systemPipeRdata: NGS workflow templates and sample data

Author: Daniela Cassol (danielac@ucr.edu) and Thomas Girke (thomas.girke@ucr.edu)

Last update: 01 December, 2018

Package

systemPipeRdata 1.9.1

Contents

Note: the most recent version of this vignette can be found here and a short overview slide show here.

1 Introduction

systemPipeRdata is a helper package to generate with a single command NGS workflow templates that are intended to be used by its parent package *systemPipeR* (H Backman and Girke 2016). The latter is an environment for building *end-to-end* analysis pipelines with automated report generation for next generation sequence (NGS) applications such as RNA-Seq, Ribo-Seq, ChIP-Seq, VAR-Seq and many others. The directory structure of the workflow templates and the sample data used by *systemPipeRdata* are described here.

2 Getting Started

2.1 Installation

The R software for using *systemPipeRdata* can be downloaded from CRAN. The *systemPipeRdata* package can be installed from within R as follows:

```
if (!requireNamespace("BiocManager", quietly = TRUE)) install.packages("BiocManager")
BiocManager::install("systemPipeRdata")  # Installs from Bioconductor once
# available there
BiocManager::install("tgirke/systemPipeR", build_vignettes = TRUE,
    dependencies = TRUE)  # Installs from github
```

2.2 Loading package and documentation

```
library("systemPipeRdata")  # Loads the package
```

```
library(help = "systemPipeRdata")  # Lists package info
vignette("systemPipeRdata")  # Opens vignette
```

2.3 Generate workflow template

Load one of the available NGS workflows into your current working directory. The following does this for the *varseq* template. The name of the resulting workflow directory can be specified under the *mydirname* argument. The default *NULL* uses the name of the chosen workflow. An error is issued if a directory of the same name and path exists already.

```
genWorkenvir(workflow = "varseq", mydirname = NULL)
setwd("varseq")
```

On Linux and OS X systems the same can be achieved from the command-line of a terminal with the following commands.

```
Rscript -e "systemPipeRdata::genWorkenvir(workflow='varseq', mydirname=NULL)"
```

The workflow templates generated by `genWorkenvir` contain the following preconfigured directory structure:

- ***workflow/*** (*e.g. rnaseq/*)
  - This is the directory of the R session running the workflow.
  - Run script (**.Rmd* or **.Rnw*) and sample annotation (*targets.txt*) files are located here.
  - Note, this directory can have any name (*e.g.* ***rnaseq***, ***varseq***). Changing its name does not require any modifications in the run script(s).
  - Important subdirectories:
    - ***param/***
      - Stores parameter files such as: **.param*, **.tmpl* and **_run.sh*.
    - ***data/***
      - FASTQ samples
      - Reference FASTA file
      - Annotations
      - etc.
    - ***results/***
      - Alignment, variant and peak files (BAM, VCF, BED)
      - Tabular result files
      - Images and plots
      - etc.

2.4 Run workflows

Next, run from within R the chosen sample workflow by executing the code provided in the corresponding `*.Rnw` template file. If preferred the corresponding `*.Rmd` or `*.R` versions can be used instead. Alternatively, one can run an entire workflow from start to finish with a single command by executing from the command-line `'make -B'` within the workflow directory (here `'varseq'`). Much more detailed information on running and customizing `systemPipeR` workflows is available in its overview vignette here. This vignette can also be opened from R with the following command.

```
library("systemPipeR")
# Loads systemPipeR which needs to be installed via
# BiocManager from Bioconductor
```

```
vignette("systemPipeR", package = "systemPipeR")
```

2.5 Return paths to sample data

The location of the sample data provided by `systemPipeRdata` can be returned as a `list`.

```
pathList()
## $targets
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/param/targets.txt"
##
```

```
## $targetsPE
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/param/targetsPE.txt"
##
## $annotationdir
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/annotation/"
##
## $fastqdir
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/fastq/"
##
## $bamdir
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/bam/"
##
## $paramdir
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/param/"
##
## $workflows
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/workflows/"
##
## $chipseq
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/workflows/chipseq/"
##
## $rnaseq
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/workflows/rnaseq/"
##
## $riboseq
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/workflows/riboseq/"
##
## $varseq
## [1] "/home/dcassol/R/x86_64-pc-linux-gnu-library/3.5/systemPipeRdata/extdata/workflows/varseq/"
```

3 Version information

```
sessionInfo()
## R version 3.5.1 (2018-07-02)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 18.04.1 LTS
##
## Matrix products: default
## BLAS: /usr/lib/x86_64-linux-gnu/blas/libblas.so.3.7.1
## LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.7.1
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
```

```
##
## attached base packages:
## [1] stats4    parallel  stats     graphics  grDevices
## [6] utils     datasets  methods   base
##
## other attached packages:
##  [1] systemPipeRdata_1.9.1       systemPipeR_1.15.5
##  [3] ShortRead_1.40.0            GenomicAlignments_1.18.0
##  [5] SummarizedExperiment_1.12.0 DelayedArray_0.8.0
##  [7] matrixStats_0.54.0          Biobase_2.42.0
##  [9] BiocParallel_1.16.2         Rsamtools_1.34.0
## [11] Biostrings_2.50.1           XVector_0.22.0
## [13] GenomicRanges_1.34.0        GenomeInfoDb_1.18.1
## [15] IRanges_2.16.0              S4Vectors_0.20.1
## [17] BiocGenerics_0.28.0         BiocStyle_2.10.0
##
## loaded via a namespace (and not attached):
##  [1] Category_2.48.0        bitops_1.0-6
##  [3] bit64_0.9-7            RColorBrewer_1.1-2
##  [5] progress_1.2.0         httr_1.3.1
##  [7] rprojroot_1.3-2        Rgraphviz_2.26.0
##  [9] tools_3.5.1            backports_1.1.2
## [11] R6_2.3.0               DBI_1.0.0
## [13] lazyeval_0.2.1         colorspace_1.3-2
## [15] withr_2.1.2            tidyselect_0.2.5
## [17] prettyunits_1.0.2      bit_1.1-14
## [19] compiler_3.5.1         graph_1.60.0
## [21] formatR_1.5            rtracklayer_1.42.1
## [23] bookdown_0.7           checkmate_1.8.5
## [25] scales_1.0.0           genefilter_1.64.0
## [27] RBGL_1.58.1            rappdirs_0.3.1
## [29] stringr_1.3.1          digest_0.6.18
## [31] rmarkdown_1.10         AnnotationForge_1.24.0
## [33] pkgconfig_2.0.2        htmltools_0.3.6
## [35] limma_3.38.2           rlang_0.3.0.1
## [37] RSQLite_2.1.1          bindr_0.1.1
## [39] GOstats_2.48.0         hwriter_1.3.2
## [41] dplyr_0.7.8            RCurl_1.95-4.11
## [43] magrittr_1.5           GO.db_3.7.0
## [45] GenomeInfoDbData_1.2.0 Matrix_1.2-15
## [47] Rcpp_1.0.0             munsell_0.5.0
## [49] stringi_1.2.4          yaml_2.2.0
## [51] edgeR_3.24.0           zlibbioc_1.28.0
## [53] plyr_1.8.4             grid_3.5.1
## [55] blob_1.1.1             crayon_1.3.4
## [57] lattice_0.20-38        splines_3.5.1
## [59] GenomicFeatures_1.34.1 annotate_1.60.0
## [61] hms_0.4.2              batchtools_0.9.11
## [63] locfit_1.5-9.1         knitr_1.20
## [65] pillar_1.3.0           rjson_0.2.20
## [67] base64url_1.4          codetools_0.2-15
```

```
## [69] biomaRt_2.38.0         XML_3.98-1.16
## [71] glue_1.3.0             evaluate_0.12
## [73] latticeExtra_0.6-28    data.table_1.11.8
## [75] BiocManager_1.30.4     gtable_0.2.0
## [77] purrr_0.2.5            assertthat_0.2.0
## [79] ggplot2_3.1.0          xfun_0.4
## [81] xtable_1.8-3           survival_2.43-3
## [83] pheatmap_1.0.10        tibble_1.4.2
## [85] AnnotationDbi_1.44.0   memoise_1.1.0
## [87] bindrcpp_0.2.2         brew_1.0-6
## [89] GSEABase_1.44.0
```

4 Funding

This project was supported by funds from the National Institutes of Health (NIH) and the National Science Foundation (NSF).

References

H Backman, Tyler W, and Thomas Girke. 2016. "systemPipeR: NGS workflow and report generation environment." *BMC Bioinformatics* 17 (1): 388. doi:10.1186/s12859-016-1241-0.