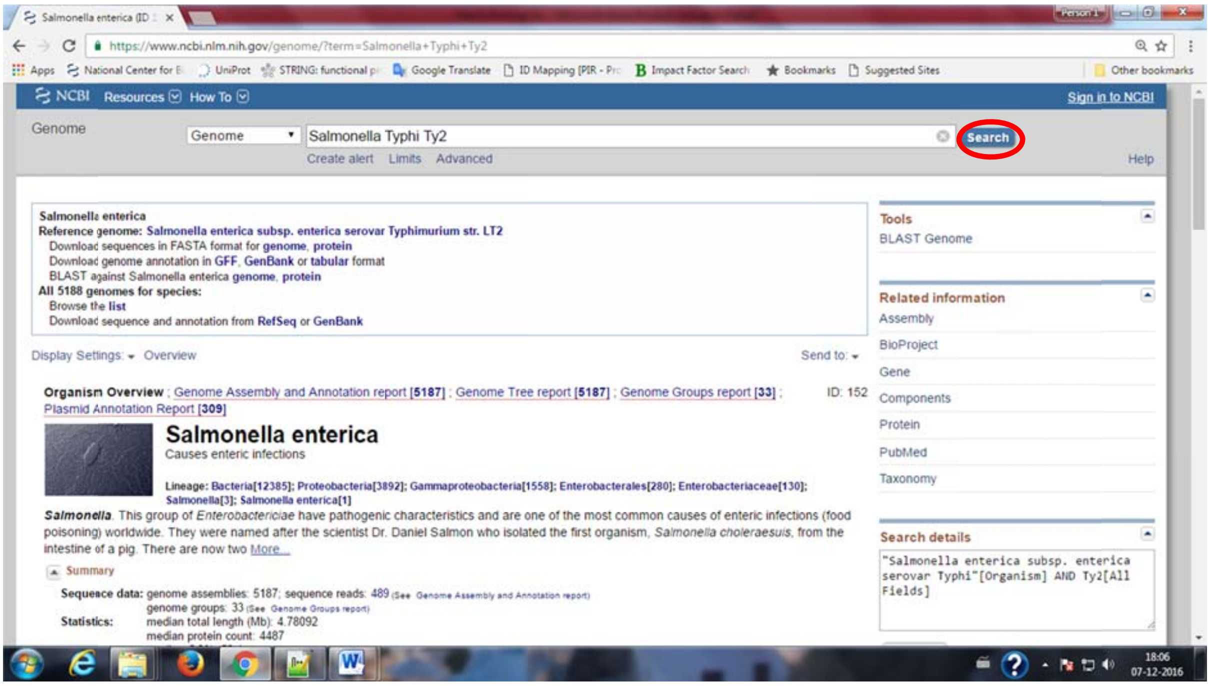
Instruction for parsing protein coding table from coding sequence (CDS)

As we have mentioned in HelpForPredictingsRNAsByBarmanetal.pdf (Step2) that in some cases direct protein coding table is not available in NCBI, so we need to parse all CDS from NCBI in such cases.

Ex : If you search *Salmonella* Typhi (*S.* Typhi) Ty2 in NCBI genome section, you will find protein coding table of *S.* Typhi CT18. Therefore we have to find all CDS of *S.* Typhi Ty2. You can find all CDS of *S.* Typhi Ty2 from NCBI nucleotide section. Therefore you have to copy the all CDS related information and paste in notepad or notepad++ or editplus and save in .txt format.

Finally we have developed ParsingCDsFromCompleteGenome.r code to parse all the forward coding start and stop positions of *S.* Typhi Ty2. Then you have to create similar protein coding table mention in **Step4** of HelpForPredictingsRNAsByBarmanetal.pdf by using start and stop positions of all CDS. The other field including Protein name, GeneID, Protein product and Length can be filled by "NA" since they are not playing any role to finding intergenic regions of *S.* Typhi Ty2.

Please see the below screenshots wisely.



Salmonella enterica
Reference genome: Salmonella enterica subsp. enterica serovar Typhimurium str. LT2
Download sequences in FASTA format for genome, protein
Download genome annotation in GFF, GenBank or tabular format
BLAST against Salmonella enterica genome, protein
All 5188 genomes for species:
Browse the list
Download sequence and annotation from RefSeq or GenBank

Organism Overview ; Genome Assembly and Annotation report [5187] ; Genome Tree report [5187] ; Genome Groups report [33] ; Plasmid Annotation Report [309]

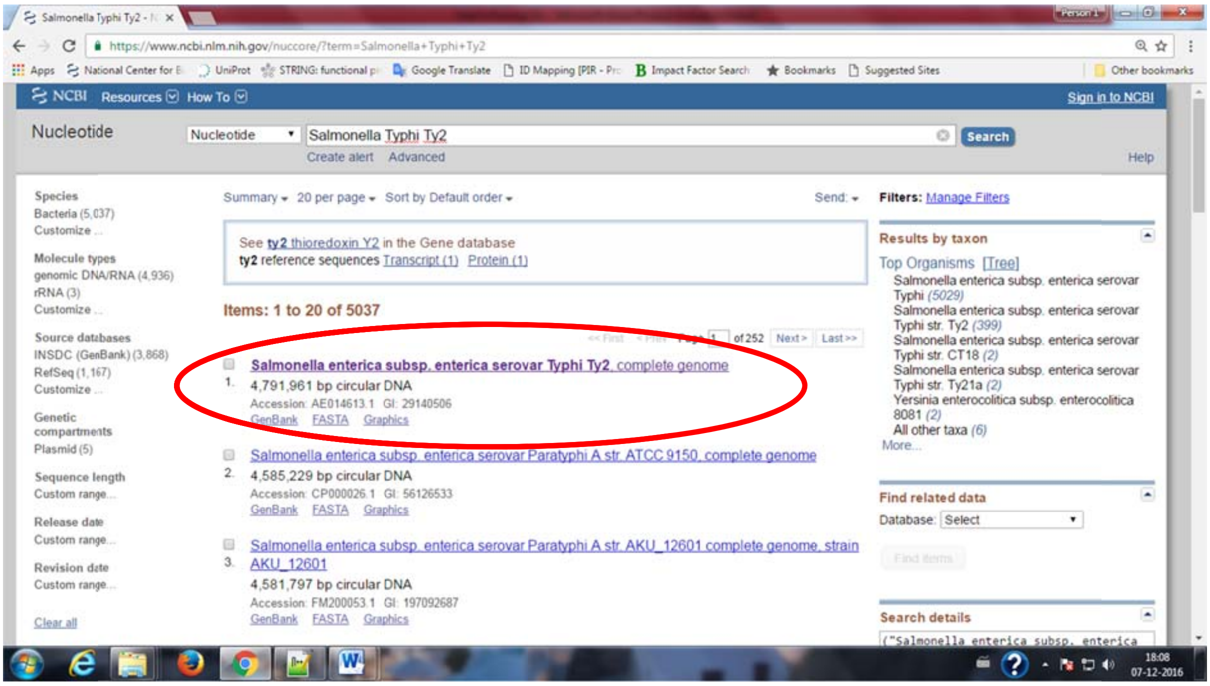
Salmonella enterica
Causes enteric infections

Lineage: Bacteria[12385]; Proteobacteria[3892]; Gammaproteobacteria[1558]; Enterobacterales[280]; Enterobacteriaceae[130]; Salmonella[3]; Salmonella enterica[1]

Salmonella. This group of *Enterobactericiae* have pathogenic characteristics and are one of the most common causes of enteric infections (food poisoning) worldwide. They were named after the scientist Dr. Daniel Salmon who isolated the first organism, *Salmonella choleraesuis*, from the intestine of a pig. There are now two More...

Sequence data: genome assemblies: 5187; sequence reads: 489 (See Genome Assembly and Annotation report)
genome groups: 33 (See Genome Groups report)
Statistics: median total length (Mb): 4.78092
median protein count: 4487

"Salmonella enterica subsp. enterica serovar Typhi"[Organism] AND Ty2[All Fields]



Nucleotide: Salmonella Typhi Ty2

Species
Bacteria (5,037)
Customize ...

Molecule types
genomic DNA/RNA (4,936)
rRNA (3)
Customize ...

Source databases
INSDC (GenBank) (3,868)
RefSeq (1,167)
Customize ...

Genetic compartments
Plasmid (5)

Sequence length
Custom range...

Release date
Custom range...

Revision date
Custom range...

Clear all

See ty2 thioredoxin Y2 in the Gene database
ty2 reference sequences Transcript (1) Protein (1)

Items: 1 to 20 of 5037

<< First < Prev Page 1 of 252 Next > Last >>

1. Salmonella enterica subsp. enterica serovar Typhi Ty2, complete genome
4,791,961 bp circular DNA
Accession: AE014613.1 GI: 29140506
GenBank FASTA Graphics

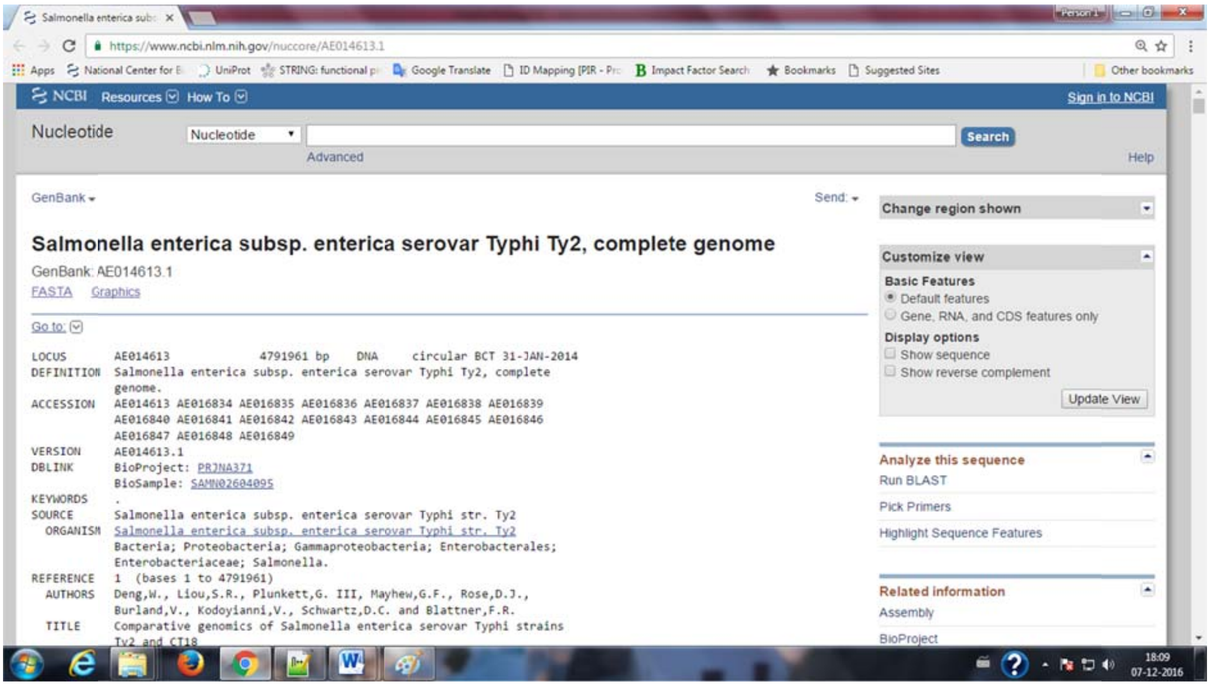
2. Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150, complete genome
4,585,229 bp circular DNA
Accession: CP000026.1 GI: 56126533
GenBank FASTA Graphics

3. Salmonella enterica subsp. enterica serovar Paratyphi A str. AKU_12601 complete genome, strain AKU_12601
4,581,797 bp circular DNA
Accession: FM200053.1 GI: 197092687
GenBank FASTA Graphics

Results by taxon
Top Organisms [Tree]
Salmonella enterica subsp. enterica serovar Typhi (5029)
Salmonella enterica subsp. enterica serovar Typhi str. Ty2 (399)
Salmonella enterica subsp. enterica serovar Typhi str. CT18 (2)
Salmonella enterica subsp. enterica serovar Typhi str. Ty21a (2)
Yersinia enterocolitica subsp. enterocolitica 8081 (2)
All other taxa (6)
More...

Find related data
Database: Select
Find items

Search details
("Salmonella enterica subsp. enterica



Salmonella enterica subsp. enterica serovar Typhi Ty2, complete genome

GenBank: AE014613.1
FASTA Graphics

Go to:

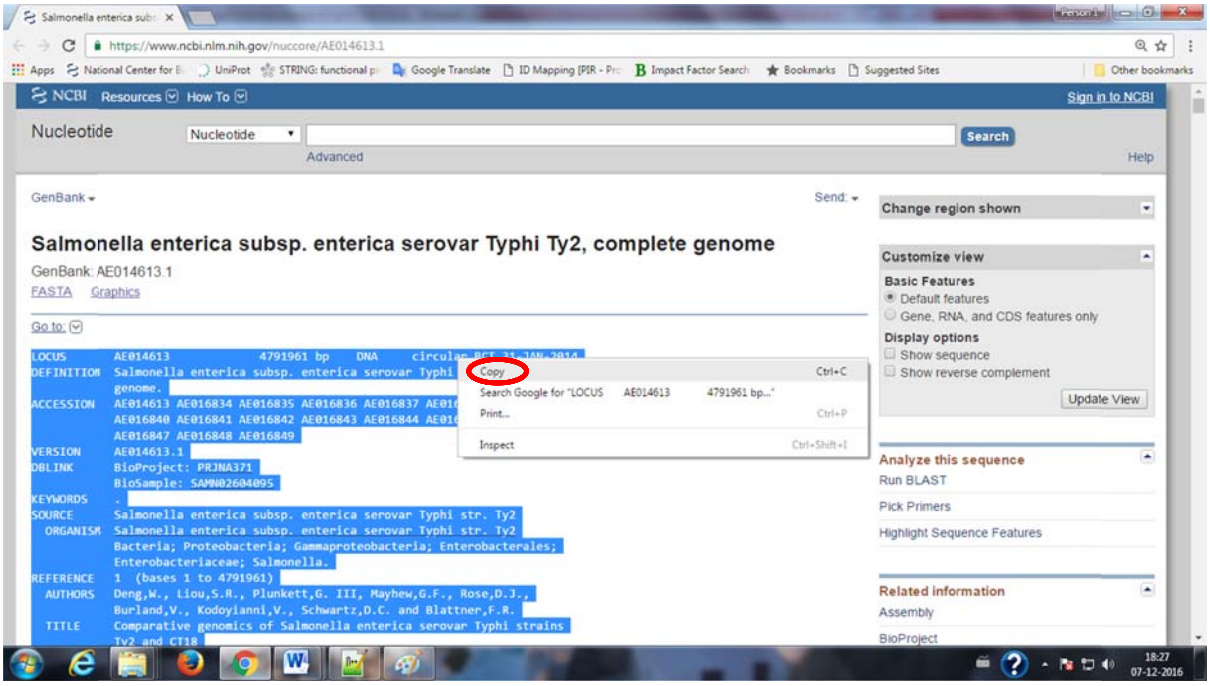
```
LOCUS       AE014613             4791961 bp    DNA     circular BCT 31-JAN-2014
DEFINITION  Salmonella enterica subsp. enterica serovar Typhi Ty2, complete
            genome.
ACCESSION   AE014613 AE016834 AE016835 AE016836 AE016837 AE016838 AE016839
            AE016840 AE016841 AE016842 AE016843 AE016844 AE016845 AE016846
            AE016847 AE016848 AE016849
VERSION     AE014613.1
DBLINK      BioProject: PRJNA371
            BioSample: SAMN02604095
KEYWORDS    .
SOURCE      Salmonella enterica subsp. enterica serovar Typhi str. Ty2
  ORGANISM  Salmonella enterica subsp. enterica serovar Typhi str. Ty2
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Salmonella.
REFERENCE   1  (bases 1 to 4791961)
  AUTHORS   Deng,W., Liou,S.R., Plunkett,G. III, Mayhew,G.F., Rose,D.J.,
            Burland,V., Kodoyianni,V., Schwartz,D.C. and Blattner,F.R.
  TITLE     Comparative genomics of Salmonella enterica serovar Typhi strains
            Ty2 and CT18
```

Change region shown

Customize view
Basic Features
● Default features
○ Gene, RNA, and CDS features only
Display options
□ Show sequence
□ Show reverse complement
Update View

Analyze this sequence
Run BLAST
Pick Primers
Highlight Sequence Features

Related information
Assembly
BioProject



Salmonella enterica subsp. enterica serovar Typhi Ty2, complete genome

GenBank: AE014613.1

FASTA Graphics

Go to:

```
LOCUS       AE014613             4791961 bp    DNA     circular BCT 31-JAN-2014
DEFINITION  Salmonella enterica subsp. enterica serovar Typhi
            genome.
ACCESSION   AE014613 AE016834 AE016835 AE016836 AE016837 AE016
            AE016840 AE016841 AE016842 AE016843 AE016844 AE016
            AE016847 AE016848 AE016849
VERSION     AE014613.1
DBLINK      BioProject: PRJNA371
            BioSample: SAMN02604095
KEYWORDS    .
SOURCE      Salmonella enterica subsp. enterica serovar Typhi str. Ty2
  ORGANISM  Salmonella enterica subsp. enterica serovar Typhi str. Ty2
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Salmonella.
REFERENCE   1  (bases 1 to 4791961)
  AUTHORS   Deng,W., Liou,S.R., Plunkett,G. III, Mayhew,G.F., Rose,D.J.,
            Burland,V., Kodoyianni,V., Schwartz,D.C. and Blattner,F.R.
  TITLE     Comparative genomics of Salmonella enterica serovar Typhi strains
            Ty2 and CT18
```

Basic Features

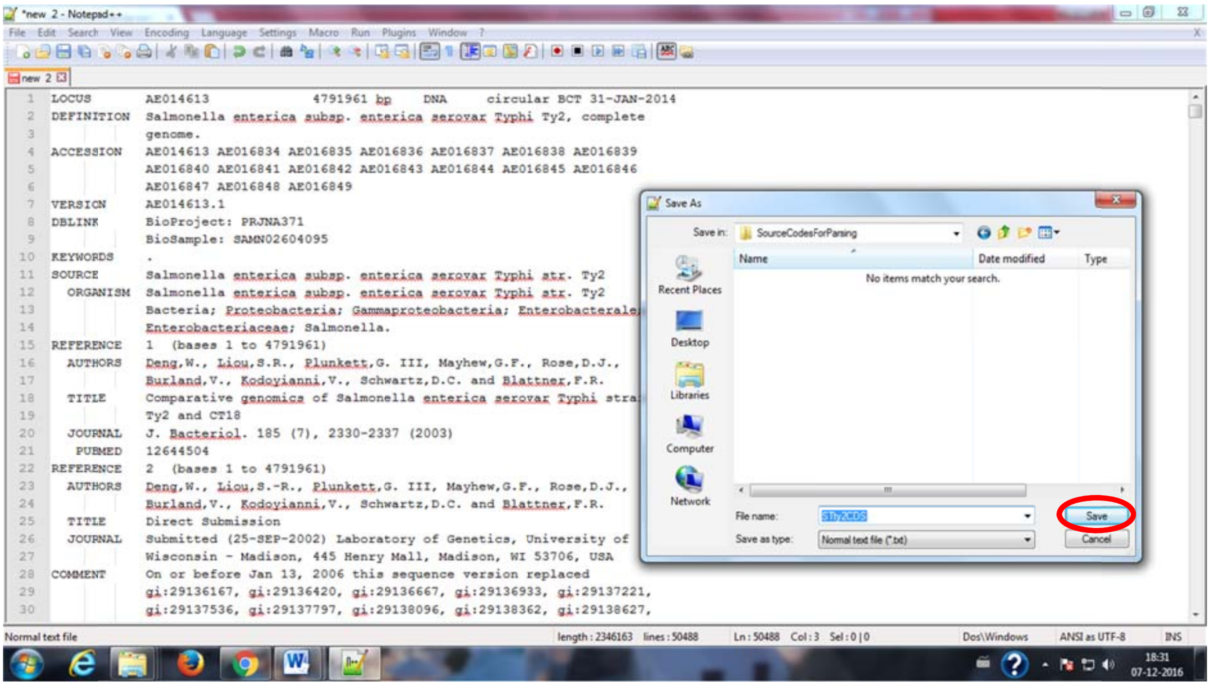
● Default features

○ Gene, RNA, and CDS features only

Display options

□ Show sequence

□ Show reverse complement



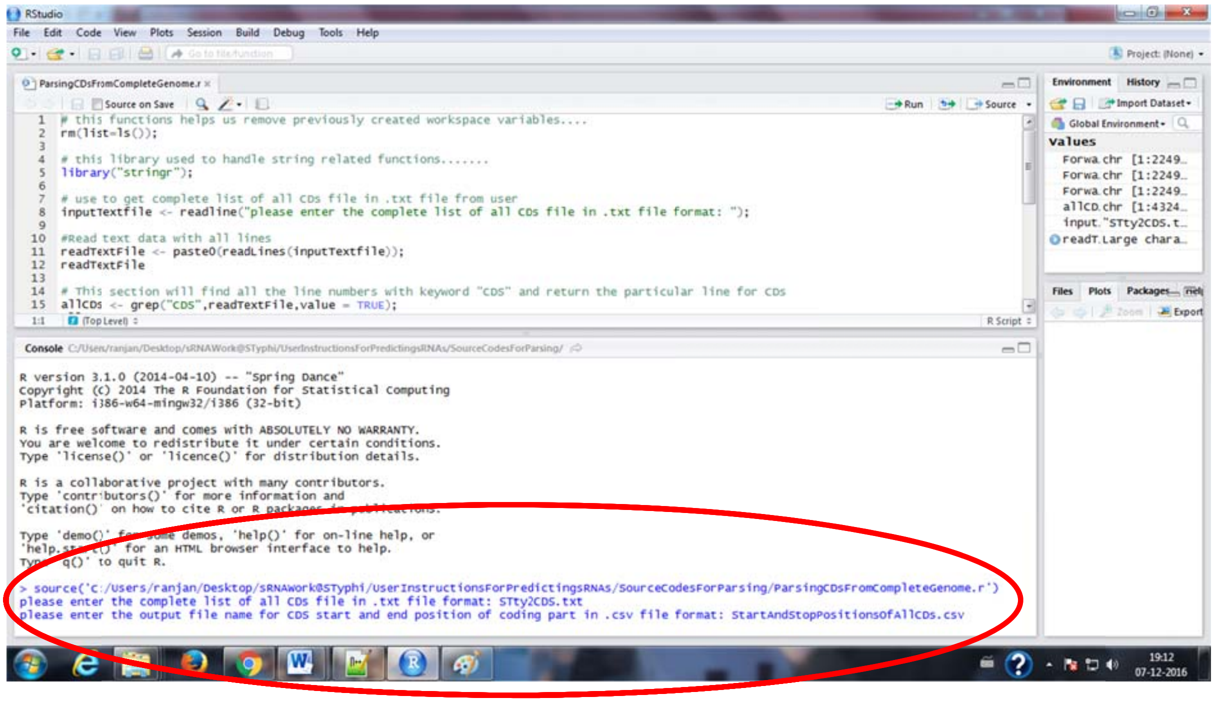
```
LOCUS       AE014613             4791961 bp    DNA     circular BCT 31-JAN-2014
DEFINITION  Salmonella enterica subsp. enterica serovar Typhi Ty2, complete
            genome.
ACCESSION   AE014613 AE016834 AE016835 AE016836 AE016837 AE016838 AE016839
            AE016840 AE016841 AE016842 AE016843 AE016844 AE016845 AE016846
            AE016847 AE016848 AE016849
VERSION     AE014613.1
DBLINK      BioProject: PRJNA371
            BioSample: SAMN02604095
KEYWORDS    .
SOURCE      Salmonella enterica subsp. enterica serovar Typhi str. Ty2
  ORGANISM  Salmonella enterica subsp. enterica serovar Typhi str. Ty2
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterale
            Enterobacteriaceae; Salmonella.
REFERENCE   1  (bases 1 to 4791961)
  AUTHORS   Deng,W., Liou,S.R., Plunkett,G. III, Mayhew,G.F., Rose,D.J.,
            Burland,V., Kodoyianni,V., Schwartz,D.C. and Blattner,F.R.
  TITLE     Comparative genomics of Salmonella enterica serovar Typhi stra
            Ty2 and CT18
  JOURNAL   J. Bacteriol. 185 (7), 2330-2337 (2003)
   PUBMED   12644504
REFERENCE   2  (bases 1 to 4791961)
  AUTHORS   Deng,W., Liou,S.-R., Plunkett,G. III, Mayhew,G.F., Rose,D.J.,
            Burland,V., Kodoyianni,V., Schwartz,D.C. and Blattner,F.R.
  TITLE     Direct Submission
  JOURNAL   Submitted (25-SEP-2002) Laboratory of Genetics, University of
            Wisconsin - Madison, 445 Henry Mall, Madison, WI 53706, USA
COMMENT     On or before Jan 13, 2006 this sequence version replaced
            gi:29136167, gi:29136420, gi:29136667, gi:29136933, gi:29137221,
            gi:29137536, gi:29137797, gi:29138096, gi:29138362, gi:29138627,
```

Save in: SourceCodesForParsing

File name: STy2CDS

Save as type: Normal text file (*.txt)

Normal text file length : 2346163 lines : 50488 Ln : 50488 Col : 3 Sel : 0 | 0 Dos\Windows ANSI as UTF-8 INS

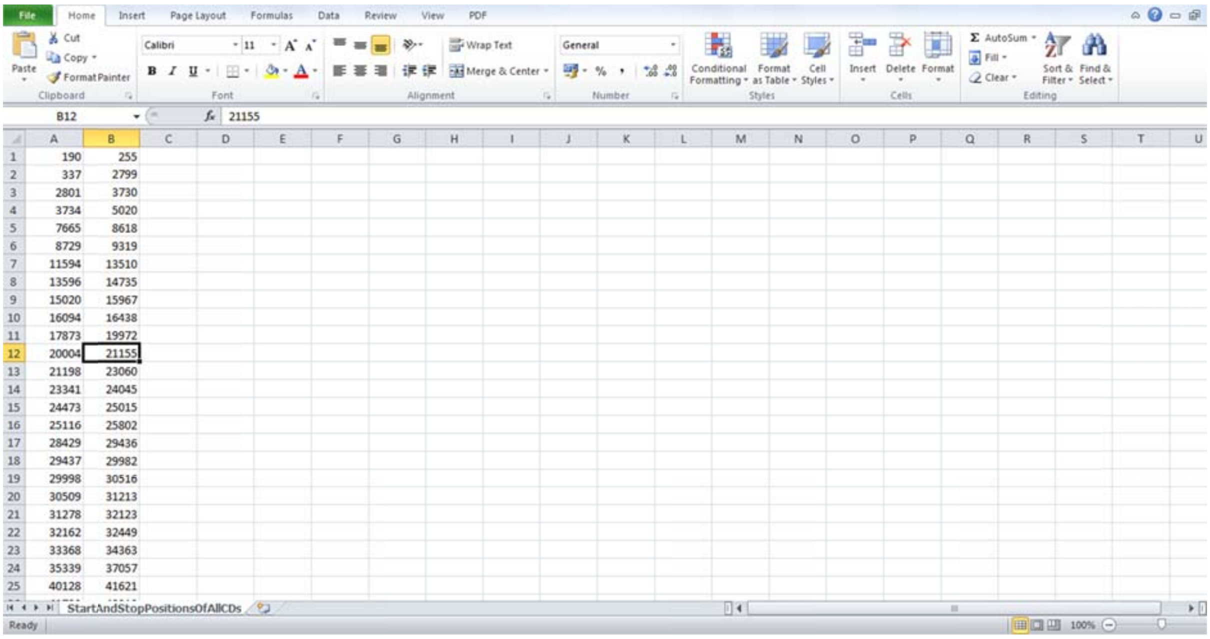


```r
# this functions helps us remove previously created workspace variables....
rm(list=ls());

# this library used to handle string related functions.......
library("stringr");

# use to get complete list of all CDs file in .txt file from user
inputTextfile <- readline("please enter the complete list of all CDs file in .txt file format: ");

#Read text data with all lines
readTextFile <- paste0(readLines(inputTextfile));
readTextFile

# This section will find all the line numbers with keyword "CDS" and return the particular line for CDs
allCDs <- grep("CDS",readTextFile,value = TRUE);
```

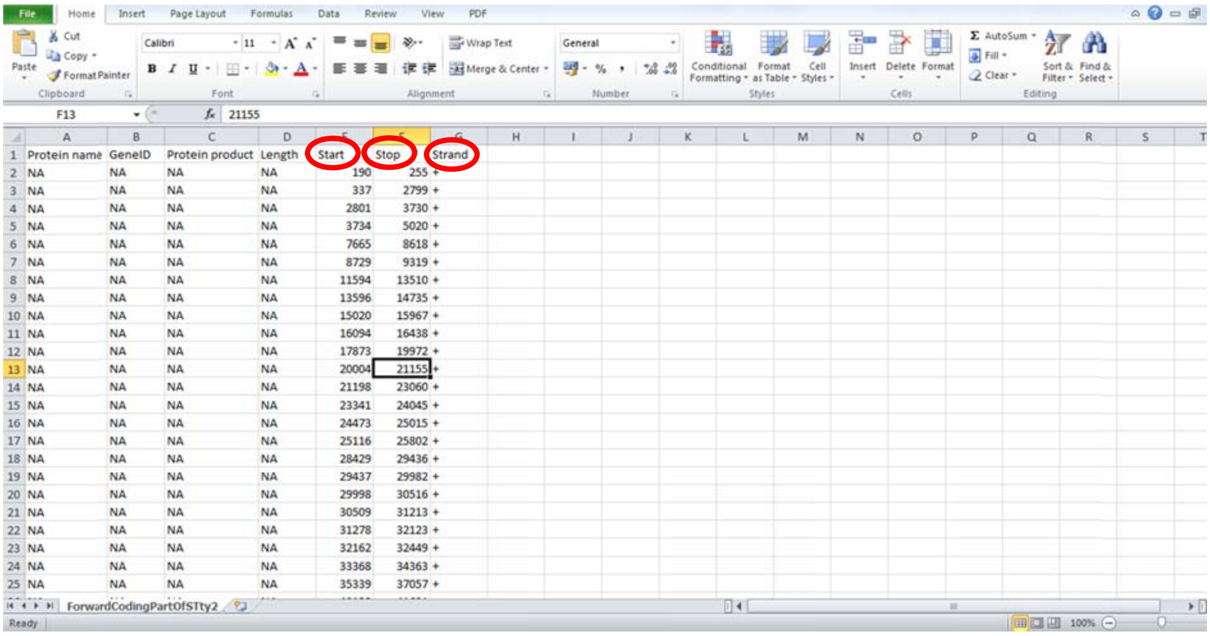
```
values
  Forwa.chr   [1:2249…
  Forwa.chr   [1:2249…
  Forwa.chr   [1:2249…
  allCD.chr   [1:4324…
  input. "STty2CDS.t…
 readT.Large chara…
```

Console C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodesForParsing/

```
R version 3.1.0 (2014-04-10) -- "Spring Dance"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

R is a collaborative project with many contributors.
Type 'contributors()' for more information and
'citation()' on how to cite R or R packages in publications.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> source('C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodesForParsing/ParsingCDsFromCompleteGenome.r')
please enter the complete list of all CDs file in .txt file format: STty2CDS.txt
please enter the output file name for CDS start and end position of coding part in .csv file format: StartAndStopPositionsOfAllCDs.csv
```



	A	B
1	190	255
2	337	2799
3	2801	3730
4	3734	5020
5	7665	8618
6	8729	9319
7	11594	13510
8	13596	14735
9	15020	15967
10	16094	16438
11	17873	19972
12	20004	21155
13	21198	23060
14	23341	24045
15	24473	25015
16	25116	25802
17	28429	29436
18	29437	29982
19	29998	30516
20	30509	31213
21	31278	32123
22	32162	32449
23	33368	34363
24	35339	37057
25	40128	41621



	A	B	C	D	E	F	G
1	Protein name	GeneID	Protein product	Length	Start	Stop	Strand
2	NA	NA	NA	NA	190	255	+
3	NA	NA	NA	NA	337	2799	+
4	NA	NA	NA	NA	2801	3730	+
5	NA	NA	NA	NA	3734	5020	+
6	NA	NA	NA	NA	7665	8618	+
7	NA	NA	NA	NA	8729	9319	+
8	NA	NA	NA	NA	11594	13510	+
9	NA	NA	NA	NA	13596	14735	+
10	NA	NA	NA	NA	15020	15967	+
11	NA	NA	NA	NA	16094	16438	+
12	NA	NA	NA	NA	17873	19972	+
13	NA	NA	NA	NA	20004	21155	+
14	NA	NA	NA	NA	21198	23060	+
15	NA	NA	NA	NA	23341	24045	+
16	NA	NA	NA	NA	24473	25015	+
17	NA	NA	NA	NA	25116	25802	+
18	NA	NA	NA	NA	28429	29436	+
19	NA	NA	NA	NA	29437	29982	+
20	NA	NA	NA	NA	29998	30516	+
21	NA	NA	NA	NA	30509	31213	+
22	NA	NA	NA	NA	31278	32123	+
23	NA	NA	NA	NA	32162	32449	+
24	NA	NA	NA	NA	33368	34363	+
25	NA	NA	NA	NA	35339	37057	+

Rest of the steps are same as mention in HelpForPredictingsRNAsByBarmanetal.pdf.