BMC Bioinformatics

SOFTWARE Open Access

SCOPIT: sample size calculations for single-cell sequencing experiments

Alexander Davis[1,2], Ruli Gao[1] and Nicholas E. Navin[1,3*]

Abstract

Background: In single cell DNA and RNA sequencing experiments, the number of cells to sequence must be decided before running an experiment, and afterwards, it is necessary to decide whether sufficient cells were sampled. These questions can be addressed by calculating the probability of sampling at least a defined number of cells from each subpopulation (cell type or cancer clone).

Results: We developed an interactive web application called SCOPIT (Single-Cell One-sided Probability Interactive Tool), which calculates the required probabilities using a multinomial distribution (www.navinlab.com/SCOPIT). In addition, we created an R package called pmultinom for scripting these calculations.

Conclusions: Our tool for fast multinomial calculations provide a simple and intuitive procedure for prospectively planning single-cell experiments or retrospectively evaluating if sufficient numbers of cells have been sequenced. The web application can be accessed at navinlab.com/SCOPIT.

Keywords: Single cell sequencing, Sample size, Multinomial distributions

Background

Biological tissues consist of a heterogeneous mixture of cells, including a variety of cell types in normal tissue or subclones in tumor tissue. This heterogeneity can be resolved using single-cell DNA or RNA sequencing methods [1, 2]. Single-cell sequencing studies require sufficiently many cells to be sampled so that normal cell types or cancer subclones of interest (both hereafter referred to as "subpopulations") are represented in the sample. In most studies, however, the total number of cells is determined arbitrarily by the limits of an instrumentation run, or by budget constraints, which may result in the sampling of too few or too many cells. Here, we have developed an interactive web tool, called SCOPIT (Single-Cell One-sided Probability Interactive Tool), which provides assistance for planning experiments, using calculations from a multinomial distribution.

Implementation

The first fact used for calculating multinomial probabilities is the well-known equivalence between the probability mass function of a multinomial distribution and conditional probabilities of a Poisson distribution. This equivalence was first noted, to our knowledge, by Fisher [3].

Theorem 1 *Assume that*

$$N\sim\text{Multinomial}(p,n)$$

where N and p are length k vectors, and $\sum_{i=1}^{k} p_i = 1$. *Also assume that*

$$X_i\sim\text{Poisson}(\lambda_i)$$

for $i = 1$ *to k, where* $\lambda_i = \alpha p_i$ *for some α. Furthermore, assume that* $X_1 \ldots X_k$ *are independent. Then for any event E,*

$$P(N\in E) = P\left(X\in E \middle| \sum_{i=1}^{k} X_i = n\right)$$

The second fact is a relationship between conditional Poisson probabilities, and an expression involving the sum of truncated Poisson random variables.

* Correspondence: nnavin@mdanderson.org
[1] Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[3] Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA
Full list of author information is available at the end of the article

The following is a slight variant of a theorem due to Levin [4].

Theorem 2 *Let $X_i^{(a_i,b_i)}$ be a truncated Poisson random variable, with probability mass function*

$$P(X_i^{(a_i,b_i)} = x) = P(X_i = x | a_i < X_i \leq b_i)$$

where X_i is a Poisson random variable with rate λ_i. For vectors a and b, let $X^{(a,b)}$ be the vector containing all of these truncated Poisson random variables. Let E be the set of vectors x such that $a_i < x_i \leq b_i$. Then,

$$P\left(X \in E \middle| \sum_{i=1}^{k} X_i = n\right) = \prod_{i=1}^{k} P(a_i < X_i \leq b_i) \frac{P\left(\sum_{i=1}^{k} X_i^{(a_i,b_i)} = n\right)}{P\left(\sum_{i=1}^{k} X_i = n\right)}$$

Proof: By Bayes' theorem,

$$P\left(X \in E \middle| \sum_{i=1}^{k} X_i = n\right) = P(X \in E) \frac{P\left(\sum_{i=1}^{k} X_i = n \middle| X \in E\right)}{P\left(\sum_{i=1}^{k} X_i = n\right)}$$

Substituting $P(\sum_{i=1}^{k} X_i^{(a_i,b_i)} = n)$ for $P(\sum_{i=1}^{k} X_i = n | X \in E)$ and $\prod_{i=1}^{k} P(a_i < X_i \leq b_i)$ for $P(X \in E)$ yields the theorem. □

This theorem enables a fast calculation of the multinomial probability. The rate-limiting step is calculation of the probability distribution of $\sum_{i=1}^{k} X_i^{(a_i,b_i)}$. Levin [4] provided two suggestions for computing this probability distribution: the first by convolution of the distributions of each $X_i^{(a_i,b_i)}$, and the second using an Edgeworth expansion of the probability distribution of $\sum_{i=1}^{k} X_i^{(a_i,b_i)}$. We implemented both suggestions, which are used for different values of *n*. For small values of *n*, convolution is performed, using The Fastest Fourier Transform In The West algorithm [5]. For large values of *n*, an Edgeworth expansion is used. However, whereas Levin [4] used the first four terms in the expansion, we continue adding terms until the last term added is sufficiently small.

SCOPIT also computes Bayesian posterior probability distributions for the multinomial probabilities. The multinomial probabilities described above are a function of the population frequencies. When the true population frequencies are not known, but observed frequencies from a previous experiment are available, SCOPIT computes a posterior distribution for the frequencies. The prior used for the frequencies is Dirichlet$(0, \ldots, 0)$, following Jaynes [6] for an experiment in which the possible outcomes are not known in advance. The resulting posterior is Dirichlet$(n_1, \ldots, n_k)$, where n_i is the number of cells observed from population *i*. Possible frequency vectors are randomly drawn from this posterior using the R package rBeta2009 [7, 8]. Then, the desired multinomial probability is calculated from each sampled frequency vector, resulting in samples from the posterior distribution of possible multinomial probabilities. A posterior distribution over the number of cells required is calculated in the same way.

Results

Estimating required sample size using the multinomial distribution

We make the simplifying assumption that a successful experiment requires sampling a sufficient number of representatives from each subpopulation of interest in the tissue. Defining *c* as the required number of representatives from each subpopulation, N_i as the number of cells of subpopulation *i* which are sampled, and *k* as the number of subpopulations of interest, then the probability of meeting this condition is

$$P(N_1 \geq c, N_2 \geq c, \ldots, N_k \geq c)$$

Assuming that a fixed number of cells are chosen at random from the population, the distribution of $N_1, \ldots, N_k$ is multinomial. To calculate this probability, we created an R implementation of a previously described algorithm [4], described further in the Implementation section. Our implementation is available for R scripting in the package "pmultinom", available from CRAN (Table 1).

Our web tool, SCOPIT, provides an interactive interface for multinomial calculations. SCOPIT provides both prospective and retrospective calculations, described below.

Table 1 Package functions for pmultinom. This table lists the R functions for the package "pmultinom" for calculating multinomial probabilities

Function	Arguments	Description
pmultinom	lower, upper, size, probs, method	Probability that a multinomial random vector is elementwise greater than "lower" and elementwise less than or equal to "upper". "size" and "probs" specify the parameters of the multinomial distribution. Either "lower" or "upper" may be left unspecified.
invert.pmultinom	lower, upper, probs, target.prob, method	Returns the "size" parameter required for pmultinom to reach the target probability "target.prob".

Prospective calculations

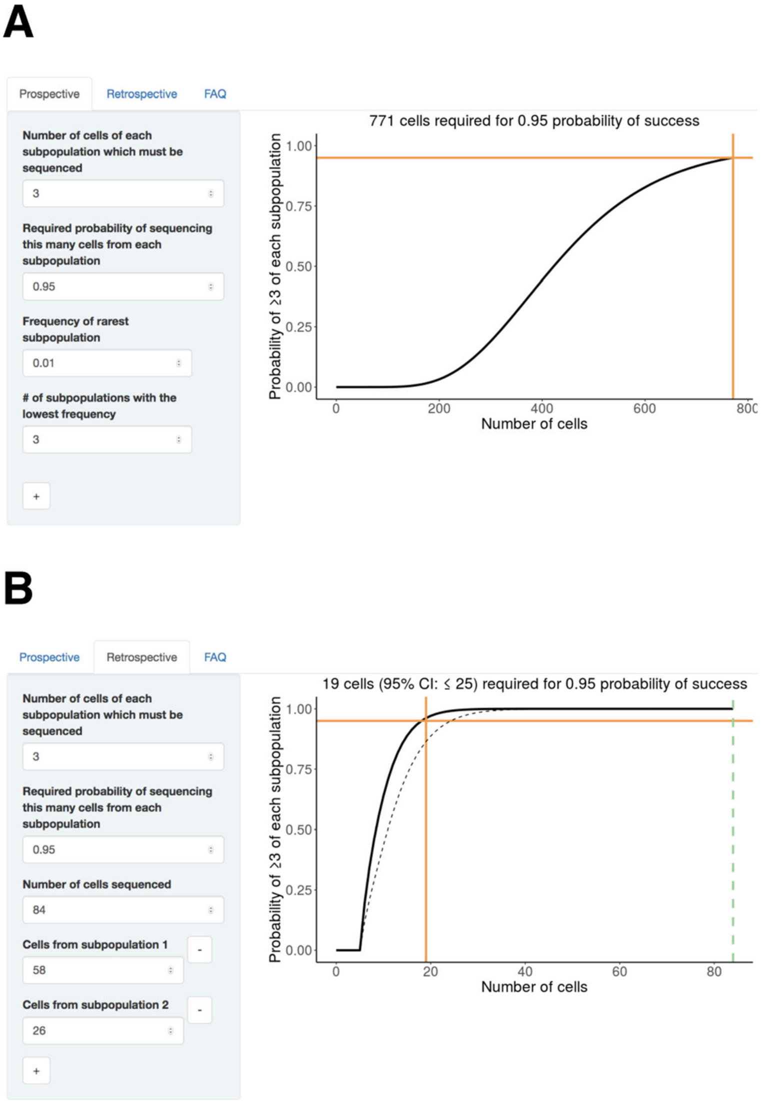
SCOPIT's prospective mode is intended to estimate the number of cells that must be sampled in a single-cell sequencing experiment. Ideally, the number of cells can be decided by finding a number of cells, n^*, such that the above multinomial probability is above a specified success probability, p^*. Such a calculation would require specifying the frequency of each subpopulation of cells in the tissue, but the precise subpopulation frequencies are usually unknown before performing the experiment.

The strategy implemented in the prospective mode is to specify the frequency of the rarest subpopulations that the researcher intends to find, as well as k, the number of populations with approximately this frequency. Both numbers are relevant, since it is harder to find, for example, 10 subpopulations with frequency 1%, than it is to find only one.

The required number of cells is defined as follows:

$$n^* = \min\{n \mid P(N_1 \geq c, N_2 \geq c, \dots, N_k \geq c) \geq p^*\}$$

SCOPIT reports n^* along with a plot of the probability as a function of the number of cells sequenced (Fig. 1a).


A
Prospective
Retrospective
FAQ
771 cells required for 0.95 probability of success
Probability of ≥3 of each subpopulation
1.00
0.75
0.50
0.25
0.00
0
200
400
600
800
Number of cells
B
Prospective
Retrospective
FAQ
19 cells (95% CI: ≤ 25) required for 0.95 probability of success
Probability of ≥3 of each subpopulation
1.00
0.75
0.50
0.25
0.00
0
20
40
60
80
Number of cells

A

- Number of cells of each subpopulation which must be sequenced: 3
- Required probability of sequencing this many cells from each subpopulation: 0.95
- Frequency of rarest subpopulation: 0.01
- # of subpopulations with the lowest frequency: 3
- +

B

- Number of cells of each subpopulation which must be sequenced: 3
- Required probability of sequencing this many cells from each subpopulation: 0.95
- Number of cells sequenced: 84
- Cells from subpopulation 1: 58 -
- Cells from subpopulation 2: 26 -
- +

Fig. 1 SCOPIT interface. **a**. Interface for prospective calculations. Orange lines identify the number of cells required and the target probability of detecting a specified number of each subpopulation. **b**. Interface for retrospective calculations. The number of cells which were sequenced is entered, and is marked on the plot with a dotted green line. In this example, the orange line is far to the left of the dotted green line, suggesting that more cells were sequenced than required to detect these three subpopulations. To quantify confidence in the results, a dotted black line is plotted that shows the lower end of a 95% credible interval for the probability. The plot title states the upper end of a 95% credible interval for the number of cells required

This mode requires only one subpopulation frequency to be specified: the minimum frequency among all subpopulations of interest. The SCOPIT interface does enable the user to add additional subpopulations with higher frequencies, but the user will find that these additional subpopulations have negligible effects on n^*, unless they are very close in frequency to the rarest subpopulations. This phenomenon justifies specifying only the lowest frequency.

Retrospective calculations

After an experiment has been performed, estimates of the subpopulation frequencies are available as input parameters. It is then possible to use SCOPIT in retrospective mode to estimate how many cells would be required, in a hypothetical replicate experiment, to detect all k observed subpopulations, with c representatives from each. In retrospective mode, the information required from the user consists of the total number of cells sequenced in a previous experiment, and the number of cells observed from each subpopulation. With this information, SCOPIT will calculate, for each number of cells n, the probability $P(N_1 \geq c, N_2 \geq c, \ldots, N_k \geq c)$, assuming the true subpopulation frequencies are equal to the empirically observed ones. For example, in Fig. 1b, we use single cell DNA data from a triple-negative breast tumor [9] in which the authors sequenced $N = 84$ single cells and detected two major clonal subpopulations. Using SCOPIT we estimated that only 19 cells were required to detect the two subpopulations with a 0.95 probability, suggesting that this study sequenced about 4 times the number of cells that were necessary.

Because the retrospective analysis involves uncertainty about the true frequencies of each population, SCOPIT provides measures of uncertainty using Bayesian credible intervals at a 95% confidence level. For the number of cells required, SCOPIT reports the upper end of a one-sided credible interval, which is interpretable as the highest number of cells consistent with the data. For the probability of obtaining a sufficient number of cells from each population, SCOPIT plots the lower end of a one-sided credible interval, interpretable as the lowest probability consistent with the data. In the example described above, the credible interval boundaries were close to the estimated values, indicating that the estimated values were strongly supported by the data provided.

The retrospective tool is useful for planning a second experiment, assuming that all the subpopulations of interest were observed in the first experiment, and that the underlying subpopulation frequencies are consistent in both experiments. Although the exact subpopulation frequencies are not known, overconfident conclusions on the basis of limited information can be avoided using the credible intervals provided by the retrospective tool.

Comparison with independence approximation

Another previous software tool for estimating single cell sample sizes is an unpublished web application (https://satijalab.org/howmanycells). The previous tool is based upon two simplifying assumptions: that the subpopulations have equal frequencies, and that the observed frequencies of each subpopulation are statistically independent. Under these assumptions:

$$P(N_1 \geq c, N_2 \geq c, \ldots, N_k \geq c) = P(N \geq c)^k$$

where N represents the number of cells sampled from an arbitrary subpopulation. To compare the independence approximation method to SCOPIT, the required number of cells was calculated with and without the independence assumption (Table 2). The calculations performed under the independence assumption underestimated the required number of cells by at most 1 cell and were highly similar. These data suggests that using independence approximation is an alternative approach that can also be used for estimating single cell sample sizes.

Discussion

SCOPIT's function is to calculate the number of cells that must be sampled in a single-cell sequencing experiment, on the basis of input subpopulation frequencies, and under the assumption of random sampling. To achieve this goal, we implemented a fast multinomial probability calculation approach that is provided as open

Table 2 Comparison of Independent Approximation and Exact Calculations.

Subpopulation frequency	# of subpopulations	Cells required (exact)	Cells required (approx.)
0.1	6	186	186
0.2	3	85	85
0.3	2	53	53
0.1	8	191	191
0.2	4	87	87
0.4	2	39	39
0.1	9	193	193
0.3	3	55	55
0.1	10	195	194
0.2	5	89	89
0.5	2	30	30

The number of cells required to achieve a 95% certainty of sampling sufficiently many cells from each subpopulation. The number of cells was calculated in two ways: by an exact calculation, and by an approximate calculation in which the counts of different subpopulations were assumed to be independent access software through the R package 'pmultinom'. This method enables calculations at speeds sufficient for interactive plotting. The retrospective sample size calculation performed by SCOPIT is distinct from estimation of the number of undiscovered subpopulations [10] or the number likely to be discovered in further sampling [11], and can instead be interpreted as the required sample size of a replicate experiment which would detect the same subpopulations as the original experiment.

To determine the number of cells required, SCOPIT calculates the probability of sampling sufficiently many representatives of each subpopulation. The probability calculated by SCOPIT is relevant to a wide variety of analyses and technologies, but specific technologies introduce additional experimental design considerations. For example, in single-cell differential expression analysis, it is important not only to sample sufficiently many cells, but also to sample sufficiently many transcripts from each cell. Other tools have been developed to calculate the probability of detecting a specific transcript [12], to calculate the power to detect differential expression [13], and to determine the number of cells and reads required to find accurate low-dimensional representations of single-cell RNA sequencing data [14]. Accommodating the unique aspects of other technologies and analyses is an important topic for future research in the design of single-cell sequencing experiments.

A previous tool is available for calculating the number of cells to sequence (https://satijalab.org/howmanycells) and a direct comparison to SCOPIT shows that it generates results that are highly similar to SCOPIT, despite using independent approximations instead of exact probabilities. However SCOPIT offers several additional features, including the ability to enter multiple cell type frequencies, and interfaces to perform both prospective estimates of the sample sizes for planning experiments and retrospective calculations which include measures of confidence in the result.

While SCOPIT can be used to decide how many cells to sample from a tissue, another important question is how many spatial regions to sample to capture the diversity of the population. In the case of sampling from tumor tissue, the question of how widely to sample can be addressed by simulating the generation of intratumor heterogeneity [15], followed by simulating sampling. However, simpler statistical calculations which avoid detailed simulations are currently not available and represent an important future direction.

Conclusions

This study reports a useful tool for estimating sample size calculations for planning single cell sequencing experiments prospectively and retrospectively. We expect that SCOPIT will have applications in many diverse areas of biology, and for planning experiments on a variety of single cell technologies (scDNA, scRNA and scATAC-seq).

Availability and requirements

Project name: SCOPIT
Project homepage: https://github.com/navinlabcode/scopit
Web interface: http://www.navinlab.com/SCOPIT
Operating system: Platform independent
Programming language: R
License: AGPL v3

Abbreviations

SCOPIT: Single-cell one-sided probability interactive tool

Acknowledgments

We thank Aislyn Schalck for assistance with the project.

Authors' contributions

AD was involved in all aspects. RG conceptualized the project, and NN wrote the manuscript. All authors have read and approved the final manuscript.

Funding

This work was supported by grants to NN from the American Cancer Society (129098-RSG-16-092-01-TBG), the Chan-Zuckerberg Initiative Human Cell Atlas Seed Network Grant (HCA3–0000000147) and the CPRIT Single Cell Genomics Center Grant (RP180684). This work was also supported by Susan Komen (PDF17487910) and AACR Basic Cancer Research Fellowships (174042GAO) to RG. AD was supported by an NLM fellowship (4T15LM007093–25). These funding agencies did not play any role in the design of the study, the collection, analysis, and interpretation of data, or in the writing of the manuscript.

Availability of data and materials

The source code used in the current study are available from GitHub, https://github.com/navinlabcode/

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Genetics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [2]The University of Texas MD Anderson Cancer Center UTHealth Graduate School of Biomedical Sciences, Houston, TX, USA. [3]Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, TX, USA.

Received: 20 February 2019 Accepted: 21 October 2019
Published online: 12 November 2019

References

1. Baran-Gale J, Chandra T, Kirschner K. Experimental design for single-cell RNA sequencing. Brief Funct Genomics. 2018;17(4):233–9. https://doi.org/10.1093/bfgp/elx035.
2. Navin NE. The first five years of single-cell cancer genomics and beyond. Genome Res. 2015;25(10):1499–507. https://doi.org/10.1101/gr.191098.115.
3. Fisher RA. On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. J R Stat Soc. 1922;85(1). https://doi.org/10.2307/2340521.

4. Levin B. A representation for multinomial cumulative distribution functions. Ann Stat. 1981;9(5):1123–6. https://doi.org/10.1214/aos/1176345593.
5. Frigo M, Johnson SG. The design and implementation of FFTW3. Proc IEEE. 2005;93(2):216–31. https://doi.org/10.1109/JPROC.2004.840301.
6. Jaynes ET. Probability theory: the logic of science: Cambridge University Press; 2003. https://books.google.com/books?id=UjsgAwAAQBAJ. Accessed 11 Sept 2009.
7. Cheng CW, Hung YC, Balakrishnan N. rBeta2009: The Beta Random Number and Dirichlet Random Vector Generating Functions. 2012. https://cran.r-project.org/package=rBeta2009. Accessed 11 Sept 2009.
8. Hung YC, Balakrishnan N, Cheng CW. Evaluation of algorithms for generating Dirichlet random vectors. J Stat Comput Simul. 2011;81(4):445–59. https://doi.org/10.1080/00949650903409999.
9. Gao R, Davis A, McDonald TO, Sei E, Shi X, Wang Y, Tsai P-C, et al. Punctuated copy number evolution and clonal stasis in triple-negative breast cancer. Nat Genet. 2016;48:1–15. https://doi.org/10.1038/ng.3641.
10. Gotelli NJ, Colwell RK. Estimating species richness. In: Biological Diversity. Frontiers in Measurement and Assessment, vol. 2; 2011. p. 39–54. https://doi.org/10.2307/3547060.
11. Shen TJ, Chao A, Lin CF. Predicting the number of new species in further taxonomic sampling. Ecology. 2003;84(3):798–804. https://doi.org/10.1890/0012-9658(2003)084[0798:PTNONS]2.0.CO;2.
12. Svensson V, Natarajan KN, Ly LH, Miragaia RJ, Labalette C, Macaulay IC, Cvejic A, Teichmann SA. Power analysis of single-cell RNA-sequencing experiments. Nat Methods. 2017;14(4):381–7. https://doi.org/10.1038/nmeth.4220.
13. Jenkins D, Faits T, Khan MM, Briars E, Pro SC, Johnson WE. singleCellTK: Interactive Analysis of Single Cell RNA-Seq Data. 2018. https://doi.org/10.18129/B9.bioc.singleCellTK.
14. Svensson V, da Veiga Beltrame E, Pachter L. Quantifying the tradeoff between sequencing depth and cell number in single-cell RNA-seq. bioRxiv. 2019. https://doi.org/10.1101/762773.
15. Sun R, Hu Z, Sottoriva A, Graham TA, Harpak A, Ma Z, Fischer JM, Shibata D, Curtis C. Between-region genetic divergence reflects the mode and tempo of tumor evolution. Nat Genet. 2017:1–13. https://doi.org/10.1038/ng.3891.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.