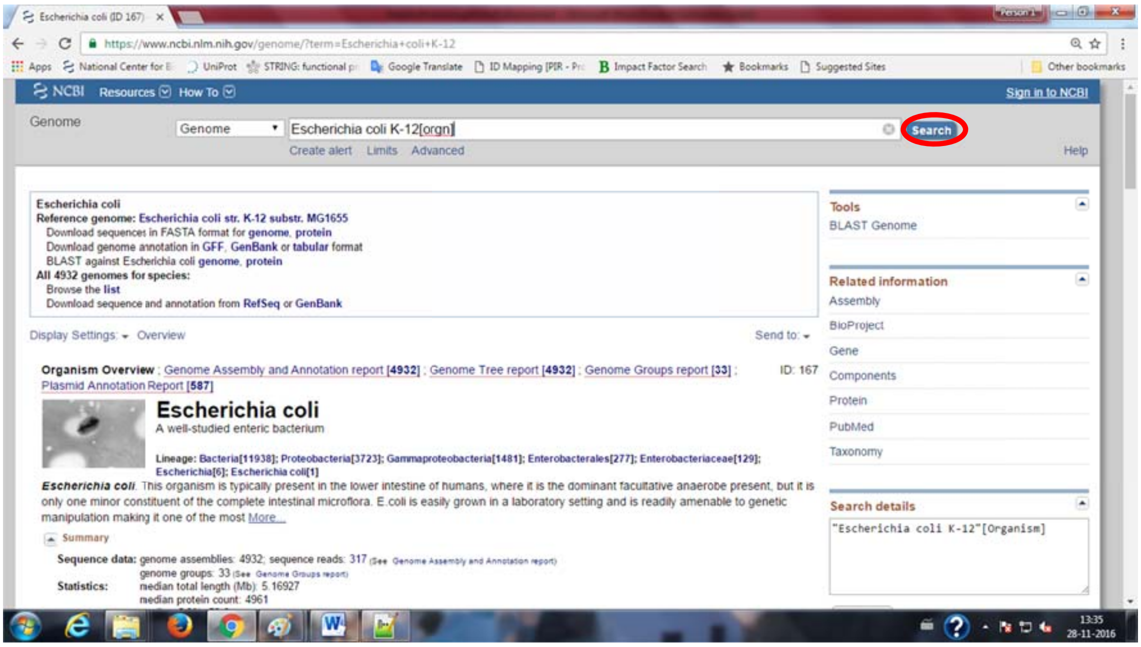
Instructions for predicting bacterial small non-coding RNAs (sRNAs) from complete genome

Step1: Please download complete genome of your bacteria of interest (.fasta extension) from NCBI nucleotide section.

> **Ex:** We have downloaded complete genome of *Escherichia coli* (*E. coli*) K-12 from https://www.ncbi.nlm.nih.gov/nuccore/556503834?report=fasta. Please see the below screenshots.



https://www.ncbi.nlm.nih.gov/genome/?term=Escherichia+coli+K-12

Genome

Genome: Escherichia coli K-12[orgn]

Escherichia coli
Reference genome: Escherichia coli str. K-12 substr. MG1655
Download sequences in FASTA format for **genome**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Escherichia coli **genome**, **protein**
All 4932 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**

Organism Overview ; Genome Assembly and Annotation report **[4932]** ; Genome Tree report **[4932]** ; Genome Groups report **[33]** ; Plasmid Annotation Report **[587]** ID: 167

Escherichia coli
A well-studied enteric bacterium

Lineage: Bacteria[11938]; Proteobacteria[3723]; Gammaproteobacteria[1481]; Enterobacterales[277]; Enterobacteriaceae[129]; Escherichia[6]; Escherichia coli[1]

Escherichia coli. This organism is typically present in the lower intestine of humans, where it is the dominant facultative anaerobe present, but it is only one minor constituent of the complete intestinal microflora. E.coli is easily grown in a laboratory setting and is readily amenable to genetic manipulation making it one of the most More...

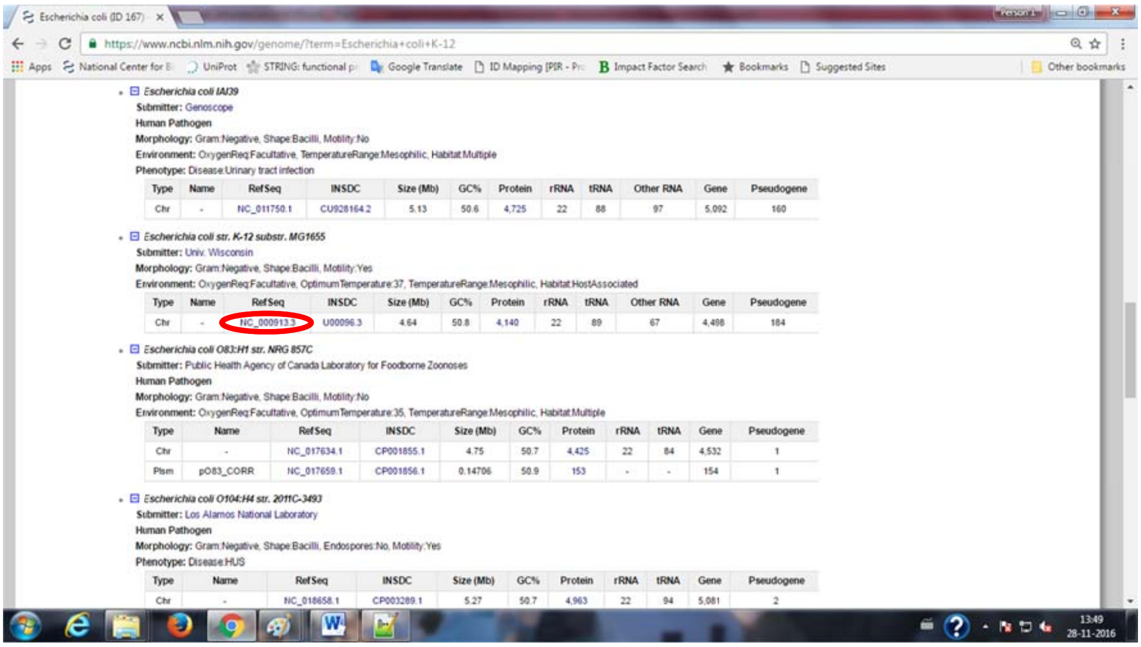
Summary

Sequence data: genome assemblies: 4932; sequence reads: 317 (See Genome Assembly and Annotation report)
genome groups: 33 (See Genome Groups report)
Statistics: median total length (Mb): 5.16927
median protein count: 4961

Tools
BLAST Genome

Related information
Assembly
BioProject
Gene
Components
Protein
PubMed
Taxonomy

Search details
"Escherichia coli K-12"[Organism]



https://www.ncbi.nlm.nih.gov/genome/?term=Escherichia+coli+K-12

- *Escherichia coli IAI39*
  Submitter: Genoscope
  Human Pathogen
  Morphology: Gram:Negative, Shape:Bacilli, Motility:No
  Environment: OxygenReq:Facultative, TemperatureRange:Mesophilic, Habitat:Multiple
  Phenotype: Disease:Urinary tract infection

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_011750.1	CU928164.2	5.13	50.6	4,725	22	88	97	5,092	160

- *Escherichia coli str. K-12 substr. MG1655*
  Submitter: Univ. Wisconsin
  Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes
  Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated

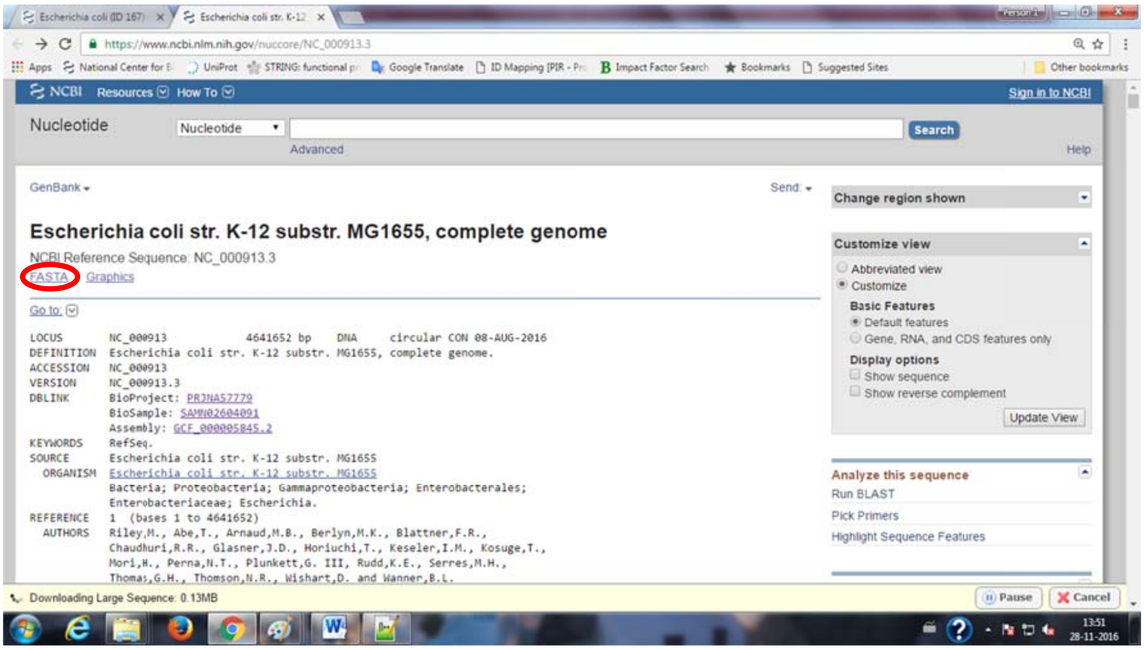
Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_000913.3	U00096.3	4.64	50.8	4,140	22	89	67	4,498	184

- *Escherichia coli O83:H1 str. NRG 857C*
  Submitter: Public Health Agency of Canada Laboratory for Foodborne Zoonoses
  Human Pathogen
  Morphology: Gram:Negative, Shape:Bacilli, Motility:No
  Environment: OxygenReq:Facultative, OptimumTemperature:35, TemperatureRange:Mesophilic, Habitat:Multiple

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_017634.1	CP001855.1	4.75	50.7	4,425	22	84	4,532	1
Plsm	pO83_CORR	NC_017659.1	CP001856.1	0.14706	50.9	153	-	-	154	1

- *Escherichia coli O104:H4 str. 2011C-3493*
  Submitter: Los Alamos National Laboratory
  Human Pathogen
  Morphology: Gram:Negative, Shape:Bacilli, Endospores:No, Motility:Yes
  Phenotype: Disease:HUS

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_018658.1	CP003289.1	5.27	50.7	4,963	22	94	5,081	2



Escherichia coli str. K-12 substr. MG1655, complete genome

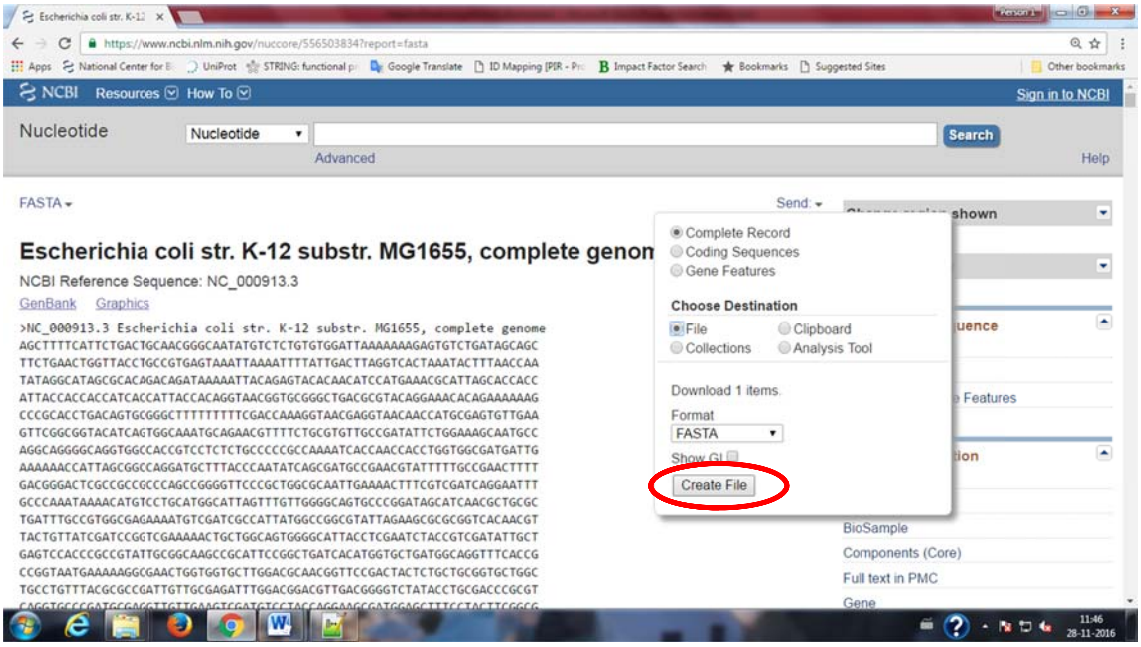
NCBI Reference Sequence: NC_000913.3

FASTA Graphics

```
LOCUS       NC_000913            4641652 bp    DNA     circular CON 08-AUG-2016
DEFINITION  Escherichia coli str. K-12 substr. MG1655, complete genome.
ACCESSION   NC_000913
VERSION     NC_000913.3
DBLINK      BioProject: PRJNA57779
            BioSample: SAMN02604091
            Assembly: GCF_000005845.2
KEYWORDS    RefSeq.
SOURCE      Escherichia coli str. K-12 substr. MG1655
  ORGANISM  Escherichia coli str. K-12 substr. MG1655
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacterales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 4641652)
  AUTHORS   Riley,M., Abe,T., Arnaud,M.B., Berlyn,M.K., Blattner,F.R.,
            Chaudhuri,R.R., Glasner,J.D., Horiuchi,T., Keseler,I.M., Kosuge,T.,
            Mori,H., Perna,N.T., Plunkett,G. III, Rudd,K.E., Serres,M.H.,
            Thomas,G.H., Thomson,N.R., Wishart,D. and Wanner,B.L.
```

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

Downloading Large Sequence: 0.13MB



Escherichia coli str. K-12 substr. MG1655, complete genon

NCBI Reference Sequence: NC_000913.3

GenBank Graphics

```
>NC_000913.3 Escherichia coli str. K-12 substr. MG1655, complete genome
AGCTTTTCATTCTGACTGCAACGGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGC
TTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAGGTCACTAAATACTTTAACCAA
TATAGGCATAGCGCACAGACAGATAAAAATTACAGAGTACACAACATCCATGAAACGCATTAGCACCACC
ATTACCACCACCATCACCATTACCACAGGTAACGGTGCGGGCTGACGCGTACAGGAAACACAGAAAAAAG
CCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACCATGCGAGTGTTGAA
GTTCGGCGGTACATCAGTGGCAAATGCAGAACGTTTTCTGCGTGTTGCCGATATTCTGGAAAGCAATGCC
AGGCAGGGGCAGGTGGCCACCGTCCTCTCTGCCCCCGCCAAAATCACCAACCACCTGGTGGCGATGATTG
AAAAAACCATTAGCGGCCAGGATGCTTTACCCAATATCAGCGATGCCGAACGTATTTTTGCCGAACTTTT
GACGGGACTCGCCGCCCGCCCAGCCGGGGTTCCCGCTGGCGCAATTGAAAACTTTCGTCGATCAGGAATTT
GCCCAAATAAAACATGTCCTGCATGGCATTAGTTTGTTGGGGCAGTGCCCGGATAGCATCAACGCTGCGC
TGATTTGCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTATTAGAAGCGCGCGGTCACAACGT
TACTGTTATCGATCCGGTCGAAAAACTGCTGGCAGTGGGGCATTACCTCGAATCTACCGTCGATATTGCT
GAGTCCACCCGCCGTATTGCGGCAAGCCGCATTCCGGCTGATCACATGGTGCTGATGGCAGGTTTCACCG
CCGGTAATGAAAAAGGCGAACTGGTGGTGCTTGGACGCAACGGTTCCGACTACTCTGCTGCGGTGCTGGC
TGCCTGTTTACGCGCCGATTGTTGCGAGATTTGGACGGACGTTGACGGGGTCTATACCTGCGACCCGCGT
```

● Complete Record
○ Coding Sequences
○ Gene Features

Choose Destination
● File ○ Clipboard
○ Collections ○ Analysis Tool

Download 1 items.

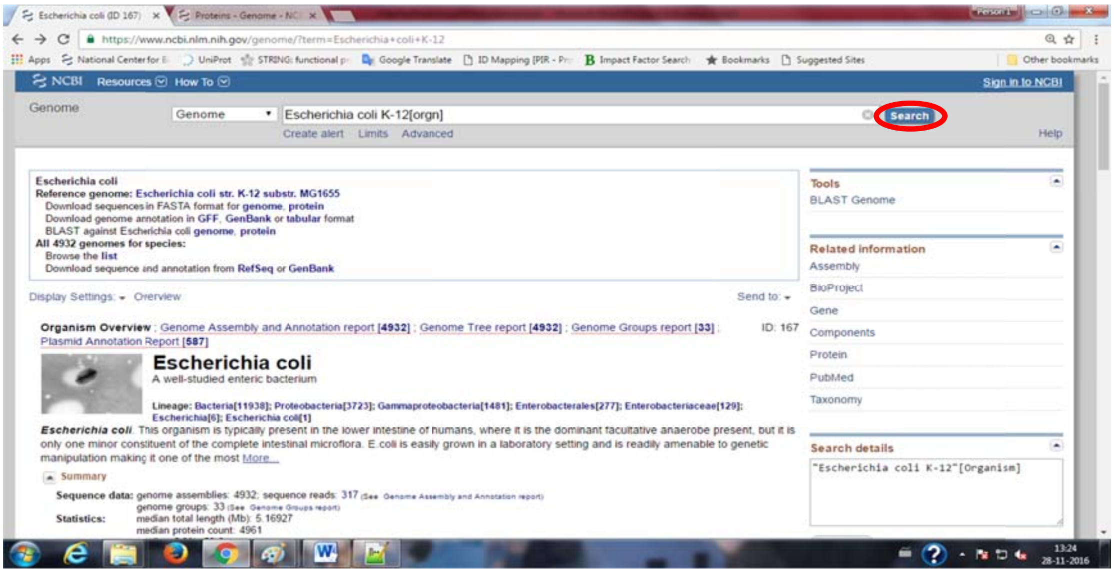
Format
FASTA

Show GI □

Create File

Step2: Please download protein coding table of your bacteria of interest from NCBI genome section. In some cases direct protein coding table is not available in NCBI. Therefore you have to parse the coding table from all coding sequence (CDS) for your bacteria of interest. Please see the HelpForParsingCDS.pdf for parsing all CDS from NCBI.

Ex: We have downloaded protein coding table of *E. coli* K-12 from https://www.ncbi.nlm.nih.gov/genome/proteins/167?genome_assembly_id=161521&gi=556503834. Please see the below screenshots.



Escherichia coli
Reference genome: Escherichia coli str. K-12 substr. MG1655
Download sequences in FASTA format for **genome**, **protein**
Download genome annotation in **GFF**, **GenBank** or **tabular** format
BLAST against Escherichia coli **genome**, **protein**
All 4932 genomes for species:
Browse the **list**
Download sequence and annotation from **RefSeq** or **GenBank**

Organism Overview ; Genome Assembly and Annotation report [**4932**] ; Genome Tree report [**4932**] ; Genome Groups report [**33**] ; Plasmid Annotation Report [**587**]

ID: 167

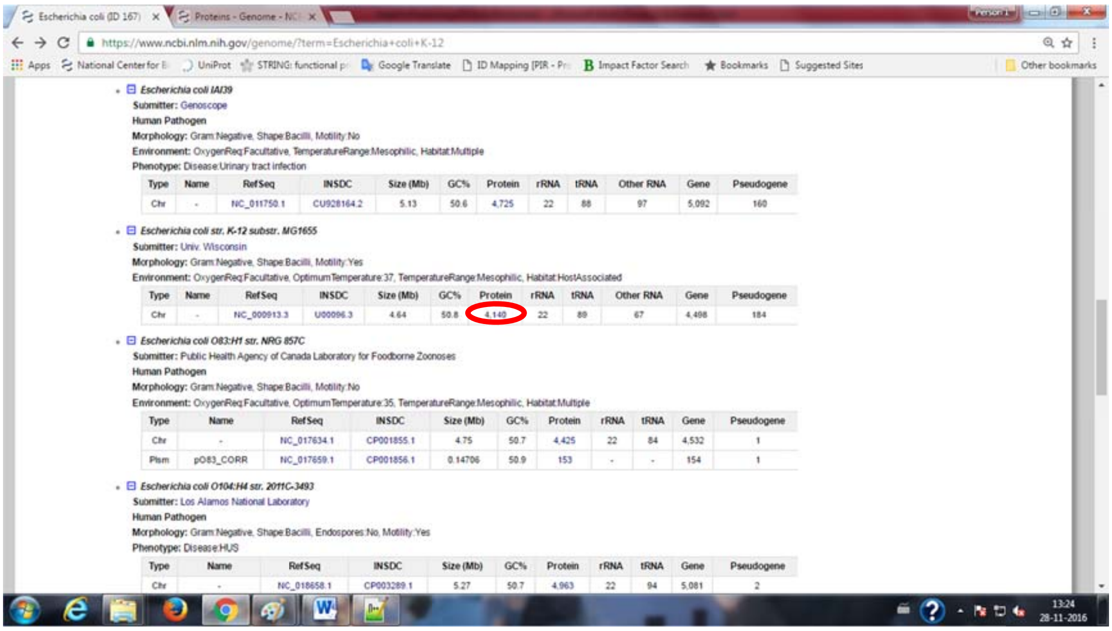
Escherichia coli
A well-studied enteric bacterium

Lineage: Bacteria[11938]; Proteobacteria[3723]; Gammaproteobacteria[1481]; Enterobacterales[277]; Enterobacteriaceae[129]; Escherichia[6]; Escherichia coli[1]

Escherichia coli. This organism is typically present in the lower intestine of humans, where it is the dominant facultative anaerobe present, but it is only one minor constituent of the complete intestinal microflora. E.coli is easily grown in a laboratory setting and is readily amenable to genetic manipulation making it one of the most More...

Summary

Sequence data: genome assemblies: 4932; sequence reads: 317 (See Genome Assembly and Annotation report)
genome groups: 33 (See Genome Groups report)
Statistics: median total length (Mb): 5.16927
median protein count: 4961



- ***Escherichia coli IAI39***

Submitter: Genoscope
Human Pathogen
Morphology: Gram:Negative, Shape:Bacilli, Motility:No
Environment: OxygenReq:Facultative, TemperatureRange:Mesophilic, Habitat:Multiple
Phenotype: Disease:Urinary tract infection

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_011750.1	CU928164.2	5.13	50.6	4,725	22	88	97	5,092	160

- ***Escherichia coli str. K-12 substr. MG1655***

Submitter: Univ. Wisconsin
Morphology: Gram:Negative, Shape:Bacilli, Motility:Yes
Environment: OxygenReq:Facultative, OptimumTemperature:37, TemperatureRange:Mesophilic, Habitat:HostAssociated

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Other RNA	Gene	Pseudogene
Chr	-	NC_000913.3	U00096.3	4.64	50.8	4,140	22	89	67	4,498	184

- ***Escherichia coli O83:H1 str. NRG 857C***

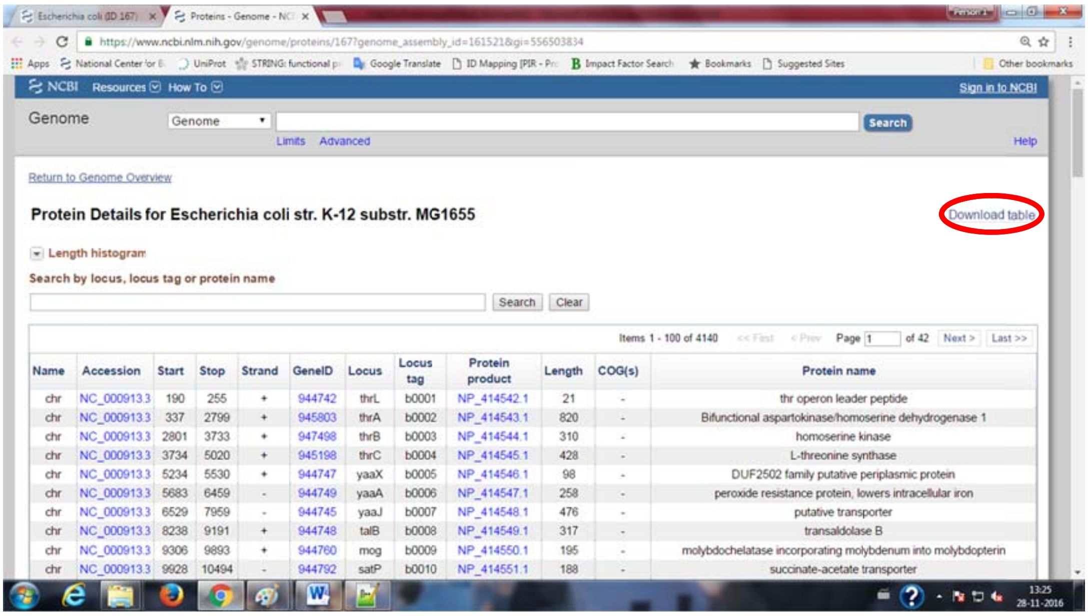
Submitter: Public Health Agency of Canada Laboratory for Foodborne Zoonoses
Human Pathogen
Morphology: Gram:Negative, Shape:Bacilli, Motility:No
Environment: OxygenReq:Facultative, OptimumTemperature:35, TemperatureRange:Mesophilic, Habitat:Multiple

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_017634.1	CP001855.1	4.75	50.7	4,425	22	84	4,532	1
Plsm	pO83_CORR	NC_017659.1	CP001856.1	0.14706	50.9	153	-	-	154	1

- ***Escherichia coli O104:H4 str. 2011C-3493***

Submitter: Los Alamos National Laboratory
Human Pathogen
Morphology: Gram:Negative, Shape:Bacilli, Endospores:No, Motility:Yes
Phenotype: Disease:HUS

Type	Name	RefSeq	INSDC	Size (Mb)	GC%	Protein	rRNA	tRNA	Gene	Pseudogene
Chr	-	NC_018658.1	CP003289.1	5.27	50.7	4,963	22	94	5,081	2



Return to Genome Overview

Protein Details for Escherichia coli str. K-12 substr. MG1655

Length histogram

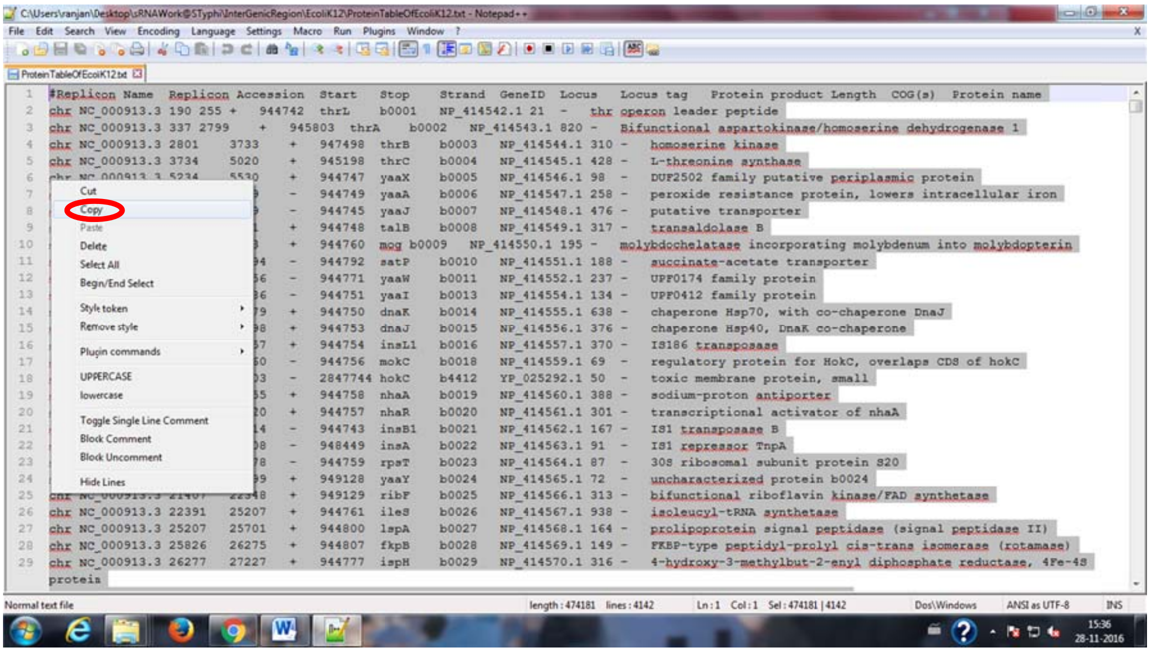
Search by locus, locus tag or protein name

Items 1 - 100 of 4140

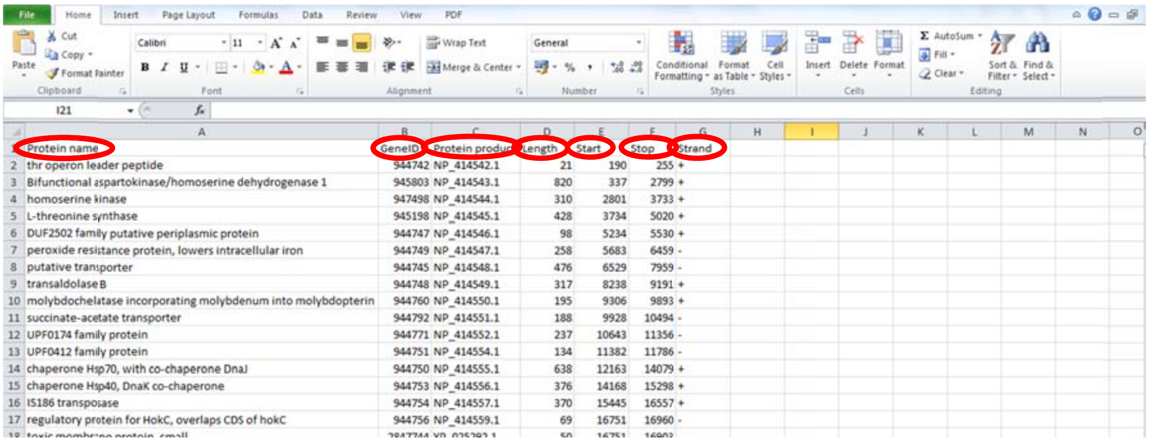
Name	Accession	Start	Stop	Strand	GeneID	Locus	Locus tag	Protein product	Length	COG(s)	Protein name
chr	NC_000913.3	190	255	+	944742	thrL	b0001	NP_414542.1	21	-	thr operon leader peptide
chr	NC_000913.3	337	2799	+	945803	thrA	b0002	NP_414543.1	820	-	Bifunctional aspartokinase/homoserine dehydrogenase 1
chr	NC_000913.3	2801	3733	+	947498	thrB	b0003	NP_414544.1	310	-	homoserine kinase
chr	NC_000913.3	3734	5020	+	945198	thrC	b0004	NP_414545.1	428	-	L-threonine synthase
chr	NC_000913.3	5234	5530	+	944747	yaaX	b0005	NP_414546.1	98	-	DUF2502 family putative periplasmic protein
chr	NC_000913.3	5683	6459	-	944749	yaaA	b0006	NP_414547.1	258	-	peroxide resistance protein, lowers intracellular iron
chr	NC_000913.3	6529	7959	-	944745	yaaJ	b0007	NP_414548.1	476	-	putative transporter
chr	NC_000913.3	8238	9191	+	944748	talB	b0008	NP_414549.1	317	-	transaldolase B
chr	NC_000913.3	9306	9893	+	944760	mog	b0009	NP_414550.1	195	-	molybdochelatase incorporating molybdenum into molybdopterin
chr	NC_000913.3	9928	10494	-	944792	satP	b0010	NP_414551.1	188	-	succinate-acetate transporter

Step3: Please copy all the tab delimitated .txt file information and paste it in Microsoft excel file (.xlsx). Consider only Protein name, GeneID, Protein product, Length, Start, Stop and Strand columns.

Ex: Please see the below screenshots.


C:\Users\ranjan\Desktop\sRNAWork@STyphi\InterGenicRegion\EcoliK12\ProteinTableOfEcoliK12.txt - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
ProteinTableOfEcoliK12.txt
Cut
Copy
Paste
Delete
Select All
Begin/End Select
Style token
Remove style
Plugin commands
UPPERCASE
lowercase
Toggle Single Line Comment
Block Comment
Block Uncomment
Hide Lines
Normal text file
length : 474181 lines : 4142
Ln : 1 Col : 1 Sel : 474181 | 4142
Dos\Windows
ANSI as UTF-8
INS
15:36
28-11-2016

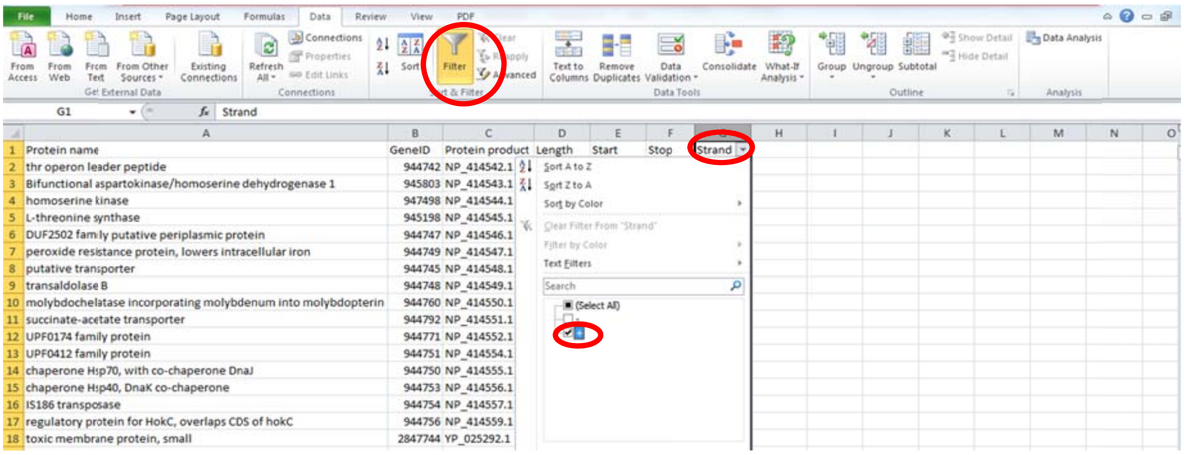
```
#Replicon Name  Replicon Accession  Start   Stop    Strand  GeneID  Locus   Locus tag   Protein product Length  COG(s)  Protein name
chr NC_000913.3 190 255 +   944742  thrL    b0001   NP_414542.1 21  -   thr operon leader peptide
chr NC_000913.3 337 2799    +   945803  thrA    b0002   NP_414543.1 820 -   Bifunctional aspartokinase/homoserine dehydrogenase 1
chr NC_000913.3 2801    3733    +   947498  thrB    b0003   NP_414544.1 310 -   homoserine kinase
chr NC_000913.3 3734    5020    +   945198  thrC    b0004   NP_414545.1 428 -   L-threonine synthase
chr NC_000913.3 5234    5530    +   944747  yaaX    b0005   NP_414546.1 98  -   DUF2502 family putative periplasmic protein
                        9   -   944749  yaaA    b0006   NP_414547.1 258 -   peroxide resistance protein, lowers intracellular iron
                        9   -   944745  yaaJ    b0007   NP_414548.1 476 -   putative transporter
                        1   +   944748  talB    b0008   NP_414549.1 317 -   transaldolase B
                        3   +   944760  mog b0009   NP_414550.1 195 -   molybdochelatase incorporating molybdenum into molybdopterin
                       94   -   944792  satP    b0010   NP_414551.1 188 -   succinate-acetate transporter
                       56   -   944771  yaaW    b0011   NP_414552.1 237 -   UPF0174 family protein
                       36   -   944751  yaaI    b0013   NP_414554.1 134 -   UPF0412 family protein
                       79   +   944750  dnaK    b0014   NP_414555.1 638 -   chaperone Hsp70, with co-chaperone DnaJ
                       98   +   944753  dnaJ    b0015   NP_414556.1 376 -   chaperone Hsp40, DnaK co-chaperone
                       57   +   944754  insL1   b0016   NP_414557.1 370 -   IS186 transposase
                       50   -   944756  mokC    b0018   NP_414559.1 69  -   regulatory protein for HokC, overlaps CDS of hokC
                       03   -   2847744 hokC    b4412   YP_025292.1 50  -   toxic membrane protein, small
                       55   +   944758  nhaA    b0019   NP_414560.1 388 -   sodium-proton antiporter
                       20   +   944757  nhaR    b0020   NP_414561.1 301 -   transcriptional activator of nhaA
                       14   -   944743  insB1   b0021   NP_414562.1 167 -   IS1 transposase B
                       08   -   948449  insA    b0022   NP_414563.1 91  -   IS1 repressor TnpA
                       78   -   944759  rpsT    b0023   NP_414564.1 87  -   30S ribosomal subunit protein S20
                       99   +   949128  yaaY    b0024   NP_414565.1 72  -   uncharacterized protein b0024
chr NC_000913.3 21407   22348   +   949129  ribF    b0025   NP_414566.1 313 -   bifunctional riboflavin kinase/FAD synthetase
chr NC_000913.3 22391   25207   +   944761  ileS    b0026   NP_414567.1 938 -   isoleucyl-tRNA synthetase
chr NC_000913.3 25207   25701   +   944800  lspA    b0027   NP_414568.1 164 -   prolipoprotein signal peptidase (signal peptidase II)
chr NC_000913.3 25826   26275   +   944807  fkpB    b0028   NP_414569.1 149 -   FKBP-type peptidyl-prolyl cis-trans isomerase (rotamase)
chr NC_000913.3 26277   27227   +   944777  ispH    b0029   NP_414570.1 316 -   4-hydroxy-3-methylbut-2-enyl diphosphate reductase, 4Fe-4S
protein
```



	A	B	C	D	E	F	G
1	Protein name	GeneID	Protein produc	Length	Start	Stop	Strand
2	thr operon leader peptide	944742	NP_414542.1	21	190	255	+
3	Bifunctional aspartokinase/homoserine dehydrogenase 1	945803	NP_414543.1	820	337	2799	+
4	homoserine kinase	947498	NP_414544.1	310	2801	3733	+
5	L-threonine synthase	945198	NP_414545.1	428	3734	5020	+
6	DUF2502 family putative periplasmic protein	944747	NP_414546.1	98	5234	5530	+
7	peroxide resistance protein, lowers intracellular iron	944749	NP_414547.1	258	5683	6459	-
8	putative transporter	944745	NP_414548.1	476	6529	7959	-
9	transaldolase B	944748	NP_414549.1	317	8238	9191	+
10	molybdochelatase incorporating molybdenum into molybdopterin	944760	NP_414550.1	195	9306	9893	+
11	succinate-acetate transporter	944792	NP_414551.1	188	9928	10494	-
12	UPF0174 family protein	944771	NP_414552.1	237	10643	11356	-
13	UPF0412 family protein	944751	NP_414554.1	134	11382	11786	-
14	chaperone Hsp70, with co-chaperone DnaJ	944750	NP_414555.1	638	12163	14079	+
15	chaperone Hsp40, DnaK co-chaperone	944753	NP_414556.1	376	14168	15298	+
16	IS186 transposase	944754	NP_414557.1	370	15445	16557	+
17	regulatory protein for HokC, overlaps CDS of hokC	944756	NP_414559.1	69	16751	16960	-
18	toxic membrane protein, small	2847744	YP_025292.1	50	16751	16903	

Step4: Please consider only forward strand since the proposed method has been developed on forward strand specific. Filter the “Strand” column by “+” sign only.

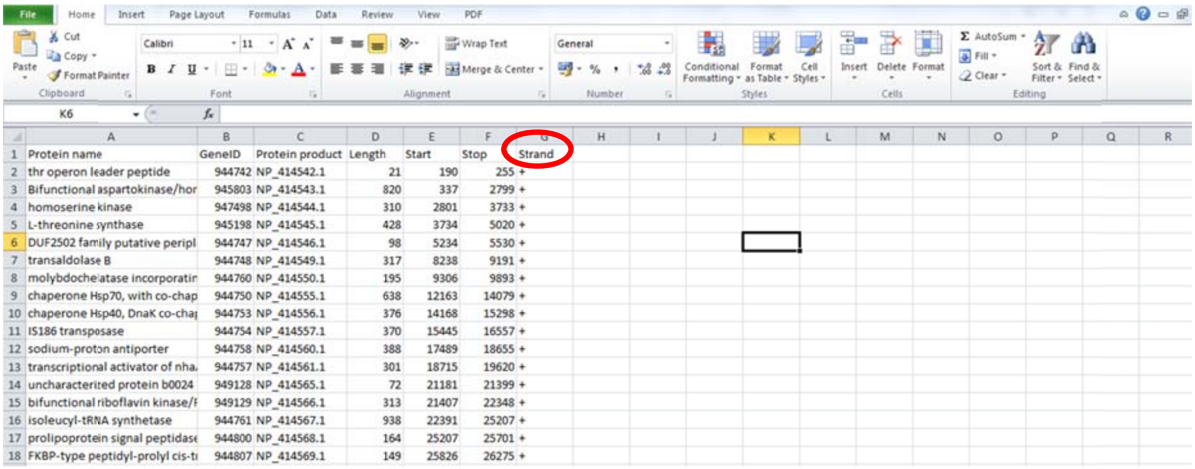
Ex: Please see the below screenshots.



G1 Strand

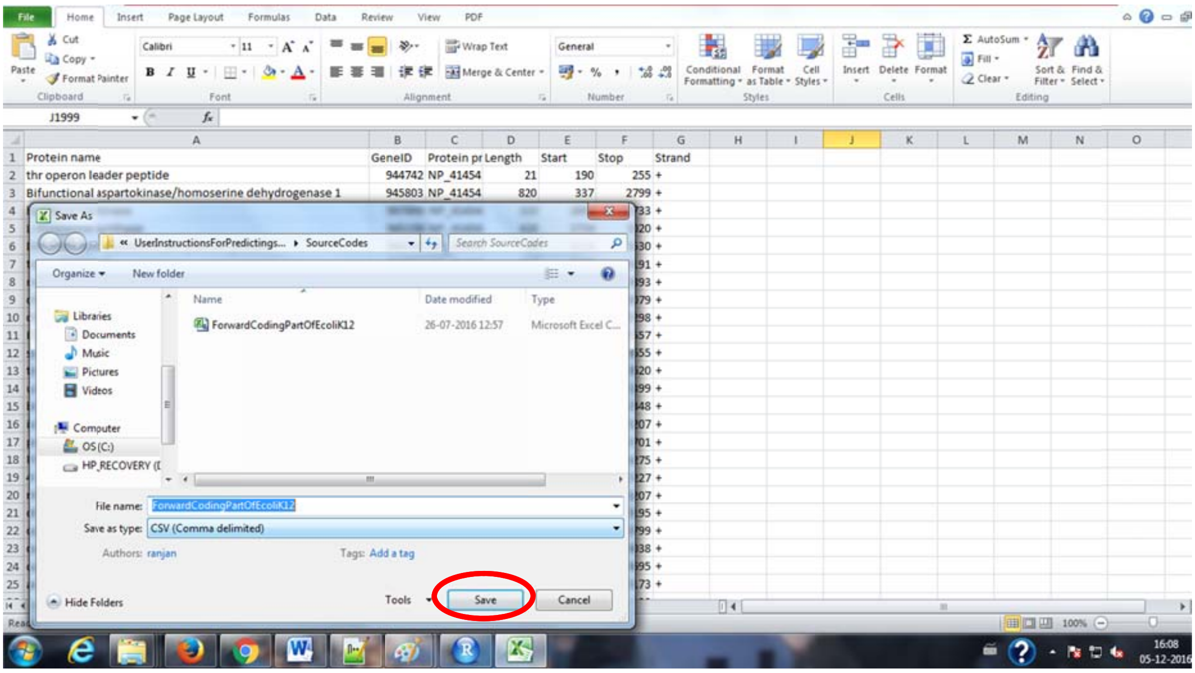
	A	B	C	D	E	F	G
1	Protein name	GeneID	Protein product	Length	Start	Stop	Strand
2	thr operon leader peptide	944742	NP_414542.1				
3	Bifunctional aspartokinase/homoserine dehydrogenase 1	945803	NP_414543.1				
4	homoserine kinase	947498	NP_414544.1				
5	L-threonine synthase	945198	NP_414545.1				
6	DUF2502 family putative periplasmic protein	944747	NP_414546.1				
7	peroxide resistance protein, lowers intracellular iron	944749	NP_414547.1				
8	putative transporter	944745	NP_414548.1				
9	transaldolase B	944748	NP_414549.1				
10	molybdochelatase incorporating molybdenum into molybdopterin	944760	NP_414550.1				
11	succinate-acetate transporter	944792	NP_414551.1				
12	UPF0174 family protein	944771	NP_414552.1				
13	UPF0412 family protein	944751	NP_414554.1				
14	chaperone Hsp70, with co-chaperone DnaJ	944750	NP_414555.1				
15	chaperone Hsp40, DnaK co-chaperone	944753	NP_414556.1				
16	IS186 transposase	944754	NP_414557.1				
17	regulatory protein for HokC, overlaps CDS of hokC	944756	NP_414559.1				
18	toxic membrane protein, small	2847744	YP_025292.1				

■ (Select All)
□ -
☑ +



K6

	A	B	C	D	E	F	G
1	Protein name	GeneID	Protein product	Length	Start	Stop	Strand
2	thr operon leader peptide	944742	NP_414542.1	21	190	255	+
3	Bifunctional aspartokinase/hor	945803	NP_414543.1	820	337	2799	+
4	homoserine kinase	947498	NP_414544.1	310	2801	3733	+
5	L-threonine synthase	945198	NP_414545.1	428	3734	5020	+
6	DUF2502 family putative peripl	944747	NP_414546.1	98	5234	5530	+
7	transaldolase B	944748	NP_414549.1	317	8238	9191	+
8	molybdochelatase incorporatir	944760	NP_414550.1	195	9306	9893	+
9	chaperone Hsp70, with co-chap	944750	NP_414555.1	638	12163	14079	+
10	chaperone Hsp40, DnaK co-chaj	944753	NP_414556.1	376	14168	15298	+
11	IS186 transposase	944754	NP_414557.1	370	15445	16557	+
12	sodium-proton antiporter	944758	NP_414560.1	388	17489	18655	+
13	transcriptional activator of nha.	944757	NP_414561.1	301	18715	19620	+
14	uncharacterized protein b0024	949128	NP_414565.1	72	21181	21399	+
15	bifunctional riboflavin kinase/F	949129	NP_414566.1	313	21407	22348	+
16	isoleucyl-tRNA synthetase	944761	NP_414567.1	938	22391	25207	+
17	prolipoprotein signal peptidase	944800	NP_414568.1	164	25207	25701	+
18	FKBP-type peptidyl-prolyl cis-ti	944807	NP_414569.1	149	25826	26275	+



J1999

	A	B	C	D	E	F	G
1	Protein name	GeneID	Protein pr	Length	Start	Stop	Strand
2	thr operon leader peptide	944742	NP_41454	21	190	255	+
3	Bifunctional aspartokinase/homoserine dehydrogenase 1	945803	NP_41454	820	337	2799	+

« UserInstructionsForPredictings... ▸ SourceCodes

Search SourceCodes

Name	Date modified	Type
ForwardCodingPartOfEcoliK12	26-07-2016 12:57	Microsoft Excel C...

File name: ForwardCodingPartOfEcoliK12

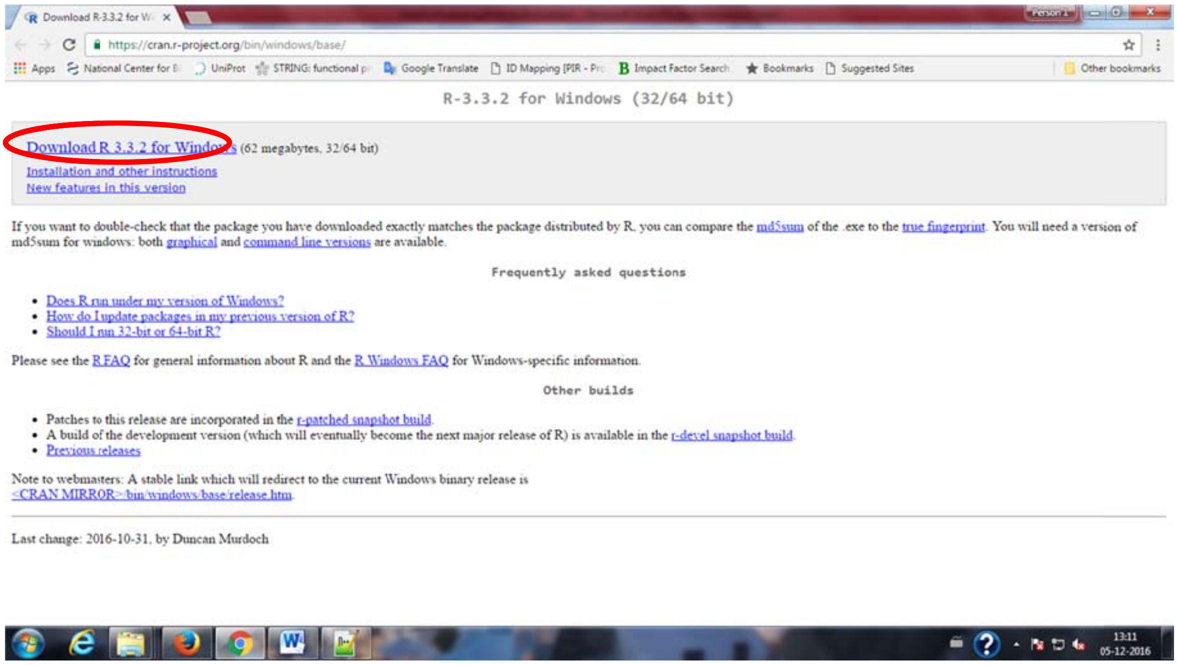
Save as type: CSV (Comma delimited)

Authors: ranjan

Tags: Add a tag

Step5: We have developed all the codes in R programing language. Therefore we request you to please download latest version of R and install it in your local desktop. We have downloaded latest version of R (3.3.2) from https://cran.r-project.org/bin/windows/base/. You can also find R for Linux and (Mac) OS X from https://cran.r-project.org/.

Ex: Please see the below screenshot.



R-3.3.2 for Windows (32/64 bit)

Download R 3.3.2 for Windows (62 megabytes, 32/64 bit)
Installation and other instructions
New features in this version

If you want to double-check that the package you have downloaded exactly matches the package distributed by R, you can compare the md5sum of the .exe to the true fingerprint. You will need a version of md5sum for windows: both graphical and command line versions are available.

Frequently asked questions

- Does R run under my version of Windows?
- How do I update packages in my previous version of R?
- Should I run 32-bit or 64-bit R?

Please see the R FAQ for general information about R and the R Windows FAQ for Windows-specific information.

Other builds

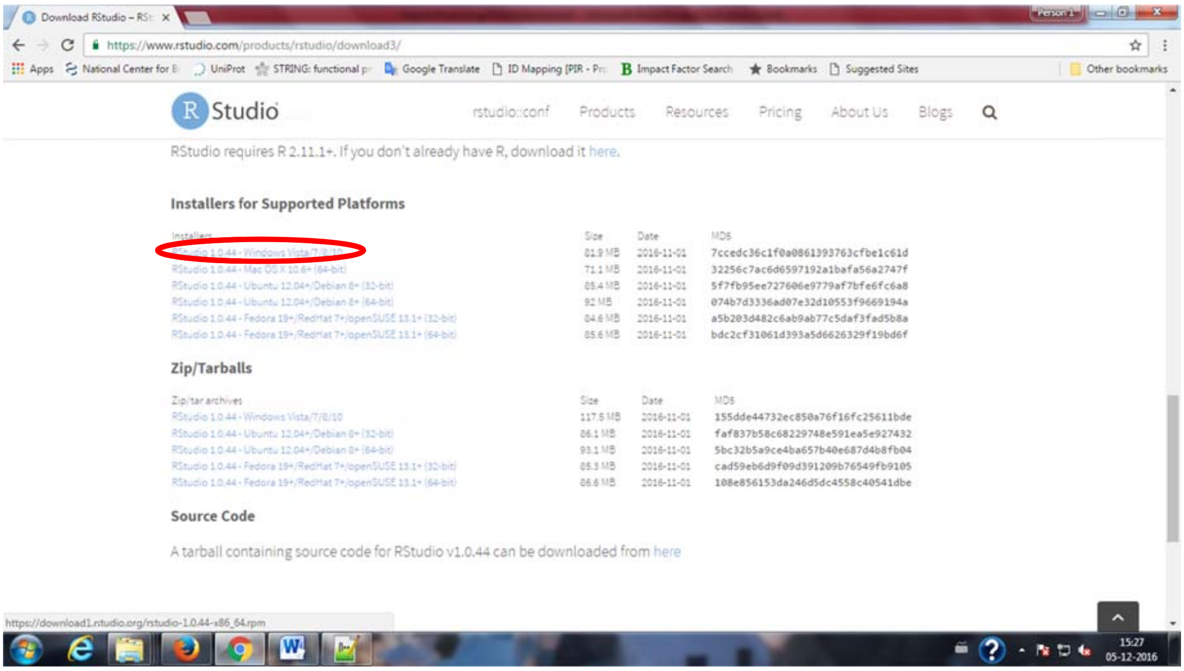
- Patches to this release are incorporated in the r-patched snapshot build.
- A build of the development version (which will eventually become the next major release of R) is available in the r-devel snapshot build.
- Previous releases

Note to webmasters: A stable link which will redirect to the current Windows binary release is
<CRAN MIRROR>/bin/windows/base/release.htm.

Last change: 2016-10-31, by Duncan Murdoch

Step6: We always prefer to write code in integrated development environment (IDE). RStudio is a free and open-source IDE for R. You can download RStudio from https://www.rstudio.com/products/rstudio/download3/.

Ex: Please see the below screenshot.



RStudio requires R 2.11.1+. If you don't already have R, download it here.

Installers for Supported Platforms

Installers	Size	Date	MD5
RStudio 1.0.44 - Windows Vista/7/8/10	81.9 MB	2016-11-01	7ccedc36c1f0a0861393763cfbe1c61d
RStudio 1.0.44 - Mac OS X 10.6+ (64-bit)	71.1 MB	2016-11-01	32256c7ac6d6597192a1bafa56a2747f
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (32-bit)	85.4 MB	2016-11-01	5f7fb95ee727606e9779af7bfe6fc6a8
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (64-bit)	92 MB	2016-11-01	074b7d3336ad07e32d10553f9669194a
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	84.6 MB	2016-11-01	a5b203d482c6ab9ab77c5daf3fad5b8a
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	85.6 MB	2016-11-01	bdc2cf31061d393a5d6626329f19bd6f

Zip/Tarballs

Zip/tar archives	Size	Date	MD5
RStudio 1.0.44 - Windows Vista/7/8/10	117.5 MB	2016-11-01	155dde44732ec850a76f16fc25611bde
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (32-bit)	86.1 MB	2016-11-01	faf837b58c68229748e591ea5e927432
RStudio 1.0.44 - Ubuntu 12.04+/Debian 8+ (64-bit)	93.1 MB	2016-11-01	5bc32b5a9ce4ba657b40e687d4b8fb04
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (32-bit)	85.3 MB	2016-11-01	cad59eb6d9f09d391209b76549fb9105
RStudio 1.0.44 - Fedora 19+/RedHat 7+/openSUSE 13.1+ (64-bit)	86.6 MB	2016-11-01	108e856153da246d5dc4558c40541dbe

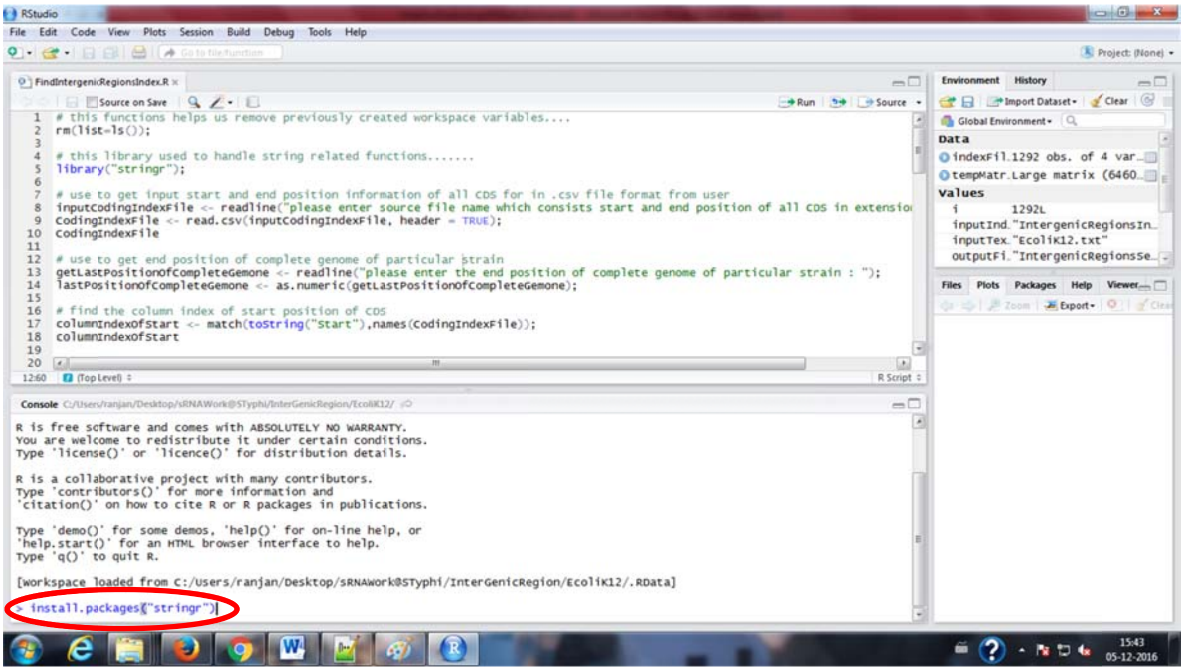
Source Code

A tarball containing source code for RStudio v1.0.44 can be downloaded from here

https://download1.rstudio.org/rstudio-1.0.44-x86_64.rpm

Step7: We have used R package ‘stringr’ for handling string operation. Therefore we request you to please install R package ‘stringr’ from console of Rstudio by using install.packages("stringr") comment.

Ex: Please see the below screenshot.


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
FindIntergenicRegionsIndex.R
Source on Save
Run
Source
12:60
(Top Level)
R Script
Console C:/Users/ranjan/Desktop/sRNAWork@STyphi/InterGenicRegion/EcoliK12/
Environment History
Import Dataset
Clear
Global Environment
Files Plots Packages Help Viewer
Zoom
Export
15:43
05-12-2016

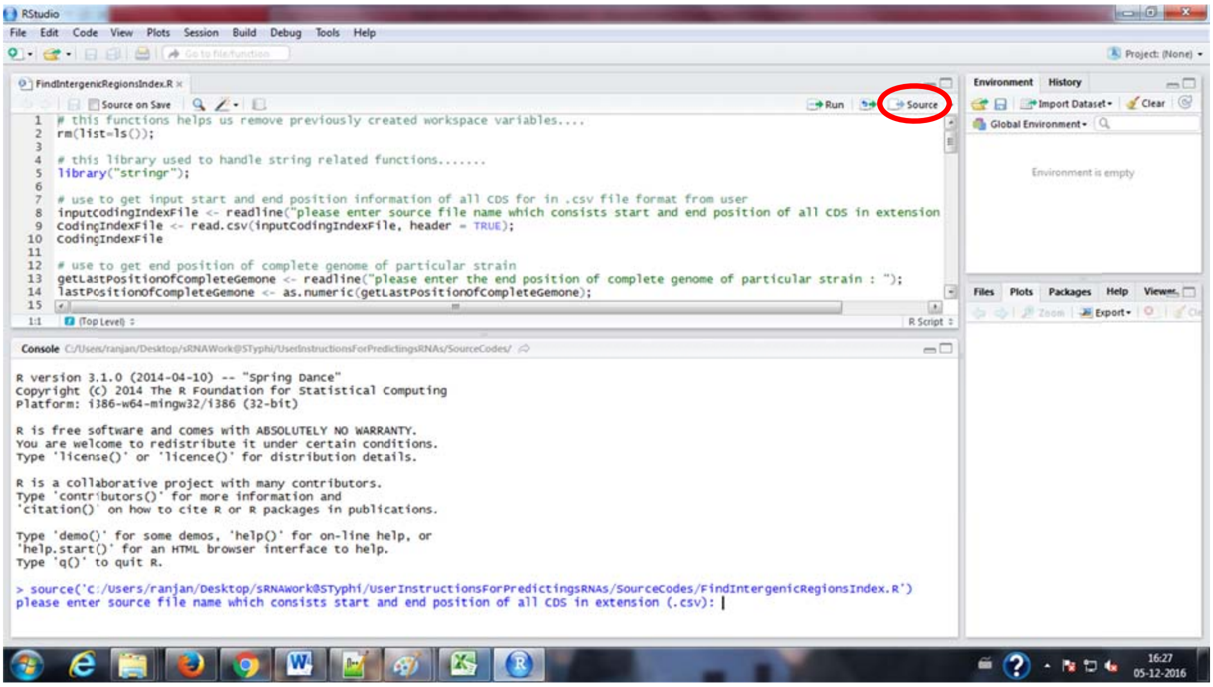
```
# this functions helps us remove previously created workspace variables....
rm(list=ls());

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

# use to get input start and end position information of all CDS for in .csv file format from user
inputCodingIndexFile <- readline("please enter source file name which consists start and end position of all CDS in extensio
CodingIndexFile <- read.csv(inputCodingIndexFile, header = TRUE);
CodingIndexFile

# use to get end position of complete genome of particular strain
getLastPositionofCompleteGemone <- readline("please enter the end position of complete genome of particular strain : ");
lastPositionofCompleteGemone <- as.numeric(getLastPositionofCompleteGemone);

# find the column index of start position of CDS
columnIndexofStart <- match(toString("Start"),names(CodingIndexFile));
columnIndexofStart
```

```
R is free scftware 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.

[Workspace loaded from C:/Users/ranjan/Desktop/sRNAWork@STyphi/InterGenicRegion/EcoliK12/.RData]

> install.packages("stringr")
```

```
Data
indexFil… 1292 obs. of 4 var…
tempMatr… Large matrix (6460…
Values
i         1292L
inputInd… "IntergenicRegionsIn…
inputTex… "EcoliK12.txt"
outputFi… "IntergenicRegionsSe…
```

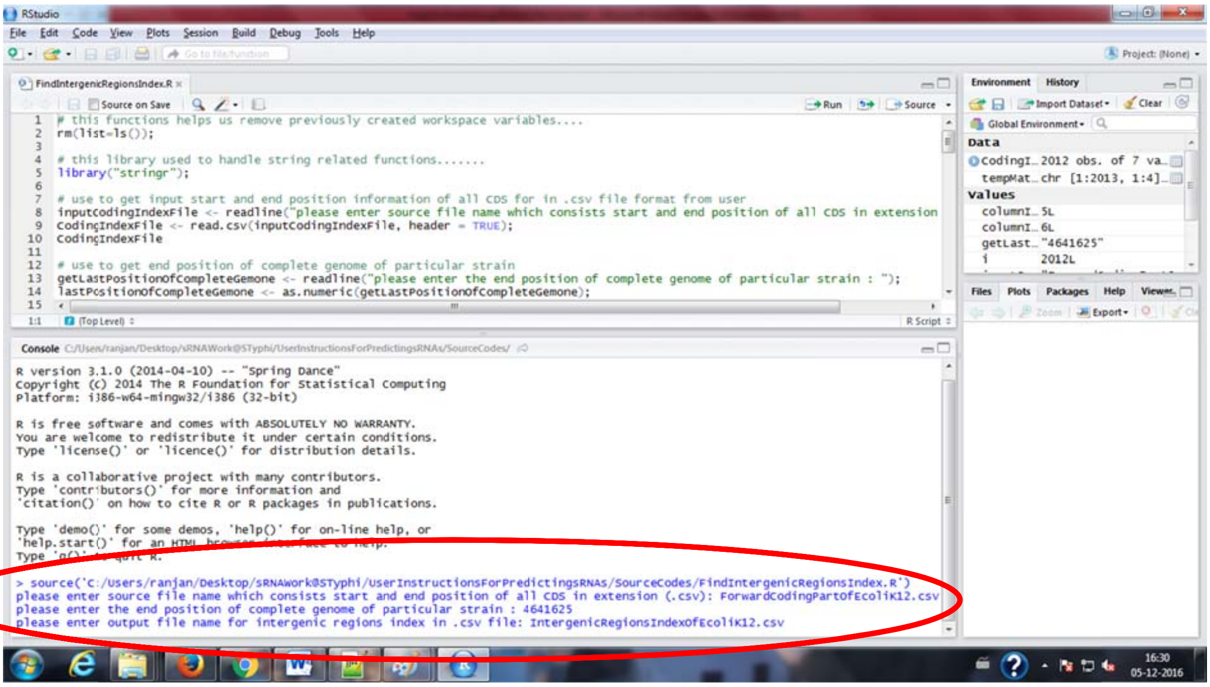
Step8: We have developed FindIntergenicRegionsIndex.r code to find all the intergenic regions of a particular strain. This code will accept input as protein coding table (created in **Step4**) and end position of complete genome of particular strain. This code will produce all the intergenic index of particular strain.

Ex: Please see the below screenshots.


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Go to file/function
Project: (None)
FindIntergenicRegionsIndex.R
Source on Save
Run
Source
1:1
(Top Level)
R Script
Console C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/
Environment History
Import Dataset
Clear
Global Environment
Environment is empty
Files Plots Packages Help Viewer
Zoom
Export
16:27
05-12-2016

```
# 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 input start and end position information of all CDS for in .csv file format from user
inputCodingIndexFile <- readline("please enter source file name which consists start and end position of all CDS in extension
CodingIndexFile <- read.csv(inputCodingIndexFile, header = TRUE);
CodingIndexFile

# use to get end position of complete genome of particular strain
getLastPositionofCompleteGemone <- readline("please enter the end position of complete genome of particular strain : ");
lastPositionofCompleteGemone <- as.numeric(getLastPositionofCompleteGemone);
```

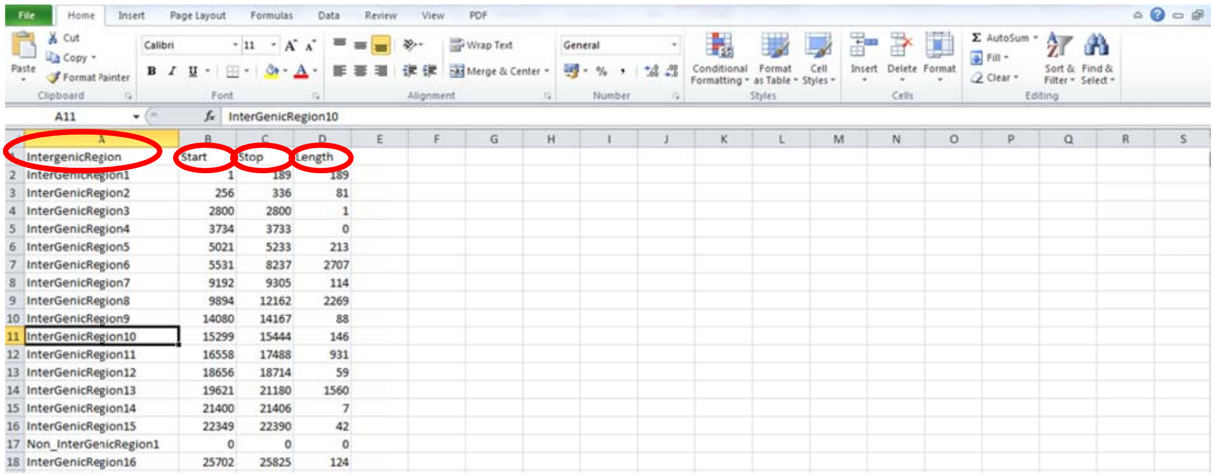
```
R version 3.1.0 (2014-04-10) -- "Spring Dance"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

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/SourceCodes/FindIntergenicRegionsIndex.R')
please enter source file name which consists start and end position of all CDS in extension (.csv): 
```


RStudio
File Edit Code View Plots Session Build Debug Tools Help
FindIntergenicRegionsIndex.R
Source on Save
Run
Source
Environment History
Import Dataset
Clear
Global Environment
Files Plots Packages Help Viewer
Export
Project: (None)
(Top Level)
R Script
Console C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/
16:30
05-12-2016

```
# 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 input start and end position information of all CDS for in .csv file format from user
inputCodingIndexFile <- readline("please enter source file name which consists start and end position of all CDS in extension
CodingIndexFile <- read.csv(inputCodingIndexFile, header = TRUE);
CodingIndexFile

# use to get end position of complete genome of particular strain
getLastPositionOfCompleteGemone <- readline("please enter the end position of complete genome of particular strain : ");
lastPositionOfCompleteGemone <- as.numeric(getLastPositionOfCompleteGemone);
```

```
Data
CodingI… 2012 obs. of 7 va…
tempMat… chr [1:2013, 1:4]…
Values
columnI… 5L
columnI… 6L
getLast… "4641625"
i        2012L
```

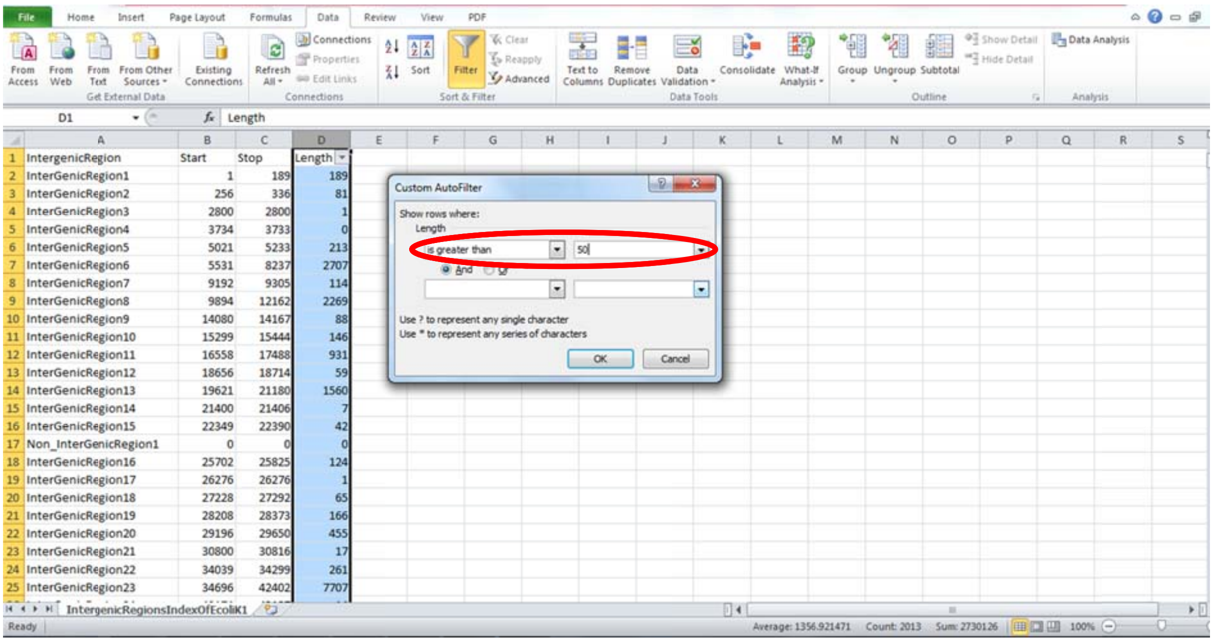
```
R version 3.1.0 (2014-04-10) -- "Spring Dance"
Copyright (C) 2014 The R Foundation for Statistical Computing
Platform: i386-w64-mingw32/i386 (32-bit)

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/SourceCodes/FindIntergenicRegionsIndex.R')
please enter source file name which consists start and end position of all CDS in extension (.csv): ForwardCodingPartOfEcoliK12.csv
please enter the end position of complete genome of particular strain : 4641625
please enter output file name for intergenic regions index in .csv file: IntergenicRegionsIndexOfEcoliK12.csv
```



IntergenicRegion	Start	Stop	Length
InterGenicRegion1	1	189	189
InterGenicRegion2	256	336	81
InterGenicRegion3	2800	2800	1
InterGenicRegion4	3734	3733	0
InterGenicRegion5	5021	5233	213
InterGenicRegion6	5531	8237	2707
InterGenicRegion7	9192	9305	114
InterGenicRegion8	9894	12162	2269
InterGenicRegion9	14080	14167	88
InterGenicRegion10	15299	15444	146
InterGenicRegion11	16558	17488	931
InterGenicRegion12	18656	18714	59
InterGenicRegion13	19621	21180	1560
InterGenicRegion14	21400	21406	7
InterGenicRegion15	22349	22390	42
Non_InterGenicRegion1	0	0	0
InterGenicRegion16	25702	25825	124

Step9: We have considered only those intergenic regions having length greater than 50nts. You can filter intergenic regions by length greater than 50.

Ex: Please see the below screenshots.



D1 Length

	A	B	C	D
1	IntergenicRegion	Start	Stop	Length
2	InterGenicRegion1	1	189	189
3	InterGenicRegion2	256	336	81
4	InterGenicRegion3	2800	2800	1
5	InterGenicRegion4	3734	3733	0
6	InterGenicRegion5	5021	5233	213
7	InterGenicRegion6	5531	8237	2707
8	InterGenicRegion7	9192	9305	114
9	InterGenicRegion8	9894	12162	2269
10	InterGenicRegion9	14080	14167	88
11	InterGenicRegion10	15299	15444	146
12	InterGenicRegion11	16558	17488	931
13	InterGenicRegion12	18656	18714	59
14	InterGenicRegion13	19621	21180	1560
15	InterGenicRegion14	21400	21406	7
16	InterGenicRegion15	22349	22390	42
17	Non_InterGenicRegion1	0	0	0
18	InterGenicRegion16	25702	25825	124
19	InterGenicRegion17	26276	26276	1
20	InterGenicRegion18	27228	27292	65
21	InterGenicRegion19	28208	28373	166
22	InterGenicRegion20	29196	29650	455
23	InterGenicRegion21	30800	30816	17
24	InterGenicRegion22	34039	34299	261
25	InterGenicRegion23	34696	42402	7707

Custom AutoFilter

Show rows where:

Length

is greater than 50

● And ○ Or

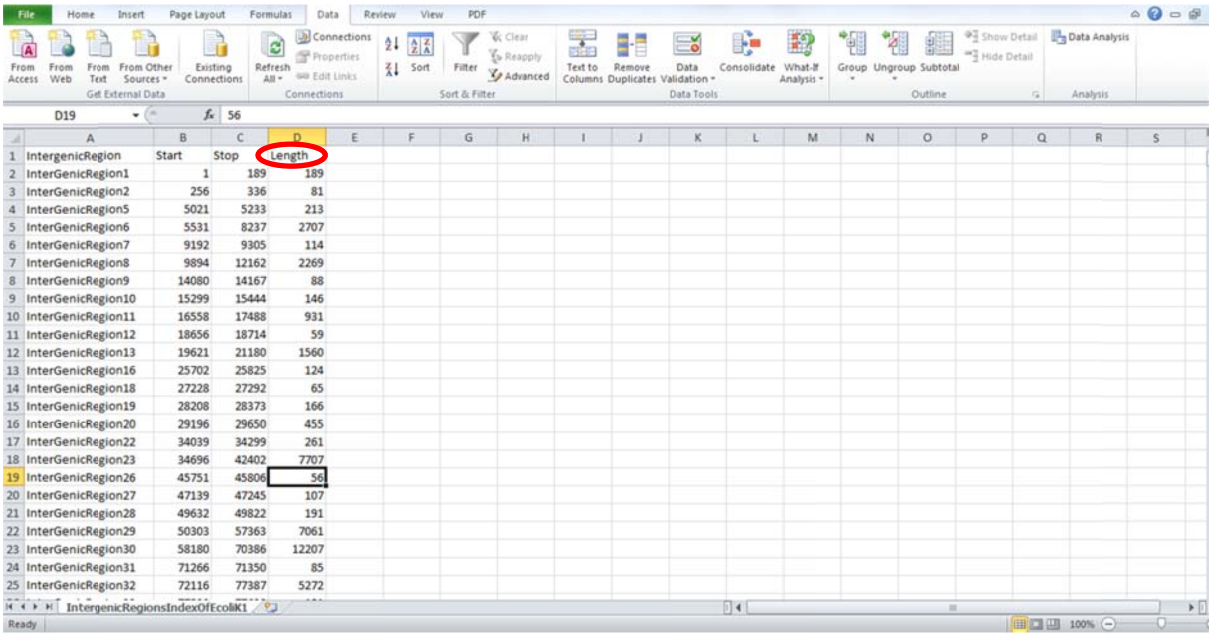
Use ? to represent any single character

Use * to represent any series of characters

OK Cancel

IntergenicRegionsIndexOfEcoliK1

Ready Average: 1356.921471 Count: 2013 Sum: 2730126 100%



D19 56

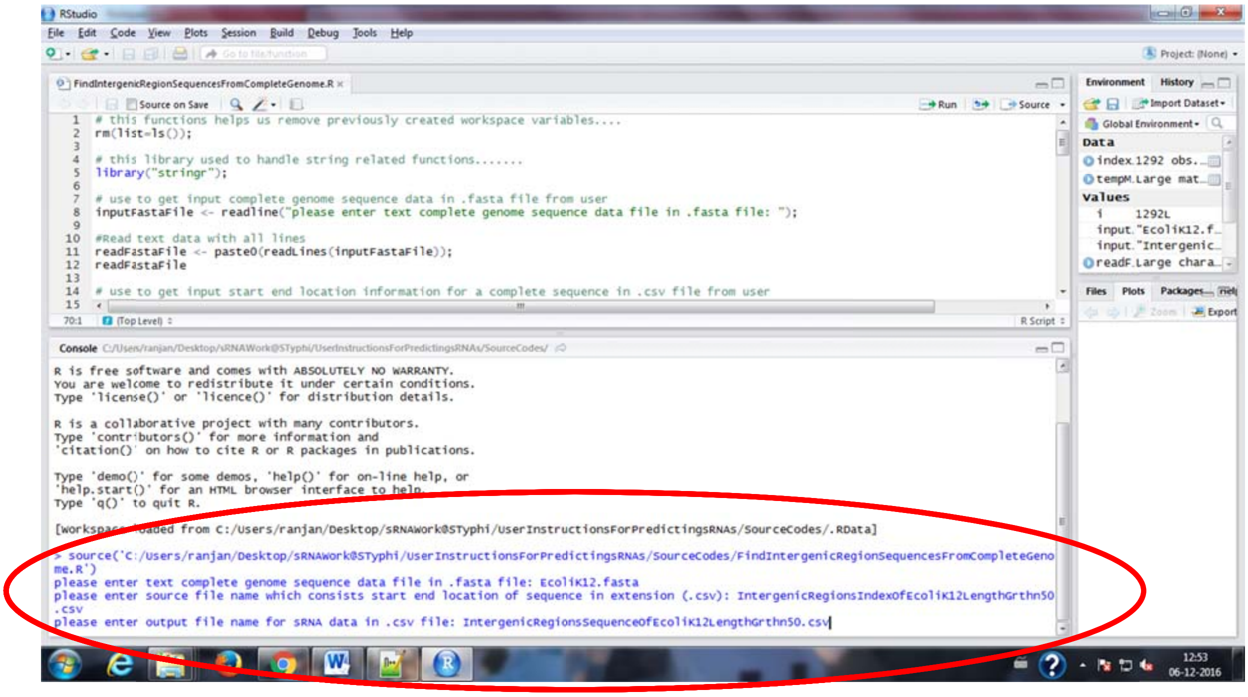
	A	B	C	D
1	IntergenicRegion	Start	Stop	Length
2	InterGenicRegion1	1	189	189
3	InterGenicRegion2	256	336	81
4	InterGenicRegion5	5021	5233	213
5	InterGenicRegion6	5531	8237	2707
6	InterGenicRegion7	9192	9305	114
7	InterGenicRegion8	9894	12162	2269
8	InterGenicRegion9	14080	14167	88
9	InterGenicRegion10	15299	15444	146
10	InterGenicRegion11	16558	17488	931
11	InterGenicRegion12	18656	18714	59
12	InterGenicRegion13	19621	21180	1560
13	InterGenicRegion16	25702	25825	124
14	InterGenicRegion18	27228	27292	65
15	InterGenicRegion19	28208	28373	166
16	InterGenicRegion20	29196	29650	455
17	InterGenicRegion22	34039	34299	261
18	InterGenicRegion23	34696	42402	7707
19	InterGenicRegion26	45751	45806	56
20	InterGenicRegion27	47139	47245	107
21	InterGenicRegion28	49632	49822	191
22	InterGenicRegion29	50303	57363	7061
23	InterGenicRegion30	58180	70386	12207
24	InterGenicRegion31	71266	71350	85
25	InterGenicRegion32	72116	77387	5272

IntergenicRegionsIndexOfEcoliK1

Ready 100%

Step10: Next we have developed FindIntergenicRegionSequencesFromCompleteGenome.r code to find corresponding sequence of intergenic regions. This code will accept input as complete genome sequence (created in **Step1**) and index file of start and stop position of intergenic regions (created in **Step9**). This code will produce sequence corresponding to the intergenic region. Please note that this code will take few minutes since it will operate on complete genome.

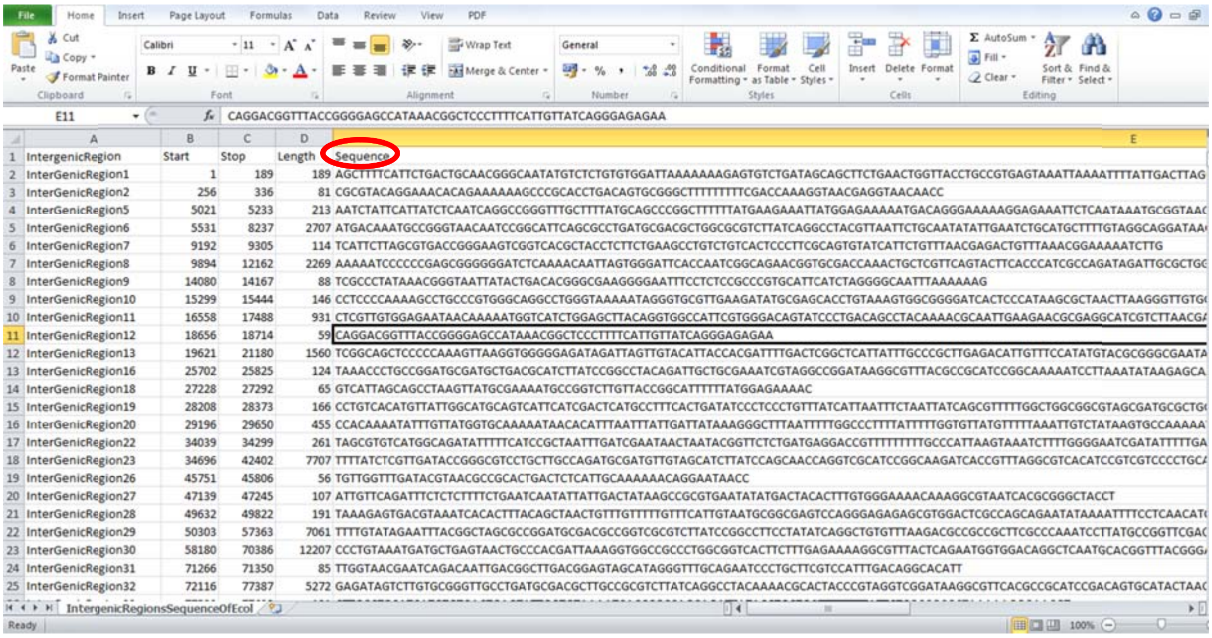
Ex: Please see the below screenshots.



```
# 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 input complete genome sequence data in .fasta file from user
inputFastaFile <- readline("please enter text complete genome sequence data file in .fasta file: ");

#Read text data with all lines
readFastaFile <- paste0(readLines(inputFastaFile));
readFastaFile

# use to get input start end location information for a complete sequence in .csv file from user
```

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

```
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.

[Workspace loaded from C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/.RData]

> source('C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/FindIntergenicRegionSequencesFromCompleteGeno
me.R')
please enter text complete genome sequence data file in .fasta file: Ecolik12.fasta
please enter source file name which consists start end location of sequence in extension (.csv): IntergenicRegionsIndexOfEcoliK12LengthGrthn50
.csv
please enter output file name for sRNA data in .csv file: IntergenicRegionsSequenceOfEcoliK12LengthGrthn50.csv
```



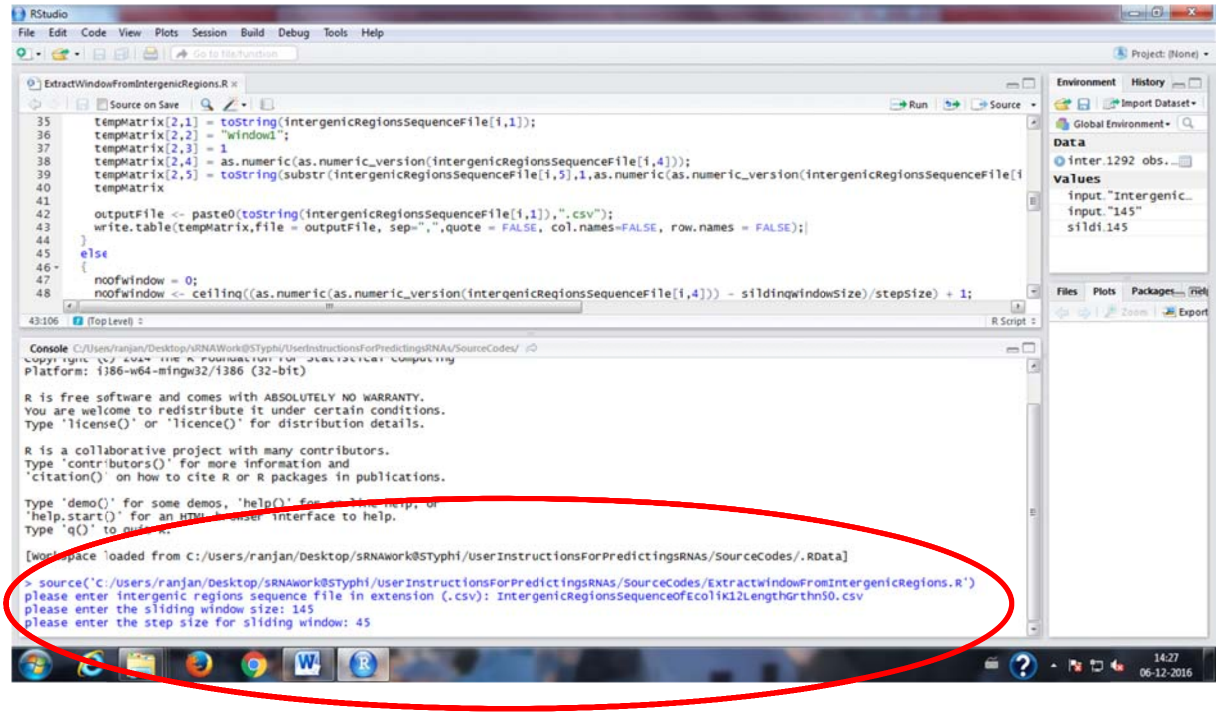
File Home Insert Page Layout Formulas Data Review View PDF
Paste Cut Copy Format Painter
Clipboard
Calibri 11
Font
Wrap Text Merge & Center
Alignment
General
Number
Conditional Formatting Format as Table Cell Styles
Styles
Insert Delete Format
Cells
AutoSum Fill Clear Sort & Filter Find & Select
Editing
IntergenicRegionsSequenceOfEcol
Ready
100%

E11 CAGGACGGTTTACCGGGGAGCCATAAACGGCTCCCTTTTCATTGTTATCAGGGAGAGAA

	A	B	C	D	E
1	IntergenicRegion	Start	Stop	Length	Sequence
2	InterGenicRegion1	1	189	189	AGCTTTTCATTCTGACTGCAACGGCAATATGTCTCTGTGTGGATTAAAAAAAGAGTGTCTGATAGCAGCTTCTGAACTGGTTACCTGCCGTGAGTAAATTAAAATTTTATTGACTTAG
3	InterGenicRegion2	256	336	81	CGCGTACAGGAAACACAGAAAAAAGCCCGCACCTGACAGTGCGGGCTTTTTTTTTCGACCAAAGGTAACGAGGTAACAACC
4	InterGenicRegion5	5021	5233	213	AATCTATTCATTATCTCAATCAGGCCGGGTTTGCTTTTATGCAGCCCGGCTTTTTTATGAAGAAATTATGGAGAAAAATGACAGGGAAAAAGGAGAAATTCTCAATAAATGCGGTAAC
5	InterGenicRegion6	5531	8237	2707	ATGACAAATGCCGGGTAACAATCCGGCATTCAGCGCCTGATGCGACGCTGGCGCGTCTTATCAGGCCTACGTTAATTCTGCAATATATTGAATCTGCATGCTTTTGTAGGCAGGATAA
6	InterGenicRegion7	9192	9305	114	TCATTCTTAGCGTGACCGGGAAGTCGGTCACGCTACCTCTTCTGAAGCCTGTCTGTCACTCCCTTCGCAGTGTATCATTCTGTTTAACGAGACTGTTTAAACGGAAAAATCTTG
7	InterGenicRegion8	9894	12162	2269	AAAAATCCCCCGAGCGGGGGATCTCAAAACAATTAGTGGGATTCACCAATCGGCAGAACGGTGCGACCAAACTGCTCGTTCAGTACTTCACCCATCGCCAGATAGATTGCGCTGG
8	InterGenicRegion9	14080	14167	88	TCGCCCTATAAACGGGTAATTATACTGACACGGGCGAAGGGGAATTTCCTCTCCGCCCGTGCATTCATCTAGGGGCAATTTAAAAAAG
9	InterGenicRegion10	15299	15444	146	CCTCCCCAAAAGCCTGCCCGTGGGCAGGCCTGGGTAAAAATAGGGTGCGTTGAAGATATGCGAGCACCTGTAAAGTGGCGGGGATCACTCCCATAAGCGCTAACTTAAGGGTTGTG
10	InterGenicRegion11	16558	17488	931	CTCGTTGTGGAGAATAACAAAAATGGTCATCTGGAGCTTACAGGTGGCCATTCGTGGGACAGTATCCCTGACAGCCTACAAAACGCAATTGAAGAACGCGAGGCATCGTCTTAACGA
11	InterGenicRegion12	18656	18714	59	CAGGACGGTTTACCGGGGAGCCATAAACGGCTCCCTTTTCATTGTTATCAGGGAGAGAA
12	InterGenicRegion13	19621	21180	1560	TCGGCAGCTCCCCCAAAGTTAAGGTGGGGGAGATAGATTAGTTGTACATTACCACGATTTTGACTCGGCTCATTATTTGCCCGCTTGAGACATTGTTTCCATATGTACGCGGGCGAATA
13	InterGenicRegion16	25702	25825	124	TAAACCCTGCCGGATGCGATGCTGACGCATCTTATCCGGCCTACAGATTGCTGCGAAATCGTAGGCCGGATAAGGCGTTTACGCCGCATCCGGCAAAAATCCTTAAATATAAGAGCA
14	InterGenicRegion18	27228	27292	65	GTCATTAGCAGCCTAAGTTATGCGAAAATGCCGGTCTTGTTACCGGCATTTTTTATGGAGAAAAC
15	InterGenicRegion19	28208	28373	166	CCTGTCACATGTTATTGGCATGCAGTCATTCATCGACTCATGCCTTTCACTGATATCCCTCCCTGTTTATCATTAATTTCTAATTATCAGCGTTTTTGGCTGGCGGCGTAGCGATGCGCTG
16	InterGenicRegion20	29196	29650	455	CCACAAAATATTTGTTATGGTGCAAAAATAACACATTTAATTTATTGATTATAAAGGGCTTTAATTTTTGGCCCTTTTATTTTTGGTGTTATGTTTTTAAATTGTCTATAAGTGCCAAAAA
17	InterGenicRegion22	34039	34299	261	TAGCGTGTCATGGCAGATATTTTTCATCCGCTAATTTGATCGAATAACTAATACGGTTCTCTGATGAGGACCGTTTTTTTTTGCCCATTAAGTAAATCTTTTGGGGAATCGATATTTTTGA
18	InterGenicRegion23	34696	42402	7707	TTTTATCTCGTTGATACCGGGCGTCCTGCTTGCCAGATGCGATGTTGTAGCATCTTATCCAGCAACCAGGTCGCATCCGGCAAGATCACCGTTTAGGCGTCACATCCGTCGTCCCCTGCA
19	InterGenicRegion26	45751	45806	56	TGTTGGTTTGATACGTAACGCCGCACTGACTCTCATTGCAAAAAACAGGAATAACC
20	InterGenicRegion27	47139	47245	107	ATTGTTCAGATTTCTCTCTTTTCTGAATCAATATTATTGACTATAAGCCGCGTGAATATATGACTACACTTTGTGGGAAAACAAAGGCGTAATCACGCGGGCTACCT
21	InterGenicRegion28	49632	49822	191	TAAAGAGTGACGTAAATCACACTTTACAGCTAACTGTTTGTTTTTGTTTCATTGTAATGCGGCGAGTCCAGGGAGAGAGCGTGGACTCGCCAGCAGAATATAAAATTTTCCTCAACAT
22	InterGenicRegion29	50303	57363	7061	TTTTGTATAGAATTTACGGCTAGCGCCGGATGCGACGCCGGTCGCGTCTTATCCGGCCTTCCTATATCAGGCTGTGTTTAAGACGCCGCCGCTTCGCCCAAATCCTTATGCCGGTTCGAC
23	InterGenicRegion30	58180	70386	12207	CCCTGTAAATGATGCTGAGTAACTGCCCACGATTAAAGGTGGCCGCCCTGGCGGTCACTTCTTTGAGAAAAGGCGTTTACTCAGAATGGTGGACAGGCTCAATGCACGGTTTACGGG
24	InterGenicRegion31	71266	71350	85	TTGGTAACGAATCAGACAATTGACGGCTTGACGGAGTAGCATAGGGTTTGCAGAATCCCTGCTTCGTCCATTTGACAGGCACATT
25	InterGenicRegion32	72116	77387	5272	GAGATAGTCTTGTGCGGGTTGCCTGATGCGACGCTTGCCGCGTCTTATCAGGCCTACAAAACGCACTACCCGTAGGTCGGATAAGGCGTTCACGCCGCATCCGACAGTGCATACTAAC

Step11: Following this we have developed ExtractWindowFromIntergenicRegions.r code to extract window from intergenic region. This code will accept input as intergenic region sequence file (created in **Step10**), window size and step size of the window. This code will produce single and multiple windows based on intergenic region length, window size and step size.

Ex: Please see the below screenshots.

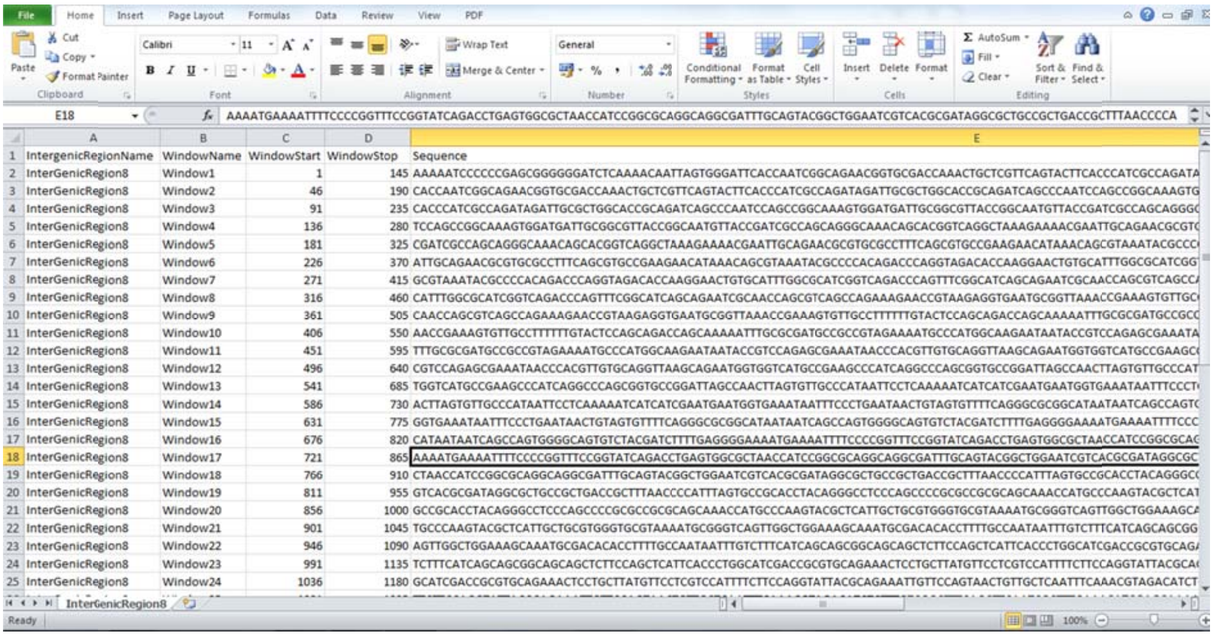


Data
inter 1292 obs. …
Values
input "Intergenic…
input "145"
sildi 145

```
    tempMatrix[2,1] = toString(intergenicRegionsSequenceFile[i,1]);
    tempMatrix[2,2] = "window1";
    tempMatrix[2,3] = 1
    tempMatrix[2,4] = as.numeric(as.numeric_version(intergenicRegionsSequenceFile[i,4]));
    tempMatrix[2,5] = toString(substr(intergenicRegionsSequenceFile[i,5],1,as.numeric(as.numeric_version(intergenicRegionsSequenceFile[i
    tempMatrix

    outputFile <- paste0(toString(intergenicRegionsSequenceFile[i,1]),".csv");
    write.table(tempMatrix,file = outputFile, sep=",",quote = FALSE, col.names=FALSE, row.names = FALSE);
  }
  else
  {
    ncofwindow = 0;
    ncofwindow <- ceiling((as.numeric(as.numeric_version(intergenicRegionsSequenceFile[i,4])) - sildingwindowSize)/stepSize) + 1;
```

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

```
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.

[Workspace loaded from C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/.RData]

> source('C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/ExtractWindowFromIntergenicRegions.R')
please enter intergenic regions sequence file in extension (.csv): IntergenicRegionsSequenceOfEcoliK12LengthGrthn50.csv
please enter the sliding window size: 145
please enter the step size for sliding window: 45
```

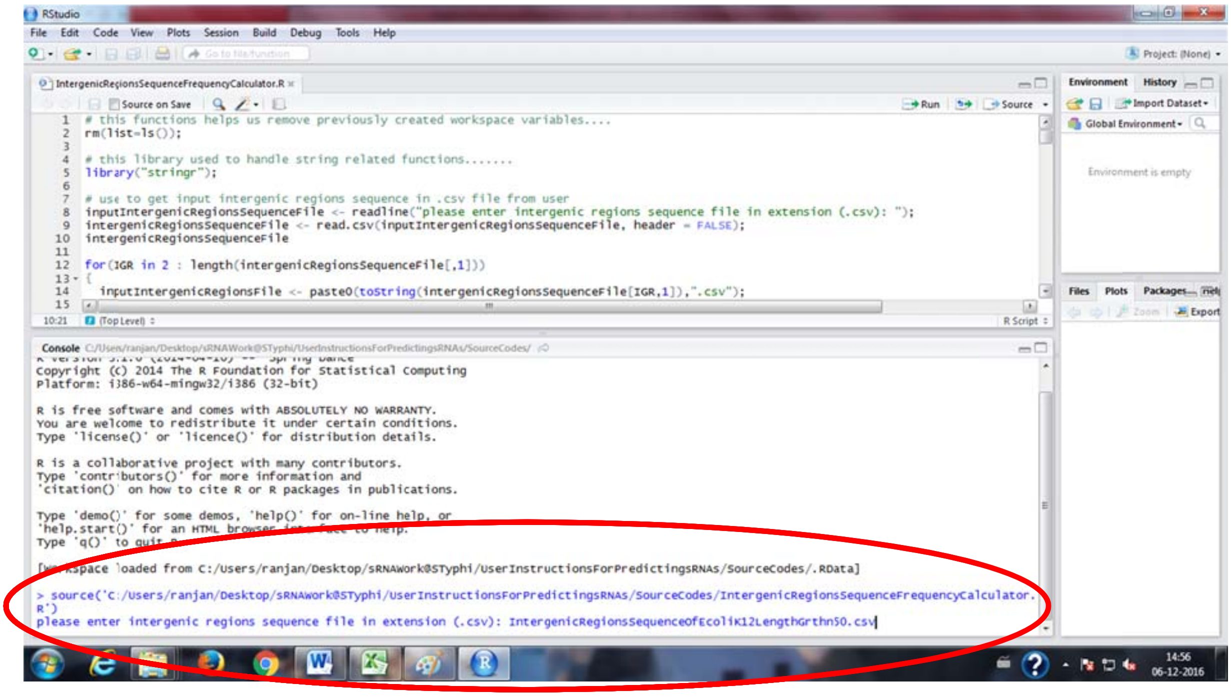


E18 AAAATGAAAATTTTCCCCGGTTTCCGGTATCAGACCTGAGTGGCGCTAACCATCCGGCGCAGGCAGGCGATTTGCAGTACGGCTGGAATCGTCACGCGATAGGCGCTGCCGCTGACCGCTTTAACCCCA

	A	B	C	D	E
1	IntergenicRegionName	WindowName	WindowStart	WindowStop	Sequence
2	InterGenicRegion8	Window1	1	145	AAAAATCCCCCGAGCGGGGGGATCTCAAAACAATTAGTGGGATTCACCAATCGGCAGAACGGTGCGACCAAACTGCTCGTTCAGTACTTCACCCATCGCCAGATA
3	InterGenicRegion8	Window2	46	190	CACCAATCGGCAGAACGGTGCGACCAAACTGCTCGTTCAGTACTTCACCCATCGCCAGATAGATTGCGCTGGCACCGCAGATCAGCCCAATCCAGCCGGCAAAGTG
4	InterGenicRegion8	Window3	91	235	CACCCATCGCCAGATAGATTGCGCTGGCACCGCAGATCAGCCCAATCCAGCCGGCAAAGTGGATGATTGCGGCGTTACCGGCAATGTTACCGATCGCCAGCAGGG
5	InterGenicRegion8	Window4	136	280	TCCAGCCGGCAAAGTGGATGATTGCGGCGTTACCGGCAATGTTACCGATCGCCAGCAGGGCAAACAGCACGGTCAGGCTAAAGAAAACGAATTGCAGAACGCGT
6	InterGenicRegion8	Window5	181	325	CGATCGCCAGCAGGGCAAACAGCACGGTCAGGCTAAAGAAAACGAATTGCAGAACGCGTGCGCCTTTCAGCGTGCCGAAGAACATAAACAGCGTAAATACGCCC
7	InterGenicRegion8	Window6	226	370	ATTGCAGAACGCGTGCGCCTTTCAGCGTGCCGAAGAACATAAACAGCGTAAATACGCCCCACAGACCCAGGTAGACACCAAGGAACTGTGCATTTGGCGCATCGG
8	InterGenicRegion8	Window7	271	415	GCGTAAATACGCCCCACAGACCCAGGTAGACACCAAGGAACTGTGCATTTGGCGCATCGGTCAGACCCAGTTTCGGCATCAGCAGAATCGCAACCAGCGTCAGCC
9	InterGenicRegion8	Window8	316	460	CATTTGGCGCATCGGTCAGACCCAGTTTCGGCATCAGCAGAATCGCAACCAGCGTCAGCCAGAAAGAACCGTAAGAGGTGAATGCGGTTAAACCGAAAGTGTTGC
10	InterGenicRegion8	Window9	361	505	CAACCAGCGTCAGCCAGAAAGAACCGTAAGAGGTGAATGCGGTTAAACCGAAAGTGTTGCCTTTTTTGTACTCCAGCAGACCAGCAAAAATTTGCGCGATGCCGCC
11	InterGenicRegion8	Window10	406	550	AACCGAAAGTGTTGCCTTTTTTGTACTCCAGCAGACCAGCAAAAATTTGCGCGATGCCGCCGTAGAAAATGCCCATGGCAAGAATAATACCGTCCAGAGCGAAATA
12	InterGenicRegion8	Window11	451	595	TTTGCGCGATGCCGCCGTAGAAAATGCCCATGGCAAGAATAATACCGTCCAGAGCGAAATAACCCACGTTGTGCAGGTTAAGCAGAATGGTGGTCATGCCGAAGC
13	InterGenicRegion8	Window12	496	640	CGTCCAGAGCGAAATAACCCACGTTGTGCAGGTTAAGCAGAATGGTGGTCATGCCGAAGCCCATCAGGCCCAGCGGTGCCGGATTAGCCAACTTAGTGTTGCCCAT
14	InterGenicRegion8	Window13	541	685	TGGTCATGCCGAAGCCCATCAGGCCCAGCGGTGCCGGATTAGCCAACTTAGTGTTGCCCATAATTCCTCAAAAATCATCATCGAATGAATGGTGAAATAATTTCCCT
15	InterGenicRegion8	Window14	586	730	ACTTAGTGTTGCCCATAATTCCTCAAAAATCATCATCGAATGAATGGTGAAATAATTTCCCTGAATAACTGTAGTGTTTTCAGGGCGCGGCATAATAATCAGCCAGT
16	InterGenicRegion8	Window15	631	775	GGTGAAATAATTTCCCTGAATAACTGTAGTGTTTTCAGGGCGCGGCATAATAATCAGCCAGTGGGGCAGTGTCTACGATCTTTTGAGGGGAAAATGAAAATTTTCCC
17	InterGenicRegion8	Window16	676	820	CATAATAATCAGCCAGTGGGGCAGTGTCTACGATCTTTTGAGGGGAAAATGAAAATTTTCCCCGGTTTCCGGTATCAGACCTGAGTGGCGCTAACCATCCGGCGCAG
18	InterGenicRegion8	Window17	721	865	AAAATGAAAATTTTCCCCGGTTTCCGGTATCAGACCTGAGTGGCGCTAACCATCCGGCGCAGGCAGGCGATTTGCAGTACGGCTGGAATCGTCACGCGATAGGCGC
19	InterGenicRegion8	Window18	766	910	CTAACCATCCGGCGCAGGCAGGCGATTTGCAGTACGGCTGGAATCGTCACGCGATAGGCGCTGCCGCTGACCGCTTTAACCCCATTTAGTGCCGCACCTACAGGGC
20	InterGenicRegion8	Window19	811	955	GTCACGCGATAGGCGCTGCCGCTGACCGCTTTAACCCCATTTAGTGCCGCACCTACAGGGCCTCCCAGCCCCGCGCCGCGCAGCAAACCATGCCCAAGTACGCTCAT
21	InterGenicRegion8	Window20	856	1000	GCCGCACCTACAGGGCCTCCCAGCCCCGCGCCGCGCAGCAAACCATGCCCAAGTACGCTCATTGCTGCGTGGGTGCGTAAAATGCGGGTCAGTTGGCTGGAAAGC
22	InterGenicRegion8	Window21	901	1045	TGCCCAAGTACGCTCATTGCTGCGTGGGTGCGTAAAATGCGGGTCAGTTGGCTGGAAAGCAAATGCGACACACCTTTTGCCAATAATTTGTCTTTCATCAGCAGCGG
23	InterGenicRegion8	Window22	946	1090	AGTTGGCTGGAAAGCAAATGCGACACACCTTTTGCCAATAATTTGTCTTTCATCAGCAGCGGCAGCAGCTCTTCCAGCTCATTCACCCTGGCATCGACCGCGTGCAG
24	InterGenicRegion8	Window23	991	1135	TCTTTCATCAGCAGCGGCAGCAGCTCTTCCAGCTCATTCACCCTGGCATCGACCGCGTGCAGAAACTCCTGCTTATGTTCCTCGTCCATTTTCTTCCAGGTATTACGCA
25	InterGenicRegion8	Window24	1036	1180	GCATCGACCGCGTGCAGAAACTCCTGCTTATGTTCCTCGTCCATTTTCTTCCAGGTATTACGCAGAAATTGTTCCAGTAACTGTTGCTCAATTTCAAACGTAGACATCT

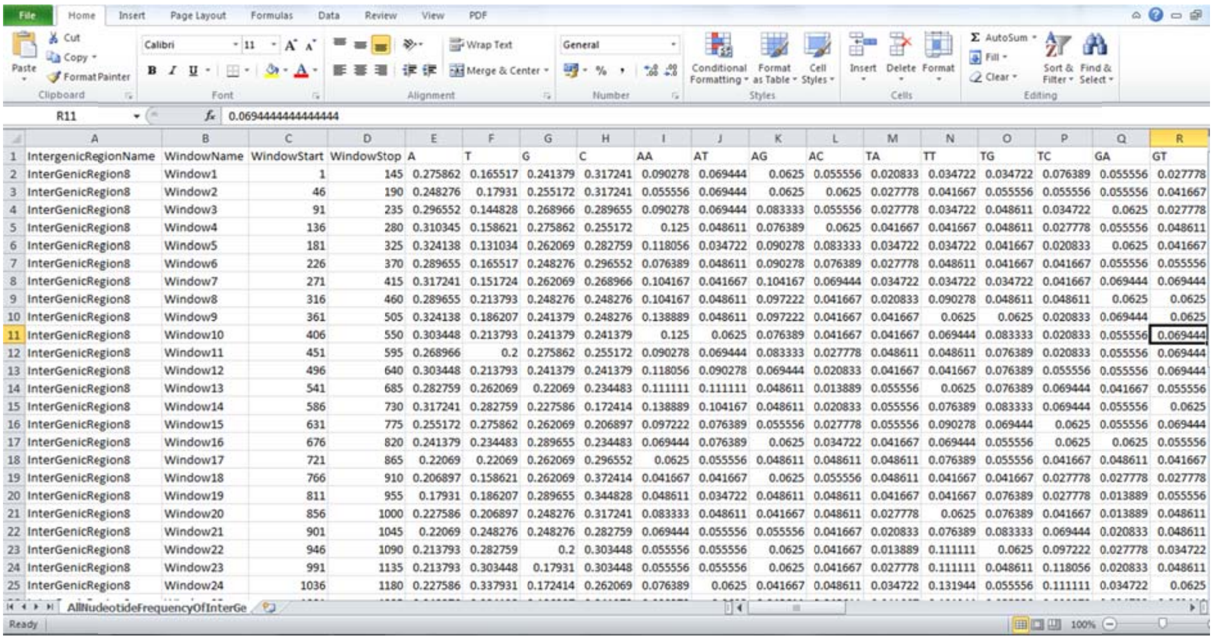
Step12: Afterwards we have developed IntergenicRegionsSequenceFrequencyCalculator.r code to calculate mono (4), di (16) and tri (64) nucleotide composition features for each window of intergenic region. This code will accept input as an intergenic region sequence file (created in **Step10**) and produce sequence composition (84) for all windows. Please note that this code will take few hours for calculating composition features for each window of each intergenic region. The execution time is directly proportional to the number of intergenic region and length of the intergenic region.

Ex: Please see the below screenshots.


RStudio
File Edit Code View Plots Session Build Debug Tools Help
IntergenicRegionsSequenceFrequencyCalculator.R
Source on Save
Run
Source
Environment History
Import Dataset
Global Environment
Environment is empty
Files Plots Packages
Zoom Export
Project: (None)
10:21 (Top Level)
R Script
Console C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/
14:56
06-12-2016

```
# 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 input intergenic regions sequence in .csv file from user
inputIntergenicRegionsSequenceFile <- readline("please enter intergenic regions sequence file in extension (.csv): ");
intergenicRegionsSequenceFile <- read.csv(inputIntergenicRegionsSequenceFile, header = FALSE);
intergenicRegionsSequenceFile

for(IGR in 2 : length(intergenicRegionsSequenceFile[,1]))
{
  inputIntergenicRegionsFile <- paste0(toString(intergenicRegionsSequenceFile[IGR,1]),".csv");
```

```
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.

[Workspace loaded from C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/.RData]

> source('C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/IntergenicRegionsSequenceFrequencyCalculator.R')
please enter intergenic regions sequence file in extension (.csv): IntergenicRegionsSequenceOfEcoliK12LengthGrthn50.csv
```

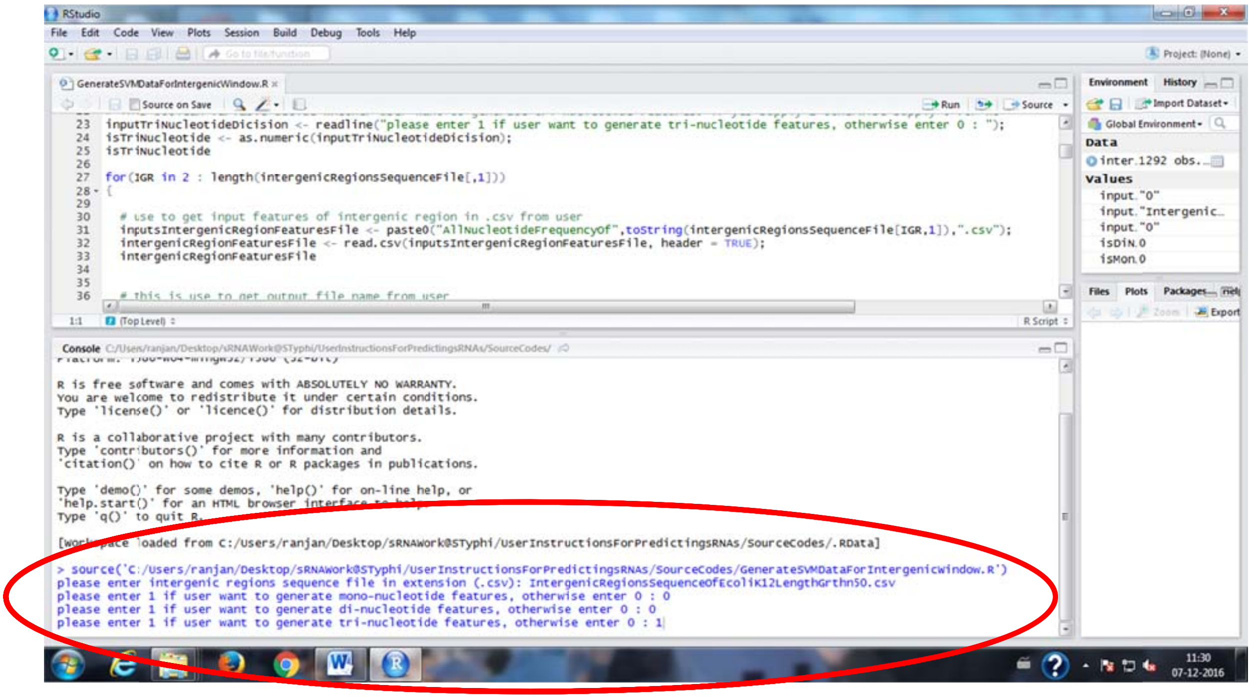


R11 0.0694444444444444

IntergenicRegionName	WindowName	WindowStart	WindowStop	A	T	G	C	AA	AT	AG	AC	TA	TT	TG	TC	GA	GT
InterGenicRegion8	Window1	1	145	0.275862	0.165517	0.241379	0.317241	0.090278	0.069444	0.0625	0.055556	0.020833	0.034722	0.034722	0.076389	0.055556	0.027778
InterGenicRegion8	Window2	46	190	0.248276	0.17931	0.255172	0.317241	0.055556	0.069444	0.0625	0.0625	0.027778	0.041667	0.055556	0.055556	0.055556	0.041667
InterGenicRegion8	Window3	91	235	0.296552	0.144828	0.268966	0.289655	0.090278	0.069444	0.083333	0.055556	0.027778	0.034722	0.048611	0.034722	0.0625	0.027778
InterGenicRegion8	Window4	136	280	0.310345	0.158621	0.275862	0.255172	0.125	0.048611	0.076389	0.0625	0.041667	0.041667	0.048611	0.027778	0.055556	0.048611
InterGenicRegion8	Window5	181	325	0.324138	0.131034	0.262069	0.282759	0.118056	0.034722	0.090278	0.083333	0.034722	0.034722	0.041667	0.020833	0.0625	0.041667
InterGenicRegion8	Window6	226	370	0.289655	0.165517	0.248276	0.296552	0.076389	0.048611	0.090278	0.076389	0.027778	0.048611	0.041667	0.041667	0.055556	0.055556
InterGenicRegion8	Window7	271	415	0.317241	0.151724	0.262069	0.268966	0.104167	0.041667	0.104167	0.069444	0.034722	0.034722	0.034722	0.041667	0.069444	0.069444
InterGenicRegion8	Window8	316	460	0.289655	0.213793	0.248276	0.248276	0.104167	0.048611	0.097222	0.041667	0.020833	0.090278	0.048611	0.048611	0.0625	0.0625
InterGenicRegion8	Window9	361	505	0.324138	0.186207	0.241379	0.248276	0.138889	0.048611	0.097222	0.041667	0.041667	0.0625	0.0625	0.020833	0.069444	0.0625
InterGenicRegion8	Window10	406	550	0.303448	0.213793	0.241379	0.241379	0.125	0.0625	0.076389	0.041667	0.041667	0.069444	0.083333	0.020833	0.055556	0.069444
InterGenicRegion8	Window11	451	595	0.268966	0.2	0.275862	0.255172	0.090278	0.069444	0.083333	0.027778	0.048611	0.048611	0.076389	0.020833	0.055556	0.069444
InterGenicRegion8	Window12	496	640	0.303448	0.213793	0.241379	0.241379	0.118056	0.090278	0.069444	0.020833	0.041667	0.041667	0.076389	0.055556	0.055556	0.069444
InterGenicRegion8	Window13	541	685	0.282759	0.262069	0.22069	0.234483	0.111111	0.111111	0.048611	0.013889	0.055556	0.0625	0.076389	0.069444	0.041667	0.055556
InterGenicRegion8	Window14	586	730	0.317241	0.282759	0.227586	0.172414	0.138889	0.104167	0.048611	0.020833	0.055556	0.076389	0.083333	0.069444	0.055556	0.0625
InterGenicRegion8	Window15	631	775	0.255172	0.275862	0.262069	0.206897	0.097222	0.076389	0.055556	0.027778	0.055556	0.090278	0.069444	0.0625	0.055556	0.069444
InterGenicRegion8	Window16	676	820	0.241379	0.234483	0.289655	0.234483	0.069444	0.076389	0.0625	0.034722	0.041667	0.069444	0.055556	0.0625	0.0625	0.055556
InterGenicRegion8	Window17	721	865	0.22069	0.22069	0.262069	0.296552	0.0625	0.055556	0.048611	0.048611	0.048611	0.076389	0.055556	0.041667	0.048611	0.041667
InterGenicRegion8	Window18	766	910	0.206897	0.158621	0.262069	0.372414	0.041667	0.041667	0.0625	0.055556	0.048611	0.041667	0.041667	0.027778	0.027778	0.027778
InterGenicRegion8	Window19	811	955	0.17931	0.186207	0.289655	0.344828	0.048611	0.034722	0.048611	0.048611	0.041667	0.041667	0.076389	0.027778	0.013889	0.055556
InterGenicRegion8	Window20	856	1000	0.227586	0.206897	0.248276	0.317241	0.083333	0.048611	0.041667	0.048611	0.027778	0.0625	0.076389	0.041667	0.013889	0.048611
InterGenicRegion8	Window21	901	1045	0.22069	0.248276	0.248276	0.282759	0.069444	0.055556	0.055556	0.041667	0.020833	0.076389	0.083333	0.069444	0.020833	0.048611
InterGenicRegion8	Window22	946	1090	0.213793	0.282759	0.2	0.303448	0.055556	0.055556	0.0625	0.041667	0.013889	0.111111	0.0625	0.097222	0.027778	0.034722
InterGenicRegion8	Window23	991	1135	0.213793	0.303448	0.17931	0.303448	0.055556	0.055556	0.0625	0.041667	0.027778	0.111111	0.048611	0.118056	0.020833	0.048611
InterGenicRegion8	Window24	1036	1180	0.227586	0.337931	0.172414	0.262069	0.076389	0.0625	0.041667	0.048611	0.034722	0.131944	0.055556	0.111111	0.034722	0.0625

Step13: Next we have developed GenerateSVMDataForIntergenicWindow.r code to generate SVM format data for each window of each intergenic region. This code will take input as intergenic region sequence file (created in **Step10**) and decision (1 and 0 indicates Yes and No respectively) of generating mono, di and tri nucleotide features. This code will produce data in SVM format. We found that trinucleotide composition features performed well, therefore we have generated only tri nucleotide features in this step.

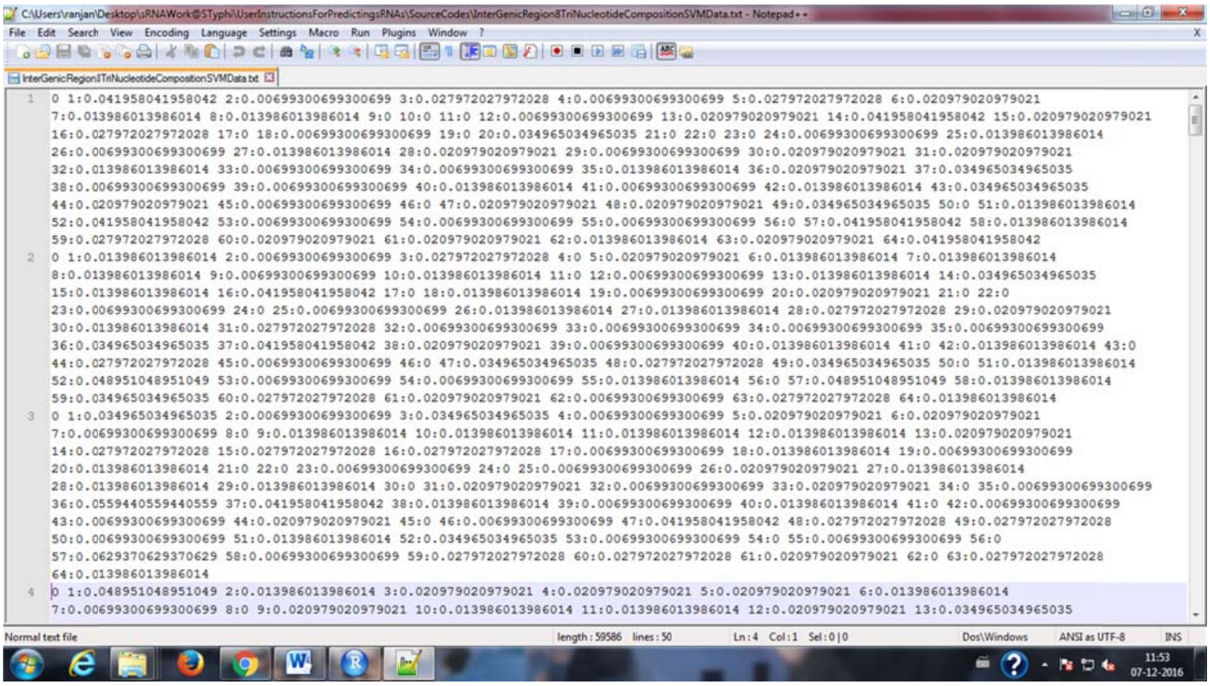
Ex: Please see the below screenshots.



RStudio
File Edit Code View Plots Session Build Debug Tools Help
GenerateSVMDataForIntergenicWindow.R
Source on Save
Run
Source
Environment History
Import Dataset
Global Environment
Data
inter 1292 obs.
Values
input "0"
input "Intergenic...
input "0"
isDiN 0
isMon 0
Files Plots Packages Help
Zoom Export
(Top Level)
R Script
Console C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/
11:30
07-12-2016

```
inputTriNucleotideDicision <- readline("please enter 1 if user want to generate tri-nucleotide features, otherwise enter 0 : ");
isTriNucleotide <- as.numeric(inputTriNucleotideDicision);
isTriNucleotide

for(IGR in 2 : length(intergenicRegionsSequenceFile[,1]))
{

  # use to get input features of intergenic region in .csv from user
  inputsIntergenicRegionFeaturesFile <- paste0("AllNucleotideFrequencyOf",toString(intergenicRegionsSequenceFile[IGR,1]),".csv");
  intergenicRegionFeaturesFile <- read.csv(inputsIntergenicRegionFeaturesFile, header = TRUE);
  intergenicRegionFeaturesFile



  # This is use to get output file name from user
```

```
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.

[Workspace loaded from C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/.RData]

> source('C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/GenerateSVMDataForIntergenicWindow.R')
please enter intergenic regions sequence file in extension (.csv): IntergenicRegionsSequenceOfEcoliK12LengthGrthn50.csv
please enter 1 if user want to generate mono-nucleotide features, otherwise enter 0 : 0
please enter 1 if user want to generate di-nucleotide features, otherwise enter 0 : 0
please enter 1 if user want to generate tri-nucleotide features, otherwise enter 0 : 1
```



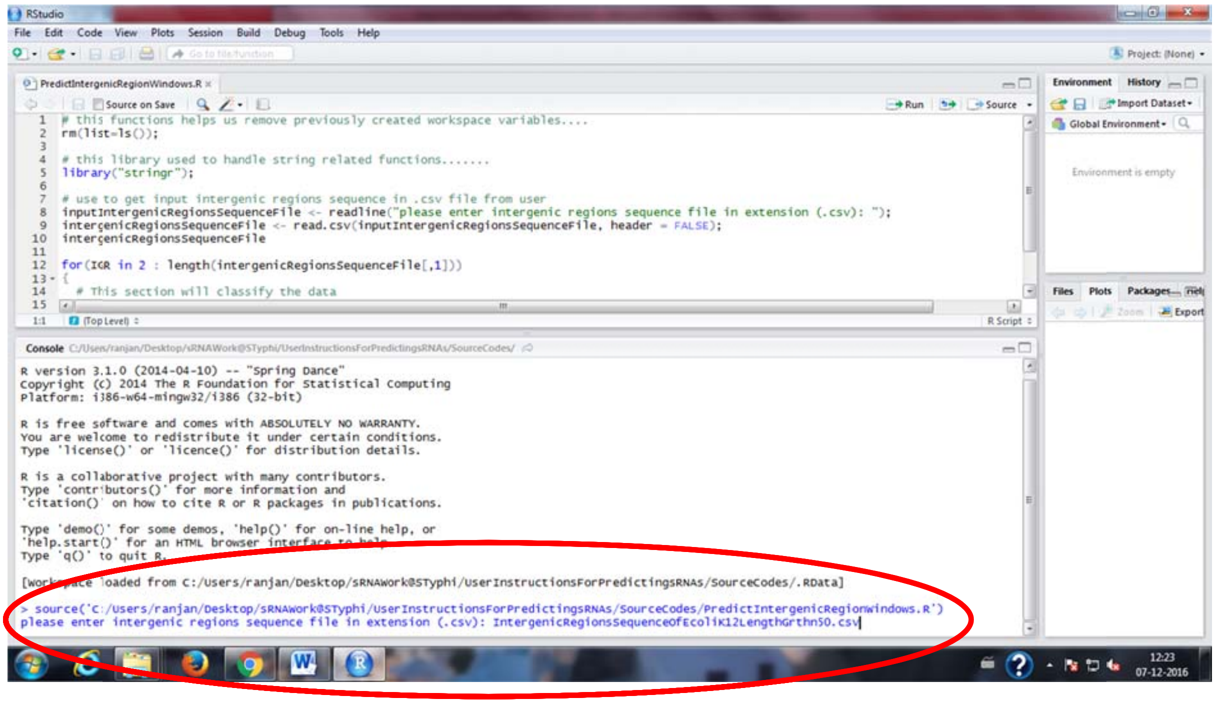
C:\Users\ranjan\Desktop\sRNAWork@STyphi\UserInstructionsForPredictingsRNAs\SourceCodes\InterGenicRegion8TriNucleotideCompositionSVMData.txt - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
InterGenicRegion8TriNucleotideCompositionSVMData.txt
Normal text file
length : 59586 lines : 50
Ln : 4 Col : 1 Sel : 0 | 0
Dos\Windows
ANSI as UTF-8
INS
11:53
07-12-2016

```
0 1:0.041958041958042 2:0.00699300699300699 3:0.027972027972028 4:0.00699300699300699 5:0.027972027972028 6:0.020979020979021
7:0.013986013986014 8:0.013986013986014 9:0 10:0 11:0 12:0.00699300699300699 13:0.020979020979021 14:0.041958041958042 15:0.020979020979021
16:0.027972027972028 17:0 18:0.00699300699300699 19:0 20:0.034965034965035 21:0 22:0 23:0 24:0.00699300699300699 25:0.013986013986014
26:0.00699300699300699 27:0.013986013986014 28:0.020979020979021 29:0.00699300699300699 30:0.020979020979021 31:0.020979020979021
32:0.013986013986014 33:0.00699300699300699 34:0.00699300699300699 35:0.013986013986014 36:0.020979020979021 37:0.034965034965035
38:0.00699300699300699 39:0.00699300699300699 40:0.013986013986014 41:0.00699300699300699 42:0.013986013986014 43:0.034965034965035
44:0.020979020979021 45:0.00699300699300699 46:0 47:0.020979020979021 48:0.020979020979021 49:0.034965034965035 50:0 51:0.013986013986014
52:0.041958041958042 53:0.00699300699300699 54:0.00699300699300699 55:0.00699300699300699 56:0 57:0.041958041958042 58:0.013986013986014
59:0.027972027972028 60:0.020979020979021 61:0.020979020979021 62:0.013986013986014 63:0.020979020979021 64:0.041958041958042
0 1:0.013986013986014 2:0.00699300699300699 3:0.027972027972028 4:0 5:0.020979020979021 6:0.013986013986014 7:0.013986013986014
8:0.013986013986014 9:0.00699300699300699 10:0.013986013986014 11:0 12:0.00699300699300699 13:0.013986013986014 14:0.034965034965035
15:0.013986013986014 16:0.041958041958042 17:0 18:0.013986013986014 19:0.00699300699300699 20:0.020979020979021 21:0 22:0
23:0.00699300699300699 24:0 25:0.00699300699300699 26:0.013986013986014 27:0.013986013986014 28:0.027972027972028 29:0.020979020979021
30:0.013986013986014 31:0.027972027972028 32:0.00699300699300699 33:0.00699300699300699 34:0.00699300699300699 35:0.00699300699300699
36:0.034965034965035 37:0.041958041958042 38:0.020979020979021 39:0.00699300699300699 40:0.013986013986014 41:0 42:0.013986013986014 43:0
44:0.027972027972028 45:0.00699300699300699 46:0 47:0.034965034965035 48:0.027972027972028 49:0.034965034965035 50:0 51:0.013986013986014
52:0.048951048951049 53:0.00699300699300699 54:0.00699300699300699 55:0.013986013986014 56:0 57:0.048951048951049 58:0.013986013986014
59:0.034965034965035 60:0.027972027972028 61:0.020979020979021 62:0.00699300699300699 63:0.027972027972028 64:0.013986013986014
0 1:0.034965034965035 2:0.00699300699300699 3:0.034965034965035 4:0.00699300699300699 5:0.020979020979021 6:0.020979020979021
7:0.00699300699300699 8:0 9:0.013986013986014 10:0.013986013986014 11:0.013986013986014 12:0.013986013986014 13:0.020979020979021
14:0.027972027972028 15:0.027972027972028 16:0.027972027972028 17:0.00699300699300699 18:0.013986013986014 19:0.00699300699300699
20:0.013986013986014 21:0 22:0 23:0.00699300699300699 24:0 25:0.00699300699300699 26:0.020979020979021 27:0.013986013986014
28:0.013986013986014 29:0.013986013986014 30:0 31:0.020979020979021 32:0.00699300699300699 33:0.020979020979021 34:0 35:0.00699300699300699
36:0.0559440559440559 37:0.041958041958042 38:0.013986013986014 39:0.00699300699300699 40:0.013986013986014 41:0 42:0.00699300699300699
43:0.00699300699300699 44:0.020979020979021 45:0 46:0.00699300699300699 47:0.041958041958042 48:0.027972027972028 49:0.027972027972028
50:0.00699300699300699 51:0.013986013986014 52:0.034965034965035 53:0.00699300699300699 54:0 55:0.00699300699300699 56:0
57:0.0629370629370629 58:0.00699300699300699 59:0.027972027972028 60:0.027972027972028 61:0.020979020979021 62:0 63:0.027972027972028
64:0.013986013986014
0 1:0.048951048951049 2:0.013986013986014 3:0.020979020979021 4:0.020979020979021 5:0.020979020979021 6:0.013986013986014
7:0.00699300699300699 8:0 9:0.020979020979021 10:0.013986013986014 11:0.013986013986014 12:0.020979020979021 13:0.034965034965035
```

Step14: Please download our proposed best model (ModelTriNucleotideCompositionOneistoTwo7_7_4_3) and svm_classify.exe files from SourceCodes. You can also download SVM^{light} (svm_learn.exe and svm_classify.exe)from http://svmlight.joachims.org/ (under Source Code and Binaries).

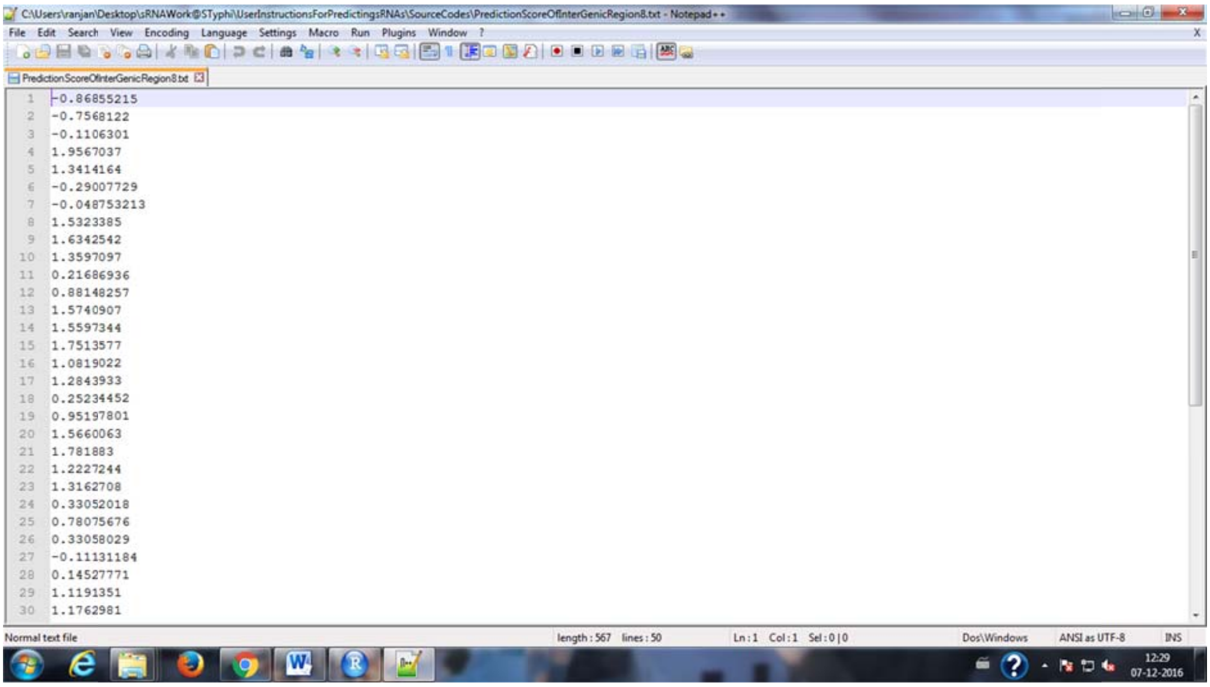
Step15: Finally we have developed PredictIntergenicRegionWindows.r code to predict each window of each intergenic region using proposed best model. This code will take only input as an intergenic region sequence file (created in **Step10**). This code will produce prediction score of each window of each intergenic region.

Ex: Please see the below screenshots.


RStudio
File Edit Code View Plots Session Build Debug Tools Help
Project: (None)
PredictIntergenicRegionWindows.R
Source on Save
Run
Source
Environment History
Import Dataset
Global Environment
Environment is empty
Files Plots Packages Help
Zoom Export
1:1 (Top Level)
R Script
Console C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/
12:23
07-12-2016

```
# 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 input intergenic regions sequence in .csv file from user
inputIntergenicRegionsSequenceFile <- readline("please enter intergenic regions sequence file in extension (.csv): ");
intergenicRegionsSequenceFile <- read.csv(inputIntergenicRegionsSequenceFile, header = FALSE);
intergenicRegionsSequenceFile

for(IGR in 2 : length(intergenicRegionsSequenceFile[,1]))
{
  # This section will classify the data
```

```
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.

[Workspace loaded from C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/.RData]

> source('C:/Users/ranjan/Desktop/sRNAWork@STyphi/UserInstructionsForPredictingsRNAs/SourceCodes/PredictIntergenicRegionWindows.R')
please enter intergenic regions sequence file in extension (.csv): IntergenicRegionsSequenceOfEcoliK12LengthGrthn50.csv
```


C:\Users\ranjan\Desktop\sRNAWork@STyphi\UserInstructionsForPredictingsRNAs\SourceCodes\PredictionScoreOfInterGenicRegion8.txt - Notepad++
File Edit Search View Encoding Language Settings Macro Run Plugins Window ?
PredictionScoreOfInterGenicRegion8.txt
Normal text file
length : 567 lines : 50
Ln : 1 Col : 1 Sel : 0 | 0
Dos\Windows
ANSI as UTF-8
INS
12:29
07-12-2016

```
-0.86855215
-0.7568122
-0.1106301
1.9567037
1.3414164
-0.29007729
-0.048753213
1.5323385
1.6342542
1.3597097
0.21686936
0.88148257
1.5740907
1.5597344
1.7513577
1.0819022
1.2843933
0.25234452
0.95197801
1.5660063
1.781883
1.2227244
1.3162708
0.33052018
0.78075676
0.33058029
-0.11131184
0.14527771
1.1191351
1.1762981
```