Outreach
Stephanie Hicks
Orchestrating single-cell analysis with Bioconductor

Robert A. Amezquita, Aaron T. Lun, Etienne Becht, Vince J. Carey, Lindsay N. Carpp, Ludwig Geistlinger, Federico Marini, Kevin Rue-Albrecht, Davide Risso, Charlotte Soneson, Levi Waldron, Hervé Pagès, Mike L. Smith, Wolfgang Huber, Martin Morgan, Raphael Gottardo and Stephanie C. Hicks

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 has attracted a rich community of developers and users from diverse scientific fields, driving the development of open-source software packages for 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
Standardized ways to store single-cell data

Bioconductor workflows

<table>
<thead>
<tr>
<th>Workflow</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Planning</td>
<td>Experimental metadata is recorded for downstream annotation</td>
</tr>
<tr>
<td>Preprocessing</td>
<td>Sample processing and sequencing</td>
</tr>
<tr>
<td></td>
<td>Read alignment</td>
</tr>
<tr>
<td></td>
<td>Quantification into raw counts matrix</td>
</tr>
<tr>
<td>Import to R</td>
<td>Construction of SingleCellExperiment</td>
</tr>
<tr>
<td></td>
<td>Sample metadata specified as colData (sce)</td>
</tr>
<tr>
<td></td>
<td>Reference genome specified as rowData (sce)</td>
</tr>
<tr>
<td></td>
<td>Primary data specified as assay (sce, 'counts')</td>
</tr>
<tr>
<td>Data processing</td>
<td>Quality control metrics added to colData (sce) and rowData (sce)</td>
</tr>
<tr>
<td></td>
<td>Normalized data added into assays as assay (sce, 'counts')</td>
</tr>
<tr>
<td></td>
<td>Feature selection</td>
</tr>
<tr>
<td></td>
<td>Integrating datasets</td>
</tr>
<tr>
<td></td>
<td>Dimensionality reduction</td>
</tr>
<tr>
<td>Downstream analysis</td>
<td>Clustering</td>
</tr>
<tr>
<td></td>
<td>Differential expression</td>
</tr>
<tr>
<td></td>
<td>Trajectory analysis</td>
</tr>
<tr>
<td></td>
<td>Annotation</td>
</tr>
<tr>
<td></td>
<td>Cell-level results such as clusters, cell labels and trajectory-based cell order added to colData (sce)</td>
</tr>
<tr>
<td></td>
<td>Gene-level results such as differential expression and pathway annotations added to rowData (sce)</td>
</tr>
<tr>
<td>Interactive visualization and report generation</td>
<td>Interactive data visualization</td>
</tr>
<tr>
<td></td>
<td>Report generation</td>
</tr>
</tbody>
</table>
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
10.2 What is the “true clustering”?*

At this point, it is worth stressing the distinction between clusters and subtypes. The former is an empirical construct while the latter is a biological truth (albeit a vaguely defined one). For this reason, questions like “what is the true number of clusters?” are usually meaningless. We can define as many clusters as we like, with whatever algorithm we like - each clustering will represent a 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
simpleSingleCell to trigger propagation or

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>&lt;code&gt;sce&lt;/code&gt;</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).&lt;/p&gt;
<div class="sourceCode" id="cb10"><pre class="sourceCode r"><code class="sourceCode r">&lt;span id="cb10-1"&gt;&lt;a href="data-infrastructure.html#cb10-1" aria-hidden="true"&gt;&lt;/a&gt;&lt;/span&gt;&lt;/code&gt;</pre></div><pre><code>## class: SingleCellExperiment
## dim: 10 3
## metadata(0):
## assays(1): counts
</code></pre>
Integrative analysis of sc data

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
6 Keynotes, 16 Contributed talks, 9 Brainstorming sessions

Guo-Cheng Yuan & Ruben Dries
Dana-Farber Cancer Institute, Harvard TH Chan School of Public Health & Boston University

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

Ricard Arguelaghet & Oliver Stegle
German Cancer Research Center & EMBL

Kim-Anh Lê Cao & Casey Green
University of Melbourne & Uni Pennsylvania

Mike Love & Matt Ritchie
University of North Carolina-Chapel Hill & Walter and Eliza Hall Institute

Susan Holmes
Stanford University

Vincent Carey
Harvard Medical School and Brigham & Women’s Hospital

Elana Fertig
Johns Hopkins University
Goals of this workshop

Multi-omics integration of single cell data

- is an active and emerging field
- May provide insight that cannot be obtained from single datasets
- lacks established performance benchmarks, gold standard datasets, assessment standards.

Bring together interdisciplinary computational scientists

- to examine cutting edge techniques for integrative analysis of diverse multi-omics.
- Provide & assess open source resources for multi-platform analysis
- Formulate goals and future directions to advance multi-omics analysis

Products: Guidelines, build collaboration, code & datasets, a white paper

#Bioc2020

http://bioc2020.bioconductor.org/

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

Now Virtual