Orchestrating Single-Cell Analysis

Bioconductor

2019-01-07

Contents

Chapter 1

Prerequisites

This is a *sample* book written in **Markdown**. You can use anything that Pandoc's Markdown supports, e.g., a math equation $a^2 + b^2 = c^2$.

The **bookdown** package can be installed from CRAN or Github:

```
install.packages("bookdown")
# or the development version
# devtools::install_github("rstudio/bookdown")
```

Remember each Rmd file contains one and only one chapter, and a chapter is defined by the first-level heading `#`.

To compile this example to PDF, you need XeLaTeX. You are recommended to install TinyTeX (which includes XeLaTeX): https://yihui.name/tinytex/.

Chapter 2

Introduction

You can label chapter and section titles using `{#label}` after them, e.g., we can reference Chapter 2. If you do not manually label them, there will be automatic labels anyway, e.g., Chapter **??**.

Figures and tables with captions will be placed in `figure` and `table` environments, respectively.

```
par(mar = c(4, 4, .1, .1))
plot(pressure, type = 'b', pch = 19)
```

Reference a figure by its code chunk label with the `fig:` prefix, e.g., see Figure 2.1. Similarly, you can reference tables generated from `knitr::kable()`, e.g., see Table 2.1.

```
knitr::kable(
  head(iris, 20), caption = 'Here is a nice table!',
  booktabs = TRUE
)
```

You can write citations, too. For example, we are using the **bookdown** package (Xie, 2018) in this sample book, which was built on top of R Markdown and **knitr** (Xie, 2015).

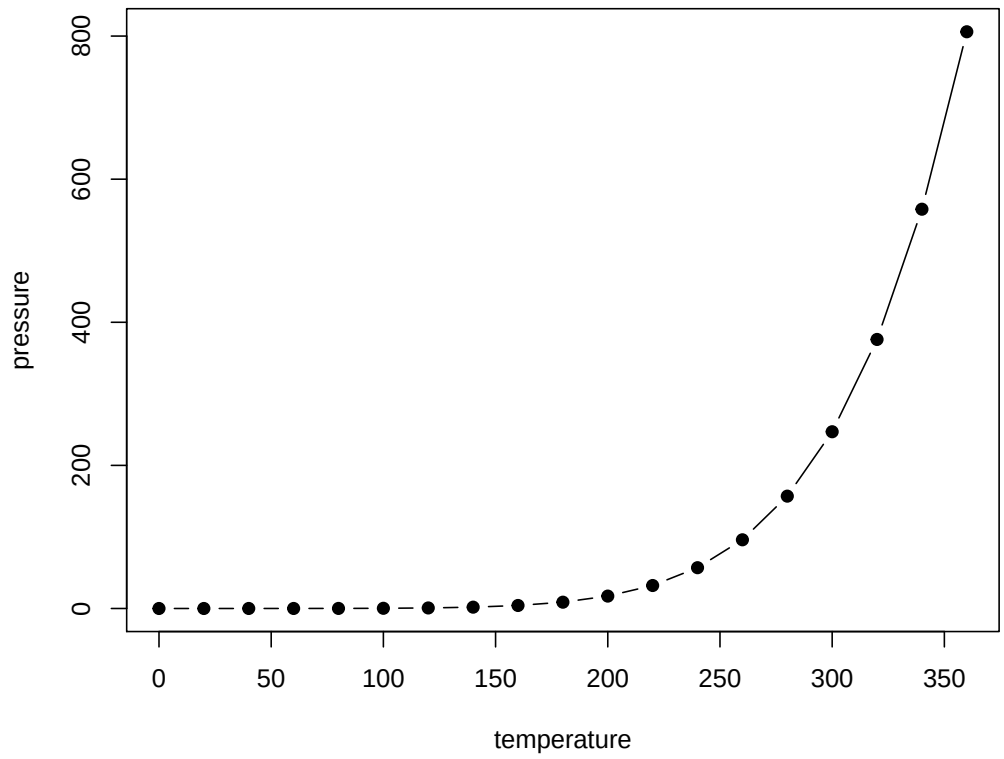

Figure 2.1: Here is a nice figure!

Table 2.1: Here is a nice table!

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
4.6	3.1	1.5	0.2	setosa
5.0	3.6	1.4	0.2	setosa
5.4	3.9	1.7	0.4	setosa
4.6	3.4	1.4	0.3	setosa
5.0	3.4	1.5	0.2	setosa
4.4	2.9	1.4	0.2	setosa
4.9	3.1	1.5	0.1	setosa
5.4	3.7	1.5	0.2	setosa
4.8	3.4	1.6	0.2	setosa
4.8	3.0	1.4	0.1	setosa
4.3	3.0	1.1	0.1	setosa
5.8	4.0	1.2	0.2	setosa
5.7	4.4	1.5	0.4	setosa
5.4	3.9	1.3	0.4	setosa
5.1	3.5	1.4	0.3	setosa
5.7	3.8	1.7	0.3	setosa
5.1	3.8	1.5	0.3	setosa

Chapter 3

Clustering and Differential Expression Workflow

Author: Robert A. Amezquita, Fred Hutchinson Cancer Research Center

3.1 Introduction

Introduce the main purpose of this vignette in paragraph form, why these steps are important, what the end products allow for.

3.2 Learning objectives

- In bullet form, illustrate what the main outputs are of this vignette
- A couple of bullet points are fine

3.3 Package requirements

Packages that are required to go through the vignette and how to install them, should largely be through BiocManager interface.

```
BiocManager::install(c('SC3', 'clusterExperiment', 'BiocParallel', 'BEARscc', 'zinbwave', 'edgeR', 'DES
```

3.4 Loading the data

Here the data should be loaded if in package form and briefly explored to illustrate the contents of the data.

If data comes from "raw" form, then a simple pipeline for creating the processed data should be illustrated as succinctly as possible.

```
library(TENxPBMCData)
```

```
## Loading required package: SingleCellExperiment
```

```
## Loading required package: SummarizedExperiment

## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, Map, mapply, match, mget, order, paste,
##     pmax, pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,
##     rowMeans, rownames, rowSums, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## Loading required package: Biobase
```

```
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.

## Loading required package: DelayedArray

## Loading required package: matrixStats

##
## Attaching package: 'matrixStats'

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

## Loading required package: BiocParallel

##
## Attaching package: 'DelayedArray'

## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges

## The following objects are masked from 'package:base':
##
##     aperm, apply

## Loading required package: HDF5Array

## Loading required package: rhdf5
```

```
TENxPBMCData('pbmc4k')
```

```
## snapshotDate(): 2019-01-04

## see ?TENxPBMCData and browseVignettes('TENxPBMCData') for documentation

## downloading 0 resources

## loading from cache
##     '/Users/ramezqui//.ExperimentHub/1613'

## class: SingleCellExperiment
## dim: 33694 4340
## metadata(0):
## assays(1): counts
## rownames(33694): ENSG00000243485 ENSG00000237613 ...
##   ENSG00000277475 ENSG00000268674
## rowData names(3): ENSEMBL_ID Symbol_TENx Symbol
## colnames: NULL
## colData names(11): Sample Barcode ... Individual Date_published
## reducedDimNames(0):
## spikeNames(0):
```

3.5 Preprocessing

All steps required to preprocess the data into the clean expression matrix, split it up into subsections as necessary. Subsection splits should be informed by the paper (sub)sections. This is all steps prior to the main strategy of interest.

3.5.1 Cell and gene quality control

Removal of "bad cells", low abundance genes.

3.5.2 Normalization

Application of normalization approach, even if its just a log normalization on the count data. Cell cycle normalization may be applicable here.

3.5.3 Feature selection

Identifying the subset of the clean matrix to work with.

3.5.4 Dimensionality reduction

Calculating PCs and UMAP/tSNE representation.

3.6 Clustering

Strategy 1 of interest. Use BiocParallel. May need to save intermediate data within package to facilitate fast construction of Rmd.

- SC3
- clusterExperiment
- clustree
- BEARscc # uses ERCC spike ins
- BiocNeighbors # more for devs

3.7 Differential Expression

Strategy 2 of interest. Differential expression approaches. Use one of the outputs from above to perform differential expression analyses comparison.

- zinbwave + edgeR/DESeq2 (may exclude as workflow is very different or highlight as an "advanced section")
- MAST
- scDD
- SC3

3.8 Session Info

Chapter 4

Trajectory Analysis Workflow

Author: Robert A. Amezquita, Fred Hutchinson Cancer Research Center

4.1 Introduction

Introduce the main purpose of this vignette in paragraph form, why these steps are important, what the end products allow for.

4.2 Learning objectives

- In bullet form, illustrate what the main outputs are of this vignette
- A couple of bullet points are fine

4.3 Package requirements

Packages that are required to go through the vignette and how to install them, should largely be through BiocManager interface.

```
BiocManager::install(c('SC3', 'clusterExperiment', 'BiocParallel', 'BEARscc', 'zinbwave', 'edgeR', 'DES
```

4.4 Loading the data

Here the data should be loaded if in package form and briefly explored to illustrate the contents of the data.

If data comes from "raw" form, then a simple pipeline for creating the processed data should be illustrated as succinctly as possible.

```
library(TENxPBMCData)
```

```
## Loading required package: SingleCellExperiment
```

```
## Loading required package: SummarizedExperiment
```

```
## Loading required package: GenomicRanges

## Loading required package: stats4

## Loading required package: BiocGenerics

## Loading required package: parallel

##
## Attaching package: 'BiocGenerics'

## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB

## The following objects are masked from 'package:stats':
##
##     IQR, mad, sd, var, xtabs

## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, basename, cbind,
##     colMeans, colnames, colSums, dirname, do.call, duplicated,
##     eval, evalq, Filter, Find, get, grep, grepl, intersect,
##     is.unsorted, lapply, Map, mapply, match, mget, order, paste,
##     pmax, pmax.int, pmin, pmin.int, Position, rank, rbind, Reduce,
##     rowMeans, rownames, rowSums, sapply, setdiff, sort, table,
##     tapply, union, unique, unsplit, which, which.max, which.min

## Loading required package: S4Vectors

##
## Attaching package: 'S4Vectors'

## The following object is masked from 'package:base':
##
##     expand.grid

## Loading required package: IRanges

## Loading required package: GenomeInfoDb

## Loading required package: Biobase

## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
## Loading required package: DelayedArray

## Loading required package: matrixStats

##
## Attaching package: 'matrixStats'

## The following objects are masked from 'package:Biobase':
##
##     anyMissing, rowMedians

## Loading required package: BiocParallel

##
## Attaching package: 'DelayedArray'

## The following objects are masked from 'package:matrixStats':
##
##     colMaxs, colMins, colRanges, rowMaxs, rowMins, rowRanges

## The following objects are masked from 'package:base':
##
##     aperm, apply

## Loading required package: HDF5Array

## Loading required package: rhdf5
```

```
TENxPBMCData('pbmc4k')
```

```
## snapshotDate(): 2019-01-04

## see ?TENxPBMCData and browseVignettes('TENxPBMCData') for documentation

## downloading 0 resources

## loading from cache
##     '/Users/ramezqui//.ExperimentHub/1613'

## class: SingleCellExperiment
## dim: 33694 4340
## metadata(0):
## assays(1): counts
## rownames(33694): ENSG00000243485 ENSG00000237613 ...
##   ENSG00000277475 ENSG00000268674
## rowData names(3): ENSEMBL_ID Symbol_TENx Symbol
## colnames: NULL
## colData names(11): Sample Barcode ... Individual Date_published
## reducedDimNames(0):
## spikeNames(0):
```

4.5 Preprocessing

All steps required to preprocess the data into the clean expression matrix, split it up into subsections as necessary. Subsection splits should be informed by the paper (sub)sections. This is all steps prior to the main strategy of interest.

4.5.1 Cell and gene quality control

Removal of "bad cells", low abundance genes.

4.5.2 Normalization

Application of normalization approach, even if its just a log normalization on the count data. Cell cycle normalization may be applicable here.

4.5.3 Feature selection

Identifying the subset of the clean matrix to work with.

4.5.4 Dimensionality reduction

Calculating PCs and UMAP/tSNE representation. This may not be necessary for certain software or may differ between packages.

4.6 Trajectory Analysis

- slingshot
- ouija
- celltrails

4.7 Session Info

Bibliography

Xie, Y. (2015). *Dynamic Documents with R and knitr*. Chapman and Hall/CRC, Boca Raton, Florida, 2nd edition. ISBN 978-1498716963.

Xie, Y. (2018). *bookdown: Authoring Books and Technical Documents with R Markdown*. R package version 0.9.