R 프로그래밍 #12

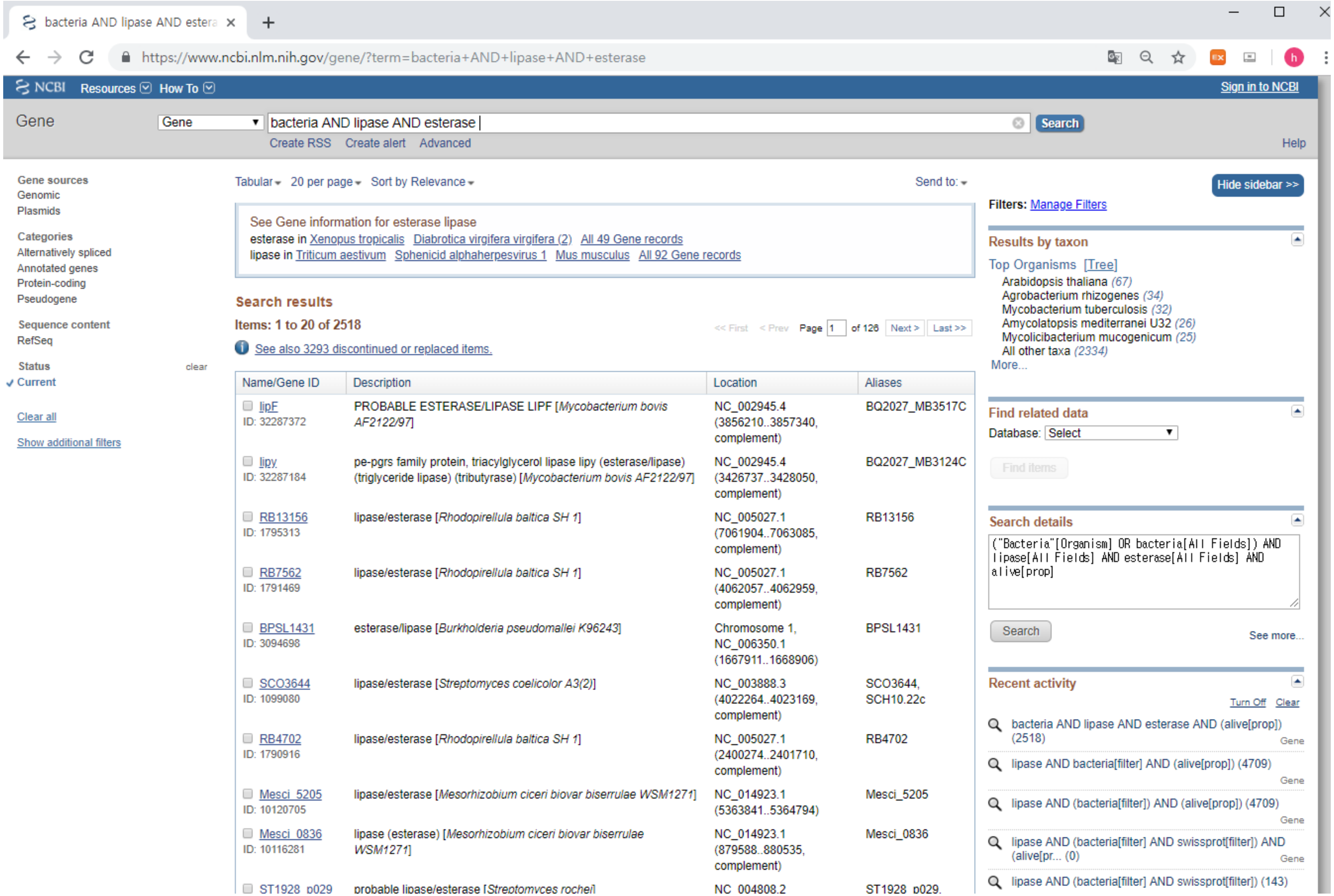
2019. 6. 4

한국생명공학연구원
김하성

Sequence analysis III

- Get sequences of 20 genes by searching “esterase & lipase & bacteria” from NCBI
- Align and visualize the sequences

Download sequence dataset



https://www.ncbi.nlm.nih.gov/gene/?term=bacteria+AND+lipase+AND+esterase

Gene: bacteria AND lipase AND esterase

Gene sources
Genomic
Plasmids

Categories
Alternatively spliced
Annotated genes
Protein-coding
Pseudogene

Sequence content
RefSeq

Status clear
✓ Current

Clear all

Show additional filters

See Gene information for esterase lipase
esterase in Xenopus tropicalis Diabrotica virgifera virgifera (2) All 49 Gene records
lipase in Triticum aestivum Sphenicid alphaherpesvirus 1 Mus musculus All 92 Gene records

Search results

Items: 1 to 20 of 2518

Page 1 of 126

See also 3293 discontinued or replaced items.

Name/Gene ID	Description	Location	Aliases
□ lipF ID: 32287372	PROBABLE ESTERASE/LIPASE LIPF [*Mycobacterium bovis AF2122/97*]	NC_002945.4 (3856210..3857340, complement)	BQ2027_MB3517C
□ lipy ID: 32287184	pe-pgrs family protein, triacylglycerol lipase lipy (esterase/lipase) (triglyceride lipase) (tributyrase) [*Mycobacterium bovis AF2122/97*]	NC_002945.4 (3426737..3428050, complement)	BQ2027_MB3124C
□ RB13156 ID: 1795313	lipase/esterase [*Rhodopirellula baltica SH 1*]	NC_005027.1 (7061904..7063085, complement)	RB13156
□ RB7562 ID: 1791469	lipase/esterase [*Rhodopirellula baltica SH 1*]	NC_005027.1 (4062057..4062959, complement)	RB7562
□ BPSL1431 ID: 3094698	esterase/lipase [*Burkholderia pseudomallei K96243*]	Chromosome 1, NC_006350.1 (1667911..1668906)	BPSL1431
□ SCO3644 ID: 1099080	lipase/esterase [*Streptomyces coelicolor A3(2)*]	NC_003888.3 (4022264..4023169, complement)	SCO3644, SCH10.22c
□ RB4702 ID: 1790916	lipase/esterase [*Rhodopirellula baltica SH 1*]	NC_005027.1 (2400274..2401710, complement)	RB4702
□ Mesci_5205 ID: 10120705	lipase/esterase [*Mesorhizobium ciceri biovar biserrulae WSM1271*]	NC_014923.1 (5363841..5364794)	Mesci_5205
□ Mesci_0836 ID: 10116281	lipase (esterase) [*Mesorhizobium ciceri biovar biserrulae WSM1271*]	NC_014923.1 (879588..880535, complement)	Mesci_0836
□ ST1928_p029	probable lipase/esterase [*Streptomyces rochei*]	NC_004808.2	ST1928_p029,

Filters: Manage Filters

Results by taxon

Top Organisms [Tree]

Arabidopsis thaliana	(67)
Agrobacterium rhizogenes	(34)
Mycobacterium tuberculosis	(32)
Amycolatopsis mediterranei U32	(26)
Mycolicibacterium mucogenicum	(25)
All other taxa	(2334)

More...

Find related data

Database: Select

Search details

("Bacteria"[Organism] OR bacteria[All Fields]) AND lipase[All Fields] AND esterase[All Fields] AND alive[prop]

Recent activity

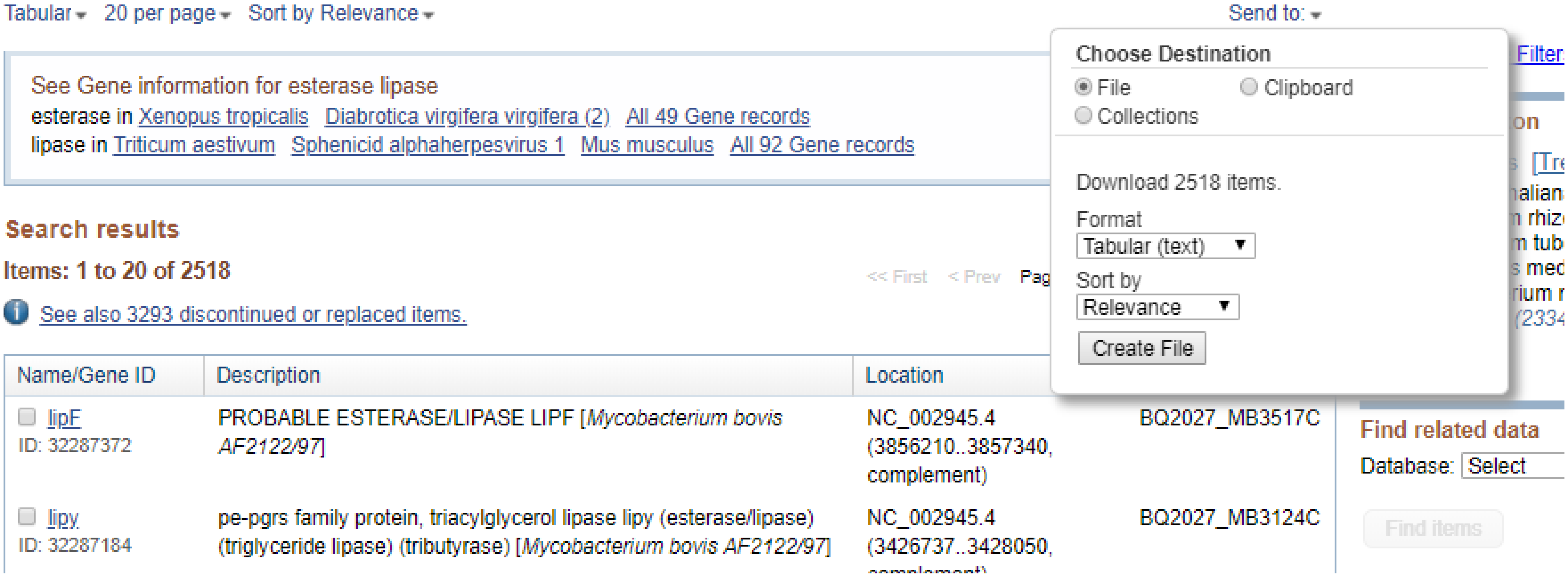
bacteria AND lipase AND esterase AND (alive[prop]) (2518) Gene

lipase AND bacteria[filter] AND (alive[prop]) (4709) Gene

lipase AND (bacteria[filter]) AND (alive[prop]) (4709) Gene

lipase AND (bacteria[filter] AND swissprot[filter]) AND (alive[pr... (0) Gene

lipase AND (bacteria[filter] AND swissprot[filter]) (143)



See Gene information for esterase lipase
esterase in Xenopus tropicalis Diabrotica virgifera virgifera (2) All 49 Gene records
lipase in Triticum aestivum Sphenicid alphaherpesvirus 1 Mus musculus All 92 Gene records

Search results

Items: 1 to 20 of 2518

See also 3293 discontinued or replaced items.

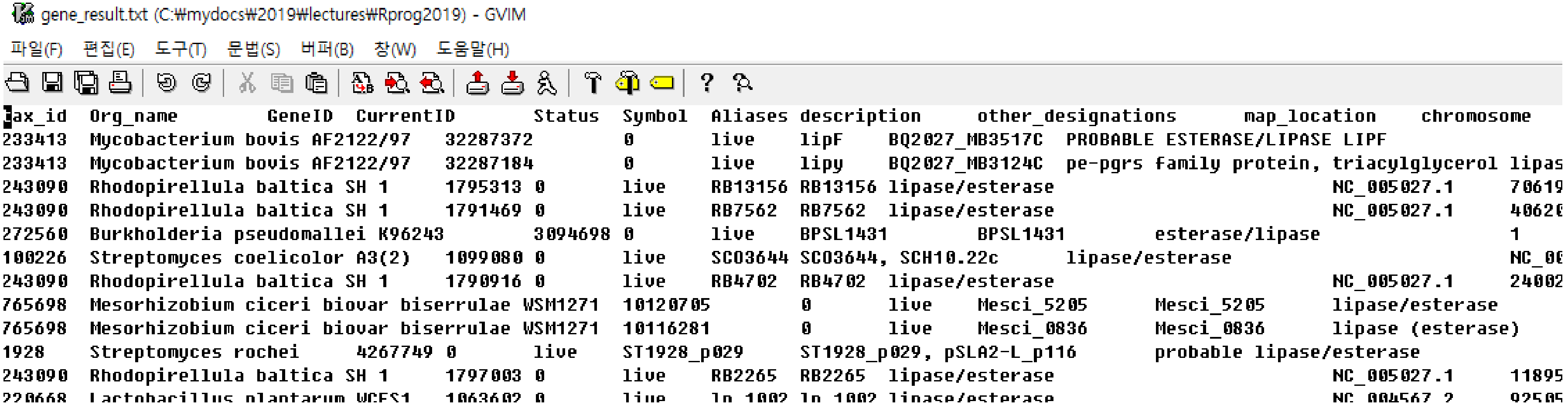
Name/Gene ID	Description	Location	
□ lipF ID: 32287372	PROBABLE ESTERASE/LIPASE LIPF [*Mycobacterium bovis AF2122/97*]	NC_002945.4 (3856210..3857340, complement)	BQ2027_MB3517C
□ lipy ID: 32287184	pe-pgrs family protein, triacylglycerol lipase lipy (esterase/lipase) (triglyceride lipase) (tributyrase) [*Mycobacterium bovis AF2122/97*]	NC_002945.4 (3426737..3428050, complement)	BQ2027_MB3124C

Choose Destination
● File ○ Clipboard
○ Collections

Download 2518 items.

Format: Tabular (text)
Sort by: Relevance

Create File
Filter
on
[Tre
alian
rhiz
m tub
s med
rium r
(2334
Find related data
Database: Select
Find items

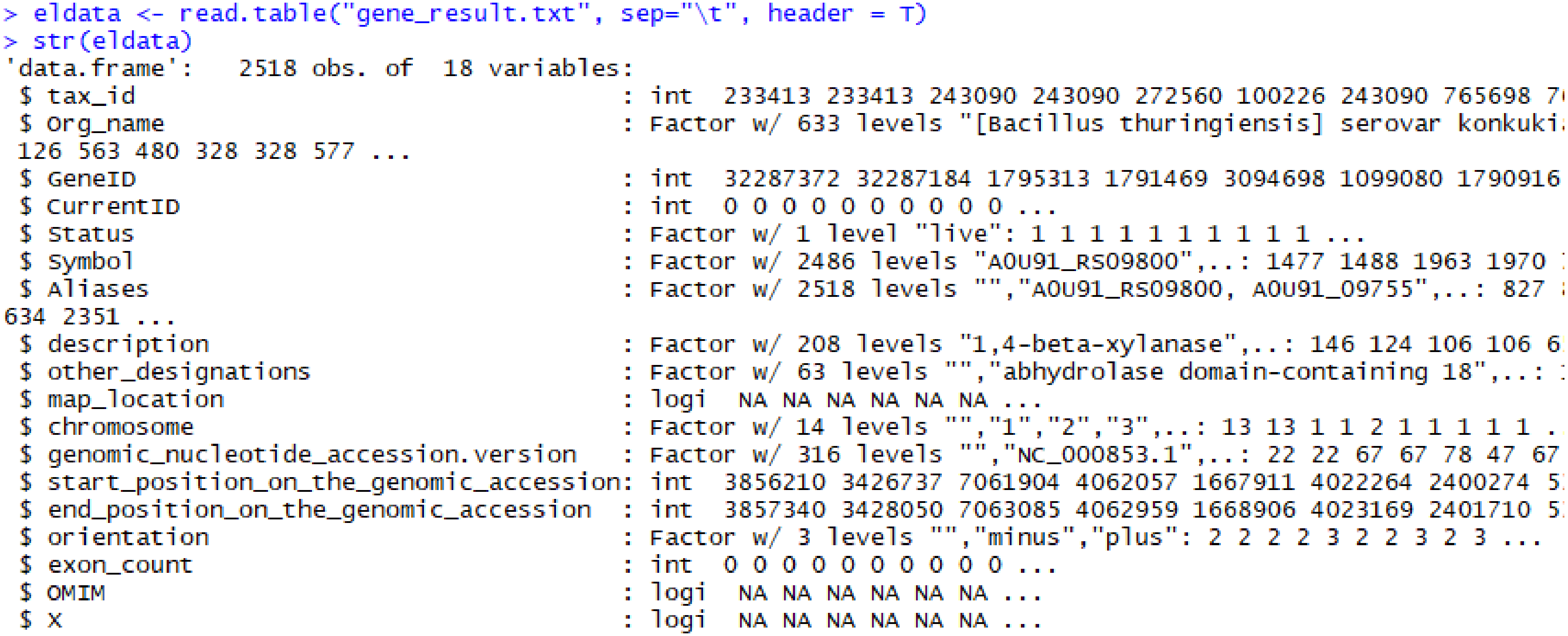


gene_result.txt (C:₩mydocs₩2019₩lectures₩Rprog2019) - GVIM
파일(F) 편집(E) 도구(T) 문법(S) 버퍼(B) 창(W) 도움말(H)

```
tax_id  Org_name        GeneID  CurrentID       Status  Symbol  Aliases description     other_designations      map_location    chromosome
233413  Mycobacterium bovis AF2122/97   32287372        0       live    lipF    BQ2027_MB3517C  PROBABLE ESTERASE/LIPASE LIPF
233413  Mycobacterium bovis AF2122/97   32287184        0       live    lipy    BQ2027_MB3124C  pe-pgrs family protein, triacylglycerol lipas
243090  Rhodopirellula baltica SH 1     1795313 0       live    RB13156 RB13156 lipase/esterase                         NC_005027.1     70619
243090  Rhodopirellula baltica SH 1     1791469 0       live    RB7562  RB7562  lipase/esterase                         NC_005027.1     40620
272560  Burkholderia pseudomallei K96243        3094698 0       live    BPSL1431        BPSL1431        esterase/lipase                 1
100226  Streptomyces coelicolor A3(2)   1099080 0       live    SCO3644 SCO3644, SCH10.22c      lipase/esterase                 NC_00
243090  Rhodopirellula baltica SH 1     1790916 0       live    RB4702  RB4702  lipase/esterase                         NC_005027.1     24002
765698  Mesorhizobium ciceri biovar biserrulae WSM1271  10120705        0       live    Mesci_5205      Mesci_5205      lipase/esterase
765698  Mesorhizobium ciceri biovar biserrulae WSM1271  10116281        0       live    Mesci_0836      Mesci_0836      lipase (esterase)
1928    Streptomyces rochei     4267749 0       live    ST1928_p029     ST1928_p029, pSLA2-L_p116       probable lipase/esterase
243090  Rhodopirellula baltica SH 1     1797003 0       live    RB2265  RB2265  lipase/esterase                         NC_005027.1     11895
220668  Lactobacillus plantarum WCFS1   1063602 0       live    lp_1002 lp_1002 lipase/esterase                         NC_004567.2     92505
```

Esterase & lipase in bacteria

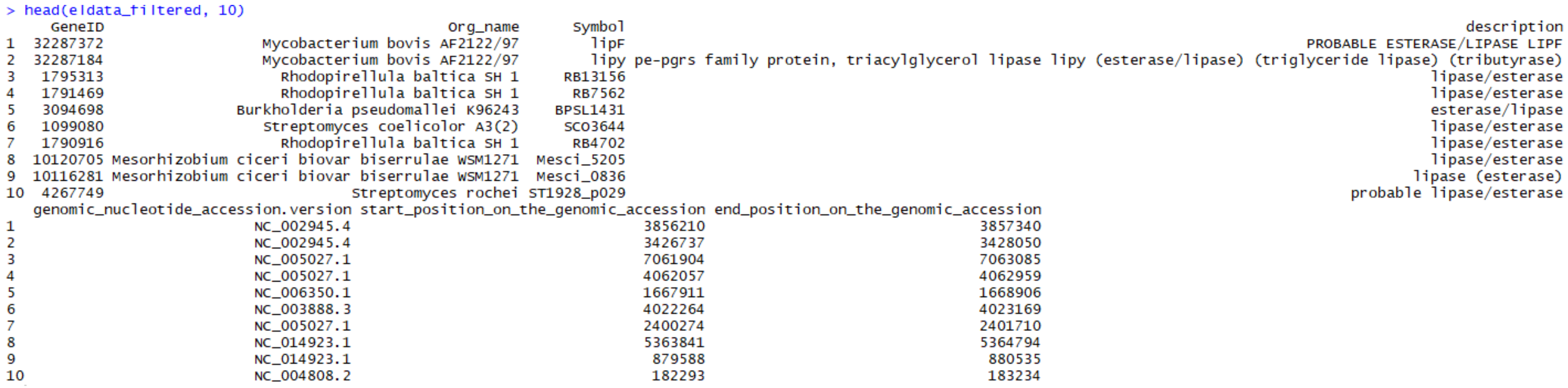
```
eldata <- read.table("gene_result.txt", sep="\t", header = T)
str(eldata)
```



```
> eldata <- read.table("gene_result.txt", sep="\t", header = T)
> str(eldata)
'data.frame':	2518 obs. of  18 variables:
 $ tax_id                                 : int  233413 233413 243090 243090 272560 100226 243090 765698 7
 $ Org_name                               : Factor w/ 633 levels "[Bacillus thuringiensis] serovar konkuki
 126 563 480 328 328 577 ...
 $ GeneID                                 : int  32287372 32287184 1795313 1791469 3094698 1099080 1790916
 $ CurrentID                              : int  0 0 0 0 0 0 0 0 0 0 ...
 $ Status                                 : Factor w/ 1 level "live": 1 1 1 1 1 1 1 1 1 1 ...
 $ Symbol                                 : Factor w/ 2486 levels "A0U91_RS09800",..: 1477 1488 1963 1970 
 $ Aliases                                : Factor w/ 2518 levels "","A0U91_RS09800, A0U91_09755",..: 827 
634 2351 ...
 $ description                            : Factor w/ 208 levels "1,4-beta-xylanase",..: 146 124 106 106 6
 $ other_designations                     : Factor w/ 63 levels "","abhydrolase domain-containing 18",..: 
 $ map_location                           : logi  NA NA NA NA NA NA ...
 $ chromosome                             : Factor w/ 14 levels "","1","2","3",..: 13 13 1 1 2 1 1 1 1 1 .
 $ genomic_nucleotide_accession.version   : Factor w/ 316 levels "","NC_000853.1",..: 22 22 67 67 78 47 67
 $ start_position_on_the_genomic_accession: int  3856210 3426737 7061904 4062057 1667911 4022264 2400274 5
 $ end_position_on_the_genomic_accession  : int  3857340 3428050 7063085 4062959 1668906 4023169 2401710 5
 $ orientation                            : Factor w/ 3 levels "","minus","plus": 2 2 2 2 3 2 2 3 2 3 ...
 $ exon_count                             : int  0 0 0 0 0 0 0 0 0 0 ...
 $ OMIM                                   : logi  NA NA NA NA NA NA ...
 $ X                                      : logi  NA NA NA NA NA NA ...
```

Data selection, filtering

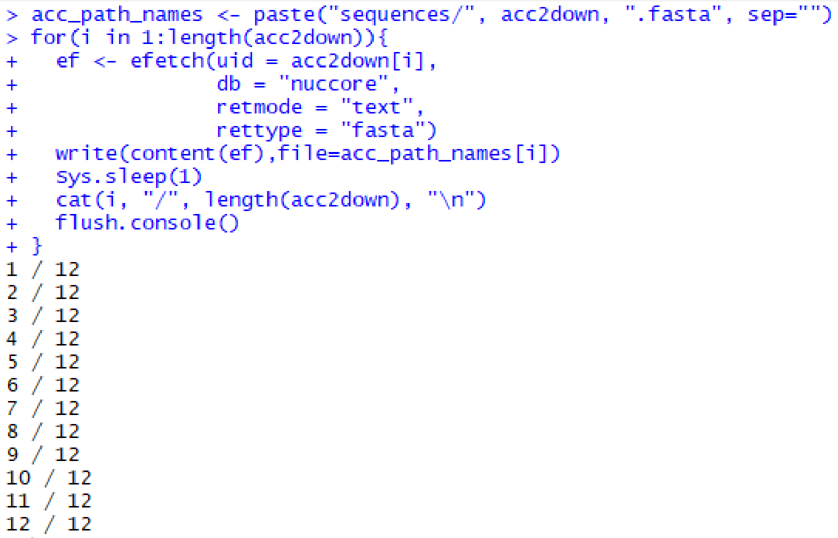
```
library(dplyr)

eldata_filtered <- eldata %>%
  select(GeneID,
         Org_name,
         Symbol,
         description,
         genomic_nucleotide_accession.version,
         start_position_on_the_genomic_accession,
         end_position_on_the_genomic_accession) %>%
  filter(genomic_nucleotide_accession.version != "")
```



```
> head(eldata_filtered, 10)
     GeneID                                       Org_name      Symbol                                                                                               description
1  32287372                  Mycobacterium bovis AF2122/97        lipF                                                                             PROBABLE ESTERASE/LIPASE LIPF
2  32287184                  Mycobacterium bovis AF2122/97        lipy pe-pgrs family protein, triacylglycerol lipase lipy (esterase/lipase) (triglyceride lipase) (tributyrase)
3   1795313                    Rhodopirellula baltica SH 1     RB13156                                                                                           lipase/esterase
4   1791469                    Rhodopirellula baltica SH 1      RB7562                                                                                           lipase/esterase
5   3094698               Burkholderia pseudomallei K96243    BPSL1431                                                                                           esterase/lipase
6   1099080                  Streptomyces coelicolor A3(2)     SCO3644                                                                                           lipase/esterase
7   1790916                    Rhodopirellula baltica SH 1      RB4702                                                                                           lipase/esterase
8  10120705 Mesorhizobium ciceri biovar biserrulae WSM1271  Mesci_5205                                                                                           lipase/esterase
9  10116281 Mesorhizobium ciceri biovar biserrulae WSM1271  Mesci_0836                                                                                         lipase (esterase)
10  4267749                            Streptomyces rochei ST1928_p029                                                                                  probable lipase/esterase
   genomic_nucleotide_accession.version start_position_on_the_genomic_accession end_position_on_the_genomic_accession
1                           NC_002945.4                                 3856210                               3857340
2                           NC_002945.4                                 3426737                               3428050
3                           NC_005027.1                                 7061904                               7063085
4                           NC_005027.1                                 4062057                               4062959
5                           NC_006350.1                                 1667911                               1668906
6                           NC_003888.3                                 4022264                               4023169
7                           NC_005027.1                                 2400274                               2401710
8                           NC_014923.1                                 5363841                               5364794
9                           NC_014923.1                                  879588                                880535
10                          NC_004808.2                                  182293                                183234
```

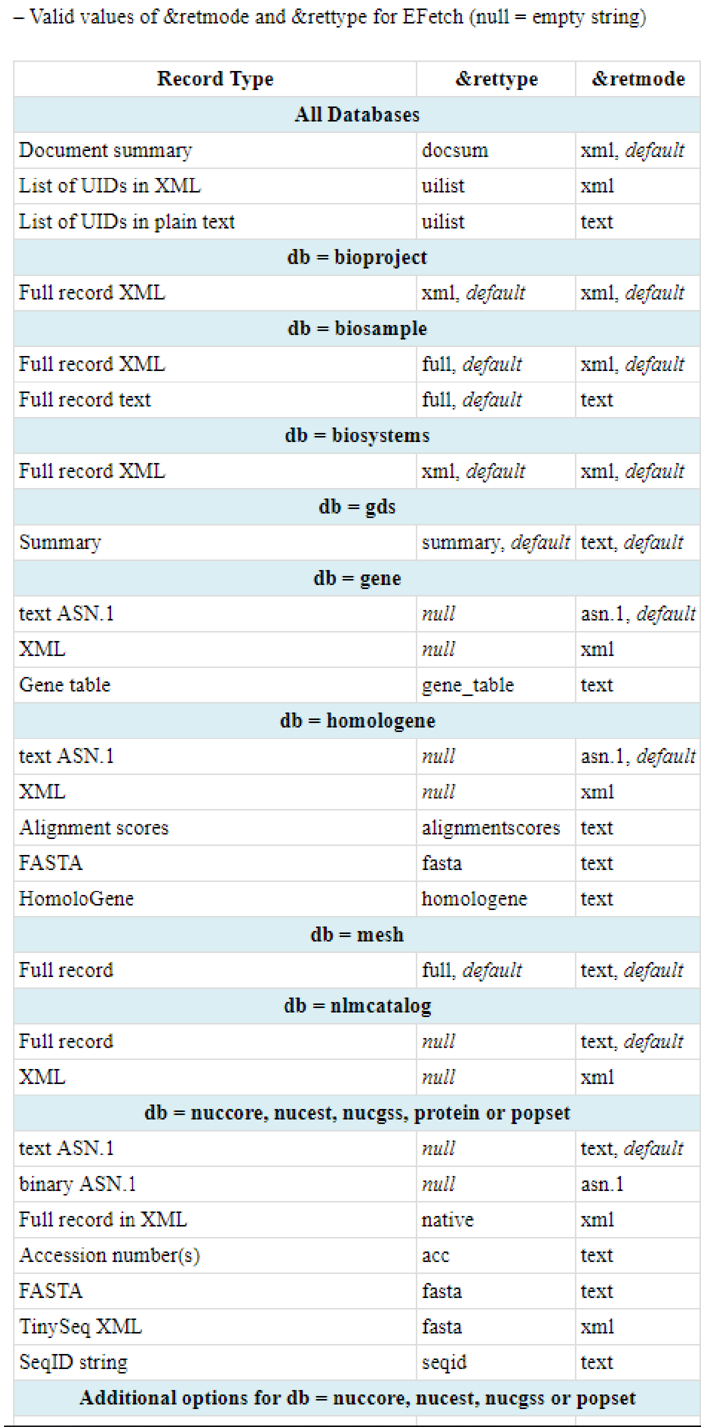
Download fasta files

```
eldata_filtered2 <- eldata_filtered[1:20,]
acc <- eldata_filtered2$genomic_nucleotide_accession.version
acc2 <- as.character(acc)
acc2down <- acc2[!duplicated(acc2)]
acc_path_names <- paste("sequences/", acc2down, ".fasta", sep="")
for(i in 1:length(acc2down)){
  ef <- efetch(uid = acc2down[i],
               db = "nuccore",
               retmode = "text",
               rettype = "fasta")
  write(content(ef),file=acc_path_names[i])
  Sys.sleep(1)
  cat(i, "/", length(acc2down), "\n")
  flush.console()
}
```



```
> acc_path_names <- paste("sequences/", acc2down, ".fasta", sep="")
> for(i in 1:length(acc2down)){
+   ef <- efetch(uid = acc2down[i],
+                db = "nuccore",
+                retmode = "text",
+                rettype = "fasta")
+   write(content(ef),file=acc_path_names[i])
+   Sys.sleep(1)
+   cat(i, "/", length(acc2down), "\n")
+   flush.console()
+ }
1 / 12
2 / 12
3 / 12
4 / 12
5 / 12
6 / 12
7 / 12
8 / 12
9 / 12
10 / 12
11 / 12
12 / 12
```

https://www.ncbi.nlm.nih.gov/books/NBK25499/table/chapter4.T._valid_values_of__retmode_and/?report=objectonly



– Valid values of &retmode and &rettype for EFetch (null = empty string)

Record Type	&rettype	&retmode
All Databases		
Document summary	docsum	xml, *default*
List of UIDs in XML	uilist	xml
List of UIDs in plain text	uilist	text
db = bioproject		
Full record XML	xml, *default*	xml, *default*
db = biosample		
Full record XML	full, *default*	xml, *default*
Full record text	full, *default*	text
db = biosystems		
Full record XML	xml, *default*	xml, *default*
db = gds		
Summary	summary, *default*	text, *default*
db = gene		
text ASN.1	*null*	asn.1, *default*
XML	*null*	xml
Gene table	gene_table	text
db = homologene		
text ASN.1	*null*	asn.1, *default*
XML	*null*	xml
Alignment scores	alignmentscores	text
FASTA	fasta	text
HomoloGene	homologene	text
db = mesh		
Full record	full, *default*	text, *default*
db = nlmcatalog		
Full record	*null*	text, *default*
XML	*null*	xml
db = nuccore, nucest, nucgss, protein or popset		
text ASN.1	*null*	text, *default*
binary ASN.1	*null*	asn.1
Full record in XML	native	xml
Accession number(s)	acc	text
FASTA	fasta	text
TinySeq XML	fasta	xml
SeqID string	seqid	text
Additional options for db = nuccore, nucest, nucgss or popset		

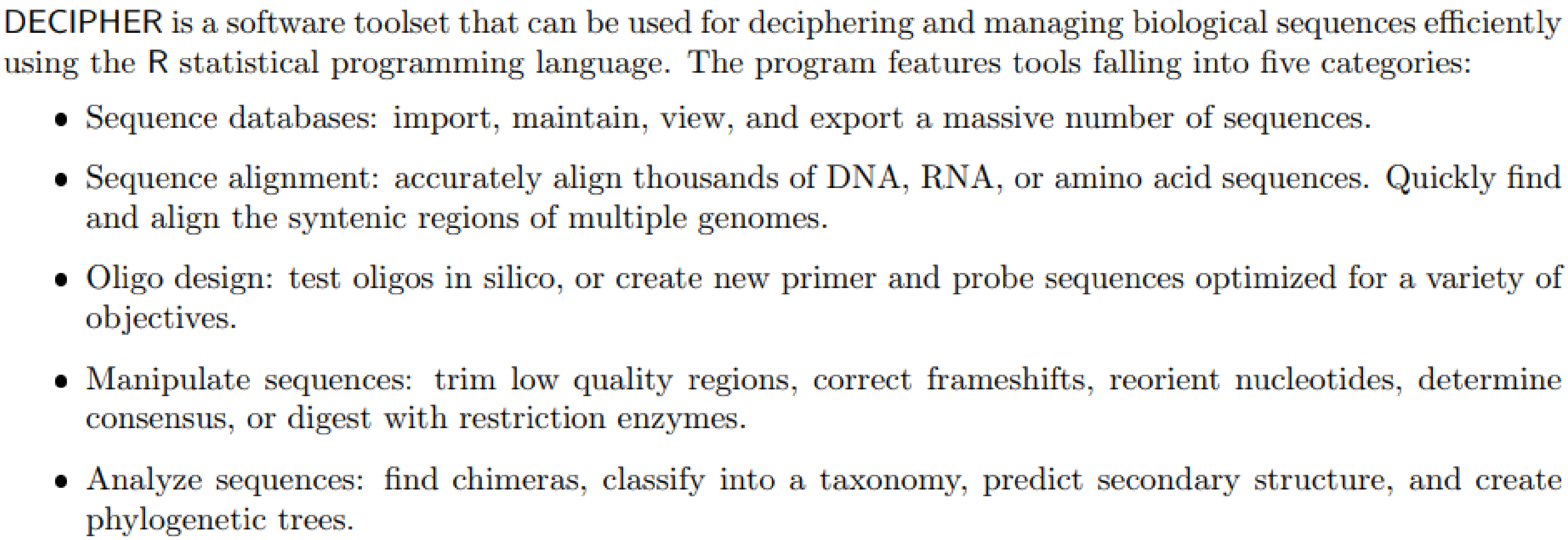
Extract gene sequences

```
library(Biostrings)

genomeseq <- readDNAStringSet(acc_path_names)

tmp <- strsplit(names(genomeseq), split=" ")
tmp2 <- lapply(tmp , function(x){x[1]})
names(genomeseq) <- unlist(tmp2)

acc_ids <-
as.character(eldata_filtered2$genomic_nucleotide_accession.version)
startpos <- eldata_filtered2$start_position_on_the_genomic_accession
endpos <- eldata_filtered2$end_position_on_the_genomic_accession
```

Exercise 12-1

- Make a list type variable “myseq” with length 20
- Use ‘for’ to read all the lipase/esterase sequences
- change the type of “myseq” to DNAStringSet

DECIPHER

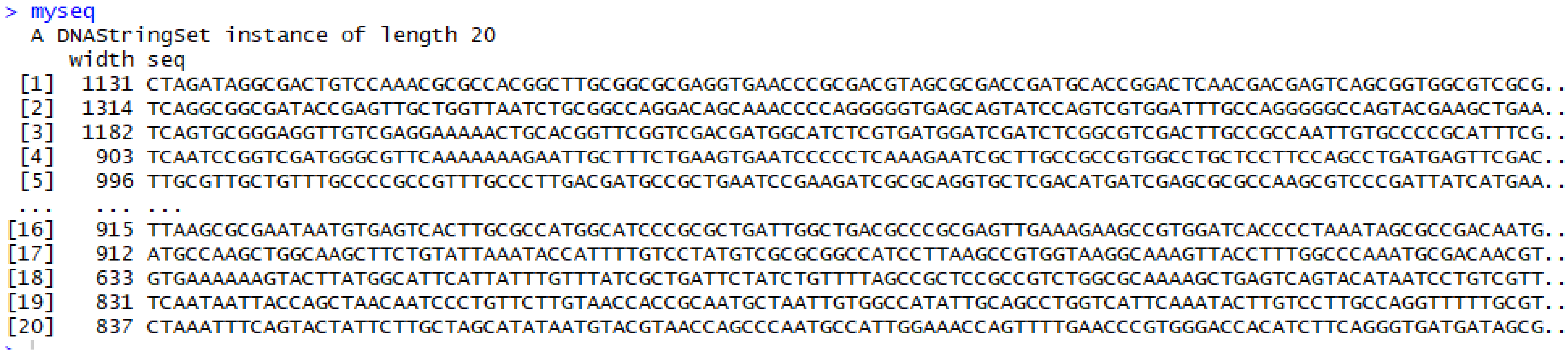


DECIPHER is a software toolset that can be used for deciphering and managing biological sequences efficiently using the R statistical programming language. The program features tools falling into five categories:

- Sequence databases: import, maintain, view, and export a massive number of sequences.
- Sequence alignment: accurately align thousands of DNA, RNA, or amino acid sequences. Quickly find and align the syntenic regions of multiple genomes.
- Oligo design: test oligos in silico, or create new primer and probe sequences optimized for a variety of objectives.
- Manipulate sequences: trim low quality regions, correct frameshifts, reorient nucleotides, determine consensus, or digest with restriction enzymes.
- Analyze sequences: find chimeras, classify into a taxonomy, predict secondary structure, and create phylogenetic trees.

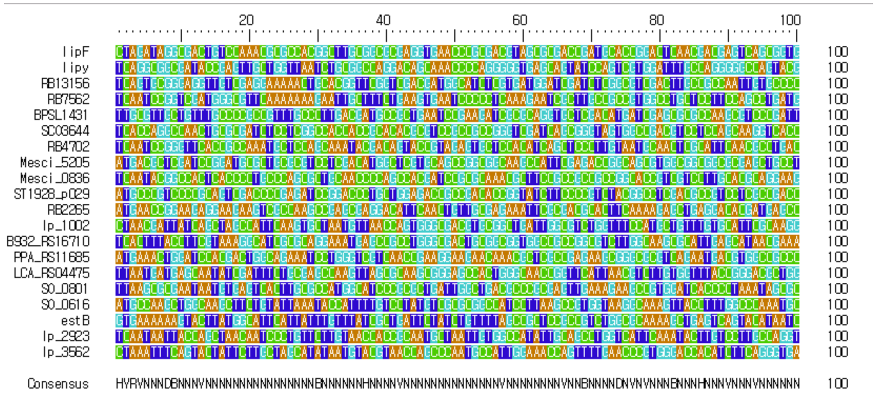
https://bioconductor.org/packages/release/bioc/html/DECIPHER.html

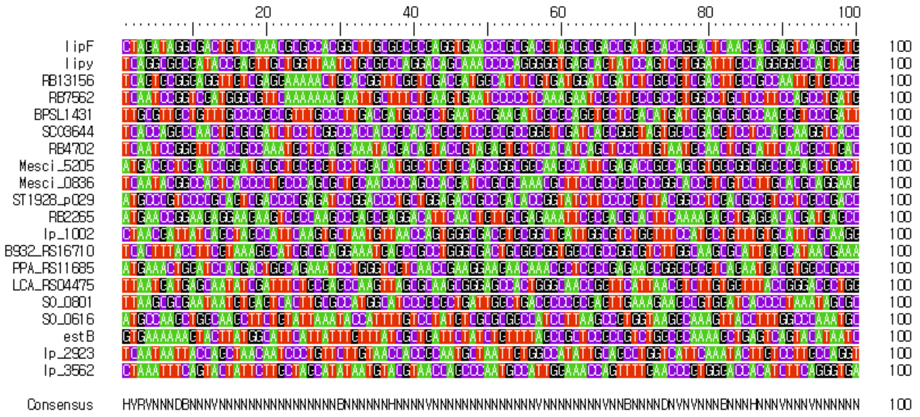
Browse sequences



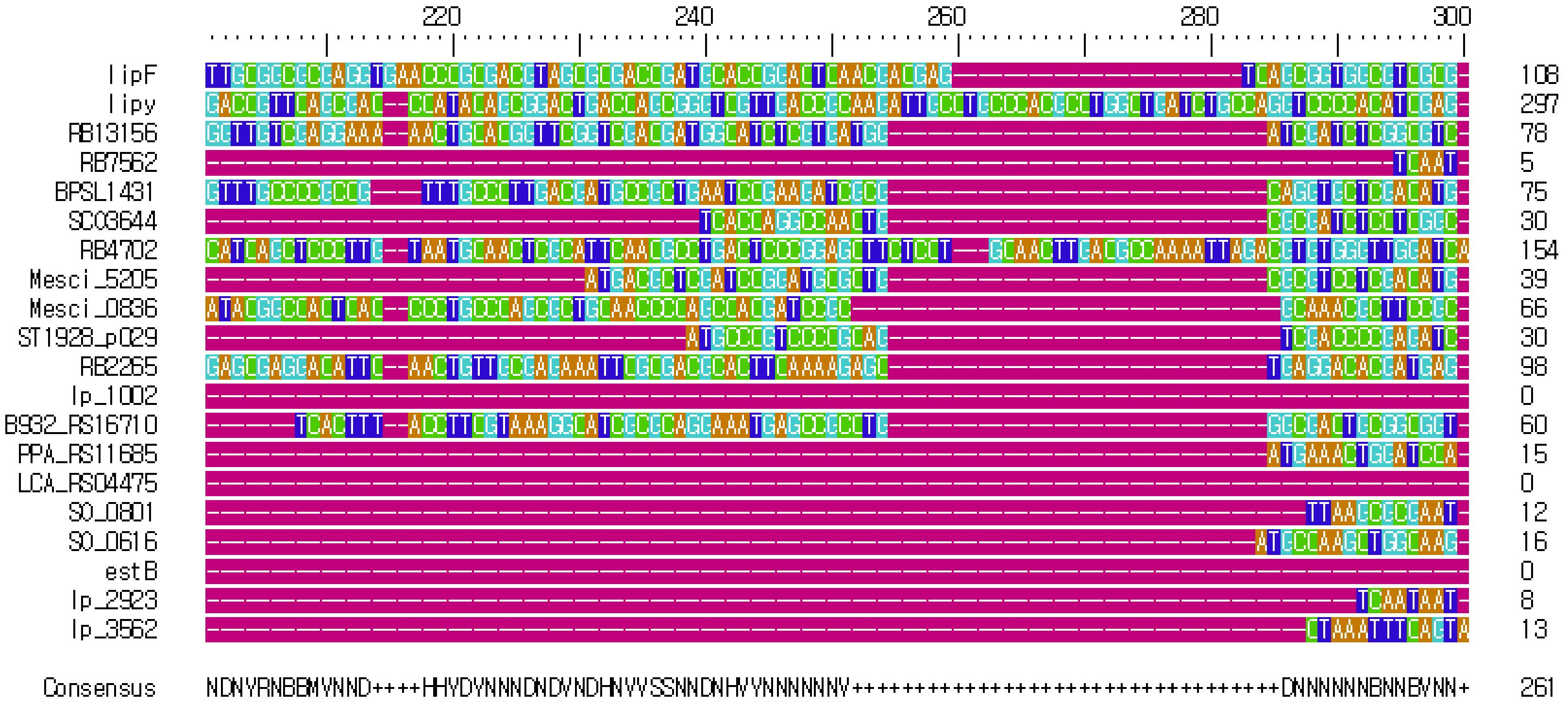
```
> myseq
  A DNAStringSet instance of length 20
     width seq
 [1]  1131 CTAGATAGGCGACTGTCCAAACGCGCCACGGCTTGCGGCGCGAGGTGAACCCGCGACGTAGCGCGACCGATGCACCGGACTCAACGACGAGTCAGCGGTGGCGTCGCG...
 [2]  1314 TCAGGCGGCGATACCGAGTTGCTGGTTAATCTGCGGCCAGGACAGCAAACCCCAGGGGGTGAGCAGTATCCAGTCGTGGATTTGCCAGGGGGCCAGTACGAAGCTGAA...
 [3]  1182 TCAGTGCGGGAGGTTGTCGAGGAAAAACTGCACGGTTCGGTCGACGATGGCATCTCGTGATGGATCGATCTCGGCGTCGACTTGCCGCCAATTGTGCCCCGCATTTCG...
 [4]   903 TCAATCCGGTCGATGGGCGTTCAAAAAAAGAATTGCTTTCTGAAGTGAATCCCCCTCAAAGAATCGCTTGCCGCCGTGGCCTGCTCCTTCCAGCCTGATGAGTTCGAC...
 [5]   996 TTGCGTTGCTGTTTGCCCCGCCGTTTGCCCTTGACGATGCCGCTGAATCCGAAGATCGCGCAGGTGCTCGACATGATCGAGCGCGCCAAGCGTCCCGATTATCATGAA...
...    ... ...
[16]   915 TTAAGCGCGAATAATGTGAGTCACTTGCGCCATGGCATCCCGCGCTGATTGGCTGACGCCCGCGAGTTGAAAGAAGCCGTGGATCACCCCTAAATAGCGCCGACAATG...
[17]   912 ATGCCAAGCTGGCAAGCTTCTGTATTAAATACCATTTTGTCCTATGTCGCGCGGCCATCCTTAAGCCGTGGTAAGGCAAAGTTACCTTTGGCCCAAATGCGACAACGT...
[18]   633 GTGAAAAAAGTACTTATGGCATTCATTATTTGTTTATCGCTGATTCTATCTGTTTTAGCCGCTCCGCCGTCTGGCGCAAAAGCTGAGTCAGTACATAATCCTGTCGTT...
[19]   831 TCAATAATTACCAGCTAACAATCCCTGTTCTTGTAACCACCGCAATGCTAATTGTGGCCATATTGCAGCCTGGTCATTCAAATACTTGTCCTTGCCAGGTTTTTGCGT...
[20]   837 CTAAATTTCAGTACTATTCTTGCTAGCATATAATGTACGTAACCAGCCCAATGCCATTGGAAACCAGTTTTGAACCCGTGGGACCACATCTTCAGGGTGATGATAGCG...
```

```
library(DECIPHER)
BrowseSeqs(myseq, htmlFile="myseq.html", colWidth=100)
dnacolors <- c("#1E90FF", "#32CD32", "#9400D3", "black", "#EE3300")
BrowseSeqs(myseq, htmlFile="myseq.html", colors=dnacolors, colWidth=100)
```





Sequence alignment

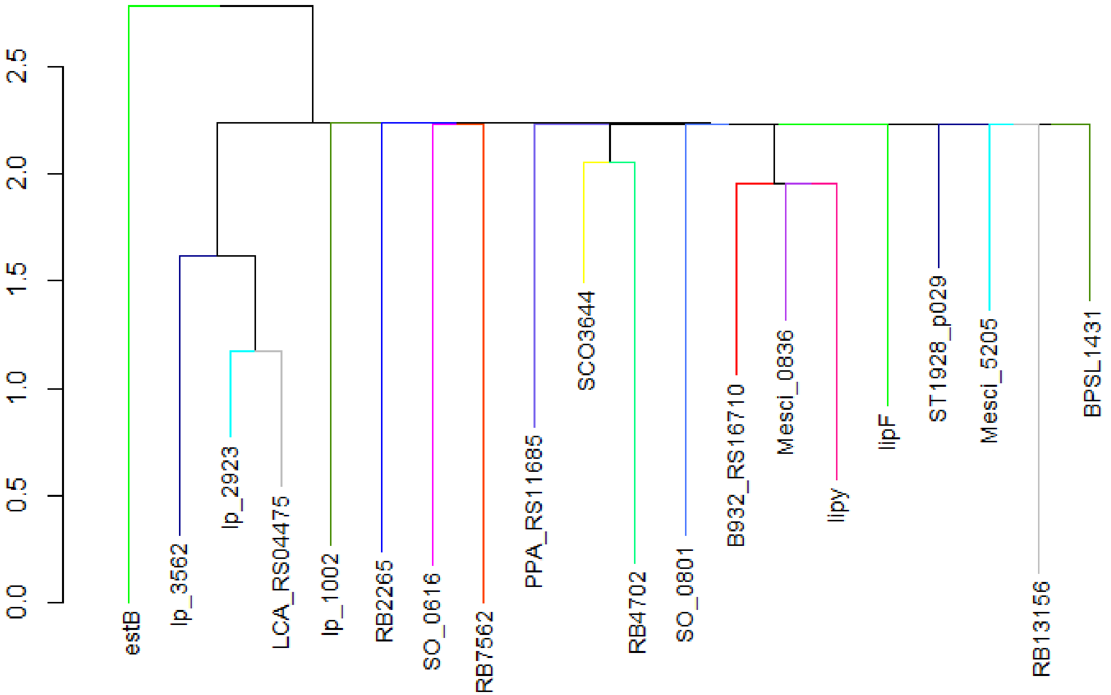


```
                                    220                 240                 260                 280                 300
              ..........|.........|.........|.........|.........|.........|.........|.........|.........|.........|
        lipF  TTGCGGCGCGAGGTGAACCCGCGACGTAGCGCGACCGATGCACCGGACTCAACGACGAG-----------------------TCAGCGGTGGCGTCGCG-  108
        lipy  GACCGTTCAGCGAC--CCATACAGCGGACTGACCAGCGGGTCGTTGACCGCAAGATTGCCTGCCCACGCCTGGCTGATCTGCCAGGTCCCCACATCGAG-  297
     RB13156  GGTTGTCGAGGAAA--AACTGCACGGTTCGGTCGACGATGGCATCTCGTGATGG------------------------------ATCGATCTCGGCGTC-  78
      RB7562  ----------------------------------------------------------------------------------------------TCAAT-  5
    BPSL1431  GTTTGCCCCGCCG---TTTGCCCTTGACGATGCCCGCTGAATCCGAAGATCGCG------------------------------CAGGTGCTCGACATG-  75
     SCO3644  ---------------------------------------TCACCAGGCCAACTG------------------------------CGCCATCTCCTCGGC-  30
      RB4702  CATCAGCTCCCTTG--TAATGCAACTCGCATTCAACGCCTGACTCCCGGAGCTTCTCCT---GCAACTTGACGCCAAAATTAGACGTGTGGGTTGGATCA  154
  Mesci_5205  ------------------------------ATGACGCTCGATCCGGATGCGCTG------------------------------CGCGTCCTCGACATG-  39
  Mesci_0836  ATACGGCCACTCAC--CCCTGCCCAGCGCTGCAACCCCAGCCACGATCCGC----------------------------------GCAAACGCTTCCGC-  66
ST1928_p029  --------------------------------------ATGCCCGTCCCCGCAG-------------------------------TCGACCCCGAGATC-  30
      RB2265  GAGCGAGGACATTC--AACTGTTGCCGAGAAATTCGCGACGCACTTCAAAAGAGC------------------------------TCAGGACACCATGAG-  98
     lp_1002  ----------------------------------------------------------------------------------------------------  0
B932_RS16710  -------TCACTTT--ACCTTCGTAAAGGCATCGCGCAGGAAATGAGCCGCCTG------------------------------GGCGACTGCGGCGGT-  60
PPA_RS11685  ------------------------------------------------------------------------------------ATGAAACTGGATCCA-  15
LCA_RS04475  ----------------------------------------------------------------------------------------------------  0
     SO_0801  ---------------------------------------------------------------------------------------TTAAGCGCGAAT-  12
     SO_0616  -----------------------------------------------------------------------------------ATGCCAAGCTGGCAAG-  16
        estB  ----------------------------------------------------------------------------------------------------  0
     lp_2923  -------------------------------------------------------------------------------------------TCAATAAT-  8
     lp_3562  ---------------------------------------------------------------------------------------CTAAATTTCAGTA  13

   Consensus  NDNYRNBBMVNND++++HHYDVNNNDNDVNDHNVVSSNNDNHVYNNNNNNV++++++++++++++++++++++++++++++++++DNNNNNNBNNBVNN+  261
```

```
aln <- AlignSeqs(myseq) # output alignment
BrowseSeqs(aln, htmlFile="myaln.html", colWidth=100)
```

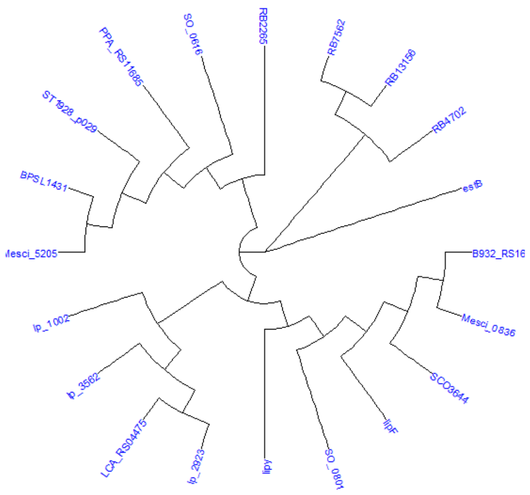
Clustering and tree

```
d <- DistanceMatrix(aln, correction="Jukes-Cantor", verbose=FALSE)
c <- IdClusters(d, method="ML", cutoff=.05, showPlot=TRUE, myXStringSet=aln)
```



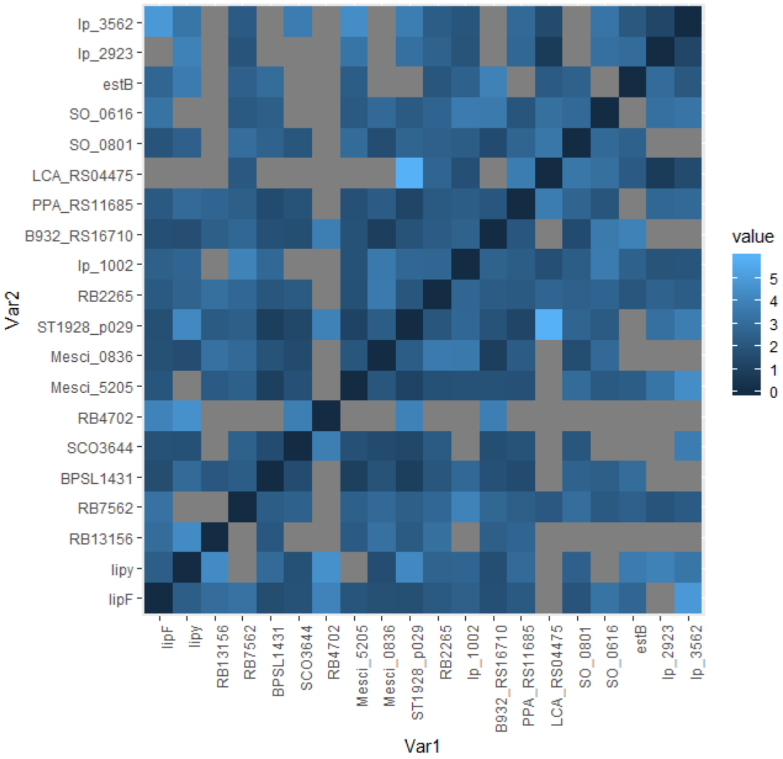
Clustering and tree II

```
library(msa)
library(ape)
library(seqinr)
library(ggtree)

myaln<-msa(myseq, method="ClustalOmega", type="dna")
myaln2 <- msaConvert(myaln, type="seqinr::alignment")
d <- dist.alignment(myaln2, "identity")
mytree <- njs(d)

ggtree(mytree) +
  geom_tiplab() +
  xlim(-1, 15)

ggtree(mytree, branch.length="none") +
  geom_tiplab() +
  xlim(-1, 15)

ggtree(mytree, layout="circular") +
  geom_tiplab2(color='blue', size=3)

ggtree(mytree, layout="circular", branch.length="none",
  geom_tiplab2(aes(angle=angle), color='blue', size=3)
```



Heatmap with ggplot2

```
library(reshape2)
library(ggplot2)
d <- DistanceMatrix(aln, correction="Jukes-Cantor", verbose=FALSE)

d_melt <- melt(d)
ggplot(d_melt, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile()

ggplot(d_melt, aes(x=Var1, y=Var2, fill=value)) +
  geom_tile() +
  theme(axis.text.x = element_text(angle = 90, hjust = 1))
```



Next

- Sequence analysis IV
- Case study
- R with blast