False Discovery Rate (FDR) Estimation with qvalue

EICC : Jessica Randall

The original paper from Storey and Tibshirani introducing the concepts implemented in this package is available here

Briefly, qvalue takes in p-values and adjusts them to account for the multiple hypothesis tests. If you are new to multiple hypothesis testing, check out this article from Noble (2009) for a succint overview of the concept.

Multiple hypothesis testing is inherent to RNA-seq differential expression (DE) analysis and we will use this example to illustrate how we use qvalue to provide actionable results to our clients.

Accurate interpretation of unadjusted p-values assumes that each gene is assessed for DE on its own. However, most RNA-seq experiments assess more than one gene at a time. In order to account for the number of genes we are testing, we must calculate and interpret the adjusted p-value for each gene.

In DE analysis, a single p-value tells you how likely it is that a single gene is differentially expressed between at least two groups (ex: a control and a treatment group) due to some actual difference between the groups as opposed to random chance.

Definition of terms

0.0.0.1 False Discovery Rate (FDR) This tells you how likely it is that all genes identified as DE are false positives. A FDR of 5% means that among all genes called DE, an average of 5% of those are truly not DE. The q-value is the significance threshold adjusted for the fact that we have assessed muliple genes.

0.0.0.1.1 Quasi-likeliehood F Test: Unadjusted p values vs adjusted p values/(FDR): In DE analysis, a single p-value tells you how likely it is that a single gene is differentially expressed between at least two groups (ex: a control and a treatment group) due to some actual difference between the groups as opposed to random chance. False Discovery Rate (FDR) tells you how likely it is that all genes identified as DE are false positives. A FDR of 5% means that among all genes called DE, an average of 5% of those are truly not DE.DE genes are only considered significantly so if they meet the adjusted p value, not only the unadjusted p value.

Loading R packages

Our very first step is to load the qvalue package rom Bioconductor. Please see Bioconductor for information about installation and use of Bioconductor and its packages.

```
require("pacman")
p_load("readr", "here", "dplyr", "ggplot2", "qvalue")


files <- list(
  results = here("Miscellaneous/data/results.txt")
)
```

Loading p-values

qvalue gives us a few options for calculating q-values but to keep our example simplified we import a table of results from a previous differential expression analysis and extract the nominal pvalues as a vector called pvalues.

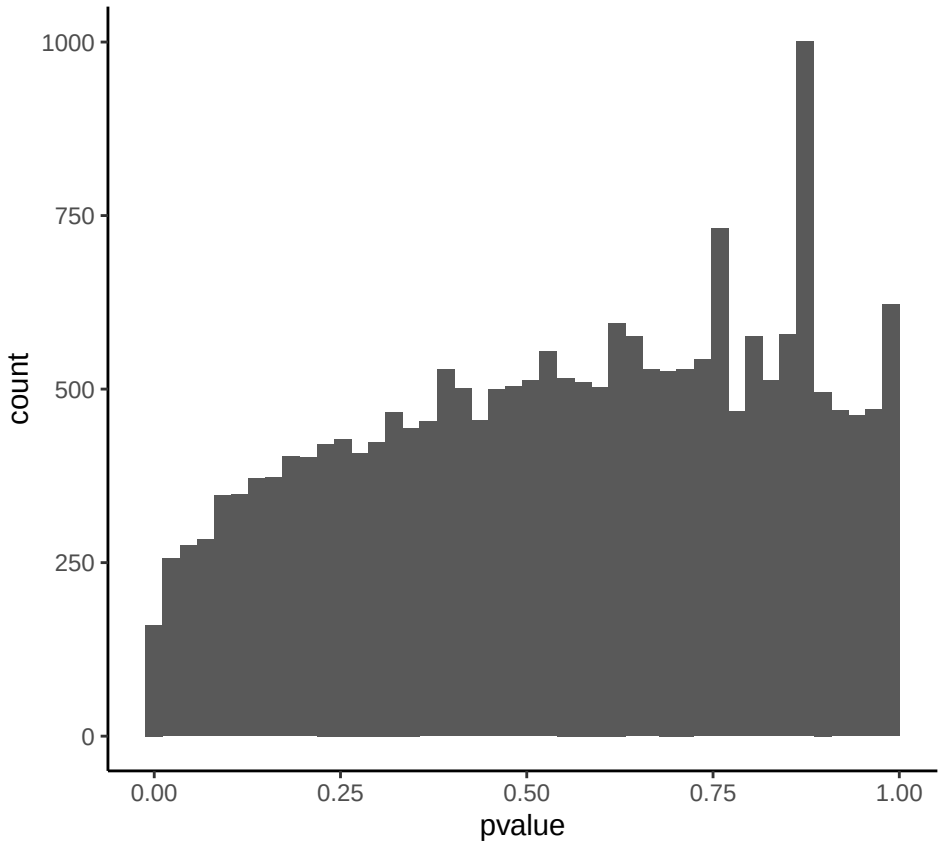
```
results <- read_tsv(files$results)

head(results)
```

```
## # A tibble: 6 x 3
##   genes     pvalue      padj
##   <chr>      <dbl>     <dbl>
## 1 Pcyt1a 3.99e-109 8.39e-105
## 2 Senp5  6.40e-108 6.73e-104
## 3 Bdh1   2.18e- 96 1.53e- 92
## 4 Fbxo45 1.34e- 92 7.02e- 89
## 5 Rnf168 5.16e- 88 2.17e- 84
## 6 Pak2   1.62e- 76 5.67e- 73
```

Checking assumptions

Next, we check out the histogram of the p-values. This allows us to examine the distribution of the p-values and check the package assumption that they should be relatively uniform. This means they will look like one big slope with a high frequency of values of 0.0 (left side) and lower and lower values towards the right of the graph. If this assumption is not met and your p-values look like a U shape, sine curve, or any other funky thing, this package will provide you with faulty analysis if you do not do something to make that assumption hold.

Please contact EICC and depending on why your p-values are looking unusual we may be able to fix it for you.



In our example, the p-values are approximately uniform.

Creating the qvalue object

Next, we create the qvalue object. qvalue gives us many options to do this. These all matter and all vary by project as to which are appropriate, the default options are not going to work for everyone, if anyone.

For this example we have done our literature review and decided (prior to running our sequencing) that we are comfortable with FDR of 0.1 (10%). Recall that this means that of all genes called DE, an average of 10% of those are false positives. In our case this would mean if we have 100 genes that are called DE, we are comfortable with up to 10 of them being false positives. This is the assumption that a popular DE analysis package, edgeR, makes and the accuracy of the results you get from edgeR depend on these assumptions.

Controlling vs. Estimating FDR

In controlling FDR, you set a FDR you are comfortable with a priori, you run your sequencing, you do your pairwise comparisons, and you see which genes meet that pre-specified FDR threshold. In estimating FDR, you run your sequencing, do your pairwise comparisons tests, and afterwards, you specify how comfortable you are with the possibility that a gene or group of genes which have been called DE is/are false positives.

Estimating FDR means that you run the risk of missing DE genes you may be interested in investivating. Controlling FDR means that you may have more genes identifed as DE than you can make sense of and many of them may be false positives. The choice you make depends on whether your experiment is exploratory or looking to assess one or a group of genes in particular. EICC would be happy to work with you to design your experiment to meet your goals.

Summarizing

Below is a summary of some of the information available to us in the qvalue object.

```
summary(qobj)
```

```
##
## Call:
## qvalue(p = pvalues, fdr.level = 0.1, lambda = 0, pi0.method = "smoother")
##
## pi0: 1
##
## Cumulative number of significant calls:
##
##           <1e-04 <0.001 <0.01 <0.025 <0.05 <0.1    <1
## p-value       25     39   147    317   592 1261 21028
## q-value       18     18    19     21    24   26 21028
## local FDR     18     18    18     18    18   18    26
```

From this summary we see that at our pre-specified FDR of 0.1 we determine that all genes with q-values less than or equal to 0.1 are significant. This gives us 7874 significant genes with the understanding that up to 787 DE genes are false positives.

Controlling FDR gave us a large number of genes to investigate further. We decided that we would risk up to 10% of these being false positives if we could find as many DE genes as possible. We do not know how likely it is that any one of the genes we found is a false positive but we know it could be any or all 787 of them.

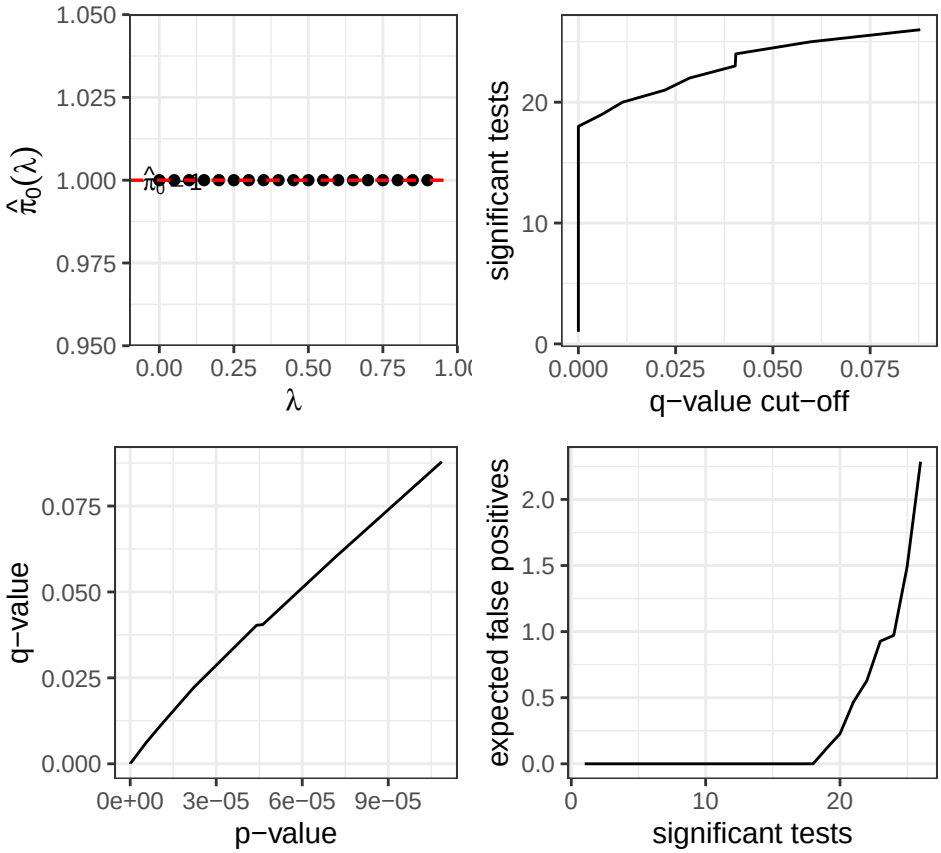
If we had decided at the beginning to estimate FDR, we could have used the q-value of <0.001 and had a list of 911 genes to follow up on along with an estimate of how likely it is that each gene is a false positive, and comfortable with up to 9 of them being false positives.

This choice of controlling or estimating FDR is one EICC is happy to help you make to get the most actionable results from your projects.

Visualizing

qvalue also offers a number of visualization options to assess your findings. Here is one example of the types of plots you can generate.

Overall, these graphs show us how closely our estimate of the null hypothesis (that no genes are DE) resembles reality, how many genes are DE, and how many false positives we can expect from our q-value cut-off.



For a more specific application and interpretation of these or additional tools and visualizations for your own data, please find our contact information on our website

Session information and References

```
## [1] "Fri Jan 31 13:29:24 2020"

## R version 3.6.2 (2019-12-12)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 18363)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.1252
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] qvalue_2.18.0 ggplot2_3.2.1 dplyr_0.8.3   here_0.1      readr_1.3.1
```

```
## [6] pacman_0.5.1
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.3       plyr_1.8.5       pillar_1.4.3     compiler_3.6.2
##  [5] tools_3.6.2      digest_0.6.23    evaluate_0.14    tibble_2.1.3
##  [9] lifecycle_0.1.0  gtable_0.3.0     pkgconfig_2.0.3  png_0.1-7
## [13] rlang_0.4.4      cli_2.0.1        yaml_2.2.0       xfun_0.12
## [17] withr_2.1.2      stringr_1.4.0    knitr_1.27       vctrs_0.2.2
## [21] hms_0.5.3        rprojroot_1.3-2  grid_3.6.2       tidyselect_1.0.0
## [25] glue_1.3.1       R6_2.4.1         fansi_0.4.1      rmarkdown_2.1
## [29] bookdown_0.17    farver_2.0.3     reshape2_1.4.3   purrr_0.3.3
## [33] magrittr_1.5     splines_3.6.2    backports_1.1.5  scales_1.1.0
## [37] htmltools_0.4.0  assertthat_0.2.1 colorspace_1.4-1 labeling_0.3
## [41] utf8_1.1.4       stringi_1.4.5    lazyeval_0.2.2   munsell_0.5.0
## [45] crayon_1.3.4

##
## To cite the 'bookdown' package in publications use:
##
##   Yihui Xie (2020). bookdown: Authoring Books and Technical Documents
##   with R Markdown. R package version 0.17.
##
##   Yihui Xie (2016). bookdown: Authoring Books and Technical Documents
##   with R Markdown. Chapman and Hall/CRC. ISBN 978-1138700109
##
## To see these entries in BibTeX format, use 'print(<citation>,
## bibtex=TRUE)', 'toBibtex(.)', or set
## 'options(citation.bibtex.max=999)'.

##
## To cite ggplot2 in publications, please use:
##
##   H. Wickham. ggplot2: Elegant Graphics for Data Analysis.
##   Springer-Verlag New York, 2016.
##
## A BibTeX entry for LaTeX users is
##
##   @Book{,
##     author = {Hadley Wickham},
##     title = {ggplot2: Elegant Graphics for Data Analysis},
##     publisher = {Springer-Verlag New York},
##     year = {2016},
##     isbn = {978-3-319-24277-4},
##     url = {https://ggplot2.tidyverse.org},
##   }

##
## To cite package 'readr' in publications use:
##
##   Hadley Wickham, Jim Hester and Romain Francois (2018). readr: Read
##   Rectangular Text Data. R package version 1.3.1.
##   https://CRAN.R-project.org/package=readr
##
```

```
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {readr: Read Rectangular Text Data},
##     author = {Hadley Wickham and Jim Hester and Romain Francois},
##     year = {2018},
##     note = {R package version 1.3.1},
##     url = {https://CRAN.R-project.org/package=readr},
##   }

##
## To cite package 'dplyr' in publications use:
##
##   Hadley Wickham, Romain François, Lionel Henry and Kirill Müller
##   (2019). dplyr: A Grammar of Data Manipulation. R package version
##   0.8.3. https://CRAN.R-project.org/package=dplyr
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {dplyr: A Grammar of Data Manipulation},
##     author = {Hadley Wickham and Romain François and Lionel Henry and Kirill Müller},
##     year = {2019},
##     note = {R package version 0.8.3},
##     url = {https://CRAN.R-project.org/package=dplyr},
##   }

##
## To cite package 'qvalue' in publications use:
##
##   John D. Storey, Andrew J. Bass, Alan Dabney and David Robinson
##   (2019). qvalue: Q-value estimation for false discovery rate control.
##   R package version 2.18.0. http://github.com/jdstorey/qvalue
##
## A BibTeX entry for LaTeX users is
##
##   @Manual{,
##     title = {qvalue: Q-value estimation for false discovery rate control},
##     author = {John D. Storey and Andrew J. Bass and Alan Dabney and David Robinson},
##     year = {2019},
##     note = {R package version 2.18.0},
##     url = {http://github.com/jdstorey/qvalue},
##   }
```