BGDMdocker: a Docker workflow for analysis and visualization pan-genome and biosynthetic gene clusters of bacterial

Gong Cheng 1, 2, Quan Lu 3, Zongshan Zhou 1,*, Ling Ma 4, Guocai Zhang 4, Yilei WU 5,
Chao Chen 6

1 Research Institute of Pomology, Chinese Academy of Agricultural Sciences, Xingcheng, Liaoning Province 125100, China, 2 Forest Protection Research Institute of Heilongjiang Province, Harbin Heilongjiang Province 150040, China, 3 Research Institute of Forest Ecology, Environment and Protection, Chinese Academy of Forestry, Beijing 100091, China, 4 Northeast Forestry University, Harbin Heilongjiang Province 150040, 5 Beijing Hygenoimes CO., LTD, Beijing 102209, China, 6 Tang Tang Tianxia Biomedical Information Technology CO., LTD, Beijing 100094, China.

ABSTRACT

Motivation: At present Docker technology has received increasing level of attention throughout the bioinformatics community. However, its implementation details have not yet been mastered by most biologists and applied widely in biological researches. In order to popularizing this technology in the bioinformatics and sufficiently use plenty of public resources of bioinformatics tools (Dockerfile and image of scommunity, officially and privately) in Docker Hub Registry and other Docker sources based on Docker, we introduced full and accurate instance of a bioinformatics workflow based on Docker to analyse and visualize pan-genome and biosynthetic gene clusters of a bacteria in this article, provided the solutions for mining bioinformatics big data from various public biology databases. You could be guided step-by-step through the workflow process from docker file to build up your own images and run an container fast creating an workflow.

Results: We presented a BGDMdocker (bacterial genome data mining docker-based) workflow based on docker. The workflow consists of three integrated toolkits, Prokka v1.11, panX, and antiSMASH3.0. The dependencies were all written in Dockerfile, to build docker image and run container for analysing pan-genome of total 44 Bacillus amyloliquefaciens strains, which were retrieved from public ??? database. The pan-genome totally includes 172,432 gene, 2,306 Core gene cluster. The visualized pan-genomic data such as alignment, phylogenetic trees, maps mutations within that cluster to the branches of the tree, infers loss and gain of genes on the core-genome phylogeny for each gene cluster were presented. Besides, 997 known (MIBiG database) and 553 unknown (antiSMASH-predicted clusters and Pfam database ) genes of biosynthesis gene clusters types and orthologous groups were mined in all strains. This workflow could also be used for other species pan-genome analysis and visualization. The display of visual data can completely duplicated as well as done in this paper. All result data and relevant tools and files can be downloaded from our website with no need to register. The pan-genome and biosynthetic gene clusters analysis and visualization can be fully
reusable immediately in different computing platforms (Linux, Windows, Mac and deployed in the cloud), achieved cross platform deployment flexibility, rapid development integrated software package.

**Availability and implementation:** BGDMdocker is available at http://42.96.173.25/bapgd/ and the source code under GPL license is available at https://github.com/cgwxy/debian_prokka_panx_antismash biodocker.

**Contact:** chenggongwyx@foxmail.com

**Supplementary information:** Supplementary data are available at biorxiv online.

**Keywords:** Docker, pan-genome, biosynthesis gene clusters, Bacteria, *Bacillus amyloliquefaciens*,

1 INTRODUCTION

Bioinformatics academic free softwares generally have the shortcomings of installing and configuration difficulty, large dependencies, limit size of the data uploading to online servers and so on. Therefore a lot of excellent softwares could not be fully used by biologists. Sharing bioinformatics tools with Docker can reproducibly and conveniently build all kinds of bioinformatics workflows. It gives programmers, development teams and biologists of bioinformatics the common toolbox they need to take advantage of the use, and building, shipping and running any app, anywhere your distributed applications.

Thus, Dockers technology is very suitable for the application in the field of bioinformatics, because of its advantages and characteristics that allow applications to run in an isolated, self-contained package that can be efficiently distributed and executed in a portable manner across a wide range of computing platforms. ([Belmann et al., 2015; Hosny et al., 2016; Aranguren and Wilkinson, 2015](#)). At present, there are many bioinformatics tools based on docker are developed and published, such as perl and bioperl (Martini, 2016), python and biopython ([Moghedrin et al., 2016](#)), R and Bioconductor ([Eddelbuettel et al., 2016](#)), contribute their Official Docker image; famous Galaxy also contribute docker galaxy (Björn, 2016). It’s reasonable to predict that docker will become more and more extensive in the field of bioinformatics. We used docker technology to rapidly construct a pan-genome analysis process, which can be set up in the Linux, windows and MAC systems (64-bit), can also be deployed in the cloud such as Amazon EC2 or other cloud providers. The workflow will greatly facilitate biologists to apply in bioinformatics fields. Docker containers have only a minor impact on the performance of common genomic pipelines ([Tommaso et al., 2015](#)).

*Bacillus amyloliquefaciens* has been researched and explored extensively and intensively with the abilities to inhibit fungi and bacteria ([Nam et al., 2016](#)). [Based on the Docker we quickly run an container (on ubuntu 16.04 and win10 host) of analysis and reveal the pan-genome and biosynthetic gene clusters basic features of 44 *B. amyloliquefaciens* strains, and achieve visualizations result. The analytical workflow consists of three toolkits: Prokka v1.11 ([Seemann, 2014](#)) prokaryotic genome annotation, panX ([Ding et al., 2016](#)) anlyasis and visualization of pan genome and antiSMASH3.0 ([Weber et al., 2015](#)) biosynthesis gene clusters, which we wrote Dockerfile of BGDMdocker with all of the application and its dependencies, (we also wrote standalone Dockerfile of Prokka, panX and antiMASH in order to meet different application
requirements of user. recommended in this way, see Supplement(以上也看不懂) Here we described in details how to build workflow and manipulate the analysis.

2 METHODS AND IMPLEMENTATION DETAILS

2.1 Installing docker

Docker-engine was installed on Ubuntu Xenial 16.04 (LTS) and win10 Enterprise operating system. Docker requires a 64-bit installation regardless of your Ubuntu version. Additionally, the kernel must be 3.10 at minimum. More operating systems installation see here.

Installation latest docker (docker-engine 1.12.5-0~ubuntu-xenial) on Ubuntu Xenial 16.04 (LTS) (Install Docker on Ubuntu, 2016):

To copy the following commands meant quickly & easily installing via latest docker-engine (ubuntu, debian, raspbian, fedora, centos, redhat, suse and oracle linux et al., all applicable):

```
$ curl -sSL https://get.docker.com/ | bash -x
```
or:
```
$ wget -qO- https://get.docker.com/ | bash -x
```

Type the following commands at your shell prompt, If output for docker version it means your installation is successful:

```
$ docker version
```

Installation latest docker on Windows 10 Enterprise(Install Docker on Windows,2016):

The current version of Docker for Windows runs on 64 bit Windows 10 Pro, Enterprise and Education.

Step 1. Download Docker InstallDocker.msi (https://download.docker.com/win/stable/InstallDocker.msi)

Step 2. Install Docker(1.12.0-rc2) for Windows

Double-click InstallDocker.msi to run the installer. Follow the install wizard to accept the license, authorize the installer, and proceed with the install.

Type the following commands at your shell prompt (cmd.exe or PowerShell), will output for docker version it means your installation is successful.

```
$ docker version
```

2.2 build images of workflow and run container login interaction patterns

Dockerfile of BGDMdocker workflow have been submitted to Github. On your host type the following commands line will login a BGDMdocker container:

```
$ git clone https://github.com/cgwyx/debian_prokka_panx_antismash_biodocker.git
```

Or download “.zip” file.

```
$ unzip debian_prokka_panx_antismash_biodocker-master.zip
```

Build images of workflow:

```
$ cd ./debian_prokka_panx_antismash_biodocker/prokka_panx_antismash_dockerfile
$ sudo docker build -t BGDMdocker:latest .
```

Run an container from image of BGDMdocker:atest:

```
$ sudo docker run -it --rm --v home:home -p 8000:8000 --name=BGDMdocker BGDMdocker:latest
```

“-v home:home” parameter, Docker will mount the local folder /home into the
Container under /home, Store all your data in one directory of home of host operating system, then you may access those directory of home inside of container.

Check out local images and container:
$ sudo docker images
$ sudo docker ps -a

2.2.1 Run Prokka annotation genome in container in interaction patterns (if you have own sequences of genome need this step generates.gbff annotation files):

Check out help documentation and command parameters:
$ prokka --help

Copy the following commands run the analysis instances of annotation genome of Ba_xx strains in command-line interface of container (boldface please enter your data):

$ prokka --kingdom Bacteria --gcode 11 --genus Bacillus\n  --species Amyloliquefaciens\n  --strain Ba_xx --locustag Ba_xx --prefix Ba_xx --rfam\n  --rawproduct --outdir /home/manager/PRJNA291327\n/home/manager/Ba_xx.fasta

“Ba_xx.fasta” is sequence of genome, “PRJNA291327” is output folder of result.

2.2.2 Run panX anlysis pan-genome in container in interaction patterns:

panX starts with a set of annotated sequences files.gbff(.gbk) (e.g. NCBI RefSeq or GenBank) of a bacterial species genomes, If using own GenBank files(or you have download these files), step 02 can be skipped. The detailed parameters see here.

Download all “*genomic.gbff.gz”of specified species from RefSeq or GenBank Database(boldface please enter your species):

Installion scrip on your host
$ wget ftp://ftp.ncbi.nlm.nih.gov/entrez/entrezdirect/versions/current/edirect.zip
$ unzip -u -q edirect.zip
$ export PATH=$PATH:$HOME/edirect
$ ./edirect/setup.sh

Download “*genomic.gbff.gz” of all strains of Bacillus amyloliquefaciens from GenBank

$ esearch -db assembly -query "Bacillus amyloliquefaciens [ORGN]" | efetch -format docsum | xtract -pattern "DocumentSummary" -element FtpPath_GenBank | sed 's/$/\n*genomic.gbff.gz/\nxargs wget -c -nd;sleep 3s;

or

Download “*genomic.gbff.gz” of all strains of Bacillus amyloliquefaciens from RefSeq

$ esearch -db assembly -query "Bacillus amyloliquefaciens[ORGN]" | efetch -format docsum | xtract -pattern "DocumentSummary" -element FtpPath_RefSeq | sed 's/$/\n*genomic.gbff.gz/\nxargs wget -c -nd;sleep 3s;

Check out help documentation and command parameters of panX:
$ sh run.sh

Copy the following commands run the analysis instances of pan-genome of 44 B. amyloliquefaciens strains in command-line interface of container (boldface please enter your data):
 dollar cd /pan-genome-analysis

 dollar python ./scripts/run-pipeline.py -fn ./data/ B_amy -sl B_amy-RefSeq.txt

- st 1 3 4 5 6 7 8 9 10 11

*.gbff (GenBank files) and B_amy-RefSeq.txt (accession list for strains) should reside in “./data/B_amy” folder, result will also output “./data/B_amy” folder.

Visualization pan-genome of 44 B. amyloliquefaciens strains (run in container):

 dollar python link-to-server.py B_amy

 dollar add-new-pages-repo.sh B_amy

 dollar gulp

http://localhost:8000/B_amy

Create a new image from a container’s changes of we have process data in order to save results in image (run in host):

dollar sudo docker commit <ID of container> <name of new images>

2.2.3 Run antiSAMSH search gene clusters of every strain in container in interaction patterns:

 Check out help documentation and command parameters:

dollar run_antismash.py --help

 Copy the following commands run the analysis instances of Biosynthesis gene clusters of Y2(Y2.gbff) strains in command-line interface of container (boldface please enter your data):

dollar cd /antismash-3.0.5/antismash/

 dollar run_antismash.py /home/ manager/input/Y2.gbff

--outputfolder /home/manager/output/Y2_out

--clusterblast /generic_modules/clusterblast

--pfamdir /generic_modules/fullhmmer --input-type nucl --knownclusterblast

--clusterblast --subclusterblast --inclusive --full-hmmer --smcogs --verbose --asn

--borderpredict

*.gbff (GenBank files) reside in “input” folder, “Y2_out” is output folder of result.

3 RESULTS AND CONCLUSIONS

3.1 Fast and repeatable buildup of BGDMdocker workflow based on Docker across computing platform

 Based on docker technology, using Dockerfile script file, build images and run a containers in seconds or milliseconds on Linux and Windows (also can be deployed in Mac, cloud such as Amazon EC2 or other cloud providers). Dockerfile is just a small plain text file that can be easily stored and shared. The user does not have to deal with installing and configuring.

 In this instance, all image based on Debian 8.0 (Jessie), establishment of a novel Docker-based bioinformatics platform for the study of microbes genomes. The use of workflow containers with standardised interfaces has the potential of making the work of biologists easier by creating simple to use, inter-changeable tools to excavate the biological meaning contained in data from biological experiments and Studied on the biological sequence data obtained from biological database through internet. Depending on the sequence obtained, we focused on the information mining in the sequence. We have uploaded this dockerfile to GitHub for the use of relevant scientific
3.2 Result of pan genomes of \textit{B. amyloliquefaciens}

BGDMdocker workflow analyzed pan genomes of \textit{B. amyloliquefaciens}. In order to explore high dimensional data, we built a website for interactive exploration of the pan-genome and biosynthetic gene clusters. The visualization allows rapid filtering of and searching for genes. For each gene cluster, panX displays an alignment, a phylogenetic tree, maps mutations within that cluster to the branches of the tree, and infers loss and gain of genes on the core-genome phylogeny. Here we only lists summary statistics results of pan genomes (Table 1), phylogenetic relationship of 44 \textit{B. amyloliquefaciens} strains (Figure 1). All detail data can be visualized and downloaded without registration.

| Strains | Gene numbers in pan-genome of \textit{B. amyloliquefaciens} (172,432 Total gene; 2306 Core gene cluster) | Gene of strain genomes |
|---------|-------------------------------------------------------------------------------------------------|------------------------|
| CYHL01000001     | CYL5 | 3856 | 2310 | 1546 | 57 | 3870 | 3863 |
| CYHP01000001     | CYL8 | 3994 | 2311 | 1683 | 118 | 4016 | 4006 |
| NC_014551        | DSM7 | 3935 | 2307 | 1628 | 21 | 4030 | 3811 |
| NC_017188        | TA208 | 3935 | 2307 | 1628 | 1 | 3974 | 3847 |
| NC_017190        | LL3  | 3981 | 2308 | 1673 | 19 | 4037 | 3887 |
| NC_017191        | XH7  | 3942 | 2307 | 1635 | 6 | 3983 | 3846 |
| NC_017912        | Y2   | 4099 | 2310 | 1789 | 46 | 4148 | 3983 |
| NC_020272        | IT-45 | 3803 | 2310 | 1493 | 4 | 3832 | 3678 |
| NC_022653        | CC178 | 3754 | 2310 | 1444 | 19 | 3795 | 3641 |
| NC_023073        | LFB112 | 3761 | 2308 | 1453 | 19 | 3801 | 3637 |
| NZ_AUNG01000001  | Lx-11 | 3700 | 2309 | 1391 | 5 | 3742 | 3619 |
| NZ_AUWK01000001  | HB-26 | 3797 | 2311 | 1486 | 30 | 3842 | 3714 |
| NZ_AVQH01000001  | EGD-AQ141 | 4079 | 2311 | 1768 | 54 | 4121 | 3995 |
| NZ_AWQY01000001  | UASWS BA1 | 3794 | 2309 | 1485 | 8 | 3806 | 3681 |
| NZ_CP006058      | UMA6639 | 3825 | 2311 | 1514 | 20 | 3879 | 3716 |
| NZ_CP006960      | UMA6614 | 3804 | 2311 | 1493 | 13 | 3850 | 3695 |
| NZ_CP007242      | KH912 | 3775 | 2310 | 1465 | 19 | 3816 | 3658 |
| NZ_CP010556      | L-H15 | 3724 | 2309 | 1415 | 6 | 3769 | 3615 |
| NZ_CP011278      | L-S60 | 3728 | 2310 | 1418 | 19 | 3773 | 3611 |
| NZ_CP013727      | MBE1283 | 3794 | 2314 | 1480 | 24 | 3856 | 3681 |
| NZ_CP014700      | S499 | 3776 | 2310 | 1466 | 5 | 3819 | 3671 |
| NZ_CP014783      | B15  | 3820 | 2315 | 1505 | 13 | 3875 | 3704 |
| NZ_CP016913      | RD7-7 | 3597 | 2308 | 1289 | 39 | 3656 | 3483 |
| NZ_DF36901       | CMW1 | 3771 | 2311 | 1460 | 128 | 3901 | 3706 |
| NZ_ICOC01000001  | EHL11 | 3733 | 2308 | 1425 | 167 | 3598 | 3604 |
| NZ_JE4G01000001  | B1895 | 3824 | 2306 | 1518 | 167 | 4026 | 3623 |
| NZ_JQNZ01000001  | X1   | 3724 | 2309 | 1415 | 3 | 3766 | 3619 |
| NZ_JT1G01000001  | JCC3M | 3888 | 2309 | 1579 | 121 | 3952 | 3796 |
| NZ_JXAT01000001  | LPL-K103 | 3709 | 2309 | 1400 | 15 | 3743 | 3637 |
| NZ_JX2D01000001  | 12B2 | 8166 | 2354 | 5812 | 4040 | 8194 | 7906 |
| NZ_KB2P01000001  | DC-12 | 3910 | 2311 | 1599 | 50 | 3894 | 3842 |
| NZ_KN723307      | TF281 | 3640 | 2312 | 1328 | 5 | 3782 | 3571 |
| NZ_LGYP01000001  | 629  | 3536 | 2313 | 1223 | 11 | 3785 | 3427 |
| NZ_LJAU01000001  | BS06 | 4042 | 2312 | 1730 | 46 | 4074 | 3969 |
| NZ_LJD1010000020 | XK-4-1 | 3799 | 2310 | 1489 | 14 | 3821 | 3701 |
| NZ_LMG01000001   | RHNK22 | 3781 | 2309 | 1472 | 37 | 3837 | 3698 |
| NZ_LMAT01000001  | Jxnuwx-1 | 3930 | 2309 | 1621 | 246 | 4008 | 3870 |
| NZ_LMUC010000016 | H57  | 3816 | 2310 | 1506 | 42 | 3839 | 3732 |
| NZ_LPUP010000011 | 11B91 | 3790 | 2311 | 1479 | 49 | 3892 | 3702 |
| NZ_LQQW01000001  | M49  | 3694 | 2311 | 1383 | 21 | 3741 | 3617 |
| NZ_LQY01000001   | B4100 | 3771 | 2307 | 1464 | 49 | 3847 | 3713 |
| NZ_LQYP01000001  | B425 | 3921 | 2310 | 1611 | 39 | 4034 | 3844 |
| NZ_LYUG01000001  | SRCM101266 | 3724 | 2306 | 1418 | 15 | 3781 | 3628 |
| NZ_LZZ01000001   | SRCM101294 | 3946 | 2308 | 1638 | 175 | 3982 | 3850 |

Genome sequences of \textit{44 B. amyloliquefaciens} strains (downloaded from GenBank RefSeq database) used in this study, strains name, accession numbers, and gene numbers of pan-genome and genome summary statistics.

“Acc gene” is accessory gene (dispensable gene), “Uni gene” is unique gene (strain-specific gene), “All-genes” is gene of *.gbff files recoder, incloud Pseudo Genes, “Total genes” is involved in the pan genome anal-
ysis, not Pseudo Genes.

Fig. 1. Phylogenetic tree of 44 \( B. \text{amyloliquefaciens} \) strains. The tree was constructed using all of the genes shared between all 44 strains (2306 core genes).

### 3.3 Result of analysis of biosynthetic gene clusters

BGDMdocker workflow identification and analysis results of biosynthetic gene clusters of genomes of 44 \( B. \text{amyloliquefaciens} \) strains have been uploaded in our website, no need to register and visit and download all detail data.

In this paper we only give a brief summary statistics of biosynthetic gene clusters of all 44 strains (Table 2), and shows an instance of biosynthetic gene clusters type and number result of Y2 strain, there are total 31 gene clusters in Genome of Y2 strain. 21 gene clusters show similarity known cluster MIBiG, like this Surfactin, Mersacidin, Fengycin, and so on, other 10 gene clusters are unknown (Table 3).

| Accession          | Strains | Biosynthesis geneclusters | Genome of strains | Host |
|--------------------|---------|----------------------------|-------------------|------|
|                    |         | Total | known | Unknown | Type | Size(Mb) | Gene | Protein |
| CYHL010000001      | JRS5    | 38    | 27    | 11      | 12   | 4.03148 | 3870 | 3863     |
| CYHP010000001      | JRS8    | 42    | 26    | 16      | 11   | 4.0909  | 4016 | 4006     |
| NC_014551          | DSM7    | 31    | 18    | 13      | 9    | 3.9802  | 4030 | 3811     |
| NC_017188          | TA208   | 29    | 17    | 12      | 10   | 3.93751 | 3974 | 3847     |
| NC_017190          | LL3     | 29    | 17    | 12      | 9    | 4.00199 | 4037 | 3887     |
| NC_017191          | XH7     | 29    | 17    | 12      | 10   | 3.9362  | 3983 | 3846     |
| NC_017912          | Y2      | 31    | 21    | 10      | 11   | 4.23862 | 4148 | 3983     |
| NC_020272          | IT-45   | 32    | 19    | 13      | 11   | 3.93687 | 3832 | 3678     |
| NC_022653          | CC178   | 35    | 19    | 16      | 9    | 3.91683 | 3795 | 3641     |
| NC_023073          | LFBI12  | 35    | 20    | 15      | 10   | 3.94275 | 3801 | 3637     |
| NZ_AUNG01000001    | Lx-11   | 37    | 25    | 12      | 11   | 3.88689 | 3742 | 3619     |
| NZ_AUWK01000001    | HB-26   | 45    | 30    | 15      | 9    | 3.98936 | 3842 | 3714     |
| NZ_AVHQ010000001   | EGDAQ141| 36    | 26    | 10      | 12   | 4.22259 | 4121 | 3995     |
| NZ_AWQY010000001   | UASWS BA1| 37   | 25    | 12      | 11   | 3.94409 | 3806 | 3681     |
| NZ_CP006058        | UMAF6639| 35    | 21    | 14      | 10   | 4.03464 | 3879 | 3716     |
| NZ_CP006960        | UMAF6614| 32    | 20    | 12      | 10   | 4.00514 | 3850 | 3695     |
the MIBiG (Minimum Information about a Biosynthetic Gene cluster).

Clusters of unknown types, detected by Cluster Finder are further categorized into "putative ('Cf putative') biosynthetic types.

Cluster Finder now allows using this probabilistic algorithm to detect putative gene clusters of Y2 strain from the MIBiG (Minimum Information about a Biosynthetic Gene cluster). "Unknown" of Biosynthesis gene clusters detected by Cluster Finder are further categorized into putative gene clusters of unknown types, "-" of host is unrecorded.

**Table 3. Biosynthetic gene clusters of Y2 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 | Pectavinc_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 |
| Cluster 7 | Cf putative | - | - |
| Cluster 8 | Others | - | - |
| Cluster 9 | Cf fatty_acid | - | - |
| Cluster 10 | Cf putative | - | - |
| Cluster 11 | Terpene | - | - |
| Cluster 12 | Cf fatty_acid | - | - |
| Cluster 13 | Cf putative | - | - |
| Cluster 14 | Cf fatty_acid | - | - |
| Cluster 15 | Transatpks | Macrolactin_biosynthetic_gene_cluster (90% of genes show similarity) | BGC0000181_c1 |
| Cluster 16 | Nrps-Transatpks | Bacillaene_biosynthetic_gene_cluster (85% of genes show similarity) | BGC00001089_c1 |
| Cluster 17 | Nrps-Transatpks | Fengycin_biosynthetic_gene_cluster (93% of genes show similarity) | BGC00001095_c1 |
| Cluster 18 | Terpene | - | - |
| Cluster 19 | Cf saccharide-T3pks | Difftacidin_biosynthetic_gene_cluster (100% of genes show similarity) | BGC0000176_c1 |
| Cluster 20 | Transatpks | - | - |
| Cluster 21 | Cf putative | - | - |
| Cluster 22 | Nrps-Bacteriocin | Bacillobactin_biosynthetic_gene_cluster (100% of genes show similarity) | BGC0000309_c1 |
| Cluster 23 | Cf saccharide | - | - |
| Cluster 24 | Nrps | - | - |
Cluster 25: Cf_saccharide
- Tethylchuronic_acid_biosynthetic_gene_cluster (100% of genes show similarity)
- BGC0000868_c1

Cluster 26: Cf_putative
- Bacilysin_biosynthetic_gene_cluster
- BGC0001184_c1

Cluster 27: Cf_saccharide
- Mersacidin_biosynthetic_gene_cluster (90% of genes show similarity)
- BGC0000527_c1

Cluster 28: Cf_putative
- Bacilysin_biosynthetic_gene_cluster
- BGC0001184_c1

Cluster 29: Lantipeptide
- Mersacidin_biosynthetic_gene_cluster (90% of genes show similarity)
- BGC0000527_c1

Cluster 30: Cf_saccharide
- Mersacidin_biosynthetic_gene_cluster (90% of genes show similarity)
- BGC0000527_c1

Cluster 31: Cf_putative
- Mersacidin_biosynthetic_gene_cluster (90% of genes show similarity)
- BGC0000527_c1

“Cf putative” is putative biosynthetic types (unknown types) detected by Cluster Finder are further categorized, known types from the MIBiG (Minimum Information about a Biosynthetic Gene cluster, http://mibig.secondarymetabolites.org).

4 DISCUSSIONS

Docker advantages as follow:

1. Dockerfile convenient for deployed and shared; make it is easy for other users to customize the image by editing the Dockerfile directly. It very unlikely Makefile and other installation that the resulting build will differ when being built on different machines (Boettiger, 2014). Through Dockerfile, can maintain and update related adjustment, further more rapid rollback in the event of failure of System, according to the demand to control version and build the best application environment.

2. Portability, Modular reuse; Most bioinformatics tools is written in different languages, require different operating environment configuration and cross platform, Docker provide the same functions and services in different environments without additional configuration (Folarin et al., 2015), so it is possible that the results of repetition and reuse tools. By constructing pipelines with different tools, bioinformatics can automatically and effectively analyze scientific problem they are concern at of biological.

3. Application solution, efficiency, exible: Docker can be run independently containers of every applications, and Management operations (start, stop, boot, etc.) of Containers in seconds or milliseconds; we may run more than hundreds of containers on a single host (Ali et al., 2016). This ensure that the failure of one task does not cause disruption of the entire process, and quickly start a new container to continue to perform this task until the completion of the whole process, improve the overall efficiency.

Docker limitations as follow:

1. Docker is limited to 64-bit host machines, then making it can not to run on 32-bit older hardware.

2. On Mac and Windows OS, Docker must run boot2docker tool in a fully virtualized environment.

References
Belmann, P et al. (2015) Bioboxes: standardised containers for interchangeable bioinformatics software. GigaScience, 4, 47 DOI 10.1186/s13742-015-0087-0

Abdelrahman, H et al. (2016) AlgoRun: a Docker-based packaging system for platform-agnostic implemented algorithms. Bioinformatics, 32, 2396-2398.

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

Martini, P. (2016). https://hub.docker.com/_/perl/
Robenolt, M. (2016). https://hub.docker.com/_/python/
Boettiger, C and Edelbuettel, D. (2016). https://hub.docker.com/_/r-base/
Björn, A. and Grüning, B. (2016). https://hub.docker.com/r/bgruening/galaxy-stable/
Tommaso, D et al. (2015). The impact of Docker containers on the performance of genomic pipelines. PeerJ, 3, e1273. http://doi.org/10.7717/peerj.1273
Nam, H et al. (2016) Biological Control Potential of Bacillus amyloliquefaciens KB3 Isolated from the Feces of Allomyrina dichotoma Larvae. The Plant Pathology Journal. 32(3): 273-280. doi:10.5423/PPJ.NT.12.2015.0274.
Torsten S. (2014) Prokka: rapid prokaryotic genome annotation. Bioinformatics 30 (14): 2068-2069 doi:10.1093/bioinformatics/btu153
Wei, D et al. (2016) panX: pan-genome analysis and exploration. bioRxiv 072082; doi: http://dx.doi.org/10.1101/072082
Weber, T et al. (2015) antiSMASH 3.0- a comprehensive resource for the genome mining of biosynthetic gene clusters. [J]. Nucleic Acids Research, 43(W1):237-43.
Docker Inc. (2016) Install Docker on Windows. https://docs.docker.com/docker-for-windows/
Docker Inc. (2016) Install Docker on Ubuntu. https://docs.docker.com/engine/installation/linux/ubuntulinux/
Boettiger, C. (2014) 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.
Folarin, A et al. (2015) NGSeasy: a next generation sequencing pipeline in Docker containers [version 1; referees: awaiting peer review], F1000Research 4(ISCB Comm J):997 (doi: 10.12688/f1000research.7104.1)
Ali, A.et al. (2016). The Case for Docker in Multicloud Enabled Bioinformatics Applications. Bioinformatics and Biomedical Engineering. Springer International Publishing. Volume 9656,587-601