episummer
@columbia

Introduction to Bioconductor

CLASS SESSIONS

Digital workshop, 5 hours
Materials will be available online from June 1st, 2020 - June 30th, 2020

INSTRUCTORS:

Dr. Levi Waldron
Associate Professor of Biostatistics
Email: levi.waldon@sph.cuny.edu

Dr. Ludwig Geistlinger
Post-doctoral fellow in cancer genomics
Email: ludwig.geistlinger@sph.cuny.edu

Institute for Implementation Science in Population Health
Graduate School of Public Health and Health Policy
City University of New York

COURSE DESCRIPTION

The *Bioconductor* project provides open-source software based on the *R* programming language for statistical analysis and visualization of high-throughput genomic data. This course provides a broad introduction to the project, from navigating its large collection of packages to its core functionality for representation, manipulation, and visualization of genomic data. We will learn how to efficiently analyze genomic intervals and SNPs, how to manage experiments of one or more genomic data type with clinical and pathological data, and how to visualize genomic data. This workshop equips participants with essential background for a wide range of applications in statistical genomics and genetic epidemiology, such as GWAS, RNA-seq, DNA methylation, ChIP-seq, metagenomics, and multi'omic experiments.

PREREQUISITES

This workshop is accessible for those with little or no experience using *Bioconductor*, although even more experienced users can benefit from the broad overview of *Bioconductor* paradigms. The workshop assumes elementary knowledge of R, which can be gained in advance or simultaneously from other courses such as the introductory course from DataCamp. A basic understanding of genome biology and statistical analysis is helpful, but specific prerequisites are not needed.

TECHNICAL REQUIREMENTS

R and Bioconductor: www.bioconductor.org/install
R Studio: https://www.rstudio.com/products/rstudio/download3/

COURSE OBJECTIVES

Part 1: *Bioconductor*

- Find, install, and learn how to use *Bioconductor* packages.
- Import and manipulate genomic files and *Bioconductor* data objects.
- Start an RNA-seq differential expression workflow.

Part 2: *Data structures for representing 'omics experiments*

- Use the ExpressionSet data structure to represent, manipulate, and analyze microarray data
- Use the SummarizedExperiment data structure to represent, manipulate, and analyze RNA-seq data
- Understand similarities and differences between the two data structures
- Create both data structures from public data resources
- Use the MultiAssayExperiment data structure to coordinate multi'omics experiments

Part 3: *GenomicRanges*

- Understand how to apply the *Ranges infrastructure to solve common bioinformatic challenges in genomic research
- Gain insight into the design principles of the infrastructure and how it is meant to be used
- Learn basics of genomic region algebra and how to carry out intra- and inter-region operations

Part 4: *Visualizing genomic data*

- Understand basic principles of the grammar of graphics used in *R/Bioconductor*
- Learn how to display heatmaps for genomic data exploration
- Learn how to display genomic data tracks in a genome browser view

DETAILED COURSE OBJECTIVES

Part 1: *Introduction to Bioconductor*

- Discover, install, and read the vignette of the *DESeq2* package.
- Discover the 'single cell sequencing' vignette
- Import BED and GTF files into *Bioconductor*
- Find regions of overlap between the BED and GTF files.
- Import a matrix and data.frame into *Bioconductor*'s SummarizedExperiment object for RNA-Seq differential expression.

Part 2: *Data structures for representing ‘omics experiments*

- Load, inspect, subset, and manipulate the ExpressionSet from the *ALL* package
- Load, inspect, subset, and manipulate the SummarizedExperiment from the *airway* package
- Understand similarities between *exprs, pData, fData* (ExpressionSet) and *assay, colData, rowData* (SummarizedExperiment)
- Understand the extended design of the SummarizedExperiment for dealing with big data and genomic ranges
- Create an ExpressionSet using the *GEOquery* package
- Create a SummarizedExperiment using the *recount* package
- Create and manipulate a MultiAssayExperiment

Part 3: *GenomicRanges*

- Understand *S4Vectors*, *IRanges*, and *GRanges*
- Use *GRanges* to represent, manipulate, and analyze genomic ranges
- Use the ranges algebra to carry out basic operations on genomic ranges
- Find overlaps between genomic ranges

Part 4: *Visualizing genomic data*

- Understand basic principles of plotting with *base R* and *ggplot2*
- Use the *ComplexHeatmap* package for displaying a differential expression setup
- Use the *Gviz* package to display genomic data tracks from UCSC

COURSE SCHEDULE

Activity	**Time**
Introduction to *Bioconductor*	75m
- Project history	
- Discovering and using packages	
- Working with objects	

- Exercises	

ExpressionSet / SummarizedExperiment	75m
- ExpressionSet	
- SummarizedExperiment	
- Similarities & differences	
- Construction from public data	
- Exercises	

Genomic Ranges	75m
- S4Vectors, IRanges, GRanges	
- Ranges algebra	
- Basic operations	
- Finding overlaps	
- Exercises	

Visualizing genomic data	75m

- Graphics overview	
- Creating heatmaps	
- Genomic data tracks	
- Exercises	