

# Using the ONCOSCORE package

Luca De Sano\*   Carlo Gambacorti Passerini†   Rocco Piazza†   Daniele Ramazzotti\*  
Roberta Spinelli†

September 29, 2017

**Overview.** ONCOSCORE is a tool to measure the association of genes to cancer based on citation frequency in biomedical literature. The score is evaluated from PubMed literature by dynamically updatable web queries.

*In this vignette, we give an overview of the package by presenting some of its main functions.*

The ONCOSCORE analysis consists of two parts. One can estimate a score to assess the oncogenic potential of a set of genes, given the literature knowledge, at the time of the analysis, or one can study the trend of such score over time.

We next present the two analysis and we conclude with showing the capabilities of the tool to visualize the results.

**Requirements.** First we load the library.

```
library(OncoScore)
```

**OncoScore analysis.** The query that we show next retrieves from PubMed the citations, at the time of the query, for a list of genes in cancer related and in all the documents.

```
query = perform.query(c("ASXL1", "IDH1", "IDH2", "SETBP1", "TET2"))
```

```
### Starting the queries for the selected genes.
```

```
### Performing queries for cancer literature
```

```
Number of papers found in PubMed for ASXL1 was: 309  
Number of papers found in PubMed for IDH1 was: 1450  
Number of papers found in PubMed for IDH2 was: 557  
Number of papers found in PubMed for SETBP1 was: 69  
Number of papers found in PubMed for TET2 was: 560
```

```
### Performing queries for all the literature
```

```
Number of papers found in PubMed for ASXL1 was: 350  
Number of papers found in PubMed for IDH1 was: 1581  
Number of papers found in PubMed for IDH2 was: 648  
Number of papers found in PubMed for SETBP1 was: 89  
Number of papers found in PubMed for TET2 was: 717
```

---

\*Dipartimento di Informatica Sistemistica e Comunicazione, Università degli Studi Milano Bicocca Milano, Italy.

†Dipartimento di Medicina e Chirurgia, Università degli Studi Milano Bicocca Milano, Italy.

OncoScore provides a function to merge gene names if requested by the user. This function is useful when there are aliases in the gene list.

```
combine.query.results(query, c('IDH1', 'IDH2'), 'new_gene')

## CitationsGene CitationsGeneInCancer
## ASXL1 350 309
## SETBP1 89 69
## TET2 717 560
## new_gene 2229 2007
```

OncoScore also provides a function to retrieve the names of the genes in a given portion of a chromosome that can be exploited if we are dealing, e.g., with copy number alterations hitting regions rather than specific genes.

```
chr13 = get.genes.from.biomart(chromosome=13,start=54700000,end=72800000)
head(chr13)
```

```
[1] "LINC00374" "RNA5SP30" "RNU7-87P" "HNF4GP1" "RN7SL375P" "BORA"
```

Furthermore, one can also automatically perform the OncoScore analysis on chromosomal regions as follows:

```
result = compute.oncoscore.from.region(10, 100000, 500000)
```

```
### Performing query on BioMart
### Performing web query on: RNA5SP297 RNA5SP298 RN7SL754P ZMYND11 DIP2C
### Starting the queries for the selected genes.
```

```
### Performing queries for cancer literature
Number of papers found in PubMed for RNA5SP297 was: -1
Number of papers found in PubMed for RNA5SP298 was: -1
Number of papers found in PubMed for RN7SL754P was: -1
Number of papers found in PubMed for ZMYND11 was: 27
Number of papers found in PubMed for DIP2C was: 1
```

```
### Performing queries for all the literature
Number of papers found in PubMed for RNA5SP297 was: -1
Number of papers found in PubMed for RNA5SP298 was: -1
Number of papers found in PubMed for RN7SL754P was: -1
Number of papers found in PubMed for ZMYND11 was: 45
Number of papers found in PubMed for DIP2C was: 3
```

```
### Processing data
### Computing frequencies scores
### Estimating oncogenes
### Results:
RNA5SP297 -> 0
RNA5SP298 -> 0
RN7SL754P -> 0
ZMYND11 -> 49.07473
DIP2C -> 12.30234
```

We now compute a score for each of the genes, to estimate their oncogenic potential.

```
result = compute.oncoscore(query)

## ### Processing data
## ### Computing frequencies scores
## ### Estimating oncogenes
## ### Results:
## ASXL1 -> 77.8392
## IDH1 -> 83.08351
## IDH2 -> 76.75356
## SETBP1 -> 65.556
## TET2 -> 69.86954
```

**OncoScore timeline analysis.** The query that we show next retrieves from PubMed the citations, at specified time points, for a list of genes in cancer related and in all the documents.

```
query.timepoints = perform.query.timeseries(c("ASXL1", "IDH1", "IDH2", "SETBP1", "TET2"),
      c("2012/03/01", "2013/03/01", "2014/03/01", "2015/03/01", "2016/03/01"))
```

```
### Starting the queries for the selected genes.
```

```
### Querying PubMed for timepoint 2012/03/01
```

```
### Performing queries for cancer literature
```

```
Number of papers found in PubMed for ASXL1 was: 83
```

```
Number of papers found in PubMed for IDH1 was: 408
```

```
Number of papers found in PubMed for IDH2 was: 172
```

```
Number of papers found in PubMed for SETBP1 was: 5
```

```
Number of papers found in PubMed for TET2 was: 169
```

```
### Performing queries for all the literature
```

```
Number of papers found in PubMed for ASXL1 was: 91
```

```
Number of papers found in PubMed for IDH1 was: 488
```

```
Number of papers found in PubMed for IDH2 was: 234
```

```
Number of papers found in PubMed for SETBP1 was: 10
```

```
Number of papers found in PubMed for TET2 was: 196
```

```
### Querying PubMed for timepoint 2013/03/01
```

```
### Performing queries for cancer literature
```

```
Number of papers found in PubMed for ASXL1 was: 132
```

```
Number of papers found in PubMed for IDH1 was: 662
```

```
Number of papers found in PubMed for IDH2 was: 267
```

```
Number of papers found in PubMed for SETBP1 was: 11
```

```
Number of papers found in PubMed for TET2 was: 254
```

```
### Performing queries for all the literature
```

```
Number of papers found in PubMed for ASXL1 was: 149
```

```
Number of papers found in PubMed for IDH1 was: 753
```

```
Number of papers found in PubMed for IDH2 was: 336
```

```
Number of papers found in PubMed for SETBP1 was: 18
```

```
Number of papers found in PubMed for TET2 was: 302
```

```
### Querying PubMed for timepoint 2014/03/01
```

```
### Performing queries for cancer literature
```

```
Number of papers found in PubMed for ASXL1 was: 185
```

```
Number of papers found in PubMed for IDH1 was: 903
```

```
Number of papers found in PubMed for IDH2 was: 364
```

```
Number of papers found in PubMed for SETBP1 was: 29
```

```
Number of papers found in PubMed for TET2 was: 342
```

```
### Performing queries for all the literature
```

```
Number of papers found in PubMed for ASXL1 was: 208
```

```
Number of papers found in PubMed for IDH1 was: 1002
```

```
Number of papers found in PubMed for IDH2 was: 439
```

```
Number of papers found in PubMed for SETBP1 was: 36
```

```
Number of papers found in PubMed for TET2 was: 430
```

```
### Querying PubMed for timepoint 2015/03/01
```

```
### Performing queries for cancer literature
```

```
Number of papers found in PubMed for ASXL1 was: 250
```

```
Number of papers found in PubMed for IDH1 was: 1188
```

```
Number of papers found in PubMed for IDH2 was: 467
```

```
Number of papers found in PubMed for SETBP1 was: 49
```

```
Number of papers found in PubMed for TET2 was: 452
```

```
### Performing queries for all the literature
```

```
Number of papers found in PubMed for ASXL1 was: 283
```

```
Number of papers found in PubMed for IDH1 was: 1300
```

```
Number of papers found in PubMed for IDH2 was: 550
```

```
Number of papers found in PubMed for SETBP1 was: 64
```

```
Number of papers found in PubMed for TET2 was: 576
```

```

### Querying PubMed for timepoint 2016/03/01
### Performing queries for cancer literature
Number of papers found in PubMed for ASXL1 was: 309
Number of papers found in PubMed for IDH1 was: 1446
Number of papers found in PubMed for IDH2 was: 557
Number of papers found in PubMed for SETBP1 was: 69
Number of papers found in PubMed for TET2 was: 558
### Performing queries for all the literature
Number of papers found in PubMed for ASXL1 was: 350
Number of papers found in PubMed for IDH1 was: 1576
Number of papers found in PubMed for IDH2 was: 648
Number of papers found in PubMed for SETBP1 was: 89
Number of papers found in PubMed for TET2 was: 715

```

We now compute a score for each of the genes, to estimate their oncogenic potential at specified time points.

```

result.timeseries = compute.oncoscore.timeseries(query.timepoints)

## ### Computing oncoscore for timepoint 2012/03/01
## ### Processing data
## ### Computing frequencies scores
## ### Estimating oncogenes
## ### Results:
## ASXL1 -> 77.19348
## IDH1 -> 74.24489
## IDH2 -> 64.1649
## SETBP1 -> 34.9485
## TET2 -> 74.90108
## ### Computing oncoscore for timepoint 2013/03/01
## ### Processing data
## ### Computing frequencies scores
## ### Estimating oncogenes
## ### Results:
## ASXL1 -> 76.31902
## IDH1 -> 78.71551
## IDH2 -> 69.99559
## SETBP1 -> 46.4559
## TET2 -> 73.89695
## ### Computing oncoscore for timepoint 2014/03/01
## ### Processing data
## ### Computing frequencies scores
## ### Estimating oncogenes
## ### Results:
## ASXL1 -> 77.39202
## IDH1 -> 81.07946
## IDH2 -> 73.46995
## SETBP1 -> 64.97398
## TET2 -> 70.44331
## ### Computing oncoscore for timepoint 2015/03/01
## ### Processing data
## ### Computing frequencies scores
## ### Estimating oncogenes
## ### Results:
## ASXL1 -> 77.49295
## IDH1 -> 82.55032
## IDH2 -> 75.58179
## SETBP1 -> 63.80208
## TET2 -> 69.91466
## ### Computing oncoscore for timepoint 2016/03/01
## ### Processing data

```

```
## ### Computing frequencies scores
## ### Estimating oncogenes
## ### Results:
## ASXL1 -> 77.8392
## IDH1 -> 83.11346
## IDH2 -> 76.75356
## SETBP1 -> 65.556
## TET2 -> 69.81125
```

**Visualization of the results.** We next plot the scores measuring the oncogenetic potential of the considered genes as a barplot.

```
plot.oncoscore(result, col = 'darkblue')
```

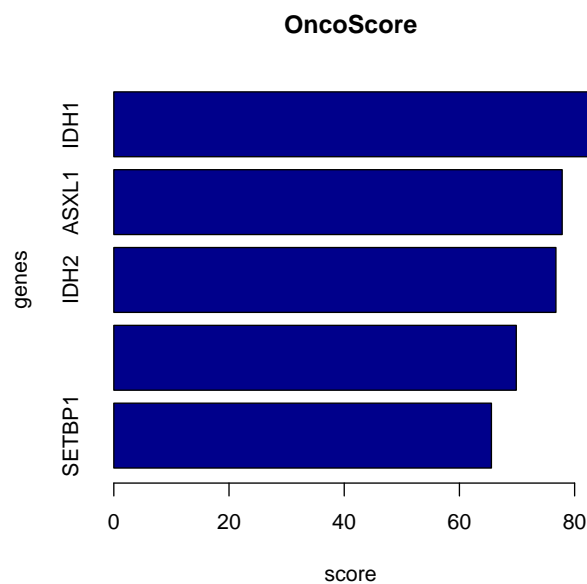


Figure 1: **Oncogenetic potential of the considered genes.**

We finally plot the trend of the scores over the considered times as absolute and values and as variations.

```
plot.oncoscore.timeseries(result.timeseries)
```

```
plot.oncoscore.timeseries(result.timeseries,
  incremental = TRUE,
  ylab='absolute variation')
```

```
plot.oncoscore.timeseries(result.timeseries,
  incremental = TRUE,
  relative = TRUE,
  ylab='relative variation')
```

```
sessionInfo()
```

- R version 3.4.1 (2017-06-30), x86\_64-w64-mingw32
- Locale: LC\_COLLATE=C, LC\_CTYPE=English\_United States.1252, LC\_MONETARY=English\_United States.1252, LC\_NUMERIC=C, LC\_TIME=English\_United States.1252
- Running under: Windows Server 2012 R2 x64 (build 9600)
- Matrix products: default
- Base packages: base, datasets, grDevices, graphics, methods, stats, utils

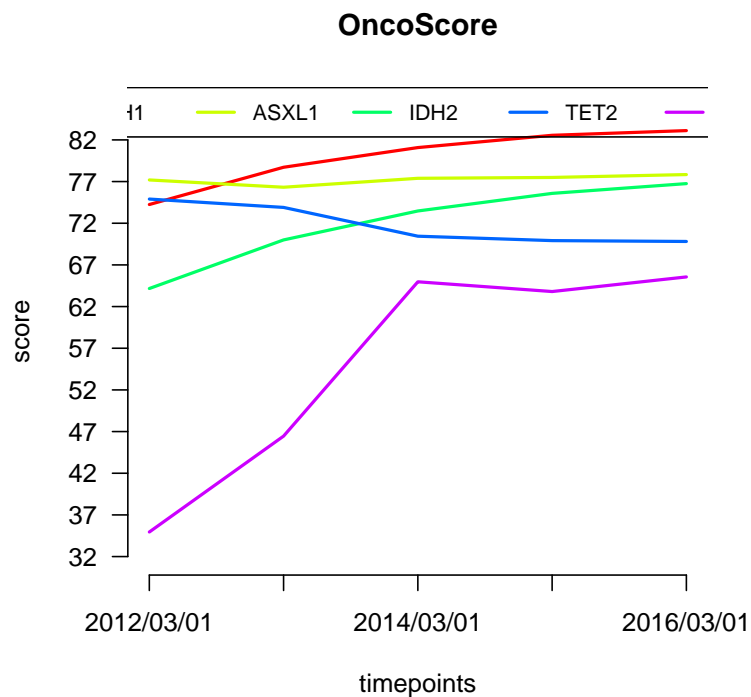


Figure 2: **Absolute values of the oncogenetic potential of the considered genes over times.**

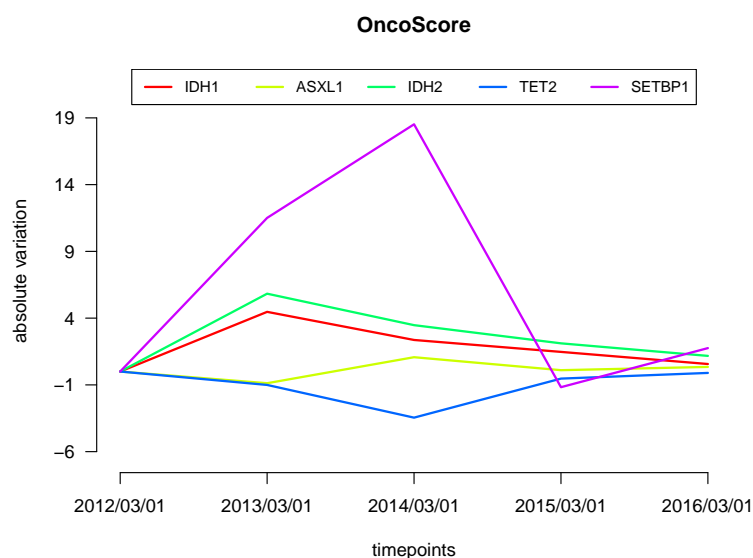


Figure 3: **Variations of the oncogenetic potential of the considered genes over times.**

- Other packages: OncoScore 1.4.2
- Loaded via a namespace (and not attached): AnnotationDbi 1.38.2, Biobase 2.36.2, BiocGenerics 0.22.0, BiocStyle 2.4.1, DBI 0.7, IRanges 2.10.3, Rcurl 1.95-4.8, RSQLite 2.0, Rcpp 0.12.13, S4Vectors 0.14.5, XML 3.98-1.9, backports 1.1.1, biomaRt 2.32.1, bit 1.1-12, bit64 0.9-7, bitops 1.0-6, blob 1.1.0, compiler 3.4.1, digest 0.6.12, evaluate 0.10.1, highr 0.6, htmltools 0.3.6, knitr 1.17, magrittr 1.5, memoise 1.1.0, parallel 3.4.1, rlang 0.1.2, rmarkdown 1.6, rprojroot 1.2, stats4 3.4.1, stringi 1.1.5, stringr 1.2.0, tibble 1.3.4, tools 3.4.1, yaml 2.1.14

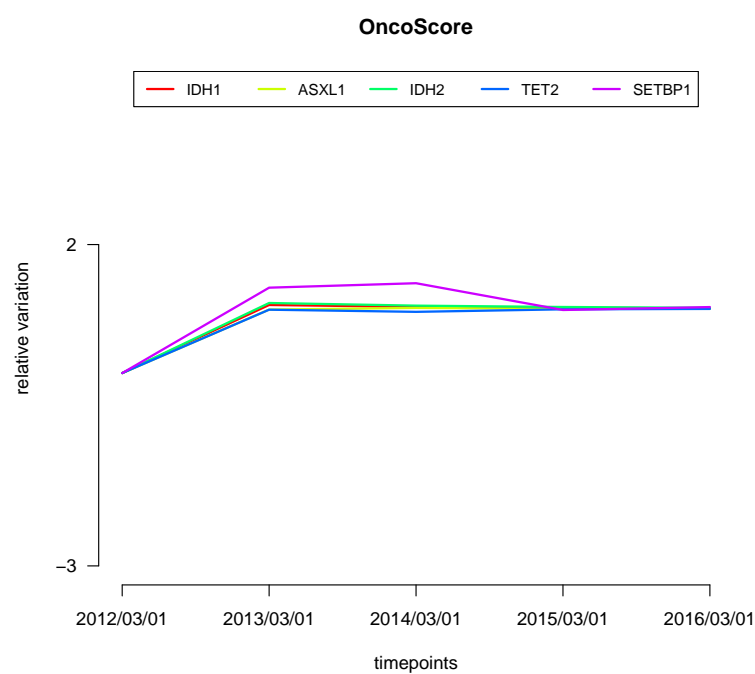


Figure 4: Variations as relative values of the oncogenetic potential of the considered genes over times.