systemPipeR: NGS workflow and report generation environment

Thomas Girke
Email contact: thomas.girke@ucr.edu

December 1, 2014

1 Introduction

systemPipeR provides utilities for building end-to-end analysis workflows with automated report generation for next generation sequence (NGS) applications such as RNA-Seq, ChIP-Seq, VAR-Seq and many others (Girke, 2014). An important feature is support for running command-line software, such as NGS aligners, on both single machines or compute clusters. This includes both interactive job submissions or batch submissions to queuing systems of clusters. For instance, systemPipeR can be used with most command-line aligners such as BWA (Li, 2013; Li and Durbin, 2009), TopHat 2 (Kim et al., 2013) and Bowtie 2 (Langmead and Salzberg, 2012), as well as the R-based NGS aligner Rsubread (Liao et al., 2013). Efficient handling of complex sample sets and experimental designs is facilitated by a well-defined sample annotation infrastructure which improves reproducibility and user-friendliness of many typical analysis workflows in the NGS area (Lawrence et al., 2013).

Templates for setting up custom project reports are provided as *.Rnw files in the vignettes subdirectory of this package. The corresponding PDFs of these report templates are linked here: systemPipeRNAseq, systemPipeChIPseq and systemPipeVARseq.

Contents

1 Introduction 1

2 Getting Started 2
   2.1 Installation .................................................. 2
   2.2 Loading the Package and Documentation ..................... 2
   2.3 Sample FASTQ Files ........................................ 2

3 Structure of targets file 2

4 Structure of param file and SYSargs container 3

5 Workflow 4
   5.1 Define environment settings and samples ........................ 4
   5.2 FASTQ quality report ......................................... 5
   5.3 Alignment with Tophat 2 ...................................... 5
   5.4 Read and alignment count stats ................................ 6
   5.5 Create symbolic links for viewing BAM files in IGV ............ 6
   5.6 Alignment with Bowtie 2 (here for miRNA profiling experiment) 6
   5.7 Read counting for mRNA profiling experiments .................. 7
   5.8 Read counting for miRNA profiling experiments ................ 7
   5.9 Correlation analysis of samples .................................. 7
   5.10 DEG analysis with edgeR ..................................... 8
2 Getting Started

2.1 Installation

The R software for running systemPipeR can be downloaded from CRAN (http://cran.at.r-project.org/). The systemPipeR package can be installed from R using the biocLite install command.

```r
> source("http://bioconductor.org/biocLite.R") # Sources the biocLite.R installation script
> biocLite("systemPipeR") # Installs the package
```

2.2 Loading the Package and Documentation

```r
> library("systemPipeR") # Loads the package
> library(help="systemPipeR") # Lists all functions and classes
> vignette("systemPipeR") # Opens this PDF manual from R
```

2.3 Sample FASTQ Files

The mini sample FASTQ files used by this overview vignette as well as the associated workflow reporting vignettes can be downloaded from here. The chosen data set SRP010938 contains 18 paired-end (PE) read sets from Arabidopsis thaliana (Howard et al., 2013). To minimize processing time during testing, each FASTQ file has been subsetted to 90,000-100,000 randomly sampled PE reads that map to the first 100,000 nucleotides of each chromosome of the A. thaliana genome. The corresponding reference genome sequence (FASTA) and its GFF annotation files (provided in the same download) have been truncated accordingly. This way the entire test sample data set is less than 200MB in storage space. A PE read set has been chosen for this test data set for flexibility, because it can be used for testing both types of analysis routines requiring either SE (single end) reads or PE reads.

3 Structure of targets file

The targets file defines all FASTQ files and sample comparisons of an analysis workflow. The following shows the format of a sample targets file provided by this package. In target files for single end (SE) data sets, the first three columns are mandatory including their column names, while it is four mandatory columns for PE data. All subsequent columns are optional and any number of additional columns can be added as needed.

```r
> library(systemPipeR)
> targetspath <- system.file("extdata", "targets.txt", package="systemPipeR")
> read.delim(targetspath, comment.char = "#")
```
Structure of targets file for paired end (PE) samples.

```r
> library(systemPipeR)
> targetspath <- system.file("extdata", "targetsPE.txt", package="systemPipeR")
> read.delim(targetspath, comment.char = ">#")[1:2,1:6]
```

```text
FileName1  FileName2  SampleName  Factor  SampleLong  Experiment  
1  ./data/SRR446027_1.fastq  ./data/SRR446027_2.fastq  M1A  M1  Mock.1h.A  1  
2  ./data/SRR446028_1.fastq  ./data/SRR446028_2.fastq  M1B  M1  Mock.1h.B  1  
3  ./data/SRR446033_1.fastq  ./data/SRR446033_2.fastq  V6A  V6  Vir.6h.A  1  
4  ./data/SRR446034_1.fastq  ./data/SRR446034_2.fastq  V6B  V6  Vir.6h.B  1  
```

Sample comparisons are defined in the header lines of the targets file starting with '# <CMP>'. The function readComp imports the comparison and stores them in a list. These lines are also optional. They are mainly useful for controlling comparative analysis according to certain biological expectations, such as simple pairwise comparisons in RNA-Seq experiments.

```r
> readComp(file=targetspath, format="vector", delim="-")
```

```
$CMPset1

$CMPset2
[19] "V1-M12"  "V1-A12"  "V1-V12"  "M6-A6"  "M6-V6"  "M6-M12"  "M6-A12"  "M6-V12"  "A6-V6"  
[28] "A6-M12"  "A6-A12"  "A6-V12"  "V6-M12"  "V6-A12"  "V6-V12"  "M12-A12"  "M12-V12"  "A12-V12"
```

4 Structure of param file and SYSargs container

The param file defines the parameters of the command-line software. The following shows the format of a sample param file provided by this package.

```r
> parampath <- system.file("extdata", "tophat.param", package="systemPipeR")
> read.delim(parampath, comment.char = "#")
```

```
PairSet Name Value
1 modules <NA> bowtie2/2.1.0
```
2 modules <NA> tophat/2.0.8b
3 software <NA> tophat
4 cores -p 4
5 other <NA> -g 1 --segment-length 25 -i 30 -I 3000
6 outfile1 -o <FileName1>
7 outfile1 path ./results/
8 outfile1 remove <NA>
9 outfile1 append .tophat
10 outfile1 outextension .tophat/accepted_hits.bam
11 reference <NA> ./data/tair10.fasta
12 infile1 <NA> <FileName1>
13 infile1 path <NA>
14 infile2 <NA> <FileName2>
15 infile2 path <NA>

The systemArgs function imports the definitions of both the param file and the targets file, and stores all relevant information as SYSargs object.

> args <- systemArgs(sysma=parampath, mytargets=targetspath)
> args

An instance of 'SYSargs' for running 'tophat' on 18 samples

Several accessor functions are available that are named after the slot names of the SYSargs object class.

> names(args)
[1] "modules" "software" "cores" "other" "reference" "results" "infile1"
[8] "infile2" "outfile1" "sysargs" "outpaths"

> modules(args)
[1] "bowtie2/2.1.0" "tophat/2.0.8b"
> cores(args)
[1] 4
> outpaths(args)[1]

"/Users/tgirke/Desktop/Rrnaseq/results/SRR446027_1.fastq.tophat/accepted_hits.bam"

> sysargs(args)[1]

"tophat -p 4 -g 1 --segment-length 25 -i 30 -I 3000 -o /Users/tgirke/Desktop/Rrnaseq/results/SRR446027_1.fastq"

The content of the param file can be returned as JSON object as follows (requires rjson package).

> systemArgs(sysma=parampath, mytargets=targetspath, type="json")
[1] "{"modules":{"n1":"\"","v2":\"bowtie2/2.1.0\","n1":\"","v2":\"tophat/2.0.8b\"},"software"}"}

5 Workflow

5.1 Define environment settings and samples

Load package:
> library(systemPipeR)
Construct SYSargs object from param and targets files.

```r
args <- systemArgs(sysma="tophat.param", mytargets="targetsPE.txt")
```

### 5.2 FASTQ quality report

The following `seeFastq` and `seeFastqPlot` functions generate and plot a series of useful quality statistics for a set of FASTQ files including per cycle quality box plots, base proportions, base-level quality trends, relative k-mer diversity, length and occurrence distribution of reads, number of reads above quality cutoffs and mean quality distribution.

```r
fqlist <- seeFastq(fastq=infile1(args), batchsize=10000, klength=8)
pdf("fastqReport.pdf", height=18, width=4*length(fqlist))
seeFastqPlot(fqlist)
dev.off()
```

![Figure 1: QC report for 18 FASTQ files.](image)

### 5.3 Alignment with Tophat 2

Build Bowtie 2 index.

```r
moduleload(modules(args)) # Skip if module system is not available
system("bowtie2-build data/tair10.fa data/tair10.fa")
```

Execute SYSargs on a single machine without submitting to a queuing system of a compute cluster. This way the input FASTQ files will be processed sequentially. If available, multiple CPU cores can be used for processing each file. The number of CPU cores (here 4) to use for each process is defined in the `*.param` file. With `cores(args)` one can return this value from the SYSargs object. Note, if a module system is not installed or used, then the corresponding `*.param` file needs to be edited accordingly by either providing an empty field in the line(s) starting with `module` or by deleting these lines.

```r
bampaths <- runCommandline(args=nums)
```

Alternatively, the computation can be greatly accelerated by processing many files in parallel using several compute nodes of a cluster, where a scheduling/queuing system is used for load balancing. To avoid over-subscription of CPU cores on the compute nodes, the value from `cores(args)` is passed on to the submission command, here `nodes` in the `resources` list object. The number of independent parallel cluster processes is defined under the `Njobs` argument. The following example will run 18 processes in parallel using for each 4 CPU cores. If the resources available on a cluster allow to run all 18 processes at the same time then the shown sample submission will utilize in total 72 CPU cores. Note, `runCluster` can be used with most queueing systems as it is based on utilities from the `BatchJobs` package which supports the use of template files (`*.tmpl`) for defining the run parameters of different schedulers. To run the following code, one needs to have both a conf file (see .BatchJob samples [here](#)) and a template file (see `*.tmpl` samples [here](#)) for the queueing
available on a system. The following example uses the sample conf and template files for the Torque scheduler provided by this package.

```r
> file.copy(system.file("extdata", ".BatchJobs.R", package="systemPipeR"), ".")
> file.copy(system.file("extdata", "torque.tmpl", package="systemPipeR"), ".")
> resources <- list(walltime="20:00:00", nodes=paste0("1:ppn=", cores(args)), memory="10gb")
> reg <- clusterRun(args, conffile=".BatchJobs.R", template="torque.tmpl", Njobs=18, runid="01",
+   resourceList=resources)
```

Useful commands for monitoring progress of submitted jobs

```r
> showStatus(reg)
> file.exists(outpaths(args))
> sapply(1:length(args), function(x) loadResult(reg, x))  # Works after job completion
```

## 5.4 Read and alignment count stats

Generate table of read and alignment counts for all samples.

```r
> read_statsDF <- alignStats(args)
> write.table(read_statsDF, "results/alignStats.xls", row.names=FALSE, quote=FALSE, sep="\t")
```

The following shows the first four lines of the sample alignment stats file provided by the `systemPipeR` package. For simplicity the number of PE reads is multiplied here by 2 to approximate proper alignment frequencies where each read in a pair is counted.

```r
> read.table(system.file("extdata", "alignStats.xls", package="systemPipeR"), header=TRUE)[1:4,]

        FileName Nreads2x Nalign Perc_Aligned Nalign_Primary Perc_Aligned_Primary
1         M1A   192918  177961  92.24697      177961         92.24697
2         M1B   197484  159378  80.70426      159378         80.70426
3         A1A   189870  176055  92.72397      176055         92.72397
4         A1B   188854  147768  78.24457      147768         78.24457
```

## 5.5 Create symbolic links for viewing BAM files in IGV

The genome browser IGV supports reading indexed/sorted BAM via web URLs. This way no unnecessary copies of these large files need to be generated. To enable this approach, an HTML directory with http access needs to be available in the user account (e.g. `/public_html`) of a system. If this is not the case then the BAM files need to be moved or copied to the system where IGV runs. In the following, `htmldir` defines the path to the HTML directory with http access where the symbolic links to the BAM files will be stored. The corresponding URLs will be written to a text file specified under the `urlfile` argument.

```r
> symLink2bam(sysargs=args, htmldir=c("~/.html/", "somedir/"),
+    urlbase="http://myserver.edu/~username/",
+    urlfile="IGVurl.txt")
```

## 5.6 Alignment with Bowtie 2 (here for miRNA profiling experiment)

Run as single process without submitting to cluster, e.g. via `qsub -I`.

```r
> args <- systemArgs(sysma="bowtieSE.param", mytargets="targets.txt")
> bampaths <- runCommandline(args=args)
```

Alternatively, submit the job to compute nodes.
5.7 Read counting for mRNA profiling experiments

Create txdb (needs to be done only once)

```r
> library(GenomicFeatures)
> txdb <- makeTranscriptDbFromGFF(file="data/tair10.gff", format="gff", dataSource="TAIR", species="A. thaliana")
> saveDb(txdb, file='./data/tair10.sqlite')
```

Read counting with `summarizeOverlaps` in parallel mode with multiple cores

```r
> library(BiocParallel)
> txdb <- loadDb('./data/tair10.sqlite')
> eByg <- exonsBy(txdb, by="gene")
> bfl <- BamFileList(outpaths(args), yieldSize=50000, index=character())
> multicoreParam <- MulticoreParam(workers=4); register(multicoreParam); registered()
> counteByg <- bplapply(bfl, function(x) summarizeOverlaps(eByg, x, mode="Union", ignore.strand=TRUE, inter.feature=TRUE, singleEnd=TRUE)) # Note: for strand-specific RNA-Seq set 'ignore.strand=FALSE' and for PE data set 'singleEnd=FALSE'
> countDFeByg <- sapply(seq(along=counteByg), function(x) assays(counteByg[[x]])$counts)
> rownames(countDFeByg) <- names(rowData(counteByg[[1]])); colnames(countDFeByg) <- names(bfl)
> countDFeByg <- apply(countDFeByg, 2, function(x) returnRPKM(counts=x, ranges=eByg))
> write.table(countDFeByg, "results/countDFeByg.xls", col.names=NA, quote=FALSE, sep="\t")
```

5.8 Read counting for miRNA profiling experiments

Download miRNA genes from miRBase

```r
> system("wget ftp://mirbase.org/pub/mirbase/19/genomes/My_species.gff3 -P ./data/")
> gff <- import.gff("./data/My_species.gff3", asRangedData=FALSE)
> gff <- split(gff, elementMetadata(gff)$ID)
> bams <- names(bampaths); names(bams) <- targets$SampleName
> bfl <- BamFileList(bams, yieldSize=50000, index=character())
> countDFmiR <- summarizeOverlaps(gff, bfl, mode="Union", ignore.strand=FALSE, inter.feature=FALSE) # Note: inter.feature=FALSE important since pre and mature miRNA ranges overlap
> rpkmDFmiR <- apply(countDFmiR, 2, function(x) returnRPKM(counts=x, gffsub=gff))
> write.table(rpkmiDFmiR, "results/rpkmDFmiR.xls", col.names=NA, quote=FALSE, sep="\t")
```

5.9 Correlation analysis of samples

The following computes the sample-wise Spearman correlation coefficients from the RPKM normalized expression values. After transformation to a distance matrix, hierarchical clustering is performed with the `hclust` function and the result is plotted as a dendrogram (`sample_tree.pdf`).

```r
> library(ape)
> rpkmDFeByg <- read.table("./results/rpkmDFeByg.xls", check.names=FALSE)
> rpkmDFeByg <- rpkmDFeByg[rowMeans(rpkmiDFmiR) > 50,]
> d <- cor(rpkmiDFmiR, method="spearman")
> hc <- hclust(as.dist(1-d))
> plot.phylo(as.phylo(hc), type="p", edge.col="blue", edge.width=2, show.node.label=TRUE, no.margin=TRUE)
```
5.10 DEG analysis with edgeR

```r
> targets <- read.delim(targetspath, comment="#")
> cmp <- readComp(file=targetspath, format="matrix", delim="-")
> cmp[[1]]

[,1] [,2]
[1,] "M1" "A1"
[2,] "M1" "V1"
[3,] "A1" "V1"
[4,] "M6" "A6"
[5,] "M6" "V6"
[6,] "A6" "V6"
[7,] "M12" "A12"
[8,] "M12" "V12"
[9,] "A12" "V12"

Run edgeR

```r
> countDFeByg <- read.delim("./results/countDFeByg.xls", row.names=1)
> edgeDF <- run_edgeR(countDF=countDFeByg, targets=targets, cmp=cmp[[1]], independent=FALSE, mdsplot="")

Filter and plot DEG results for up and down regulated genes. Because of the small size of the toy data set used by this vignette, the FDR value has been set to a relatively high threshold (here 10%). More commonly used FDR cutoffs are 1% or 5%.

```r
> DEG_list <- filterDEGs(degDF=edgeDF, filter=c(Fold=2, FDR=10))
```
## 5.11 GO term enrichment analysis of DEGs

### 5.11.1 Obtain gene-to-GO mappings

The following shows how to obtain gene-to-GO mappings from `biomaRt` (here for *A. thaliana*) and how to organize them for the downstream GO term enrichment analysis. Alternatively, the gene-to-GO mappings can be obtained for many organisms from Bioconductor’s *db* genome annotation packages or GO annotation files provided by various genome databases. For each annotation this relatively slow preprocessing step needs to be performed only once. Subsequently, the preprocessed data can be loaded with the `load` function as shown in the next subsection.

```r
> library("biomaRt")
> listMarts() # To choose BioMart database
> m <- useMart("ENSEMBL_MART_PLANT"); listDatasets(m)
> m <- useMart("ENSEMBL_MART_PLANT", dataset="athaliana_eg_gene")
> listAttributes(m) # Choose data types you want to download
> go <- getBM(attributes=c("go_accession", "tair_locus", "go_namespace_1003"), mart=m)
> go <- go[go[,3]!="",]; go[,3] <- as.character(go[,3])
> dir.create("./data/GO")
> write.table(go, "data/GO/GOannotationsBiomart_mod.txt", quote=FALSE, row.names=FALSE, col.names=FALSE, sep="\t")
> catdb <- makeCATdb(myfile="data/GO/GOannotationsBiomart_mod.txt", lib=NULL, org="", colno=c(1,2,3), idconv=NULL)
> save(catdb, file="data/GO/catdb.RData")
```
5.11.2 Batch GO term enrichment analysis

Apply the enrichment analysis to the DEG sets obtained in the above differential expression analysis. Note, in the following example the FDR filter is set here to an unreasonably high value, simply because of the small size of the toy data set used in this vignette. Batch enrichment analysis of many gene sets is performed with the GOCluster_Report function. When method="all", it returns all GO terms passing the p-value cutoff specified under the cutoff arguments. When method="slim", it returns only the GO terms specified under the myslimv argument. The given example shows how one can obtain such a GO slim vector from BioMart for a specific organism.

```r
> load("data/GO/catdb.RData")
> DEG_list <- filterDEGs(degDF=edgeDF, filter=c(Fold=2, FDR=50), plot=FALSE)
> up_down <- DEG_list$UporDown; names(up_down) <- paste(names(up_down), ".up_down", sep="")
> up <- DEG_list$Up; names(up) <- paste(names(up), ".up", sep="")
> down <- DEG_list$Down; names(down) <- paste(names(down), ".down", sep="")
> DEGlist <- c(up_down, up, down)
> DEGlist <- DEGlist[sapply(DEGlist, length) > 0]
> BatchResult <- GOCluster_Report(catdb=catdb, setlist=DEGlist, method="all", id_type="gene", CLSZ=2, cutoff=0.9, gocats=c("MF", "BP", "CC"), recordSpecGO=NULL)
> library("biomaRt"); m <- useMart("ENSEMBL_MART_PLANT", dataset="athaliana_eg_gene")
> goslimvec <- as.character(getBM(attributes=c("goslim_goa_accession"), mart=m)[,1])
> BatchResultslim <- GOCluster_Report(catdb=catdb, setlist=DEGlist, method="slim", id_type="gene", myslimv=goslimvec)
```

5.11.3 Plot batch GO term results

The data.frame generated by GOCluster_Report can be plotted with the goBarplot function. Because of the variable size of the sample sets, it may not always be desirable to show the results from different DEG sets in the same bar plot. Plotting single sample sets is achieved by subsetting the input data frame as shown in the first line of the following example.

```r
> gos <- BatchResultslim[grep("M6-V6_up_down", BatchResultslim$CLID), ]
> gos <- BatchResultslim
> pdf("GOslimbarplotMF.pdf", height=8, width=10); goBarplot(gos, gocat="MF"); dev.off()
> goBarplot(gos, gocat="BP")
> goBarplot(gos, gocat="CC")
```
5.12 Clustering and heat maps

The following example performs hierarchical clustering on the RPKM normalized expression matrix subsetted by the DEGs identified in the above differential expression analysis. It uses a Pearson correlation-based distance measure and complete linkage for cluster joining.

```r
> library(pheatmap)
> geneids <- unique(as.character(unlist(DEG_list[[1]])))
> y <- rpkmDFeByg[geneids,]
> pdf("heatmap1.pdf")
> pheatmap(y, scale="row", clustering_distance_rows="correlation", clustering_distance_cols="correlation")
> dev.off()
```
Figure 5: Heat map with hierarchical clustering dendrograms of DEGs.

6 Version Information

> toLatex(sessionInfo())

- R version 3.1.2 (2014-10-31), x86_64-apple-darwin13.4.0
- Base packages: base, datasets, graphics, grDevices, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.28.1, Biobase 2.26.0, BiocGenerics 0.12.1, BiocParallel 1.0.0, Biostrings 2.34.0, DBI 0.3.1, GenomInfoDb 1.2.3, GenomicAlignments 1.2.1, GenomicRanges 1.18.3, IRanges 2.0.0, Rsamtools 1.18.2, RSQLite 1.0.0, S4Vectors 0.4.0, ShortRead 1.24.0, systemPipeR 1.0.0, XVector 0.6.0
- Loaded via a namespace (and not attached): annotate 1.44.0, AnnotationForge 1.8.1, base64enc 0.1-2, BatchJobs 1.5, BBmisc 1.8, BiocStyle 1.4.1, bitops 1.0-6, brew 1.0-6, Category 2.32.0, checkmate 1.5.0, codetools 0.2-9, colorspace 1.2-4, digest 0.6.4, edgeR 3.8.5, fail 1.2, foreach 1.4.2, genefilter 1.48.1, ggplot2 1.0.0, GO.db 3.0.0, GOstats 2.32.0, grid 3.1.2, GSEABase 1.28.0, gtable 0.1.2, hwriter 1.3.2, iterators 1.0.7, lattice 0.20-29, latticeExtra 0.6-26, limma 3.22.1, MASS 7.3-35, Matrix 1.1-4, munsell 0.4.2, pheatmap 0.7.7, plyr 1.8.1, proto 0.3-10, RBGL 1.42.0, RColorBrewer 1.0-5, Rcpp 0.11.3, reshape2 1.4, rjson 0.2.15, scales 0.2.4, sendmailR 1.2-1, splines 3.1.2, stringr 0.6.2, survival 2.37-7, tools 3.1.2, XML 3.98-1.1, xtable 1.7-4, zlibbioc 1.12.0
7 Funding

This software was developed with funding from the Agriculture and Food Research Institute of the National Institute of Food and Agriculture of the USDA (2011-68004-30154), the National Science Foundation (MCB-1021969) and the National Institutes of Health/National Institute of Allergy and Infectious Diseases (5R01 AI036959).

8 References


