Sensei: Sample size determination for comparing proportion of cells under different conditions

Shaoheng Liang, Jinzhuang Dou, Chenjie Lyu, Ken Chen

# Abstract

## Background

Single-cell RNA sequencing (scRNA-seq) allows for studying compositional changes of cell types at the cell level. Determining number of samples to be gathered, and number of cells in each sample to be sequenced, is critical to such studies. The estimation should be based on limited prior knowledge of what portion the cell type may take in each sample.

## Results

We developed an interactive web application called Sensei, which calculates the statistical power of a scRNA-seq study design. Using the aforementioned prior knowledge, the tool models the proportion of cells in a type of interest through beta binomial conjugacy and calculates the statistical power for t-test by normal approximation.

## Conclusions

Using conjugate prior and approximation, our tool generates a table to illustrate the relationship of sample size and statistical power to help researchers design a study. The web application can be accessed at <https://kchen-lab.github.io/sensei/table_beta.html>.

# Background

Biological samples consist of various types of cells. The proportion taken by a cell type may vary as a result of various physiological and pathological conditions. These changes, if observed, can inspire biological discoveries. For example, more T cells are found in colon tumors resulting from Lynch syndrome than sporadic ones, and the difference suggests different levels of tumor infiltration [1]. This type of studies, known as compositional analysis, have been done on cytometry [2] and microbiome [3] datasets, and is on the rise for scRNA-seq data [4, 5]. To confirm such differences, samples from patients need to be gathered and sequenced. However, it can be costly to recruit participants and sequencing, and even more so when an insufficient number of samples lead to a false negative result. Thus, the sample size should be determined in advance.

For single-cell study, tools like SCOPIT [6] and howmanycells (<https://satijalab.org/howmanycells>) have been developed to estimate how many cells in total should be sequenced to obtain desired abundance a given cell type. However, to study changes of proportions of cells, an estimation of number of participants is also needed. Here, we develop an interactive web tool, Sensei, which rapidly calculates statistical power (or, equivalently, false negative rate) for a set of study designs, to assist researcher in planning experiments, using beta-binomial conjugacy and normal approximation. Beta-binomial has been used in analysis of proportion [7, 8].

# Implementation

The statistical power estimation is based on a t-test, which is widely used in comparing two samples that are roughly normally distributed. Naturally, a single-cell study design includes number of samples (participants) in each (control “” or experimental “”) group, and number of single cells in each sample . We also need an estimation of the proportion of the cell type of interest in each group, and the standard deviation of it as a result of biological variances. Finally, a significant level (false positive rate) should be assigned, to calculate the false negative rate , or, equivalently, the statistical power . The required parameters are shown in Table 1.

|  |  |  |
| --- | --- | --- |
| Table 1. Required parameters | | |
| Parameter | Notation | Annotation |
| Number of samples under condition |  |  |
| Number of cells in each sample |  |  |
| Mean and standard deviation of proportions |  | Parameters for beta distribution. |
| False positive rate |  | Significant value. |

We assume that t-test is used, where the null hypothesis is

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | . | (1) |

for two-sided test, where denotes the cell proportion in sample from group . We assume that in single-cell sequencing, the probability of a cell being of the type of interest, is . For the th sample in group , The total number of such cells is a random variable . it self is a random variable, which represents the variances among samples in the same group. It is natural to then model with , which is conveniently defined on . The and can be reparametrized from the user-defined mean and standard deviation and . Formally,

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | . | (2) |

Practically, we require that the resulting and to be both greater than to confine the beta distribution to be of unimodal. Because beta distribution is a conjugate prior of binomial distribution, , whose mean and variance are

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | . | (3) |

The proportion is simply normalized by the total number of cells . Thus,

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | . | (4) |

We now assume that the beta binomial distribution can be approximated by a normal distribution, awhere

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | . | (5) |

To justify such an approximation, we compare the two distributions under different conditions. In specific, Fig. 1a shows that the L1 distance between a beta binomial distribution is smaller when is closer to 0.5, is smaller, and is larger. Fig. 1b further shows two representative examples. For , , the underlying beta distribution is already skewed to the left and deviates from normal distribution, this results in a slightly unprecise, but still largely acceptable normal approximation. For , , as the beta distribution itself is visually like a normal distribution, and the generated can be perfectly approximated by a normal distribution. Note that we only allow the unimodal subset of beta distribution to be used.

|  |  |
| --- | --- |
| a | b |
|  |  |
| **Fig. 1** Comparison of beta binomial distribution and the normal estimation.  (a) L1 distance between beta distribution and normal distribution. Subpanel: number of cells. Color: mean; X-axis: standard deviation; Y-axis: L1 distance between cumulative distribution function (CDF) of a beta binomial distribution and its normal approximation.  (b) Four representative cases. First row: corresponding beta distribution. Second row and third row: CDF for four cases. | |

For a t-test with possibly different variances in two samples [9], the t-value is

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | , | (6) |

Thus, the false negative rate is

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | , | (7) |

where , as , for a two-sided test, or for a one-sided test, where

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | . | (8) |

which is equivalent to when and . For a fast estimation, we do not account for the variability brought by estimating and from samples in the and in the . This is a common practice in power estimation [10]. A more precise estimation can only be done by sampling, which is too costly for a web application. Nevertheless, we provide MATLAB scripts for more comprehensive estimations.

# Results

## Sensei - the web application

We created a neat standalone web application using HTML and JavaScript. It runs purely on the front-end, and no link to any server is needed. The tool is based on jstat, a JavaScript statistics library, and ploty, a plotting library. A screen shot is shown in Fig. 2. Sensei asks for the parameters detailed above, and an additional cutoff for the false negative rate .

|  |  |
| --- | --- |
| a | b |
|  | |
| c | |
|  | |
| **Fig. 2** Screenshot of Sensei.  (a, b) The table for negative rate under various and in one-sided and two-sided t-tests. Shaded entries correspond to . Parameters are set to . This is the main output of the application.  (b) The visualization for the beta distributions in (a). This is given when a user sets the parameters for the distribution of as a visual confirmation of what is desired. | |

## Generating a table for various study designs

To further help researcher choose a study design, we allow user to set a range for feasible number of samples. Sensei then rapidly calculate false negative rate for all pairs of feasible and . Fig. 2 shows an example where we set . Two tables are generated to show for each pair of and for one-sided test and two-sided test. This presentation allows for comparison among multiple combinations of and , where there is a trade-off in between. For example, in Fig. 2a, to obtain , the researcher may gather either control samples and experimental samples, or control samples and experimental samples. The final decision will be based on the difference in the cost of control and experimental samples.

## Reparametrizing and visualizing beta distributions

Sensei models the distribution of the portion of cells using beta distribution. To give the user a more intuitive interface, Sensei supports both parametrization of beta distribution, i.e., and , or and . User may use either set and the other set will be updated accordingly. The beta distribution is further visualized (Fig. 2b) to help the user determine if the shape is desired.

## Simulation verification

Because several approximation was made in the derivation, to further verify the program, we perform a simulation study for , the same as that in Fig. 2. The results are shown in Table 2 and the difference from Fig. 2b is shown in table 3. The difference is generally small. In fact, the error is mostly from ignoring the variance in the variance term in the t-test, which is a common practice [10].

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2. Simulation results for two-sided test | | | | | | | | |
|  | Experimental | | | | | | | |
| Control | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 5 | 0.4540 | 0.4299 | 0.4213 | 0.4162 | 0.4007 | 0.3953 | 0.4040 | 0.4025 |
| 6 | 0.3898 | 0.3596 | 0.3411 | 0.3289 | 0.3233 | 0.3161 | 0.3099 | 0.3126 |
| 7 | 0.3347 | 0.3012 | 0.2702 | 0.2509 | 0.2543 | 0.2331 | 0.2311 | 0.2395 |
| 8 | 0.2789 | 0.2513 | 0.2204 | 0.2115 | 0.2027 | 0.1920 | 0.1835 | 0.1844 |
| 9 | 0.2568 | 0.2158 | 0.1889 | 0.1691 | 0.1581 | 0.1555 | 0.1478 | 0.1420 |
| 10 | 0.2257 | 0.1926 | 0.1574 | 0.1390 | 0.1304 | 0.1251 | 0.1087 | 0.1099 |
| 11 | 0.2103 | 0.1617 | 0.1417 | 0.1083 | 0.1020 | 0.0981 | 0.0907 | 0.0788 |
| 12 | 0.1894 | 0.1408 | 0.1077 | 0.1000 | 0.0863 | 0.0734 | 0.0715 | 0.0626 |

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table 2. Simulation results for two-sided test | | | | | | | | |
|  | Experimental | | | | | | | |
| Control | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 5 | 0.0062 | 0.0095 | 0.0060 | 0.0037 | 0.0144 | 0.0167 | 0.0060 | 0.0061 |
| 6 | -0.0248 | -0.0252 | -0.0260 | -0.0269 | -0.0308 | -0.0306 | -0.0298 | -0.0368 |
| 7 | -0.0338 | -0.0366 | -0.0289 | -0.0258 | -0.0410 | -0.0288 | -0.0338 | -0.0479 |
| 8 | -0.0234 | -0.0356 | -0.0301 | -0.0388 | -0.0428 | -0.0418 | -0.0410 | -0.0480 |
| 9 | -0.0348 | -0.0358 | -0.0352 | -0.0334 | -0.0354 | -0.0426 | -0.0426 | -0.0429 |
| 10 | -0.0293 | -0.0394 | -0.0308 | -0.0303 | -0.0344 | -0.0386 | -0.0295 | -0.0365 |
| 11 | -0.0339 | -0.0291 | -0.0356 | -0.0196 | -0.0256 | -0.0307 | -0.0302 | -0.0236 |
| 12 | -0.0290 | -0.0245 | -0.0174 | -0.0266 | -0.0245 | -0.0201 | -0.0245 | -0.0205 |

# Discussion

Sensei estimates the participants needed in a single-cell study in order to obtain a statistical meaningful comparison of proportion of a cell type between two samples. Although multiple tools are available to estimate number of cells needed [11]. Sensei is by far the first tool we know that estimates the number of participants. Although necessary approximations are used in order to fit Sensei in a light-weight web application, its estimation is consistent with the result of a computationally intense simulation.

In some cases, researches may want to test if the portion has no meaningful difference between two samples. Such a test can be performed by using two t-tests showing that the portion is confined in a small range that has no biological meaning. This tool can also be used in such cases as t-test is still the basis.

Composition effect may confound the differential abundance analysis. Although as a rough estimation based on limited prior knowledge, Sensei may not be influenced as much, we suggest that user to exclude the highly variable cell types that introduced the composition effect when estimating the number of cells in a sample, namely . For example, if number of tumor cells is highly variable, and normally takes around 1000 ~ 3000 cells in a sample with 5000 cells, we suggest that a user use . Here, there are effectively three types of cells, i.e., cell of interest, reliable background cells, and highly variable cells. To more precisely model the three types of cells, Dirichlet-multinomial distribution may be useful. It is also possible to extend the model by using finite mixture model [8]. However, our current model allows for a closed-form representation, which is essential to a light-weight web-based application providing a fast estimation of sample sizes.

It also should be noted that there are different kinds of t-tests, and the one factors in the different variances of two samples should be used when the variances are indeed dissimilar [9].

# Conclusions

This study reports a user-friendly web application for estimating sample size and statistical power to help researchers design a single-cell study. We expect that Sensei will have applications in different scRNA studies involving differential abundance analysis.

# Bibliography

1. Lynch HT, Drescher KM, de la Chapelle A. Immunology and the Lynch Syndrome. Gastroenterology. 2008;134:1246–9.

2. Lun ATL, Richard AC, Marioni JC. Testing for differential abundance in mass cytometry data. Nat Methods. 2017;14:707–9.

3. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome Datasets Are Compositional: And This Is Not Optional. Front Microbiol. 2017;8. doi:10.3389/fmicb.2017.02224.

4. Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. Mol Syst Biol. 2019;15:e8746.

5. Cao Y, Lin Y, Ormerod JT, Yang P, Yang JYH, Lo KK. scDC: single cell differential composition analysis. BMC Bioinformatics. 2019;20:721.

6. Davis A, Gao R, Navin NE. SCOPIT: sample size calculations for single-cell sequencing experiments. BMC Bioinformatics. 2019;20:566.

7. Crowder MJ. Beta-Binomial Anova for Proportions. J R Stat Soc Ser C Appl Stat. 1978;27:34–7.

8. Brooks SP, Morgan BJT, Ridout MS, Pack SE. Finite Mixture Models for Proportions. Biometrics. 1997;53:1097–115.

9. Ruxton GD. The unequal variance t-test is an underused alternative to Student’s t-test and the Mann–Whitney U test. Behav Ecol. 2006;17:688–90.

10. Livingston EH, Cassidy L. Statistical Power and Estimation of the Number of Required Subjects for a Study Based on the t-Test: A Surgeon’s Primer. J Surg Res. 2005;126:149–59.

11. Plotly Technologies Inc. Collaborative data science. 2015. https://plot.ly.