systemPipeR: a generic workflow design and reporting environment in R with built-in support for command-line software

**Authors**: Daniela Cassol ([danielac@ucr.edu](mailto:danielac@ucr.edu)), Le Zhang (“[le.zhang001@email.ucr.edu](mailto:le.zhang001@email.ucr.edu)”), Thomas Girke ([thomas.girke@ucr.edu](mailto:thomas.girke@ucr.edu)).

**Institution**: Institute for Integrative Genome Biology, University of California, Riverside, California, USA.

**Last modified**: 8 March, 2021.

# Overview

## Workshop Description

This workshop demonstrates the use of the *systemPipeR* project environment to designing, configuring, and running end-to-end reproducible analysis workflows while generating, at the same time, publication-quality analysis reports.

The introduction of the workshop will provide an overview of *systemPipeR* major features. We will help participants understand the package’s components environment, such as built-in support for command-line software, integrating both command-line and R/Bioconductor individual software steps, and the flexibility to choose and run a single or any number of complex workflow steps while monitoring its execution status and tracking all metadata associated with a project. We will explore the latter, with the new workflow control class (S4) added to the *systemPipeR* package.

A short use-case demonstration will guide users through the basic usage of a pre-configured workflow template provided by the *systemPipeRdata* package. This workshop will overview a small-RNA-Seq profiling workflow that will include reading mappings with variable parameter settings of two popular short-read aligners (HISAT2 and STAR), read quantification, and statistical analysis steps, as well as automation routines for running NGS workflows from start to finish with a single command. This workshop will detail how to parallelize the analysis of relatively large data sets on multiple CPU cores of single machines as well as computer clusters with a scheduler (e.g., Slurm). Simultaneously, the hands-on components will cover the customization and demonstration of the construction of new workflows.

The last part will demonstrate *systemPipeShiny* (SPS), the new graphical user interface for workflow management and visualization. SPS has been designed as a general purpose framework for interacting with other R packages in an intuitive manner.

## Pre-requisites

* Basic knowledge of R and usage of Bioconductor packages for NGS analysis
* Basic knowledge of running command-line software
* Basic knowledge of parallelization concepts

Non-essential background reading:

* [systemPipeR vignette](https://bioconductor.org/packages/devel/bioc/html/systemPipeR.html)
* [systemPipeShiny vignette](https://bioconductor.org/packages/devel/bioc/html/systemPipeShiny.html)
* [systemPipeRdata vignette](http://bioconductor.org/packages/devel/data/experiment/html/systemPipeRdata.html)
* [R Markdown tutorial](https://rmarkdown.rstudio.com/lesson-2.html)

## Workshop Participation

Participants will be able to perform all analysis components of this workshop hands-on. Active user participation throughout the event is highly encouraged including but not limited to lecture material, hands-on sections and final discussion about package improvements. Participants are encouraged to ask questions at any time during the workshop.

## *R* / *Bioconductor* packages used

* [systemPipeR](http://www.bioconductor.org/packages/release/bioc/html/systemPipeR.html)
* [systemPipeRdata](http://www.bioconductor.org/packages/release/data/experiment/html/systemPipeRdata.html)
* [R Markdown tutorial](https://rmarkdown.rstudio.com/lesson-2.html)

## Time outline

1h 45m total

|  |  |
| --- | --- |
| Activity | Time |
| Introduction | 5m |
| Overview of *systemPipeR* package | 10m |
| Introduction to new S4 class: SYSargs2 and SYSargsList | 20m |
| Showcase small RNA-Seq workflow | 20m |
| Parallelization on single machines and clusters | 10m |
| Building a custom workflow | 20m |
| Overview of *systemPipeShiny* package | 20m |

## Workshop goals and objectives

### Learning goals

* Recognize the benefits of generic R-based workflow construction environment that is both scalable and reproducible
* Integration of command-line tools via the CWL community standard
* Rendering of R markdown reports and critical assessment of scientific analysis reports
* Parallelization of big data analysis tasks

### Learning objectives

* Identify and practice how to make analysis workflows more robust, reproducible and portable across heterogeneous computing systems
* Usage of new workflow control class
* Optimize and debug workflows
* Inspection of technical reports and log files
* Design of new and fully customized workflows

# Workshop

## Workshop setup with Docker