covRNA - Tutorial

Lara Urban

This tutorial exemplifies how the covRNA package uses fourthcorner analysis and RLQ to discover covariate associations in transcriptomic data. Here, an RNA-Seq dataset of *Bacillus anthracis* containing different stress conditions as sample covariates and COG annotations (Clusters of Orthologous Groups) as gene covariates will be analysed. Further, it will be shown how gene covariates can be assigned to the dataset by using other R packages.

1. Overview of the Analysis

Transcriptomic data normally comes with covariates of the samples and of the genes. To analyse associations between sample covariates and gene covariates, the fourthcorner analysis tests the statistical significance of the associations by permutation tests (Legendre *et al.*, 1997) and the RLQ visualizes associations within and between the covariates (Doledec *et al.*, 1996).

This package contains fast and user-friendly alternatives to the functions fourthcorner and rlq of the ade4 package (Dray *et al.*, 2007, and Dray *et al.*, 2014) for the analysis of large-scale transcriptomic data. The functions stat and ord of this package can be used for fourthcorner analys and rlq, respectively, and hereby 1. significantly reduce runtime and storage space; 2. account for transcriptome-specific shapes of the empirical permutation distributions (according to Chihara and Hesterberg, 2011); 3. avoid redundancy and 4. render the analysis more user-friendly by supplying automatation, direct modification of the plots and unsupervised filtering of the genes.

Please refer to the manpages for details about the functions. The package covRNA is implemented to be easily combinable with other packages and objects of the Bioconductor project (Gentleman *et al.*, 2004).

input An ExpressionSet object of the package Biobase can be used as input. Then, the ExpressionSet has to contain transcriptomic data in its argument assayData, the sample covariates in phenoData and the gene covariates in featureData.

Alternatively to an ExpressionSet, the three data.frames R, L and Q can be used as input. Here, data.frame L contains transcriptomic data, Q the sample covariates and R the gene functions.

stat The function stat takes each combination between one sample covariate and one gene covariate and calculates a statistic. If at least one the covariates is quantitative, a correlation coeffcient is calculated. If both covariates are categorical, a Chi-Square test (Fisher, 1922) related statistic is calculated. Significance of the associations is assessed by permutation tests. By default, multiple testing correction according to Benjamini and Hochberg (1995) is applied. The resulting p-values are plotted as cross-tabulation of the sample covariates and the gene functions; by default, red and blue cells show negative and positive significant associations at α=0.05, respectively.

ord The function ord automatically applies singular matrix ordination to each of the three data.frames R, L and Q. Correspondence Analysis (CA) is applied to L. Principal Component Analysis (PCA) or Hillsmith Analysis (HA) or Multiple Correspondence Analysis (MCA) are applied to R and Q, depending on the type of variables they contain. Then, the rlq function of the ade4 package is applied and the results can be plotted.

Let's install the package and then analyse an RNA-Seq dataset.

```
> source("https://bioconductor.org/biocLite.R")
> biocLite("covRNA")
```

2. Analysis of an RNA-Seq Dataset

Here, an RNA-Seq dataset of *Bacillus anthracis* (`ExpressionSet Baca`) containing different stress conditions as sample covariates and COG annotations (Clusters of Orthologous Groups, Tatusov *et al.*, 2000) as gene covariates will be analysed.

2.1 Preparation of the dataset

We load the `covRNA` package and the integrated `Baca` dataset, which contains the `ExpressionSet Baca`. The `assayData` contains deep sequenced RNA-Seq data of *B. anthracis* under four stress conditions (with four replicates per stress conditions). The raw sequence reads derive from Passalacqua *et al.* (2012) and are availaible at Gene Expression Omnibus (GEO, accession number GSE36506). We have already mapped, counted and `DESeq2` (Love *et al.*, 2014) normalised these counts. The `phenoData` assigns the stress condition, i.e. ctrl, cold, salt and alcohol stress, to the samples. The `featureData` contains COG annotations of the genes.

```
library(covRNA)
data(Baca)
```

2.2 Fourthcorner Analysis with `stat`

We use the function `stat` to statistically analyse associations between gene and sample covariates.

```
statBaca <- stat(ExprSet = Baca, npermut = 999, padjust = "BH", nrcor = 2, exprvar = 1)
```

`statBaca` is then an object of type `stat`. As a list, it saves all results as well as the input of the function. For instance, we can access the adjusted p-values of all covariate combinations.

```
ls(statBaca)
adjp <- statBaca$adj.pvalue; adjp
```

To visualise the results, the `stat` object can be plotted (Figure 1). If the plot shall be shown in high quality, we advise to use the default setting `pdf=TRUE`.

```
plot(statBaca, xnames = c('cold','ctrl','etoh','salt'), shiftx = -0.1)
```

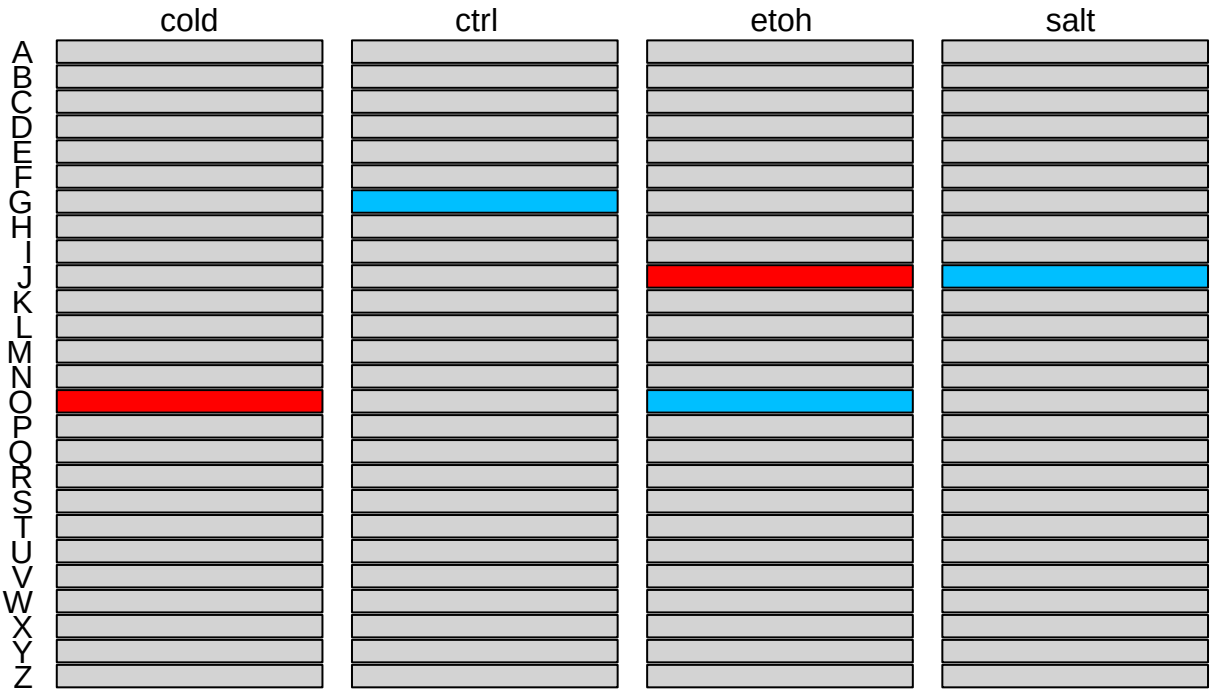

Figure 1: Cross-tabulation of the adjusted p-values between the sample covariates and the gene functions of the `ExpressionSet Baca`. Red and blue cells show negative and positive significant associations at α=0.05, respectively.

The cross-tabulation of the sample covariates and the gene functions visualises negative and positive significant associations at α=0.05. Five significant associations can be discovered.

2.3 `RLQ` with `ord`

We use the function `ord` to visualise sample and gene covariates in one coordinate system.

```
ordBaca <- ord(Baca)
```

`ordBaca` is then an object of type `ord`. Different features of this object can be plotted by using the `feature` argument of the `plot` function (see manpages for more information). For instance, we can plot the amount of variance explained by each axis (Figure 2).

```
plot(ordBaca, feature = "variance")
```

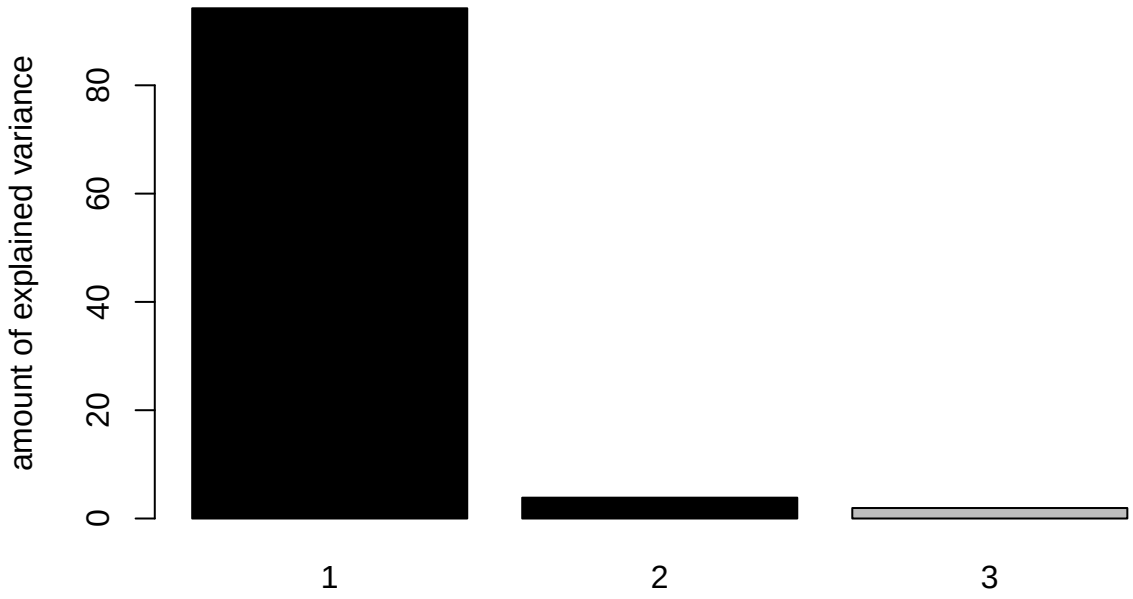

Figure 2: Barplot of the amount of variance explained by each axis of the ordination `ordBaca`. As the number of axes to be taken into account by ordination is by default set to 2, the bars of the first two axes are shown in black.

The first two axes of the RLQ explain a large amount of the variance of the data (93.81% and 4.09%, respectively).

2.4 Combination of Results

The results of the functions `stat` and `ord` can be simultaneously visualised by the function `vis` (Figure 3).

```
vis(Stat = statBaca, Ord = ordBaca, rangex=4, rangey=4)
```

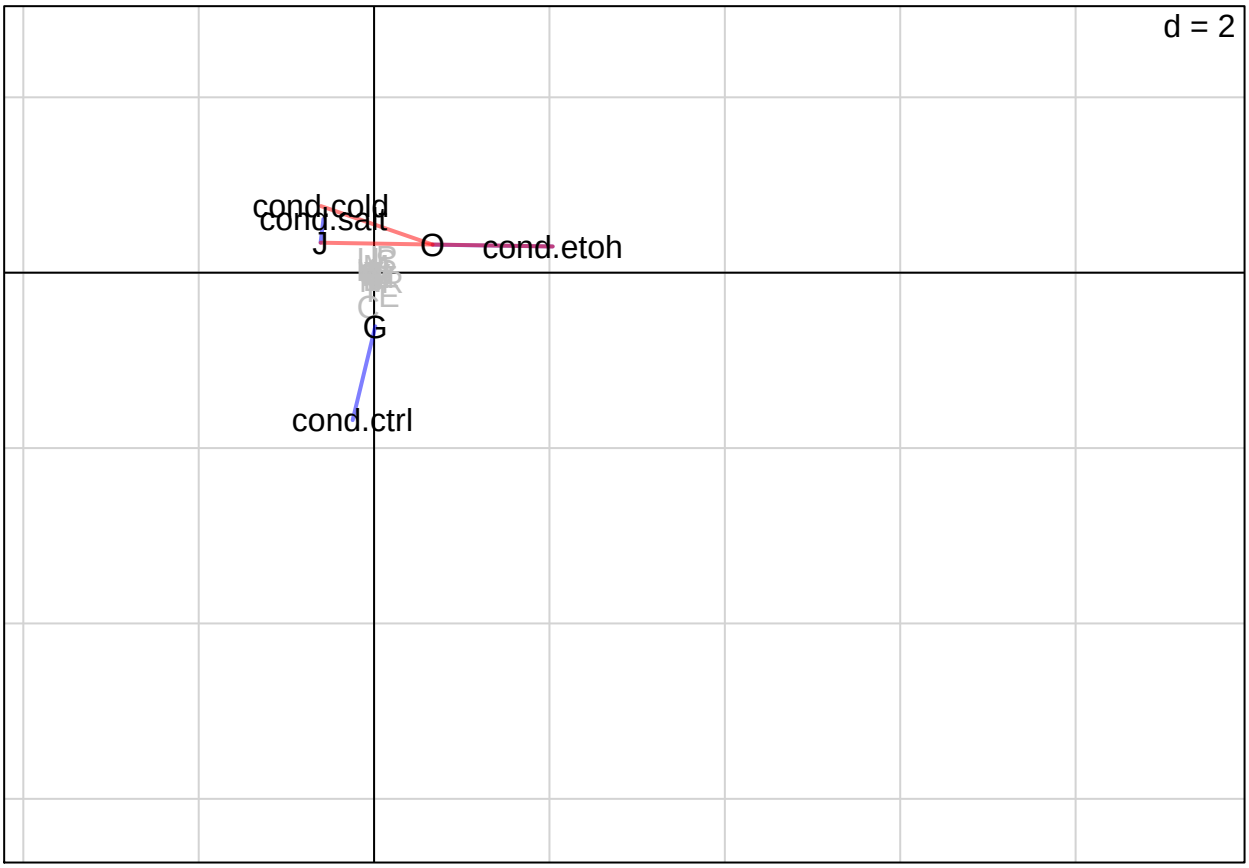

Figure 3: Simultaneous visualisation of the statistical analysis and of the ordination of `Baca`. Blue and red lines between the covariates represent positive and negative significant associations, respectively.

Here, covariates involved in at least one significant association are shown in black, others are shown in gray. All significant covariates are connected by lines which colour represents the character of their association. As expected, positively associated covariates are situated close to each other and at similar angles from the origin. On the contrary, negatively associated covariates are distant from each other.

We can further observe that the first axis seems to be spanned by the difference between the classes J and O. The second axis seems to be spanned between ctrl and the other treatments. Spatial proximity of cold and salt treatment in the second quadrant suggest that they have similar functional effects on the gene expression.

2.5 Comparison with Other Methods

To validate our results of the analysis of `Baca`, we compare them to traditional approaches like hypergeometric test (HG), Mann-Whitney rank test (RANK) und gene set enrichment analysis (GSEA, Subramanian *et al.*, 2005) by using the `R` package `BOG` (see Park *et al.*, 2015, for further details).

RANK and GSEA discover class J as significantly enriched ($p=6.40e^{-11}$ and $p=0.02$, respectively). HG does not detect any significant gene functions.

3. Gene Annotation

If your dataset contains no gene covariates, but you would like to analyse the associations between sample covariates and gene functions, `Bioconductor` offers multiple ways to assign gene covariates to the genes. We propose to use the `Bioconductor` package `biomaRt` (Durinck *et al.*, 2009).

```
> source("https://bioconductor.org/biocLite.R")
> biocLite("biomaRt")
> library(biomaRt)
```

Via the `biomaRt` package, different databases can be accessed. By using `listEnsembl()`, for example, all available ENSEMBL databases can be listed (Hubbard *et al.*, 2002). After choosing a database, a dataset can be selected. This dataset will contain different gene functions and other information about genes which can be accessed by `listAttributes()`.

```
> ensembl <- useEnsembl(biomart = "ensembl")
> listDatasets(ensembl)
> ensemblhuman <- useEnsembl(biomart = "ensembl", dataset = "hsapiens_gene_ensembl")
> listAttributes(ensemblhuman)
```

If the gene identifiers do not correspond to each other, the `Bioconductor` package `annotate` can be used to assign identifier to each other.

Like this, we receive a fully annotated dataset which can be analysed by functions of the `covRNA` package.

4. Installation

The `covRNA` package is freely available from `Bioconductor` at http://www.bioconductor.org.

References

Benjamini, Y. and Hochberg, Y. (1995). *Cotrlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the Royal Statistical Society. Series B (Methodological). 289:300.

Chihara, L. M. and Hesterberg, T. C. (2011) *Mathematical Statistics with Resampling and R.* John Wiley & Sons, 35:75.

Doledec, S., Chessel, D., Ter Braak, C., and Champely, S. (1996) *Matching species traits to environmental variables: a new three-table ordination method.* Environmental and Ecological Statistics, 3(2), 143:166.

Dray, S., Choler, P., Doledec, S., Peres-Neto, P. R., Thuiller, W., Pavoine, S., and ter Braak, C. J. (2014) *Combining the fourth-corner and the rlq methods for as-sessing trait responses to environmental variation.* Ecology, 95(1), 14:21.

Dray, S., Dufour, A.-B., et al. (2007) *The ade4 package: implementing the duality diagram for ecologists.* Journal of statistical software, 22(4), 1:20.

Durinck, S., Spellman, P., Birney, E., Huber, W. (2009) *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt.* Nature Protocols, 4, 1184:1191.

Fisher, R. A. (1922) *On the interpretation of Chi2 from contingency tables, and the calculation of P.* Journal of the Royal Statistical Society, 85 (1), 87:94.

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., Gautier, L., Ge, Y., Gentry, J., et al. (2004) *Bioconductor: open soft-ware development for computational biology and bioinformatics.* Genome Biology, 5(10), R80.

Hubbard, T., Barker, D., Birney, E., Cameron, G., Chen, Y. et al. (2002) *The Ensembl genome database project.* Nucleic Acids Research 2002 30(1), 38:41.

Legendre, P., Galzin, R., and Harmelin-Vivien, M. L. (1997) *Relating behavior to habitat: solutions to the fourth-corner problem.* Ecology, 78(2), 547:562.

Love, M. I., Huber, W., Anders, S. (2014) *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2.* Genome Biology, 15(12), 550.

Park, J., Taslim, C., Lin, S. (2015) *BOG: R-package for Bacterium and virus analysis of Orthologous Groups.* Elsevier Computational and Structural Biotechnology Journal, 13, 366:369.

Passalacqua, K. D., Varadarajan, A., Weist, C., Ondov, B. D., Byrd, B. et al. (2012) *Strand-Specific RNA-Seq Reveals Ordered Patterns of Sense and Antisense Transcription in Bacillus anthracis.* PLoS ONE, 7(8), e43350.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S., et al. (2005) *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proceedings of the National Academy of Sciences of the United States of America, 102(43), 15545:15550.

Tatusov, R. L., Galperin, M. Y., Natale, D. A., Koonin, E. V. (2000) *The COG database: a tool for genome-scale analysis of protein functions and evolution.* Nucleic Acids Research, 28(1), 33:36.