The Bioconductor 2018 Workshop Compilation

Contents

Chapter 1

Introduction

Author: Martin Morgan[1]. Last modified: 22 May, 2018.

1.1 For Everyone

This book contains workshops used in *R* / *Bioconductor* training. The workshops are divided into 3 sections:

- **Learn** (100-series chapters) contains material for beginning users of *R* and *Bioconductor*. The *Bioconductor*-related material is relevant even for experienced *R* users who are new to *Bioconductor*.
- **Use** (200-series chapters) contains workshops emphasizing use of *Bioconductor* for common tasks, e.g., bulk RNA-seq differential expression, ChIP-seq, single-cell analysis, gene set enrichment, and annotation.
- **Develop** (500-series chapters) contains workshops to help expert users hone their skills and contribute their domain-specific knowledge to the *Bioconductor* community.

1.2 For Workshop Authors

To contribute a new workshop, open a BiocWorkshops issue asking to be added as a collaborator.

Write your workshop as a stand-alone markdown document, using the _template.Rmd file as a starting point. Follow the numbering scheme for classifying your workshop.

Also update the DESCRIPTION file adding packages utilized in your workshop to the Imports field. Please be kind and don't remove anyone elses packages from the DESCRIPTION as this is a shared file for all workshops.

See bookdown instructions for authoring documents; we are using the 'knit-then-merge' strategy. You'll need to install the bookdown package package from CRAN, as well as pandoc. Render your chapter with the `preview=` option to `render_book()`, e.g.,

```
Rscript -e "bookdown::render_book(                              \
    'xxx_Your_Workshop.Rmd', 'bookdown::gitbook', preview=TRUE \
)"
```

[1]Roswell Park Comprehensive Cancer Center, Buffalo, NY

As this is a shared space for all workshop contributors, in order to use the above command in the BiocWorkshops directory, the index has to be built at least once, which can be time consuming depending on how many workshops have already been submitted.

```
Rscript -e "bookdown::render_book(                              \
    'index.Rmd', 'bookdown::gitbook')"
```

To avoid having to build all workshops but still be able to preview your individual workshop we recommend creating a soft link to your .Rmd file. We recommend having the file in the `BiocWorkshop/` and the soft link in any other directory on your system. By default, this will generate an html file in `_book/` wherever this command is run.

```
mkdir tmp
cd tmp/
ln -s ../xxx_Your_Workshop.Rmd
Rscript -e "bookdown::render_book(                              \
    'xxx_Your_Workshop.Rmd', 'bookdown::gitbook', preview=TRUE \
)"
```

Push **only** your .Rmd file to the BiocWorkshop repository; the book will be rebuilt manually or automatically. Eventually the output will be available for end-users at https://bioconductor.github.io/BiocWorkshops .The master branch will not contain the built version of the book. Switching to the gh-pages branch will show built output.

1.3 Deadlines for Bioc2018

Please be aware of the following deadlines for the Bioconductor 2018 Conference in Toronto

- **Fri June 29:** draft workshop materials submitted to this Bioconductor GitHub bookdown site
- **Fri July 6:** feedback period completes
- **Weds July 18:** workshops must pass checks without errors or warnings (All materials will be checked by Continuous Integration)
- **Thurs / Fri July 26-27:** Bioc2018

Chapter 2

Variant Functional Annotation using StatePaintR, FunciVar & MotifBreakR

2.1 Authors:

- Simon G. Coetzee (scoetzee@gmail.com)
- Dennis J. Hazelett (dennis.hazelett@csmc.edu)

2.2 Workshop Description

This workshop will entail lecture and live demo of StateHub/ StatePaintR and funciVar bioconductor packages. ### Pre-requisites It is recommended that workshop participants have: * Basic concepts in epigenomics and GWAS * Intermediate R skills, familiarity with GRanges

Relevant background reading for the workshop.

- StateHub/StatePaintR
- BioC2017 Workshop
- MotifBreakR

2.2.1 Workshop Participation

Describe how students will be expected to participate in the workshop.

2.2.2 *R / Bioconductor* packages used

StatePaintR FunciVar MotifBreakR

2.2.3 Time outline

1 hour Workshop:

Activity	Time
StatePaintR	30m
intro and theory	(10m)
Generate Decision Matrix	(10m)
live demo	(10m)
FunciVar	15m
intro	(10m)
live demo	(5m)
MotifBreakR	15m
intro & demo	(10m)
Shiny webapp demo	(5m)

2.3 Workshop goals and objectives

This workshop focuses on bioconductor based tools for non-coding variant annotation. I will present a high level overview of these concepts and the bioconductor tools our group uses to address them *in silico*. These tools can be used for germline or somatic variants, but as we will see, can also be used for any other type of feature represented as GRanges objects.

2.3.1 Key Concepts

I will provide brief overview with examples of:

- Genome segmentation and Chromatin State
- Feature enrichment analysis
- Transcription factors and Motif Disruption

2.3.2 Learning objectives

At the end of this workshop students will understand the basic inputs and outputs of all these analyses, as well as an intuitive understanding of the tools and where to find them:

- The StateHub website
- StatePaintR genome segmentation software Bioconductor
- FunciVar for variant annotation and enrichment
- MotifBreakR Transcription Factor motif disruption Bioconductor

Chapter 3

Working with Genomic Data in R with the DECIPHER package

3.1 Authors:

- Nicholas Cooley (npc19@pitt.edu)
- Erik Wright (eswright@pitt.edu)

3.2 Workshop Description

In this workshop we will give an introduction to working with biological sequence data in R using the Biostrings and DECIPHER packages. We will cover:

- Importing, viewing, and manipulating genomes in R
- Construction of sequence databases for organizing genomes
- Mapping syntenic regions between genomes with the FindSynteny function
- Understanding, accessing, and viewing objects of class Synteny
- Using syntenic information to predict orthologous genes
- Alignment of sequences and genomes with DECIPHER
- Downstream analyses enabled by syntenic mapping

3.2.1 Pre-requisites

- Familiarity with Biostrings
- Familiarity with DECIPHER Databases (Ref. 1)

1. Wright, E. S. The R Journal 2016, 8 (1), 352–359.

3.2.2 Workshop Participation

This will be a lab where participants follow along on their computers.

3.2.3 *R / Bioconductor* packages used

- Biostrings
- DECIPHER

3.2.4 Time outline

Activity	Time
Packages/Introduction	5m
Basic commands for sequences	5m
Sequence databases	10m
Demonstration of synteny mapping	10m
Explanation of function arguments	10m
Dissecting Synteny objects	10m
Visualization of syntenic blocks	10m
Alignment of syntenic regions	10m
Ortholog prediction from Synteny	10m
Constructing phylogenetic trees	10m

3.3 Workshop goals and objectives

3.3.1 Learning goals

- Understand a simple workflow for analysis of sequences in R and DECIPHER
- Learn the basic use and appropriate application of functions within DECIPHER

3.3.2 Learning objectives

- Learn basic commands for working with sequences in R
- Import genomes from online repositories or local files
- Map synteny between genomes
- Analyze a synteny map among multiple genomes
- Develop an understanding of the data structures involved
- Predict orthologs from syntenic maps
- Select core and pan genomes from predicted orthologs
- Construct and interpret phylogenetic trees

Chapter 4

Analysis of single-cell RNA-seq data: Dimensionality reduction, clustering, and lineage inference

4.1 Instructor(s) name(s) and contact information

- Diya Das (@diyadas, diyadas@berkeley.edu)
- Davide Risso (@drisso)
- Kelly Street (@kstreet13)

4.2 Workshop Description

This workshop will be presented as a lab session (brief introduction followed by hands-on coding) that instructs participants in a Bioconductor workflow for the analysis of single-cell RNA-sequencing data, in three parts: 1. dimensionality reduction that accounts for zero inflation, over-dispersion, and batch effects 2. cell clustering that employs a resampling-based approach resulting in robust and stable clusters 3. lineage trajectory analysis that uncovers continuous, branching developmental processes

We will provide worked examples for lab sessions, and a set of stand-alone notes in this repository.

Note to organizers: A previous version of this workshop was well-attended at BioC 2017, but the tools presented have been significantly updated for interoperability (most notably, through the use of the `SingleCellExperiment` class), and we have been receiving many requests to provide an updated workflow. We plan to incorporate feedback from this workshop into a revised version of our F1000 Workflow.

4.2.1 Pre-requisites

We expect basic knowledge of R syntax. Some familiarity with S4 objects may be helpful, but not required. More importantly, participants should be familiar with the concept and design of RNA-sequencing experiments. Direct experience with single-cell RNA-seq is not required, and the main challenges of single-cell RNA-seq compared to bulk RNA-seq will be illustrated.

4.2.2 Workshop Participation

This will be a hands-on workshop, in which each student, using their laptop, will analyze a provided example datasets. The workshop will be a mix of example code that the instructors will show to the students (available through this repository) and short exercises.

4.2.3 *R / Bioconductor* packages used

1. *zinbwave* : https://bioconductor.org/packages/zinbwave
2. *clusterExperiment*: https://bioconductor.org/packages/clusterExperiment
3. *slingshot*: https://github.com/kstreet13/slingshot (will be submitted to Bioconductor shortly)

4.2.4 Time outline

2 hr workshop:

Activity	Time
Intro to single-cell RNA-seq analysis	15m
zinbwave (dimensionality reduction)	30m
clusterExperiment (clustering)	30m
slingshot (lineage trajectory analysis)	30m
Questions / extensions	15m

4.3 Workshop goals and objectives

4.3.1 Learning goals

- describe the goals of single-cell RNA-seq analysis
- identify the main steps of a typical single-cell RNA-seq analysis
- evaluate the results of each step in terms of model fit
- synthesize results of successive steps to interpret biological significance and develop biological models
- apply this workflow to carry out a complete analysis of other single-cell RNA-seq datasets

4.3.2 Learning objectives

- compute and interpret low-dimensional representations of single-cell data
- identify and remove sources of technical variation from the data
- identify sub-populations of cells (clusters) and evaluate their robustness
- infer lineage trajectories corresponding to differentiating cells
- order cells by developmental “pseudotime”
- identify genes that play an important role in cell differentiation

Chapter 5

Functional enrichment analysis of high-throughput omics data

5.1 Instructor(s) name(s) and contact information

- Ludwig Geistlinger (Ludwig.Geistlinger@sph.cuny.edu)
- Levi Waldron

CUNY School of Public Health 55 W 125th St, New York, NY 10027

5.2 Workshop Description

This workshop gives an in-depth overview of existing methods for enrichment analysis of gene expression data with regard to functional gene sets, pathways, and networks. The workshop will help participants understand the distinctions between assumptions and hypotheses of existing methods as well as the differences in objectives and interpretation of results. It will provide code and hands-on practice of all necessary steps for differential expression analysis, gene set- and network-based enrichment analysis, and identification of enriched genomic regions and regulatory elements, along with visualization and exploration of results. A previous well-attended (>50 participants) version of this workshop was presented at Bioc2017 (https://www.bioconductor.org/help/course-materials/2017/BioC2017/Day1/Workshops/OmicsData/doc/enrichOmics.html).

5.2.1 Pre-requisites

- Basic knowledge of R syntax
- Familiarity with the SummarizedExperiment class
- Familiarity with the GenomicRanges class
- Familiarity with high-throughput gene expression data as obtained with microarrays and RNA-seq
- Familiarity with the concept of differential expression analysis (with e.g. limma, edgeR, DESeq2)

5.2.2 Workshop Participation

Execution of example code and hands-on practice

5.2.3 *R / Bioconductor* packages used

- EnrichmentBrowser
- regioneR

5.2.4 Time outline

Activity	Time
Background	30m
Differential expression analysis	15m
Gene set analysis	30m
Gene network analysis	15m
Genomic region analysis	15m

5.3 Goals and objectives

Theory * Gene sets, pathways & regulatory networks * Resources * Gene set analysis vs. gene set enrichment analysis * Underlying null: competitive vs. self-contained * Generations: ora, fcs & topology-based

Practice: * Data types: microarray vs. RNA-seq * Differential expression analysis * Defining gene sets according to GO and KEGG * GO/KEGG overrepresentation analysis * Functional class scoring & permutation testing * Network-based enrichment analysis * Genomic region enrichment analysis

Chapter 6

Effectively using the DelayedArray framework to support the analysis of large datasets

6.1 Instructor(s) name(s) and contact information

- Peter Hickey (GitHub/Twitter: @PeteHaitch, email: peter.hickey@gmail.com)

6.2 Workshop Description

This workshop will teach the fundamental concepts underlying the DelayedArray framework and related infrastructure. It is intended for package developers who want to learn how to use the DelayedArray framework to support the analysis of large datasets, particularly through the use of on-disk data storage. The first part of the workshop will provide an overview of the DelayedArray infrastructure and introduce computing on DelayedArray objects using delayed operations and block-processing. The second part of the workshop will present strategies for adding support for DelayedArray to an existing package and extending the DelayedArray framework. Students can expect a mixture of lecture and question-and-answer session to teach the fundamental concepts. There will be plenty of examples to illustrate common design patterns for writing performant code, although we will not be writing much code during the workshop.

6.2.1 Pre-requisites

- Solid understanding of R
- Familiarity with common operations on arrays (e.g., `colSums()` and those available in the matrixStats package)
- Familiarity with object oriented programming, particularly S4, will be useful but is not essential
- No familiarity required of technical details of particular data storage backends (e.g., HDF5, sparse matrices)
- No familiarity required of particular biological techniques (e.g., single-cell RNA-seq)

6.2.2 Workshop Participation

Questions and discussion are encouraged! This will be especially important to guide the second half of the worshop which focuses on integrating DelayedArray into an existing or new Bioconductor package. Students will be expected to be able to follow and reason about example R code.

6.2.3 *R / Bioconductor* packages used

- DelayedArray
- HDF5Array
- SummarizedExperiment
- DelayedMatrixStats
- beachmat

6.2.4 Time outline

Activity	Time
Introductory slides	15m
Part 1: Overview of DelayedArray framework	45m
Part 2: Incoporating DelayedArray into a package	45m
Questions and discussion	15m

6.3 Workshop goals and objectives

6.3.1 Learning goals

- Identify when it is useful to use a DelayedArray instead of an ordinary array or other array-like data structure.
- Become familiar with the fundamental concepts of delayed operations, block-processing, and realization.
- Learn of existing functions and packages for constructing and computing on DelayedArray objects, avoiding the need to re-invent the wheel.
- Learn common design patterns for writing performant code that operates on a DelayedArray.
- Evaluate whether an existing function that operates on an ordinary array can be readily adapted to work on a DelayedArray.
- Reason about potential bottlenecks in algorithms operating on DelayedArray objects.

6.3.2 Learning objectives

- Understand the differences between a DelayedArray instance and an instance of a subclass (e.g., HDF5Array, RleArray).
- Know what types of operations 'degrade' an instance of a DelayedArray subclass to a DelayedArray, as well as when and why this matters.
- Construct a DelayedArray:
  - From an in-memory array-like object.
  - From an on-disk data store (e.g., HDF5).
  - From scratch by writing data to a RealizationSink.
- Take a function that operates on rows or columns of a matrix and apply it to a DelayedMatrix.
- Use block-processing on a DelayedArray to compute:

  - A univariate (scalar) summary statistic (e.g., `max()`).
  - A multivariate (vector) summary statistic (e.g., `colSum()` or `rowMean()`).
  - A multivariate (array-like) summary statistic (e.g., `rowRanks()`).
- Design an algorithm that imports data into a DelayedArray.

Chapter 7

Cytoscape Automation in R using Rcy3

7.1 Instructor(s) name(s) and contact information

- Ruth Isserlin ruth.isserlin@utoronto.ca
- Brendan Innes brendan.innes@mail.utoronto.ca
- Jeff Wong jvwong@gmail.com
- Gary Bader gary.bader@utoronto.ca

7.2 Workshop Description

Cytoscape is one of the most popular applications for network analysis and visualization. In this workshop, we will demonstrate new capabilities to integrate Cytoscape into programmatic workflows and pipelines using R. We will begin with an overview of network biology themes and concepts, and then we will translate these into Cytoscape terms for practical applications. The bulk of the workshop will be a hands-on demonstration of accessing and controlling Cytoscape from R to perform a network analysis of tumor expression data.

7.2.1 Pre-requisites

7.2.1.1 Workshop prerequisites:

- Basic knowledge of R syntax
- Basic knowledge of Cytoscape software
- Familiarity with network biology concepts

7.2.1.2 Background:

- "How to visually interpret biological data using networks." Merico D, Gfeller D, Bader GD. Nature Biotechnology 2009 Oct 27, 921-924 - http://baderlab.org/Publications?action=AttachFile&do=view&target=2009_Merico_Primer_NatBiotech_Oct.pdf
- "CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API". Keiichiro Ono, Tanja Muetze, Georgi Kolishovski, Paul Shannon, Barry Demchak.F1000Res. 2015 Aug 5;4:478. - https://f1000research.com/articles/4-478/v1

7.2.2 Workshop Participation

Participants are required to bring a laptop with Cytoscape, R, and RStudio installed. Installation instructions will be provided in the weeks preceding the workshop. The workshop will consist of a lecture and lab.

7.2.3 R / Bioconductor packages used

- RCy3

7.2.4 Time outline

Activity	Time
Introduction	15m
Driving Cytoscape from R	15m
Creating, retrieving and manipulating networks	15m
Summary	10m

7.3 Workshop goals and objectives

7.3.1 Learning goals

- Know when and how to use Cytoscape in your research area
- Generalize network analysis methods to multiple problem domains
- Integrate Cytoscape into your bioinformatics pipelines

7.3.2 Learning objectives

- Programmatic control over Cytoscape from R
- Publish, share and export networks

Chapter 8

RNA-seq analysis is easy as 1-2-3 with limma, Glimma and edgeR

8.1 Author:

- Charity Law (law@wehi.edu.au)

8.2 Workshop Description

In this instructor-led live demo, we analyse RNA-sequencing data from the mouse mammary gland, demonstrating use of the popular **edgeR** package to import, organise, filter and normalise the data, followed by the **limma** package with its voom method, linear modelling and empirical Bayes moderation to assess differential expression and graphical representations. This pipeline is further enhanced by the **Glimma** package which enables interactive exploration of the results so that individual samples and genes can be examined by the user. The complete analysis offered by these three packages highlights the ease with which researchers can turn the raw counts from an RNA-sequencing experiment into biological insights using Bioconductor. The complete workflow is available at http://master.bioconductor.org/packages/release/workflows/html/RNAseq123.html .

8.2.1 Pre-requisites

- Basic knowledge of RNA-sequencing
- Basic knowledge of R syntax, R object classes and object manipulation

8.2.2 Workshop Participation

Participants can watch the live demo, or may prefer to follow the demonstration by bringing their laptops along. To follow the analysis on their own laptops, participants need to install the *RNAseq123* workflow by running

```
source("https://bioconductor.org/biocLite.R")
biocLite("RNAseq123")
```

in R. The relevant sequencing data should also be download in advance.

```
url <- "https://www.ncbi.nlm.nih.gov/geo/download/?acc=GSE63310&format=file"
utils::download.file(url, destfile="GSE63310_RAW.tar", mode="wb")
utils::untar("GSE63310_RAW.tar", exdir = ".")
files <- c("GSM1545535_10_6_5_11.txt", "GSM1545536_9_6_5_11.txt", "GSM1545538_purep53.txt",
  "GSM1545539_JMS8-2.txt", "GSM1545540_JMS8-3.txt", "GSM1545541_JMS8-4.txt",
  "GSM1545542_JMS8-5.txt", "GSM1545544_JMS9-P7c.txt", "GSM1545545_JMS9-P8c.txt")
for(i in paste(files, ".gz", sep=""))
  R.utils::gunzip(i, overwrite=TRUE)
```

Due to time restraints, extra help regarding R package installation and coding errors will not be addressed during the workshop.

8.2.3 *R / Bioconductor* packages used

Bioconductor: limma, Glimma, edgeR, Mus.musculus
CRAN: RColorBrewer, gplots

8.2.4 Time outline

Activity	Time
Introduction	5mins
Data packaging	10mins
Data pre-processing	15mins
Differential expression analysis	30mins

8.3 Workshop goals and objectives

The key steps to RNA-seq data analysis are described in this workshop with basic statistical theory of methods used. The goal is to allow beginner-analysts of RNA-seq data to become familiar with each of the steps involved, as well as completing a standard analysis pipeline from start to finish.

8.3.1 Learning goals

- learn how to analyse RNA-seq data
- identify methods for pre-processing data
- understand linear models used in differential expression analysis
- examine plots for data exploration and result representation

8.3.2 Learning objectives

- read in count data and format as a DGEList-object
- annotate Entrez gene identifiers with gene information
- filter out lowly expressed genes
- normalise gene expression values
- unsupervised clustering of samples (standard and interactive plots)
- linear modelling for comparisons of interest
- remove heteroscedascity
- examine the number of differentially expressed genes

- mean-difference plots (standard and interactive plots)
- heatmaps

Chapter 9

Solving common bioinformatic challenges using GenomicRanges

9.1 Instructor(s) name(s) and contact information

- Michael Lawrence (michafla@gene.com)

9.2 Workshop Description

We will introduce the fundamental concepts underlying the GenomicRanges package and related infrastructure. After a structured introduction, we will follow a realistic workflow, along the way exploring the central data structures, including GRanges and SummarizedExperiment, and useful operations in the ranges algebra. Topics will include data import/export, computing and summarizing data on genomic features, overlap detection, integration with reference annotations, scaling strategies, and visualization. Students can follow along, and there will be plenty of time for students to ask questions about how to apply the infrastructure to their particular use case. Michael Lawrence (Genentech).

9.2.1 Pre-requisites

- Solid understanding of R
- Basic familiarity with GRanges objects
- Basic familiarity with packages like S4Vectors, IRanges, GenomicRanges, rtracklayer, etc.

9.2.2 Workshop Participation

Describe how students will be expected to participate in the workshop.

9.2.3 *R / Bioconductor* packages used

- S4Vectors
- IRanges
- GenomicRanges
- rtracklayer

- GenomicFeatures
- SummarizedExperiment
- VariantAnnotation
- GenomicAlignments

9.2.4 Time outline

Activity	Time
Intro slides	30m
Workflow(s)	1hr
Remaining questions	30m

9.3 Workshop goals and objectives

9.3.1 Learning goals

- Understand how to apply the *Ranges infrastructure to real-world problems
- Gain insight into the design principles of the infrastructure and how it was meant to be used

9.3.2 Learning objectives

- Manipulate GRanges and related objects
- Use the ranges algebra to analyze genomic ranges
- Implement efficient workflows based on the *Ranges infrastructure

Chapter 10

Fluent genomic data analysis with plyranges

10.1 Instructor(s) name(s) and contact information

- Stuart Lee (lee.s@wehi.edu.au)
- Michael Lawrence (lawremi@gmail.com)

10.2 Workshop Description

In this workshop, we will give an overview of how to perform low-level analyses of genomic data using the grammar of genomic data transformation defined in the plyranges package. We will cover:

- introduction to GRanges
- overview of the core verbs for arithmetic, restriction, and aggregation of GRanges objects
- performing joins between GRanges objects
- designing pipelines to quickly explore data from AnnotationHub
- reading BAM and other file types as GRanges objects

The workshop will be a computer lab, in which the participants will be able to ask questions and interact with the instructors.

10.2.1 Pre-requisites

This workshop is mostly self-contained however familiarity with the following would be useful:

- plyranges vignette
- the GenomicRanges and IRanges packages
- tidyverse approaches to data analysis

10.2.2 Workshop Participation

Students will work through an Rmarkdown document while the instructors respond to any questions they have.

10.2.3 *R / Bioconductor* packages used

- plyranges
- AnnotationHub
- GenomicRanges
- IRanges
- S4Vectors

10.2.4 Time outline

Activity	Time
Overview of GRanges	5m
The plyranges grammar	20m
I/O and data pipelines	20m

10.3 Workshop goals and objectives

10.3.1 Learning goals

- Understand that GRanges follows tidy data principles
- Apply the plyranges grammar to genomic data analysis

10.3.2 Learning objectives

- Use AnnotationHub to find and summarise data
- Read files into R as GRanges objects
- Perform coverage analysis
- Build data pipelines for analysis based on GRanges

Chapter 11

RNA-seq data analysis with DESeq2

11.1 Instructor(s) name(s) and contact information

- Michael Love (michaelisaiahlove@gmail.com)

11.2 Workshop Description

In this workshop, we will give a quick overview of the most useful functions in the DESeq2 package, and a basic RNA-seq analysis. We will cover: how to quantify transcript expression from FASTQ files using Salmon, import quantification from Salmon with tximport and tximeta, generate plots for quality control and exploratory data analysis EDA (also using MultiQC), perform differential expression (DE) (also using apeglm), overlap with other experimental data (using AnnotationHub), and build reports (using ReportingTools and Glimma). We will give a short example of integration of DESeq2 with the zinbwave package for single-cell RNA-seq differential expression. The workshop is designed to be a lab with plenty of time for questions throughout the lab.

11.2.1 Pre-requisites

- Basic knowledge of R syntax

Non-essential background reading:

- DESeq2 paper: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4302049/
- tximport paper: https://f1000research.com/articles/4-1521/v2
- apeglm paper: https://www.biorxiv.org/content/early/2018/04/17/303255

11.2.2 Workshop Participation

Students will participate by following along an Rmarkdown document, and asking questions throughout the workshop.

11.2.3 *R / Bioconductor* packages used

- DESeq2
- tximport

- apeglm
- AnnotationHub
- ReportingTools
- Glimma
- zinbwave

11.2.4 Time outline

Activity	Time
Overview of packages	20m
Quantification and import	20m
EDA and DE	20m
Downstream analysis & reports	20m
ZINB-WaVE integration	20m
Additional questions	20m

11.3 Workshop goals and objectives

Learning goals

- Visually assess quality of RNA-seq data
- Perform basic differential analysis of RNA-seq data
- Compare RNA-seq results with other experimental data

Learning objectives

- Quantify transcript expression from FASTQ files
- Import quantification into R/Bioconductor
- Perform quality control and exploratory data analysis
- Perform differential expression
- Overlap with other experimental data
- Build dynamic reports
- Integrate DESeq2 and zinbwave for single-cell RNA-seq data

Chapter 12

Introduction to Bioconductor annotation resources

12.1 Instructors

- James W. MacDonald (jmacdon@uw.edu)
- Lori Shepherd (lori.shepherd@roswellpark.org)

12.2 Workshop Description

There are various annotation packages provided by the Bioconductor project that can be used to incorporate additional information to results from high-throughput experiments. This can be as simple as mapping Ensembl IDs to corresponding HUGO gene symbols, to much more complex queries involving multiple data sources. In this workshop we will cover the various classes of annotation packages, what they contain, and how to use them efficiently.

12.2.1 Prerequisites

- Basic knowledge of R syntax
- Basic understanding of the various annotation sources (NCBI, EBI/EMBL)

Useful background reading

- The AnnotationDbi vignette.
- The biomaRt vignette.
- The GenomicFeatures vignette.

12.2.2 Workshop Participation

After each type of annotation package is introduced, students will be given the opportunity to practice making their own queries.

12.2.3 *R / Bioconductor* packages used

- AnnotationDbi
- AnnotationHub
- BSgenome
- biomaRt
- ensembldb
- org.Hs.eg.db
- TxDb.Hsapiens.UCSC.hg19.knownGene
- EnsDb.Hsapiens.v79
- EnsDb.Mmusculus.v79
- Homo.sapiens
- BSgenome.Hsapiens.UCSC.hg19
- hugene20sttranscriptcluster.db

12.3 Workshop goals and objectives

Annotating data is a complex task. For any high-throughput experiment the analyst usually starts with a set of identifiers for each thing that was measured, and in order to make the results useful to collaborators these identifiers need to be mapped to other identifiers that are either more familiar to collaborators, or that can be used for further analyses. As an example, RNA-Seq data may only have Entrez Gene IDs for each gene measured, and as part of the output you may want to include the gene symbols, which are more likely to be familiar to a Biologist.

12.3.1 Learning goals

- Understand what sort of annotation data are available
- Understand the difference between annotation sources (NCBI and EBI/EMBL)
- Gain familiarity with the various ways to query annotation packages
- Get some practice making queries

12.3.2 Learning objectives

- Be able to use select and mapIds to map between identifiers
- Be able to extract data from TxDb and EnsDb packages
- Be able to make queries using biomaRt
- Extract and utilize various data from AnnotationHub

Chapter 13

Workflow for Multi-omics Analysis with MultiAssayExperiment

13.1 Instructor names and contact information

- Marcel Ramos
- Ludwig Geistlinger
- Levi Waldron

13.2 Workshop Description

This workshop demonstrates data management and analyses of multiple assays associated with a single set of biological specimens, using the `MultiAssayExperiment` data class and methods. It introduces the `RaggedExperiment` data class, which provides efficient and powerful operations for representation of copy number and mutation and variant data that are represented by different genomic ranges for each specimen.

13.2.1 Pre-requisites

List any workshop prerequisites, for example:

- Basic knowledge of R syntax
- Familiarity with the GRanges and SummarizedExperiment classes
- Familiarity with 'omics data types including copy number and gene expression

13.2.2 workshop Participation

Students will have a chance to build a `MultiAssayExperiment` object from scratch, and will also work with more complex objects provided by the `curatedTCGAData` package.

13.2.3 R/Bioconductor packages used

- [MultiAssayExperiment]
- [RaggedExperiment]
- [curatedTCGAData]

- [SummarizedExperiment]

13.2.4 Time outline

1h 45m total

Activity	Time
Overview of key data classes	25m
Working with RaggedExperiment	20m
Building a MultiAssayExperiment from scratch	10m
Creating and TCGA multi-assay dataset	10m
Subsetting and reshaping multi-assay data	20m
Plotting, correlation, and other analyses	20m

13.3 Workshop goals and objectives

13.3.1 Learning goals

- identify appropriate data structures for different 'omics data types
- gain familiarity with GRangesList and RaggedExperiment

13.3.2 Learning objectives

- use curatedTCGAData to create custom TCGA MultiAssayExperiment objects
- create a MultiAssayExperiment for TCGA or other multi'omics data
- perform subsetting, reshaping, growing, and extraction of a MultiAssayExperiment
- link MultiAssayExperiment data with packages for differential expression, machine learning, and plotting

Chapter 14

Biomarker discovery from large pharmacogenomics datasets

14.1 Instructors:

- Zhaleh Safikhani (zhaleh.safikhani@utoront.ca)
- Petr Smirnov (petr.smirnov@mail.utoronto.ca)
- Benjamin Haib-Kains (benjamin.haibe.kains@utoronto.ca)

14.2 Workshop Description

This workshop will focus on the challenges encountered when applying machine learning techniques in complex, high dimensional biological data. In particular, we will focus on biomarker discovery from pharmacogenomic data, which consists of developing predictors of response of cancer cell lines to chemical compounds based on their genomic features. From a methodological viewpoint, biomarker discovery is strongly linked to variable selection, through methods such as Supervised Learning with sparsity inducing norms (e.g., ElasticNet) or techniques accounting for the complex correlation structure of biological features (e.g., mRMR). Yet, the main focus of this talk will be on sound use of such methods in a pharmacogenomics context, their validation and correct interpretation of the produced results. We will discuss how to assess the quality of both the input and output data. We will illustrate the importance of unified analytical platforms, data and code sharing in bioinformatics and biomedical research, as the data generation process becomes increasingly complex and requires high level of replication to achieve robust results. This is particularly relevant as our portfolio of machine learning techniques is ever enlarging, with its set of hyperparameters that can be tuning in a multitude of ways, increasing the risk of overfitting when developing multivariate predictors of drug response.

14.2.1 Pre-requisites

- Basic knowledge of R syntax
- Familiarity with the machine learning concept and at least a few approaches

Following resources might be useful to read:

- https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/btv723
- https://academic.oup.com/nar/article/46/D1/D994/4372597
- https://web.stanford.edu/~hastie/Papers/ESLII.pdf

14.2.2 Workshop Participation

Participants expected to have the following required packages installed on their machines to be able to run the commands along with the instructors. * PharmacoGx and Biobase from Bioconductor * mRMRe, caret, glmnet, randomForest from cran * bhklab/mci and bhklab/PharmacoGx-ML from github

14.2.3 *R / Bioconductor* packages used

- https://bioconductor.org/packages/release/bioc/html/PharmacoGx.html

14.2.4 Time outline

An example for a 45-minute workshop:

Activity	Time
Introduction	10m
Basic functionalities of PharmacoGx	15m
Consistency assessment between datasets	15m
Machine learning and biomarker discovery	20m

14.3 Workshop goals and objectives

14.3.1 Learning goals

- describe the pharmacogenomic datasets and their usefulness
- learn how to extract information from these datasets and to intersect them over their common features
- identify functionalities available in PharmacoGx package to work with the high dimensional pharmacogenomics data
- assess reproducibility and replication of pharmacogenomics studies
- understand how to handle the biomarker discovery as a pattern recognition problem in the domain of pharmacogenomics studies

14.3.2 Learning objectives

- list available standardized pharmacogenomic datasets and download them
- understand the structure of these darasest and how to access the features and response quantifications
- create drug-dose response plots
- Measure the consistency across multiple datasets and how to improve such measurements
- Assess whether known biomarkers are reproduced within these datasets
- Predict new biomarkers by applying different machine learning methods

Chapter 15

Maintaining your Bioconductor package

15.1 Authors:

- Nitesh Turaga (nitesh.turaga@roswellpark.org)

15.2 Workshop Description:

Once an R package is accepted into Bioconductor, maintaining it is an active responsibility undertaken by the package developers and maintainers. In this short workshop, we will cover how to maintain a Bioconductor package using existing infrastructure. Bioconductor hosts a range of tools which maintainers can use such as daily build reports, landing page badges, RSS feeds, download stats, support site questions, and the bioc-devel mailing list. Some packages have their own continuous integration hook setup on their github pages which would be an additional tool maintainers can use to monitor their package. We will also highlight one particular git practice which need to be done while updating and maintaining your package on out git system.

15.3 Pre-requisites:

Accepted Bioconductor package or plans to contribute a Bioconductor package in the future.

15.4 Participation:

Students will be expected to follow along with their laptops if they choose to, although it is not needed.

15.5 Time outline: 50 mins short workshop

Introduction: 10 mins Why maintain your package Infrastructure used for maintenance Outline how to use infrastructure: 35 mins Build report Landing page badges RSS feeds Download statistics Support site,

Github issues, Bioc-devel mailing list Sync package with Github / Bioconductor before every change GitHub + webhooks Round up of resources available: 5 mins

15.6 Workshop goals and objectives:

- Gain knowledge of how to track Bioconductor packages via download statistics, RSS feeds.
- Understand the importance of supporting their user community and how to do so using the support site and bioc-devel mailing list.
- Maintain their package and fix bugs using build reports, continuous integration tools.
- Update their package consistently on both Github and Bioconductor private git server.

Chapter 16

Public Data Resources and Bioconductor

16.1 Instructors

- Levi Waldron, City University of New York, New York, NY, USA
- Benjamin Haibe-Kain, Princess Margaret Cancer Center, Toronto, Canada
- Sean Davis, Center for Cancer Research, National Cancer Institute, National Institutes of Health, Bethesda, MD, USA

16.2 Workshop Description

The goal of this workshop is to introduce Bioconductor packages for finding, accessing, and using large-scale public data resources including the Gene Expression Omnibus GEO, Sequence Read Archive SRA, the Genomic Data Commons GDC, and Bioconductor-hosted curated data resources for metagenomics, pharmacogenomics, and The Cancer Genome Atlas.

16.2.1 Pre-requisites

- Basic knowledge of R syntax
- Familiarity with the ExpressionSet and SummarizedExperiment classes
- Basic familiarity with 'omics technologies such as microarray and NGS sequencing

Interested students can prepare by reviewing vignettes of the packages listed in "R/Bioconductor packages used" to gain background on aspects of interest to them.

Some more general background on these resources is published in:

Kannan L, Ramos M, Re A, El-Hachem N, Safikhani Z, Gendoo DMA, Davis S, Gomez-Cabrero D, Castelo R, Hansen KD, Carey VJ, Morgan M, Culhane AC, Haibe-Kains B, Waldron L: **Public data and open source tools for multi-assay genomic investigation of disease.** *Brief. Bioinform.* 2016, 17:603–615. (link)

16.2.2 Workshop Participation

Each component will include runnable examples of typical usage that students are encouraged to run during demonstration of that component.

16.2.3 R/Bioconductor packages used

- GEOquery: Access to the NCBI Gene Expression Omnibus (GEO), a public repository of gene expression (primarily microarray) data.
- GenomicDataCommons: Access to the NIH / NCI Genomic Data Commons RESTful service.
- SRAdb: A compilation of metadata from the NCBI Sequence Read Archive, the largest public repository of sequencing data from the next generation of sequencing platforms, and tools
- curatedTCGAData: Curated data from The Cancer Genome Atlas (TCGA) as MultiAssayExperiment Objects
- curatedMetagenomicData: Curated metagenomic data of the human microbiome
- HMP16SData: Curated metagenomic data of the human microbiome
- PharmacoGx: Analysis of large-scale pharmacogenomic data

16.2.4 Time outline

This is a 1h45m workshop.

Activity	Time
Overview	10m
GEOquery	15m
GenomicDataCommons	20m
Sequence Read Archive	20m
curatedTCGAData	10m
curatedMetagenomicData and HMP16SData	15m
PharmacoGx	20m

16.3 workshop goals and objectives

Bioconductor provides access to significant amounts of publicly available experimental data. This workshop introduces students to Bioconductor interfaces to the NCBI's Gene Expression Omnibus, Genomic Data Commons, Sequence Read Archive and PharmacoDB. It additionally introduces curated resources providing The Cancer Genome Atlas, the Human Microbiome Project and other microbiome studies, and major pharmacogenomic studies, as native Bioconductor objects ready for analysis and comparison to in-house datasets.

16.3.1 Learning goals

- search NCBI resources for publicly available 'omics data
- quickly use data from the TCGA and the Human Microbiome Project

16.3.2 Learning objectives

- find and download processed microarray and RNA-seq datasets from the Gene Expression Omnibus

- find and download 'omics data from the Genomic Data Commons and Sequence Read Archive
- download and manipulate data from The Cancer Genome Atlas and Human Microbiome Project
- download and explore pharmacogenomics data