

# SeqRetriever

David R. Hill and Shrikar Thodla

March 1, 2016

## Contents

<b>1</b>	<b>Aims</b>	<b>1</b>
<b>2</b>	<b>Design criteria</b>	<b>1</b>
<b>3</b>	<b>Work flow</b>	<b>2</b>
<b>4</b>	<b>Examples</b>	<b>2</b>
4.1	Setup database . . . . .	2
4.2	Select genes . . . . .	2
4.3	Print boxplot . . . . .	2
4.4	Print Heatmap . . . . .	3
4.5	Print boxplot showing only genes that differ significantly between "HLO" and "Lung <sub>A</sub> " . . . . .	4
<b>5</b>	<b>Install from Github</b>	<b>4</b>

## 1 Aims

High throughput RNA sequencing is an incredibly powerful tool for characterizing gene expression events ranging in scale from the complete transcriptome of an organism or tissue to the expression of a single rare RNA isoform accounting for the tiniest fraction of all transcriptional activity in a cell.

SeqRetriever is designed to extract gene abundance data from the raw output of the [Tophat/Cufflinks](#) transcriptome assembly and abundance estimation suite and convert it into R dataframes, commonly used plots, and universal CSV output files. This allows you to ask simple questions about differences in gene expression and generate output that you can share with your colleagues, advisors, collaborators, and reviewers without requiring a degree in computer science or lengthy immersion in boring programming blogs. We get it - you have papers that needed to be written yesterday and you don't have time to learn to program. We're even okay with you analyzing these output files in Excel if you want.

There are a variety of tools available for analyzing RNA-seq datasets in the R environment, including the outstanding [CummeRbund](#) package designed for analyzing Cufflinks output which we highly recommend. SeqRetriever has far fewer features, focusing exclusively on differential expression analysis of assembled RNA-seq datasets. SeqRetriever aims to be the first R package you run to answer a brand new question about gene expression differences, but probably not the last one. This package is designed to get you asking, answering, refining, and re-asking your questions as quickly as possible.

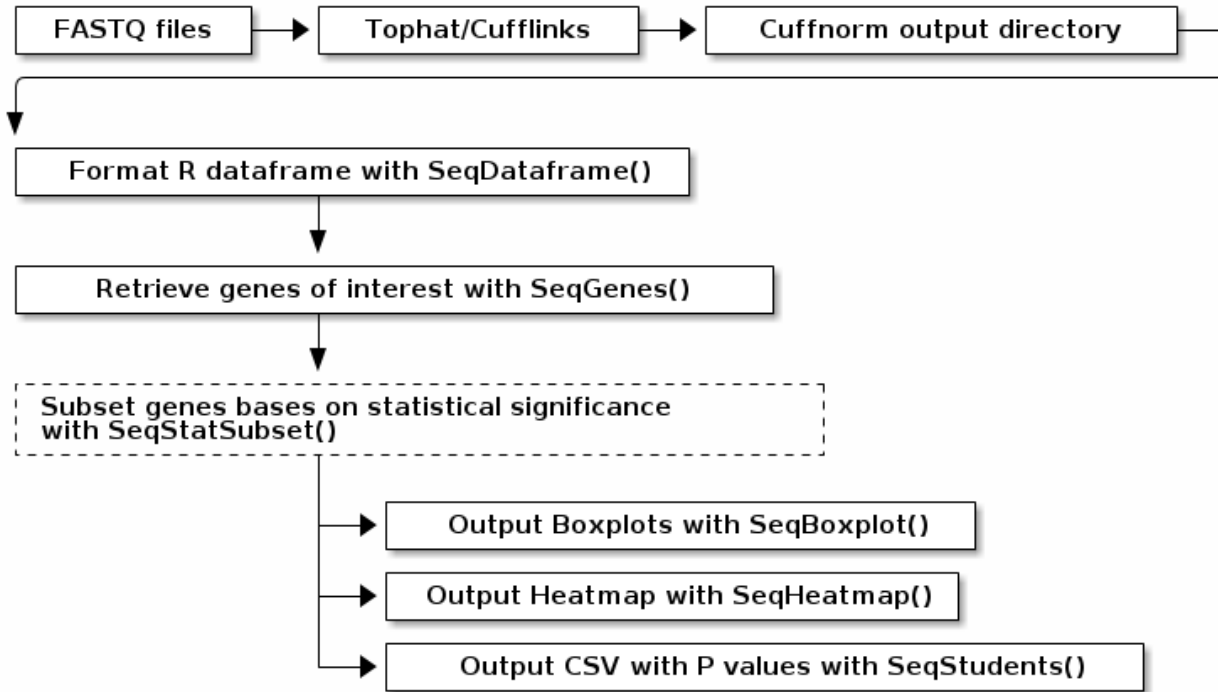
## 2 Design criteria

All functions in the SeqRetriever package are designed with the following criteria in mind:

1. Approachable - Anyone can generate meaningful, high quality data from RNA-seq datasets.
2. Productive - Functions should be simple and intuitive. You shouldn't have to pull up `?SeqDataframe` in every R session. SeqRetriever functions should free your mind to think about science, not programming.

3. Readable - Good R script is easy to follow and, therefore, easier to share and debug. While discrete functions in SeqRetriever can be sequenced together in a pipeline to perform complex tasks, the individual SeqRetriever functions perform simple tasks and can be used independently of other SeqRetriever functions.

### 3 Work flow



## 4 Examples

### 4.1 Setup database

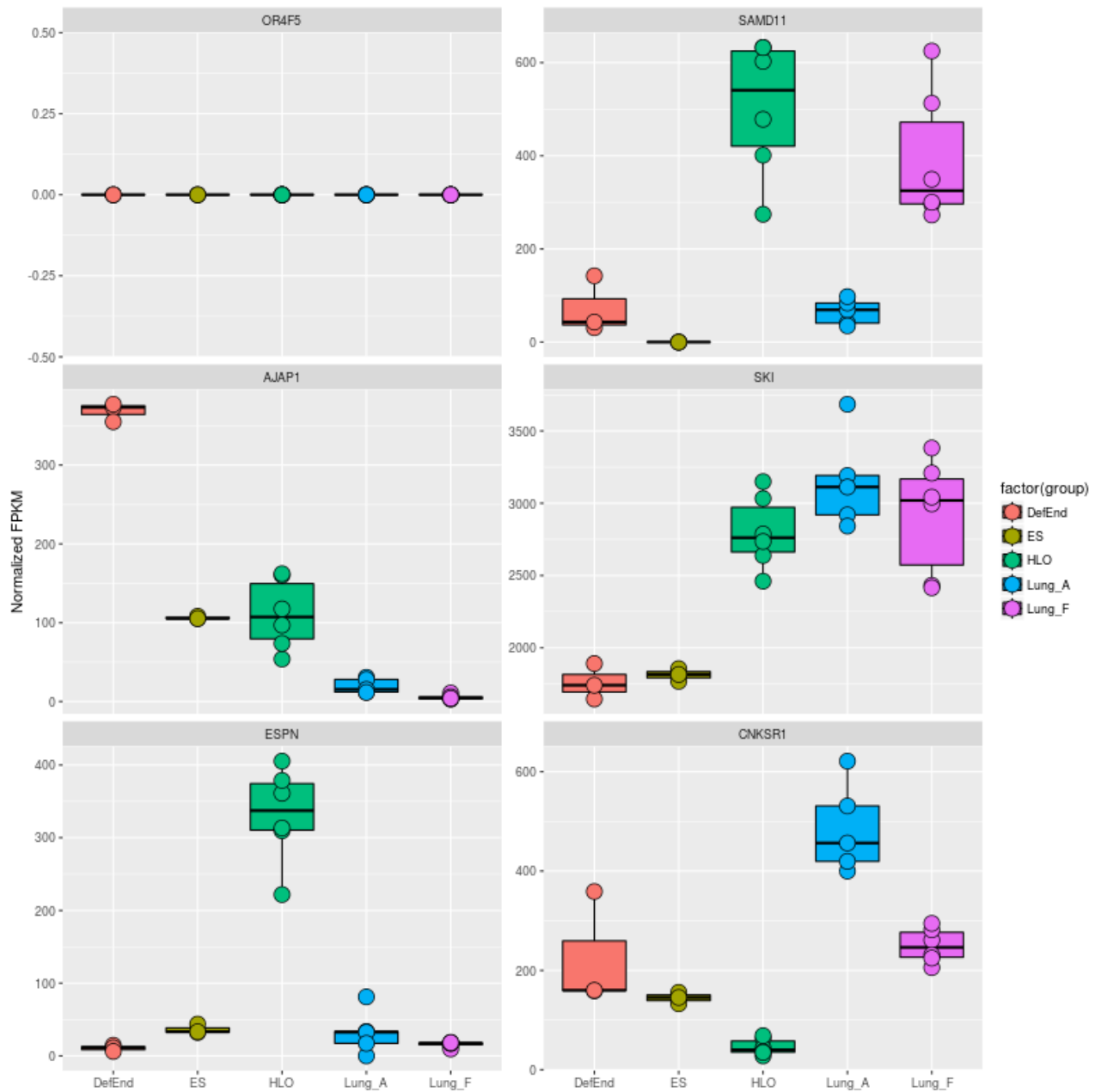
```
library(SeqRetriever)
getSRexample() # Downloads and unpacks example dataset in working directory
testdf <- SeqDataframe(dir = "./norm_out") # format dataframe
```

### 4.2 Select genes

```
genes <- SeqGenes(gene.names = c("OR4F5", "SAMD11", "AJAP1", "SKI", "ESPN", "CNKSR1"), df = testdf)
```

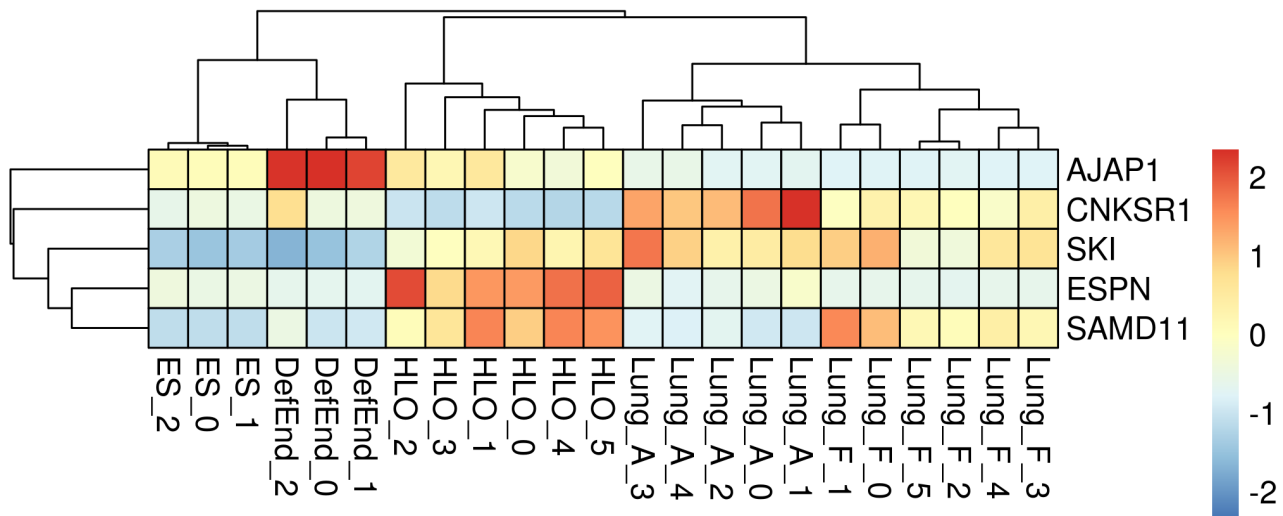
### 4.3 Print boxplot

```
plot <- SeqBoxplot(genes)
print(plot)
```



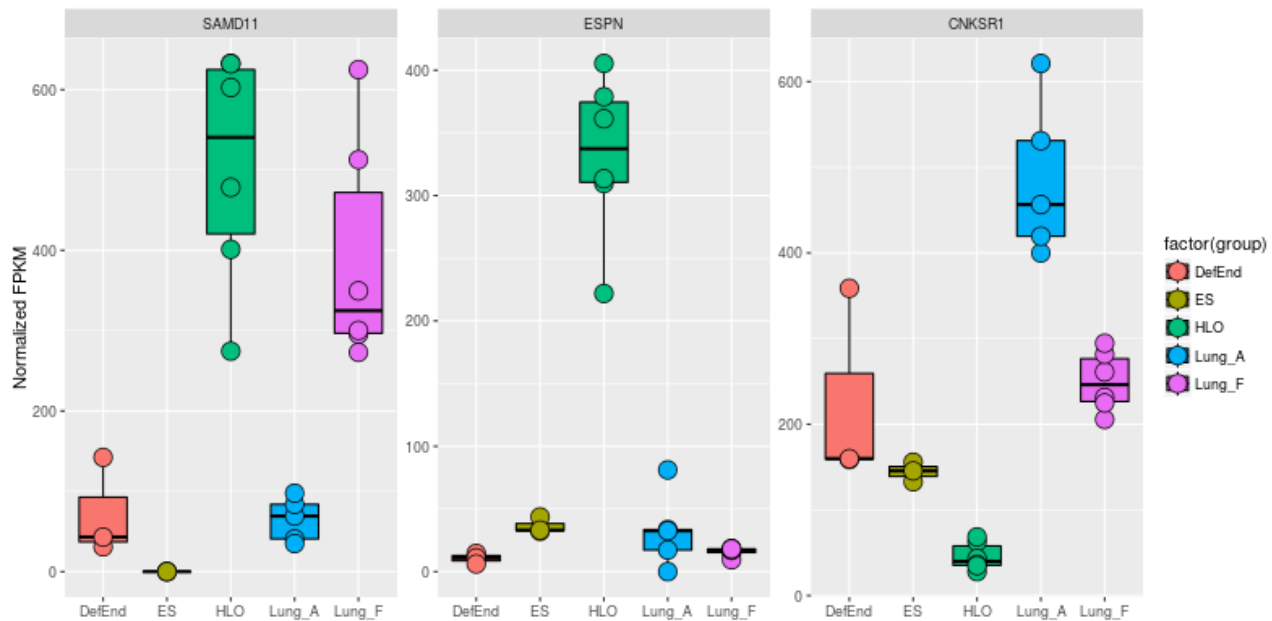
#### 4.4 Print Heatmap

```
SeqHeatmap(genes, hm.name = "heatmap.png")
```



#### 4.5 Print boxplot showing only genes that differ significantly between "HLO" and "Lung<sub>A</sub>"

```
sig.genes <- SeqStatSubset(genes, limit = 0.001, group1 = "HLO", group2 = "Lung_A")
plot2 <- SeqBoxplot(sig.genes, nrow = 1)
print(plot2)
```



## 5 Install from Github

```
library(devtools)
devtools::install_github("hilldr/SeqRetriever/SeqRetriever")
```