A Guide to Writing R Packages

The Statomics Lab
statomics.com

By: Kyle Leckett
Sponsored by: Dr. David Bickel

UROP Winter 2012

Introduction

In the growing field of genomics there is a growing need for more accurate and efficient methods of analyzing this data. Statisticians working with genomics data use a programming language called R to develop models that can effectively analyze this data and draw more accurate conclusions. These models can be very complex and tedious to decipher, so every statistical model must be accompanied by a documentation package that will help researchers worldwide use the model for their research. Creating R packages requires the use of R software, a programming editor, and a Linux server that will produce a final product of a universally accepted R package.

Goals

The goal of this project was to compile a user-friendly documentation package for Dr. Bickel's statistical model of the LFDR and upload it to CRAN. A second goal was to provide a chart that can easily display the drastically different results obtained from using different methods to calculate the LFDR

Methods with a sample

Since many functions were documented, we will just provide a sample. Starting with the R model LFDR-MLE.R, there is a function within called get_lfdr:

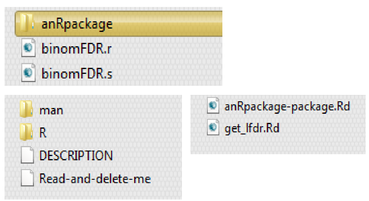
```
get_lfdr <- function(pval,qFUN,W,p0,d_alt,dFUN=dchisq,df=1){
 if(missing(W) && is.function(qFUN)){
  W <- qFUN(pval, df = df,lower.tail=FALSE)
 }
 assert.is(W, "numeric")
 if(p0 %in% 0:1) return(rep(p0, length(W)))
 log_odds <- log(p0) - log(1 - p0) + dFUN(W,df=df,ncp=0, log = TRUE) - dFUN(W,df=df,ncp=d_alt, log = TRUE)
 LFDR.hat <- exp(log_odds)/(1 + exp(log_odds))
 LFDR.hat[is.infinite(log_odds)] <- ifelse(sign(log_odds[is.infinite(log_odds)]) < 0, 0, 1)
 LFDR.hat
```

Before we make the skeleton file, we must load (source) LFDR-MLE.R into R with the command `source("LFDR-MLE.R")`. A long list of functions is produced, but a small sample can be seen here:

```
> setwd("C:/Users/Kyle/Documents/uOttawa/uOttawa/UROP/KylesWork/LFDRMLE")
> source("LFDR-MLE.R")

Welcome to Bioconductor

  Vignettes contain introductory material. To view, type
  'browseVignettes()'. To cite Bioconductor, see
  'citation("Biobase")' and for packages 'citation("pkgname")'.


Attaching package: 'Biobase'

The following object(s) are masked from 'package:Hmisc':

    combine, contents

Creating a new generic function for "plot" in ".GlobalEnv"
Creating a new generic function for "print" in ".GlobalEnv"
Creating a new generic function for "logb" in ".GlobalEnv"
Creating a new generic function for "stripchart" in ".GlobalEnv"
Creating a new generic function for "lines" in ".GlobalEnv"
```

Next, we run the function that produces the skeleton .Rd file. The resulting folders will be in the current working directory:

```
> package.skeleton(list=c("get_lfdr"))
Creating directories ...
Creating DESCRIPTION ...
Creating Read-and-delete-me ...
Saving functions and data ...
Making help files ...
Done.
```



'get_lfdr.Rd' can be found in the 'man' folder within the 'anRpackage' folder. From here, the fields will be manually entered. Here is a small sample of the .Rd file:

Example continued

```
\name{get_lfdr}
\alias{get_lfdr}
\alias{dchisq}
\title{
get_lfdr
}
\description{
Calculate the local false discovery rate
}
\usage{
get_lfdr(pval, qFUN, W, p0, d_alt, dFUN = dchisq, df = 1)
}

\arguments{
  \item{Pval}{vector of p values}
  \item{qFUN}{density function used to compute test statistics}
  \item{W}{Test statistic}
  \item{p0}{proportion/probability of non-affected features}
  \item{d_alt}{noncentrality parameter of the alternative hypothesis}
  \item{dFUN}{Denisty function (default: dchisq)}
  \item{df}{degree of freedom (default = 1)}
}
```

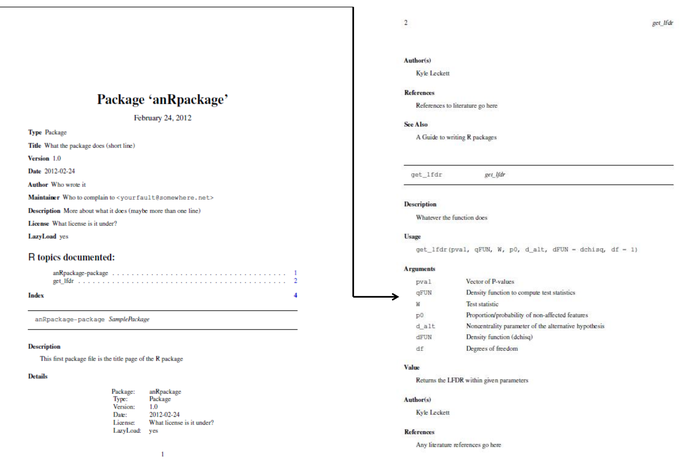
Once finished, the .Rd file must be uploaded to a Linux server (we used FileZilla) and built into a package through an X11 terminal (note that just a sample of the screen is presented):

```
[kleck096@statomic1 ~]$ R CMD build anRpackage
* checking for file 'anRpackage/DESCRIPTION' ... OK
* preparing 'anRpackage':
* checking DESCRIPTION meta-information ... WARNING
Non-standard license specification:
  What license is it under?
Standardizable: FALSE
* removing junk files
* checking for LF line-endings in source and make files
* checking for empty or unneeded directories
* building 'anRpackage_1.0.tar.gz'
```

Completing R's automated checks will produce a final PDF file (only if the package passes all of the tests of course):

```
[kleck096@statomic1 ~]$ R CMD check anRpackage
* checking for working pdflatex ... OK
* using log directory '/mnt/raid/home/kleck096/anRpackage.Rcheck'
* using R version 2.10.0 (2009-10-26)
* using session charset: UTF-8
```

Once all of the checks have passed, the final product will be a PDF manual that is ready to be uploaded to CRAN. Note that the manuals produced by this project were approximately 25 pages in length, however, this sample package only contains two pages. The first .Rd file is a title page that precedes the remainder of the manual. Our sample function 'get_lfdr' is presented:



Package 'anRpackage'

February 24, 2012

Type Package
Title What the package does (short line)
Version 1.0
Date 2012-02-24
Author Who wrote it
Maintainer Who to complain to <yourfault@somewhere.net>
Description More about what it does (maybe more than one line)
License What license is it under?
LazyLoad yes

R topics documented:

anRpackage-package 1
get_lfdr 2
Index 4

anRpackage-package *SampleRPackage*

Description
This first package file is the title page of the R package

Details

Package:	anRpackage
Type:	Package
Version:	1.0
Date:	2012-02-24
License:	What license is it under?
LazyLoad:	yes

1

2 get_lfdr

Author(s)
Kyle Leckett

References
References to literature go here

See Also
A Guide to writing R packages

get_lfdr get_lfdr

Description
Whatever the function does

Usage
get_lfdr(pval, qFUN, W, p0, d_alt, dFUN = dchisq, df = 1)

Arguments

pval	Vector of P-values
qFUN	Density function to compute test statistics
W	Test statistic
p0	Proportion/probability of non-affected features
d_alt	Noncentrality parameter of the alternative hypothesis
dFUN	Density function (dchisq)
df	Degrees of freedom

Value
Returns the LFDR within given parameters

Author(s)
Kyle Leckett

References
Any literature references go here

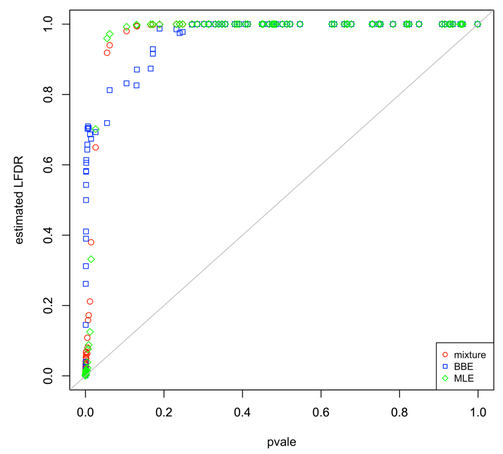
Note how the 'argument's code reflects the automatic formatting of the R package. The subheadings 'usage' and 'description' can also be easily seen in this sample. This sample only documented the function 'get_lfdr' to provide a short sample. A real R package may have more than 25 pages.

Simulated data

The underlying purpose behind Dr. Bickel's model is to show that using different methods to calculate the local false discovery rate (LFDR) may yield drastically different results; so different that one method may lead to rejecting a hypothesis and another method may lead to accepting a hypothesis. A short definition of the LFDR is the probability of falsely identifying the presence a gene. With this worked example, we will explore how estimating the LFDR with three different methods, namely mixture, maximum likelihood estimation, and binomial based estimation. The following code provides an analysis methods:

```
setwd("C:/Users/Kyle/Documents/uOttawa/uOttawa/UROP/KylesWork/LFDRMLE")
source("PsiHat.R")

nfeature<-5;
x.size<-5
x.mean1 <- rep(5,1)
x <- r.xprnSetObject(alpha =0, nfeature = nfeature,  x.size = x.size,
y.size = x.size, x.mean1 = x.mean1, y.mean1 = 0,J=nfeature)
pv<-stat(x, FUN = function(x, y,na.rm=na.rm){fct.pvalue(x = x, y = y,...)$p.value})
#OR:
x1<-exprs(x@x);x2<-exprs(x@y)
pv<-sapply(1:nrow(x1),FUN=function(i){t.test(x=x1[i,],y=x2[i,],
alternative = "two.sided", var.equal = F,paired=F)$p.value})
#------
z.bbe<-bbe(pv)
z.bbe1<-bbe1(pv)
z.rval<-rval(pv)
z.rval1<-rval1(pv)

post<-posteriorP0(x = x, lower.ncp = 1/1e3, lower.P0 = 0)

lfdr.bbe<-z.bbe$lfdr
lfdr.bbe1<-z.bbe1$lfdr
lfdr.rval<-z.rval$lfdr
lfdr.post<-as(post,'numeric')
parameters.post<- unknownParam(post)
parameters.bbe<-z.bbe$estim.p0
parameters.bbe1<-z.bbe1$estim.p0
parameters.rval<-z.rval$estim.p0
parameters.rval1<-z.rval1$estim.p0
```

This code reflects internal functions created by Dr. Bickel used to calculate the LFDR with by the three methods mentioned above.



This plot displays the estimated LFDR against the p-value with the three different estimation methods. P-values determine whether one rejects or does not reject a null hypothesis based on a significance level. Though a trend appears to be followed, there are significant differences between p-values of 0.1 and 0.2.

Acknowledgements

This project was completed by Kyle Leckett. Marta Padilla, Ph.D provided help with developing a systematic approach to making R Packages, examples of the LFDR model, translating and explaining the code, refining the completed R packages, and creating simulated data and associated graphics. Dr. David Bickel sponsored this project for the Undergraduate Research Opportunity Program of winter 2012.