Community-driven ELIXIR activities in single-cell omics

Paulo Czarnewski1, Ahmed Mahfouz2, Raffaele A. Calogero3, Patricia M. Palagi4, Laura Portell-Silva5, Asier Gonzalez-Uriarte6, Charlotte Soneson4,6, Tony Burdett7, Barbara Szomolay8, Pavankumar Videm9, Hans-Rudolf Hotz6, Irene Papaetheodorou7, John M. Hancock10, Björn Grüning9, Wilfried Haerty11, Roland Krause12, Salvador Capella-Gutierrez5, Brane Leskošek10, Luca Alessandri3, Maddalena Arigoni3, Tadeja Rezen10, Alexander Botzki13, Polonca Ferk10, Jessica Lindvall1, Katharina F. Heil14, Naveed Ishaque15, Eija Korpelainen16

1Science for Life Laboratory, Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Stockholm University, Solna, Sweden
2Department of Human Genetics, The Netherlands & Delft Bioinformatics Lab, Leiden University Medical Center, Delft University of Technology, Leiden, Delft, The Netherlands
3Bioinformatics and Genomics unit, Dept. Molecular Biotechnology and Health Science, University of Torino, Torino, Italy
4SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland
5Barcelona Supercomputing Center, Barcelona, Spain
6Friedrich Miescher Institute for Biomedical Research, Basel, Switzerland
7European Bioinformatics Institute, Hinxton, UK
8Division of Infection and Immunity, School of Medicine, Cardiff University, Wales, UK
9Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany
10Institute of Biochemistry and Molecular Genetics, Faculty of Medicine, University of Ljubljana, Ljubljana, Slovenia
11Earlham Institute, Norwich, UK
12Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Luxembourg, Luxembourg
13VIB Bioinformatics Core, Ghent, Belgium
14ELIXIR Hub, Wellcome Genome Campus, Hinxton, UK
15Digital Health Center, Berlin Institute of Health at Charité, Universitätsmedizin Berlin, Berlin, Germany
16CSC – IT center for science, Espoo, Finland

First published: 29 Jul 2022, 11(ELIXIR):869
https://doi.org/10.12688/f1000research.122312.1

Abstract
Single-cell omics (SCO) has revolutionized the way and the level of resolution by which life science research is conducted, not only impacting our understanding of fundamental cell biology but also providing novel solutions in cutting-edge medical research. The rapid development of single-cell technologies has been accompanied by the active development of data analysis methods, resulting in a plethora of new analysis tools and strategies every year. Such a rapid
development of SCO methods and tools poses several challenges in standardization, benchmarking, computational resources and training. These challenges are in line with the activities of ELIXIR, the European coordinated infrastructure for life science data. Here, we describe the current landscape of and the main challenges in SCO data, and propose the creation of the ELIXIR SCO Community, to coordinate the efforts in order to best serve SCO researchers in Europe and beyond. The Community will build on top of national experiences and pave the way towards integrated long-term solutions for SCO research.

Keywords
Single cell, multi-omics, spatial transcriptomics, FAIR, data analysis, data standards, training, computing infrastructure

Corresponding authors: Paulo Czarnewski (paulo.czarnewski@scilifelab.se), Naveed Ishaque (naveed.ishaque@bih-charite.de), Eija Korpelainen (eija.korpelainen@csc.fi)

Author roles: Czarnewski P: Data Curation, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; Mahfouz A: Writing – Original Draft Preparation, Writing – Review & Editing; Calogero RA: Writing – Original Draft Preparation; Palagi PM: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing; Portelli-Silva L: Writing – Original Draft Preparation, Writing – Review & Editing; Soneson C: Writing – Review & Editing; Burdett T: Conceptualization; Szomolay B: Writing – Original Draft Preparation, Writing – Review & Editing; Videm P: Writing – Review & Editing; Hotz HR: Writing – Review & Editing; Papatheodorou I: Conceptualization; Hancock JM: Writing – Review & Editing; Grüning B: Writing – Review & Editing; Haerty W: Writing – Review & Editing; Krause R: Writing – Review & Editing; Capella-Gutierrez S: Writing – Original Draft Preparation; Leskošek B: Writing – Review & Editing; Alessandri L: Writing – Original Draft Preparation; Arigoni M: Writing – Original Draft Preparation; Rezen T: Writing – Review & Editing; Botzki A: Writing – Review & Editing; Ferk P: Writing – Review & Editing; Lindvall J: Conceptualization; Heil KF: Conceptualization, Writing – Review & Editing; Ishaque N: Writing – Original Draft Preparation, Writing – Review & Editing; Korpelainen E: Conceptualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was supported by ELIXIR, the research infrastructure for life-science data. P.C. is financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2022 Czarnewski P et al. This is an open access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

How to cite this article: Czarnewski P, Mahfouz A, Calogero RA et al. Community-driven ELIXIR activities in single-cell omics [version 1; peer review: awaiting peer review] F1000Research 2022, 11(ELIXIR):869 https://doi.org/10.12688/f1000research.122312.1

First published: 29 Jul 2022, 11(ELIXIR):869 https://doi.org/10.12688/f1000research.122312.1
Introduction

Single-cell omics (SCO) is an umbrella term that encompasses multiple technologies that are able to profile various omic modalities at the single-cell level. These high-throughput single-cell approaches have rapidly become the method of choice over traditional bulk methods which average data across a population of cells. Single-cell approaches are better suited for characterizing many biological phenomena and exploring cellular heterogeneity such as characterisation of rare cell types and diverse cell states. Besides single-cell/nucleus RNA sequencing, SCO approaches include nuclear epigenetic profiling such as chromatin accessibility, histone profiling, DNA methylation, chromatin conformation as well as high throughput single-cell proteomics. Recent developments allow simultaneous profiling of two or more of the aforementioned modalities, opening up unprecedented opportunities to study diverse processes such as development, gene expression dynamics, tissue heterogeneity and disease pathogenesis. More recently, several approaches have been developed to deliver spatial resolution of single cell expression within tissues, adding another layer of complexity. While several grand challenges in exploratory data analysis remain, a parallel issue is the provision of infrastructure to support such analysis in the rapidly developing field.

New SCO profiling technologies are mushrooming, and new analysis methods are published weekly (Figure 1a). As the scale and modality of data sets grow, new computational methods are required. The data also need to be stored and annotated in a standardized manner in order to enable their reuse. This in turn adds an extra challenge and makes it hard for most institutes to handle alone, and calls for international collaboration on training, tools, compute, data, interoperability and standardization in SCO.

ELIXIR, the European infrastructure for life sciences data, brings Europe’s national centers and core bioinformatics resources into a single, coordinated infrastructure. This intergovernmental organization currently has 23 Nodes, and facilitates collaboration between its member institutes and researchers with two intersecting organizational groupings: Platforms and Communities. ELIXIR Platforms (Data, Interoperability, Tools, Compute and Training) provide services,

---

Figure 1. The current landscape in SCO surveyed up until January 2022. (a) Current count of articles using SCO technologies and cumulative number of cells sequenced and deposited in public databases. (b) Number of tools developed specifically to work with SCO. (c) Most common SCO molecular profiling technologies mentioned in publications. (d) Top 15 most targeted categories for software development in SCO. (e) Number of tools developed for SCO, split by which scripting languages are used. Data were taken from public databases. See the Data and Software Availability section for details.
Table 1. Overview of past single-cell course collaborations between ELIXIR Nodes. Tutorial training style refers to teaching bioinformatics by following pre-made sequential analysis steps with code. PBL training style refers to teaching using project-based learning, where students develop their own analysis code to solve analysis tasks. All courses listed below were taught in English.

<table>
<thead>
<tr>
<th>Course name, link and year</th>
<th>ELIXIR node</th>
<th>Training style</th>
<th>Computing Language</th>
<th>Recorded lectures</th>
<th>FAIR Tool</th>
<th>Length</th>
<th>Technologies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Advanced topics in single cellomics (2021)</td>
<td>SE, CH</td>
<td>PBL</td>
<td>R and python</td>
<td>Yes</td>
<td>Docker</td>
<td>5 days</td>
<td>scRNAseq, scATACseq, ST, Deep Learning</td>
</tr>
<tr>
<td>Single Cell School (2019)</td>
<td>SE, CH</td>
<td>Tutorial</td>
<td>R</td>
<td>No</td>
<td>-</td>
<td>-</td>
<td>scRNAseq, CyTOF, scProteomics</td>
</tr>
<tr>
<td>Galaxy single-cell omics training materials (ongoing)</td>
<td>DE, UK, CH</td>
<td>Tutorial</td>
<td>-</td>
<td>Yes</td>
<td>Galaxy</td>
<td>3-5 days</td>
<td>scRNAseq</td>
</tr>
<tr>
<td>Single-cell RNAseq analysis using R (2021)</td>
<td>UK EMBL-EBI</td>
<td>Tutorial</td>
<td>R</td>
<td>No</td>
<td>-</td>
<td>5 days</td>
<td>scRNAseq</td>
</tr>
<tr>
<td>Gene Expression at Spatial Resolution (2021)</td>
<td>EMBL-EBI SE</td>
<td>Tutorial</td>
<td>-</td>
<td>Yes</td>
<td>-</td>
<td>4 Days</td>
<td>Spatial transcriptomics (10X Visium)</td>
</tr>
<tr>
<td>Single-cell RNAseq data analysis with R (2019)</td>
<td>FI, SE, NL, DE, NO, FR</td>
<td>Tutorial</td>
<td>R</td>
<td>Yes</td>
<td>Conda</td>
<td>3 days</td>
<td>scRNAseq</td>
</tr>
<tr>
<td>Single-cell RNAseq data analysis with Chipster (2020)</td>
<td>FI, LU</td>
<td>Tutorial</td>
<td>-</td>
<td>Yes</td>
<td>Chipster</td>
<td>3 days</td>
<td>scRNAseq</td>
</tr>
<tr>
<td>Single-cell RNAseq data analysis with Chipster (ongoing)</td>
<td>FI, NL, SE, FR</td>
<td>Tutorial, eLearning</td>
<td>-</td>
<td>Yes</td>
<td>Chipster</td>
<td>3 days</td>
<td>scRNAseq</td>
</tr>
<tr>
<td>Single cell RNA-seq analysis workshop (2022)</td>
<td>SE, NL</td>
<td>Tutorial</td>
<td>R and Python</td>
<td>Yes</td>
<td>Conda</td>
<td>5 days</td>
<td>scRNAseq, ST</td>
</tr>
</tbody>
</table>
and ELIXIR Communities identify the needs of domain- or technology-specific research around a theme. There are currently 13 Communities ranging from Metabolomics and Proteomics to Federated Human Data and Galaxy.

Many ELIXIR Nodes already have single-cell facilities, and others are setting them up. The Nodes are facing a huge demand for single-cell data analysis and training, and some knowledge transfer between the Nodes already exists. Examples of past data analysis courses co-organized by ELIXIR Nodes are listed in Table 1. The Nodes have also co-organized workshops to discuss FAIR data management and best training practices for SCO. Together, 17 ELIXIR Nodes proposed to create the ELIXIR SCO Community to connect these grass-roots efforts and strengthen European and international cohesion in SCO.

**Landscape of SCOs**

**Technologies**

Single-cell omics technologies have seen widespread adoption since its announcement as *Nature* Method of the Year in 2013. The most widely used SCO technologies are single-cell RNA-seq, single-nucleus RNA-seq, and single-cell ATAC-seq. The early days of these technologies were dominated by heterogeneous implementations of handling, preparation and sequencing protocols, which left its mark in the large number of software tools that had been developed, which in part led to lack of standardization of data and metadata. In recent years, we have seen a number of technology providers prevail for instrumentation (e.g. Fluidigm, 10x Genomics), reagents (e.g. ThermoFisher, QiaGen, Roche), and sequencing (e.g. Illumina, MGI, ONT and PacBio). This has led to the community converging around a few workflows based around the popular 10x Genomics Chromium and SmartSeq chemistries (Figure 1c), which have been exploited for large scale sequencing efforts such as the Human Cell Atlas (HCA) and Human BioMolecular Atlas (HuBMAP), that, in turn, have resulted in large investments into solving the sample handling, data integration and data management problems underpinning the vast array of data being generated.

While it has been a decade since the SCO technologies have taken the center stage in unraveling bio-molecular heterogeneity, technology developers are far from stagnant and we are seeing rapid evolution of these technologies. We witness the adoption of single-cell genomics (e.g. MissionBio Tapestri) and some progress in the field of single-cell proteomics. Single-cell multimodal omics was announced as *Nature* Method of the Year 2019. These assays provide multiple readouts that can be used to define cells. For example, antibody profiling (e.g. CITE-seq) allows scientists to contextualize novel cell types and states in the context of well-established cell biology markers; immune repertoire profiling can also link how transcriptional profiles of immune cells differ based on the receptor specificity; true multimodal omics such as simultaneous profiling of chromatin accessibility, DNA methylation and transcriptomics (e.g. scNMT-seq) allows us to also decipher the regulatory changes that underpin the transcriptional landscape of cells.

In 2020, spatially resolved transcriptomics (SRT) was announced as *Nature* Method of the Year. These technologies fall under three broad areas of laser capture microscopy combined with single-cell sequencing (e.g. GeoMX DSP, Tomo-seq), in situ capture arrays (e.g. ST, Visium, Slide-seq, HDST), and image-based single-molecule expression quantifications (e.g. *in situ* sequencing, seqFISH+, Molecular Cartography, MERFISH+). Despite its youth, there are over 20 SRT profiling technologies, which are also expanding into other omics modalities, such as proteomics (e.g. CODEX) and metabolomics. However, the community quickly realized the potential of this technology and early efforts were pushed by the Chan Zuckerberg Initiative to harmonize the field by funding efforts such as StarFish, a platform to uniformly process raw single molecule SRT data, and the SpaceTX consortium, which aim to benchmark and harmonize data from various SRT platforms and analytics methods. Efforts to make comprehensive cell atlases available to the community now facilitate more single-cell study designs to include perturbation and lineage tracing experiments (e.g. via CRISPR).

While the standards for scRNA-seq have by-and-large converged, the extension of single-cell technologies to new modalities and experimental setups places even more emphasis on establishing adaptable and extensible standards.

**Analysis tools**

There is a large number of single-cell analysis methods and tools available that cover a wide range of analysis steps. In January 2022, the scRNA-tools database recorded nearly 1,200 tools divided over more than 30 categories (Figure 1b and d). Computational and analytical challenges in single-cell genomics have been discussed extensively. The analysis steps vary depending on the modality of single-cell data. For the most widely used modality, transcriptomic data, the community has converged on a consensus regarding the analysis steps. Yet, even some foundational steps remain active areas for research, such as how to normalize scRNA-seq data or how best to perform differential expression analysis. Also, annotation of cell types varies drastically between studies, with many resorting to *ad hoc* decisions. A complete atlas of all cell types would be required to improve the standardization of cell-type nomenclature and
ontologies (e.g. Cell Ontology, UBERON).52 Scientists analyzing SCO data need to have sufficient information on the strengths and limitations of the available analysis tools in order to select the most suitable ones for their data and purpose. However, systematic comparison of these tools is challenging, especially given the ever-increasing number of methods and their parameter combinations.

Studies have identified essential guidelines for benchmarking computational methods43–49 and reviewed published benchmarking studies of computational tools for omics data, highlighting the advantages and limitations of benchmarking across various domains of the life sciences.50 Systematic benchmarking frameworks can enable crowdsourcing and community challenges, which have been a successful means for fostering community creativity and expertise to address open problems.51 In a concerted effort to address “grand challenges” for the SCO community,11 the Open Problem in Single-Cell Analysis group57 is devising competitions to address those challenges, e.g. the Multimodal Single-Cell Data Integration competition (NeurIPS 2021).

Despite community efforts, the major challenges facing SCO benchmarking studies are the lack of appropriate experimental data and/or realistic simulated data that can be used for benchmarking, as well as the lack of agreed-upon measures to evaluate different methods. There is also a need for a common platform to conduct benchmark studies. The Open Problems NeurIPS challenge provides a leading example for evaluating methods using common datasets, performance metrics, as well as providing a compute infrastructure to run these methods. However, there is still a need for platforms that allow for continuous update of results as new tools and/or metrics become available and to dynamically respond to the needs of individual communities within the life sciences.

Currently available tools and pipelines differ in their usability. While the majority require programming knowledge, several pipelines provide GUIs for users without programming experience (e.g. Galaxy, Chipster).53,54 Most tools are available as R and Python packages or as a collection of scripts on GitHub (Figure 1c). To keep up with the technology developments, these methods and tools are continuously updated. Yet, maintaining tools and providing support is often challenging for research groups. Interoperability between methods and tools is limited despite efforts by popular packages such as Seurat55 to provide wrappers around other tools. However, frequent updates to tools to keep up with technology developments (e.g. updating single-cell objects to cater for multi-modal data) limits interoperability, emphasizing the importance of a concerted effort to address robust data and metadata standards.

Standards and research data management

The major factor in realizing interoperability is the definition and adoption of robust data format standards. While more than 1,000 SCO tools exist, there is broad acceptance of widely adopted raw data standards (e.g. FASTQ, FAST5, BAM, CRAM) and convergence to a few processed data formats (e.g. tab-separated files, AnnData, HDF5, loom, SingleCellExperiment, Seurat). The data formats and structures employed by some of the most popular tools for SCO data analysis55,56 have had to change to adapt to new technologies that rendered previous formats inadequate. While these changes in data formats are frustrating for maintaining data analysis workflows, they are necessary for keeping up to date with the rapid technological developments in this field. This places a strong emphasis on planning to adapt to changes by employing extensible structures that do not break the chain of backwards compatibility.

Furthermore, metadata standards and minimal reporting guidelines enable the appropriate archiving and subsequent reuse of SCO data. For some specific library construction or sequencing technologies, provision of platform specific metadata is routine and standardized (e.g. for the 10x Chromium), however additional care is required for in-house solutions and for reporting metadata for other parts of the experimental design. Establishing the Minimum INformation about a SEQuencing Experiment (MINSEQE) guideline was an important achievement for reporting metadata for sequencing data.57 Recently the Minimum Information about a Single-Cell Experiment (minSCe) guidelines were established,58 which defines 48 attributes that describe the biosource, isolation method, protocols, library construction, SingleCellExperiment, Seurat). The data formats and structures employed by some of the most popular tools for SCO data analysis55,56 have had to change to adapt to new technologies that rendered previous formats inadequate. While these changes in data formats are frustrating for maintaining data analysis workflows, they are necessary for keeping up to date with the rapid technological developments in this field. This places a strong emphasis on planning to adapt to changes by employing extensible structures that do not break the chain of backwards compatibility.

Standards and research data management

The major factor in realizing interoperability is the definition and adoption of robust data format standards. While more than 1,000 SCO tools exist, there is broad acceptance of widely adopted raw data standards (e.g. FASTQ, FAST5, BAM, CRAM) and convergence to a few processed data formats (e.g. tab-separated files, AnnData, HDF5, loom, SingleCellExperiment, Seurat). The data formats and structures employed by some of the most popular tools for SCO data analysis55,56 have had to change to adapt to new technologies that rendered previous formats inadequate. While these changes in data formats are frustrating for maintaining data analysis workflows, they are necessary for keeping up to date with the rapid technological developments in this field. This places a strong emphasis on planning to adapt to changes by employing extensible structures that do not break the chain of backwards compatibility.

Furthermore, metadata standards and minimal reporting guidelines enable the appropriate archiving and subsequent reuse of SCO data. For some specific library construction or sequencing technologies, provision of platform specific metadata is routine and standardized (e.g. for the 10x Chromium), however additional care is required for in-house solutions and for reporting metadata for other parts of the experimental design. Establishing the Minimum INformation about a SEQuencing Experiment (MINSEQE) guideline was an important achievement for reporting metadata for sequencing data.57 Recently the Minimum Information about a Single-Cell Experiment (minSCe) guidelines were established,58 which defines 48 attributes that describe the biosource, isolation method, protocols, library construction, SingleCellExperiment, Seurat). The data formats and structures employed by some of the most popular tools for SCO data analysis55,56 have had to change to adapt to new technologies that rendered previous formats inadequate. While these changes in data formats are frustrating for maintaining data analysis workflows, they are necessary for keeping up to date with the rapid technological developments in this field. This places a strong emphasis on planning to adapt to changes by employing extensible structures that do not break the chain of backwards compatibility.

Given the fast pace of technological developments in the SCO field, the community has identified that both adaptability and extensibility are key considerations in defining sustainable standards. This has been achieved in the field of medical
imaging with the Digital Imaging and Communications in Medicine (DICOM) format,\textsuperscript{60} which has been constantly extended and updated without breaking backwards compatibility for nearly 30 years. However, this level of flexibility was only achieved by the third version of the DICOM standard, 10 years after its initial inception, and it was a concerted effort between medical and trade associations. Part of the successful adoption of the DICOM format is that despite all major medical imaging players having their own proprietary formats, they provide an interface to the DICOM format. In order for the SCO community to reach a similar level of interoperability as has been achieved in medical imaging, technology providers and tool developers should also either adopt the most common standards in the SCO community, or provide interfaces to them. While it is not clear how the current landscape of SCO data and metadata standards will stand the test of time, some aspects that will determine their success with be their ability to adapt to change (e.g. through using extensible formats such a JSON), used of controlled nomenclature (e.g. utilising ontologies for defining attributes), and adopting versioning (e.g. semantic versioning).

**Training**

Upskilling life scientists to analyze SCO data is a moving target, given the fast development of the field. The cutting edge analysis methods for SCO data tend to be rather computationally complex, making them harder to grasp for life scientists who typically lack a solid background in mathematics, statistics and machine learning and often R/Python skills too.

Trainers, on the other hand, find themselves updating training materials constantly and, in general, struggle to keep up with the fast development of new analysis methods in order to choose what to teach. To make things worse, often only a small fraction of their working time is dedicated to training, or training is offered on a voluntary basis on top of their workload. It is therefore not surprising that even though single-cell courses are offered by several ELIXIR Nodes, many flavors of SCO are not yet covered. For example, courses on single-cell epigenetic, multi-omics as well as image-based spatially resolved SCOs are still rare. The demand for training continues to grow, but the lack of competent experts with enough training experience and time available is a major bottleneck in scaling up training provision. While pedagogical train-the-trainer (TtT) courses\textsuperscript{61,62} can empower experts to feel more comfortable to teach, the constant evolution of the SCO field can intimidate newcomers.

There are also more practical hurdles: the analysis of single-cell data requires a sophisticated computational environment with many tools and their dependencies, often requiring high-end computational resources. These environments have to be ready-to-go or at least easy to set up, and reproducible across heterogeneous hardware infrastructure, allowing the participants to re-run the practical and to analyze their (probably much larger) own data in their own setting. It is also challenging to find good training datasets that are small enough to be run in a class but meaningful enough to prove the concepts.

**Alignment with ELIXIR Platforms and Communities**

The ELIXIR SCO Community will bring together current efforts and produce guidelines and training. It creates a communication channel to exchange experiences, collect user requests and feedback and push for standards. Given its needs for training, tools, compute, data and interoperability, the SCO Community aligns well with all the ELIXIR Platforms. It also has synergies with the ELIXIR Human Data Communities and the Galaxy Community, as well as some ELIXIR Focus Groups like Cancer Data and FAIR Training.

**Training platform**

Upskilling scientists in SCO data analysis and standards lies at the heart of the ELIXIR SCO Community, and particular efforts will be made to make the training scalable and FAIR in coordination with the ELIXIR Training platform (Table 2).

The SCO Community will ensure that training materials and expertise are shared efficiently and following FAIR and open research principles.\textsuperscript{63,64} We will collaborate with ELIXIR’s Training Portal TeSS\textsuperscript{55} to establish a well-curated SCO training portal, listing national and international training providers, web resources and upcoming training events. To help the current trainers and encourage new ones, we will annotate training materials with appropriate metadata, curate training datasets, provide detailed explanation on how to run courses, and share best practices and best ways to teach the more advanced concepts. In order to identify SCO areas which lack sufficient training, we will participate in designing the annual training gap survey by the Training Platform, and also perform more detailed SCO training surveys if needed. We will regularly host trainer workshops targeting the areas identified as lacking sufficient training to exchange experiences and discuss materials.

Anyone should be able to learn about SCO data analysis independently of time and place. To make the training scalable, lectures and video tutorials will be recorded for asynchronous learning, and combined into modular eLearning courses. Resources will be gathered and annotated on a single site for easy discovery. In addition to organizing training in SCO
data analysis and standards to complement ELIXIR Nodes activities, we will provide training in best practices for trainers (TtT) to increase the number of expert trainers.

The course software installation challenge will be addressed together with the ELIXIR Tools Platform, as described below, using Conda environments, containers and Notebooks. Both Galaxy\textsuperscript{66} and Chipster\textsuperscript{54} offer specific training access and a comprehensive collection of training materials. There is no setup required, and the same environment is available when analyzing one’s own data after the course.

**Tools platform**

SCO data analysis typically requires a large number of tools and their dependencies. The installation challenge can be eased by providing Conda environments\textsuperscript{54} and containers\textsuperscript{58} for SCO, in alignment with the work developed in the Tools Platform.
Platform’s Packaging, containerisation and deployment activity. Also, RStudio or Jupyter Lab based SCO Notebooks can be made to support courses and self-study. The Community will develop cloud-deployable analysis pipelines for SCO data and make them available also through the web-based Galaxy Single Cell Omics, Galaxy Human Cell Atlas project, and Chipster analysis platforms for researchers lacking programming skills. The analysis pipelines will be deposited in WorkflowHub for easy discovery, re-use and assessment.

The SCO Community will take several actions to address the aforementioned challenges in benchmarking. Liaising with data analysis experts, we will carefully curate data collections suitable for addressing specific tasks within the SCO data analysis workflow (e.g. multi-modal data integration, deconvolution of bulk data). For this, we will survey the landscape of existing benchmarking studies and identify the datasets they used and how they were evaluated. Whenever possible, our focus will be on real datasets rather than simulated ones, given the bias introduced by simulated data towards methods using the same underlying model. In order to address the lack of agreed-upon performance metrics to evaluate different types of methods, we will collect and curate existing metrics, and develop/suggest new measures when necessary (Table 2).

Regarding the need for a common platform to conduct benchmark studies, we will explore using OpenEBench. This ELIXIR benchmarking platform offers a flexible computational framework that allows individual communities to design and perform their benchmarking experiments. Communities are responsible for defining the reference datasets and the evaluation metrics and designing and developing evaluation workflows. Software developers are then able to use these workflows to evaluate their tools against the reference datasets, and the computed metrics are compiled, analyzed and publicly exposed in tables and visualizations. The results of the evaluation are then used by the community or any other OpenEBench user to decide which is the most suitable tool to do their analysis. The SCO Community will provide guidelines for the setup of single-cell benchmarking experiments. The guidelines have to cover three topics: 1) the scope of the benchmark, 2) the evaluation metrics that will be used to measure the performance of the tools and 3) the reference or gold standard datasets. The SCO Community will establish a benchmarking environment for SCO data analysis tools within the OpenEBench infrastructure, to facilitate a variety of community-driven challenges to address the diversity of the SCO applications.

The SCO community will perform periodic reviews of highly performant and rapidly adopted methods for registration in the bio.tools catalogue. To this end we will work closely also with the EDAM ontology to define single-cell specific keywords, which will help us not only to annotate the tools but also tag courses in TeSS.

Compute platform
The computing resource requirements of SCO data analysis increase constantly as the scale and modality of the data sets grow. The discussion between the SCO Community and the ELIXIR Compute Platform is therefore vital to ensure sufficient resources. The Community will also benefit from the Compute Platform’s Container Orchestration task, which will allow execution of containerised software tools and workflow workloads supporting public and sensitive data across ELIXIR Nodes. The ELIXIR Authentication and Authorisation Infrastructure (AAI) will be supported in the context of sensitive SCO data and whenever controlled access will be needed, we count on learning from the HCA’s experience on this matter.

Interoperability platform
The ELIXIR SCO Community will promote the development and usage of standards of metadata and file formats to ensure reproducibility of analyses and data reuse across biological and bioinformatics research communities. We will support existing efforts for aggregating and disseminating related metadata standards, e.g. from ArrayExpress, the HCA and further efforts. This is particularly important for emerging spatially resolved data, in line with current efforts in imaging and omics databases (Table 2).

Data platform
An important consideration for sensitive human data is the General Data Protection Regulation (GDPR). To comply with the GDPR, raw human sequencing data deposited in EGA is protected and requires approval of the Data Access Committee Officer (DACO) as well as Data Transfer Agreements (DTA) outlining the conditions for allowing access to sensitive data. However, there is heterogeneous interpretation of the GDPR across Europe, and to facilitate this there have been a number of nationally Federated EGAs being established. Other non-human raw sequencing data would be deposited in ENA and would not be subject to these restrictions.

The ELIXIR SCO Community will encourage all data generators to ensure their data is available from an ELIXIR core data resource, or is deposited with a suitable ELIXIR core deposition resource, wherever possible, to ensure maximum
data reuse and long term sustainability of all SCO data across the broader community. Via connections to key ELIXIR resources at EMBL-EBI (ENA, EGA, ArrayExpress and BioSamples database), we will promote discussions with these data resources to encourage the adoption and development of standards, where needed, to support the rapid pace of technology change in the single-cell field. We will leverage EMBL-EBI connections to the HCA Data Coordination Platform to broker the HCA data to ELIXIR core data resources and deposition resources.

Alignment with other European and global SCO initiatives
The SCO Community will bring together data standardization efforts across Europe and combine them with global collaborations. The EMBL-EBI Node is a member of the global HCA community, whose mission is to create comprehensive reference maps of all human cells as a basis for both understanding human health and diagnosing, monitoring, and treating disease.15 It is also involved in the NIH-supported HuBMAP consortium,16 which develops tools to create an open, global atlas of the human body at the cellular level. The EMBL-EBI has already led an international effort to define the first guidelines for metadata standards of scRNA-seq experiments,58 involving members of the HCA and HuBMAP data platforms. As SCO techniques develop, we expect these guidelines to evolve to enable reproducible analysis of other methods, such as scATAC-seq, CITE-seq, single-cell HiC, to name a few.

Importantly, the ELIXIR SCO Community will align its activities with the LifeTime FET initiative,71 which combines single-cell multi-omics technologies with artificial intelligence and machine learning in order to revolutionize healthcare by tracking, understanding, and treating human cells during diseases. The LifeTime consortium includes over 90 research institutes and 70 supporting companies across Europe. Scientists from some of ELIXIR Nodes belong to both the LifeTime initiative and the SCO Community, thereby providing a direct link between them.

While HCA, HuBMAP and LifeTime focus on human cells, it is important to note that SCO technologies are used for different organisms, and thereby the ELIXIR SCO Community is not limited to human research. For example, the EMBL-EBI is also involved in the Fly Cell Atlas consortium.72

The training activities of the SCO Community will be enriched by collaboration with the Global Organization for Bioinformatics Learning, Education and Training (GOBLET).73 GOBLET’s mission is to cultivate the global bioinformatics trainer community, set standards and provide high-quality resources to support learning, education and training. The emerging SCO Community and GOBLET co-organized a global workshop for single-cell RNA-seq data analysis trainers in 2021. Sharing information about different training approaches, materials and datasets was considered very useful by the participants, and follow-up workshops are planned.

Finally, the SCO Community is discussing with the emerging SCO Community of Australian BioCommons, which is currently collecting user needs and finding solutions to the challenges identified, similar to us.

Conclusions
The SCO paradigm represents a revolution in the life sciences that pushes the boundaries of what can be explored, creating both new opportunities and challenges. We are witnessing increasing numbers of individual- and multi-omics modalities, and spatio-temporally resolved read outs. Both the rapid pace of advancement and adoption indicate that SCO will become the new normal in the life sciences. In the past five years, many ELIXIR Nodes have been working to assemble resources with the goal of developing future-proof guidelines and infrastructure as well as delivering training to SCO scientists. Here, we defined key goals at different infrastructural areas in order to create the ELIXIR SCO Community (Table 2) to ultimately strengthen current and foster new collaborations, and establish sustainable European and global frameworks for SCO research.

Data and Software Availability
Data on scientific publications on SCO and number of cells sequenced was obtained from the Single-cell studies database.74 Data on SCO tools was taken from the publicly available repository of the scRNA-tools database.34

Author contributions
PC, PMP, IP, TB, JL, KH and EK conceptualized the study. PC performed data curation and visualization. KH performed project administration. PC, AM, RAC, PMP, LPS, AGU, LA, SCG, BS, MA, NI and EK wrote the original draft of the manuscript. PC, AM, PMP, LPS, AGU, CS, BS, PV, H-RH, BL, JMH, BG, WH, RK, TR, AB, PF, KH, NI, and EK reviewed and edited the manuscript.
Acknowledgments
We would like to thank several members of ELIXIR Nodes for useful discussions and input: Ana Melo (PT), Andrei Zinovyev (FR), Åsa Björklund (SE), Celia van Gelder (NL), Ernesto Picardi (IT), Philip Lijnzaad (NL), Jan Korbel (DE), Joaquim Dopazo (ES) Loredana Le Pera (IT), Priti Adler (EE), Ricardo Leite (PT), Silvie Fexova (EMBL-EBI), Ståle Nygård (NO), Victoria Dominguez del Angel (FR).

References


The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com