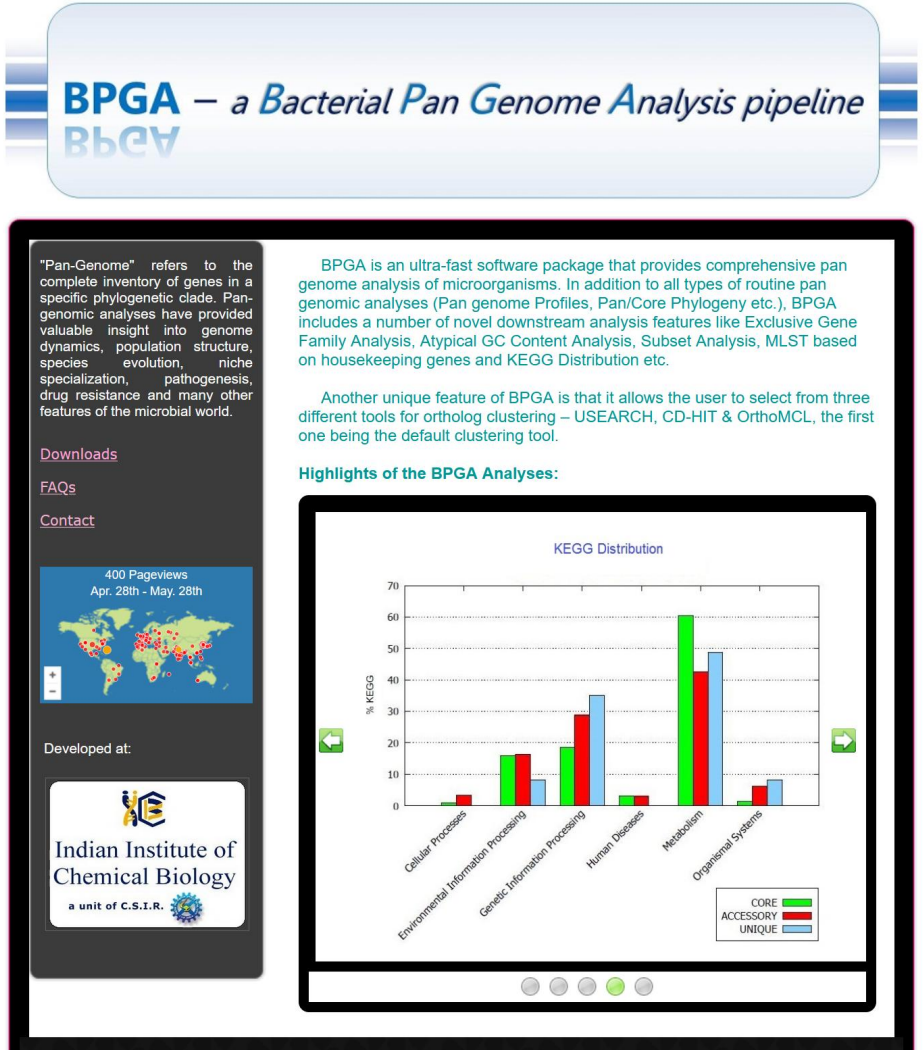
Novogene 诺禾致源

BPGA软件分享（windows版）

Novogene 诺禾致源

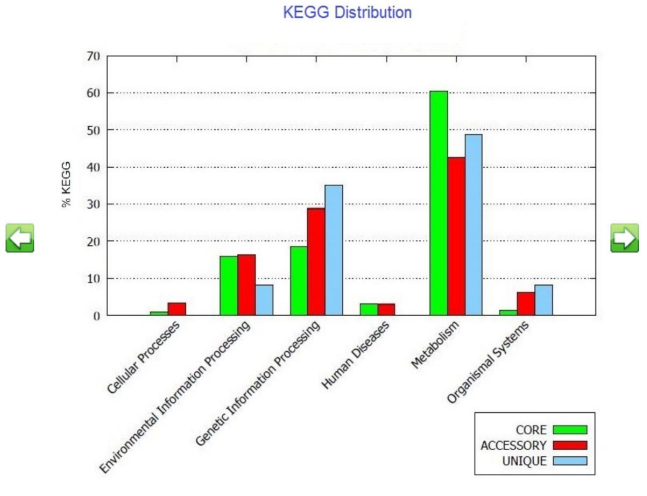
➢ 软件介绍（https://iicb.res.in/bpga/index.html）

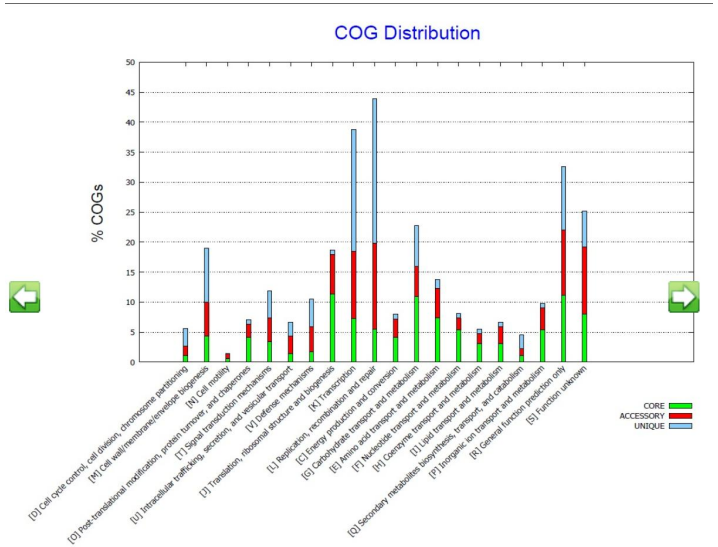


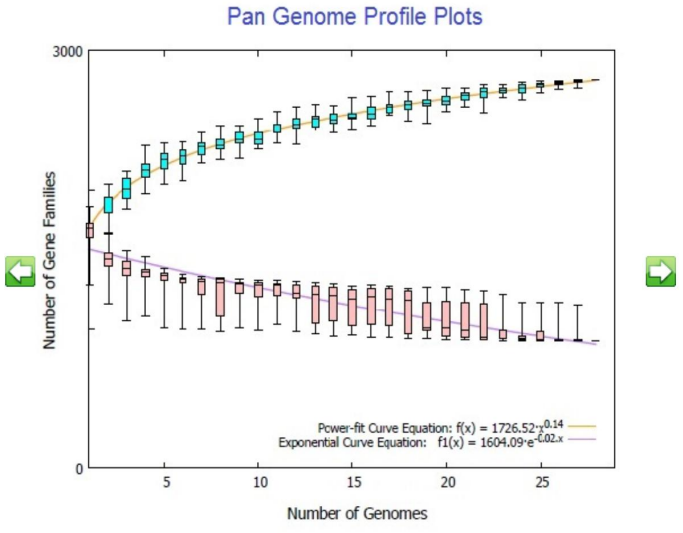
BPGA 是一种超快速的软件包，为微生物提供全面的泛基因组分析。除了所有类型的常规泛基因组分析（如泛基因组图谱、泛/核心系统发育树等）外，BPGA 还包含许多新颖的下游分析功能，如特有基因家族分析、非典型 GC 含量分析、子集分析、基于管家基因的 MLST 以及 KEGG 分布等。

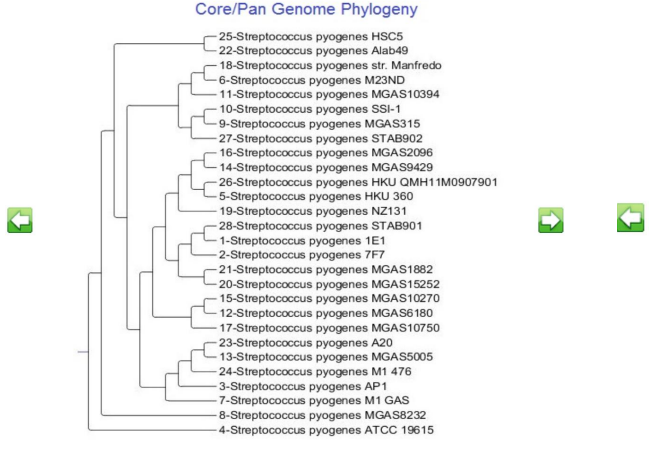
Novogene 诺禾致源

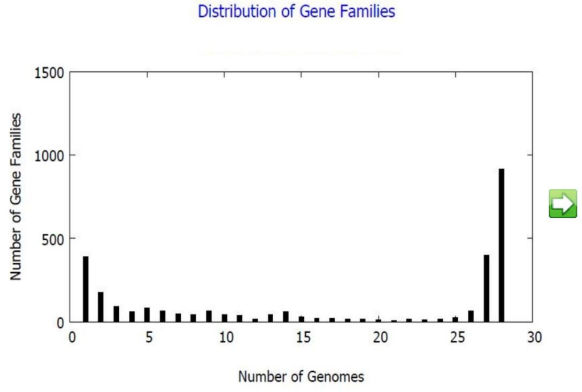
➢ 软件介绍





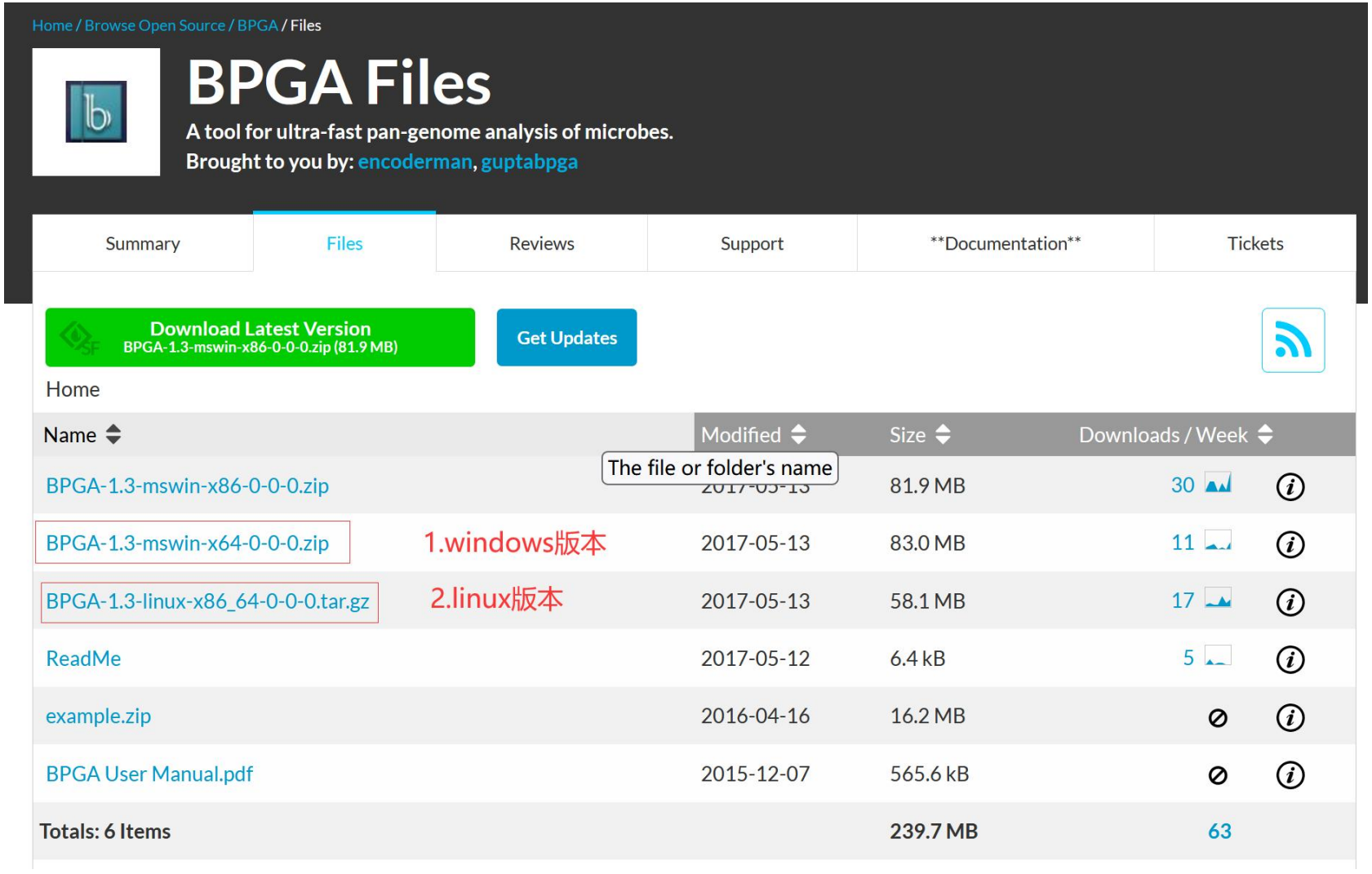






Novogene 诺禾致源

➢ 软件下载



- 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

Novogene 诺禾致源

➢ 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 Gnupot_Win-64bit_Version or Download_Gnupot_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)

Novogene 诺禾致源

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="MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVDKHIATI
                     ...
```

- NCBI Protein FASTA files sample: (.faa) 2 .faa文件

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

(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]
MRTNFKVSFYLRSNYENKEGKSPVMLRVFLNGEMSNFG
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVD
```

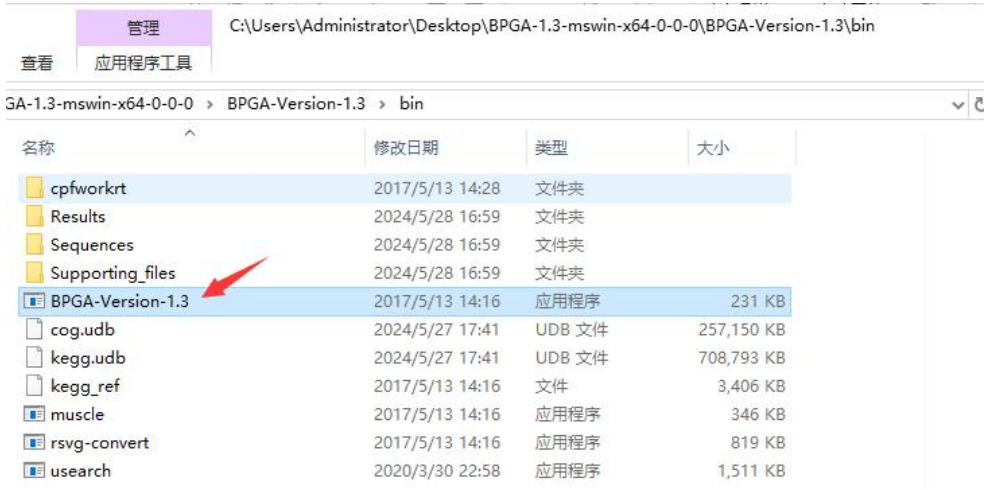
- Any Protein FASTA files sample:(.fasta) 4 .fasta文件

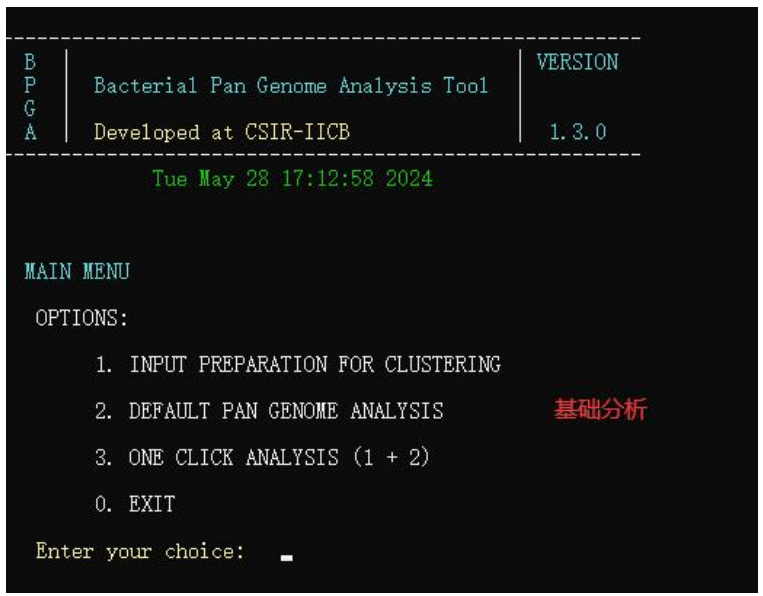
```
>Any_header_information
MRTNFKVSFYLRSNYENKEGKSPVMLRVFLNGEMSNFG
MTENEQIFWNRVLELAQSQLKQATYEFFVHDARLLKVD
```

- 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

Novogene 诺禾致源

➢ 软件分析(基础分析)



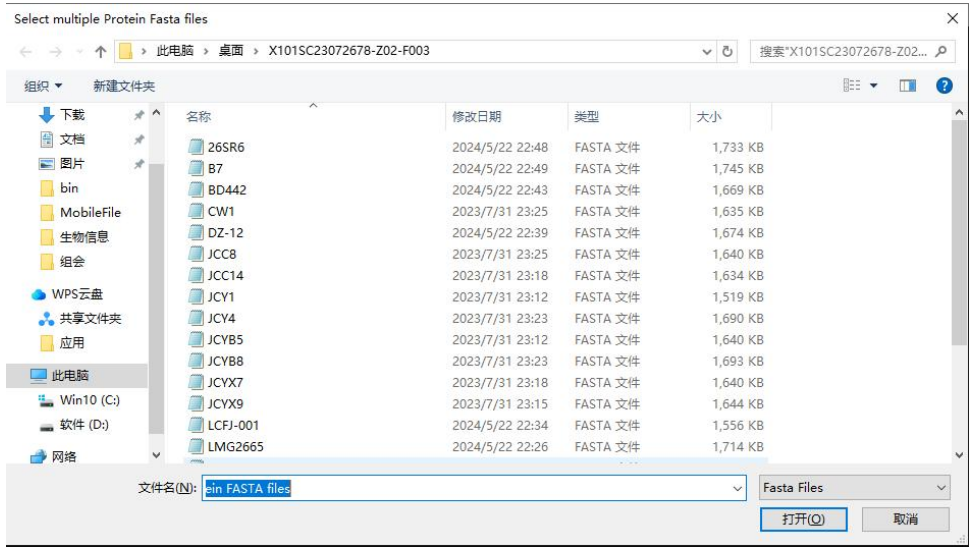


✓ step1：运行主程序，可以选择3一次完成基础分析，也可以分步骤分析

Novogene 诺禾致源

➢ 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..
```

Novogene 诺禾致源

➢ 软件分析（基础分析）

```
-------------------------------------------------------
B |                                   | VERSION
P |  Bacterial Pan Genome Analysis Tool |
G |                                   |
A |  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
```

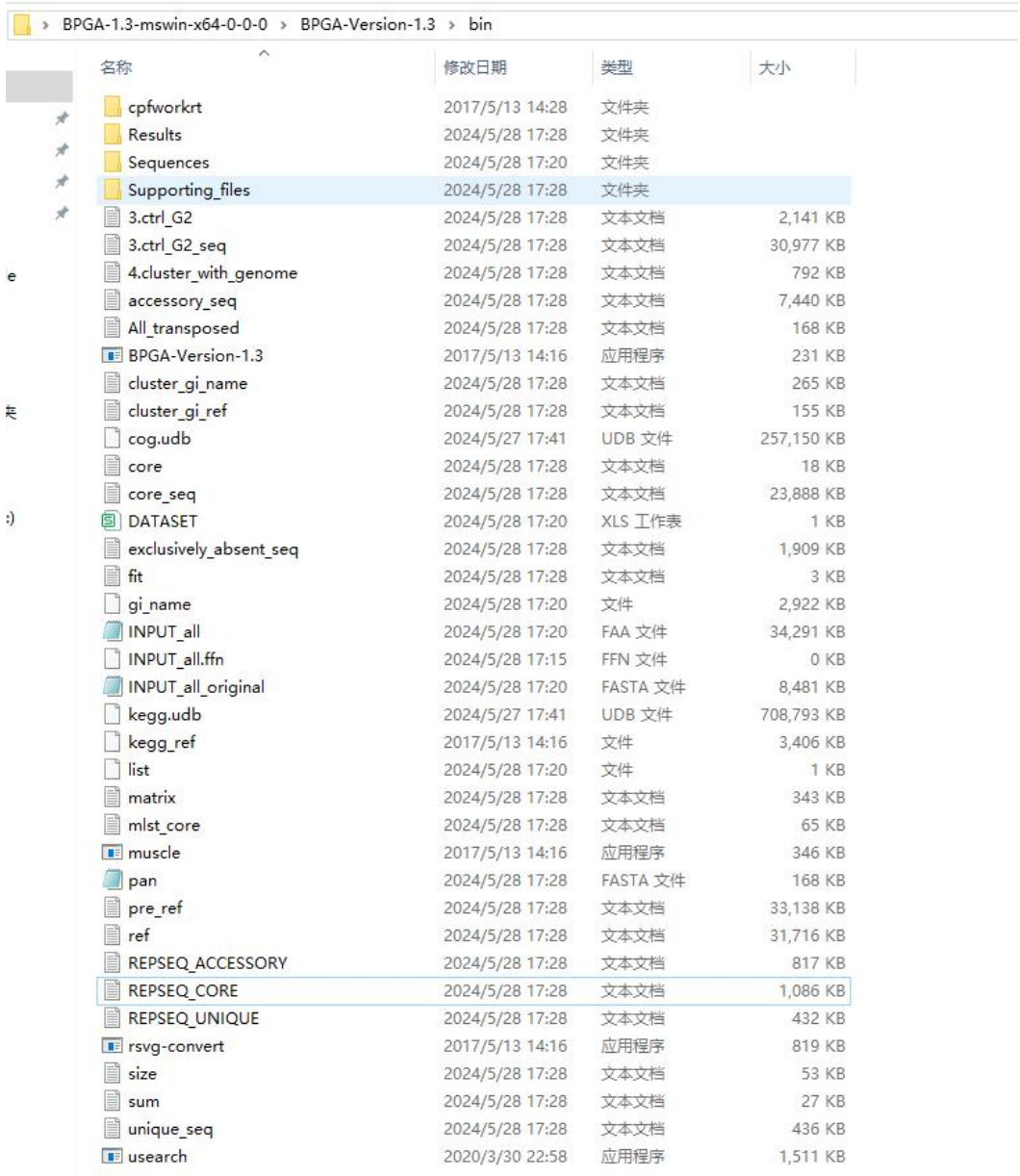
✓ step3：PAN基因分析

```
-DEFAULT PAN GENOME ANALYSIS:

  1. Use USEARCH Clustring 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 Clustring 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
```

Novogene 诺禾致源

➢ 软件分析（基础分析）



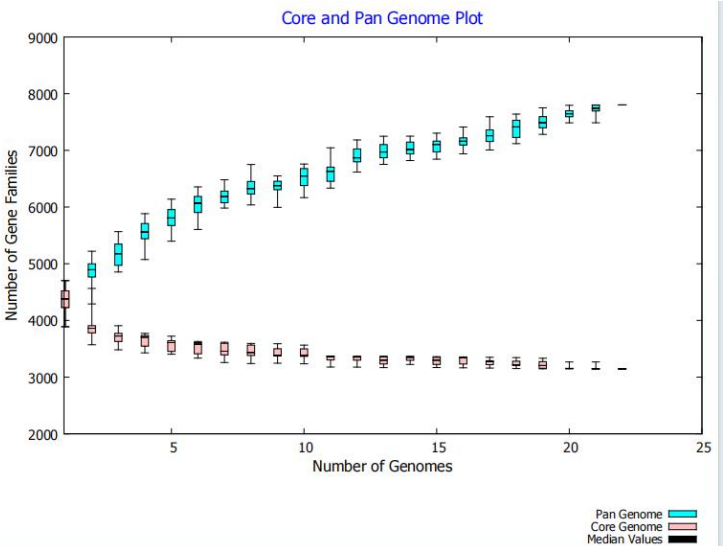
Novogene 诺禾致源

➢ 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:
```

✓ 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。

```
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
```



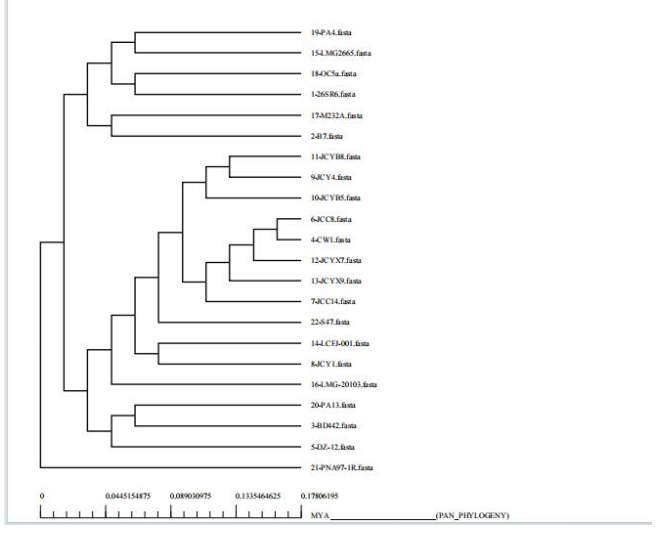
Novogene 诺禾致源

➢ 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
```

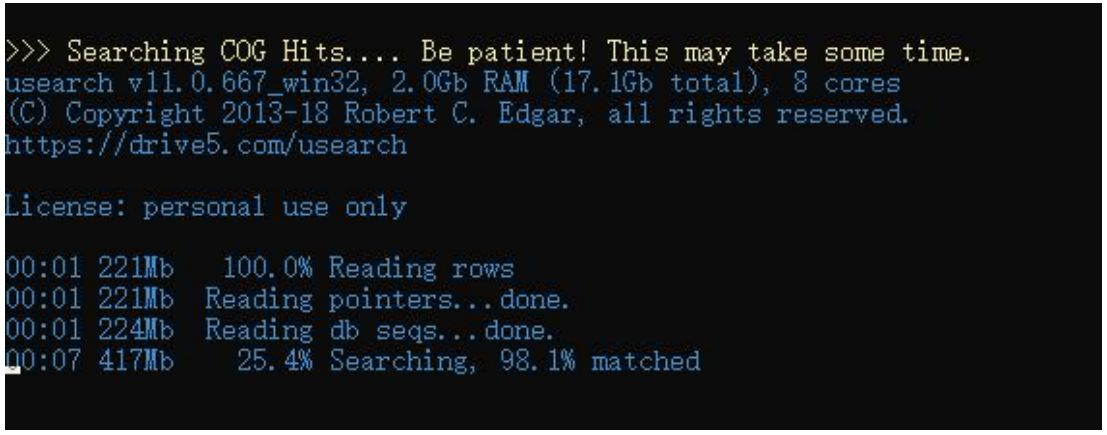
```
      Pan Phylogeny is over!

      For MLST based core phylogeny, refer tutorial.

      Press enter to continue.

.................................Alingning cluster number 001.

IUSCLE v3.8.31 by Robert C. Edgar

ttp://www.drive5.com/muscle
his software is donated to the public domain.
lease cite: Edgar, R.C. Nucleic Acids Res 32(5), 1792-97.

seq_001 22 seqs, max length 4492, avg  length 4304
0:00:00      5 MB(0%)  Iter   1  100.00%  K-bit distance matrix
0:00:00     14 MB(1%)  Iter   1    4.76%  Align node
```

✓ 进化树分析



Novogene 诺禾致源

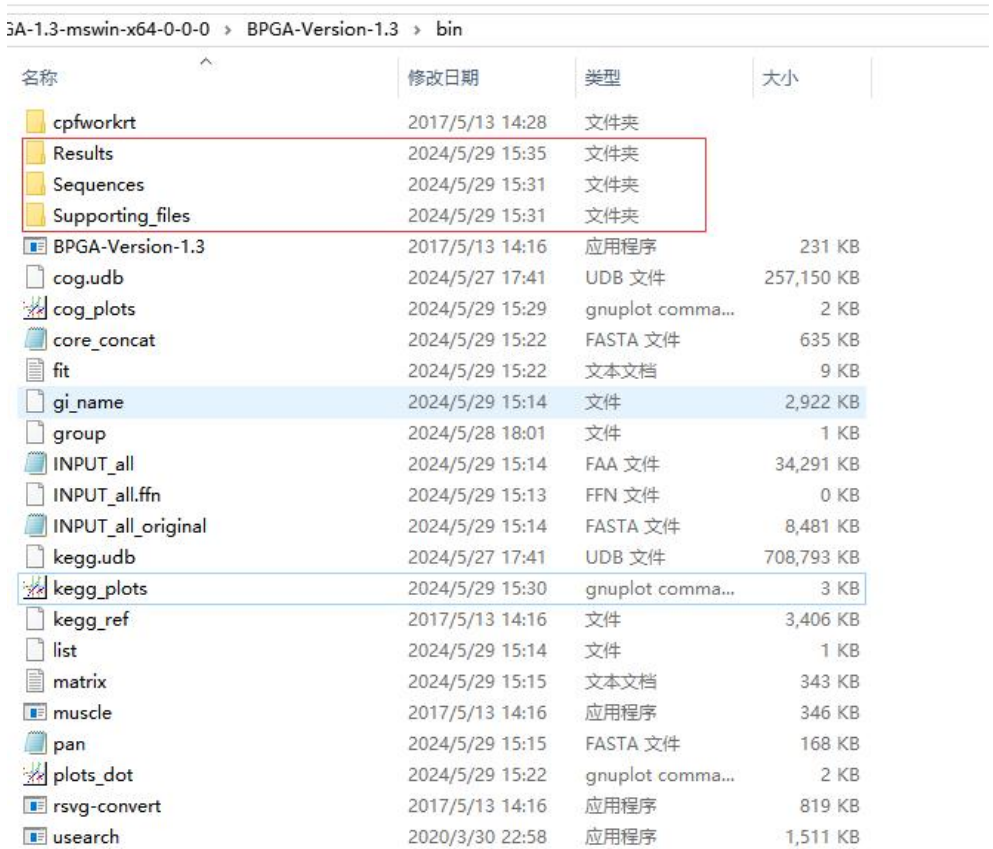
➢ 软件分析（高级分析）



✓ 功能分析（内置KEGG/COG数据库）

Novogene 诺禾致源

➢ 软件分析（结果目录）



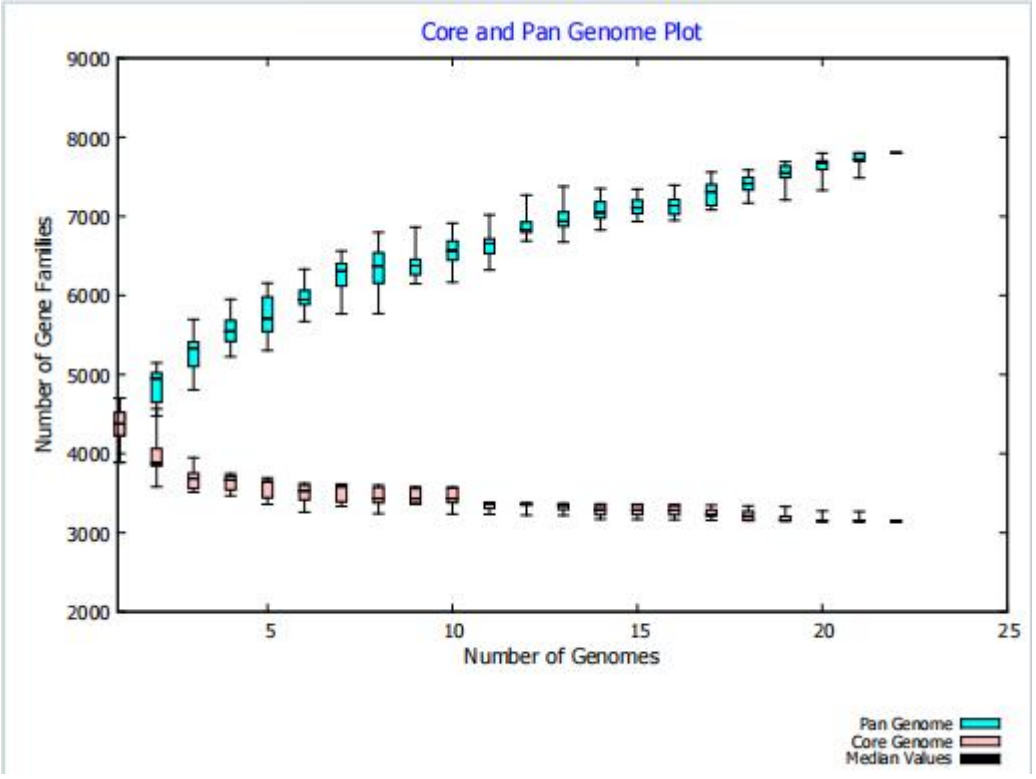
① Results文件夹中是分析的结果图片

② Sequences中是core，unique等基因的序列文件

③ Supporting_files是分析过程文件

Novogene 诺禾致源

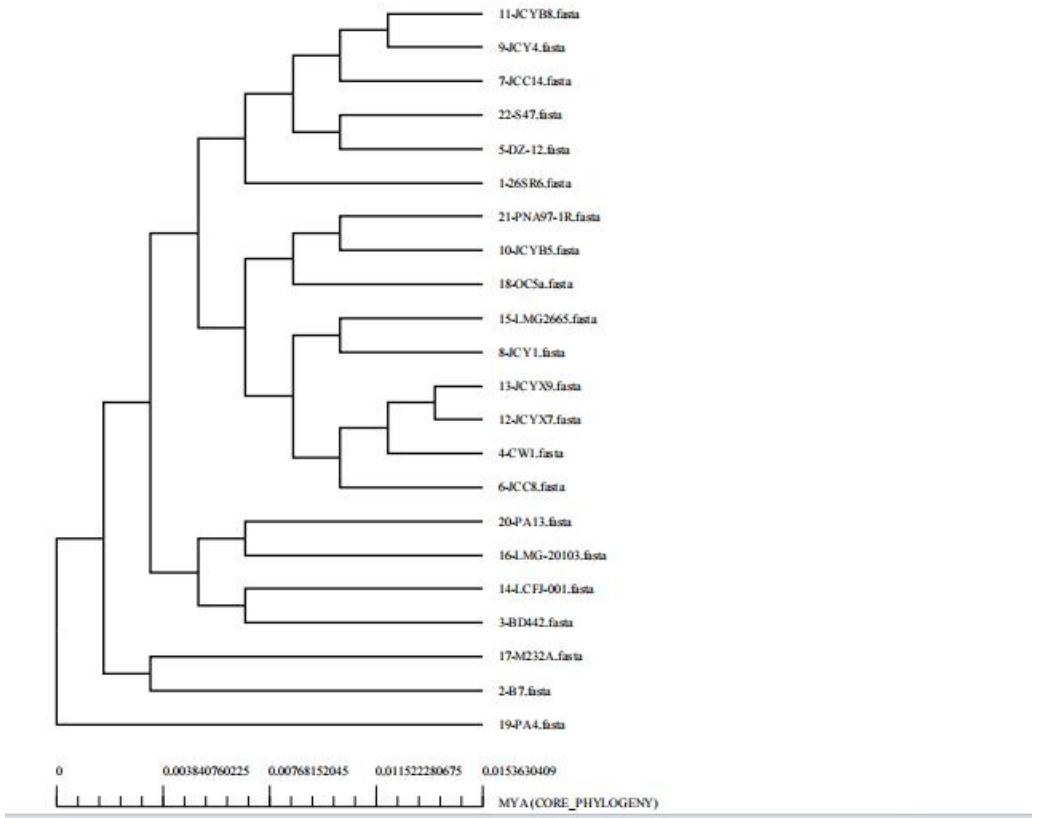
➢ 软件分析（结果目录）



⟶ corepan基因箱型图，类似meta的corepan图

Novogene 诺禾致源

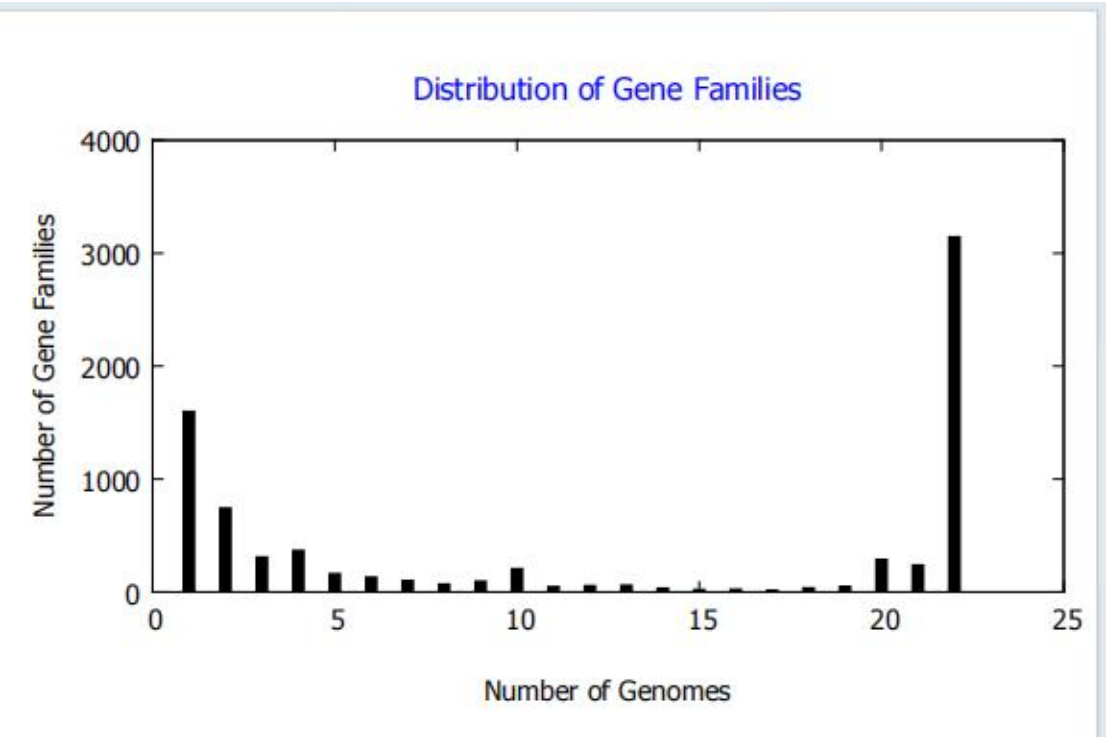
➢ 软件分析（结果目录）



→ core基因进化树图

Novogene 诺禾致源

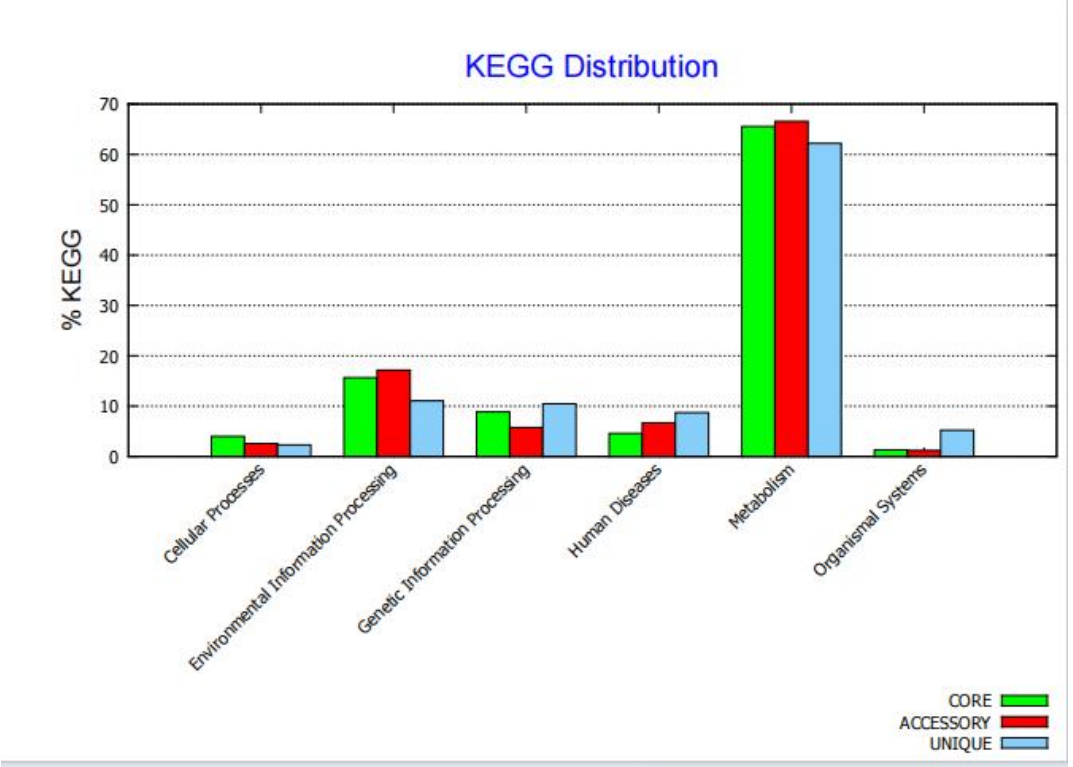
➢ 软件分析（结果目录）



→ 各个样本中基因数目统计图

Novogene 诺禾致源

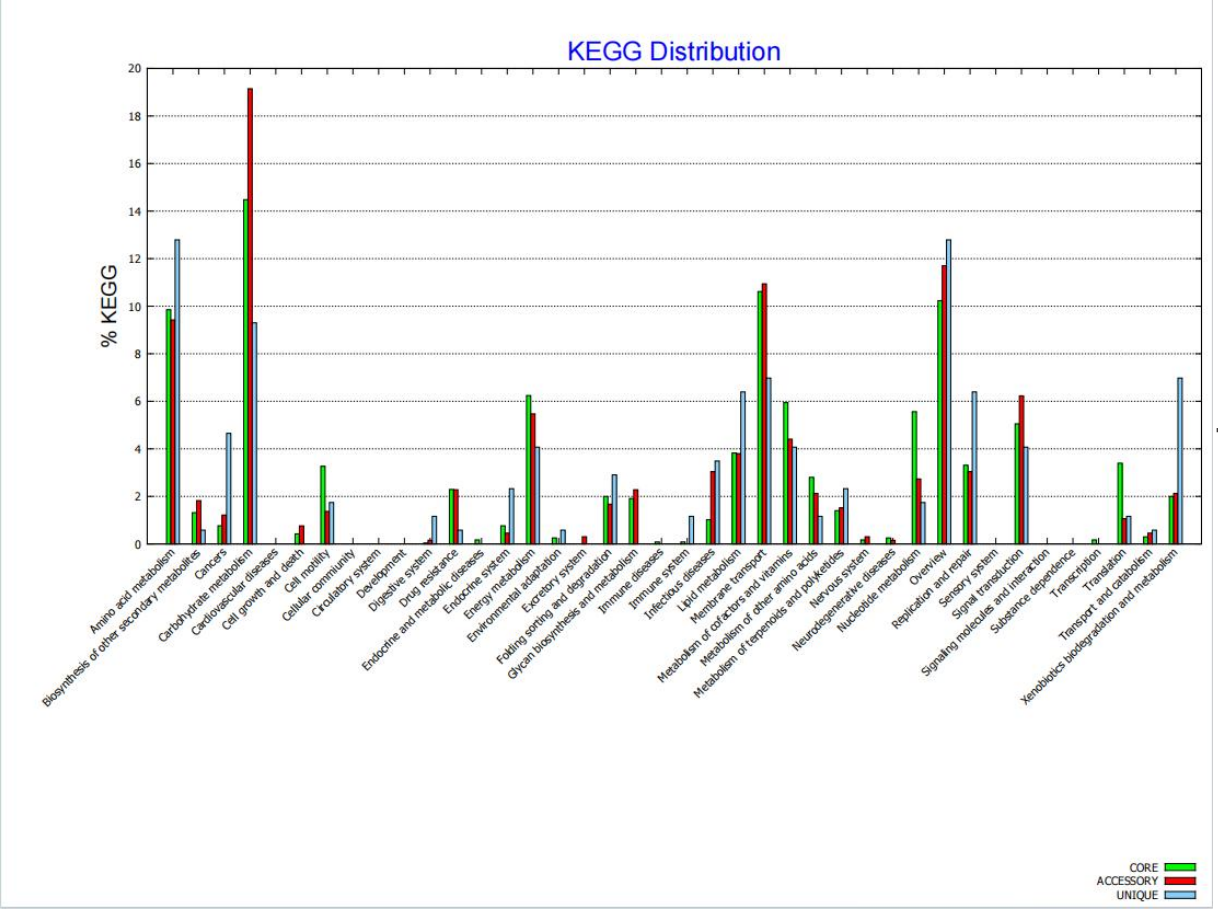
➢ 软件分析（结果目录）



→ 各类基因在KEGG数据库level1中的功能分布情况

Novogene 诺禾致源

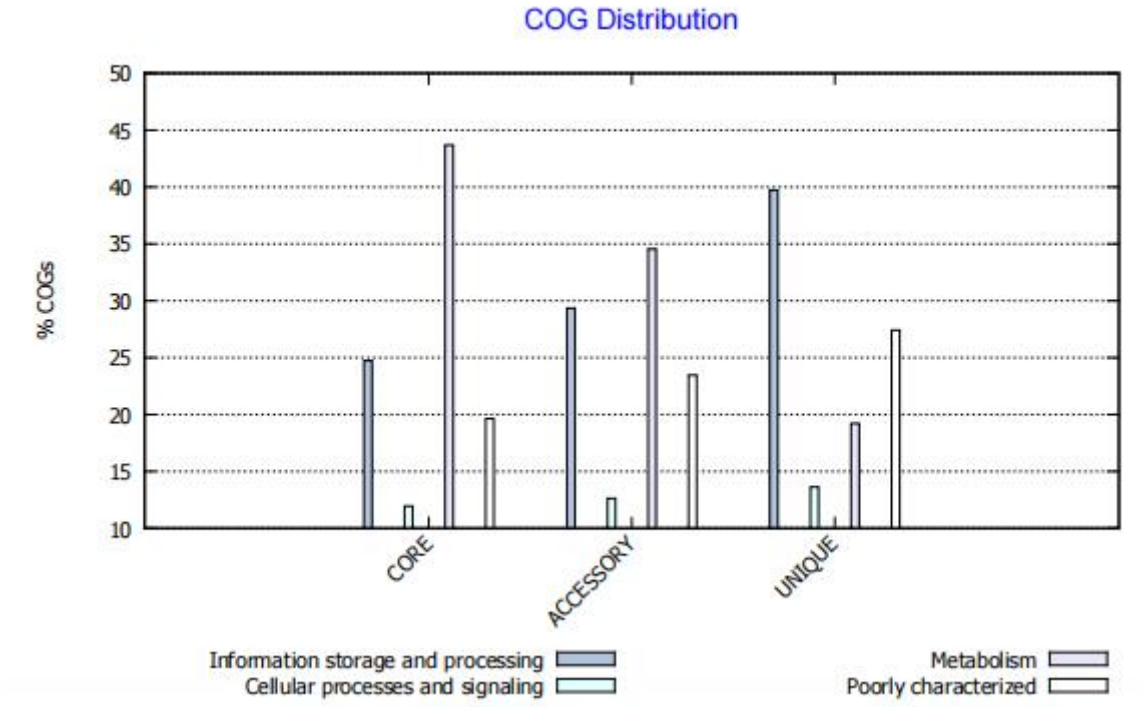
➢ 软件分析（结果目录）



→ 各类基因在KEGG数据库level2中的功能分布情况

Novogene 诺禾致源

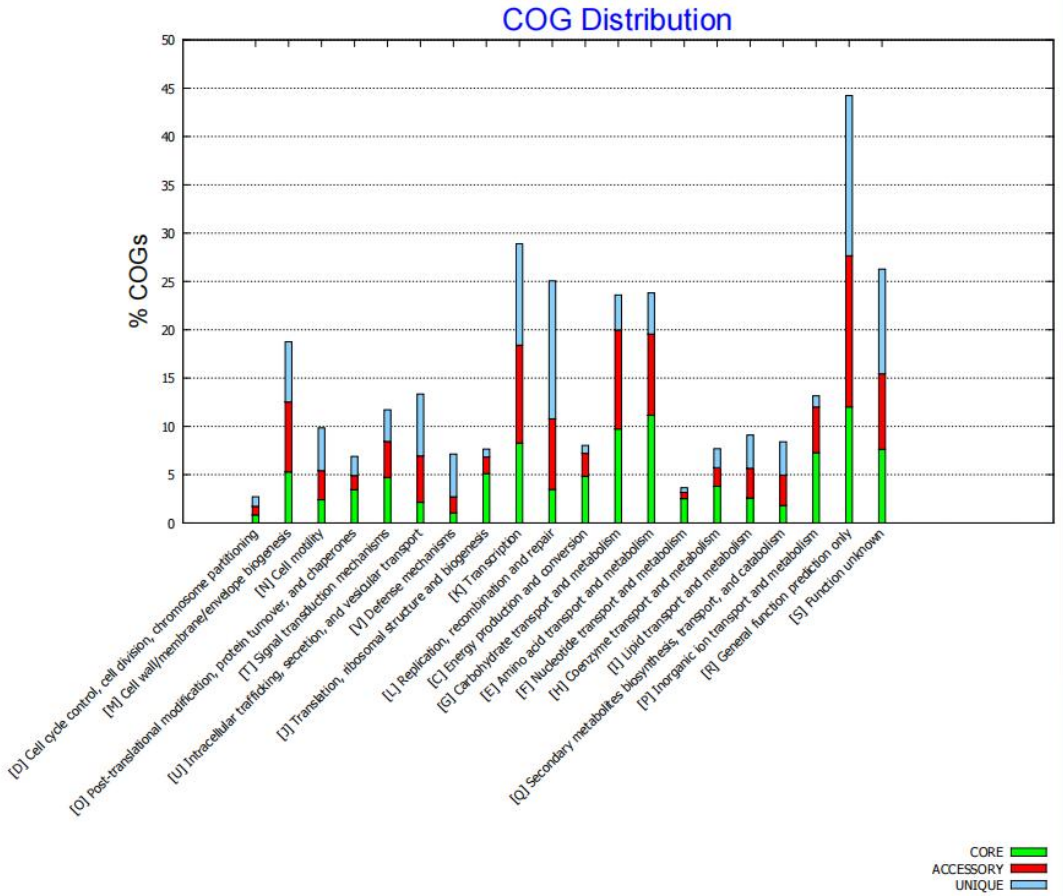
➢ 软件分析（结果目录）



→ 各类基因在COG数据库level1中的功能分布情况

Novogene 诺禾致源

➢ 软件分析（结果目录）



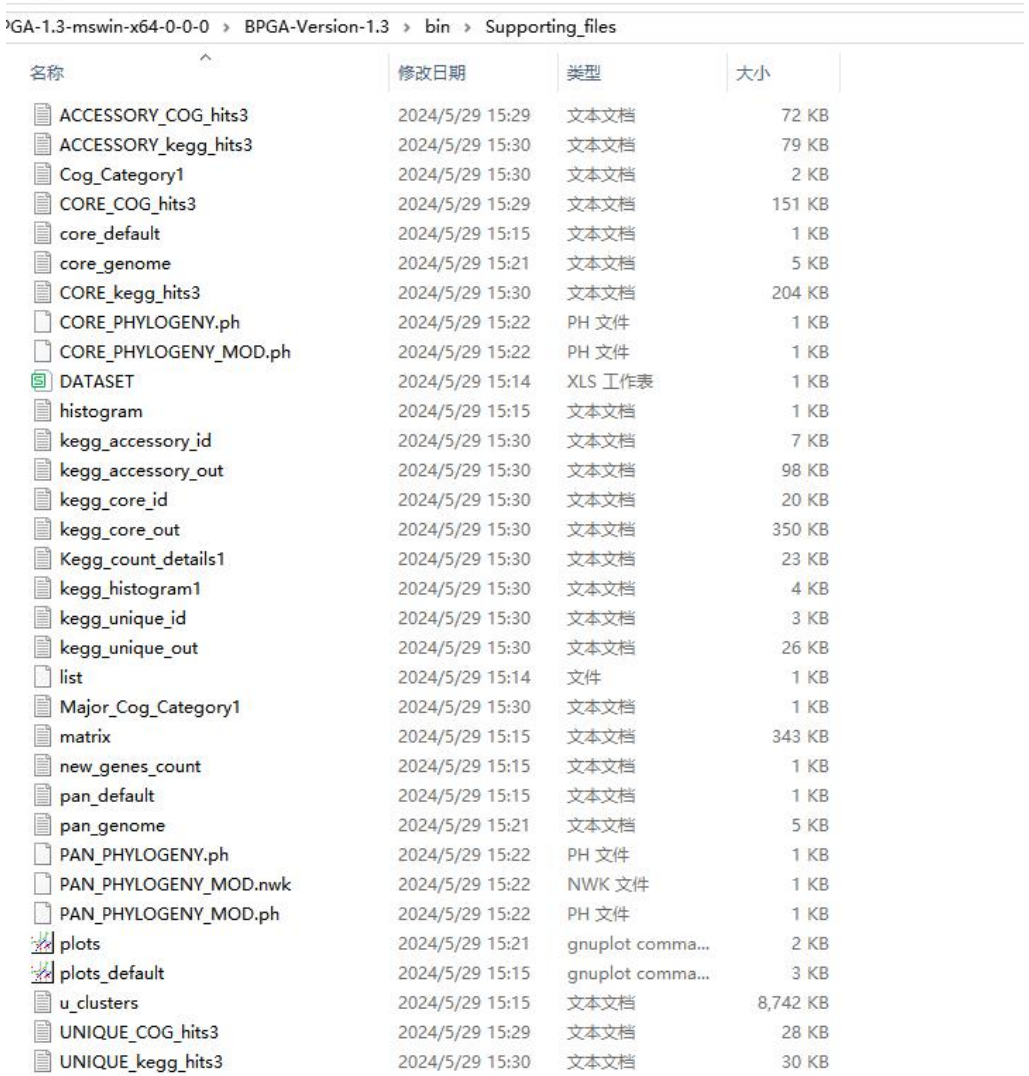
→ 各类基因在COG数据库level2中的功能分布情况

Novogene 诺禾致源

➢ 软件分析（结果目录）

Novogene 诺禾致源

➢ 软件分析（结果目录）



Novogene 诺禾致源

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