Weighted Kolmogorov-Smirnov test: using the R functions

K. Charmpi, B. Ycart

The file `wks.tgz` unzips into a directory `/wks/` which contains:

- `wks_manual.pdf`: this file;
- `wks.r`: the script file. Open R, set `/wks/` as working directory, then source the script: `source("wks.r")`.
- `C2.rds`: the MSig database C2 from Subramanian et al. (2005). It is encoded as a list of vectors of character strings: see Ycart et al. (2014). Each vector in the list is a gene set. Load the database by:

  ```
  C2 <- readRDS("C2.rds")
  ```

- `F1.rds`, `F1os.rds`: pre-calculated distribution function for WKS test applied to rank statistics;
- `breast.rds`, `kidney.rds`, `liver.rds`: three sample vectors of expression levels, from the GEO dataset GSE36133 of Barretina et al. (2012): GSM887033 (breast), GSM886844 (kidney), GSM887630 (liver). Load them by:

  ```
  b <- readRDS("breast.rds")
  k <- readRDS("kidney.rds")
  l <- readRDS("liver.rds")
  ```

Two test functions are provided: `WKS.test` and `GSEA.test`. Both take a vector and a database as their first two entries. Default values are provided for the other variables. The output is a sorted vector of p-values, indexed by the gene set names. The three alternatives are `"greater"`, `"less"`, and `"two.sided"` (default). For `"alternative="greater"`, the more enriched the gene set, the smaller the p-value. Respectively, for `"alternative="less"`, the more depleted the gene set, the smaller the p-value.

```
                                # default: two.sided
pvb <- WKS.test(b,C2)
pvk <- WKS.test(k,C2)
pvl <- WKS.test(l,C2)
```

```
                                  # enriched
pvbg <- WKS.test(b,C2,alternative="greater")
pvkg <- WKS.test(k,C2,alternative="greater")
pvlg <- WKS.test(l,C2,alternative="greater")

                                  # depleted
pvbl <- WKS.test(b,C2,alternative="less")
pvkl <- WKS.test(k,C2,alternative="less")
pvll <- WKS.test(l,C2,alternative="less")
```

P-values can be visualized using `paired.pvalues`. The entries are two vectors of p-values and a keyword, to be searched among gene set names. Negative logarithms of p-values are plotted. The gene sets whose names match the keyword appear as triangles. They can be identified by clicking on the plot, if `nid` is provided. Title and axes labels are optional.

```
paired.pvalues(pvb,pvl)
paired.pvalues(pvb,pvl,"breast")
paired.pvalues(pvb,pvl,"liver",nid=3)
paired.pvalues(pvbg,pvlg,"liver",nid=3)
paired.pvalues(pvbl,pvll,"liver",nid=3)
{paired.pvalues(pvb,pvl,"liver",
   title="breast vs. liver",xlab="breast",ylab="liver")}
```

False Detection Rate adjustment of p-value vectors is obtained by the R function `p.adjust`.

```
pvba <- p.adjust(pvb,method="BY")
pvka <- p.adjust(pvk,method="BY")
pvla <- p.adjust(pvl,method="BY")
```

Significant p-values can be listed as follows.

```
pvba[pvba<0.05]
pvka[pvka<0.05]
pvla[pvla<0.05]
```

By default, `WKS.test` calculates over ten thousand simulations. This can be modified using the variable `nsim`. The default value for the number of discretization points `ndiscr` is the size of the vector.

```
pvb1 <- WKS.test(b,C2,nsim=1e5,ndiscr=1000)
paired.pvalues(pvb1,pvb)
```

The initial vector can be replaced by its rank statistics. In that case, a more precise calculation is done, using the values of `F1.rds`, which have been pre-calculated over one million simulations.

```
pvbr <- WKS.test(b,C2,ranked=TRUE)
paired.pvalues(pvb,pvbr,"breast")
pvkr <- WKS.test(k,C2,ranked=TRUE)
paired.pvalues(pvk,pvkr,"kidney")
pvlr <- WKS.test(l,C2,ranked=TRUE)
paired.pvalues(pvl,pvlr,"liver")
```

The function `GSEA.test` reproduces the code from Subramanian et al. (2005, 2007). The default value for the number of simulations is one thousand. Even for that reduced number, it is much slower than `WKS.test`. Figure 3 in the paper has been produced by the following commands (`GSEA.test` with ten thousand simulations for each gene set takes very long).

```
pvlW <- WKS.test(l,C2,nsim=1e5)
pvlG <- GSEA.test(l,C2,nsim=1e4)
paired.pvalues(pvlG,pvlW,"liver")
```

Two visualizations of enrichment scores are proposed: `cumulated.weights` and `enrichment.plot`. The first one plots the cumulated proportions of weights (function $S_n(t)$ in the paper), which is used in both WKS and GSEA tests. The entries are the vector and the database to be tested, plus a vector of gene set names. All curves are superposed on the same graphic. The second function, `enrichment.plot` plots for each gene set, the difference between the cumulated weights and their expectation under the null hypothesis.

```
               # two-sided, low p-values
gs <- names(pvb)[1:10]
cumulated.weights(b,C2,gs)
enrichment.plot(b,C2,gs)
               # two-sided, high p-values
gs <- names(tail(pvb,10))
cumulated.weights(b,C2,gs)
enrichment.plot(b,C2,gs)
```

```
                # right-sided, low p-values = enriched
gs <- names(pvbg)[1:10]
cumulated.weights(b,C2,gs)
enrichment.plot(b,C2,gs)
                # right-sided, high p-values = depleted
gs <- names(tail(pvbg,10))
cumulated.weights(b,C2,gs)
enrichment.plot(b,C2,gs)
                # left-sided, low p-values = depleted
gs <- names(pvbl)[1:10]
cumulated.weights(b,C2,gs)
enrichment.plot(b,C2,gs)
                # left-sided test, high p-values = enriched
gs <- names(tail(pvbl,10))
cumulated.weights(b,C2,gs)
enrichment.plot(b,C2,gs)

cumulated.weights(l,C2,"ACEVEDO_METHYLATED_IN_LIVER_CANCER_DN")
enrichment.plot(l,C2,"ACEVEDO_METHYLATED_IN_LIVER_CANCER_DN")
```

This is only a script, and not a R package. The entries have not been protected and some functions may fail on extreme entries. We have tried to respect the spirit and scope of R but we have also focused on clarity and readability of the codes. There is certainly room for gain in precision and computing time. You are welcome to read and modify the code for your own usage, and get back to us for possible improvement.

References

Barretina J., Caponigro G., Stransky N., Venkatesan K., and others (2012): "The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity," *Nature*, 483(7391), 603–7.

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. and Mesirov J. P. (2005): "Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles," *PNAS*, 102, 15545–50, URL `http://www.pnas.org/content/102/43/15545.full`.

Subramanian A., Kuehn H., Gould J., Tamayo P., and Mesirov J. P. (2007): "GSEA-P: a desktop application for Gene Set Enrichment Analysis," *Bioinformatics*, 23(23), 3251–3.

Ycart B., Pont F., and Fournié J. J. (2014): "Curbing false discovery rates in interpretation of genome-wide expression profiles," *J Biomed Inform.*, 47, 58–61.