Estimating the completeness 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. Capitalizing on 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 various organs and their cellular compositions, we lack metrics to assess the completeness of the sequencing projects. In this cellular biodiversity analysis, we made use of the increasingly available single-cell data together with statistical methods, originally developed for assessing the species richness of ecological communities, to estimate the cellular diversity of any organ based on data from single-cell profiling technologies. Deriving from such cellular richness estimates, we established a statistical framework that allows assessment of the completeness of any large-scale single-cell profiling project, 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 single-cell sequencing projects, hence guiding more cost-efficient completion of the profiling on various human tissues.

Statistical estimator | Single-cell RNA sequencing | Cellular biodiversity | Log-normal distribution | Chao2 estimator

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 and 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. Even tough there does not exist a clear consensus on the level of granularity at which cell types should be classified, such cataloguing efforts and 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 most intuitive description of biodiversity, which 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. Since the number of species present in a sample is almost always an underestimate of the true population value, many statistical methods have been proposed to correct this bias and provide more accurate estimates of the species richness of an ecological population (Gwinn et al., 2016). We argue that such estimates could be used also to assess the richness of cellular ecosystems, and to provide completeness estimates for single-cell atlases, provided the statistical 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, along with upsampling methods (Schmitz & Rahmann, 2024). While these approaches have extensively been 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 or single-nucleus 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 consensus on how to best model cellular communities and species, selecting a suitable and accurate method for estimating species richness of cellular systems needs to be done with a great care.

As single-cell 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 methods to describe their biodiversity characteristics, such as the number and distribution of cellular species across tissues. Our systematic analysis reveals many key insights into how cell types are distributed along different organ systems, which together with statistical modelling provides an informative means to guide experimental single-cell profiling efforts. We show that cell communities in human can be generally modelled with a log-linear distribution. Importantly, using single-cell data together with statistical

species richness estimation methods, we investigate the completeness of currently available single-cell resources and suggest ways on how statistical analysis can enrich the ongoing efforts to build most comprehensive and complete single cell atlases.

Results

Cellular communities in human follow a log-linear 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, as of June 2025) across the 20 organs. Such an organ-level analysis demonstrated that there exist various degrees of both common and unique cell types across the organs (Figure 1A-B), and provides a principled means to understand organ development through organ-specific and shared cell types between organs. For example, the large number of different cell types both in the eye and lung highlight the complex functions of these organs (Figure 1A-B), whereas the large number of unique, organ-specific cell types in the eye reflects the evolutionary time needed to develop such a specialized organ (Figure 1A-B). To compare, the eye had the most specialized repertoire of unique cell types (58.1% unique cell types), while the lymph nodes had the least unique cell types (4.8%). This was also true for the 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 1C). 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 1C). Notably, most cell types (64%) were highly specialized and thus unique to a single organ.

Since the cell type uniqueness analysis demonstrated that cell types can be either shared or unique across organs, we next assessed the similarity of human organs. Indeed, the organ systems analysed from Tabula Sapiens data showed variable degrees of correlations in terms of the shared cell type compositions (Figure 1D). Although this 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 showed unexpectedly high correlation although not being closely related in terms of their primary function or anatomical location (Figure 1D). Notably, the cell type abundances in the organs displayed a heavy-tailed distribution, and since cell abundances in other cellular ecosystems in bacterial communities have been previously described to follow a log-linear distribution (Grilli, 2020), we tested whether cell type abundances in 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 (SFigure 1A), including the eye (Figure 1E). Finally, we wanted to understand the depth at which the individual organs have been profiled. Although sample coverage (the amount of population profiled) is a crucial piece of information to assess the statistical properties of the sample, this is not generally reported by any of the current single-cell atlas projects. We estimated the sample coverage for each of the 20 organs in Tabula Sapiens and observed drastic differences in the levels at which different organs have currently been profiled (Figure 1F). These differences underscore the need for standardized reporting of sample coverage to improve the cross-organ comparisons and highlight the organs systems that benefit from additional sampling.

In summary, our analysis of the Tabula Sapiens cell type abundance data gives an overview of the unique and shared distributions of individual cell types for better understanding of these cellular communities. Importantly, we show that organ systems in human are cellular ecosystems whose cell type abundances are approximately log-normally distributed. In light of previous research done on bacterial species (Grilli, 2020), our analysis supports the log-normal distribution as a natural law from which cellular communities can be derived.

Statistical framework for estimating cellular species richness. Currently, the completeness of most single-cell atlases is defined simply by 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 a non-uniform manner, such estimates are not appropriate for assessing atlas completeness. The count metrics could be improved by displaying them organ-by-organ, but even this approach cannot assess how complete the profiling efforts are. Since one of the immediate goals of single-cell profiling efforts is to exhaustively catalogue all human cell types, we argue that the most relevant way to assess any single-cell atlas completeness is to estimate how many cell types (species) remain to be unrecognized by the current profiling data. Because there is not yet an established consensus on the definition of cell types, in this work, we refer to distinct cell types or cell states as cellular species. To solve this statistical estimation problem (Figure 2A), we assessed the accuracy of different species richness estimation methods when applied to single-cell atlas data using a simulation-based statistical framework (Figure 2B).

We first established a simulated human body with defined organ systems and cellular species. Since the cell type uniqueness analysis demonstrated that cell types (species) can be either shared or unique across different organs (Figure 1A-C), we simulated realistically distributed and correlated cellular species across all the organs using statistical parameters inferred from Tabula Sapiens (see Methods).

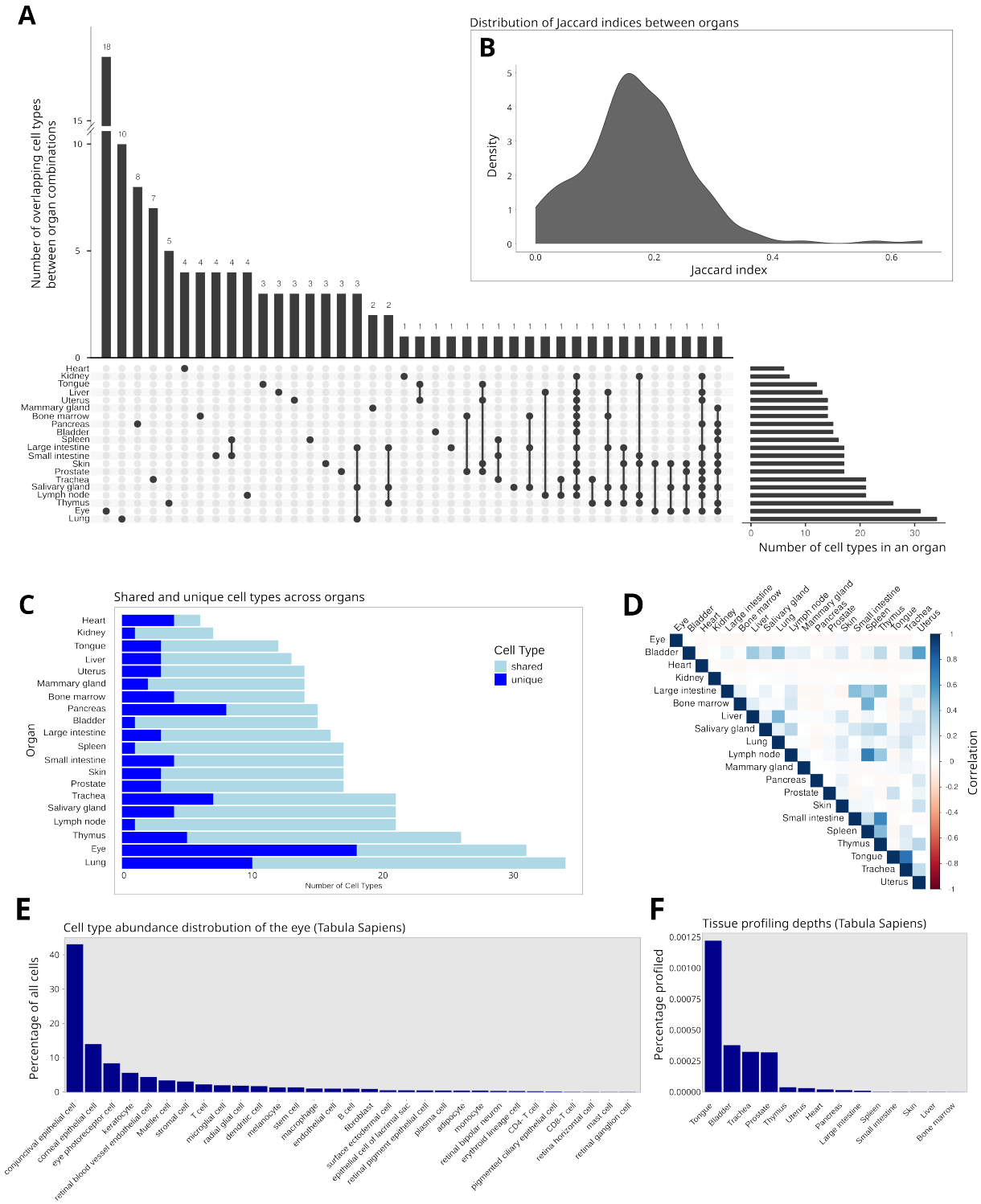

Fig. 1. Cellular biodiversity of human organ systems according to Tabula Sapiens data. (A) UpSet plot shows how cell types are distributed across the 20 organ systems in Tabula Sapiens. (B) As a metric for the organ uniqueness, distribution of pairwise Jaccard indices are displayed (0 indicates completely dissimilar organs, and 1 completely identical organs). (C) The number of shared and unique cell types across the 20 organs in Tabula Sapiens. (D) Pairwise Pearson correlations between organ pairs in Tabula Sapiens dataset, based on the counts of overlapping cell types between the two organs. (E) Log-normal distribution of the eye as an example. KS-statistic of 0.07, P = 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. The estimates of organ total cell numbers are from https://humancelltreemap.mis.mpg.de/ (SFigure 1B). Breast tissue is missing since estimates of tissue size were not available (as of June 2024).

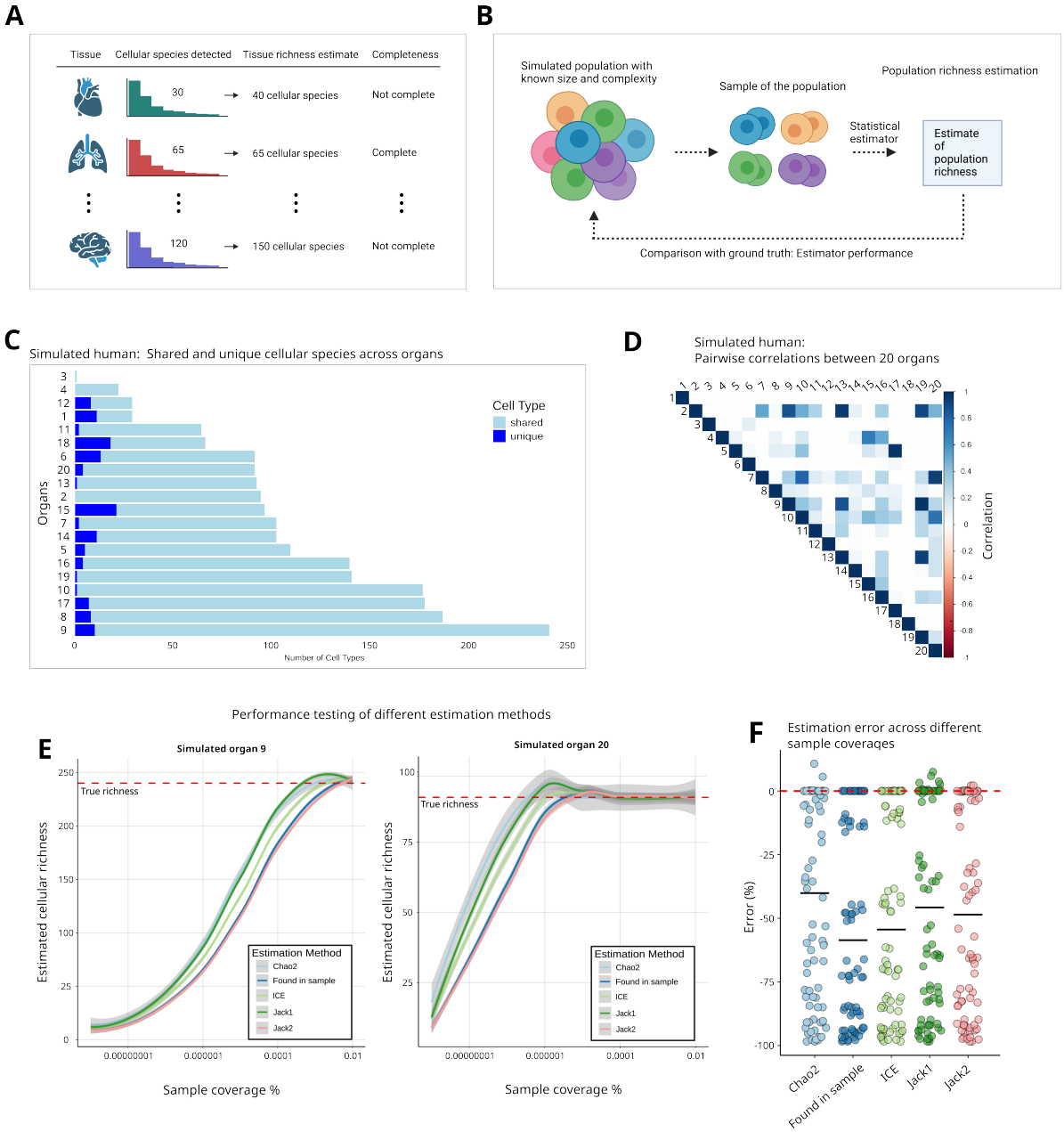

Fig. 2. Statistical framework to assess the accuracy of species richness estimators. (A) Schematic presentation of the estimation problem. Single-cell sample of a tissue is almost always an underrepresentation of the true tissue complexity. (B) Framework to assess the performance of different species richness estimators. (C) Simulated human: The number of shared and unique cellular species across different organs. (D) Simulated human: pairwise Pearson correlations between different organ pairs, based on the counts of overlapping cellular species. (E) Accuracy of different species richness estimators across different sample sizes from the simulated human. Shown for two simulated organs as examples, with different cell species compositions. The true richness indicates the number of cellular species present in the simulated organ. (F) Error of species richness estimators in a simulated organ profiled at different sample coverages ranging between 0.00000001%-0.01% using six samples. The horizontal lines indicate average errors across sample coverages.

We next used the simulated human as the ground truth, from which samples are drawn while accounting for the population complexity and number of cellular species. As expected, the cell species across simulated organs had distributions similar to those in Tabula Sapiens (Figure 2C). Interestingly, the fraction of cellular species unique to each organ was significantly higher than in the Tabula Sapiens data, providing evidence that many organ-specific species are yet to be discovered (Figure 2C). This is likely true for organs which are known to be highly specialized, such as the kidney, which in Tabula Sapiens currently lack granular cell definitions (Figure 1B). In addition, cellular species compositions correlated between organs ranging from low to high depending on the organs and was similar to the Tabula Sapiens data (Figure 2D, SFigure 1C).

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). We assessed the performance of four estimators by drawing differently sized triplicate samples to assess the number of unique cellular species present in the ground truth simulated organ. We performed this for organs with different and partly overlapping cellular compositions. Notably, after sampling 0.001% of the population, all the tested estimators started to stabilise towards the ground truth richness level, indicating that the estimators can efficiently estimate the true population value from an extremely small sample size, when 3 samples are available (Figure 2E). Among the tested estimators, Chao2 showed the most promising results by being the fastest to plateau towards the true population value, while not being prone to overestimation. We also studied the error percentage for each cellular richness estimator by drawing differently sized samples from the simulated organs (Figure 2F). Notably, the Chao2 estimator in this case resulted in a median reduction of 18% error, compared to the estimates based solely on observed cellular species (species recorded in the samples). Having more samples available had a positive impact on estimation accuracy, and showed that a fully accurate estimation of differently sized organs is possible, given enough samples were available (SFigure 1D). We therefore selected Chao2 for further use in case studies due to its desirable statistical properties.

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, log-normally 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 favourable statistical features when estimating cellular richness from a log-normally distributed population.

Estimating the completeness of large single-cell sequencing projects. Our previous results indicated that it is possible to confidently estimate the cellular richness using sample sizes of 0.001% or larger of the whole cellular population. Since many ongoing large-scale tissue profiling projects have recently made use of substantial sample sizes, we next wanted to assess in more detail the cellular richness of an organ, using repeated sampling from a real single-cell data set. We argue that estimating the cellular richness of an organ and comparing it with the currently detected amount of cellular species allows for completeness estimation of any large-scale single-cell sequencing project.

We used the open data provided in Yao et al. 2023, in which the whole mouse brain is dissected into different parts and profiled separately with repeated sampling. Specifically, the study provides four samples from the mouse isocortex, profiled using 10x genomics chromium (V2 chemistry), with approximately 250,000 cells per sample, making it one of the largest datasets when assessed in terms of 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 4% of the mouse isocortex. We first assessed the variability in the sampling after filtering out poor quality cells using standard metrics (see Methods). After filtering, cells were plotted and clustered, showing marked differences in sample complexity between the four samples (Figure 3A). Notably, even smaller cell clusters were often detected in multiple samples, indicating high completeness and good mixing between the different samples of the dataset (Figure 3B, SFigure 2A). To further test for possible batch effects between samples, we visualized GABAergic neurons, endothelial cells and oligodendrocytes, each of which was detected as a single cell cluster. This indicates good mixing between the samples and suggests minimal batch effects (Figure 3C).

Species richness estimation requires a criteria on how the cell types, here cellular species, are defined. Since there currently exists a variety of criteria for cell type classification, for the purpose of assessing the completeness of the mouse isocortex dataset, we simply considered the detected cell clusters as distinct cellular species. To study the frequency at which cell species are found in the four samples, we cross-tabulated the presence of each detected cell cluster across samples, creating a contingency table of the cell species frequency (Figure 3D, upper heatmap). This allowed us to assess both the consistency of cell species detection across samples and the extent of rare or sample-specific cell populations. As expected, the majority of cellular species were detected in most of the samples due to their larger size. However, we also observed sample-unique cellular clusters that were absent from the other samples (Figure 3D). Even though the Chao2 estimator uses the presence/absence data, we also investigated the cellular abundances showing that, occasionally and due to sampling, even abundant cellular species can be absent from a sample. However, absent cellular species are more common in less abundant cell species.

After defining how the different types of cells distribute across different samples, we used the Chao2 estimator to estimate the number of cellular species in the mouse isocortex using all the four samples available. Although the coverage needed to find all possible cellular species in the organ varies from organ to organ, on the basis of the simulation studies we expected 4% sample coverage and 3 samples to be generally

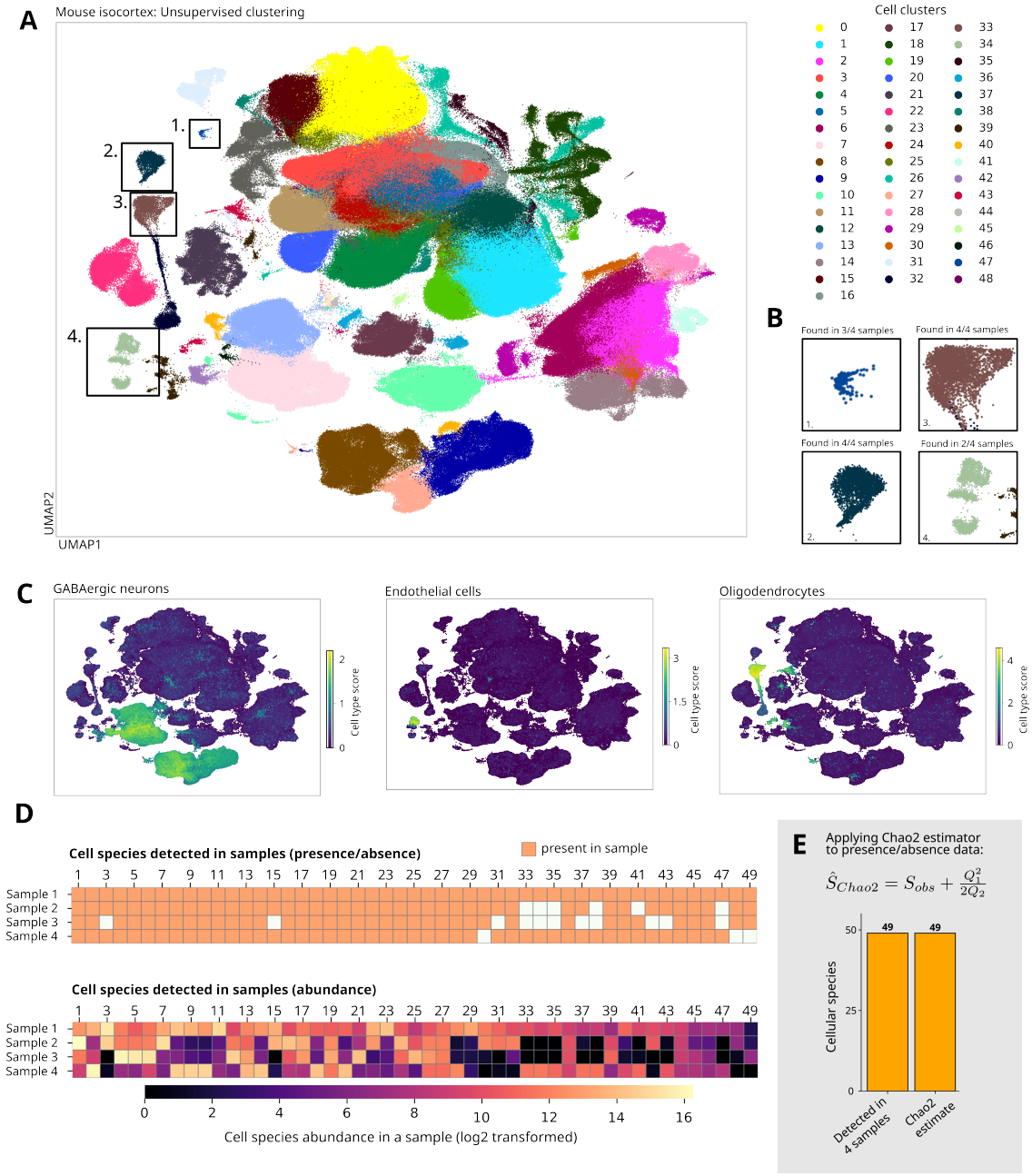

Fig. 3. Completeness estimation of a large single-cell mouse isocortex dataset. (A) one million cell dataset with 4 individual samples analyzed with 10x chromium V2 chemistry of the mouse isocortex from Yao et al., 2023. Leiden clusters are shown. Colors indicate different cell clusters (49 total clusters detected across 4 samples). (B) Smaller clusters labelled in panel A within the integrated dataset are detected in multiple samples, indicating high dataset completeness. (C) Individual clusters identified with RNA markers and matched to common cell types found in mouse isocortex. The cell type score is calculated by averaging gene expression values of the marker genes separately for each cell type (see Methods). (D) Cellular species detection in the four isocortex samples (upper panel: presence/absence and lower panel: abundance data). Cellular species correspond here to to the detected cell clusters. To avoid undefined values for zero counts, a small constant (ϵ) was added before transformation (log_2(x + ϵ)) (E) Chao2 estimate of the cellular richness of mouse isocortex dataset using data from Yao et a., 2023. S_{obs} is the observed cellular 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.

enough to detect all the cellular species in the mouse isocortex. Indeed, applying the Chao2 estimator on the mouse isocortex data led to the conclusion that this organ has a total of 49 cellular species (Figure 3E). This suggests that when the cellular species in mouse isocortex are defined as here (i.e., all detected cell clusters), single-cell profiling of more samples will not reveal additional cellular species. We therefore consider that the healthy mouse isocortex has been fully profiled. The same analysis can be applied to any ongoing single-cell sequencing or large-scale atlas projects, providing estimates of their current completeness stage.

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 remains challenging. Previous works have estimated different aspects of the human cell populations, such as cellular size distribution and the number of total cells in the body (Bianconi et al., 2013; Hatton et al., 2023). In contrast, we focused here on combining single-cell datasets with statistical species richness estimation to provide a framework that enables cellular richness estimates (the number of different cell species). In particular, we provided methods for completeness estimation of any large-scale single-cell profiling project.

Currently, substantial investments are being made in mapping various cell types, their locations, and gene regulation across human tissues using single-cell RNA sequencing and related technologies. Although single-cell and single-nucleus RNA sequencing are currently the most practical and scalable technologies 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). In addition, different protein-based techniques which do not suffer from RNA drop-outare rapidly emerging. Currently, cell type and state definitions are hard to disentangle using only data-driven approaches, especially in heterogeneous and dynamical 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, today there still exist multiple definitions of cell types and states. Therefore, 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 of a cell type.

Despite the significant interest in profiling various types of healthy and diseased tissues, there has so far been little efforts to estimate the total number of unique cellular species within these systems. Most studies have focused 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 the true cellular complexity of a tissue, we here sought to benchmark the 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 in other applications (Schmitz & Rahmann, 2024), herein we focused on benchmarking estimators that would work especially with single-cell profiling data.

Our analyses suggest that the cellular species in most, if not all, human tissue systems are log-normally distributed. This led us to test different species richness estimators on log-normally 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 made use of a previously published, high-quality 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 is expected to make the estimates more accurate.

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 mainly depends on the abundance of the cell. This is especially true for single-nucleus profiling technologies. In addition, assessing sample coverage for cells is much easier than for animal species, which is expected to lead to more robust estimates. Importantly, derived estimates of a dataset or atlas completeness are especially useful when building large single-cell resources, and we therefore suggest continuously running species richness estimates in regular intervals on every profiled organ, with the aim of understanding the level of completeness for each organ that is being profiled. As single-cell resources continue to expand and the profiling depths across organs become more balanced, this work will provide the basis of future estimates of even the whole human body.

Supplementary Note 1: Methods

A. Simulations of organs and human. Since the organs are partially correlated, we simulated a human body with a defined number of organs (n=20), and cell types by drawing a sample of size 5 trillion cells from a multivariate log-normal distribution. To define organ variable dependencies, the covariance structure of this distribution was derived from

the empirical covariance matrix of Tabula Sapiens. The user-defined parameters in the simulation were:

- **n**: number of cells in the whole organism
- **m**: number of organs
- **k**: number of cell types

To simulate a human body with an organ system, we first generated data in log-transformed space, where the organ abundances followed a multivariate normal distribution. This was done using the mvrnorm() function from the MASS library in R. We then applied an exponential transformation to obtain log-normally distributed organ abundances.
To simulate realistic dependencies between organs, we analyzed Tabula Sapiens data, assessing its organ means and covariance structure in both normal and log-normal spaces. We found that the means of log-transformed organ abundances followed a normal distribution, allowing us to simulate a corresponding vector of means for the normally-distributed space. However, because exponentiation is a nonlinear transformation, it alters dependencies between the variables.
To retain the empirical covariance structure after back-transformation, we derived a mathematical relationship between the covariance matrices in normal and log-normal spaces. For this, we considered Y as a multivariate log-normal random vector:

$$Y_i = \exp(X_i), \quad \text{where } X \sim \text{MVN}(\mu, \Sigma). \tag{1}$$

The covariance for two dependent variables (Y_i, Y_j) in logarithmic space is by definition:

$$\text{Cov}(Y_i, Y_j) = \mathbb{E}[Y_i Y_j] - \mathbb{E}[Y_i]\mathbb{E}[Y_j]. \tag{2}$$

Because $Y_i = \exp(X_i)$, the expected value of $Y_i Y_j$ is:

$$\mathbb{E}[Y_i Y_j] = e^{\mu_i + \mu_j + \frac{1}{2}(\sigma_i^2 + \sigma_j^2 + 2\sigma_{ij})}, \tag{3}$$

where σ_{ij} is the covariance in normal space. The relationship between and the covariance in log-normal space depends on the covariance in normal space:

$$\text{Cov}(Y_i, Y_j) = e^{\mu_i + \mu_j}(e^{\sigma_{ij}} - 1)e^{\frac{1}{2}(\sigma_i^2 + \sigma_j^2)}. \tag{4}$$

Rearranging equation (4), we can express covariance in normal space as:

$$\sigma_{ij} = \log\left(\frac{\text{Cov}(Y_i, Y_j)}{e^{\mu_i + \mu_j} e^{\frac{1}{2}(\sigma_i^2 + \sigma_j^2)}} + 1\right) \tag{5}$$

This enables covariance matrix (Σ) back-transformation while retaining dependencies between the variables.

B. Species Richness Estimators. We reasoned that statistical estimators which use presence/absence data, rather than abundance data, are more suitable for the use with single-cell or single-nucleus RNA sequencing data. Many abundance data-based estimators rely on singlet-detection, which can be difficult with current RNA-based methods. Instead, we focused on benchmarking statistical estimators which use presence/absence data and that rely on repetitive sampling.

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

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

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 (Chao, 1987).

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

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 for estimating 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 the statistical analyses were run using `R` with `R Studio`. Different statistical estimators were used through `R` with the CRAN package `fossil` (Vavrek, 2011).

Supplementary Note 2: Single-Cell Analysis Pipeline

The single-cell analysis was conducted using **Python** and **Scanpy** (Wolf et al., 2018). Briefly, raw sequencing data was filtered, retaining only cells with mitochondrial gene content below <5% and with at least 200 detected genes.

A. Clustering and Cell Type Annotation. Data clustering was performed using the **Louvain algorithm** implemented in Scanpy. Each identified cluster was defined using a marker panel via the **scType Python implementation**, which allows for consistent cluster identification across parallel samples (Nader et al., 2024).

B. Data Integration and Cell Type Scores. To assess data integration across multiple samples, cell type scores were calculated by averaging the expression of a predefined set of RNA markers for each cell type. The marker genes used were:

- **Inhibitory Neurons (GABAergic neurons)**: Gad1, Gad2, Pvalb, Sst, Vip, Nkx2-1
- **Oligodendrocytes**: Mog, Mbp, Cnp, Plp1, Olig1
- **Endothelial Cells**: Pecam1, Cldn5, Kdr

C. 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
- **Human cell treemap**: Organ size estimates. Data accessed on 30.05.2024
- **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.

Supplementary material

Supplementary material can be found from here.

ACKNOWLEDGEMENTS
Tha authors thank Yunus Vilkkavaara-Cankocak and Yidian Chu for their support while preparing this manuscript. All single-cell data analytics were performed using a computing cluster. The authors acknowledge the **CSC – IT Center for Science, Finland**, for providing the computational resources.

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