Completeness estimation of large-scale single-cell sequencing projects

Mitro Miihkinen[1,2], **Sara Vakkilainen**[1], **and Tero Aittokallio**[1,2,3,4]

[1]Institute for Molecular Medicine Finland (FIMM), HiLIFE, University of Helsinki, Helsinki 00290, Finland
[2]iCAN Digital Precision Cancer Medicine Flagship, University of Helsinki and Helsinki University Hospital, Helsinki 00290, Finland
[3]Institute for Cancer Research, Department of Cancer Genetics, Oslo University Hospital, Oslo 0310, Norway
[4]Oslo Centre for Biostatistics and Epidemiology (OCBE), Faculty of Medicine, University of Oslo, Oslo 0372, Norway

During embryonic development, cells undergo differentiation into highly specialized cell types. Taking advantage of single-cell RNA sequencing, substantial resources have been invested towards cataloguing these differentiated cell types by their transcriptomic profiles. Despite the extensive efforts to profile different organs and their cellular compositions, metrics to assess the completeness of such datasets have been lacking. In this cellular biodiversity analysis, we used the increasingly available single-cell data together with statistical methods, originally developed for assessing the species richness of ecological communities, to provide cellular diversity estimates for any organ based on single-cell RNA sequencing data, as well as estimating how cell types are distributed along different organ systems. Deriving from such cellular richness estimates, we establish a statistical framework that allows estimation of the completeness of any large-scale single-cell profiling projects, after which additional sequencing efforts do not anymore reveal new insights into an organ's cellular composition. Such estimates can serve as stoppage-points for the ongoing sequencing projects for more cost-efficient completion of the remaining profiling on various human tissues.

Statistical estimator | Single-cell RNA sequencing | Completeness estimation | Chao2

Correspondence: ***tero.aittokallio @helsinki.fi***

Introduction

Each organ in a human body can be thought of as a natural community, an ecosystem, composed of various, morphologically and functionally diverse cell types. In an adult human, differentiated cells are defined by their cell type specific spatial location, function as well as gene expression which both defines and maintains cell identity and function (Zeng, 2022). For this reason, the emergence of single-cell RNA sequencing has paved the way for profiling and cataloguing cells from different organs, placing them into increasingly granular cell type categories. Although there does not exist a clear consensus on the level of granularity at which cell types should be classified, based on such cataloguing efforts ambitious international initiatives, such as the human cell atlas, are aiming to build human cell type reference maps as a basis for understanding human health and disease (Rood et al., 2024). However, the lack of accurate estimates of the number of unique cell types make it challenging to know how complete such cataloguing efforts currently are.

The cellular communities that compose different organs can be analysed using biodiversity metrics. "Species richness" of an ecosystem, or any natural community, is the simplest and most intuitive description of biodiversity, and it is often related to ecosystem function and resilience (Gwinn et al., 2016). However, identifying and cataloguing every individual in a population to understand species richness is often impractical, expensive and unnecessary (Magurran, 1988). A more effective approach to study the number of species is through replicate sampling of the ecosystem. However, it is important to note that the number of species present in a sample is almost always an underestimate of the true population value. To correct this bias, many statistical methods have been proposed to provide more accurate estimates of the species richness of a population (Gwinn et al., 2016). We argue that such estimates could be used to assess the richness of cellular ecosystems, and to provide completeness estimates for single-cell atlases, provided the methods are applicable to cellular systems.

Methods to infer species richness from a sample of an ecosystem range from purely data-analytical approaches to parametric and non-parametric statistical models as well as to upsampling methods (Schmitz & Rahmann, 2024). While these approaches have been extensively applied to ecological populations—such as animal or plant species—special considerations are required when selecting appropriate estimators for cellular communities, particularly in the context of single-cell RNA sequencing data. Importantly, most of the statistical models pose underlying assumptions on the population structure, which in many occasions make them unsuitable for estimating species richness of a cellular community (Chiu, 2023; Schmitz & Rahmann, 2024). In the absence of a clear consensus on the best models for species richness estimation in cellular systems, selecting suitable and accurate models among the best-performing alternatives needs to be done with great care.

As single-cell resolution data from different human tissues becomes increasingly available, we considered human tissues as natural communities consisting of cells with defined cell types and used statistical models to describe their biodiversity characteristics, such as number and distribution across tissues. Our analysis reveals many key insights into how cell types are distributed along different organ systems, which together with statistical modelling give informative means to guide experimental single-cell profiling efforts. Importantly, using single-cell data together with statistical species richness estimation models, we investigate the completeness of

currently available single-cell resources and suggest ways on how statistical analysis can enrich the ongoing efforts to build comprehensive single cell atlases.

Results

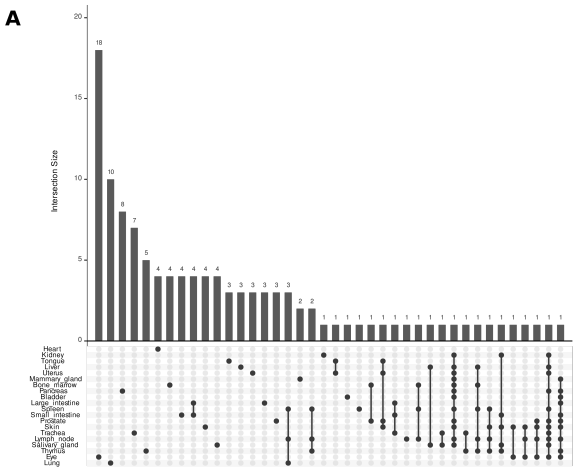
Cell communities in human follow a loglinear distribution. To investigate the cellular biodiversity among different organ systems in a human body, we exploited data of cell type abundances across 20 human organs from Tabula Sapiens (Tabula Sapiens Consortium* et al., 2022), and assessed the statistical properties of each detected cell type (137 cell types in total) across all 20 organs. To begin, we first describe how individual cell types can be either organ-specific or shared across different organs. Such an organ-level analysis demonstrates that there exist various degrees of both common and unique cell types across organs (Figure 1A), and provides a way to understand organ development through cell types. For example, the large number of different cell types in the eye and lung highlight the complex functions of both of these organs (Figure 1B), whereas the high number of unique, organ-specific cell types in the eye reflects the evolutionary time needed to develop such an organ (Figure 1B). More specifically, the eye had the most specialized repertoire of unique cell types (58.1% unique cell types), and lymph nodes having the least (4.8%). This was also true for other organs of the lymphatics system, such as speen, with only 5.9% of unique cell types, reflecting the migratory nature of these cells (Figure 1A-B). Our analysis revealed different immune cell populations and fibroblasts as the cell types most common among the organs, reflecting the importance of fibroblasts and the immune system for all organ development and function (Figure 1A). Interestingly, most cell types (64%) were highly specilialized and thus organ-unique.

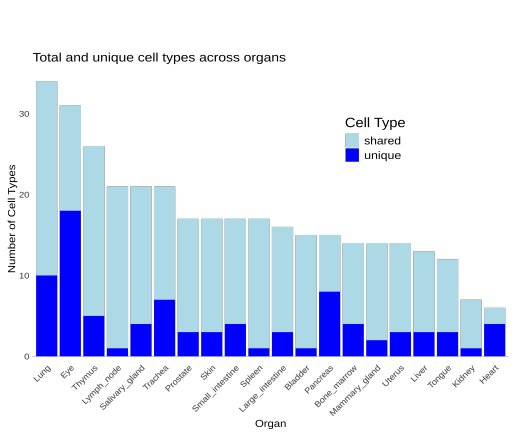
Since the cell type uniqueness analysis (Figure 1A) demonstrated cell types can be either shared or unique across different organs, we first assessed how different human organs are correlated in terms of their cell type compositions. Indeed, the organ systems analysed from Tabula Sapiens data show variable degrees of correlations in terms of the shared cell populations (Figure 1C). Although such analysis shows many expected correlations between similar organs (e.g., organs of the lymphatics system, or different parts of the intestine), surprisingly, small intestine and thymus show unexpectedly high correlation although not being closely related in terms of their primary function or anatomical location (Figure 1C). Notably, cell type abundances in all 20 organs displayed a heavy-tailed distribution and since cell abundances in other cellular ecosystems in bacterial communities have been previously described to follow a loglinear distribution (Grilli, 2020), we tested whether cell type abundances in different human organ systems follow this distribution as well. Indeed, Kolmogorov-Smirnov test showed lognormal distribution to be a valid approximation across all the 20-organs tested (Figure 1D), such as the eye (Figure 1E). Finally, we wanted to understand the depth at which all of the organs have been profiled. Although sample coverage (amount of population profiled) is a crucial piece of information to assess the statistical properties of the sample, this generally is not reported by any of the current atlas projects. We estimated the sample coverage for each of the 20 organs in Tabula Sapiens and observed drastic differences at the levels which different organs have currently been profiled. These differences underscore the need for standardized reporting of sample coverage to improve better cross-organ comparisons and highlight the organs systems benefitting for more sampling.

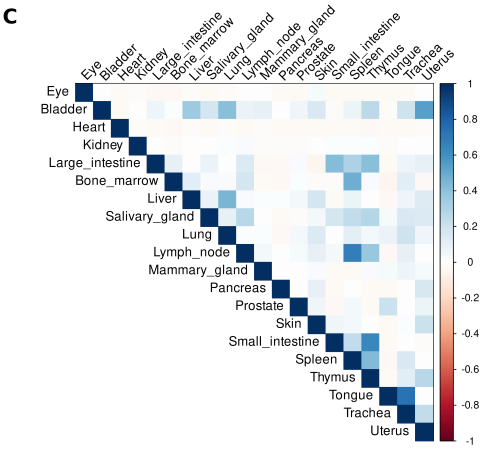
In summary, our analysis of Tabula Sapiens cell type abundance data gives an overview of the unique and shared distributions of individual cell types, enabling an intuitive way to understand these cellular communities. Importantly, we show that organ systems and a complete human are cellular ecosystems whose cell type abundances are approximately lognormally distributed. In light of previous research done on bacterial species (Grilli, 2020), our analysis supports lognormal distribution as a natural law from which cellular communities can be derived.

Statistical framework for estimating the accuracy of species richness estimators. Currently, the completeness of most single-cell atlases is defined as the number of cells stored in the database. By such count metrics, the currently largest dataset, CELLxGENE (Program et al., 2023), stores data of 96.1 million cells. Since the human body consists of 35 trillion cells, the completeness of CELLxGENE would be approximately 0.03%. However, since these datasets store data from different organs in an non-uniform manner, such estimates are not appropriate for atlas completeness (Figure 1F). The count metrics could be improved by displaying them organ-by-organ, but even this approach is not able to assess how complete the profiling efforts are. Since one of the immediate goals of single-cell profiling efforts is to exhaustively catalogue all cell types, we argue that the most relevant way to assess any single-cell atlas completeness is to estimate how many cell types remain to be unrecognized by the current profiling data. To solve this statistical problem (Figure 2A), we assessed the accuracy of different species richness estimation methods when applied to single-cell atlas data from different organ systems.

To enable performance testing of different statistical estimators, we first created a simulated human body with defined organ systems and cell types. Since the cell type uniqueness analysis (Figure 1A) demonstrated cell types can be either shared or unique across different organs (Figure 1A-C), we simulated realistically distributed cell types across all the organs (see Methods). We next used the simulated human as the ground truth, from which samples are drawn while understanding the population complexity and number of cell types. Importantly, the simulated organs had cell type distributions similar to those in Tabula Sapiens (Figure 2B) and when assessed over all organs, the cell types showed lognormal distribution (Figure 2C). Interestingly, when 20 human organs were simulated based on the parameters from Tabula Sapiens data, the fraction of organ-unique cell types in each organ was significantly higher than in Tabula Sapiens data, providing some evidence that many organ-specific cell types are yet to be discovered (Figure 2B).

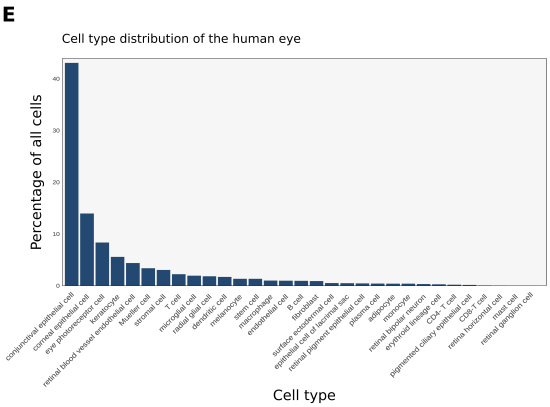






D

Organ	KS Statistic	P-Value	Distribution
Eye	0.0752801	0.9946675	LogNormal(4.5, 1.6)
Bladder	0.1121061	0.9807298	LogNormal(6.4, 1.7)
Heart	0.2147759	0.8940697	LogNormal(6.5, 1.6)
Kidney	0.3540887	0.2735233	LogNormal(5.8, 1.4)
Large_intestine	0.1642162	0.7221118	LogNormal(5.7, 1.6)
Bone_marrow	0.2203980	0.4418320	LogNormal(6.2, 1.6)
Liver	0.1094668	0.9976911	LogNormal(5.2, 1.3)
Salivary_gland	0.1490187	0.6852977	LogNormal(6.2, 1.4)
Lung	0.0743019	0.9919100	LogNormal(5.5, 1.7)
Lymph_node	0.1214936	0.8795552	LogNormal(5.9, 2.3)
Mammary_gland	0.1412345	0.9062482	LogNormal(5.2, 1.7)
Pancreas	0.1870638	0.6701663	LogNormal(5.3, 1.8)
Prostate	0.1446297	0.8206159	LogNormal(5.5, 1.9)
Skin	0.2128894	0.3390233	LogNormal(5.2, 1.7)
Small_intestine	0.1532172	0.7651580	LogNormal(5.5, 1.4)
Spleen	0.2142533	0.3634772	LogNormal(6.4, 1.9)
Thymus	0.1437035	0.6056948	LogNormal(5.9, 2.0)
Tongue	0.1806184	0.7661923	LogNormal(5.5, 2.0)
Trachea	0.1247401	0.8599755	LogNormal(5.1, 1.5)
Uterus	0.1069748	0.9916980	LogNormal(5.0, 1.6)



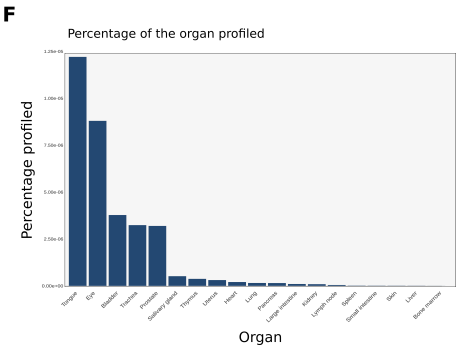


Fig. 1. Cellular biodiversity of human organ systems according to Tabula Sapiens. (A) UpSet plot show how cell types are distributed across the 20 organ systems in Tabula Sapiens (B) The number of cell types and unique cell types across 20 organs in Tabula Sapiens (C) Pearson pairwise correlations between different organ pairs in Tabula Sapiens dataset, based on the counts of different cell types (D) Statistical loglinear goodness-of-fit testing of cell type abundances from different organs. The table reports test statistics from Kolmogorov-Smirnov test and calculated p-values (E) Lognormal distribution of the eye. KS-statistic 0.07, P-value 0.99. (F) Depth at which organs in Tabula Sapiens have currently been profiled. Calculated as a percentage of the number of cells profiled from the total number of cells in each organ. Missing are breast tissue for which estimates of tissue size were not available.

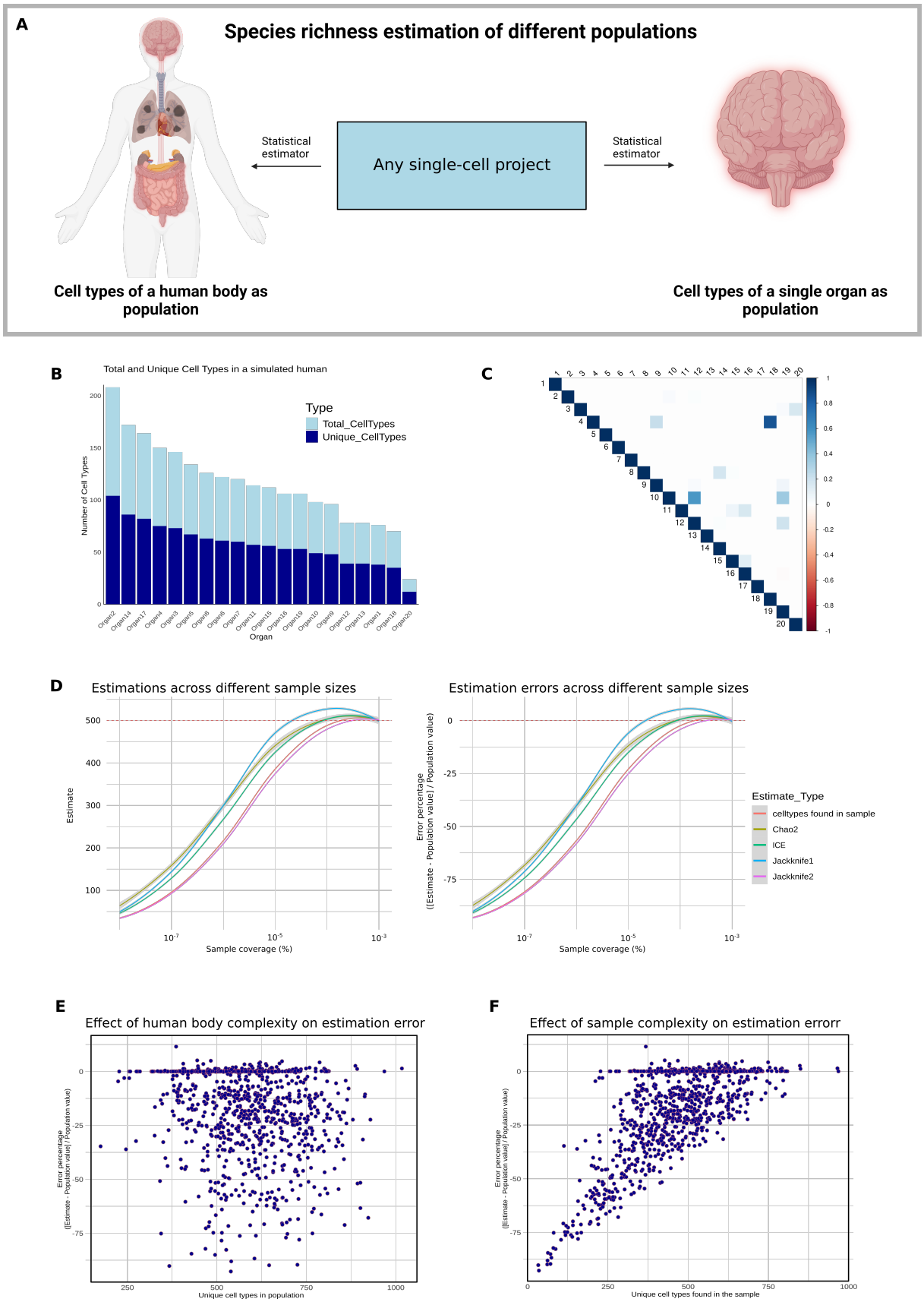


Fig. 2. Statistical framework to assess the accuracy of species richness estimators (A) Schematic representation of the estimation problem. Created using biorender.com (B) Simulated human: The number of cell types and unique cell types across different organs (C) Simulated human: Pearson pairwise correlations between different organ pairs, based on the counts of different cell types (D) Accuracy and error of different species richness estimators across different sample sizes from the simulated human. The simulated human body here consisted of 500 cell types across 20 organs. (E) Simulation of 1000 individual humans with variable body compositions. Measuring the effect of body composition on estimation error (no effect). (F) Simulation of 1000 individual humans with variable body compositions. Measuring the effect of random sampling on estimation error.

This is likely true for organs which are known to be highly specialized, such as the kidney, which in Tabula Sapiens currently has a low number of defined, organ-specific cell types (Figure 1B).

Since there is no consensus model to estimate species richness (hereafter referred to as cellular richness), especially for cellular ecosystems, we first compared the accuracy of various non-parametric statistical estimators (see Methods). Since the currently available single-cell profiling methodologies are unable to detect cellular species from data, if clustering is not performed, the profiling assays need to capture at least 10 cells to be able to define it as cell type. By using the simulated human(s) as ground truth, we next assessed the accuracy of different cellular richness estimators under varying conditions; in particular, we wanted to understand how the sampling sizes and cell population complexity affect cellular richness estimation. To establish this, we created simulated humans at different levels of cellular complexity (from 200 to 1020 unique cell types), where the cell types were distributed across the organ systems based on the correlation structure defined previously (Figure 1A-C). From each simulated individual, we then drew different sized triplicate samples and used them to assess the number of unique cell types present in the ground truth simulated human. Importantly, after sampling 0.00001% of the population, all the tested estimators started to stabilise towards the ground truth value, indicating that the estimators can efficiently estimate the true population value from an extremely small sample size, when 3 samples are available (Figure 2D).

Among the tested estimators, Chao2 showed the most promising features; while it was not the fastest to plateau towards the true population value, once it reached the plateau, it did not vastly overestimate. We also estimated error percentage for each cellular richness estimator, which after reaching the sample size of 0.00001% was less than 2% for Chao2 estimator (Figure 2D). Due to its desirable statistical properties (low sample size needed, no significant overestimation, low error), we selected Chao2 estimator for further testing, where we simulated 1000 humans with different amounts of unique cell types in their bodies, aiming to measure the performance of Chao2 estimator across humans with potentially variable cellular complexities (Figure 2E). Importantly, we observed no effect of different body compositions on estimator error (Figure 2E). Finally, in a similar 1000 human simulation, we assessed how random sampling affected the estimation error. As predicted, due to random chance, some samples can contain low detection of cell types, leading to higher estimation error (Figure 2F). However, even in these cases, it is important to note that the estimations made using the Chao2 estimator were always more accurate than merely trying to judge the tissue complexity from the amount of cell types detected in a sample (Figure 2D).

As a conclusion, this simulation study and the benchmarking of existing statistical species richness estimators show that cellular richness estimation of organ systems is possible from existing, lognormally distributed single-cell atlas data, when 3 or more samples from a population are available. Importantly, the best-performing Chao2 estimator is simple to use and offers extremely favourable features when estimating cellular richness from a lognormally distributed population. We have here shown that cellular richness can be assessed using Chao2 estimator, which presents a valid option for estimating the richness of lognormally distributed cellular communities.

Estimating the completeness of large single-cell projects. Our previous results indicate that it is possible to confidently estimate the cellular richness from using sample sizes of 0.00001% or larger. Since recently many large scale tissue profiling projects have made use of substantial sample sizes, we next wanted to assess the cellular richness of an organ, using repeated sampling with sufficient sample sizes. We argue further that estimating cellular richness of an organ and comparing it with the currently detected amount of cell types would allow completeness estimation of any large scale single-cell sequencing project.

To begin assessing the completeness of a large-scale single-cell sequencing project, we used open data provided in Yao et al. 2023 (A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain), in which mouse brain is dissected into different parts and profiled separately with repeated sampling. Specifically, Yao et al. provide six samples from the mouse isocortex, with about 250,000 cells per sample, making it one of the largest datasets when assessing it based on sample coverage (sample size / organ size). Since the mouse isocortex has about 25 million cells in total (Herculano-Houzel et al., 2013), the size of the whole dataset thus covers an impressive 6% of the mouse isocortex. To begin analysing mouse isocortex, we assessed the variability in the sampling after filtering out poor quality cells using standard metrics such as gene detection rates and percentage of mitochondrial genes detected in a cell (methods). After filtering, each sample was plotted as a UMAP with detected Leiden clusters showing marked differences in sample complexity between the samples (Fig. 3A).

Species richness estimation requires the criteria on how species, here cell types are defined. Since there are a variety of criteria for cell types classification, for the purpose of assessing the completeness of the isocortex dataset, we simply considered different detected cell clusters as cellular species and labelled them based on their scRNA-seq profiles. To study the frequency at which cell types are found in samples 1-6, a marker set consisting of positive and negative cell type markers was defined for each of the cell types, showing that it was possible to find 29 different cell types in the data, some of which were present in all the 6 samples (cell types 8 & 18) (Fig. 3B). However, the majority of cell types were found on average from 3 out of 6 samples, highlighting the importance of repetitive sampling when building large single-cell atlases (Fig 3B). After defining how the cell types distribute across different samples, we used the Chao2 estimator to estimate the number of cell types in the mouse isocortex using all six samples available.

Although the coverage needed to find all possible cellular species in the organ varies from organ to organ, based on the simulation studies we expected 3% sample coverage to

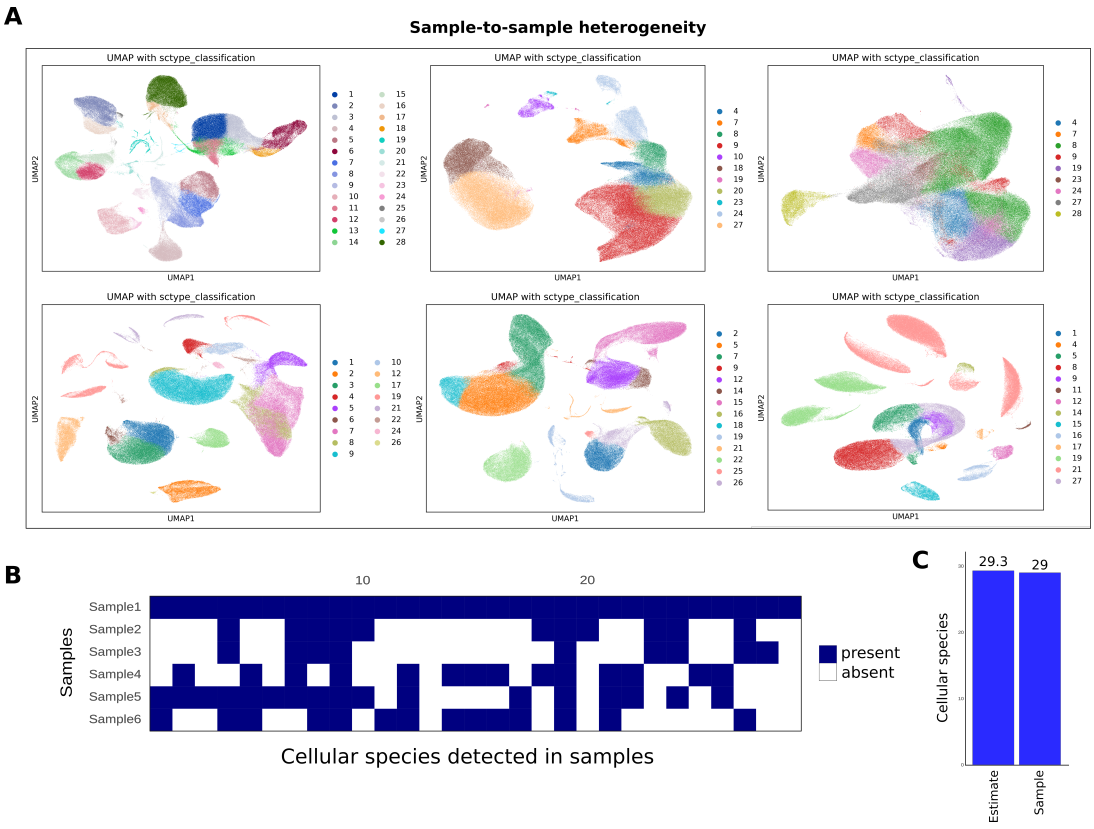


Fig. 3. Completeness estimation of a large single-cell mouse isocortex dataset (A) 6 samples of the mouse isocortex from Yao et al., 2023 (B) Cell type incidence (present/abcent) across the 6 mouse isocortex single-cell samples (C) Chao2 estimation of cellular richness of mouse isocortex dataset using data from Yao et a., 2023.

be enough to detect all possible cellular species in the mouse isocortex. Indeed, applying chao2 estimator on the mouse isocortex data estimated that this organ to have 29.3 cellular species. This suggests that profiling more samples will not anymore reveal additional cellular species, and we therefore assess that the healthy mouse isocortex has been fully profiled.

Conclusions

Species richness estimation is a fundamental problem in ecology and biodiversity studies, where the goal is to determine the total number of species within a given habitat or ecosystem. The problem arises with the practical challenge of trying to observe all species, either plant or animal, which is in most cases difficult or impossible. The era of single-cell biology makes it relatively easy to understand tissues through their cellular composition, but similar to biodiversity studies, assessing the complexity of the whole tissue is challenging. Although previous work has estimated different aspects of human cell populations such as cellular size distribution or number of total cells in the body (Bianconi et al., 2013; Hatton et al., 2023), here we focused on combining single-cell methodologies with statistical species richness estimation to provide a framework enabling cellular richness estimates (number of different cell species). In addition, we suggest derived completeness estimations for any large-scale single-cell profiling project.

Currently, substantial investments are being made in mapping cell types, their locations, and gene regulation across various human tissues using single-cell RNA sequencing and related technologies. Although scRNA-seq is currently the most practical and scalable technology for defining cell types and their states, there exists versatile technologies and criteria on how cell types and cell states can be defined, and how to separate these two from each other (Dance, 2024). Currently, cell type and state definitions are hard to disentagle using only data-analytical means, especially in heterogeneous and dynamically adjusting tissue systems. Although it is likely that future advancements in technology and integrative multi-modal analysis will lead to sharper and more universally accepted definitions, due to the still multiple definitions of cell types and states, here we simply refer to all the detected cellular clusters as to cellular species. In subsequent analyses, it is possible to perform the same analytics and statistical estimations using more nuanced definitions for a cell type, either broad or grained.

Despite significant interest in profiling different healthy and diseased tissues, there has so far been little effort to estimate the total number of unique cellular species within these systems. Most studies focus on cataloging cell populations based on the cells observed in a given dataset, without addressing the question of how many distinct cellular species—encompassing both cell types and cell states—remain undiscovered. Because a sample is almost always a poor estimate for true cellular complexity in a tissue, we here sought to benchmark different available species richness estimators, their utility and sampling requirements

to understand cellular diversity in different tissue systems. Although there exists recent extensive reviews about species richness estimators (Schmitz & Rahmann, 2024), herein we focused on finding estimators that would work together especially with single-cell RNA sequencing.

Our analyses suggest that the cellular species in most, if not all, human tissue systems are lognormally distributed. This led us to test different species richness estimators on lognormally distributed data, showing that many of richness estimators are robust when the sample sizes are large enough. Importantly, such sample sizes are already available in many existing single-cell projects, particularly in large-scale initiatives, such as the Human Cell Atlas and various tissue-specific profiling efforts, which often include millions of cells. To apply these statistical estimators on real-world data, we took the advantage of a previously published mouse isocortex dataset with six individual samples (Yao et al., 2023) This dataset revealed notable heterogeneity in cellular composition between the samples, strongly advising towards using multiple samples when attempting to assess cellular biodiversity. Furthermore, performing many rounds of richness estimations when more samples become available will make the estimations more accurate.

Here, we reasoned that it is possible to understand single-cell atlas resources as a sample of the population and make statistical inference of the human body based on these somewhat large samples - similar to ecological populations. However, unlike for animal species, cells exhibit many favourable statistical properties. In particular, the probability of finding a cell from a tissue sample only depends on the abundance of the cell. In addition, assessing sample coverage for cells is much easier than for animal species, which can lead to more robust estimations. Importantly, derived estimations of a dataset or atlas completeness can become useful when building large single-cell resources, and we therefore suggest running species richness estimations in regular intervals on every profiled organ for understanding the level of completeness of how each organ has been profiled.

We also explored the possibility of estimating the total number of cell types in humans. However, we found that current data, with its highly uneven profiling depths across different tissues, presents limitations for such estimations. However, as large single-cell resources continue to expand and profiling depths across organs become more balanced, this first work will provide the basis of future estimations of even the whole human body. Finally, since single-cell RNA sequencing is becoming a widely used technique also in the study of prokaryotes, we suggest similar strategies could be used to assess bacterial population richness based on a sample.

Supplementary Note 1: Methods

A. Simulations of Organs and Human. Since the variables (organs) are partially correlated, we simulated a human with a defined number of organs (20) and cell types by drawing a sample of size 7.4 trillion cells from a multivariate lognormal distribution with a covariance matrix similar to the empirical covariance structure in Tabula Sapiens. The user-defined parameters in the simulation were:

- **n**: number of cells in the whole organism (set to 7.4 trillion)
- **m**: number of organs
- **k**: number of cell types

Since a multivariate lognormal distribution can be simulated by drawing a sample from a multivariate normal distribution followed by exponential transformation, we first log-transformed the Tabula Sapiens abundance data to understand its central tendency and covariance structure when normally distributed. We observed that for log-transformed data, the non-diagonal elements in the empirical covariance matrix followed a lognormal distribution, while the diagonal elements followed a normal distribution. The central tendency (means of the variables/organs) of the data appeared to follow a normal distribution. Based on this, we simulated an *m*-vector of means and a covariance structure (*m*x*m* positively definite matrix) where variances and covariances follow the observed distributions.

As the final step, the `mvrnorm` function from the `MASS` library in R was used to generate multivariate normally distributed abundance data for a simulated human, which was then back-transformed exponentially to yield lognormally distributed, multivariate data with a defined correlation/covariance structure. Organ sizes, π, were defined based on previous estimates as well as the number of cells in a human body (Bianconi et al., 2013).

B. Species Richness Estimators. Due to limitations imposed by single-cell RNA sequencing on species richness estimation, we focused on testing statistical estimators that use presence/absence data rather than abundance data. Current single-cell profiling methodologies are unable to detect cellular species from data without clustering, requiring assays to capture at least $\sim$10 cells to confidently detect a cell type from a sample. All described estimators are easily accessible through R package fossil.

Chao2 (For Incidence Data) The Chao2 estimator is a non-parametric method used to estimate species richness based on incidence data. It accounts for unseen species by leveraging the frequency of rare species across samples. The formula for Chao2 is:

$$\text{Chao2} = S_{\text{obs}} + \frac{Q_1^2}{2Q_2}$$

where S_{obs} is the observed species count, Q_1 is the number of species observed in exactly one sample, and Q_2 is the number of species observed in exactly two samples (Chao, 1984).

ICE (Incidence-based Coverage Estimator) The ICE estimator is a non-parametric statistical method that uses presence/absence data to infer the number of unseen species. It emphasizes the contribution of rare species across samples. The formula for ICE (Chao, 1987) is:

$$S_{\text{ICE}} = S_{\text{obs}} + \frac{S_{\text{rare}}}{C_{\text{rare}}}$$

Where:

- S_{ICE}: The estimated total species richness.
- S_{obs}: The number of species observed in the dataset.
- S_{rare}: The number of species classified as rare (by default those appearing in ≤ 10 samples).
- C_{rare}: The sample coverage estimate for rare species, calculated as:

$$C_{\text{rare}} = 1 - \frac{Q_1}{\sum_{i=1}^{10} i \cdot Q_i}$$

Here:

- Q_1: The number of species that occur in only one sample.
- Q_i: The number of species that occur in exactly i samples.

Jackknife1 (For Incidence Data) The Jackknife1 estimator is a resampling technique used to estimate species richness. For incidence data, it is calculated as:

$$\text{Jackknife1} = S_{\text{obs}} + \frac{(n-1)Q_1}{n}$$

where S_{obs} is the observed species count, n is the number of samples, and Q_1 is the number of species observed in exactly one sample (Heltshe & Forrester, 1983).

Jackknife2 (For Incidence Data) The Jackknife2 estimator refines the richness estimate by considering species observed in only one or two samples. It is calculated as:

$$\text{Jackknife2} = S_{\text{obs}} + \left(\frac{2n-3}{n}\right) Q_1 - \left(\frac{n-2}{n}\right) Q_2$$

where Q_1 and Q_2 are the number of species observed in exactly one and exactly two samples, respectively (Heltshe & Forrester, 1983).

C. Statistical Methods. Goodness of fit testing for lognormal distributions was done using the Kolmogorov-Smirnov test. All statistical analyses were run using `R` with `R Studio`. Different statistical estimators were used through `R` with the CRAN package `fossil` (Vavrek, 2011).

D. Single-Cell Analytics. All single-cell analysis was performed using `Python` and the `Scanpy` library (Wolf et al., 2018). Briefly, the raw data was filtered, retaining cells with less than 5% mitochondrial genes and more than 200 genes detected. Clustering of the data was performed using `Scanpy`'s Louvain algorithm, and each detected cluster was defined by a marker panel using the `scType` Python implementation (Nader et al., 2024). This allowed for the detection of the same clusters in parallel samples.

E. Data Availability. The codes to produce simulated data can be found on GitHub upon completion of peer review. The real-world cell type data was extracted from the following open databases:

- **Tabula Sapiens**: Data accessed on 17.05.2024
- **CELLxGENE**: Data accessed on 30.07.2024
- **Mouse isocortex**: Data available from https://assets.nemoarchive.org/dat-qg7n1b0 (Yao et al.,2023)

Code Availability

All codes to reproduce the analyses presented in this publication will be made available on GitHub upon completion of peer review for this manuscript.

ACKNOWLEDGEMENTS
We want to thank Yunus Vilkkavaara-Cankocak for their support while preparing this manuscript.

Bibliography

1. Bianconi, E., Piovesan, A., Facchin, F., Beraudi, A., Casadei, R., Frabetti, F., Vitale, L., Pelleri, M. C., Tassani, S., Piva, F., Perez-Amodio, S., Strippoli, P., & Canaider, S. (2013). An estimation of the number of cells in the human body. *Annals of Human Biology*. Retrieved from https://www.tandfonline.com/doi/abs/10.3109/03014460.2013.807878
2. Chao, A. (1984). Nonparametric Estimation of the Number of Classes in a Population. *Scandinavian Journal of Statistics, 11*(4), 265–270.
3. Chao, A. (1987). Estimating the Population Size for Capture-Recapture Data with Unequal Catchability. *Biometrics, 43*(4), 783–791. https://doi.org/10.2307/2531532
4. Chiu, C.-H. (2023). Sample coverage estimation, rarefaction, and extrapolation based on sample-based abundance data. *Ecology, 104*(8), e4099. https://doi.org/10.1002/ecy.4099
5. Dance, A. (2024). What is a cell type, really? The quest to categorize life's myriad forms. *Nature, 633*(8031), 754–756. https://doi.org/10.1038/d41586-024-03073-2
6. Grilli, J. (2020). Macroecological laws describe variation and diversity in microbial communities. *Nature Communications, 11*(1), 4743. https://doi.org/10.1038/s41467-020-18529-y
7. Gwinn, D. C., Allen, M. S., Bonvechio, K. I., Hoyer, M. V., & Beesley, L. S. (2016). Evaluating estimators of species richness: The importance of considering statistical error rates. *Methods in Ecology and Evolution, 7*(3), 294–302. https://doi.org/10.1111/2041-210X.12462
8. Hatton, I. A., Galbraith, E. D., Merleau, N. S. C., Miettinen, T. P., Smith, B. M., & Shander, J. A. (2023). The human cell count and size distribution. *Proceedings of the National Academy of Sciences, 120*(39), e2303077120. https://doi.org/10.1073/pnas.2303077120
9. Heltshe, J. F., & Forrester, N. E. (1983). Estimating Species Richness Using the Jackknife Procedure. *Biometrics, 39*(1), 1–11. https://doi.org/10.2307/2530802
10. Herculano-Houzel, S., Watson, C. R., & Paxinos, G. (2013). Distribution of neurons in functional areas of the mouse cerebral cortex reveals quantitatively different cortical zones. *Frontiers in Neuroanatomy, 7*. https://doi.org/10.3389/fnana.2013.00035
11. Magurran, A. E. (1988). *Ecological Diversity and Its Measurement*. Springer Netherlands. https://doi.org/10.1007/978-94-015-7358-0
12. Nader, K., Tasci, M., Ianevski, A., Erickson, A., Verschuren, E. W., Aittokallio, T., & Miihkinen, M. (2024). ScType enables fast and accurate cell type identification from spatial transcriptomics data. *Bioinformatics, 40*(7), btae426. https://doi.org/10.1093/bioinformatics/btae426
13. Vavrek, M. J. (2011). fossil: palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica, 14*(1), 1T. R package version 0.4.0.
14. Program, C. S.-C. B., Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S. M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., Cherry, J. M., Chi, T., Chien, J., Dorman, L., Garcia-Nieto, P., Gloria, N., Hastie, M., Hegeman, D., Hilton, J., ... Carr, A. (2023). CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data (p. 2023.10.30.563174). *bioRxiv*. https://doi.org/10.1101/2023.10.30.563174
15. Rood, J. E., Wynne, S., Robson, L., Hupalowska, A., Randell, J., Teichmann, S. A., & Regev, A. (2024). The Human Cell Atlas from a cell census to a unified foundation model. *Nature*, 1–2. https://doi.org/10.1038/s41586-024-08338-4
16. Schmitz, J. E., & Rahmann, S. (2024). A Review and Evaluation of Species Richness Estimation (p. 2024.10.09.615408). *bioRxiv*. https://doi.org/10.1101/2024.10.09.615408
17. Tabula Sapiens Consortium*, Jones, R. C., Karkanias, J., Krasnow, M. A., Pisco, A. O., Quake, S. R., Salzman, J., Yosef, N., Bulthaup, B., Brown, P., Harper, W., Hemenez, M., Ponnusamy, R., Salehi, A., Sanagavarapu, B. A., Spallino, E., Aaron, K. A., Concepcion, W.,

Gardner, J. M., ... Wyss-Coray, T. (2022). The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science (New York, N.Y.), 376*(6594), eabl4896. https://doi.org/10.1126/science.abl4896

18. Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: Large-scale single-cell gene expression data analysis. *Genome Biology, 19*(1), 15. https://doi.org/10.1186/s13059-017-1382-0
19. Yao, Z., van Velthoven, C. T. J., Kunst, M., Zhang, M., McMillen, D., Lee, C., Jung, W., Goldy, J., Abdelhak, A., Aitken, M., Baker, K., Baker, P., Barkan, E., Bertagnolli, D., Bhandiwad, A., Bielstein, C., Bishwakarma, P., Campos, J., Carey, D., ... Zeng, H. (2023). A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature, 624*(7991), 317–332. https://doi.org/10.1038/s41586-023-06812-z
20. Zeng, H. (2022). What is a cell type and how to define it? *Cell, 185*(15), 2739–2755. https://doi.org/10.1016/j.cell.2022.06.031