PanGeT v.1.0

User Manual

Last modified on October, 2016

TABLE OF CONTENTS

1. Introduction:

PanGeT is a user friendly pan-genome tool. Theoretically, PanGeT can work with a single query and 'N' number reference genomes given by the users. It differentiates the annotated CDS/protein based on the homology score or H-value cut-off scores (H > = 0.42). It computes the core, strain specific and dispensable/accessory genes based on the above threshold and computes the pan-genome. Further, the computed core and strain specific genes will be graphically displayed in the form of a 'Flower plot'. Link for the list of dispensable genes are also provided in the flower plot.

Tested Operating Systems: Ubuntu 11.04, Fedora core 14, Mac os X EI Captain, Windows 8

Architecture support: 64 bit

2. Dependencies for Linux

(1) Make sure that you have installed BLAST 2.4.0+ in your machine.

Installation commands

---

rpm -ivh ncbi-blast-2.4.0+-2.x86_64.rpm

OR

sudo apt-get install ncbi-blast+

(in case of Ubuntu)

---

(2) Make sure that your machine has latex installed.

Check it by typing command 'latex' in the terminal.

This is pdfTeXk, Version 3.141592-1.40.3 (Web2C 7.5.6).

If it is not installed in your machine. Install it manually.

(3) In case of Ubuntu type,

---

sudo apt-get install texlive-full

---

3. Dependencies for Windows

Download the following packages and make the default installation

ActivePerl-5.16.3.1604-MSWin32-x64-298023.exe

basic-miktex-2.9.5105-x64.exe

ncbi-blast-2.4.0+-win64

4. Dependencies for Mac

Download the following packages and make the default installation

Basic TeX

Website: (https://tug.org/mactex/)

ncbi-blast-2.4.0+

5. Running procedure

5a. PanGeT BLASTN mode

(1) Create folder in the name of a genus (or any other name) that you want to analyze. For example: Salmonella, Ecoli, Brucella etc.,

(2) The '.fna' files and '.ptt' files of strains you want to compare in a genus should be kept inside the created Genus folder. Specify input and output folder paths, H-value, E-value cutoffs in the parameter file.

(Ex: /home/ubuntu/input_files/**Salmonella**)

(Ex: C:\Users\username\input_files**Salmonella** in windows)

(3) Example : Parameter File for Linux / Mac OS

Input_directory /home/user/Documents/*Salmonella*

Output_directory /home/user/Documents/Salmonella_output

Hvalue 0.42

Evalue 1e-5

Threads 4

(4) Example : Parameter File for windows users

Input_directory C:\Users\username\Documents\S*almonella*

Output_directory C:\Users\username\Documents\Salmonella_result

Hvalue 0.42

Evalue 1e-5

Threads 4

(5) Run the executable file PanGeT using the following command in the terminal

./PanGeT_BLASTN parameter.txt or perl PanGeT_blastn.pl parameter.txt

(6) An entire list of RefSeq Ids (Ex: NC_000913) of the genus will be listed down to choose the reference strain

0. NC_003197

1 .NC_003198

2 .NC_004631

3. NC_006511

4. NC_006905

5. NC_010067

6. NC_011080

7. NC_011083

8. NC_011094

9. NC_011147

10. NC_011149

Enter the corresponding number entry to keep it as 'Query' strain

(7) The Program will create two folders 'process_genus_name' and 'output_genus_name'. (eg. process_Salmonella and output_Salmonella).

(8) Genus output folder (eg. output_Salmonella) has Result.pdf, which contains the pan-genome flower plot with links, core genes output named CORE and many sub-folders with the name of strain specific RefSeq ids, which has different outputs, such as list of conserved genes, dispensable genes, unique genes and the total genes.

(9) For further output details, please go through **Sample results** section in the manual.

5b. PanGeT BLASTP mode

(1) Create folder in the name of a genus (or any other name) that you want to analyze. For example: Salmonella, Ecoli, Brucella etc.,

(4) The '.faa' of strains you want to compare in a genus should be kept inside the created Genus folder. Specify input and output folder paths, H-value, E-value cutoffs in the parameter file.

(Ex: /home/ubuntu/input_files/**Salmonella**)

(Ex: C:\Users\username\input_files**Salmonella** in windows)

(3) Example : Parameter File for Linux / Mac OS

Input_directory /home/user/Documents/*Salmonella*

Output_directory /home/user/Documents/Salmonella_output

Hvalue 0.42

Evalue 1e-5

Threads 4

(4) Example : Parameter File for windows users

Input_directory C:\Users\username\Documents\S*almonella*

Output_directory C:\Users\username\Documents\Salmonella_result

Hvalue 0.42

Evalue 1e-5

Threads 4

(5) Run the executable file PanGeT using the following command in the terminal

./PanGeT_BLASTP parameter.txt or perl PanGeT_blastp.pl parameter.txt

(6) An entire list of RefSeq Ids (Ex: NC_000913) of the genus will be listed down to choose the reference strain

0. NC_003197

1 .NC_003198

2 .NC_004631

3. NC_006511

4. NC_006905

5. NC_010067

6. NC_011080

7. NC_011083

8. NC_011094

9. NC_011147

10. NC_011149

Enter the corresponding number entry to keep it as ‘Query’ strain

(7) The Program will create two folders 'process_genus_name' and 'output_genus_name'. (eg. process_Salmonella and output_Salmonella).

(8) Genus output folder (eg. output_Salmonella) has Result.pdf, which contains the pan-genome flower plot

with links, core genes output named CORE and many sub-folders with the name of strain specific RefSeq ids, which has different outputs, such as list of conserved genes, dispensable genes, unique genes and the total genes.

(9) For further output details go through **Sample results** section in the manual.

6. Sample results:

The **'Flower plot' (Result.pdf)** generated by **'PanGeT'** is an output, which describes the number of 'core genes' found within the entire 'Genus' or 'Species' while selecting the specific 'Query' and 'Reference' genomes. The flower pots will be opened in 'Adobe Reader' or 'Document viewer' utilities. The list of 'Dispensable genes' is shown through a link at the bottom right side of the flower plots.

Genus output folder

Genus output folder (eg. output_Salmonella) contains the following files

Result.pdf

Result.pdf has a flower plot which has hyper-links for the list of core, dispensable and strain specific genes (see example: Fig 1). When you click on the hyper-links, the core genes.html (see example: Fig 2), Strain specific genes.html (see example: Fig 3) will open in your default browser. Further, genes.html pages where users can download each core sequence through sequence link (see example: Fig 4) and also retrieve annotations from KEGG for every gene sequence (see example: Fig 5).

CORE

This file contains the core orthologue genes present in all the genomes found by reciprocal blast hits

Strain Folders

(Eg. NC_003917)

It will create sub-folders with the name of strain specific RefSeq ids. Each folder contains

1. **Conserved genes files**: conserved with one genome to 'n'genome you have run

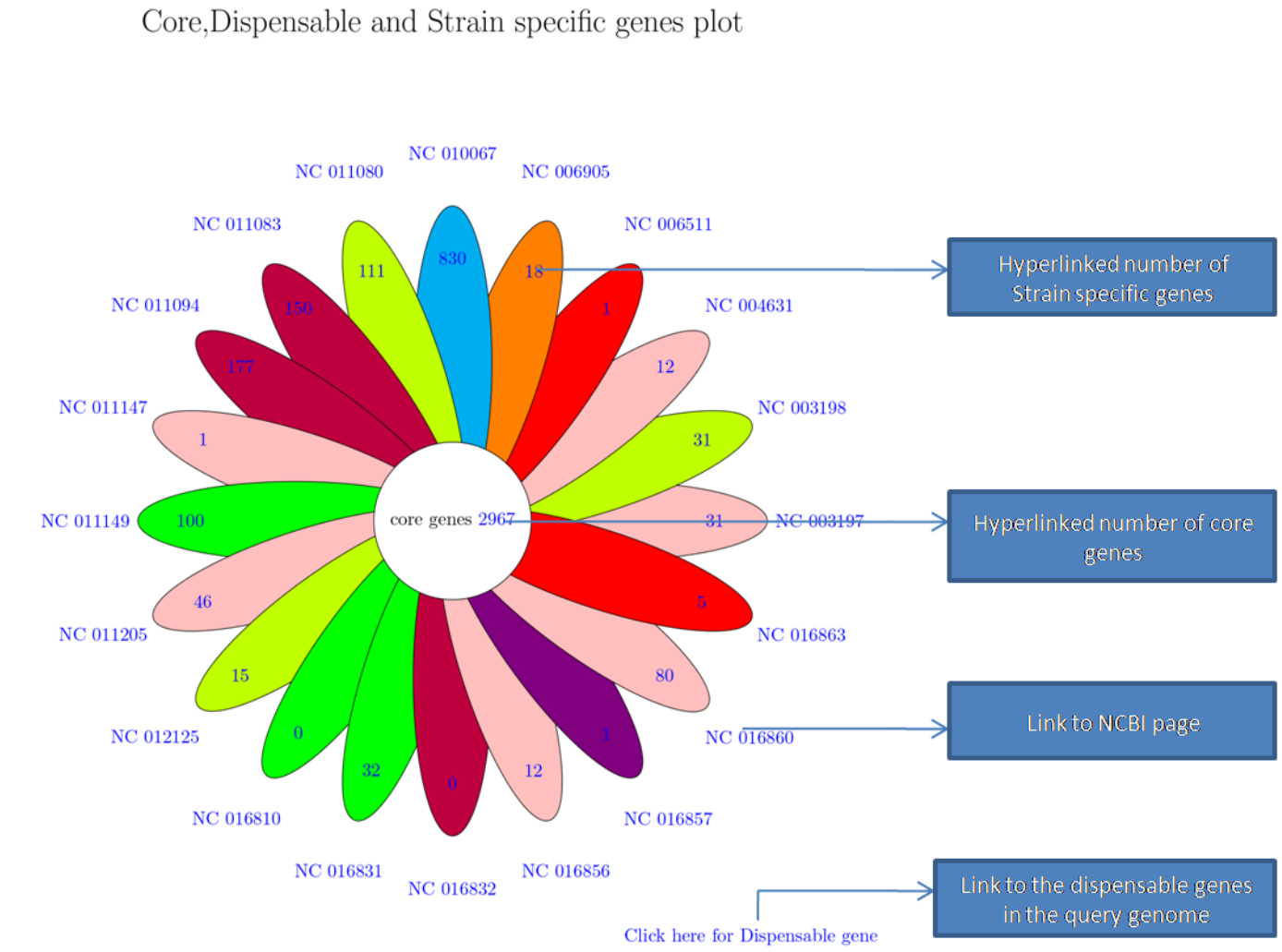
2. **Dispensable genes file**: contains genes present in at least 1 to n-1 genomes but not in all genomes.

3. **Conserved in all genomes files**: contains genes present in all genomes, but not reciprocal blast hits to every other genome.

4. **Unique genes file**: genes specific to the particular strain

5. **Total genes file:** All the genes with their H-value (Homology value) with other genome.

OUTPUT

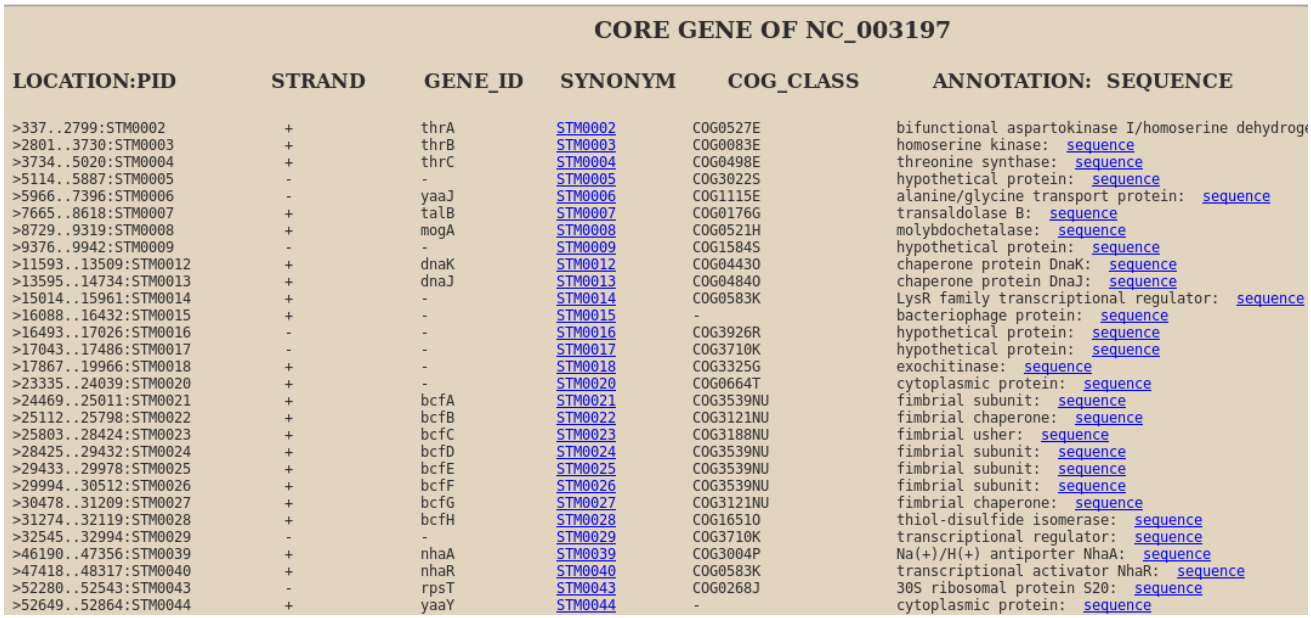


;

* click on the numbers to see the unique gene sequences
* click on the NCBI-IDs to go to NCBI genome page

1

Fig 1: Core, Dispensable and strain specific genes plot .

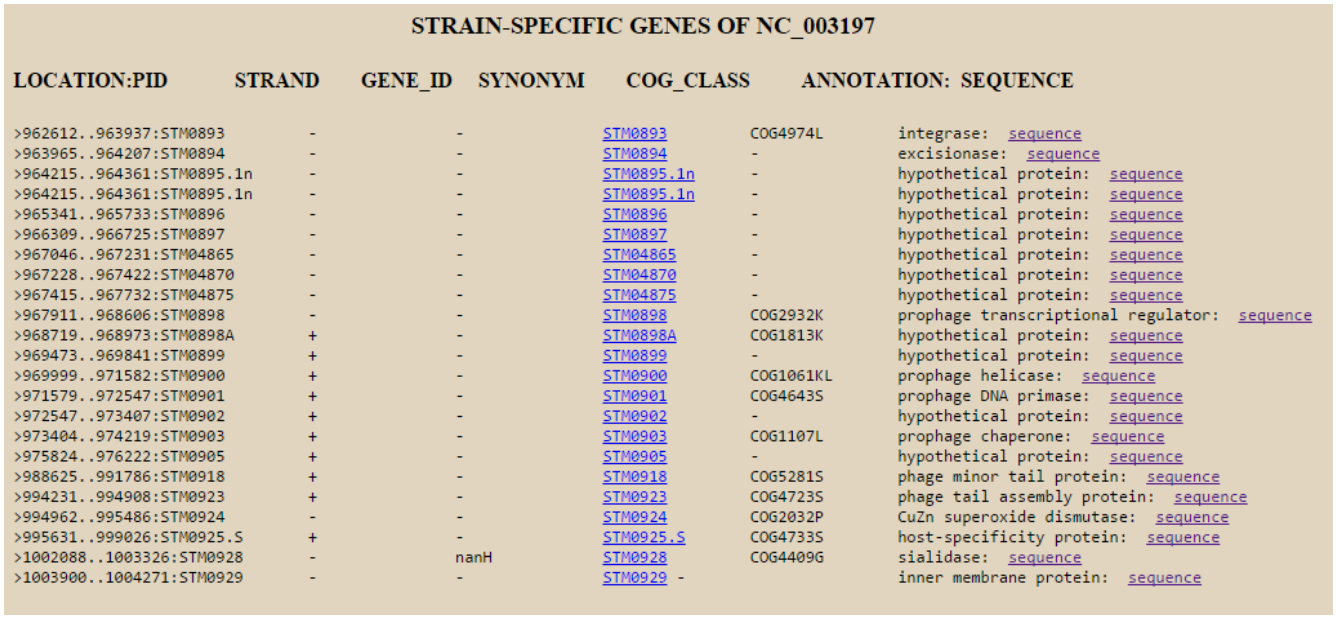


CORE GENE OF NC_003197

LOCATION:PID	STRAND	GENE_ID	SYNONYM	COG_CLASS	ANNOTATION: SEQUENCE
>337..2799:STM0002	+	thrA	STM0002	COG0527E	bifunctional aspartokinase I/homoserine dehydroge
>2801..3730:STM0003	+	thrB	STM0003	COG0083E	homoserine kinase: sequence
>3734..5020:STM0004	+	thrC	STM0004	COG0498E	threonine synthase: sequence
>5114..5887:STM0005	-	-	STM0005	COG3022S	hypothetical protein: sequence
>5966..7396:STM0006	-	yaaJ	STM0006	COG1115E	alanine/glycine transport protein: sequence
>7665..8618:STM0007	+	talB	STM0007	COG0176G	transaldolase B: sequence
>8729..9319:STM0008	+	mogA	STM0008	COG0521H	molybdochetalase: sequence
>9376..9942:STM0009	-	-	STM0009	COG1584S	hypothetical protein: sequence
>11593..13509:STM0012	+	dnaK	STM0012	COG0443O	chaperone protein DnaK: sequence
>13595..14734:STM0013	+	dnaJ	STM0013	COG0484O	chaperone protein DnaJ: sequence
>15014..15961:STM0014	+	-	STM0014	COG0583K	LysR family transcriptional regulator: sequence
>16088..16432:STM0015	+	-	STM0015	-	bacteriophage protein: sequence
>16493..17026:STM0016	-	-	STM0016	COG3926R	hypothetical protein: sequence
>17043..17486:STM0017	-	-	STM0017	COG3710K	hypothetical protein: sequence
>17867..19966:STM0018	+	-	STM0018	COG3325G	exochitinase: sequence
>23335..24039:STM0020	+	-	STM0020	COG0664T	cytoplasmic protein: sequence
>24469..25011:STM0021	+	bcfA	STM0021	COG3539NU	fimbrial subunit: sequence
>25112..25798:STM0022	+	bcfB	STM0022	COG3121NU	fimbrial chaperone: sequence
>25803..28424:STM0023	+	bcfC	STM0023	COG3188NU	fimbrial usher: sequence
>28425..29432:STM0024	+	bcfD	STM0024	COG3539NU	fimbrial subunit: sequence
>29433..29978:STM0025	+	bcfE	STM0025	COG3539NU	fimbrial subunit: sequence
>29994..30512:STM0026	+	bcfF	STM0026	COG3539NU	fimbrial subunit: sequence
>30478..31209:STM0027	+	bcfG	STM0027	COG3121NU	fimbrial chaperone: sequence
>31274..32119:STM0028	+	bcfH	STM0028	COG1651O	thiol-disulfide isomerase: sequence
>32545..32994:STM0029	-	-	STM0029	COG3710K	transcriptional regulator: sequence
>46190..47356:STM0039	+	nhaA	STM0039	COG3004P	Na(+)/H(+) antiporter NhaA: sequence
>47418..48317:STM0040	+	nhaR	STM0040	COG0583K	transcriptional activator NhaR: sequence
>52280..52543:STM0043	-	rpsT	STM0043	COG0268J	30S ribosomal protein S20: sequence
>52649..52864:STM0044	+	yaaY	STM0044	-	cytoplasmic protein: sequence

Fig 2: core genes output

Sequences of the core genes retrieved from NC_003197 genome through hyperlinked petals of the flower plot. The list of core genes identified from the PanGeT analysis is displayed with their genome location, strand orientation, Gene_ID, synonym, COG functional class, annotation and link to the corresponding sequence, as shown above. The synonyms were given with hyperlinks to the KEGG GENES database to get more information.



STRAIN-SPECIFIC GENES OF NC_003197

LOCATION:PID	STRAND	GENE_ID	SYNONYM	COG_CLASS	ANNOTATION: SEQUENCE
>962612..963937:STM0893	-	-	STM0893	COG4974L	integrase: sequence
>963965..964207:STM0894	-	-	STM0894	-	excisionase: sequence
>964215..964361:STM0895.1n	-	-	STM0895.1n	-	hypothetical protein: sequence
>964215..964361:STM0895.1n	-	-	STM0895.1n	-	hypothetical protein: sequence
>965341..965733:STM0896	-	-	STM0896	-	hypothetical protein: sequence
>966309..966725:STM0897	-	-	STM0897	-	hypothetical protein: sequence
>967046..967231:STM04865	-	-	STM04865	-	hypothetical protein: sequence
>967228..967422:STM04870	-	-	STM04870	-	hypothetical protein: sequence
>967415..967732:STM04875	-	-	STM04875	-	hypothetical protein: sequence
>967911..968606:STM0898	-	-	STM0898	COG2932K	prophage transcriptional regulator: sequence
>968719..968973:STM0898A	+	-	STM0898A	COG1813K	hypothetical protein: sequence
>969473..969841:STM0899	+	-	STM0899	-	hypothetical protein: sequence
>969999..971582:STM0900	+	-	STM0900	COG1061KL	prophage helicase: sequence
>971579..972547:STM0901	+	-	STM0901	COG4643S	prophage DNA primase: sequence
>972547..973407:STM0902	+	-	STM0902	-	hypothetical protein: sequence
>973404..974219:STM0903	+	-	STM0903	COG1107L	prophage chaperone: sequence
>975824..976222:STM0905	+	-	STM0905	-	hypothetical protein: sequence
>988625..991786:STM0918	+	-	STM0918	COG5281S	phage minor tail protein: sequence
>994231..994908:STM0923	+	-	STM0923	COG4723S	phage tail assembly protein: sequence
>994962..995486:STM0924	-	-	STM0924	COG2032P	CuZn superoxide dismutase: sequence
>995631..999026:STM0925.S	+	-	STM0925.S	COG4733S	host-specificity protein: sequence
>1002088..1003326:STM0928	-	nanH	STM0928	COG4409G	sialidase: sequence
>1003900..1004271:STM0929	-	-	STM0929	-	inner membrane protein: sequence

Fig 3: Strain specific genes output

Sequences of the strain specific genes retrieved from NC_003197 genome through hyperlinked petals of the flower plot. The list of strain specific genes identified from the PanGeT analysis is displayed with their genome location, strand orientation, Gene_ID, synonym, COG functional class, annotation and link to the corresponding sequence as shown above. The synonyms are given with hyperlinks to the KEGG GENES database to get more information.

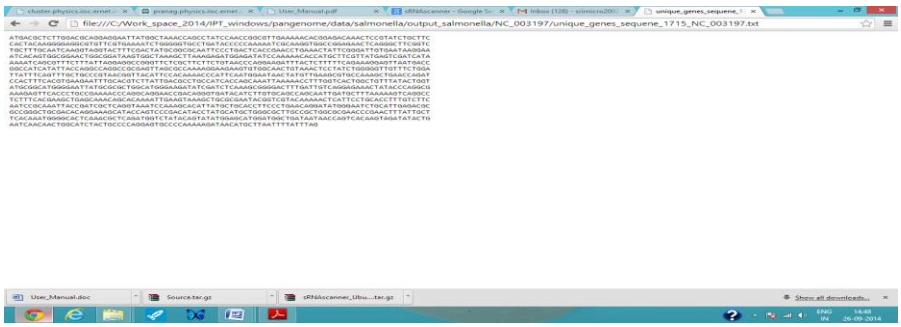


Fig 4: The sequence retrieved for the gene id 'STM_0893' is shown above.

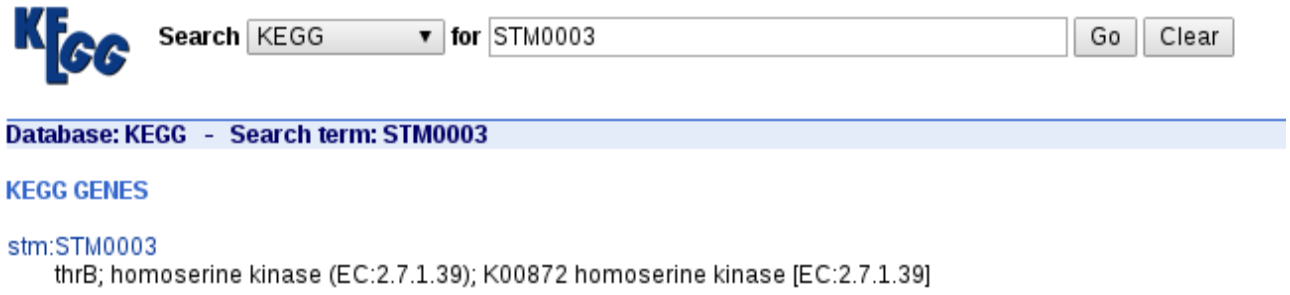


Fig 5: KEGG Genes database entry for 'STM0003' gene entry.

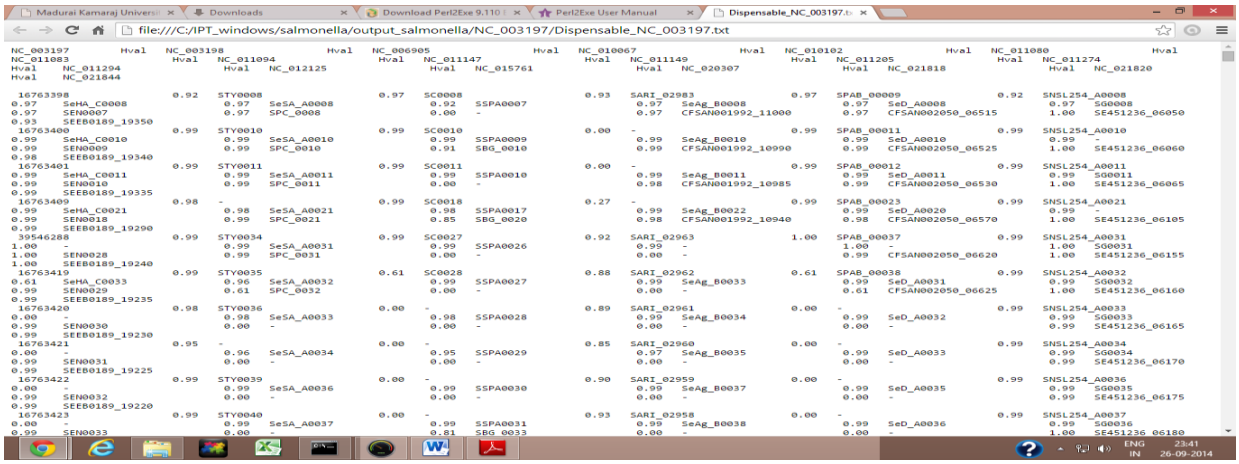


Fig 6 : List of Dispensable genes output

The list of 'Dispensable genes' identified by PanGeT is shown above with their respective homology scores in the reference genomes. The above list may be imported into MS-EXCEL to get more information.

7. Pathogenic and non pathogenic strains comparison

If the user want to find the genes which are conserved only in few pathogens and absent in non pathogens, user can run the program *'Find_genes_conserved_few_while_comparing_many.pl* '. Instructions for running is given inside the program file.