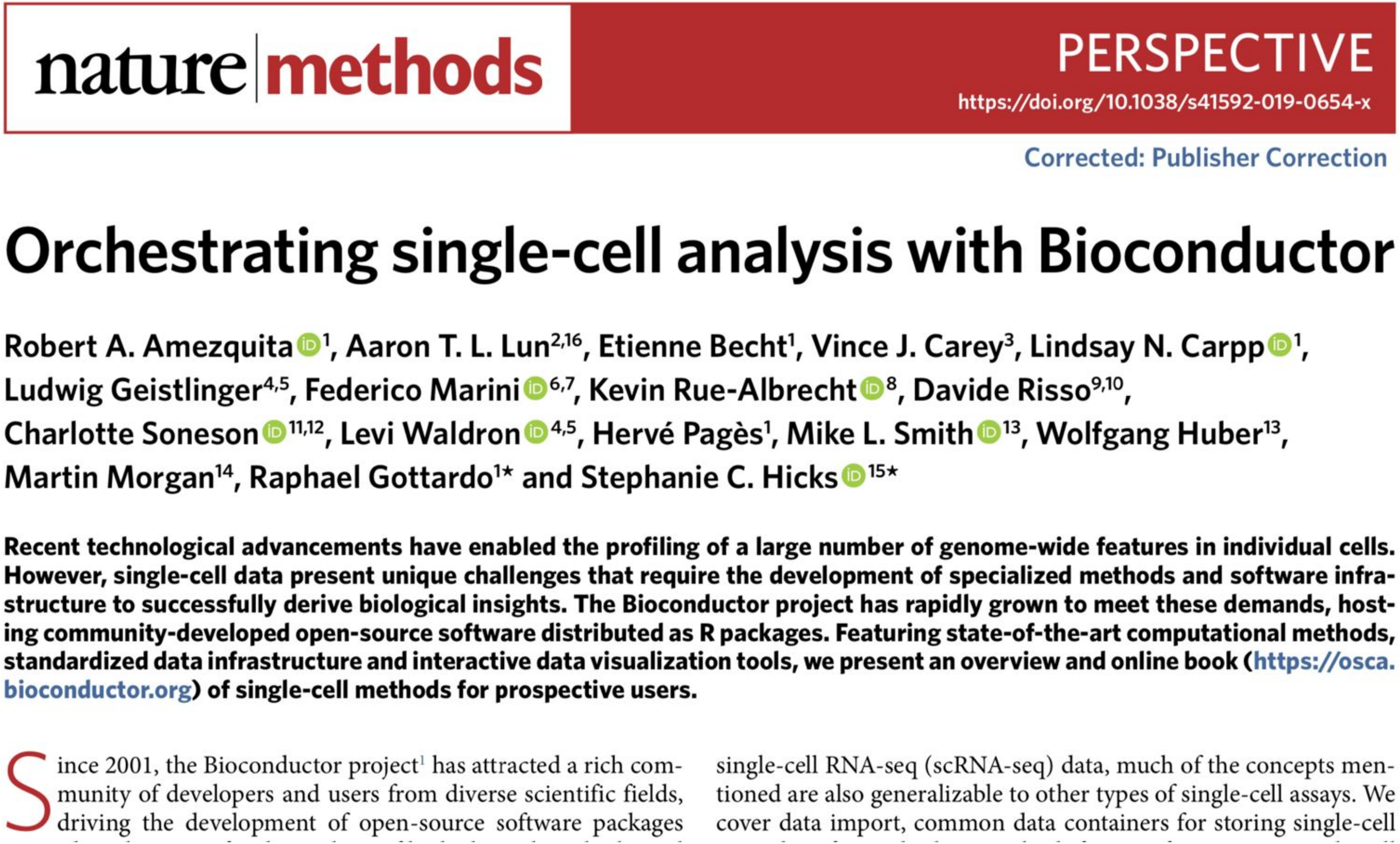
Outreach

Stephanie Hicks



nature methods

PERSPECTIVE

https://doi.org/10.1038/s41592-019-0654-x

Corrected: Publisher Correction

Orchestrating single-cell analysis with Bioconductor

Robert A. Amezquita[1], Aaron T. L. Lun[2,16], Etienne Becht[1], Vince J. Carey[3], Lindsay N. Carpp[1], Ludwig Geistlinger[4,5], Federico Marini[6,7], Kevin Rue-Albrecht[8], Davide Risso[9,10], Charlotte Soneson[11,12], Levi Waldron[4,5], Hervé Pagès[1], Mike L. Smith[13], Wolfgang Huber[13], Martin Morgan[14], Raphael Gottardo[1*] and Stephanie C. Hicks[15*]

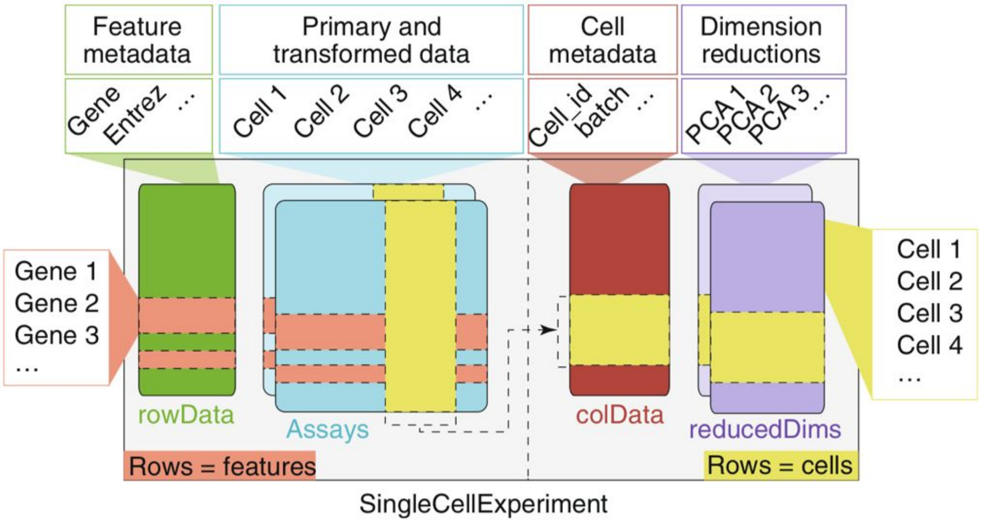
Recent technological advancements have enabled the profiling of a large number of genome-wide features in individual cells. However, single-cell data present unique challenges that require the development of specialized methods and software infrastructure to successfully derive biological insights. The Bioconductor project has rapidly grown to meet these demands, hosting community-developed open-source software distributed as R packages. Featuring state-of-the-art computational methods, standardized data infrastructure and interactive data visualization tools, we present an overview and online book (https://osca.bioconductor.org) of single-cell methods for prospective users.

Since 2001, the Bioconductor project[1] has attracted a rich community of developers and users from diverse scientific fields, driving the development of open-source software packages

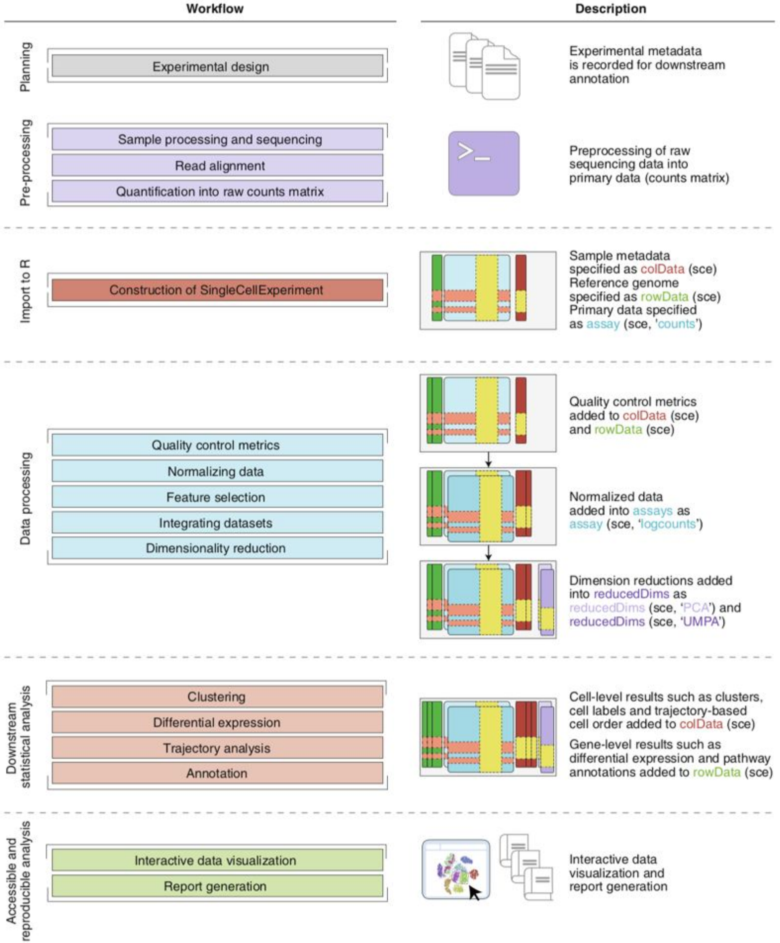
single-cell RNA-seq (scRNA-seq) data, much of the concepts mentioned are also generalizable to other types of single-cell assays. We cover data import, common data containers for storing single-cell

https://osca.bioconductor.org

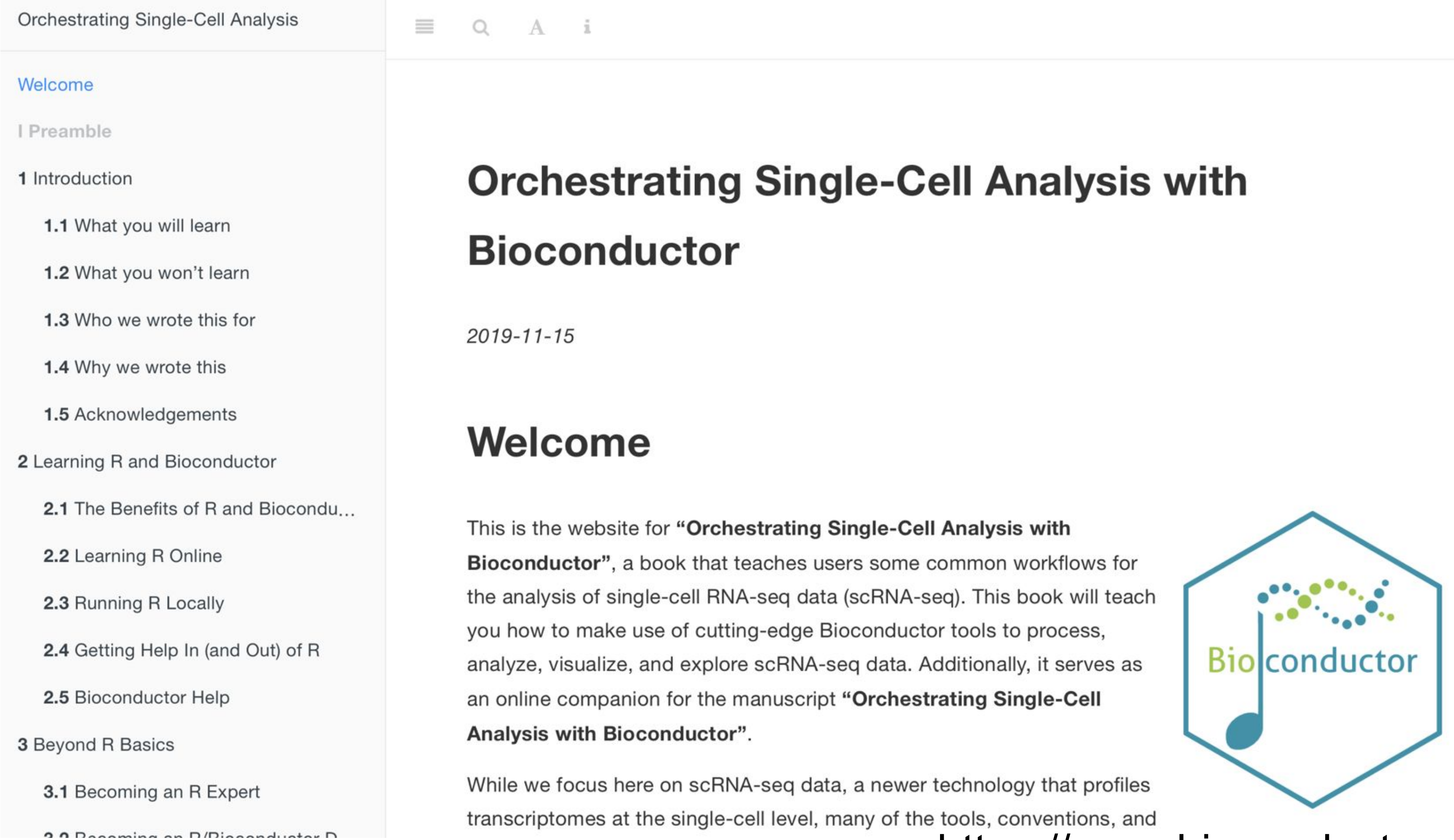
Standardized ways to store single-cell data



Bioconductor workflows



	Workflow	Description
Planning	Experimental design	Experimental metadata is recorded for downstream annotation
Pre-processing	Sample processing and sequencing Read alignment Quantification into raw counts matrix	Preprocessing of raw sequencing data into primary data (counts matrix)
Import to R	Construction of SingleCellExperiment	Sample metadata specified as colData (sce) Reference genome specified as rowData (sce) Primary data specified as assay (sce, 'counts')
Data processing	Quality control metrics Normalizing data Feature selection Integrating datasets Dimensionality reduction	Quality control metrics added to colData (sce) and rowData (sce) Normalized data added into assays as assay (sce, 'logcounts') Dimension reductions added into reducedDims as reducedDims (sce, 'PCA') and reducedDims (sce, 'UMPA')
Downstream statistical analysis	Clustering Differential expression Trajectory analysis Annotation	Cell-level results such as clusters, cell labels and trajectory-based cell order added to colData (sce) Gene-level results such as differential expression and pathway annotations added to rowData (sce)
Accessible and reproducible analysis	Interactive data visualization Report generation	Interactive data visualization and report generation



Orchestrating Single-Cell Analysis

Orchestrating Single-Cell Analysis with Bioconductor

2019-11-15

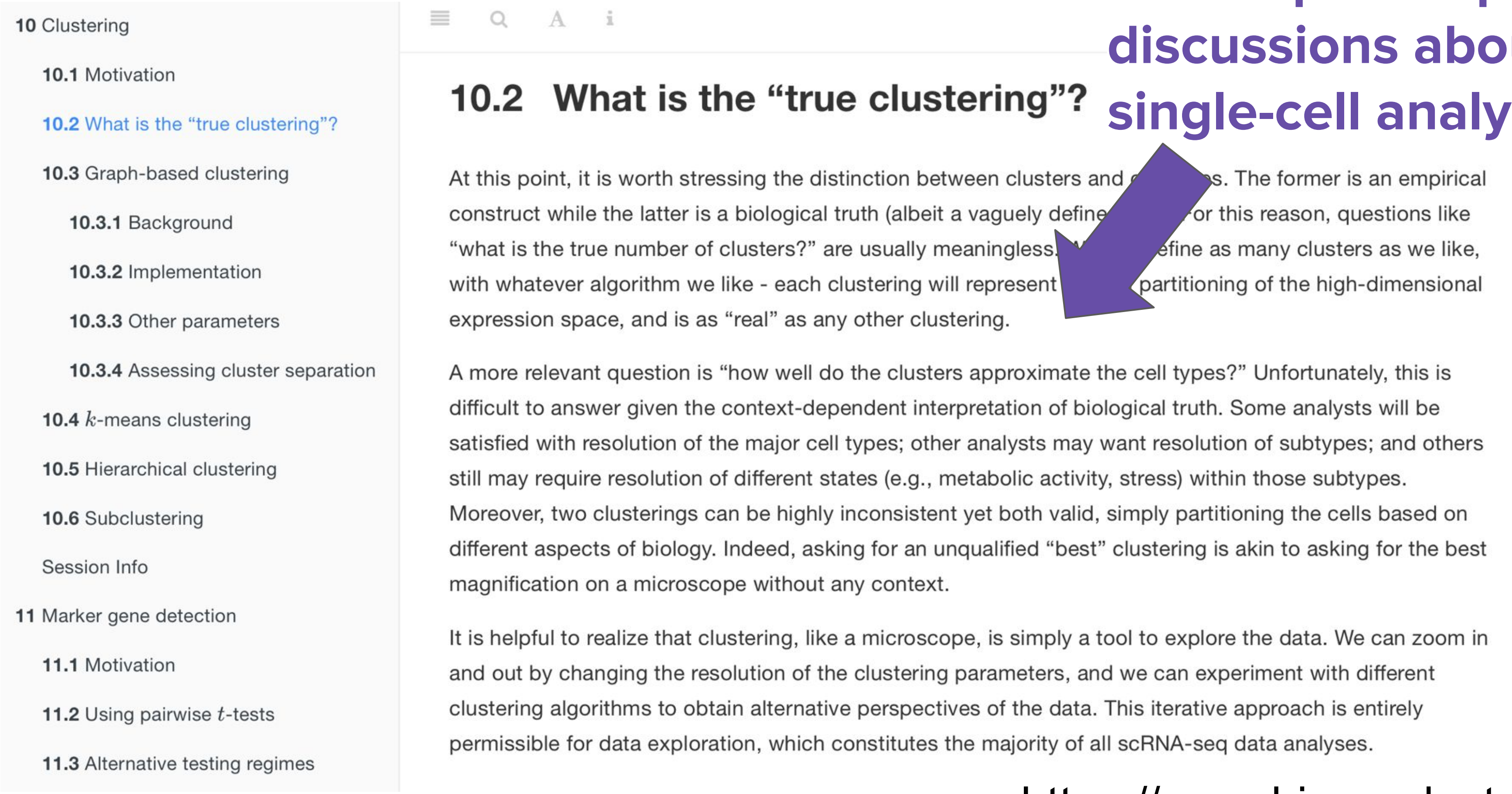
Welcome

This is the website for **“Orchestrating Single-Cell Analysis with Bioconductor”**, a book that teaches users some common workflows for the analysis of single-cell RNA-seq data (scRNA-seq). This book will teach you how to make use of cutting-edge Bioconductor tools to process, analyze, visualize, and explore scRNA-seq data. Additionally, it serves as an online companion for the manuscript **“Orchestrating Single-Cell Analysis with Bioconductor”**.

While we focus here on scRNA-seq data, a newer technology that profiles transcriptomes at the single-cell level, many of the tools, conventions, and

https://osca.bioconductor.org

Lots of philosophical discussions about single-cell analyses



10.2 What is the “true clustering”?

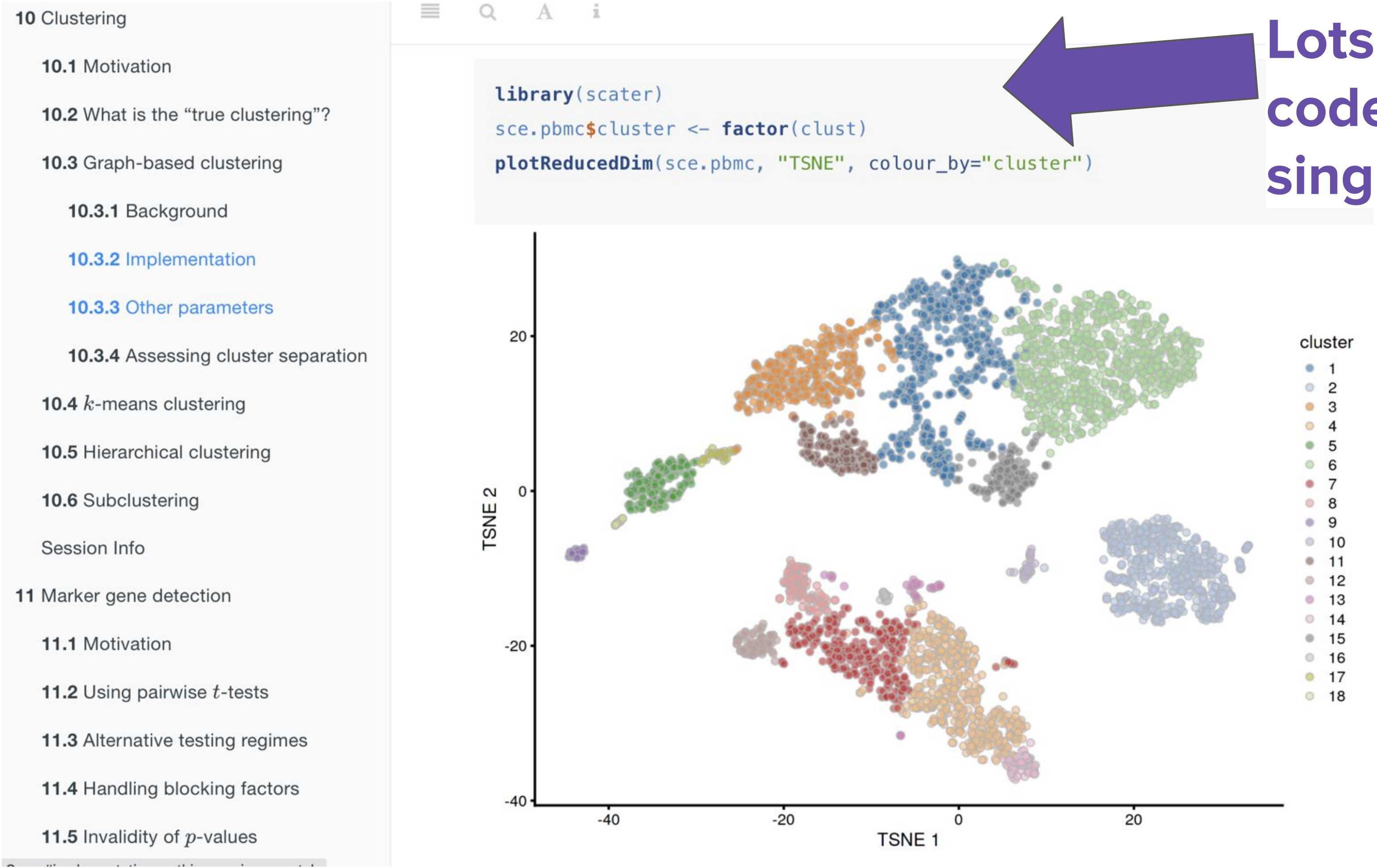
At this point, it is worth stressing the distinction between clusters and c s. The former is an empirical construct while the latter is a biological truth (albeit a vaguely define or this reason, questions like “what is the true number of clusters?” are usually meaningless. efine as many clusters as we like, with whatever algorithm we like - each clustering will represent partitioning of the high-dimensional expression space, and is as “real” as any other clustering.

A more relevant question is “how well do the clusters approximate the cell types?” Unfortunately, this is difficult to answer given the context-dependent interpretation of biological truth. Some analysts will be satisfied with resolution of the major cell types; other analysts may want resolution of subtypes; and others still may require resolution of different states (e.g., metabolic activity, stress) within those subtypes. Moreover, two clusterings can be highly inconsistent yet both valid, simply partitioning the cells based on different aspects of biology. Indeed, asking for an unqualified “best” clustering is akin to asking for the best magnification on a microscope without any context.

It is helpful to realize that clustering, like a microscope, is simply a tool to explore the data. We can zoom in and out by changing the resolution of the clustering parameters, and we can experiment with different clustering algorithms to obtain alternative perspectives of the data. This iterative approach is entirely permissible for data exploration, which constitutes the majority of all scRNA-seq data analyses.

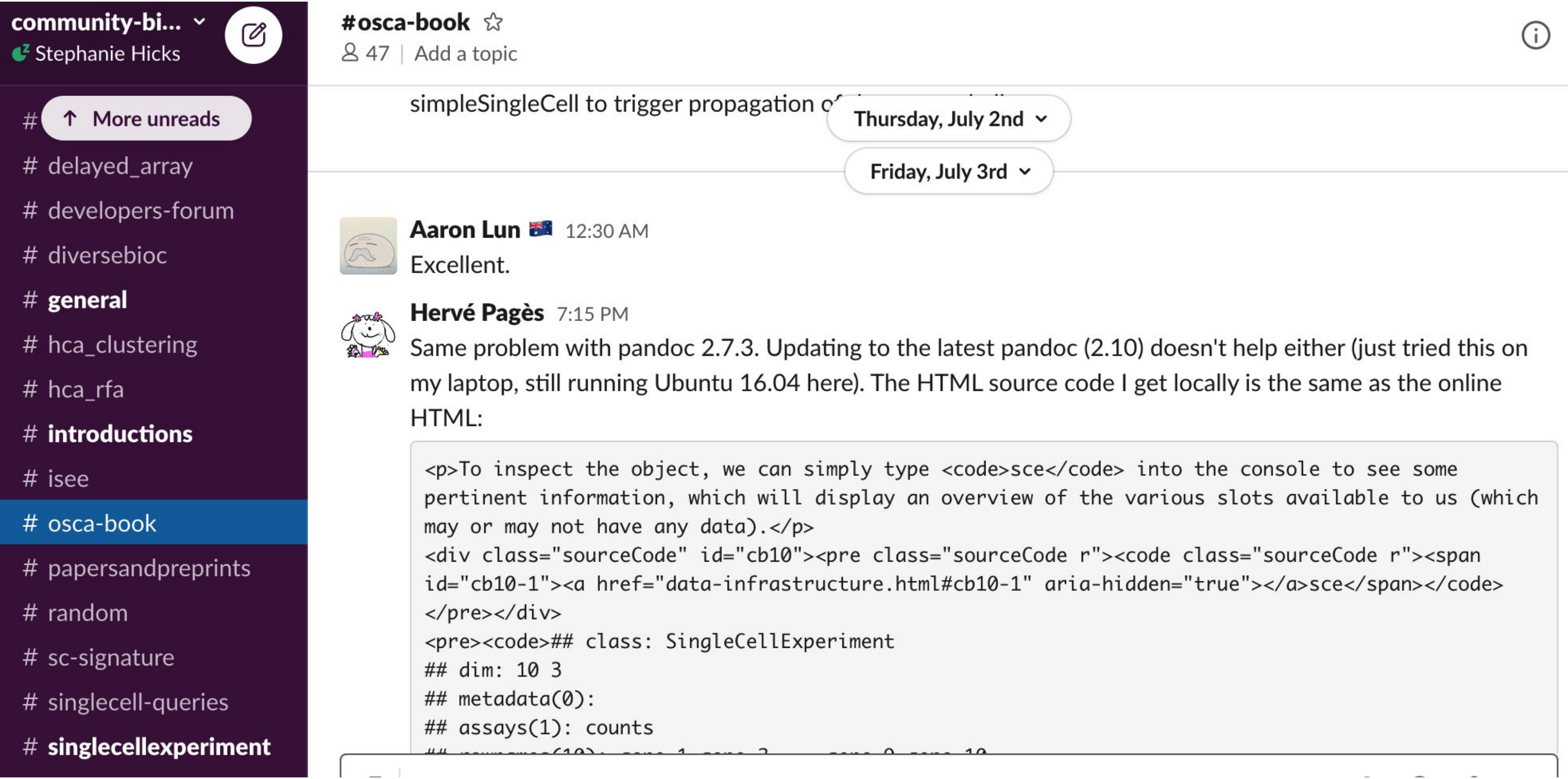
https://osca.bioconductor.org

Lots of R/Bioc code to analyze single-cell data



```r
library(scater)
sce.pbmc$cluster <- factor(clust)
plotReducedDim(sce.pbmc, "TSNE", colour_by="cluster")
```

https://osca.bioconductor.org



simpleSingleCell to trigger propagation o

Aaron Lun 12:30 AM

Excellent.

Hervé Pagès 7:15 PM

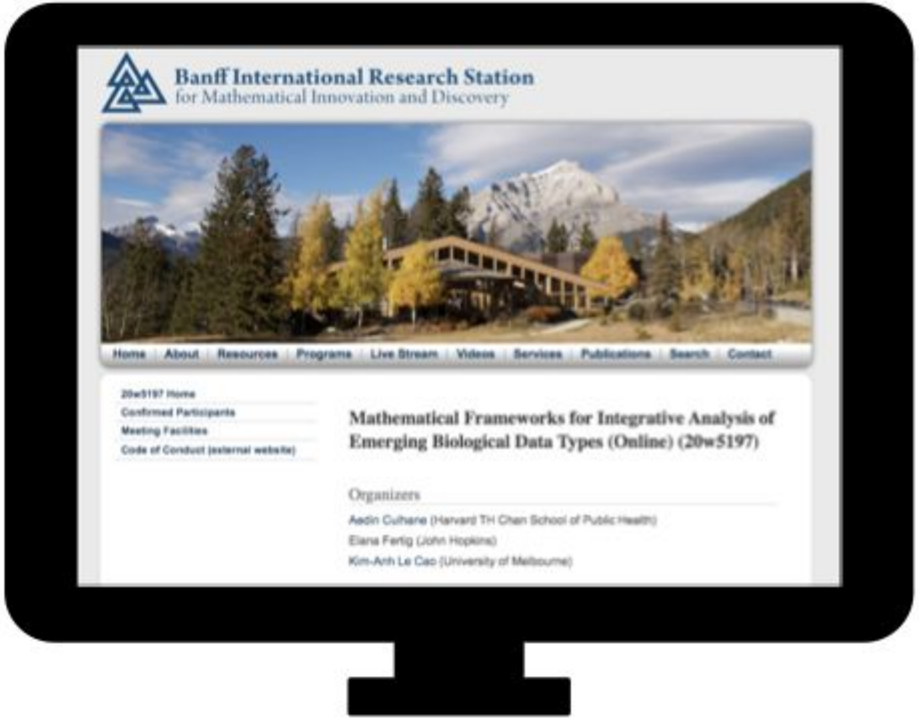
Same problem with pandoc 2.7.3. Updating to the latest pandoc (2.10) doesn't help either (just tried this on my laptop, still running Ubuntu 16.04 here). The HTML source code I get locally is the same as the online HTML:

```
<p>To inspect the object, we can simply type <code>sce</code> into the console to see some
pertinent information, which will display an overview of the various slots available to us (which
may or may not have any data).</p>
<div class="sourceCode" id="cb10"><pre class="sourceCode r"><code class="sourceCode r"><span
id="cb10-1"><a href="data-infrastructure.html#cb10-1" aria-hidden="true"></a>sce</span></code>
</pre></div>
<pre><code>## class: SingleCellExperiment
## dim: 10 3
## metadata(0):
## assays(1): counts
## rownames(10): gene-1 gene-2 ... gene-9 gene-10
```

Integrative analysis of sc data

#BIRSBioIntegration

Banff International Research Station
for Mathematical Innovation and Discovery


Banff International Research Station
for Mathematical Innovation and Discovery
Home | About | Resources | Programs | Live Stream | Videos | Services | Publications | Search | Contact
20w5197 Home
Confirmed Participants
Meeting Facilities
Code of Conduct (external website)

Mathematical Frameworks for Integrative Analysis of Emerging Biological Data Types (Online) (20w5197)

Organizers

Aedin Culhane (Harvard TH Chan School of Public Health)

Elana Fertig (John Hopkins)

Kim-Anh Le Cao (University of Melbourne)

Aedín Culhane
Dana-Farber
Cancer Institute/
Harvard Chan

Elana Fertig
Johns Hopkins
University

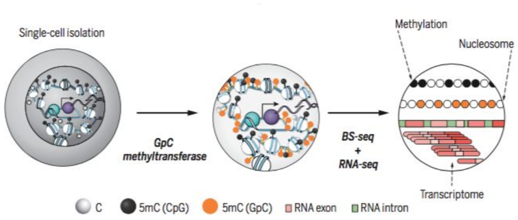
Kim-Anh Lê Cao
The University of
Melbourne

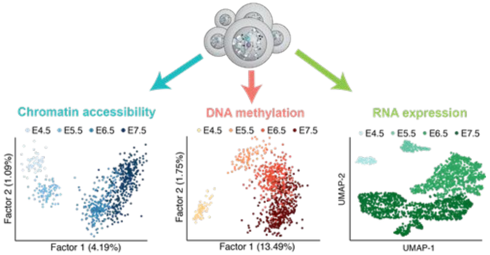
https://www.birs.ca/events/2020/5-day-workshops/20w5197

3 Hackathon Challenges

Gastrulation (scNMT)

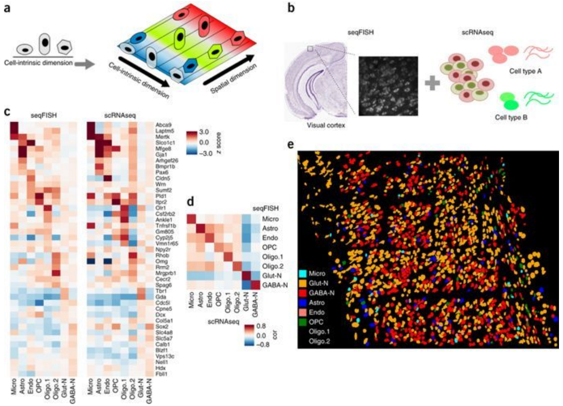
826 cells matching across all data sets (transcriptome, DNA accessibility and DNA methylation) after quality control and filtering.





Adult mouse visual cortex seqFISH, scRNAseq

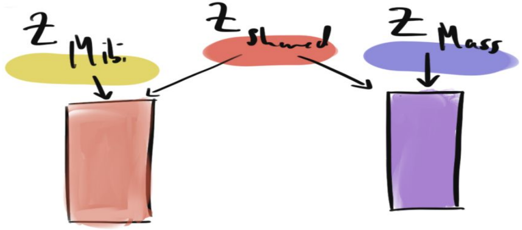
- seqFISH - 1,597 single cells x 125 genes mapped (Zhu *et al* 2018)
- scRNA-seq. ~1,600 cells (Tasic *et al* 2016)



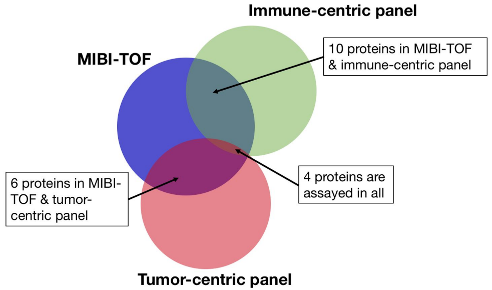
Breast Cancer sc Proteomics

Non-overlapping patients

MIBI 40 TN, Mass Tag 7 TN



… with 20 overlapping proteins

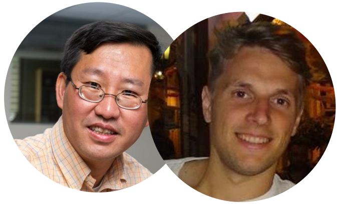


10 proteins in MIBI-TOF & immune-centric panel

6 proteins in MIBI-TOF & tumor-centric panel

4 proteins are assayed in all

6 Keynotes, 16 Contributed talks, 9 Brainstorming sessions



seqfish_theme

Guo-Cheng Yuan &
Ruben Dries

Dana-Farber Cancer Institute,
Harvard TH Chan School of Public
Health & Boston University

sc_targ_proteomics_theme

Aedin Culhane &
Olga Vitek

Dana-Farber Cancer Institute,
Harvard TH Chan School of Public
Health & Northeastern University

scNMT-seq_theme

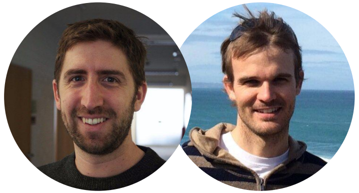
Ricard Arguelaget &
Oliver Stegle

German Cancer Research Center &
EMBL

summary_analyses_theme

Kim-Anh Lê Cao &
Casey Green

University of Melbourne & Uni
Pennsylvania



benchmark_theme

Mike Love &
Matt Ritchie

University of North Carolina-Chapel Hill
& Walter and Eliza Hall Institute

Susan Holmes

Stanford University

interpretation_theme

Vincent Carey

Harvard Medical School and
Brigham & Women’s Hospital

software_theme

Elana Fertig

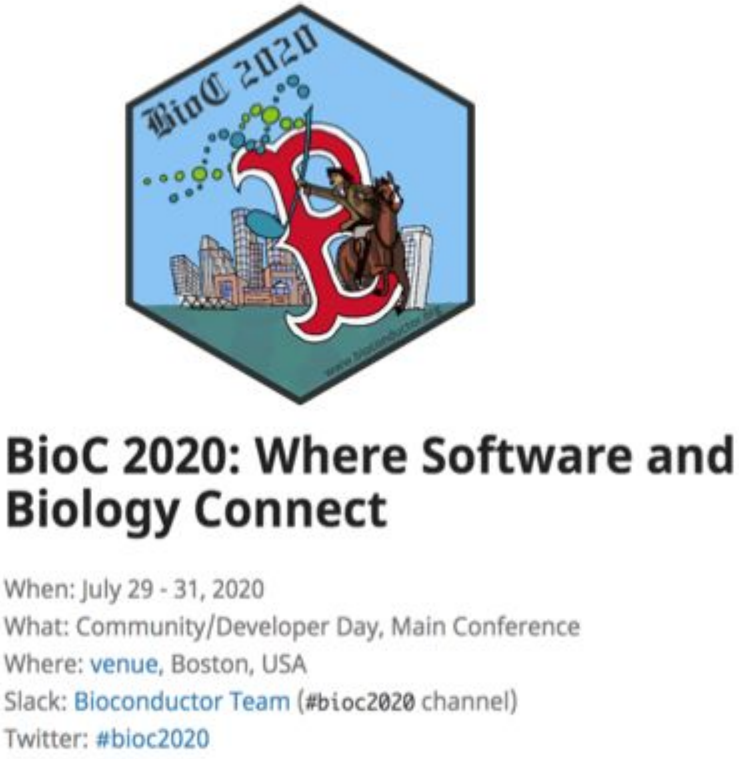
Johns Hopkins University

future_theme

#Bioc2020

http://bioc2020.bioconductor.org/

Now Virtual



BioC 2020: Where Software and Biology Connect

When: July 29 - 31, 2020
What: Community/Developer Day, Main Conference
Where: venue, Boston, USA
Slack: Bioconductor Team (`#bioc2020` channel)
Twitter: #bioc2020