

hpar: The Human Protein Atlas in *R*

Laurent Gatto*

May 15, 2016

Abstract

The Human Protein Atlas (HPA) is a systematic study of the human proteome using antibody-based proteomics. Multiple tissues and cell lines are systematically assayed with affinity-purified antibodies and confocal microscopy. The *hpar* package is an *R* interface to the HPA project. It distributes three data sets, provides functionality to query these and to access detailed information pages, including confocal microscopy images available on the HPA web page.

Keywords: infrastructure, bioinformatics, proteomics, microscopy

Contents

*<http://cpu.sysbiol.cam.ac.uk>

1 Introduction

1.1 The HPA project

From the Human Protein Atlas¹ [?, ?] site:

The Swedish Human Protein Atlas project, funded by the Knut and Alice Wallenberg Foundation, has been set up to allow for a systematic exploration of the human proteome using Antibody-Based Proteomics. This is accomplished by combining high-throughput generation of affinity-purified antibodies with protein profiling in a multitude of tissues and cells assembled in tissue microarrays. Confocal microscopy analysis using human cell lines is performed for more detailed protein localisation. The program hosts the Human Protein Atlas portal with expression profiles of human proteins in tissues and cells.

The *hpar* package provides functionality to use HPA data from the *R*interface. It also distributes three data sets available from the HPA site.

Normal tissue data Expression profiles for proteins in human tissues based on immunohistochemistry using tissue microarrays. The dataframe includes Ensembl gene identifier ("Gene"), tissue name ("Tissue"), annotated cell type ("Cell.type"), expression value ("Level"), the type of annotation (annotated protein expression (APE), based on more than one antibody, or staining, based on one antibody only) ("Expression.type"), and the reliability or validation of the expression value ("Reliability").

Subcellular location data Subcellular localisation of proteins based on immunofluorescently stained cells. The dataframe includes Ensembl gene identifier ("Gene"), main subcellular location of the protein ("Main.location"), other locations ("Other.location"), the type of annotation (annotated protein expression (APE), based on more than one antibody, or staining, based on one antibody only) ("Expression.type"), and the reliability or validation of the expression value ("Reliability").

RNA data RNA levels in three different cell lines, based on RNA-seq. The dataframe includes Ensembl gene identifier ("Gene"), analysed cell line ("Cell.line"), number of reads per kilobase gene model and million reads ("RPKM"), and abundance class ("Abundance").

1.2 HPA data usage policy

The use of data and images from the HPA in publications and presentations is permitted provided that the following conditions are met:

- The publication and/or presentation are solely for informational and non-commercial purposes.
- The source of the data and/or image is referred to the HPA site (www.proteinatlas.org) and/or one or more of our publications are cited.

1.3 Installation

hpar is available through the Bioconductor project. Details about the package and the installation procedure can be found on its page². To install using the dedicated Bioconductor infrastructure, run :

```
source("http://bioconductor.org/biocLite.R")
## or, if you have already used the above before
library("BiocInstaller") ## and to install the package
biocLite("hpar")
```

After installation, *hpar* will have to be explicitly loaded with

¹<http://www.proteinatlas.org/>

²<http://bioconductor.org/packages/devel/bioc/html/hpar.html>

```
library("hpar")

## This is hpar 1.14.2. For more information,
## please type '?hpar' or 'vignette('hpar')'.
```

so that all the package's functionality and data is available to the user.

2 The hpar package

2.1 Data sets

The three data sets, named `hpaNormalTissue`, `hpaSubcellularLoc` and `hpaRna` in the package can be loaded with the `data` function, as illustrated below for `hpaNormalTissue` below. Each data set is a dataframe and can be easily manipulated using standard R functionality. The code chunk below illustrates some of its properties.

```
data(hpaNormalTissue)
dim(hpaNormalTissue)

## [1] 1319440      6

names(hpaNormalTissue)

## [1] "Gene"          "Tissue"         "Cell.type"      "Level"
## [5] "Expression.type" "Reliability"

## Number of genes
length(unique(hpaNormalTissue$Gene))

## [1] 16613

## Number of cell types
length(unique(hpaNormalTissue$Cell.type))

## [1] 44

head(levels(hpaNormalTissue$Cell.type))

## [1] "adipocytes"          "bile duct cells"
## [3] "cells in endometrial stroma" "cells in glomeruli"
## [5] "cells in granular layer" "cells in molecular layer"

## Number of tissues
length(unique(hpaNormalTissue$Tissue))

## [1] 48

head(levels(hpaNormalTissue$Tissue))

## [1] "adrenal gland" "appendix"      "bone marrow"   "breast"        "bronchus"
## [6] "cerebellum"

table(hpaNormalTissue$Expression.type)

##
##      APE
## 1319440
```

2.2 HPA interface

The package provides a interface to the HPA data. The `getHpa` allows to query the data sets described in section ???. It takes three arguments, `id`, `hpadata` and `type`, that control the query, what data set to interrogate and how to report results respectively. The HPA data uses Ensembl gene identifiers and `id` must be a valid identifier. `hpadata` must be one of "NormalTissue", "Rna" or "SubcellularLoc". `type` can be data or details. The former is the default and returns a dataframe containing the information relevant to `id`. It is also possible to obtained detailed information, (including cell images) as web pages, directly from the HPA web page, using `details`.

We will illustrate this functionality with using the TSPAN6 (tetraspanin 6) gene (ENSG00000000003) as example.

```
id <- "ENSG00000000003"
head(getHpa(id, hpadata = "NormalTissue"))
```

##	Gene	Tissue	Cell.type	Level	Expression.type
## 1	ENSG00000000003	adrenal gland	glandular cells	Not detected	APE
## 2	ENSG00000000003	appendix	glandular cells	Medium	APE
## 3	ENSG00000000003	appendix	lymphoid tissue	Not detected	APE
## 4	ENSG00000000003	bone marrow	hematopoietic cells	Not detected	APE
## 5	ENSG00000000003	breast	adipocytes	Not detected	APE
## 6	ENSG00000000003	breast	glandular cells	High	APE

```
## Reliability
## 1 Supportive
## 2 Supportive
## 3 Supportive
## 4 Supportive
## 5 Supportive
## 6 Supportive

getHpa(id, hpadata = "SubcellularLoc")
```

##	Gene	Main.location	Other.location	Expression.type	Reliability
## 1	ENSG00000000003	Cytoplasm		APE	Uncertain

```
head(getHpa(id, hpadata = "Rna"))
```

##	Gene	Sample	Value	Unit	Abundance
## 1	ENSG00000000003	A-431	21.3	FPKM	Medium
## 2	ENSG00000000003	A549	32.5	FPKM	Medium
## 3	ENSG00000000003	AN3-CA	38.2	FPKM	Medium
## 4	ENSG00000000003	BEWD	31.4	FPKM	Medium
## 5	ENSG00000000003	CACO-2	63.9	FPKM	High
## 6	ENSG00000000003	CAPAN-2	34.2	FPKM	Medium

If we ask for detail, a browser page pointing to the relevant page is open (see figure ??)

```
getHpa(id, type = "details")
```

If a user is interested specifically in one data set, it is possible to set `hpadata` globally and omit it in `getHpa`. This is done by setting the `hpar` options `hpadata` with the `setHparOptions` function. The current default data set can be tested with `getHparOptions`.

```
getHparOptions()

## $hpar
## $hpar$hpadata
## [1] "NormalTissue"

setHparOptions(hpadata = "SubcellularLoc")
getHparOptions()
```

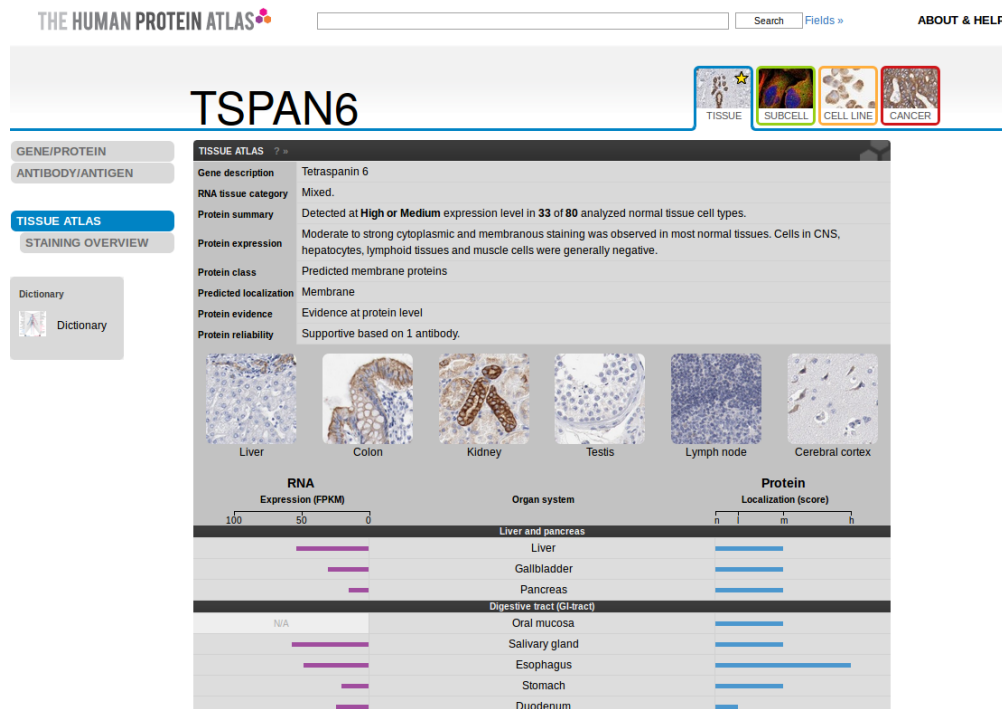


Figure 1: The HPA web page for the tetraspanin 6 gene (ENSG00000000003).

```
## $hpar
## $hpar$hpadata
## [1] "SubcellularLoc"

getHpa(id)

##           Gene Main.location Other.location Expression.type Reliability
## 1 ENSG00000000003      Cytoplasm                APE      Uncertain
```

2.3 HPA release information

Information about the HPA release used to build the installed *hpar* package can be accessed with `getHpaVersion`, `getHpaDate` and `getHpaEnsembl`. Full release details can be found on the HPA release history³ page.

```
getHpaVersion()

## [1] "Protein Atlas version 13"

getHpaDate()

## [1] "2014.11.06"

getHpaEnsembl()

## [1] "75.37"
```

³<http://www.proteinatlas.org/about/releases>

3 A small use case

Let's compare the subcellular localisation annotation obtained from the HPA subcellular location data set and the information available in the Bioconductor annotation packages.

```
id <- "ENSG00000001460"
getHpa(id, "SubcellularLoc")

##           Gene Main.location   Other.location Expression.type Reliability
## 6 ENSG00000001460      Nucleus Nuclear membrane           APE Supportive
```

Below, we first extract all cellular component GO terms available for ENSG00000001460 from the *org.Hs.eg.db* human annotation and then retrieve their term definitions using the *GO.db* database.

```
library(org.Hs.eg.db)
library(GO.db)
ans <- select(org.Hs.eg.db, keys = id,
              columns = c("ENSEMBL", "GO", "ONTOLOGY"),
              keytype = "ENSEMBL")

## 'select()' returned 1:many mapping between keys and columns

ans <- ans[ans$ONTOLOGY == "CC", ]
ans

##           ENSEMBL           GO EVIDENCE ONTOLOGY
## 1 ENSG00000001460 GO:0005634          IEA        CC
## 2 ENSG00000001460 GO:0005737          IEA        CC

sapply(as.list(GOTERM[ans$GO]), slot, "Term")

##  GO:0005634  GO:0005737
##  "nucleus"  "cytoplasm"
```

Session information

- R version 3.3.0 (2016-05-03), x86_64-apple-darwin13.4.0
- Locale: C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
- Base packages: base, datasets, grDevices, graphics, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.34.2, Biobase 2.32.0, BiocGenerics 0.18.0, GO.db 3.3.0, IRanges 2.6.0, S4Vectors 0.10.0, hpar 1.14.2, org.Hs.eg.db 3.3.0
- Loaded via a namespace (and not attached): BiocStyle 2.0.2, DBI 0.4-1, RSQLite 1.0.0, evaluate 0.9, formatR 1.4, highr 0.6, knitr 1.13, magrittr 1.5, stringi 1.0-1, stringr 1.0.0, tools 3.3.0