf and Thomas Schwarzmayr for help in sequencing. We thank the BeCOME study team at the Max Planck Institute for Psychiatry, including the BioPrep core unit for their contribution to control individuals recruitment and characterizations, as well as collection of PBMCs. We thank Maren Büttner and Barbara Höllbacher for insightful discussion and proofreading of the manuscript.

## Bibliography

1. Khan, J. *et al.* Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays. *Cancer Res.* **58**, 5009–5013 (1998).
2. Debouck, C. & Goodfellow, P. N. DNA microarrays in drug discovery and development. *Nat. Genet.* **21**, 48–50 (1999).
3. Claverie, J. M. Computational methods for the identification of differential and coordinated gene expression. *Hum. Mol. Genet.* **8**, 1821–1832 (1999).
4. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
5. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
6. Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. Mapping complex disease traits with global gene expression. *Nat. Rev. Genet.* **10**, 184–194 (2009).
7. Civelek, M. & Lusis, A. J. Systems genetics approaches to understand complex traits. *Nat. Rev. Genet.* **15**, 34–48 (2014).
8. GTEx Consortium *et al.* Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
9. Aguet, F. *et al.* The GTEx Consortium atlas of genetic regulatory effects across human tissues. *bioRxiv* 787903 (2019) doi:10.1101/787903.
10. GTEx Consortium. The GTEx Consortium atlas of genetic regulatory effects across human tissues. *Science* **369**, 1318–1330 (2020).
11. Tang, F. *et al.* mRNA-Seq whole-transcriptome analysis of a single cell. *Nat. Methods* **6**,

- 377–382 (2009).
12. Stegle, O., Teichmann, S. A. & Marioni, J. C. Computational and analytical challenges in single-cell transcriptomics. *Nat. Rev. Genet.* **16**, 133–145 (2015).
  13. Angerer, P. *et al.* Single cells make big data: New challenges and opportunities in transcriptomics. *Current Opinion in Systems Biology* **4**, 85–91 (2017).
  14. Svensson, V., Vento-Tormo, R. & Teichmann, S. A. Exponential scaling of single-cell RNA-seq in the past decade. *Nat. Protoc.* **13**, 599–604 (2018).
  15. Stark, R., Grzelak, M. & Hadfield, J. RNA sequencing: the teenage years. *Nat. Rev. Genet.* **20**, 631–656 (2019).
  16. Kharchenko, P. V., Silberstein, L. & Scadden, D. T. Bayesian approach to single-cell differential expression analysis. *Nat. Methods* **11**, 740–742 (2014).
  17. 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).
  18. Soneson, C. & Robinson, M. D. Bias, robustness and scalability in single-cell differential expression analysis. *Nat. Methods* **15**, 255–261 (2018).
  19. Lähnemann, D. *et al.* Eleven grand challenges in single-cell data science. *Genome Biol.* **21**, 31 (2020).
  20. Cohen, J. *Statistical power analysis for the behavioral sciences*. (Hillsdale, 1989).
  21. Yang, Y. H. & Speed, T. P. Design and analysis of comparative microarray experiments. *Statistical analysis of gene expression microarray data* **35**, 91 (2003).
  22. Jung, S.-H. Sample size for FDR-control in microarray data analysis. *Bioinformatics* **21**, 3097–3104 (2005).
  23. Pounds, S. & Cheng, C. Sample size determination for the false discovery rate. *Bioinformatics* **21**, 4263–4271 (2005).
  24. Liu, P. & Hwang, J. T. G. Quick calculation for sample size while controlling false discovery rate with application to microarray analysis. *Bioinformatics* **23**, 739–746 (2007).

25. Hart, S. N., Therneau, T. M., Zhang, Y., Poland, G. A. & Kocher, J.-P. Calculating sample size estimates for RNA sequencing data. *J. Comput. Biol.* **20**, 970–978 (2013).
26. Li, C.-I. & Shyr, Y. Sample size calculation based on generalized linear models for differential expression analysis in RNA-seq data. *Stat. Appl. Genet. Mol. Biol.* **15**, 491–505 (2016).
27. van Iterson, M., van de Wiel, M. A., Boer, J. M. & de Menezes, R. X. General power and sample size calculations for high-dimensional genomic data. *Stat. Appl. Genet. Mol. Biol.* **12**, 449–467 (2013).
28. Busby, M. A., Stewart, C., Miller, C. A., Grzeda, K. R. & Marth, G. T. Scotty: a web tool for designing RNA-Seq experiments to measure differential gene expression. *Bioinformatics* **29**, 656–657 (2013).
29. Bi, R. & Liu, P. Sample size calculation while controlling false discovery rate for differential expression analysis with RNA-sequencing experiments. *BMC Bioinformatics* **17**, 146 (2016).
30. Ching, T., Huang, S. & Garmire, L. X. Power analysis and sample size estimation for RNA-Seq differential expression. *RNA* **20**, 1684–1696 (2014).
31. Wu, H., Wang, C. & Wu, Z. PROPER: comprehensive power evaluation for differential expression using RNA-seq. *Bioinformatics* **31**, 233–241 (2015).
32. Poplawski, A. & Binder, H. Feasibility of sample size calculation for RNA-seq studies. *Brief. Bioinform.* **19**, 713–720 (2018).
33. Monaco, G. *et al.* RNA-Seq Signatures Normalized by mRNA Abundance Allow Absolute Deconvolution of Human Immune Cell Types. *Cell Rep.* **26**, 1627–1640.e7 (2019).
34. Wu, A. R. *et al.* Quantitative assessment of single-cell RNA-sequencing methods. *Nat. Methods* **11**, 41–46 (2014).
35. Svensson, V. *et al.* Power analysis of single-cell RNA-sequencing experiments. *Nat. Methods* **14**, 381–387 (2017).
36. Ziegenhain, C. *et al.* Comparative Analysis of Single-Cell RNA Sequencing Methods.

- Mol. Cell* **65**, 631–643.e4 (2017).
37. Hafemeister, C. How Many Cells. <https://satijalab.org/howmanycells> (2019).
  38. Abrams, D., Kumar, P., Karuturi, R. K. M. & George, J. A computational method to aid the design and analysis of single cell RNA-seq experiments for cell type identification. *BMC Bioinformatics* **20**, 275 (2019).
  39. Davis, A., Gao, R. & Navin, N. E. SCOPIT: sample size calculations for single-cell sequencing experiments. *BMC Bioinformatics* **20**, 566 (2019).
  40. Vieth, B., Ziegenhain, C., Parekh, S., Enard, W. & Hellmann, I. powsimR: power analysis for bulk and single cell RNA-seq experiments. *Bioinformatics* **33**, 3486–3488 (2017).
  41. Crowell, H. L. *et al.* On the discovery of population-specific state transitions from multi-sample multi-condition single-cell RNA sequencing data. *bioRxiv* 713412 (2019) doi:10.1101/713412.
  42. Li, W. V. & Li, J. J. A statistical simulator scDesign for rational scRNA-seq experimental design. *Bioinformatics* **35**, i41–i50 (2019).
  43. Su, K., Wu, Z. & Wu, H. Simulation, power evaluation and sample size recommendation for single-cell RNA-seq. *Bioinformatics* **36**, 4860–4868 (2020).
  44. 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).
  45. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
  46. Law, C. W., Chen, Y., Shi, W. & Smyth, G. K. voom: Precision weights unlock linear model analysis tools for RNA-seq read counts. *Genome Biol.* **15**, R29 (2014).
  47. Kang, H. M. *et al.* Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. *Nat. Biotechnol.* **36**, 89–94 (2018).
  48. Sarkar, A. K. *et al.* Discovery and characterization of variance QTLs in human induced pluripotent stem cells. *PLoS Genet.* **15**, e1008045 (2019).

49. Cuomo, A. S. E. *et al.* Publisher Correction: Single-cell RNA-sequencing of differentiating iPS cells reveals dynamic genetic effects on gene expression. *Nat. Commun.* **11**, 1572 (2020).
50. Mandric, I. *et al.* Optimized design of single-cell RNA sequencing experiments for cell-type-specific eQTL analysis. *Nat. Commun.* **11**, 5504 (2020).
51. Enge, M. *et al.* Single-Cell Analysis of Human Pancreas Reveals Transcriptional Signatures of Aging and Somatic Mutation Patterns. *Cell* vol. 171 321–330.e14 (2017).
52. Vieira Braga, F. A. *et al.* A cellular census of human lungs identifies novel cell states in health and in asthma. *Nat. Med.* **25**, 1153–1163 (2019).
53. Rendeiro, A. F. *et al.* Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat. Commun.* **7**, 11938 (2016).
54. Moreno-Moral, A. *et al.* Changes in macrophage transcriptome associate with systemic sclerosis and mediate GSDMA contribution to disease risk. *Ann. Rheum. Dis.* **77**, 596–601 (2018).
55. Nicodemus-Johnson, J. *et al.* DNA methylation in lung cells is associated with asthma endotypes and genetic risk. *JCI Insight* **1**, e90151 (2016).
56. Arda, H. E. *et al.* Age-Dependent Pancreatic Gene Regulation Reveals Mechanisms Governing Human  $\beta$  Cell Function. *Cell Metab.* **23**, 909–920 (2016).
57. Chen, L. *et al.* Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell* **167**, 1398–1414.e24 (2016).
58. Wolf, F. A. *et al.* PAGA: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome Biol.* **20**, 59 (2019).
59. Baran, Y. *et al.* MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).
60. Anders, S. & Huber, W. Differential expression analysis for sequence count data. *Genome Biol.* **11**, R106 (2010).
61. Regev, A. *et al.* The Human Cell Atlas. *Elife* **6**, (2017).

62. Jaakkola, M. K., Seyednasrollah, F., Mehmood, A. & Elo, L. L. Comparison of methods to detect differentially expressed genes between single-cell populations. *Brief. Bioinform.* **18**, 735–743 (2017).
63. Wang, T., Li, B., Nelson, C. E. & Nabavi, S. Comparative analysis of differential gene expression analysis tools for single-cell RNA sequencing data. *BMC Bioinformatics* **20**, 40 (2019).
64. Luecken, M. D. & Theis, F. J. Current best practices in single-cell RNA-seq analysis: a tutorial. *Mol. Syst. Biol.* **15**, e8746 (2019).
65. Zhu, H. & Lakkis, H. Sample size calculation for comparing two negative binomial rates. *Stat. Med.* **33**, 376–387 (2014).
66. Dunn, O. J. Multiple Comparisons among Means. *J. Am. Stat. Assoc.* **56**, 52–64 (1961).
67. GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.* **45**, 580–585 (2013).
68. Lappalainen, T. *et al.* Transcriptome and genome sequencing uncovers functional variation in humans. *Nature* **501**, 506–511 (2013).
69. Wolock, S. L., Lopez, R. & Klein, A. M. Scrublet: Computational Identification of Cell Doublets in Single-Cell Transcriptomic Data. *Cell Syst* **8**, 281–291.e9 (2019).
70. 10X Genomics. User Guides - 10x Genomics. *10x Genomics* <https://www.10xgenomics.com/resources/user-guides/> (2019).
71. van der Wijst, M. G. P. *et al.* Single-cell RNA sequencing identifies celltype-specific cis-eQTLs and co-expression QTLs. *Nat. Genet.* **50**, 493–497 (2018).
72. Heinrich, V. *et al.* The allele distribution in next-generation sequencing data sets is accurately described as the result of a stochastic branching process. *Nucleic Acids Res.* **40**, 2426–2431 (2012).
73. Lafzi, A., Moutinho, C., Picelli, S. & Heyn, H. Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. *Nat. Protoc.* **13**, 2742–2757 (2018).
74. 10x Genomics. What is the recommended sequencing depth for Single Cell 3' and 5' Gene Expression libraries? *10X Genomics* <https://kb.10xgenomics.com/hc/en->

- us/articles/115002022743-What-is-the-recommended-sequencing-depth-for-Single-Cell-3-and-5-Gene-Expression-libraries- (2020).
75. Heimberg, G., Bhatnagar, R., El-Samad, H. & Thomson, M. Low Dimensionality in Gene Expression Data Enables the Accurate Extraction of Transcriptional Programs from Shallow Sequencing. *Cell Syst* **2**, 239–250 (2016).
  76. Bourgon, R., Gentleman, R. & Huber, W. Independent filtering increases detection power for high-throughput experiments. *Proc. Natl. Acad. Sci. U. S. A.* **107**, 9546–9551 (2010).
  77. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat. Biotechnol.* **32**, 903–914 (2014).
  78. Regev, A. *et al.* The Human Cell Atlas White Paper. *arXiv [q-bio.TO]* (2018).
  79. Zheng, G. X. Y. *et al.* Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
  80. Samuel L. Wolock Romain Lopez, A. M. K. Scrublet: computational identification of cell doublets in single-cell transcriptomic data. *bioRxiv* 1–18 (2018).
  81. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 15 (2018).
  82. Preprocessing and clustering 3k PBMCs — Scanpy documentation. <https://scanpy-tutorials.readthedocs.io/en/latest/pbmc3k.html>.
  83. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* vol. 2008 P10008 (2008).
  84. fastq-tools. <https://homes.cs.washington.edu/~dcjones/fastq-tools/>.
  85. Zappia, L., Phipson, B. & Oshlack, A. Splatter: Simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 1–15 (2017).
  86. Lyles, R. H., Lin, H.-M. & Williamson, J. M. A practical approach to computing power for generalized linear models with nominal, count, or ordinal responses. *Stat. Med.* **26**,

- 1632–1648 (2007).
87. Korthauer, K. D. *et al.* A statistical approach for identifying differential distributions in single-cell RNA-seq experiments. *Genome Biology* vol. 17 (2016).
  88. Macosko, E. Z. *et al.* Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell* **161**, 1202–1214 (2015).
  89. Lun, A. T. L. *et al.* EmptyDrops: distinguishing cells from empty droplets in droplet-based single-cell RNA sequencing data. *Genome Biol.* **20**, 63 (2019).
  90. Bio-Rad. Cell frequencies in common samples - Flow Cytometry analysis | Bio-Rad. *Bio-Rad* <https://www.bio-rad-antibodies.com/flow-cytometry-cell-frequency.html>.
  91. Burnham, K. P., Anderson, D. R. & Huyvaert, K. P. AIC model selection and multimodel inference in behavioral ecology: some background, observations, and comparisons. *Behav. Ecol. Sociobiol.* **65**, 23–35 (2011).

# Figures

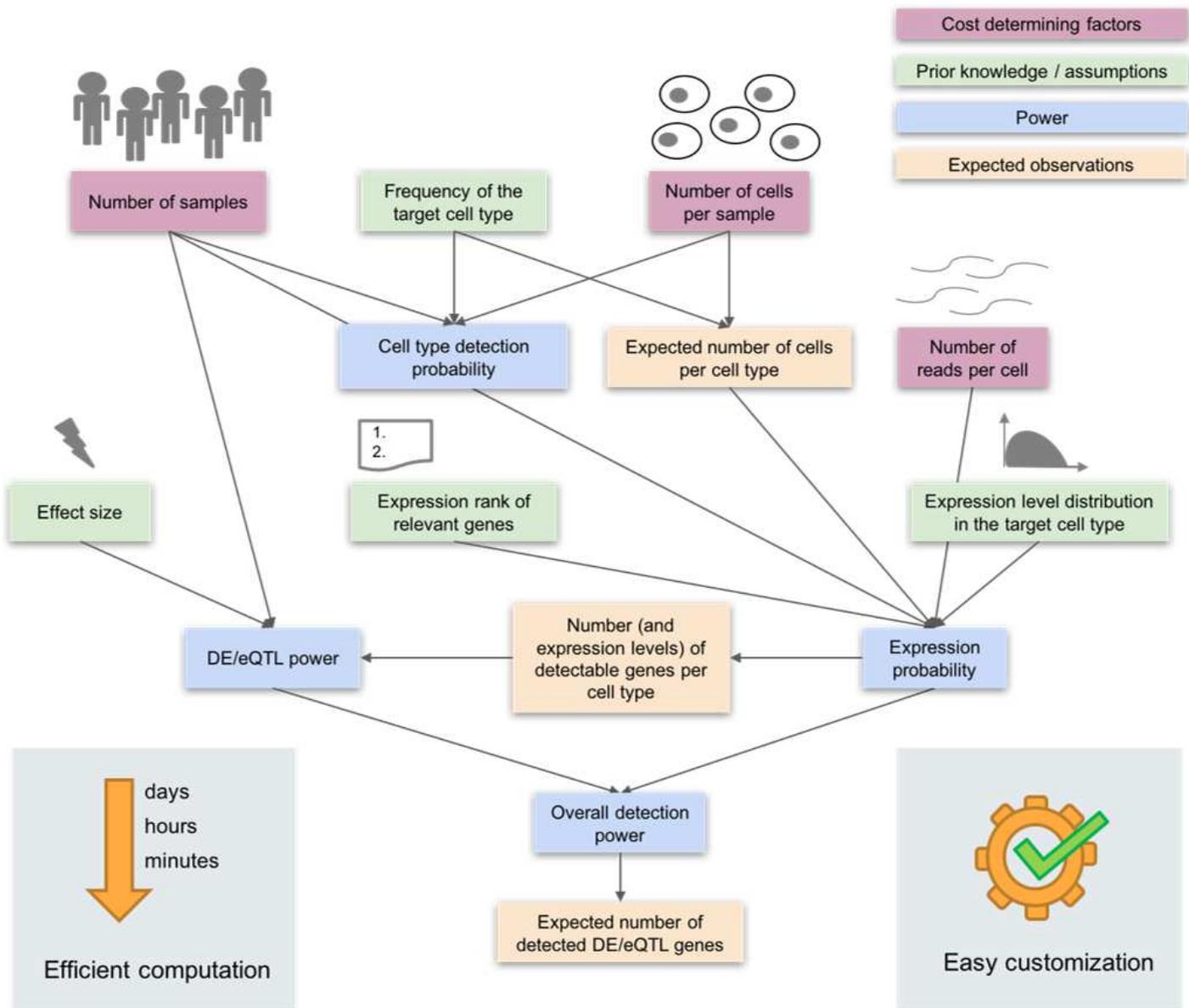
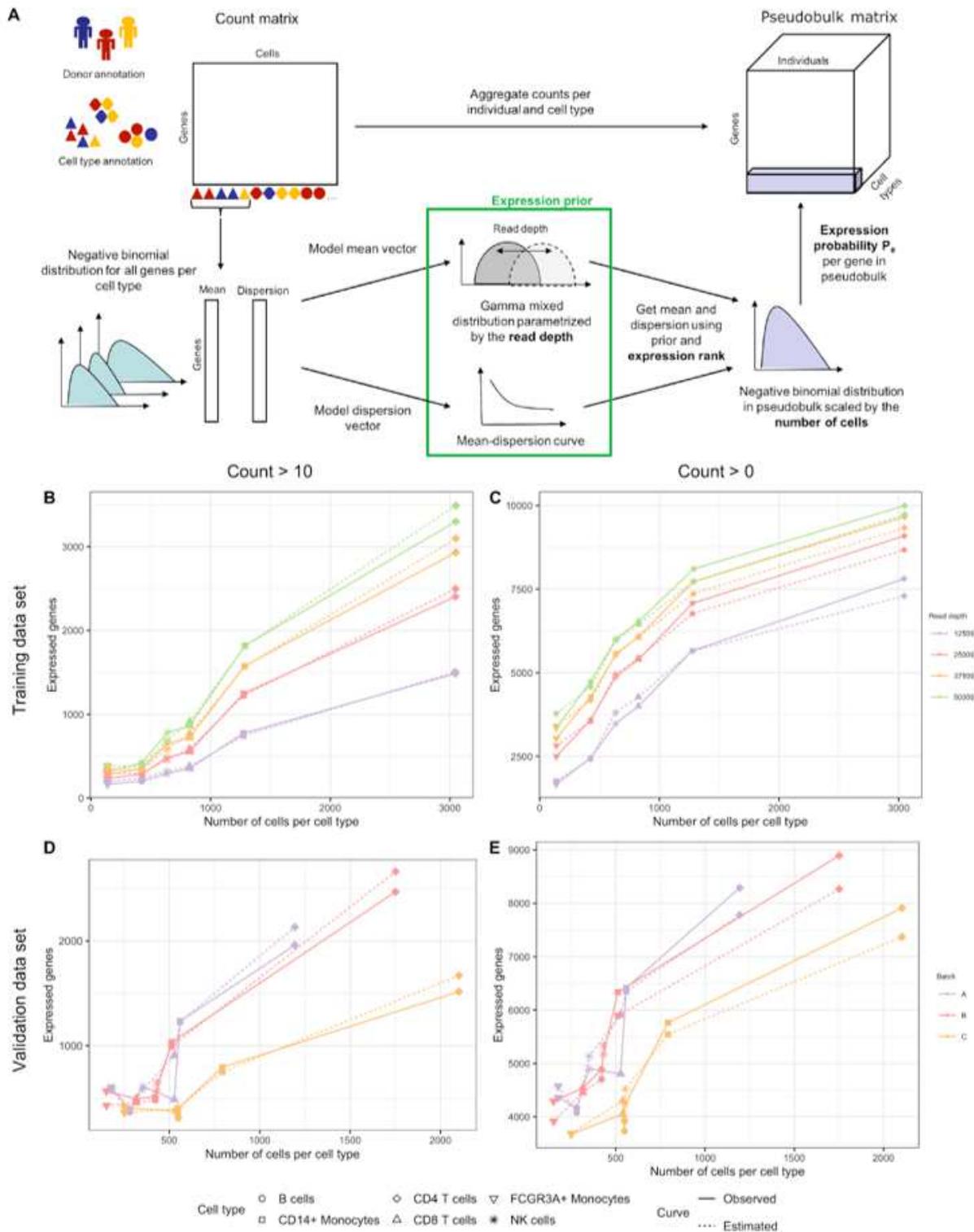


Figure 1

Dependence of experimental design parameters. The cost determining factors (purple: number of samples, number of cells per sample and number of reads per cell) are related to detection power (blue) and expected number of observations (orange). In addition, power and expected observations also depend on prior knowledge or assumptions (green). Our model enables fast power calculation, independent of the chosen experimental parameters, and easy adaptation to different use-cases through reference priors.



**Figure 2**

Expression probability model parameterized by UMI counts per cell. Panel A visualizes how the count distribution in the pseudobulk is modelled using negative binomial distributions with mean and dispersion parameters derived from cell type specific expression priors. Panel B and C show the observed number of expressed genes (solid line) and the number of expressed genes expected under our model (dashed line) on the y-axis and the number of cells per cell type (cell type indicated by the point symbol)

on the x-axis for one batch of the trainings PBMC data set (Supplementary Table S2). The data is subsampled to different read depths (indicated by the color). Panel D and E show expressed genes for the three batches of an independent validation PBMC data set 47. A gene is called expressed with count > 10 (B,D) or count > 0 (C,E) in more than 50% of the individuals.

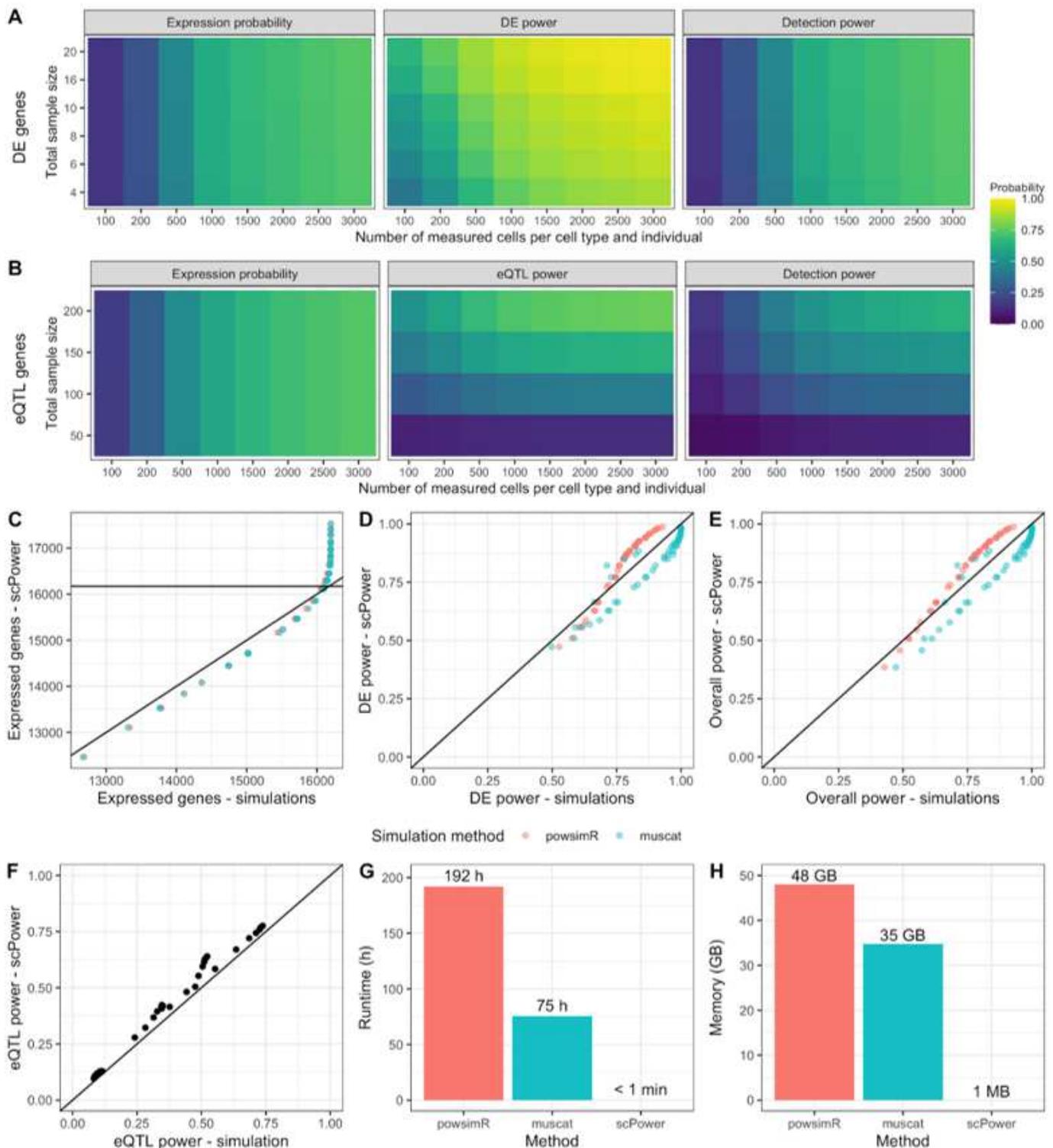
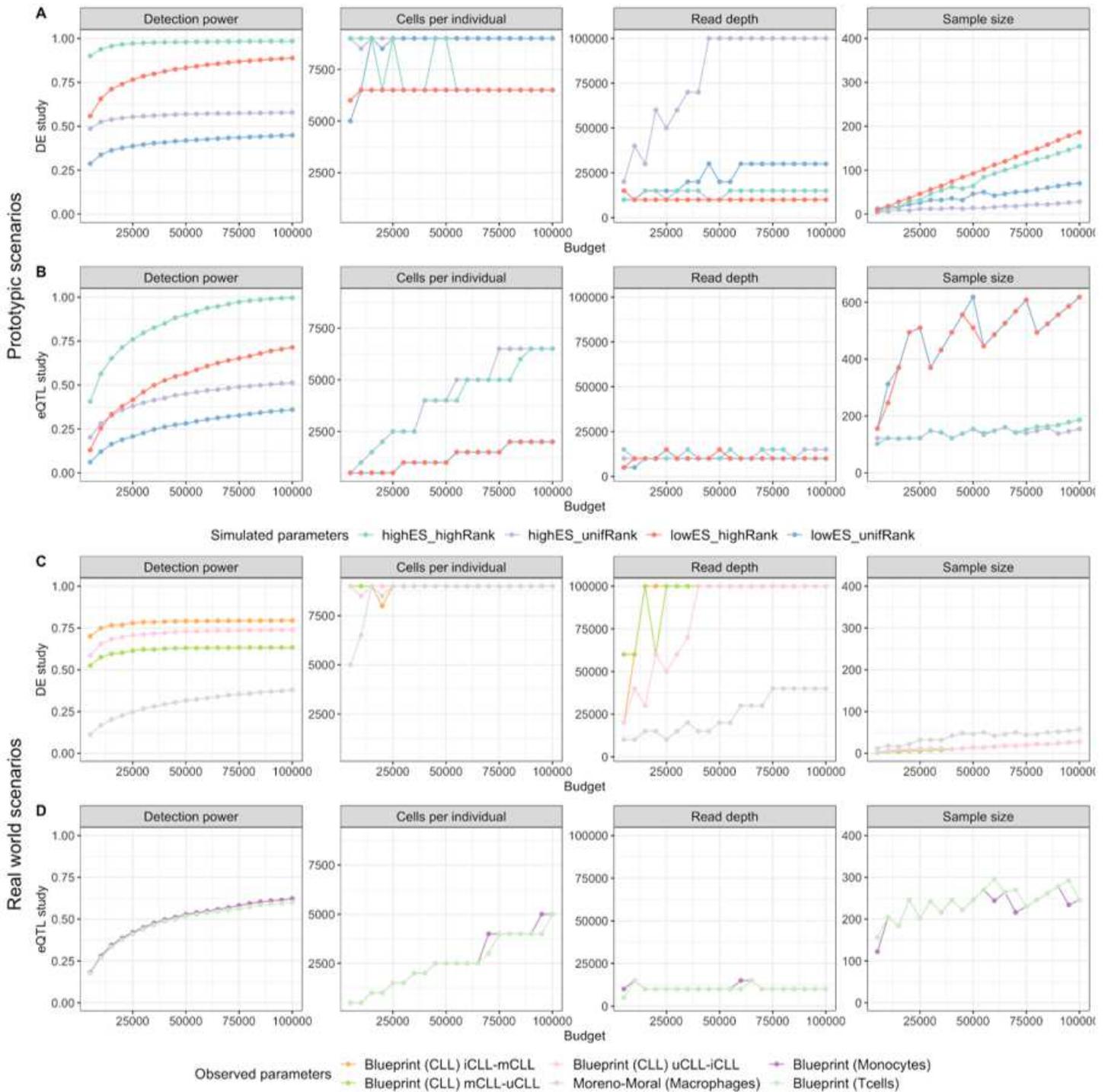


Figure 3

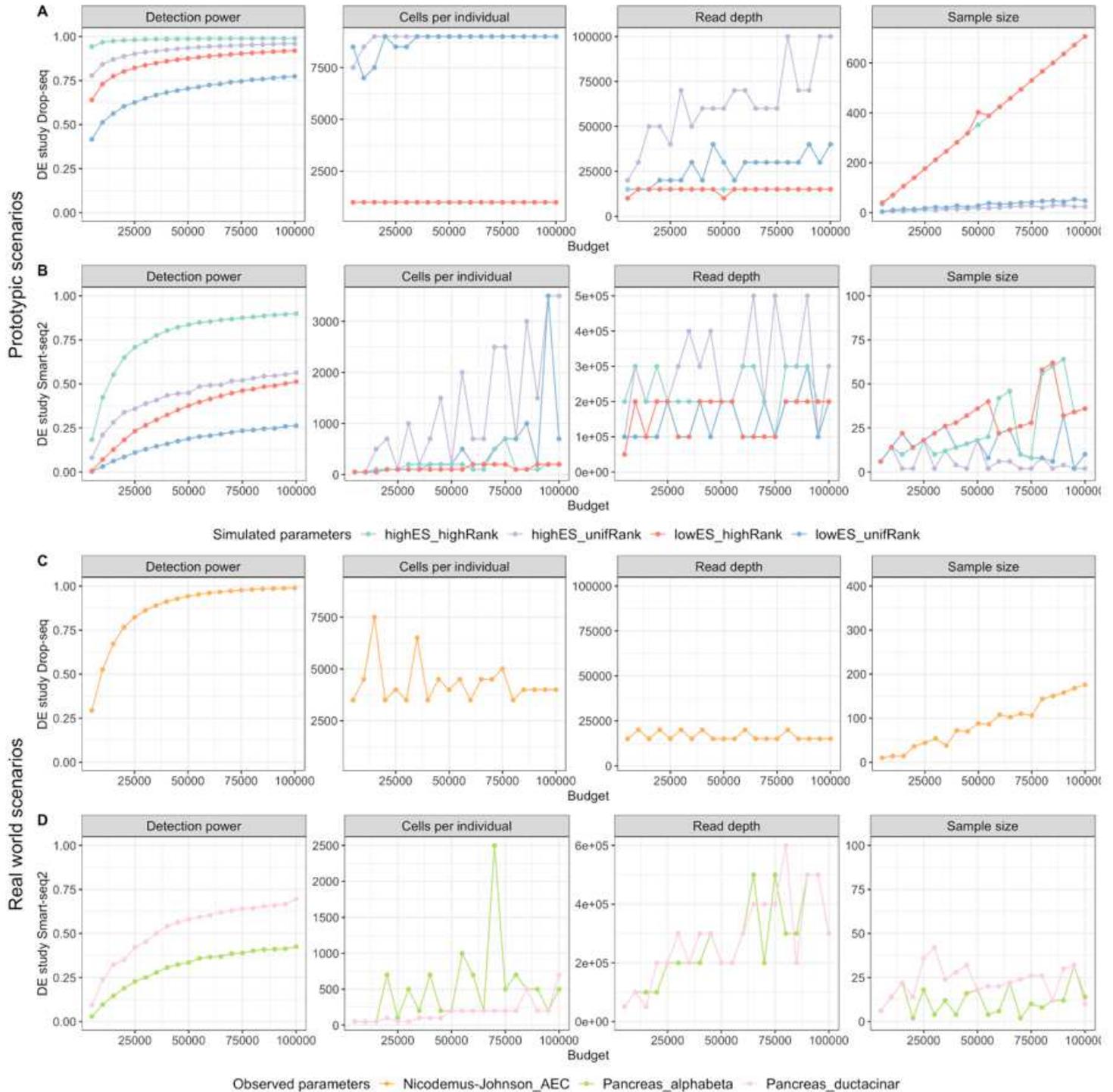
Expression probability, DE/eQTL power and overall detection power and their validation in simulation studies. Power estimation using data driven priors for A. DE genes and B. eQTL genes dependent on the total sample size and the number of measured cells per cell type. The detection power is the product of the expression probability and the power to detect the genes as DE or eQTL genes, respectively. The fold change for DEGs and the R<sup>2</sup> for eQTL genes were taken from published studies, together with the expression rank of the genes. For A, the Blueprint CLL study with comparison iCLL vs mCLL was used, for B, the Blueprint T cell study. The expression profile and expression probabilities in a single cell experiment with a specific number of samples and measured cells was estimated using our expression prior, setting the definition for expressed to > 10 counts in more than 50% of the individuals. Multiple testing correction was performed by using FDR adjusted p-values for DE power and FWER adjusted p-values for eQTL power. The probabilities calculated in A were verified by the simulation-based methods powsimR and muscat (C-E) with each point representing one parameter combination. Also the eQTL power of B could be replicated with a self-implemented simulation (F). Runtime (G) and memory requirements (H) were drastically higher in the simulations than for our tool scPower during the evaluations of C-E, showing the clear strength of our analytic model.



**Figure 4**

Optimal parameters for varying budgets and 10X Genomics data. The figure shows the maximal reachable detection power (y-axis, first column) for a given experimental budget (x-axis) and the corresponding optimal parameter combinations for that budget (y-axis, second till fourth column). The colored lines indicate different effect sizes and gene expression rank distributions. Panels A-B visualize different simulated effect sizes and rank distributions (simulation names see text) for DEG studies (A) and eQTL studies (B) with models fitted on 10X PBMC data. Panels C-D visualize effect sizes and rank

distributions observed in cell type sorted bulk RNA-seq DEG studies (C) and eQTL studies (D) with model fits analogously to A-B.



**Figure 5**

Optimal parameters for varying budgets and Drop-seq and Smart-seq2 data. The figure shows the maximal reachable detection power (y-axis, first column) for a given experimental budget (x-axis) and the corresponding optimal parameter combinations for that budget (y-axis, second till fourth column). The colored lines indicate different effect sizes and gene expression rank distributions. Panels A-B visualize

different simulated effect sizes and rank distributions (simulation names see text) for DE studies with models fitted on Drop-seq lung data (A) and Smart-seq2 pancreas data (B). Panels C-D visualize effect sizes and rank distributions observed in cell type sorted bulk RNA-seq DE studies with model fits analogously to A-B.

## Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [ExperimentalDesignSupplement.pdf](#)