Rasp: R package for alternative splicing differential analysis

Mikel Esnaola[1] `mesnaola@creal.cat`

Juan Ramón González[1,2] `jrgonzalez@creal.cat`

April 25, 2017

[1]Centre for Research in Environmental Epidemiology (CREAL), Barcelona, Spain
`http://www.creal.cat/jrgonzalez/software.htm`
[2]Hospital del Mar Research Institute (IMIM), Barcelona, Spain

1 Getting started

This document gives an overview of the R package rasp, which provides statistical procedures to study differential splicing among conditions in the context of annotated genomes. That is, isoform information is summarized using the available annotation (at exon or transcript level) using any of the existing tools. This information can provided in RPKM as for instance FluxCapacitor (`http://big.crg.cat/services/flux_capacitor`) does, or using read counts on exons. The statistical methods used in rasp are described in Gonzalez-Porta et al. (2012). The method is based on comparing the variability of the relative expression among conditions using the methodology described in Anderson (2006) that can be seen as an ANOVA for the case of having multivariate data (multiple isoforms or exons per gene). The current vesion of the package also allows the user to compare splicing patterns for more than two conditions

We should start the R session by loading the library as follows:

```
> library(rasp)
```

We will start loading into the workspace the data corresponding to the initial table of transcipt expression estimates provided by Gonzalez-Porta et al. (2012). The RNA-seq data correspond to lymphoblastoid cell lines derived from 69 unrelated Nigerian individuals produced by Pickrell et al. (2010). More than half billion 35-bp-long reads were sequenced. Reads were mapped to the human genome version hg19 using GEM mapper Marco-Sola et al. (2012). After that, mapped reads were used to obtain transcipt expression estimates using GENCODE version 3c produced in the framework of the ENCODE project as the reference genome annotation. Finally, Flux Capacitor was used to produce estimates of transcript abundances measured as RPKMs. Data was filtered to have only information about protein-coding genes expressed in all individuals, having at least two annotated splice forms and isoforms with an expresion level of at least 1 RPKM in at least one individual. These filters lead to a total of 1654 genes and 4668 associated transcripts Gonzalez-Porta et al. (2012). The object containing this information can be loaded by typing:

```
> data(YRI)
> dim(YRI)

[1] 4673   71
```

The object also contains information about gender of each sample that will be used as a grouping factor to illustrate comparison among conditions:

```
> genderYRI <- attr(YRI, "gender")
> head(genderYRI)

[1] male   male   female male   female male
Levels: female male

> table(genderYRI)

genderYRI
female   male
    40     29
```

2 Data Visualization

For each gene n alternative splice forms (transcripts or exons) are observed. Therfore, for a given gene, and for each individual j, the relative expression of each transcipt, i $i = 1,\ldots,n$ is computed as:

$$f_{ij} = \frac{x_{ij}}{\sum_{i=1}^{n} x_{ij}} \tag{1}$$

where x_{ij} denotes the transcript expression estimates (either as RPKMs or counts). Therefore, we consider as a outcome the vector of splicing proportions for the j individual, $f_j = (f_{1j},\ldots,f_{nj})$. This multivariate nature of the data is the reason why Gonzalez-Porta et al. (2012) proposed to use the methodology described in Anderson (2006) to compare groups.

This type of data is also known as compositional data and can be represented in the simplex using the function `plotTernary` as illustrate Figure 2 that can be obtained by typing

```
> plotTernary(YRI, "ENSG00000160741", transcripts=1:3)
```

Notice that the argument `transcript` indicates the three transcripts to be represented in the Ternary plot.

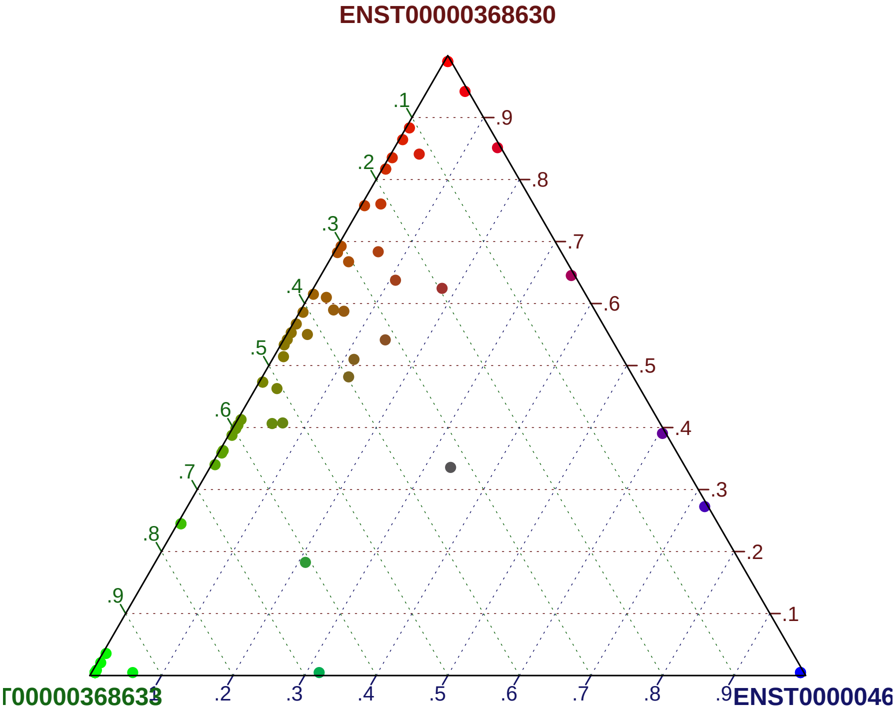

Figure 1. Ternary plot corresponding to the splicing ratios of three transcripts (1st, 2nd and 3rd) belonging to gene ENSG00000160741.

An obvious limitation of the Ternary Plot is that it can oly plot data for three different transcripts. In order to plot the splicing ratios for any number of transcripts one can use the function `plotAllIso`. The result is a plot showing the splicing ratios for all selected transcripts and individuals.

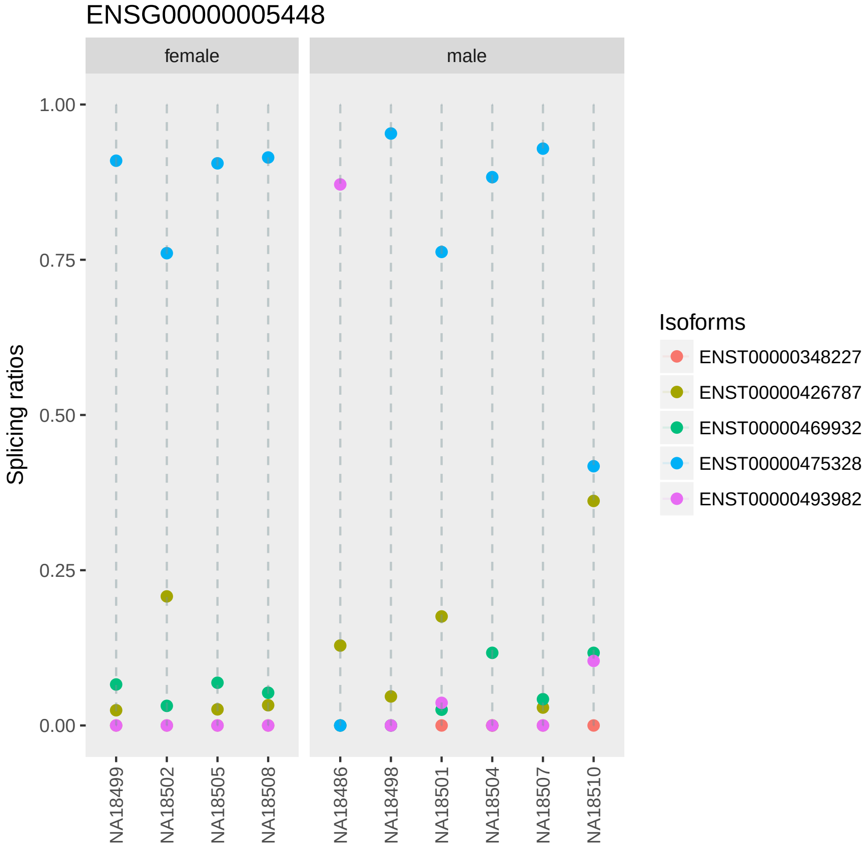

Figure 2. Plot showing splicing ratios of all transcripts in gene ENSG00000005448.

3 Data analysis

A single gene can be analyzed using the `testRasp` function. For instance, let us investigate whether the splicing ratios of the gene ENSG00000160741 containing 3 different transcripts are different between males and females (Figure 3).

Figure 3. Ternary plot corresponding to the splicing ratios of three transcripts (1st, 2nd and 3rd) belonging to gene ENSG00000160741 for males and females.

The p-value can be obtained by typing

```
> gene <- YRI[YRI$gene_id=="ENSG00000160741", 3:ncol(YRI)]
> mod <- testRasp(gene, genderYRI)
> mod
```

```
       female           male p.homogeneity         pvalue
   0.42113098     0.42765633     0.93280613     0.00209667
```

The first of the two p-values, `p.homogeneity`, represents the p-value resulting from the test for dispersion while the second one, `pvalue`, is the resulting p-value from the location test. The location test is sensitive to hetereogeneity in the dispersion of points of each group and, as a result, both p-values should always be taken into account. In this case, the dispersion p-value is highly non-significant, while the location p-value shows that there are significant differences in the splicing patterns among males and females. The ternary plot in Figure 4 also shows some differences across sexes. One can observe that females tend to have more RPKMs in the transcript ENST0000046.

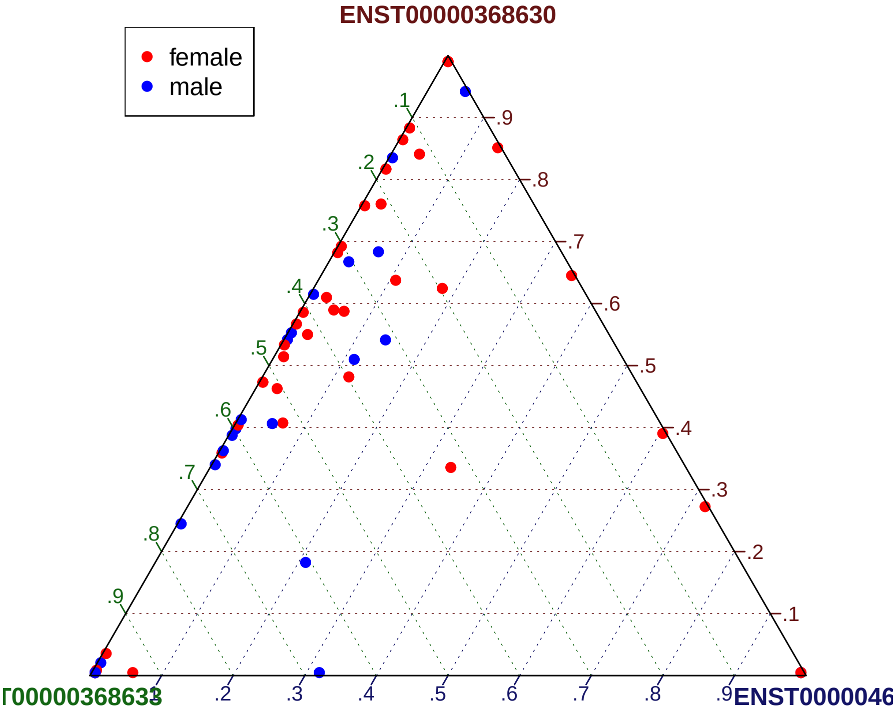

Figure 4. Ternary plot corresponding to the splicing ratios of three transcripts (1st, 2nd and 3rd) belonging to gene ENSG00000160741 for males and females.

This process can be performed using multiple processors when analyzing several genes. This procedure is implemented in the main function of Rasp called `rasp()`:

```
> ngenes <- 5
> sel.genes <-  names(table(YRI[,1]))[1:ngenes]
> res <- rasp(YRI[YRI[,1]%in%sel.genes, ], genderYRI, mc.cores = 1,
+             geneidCol = 1, expressionCols = 3:ncol(YRI))

  |
  |                                                                      |   0%
  |
  |==============                                                        |  20%
  |
  |============================                                          |  40%
  |
  |==========================================                            |  60%
  |
  |========================================================              |  80%
  |
  |======================================================================| 100%
```

Note that we have only selected 5 genes among the existing ones in order to avoid computational problems. Also notice that the function has several arguments beside the data and group objects. Firstly, it allows the user to use several cores during the computations (`mc.cores` argument). Secondly, three diffent tests are allowed. Finally, three arguments must be specified to tell which columns of the original dataset are the *gene id*, *transcript id* and *expression* respectively.

The centroids of each group, the p-values and BH adjusted p-values can be obtained by executing:

```
> print(res)

     female      male p.homogeneity    pvalue   padjust
1 0.5060722 0.4746152     0.4326448 0.3115221 0.4691305
2 0.6264664 0.3898278     0.1418535 0.1380508 0.3451269
3 0.3076060 0.3035915     0.9323146 0.3753044 0.4691305
4 0.4743753 0.4844921     0.9108994 0.9937647 0.9937647
5 0.2380074 0.2856032     0.4137914 0.1290607 0.3451269
```

3.1 Test for groups with more than two levels

As shown previously `rasp` performs a test that looks for differences in splicing ratios across all possible conditions of a certaing group variable. Sometimes researchers might be interested in testing whether the individuals of a specific subgroup or condition have significant differences in the splicing ratios compared to the rest of conditions. We have developed a test that, for a given group with more than two levels, performs a permutation test comparing each of the conditions against the rest. This can be done specifying `testGroup = TRUE` in both the `rasp` and `testRasp` functions, as shown in the following code.

```
> set.seed(1234)
> g <- as.character(genderYRI)
> g[g == "female"] <- ifelse(runif(sum(g == "female"))<.5, "female1", "female2")
> resG <- rasp(YRI[YRI[,1]%in%sel.genes, ], factor(g), mc.cores = 1,
+              geneidCol = 1, expressionCols = 3:ncol(YRI), testGroup = TRUE)

> print(resG)

    female1   female2      male p.homogeneity    pvalue pvalue_female1
1 0.5025654 0.4862753 0.4746152     0.8459983 0.2434855          0.817
2 0.5357731 0.6812374 0.3898278     0.2253030 0.2122954          0.491
3 0.3207199 0.2875963 0.3035915     0.8691705 0.5961924          0.682
4 0.5666431 0.3710038 0.4844921     0.2307668 0.4286361          0.984
5 0.2777184 0.1931642 0.2856032     0.3809330 0.3072558          0.845
  pvalue_female2 pvalue_male   padjust padjust_female1 padjust_female2
1          0.359       0.283 0.5120930               1           0.982
2          0.968       0.069 0.5120930               1           0.982
3          0.321       0.560 0.5961924               1           0.982
4          0.042       0.532 0.5357952               1           0.315
5          0.006       0.897 0.5120930               1           0.090
  padjust_male
1            1
2            1
3            1
4            1
5            1
```

4 Session info

```
> sessionInfo()

R version 3.3.2 (2016-10-31)
Platform: x86_64-w64-mingw32/x64 (64-bit)
Running under: Windows 10 x64 (build 14393)

locale:
[1] LC_COLLATE=Spanish_Spain.1252  LC_CTYPE=Spanish_Spain.1252
[3] LC_MONETARY=Spanish_Spain.1252 LC_NUMERIC=C
[5] LC_TIME=Spanish_Spain.1252

attached base packages:
[1] stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] rasp_1.2.2         CompQuadForm_1.4.3 ggplot2_2.2.1      DirichletReg_0.6-3
 [5] rgl_0.98.1         Formula_1.2-1      NMFN_2.0           vegan_2.4-3
 [9] lattice_0.20-34    permute_0.9-4

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.9      plyr_1.8.4       miscTools_0.6-22 tools_3.3.2
 [5] digest_0.6.11    jsonlite_1.2     tibble_1.2       nlme_3.1-128
 [9] gtable_0.2.0     mgcv_1.8-15      Matrix_1.2-7.1   shiny_1.0.0
[13] maxLik_1.3-4     parallel_3.3.2   stringr_1.2.0    cluster_2.0.5
[17] knitr_1.15.1     htmlwidgets_0.8  grid_3.3.2       R6_2.2.0
[21] reshape2_1.4.2   magrittr_1.5     scales_0.4.1     htmltools_0.3.5
[25] MASS_7.3-45      assertthat_0.1   mime_0.5         xtable_1.8-2
[29] colorspace_1.3-2 httpuv_1.3.3     labeling_0.3     sandwich_2.3-4
[33] stringi_1.1.2    lazyeval_0.2.0   munsell_0.4.3    zoo_1.7-14
```

References

Anderson, M. (2006). Distance-based tests for homogeneity of multivariate dispersion. *Biometrics*, 62:245–253.

Gonzalez-Porta, M., Calvo, M., Sammeth, M., and R, G. (2012). Estimation of alternative splicing variability in human populations. *Genome Research*, 22(3):528–38.

Marco-Sola, S., Sammeth, M., Guigó, R., and Ribeca, P. (2012). The gem mapper: fast, accurate and versatile alignment by filtration. *Nat Methods*, 9(12):1185–8.

Pickrell, J., Marioni, J., Pai, A., Degner, J., Engelhardt, B., Nkadori, E., Veyrieras, J., Stephens, M., Gilad, Y., and Pritchard, J. (2010). Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature*, 464:768–772.