

BPGA软件分享（windows版）



➤ 软件介绍 (<https://iicb.res.in/bpga/index.html>)

BPGA – a *Bacterial Pan Genome Analysis pipeline*

"Pan-Genome" refers to the complete inventory of genes in a specific phylogenetic clade. Pan-genomic analyses have provided valuable insight into genome dynamics, population structure, species evolution, niche specialization, pathogenesis, drug resistance and many other features of the microbial world.

[Downloads](#)
[FAQs](#)
[Contact](#)

400 Pageviews
Apr. 28th - May. 28th



Developed at:

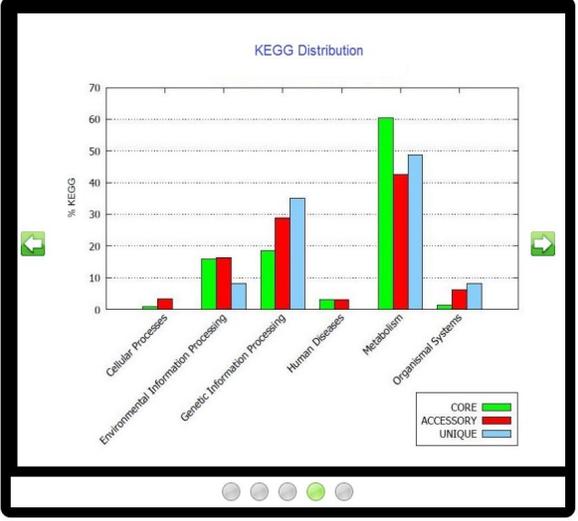


Indian Institute of Chemical Biology
a unit of C.S.I.R.

BPGA is an ultra-fast software package that provides comprehensive pan genome analysis of microorganisms. In addition to all types of routine pan genomic analyses (Pan genome Profiles, Pan/Core Phylogeny etc.), BPGA includes a number of novel downstream analysis features like Exclusive Gene Family Analysis, Atypical GC Content Analysis, Subset Analysis, MLST based on housekeeping genes and KEGG Distribution etc.

Another unique feature of BPGA is that it allows the user to select from three different tools for ortholog clustering – USEARCH, CD-HIT & OrthoMCL, the first one being the default clustering tool.

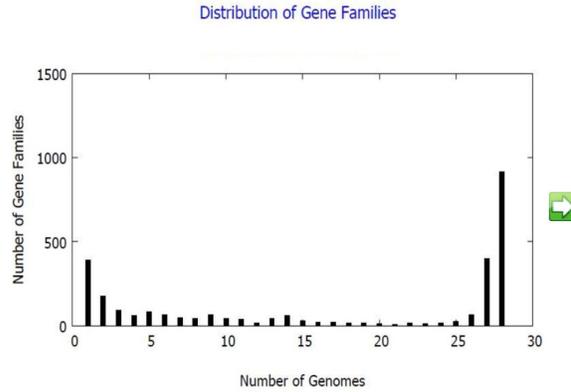
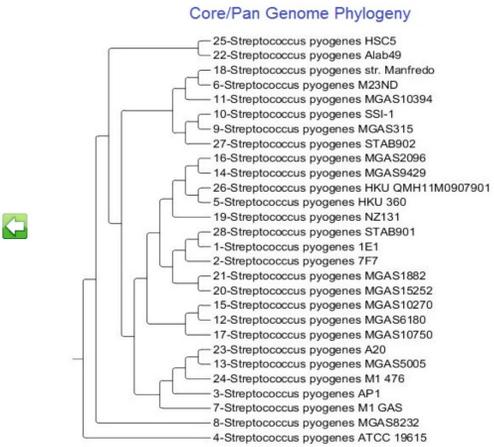
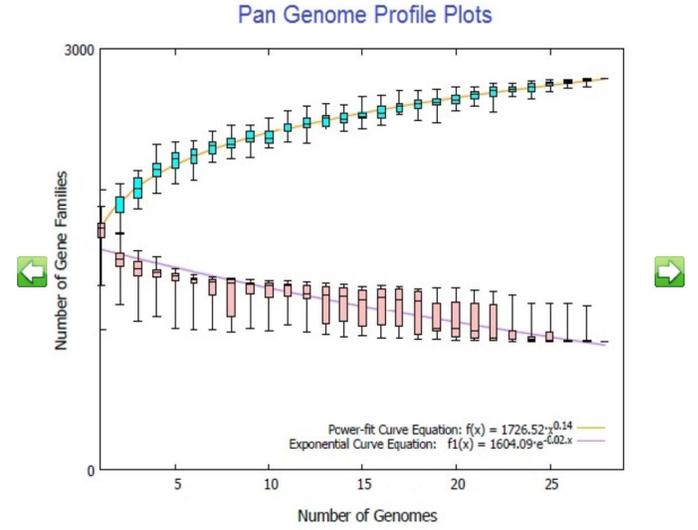
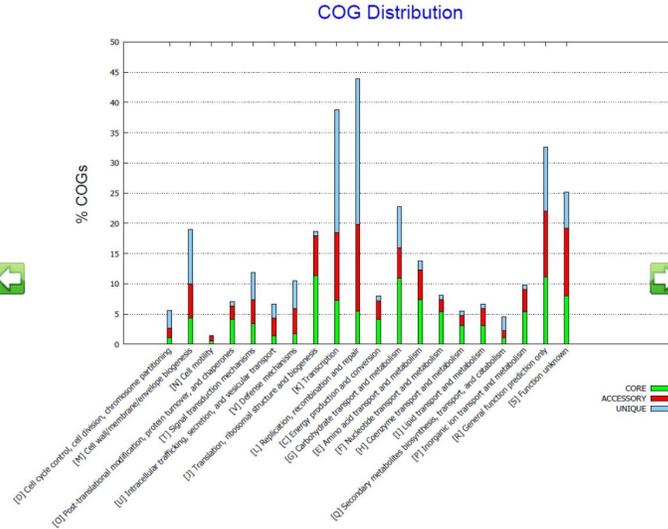
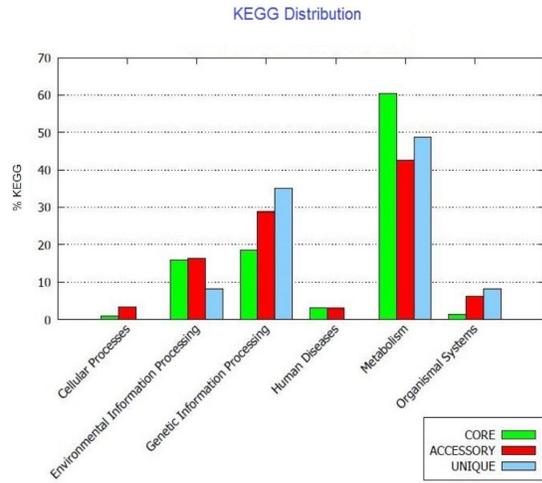
Highlights of the BPGA Analyses:



Category	CORE (%)	ACCESSORY (%)	UNIQUE (%)
Cellular Processes	~2	~5	~1
Environmental Information Processing	~18	~15	~10
Genetic Information Processing	~20	~30	~35
Human Diseases	~5	~5	~1
Metabolism	~60	~45	~50
Organismal Systems	~2	~8	~10

BPGA 是一种超快速的软件包，为微生物提供全面的泛基因组分析。除了所有类型的常规泛基因组分析（如泛基因组图谱、泛/核心系统发育树等）外，BPGA 还包含许多新颖的下游分析功能，如特有基因家族分析、非典型 GC 含量分析、子集分析、基于管家基因的 MLST 以及 KEGG 分布等。

软件介绍



➤ 软件下载

Home / Browse Open Source / BPGA / Files



BPGA Files

A tool for ultra-fast pan-genome analysis of microbes.
Brought to you by: [encoderman](#), [guptabpga](#)

Summary
Files
Reviews
Support
Documentation
Tickets

Download Latest Version
BPGA-1.3-mswin-x86-0-0-0.zip (81.9 MB)

Get Updates

Home

Name ▾	Modified ▾	Size ▾	Downloads / Week ▾
BPGA-1.3-mswin-x86-0-0-0.zip	2017-05-13	81.9 MB	30
BPGA-1.3-mswin-x64-0-0-0.zip	2017-05-13	83.0 MB	11
BPGA-1.3-linux-x86_64-0-0-0.tar.gz	2017-05-13	58.1 MB	17
ReadMe	2017-05-12	6.4 kB	5
example.zip	2016-04-16	16.2 MB	0
BPGA User Manual.pdf	2015-12-07	565.6 kB	0
Totals: 6 Items		239.7 MB	63

- 近三个月下载次数最多的是windows-x86版本，共下载次数460次
- 交互式分析界面操作方式，windows版本依赖关系更简单，操作更便捷

➤ 软件下载

PREREQUISITES

- Windows Requirements:**
 System: Windows XP or latter
 Usearch: ** Get it from <http://www.drive5.com/usearch/>. Download and rename the Windows executables to "usearch.exe" **[case sensitive, also mind that Windows file extensions are visible]**
 gnuplot: [Download Gnuplot_Win-64bit_Version](#) or [Download_Gnuplot_Win-32bit_Version](#)
 WARNING (for Windows only) :** Please check "vcomp100.dll" system file in the system32 or system64 folder, path of this folder is as "C:\Windows\System32". If not present copy it into above said folder. This file is required for USEARCH to function properly. This dll is available [at this link](#)
- Linux Requirements:**
 Usearch : Get it from <http://www.drive5.com/usearch/>. Download and rename the Linux executable to "usearch". **[case sensitive]**
 gnuplot: Linux users need to download exact version of gnuplot for linux from this [SourceForge Page](#)
 Extract gnuplot 4.6.6. files by `tar -xzf FILENAME.tar.gz`
 cd to gnuplot 4.6.6 directory simply run:

```
sudo ./configure ,
sudo make ,
sudo make install
```

 to install gnuplot manually.
 ghostscript: run `sudo apt-get install ghostscript`
 wine (Ubuntu): `sudo add-apt-repository ppa:ubuntu-wine/ppa -y && sudo apt-get update && sudo apt-get install wine`
WARNING (for Debian only) : Make sure you have 'glibc 2.15' or higher (the C library in Debian).If not, you can carefully install higher version of glibc in parallel to 'glibc 2.14' or less. **[Do not try to remove original glibc, as other binaries may not work properly.]** As the post on one of the debian forum says: "One can install NEW glibc in parallel to OLD in some different place, e.g., in /opt. In fact, this is a common technique to make Google Chrome work on CentOS. I'm not saying this is as easy as 1-2-3, but it is certainly both doable and, if done properly, safe."
 Alternatively, the steps are discussed [here](#) for running new applications on old glibc.
For BPGA Version 1 (Linux) only: Make sure that libgif4 (library) and libtiff4 are installed. Install 'libtiff4' and 'libgif4' from Ununtu Software Center or install manually as follows. Ubuntu 14.04 LTS 64 bit release includes 'libtiff5' library, BPGA may not start execution with it. (BPGA Version 1 currently needs 'libtiff4' and 'libgif4'). You may also get **libtiff4** for different Ubuntu Releases from [this link](#) and **libgif4** from [this link](#). Install them manually using following commands (type exact file name that you downloaded) `sudo dpkg -i ./libgif4_version_details_32_or_64_bit.deb` and `sudo dpkg -i ./libtiff4_version_details_32_or_64_bit.deb`

● 依赖软件:

① Usearch(32bit for free)

② Gnuplot(命令行驱动的交互式绘图工具)

➤ 数据格式

• NCBI GenBank File (.gbk/.gbff)

1 .gbk/.gbff文件

```

LOCUS       NC_017040             1750832 bp    DNA             circular CON 06-JUL-2013
DEFINITION  Streptococcus pyogenes MGAS15252 chromosome, complete genome.
ACCESSION   NC_017040
VERSION    NC_017040.1  GI:383479207
...
FEATURES             Location/Qualifiers
     source           1..1750832
                     /organism="Streptococcus pyogenes MGAS15252"
...
     gene            232..1587
                     /gene="dnaA"
...
     CDS             232..1587
                     /gene="dnaA"
...
                     /product="chromosomal replication initiator protein DnaA"
                     /protein_id="YP_005388102.1"
                     /translation="MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVKHIATI
...
  
```

• NCBI Protein FASTA files sample: (.faa)

2 .faa文件

```

>gi|19745202|ref|NP_606338.1| protein name [Organism Name]
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVD
MRTNFKVVSFYLRSNYENKEGKSPVMLRVFLNGEMSNFG
  
```

(Note that new NCBI faa files may not have the above format, they may match the following .pep.fsa format. In that case, user needs to use the pep.fsa option while using BPGA and rename the files accordingly before proceeding.)

• HMP Protein FASTA files sample:(.pep.fsa)

3 .pep.fsa文件

```

>HMPREF9420_0006 protein name [Organism Name]
MRTNFKVVSFYLRSNYENKEGKSPVMLRVFLNGEMSNFG
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVD
  
```

• Any Protein FASTA files sample:(.fasta)

4 .fasta文件

```

>Any_header_information
MRTNFKVVSFYLRSNYENKEGKSPVMLRVFLNGEMSNFG
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVD
  
```

- 可提供任意四种数据格式中的一种，且每个样本的数据格式需一致

➤ 软件分析(基础分析)

管理 C:\Users\Administrator\Desktop\BPGA-1.3-mswin-x64-0-0-0\BPGA-Version-1.3\bin

查看 应用程序工具

GA-1.3-mswin-x64-0-0-0 > BPGA-Version-1.3 > bin

名称	修改日期	类型	大小
cpfworkrt	2017/5/13 14:28	文件夹	
Results	2024/5/28 16:59	文件夹	
Sequences	2024/5/28 16:59	文件夹	
Supporting_files	2024/5/28 16:59	文件夹	
BPGA-Version-1.3	2017/5/13 14:16	应用程序	231 KB
cog.udb	2024/5/27 17:41	UDB 文件	257,150 KB
kegg.udb	2024/5/27 17:41	UDB 文件	708,793 KB
kegg_ref	2017/5/13 14:16	文件	3,406 KB
muscle	2017/5/13 14:16	应用程序	346 KB
rsvg-convert	2017/5/13 14:16	应用程序	819 KB
usearch	2020/3/30 22:58	应用程序	1,511 KB

✓ step1: 运行主程序, 可以选择3
一次完成基础分析, 也可以分步
骤分析

```

-----
B  | Bacterial Pan Genome Analysis Tool | VERSION
P  |                                     |
G  | Developed at CSIR-IICB             | 1.3.0
A  |                                     |
-----
Tue May 28 17:12:58 2024

MAIN MENU
OPTIONS:
  1. INPUT PREPARATION FOR CLUSTERING
  2. DEFAULT PAN GENOME ANALYSIS      基础分析
  3. ONE CLICK ANALYSIS (1 + 2)
  0. EXIT

Enter your choice: _
  
```

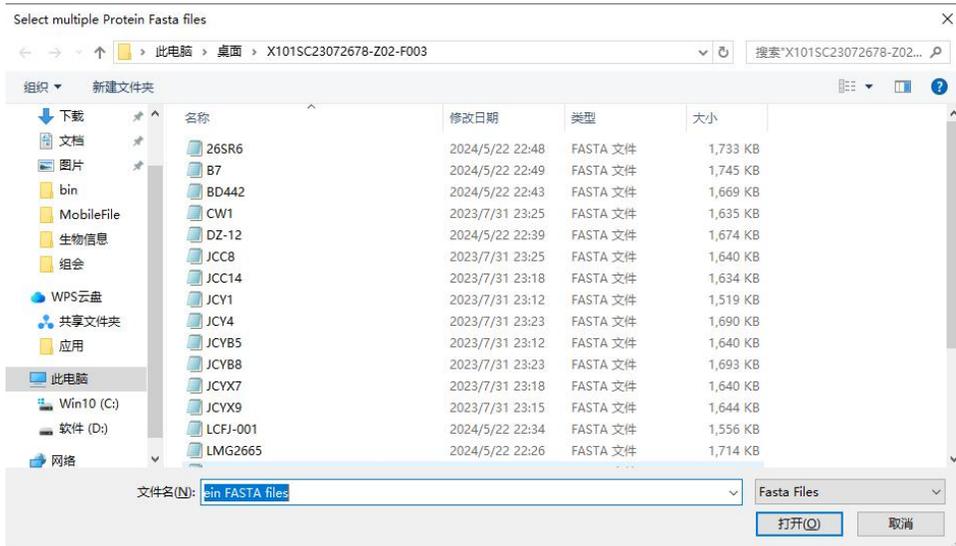
➤ 软件分析（基础分析）

INPUT PREPARATION FOR CLUSTERING:

1. Use .gbk/.gb files
2. Use NCBI .faa files
3. Use HMP .pep.fsa files
4. Use any Protein Fasta files
0. Go back

Enter your choice:

✓ step2: 选择输入文件的格式，并输入数据



Processing Fasta files...Please wait!

>> 22 files selected.

>> 22 files used.

Preparation is done!

Press Enter to go back..

➤ 软件分析（基础分析）

```
B
P
G
A | Bacterial Pan Genome Analysis Tool | VERSION
   | Developed at CSIR-IICB             | 1.3.0
-----
Tue May 28 17:20:27 2024

MAIN MENU
OPTIONS:
  1. INPUT PREPARATION FOR CLUSTERING ---DONE---
  2. DEFAULT PAN GENOME ANALYSIS
  3. ONE CLICK ANALYSIS (1 + 2)
  0. EXIT
Enter your choice: 2
```

✓ step3: PAN基因分析

-DEFAULT PAN GENOME ANALYSIS:

1. Use USEARCH Clustering Algorithm (Ultra-fast) <Press Enter>
 2. Use CD-HIT cluster file (.sorted/.clstr)
 3. Use OrthoMCL output file
 4. Use 1-0 Matrix file (Tab Seperated)
 0. Go back
- Enter your choice:

USEARCH Clustering Algorithm (Ultra-fast):

Choose Sequence Identity Cut-off for Clustering:
Identity value should be in fraction ranging from 0 to 1.
0.5 (50%) is the default.

Enter your choice: 0.5

➤ 软件分析（基础分析）

BPGA-1.3-mswin-x64-0-0-0 > BPGA-Version-1.3 > bin				
名称	修改日期	类型	大小	
cpfworkrt	2017/5/13 14:28	文件夹		
Results	2024/5/28 17:28	文件夹		
Sequences	2024/5/28 17:20	文件夹		
Supporting_files	2024/5/28 17:28	文件夹		
3.ctrl_G2	2024/5/28 17:28	文本文档	2,141 KB	
3.ctrl_G2_seq	2024/5/28 17:28	文本文档	30,977 KB	
4.cluster_with_genome	2024/5/28 17:28	文本文档	792 KB	
accessory_seq	2024/5/28 17:28	文本文档	7,440 KB	
All_transposed	2024/5/28 17:28	文本文档	168 KB	
BPGA-Version-1.3	2017/5/13 14:16	应用程序	231 KB	
cluster_gi_name	2024/5/28 17:28	文本文档	265 KB	
cluster_gi_ref	2024/5/28 17:28	文本文档	155 KB	
cog.udb	2024/5/27 17:41	UDB 文件	257,150 KB	
core	2024/5/28 17:28	文本文档	18 KB	
core_seq	2024/5/28 17:28	文本文档	23,888 KB	
DATASET	2024/5/28 17:20	XLS 工作表	1 KB	
exclusively_absent_seq	2024/5/28 17:28	文本文档	1,909 KB	
fit	2024/5/28 17:28	文本文档	3 KB	
gi_name	2024/5/28 17:20	文件	2,922 KB	
INPUT_all	2024/5/28 17:20	FAA 文件	34,291 KB	
INPUT_all.ffn	2024/5/28 17:15	FFN 文件	0 KB	
INPUT_all_original	2024/5/28 17:20	FASTA 文件	8,481 KB	
kegg.udb	2024/5/27 17:41	UDB 文件	708,793 KB	
kegg_ref	2017/5/13 14:16	文件	3,406 KB	
list	2024/5/28 17:20	文件	1 KB	
matrix	2024/5/28 17:28	文本文档	343 KB	
mlst_core	2024/5/28 17:28	文本文档	65 KB	
muscle	2017/5/13 14:16	应用程序	346 KB	
pan	2024/5/28 17:28	FASTA 文件	168 KB	
pre_ref	2024/5/28 17:28	文本文档	33,138 KB	
ref	2024/5/28 17:28	文本文档	31,716 KB	
REPSEQ_ACCESSORY	2024/5/28 17:28	文本文档	817 KB	
REPSEQ_CORE	2024/5/28 17:28	文本文档	1,086 KB	
REPSEQ_UNIQUE	2024/5/28 17:28	文本文档	432 KB	
rsvg-convert	2017/5/13 14:16	应用程序	819 KB	
size	2024/5/28 17:28	文本文档	53 KB	
sum	2024/5/28 17:28	文本文档	27 KB	
unique_seq	2024/5/28 17:28	文本文档	436 KB	
usearch	2020/3/30 22:58	应用程序	1,511 KB	

➤ 软件分析（高级分析）

```
DEFAULT ANALYSIS IS NOW COMPLETE!
```

```
-TRY ADVANCED ANALYSIS OPTIONS:
```

1. Draw Pan-Core Plot with Combinations.... <NOT DONE>
2. Perform Phylogeny (Core/Pan Phylogeny)... <NOT DONE>
3. Atypical GC Content Analysis <NOT DONE>
4. Subset Analysis (Create subgroups)..... <NOT DONE>
5. Functional Analysis(KEGG/COG)..... <NOT DONE>
0. Exit

```
Enter your choice and wait: _
```

```
Advanced Analysis Options:
```

```
Pan-Core Plot with Combinations
```

```
Set Number of Combinations:
```

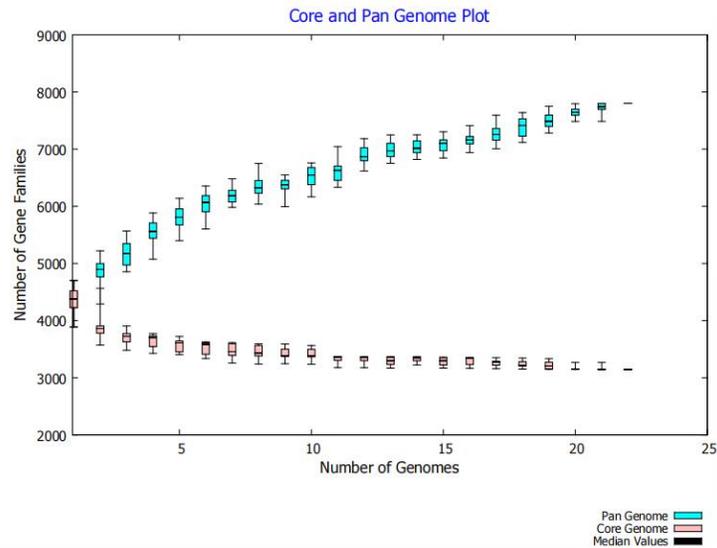
```
Equal to Number of Genomes = 22
20 .   Twenty (Press Enter for Default)
30 .   Thirty
50 .   Fifty
100 .  Time consuming
200 .  Time consuming
500 .  Time consuming
0 .    Go back
```

```
Enter 'exit' to quit.
```

```
>Note: Be patient. Plotting may take time.
```

```
Enter your choice and please wait: 20
```

- ✓ 高级分析1：根据基因组的数目选择组合数来进行 Core-pan 模型预测，少于 20 个基因组选择 30，20 - 50 个基因组选择 20。



➤ 软件分析（高级分析）

```

Select type of Pan-Phylogeny Tree:
1. Neighbour Joining Tree (NJ). <press enter for default>
2. UPGMA Tree.
0. Go Back

Enter Your Choice: 1
  
```

Pan Phylogeny is over!

For MLST based core phylogeny, refer tutorial.

Press enter to continue.

.....Aligning cluster number 001.

MUSCLE v3.8.31 by Robert C. Edgar

<http://www.drive5.com/muscle>

This software is donated to the public domain.

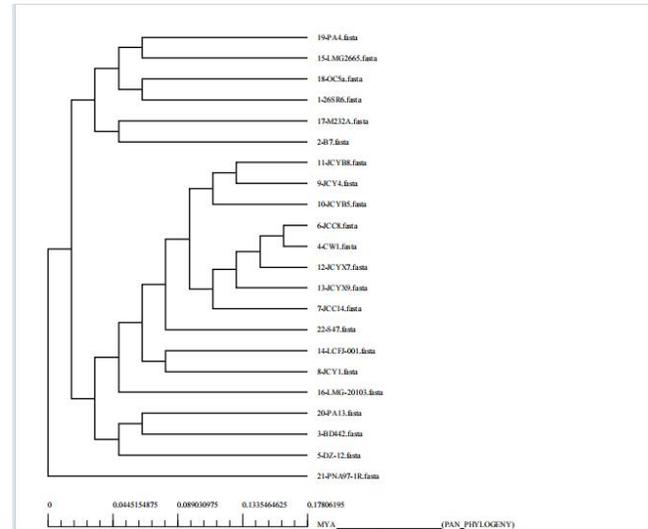
Please cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

seq_001 22 seqs, max length 4492, avg length 4304

00:00:00 5 MB(0%) Iter 1 100.00% K-bit distance matrix

00:00:00 14 MB(1%) Iter 1 4.76% Align node

✓ 进化树分析



➤ 软件分析（高级分析）

```
>>> Searching COG Hits.... Be patient! This may take some time.
usearch v11.0.667_win32, 2.0Gb RAM (17.1Gb total), 8 cores
(C) Copyright 2013-18 Robert C. Edgar, all rights reserved.
https://drive5.com/usearch
```

```
License: personal use only
```

```
00:01 221Mb 100.0% Reading rows
00:01 221Mb Reading pointers...done.
00:01 224Mb Reading db seqs...done.
00:07 417Mb 25.4% Searching, 98.1% matched
```

✓ 功能分析（内置KEGG/COG数据库）

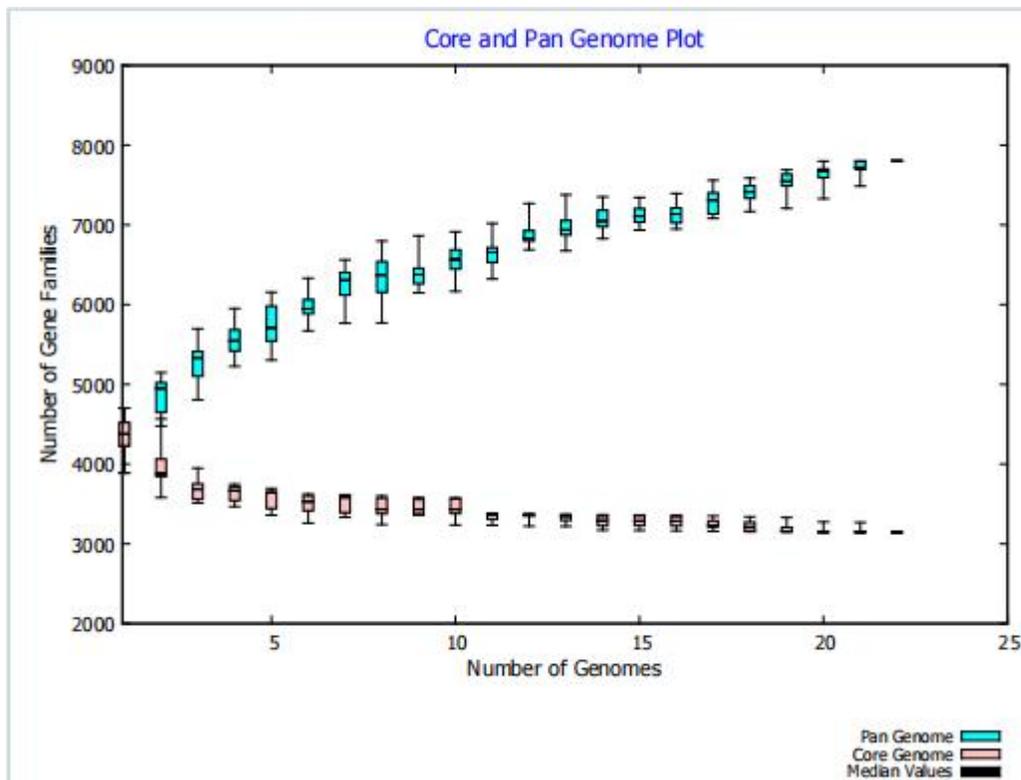
➤ 软件分析（结果目录）

3A-1.3-mswin-x64-0-0-0 > BPGA-Version-1.3 > bin

名称	修改日期	类型	大小
cpfworkrt	2017/5/13 14:28	文件夹	
Results	2024/5/29 15:35	文件夹	
Sequences	2024/5/29 15:31	文件夹	
Supporting_files	2024/5/29 15:31	文件夹	
BPGA-Version-1.3	2017/5/13 14:16	应用程序	231 KB
cog.udb	2024/5/27 17:41	UDB 文件	257,150 KB
cog_plots	2024/5/29 15:29	gnuplot comma...	2 KB
core_concat	2024/5/29 15:22	FASTA 文件	635 KB
fit	2024/5/29 15:22	文本文档	9 KB
gi_name	2024/5/29 15:14	文件	2,922 KB
group	2024/5/28 18:01	文件	1 KB
INPUT_all	2024/5/29 15:14	FAA 文件	34,291 KB
INPUT_all.ffn	2024/5/29 15:13	FFN 文件	0 KB
INPUT_all_original	2024/5/29 15:14	FASTA 文件	8,481 KB
kegg.udb	2024/5/27 17:41	UDB 文件	708,793 KB
kegg_plots	2024/5/29 15:30	gnuplot comma...	3 KB
kegg_ref	2017/5/13 14:16	文件	3,406 KB
list	2024/5/29 15:14	文件	1 KB
matrix	2024/5/29 15:15	文本文档	343 KB
muscle	2017/5/13 14:16	应用程序	346 KB
pan	2024/5/29 15:15	FASTA 文件	168 KB
plots_dot	2024/5/29 15:22	gnuplot comma...	2 KB
rsvg-convert	2017/5/13 14:16	应用程序	819 KB
usearch	2020/3/30 22:58	应用程序	1,511 KB

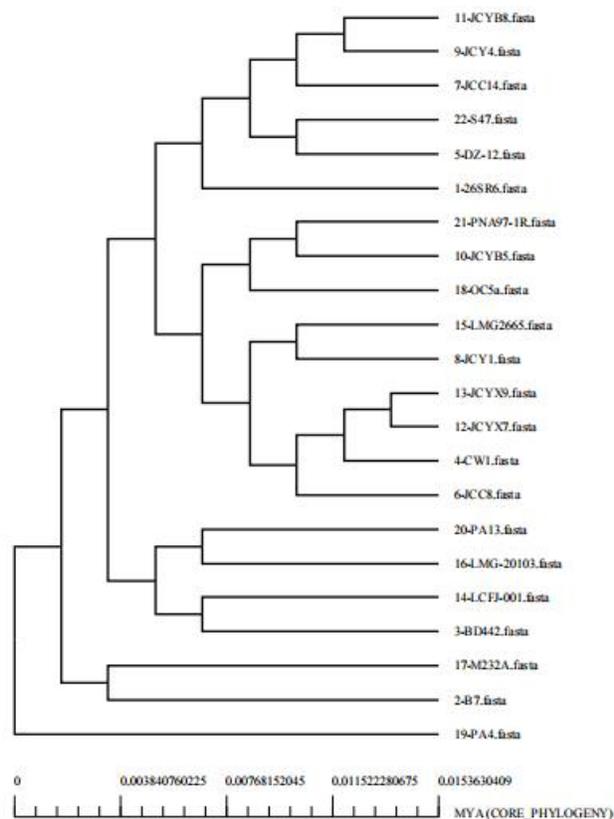
- ① Results文件夹中是分析的结果图片
- ② Sequences中是core, unique等基因的序列文件
- ③ Supporting_files是分析过程文件

➤ 软件分析（结果目录）



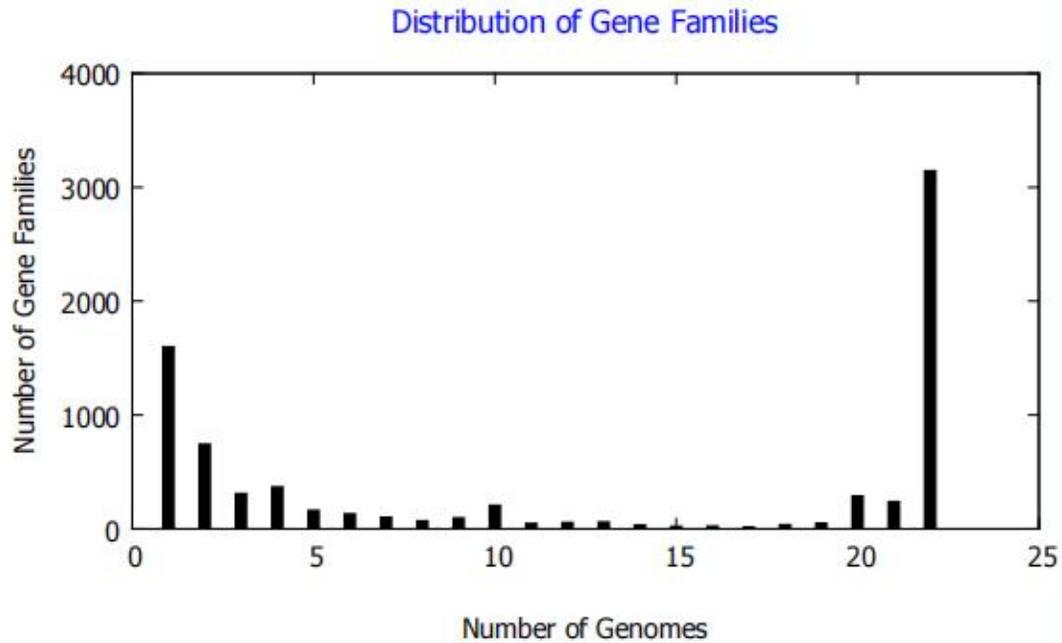
→ corepan基因箱型图，类似meta的corepan图

➤ 软件分析（结果目录）



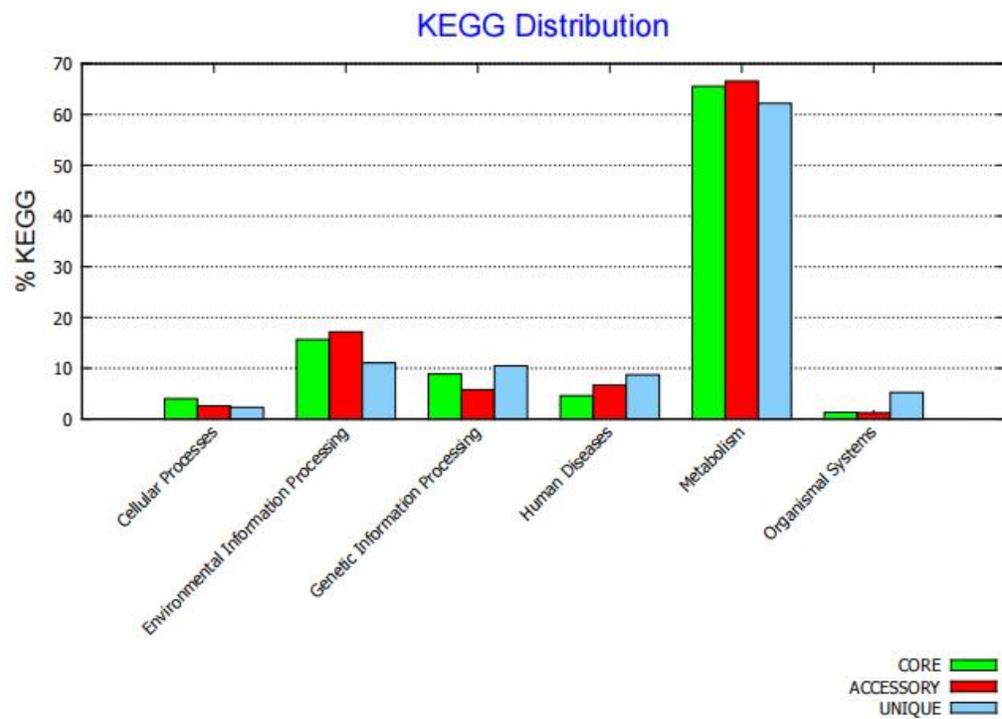
→ core基因进化树图

➤ 软件分析（结果目录）



→ 各个样本中基因数目统计图

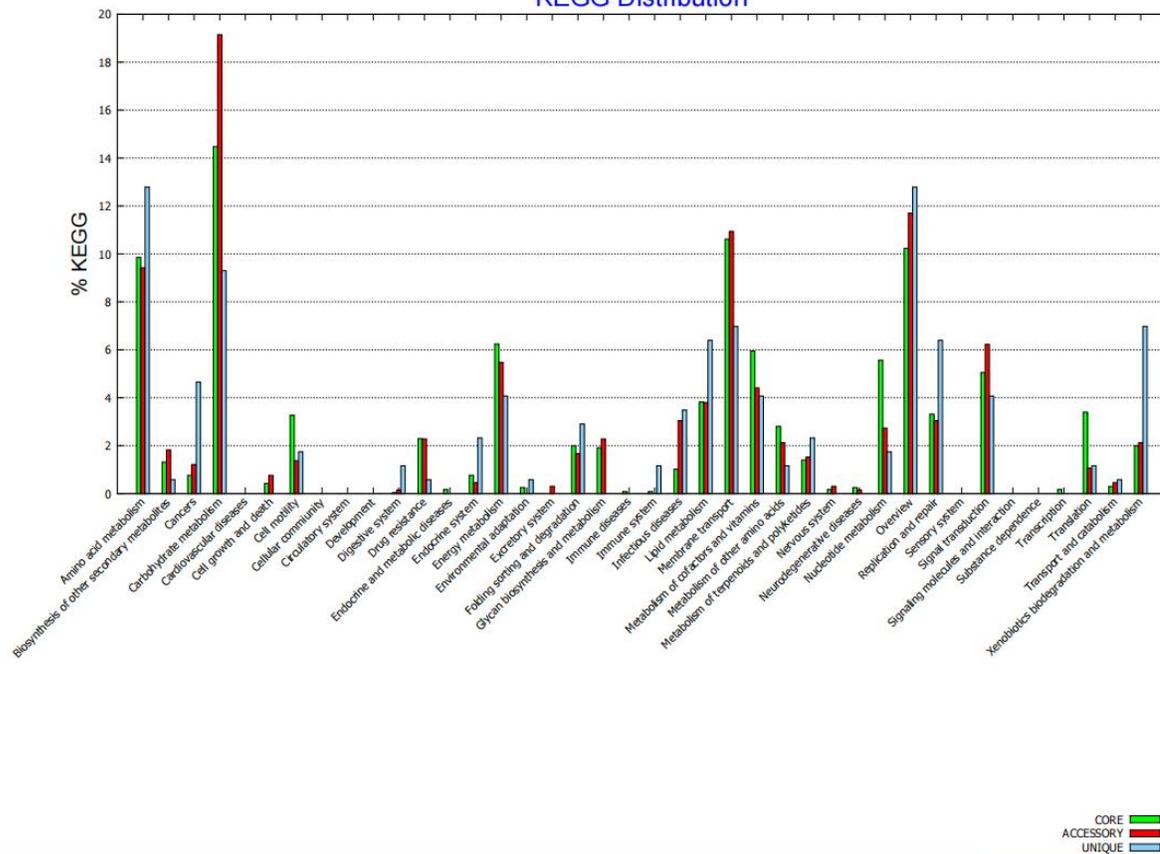
➤ 软件分析（结果目录）



→ 各类基因在KEGG数据库level1中的功能分布情况

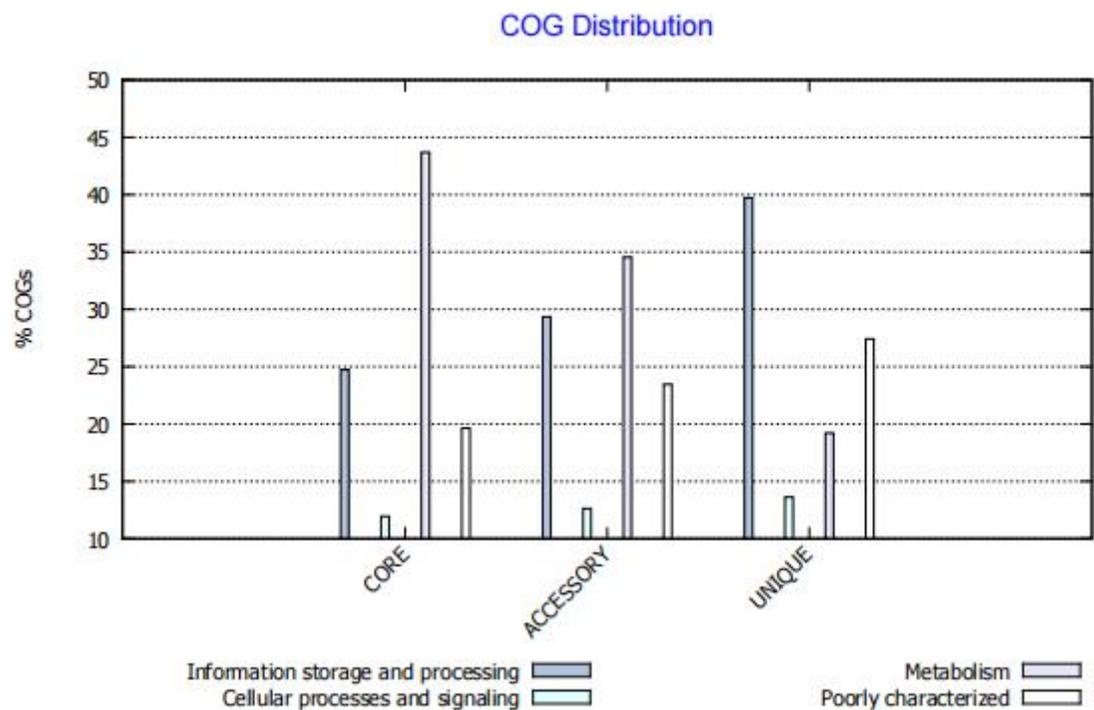
➤ 软件分析（结果目录）

KEGG Distribution



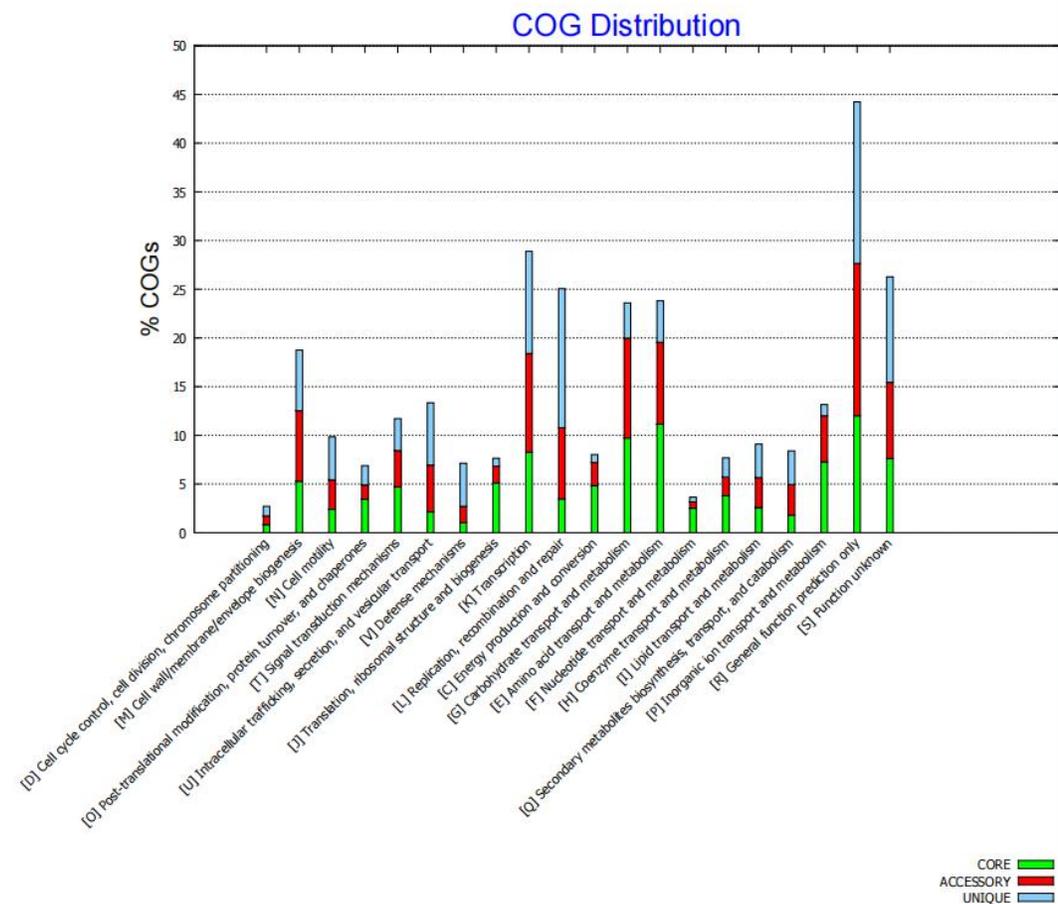
→ 各类基因在KEGG数据库level2中的功能分布情况

➤ 软件分析（结果目录）



→ 各类基因在COG数据库level1中的功能分布情况

➤ 软件分析（结果目录）



→ 各类基因在COG数据库level2中的功能分布情况

➤ 软件分析（结果目录）

IA-1.3-mswin-x64-0-0-0 > BPGA-Version-1.3 > bin > Sequences

名称	修改日期	类型	大小
 accessory_seq	2024/5/29 15:15	文本文档	7,440 KB
 core_seq	2024/5/29 15:15	文本文档	23,888 KB
 exclusively_absent_seq	2024/5/29 15:15	文本文档	1,909 KB
 REPSEQ_ACCESSORY	2024/5/29 15:15	文本文档	817 KB
 REPSEQ_CORE	2024/5/29 15:15	文本文档	1,086 KB
 REPSEQ_UNIQUE	2024/5/29 15:15	文本文档	432 KB
 unique_seq	2024/5/29 15:15	文本文档	436 KB

➤ 软件分析（结果目录）

GA-1.3-mswin-x64-0-0-0 > BPGA-Version-1.3 > bin > Supporting_files

名称	修改日期	类型	大小
 ACCESSORY_COG_hits3	2024/5/29 15:29	文本文档	72 KB
 ACCESSORY_kegg_hits3	2024/5/29 15:30	文本文档	79 KB
 Cog_Category1	2024/5/29 15:30	文本文档	2 KB
 CORE_COG_hits3	2024/5/29 15:29	文本文档	151 KB
 core_default	2024/5/29 15:15	文本文档	1 KB
 core_genome	2024/5/29 15:21	文本文档	5 KB
 CORE_kegg_hits3	2024/5/29 15:30	文本文档	204 KB
 CORE_PHYLOGENY.ph	2024/5/29 15:22	PH 文件	1 KB
 CORE_PHYLOGENY_MOD.ph	2024/5/29 15:22	PH 文件	1 KB
 DATASET	2024/5/29 15:14	XLS 工作表	1 KB
 histogram	2024/5/29 15:15	文本文档	1 KB
 kegg_accessory_id	2024/5/29 15:30	文本文档	7 KB
 kegg_accessory_out	2024/5/29 15:30	文本文档	98 KB
 kegg_core_id	2024/5/29 15:30	文本文档	20 KB
 kegg_core_out	2024/5/29 15:30	文本文档	350 KB
 Kegg_count_details1	2024/5/29 15:30	文本文档	23 KB
 kegg_histogram1	2024/5/29 15:30	文本文档	4 KB
 kegg_unique_id	2024/5/29 15:30	文本文档	3 KB
 kegg_unique_out	2024/5/29 15:30	文本文档	26 KB
 list	2024/5/29 15:14	文件	1 KB
 Major_Cog_Category1	2024/5/29 15:30	文本文档	1 KB
 matrix	2024/5/29 15:15	文本文档	343 KB
 new_genes_count	2024/5/29 15:15	文本文档	1 KB
 pan_default	2024/5/29 15:15	文本文档	1 KB
 pan_genome	2024/5/29 15:21	文本文档	5 KB
 PAN_PHYLOGENY.ph	2024/5/29 15:22	PH 文件	1 KB
 PAN_PHYLOGENY_MOD.nwk	2024/5/29 15:22	NWK 文件	1 KB
 PAN_PHYLOGENY_MOD.ph	2024/5/29 15:22	PH 文件	1 KB
 plots	2024/5/29 15:21	gnuplot comma...	2 KB
 plots_default	2024/5/29 15:15	gnuplot comma...	3 KB
 u_clusters	2024/5/29 15:15	文本文档	8,742 KB
 UNIQUE_COG_hits3	2024/5/29 15:29	文本文档	28 KB
 UNIQUE_kegg_hits3	2024/5/29 15:30	文本文档	30 KB

➤ 总结

优势： 1.软件有两种安装方式，满足不同人群使用需求
2.交互式界面，操作简单便捷，易上手
3.运行速度非常快，耗时短
4.高级分析可以直接进行功能注释，无需额外下载数据库

缺点： 1.目前更新至2017年，内置数据库版本较旧