# LowMACAAnnotation: a serie of annotation tables for LowMACA package

Stefano de Pretis , Giorgio Melloni

February 23, 2015

## Contents



## 1 Introduction

The *LowMACAAnnotation* is composed by three interconnected datasets and relative functions to retrieve them. They are used by *LowMACA* package internal functions to properly map mutations from genes to protein sequences and finally to Pfam sequences in a unambiguous manner.

## 2 Functionalities

This package contains three simple functions to retrieve three manually curated datasets. The available data.frames are:

- `myUni` a datasets of proteins with their relative gene names and their Uniprot amino acid sequence
- `myPfam` a datasets of Pfam domains and their relative boundaries in protein sequences of myUni data.frame
- `myAlias` a dataset of official gene symbols and relative aliases used by *LowMACA* internal functions to check for correct user input

We consequently show how to retrieve the data and what is the content.

```
library(LowMACAAnnotation)
myUni <- getMyUni()
str(myUni , nchar.max=10 , vec.len=2)

## 'data.frame': 20159 obs. of  9 variables:
##  $ Gene_Symbol  : chr  "A1BG" "NAT2" ...
##  $ Entrez       : int  1 10 100 1000 10000 ...
##  $ UNIPROT      : chr  "A1BG_HUMA"| __truncated__ "ARY2_HUMA"| __truncated__ ...
```

```
##  $ Entry        : chr  "P04217" "P11245" ...
##  $ HGNC         : chr  "HGNC:5" "HGNC:7646"| __truncated__ ...
##  $ Approved_Name: chr  "alpha-1-B"| __truncated__ "N-acetylt"| __truncated__ ...
##  $ Protein.name : chr  "Alpha-1B-"| __truncated__ "Arylamine"| __truncated__ ...
##  $ Chromosome   : chr  "19q13.43" "8p22" ...
##  $ AMINO_SEQ    : chr  "MSMLVVFLL"| __truncated__ "MDIEAYFER"| __truncated__ ...
```

In details, myUni is a data.frame composed by 9 columns:

- Gene_Symbol: a character vector of official Gene Symbols
- Entrez: a numeric vector of Entrez IDs
- UNIPROT: a character vector of Uniprot entries in "name_HUMAN" format
- Entry: a character vector of Uniprot entries
- HGNC: a character vector of gene names as HGNC numbers
- Approved_Name: a character vector of approved extended gene names
- Protein.name: a character vector of approved extended protein names
- Chromosome: a character vector of chromosomic cytoband positions
- Protein.name: a character vector of extended protein names
- AMINO_SEQ: a character vector of amino acid sequences for Uniprot entries

```
myPfam <- getMyPfam()
str(myPfam , nchar.max=10 , vec.len=2)

## 'data.frame': 44084 obs. of  11 variables:
##  $ Entry         : chr  "A0A5B9" "A0AUZ9" ...
##  $ Envelope_Start: chr  "14" "795" ...
##  $ Envelope_End  : chr  "107" "915" ...
##  $ Pfam_ID       : chr  "PF07654" "PF15275" ...
##  $ Pfam_Name     : chr  "C1-set" "PEHE" ...
##  $ Type          : chr  "Domain" "Family" ...
##  $ Clan_ID       : int  9 NA 52 175 175 ...
##  $ Entrez        : int  28638 151050 84561 54502 54502 ...
##  $ UNIPROT       : chr  "TRBC2_HUM"| __truncated__ "KAL1L_HUM"| __truncated__ ...
##  $ Gene_Symbol   : chr  "TRBC2" "KANSL1L" ...
##  $ Pfam_Fasta    : chr  "EPSEAEISH"| __truncated__ "ILTPSWRMV"| __truncated__ ...
```

In details, myPfam is a data.frame composed by 11 columns connected via UNIPROT/Entry to myUni dataset:

- Entry: a character vector of Uniprot entries
- Envelope_Start: a numeric vector of starts of the pfam domain relative to the reference protein
- Envelope_End: a numeric vector of ends of the pfam domain relative to the reference protein
- Pfam_ID: a character vector of Pfam IDs in the form of PF###### supported by LowMACA

- Pfam_Name: a character vector of full Pfam domain names
- Type: a character vector. One of the following: "Domain" "Family" "Repeat" or "Motif"
- Clan_ID: a numeric vector of Clan IDs, a sort of families of Pfam domains
- Entrez: a numeric vector of Entrez IDs
- UNIPROT: a character vector of Uniprot entries in format "name_HUMAN"
- Gene_Symbol: a character vector of official Gene Symbols
- Pfam_Fasta: a character vector of amino acid sequences of corresponding Pfam

```
myAlias <- getMyAlias()
str(myAlias , nchar.max=10 , vec.len=2)

## 'data.frame': 52491 obs. of  5 variables:
##  $ Alias               : chr  "NCRNA0018"| __truncated__ "A1BGAS" ...
##  $ Official_Gene_Symbol: chr  "A1BG-AS1" "A1BG-AS1" ...
##  $ Locus_Group         : chr  "non-codin"| __truncated__ "non-codin"| __truncated__ ...
##  $ Locus_Type          : chr  "RNA, long"| __truncated__ "RNA, long"| __truncated__ ...
##  $ MappedByLowMACA     : chr  "no" "no" ...
```

In details, myAlias is a data.frame composed by 2 columns:

- Alias: a character vector representing all the possible aliases and previous symbols for official Gene Symbols
- Official_Gene_Symbol: a character vector representing the approved and official Gene Symbol for HGNC database
- Locus_Group a character vector representing all the possible locus groups in HGNC database, like protein coding, RNA, pseudogene etc.
- Locus_Type a character vector representing all the possible locus types in HGNC database. It is a specification of locus group
- MappedByLowMACA a character vector of yes and no if the gene is included in myUni.RData

## 3 Datasets Curation

The three datasets presented above are the result of a manual curation of Uniprot database (http://www.uniprot.org/), Pfam-A database (http://pfam.xfam.org/) and HGNC database (http://www.genenames.org/). The entire script for the creation of the RData files can be found inside inst directory of this package.

*LowMACA* package maps mutations on residues, rather than genomic coordinates. This mapping arises a problem since a mutation is in fact a change in DNA that causes changes in all the proteins produced from that DNA piece. Our software package needs a 1 to 1 match between gene and protein as the majority of variant annotation tools does. The transformation from DNA change into unique amino acidic change on a single transcript is sometimes referred as "best effect" searching.

We follow this pipeline, in order of importance, to create our 1 gene 1 protein dataset. For every gene, we take the corresponding protein if:

1. Only one protein is known in Uniprot database

2. A "canonical" protein exists
3. There is a unique match with the protein sequence chosen by cBioPortal annotation
4. There is unique match between gene symbol and Uniprot protein symbol (like LCE6A and LCEA6_HUMAN)
5. All the Uniprot entries are classified as "Fragment", except one
6. Only one protein sequence is classified as "reviewed" by Uniprot
7. Only one protein is chosen by HGNC database
8. There is a partial match between gene symbol and Uniprot protein symbol, with a Levensthein distance that does not exceed 3 (e.g. TP53 and P53_HUMAN , distance=1). In case of ties, the "isoform 1" among them
9. The protein has the longest sequence among all the possible transcripts
10. The Uniprot name is the first in alphabetical order (there are in fact no genes in which this rule was applied)

The non protein coding genes are not included in the dataset and all the Pfam domains considered are comprised among the ones selected for myUni dataset.

# 4 Session Information

```
sessionInfo()

## R version 3.1.2 (2014-10-31)
## Platform: x86_64-apple-darwin10.8.0 (64-bit)
##
## locale:
## [1] C/it_IT.UTF-8/it_IT.UTF-8/C/it_IT.UTF-8/it_IT.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] LowMACAAnnotation_0.99.3
##
## loaded via a namespace (and not attached):
## [1] evaluate_0.5.5 formatR_1.0    highr_0.4      knitr_1.8
## [5] stringr_0.6.2  tools_3.1.2
```