systemPipeR: NGS Workflow and Report Generation Environment

Daniela Cassol & Thomas Girke

Institute for Integrative Genome Biology, University of California, Riverside, California 92521

Introduction

systemPipeR is an R/Bioconductor package for building and running automated end-to-end analysis workflows for a wide range of next generation sequence (NGS) applications. Important features include a uniform workflow interface across different NGS applications, automated report generation, and support for running both R and command-line software, such as NGS aligners or peak/variant callers, on local computers or HPC clusters. The latter supports interactive job submissions and batch submissions to queuing systems of clusters. Efficient handling of complex sample sets and experimental designs is facilitated by a well-defined sample and metadata annotation infrastructure which improves reproducibility and user-friendliness of many typical analysis workflows in the NGS area.

Recently obtained funding from NSF (Grant: ABI-1661152) allows us to continue the development of the package. Our main development efforts for the next major upgrade of *systemPipeR* include: (1) major enhancements to the workflow environment, as well as to the user and command-line interfaces; (2) release of a total of 10 ready to use and community tested workflow templates developed in collaboration with NGS domain experts; (3) addition of various convenience functions for building interactive analysis reports and visualization routines; (4) a strong focus on community integration and performance evaluations provided by *systemPipeR's* current and future users. The latter includes options for users to contribute code or entire workflows, and extensive training of the target audience to analyze NGS data with *systemPipeR* and related R/Bioconductor resources. This poster gives an overview of both existing and upcoming features of *systemPipeR*.

systemPipeR NGS solutions

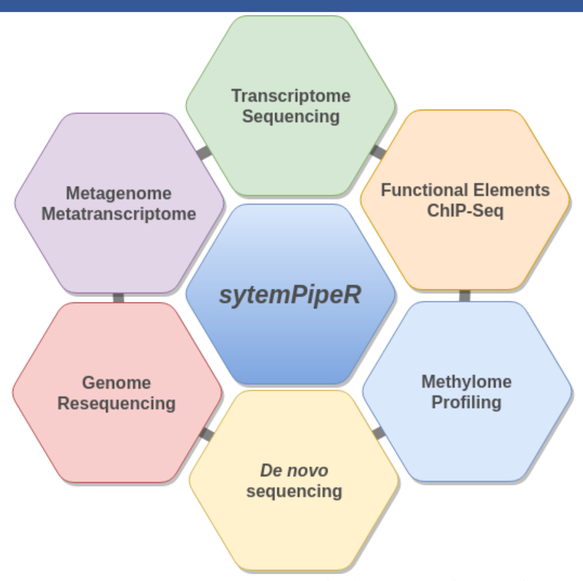

Figure 1: *systemPipeR* provides automated start-to-finish workflows for many NGS application. A common workflow environment has several advantages for improving efficiency, standardization, and reproducibility.

Important Features

Command-line software support: An important feature of *systemPipeR* is support for running command-line software directly from R on both single machines or computer clusters (Backman *et al.*, 2016). This offers several advantages such as seamless integration of most command-line software with the extensive genome analysis resources provided by R/Bioconductor (Huber *et al.*, 2015).

Parallel evaluation: The processing time for NGS experiments can be greatly reduced by making use of parallel evaluation across several CPU cores on single machines, or multiple nodes of computer clusters and cloud-based systems. *systemPipeR* simplifies these parallelization tasks without creating any limitations for users who do not have access to high-performance computer resources.

Automated Analysis Reports: *systemPipeR* generates automated analysis reports with *knitr* and *R markdown* (Xie, 2013). These modern reporting environments integrate R code with LaTeX or Markdown. During the evaluation of the R code, reports are dynamically generated in PDF or HTML format.

Overview of Important Functions

Function Name	Description
`genWorkenvir`	Generates workflow templates provided by *systemPipeRdata* helper package
`systemArgs`	Constructs `SYSargs` workflow control module (S4 object) from *targets* and *param* files
`runCommandline`	Executes command-line software on samples and parameters specified in `SYSargs`
`clusterRun`	Runs command-line software in parallel mode on a computer cluster
`preprocessReads`	Filtering and/or trimming of short reads using predefined or custom parameters
`seeFASTQ/seeFASTQplot`	Generates quality reports for any number of FASTQ files
`alignStats`	Generates alignment statistics, such as total number of reads and alignment frequency
`run_edgeR/run_DESeq2`	Runs *edgeR* or *DESeq2* for any number of pairwise sample comparisons
`filterDEGs`	Filters and plots DEG results based on user-defined parameters
`overLapper/vennPlot`	Computation of Venn intersects for 2-20 or more samples and 2-5 way Venn diagrams
`GOCluster_Report`	GO term enrichment analysis for large numbers of gene sets
`variantReport`	Generates a variant report containing genomic annotations and confidence statistics
`predORF`	Prediction of short open reading frames in DNA sequences
`featuretypeCounts`	Computes and plots read distribution for many feature types at once
`featureCoverage`	Computes and plots read depth coverage from many transcripts

Table 1: The table lists a subset of over 50 methods and functions defined by *systemPipeR*. Usage instructions are provided in the corresponding help pages and vignettes of the package.

Upcoming Features

User Interface: Several enhancements to the current workflow definition classes (S4) will allow users to execute complex workflows or their components in a summarized manner using R's bracket operator (*runWorkflowSteps[1:4]*) and/or pipes (*Step1 %>% Step2 %>% ...*).

Common Workflow Language (CWL): CWL has become a community standard for describing and executing data analysis workflows. Adopting CWL in *systemPipeR* will result in a higher level of standardization of how workflows are designed, described and executed, while broadening its user community. Additional benefits are (i) flexible options to run workflows from the command-line or from other computer languages; (ii) visualization of workflows in form of graphs with existing tools; and (iii) sharing of workflows with related environments (e.g. Galaxy or Snakemake).

Shiny Web Interface: *systemPipeR* will provide options to run workflows from easy-to-use Shiny apps.

Workflow Templates: Workflow templates for $\geq$10 NGS applications will be released. In addition, a public workflow repository will allow users to submit and share custom workflows.

Visual, Scientific and Technical Reports: *systemPipeR's* reporting infrastructure will be expanded to three types of interconnected reports each serving a different purpose: (i) a scientific report will include scientifically relevant results (R Markdown style); (ii) a technical report will document all technical information important for each workflow step including parameter settings, software versions, and warning/error messages; and (iii) a visual report will depict the entire workflow including its run status in form of a workflow graph.

Containerization: Workflow templates will be distributed as Singularity containers.

Workflow Steps and Graphical Features

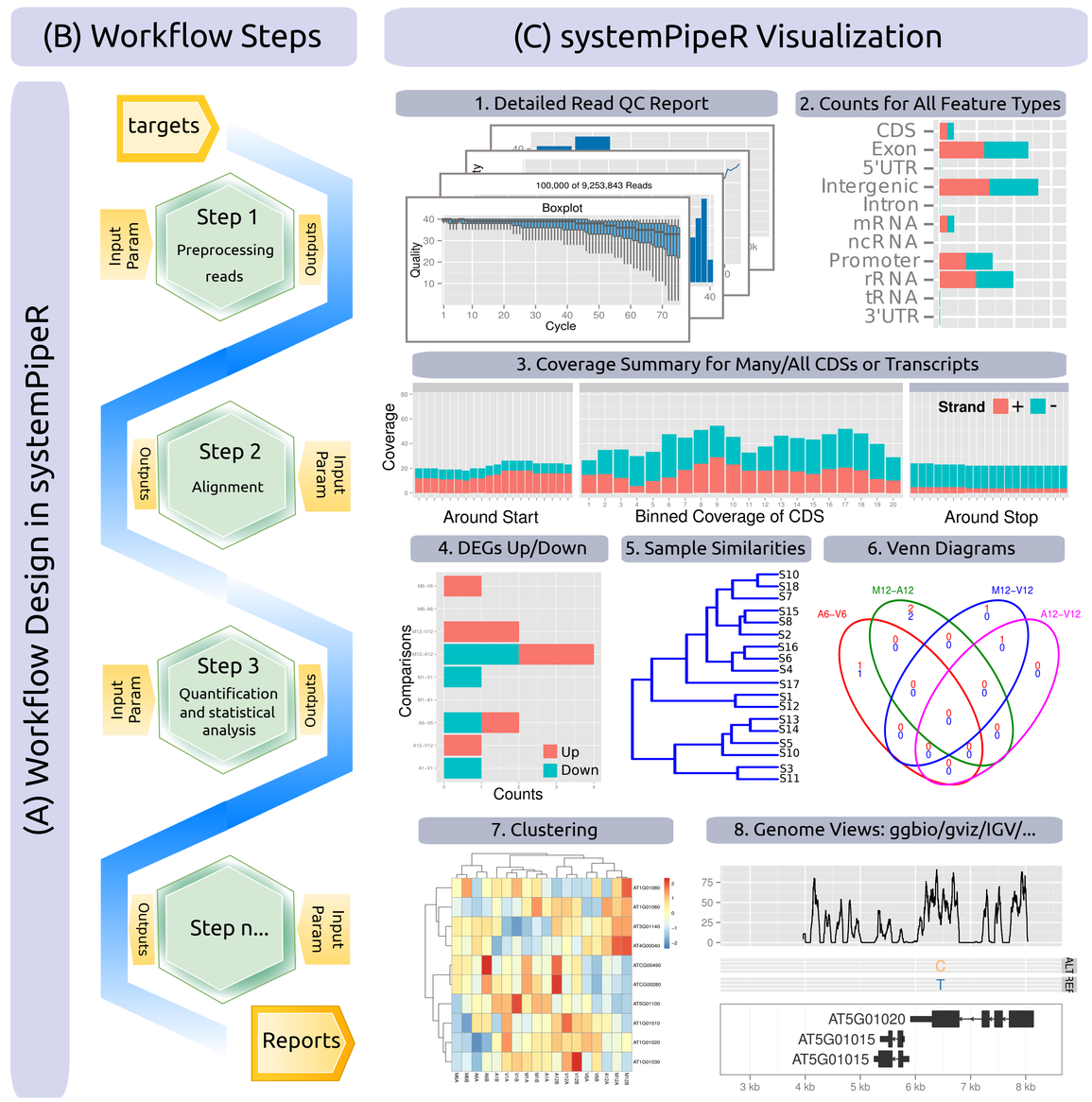

Figure 2: *Relevant workflow features in systemPipeR*. Workflow design concepts are illustrated under (A & B). Examples of *systemPipeR's* functionalities are given under (C) including: (1) plots for summarizing the quality and diversity of short reads; (2) strand-specific read count summaries for all feature types provided by a genome annotation; (3) summary plots of read depth coverage for any number of transcripts, as well as binned coverage for their coding regions; (4) enumeration of up- and down-regulated DEGs for user defined sample comparisons; (5) similarity clustering of sample profiles; (6) 2-5-way Venn diagrams for DEGs, peak and variant sets; (7) gene-wise clustering with a wide range of algorithms; and (8) support for plotting read pileups and variants in the context of genome annotations along with genome browser support.

Conclusions

systemPipeR accelerates the extraction of reproducible analysis results from NGS experiments. By combining the capabilities of many R/Bioconductor and command-line tools, it makes efficient use of existing software resources without limiting the user to a set of predefined methods or environments.

Availability: systemPipeR is freely available for all common operating systems from Bioconductor: `http://bioconductor.org/packages/systemPipeR`

Acknowlegement

We acknowledge the Bioconductor core team and community for providing valuable input for developing *systemPipeR*. Funding: This work was supported by grants from the National Science Foundation (PGRP-1546879, ABI-1661152, MCB-1021969, IOS-1546879), the National Institutes of Health (U24AG051129, R01-AI36959) and the National Institute of Food and Agriculture (2011-68004-30154).

References

Backman TW, Girke T (2016) systemPipeR: NGS workflow and report generation environment. BMC Bioinformatics: 17, 1-8.
Huber W, ..., Morgan M (2015) Orchestrating high-throughput genomic analysis with Bioconductor. Nature Methods: 12, 115-121.
Xie Y (2013) Dynamic Documents with R and Knitr (Chapman & Hall/CRC The R Series), 1edn. Boca Raton: Chapman and Hall/CRC.

danielac@ucr.edu