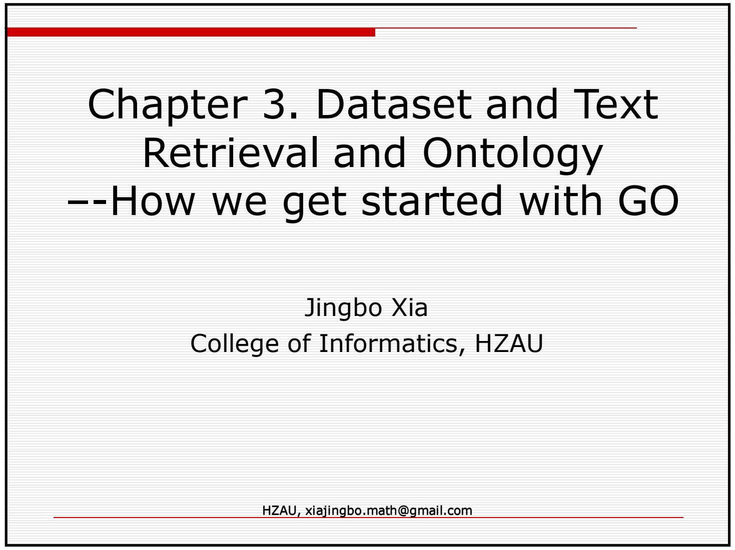
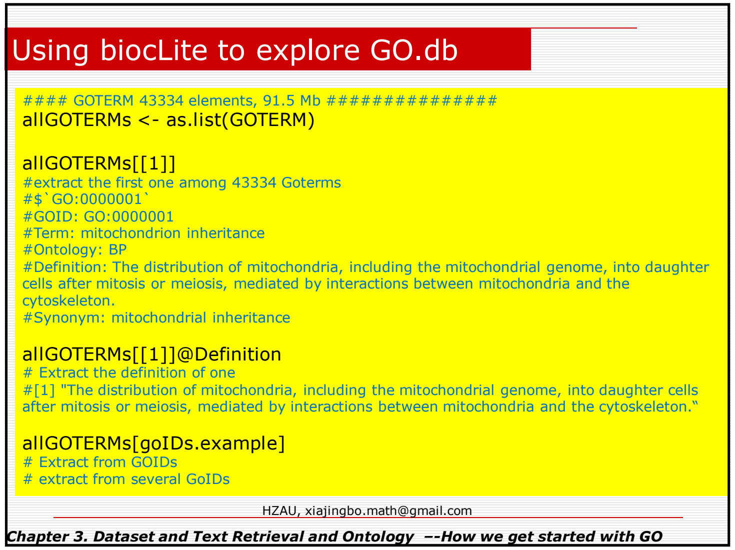


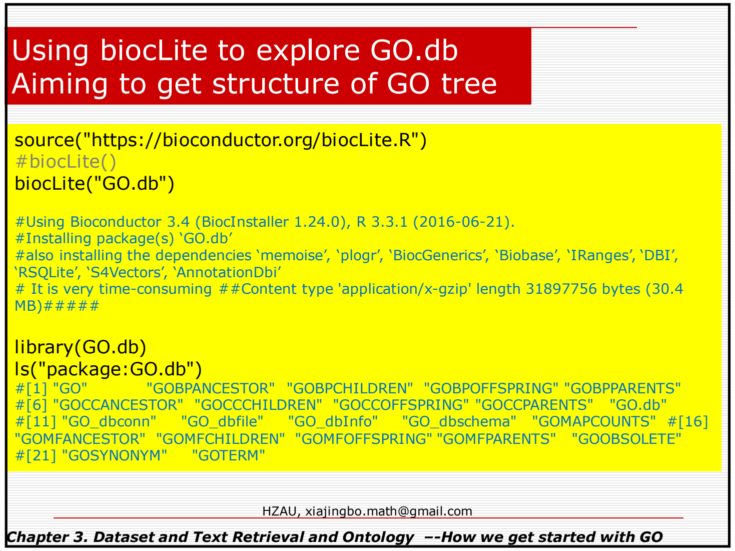
Chapter 3. Dataset and Text Retrieval and Ontology --How we get started with GO

Jingbo Xia
College of Informatics, HZAU



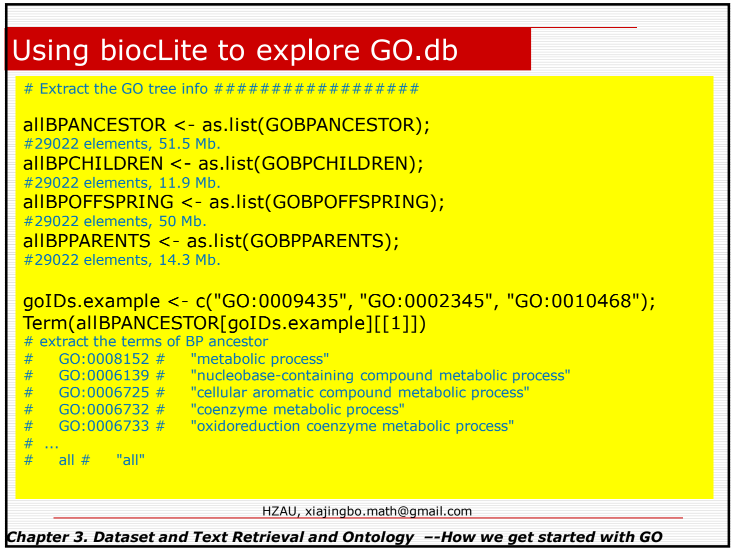
Using biocLite to explore GO.db

```
#### GOTERM 43334 elements, 91.5 Mb ###############
allGOTERMs <- as.list(GOTERM)

allGOTERMs[[1]]
#extract the first one among 43334 Goterms
#$`GO:0000001`
#GOID: GO:0000001
#Term: mitochondrion inheritance
#Ontology: BP
#Definition: The distribution of mitochondria, including the mitochondrial genome, into daughter
cells after mitosis or meiosis, mediated by interactions between mitochondria and the
cytoskeleton.
#Synonym: mitochondrial inheritance

allGOTERMs[[1]]@Definition
# Extract the definition of one
#[1] "The distribution of mitochondria, including the mitochondrial genome, into daughter cells
after mitosis or meiosis, mediated by interactions between mitochondria and the cytoskeleton."

allGOTERMs[goIDs.example]
# Extract from GOIDs
# extract from several GoIDs
```



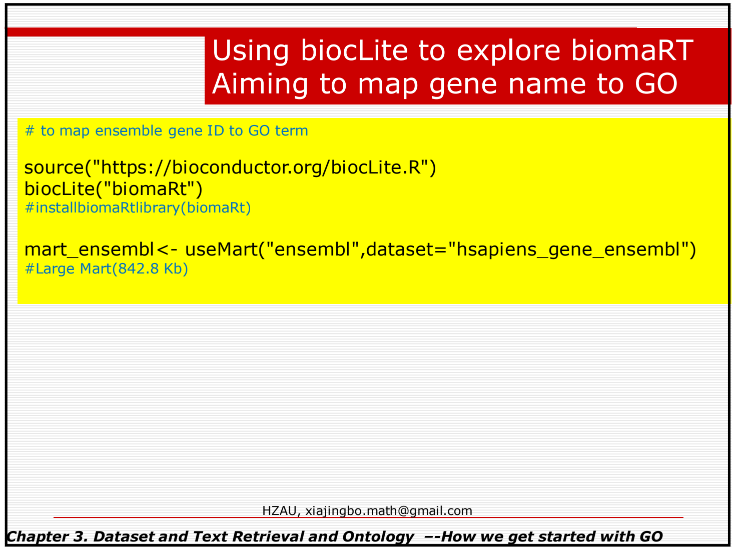
Using biocLite to explore GO.db
Aiming to get structure of GO tree

```
source("https://bioconductor.org/biocLite.R")
#biocLite()
biocLite("GO.db")

#Using Bioconductor 3.4 (BiocInstaller 1.24.0), R 3.3.1 (2016-06-21).
#Installing package(s) ‘GO.db’
#also installing the dependencies ‘memoise’, ‘plogr’, ‘BiocGenerics’, ‘Biobase’, ‘IRanges’, ‘DBI’,
‘RSQLite’, ‘S4Vectors’, ‘AnnotationDbi’
# It is very time-consuming ##Content type 'application/x-gzip' length 31897756 bytes (30.4
MB)#####

library(GO.db)
ls("package:GO.db")
#[1] "GO"             "GOBPANCESTOR"  "GOBPCHILDREN"  "GOBPOFFSPRING" "GOBPPARENTS"
#[6] "GOCCANCESTOR"  "GOCCCHILDREN"  "GOCCOFFSPRING" "GOCCPARENTS"   "GO.db"
#[11] "GO_dbconn"    "GO_dbfile"     "GO_dbInfo"     "GO_dbschema"   "GOMAPCOUNTS"  #[16]
"GOMFANCESTOR"  "GOMFCHILDREN"  "GOMFOFFSPRING" "GOMFPARENTS"   "GOOBSOLETE"
#[21] "GOSYNONYM"     "GOTERM"
```



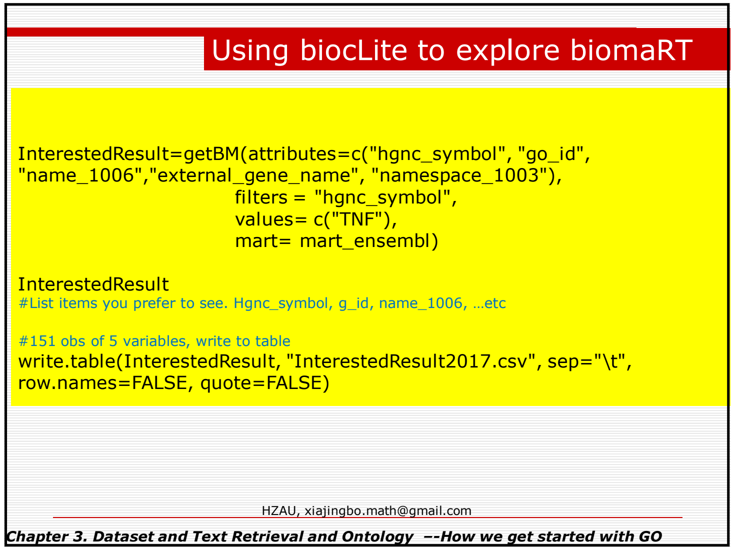
Using biocLite to explore GO.db

```
# Extract the GO tree info #################

allBPANCESTOR <- as.list(GOBPANCESTOR);
#29022 elements, 51.5 Mb.
allBPCHILDREN <- as.list(GOBPCHILDREN);
#29022 elements, 11.9 Mb.
allBPOFFSPRING <- as.list(GOBPOFFSPRING);
#29022 elements, 50 Mb.
allBPPARENTS <- as.list(GOBPPARENTS);
#29022 elements, 14.3 Mb.

goIDs.example <- c("GO:0009435", "GO:0002345", "GO:0010468");
Term(allBPANCESTOR[goIDs.example][[1]])
# extract the terms of BP ancestor
#    GO:0008152 #    "metabolic process"
#    GO:0006139 #    "nucleobase-containing compound metabolic process"
#    GO:0006725 #    "cellular aromatic compound metabolic process"
#    GO:0006732 #    "coenzyme metabolic process"
#    GO:0006733 #    "oxidoreduction coenzyme metabolic process"
#  ...
#    all #    "all"
```



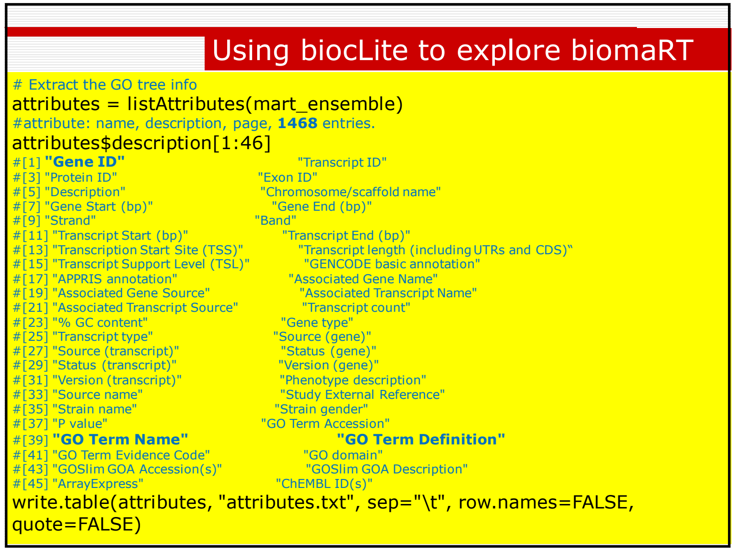
Using biocLite to explore biomaRT Aiming to map gene name to GO

```
# to map ensemble gene ID to GO term

source("https://bioconductor.org/biocLite.R")
biocLite("biomaRt")
#installbiomaRtlibrary(biomaRt)

mart_ensembl<- useMart("ensembl",dataset="hsapiens_gene_ensembl")
#Large Mart(842.8 Kb)
```



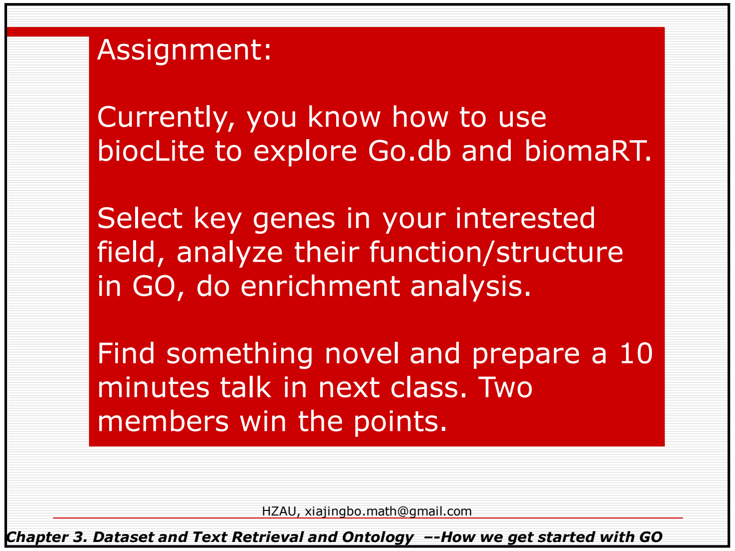
Using biocLite to explore biomaRT

```
InterestedResult=getBM(attributes=c("hgnc_symbol", "go_id",
"name_1006","external_gene_name", "namespace_1003"),
                       filters = "hgnc_symbol",
                       values= c("TNF"),
                       mart= mart_ensembl)

InterestedResult
#List items you prefer to see. Hgnc_symbol, g_id, name_1006, …etc

#151 obs of 5 variables, write to table
write.table(InterestedResult, "InterestedResult2017.csv", sep="\t",
row.names=FALSE, quote=FALSE)
```



Using biocLite to explore biomaRT

```
# Extract the GO tree info
attributes = listAttributes(mart_ensemble)
#attribute: name, description, page, 1468 entries.
attributes$description[1:46]
#[1] "Gene ID"                               "Transcript ID"
#[3] "Protein ID"                       "Exon ID"
#[5] "Description"                      "Chromosome/scaffold name"
#[7] "Gene Start (bp)"                    "Gene End (bp)"
#[9] "Strand"                          "Band"
#[11] "Transcript Start (bp)"              "Transcript End (bp)"
#[13] "Transcription Start Site (TSS)"         "Transcript length (including UTRs and CDS)"
#[15] "Transcript Support Level (TSL)"         "GENCODE basic annotation"
#[17] "APPRIS annotation"                   "Associated Gene Name"
#[19] "Associated Gene Source"                "Associated Transcript Name"
#[21] "Associated Transcript Source"           "Transcript count"
#[23] "% GC content"                      "Gene type"
#[25] "Transcript type"                   "Source (gene)"
#[27] "Source (transcript)"                 "Status (gene)"
#[29] "Status (transcript)"                "Version (gene)"
#[31] "Version (transcript)"                "Phenotype description"
#[33] "Source name"                       "Study External Reference"
#[35] "Strain name"                      "Strain gender"
#[37] "P value"                         "GO Term Accession"
#[39] "GO Term Name"                               "GO Term Definition"
#[41] "GO Term Evidence Code"                  "GO domain"
#[43] "GOSlim GOA Accession(s)"                "GOSlim GOA Description"
#[45] "ArrayExpress"                      "ChEMBL ID(s)"
write.table(attributes, "attributes.txt", sep="\t", row.names=FALSE,
quote=FALSE)
```



Assignment:

Currently, you know how to use biocLite to explore Go.db and biomaRT.

Select key genes in your interested field, analyze their function/structure in GO, do enrichment analysis.

Find something novel and prepare a 10 minutes talk in next class. Two members win the points.