PanCoreGen

Installer includes:

A. PCG.exe (the user interface)

B. script.exe (the main program)

C. makeblastdb.exe (program to format genome sequences for BLAST)

D. blastn.exe (program for standalone Blast operations)

E. Documentation.pdf (this documentation)

F. Sample_input folder (containing sample input files)

Contents:

Introduction

PanCoreGen is a user-friendly standalone application for pan- and core-genomic profiling of protein coding genes within any microbial species. The pan-genome of a species is the sum of non-redundant genomic regions from its representative genomes. Pan-genome incorporates core genomic regions present in all genomes, and accessory regions found in one or more but not in all genomes. Analysis of the pan-genomic profile has direct evolutionary significance, such as for the study of within-species population structure or cross-species metagenomics.

Features

PanCoreGen includes the following functionalities:

(1) pan-/core-genomic profiles across the entire sample set;

(2) pan-/core-genomic profiles within user-defined strain-groups;

(3) annotation of user-provided draft genomes;

(4) detection of unidentified genes in annotated genomes; and

(5) direct download of microbial genomes in Genbank (.gbk) and Fasta (.fna) format.

Applications

Based on user-defined threshold values of nucleotide sequence-identity and length-coverage for identification of orthologs, this tool distinguishes each gene as core (i.e. present in all genomes of the sample set), mosaic (i.e. present in multiple but not all genomes), or unique (i.e. present only in one of the annotated genomes) within any set of genomes. In order to build the pan-genomic database of orthologous genes, PanCoreGen performs BLAST (http://blast.ncbi.nlm.nih.gov/Blast.cgi?CMD=Web&PAGE_TYPE=BlastDocs&DOC_ TYPE=Download) against genes from each annotated genome to detect orthologs in the rest of the genomes. This algorithm provides information on genes previously unidentified in an annotated genome. Importantly, such an approach offers annotation of draft genomes or contigs based on all unique annotated genes present in the sample set. Also, to understand evolution of virulence in a specific pathotype or group of strains in a given species (e.g. extra-intestinal pathogenic *E. coli*, *Shigella*, human-restricted systemically invasive serovars of *Salmonella*, etc.), one crucial need is to extract the pan- and core-genomic information within that group.

PanCoreGen generates all such profiles based on groups defined by users. In addition, our tool outputs a sequence dataset for each gene, incorporating genomic location information on orthologs in each genome. This allows users to study genes of interest for the analysis of sequence diversity, detection of positive selection footprints etc.

Installation

Version and operating system - The present version of PanCoreGen (PCG) is available only for 64bit version of Windows systems (OS XP or higher).

Source for installer download -
https://sourceforge.net/projects/pancoregen1/

The stepwise installation procedure -

1. Download "PanCoreGen _installer" from
   https://sourceforge.net/projects/pancoregen1/

2. Run the installer to automatically install "PanCoreGen" on computer, preferably in the default C:\PanCoreGen directory. Some Windows operating systems might require administrative privilege to allow installation through User Account Control.

3. The installed folder "PanCoreGen" includes 'PCG.exe', 'script.exe', 'makeblastdb.exe', 'blastn.exe', 'Sample_input' (containing the sample genome files in GenBank and FASTA format) and 'Documentation.pdf' .

Input Files

1. Analysis of pan-genomic profile of genomes

(a) Files of annotated complete genome sequences in GenBank format (e.g. Genome1.txt through Genome5.txt sample files in “Sample_input” folder). At least two annotated genome files are required. User can download the complete genome sequence one by one of their interest, if genomic sequence information for those bacteria of interest are available from GenBank. The user also has the option to specify their own fully annotated genome sequence(s) in Genbank format.

(b) Fasta-formatted sequence files for the draft/contigs (incomplete) genomes (single/multi FASTA). In case of a multi-FASTA file, the program concatenates it to a single FASTA file for further analysis. At the end of the each file name, there should be ‘-draft’ to recognize it as a draft/contigs file (e.g. draft1-draft.txt, draft2-draft.txt in “Sample_input” folder).

2. Group-specific analysis

A maximum of 10 groups can be analyzed in a single run. This feature uses genome files provided in pan-genomic profile analysis.

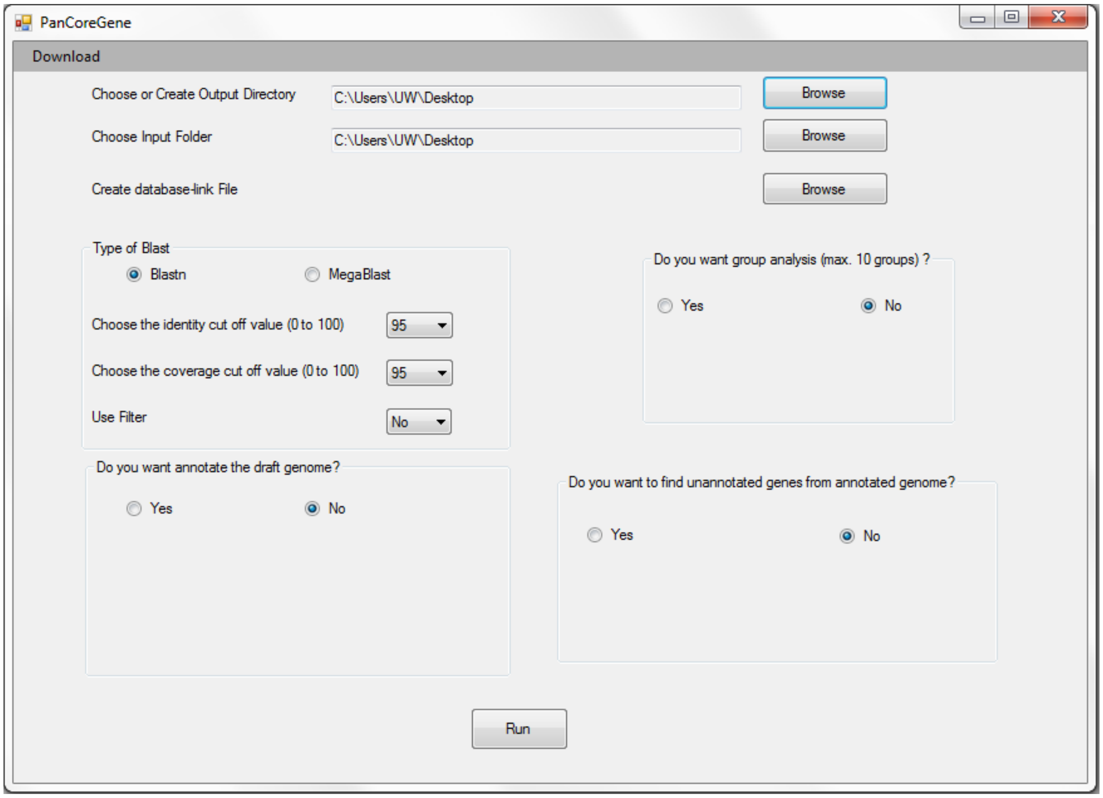
3. Annotation of draft/contigs genomes

Fasta-formatted sequence files for the draft/contigs genomes (single/multi FASTA). In case of a multi-FASTA file, the program concatenates it to a single FASTA file for further analysis. At the end of the file name there should be ‘-draft’ to recognize it as a draft/contigs file (e.g. draft1-draft.txt, draft2-draft.txt in “Sample_input” folder).

4. Identifying un-annotated genes in annotated genomes

This feature uses files of annotated genome sequences in Genbank format provided in pan-genomic profile analysis.

All the genome data files should be stored in a specific folder, which will later be used by the PanCoreGen as the “Input folder”.



Choose or Create Output Directory C:\Users\UW\Desktop

Choose Input Folder C:\Users\UW\Desktop

Create database-link File

Type of Blast ● Blastn ○ MegaBlast

Choose the identity cut off value (0 to 100) 95

Choose the coverage cut off value (0 to 100) 95

Use Filter No

Do you want group analysis (max. 10 groups) ? ○ Yes ● No

Do you want annotate the draft genome? ○ Yes ● No

Do you want to find unannotated genes from annotated genome? ○ Yes ● No

PanCoreGen interface

Running PanCoreGen

1. Start PanCoreGen by any of the following options

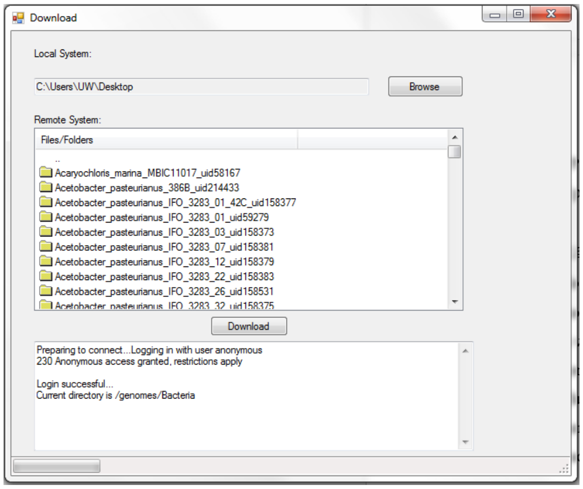
(a) By clicking a shortcut icon for PanCoreGen in Desktop or in Start Menu;

(b) By clicking "PCG" application in installed "PanCoreGen" folder;

(c) By typing "PCG" or "PCG.exe" inside the "C:\ PanCoreGen" directory from command prompt window.

2. Download fully annotated bacterial genomes from Genbank

(a) Click download toolbar Download to connect to NCBI ftp server



C:\Users\UW\Desktop

..
Acaryochloris_marina_MBIC11017_uid58167
Acetobacter_pasteurianus_386B_uid214433
Acetobacter_pasteurianus_IFO_3283_01_42C_uid158377
Acetobacter_pasteurianus_IFO_3283_01_uid59279
Acetobacter_pasteurianus_IFO_3283_03_uid158373
Acetobacter_pasteurianus_IFO_3283_07_uid158381
Acetobacter_pasteurianus_IFO_3283_12_uid158379
Acetobacter_pasteurianus_IFO_3283_22_uid158383
Acetobacter_pasteurianus_IFO_3283_26_uid158531
Acetobacter_pasteurianus_IFO_3283_32_uid158375

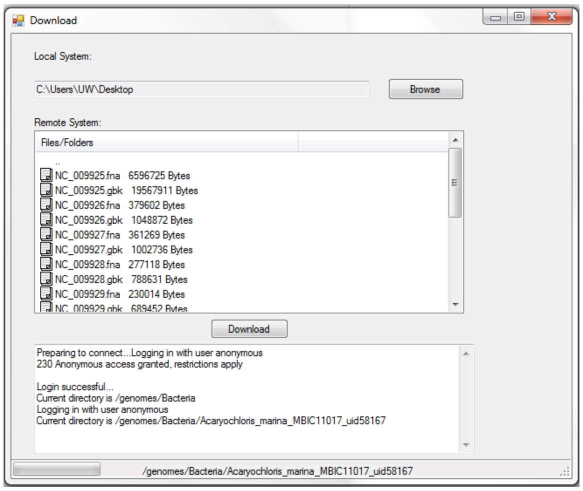
```
Preparing to connect...Logging in with user anonymous
230 Anonymous access granted, restrictions apply

Login successful...
Current directory is /genomes/Bacteria
```

(b) Create or choose folder by using browse option Browse and the downloaded files will be stored in that folder (by default the files will be downloaded in Desktop folder).

(c) Open genome folder of interest in "Remote System:" window by double-clicking on it.

(d) Only .gbk (fully annotated Genbank format) and .fna (Fasta formatted whole genome nucleotide sequence) of chromosome(s) and plasmid(s) (if present) will be shown with respective file sizes within the folder.



Local System: C:\Users\UW\Desktop

Remote System:

```
Files/Folders
..
NC_009925.fna   6596725 Bytes
NC_009925.gbk   19567911 Bytes
NC_009926.fna   379602 Bytes
NC_009926.gbk   1048872 Bytes
NC_009927.fna   361269 Bytes
NC_009927.gbk   1002736 Bytes
NC_009928.fna   277118 Bytes
NC_009928.gbk   788631 Bytes
NC_009929.fna   230014 Bytes
NC_009929.gbk   689452 Bytes
```

```
Preparing to connect...Logging in with user anonymous
230 Anonymous access granted, restrictions apply

Login successful...
Current directory is /genomes/Bacteria
Logging in with user anonymous
Current directory is /genomes/Bacteria/Acaryochloris_marina_MBIC11017_uid58167
```

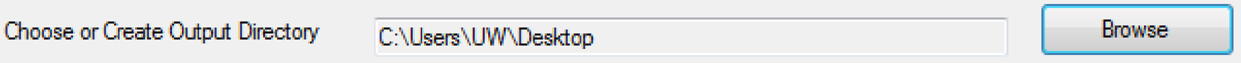
/genomes/Bacteria/Acaryochloris_marina_MBIC11017_uid58167

(e) Select the desired file by clicking on it, for selecting multiple files click on the desired files while holding down the Ctrl key.

(f) Use download button Download in order to download the desired file(s).

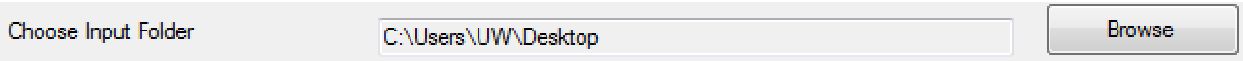
3. Choose or create input and output folders

(a) Choose or create an output folder in user specified location in order to store the outputs by clicking browse button Browse (Default is Desktop).



Choose or Create Output Directory C:\Users\UW\Desktop

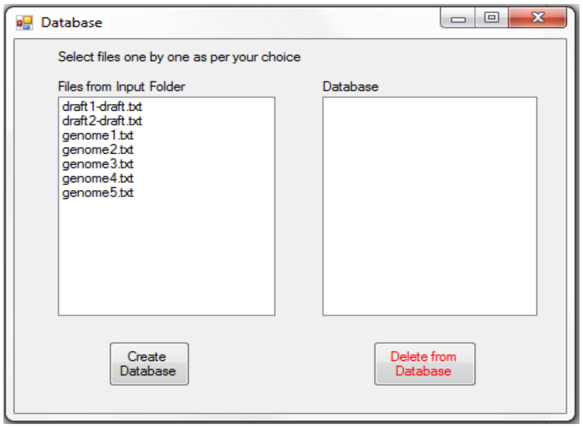
(b) Choose the input folder where all the genome data files are stored (e.g. "sample_input" folder) by clicking browse button Browse (default is Desktop).



Choose Input Folder C:\Users\UW\Desktop

4. Create database-link file

(a) Use browse button Browse in "Create database-link File" option in order to create the database-link file, which will be used for subsequent analysis. All the files present in the input folder can be seen in the left "Files from Input Folder" window from which user can select their desired files for analysis.



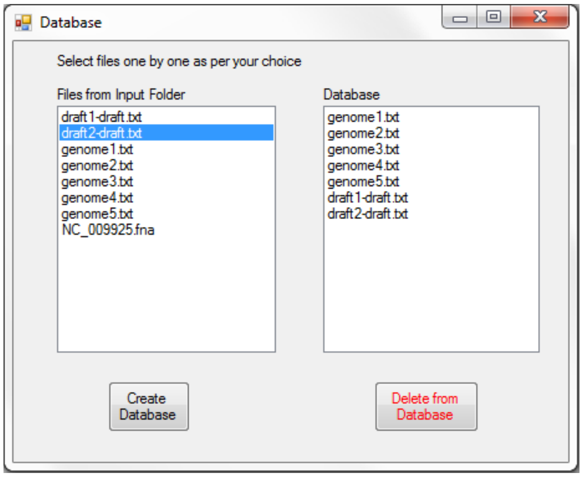
Select files one by one as per your choice

Files from Input Folder

```
draft1-draft.txt
draft2-draft.txt
genome1.txt
genome2.txt
genome3.txt
genome4.txt
genome5.txt
```

Database

(b) User can select genomes according to their choice of order from left "Files from Input Folder" window to the right "Database' window (e.g. select genome1.txt through genome5.txt then select draft1-draft.txt and draft2-draft.txt genome files one by one). User must select the annotated genome names first, then the draft/contigs genome names.



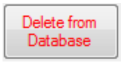
Select files one by one as per your choice

Files from Input Folder

```
draft1-draft.txt
draft2-draft.txt
genome1.txt
genome2.txt
genome3.txt
genome4.txt
genome5.txt
NC_009925.fna
```

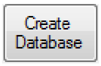
Database

```
genome1.txt
genome2.txt
genome3.txt
genome4.txt
genome5.txt
draft1-draft.txt
draft2-draft.txt
```

(c) Any wrong genome file or genome order incorporation can be deleted by using 'Delete from Database'  button after selecting that file in

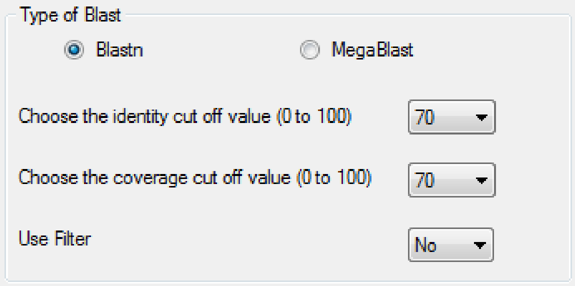
“Database” window.

(d) After selecting all the genomes, click on “Create Database’ button in order to create the database for subsequent runs. This is very important step as without creating the database-link file the program will not run further.



5. Analysis of pan-genomic profile of genomes

(a) Choose ‘Blastn’ or ‘Megablast’ (ideal for very closely-related strains) by selecting appropriate field (‘Blastn’ is the default option).



Type of Blast

● Blastn ○ MegaBlast

Choose the identity cut off value (0 to 100): 70

Choose the coverage cut off value (0 to 100): 70

Use Filter: No

(b) Select cut-off value for % sequence-identity for identification of orthologs from drop down menu of ‘Enter the identity cut off value’ (e.g. 70 for 70%; 95 is default value).

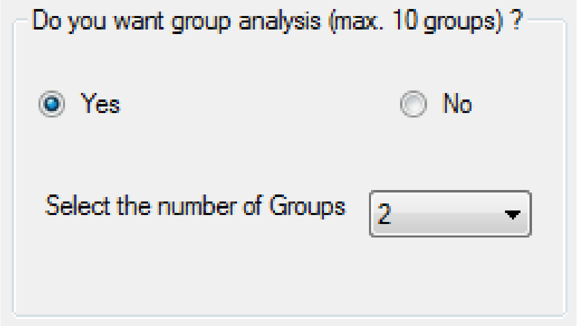
(c) Select cut-off value for % sequence-identity for identification of orthologs from drop down menu of ‘Enter the coverage cut off value’. (e.g. 70 for 70%; 95 is default value).

(d) Set “Use filter” to “Yes” to turn on the DUST filter of BLAST and to mask off the low compositional complexity regions from the query sequence. This filter is turned off as the default option in PanCoreGen to avoid possible exclusion of orthologs due to masking of any query sequence. On the other hand, enabling DUST filter can exclude hits which are statistically significant but biologically uninteresting. For further details on the use of this filter, we recommend users the

BLAST help page (http://www.ncbi.nlm.nih.gov/BLAST/blastcgihelp.shtml#filter).

6. Group specific analysis

(a) Choose this analysis by selecting 'Yes' option ('No' is default option).



Do you want group analysis (max. 10 groups) ?

● Yes ○ No

Select the number of Groups: 2

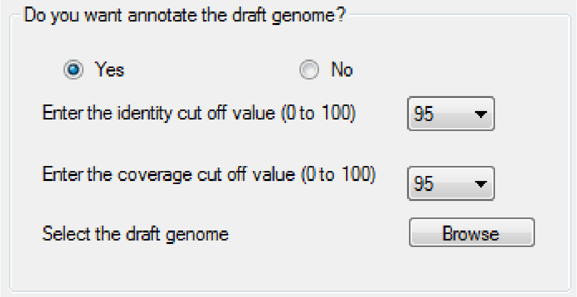
(b) Specify the number of groups from the drop-down option of 'Select the number of Groups' (maximum 10 groups can be selected)

(c) Once the number of groups is selected the window will appear to select representative genomes for each group (e.g., if you select 2 groups, 2 options 'Select Group 1' and 'Select Group 2' will appear one after another).

(c) Select representative genomes for each group by assigning respective genome files belonging to that group (e.g., user can select genome1.txt and genome2.txt for Group1 whereas, genome4.txt and genome5.txt for group2). For multiple genome selection, user can hold down Ctrl key while selecting the genomes.

7. Annotation of draft/contigs genomes

(a) Choose this analysis by selecting 'Yes' option ('No' is default option).



Do you want annotate the draft genome?

● Yes ○ No

Enter the identity cut off value (0 to 100): 95

Enter the coverage cut off value (0 to 100): 95

Select the draft genome

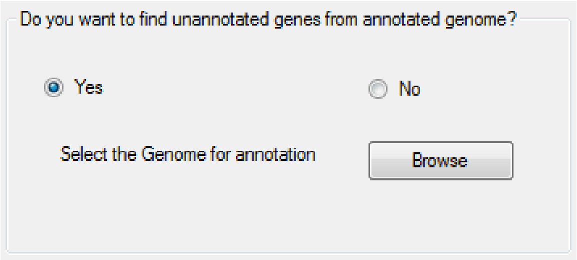
(b) Specify cut-off value for % sequence-identity for identification of orthologs (numbers only, from 0 to 100) (e.g. 95 for 95%).

(c) Specify cut-off value for % coverage of sequence-length for identification of orthologs (numbers only, from 0 to 100) (e.g. 95 for 95%).

(d) Specify the draft genomes to be annotated by selecting them using the browse option (e.g. user can select the draft genomes ‘draft1-draft.txt’ and ‘draft2-draft.txt’). For multiple genome selection, user can hold down Ctrl key while selecting the genomes.

8. Identifying un-annotated genes in annotated genomes

(a) Choose the analysis by selecting ‘Yes’ option (‘No’ is default option).



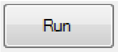
Do you want to find unannotated genes from annotated genome?

● Yes ○ No

Select the Genome for annotation

(b) Specify genomes by selecting them using the browse option (e.g. user can select the annotated genomes ‘genome4.txt’ and ‘genome5.txt’). For multiple genome selection, user can hold down Ctrl key while selecting the genomes.

9. Running PanCoreGen

After selecting all the options use run button  to start the program. Then it will show running and will be ready for next run again when the run button is shown.

Output files

Output-results will be created in the user-specified folder.

1. Analysis of pan-genomic profile of genomes

(a) A profile of core genes in a spreadsheet "Core_genes.xls".

(b) A complete profile of core, mosaic and strain-specific genes in spreadsheet "Gene_distribution.xls".

(c) A spreadsheet "Core_stop_genes.xls" for core genes with premature stop codons in one or multiple strains.

(d) "panseq.txt" file for nucleotide sequences of all unique genes based on all annotated genomes as reference.

(e) A FASTA-formatted nucleotide sequence dataset for each gene (e.g. Gene1.fasta, Gene2.fasta...GeneN.fasta) including all orthologs from sample set genomes, both annotated and un-annotated. For a given gene, the headers for orthologous sequences show respective genomic coordinates.

2. Group specific analysis

(a) Spreadsheet file for each user-assigned group with name "Group_G(1-10)_specific_genes.xls". This file contains the pan-genomic profile of that user-specified group (e.g. if user define 2 groups, then Group_G1_specific_genes.xls and Group_G2_specific_genes.xls files will be created).

3. Annotation of draft/contigs genomes

(a) A spreadsheet file for each draft/contigs genome analyzed with name "Output_<*user defined name*>.xls" (e.g. if user selects the draft genomes 'draft1-draft.txt' and 'draft2-draft.txt' then 'Output_draft1-draft.xls' and 'Output_draft2-draft.xls' files will be created).

(b) Fasta-formatted sequence file for each draft/contigs genome analyzed with name

"Sequence_<*user defined name*>.txt" (e.g. if user selects the draft genomes 'draft1-draft.txt' and 'draft2-draft.txt' then 'Sequence_draft1-draft.txt' and 'Sequence_draft2-draft.txt' files will be created).

4. Identifying un-annotated genes in annotated genomes

(a) Spreadsheet file for each annotated genome analyzed with name "Unannotated_<*user defined name*>.xls". (e.g. if user selects the annotated genomes 'genome4.txt' and 'genome5.txt' then 'Unannotated_genome4.xls' and 'Unannotated_genome5.xls' files will be created).

Alerts

1. The first 10 characters of any two genome sequence files should not be same. Only use alphanumeric characters (preferably within 10 characters) for the genome names.

2. For draft/contigs genomes analysis in pan-genome profile analysis, there should be “-draft” at the end of their genome file name.

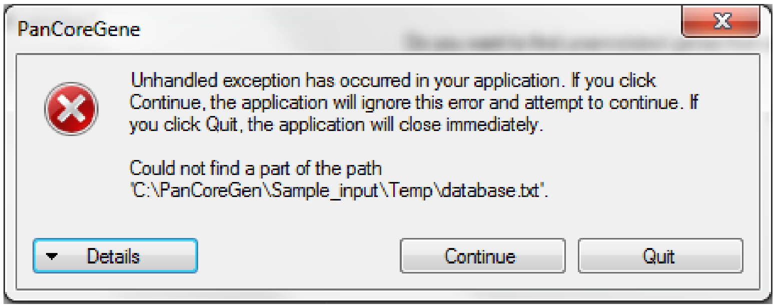
3. Before any next run, users are advised to remove all the input files of the previous run from "PanCoreGen" folder or input folder to avoid any possible redundancy in filenames.

4. In order to reduce any wrong incorporation of files in the PanCoreGen analysis and proper running of the software, users are advised to store only the genome files to be analyzed in input folder.

5. By enabling the DUST filter in BLAST option may lead to no ortholog identification for certain query genes. In that case, there will be no values in the “Representatives” and “# of seqs” field within “Gene_distribution.xls” output file for the respective gene information. Although, the genes may be present in all of the genomes analyzed.

Troubleshooting

1. Error message showing



PanCoreGene

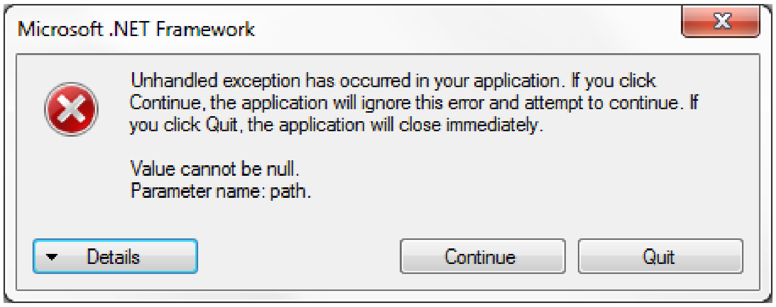
Unhandled exception has occurred in your application. If you click Continue, the application will ignore this error and attempt to continue. If you click Quit, the application will close immediately.

Could not find a part of the path 'C:\PanCoreGen\Sample_input\Temp\database.txt'.

Possible reason: No database-link file is created.

Solution: Create the database-link file as described in section 4. Create database-link file in Running PanCoreGen.

2. Error message showing



Microsoft .NET Framework

Unhandled exception has occurred in your application. If you click Continue, the application will ignore this error and attempt to continue. If you click Quit, the application will close immediately.

Value cannot be null.
Parameter name: path.

Possible reason: Option to annotate the draft genomes has been chosen but none of the genomes are selected.

Solution: Select the draft genomes to be annotated by browse option as described in section 7. Annotation of draft/contigs genomes in Running PanCoreGen.

3. No output file appears within output folder:

Possible reasons: a) Number of groups selected but no genomes assigned to any of the groups.

Solution: Assign genomes for every group selected (see section 6. Group specific analysis in Running PanCoreGen).

b) "Do you want to find unannotated genes from annotated genomes?" Option is chosen but none of the genomes are selected.

Solution: Specify genomes in order to find unannotated genes in them by selecting the genomes using the browse option (See section 8. Identifying un-annotated genes in annotated genomes in Running PanCoreGen).

4. Only one user defined group specific pan-genomic profile file created while more than one group selected:

Possible reason: Genome files are specified for only one group, no genomes selected for other groups.

Solution: Assign genomes for each group created. (See section 6. Group specific analysis in Running PanCoreGen).