goseq: Gene Ontology testing for RNA-seq datasets

Matthew D. Young Nadia Davidson Matthew J. Wakefield
Gordon K. Smyth Alicia Oshlack

8 September 2017

1 Introduction

This document gives an introduction to the use of the goseq R Bioconductor package [Young et al., 2010]. This package provides methods for performing Gene Ontology analysis of RNA-seq data, taking length bias into account [Oshlack and Wakefield, 2009]. The methods and software used by goseq are equally applicable to other category based test of RNA-seq data, such as KEGG pathway analysis.

Once installed, the goseq package can be easily loaded into R using:

```
> library(goseq)
```

In order to perform a GO analysis of your RNA-seq data, goseq only requires a simple named vector, which contains two pieces of information.

1. `Measured genes`: all genes for which RNA-seq data was gathered for your experiment. Each element of your vector should be named by a unique gene identifier.
2. `Differentially expressed genes`: each element of your vector should be either a 1 or a 0, where 1 indicates that the gene is differentially expressed and 0 that it is not.

If the organism, gene identifier or category test is currently not natively supported by goseq, it will also be necessary to supply additional information regarding the genes length and/or the association between categories and genes.

Bioconductor R packages such as Rsubread allow for the summarization of mapped reads into a table of counts, such as reads per gene. From there, several packages exist for performing differential expression analysis on summarized data (eg. edgeR [Robinson and Smyth, 2007, 2008, Robinson et al., 2010]). goseq will work with any method for determining differential expression and as such differential expression analysis is outside the scope of this document, but in order to facilitate ease of use, we will make use of the edgeR package to calculate differentially expressed (DE) genes in all the case studies in this document.

2 Reading data

We assume that the user can use appropriate in-built `R` functions (such as `read.table` or `scan`) to obtain two vectors, one containing all genes assayed in the RNA-seq experiment, the other containing all genes which are DE. If we assume that the vector of genes being assayed is named `assayed.genes` and the vector of DE genes is named `de.genes` we can construct a named vector suitable for use with `goseq` using the following:

```
> gene.vector <- as.integer(assayed.genes %in% de.genes)
> names(gene.vector) <- assayed.genes
> head(gene.vector)
```

It may be that the user can already read in a vector in this format, in which case it can then be immediately used by `goseq`.

3 GO testing of RNA-seq data

To begin the analysis, `goseq` first needs to quantify the length bias present in the dataset under consideration. This is done by calculating a Probability Weighting Function or PWF which can be thought of as a function which gives the probability that a gene will be differentially expressed (DE), based on its length alone. The PWF is calculated by fitting a monotonic spline to the binary data series of differential expression (1=DE, 0=Not DE) as a function of gene length. The PWF is used to weight the chance of selecting each gene when forming a null distribution for GO category membership. The fact that the PWF is calculated directly from the dataset under consideration makes this approach robust, only correcting for the length bias present in the data. For example, if `goseq` is run on a microarray dataset, for which no length bias exists, the calculated PWF will be nearly flat and all genes will be weighted equally, resulting in no length bias correction.

In order to account for the length bias inherent to RNA-seq data when performing a GO analysis (or other category based tests), one cannot simply use the hypergeometric distribution as the null distribution for category membership, which is appropriate for data without DE length bias, such as microarray data. GO analysis of RNA-seq data requires the use of random sampling in order to generate a suitable null distribution for GO category membership and calculate each categories significance for over representation amongst DE genes.

However, this random sampling is computationally expensive. In most cases, the Wallenius distribution can be used to approximate the true null distribution, without any significant loss in accuracy. The `goseq` package implements this approximation as its default option. The option to generate the null distribution using random sampling is also included as an option, but users should be aware that the default number of samples generated will not be enough to accurately call enrichment when there are a large number of go terms.

Having established a null distribution, each GO category is then tested for over and under representation amongst the set of differentially expressed genes and the null is used to calculate a p-value for under and over representation.

4 Natively supported Gene Identifiers and category tests

`goseq` needs to know the length of each gene, as well as what GO categories (or other categories of interest) each gene is associated with. `goseq` relies on the UCSC genome browser to provide the length information for each gene. However, because the process of fetching the length of every transcript is slow and bandwidth intensive, `goseq` relies on an offline copy of this information stored in the data package `geneLenDataBase`. To see which genome/gene identifier combinations are in the local database, simply run:

```
> supportedOrganisms()
```

The leftmost columns in the output of this command list the genomes and gene identifiers respectively. If length data exists in the local database it is indicated in the second last column. If your genome/ID combination is not in the local database, it may be downloaded from the UCSC genome browser or taken from a `TxDb` annotation package (if installed). If your genome/ID combination is not found in any database, you will have to manually specify the gene lengths. We encourage all users to manually specify their gene lengths if provided by upstream summarization programs. e.g. `featureCounts`, as these lengths will be more accurate.

In order to link GO categories to genes, `goseq` uses the organism packages from Bioconductor. These packages are named org.<Genome>.<ID>.db, where <Genome> is a short string identifying the genome and <ID> is a short string identifying the gene identifier. Currently, `goseq` will automatically retrieve the mapping between GO categories and genes from the relevant package (as long as it is installed) for commonly used genome/ID combinations. If GO mappings are not automatically available for your genome/ID combination, you will have to manually specify the relationship between genes and categories. Although the Genome/ID naming conventions used by the organism packages differ from the UCSC, `goseq` is able to convert between the two, so the user need only ever specify the UCSC genome/ID in most cases. The final column indicates whether the Genome/ID combination is supported for GO categories.

5 Non-native Gene Identifier or category test

If the organism, Gene Identifier or category test you wish to perform is not in the native `goseq` database, you will have to supply one or all of the following:

- `Length data`: the length of each gene in your gene identifier format.

- `Category mappings`: the mapping (usually many-to-many) between the categories you wish to test for over/under representation amongst DE genes and genes in your gene identifier format.

5.1 Length data format

The length data must be formatted as a numeric vector, of the same length as the main named vector specifying gene names/DE genes. Each entry should give the length of the corresponding gene in bp. If length data is unavailable for some genes, that entry should be set to NA.

5.2 Category mapping format

The mapping between category names and genes should be given as a data frame with two columns. One column should contain the gene IDs and the other the name of an associated category. As the mapping between categories and genes is usually many-to-many, this data frame will usually have multiple rows with the same gene name and category name.

Alternatively, mappings between genes and categories can be given as a list. The names of list entries should be gene IDs and the entries themselves should be a vector of category names to which the gene ID corresponds.

5.3 Some additional tips

Any organism for which there is an annotation on either Ensembl or the UCSC, can be easily turned into length data using the GenomicFeatures package. To do this, first create a TranscriptDb object using either makeTxDbFromBiomart or makeTxDbFromUCSC (see the help in the GenomicFeatures package on using these commands). Once you have a transcriptDb object, you can get a vector named by gene ID containing the median transcript length of each gene simply by using the command.

```
> txsByGene <- transcriptsBy(txdb, "gene")
> lengthData <- median(width(txsByGene))
```

The relationship between gene identifier and GO category can usually be obtained from the Gene Ontology website (www.geneontology.org) or from the NCBI. Additionally, the bioconductor AnnotationDbi library has recently added a function "makeOrgPackageFromNCBI", which can be used to create an organism package from within R, using the NCBI data. Once created, this package can then be used to obtain the mapping between genes and gene ontology.

6 Case study: Prostate cancer data

6.1 Introduction

This section provides an analysis of data from an RNA-seq experiment to illustrate the use of `goseq` for GO analysis.

This experiment examined the effects of androgen stimulation on a human prostate cancer cell line, LNCaP. The data set includes more than 17 million short cDNA reads obtained for both the treated and untreated cell line and sequenced on Illumina's 1G genome analyzer.

For each sample we were provided with the raw 35 bp RNA-seq reads from the authors. For the untreated prostate cancer cells (LNCaP cell line) there were 4 lanes totaling 10 million, 35 bp reads. For the treated cells there were 3 lanes totaling 7 million, 35 bp reads. All replicates were technical replicates. Reads were mapped to NCBI version 36.3 of the human genome using bowtie. Any read with which mapped to multiple locations was discarded. Using the ENSEMBL 54 annotation from biomart, each mapped read was associated with an ENSEMBL gene. This was done by associating any read that overlapped with any part of the gene (not just the exons) with that gene. Reads that did not correspond to genes were discarded.

6.2 Source of the data

The data set used in this case study is taken from [Li et al., 2008] and was made available from the authors upon request.

6.3 Determining the DE genes using `edgeR`

To begin with, we load in the text data and convert it the appropriate `edgeR` DGEList object.

```
> library(edgeR)
> table.summary <- read.table(system.file("extdata", "Li_sum.txt", package = "goseq"),
+   sep = "\t", header = TRUE, stringsAsFactors = FALSE
+ )
> counts <- table.summary[, -1]
> rownames(counts) <- table.summary[, 1]
> grp <- factor(rep(c("Control", "Treated"), times = c(4, 3)))
> summarized <- DGEList(counts, lib.size = colSums(counts), group = grp)
```

Next, we use `edgeR` to estimate the biological dispersion and calculate differential expression using a negative binomial model.

```
> disp <- estimateCommonDisp(summarized)
> disp$common.dispersion
```

```
[1] 0.05688364
```

```
> tested <- exactTest(disp)
> topTags(tested)
```

```
Comparison of groups:  Treated-Control
                    logFC   logCPM       PValue          FDR
ENSG00000127954 11.557868 6.680748 2.574972e-80 1.274766e-75
ENSG00000151503  5.398963 8.499530 1.781732e-65 4.410321e-61
ENSG00000096060  4.897600 9.446705 7.983755e-60 1.317479e-55
ENSG00000091879  5.737627 6.282646 1.207655e-54 1.494654e-50
ENSG00000132437 -5.880436 7.951910 2.950042e-52 2.920896e-48
ENSG00000166451  4.564246 8.458467 7.126762e-52 5.880292e-48
ENSG00000131016  5.254737 6.607957 1.066807e-51 7.544765e-48
ENSG00000163492  7.085400 5.128514 2.716461e-45 1.681014e-41
ENSG00000113594  4.051053 8.603264 9.272066e-44 5.100254e-40
ENSG00000116285  4.108522 7.864773 6.422468e-43 3.179507e-39
```

Finally, we Format the DE genes into a vector suitable for use with `goseq`

```
> genes <- as.integer(p.adjust(tested$table$PValue[tested$table$logFC != 0],
+   method = "BH"
+ ) < .05)
> names(genes) <- row.names(tested$table[tested$table$logFC != 0, ])
> table(genes)
```

```
genes
    0     1
19535  3208
```

6.4 Determining Genome and Gene ID

In order to allow for automatic data retrieval, the user has to tell `goseq` what genome and gene ID format were used to summarize the data. In our case we will use the hg19 build of the human genome, we check what code this corresponds to by running:

```
> head(supportedOrganisms())
```

```
           Genome      Id  Id Description Lengths in geneLeneDataBase
138   Arabidopsis                                                FALSE
139   E. coli K12                                                FALSE
140 E. coli Sakai                                                FALSE
141       Malaria                                                FALSE
```

```
1         anoCar1 ensGene Ensembl gene ID                          TRUE
2         anoGam1 ensGene Ensembl gene ID                          TRUE
    GO Annotation Available
138                    TRUE
139                    TRUE
140                    TRUE
141                    TRUE
1                     FALSE
2                      TRUE
```

Which lists the genome codes in the far left column, headed "Genome". As we are using "hg19" and we also know that we used ENSEMBL Gene ID to summarize our read data, we check what code this corresponds to by running:

```
> supportedOrganisms()[supportedOrganisms()$Genome == "hg19", ]
```

```
   Genome         Id  Id Description Lengths in geneLeneDataBase
25   hg19    ensGene Ensembl gene ID                         TRUE
48   hg19  knownGene  Entrez Gene ID                         TRUE
81   hg19 geneSymbol     Gene Symbol                         TRUE
   GO Annotation Available
25                    TRUE
48                    TRUE
81                    TRUE
```

The gene ID codes are listed in the column second from left, titled "Id". We find that our gene ID code is "ensGene". We will use these strings whenever we are asked for a genome or id. If the gene ID is missing for your Genome (for example this is the case for hg38), then the genome is not supported in the geneLengthDatabase. Gene lengths will either be automatically fetched from TxDB, UCSC or you will need to provide them manually. Supported Gene IDs to automatically fetch GO terms should usually either be Entrez ("knownGene"), Ensembl ("ensGene") or gene symbols ("geneSymbol").

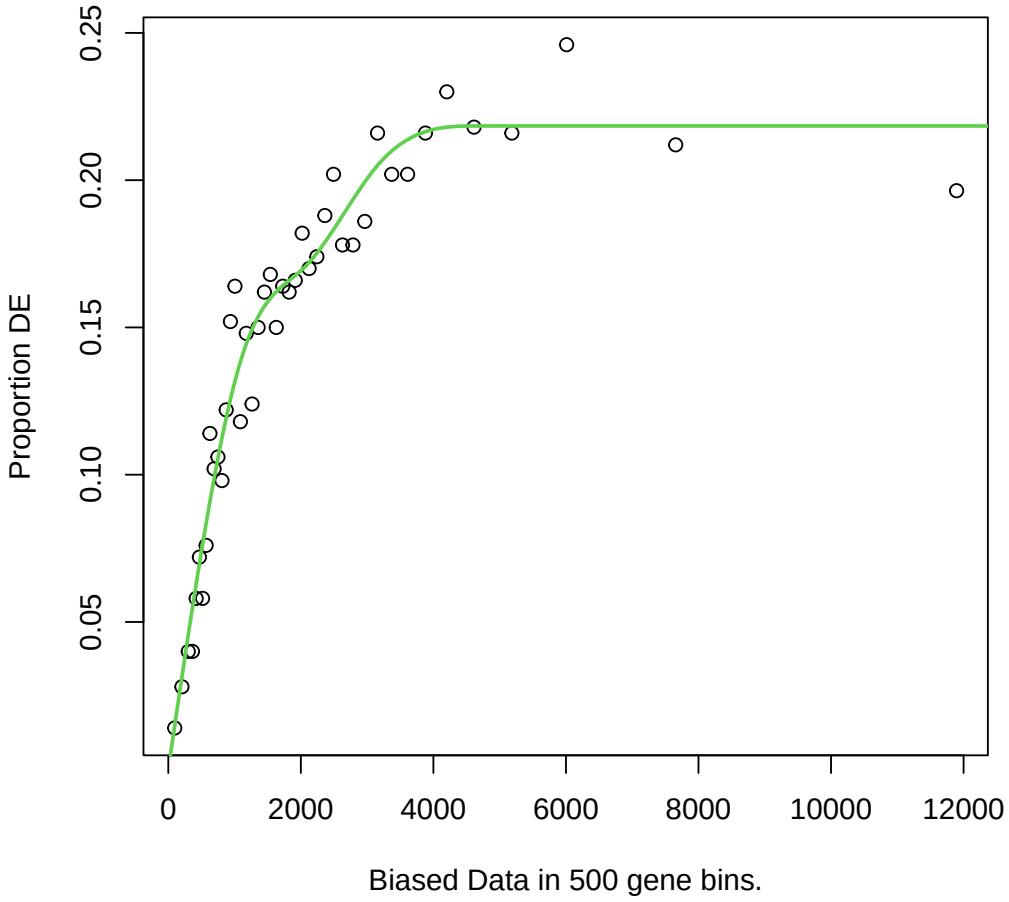
6.5 GO analysis

6.5.1 Fitting the Probability Weighting Function (PWF)

We first need to obtain a weighting for each gene, depending on its length, given by the PWF. As you may have noticed when running supportedGenomes or supportedGeneIDs, length data is available in the local database for our gene ID, "ensGene" and our genome, "hg19". We will let `goseq` automatically fetch this data from its databases.

```
> pwf <- nullp(genes, "hg19", "ensGene")
> head(pwf)
```

```
                DEgenes bias.data        pwf
ENSG00000230758       0       247 0.03757470
ENSG00000182463       0      3133 0.20436865
ENSG00000124208       0      1978 0.16881769
ENSG00000230753       0       466 0.06927243
ENSG00000224628       0      1510 0.15903532
ENSG00000125835       0       954 0.12711992
```



`nullp` plots the resulting fit, allowing verification of the goodness of fit before continuing the analysis. Further plotting of the pwf can be performed using the `plotPWF` function.

The output of nullp contains all the data used to create the PWF, as well as the PWF itself. It is a data frame with 3 columns, named "DEgenes", "bias.data" and "pwf" with the rownames set to the gene names. Each row corresponds to a gene with the DEgenes column specifying if the gene is DE (1 for DE, 0 for not DE), the bias.data column giving the numeric value of the DE bias

being accounted for (usually the gene length or number of counts) and the pwf column giving the genes value on the probability weighting function.

6.5.2 Using the Wallenius approximation

To start with we will use the default method, to calculate the over and under expressed GO categories among DE genes. Again, we allow `goseq` to fetch data automatically, except this time the data being fetched is the relationship between ENSEMBL gene IDs and GO categories.

```
> GO.wall <- goseq(pwf, "hg19", "ensGene")
> head(GO.wall)
```

```
        category over_represented_pvalue under_represented_pvalue numDEInCat
2475  GO:0005737            3.471602e-10                         1       2043
120   GO:0000278            1.110083e-08                         1        228
8929  GO:0035556            4.792590e-08                         1        573
2422  GO:0005615            5.156832e-08                         1        500
14005 GO:0070062            2.956709e-07                         1        388
10437 GO:0044281            2.982165e-07                         1        339
      numInCat                                term ontology
2475      9473                            cytoplasm       CC
120        796                   mitotic cell cycle       BP
8929      2288 intracellular signal transduction       BP
2422      2091                  extracellular space       CC
14005     1596                extracellular exosome       CC
10437     1377  small molecule metabolic process       BP
```

The resulting object is ordered by GO category over representation amongst DE genes.

6.5.3 Using random sampling

It may sometimes be desirable to use random sampling to generate the null distribution for category membership. For example, to check consistency against results from the Wallenius approximation. This is easily accomplished by using the `method` option to specify sampling and the `repcnt` option to specify the number of samples to generate:

```
> GO.samp <- goseq(pwf, "hg19", "ensGene", method = "Sampling", repcnt = 1000)
```

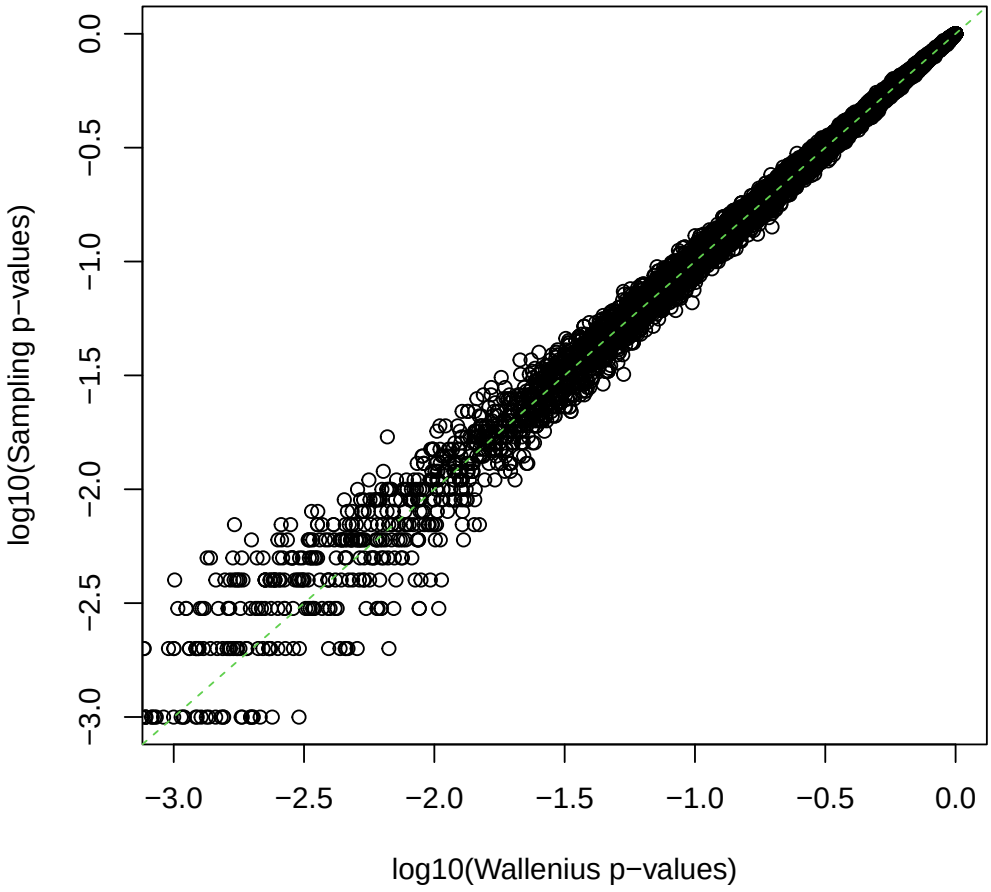
```
> head(GO.samp)
```

```
      category over_represented_pvalue under_represented_pvalue numDEInCat numInCat
31  GO:0000070            0.000999001                         1         59      181
99  GO:0000226            0.000999001                         1        155      582
120 GO:0000278            0.000999001                         1        228      796
```

```
121 GO:0000280             0.000999001                          1        109      375
241 GO:0000727             0.000999001                          1          8       12
275 GO:0000819             0.000999001                          1         69      219
                                                          term ontology
31                      mitotic sister chromatid segregation       BP
99                    microtubule cytoskeleton organization       BP
120                                     mitotic cell cycle       BP
121                                       nuclear division       BP
241 double-strand break repair via break-induced replication       BP
275                               sister chromatid segregation       BP
```

You will notice that this takes far longer than the Wallenius approximation. Plotting the p-values against one another, we see that there is little difference between the two methods. However, the accuracy of the sampling method is limited by the number of samples generated, `repcnt`, such that very low p-values will not be correctly calculated. Significantly enriched GO terms may then be missed after correcting for multiple testing.

```
> plot(log10(GO.wall[, 2]), log10(GO.samp[match(GO.wall[, 1], GO.samp[, 1]), 2]),
+   xlab = "log10(Wallenius p-values)", ylab = "log10(Sampling p-values)",
+   xlim = c(-3, 0)
+ )
> abline(0, 1, col = 3, lty = 2)
```



6.5.4 Ignoring length bias

`goseq` also allows for one to perform a GO analysis without correcting for RNA-seq length bias. In practice, this is only useful for assessing the effect of length bias on your results. You should NEVER use this option as your final analysis. If length bias is truly not present in your data, `goseq` will produce a nearly flat PWF and no length bias correction will be applied to your data and all methods will produce the same results.

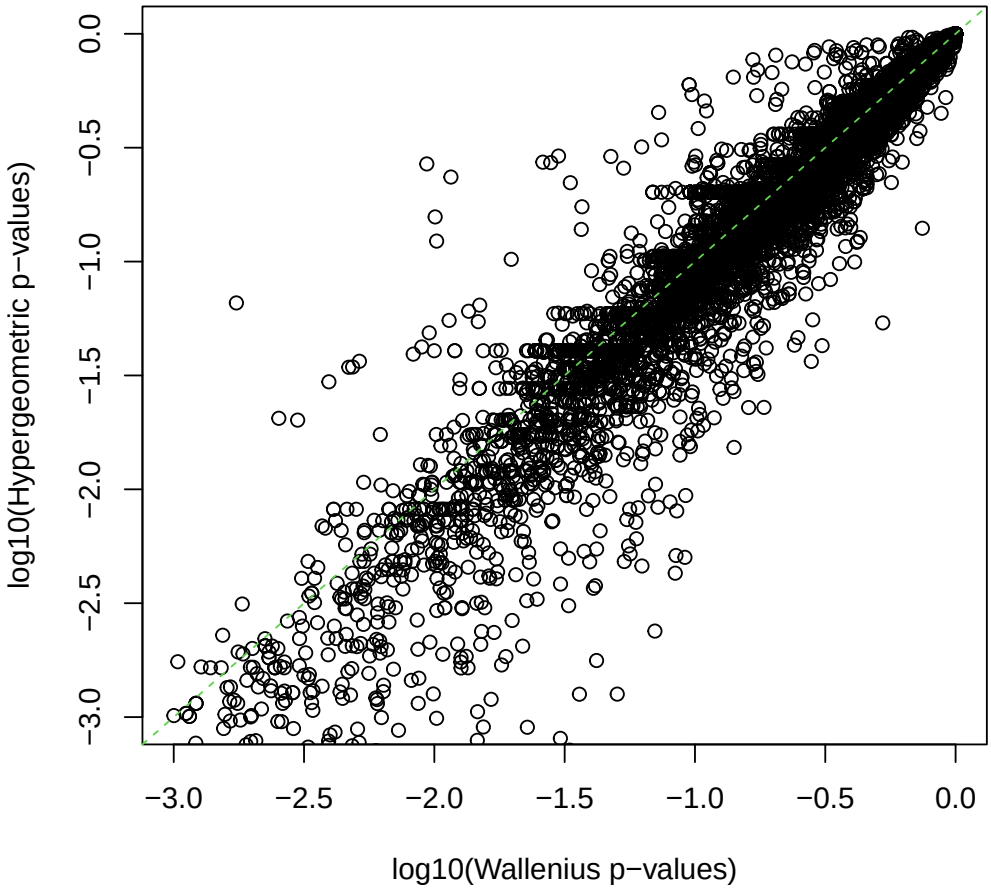
However, if you still wish to ignore length bias in calculating GO category enrichment, this is again accomplished using the `method` option.

```
> GO.nobias <- goseq(pwf, "hg19", "ensGene", method = "Hypergeometric")
> head(GO.nobias)
```

```
        category over_represented_pvalue under_represented_pvalue numDEInCat
2475  GO:0005737            8.895056e-11                         1       2043
8929  GO:0035556            2.142921e-10                         1        573
120   GO:0000278            2.173133e-09                         1        228
14959 GO:0071944            3.208516e-09                         1        867
6539  GO:0023051            6.270881e-09                         1        623
2568  GO:0005886            7.041510e-09                         1        806
      numInCat                              term ontology
2475      9473                          cytoplasm       CC
8929      2288 intracellular signal transduction       BP
120        796                mitotic cell cycle       BP
14959     3695                    cell periphery       CC
6539      2564           regulation of signaling       BP
2568      3420                  plasma membrane       CC
```

Ignoring length bias gives very different results from a length bias corrected analysis.

```
> plot(log10(GO.wall[, 2]), log10(GO.nobias[match(GO.wall[, 1], GO.nobias[, 1]), 2]),
+   xlab = "log10(Wallenius p-values)", ylab = "log10(Hypergeometric p-values)",
+   xlim = c(-3, 0), ylim = c(-3, 0)
+ )
> abline(0, 1, col = 3, lty = 2)
```



6.5.5 Limiting GO categories and other category based tests

By default, `goseq` tests all three major Gene Ontology branches; Cellular Components, Biological Processes and Molecular Functions. However, it is possible to limit testing to any combination of the major branches by using the `test.cats` argument to the goseq function. This is done by specifying a vector consisting of some combination of the strings "GO:CC", "GO:BP" and "GO:MF". For example, to test for only Molecular Function GO categories:

```
> GO.MF <- goseq(pwf, "hg19", "ensGene", test.cats = c("GO:MF"))
> head(GO.MF)
```

```
       category over_represented_pvalue under_represented_pvalue numDEInCat
1165 GO:0008092            1.430312e-05                 0.9999904        225
1003 GO:0005198            1.608456e-05                 0.9999903        135
```

```
1124 GO:0005515            2.405674e-05               0.9999816       2303
225  GO:0003735            6.731411e-05               0.9999703         40
1151 GO:0008017            8.066795e-05               0.9999554         79
2251 GO:0030215            1.578777e-04               0.9999790         11
     numInCat                                term ontology
1165      825       cytoskeletal protein binding       MF
1003      521       structural molecule activity       MF
1124    11000                    protein binding       MF
225       153 structural constituent of ribosome       MF
1151      247                microtubule binding       MF
2251       17        semaphorin receptor binding       MF
```

Native support for other category tests, such as KEGG pathway analysis are also made available via this argument. See the man `goseq` function man page for up to date information on what category tests are natively supported.

6.5.6 Making sense of the results

Having performed the GO analysis, you may now wish to interpret the results. If you wish to identify categories significantly enriched/unenriched below some p-value cutoff, it is necessary to first apply some kind of multiple hypothesis testing correction. For example, GO categories over enriched using a .05 FDR cutoff [Benjamini and Hochberg, 1995] are:

```
> enriched.GO <- GO.wall$category[p.adjust(GO.wall$over_represented_pvalue,
+   method = "BH"
+ ) < .05]
> head(enriched.GO)

[1] "GO:0005737" "GO:0000278" "GO:0035556" "GO:0005615" "GO:0070062" "GO:0044281"
```

Unless you are a machine, GO accession identifiers are probably not very meaningful to you. Information about each term can be obtained from the Gene Ontology website, `http://www.geneontology.org/`, or using the R package `GO.db`.

```
> library(GO.db)
> for (go in enriched.GO[1:10]) {
+   print(GOTERM[[go]])
+   cat("--------------------------------------\n")
+ }

GOID: GO:0005737
Term: cytoplasm
Ontology: CC
```

```
Definition: The contents of a cell excluding the plasma membrane and
    nucleus, but including other subcellular structures.
-------------------------------------
GOID: GO:0000278
Term: mitotic cell cycle
Ontology: BP
Definition: Progression through the phases of the mitotic cell cycle, the
    most common eukaryotic cell cycle, which canonically comprises four
    successive phases called G1, S, G2, and M and includes replication of
    the genome and the subsequent segregation of chromosomes into daughter
    cells. In some variant cell cycles nuclear replication or nuclear
    division may not be followed by cell division, or G1 and G2 phases may
    be absent.
Synonym: GO:0007067
Synonym: mitosis
Secondary: GO:0007067
-------------------------------------
GOID: GO:0035556
Term: intracellular signal transduction
Ontology: BP
Definition: The process in which a signal is passed on to downstream
    components within the cell, which become activated themselves to
    further propagate the signal and finally trigger a change in the
    function or state of the cell.
Synonym: GO:0007242
Synonym: GO:0007243
Synonym: GO:0023013
Synonym: GO:0023034
Synonym: intracellular signaling cascade
Synonym: intracellular signaling pathway
Synonym: intracellular signal transduction pathway
Synonym: signal transmission via intracellular cascade
Secondary: GO:0007242
Secondary: GO:0007243
Secondary: GO:0023013
Secondary: GO:0023034
-------------------------------------
GOID: GO:0005615
Term: extracellular space
Ontology: CC
Definition: That part of a multicellular organism outside the cells
    proper, usually taken to be outside the plasma membranes, and occupied
```

```
    by fluid.
Synonym: intercellular space
-------------------------------------
GOID: GO:0070062
Term: extracellular exosome
Ontology: CC
Definition: A vesicle that is released into the extracellular region by
    fusion of the limiting endosomal membrane of a multivesicular body
    with the plasma membrane. Extracellular exosomes, also simply called
    exosomes, have a diameter of about 40-100 nm.
Synonym: exosome
Synonym: extracellular vesicular exosome
-------------------------------------
GOID: GO:0044281
Term: small molecule metabolic process
Ontology: BP
Definition: The chemical reactions and pathways involving small molecules,
    any low molecular weight, monomeric, non-encoded molecule.
Synonym: small molecule metabolism
-------------------------------------
GOID: GO:0043230
Term: extracellular organelle
Ontology: CC
Definition: Organized structure of distinctive morphology and function,
    occurring outside the cell. Includes, for example, extracellular
    membrane vesicles (EMVs) and the cellulosomes of anaerobic bacteria
    and fungi.
-------------------------------------
GOID: GO:0065010
Term: extracellular membrane-bounded organelle
Ontology: CC
Definition: Organized structure of distinctive morphology and function,
    bounded by a lipid bilayer membrane and occurring outside the cell.
Synonym: extracellular membrane-enclosed organelle
-------------------------------------
GOID: GO:1903561
Term: extracellular vesicle
Ontology: CC
Definition: NA
Synonym: microparticle
-------------------------------------
GOID: GO:0005576
```

```
Term: extracellular region
Ontology: CC
Definition: The space external to the outermost structure of a cell. For
    cells without external protective or external encapsulating structures
    this refers to space outside of the plasma membrane. This term covers
    the host cell environment outside an intracellular parasite.
Synonym: extracellular
-------------------------------------
```

6.5.7 Understanding goseq internals

The situation may arise where it is necessary for the user to perform some of the data processing steps usually performed automatically by `goseq` themselves. With this in mind, it will be useful to step through the preprocessing steps performed automatically by `goseq` to understand what is happening.

To start with, when `nullp` is called, `goseq` uses the genome and gene identifiers supplied to try and retrieve length information for all genes given to the `genes` argument. To do this, it retrieves the data from the database of gene lengths maintained in the package `geneLenDataBase`. This is performed by the `getlength` function in the following way:

```
> len <- getlength(names(genes), "hg19", "ensGene")
> length(len)

[1] 22743

> length(genes)

[1] 22743

> head(len)

[1]  247 3133 1978  466 1510  954
```

After some data cleanup, the length data and the DE data is then passed to the `makespline` function to produce the PWF. The `nullp` returns a data frame which has 3 columns, the original DEgenes vector, the length bias data (in a column called bias.data) and the PWF itself (in a column named pwf). The names of the genes are also kept in this data frame as the names of the rows. If length data could not be obtained for a certain gene the corresponding entries in the "bias.data" and "pwf" columns are set to NA.

Next we call the `goseq` function to determine over/under representation of GO categories amongst DE genes. When we do this, `goseq` looks for the appropriate organism package and tries to obtain the mapping from genes to GO categories from it. This is done using the `getgo` function as follows:

```
> go <- getgo(names(genes), "hg19", "ensGene")
> length(go)

[1] 22743

> length(genes)

[1] 22743

> head(go)

$<NA>
NULL

$ENSG00000182463
 [1] "GO:0006139" "GO:0006351" "GO:0006355" "GO:0006357" "GO:0006366" "GO:0008150"
 [7] "GO:0008152" "GO:0009058" "GO:0009059" "GO:0009889" "GO:0009987" "GO:0010467"
[13] "GO:0010468" "GO:0010556" "GO:0016070" "GO:0019219" "GO:0019222" "GO:0032774"
[19] "GO:0034654" "GO:0043170" "GO:0044238" "GO:0050789" "GO:0050794" "GO:0051252"
[25] "GO:0060255" "GO:0065007" "GO:0080090" "GO:0090304" "GO:0141187" "GO:2001141"
[31] "GO:0000785" "GO:0005575" "GO:0005622" "GO:0005634" "GO:0005694" "GO:0043226"
[37] "GO:0043227" "GO:0043228" "GO:0043229" "GO:0043231" "GO:0043232" "GO:0110165"
[43] "GO:0000981" "GO:0003674" "GO:0003676" "GO:0003677" "GO:0003700" "GO:0005488"
[49] "GO:0005515" "GO:0036094" "GO:0043167" "GO:0043169" "GO:0046872" "GO:0140110"

$<NA>
NULL

$<NA>
NULL

$<NA>
NULL

$ENSG00000125835
  [1] "GO:0000245" "GO:0000375" "GO:0000377" "GO:0000387" "GO:0000398" "GO:0000966"
  [7] "GO:0001510" "GO:0006139" "GO:0006396" "GO:0006397" "GO:0006479" "GO:0008150"
 [13] "GO:0008152" "GO:0008213" "GO:0008380" "GO:0009058" "GO:0009059" "GO:0009451"
 [19] "GO:0009987" "GO:0010467" "GO:0016043" "GO:0016070" "GO:0016071" "GO:0019538"
 [25] "GO:0022607" "GO:0022613" "GO:0022618" "GO:0032259" "GO:0032774" "GO:0034654"
 [31] "GO:0036211" "GO:0036260" "GO:0036261" "GO:0043170" "GO:0043412" "GO:0043414"
 [37] "GO:0043933" "GO:0044085" "GO:0044238" "GO:0065003" "GO:0071826" "GO:0071840"
 [43] "GO:0090304" "GO:0141187" "GO:1903241" "GO:0005575" "GO:0005622" "GO:0005634"
```

```
 [49] "GO:0005654" "GO:0005681" "GO:0005682" "GO:0005683" "GO:0005684" "GO:0005685"
 [55] "GO:0005686" "GO:0005687" "GO:0005689" "GO:0005697" "GO:0005737" "GO:0005829"
 [61] "GO:0030532" "GO:0031974" "GO:0031981" "GO:0032991" "GO:0034708" "GO:0034709"
 [67] "GO:0034719" "GO:0043226" "GO:0043227" "GO:0043229" "GO:0043231" "GO:0043233"
 [73] "GO:0046540" "GO:0070013" "GO:0071004" "GO:0071005" "GO:0071007" "GO:0071010"
 [79] "GO:0071011" "GO:0071013" "GO:0071204" "GO:0097525" "GO:0097526" "GO:0110165"
 [85] "GO:0120114" "GO:0140513" "GO:1902494" "GO:1990234" "GO:1990904" "GO:0003674"
 [91] "GO:0003676" "GO:0003723" "GO:0005488" "GO:0005515" "GO:0043021" "GO:0044877"
 [97] "GO:0070034" "GO:0070990" "GO:0071208" "GO:1990446" "GO:1990447"
```

Note that some of the gene categories have been returned as "NULL". This means that a GO category could not be found in the database for one of the genes. In the `goseq` command, enrichment will only be calculated using genes with a GO category by default. However, in older versions of goseq (below 1.15.2), we counted all genes. i.e. genes with no categories still counted towards the total number of gene outside of any single category. It is possible to switch between these two behaviors using the `use_genes_without_cat` flag in `goseq`.

The first thing the getgo function does is to convert the UCSC genome/ID namings into the naming convention used by the organism packages. This is done using two hard coded conversion vectors that are included in the `goseq` package but usually hidden from the user.

```
> goseq:::.ID_MAP
```

```
 knownGene    refGene    ensGene geneSymbol        sgd     plasmo       tair
      "eg"       "eg"  "ENSEMBL"   "SYMBOL"      "sgd"   "plasmo"     "tair"
```

```
> goseq:::.ORG_PACKAGES
```

```
          anoGam      Arabidopsis           bosTau               ce
     "org.Ag.eg"    "org.At.tair"      "org.Bt.eg"      "org.Ce.eg"
          canFam               dm           danRer      E. coli K12
     "org.Cf.eg"      "org.Dm.eg"      "org.Dr.eg"   "org.EcK12.eg"
   E. coli Sakai           galGal               hg               mm
"org.EcSakai.eg"      "org.Gg.eg"      "org.Hs.eg"      "org.Mm.eg"
          rheMac          Malaria           panTro               rn
    "org.Mmu.eg"  "org.Pf.plasmo"      "org.Pt.eg"      "org.Rn.eg"
          sacCer           susScr           xenTro
    "org.Sc.sgd"      "org.Ss.eg"      "org.Xl.eg"
```

It is just as valid to run the length and GO category fetching as separate steps and then pass the result to the `nullp` and `goseq` functions using the `bias.data` and `gene2cat` arguments. Thus the following two blocks of code are equivalent:

```
> pwf <- nullp(genes, "hg19", "ensGene")
> go <- goseq(pwf, "hg19", "ensGene")
```

and

```
> gene_lengths <- getlength(names(genes), "hg19", "ensGene")
> pwf <- nullp(genes, bias.data = gene_lengths)
> go_map <- getgo(names(genes), "hg19", "ensGene")
> go <- goseq(pwf, "hg19", "ensGene", gene2cat = go_map)
```

6.6 KEGG pathway analysis

In order to illustrate performing a category test not present in the goseq database, we perform a KEGG pathway analysis. For human, the mapping from KEGG pathways to genes are stored in the package org.Hs.eg.db, in the object org.Hs.egPATH. In order to test for KEGG pathway over representation amongst DE genes, we need to extract this information and put it in a format that goseq understands. Unfortunately, the org.Hs.eg.db package does not contain direct mappings between ENSEMBL gene ID and KEGG pathway. Therefore, we have to construct this map by combining the ENSEMBL <-> Entrez and Entrez <-> KEGG mappings. This can be done using the following code:

```
> # Get the mapping from ENSEMBL 2 Entrez
> en2eg <- as.list(org.Hs.egENSEMBL2EG)
> # Get the mapping from Entrez 2 KEGG
> eg2kegg <- as.list(org.Hs.egPATH)
> # Define a function which gets all unique KEGG IDs
> # associated with a set of Entrez IDs
> grepKEGG <- function(id, mapkeys) {
+   unique(unlist(mapkeys[id], use.names = FALSE))
+ }
> # Apply this function to every entry in the mapping from
> # ENSEMBL 2 Entrez to combine the two maps
> kegg <- lapply(en2eg, grepKEGG, eg2kegg)
> head(kegg)
```

Note that this step is quite time consuming. The code written here is not the most efficient way of producing this result, but the logic is much clearer than faster algorithms. The source code for getgo contains a more efficient routine.

We produce the PWF as before. Then, to perform a KEGG analysis, we simply make use of the gene2cat option in goseq:

```
> pwf <- nullp(genes, "hg19", "ensGene")
> KEGG <- goseq(pwf, gene2cat = kegg)
> head(KEGG)
```

Note that we do not have to tell the goseq function what organism and gene ID we are using as we are manually supplying the mapping between genes and categories.

KEGG analysis is shown as an illustration of how to supply your own mapping between gene ID and category, KEGG analysis is actually natively supported by GOseq and we could have performed it with the following code.

```
> pwf <- nullp(genes, "hg19", "ensGene")
> KEGG <- goseq(pwf, "hg19", "ensGene", test.cats = "KEGG")
> head(KEGG)
```

```
    category over_represented_pvalue under_represented_pvalue numDEInCat numInCat
88     03010            6.317053e-06                0.9999980         29       87
77     00900            2.391514e-04                0.9999710         10       15
113    04115            8.164693e-04                0.9996835         26       64
175    04964            2.150726e-03                0.9995925         10       17
27     00330            3.668792e-03                0.9986594         18       44
20     00250            5.200138e-03                0.9984357         13       28
```

Noting that this time it was necessary to tell the goseq function that we are using HG19 and ENSEMBL gene ID, as the function needs this information to automatically construct the mapping from geneid to KEGG pathway.

6.7 Extracting mappings from organism packages

If you know that the information mapping gene ID to your categories of interest is contained in the organism packages, but `goseq` fails to fetch it automatically, you may want to extract it yourself and then pass it to the `goseq` function using the `gene2cat` argument. This is done in exactly the same way as extracting the KEGG to ENSEMBL mappings in the section "KEGG pathway analysis" above. This example is actually the worst case, where it is necessary to combine two mappings to get the desired list. If we had instead wanted the association between Entrez gene IDs and KEGG pathways, the following code would have been sufficient:

```
> kegg <- as.list(org.Hs.egPATH)
> head(kegg)
```

```
$`1`
[1] NA

$`2`
[1] "04610"

$`3`
```

```
[1] NA

$`9`
[1] "00232" "00983" "01100"

$`10`
[1] "00232" "00983" "01100"

$`11`
[1] NA
```

A note on fetching GO mappings from the organism. The data structure of GO is a directed acyclic graph. This means that in addition to each GO category being associated with a set of genes, it may also have children that are associated to other genes. It is important to use the org.Hs.egGO2ALLEGS and NOT the org.Hs.egGO object to create the mapping between GO categories and gene identifiers, as the latter does not include the links to genes arising from "child" GO categories. Thank you to Christopher Fjell for pointing this out.

6.8 Correcting for other biases

It is possible that in some circumstances you will wish to correct not just for length bias, but for the total number of counts. This can make sense because power to detect DE depends on the total number of counts a gene receives, which is the product of gene length and gene expression. So correcting for read count bias will compensate for all biases, known and unknown, in power to detect DE. On the other hand, it will also remove bias resulting from differences in expression level, which may not be desirable.

Correcting for count bias will produce a different PWF. Therefore, we need to tell `goseq` about the data on which the fraction DE depends when calculating the PWF using the `nullp` function. We then simply pass the result to `goseq` as usual.

So, in order to tell `goseq` to correct for read count bias instead of length bias, all you need to do is supply a numeric vector, containing the number of counts for each gene to `nullp`.

```
> countbias <- rowSums(counts)[rowSums(counts) != 0]
> length(countbias)

[1] 22743

> length(genes)

[1] 22743
```

To use the count bias when doing GO analysis, simply pass this vector to `nullp` using the `bias.data` option. Note that we have to supply "hg19" and "ensGene" to `goseq` as it is not used by nullp and hence not in the pwf.counts object.

```
> pwf.counts <- nullp(genes, bias.data = countbias)
> GO.counts <- goseq(pwf.counts, "hg19", "ensGene")
> head(GO.counts)
```

```
        category over_represented_pvalue under_represented_pvalue numDEInCat
14959 GO:0071944            3.196498e-16                         1        867
2568  GO:0005886            1.197049e-14                         1        806
6540  GO:0023052            3.635770e-09                         1        982
3306  GO:0007154            7.826356e-09                         1        981
3659  GO:0008283            1.963497e-08                         1        349
11961 GO:0048856            3.516857e-08                         1        960
      numInCat                              term ontology
14959     3695                   cell periphery       CC
2568      3420                  plasma membrane       CC
6540      4270                        signaling       BP
3306      4284                cell communication       BP
3659      1357    cell population proliferation       BP
11961     4172 anatomical structure development       BP
```

Note that if you want to correct for length bias, but your organism/gene identifier is not natively supported, then you need to follow the same procedure as above, only the numeric vector supplied will contain each gene's length instead of its number of reads.

7 Setup

This vignette was built on:

```
> sessionInfo()
```

```
R version 4.5.1 Patched (2025-08-23 r88802)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 24.04.3 LTS

Matrix products: default
BLAS:   /home/biocbuild/bbs-3.22-bioc/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.12.0  LAPACK version 3.12.0

locale:
```

```
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_GB              LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

time zone: America/New_York
tzcode source: system (glibc)

attached base packages:
[1] stats4    stats     graphics  grDevices utils     datasets  methods   base

other attached packages:
 [1] GO.db_3.21.0           org.Hs.eg.db_3.21.0    AnnotationDbi_1.71.1
 [4] IRanges_2.43.5         S4Vectors_0.47.4       Biobase_2.69.1
 [7] BiocGenerics_0.55.1    generics_0.1.4         edgeR_4.7.5
[10] limma_3.65.5           goseq_1.61.1           geneLenDataBase_1.45.0
[13] BiasedUrn_2.0.12

loaded via a namespace (and not attached):
 [1] KEGGREST_1.49.1             SummarizedExperiment_1.39.2
 [3] rjson_0.2.23                httr2_1.2.1
 [5] lattice_0.22-7              vctrs_0.6.5
 [7] tools_4.5.1                 bitops_1.0-9
 [9] curl_7.0.0                  parallel_4.5.1
[11] tibble_3.3.0                RSQLite_2.4.3
[13] blob_1.2.4                  pkgconfig_2.0.3
[15] Matrix_1.7-4                dbplyr_2.5.1
[17] lifecycle_1.0.4             stringr_1.5.2
[19] compiler_4.5.1              progress_1.2.3
[21] Rsamtools_2.25.3            Biostrings_2.77.2
[23] statmod_1.5.0               Seqinfo_0.99.2
[25] codetools_0.2-20            GenomeInfoDb_1.45.12
[27] RCurl_1.98-1.17             yaml_2.3.10
[29] pillar_1.11.1               crayon_1.5.3
[31] BiocParallel_1.43.4         DelayedArray_0.35.3
[33] cachem_1.1.0                abind_1.4-8
[35] nlme_3.1-168                locfit_1.5-9.12
[37] tidyselect_1.2.1            stringi_1.8.7
[39] dplyr_1.1.4                 restfulr_0.0.16
[41] splines_4.5.1               biomaRt_2.65.16
```

```
[43] fastmap_1.2.0              grid_4.5.1
[45] cli_3.6.5                  SparseArray_1.9.1
[47] magrittr_2.0.4             S4Arrays_1.9.1
[49] GenomicFeatures_1.61.6     XML_3.99-0.19
[51] prettyunits_1.2.0          filelock_1.0.3
[53] UCSC.utils_1.5.0           rappdirs_0.3.3
[55] bit64_4.6.0-1              XVector_0.49.1
[57] httr_1.4.7                 matrixStats_1.5.0
[59] bit_4.6.0                  hms_1.1.3
[61] png_0.1-8                  memoise_2.0.1
[63] GenomicRanges_1.61.5       BiocIO_1.19.0
[65] BiocFileCache_2.99.6       mgcv_1.9-3
[67] rtracklayer_1.69.1         txdbmaker_1.5.6
[69] rlang_1.1.6                glue_1.8.0
[71] DBI_1.2.3                  jsonlite_2.0.0
[73] R6_2.6.1                   MatrixGenerics_1.21.0
[75] GenomicAlignments_1.45.4
```

8 Acknowledgments

Christopher Fjell for a series of bug fixes and pointing out the difference between the egGO and egGO2ALLEGS objects in the organism packages.

References

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57:289–300, 1995.

Hairi Li, Michael T Lovci, Young-Soo Kwon, Michael G Rosenfeld, Xiang-Dong Fu, and Gene W Yeo. Determination of tag density required for digital transcriptome analysis: application to an androgen-sensitive prostate cancer model. *Proc Natl Acad Sci USA*, 105(51):20179–84, Dec 2008. doi: 10.1073/pnas.0807121105.

Alicia Oshlack and Matthew J Wakefield. Transcript length bias in RNA-seq data confounds systems biology. *Biol Direct*, 4:14, Jan 2009. doi: 10.1186/1745-6150-4-14.

M. D. Robinson and G. K. Smyth. Moderated statistical tests for assessing differences in tag abundance. *Bioinformatics*, 23(21):2881–2887, 2007.

M. D. Robinson and G. K. Smyth. Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332, 2008.

M. D. Robinson, D. J. McCarthy, and G. K. Smyth. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–40, Jan 2010. doi: 10.1093/bioinformatics/btp616.

M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11:R14, Feb 2010.