seq2pathway.data Vignette

Bin Wang

December 3, 2014

Contents

1 Abstract

`Seq2pathway.data` is supporting data for the `seq2pathway` package. The package includes pre-defined gene sets which are constructed from `R` package `org.Hs.eg.db`[1] for GeneOntology and the Molecular Signatures Database (MsigDB)[2] for other functional gene sets. The gene locus definitions in the package is built from the `GENCODE project`[3], including `GENCODE 20` (hg38), `GENCODE 19` (hg19), `GENCODE mmvM3` (mm10) and `GENCODE mmvM1` (mm9) currently.

2 Data

We grouped all data to nine categories. The data 2.1 to 2.5 is aimed to internal funtion use only. The data 2.6 to 2.9 is demo data for the `seq2pathway` package.

2.1 GO_BP_list; GO_MF_list; GO_CC_list

These data contains all gene symbol lists extracted from an R object `org.Hs.egGO2EG`. Note that `org.Hs.egGO2EG` "provides mappings between entrez gene identifiers and the GO identifiers that they are directly associated with. This mapping and its reverse mapping do NOT associate the child terms from the GO ontology with the gene. Only the directly evidenced terms are represented here."[1]

The list GO_BP_list includes 9407 GO biological process terms, and the names of list stand for the names of GO terms. The corresponding element of each list is the gene symbol of each term. Similarly, the list GO_MF_list includes 3529 GO molecular function terms, and the list GO_CC_list includes 1198 GO cell component terms.

Data in this category will mainly be invoked internally by the functions in the seq2pathway package.

2.2 Des_BP_list; Des_MF_list; Des_CC_list

These three lists include the description information of each GO terms. The name of each list stands for the name of GO terms. The corresponding element of each list is the description of each term. The desctiption is extracted from R object GO.db[4]. The length of the list Des_BP_list is equal to the length of the list GO_BP_list. The similiar pattern applies to the other two lists.

Data in this category will mainly be invoked internally by the functions in the seq2pathway package.

2.3 GO_GENCODE_df_hg_v19; GO_GENCODE_df_hg_v20; GO_GENCODE_df_mm_vM1; GO_GENCODE_df_mm_vM3

The gene set varies from different gene databese sources. The gene symbols of GO terms are from org.Hs.egGO2EG[1], the gene set in our annoation package is from the GENCODE[3] data sets, which attribute gene set by differenct organisms and assembly versions. These data frames record the common genes from different databases. For example, the GO_GENCODE_df_hg_v19 data frame records the common genes from org.Hs.egGO2EG and GENCODE human species 19 version. In summary, there are 17998 genes in GO_GENCODE_df_hg_v19, 18015 genes in GO_GENCODE_df_hg_v20, 14328 genes in GO_GENCODE_df_mm_vM1, and 15075 genes in GO_GENCODE_df_mm_vM3 respectively.

Data in this category will mainly be invoked internally by the functions in the seq2pathway package.

2.4 Msig_GENCODE_df_hg_v19; Msig_GENCODE_df_hg_v20; Msig_GENCODE_df_mm_vM1; Msig_GENCODE_df_mm_vM3

These four data frames record the common genes from Molecular Signatures Database (MsigDB)[2] and the GENCODE[3] data sets. MsigDB is a collection of annotated gene sets. There are 22751 genes in Msig_GENCODE_df_hg_v19, 22721 genes in Msig_GENCODE_df_hg_v20, 15420 genes in Msig_GENCODE_df_mm_vM1, and 15528 genes in Msig_GENCODE_df_mm_vM3.

Data in this category will mainly be invoked internally by the functions in the seq2pathway package.

2.5 gencode_coding

gencode_coding is an R vector which collects all protein coding gene symbols from GENCODE[3] human version20. There are 19810 unique coding genes symbols in gencode_coding object.

Data gencode_coding will mainly be invoked internally by the functions in the seq2pathway package.

2.6 MsigDB_C5

MsigDB_C5 is a gene-sets collection from MsigDB[2] in GSA.genesets format, which could be open with R package GSA[5]. MsigDB_C5 consists of genes annotated by the GO terms. As an R object, MsigDB_C5 is a list of 3 elements. The first element includes multiple sub lists. Each sub list is a gene list for one gene set. The second element records the names of genesets, and the last element is the descriptions of genesets. Seq2pathway.data is a supporting data package for the seq2pathway package. MsigDB_C5 is used as demo data for functions in the seq2pathway package. More details could be found in the vignette of seq2pathway package.

```
> data(MsigDB_C5,package="seq2pathway.data")
> class(MsigDB_C5)
[1] "GSA.genesets"
> names(MsigDB_C5)
```

```
[1] "genesets"             "geneset.names"        "geneset.descriptions"
```

2.7 gene_description

gene_description is a data frame with two columns. The data give the gene descrption based on the biomaRt[6] package. gene_description is a demo data for the seq2pathway package. More details could be found at the Vignette (**5.1.4 Add description for genes**) of the seq2pathway package.

```
> data(gene_description,package="seq2pathway.data")
> head(gene_description)
```

hgnc_symbol	description
ABCD4	ATP-binding cassette, sub-family D (ALD), member 4 [Source:HGNC Symbol;Acc:68]
ABHD12B	abhydrolase domain containing 12B [Source:HGNC Symbol;Acc:19837]
ABHD4	abhydrolase domain containing 4 [Source:HGNC Symbol;Acc:20154]
ACIN1	apoptotic chromatin condensation inducer 1 [Source:HGNC Symbol;Acc:17066]
ACOT1	acyl-CoA thioesterase 1 [Source:HGNC Symbol;Acc:33128]
ACOT2	acyl-CoA thioesterase 2 [Source:HGNC Symbol;Acc:18431]

2.8 dat_gene2path_chip; dat_gene2path_RNA

dat_gene2path_chip and dat_gene2path_RNA are demo data for functions in the seq2pathway package. Each of them is an R list object with 2 elements. The first element gene2pathway_result.2 is a list of gene2pathway test result, and the second element gene2pathway_result.FET is a list of Fisher's exact test result. More details and usage could be found at Examples section in the vignette of seq2pathway package.

Small code for checking the data dat_gene2path_chip is provided below:

```
> data(dat_gene2path_chip,package="seq2pathway.data")
> names(dat_gene2path_chip)
[1] "gene2pathway_result.2"   "gene2pathway_result.FET"
> class(dat_gene2path_chip$gene2pathway_result.2)
[1] "list"
> names(dat_gene2path_chip$gene2pathway_result.2)
[1] "GO_BP" "GO_CC" "GO_MF"
> head(dat_gene2path_chip$gene2pathway_result.2$GO_BP)
```

	Des	peakscore pathscore _Normalized	peakscore pathscore _Pvalue	Intersect _Count	Intersect _gene
GO:0000082	The mitotic cell cycle transition by which a cell in G1 commits to S phase. The process begins with the build up of G1 cyclin-dependent kinase (G1 CDK), resulting in the activation of transcription of G1 cyclins. The process ends with the positive feedback of the G1 cyclins on the G1 CDK which commits the cell to S phase, in which DNA replication is initiated.	0.32017745	0.12	11	CDKN3 GPR132 MNAT1 POLE2 PSMA3 PSMA6 PSMB5 PSMC1 PSMC6 PSME1 PSME2
GO:0000086	The mitotic cell cycle transition by which a cell in G2 commits to M phase. The process begins when the kinase activity of M cyclin/CDK complex reaches a threshold high enough for the cell cycle to proceed. This is accomplished by activating a positive feedback loop that results in the accumulation of unphosphorylated and active M cyclin/CDK complex.	-0.33586010	0.49	5	AJUBA DYNC1H1 HSP90AA1 LIN52 MNAT1
GO:0000122	Any process that stops, prevents, or reduces the frequency, rate or extent of transcription from an RNA polymerase II promoter.	-0.11535853	0.16	20	AJUBA BMP4 DACT1 DICER1 ESR2 FOXA1 GSC JDP2 NKX2-1 PPM1A PRMT5 PSEN1 RCOR1 SALL2 SIX1 SNW1 STRN3 YY1 ZBTB1 ZBTB42
GO:0000209	Addition of multiple ubiquitin groups to a protein, forming a ubiquitin chain.	0.17070465	0.11	11	ASB2 G2E3 PSMA3 PSMA6 PSMB11 PSMB5 PSMC1 PSMC6 PSME1 PSME2 RNF31
GO:0000278	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.	0.06368249	0.04	16	AJUBA DYNC1H1 HSP90AA1 LIN52 MNAT1 NEK9 POLE2 PSMA3 PSMA6 PSMB11 PSMB5 PSMC1 PSMC6 PSME1 PSME2 VRK1
GO:0000398	The joining together of exons from one or more primary transcripts of messenger RNA (mRNA) and the excision of intron sequences, via a spliceosomal mechanism, so that mRNA consisting only of the joined exons is produced.	-0.55767621	0.59	8	CPSF2 HNRNPC NOVA1 PABPN1 PAPOLA PNN SNW1 SRSF5

2.9 dat_seq2pathway_GOterms; dat_seq2pathway_Msig

`dat_seq2pathway_GOterms` and `dat_seq2pathway_Msig` are demo data for functions in the `seq2pathway` package. Each of them is an R list object with 3 elements. The first element `seq2gene_result` is a list with annotaton tables. The seconde element `gene2pathway_result.FAIME` is a list of `gene2pathway FAIME`[7] test result. And the third element `gene2pathway_result.FET` is a list of Fisher's exact test results. More details and usage could be found in the `Examples` section in the vignette of the `seq2pathway` package.

```
> data(dat_seq2pathway_Msig,package="seq2pathway.data")
> names(dat_seq2pathway_Msig)
[1] "seq2gene_result"   "gene2pathway_result.FAIME"    "gene2pathway_result.FET"
> class(dat_seq2pathway_Msig$seq2gene_result)
[1] "list"
> names(dat_seq2pathway_Msig$seq2gene_result)
[1] "seq2gene_FullResult"           "seq2gene_CodingGeneOnlyResult"
> class(dat_seq2pathway_Msig$gene2pathway_result.FAIME)
[1] "data.frame"
> class(dat_seq2pathway_Msig$gene2pathway_result.FET)
[1] "data.frame"
```

References

[1] Carlson M., *org.Hs.eg.db: Genome wide annotation for Human.*, R package version 3.0.0.

[2] Subramanian, Tamayo, et al. , *Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles*, PNAS **102** (2005), 1545–1550.

[3] Harrow J, et al., *GENCODE: The reference human genome annotation for The ENCODE Project*, Genome Res. **9** (2012), 1760–1774.

[4] Carlson M., *GO.db: A set of annotation maps describing the entire Gene Ontology.*, R package version 3.0.0.

[5] Efron, B. and Tibshirani, *R. On testing the significance of sets of genes.*, Stanford tech report rep 2006.

[6] Durinck S, Spellman P, Birney E and Huber W, *Mapping identifiers for the integration of genomic datasets with the R/Bioconductor package biomaRt*, Nature Protocols **4** (2009), 1184–1192.

[7] X. Yang, K. Regan, Y. Huang, Q. Zhang, J. Li, T. Y. Seiwert, et al., *Single sample expression-anchored mechanisms predict survival in head and neck cancer*, PLoS Comput Biol **8** (2012), e1002350.