PanGeT: A tool for Pan-Genomics studies

PanGeT v.1.0

User Manual

Last modified on October, 2016

TABLE OF CONTENTS

1. Introduction	3
2. Dependencies for Linux	3
3. Dependencies for Windows	4
4. Dependencies for Mac	4
5. Running Procedure	4
5a. PanGeT_BLASTN	4
5b. PanGeT_BLASTP	6
6. Sample Results	8
7. Pathogenic and nonpathogenic strain comparison	13

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. <u>Dependencies for Linux</u>

(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. <u>Dependencies for Windows</u>

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. <u>Running procedure</u>

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)

PanGeT: A tool for Pan-Genomics studies

(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\Salmonella					
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

(4) Example : Parameter File for windows users

4

Input_directory C:\Users\username\Documents\Salmonella

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. <u>Sample results:</u>

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

Core, Dispensable and Strain specific genes plot



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

10

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				
>3372799:STM0002	+	thrA	STM0002	C0G0527E	bifunctional aspartokinase I/homoserine dehydroge				
>28013730:STM0003	+	thrB	STM0003	C0G0083E	homoserine kinase: sequence				
>37345020:STM0004	+	thrC	STM0004	C0G0498E	threonine synthase: sequence				
>51145887:STM0005	-	-	STM0005	C0G3022S	hypothetical protein: sequence				
>59667396:STM0006	-	yaaJ	STM0006	C0G1115E	alanine/glycine transport protein: sequence				
>76658618:STM0007	+	ťalB	STM0007	C0G0176G	transaldolase B: sequence				
>87299319:STM0008	+	mogA	STM0008	C0G0521H	molybdochetalase: sequence				
>93769942:STM0009	-		STM0009	C0G1584S	hypothetical protein: sequence				
>1159313509:STM0012	+	dnaK	STM0012	C0G04430	chaperone protein DnaK: sequence				
>1359514734:STM0013	+	dnaJ	STM0013	C0G04840	chaperone protein DnaJ: sequence				
>1501415961:STM0014	+	-	STM0014	C0G0583K	LysR family transcriptional regulator: sequence				
>1608816432:STM0015	+	-	STM0015	-	bacteriophage protein: sequence				
>1649317026:STM0016	-	-	STM0016	C0G3926R	hypothetical protein: sequence				
>1704317486:STM0017	-	-	STM0017	C0G3710K	hypothetical protein: sequence				
>1786719966:STM0018	+	-	STM0018	C0G3325G	exochitinase: sequence				
>2333524039:STM0020	+	-	STM0020	C0G0664T	cytoplasmic protein: sequence				
>2446925011:STM0021	+	bcfA	STM0021	C0G3539NU	fimbrial subunit: sequence				
>2511225798:STM0022	+	bcfB	STM0022	C0G3121NU	fimbrial chaperone: sequence				
>2580328424:STM0023	+	bcfC	STM0023	C0G3188NU	fimbrial usher: <u>sequence</u>				
>2842529432:STM0024	+	bcfD	STM0024	C0G3539NU	fimbrial subunit: sequence				
>2943329978:STM0025	+	bcfE	STM0025	C0G3539NU	fimbrial subunit: <u>sequence</u>				
>2999430512:STM0026	+	bcfF	STM0026	C0G3539NU	fimbrial subunit: <u>sequence</u>				
>3047831209:STM0027	+	bcfG	STM0027	C0G3121NU	fimbrial chaperone: sequence				
>3127432119:STM0028	+	bcfH	STM0028	C0G16510	thiol-disulfide isomerase: sequence				
>3254532994:STM0029	-	-	STM0029	C0G3710K	transcriptional regulator: sequence				
>4619047356:STM0039	+	nhaA	STM0039	C0G3004P	Na(+)/H(+) antiporter NhaA: sequence				
>4741848317:STM0040	+	nhaR	STM0040	C0G0583K	transcriptional activator NhaR: sequence				
>5228052543:STM0043	-	rpsT	STM0043	C0G0268J	30S ribosomal protein S20: sequence				
>5264952864:STM0044	+	vaaY	STM0044	-	cvtoplasmic 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 S	TRAND	GENE_ID	SYNONYM	COG_CLASS	S ANNOTA	ATION: SEQUENCE				
>962612963937:STM0893 >963965964207:STM0894	-	-		STM0893 STM0894	COG4974L -	integrase: <u>sequence</u> excisionase: <u>sequence</u> hypothetical protein: cequence				
>964215964361:STM0895.1 >964215964361:STM0895.1 >965341965733:STM0896	n - -	-		<u>STM0895.1n</u> <u>STM0896</u>	:	hypothetical protein: <u>sequence</u> hypothetical protein: <u>sequence</u>				
>966309966725:STM0897 >967046967231:STM04865 >967228967422:STM04870	-	-		<u>STM0897</u> <u>STM04865</u> <u>STM04870</u>	-	hypothetical protein: <u>sequence</u> hypothetical protein: <u>sequence</u> hypothetical protein: <u>sequence</u>				
>967415967732:STM04875 >967911968606:STM0898 >968719968973:STM0898A	- - +	-		<u>STM04875</u> <u>STM0898</u> <u>STM0898A</u>	- COG2932K COG1813K	hypothetical protein: <u>sequence</u> prophage transcriptional regulator: <u>sequence</u> hypothetical protein: <u>sequence</u>				
>969473969841:STM0899 >969999971582:STM0900 >971579972547:STM0901	+ + +	-		STM0899 STM0900 STM0901	- COG1061KL COG4643S	hypothetical protein: <u>sequence</u> prophage helicase: <u>sequence</u> prophage DNA primase: <u>sequence</u>				
>972547973407:STM0902 >973404974219:STM0903 >975824976222:STM0905	+ + +	-		<u>STM0902</u> <u>STM0903</u> STM0905	- COG1107L -	hypothetical protein: <u>sequence</u> prophage chaperone: <u>sequence</u> hypothetical protein: <u>sequence</u>				
>988625991786:STM0918 >994231994908:STM0923 >994962995486:STM0924	+ + -	-		<u>STM0918</u> <u>STM0923</u> STM0924	COG5281S COG4723S COG2032P	phage minor tail protein: <u>sequence</u> phage tail assembly protein: <u>sequence</u> CuZn superoxide dismutase: <u>sequence</u>				
>995631999026:STM0925.S >10020881003326:STM0928 >10039001004271:STM0929	+ - -	- n	anH	<u>STM0925.5</u> <u>STM0928</u> <u>STM0929</u> -	C0G4733S C0G4409G	host-specificity protein: <u>sequence</u> sialidase: <u>sequence</u> inner membrane protein: <u>sequence</u>				

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.

🖊 🗅 Mađuraj Kamaraj Universit 🗴 🗸 🕭 Downloads 🛛 x X 🚷 Download Perl2Exe 9.110 1 x X 🏫 Perl2Exe User Manual 🛛 x Y 🗅 Dispensable NC_003197.ts x 🚺 🗖											
C n E file:///C/IPT_windows/salmonella/NC_003197/Dispensable_NC_003197.txt											
NC_003197 Hval NC_011083 Hval NC_011294 Hval NC_021844	NC_003198 Hval M	8 Hval NC_011094 Hval NC_012125	NC_0069 Hval	05 NC_01114 Hval	Hval NC_015761	NC_0100 Hval	67 Hval NC_011149 Hval NC_020307	NC_0101 Hval	02 Hval NC_011205 Hval NC_021818	NC_0110 Hval	80 Hval Â NC_011274 Hval Â Hval NC_021820
16763398 0.97 SeHA_C0008 0.97 SEN0007 0.93 SEEB0189_19350	0.92 1	5TY0008 0.97 SeSA_A0008 0.97 SPC_0008	0.97	SC0008 0.92 0.00	SSPA0007 -	0.93	SARI_02983 0.97 SeAg_80008 0.97 CFSAN001992_110	0.97 00	SPAB_00009 0.97 SeD_A0008 0.97 CFSAN002050_065	0.92 15	SNSL254_A0008 0.97 SG0008 1.00 SE451236_06050
16763400 0.99 SeHA_C0010 0.99 SEN0009 0.98 SEEB0189 19340	0.99 5	5TY0010 0.99 SeSA_A0010 0.99 SPC_0010	0,99	5C0010 0.99 0.91	SSPA0009 SBG_0010	0.00	- 0.99 SeAg_B0010 0.99 CFSAN001992_109	0.99 90	SPAB_00011 0.99 SeD_A0010 0.99 CFSAN002050_065	0,99 25	SNSL254_A0010 0.99 - 1.00 SE451236_06060
16763401 0.99 SeHA_C0011 0.99 SEN0010 0.99 SEER0180 10335	0.99	STY0011 0.99 SeSA_A0011 0.99 SPC_0011	0.99	SC0011 0.99 0.00	SSPA0010 -	0.00	0.99 SeAg_B0011 0.98 CFSAN001992_109	0.99 85	SPAB_00012 0.99 SeD_A0011 0.99 CFSAN002050_065	0.99 30	SNSL254_A0011 0.99 500011 1.00 SE451236_06065
16763409 0.99 SeHA_C0021 0.99 SEN0018 0.99 SEN0018	0.98 -	0.98 SeSA_A0021 0.99 SPC_0021	0.99	SC0018 0.98 0.85	SSPA0017 SBG_0020	0.27	- 0.99 SeAg_B0022 0.98 CFSAN001992_109	0.99 40	SPAB_00023 0.99 SeD_A0020 0.98 CFSAN002050_065	0.99 70	SN5L254_A0021 0.99 - 1.00 SE451236_06105
39546288 1.00 - 1.00 SEN0028	0.99 5	STY0034 0.99 SeSA_A0031 0.99 SPC_0031	0.99	SC0027 0.99 0.00	SSPA0026	0.92	SARI_02963 0.99 - 0.00 -	1.00	SPAB_00037 1.00 - 0.99 CFSAN002050_066	0.99 20	SNSL254_A0031 1.00 SG0031 1.00 SE451236_06155
16763419 0.61 SeHA_C0033 0.99 SEN0029	0.99 5	STY0035 0.96 SeSA_A0032 0.61 SPC_0032	0.61	SC0028 0.99 0.00	SSPA0027	0.88	SARI_02962 0.99 SeAg_80033 0.00 -	0.61	SPAB_00038 0.99 SeD_A0031 0.61 CFSAN002050_066	0.99 25	SNSL254_A0032 0.99 560032 1.00 5E451236_06160
0.99 SEN0030 0.99 SEN0030	0.98 5	STY0036 0.98 SeSA_A0033 0.00 -	0.00	0.98 0.00	SSPA0028	0.89	SARI_02961 0.99 SeAg_B0034 0.00 -	0.00	0.99 SeD_A0032 0.00 -	0.99	SNSL254_A0033 0.99 560033 0.99 SE451236_06165
0.00 - 0.99 SEN0031	0.95	0.96 SeSA_A0034 0.00 -	0.00	0.95	SSPA0029 -	0.85	SARI_02960 0.97 SeAg_B0035 0.00 -	0.00	0.99 SeD_A0033 0.00 -	0.99	SNSL254_A0034 0.99 SG0034 0.99 SE451236_06170
0.99 SEN0032	0.99 5	5TY0039 0.99 SeSA_A0036 0.00 -	0.00	0.99	SSPA0030 -	0.90	SARI_02959 0.99 SeAg_80037 0.00 -	0.00	0.99 SeD_A0035 0.00 -	0.99	SNSL254_A0036 0.99 SG0035 0.99 SE451236_06175
0.99 SEC80189_19220 16763423 0.00 - 0.99 SEN0033	0.99 5	5TY0040 0.99 SeSA_A0037 0.00 -	0.00	0.99 0.81	SSPA0031 SBG 0033	0.93	SARI_02958 0.99 SeAg_80038 0.00 -	0.00	0.99 SeD_A0036 0.00 -	0.99	SNSL254_A0037 0.99 560036 1.00 SE451236 06180 ~
	-			WV4						?	► 🗣 🕩 ENG 23:41 IN 26-09-2014

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.