False Discovery Rate (FDR) Estimation with qvalue

EICC : Jessica Randall

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 a crucial aspect of RNA-seq differential expression (DE) analysis. This is why you may see results reported in terms of FDR or adjusted p-vlaues rather than nominal p-values. Each gene represents an individual test of a null hypothesis with its own individual likelihood of producing a false positive result, which is usually around 20%. Multiply that by 24000 genes and the likelihood is extremely high that many of the genes found to statistically significnatly differentially expressed at a nominal p-value of 0.05 would be false positive findings which would waste the time and resources of anyone attempting to run down any results based on the p-value with a biological test.

You may have heard of the Bonferroni correction for multiple hypothesis testing. While Bonferroni is often useful, it has been found to be too struct with regard to DEG analysis. Methods like Benjamini-Hochberg correction or the Storey false discovery rate used for DEG analysis account for the library size and the inherent biological variability between each sample while Bonferroni does not. By taking these characteristics into account, these methods balance the number of truly DEG findings identified while limiting the number of false positive DEG findings.

For this example we will be using the Hendenfalk (2001) data. For this data, comparisons were made between between 3,226 genes of two mutation types, BRCA1 (7 arrays) and BRCA2 (8 arrays).

Definition of terms

0.0.0.1 * False Discovery Rate (FDR)

Nominal p values vs adjusted p values/(FDR): In DEG 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. The 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 nominal p value. The local FDR or q-value tells you the likelihood that a given gene is statistically significant due to some true difference between the groups rather than by random chance.

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.

```
pacman::p_load("tidyverse", "qvalue")

data(hedenfalk)
```

Preparing the data

We import the nominal p-values from the Hendenfalk 2001 data and use them to create the package specific qvalue data object.

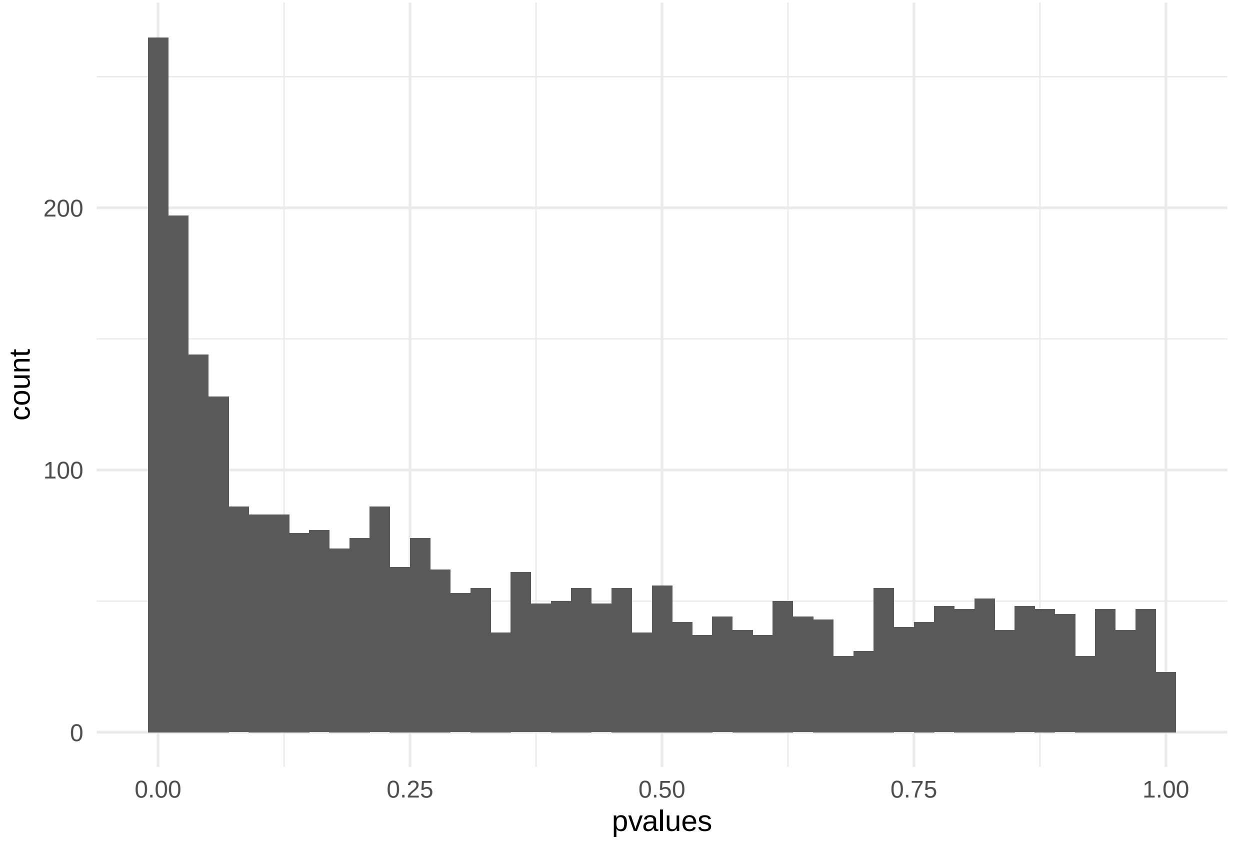
```
pvalues <- hedenfalk$p

qobj <- qvalue(p = pvalues)
```

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 largely flat along the right tail of the histogram. This suggests that the true null p-values do follow the assumed Uniform distribution. If you try this on your own data and do not see this with your p-vales at the appropriate bin width, please feel free to each out to EICC for assistance.

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 which 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.

```r
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       15     76   265    424   605  868 3170
## q-value        0      0     0     20    94  218 3170
## local FDR      0      0     1     18    53  120 1377
```

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 218 significant genes with the understanding that up to 22 DE genes may be 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.

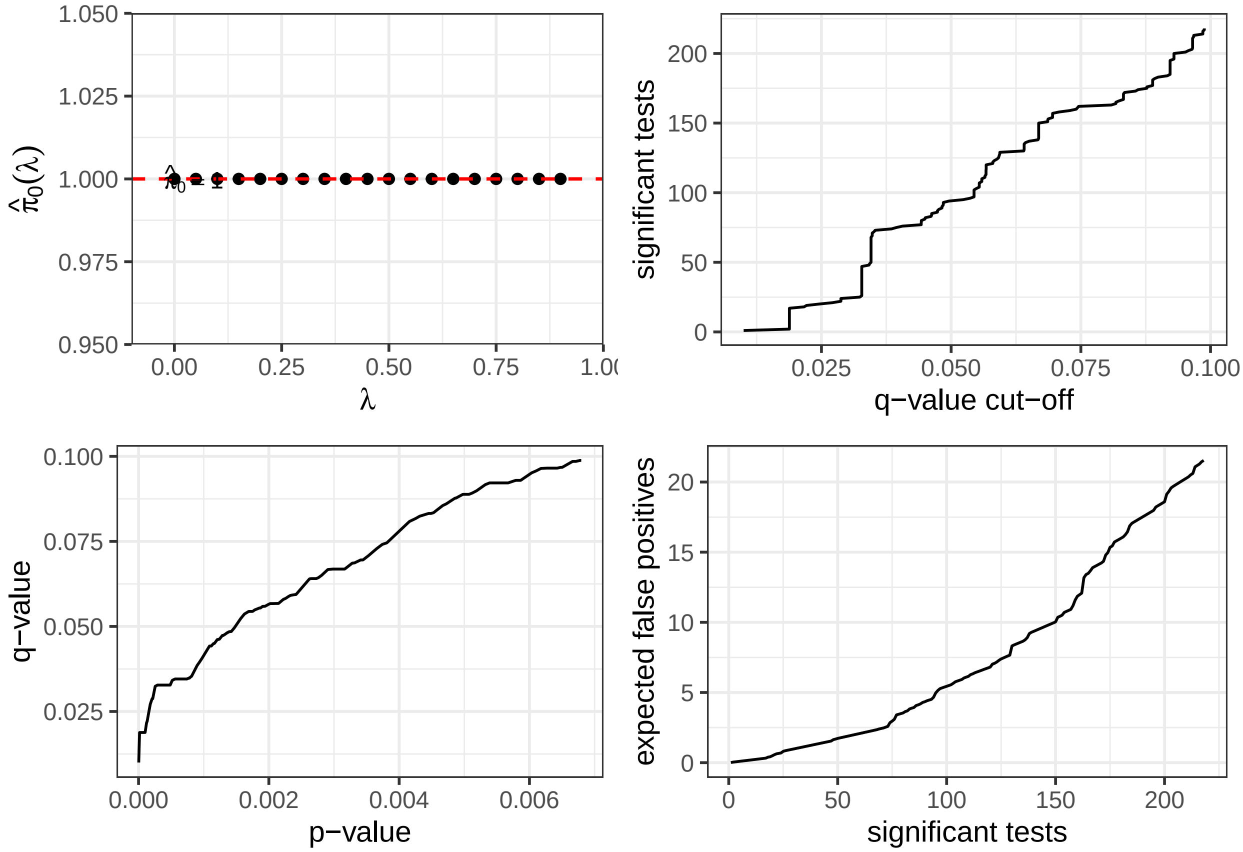
If we had decided at the beginning to estimate FDR, we could have used the q-value of <0.025 and had a list of 20 genes to follow up on along with an estimate of how likely it is that each gene is a false positive, and comfortable with up to 2 of them being false positives.

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] "Thu Mar 12 15:09:16 2020"
```

```
## R version 3.6.3 (2020-02-29)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Debian GNU/Linux 9 (stretch)
##
## Matrix products: default
## BLAS:   /usr/lib/openblas-base/libblas.so.3
## LAPACK: /usr/lib/libopenblasp-r0.2.19.so
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
##  [1] qvalue_2.18.0   forcats_0.5.0   stringr_1.4.0   dplyr_0.8.5
##  [5] purrr_0.3.3     readr_1.3.1     tidyr_1.0.2     tibble_2.1.3
##  [9] ggplot2_3.3.0   tidyverse_1.3.0
##
## loaded via a namespace (and not attached):
##  [1] tidyselect_1.0.0 xfun_0.12        reshape2_1.4.3   splines_3.6.3
##  [5] haven_2.2.0      lattice_0.20-40  colorspace_1.4-1 vctrs_0.2.4
##  [9] generics_0.0.2   htmltools_0.4.0  yaml_2.2.1       rlang_0.4.5
## [13] pillar_1.4.3     glue_1.3.2       withr_2.1.2      DBI_1.1.0
## [17] dbplyr_1.4.2     modelr_0.1.6     readxl_1.3.1     plyr_1.8.6
## [21] lifecycle_0.2.0  munsell_0.5.0    gtable_0.3.0     cellranger_1.1.0
## [25] rvest_0.3.5      evaluate_0.14    labeling_0.3     knitr_1.28
## [29] fansi_0.4.1      broom_0.5.5      Rcpp_1.0.3       scales_1.1.0
## [33] backports_1.1.5  jsonlite_1.6.1   farver_2.0.3     fs_1.3.2
## [37] hms_0.5.3        digest_0.6.25    stringi_1.4.6    bookdown_0.18
## [41] grid_3.6.3       cli_2.0.2        tools_3.6.3      magrittr_1.5
## [45] pacman_0.5.1     crayon_1.3.4     pkgconfig_2.0.3  xml2_1.2.5
## [49] reprex_0.3.0     lubridate_1.7.4  assertthat_0.2.1 rmarkdown_2.1
## [53] httr_1.4.1       rstudioapi_0.11  R6_2.4.1         nlme_3.1-145
## [57] compiler_3.6.3
```

```
##
## To cite the 'bookdown' package in publications use:
##
##   Yihui Xie (2020). bookdown: Authoring Books and Technical Documents
##   with R Markdown. R package version 0.18.
##
##   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)'.
```

```
##
##   Wickham et al., (2019). Welcome to the tidyverse. Journal of Open
##   Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
##
## A BibTeX entry for LaTeX users is
##
##   @Article{,
##     title = {Welcome to the {tidyverse}},
##     author = {Hadley Wickham and Mara Averick and Jennifer Bryan and Winston Chang and Lucy D'Agostin
##     year = {2019},
##     journal = {Journal of Open Source Software},
##     volume = {4},
##     number = {43},
##     pages = {1686},
##     doi = {10.21105/joss.01686},
##   }


##
## 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},
##   }
```