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

# 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, T cells may take different proportions in Lynch syndrome and sporadic colon tumors, and the difference may suggest different level of T-cell infiltration. To confirm such differences, samples from patients need to be gathered and sequenced. However, it can be costly to recruit participants and sequencing, but even more so when an insufficient number of samples lead to a false negative result. 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.

# 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 is natural to then model using the conjugate prior of binomial distribution, i.e., . 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. Using the properties of beta binomial distribution, we can get

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

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

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

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

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

Fig. 1 show why such approximation is justifiable. 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.

|  |  |
| --- | --- |
| 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. Otherwise: CDF for four cases. | |

For a t-test, the t-value is

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

where

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | . | (7) |

Thus, the false negative rate is

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | , | (8) |

where , as , for a two-sided test. Similarly, for a one-sided test, .

# Results

|  |
| --- |
| a |
|  |
| b |
|  |
| **Fig. 2** Screenshot of Sensei.  (a) The one-sided and two-sided table. Shaded entries correspond to . Parameters are .  (b) The visualization for the beta distributions in (a). |

## Table calculations

To further help researcher choose a study design, we allow user to set a range for feasible number of samples. Sensei then rapidly calculate for all pairs of feasible and . There is an obvious trade-off between and . For example, to obtain , the researcher may gather either 10 control samples and 7 experimental samples, or 7 control samples and 8 experimental samples. The final decision will be based on the difference of cost of control and experimental samples. One may choose the former if the control sample is much easier to get, or the latter if both samples are equally costly.

## Reparametrize beta distributions

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 it is what is expected.

# Discussion

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.

# 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