Cookbook: How to annotate using biomaRt (ensembl)
by Simon Rasmussen & H. Bjørn Nielsen

biomaRt is a package to retrieve annotation data from external resources, consequently it requires you to be online. Further details can be found here: http://www.biomart.org/

First install and load the biomaRt library
source("http://www.bioconductor.org/biocLite.R")
biocLite("biomaRt")
require(biomaRt)

Getting an overview of the available annotation resources
Use the listMarts function to see the different databases that one can accessed. Often we use ‘ensembl’ for annotating genes from higher eukaryotes. For plant annotation you may use ‘plant_mart’ and so on.

listMarts()

In this example we will use the ensembl mart.

ensMart<-useMart("ensembl")

You can then use the listDatasets function to see which data sets that are available in the database under ensemble.

listDatasets(ensMart)

Here we exemplify using the ‘homo sapiens’ data set:

ensembl_hs_mart <- useMart(biomart="ensembl", dataset="hsapiens_gene_ensembl")

Next you may list all of the different annotations by listAttributes function, there are often several hundreds of attributes so view it in steps for ease (the output has two columns, one attribute name and description)

listAttributes(ensembl_hs_mart)[1:100,]

Often these attributes are key: ensembl_gene_id, ensembl_transcript_id, entrezgene, description, go_biological_process_id, name_1006, definition_1006. In addition, gene symbols are often used, however these are named differently for different organisms. For human it is hgnc_symbol and for mice it is mgi_symbol etc.
The ensemble-mart also holds the Affymetrix GeneChips probe-sets ids, e.g. affy_hg_u133a_2. Likewise fore some Agilent arrays e.g. agilent_wholegenome (human).

Annotating an array or a set of genes
To do so you need to have a set of identifiers from the array that can be used to retrieve cross-referenced annotation from the mart.

Here is an example where ensembl_gene_id, ensembl_transcript_id, hgnc_symbol, chromosome_name and go_biological_process are retrieved:
ensembl_hs_mart <- useMart(biomart="ensembl", dataset="hsapiens_gene_ensembl")
ensembl_df <- getBM(attributes=c("ensembl_gene_id", "ensembl_transcript_id", "hgnc_symbol", "chromosome_name", "entrezgene", "go_biological_process_id"), mart=ensembl_hs_mart)

This will download the annotation to a data.frame. Then, let us assume we have the following ensembl genes that we want to annotate using the above information:

my_genes = c("ENSG000000197971", "ENSG00000153165", "ENSG00000159352", "ENSG00000146006", "ENSG00000149809", "ENSG00000204179", "ENSG00000213023", "ENSG00000115008", "ENSG00000130844", "ENSG00000155363")

You may then match these to the annotation like this

my_genes_ann = ensembl[match(my_genes, ensembl$ensembl_gene_id),]