Protocol to analyze the bacterial pangenome using PAN2HGENE software

The PAN2HGENE is a computational tool that enables two main analyses. First, the tool can identify gene products absent from the original prokaryotic genome sequence. Second, it enables automated comparative analysis for both complete and draft genomes. All analyses are performed through a simple and intuitive graphical user interface without the need for extensive and complex command lines.
Protocol to analyze the bacterial pangenome using PAN2HGENE software

Pablo Henrique Caracciolo Gomes de Sá¹,4,* Jorianne Thyeska Castro Alves,2 and Adonney Allan de Oliveira Veras3,5,*

¹Federal Rural University of Amazonia Campus Tomé-Açu (UFRA), Pará, Brazil
²Pará State University Campus Marabá (UEPA), Pará, Brazil
³Faculty of Computing, Federal University of Pará Campus Castanhal (CCAST-UFPA), Pará, Brazil
⁴Technical contact
⁵Lead contact
*Correspondence: pablo.sa@ufra.edu.br (P.H.C.G.d.S.), allanveras@ufpa.br (A.A.d.O.V.)
https://doi.org/10.1016/j.xpro.2022.101327

SUMMARY
The PAN2HGENE is a computational tool that enables two main analyses. First, the tool can identify gene products absent from the original prokaryotic genome sequence. Second, it enables automated comparative analysis for both complete and draft genomes. All analyses are performed through a simple and intuitive graphical user interface without the need for extensive and complex command lines. For complete details on the use and execution of this protocol, please refer to Silva de Oliveira (2021).

BEFORE YOU BEGIN
The PAN2HGENE is a computational tool that performs two main analyses: 1- the identification of new gene products that are not represented in the original genomic sequence. Despite the high accuracy of genome assembly software, none of them is 100% accurate in generating the result. When mapping the raw reads against an assembly result or even the file containing a complete genome, it was observed that several reads do not show any match. With the assembly of these unmapped reads followed by an annotation, it was possible to observe gene products that were absent from the original genomic sequence used as input for analysis. And 2- comparative analysis for both complete and draft genomes. Both analyses are performed automatically.

The pipeline has three execution modes, complete with all modules or executing each analysis separately (only the identification of new gene products or only the comparative analysis.).

PAN2HGENE is equipped with an intuitive graphical interface that facilitates the execution process without the need to use extensive and complex command lines, the results consist of reports and graphs aimed at helping the user’s analysis process.

KEY RESOURCES TABLE

| REAGENT or RESOURCE | SOURCE | IDENTIFIER |
|---------------------|--------|------------|
| Deposited data      |        |            |
| Fasta files         | https://www.ncbi.nlm.nih.gov | GenBank: AP012324.1, LR134348.1, NZ_JACZEM010000001.1 |
| SRA files           | https://www.ebi.ac.uk | EMBL-EBI: SRR1151287, DRR017668, SRR611381 |

(Continued on next page)
**MATERIALS AND EQUIPMENT**

This protocol was created using a desktop with a fourth-generation Core i5 processor, 16 GB of RAM, and the Ubuntu 20.04 operating system. PAN2HGENE has also been successfully tested on Debian and Mint operating systems. So, if the user wanted to reproduce the protocol in the Debian distribution, it will be necessary to log in as root and then remove the sudo command from the beginning of the execution of each protocol command.

**STEP-BY-STEP METHOD DETAILS**

**Step 1: install dependencies**

**Timing:** 1 h

The complete installation of PAN2HGENE starts with the installation of the system dependencies and then continues with the installation of other programs that are part of the pipeline.

1. To start the installation run the commands in the Box 1. Below to update the system.

   **Box 1**
   ```
   sudo apt-get install make
   sudo apt-get install build-essential
   sudo apt-get install curl
   ```

   **CRITICAL:** If you face any problem with the commands of the Box 1, we suggest update your system with the commands ‘sudo apt-get update’ and ‘sudo apt-get upgrade’. After this try to run the Box 1 again.

2. Then we will install some packages and programs with the commands below.

   **Box 2**
   ```
   sudo cpan install YAML.pm
   sudo cpan install YAML::XS
   sudo cpan install Bio::AlignIO
   sudo cpan install Statistics::LineFit
   sudo cpan install Statistics::Distributions
   ```
Step 2: installation of software that compose the pipeline

Timing: 2 h 22 min 30 s

The following are the steps to install the external software that are part of the PAN2HGENE pipeline.

3. **SPAdes Installation.** SPAdes is available at [http://cab.spbu.ru/software/spades/](http://cab.spbu.ru/software/spades/), the installation process follows below.

   ```
   Box 3
   wget http://cab.spbu.ru/files/release3.15.3/SPAdes-3.15.3-Linux.tar.gz
   tar -xzf SPAdes-3.15.3-Linux.tar.gz
   
   sudo cpan install Bio::Perl
   sudo cpan install Bio::SearchIO::hmmer3
   sudo apt-get install screen
   sudo apt-get install bowtie2
   sudo apt-get install blast2
   sudo apt-get install samtools
   sudo apt-get install python3-distutils
   sudo apt-get install python
   sudo cpan install DBI
   sudo apt-get install mafft
   sudo apt-get install mcl
   sudo apt-get install phylip
   
   mv SPAdes-3.15.3-Linux/ SPAdes/
   sudo mv SPAdes/ /opt/
   cd /opt/
   sudo chmod 777 -R SPAdes/
   
   Box 4
   sudo cpan install Bio::Perl
   sudo cpan install Bio::SearchIO::hmmer3
   sudo apt-get install screen
   sudo apt-get install bowtie2
   sudo apt-get install blast2
   sudo apt-get install samtools
   sudo apt-get install python3-distutils
   sudo apt-get install python
   sudo cpan install DBI
   sudo apt-get install mafft
   sudo apt-get install mcl
   sudo apt-get install phylip
   
   mv SPAdes-3.15.3-Linux/ SPAdes/
   sudo mv SPAdes/ /opt/
   cd /opt/
   sudo chmod 777 -R SPAdes/
   ```

   a. At the end of the installation, the user must move the SPAdes folder to the /opt directory.

   b. To validate the installation, run the command (see Box 5), shown in Figure 1.

4. **PGAP Installation.** The PGAP software is available at [https://sourceforge.net/projects/pgap/](https://sourceforge.net/projects/pgap/), after downloading it follow the steps below, it is important to note that at the time of building this protocol the most current version was PGAP-1.2.1.
a. The following step is the configuration of the PGAP script (/opt/PGAP/PGAP.pl), it can be done using gedit or other text editor preferred by the user.

```
Box 6
```

```
tar -xzf PGAP-1.2.1.tar.gz
mv PGAP-1.2.1/ PGAP/
sudo mv PGAP/ /opt/
cd /opt/
sudo chmod 777 -R PGAP/
```

b. Modify the path of the software leaving the lines of the file the same as the lines without comments (lines that do not start with the # character). It is necessary to adjust the execution path of the programs e.g., /usr/bin/formatdb according to the path where the software is in your operating system. In Box 8 below, the lines that do not start with the character # demonstrate the configuration performed in the Ubuntu operating system.

```
Box 7
```

```
sudo gedit /opt/PGAP/PGAP.pl
```

```
Box 8
```

```
### programs from BLAST
my $formatdb="/usr/bin/formatdb";
my $blastall="/usr/bin/blastall";
### programs from mcl
my $mcl="/usr/bin/mcl";
### programs from mafft
```

Figure 1. SPAdes test: Demonstrates the SPAdes test run
c. Look for the 3 lines below and replace "./" with "/opt/PGAP/". After the replacement the lines should look like Box 10.

```perl
system("perl /opt/PGAP/multiparanoid.pl -species " . join(".pep", @species) . " -pep -unique 1");
system("perl /opt/PGAP/Blast_Filter.pl All.blastpAll.pep $coverage $identity $score | $mcl -abc -I 2.0 -o All.cluster");
system("perl /opt/PGAP/inparanoid.pl $blastall $thread $formatdb $score $global $local $species[$i].pep $species[$j].pep");
```

Box 10

d. Save the file and close gedit. To validate that PGAP was installed correctly, run the command below and result should be similar to this (Figure 2).

```bash
/opt/PGAP/PGAP.pl
```

5. **Prokka Installation.** The Prokka software is available [https://github.com/tseemann/prokka](https://github.com/tseemann/prokka), to install the version for Ubuntu/Debian/Mint run the commands below.

```bash
sudo apt-get install libdatetime-perl libxml-simple-perl libdigest-md5-perl git default-jre bioperl
sudo cpan Bi::Perl
sudo git clone https://github.com/tseemann/prokka.git/opt/prokka
sudo chmod 777 -R /opt/prokka
/opt/prokka/bin/prokka -setupdb
sudo chmod 777 -R /opt/prokka
```
To verify that Prokka was installed correctly, run the command below and the result should be similar to this (Figure 3).

```
pablo@pablo-VirtualBox:~/opt/temp$ /opt/PGAP/PGAP.pl
====== Pan-Genome Analysis Pipeline (PGAP) ======
Version 1.2.1
Usage: perl PGAP.pl [Options]
Options:
   --strains    String    Input strains nicknames, and join them with '+', for example: A+B+C
   --input      String    Input data directory
   --output     String    Result output directory
   --cluster    String    Run homologous gene clustering
   --pangenome  String    Run pan-genome analysis
   --variation  String    Run homologous clusters variation analysis
   --evolution  String    Run evolution analysis
   --function   String    Run Function analysis
   --method     String    GF for GeneFamily method, and MP for MultiParanoid method
                           for GF: fast, but not very accurate
                           score, identity, coverage are employed
                           for MP: slow, but more accurate
                           score, coverage, local, global are employed
   --thread     Int       Number of processors to use in blastall. [default:1]
   --score      Int       Minimum score in blastall. [default:1e-10]
   --evalue     Decimal  Maximal E-value in blastall. [default:1e-10]
   --coverage   Decimal  Minimum alignment coverage for two homologous proteins. [default:0.5]
   --local      Decimal  Minimum local alignment overlap in MP method. [default:0.25]
   --global     Decimal  Minimum global alignment overlap in MP method. [default:0.5]
   --identity   Decimal  Minimum alignment identity for two homologous proteins. [default:0.5]
   --bootstrap  Int       Bootstrap times for phylogenetics tree. [default:1]
   -h or help    String    Display this message
```

Figure 2. PGAP test: Demonstrates the PGAP test run

- To verify that Prokka was installed correctly, run the command below and the result should be similar to this (Figure 3).

```
Box 13
/opt/prokka/bin/prokka
```

6. R Installation. The R software is installed by following the commands below.

```
Box 14
sudo apt-get update
sudo apt-get install r-base r-base-dev
```

- After installation, run R and install the libraries.

```
Box 15
sudo R
install.packages('ape')
install.packages('plotrix')
install.packages('minpack.lm')
install.packages('ctv')
```
7. **Tbl2asn Installation.** The `tbl2asn` tool is installed by running the following commands.

```
library('ctv')
install.views('Phylogenetics')
update.views('Phylogenetics')
q()
```

Box 16

```
wget -N ftp://ftp.ncbi.nih.gov/toolbox/ncbi_tools/converters/by_program/tbl2asn/
linux64.tbl2asn.gz

gzip linux64.tbl2asn.gz

sudo chmod +x linux64.tbl2asn

mv linux64.tbl2asn /usr/local/bin/tbl2asn
```

a. And to test if everything is correct, run the command below and the result should be similar to this (Figure 4).

Box 17

```
tbl2asn –help
```

8. **MySQL Installation.** To install MySQL server run the following commands.

Box 18

```
sudo apt-get install mysql-server

sudo mysql_secure_installation
```

a. MySQL will ask you to create a password for the root user. Enter the password and answer Y when asked.

```
    pablo@pablo-VirtualBox:/opt/temp$ /opt/prokka/bin/prokka
    Name: Prokka 1.14.6 by Torsten Seemann <torsten.seemann@gmail.com>
    Synopsis: rapid bacterial genome annotation
    Usage: prokka [options] contigs.fasta
    General:
    --help           This help
    --version        Print version and exit
    --citation       Print citation for referencing Prokka
    --quiet          No screen output (default OFF)
    --debug          Debug mode: keep all temporary files (default OFF)
    Setup:
    --dbdir [X]     Prokka database root folders (default '/opt/prokka/db')
    --listdb        List all configured databases
    --setdbd        Index all installed databases
    --cleandb       Remove all database indices
    --depends       List all software dependencies
    Outputs:
    --outdir [X]   Output folder [auto] (default '')
    --force         Force overwriting existing output folder (default OFF)
    --prefix [X]   Filename output prefix [auto] (default '')
```

Figure 3. Prokka test: Demonstrates the Prokka test run
b. The component checks to see if the new password is strong enough. Choose one of the three levels of password validation:
   i. Low. A password containing at least 8 characters.
   ii. Medium. A password containing at least 8 characters, including numeric, mixed case characters, and special characters.
   iii. Strong. A password containing at least 8 characters, including numeric, mixed case characters, and special characters, and compares the password to a dictionary file. Enter 0, 1, or 2 depending on the password strength you want to set.

c. The script then prompts for the following security features: Remove anonymous users? Disallow root login remotely? Remove the test database and access to it? Reload privilege tables now?

d. To check if your MySQL was installed correctly run the command below and the result should be similar to this (Figure 5).

```
Figure 5. MySQL test: Demonstrates the MySQL test run
```

```
sudo systemctl status mysql
```

**Step 3: using PAN2HGGENE**

© Timing: 1 h 27 min

After installing all dependencies, the user must download PAN2HGGENE. To start using it, follow the steps below.

**9. PAN2HGGENE Download.** The PAN2HGGENE jar package is available at [https://sourceforge.net/projects/pan2hgene-software/](https://sourceforge.net/projects/pan2hgene-software/). Download the pan2hgenev2.0.jar and lib_v2.tar.xz files leaving both in the same directory. See the example below. The PAN2HGGENE pipeline can be executed in three different ways, each one performing a specific analysis.

```
pablo@pablo-VirtualBox:~$ tbl2asn --help
tbl2asn 25.8  arguments:
   -p Path to files [String]  Optional
   -r Path for results [String]  Optional
   -i Single input file [File In]  Optional
   -o Single output file [File Out]  Optional
   -x Suffix [String]  Optional
   default = .rs
   -r Recurse [T/F]  Optional
   default = F
   -t Template file [File In]  Optional
   -a File type [Any]
```

```
Figure 4. Tbl2asn test: Demonstrates the Tbl2asn test run
```

Box 19
```
sudo systemctl status mysql
```

```
pablo@pablo-VirtualBox:~$ sudo systemctl status mysql
● mysql.service - MySQL Community Server
 Loaded: loaded (/lib/systemd/system/mysql.service; enabled; vendor preset: enabled)
 Active: active (running) since Thu 2022-01-20 08:20:36 -03; 34min ago
 Process: 759 ExecStartPre=/usr/share/mysql/mysql-systemd-start pre (code=exited, status=0/SUCCESS)
 Main PID: 929 (mysqld)
 Status: "Server is operational"
 Tasks: 37 (limit: 9468)
 Memory: 417.4M
 CGroup: /system.slice/mysql.service
         └─929 /usr/sbin/mysqld
```

```
Figure 5. MySQL test: Demonstrates the MySQL test run
```
10. **PAN2HGENE Product Identification analysis.** The input for this analysis is a pair of files in FASTA and FASTQ format, the FASTA file can contain a complete genome or a draft of the organism. An attempt is made to identify possible new gene products for the analyzed genomes.

**Note:** In this example, product identification analysis is performed using a *Bifidobacterium breve* DSM20213 genome and paired reads from the Illumina HiSeq 2000. To start the Product Identification analysis, place the fasta genome and fastq reads in the same folder, follow the steps below.

```bash
Box 20
cd /home/pablo/panTest/
tar -xf lib_v.2.tar.xz
sudo chmod 777 -R /home/pablo/panTest/
java -jar pan2hgenev2.0.jar
```

**Box 21**
```
java -jar pan2hgenev2.0.jar
```

a. If this is your first use, enter the root user and password and press the create DB button, else enter the root user and root password in the indicated fields and then click the Connect button, then click on the Next button (Figure 6).

b. On the following screen, it is necessary to enter the project name and select the type of analysis to be performed. In the following example, the name “Test1” and the Product Identification option were added, after that press the New button (Figure 7).

c. Data input is done in the following window. Press the Browse button to select the FASTA file (Remember that fasta genome and fastq reads must be in the same folder). The reads files will be displayed below, select the appropriate reads for the organism, inform the type of reads and if it is paired, inform the order and orientation (Figure 8).

d. Press Add Read button and confirmation message will be displayed (Figure 9).

e. Note that the reads are now marked as used, repeat the same process if there are more genomes. Then click Next (Figure 10).

![Figure 6. Database management: The window has all functions to handle database, like status, drop, create e connect database](image)
f. The screen below will be shown, here it is possible to modify the parameters of PAN2HGENE for Bowtie, Comparative Analysis, and Annotation process. In this case, we will use the default parameters. So just click the Save Data button, then click Next (Figure 11).
g. The screen below will be shown. To start the analysis, click on the Perform analysis button. And in the Logs field, it is possible to check the analysis steps being performed. When the analysis is finished, the message Complete Analysis will appear in the Logs field, as can be seen on the screen below. Now close PAN2HGENE and go to the folder where the data was analyzed (Figure 12).
h. Several files will be inside the folder, in addition to the fasta genome and the fastq reads used in the analysis. The result of the Product Identification analysis are the three files marked below, GenomeNameBlastResult_Products.fasta, GenomeNameBlastResult_report.pdf, and GenomeNameBlastResult_Report.txt (Figure 13).

Figure 7. Project management: In this window, the user will create the project, so just inform the name of the project and the type of analysis to be performed

Figure 8. Input data: In this window, the user will add the files in fasta and fastq format depending on the analysis chosen
11. **PAN2HGENE Comparative analysis.** In this example, three fasta genomes, *Bifidobacterium breve* DSM20213 (complete), *Bifidobacterium breve* NCTC11815 (complete) and *Bifidobacterium breve* PRL2020 (draft with six contigs) were used. To start Comparative Analysis, place all fasta genomes in the same folder as shown below.

⚠️ CRITICAL: The initial steps are the same as shown in Figures 6–12, with the exception of the input data window that changes for this specific analysis.

a. When informing the directory of the FASTA files, they will be listed as shown in the figure below (Figure 14).

b. The analysis results are organized in the pgfiles directory. The files that are the results of the Comparative Analysis are the files that start with the numbers 1, 2, 3, 4, 5 and the figures in PNG format (Figure 15).

12. **PAN2HGENE Full pipeline.** The Full Pipeline analysis performs Product Identification analysis and Comparative Analysis automatically and sequentially. Thus, the new gene products identified in the Product Identification step will be used in the Comparative Analysis step.

Note: Now follow the steps described previously in the item ‘10. PAN2HGENE Product Identification analysis’, selecting Full Pipeline analysis instead of Product Identification (Figure 16).
13. **Main graphical results.** The graphs are produced by running the comparative analysis, so the creation is included in the processing time (Figures 17 and 18).

### Step 4: patric optional usage

⊙ **Timing:** 2 min

Patric software has been integrated into the pipeline as an alternative to automatic annotation software. Thus, the user is free to choose between Patric or Prokka, for the annotation execution.

14. Your PAN2HGENE is now ready to use, if the user does not want to use Patric to make the annotation, it is not necessary to perform the following steps.

**Note:** However, the PAN2HGENE offers the option to perform all annotation analyses through PATRIC instead of Prokka (which is the default option). If you want to use PRATIC in the annotation process, follow the steps below. If you do not already have a PATRIC account, you will have to register on (https://patricbrc.org/).

a. Install PATRIC Command Line Interface.

**Box 22**

```bash
curl -o- https://github.com/PATRIC3/PATRIC-distribution/releases/download/1.034/patric-cli-1.034.deb
sudo dpkg --install patric-cli-1.034.deb
sudo apt-get -f install
```

b. If you prefer you can also install PATRIC using the tool gdebi.

**Box 23**

```bash
sudo apt-get install gdebi-core
sudo gdebi patric-cli-1.034.deb
```
c. Setting, copy the file "p3-login.pl", provided with the pan2hgene files, and replace it in the installation directory of Patric-cli.

```
cp -r <file directory> /usr/share/patric-cli/deployment/plbin/
```

d. When performing any of the PAN2HGENE analyses, it is possible to choose the Patric annotation instead of the Prokka annotation. Before saving the parameters, click on the PATRIC button (Figures 19 and 20).

**EXPECTED OUTCOMES**

Although there are several tools to perform the comparative analysis, PAN2HGENE stands out for its characteristics, presenting a simple graphical interface to facilitate the analysis, instead of complex command lines. This tool can perform the identification of possible new gene products in a genome and can also perform, unlike other tools, the comparative analysis using complete genomes and

![Figure 12. Perform analysis and Log: In this window, the user executes the analysis and follows the execution steps](image-url)
draft genomes. The results are presented graphically and/or textually, without the need for the user to use other programs to analyze or interpret the results.

Finally, it is important to note that both analyzes are performed automatically. And that the input data are fasta genomes and their reads in fastq, without the need for the user to create standardized inputs or need to manipulate input files (PAN2HGENE already does this automatically too).
LIMITATIONS
PAN2HGENE is able to perform comparative analysis using complete genomes and draft genomes as input. The PAN2HGENE will perform an automatic annotation through the Prokka or Patric. At this point, it is important to emphasize that the annotation does not depend on the PAN2HGENE, as the pipeline only uses the annotation generated by Prokka or Patric.

TROUBLESHOOTING
Problem 1
The figures generated as a result do not present information on the pangenome distribution.

Potential solution
This usually happens when few genomes are used in comparative analysis. PAN2HGENE can perform comparative analysis for a minimum of 3 genomes. In general, when 3 or 4 genomes are used in the analysis, the pangenome result has practically no distribution information. Thus, the solution is to add a greater amount of genomes in the analysis, remembering that PAN2HGENE does not have a maximum limit of genomes that it can analyze.

Problem 2
PAN2HGENE never gets to the end of the analysis, it just keeps processing.

Potential solution
Comparative analysis is a type of exponential analysis, in which the more genomes are inserted into the analysis, the greater the computational cost to perform it. What could be happening is that the computer used is not able to process the analysis. So, the solution, in this case, would be to perform an analysis with fewer genomes, to ensure that PAN2HGENE is working correctly and then try to use a more powerful computer to perform the analysis with all the selected genomes. For example, using a desktop with a fourth-generation Core i5 processor and 16 GB of RAM, when performing the comparative analysis for 10 genomes, the complete run took approximately one and a half hours.
And using the same computer to analyze 20 genomes, the complete runtime was approximately 4 h and 20 min.

Note: Problems 1 and 2 are related to the amount of genomes used and the hardware capacity used, respectively, and not to the PAN2HGENE software.

**RESOURCE AVAILABILITY**

**Lead contact**
Further information and requests should be directed to the lead contact, Allan Veras (allanveras@ufpa.br).
Materials availability
Results of the application describing the protocol are reported in Silva de Oliveira et al., 2021.

Data and code availability
The published article includes code generated and all used datasets are available at NCBI and described in the key resources table.

ACKNOWLEDGMENTS
Thanks to the Brazilian Research Council (CNPq) and the Federal University of Pará. This work is part of the research developed by BIOD (Bioinformatics, Omics, and Development research group, www.biod.ufpa.br). A.A.O.V. thanks the Federal University of Pará (UFPA), and P.H.C.G.S. thanks the Federal Rural University of Amazonia (UFRA).

Figure 17. Graphics results of pangenome: On the left we have the graph with the information about pangenome, together with the Heap’s Law calculation and on the right, in the pie chart, the information about the amount of unique gene products and genes shared between the analyzed organisms

Figure 18. Graphics results of unique genes: On the left we have the graph with information about the amount of unique genes per organism present in the analysis and on the right an example of a phylogenetic tree graph
AUTHOR CONTRIBUTIONS
A.A.O.V. conceived the idea and developed the software. P.H.C.G.S. and J.T.C.A, conducted the formal analysis and validation. A.A.O.V., P.H.C.G.S., and J.T.C.A. wrote the manuscript and generated the analytical protocols described here.

DECLARATION OF INTERESTS
The authors declare no competing interests.
REFERENCES

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A., Dvorkin, M., Kulikov, A., Lesin, V., Nikolenko, S., Pham, S., Prijibelski, A., et al. (2012). SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J. Comput. Biol. 19, 455–477.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. Bioinformatics 30, 2068–2069.

Silva de Oliveira, M., Thyseka Castro Alves, J., Henrique Caraccio Gomes de Sá, P., and Veras, A. (2021). PAN2HGENE–tool for comparative analysis and identifying new gene products. PLoS One 16, e0252414.

Wattam, A., Abraham, D., Dalay, O., Díaz, T., Driscoll, T., Gabbard, J., Gillespie, J., Gough, R., Hix, D., Kenyon, R., et al. (2014). PATRIC, the bacterial bioinformatics database and analysis resource. Nucleic Acids Res. 42, D581–D591.

Zhao, Y., Wu, J., Yang, J., Sun, S., Xiao, J., and Yu, J. (2012). PGAP: pan-genomes analysis pipeline. Bioinformatics 28, 416–418.