BPGA: Bacterial Pan Genome Analysis Pipeline

User Manual

Developed by: **Narendrakumar M. Chaudhari[1], Vinod Kumar Gupta[1] & Chitra Dutta[*]**
Structural Biology & Bioinformatics Division, CSIR- Indian Institute of Chemical Biology, 4, Raja S. C. Mullick Road, Kolkata 700032, India

Mailing addresses: NMC: naren.niper@gmail.com
VKG: vinodgupta299@gmail.com
CD: cdutta@iicb.res.in, cduta.iicb@gmail.com

[1] NMC & VKG contributed equally to this work.
[*]Corresponding author

Keywords: Bioinformatics, Pan genome analysis, Comparative genomics

Address of the Corresponding Author:
Structural Biology & Bioinformatics Division, CSIR-Indian Institute of Chemical Biology, 4, Raja S. C. Mullick Road, Kolkata 700032, India
E-mail: cdutta@iicb.res.in, cdutta.iicb@gmail.com
Phone: 91 33 2499 5812, Fax: 91 33 2472 3967, 91 33 2473 5197

1 Description

BPGA is a perl based pipeline to exploit protein clustering data for complete pan-genome analysis of bacterial species. BPGA can process outputs of three major clustering tools (USEARCH, CD-HIT and OrthoMCL) to obtain pan-genome profiles of bacterial gene pools.

Installation

Installation of BPGA is simple.

1. Download the installer for Windows 64-bit/32-bit system from our sourceforge page:
   http://sourceforge.net/projects/bpgatool/.
2. Run the installer as Administrator. It will extract files to a folder locally. Executables are present inside *bin* folder.
3. BPGA is written in perl but bundled in an executable; hence no modules are needed to be installed.

Other requirements

4. Installation of ***gnuplot*** (4.6.6) is must for plotting graphs. You can download Windows 32-bit version from here and 64-bit version from here.
5. BPGA uses **USEARCH** as a default clustering tool. Users need to get their own licensed Windows 32-bit version freely available at:
   http://www.drive5.com/usearch/download.html, rename it to ***usearch.exe*** and copy it inside the ***bin*** folder.
   Note: For USERACH to work properly, please check the required **vcomp100.dll** system file inside **Windows\System32** folder of your computer. If not, put it in this place. It is available at: http://www.drive5.com/usearch/manual/vcomp100.html .
6. **MUSCLE** is used for alignments and tree generation. It is provided with the package.
7. ***rsvg-convert.exe*** is required to handle SVG image data. It is also provided with the BPGA package.

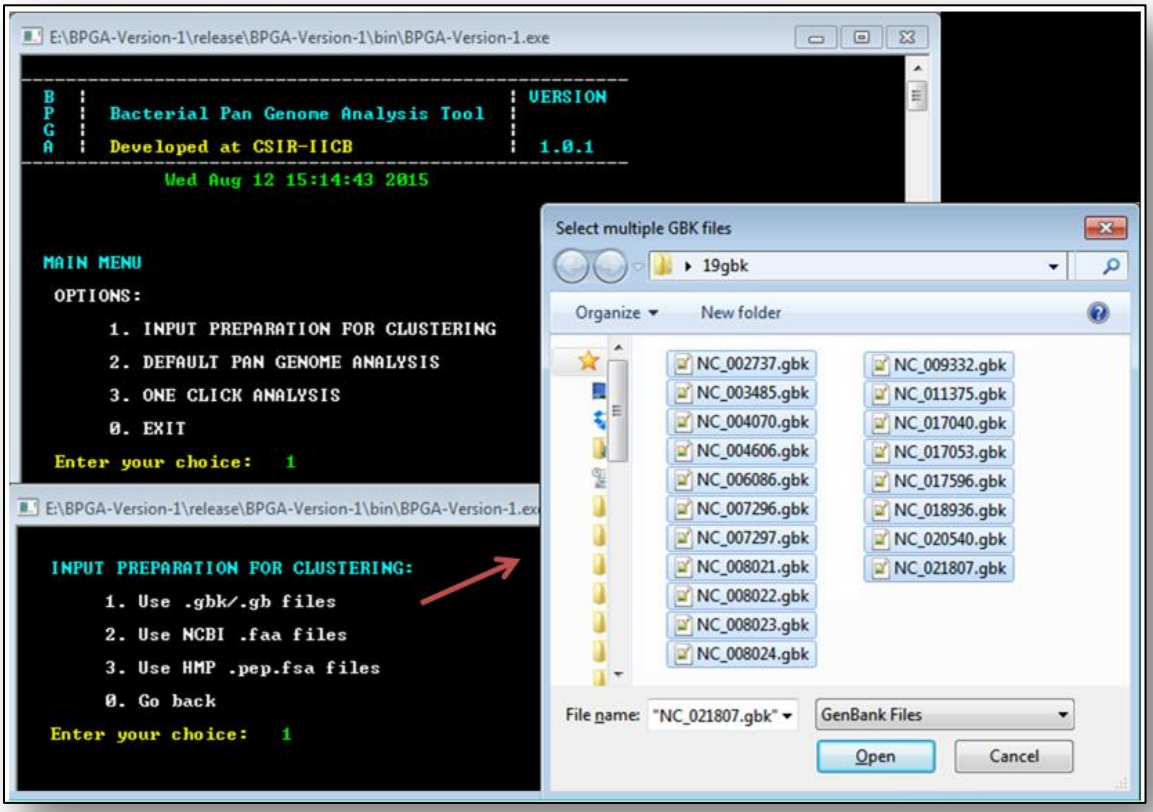
Steps for run

BPGA is easy, user friendly command line interface and it's better to run through

windows command line.

Option-1 (INPUT PREPARATION FOR CLUSTERING) allows user to prepare input for clustering using different type of sequence files. User can select multiple files from file selection dialog. Three formats are allowed (*.gbk, *.faa and *.pep.fsa). It will generate a single sequence file (*INPUT_all.faa*) that will be used for clustering. List of organisms will be written to the *list* file.

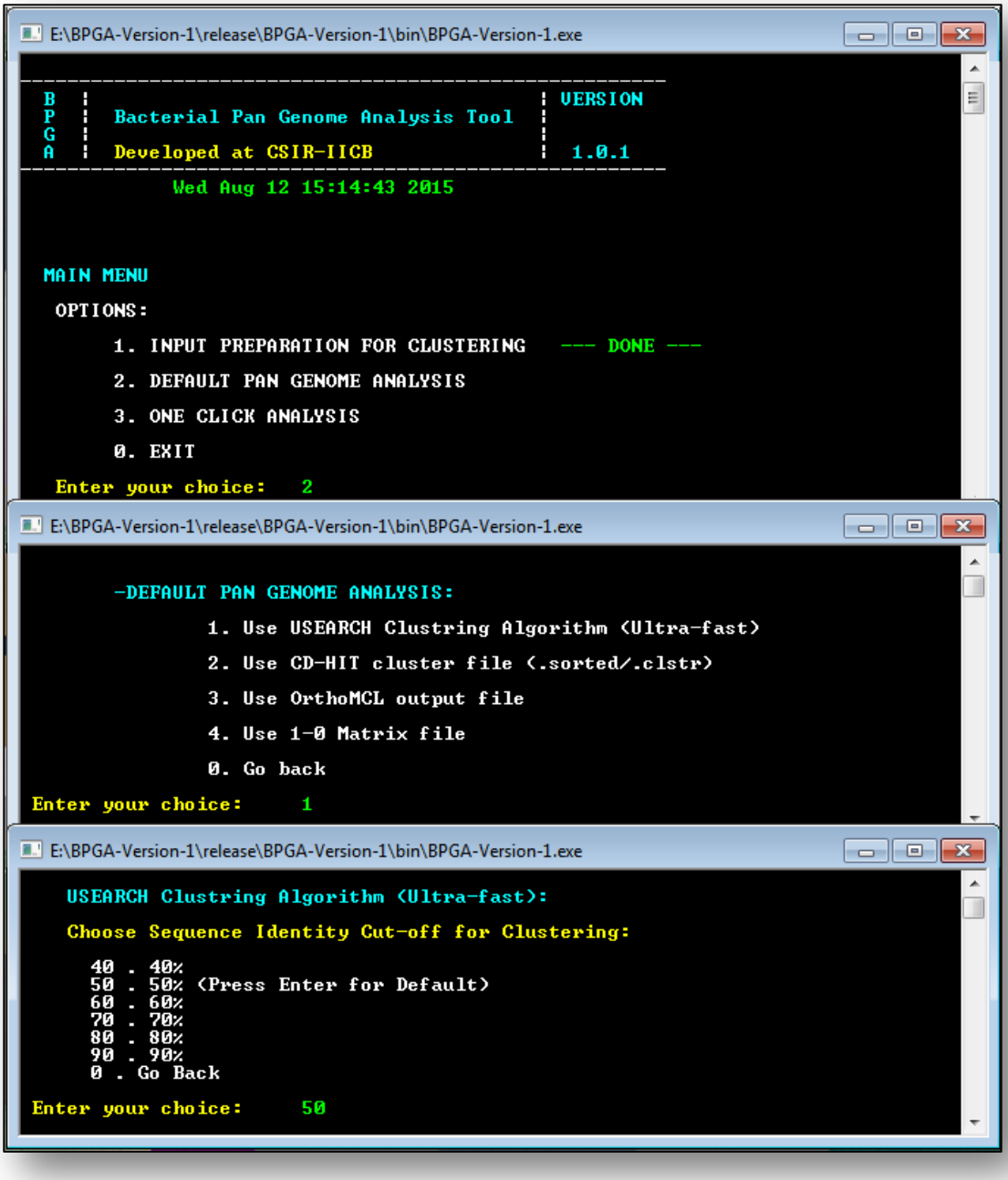
Note: BPGA treats separate files as separate organism. If there are multiple files (chromosomes) for an organism, user should concatenate all files into a single file (applicable for all three formats) for that organism.



8. *Option-2* (DEFAULT PAN GENOME ANALYSIS) allows user to perform Pan-genome analysis on the data by clustering with USEARCH or by processing pre-clustered data by CD-HIT or OrthoMCL.

Please note that, user must use *input file* (INPUT_all.faa generated by *Option-1*) for clustering with CD-HIT (online server/offline package) and with OrthoMCL pipeline with desired options. While clustering with USEARCH, identity cut off

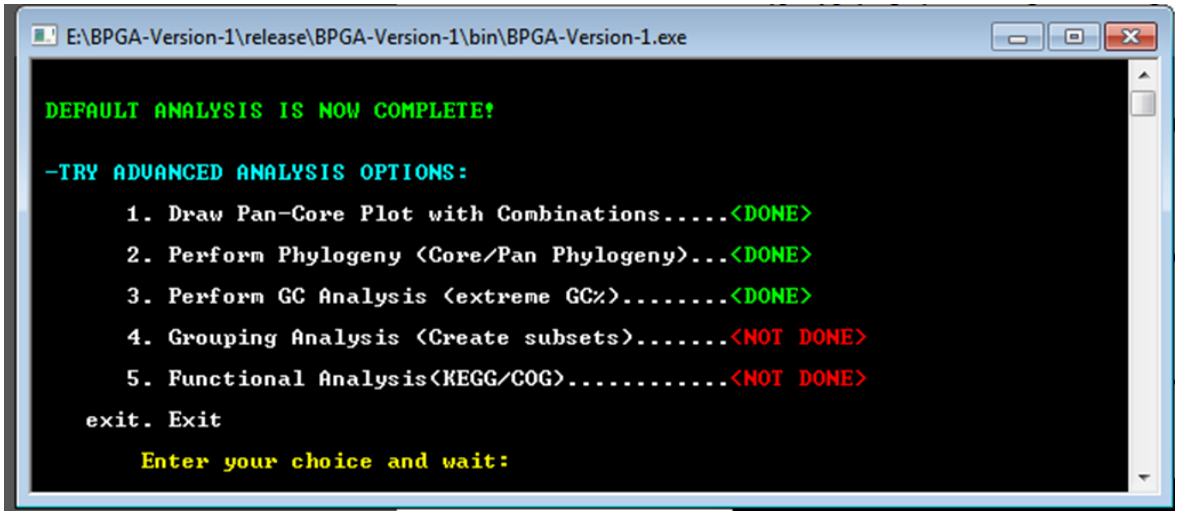
can be set by the user (see picture).



9. *Option-3* (ONE CLICK MODE) allows user to perform all the analyses in single step using all default parameters :
    - *Clustering*: USEARCH (Identity cut off = 50%)
    - *No. of combinations*: 30 for less than 20 genomes and 20 for more than 20 and less than 50 genomes.
    - *Atypical GC Content Analysis*: Extreme GC Content = 5%

- *Type of phylogeny tree*: Neighbor Joining Tree (NJ).
- *KEGG/COG Functional analysis*: will be performed if dataset contains less than 50 genomes.
- *Subset Analysis*: NA

10. In the next step, after completion of DEFAULT PAN GENOME ANALYSIS, ADVANCED ANALYSIS OPTIONS will be available.



User may perform any of the 5 analyses one by one. Completion status of each analysis will be displayed in brackets (NOT DONE or DONE).

11. After completing desired analyses, user should exit by typing '*exit*' and not closing the terminal.

Results

12. Input preparation option will give *input files* (*INPUT_all.faa*, *INPUT_all.ffn*) necessary for clustering and dataset file (*DATASET.xls*) containing organism details. A file *list*, required for further analysis is also generated.

13. The default analysis will give simple pan/core plot (*Default_Core_Pan_Plot.pdf*), distribution of gene families (*Histogram.pdf*), number of new genes added (*New_Genes_Plot.pdf*), genome wise statistics (*stats.txt*), representative sequences for core, accessory and unique gene families (*REPSEQ_*.txt*) and tab delimited pan-matrix (*matrix.txt*).

Advanced options:

1. Pan-core plot with combinations will give core and pan genome boxplot (*Core_Pan_Plot.pdf*) and dot plot (*Core_Pan_Dot_Plot.pdf*) generated using

desired number of unique combinations of genomes.

2. Phylogeny trees based on pan-matrix (*Pan_phylogeny.pdf*) and core gene/protein sequences (*Core_phylogeny.pdf*) will be generated. Respective *.ph files are provided for user to visualize using TreeView (http://taxonomy.zoology.gla.ac.uk/rod/treeview.html) and *.nwk files are also available for user to visualize using TreeGraph2 (http://treegraph.bioinfweb.info/).
3. Atypical GC analysis will give sequences of core, accessory and unique genes with atypical (extreme) GC content (**_genes_with_atypical_GC_content.txt*).
4. Subset analysis will give default results for each group in separate folder.
5. Functional analysis will give COG and KEGG distribution of the core, accessory and unique gene families based on representative sequences (*COG_DISTRIBUTION.pdf*, *COG_DISTRIBUTION_DETAILS.pdf*, *KEGG_DISTRIBUTION.pdf*, K*EGG_DISTRIBUTION_DETAILS.pdf*).

Additional instructions

For subset analysis user must create a text file having information about groups to be created. Here is the example,

	Organism id as per dataset list					
Group 1	1	2	3	4		
Group 2	6	7	8	9	13	15
Group 3	5	10	11	12	14	

Here, rows represent groups. Each number represents a genome (refer *list* file created during preparation). Blue colored labels are just for representation purpose. Actual file should contain only tab delimited values. Maximum 10 groups can be formed. There should be no repeats or wrong id.

Accepted file formats:

- **GBK**: Freshly downloaded Genbank files from NCBI or HMP databases.

```
LOCUS       NC_003485            1895017 bp    DNA     circular CON 10-JUN-2013
DEFINITION  Streptococcus pyogenes MGAS8232 chromosome, complete genome.
ACCESSION   NC_003485
VERSION     NC_003485.1  GI:19745201
DBLINK      Project: 57871
            BioProject: PRJNA57871
KEYWORDS    .
SOURCE      Streptococcus pyogenes MGAS8232
  ORGANISM  Streptococcus pyogenes MGAS8232
            Bacteria; Firmicutes; Bacilli; Lactobacillales; Streptococcaceae;
            Streptococcus.
REFERENCE   1  (bases 1 to 1895017)
  AUTHORS   Beres,S.B., Sylva,G.L., Barbian,K.D., Lei,B., Hoff,J.S.,
            Mammarella,N.D., Liu,M.Y., Smoot,J.C., Porcella,S.F., Parkins,L.D.,
            Campbell,D.S., Smith,T.M., McCormick,J.K., Leung,D.Y.,
            Schlievert,P.M. and Musser,J.M.
  TITLE     Genome sequence of a serotype M3 strain of group A Streptococcus:
            phage-encoded toxins, the high-virulence phenotype, and clone
            emergence
  JOURNAL   Proc. Natl. Acad. Sci. U.S.A. 99 (15), 10078-10083 (2002)
   PUBMED   12122206
REFERENCE   2  (bases 1 to 1895017)
  AUTHORS   Smoot,J.C., Barbian,K.D., Van Gompel,J.J., Smoot,L.M.,
            Chaussee,M.S., Sylva,G.L., Sturdevant,D.E., Ricklefs,S.M.,
            Porcella,S.F., Parkins,L.D., Beres,S.B., Campbell,D.S., Smith,T.M.,
            Zhang,Q., Kapur,V., Daly,J.A., Veasy,L.G. and Musser,J.M.
  TITLE     Genome sequence and comparative microarray analysis of serotype M18
            group A Streptococcus strains associated with acute rheumatic fever
            outbreaks
```

- **FAA**: Freshly downloaded amino acid sequence files from NCBI database.

```
>gi|19745202|ref|NP_606338.1| chromosome replication initiator DnaA [Streptococcus pyogenes MGAS8232]
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVDKHIATIYLDQMKELFWEKNLKDVILTAGFEVY
NAQISVDYVFEEDLMIEQNQTKINQKPKQQALNSLPTVTSDLNSKYSFENFIQGDENRWAVAASIAVANT
PGTTYNPLFIWGGPGLGKTHLLNAIGNSVLLENPNARIKYITAENFINEFVIHIRLDTMDELKEKFRNLD
LLLIDDIQSLAKKTLSGTQEEFFNTFNALHNNNKQIVLTSDRTPDHLNDLEDRLVTRFKWGLTVNITPPD
FETRVAILTNKIQEYNFIFPQDTIEYLAGQFDSNVRDLEGALKDISLVANFKQIDTITVDIAAEAIRARK
QDGPKMTVIPIEEIQAQVGKFYGVTVKEIKATKRTQNIVLARQVAMFLAREMTDNSLPKIGKEFGGRDHS
TVLHAYNKIKNMISQDESLRIEIETIKNKIK
>gi|19745203|ref|NP_606339.1| DNA polymerase III subunit beta [Streptococcus pyogenes MGAS8232]
MIQFSINRTLFIHALNATKRAISTKNAIPILSSIKIKVTPTGVTLTGSNGQISIENTIPVSNENAGLLIT
SPGAILLEASFFINIISSLPDISINVKEIEQHQVVLTSGKSEITLKGKDVDQYPRLQEVSTENPLILKTK
LLKSIIAETAFAASLQESRPILTGVHIVLSNHKDFKAVATDSHRMSQRLITLDNTSADFDVVIPSKSLRE
FSAVFTDDIETVEVFFSPSQILFRSEHISFYTRLLEGNYPDTDRLLMTEFETEVVFNTQSLRHAMERAFL
ISNATQNGTVKLEITQNHISAHVNSPEVGKVNEDLDIVSQSGSDLTISFNPTYLIESLKAIKSETVKIHF
LSPVRPFTLTPGDEEESFIQLITPVRTN
>gi|19745204|ref|NP_606340.1| GTP-dependent nucleic acid-binding protein EngD [Streptococcus pyogenes MGAS8232]
MALTAGIVGLPNVGKSTLFNAITKAGAEAANYPFATIDPNVGMVEVPDERLQKLTELITPKKTVPTTFEF
TDIAGIVKGASRGEGLGNKFLANIREVDAIVHVVRAFDDENVMREQGREDAFVDPIADIDTINLELILAD
LESINKRYARVEKMARTQKDKESVAEFNVLQKIKPVLEDGKSARTIEFTEEEAKVVKGLFLLTTKPVLYV
ANVDEDKVANPDGIDYVKQIRDFAATENAEVVVISARAEEEISELDDEDKEEFLEAIGLTESGVDKLTRA
AYHLLGLGTYFTAGEKEVRAWTFKRGIKAPQAAGIIHSDFERGFIRAVTMSYDDLMIYGSEKAVKEAGRL
REEGKEYVVQDGDIMEFRFNV
```

- **PEP.FSA**: Freshly downloaded amino acid sequence files from HMPDACC database.

```
>HMPREF1231_0571 ABC transporter, substrate-binding protein, family 5 [Streptococcus pyogenes GA06023]
MSKYLKYFSIITLFLTGLILVACQQQKPQTKERQRKQRPKDELVVSMGAKLPHEFDPKDR
YGVHNEGNITHSTLLKRSPELDIKGELAKTYHLSEDGLTWSFDLHDDFKFSNGEPVTADD
VKFTYDMLKADGKAWDLTFIKNVEVVGKNQVNIHLTEAHSTFTAQLTEIPIVPKKHYNDK
YKSNPIGSGPYMVKEYKAGEQAIFVRNPYWHGKKPYFKKWTWVLLDENTALAALESGDVD
MIYATPELADKKVKGTRLLDIPSNDVRGLSLPYVKKGVITDSPDGYPVGNDVTSDPAIRK
ALTIGLNRQKVLDTVLNGYGKPAYSIIDKTPFWNPKTAIKDNKVAKAKQLLTKAGWKEQA
DGSRKKGDLDAAFDLYYPTNDQLRANLAVEVAEQAKALGITIKLKASNWDEMATKSHDSA
LLYAGGRHHAQQFYESHHPSLAGKGWTNITFYNNPTVTKYLDKAMTSSDLDKANEYWKLA
QWDGKTGASTLGDLPNVWLVSLNHTYIGDKRINVGKQGVHSHGHDWSLLTNIAEWTWDES
TK
>HMPREF1231_1349 polysaccharide biosynthesis C-terminal domain protein [Streptococcus pyogenes GA06023]
MVASMGVLGYYLWKQGLLAAIFSKPDHTVSIDIKGLLLETLKESIPFIVTGSAIQAFQLI
DQWTFVNTMTLFTDSSRSQLLVLFGYFNANPAKITMVLIAVAASIGGVGIALLTENYVKK
DMKAAARLIINNIEMLVMFLLPALTGAIILARPLYSVFYGASEDRAIHLFVAVLFQTLLL
ALYTLFSPMLQALFENRKAIYYFAYGILIKLVLQIPLIYLLHAYGPLLATTIALVVPIYL
MYRRLYQVTHFNRKLLQKRLLLTLIETLLMGLVVFVANWLLGYAFKPTGRLTSLLYLLII
GGLGMTVYTALTLLTHQLDKLIGSKASRLRQKLGWH
>HMPREF1231_1350 uracil phosphoribosyltransferase [Streptococcus pyogenes GA06023]
MGKCQVISHPLIQHKLSILRRQTTSTKDFRELVNEIAMLMGYEVSRDLPLEDVDIQTPVS
KTVQKQLAGKKLAIVPILRAGIGMVDGLLSLVPAAKVGHIGMYRNEETLEPVEYLVKLPE
DINQRQIFLVDPMLATGGSAILAVDSLKKRGAANIKFVCLVAAPEGVKKLQEAHPDIDIF
TAALDDHLNEHGYIVPGLGDAGDRLFGTK
```