BGDMdocker: a Docker workflow for data mining and visualization of bacterial pan-genomes and biosynthetic gene clusters

Gong Cheng 1,2 , Quan Lu Corresp. 3, Ling Ma Corresp. 4, Guocai Zhang 4, Liang Xu 5, Zongshan Zhou Corresp. 1

1 Protection Research Center of Pomology, Research Institute of Pomology, Chinese Academy of Agricultural Sciences, Xingcheng, China
2 Forest Protection Research Institute of Heilongjiang Province, Harbin, China
3 Research Institute of Forest Ecology, Environment and Protection, Chinese Academy of Forestry, Beijing, China
4 School of Forestry, Northeast Forestry University, Harbin, China
5 Institute of Food Science and Technology, Chinese Academy of Agricultural Sciences, Beijing, China

Corresponding Authors: Quan Lu, Ling Ma, Zongshan Zhou
Email address: luquan@caf.ac.cn, maling63@163.com, zszhouqrj@163.com

Recently, Docker technology has received increasing attention throughout the bioinformatics community. However, its implementation has not yet been mastered by most biologists; accordingly, its application in biological research has been limited. In order to popularize this technology in the field of bioinformatics and to promote the use of publicly available bioinformatics tools, such as Dockerfiles and Images from communities, government sources, and private owners in the Docker Hub Registry and other Docker-based resources, we introduce here a complete and accurate bioinformatics workflow based on Docker. The present workflow enables analysis and visualization of pan-genomes and biosynthetic gene clusters of bacteria. This provides a new solution for bioinformatics mining of big data from various publicly available biological databases. The present step-by-step guide creates an integrative workflow through a Dockerfile to allow researchers to build their own Image and run Container easily.
BGDMdocker: a Docker workflow for data mining and visualization of bacterial pan-genomes and biosynthetic gene clusters

Gong Cheng¹,², Quan Lu³,*, Ling Ma⁴,*, Guocai Zhang⁴, Liang Xu⁵ and Zongshan Zhou¹,*

¹Protection Research Center of Pomology, Research Institute of Pomology, Chinese Academy of Agricultural Sciences, Xingcheng, China
²Forest Protection Research Institute of Heilongjiang Province, Harbin, China
³Research Institute of Forest Ecology, Environment and Protection, Chinese Academy of Forestry, Beijing, China
⁴School of Forestry, Northeast Forestry University, Harbin, China
⁵Institute of Food Science and Technology, Chinese Academy of Agricultural Sciences, Beijing, China

ABSTRACT

Recently, Docker technology has received increasing attention throughout the bioinformatics community. However, its implementation has not yet been mastered by most biologists; accordingly, its application in biological research has been limited. In order to popularize this technology in the field of bioinformatics and to promote the use of publicly available bioinformatics tools, such as Dockerfiles and Images from communities, government sources, and private owners in the Docker Hub Registry and other Docker-based resources, we introduce here a complete and accurate bioinformatics workflow based on Docker. The present workflow enables analysis and visualization of pan-genomes and biosynthetic gene clusters of bacteria. This provides a new solution for bioinformatics mining of big data from various publicly available biological databases. The present step-by-step guide creates an integrative workflow through a Dockerfile to allow researchers to build their own Image and run Container easily.

INTRODUCTION

Docker is an open source project and platform for building, shipping, and running any app, enabling the widespread distribution of applications. Docker allows users to package an application, along with all its dependencies, into a standardized unit for software development (Docker Inc, 2017). Docker includes three core structural compositions: Image, Container, and Repository. An image can start software as complex as a database, wait for you (or someone else) to add data, store the data for later use, and then wait for the next person. Containers afford similar resource isolation and allocation benefits as virtual machines; however, a different architectural approach allows the former to be much more portable and efficient. When an app is in Docker containers, setting up and maintaining different environments or tools for each language is not necessary (Docker Inc, 2017). The Docker Hub Registry allows users to find, manage, and pull Images from community, official, and private image libraries, and is free to use for public repositories (Docker Inc, 2017).
GitHub is a web-based source code version control repository and Internet hosting service that is mostly used for code. Compared with GitHub, Docker Hub is a cloud-based registry service of Docker, whose most notable advantages include workflow automation throughout the development pipeline based on Images and Counter of Docker.

Academic bioinformatics software programs generally suffer from limitations such as installation and configuration difficulties, large dependencies, and restrictions on the amount of data that may be uploaded to online servers. Therefore, several excellent software programs are of limited use to biologists. Bioinformatics tools may be merged with Docker technology to build reproducible and convenient types of workflows. Docker provides programmers, development teams, and bioinformaticians with a common toolbox that allows users to take full advantage of bioinformatics tools, thus helping to build, ship, and run any app, as well as distribute apps anywhere. Docker technology is suitable for use in the field of bioinformatics because of certain advantages and characteristics that allow applications to run in an isolated, self-contained package. This package may be efficiently distributed and executed in a portable manner across a wide range of computing platforms (Aranguren & Wilkinson, 2015; Belmann et al., 2015; Hosny et al., 2016). To date, numerous bioinformatics tools based on Docker have been developed and published in different programming languages such as perl and bioperl (The Perl Community, 2017; Hilmar Lapp, 2017), python and biopython (The Docker Community, 2017; Tiago, 2017), and R and Bioconductor (The Rocker Community, 2017; The Bioconductor Core Team, 2017). These projects have contributed to official Docker Images. The famous Galaxy program has also contributed to Docker Galaxy (Björn & Grüning, 2017).

Here, we used Docker technology to rapidly construct a pan-genome analysis process that may be used in Linux, Windows, or Mac environments (64-bit). The present process may also be deployed as a cloud-based system such as with Amazon EC2 or other cloud providers. This workflow should provide a useful service to biologists in the field of bioinformatics. Docker Containers have only a minor impact on the performance of common genomic pipelines (Tommaso et al., 2015).

Bacillus amyloliquefaciens has been extensively studied as an important biological control agent owing to its ability to inhibit the growth of fungi and bacteria (Nam et al., 2016). Using Docker, we rapidly executed a container (on Ubuntu 16.04 and Win10 hosts) to analyze the pan-genome and reveal biosynthetic gene cluster features of 44 B. amyloliquefaciens strains, as well as to visualize the results. The analytical workflow consisted of three toolkits: Prokka v1.11 (Seemann, 2014), panX (Ding et al., 2016), and antiSMASH3.0 (Weber et al., 2015), for prokaryotic genome annotation, pan-genome analysis and visualization, and analysis of biosynthetic gene clusters, respectively. We included all of these applications and their dependencies in a BGDMdocker (bacterial genome data mining Docker-based) to enable the workflow to be implemented online with a single run. We additionally wrote three standalone Dockerfiles for Prokka, panX, and antiSMASH in order to meet the various requirements of different users. We recommend setting up the workflow with three
independent files, each with a specific purpose. This method is presented in the Supplementary Information section. Here, we describe how to build the workflow and conduct the analysis in detail.

**MATERIALS & METHODS**

**Installation of latest Docker on your host:**

1. Copy the following commands for quickly & easily installing the latest Docker-CE (Ubuntu, Debian, Raspbian, Fedora, Centos, Redhat, Suse, Oracle, Linux etc. are all applicable):

   ```bash
   $ curl -fsSL get.docker.com -o get-docker.sh
   $ sudo sh get-docker.sh
   ```

   If user would like to use Docker as a non-root user, you should now consider adding your user to the "docker" group, e.g., using:

   ```bash
   $ sudo usermod -aG docker <user name>
   ```

   Type the following commands at your shell prompt. If this outputs the Docker version, your installation was successful.

   ```bash
   $ docker version
   ```

2. **Installing latest Docker on Windows 10 Enterprise:**

   The current version of Docker for Windows runs on 64-bit Windows 10 Pro, Enterprise, and Education editions. Download “InstallDocker.msi”. Double-click “InstallDocker.msi” to run the installer. Follow the install wizard to accept the license, authorize the installer, and proceed with the installation.

   Type the following commands at your shell prompt (cmd.exe or PowerShell). If this outputs the Docker version, your installation was successful.

   ```bash
   $ docker version
   ```

**Use Docker to build the BGDMdocker workflow:**

1. On your host (with Docker), type the following command lines to build a BGDMdocker workflow:

   ```bash
   $ git clone https://github.com/cgwyx/BGDMdocker.git
   ```

   or: download “BGDMdocker-master.zip” file
$ unzip BGDMdocker-master.zip

2. Build workflow Images:

$ cd ./BGDMdocker

$ docker build -t BGDMdocker:latest .

Or: pull Images of BGDMdocker from DockerHub, such as:

$ docker pull cgwyx/bgmdocker

3. Run a Container from the BGDMdocker Image:

$ docker run -it --rm -v home:home -p 8000:8000 BGDMdocker:latest /bin/bash

If use the "-v home:home" parameter, Docker will mount the local folder /home into the Container under /home, storing all of your data in one directory in the home folder of the host operating system; then, you may access the directories of home from inside the Container.

We analyzed the pan-genome and biosynthetic gene clusters of 44 B. amyloliquefaciens strains using the BGDMdocker workflow. For detailed commands, see Supplementary Information.

RESULTS

Fast and reproducible building of the BGDMdocker workflow across computing platforms using Docker

Using Docker technology, the Dockerfile script file can build Images and run a container in seconds or milliseconds on Linux and Windows. The file may also be deployed in Mac and cloud-based systems such as Amazon EC2 or other cloud providers. The Dockerfile is a small, plain-text file that may be easily stored and shared. Therefore, the user is not required to install and configure the programs.

Here, based on Debian 8.0 (Jessie) Image, we have established a novel Docker-based bioinformatics platform for the study of microbe genomes and pan-genomes (Fig. 1). The workflow, which offers the advantages of cross-platform and modular reuse, provides biologists with simple and standardized tools to extract biological information from their own experiments and from online sequence databases. Researchers may therefore focus solely on mining information from the obtained sequences rather than determining how to install the software package. We have uploaded this Dockerfile to GitHub for sharing with relevant scientific researchers.
Fig. 1 Schematic of BGDMdocker workflow based on Docker: Building Image and running Container from Dockerfile, then login interaction patterns to run software. This enables the user to load and save Image.tar, run and commit Container, pull from and push to Docker repository.

**Datamining and visualizing the pan-genomes of *B. amyloliquefaciens***

In order to explore the result data, a website was built for the interactive exploration of the *B. amyloliquefaciens* pan-genome and biosynthetic gene clusters using the BGDMdocker workflow. Visualization allowed for the rapid filtering and searching of genes. For each gene cluster, panX displayed an alignment and a phylogenetic tree, mapped mutations within that cluster to the branches of the tree, and inferred gene losses and gains on the core-genome phylogeny. Here, we provide the summary statistics of the pan-genome (Table 1), the phylogenetic relationships of the 44 *B. amyloliquefaciens* strains (Fig. 2), and screenshots of the website (Fig. 3, 4). All data may be visualized and downloaded without registration.

**Table 1** Summary statistics of pan-genome of 44 *B. amyloliquefaciens* strains

| Accession   | Strains | Gene numbers in pan-genome of *B. amyloliquefaciens* (Total genes, 172,388; Core gene clusters, 2,306) | Gene of strain genomes |
|-------------|---------|-------------------------------------------------------------------------------------------------|------------------------|
|             |         | Total gene  | Core gene | Acc gene | Uni gene | All gene | All protein |
| CYHL01000001 | JRS5   | 3,856       | 2,310     | 1,546    | 57       | 3,870     | 3,863       |
| CYHP01000001 | JRS8   | 3,994       | 2,311     | 1,683    | 118      | 4,016     | 4,006       |
| Reference | Genotype | NSP3 Length | NSP3 Start | NSP3 Stop | NSP3 | NSP4 |
|-----------|----------|-------------|------------|----------|------|------|
| NC_014551 | DSM7     | 3,935       | 2,307      | 1,628    | 21   | 4,030| 3,811 |
| NC_017188 | TA208    | 3,935       | 2,307      | 1,628    | 1    | 3,974| 3,847 |
| NC_017190 | LL3      | 3,981       | 2,308      | 1,673    | 19   | 4,037| 3,887 |
| NC_017191 | XH7      | 3,942       | 2,307      | 1,635    | 6    | 3,983| 3,846 |
| NC_017912 | Y2       | 4,099       | 2,310      | 1,789    | 46   | 4,148| 3,983 |
| NC_020272 | IT-45    | 3,803       | 2,310      | 1,493    | 4    | 3,832| 3,678 |
| NC_022653 | CC178    | 3,754       | 2,310      | 1,444    | 19   | 3,795| 3,641 |
| NC_023073 | LFB112   | 3,761       | 2,308      | 1,453    | 19   | 3,801| 3,637 |
| NZ_AUNGO1000001 | Lx-11    | 3,700   | 2,309    | 1,391    | 5    | 3,742| 3,619 |
| NZ_AUWKO1000001 | HB-26    | 3,797   | 2,311    | 1,486    | 30   | 3,842| 3,714 |
| NZ_AVQH01000001 | EGD-AQ141 | 4,079 | 2,311 | 1,768 | 54 | 4,121 | 3,995 |
| NZ_AWQY01000001 | UASWS BA1 | 3,794 | 2,309 | 1,485 | 8 | 3,806 | 3,681 |
| NZ_CP006058 | UMAF6639  | 3,825   | 2,311    | 1,514    | 20   | 3,879| 3,716 |
| NZ_CP006960 | UMAF6614  | 3,804   | 2,311    | 1,493    | 13   | 3,850| 3,695 |
| NZ_CP007242 | KHG19     | 3,775   | 2,310    | 1,465    | 19   | 3,816| 3,658 |
| NZ_CP010556 | L-H15     | 3,724   | 2,309    | 1,415    | 6    | 3,769| 3,615 |
| NZ_CP011278 | L-S60     | 3,728   | 2,310    | 1,418    | 7    | 3,773| 3,611 |
| NZ_CP013727 | MBE1283   | 3,794   | 2,314    | 1,480    | 24   | 3,856| 3,681 |
| NZ_CP014700 | S499      | 3,776   | 2,310    | 1,466    | 5    | 3,819| 3,671 |
| NZ_CP014783 | B15       | 3,820   | 2,315    | 1,505    | 13   | 3,875| 3,704 |
| NZ_CP016913 | RD7-7     | 3,597   | 2,308    | 1,289    | 39   | 3,656| 3,483 |
| NZ_DF836091 | CMW1      | 3,771   | 2,311    | 1,460    | 128  | 3,901| 3,706 |
| NZ_JCOC01000001 | EBL11    | 3,733   | 2,308    | 1,425    | 20   | 3,773| 3,682 |
| NZ_JMEG01000001 | B1895     | 3,824   | 2,306    | 1,518    | 167  | 4,026| 3,623 |
| NZ_JQNZ01000001 | X1        | 3,724   | 2,309    | 1,415    | 3    | 3,766| 3,619 |
| NZ_JTJG01000001 | JJC33M    | 3,888   | 2,309    | 1,579    | 121  | 3,952| 3,796 |
| NZ_JXAT01000001 | LPL-K103  | 3,709   | 2,309    | 1,400    | 15   | 3,743| 3,637 |
| NZ_JZDI01000001 | 12B       | 8,166   | 2,354    | 5,812    | 4,040| 8,194| 7,985 |
| NZ_KB206086 | DC-12     | 3,910   | 2,311    | 1,599    | 50   | 3,984| 3,842 |
| NZ_KN723307 | TF281     | 3,640   | 2,312    | 1,328    | 5    | 3,782| 3,571 |
| NZ_LGYP01000001 | 629       | 3,536   | 2,313    | 1,223    | 11   | 3,785| 3,427 |
| NZ_LJAU01000001 | Bs006     | 4,042   | 2,312    | 1,730    | 46   | 4,074| 3,969 |
| NZ_LJDI010000020 | XK-4-1    | 3,799   | 2,310    | 1,489    | 14   | 3,821| 3,701 |
| NZ_LMAG01000001 | RHNK22    | 3,781   | 2,309    | 1,472    | 37   | 3,837| 3,698 |
Genome sequences of 44 *B. amyloliquefaciens* strains downloaded from GenBank RefSeq database: “Acc gene” refers to accessory gene (dispensable gene); “Uni gene” refers to unique gene (strain-specific gene); “All genes” refers to gene of *.gbff* files recorder, including Pseudo Genes; “Total genes” refers to those used for pan-genome analysis gene of *.gbff* files recorder, excluding Pseudo Genes.

Fig. 2 Phylogenetic tree of 44 *B. amyloliquefaciens* strains

The tree was constructed using all genes shared between the 44 strains (2,306 core genes). The scale bar represents genetic distance.
Results from the identification and analysis of the biosynthetic gene clusters of 44 *B. amyloliquefaciens* strain genomes, using the BGDMdocker workflow, have been uploaded to our website. All data may be downloaded without registration.

Here, we provide brief summary statistics for the biosynthetic gene clusters of all 44 strains (Table 2), as well as an example of the type and number of biosynthetic gene clusters in the Y2 (NC_017912) strain (Table 3) and representative screenshots of the website (Fig. 5, 6). There are a total of 31 gene clusters in the genome of Y2. Among these, 21 gene clusters show similarities to known clusters in MIBiG.
such as surfactin, mersacidin, and fengycin; the remaining 10 gene clusters are unknown.

Table 2 Summary statistics of biosynthetic gene clusters of 44 B. amyloliquefaciens strains

| Accession          | Strains   | Total | known | Unknown | Type | Size (Mb) | Gene | Protein |
|--------------------|-----------|-------|-------|---------|------|-----------|------|---------|
| CYHL01000001       | JRS5      | 38    | 27    | 11      | 12   | 4.03148   | 3,870 | 3,863   |
| CYHP01000001       | JRS8      | 42    | 26    | 16      | 11   | 4.0909    | 4,016 | 4,006   |
| NC_014551          | DSM7      | 31    | 18    | 13      | 9    | 3.9802    | 4,030 | 3,811   |
| NC_017188          | TA208     | 29    | 17    | 12      | 10   | 3.93751   | 3,974 | 3,847   |
| NC_017190          | LL3       | 29    | 17    | 12      | 9    | 4.00199   | 4,037 | 3,887   |
| NC_017191          | XH7       | 29    | 17    | 12      | 10   | 3.9392    | 3,983 | 3,846   |
| NC_017912          | Y2        | 31    | 21    | 10      | 11   | 4.23862   | 4,148 | 3,983   |
| NC_020272          | IT-45     | 32    | 19    | 13      | 11   | 3.93687   | 3,832 | 3,678   |
| NC_022653          | CC178     | 35    | 19    | 16      | 9    | 3.91683   | 3,795 | 3,641   |
| NC_023073          | LFB112    | 35    | 20    | 15      | 10   | 3.94275   | 3,801 | 3,637   |
| NZ_AUNG01000001    | Lx-11     | 37    | 25    | 12      | 11   | 3.88689   | 3,742 | 3,619   |
| NZ_AUWK01000001    | HB-26     | 45    | 30    | 15      | 9    | 3.98936   | 3,842 | 3,714   |
| NZ_AVQH01000001    | EGD-AQ141 | 36    | 26    | 10      | 12   | 4.22259   | 4,121 | 3,995   |
| NZ_AWQY01000001    | UASWS BA1 | 37    | 25    | 12      | 11   | 3.94409   | 3,806 | 3,681   |
| NZ_CP006058        | UMAF6639  | 35    | 21    | 14      | 10   | 4.03464   | 3,879 | 3,716   |
| NZ_CP006960        | UMAF6614  | 32    | 20    | 12      | 10   | 4.00514   | 3,850 | 3,695   |
| NZ_CP007242        | KHG19     | 32    | 20    | 12      | 10   | 3.95336   | 3,816 | 3,658   |
| NZ_CP010556        | L-H15     | 32    | 19    | 13      | 10   | 3.90597   | 3,769 | 3,615   |
| NZ_CP011278        | L-S60     | 32    | 19    | 13      | 10   | 3.90302   | 3,773 | 3,611   |
| NZ_CP013727        | MBE1283   | 35    | 22    | 13      | 12   | 3.97993   | 3,856 | 3,681   |
| NZ_CP014700        | S499      | 33    | 19    | 14      | 11   | 3.93593   | 3,819 | 3,671   |
| NZ_CP014783        | B15       | 29    | 19    | 10      | 10   | 4.00675   | 3,875 | 3,704   |
| NZ_CP016913        | RD7-7     | 31    | 17    | 14      | 8    | 3.68821   | 3,656 | 3,483   |
| NZ_DF836091        | CMW1      | 30    | 20    | 10      | 11   | 3.90857   | 3,901 | 3,706   |
| NZ_JCOC01000001    | EBL11     | 35    | 23    | 12      | 11   | 3.92932   | 3,773 | 3,682   |
### Table 3: Biosynthetic gene clusters of Y2(NC_017912) strain

| Cluster     | Type              | Most similar known cluster                                                                 | MIBIG BGC-ID          |
|-------------|------------------|------------------------------------------------------------------------------------------|-----------------------|
| Cluster 1   | Nrps             | Surfactin_biosynthetic_gene_cluster (43% of genes show similarity)                        | BGC0000433_c1         |
| Cluster 2   | Cf_putative      | -                                                                                        | -                     |
| Cluster 3   | Cf_putative      | -                                                                                        | -                     |
| Cluster 4   | Cf_fatty_acid    | -                                                                                        | -                     |
| Cluster 5   | Phosphonate      | Pactamycin_biosynthetic_gene_cluster (3% of genes show similarity)                        | BGC0000119_c1         |
| Cluster 6   | Cf_saccharide    | Plantathiazolicin/plantazolicin_biosynthetic_gene_cluster (33% of genes show similarity) | BGC0000569_c1         |

167“Total” of Biosynthesis gene clusters includes “Known” and “Unknown”. “Known” of Biosynthesis gene clusters is inferred from the MIBiG (Minimum Information about a Biosynthetic Gene cluster, [http://mibig.secondarymetabolites.org](http://mibig.secondarymetabolites.org)). “Unknown” of Biosynthesis gene clusters is detected by Cluster Finder and further categorized into putative ("Cf_putative") biosynthetic types. A full integration of the recently published Cluster Finder algorithm now allows the use of this probabilistic algorithm to detect putative gene clusters of unknown types; “-” of host is unrecorded.
| Cluster  | Type                  | Description                                                                 | ID                |
|---------|-----------------------|-----------------------------------------------------------------------------|-------------------|
| 7       | Cf_putative          | -                                                                           | -                 |
| 8       | Otherks              | -                                                                           | -                 |
| 9       | Cf_fatty_acid        | -                                                                           | -                 |
| 10      | Cf_putative          | -                                                                           | -                 |
| 11      | Terpene              | -                                                                           | -                 |
| 12      | Cf_fatty_acid        | -                                                                           | -                 |
| 13      | Cf_putative          | -                                                                           | -                 |
| 14      | Cf_putative          | -                                                                           | -                 |
| 15      | Transatpks            | Macrolactin_biosynthetic_gene_cluster (90% of genes show similarity)       | BGC0000181_c1     |
| 16      | Nrps-Transatpks       | Bacillaene_biosynthetic_gene_cluster (85% of genes show similarity)         | BGC0001089_c1     |
| 17      | Nrps-Transatpks       | Fengycin_biosynthetic_gene_cluster (93% of genes show similarity)           | BGC0001095_c1     |
| 18      | Terpene              | -                                                                           | -                 |
| 19      | Cf_saccharide-T3pks  | -                                                                           | -                 |
| 20      | Transatpks            | Difficidin_biosynthetic_gene_cluster (100% of genes show similarity)        | BGC0000176_c1     |
| 21      | Cf_putative          | -                                                                           | -                 |
| 22      | Nrps-Bacteriocin      | Bacillibactin_biosynthetic_gene_cluster (100% of genes show similarity)    | BGC0000309_c1     |
| 23      | Cf_saccharide        | -                                                                           | -                 |
| 24      | Nrps                 | -                                                                           | -                 |
| 25      | Cf_saccharide        | Teichuronic_acid_biosynthetic_gene_cluster (100% of genes show similarity) | BGC0000868_c1     |
| 26      | Cf_putative          | -                                                                           | -                 |
| 27      | Cf_saccharide        | Bacilysin_biosynthetic_gene_cluster (100% of genes show similarity)         | BGC0001184_c1     |
| 28      | Cf_putative          | -                                                                           | -                 |
| 29      | Lantipeptide         | Mersacidin_biosynthetic_gene_cluster (90% of genes show similarity)         | BGC0000527_c1     |
| 30      | Cf_saccharide        | -                                                                           | -                 |
| 31      | Cf_putative          | -                                                                           | -                 |

*Cf putative* refers to putative biosynthetic types (unknown types) detected by Cluster Finder and further categorized, known types are from the [MIBiG](http://mibig.secondarymetabolites.org) (Minimum Information about a Biosynthetic Gene cluster).
| Cluster Type | From | To | Most similar known cluster |
|--------------|------|----|----------------------------|
| Cluster 1    | 3380357 | 358115 | Surface_biosynthetic_gene_cluster (43% of genes show similarity) |
| Cluster 2    | 382926  | 400100 | - |
| Cluster 3    | 457205  | 463870 | - |
| Cluster 4    | 553834  | 571410 | - |
| Cluster 5    | 611959  | 652640 | Pectinols_biosynthetic_gene_cluster (63% of genes show similarity) |
| Cluster 6    | 7073604 | 724559 | Plantaribiose kinase / plantaribiose biosynthetic_cluster (33% of genes show similarity) |
| Cluster 7    | 815908  | 837072 | - |
| Cluster 8    | 928306  | 969750 | - |
| Cluster 9    | 967814  | 1018791 | - |
| Cluster 10   | 1009821 | 1029470 | - |
| Cluster 11   | 1052066 | 1072806 | - |
| Cluster 12   | 1099917 | 1107373 | - |
| Cluster 13   | 1132407 | 1140612 | - |
| Cluster 14   | 1217728 | 1234302 | - |
| Cluster 15   | 1414044 | 1469984 | Maltolactin_biosynthetic_cluster (90% of genes show similarity) |
| Cluster 16   | 1766710 | 1842979 | Bacillulus_biosynthetic_cluster (55% of genes show similarity) |
| Cluster 17   | 2075576 | 2178057 | Fengycin_biosynthetic_cluster (75% of genes show similarity) |
| Cluster 18   | 2324424 | 2346307 | - |
| Cluster 19   | 2296287 | 2344898 | - |
| Cluster 20   | 2506981 | 2581912 | Difficitin_biosynthetic_cluster (100% of genes show similarity) |
| Cluster 21   | 3239477 | 3253309 | - |
| Cluster 22   | 3786690 | 3845673 | Bacilliltonin_biosynthetic_cluster (100% of genes show similarity) |
| Cluster 23   | 3533963 | 3552378 | - |
| Cluster 24   | 3634364 | 3661226 | - |
| Cluster 25   | 3683737 | 3738852 | Tacheficin_biosynthetic_cluster (100% of genes show similarity) |
| Cluster 26   | 3771770 | 3829219 | - |
| Cluster 27   | 3882481 | 3949290 | Bacillolipin_biosynthetic_cluster (100% of genes show similarity) |
| Cluster 28   | 4008100 | 4026182 | - |
| Cluster 29   | 4028151 | 4052145 | - |
| Cluster 30   | 4173935 | 4195285 | - |
| Cluster 31   | 4202706 | 4219124 | - |

**Fig. 5** Screenshot of visualization of the total number of biosynthetic gene clusters of the Y2 strain

**Fig. 6** Screenshot of website for the visualization of Transatpks type cluster of Y2 (most similar known cluster to the Difficitin_biosynthetic_gene_cluster)
DISCUSSION

The Dockerfiles of BGDMdocker scripts are convenient for deployment and sharing, and it is easy for other users to customize the Images by editing the Dockerfile directly. This is in contrast to Makefiles and other installations, for which the resulting builds differ across different machines (Boettiger, 2014). Dockerfiles can maintain and update related adjustments, rapidly recover from system failure events, control versions, and build application environments with the optimal flexibility. BGDMdocker Images enable portability and modular reuse. Bioinformatics tools are written in a variety of languages and require different operating environment configurations across platforms. Docker technology is capable of executing the same functions and services in different environments without additional configurations (Folarin et al., 2015), thus creating reproducible tools with high efficiency. By constructing pipelines with different tools, bioinformaticians may automatically and effectively analyze biological problems of interest. The BGDMdocker Container enables application isolation with high efficiency and flexibility. Applications may run Container independently with Docker technology, and each management command (start, stop, boot, etc.) may be executed in seconds or milliseconds. Hundreds or thousands of Containers may be run on a single host at the same time (Ali et al., 2016), thus ensuring that the failure of one task does not cause disruption of the entire process: new Containers may be initialized rapidly to continue the task until the completion of the entire process, thus improving overall efficiency.

In recent years, several online tools and software suites have been developed for pan-genome analysis, including Roary, PGPA, SplitMEM, PanGP, and PanTools. However, generally, the installation of these pipelines with many dependencies, but a single function, is complex and challenging. Therefore, limiting researchers’ ability directly focus on their analyses of interest (Table 4). Although the BGDMdocker workflow includes several tools, installing and running the software is quite simple. Biologists may automatically install, configure, and test the scripts, making these processes faster and the results repeatable.

Table 4. Function of BGDMdocker workflow compared with other pan-genome tools

| Tools | Automatic installation | Cross platform | Result visualization | Genome annotation | Gene Cluster mining |
|-------|------------------------|----------------|---------------------|------------------|--------------------|
| Roary | ×                      | ×              | ×                   | ✓                | ×                  |
| PGPA  | ×                      | ×              | ×                   | ×                | ×                  |
CONCLUSIONS

Here, we present a BGDMdocker workflow to achieve bacterial and viral genome annotation, pan-genome analysis, mining of biosynthetic gene clusters, and visualization of results on a local host or online. This allows researchers to browse information for every gene, including duplication, diversity, indel events, and sequence alignments, as well as for biosynthetic gene clusters, including structure, type, description, detailed annotation, and predicted core structure of the target compounds. These tools and their installation commands and dependencies were all written in a Dockerfile. We used this Dockerfile to build a Docker Image and run Container for analyzing the pan-genome of 44 B. amyloliquefaciens strains retrieved from a public database. The pan-genome included a total of 172,388 genes and 2,306 core gene clusters. The visualization of the pan-genomic data included alignments, phylogenetic trees with mutations within each cluster mapped to the branches of the tree, and inference of gene losses and gains on the core-genome phylogeny for each gene cluster. In addition, 997 known (MIBiG database) and 553 unknown (antiSMASH-predicted clusters and Pfam database) genes in biosynthetic gene clusters and orthologous groups were identified in all strains. The BGDMdocker workflow for the analysis and visualization of pan-genomes and biosynthetic gene clusters may be fully reused immediately across different computing platforms (Linux, Windows, Mac, and cloud-based systems), with flexible and rapid deployment of integrated software packages across various platforms. This workflow may also be used for other pan-genome analyses and visualization of other species. Additionally, the visual display of data provided in this study may be completely duplicated. All resulting data and relevant tools and files may be downloaded from our website with no registration required.

ACKNOWLEDGEMENTS

We thank Yilei Wu, Chao Chen, Wei Ding, and the reviewers for usability testing and valuable suggestions.

ADDITIONAL INFORMATION AND DECLARATIONS
REFERENCES

Ali AA, El-Kalioby M, Abouelhoda M. 2016. The case for Docker in multicloud enabled bioinformatics applications. In International Conference on Bioinformatics and Biomedical Engineering. Springer, Heidelberg, Vol. 9656 of the series Lecture Notes in Computer Science, pp. 587–601.

Aranguren ME, Wilkinson MD. 2015. Enhanced reproducibility of SADI web service workflows with Galaxy and Docker. GigaScience, 4, 59, doi:10.1186/s13742-015-0092-3.

Belmann P, Droge J, Bremges A, McHardy AC, Sczyrba A, Barton MD. 2015. Bioboxes: standardised Containers for interchangeable bioinformatics software. Gigascience, 4(1), 47, doi:10.1186/s13742-015-0087-0.

Boettiger CD. 2015. An introduction to Docker for reproducible research, with examples from the R environment. ACM SIGOPS Operating Systems Review, Special Issue on Repeatability and Sharing of Experimental Artifacts, 49(1), 71–79, doi:10.1145/2723872.2723882.

Docker Inc, 2017. https://docs.docker.com/

Docker Inc, 2017. https://www.docker.com/whatisdocker

Docker Inc, 2017. https://hub.docker.com/

Ding W, Baumdicker F, Neher RA. 2016. PanX: pan-genome analysis and exploration. bioRxiv, 072082, doi:10.1101/072082.

Folarin AA, Dobson RJ, Newhouse SJ. 2015. NGSeasy: a next generation sequencing pipeline in Docker Containers. F1000Research, 4, 997, doi: 10.12688/f1000research.7104.1.

Hosny A, Vera-Licona P, Laubenbacher R, Favre T. 2016. AlgoRun, a Docker-based packaging system for platform-agnostic implemented algorithms. Bioinformatics, 32(15), 2396–2398.

Nam HS, Yang HJ, Oh BJ, Anderson AJ, Kim YC. 2016. Biological control potential of Bacillus amyloliquefaciens KB3 isolated from the feces of Allomyrina dichotoma larvae. Plant Pathology Journal, 32(3), 273–280, doi:10.5423/PPJ.NT.12.2015.0274.

The Perl Community(Maintained by), 2017. https://hub.docker.com/_/perl/
Hilmar Lapp (Maintained by), 2017. https://hub.docker.com/r/bioperl/bioperl/

The Docker Community (Maintained by), 2017. https://hub.docker.com/_/python/

Tiago Antao (Maintained by), 2017. https://hub.docker.com/r/biopython/biopython/

The Rocker Community (Maintained by), 2017. https://hub.docker.com/_/r-base/

The Bioconductor Core Team (Maintained by), 2017. https://hub.docker.com/r/bioconductor/

release_base/

Björn A. Grüning BA, 2017. https://hub.docker.com/r/bgruening/galaxy-stable/~/dockerfile/

Weber T, Kai, B, Duddela S., Krug D, Kim HU, Brucoller R, Lee SY, Fischbach MA, Müller R, Wohlleben W, Breitling R, Takano E, Medema MH. 2015. AntiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, 43(W1), W237–W243.

Seemann T. 2014. Prokka: rapid prokaryotic genome annotation. Bioinformatics, 30(14): 2068–2069, doi:10.1093/bioinformatics/btu153.

Tommaso PD, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. 2015. The impact of Docker Containers on the performance of genomic pipelines. PeerJ, 3, e1273, doi:10.7717/peerj.1273.