Journal of Statistical Software

June 2004, Volume 11, Code Snippet 1. *http://www.jstatsoft.org/*

Venn Diagrams in R

Duncan J. Murdoch
University of Western Ontario

Abstract

This article describes R functions used to produce Venn diagrams and related incidence tables. The methods have applications in bioinformatics.

Keywords: Venn diagram, incidence tables.

1. Introduction

In bioinformatics, an "expression library" is a collection of thousands of samples of messenger RNA (mRNA) that have been extracted from a tissue, cloned, and annotated (identified in a standard list). Information on the function of protein derived from the mRNA can be inferred by comparing the incidence of different types of mRNA in samples drawn from different tissues or under different conditions. Biologists commonly use Venn diagrams to compare libraries: mRNA types that occur uniquely, or more prevalently, under one condition than under others appear outside the intersections in the diagram.

This note describes two functions written in R (Ihaka and Gentleman 1996) The function `incidence.table(id, category)` calculates a logical table indicating the incidence of each unique value of `id` in each `category`. An optional parameter `cutoff` (default value `1`) only sets an entry `TRUE` if at least that many replicates of the `id` occur in that `category`. This could be used to select mRNA types that are heavily expressed, for instance.

The function `venn()` draws a Venn diagram. Venn diagrams are commonly used in set theory to illustrate unions and intersections. Here the diagrams are augmented with numerical counts in each region, a graphical presentation of a cross-tabulation. We draw circles to represent the sets; this limits us to 2 or 3 sets, as the general Venn diagram with 4 or more sets requires more complicated shapes (Ruskey 2001). The function `venn()` can work with the same arguments as `incidence.table()`, in which case it passes them to that function to calculate the table, or it can work with a table calculated in some other way.

These two functions are contained in the tiny R package **venn** which accompanies this note.

They were written as part of the work described in (McKinney *et al.* 2003).

2. Example

A test dataset consisted of 1920 samples from 3 tissues. There were 1267 unique "accession numbers", or mRNA identifiers, in the dataset. Figure 1 shows three Venn diagrams of this data. On the left, all `id` values (accession numbers) are included. This figure was produced using the code

```
R> venn(accession, libname, main = "All samples")
```

where `accession` was a vector containing the codes identifying the RNA sequences, and `libname` was a vector containing the codes identifying the tissue sample (library). We see that 22 of the mRNA types were found in all three tissues, 25 in both tissues A and B, etc. In the centre, only those mRNA types for which 5 or more samples were observed are shown. We now see that there were 8 types of mRNA which appeared at least 5 times in tissue A but less frequently in the other tissues.

To narrow our attention to those 8 types, we do the following calculation:

```
R> tab <- incidence.table(accession, libname, cutoff = 5, duplicates = TRUE)
R> keep <- tab[, "A"] & !tab[, "B"] & !tab[, "C"]
```

This sets the variable keep to `TRUE` for those samples whose mRNA is one of the 8 types identified above.

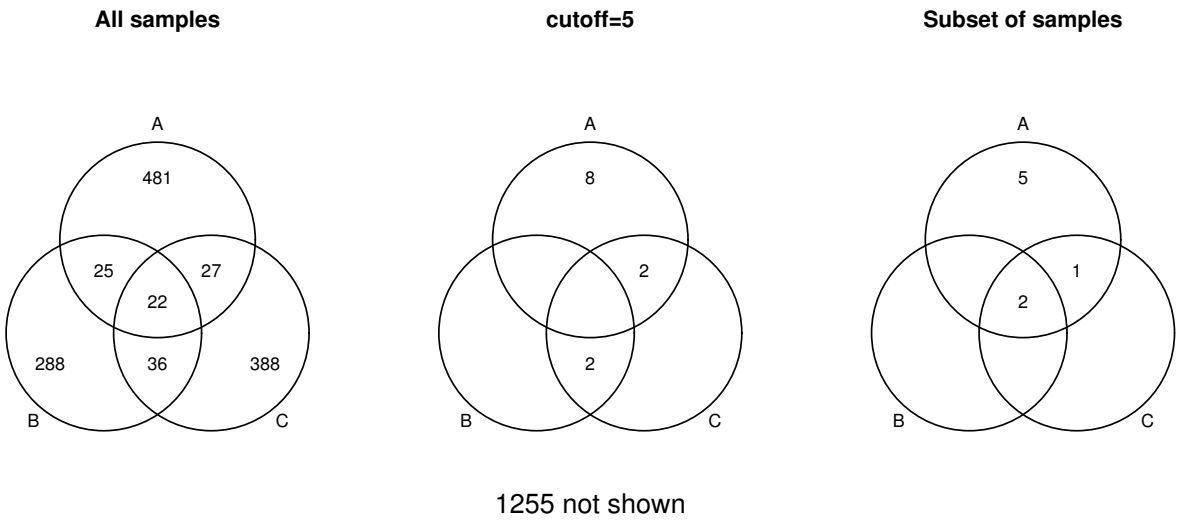

Figure 1: Venn diagrams produced by the `venn()` function. On the left, all samples; in the middle, only those which repeat at least 5 times. The figure on the right restricts attention to those 8 types identified in the previous figure to occur abundantly in tissue A but not in the other tissues.

Following this

```
R> venn(accession[keep], libname[keep], main = "Subset of samples")
```

produces the Venn diagram on the right, indicating that 5 of those mRNA types do not occur at all in the other tissues. Since they occurred frequently in the samples of tissue A, but are unique to that tissue, there is a strong suggestion that their function is related to the function of that tissue.

Acknowledgments

This work was supported in part by an NSERC Discovery Grant to Duncan Murdoch. The author is grateful to an anonymous referee for some very helpful comments.

References

Ihaka R, Gentleman R (1996). "R: A Language for Data Analysis and Graphics." *Journal of Computational and Graphical Statistics*, **5**, 299–314.

McKinney JL, Wang J, Robinson J, Biltcliffe C, Song QL, Chen B, Walker P, Murdoch D, Hegele RA (2003). "METTA: A Highly Expressed Gene Specific to Adult Heart Found Through Mining of Expressed Sequence Tags from Cardiac Tissue." In preparation.

Ruskey F (2001). "A Survey of Venn Diagrams." *The Electronic Journal of Combinatorics*, **DS #5**. URL http://www.combinatorics.org/Surveys/ds5/VennWhatEJC.html.

Affiliation:

Duncan J. Murdoch
Department of Statistical and Actuarial Sciences
University of Western Ontario, London, Ontario, Canada N6A 5B7
E-mail: murdoch@stats.uwo.ca
URL: http://fisher.stats.uwo.ca/faculty/murdoch/

Journal of Statistical Software
June 2004, Volume 11, Code Snippet 1.
http://www.jstatsoft.org/

Submitted: 2004-06-24
Accepted: 2004-06-27