A peer-reviewed version of this preprint was published in PeerJ on 30 November 2017.

View the peer-reviewed version (peerj.com/articles/3948), which is the preferred citable publication unless you specifically need to cite this preprint.

Cheng G, Lu Q, Ma L, Zhang G, Xu L, Zhou Z. 2017. BGDMdocker: a Docker workflow for data mining and visualization of bacterial pan-genomes and biosynthetic gene clusters. PeerJ 5:e3948 https://doi.org/10.7717/peerj.3948

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

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

[1] Protection Research Center of Pomology, Research Institute of Pomology, Chinese Academy of Agricultural Sciences, Xingcheng, Liaoning, China

[2] Research Institute of Forest Ecology, Environment and Protection, Chinese Academy of Forestry, Beijing, China

[3] School of Forestry, Northeast Agricultural University, Harbin, China

[4] 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, and thus its application in biological research has been limited. In order to popularize this technology in the field of bioinformatics and promote the use of publicly available bioinformatics tools, such as Dockerfiles and Images from communities, governmental, and private owners in Docker Hub Registry and other Docker-based resources, we introduce here a complete and accurate bioinformatics workflow based on Docker to analyze and visualize pan-genomes and biosynthetic gene clusters of bacteria. This provides a new solution for bioinformatics mining of big data from various public biological databases. This step-by-step guide creates an integrative workflow through a Dockerfile to allow researchers to build their own Image and run Container easily.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3040v1 | CC BY 4.0 Open Access | rec: 22 Jun 2017, publ: 22 Jun 2017

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

Gong Cheng[1,2], Quan Lu[3,*], Ling Ma[4,*], Guocai Zhang[4], Liang Xu[5] and Zongshan Zhou[1,*]

[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, China, [5]Institute of Food Science and Technology, Chinese Academy of Agricultural Sciences, Beijing 100193, China

ABSTRACT

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

Subjects Bioinformatics, Computational Biology

Keywords Docker, pan-genome, biosynthetic gene clusters, *Bacillus amyloliquefaciens*

INTRODUCTION

Bioinformatics academic software programs generally exhibit shortcomings such as installation and configuration difficulties, large dependencies, and limitations on the amount of data that can be uploaded to online servers. Therefore, several excellent software programs cannot be fully used by biologists. Bioinformatics tools can be merged with Docker technology, however, to build reproducible and convenient types of workflows. Docker provides release programmers, development teams, and bioinformaticians with a common toolbox that can be used to take full advantage of bioinformatics tools, helping to build, ship, and run any app and 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 that can 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, many bioinformatics tools based on Docker have been developed and published, such as perl and bioperl (Martini, 2016), python and biopython (Matt *et al.*, 2016), and R and Bioconductor (Boettiger *et al.*, 2016). These have all contributed to official Docker Images; the famous Galaxy program has also contributed to Docker Galaxy (Björn & Grüning, 2016). It is reasonable to assume that Docker will become more and more extensively implemented in the field of bioinformatics.

Here, we used Docker technology to rapidly construct a pan-genome analysis process that can be used in the Linux, Windows, or Mac environments (64-bit) and can also be deployed as a cloud-based system, such as with

Amazon EC2 or other cloud providers. This workflow will 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 researched and explored extensively 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 quickly ran 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 and to visualize the results. The analytical workflow consists of three toolkits: Prokka v1.11 (Seemann, 2014), panX (Ding *et al*., 2016), and antiSMASH3.0 (Weber *et al*., 2015), tools 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), so that the workflow could be implemented online with a single run. We also wrote three standalone Dockerfiles for Prokka, panX, and antiSMASH in order to meet 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. Here, we describe in detail how to build the workflow and conduct the analysis.

MATERIALS & METHODS

Installation of latest Docker on your host:

1. Copy the following commands for quickly & easily installing via latest docker-engine (Ubuntu, Debian, Raspbian, Fedora, Centos, Redhat, Suse, 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 this outputs the Docker version, your installation was successful:

```
$ 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.

```
$ 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:

```
$ git clone https://github.com/cgwyx/debian_prokka_panx_antismash_
biodocker.git
```

or: download "debian_prokka_panx_antismash_biodocker-master.zip" file

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

2. Build workflow Images:

```
$                                                                    cd
```

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

3. Run a Container from the BGDMdocker Image:

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

If you 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 with 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. It can 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 can be easily stored and shared. Therefore, the user is not required to install and configure the programs.

In this instance, 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 standardised tools to extract biological information from their own experiments and from online sequence databases. Researchers can 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.

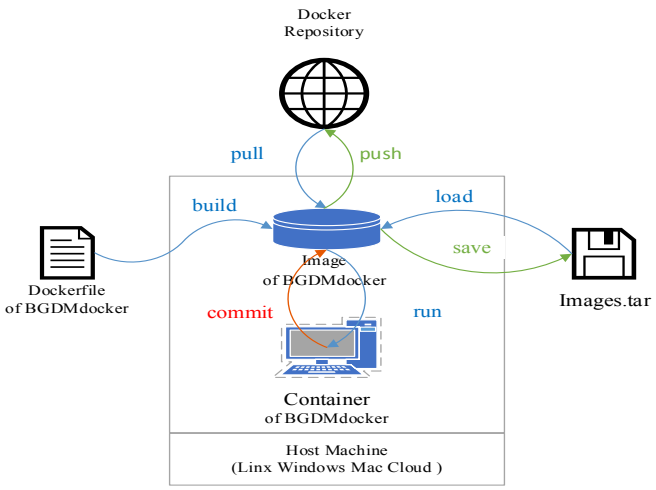

Fig. 1 Schematic of BGDMdocker workflow based on Docker. Building Image and running Container from Dockerfile, then login interaction patterns to run software. We can load and save Image.tar, run and commit Container, pull from and push to Docker repository.

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

In order to explore highly dimensional 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 can 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 172388; Core gene clusters 2306)				Gene of strain genomes	
		Total gene	Core gene	Acc gene	Uni gene	All gene	All protein
CYHL01000001	JRS5	3856	2310	1546	57	3870	3863
CYHP01000001	JRS8	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	UMAF6639	3825	2311	1514	20	3879	3716
NZ_CP006960	UMAF6614	3804	2311	1493	13	3850	3695
NZ_CP007242	KHG19	3775	2310	1465	19	3816	3658
NZ_CP010556	L-H15	3724	2309	1415	6	3769	3615
NZ_CP011278	L-S60	3728	2310	1418	7	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_DF836091	CMW1	3771	2311	1460	128	3901	3706
NZ_JCOC01000001	EBL11	3733	2308	1425	20	3773	3682
NZ_JMEG01000001	B1895	3824	2306	1518	167	4026	3623
NZ_JQNZ01000001	X1	3724	2309	1415	3	3766	3619
NZ_JTJG01000001	JJC33M	3888	2309	1579	121	3952	3796
NZ_JXAT01000001	LPL-K103	3709	2309	1400	15	3743	3637
NZ_JZDI01000001	12B	8166	2354	5812	4040	8194	7985
NZ_KB206086	DC-12	3910	2311	1599	50	3984	3842
NZ_KN723307	TF281	3640	2312	1328	5	3782	3571
NZ_LGYP01000001	629	3536	2313	1223	11	3785	3427
NZ_LJAU01000001	Bs006	4042	2312	1730	46	4074	3969
NZ_LJDI01000020	XK-4-1	3799	2310	1489	14	3821	3701
NZ_LMAG01000001	RHNK22	3781	2309	1472	37	3837	3698
NZ_LMAT01000001	Jxnuwx-1	3930	2309	1621	246	4008	3870
NZ_LMUC01000016	H57	3816	2310	1506	42	3859	3732
NZ_LPUP01000011	11B91	3790	2311	1479	49	3892	3702
NZ_LQQW01000001	M49	3694	2311	1383	21	3741	3617
NZ_LQYO01000001	B4140	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_LZZO01000001	SRCM101294	3946	2308	1638	175	3982	3850

Genome sequences of 44 B. amyloliquefaciens strains downloaded from GenBank RefSeq dadabase.“Acc gene” is accessory gene (dispensable gene), “Uni gene” is unique gene (strain-specific gene), “All genes” is gene of *.gbff files recorder, include Pseudo Genes, “Total genes” is involved in the pan genome analysis, not Pseudo Genes.

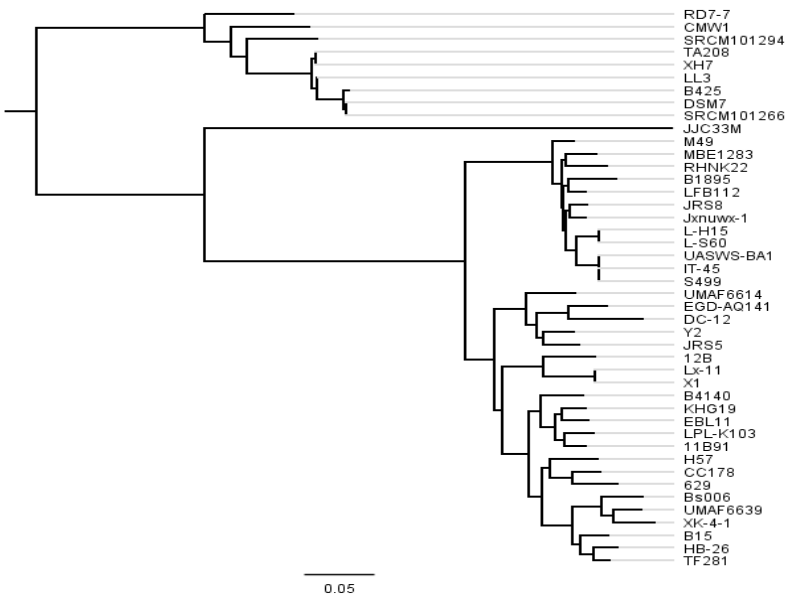

Fig. 2 Phylogenetic tree of 44 *B. amyloliquefaciens* strains. The tree was constructed using all of the genes shared between all 44 strains (2306 core genes). The scale bar represents genetic distance.

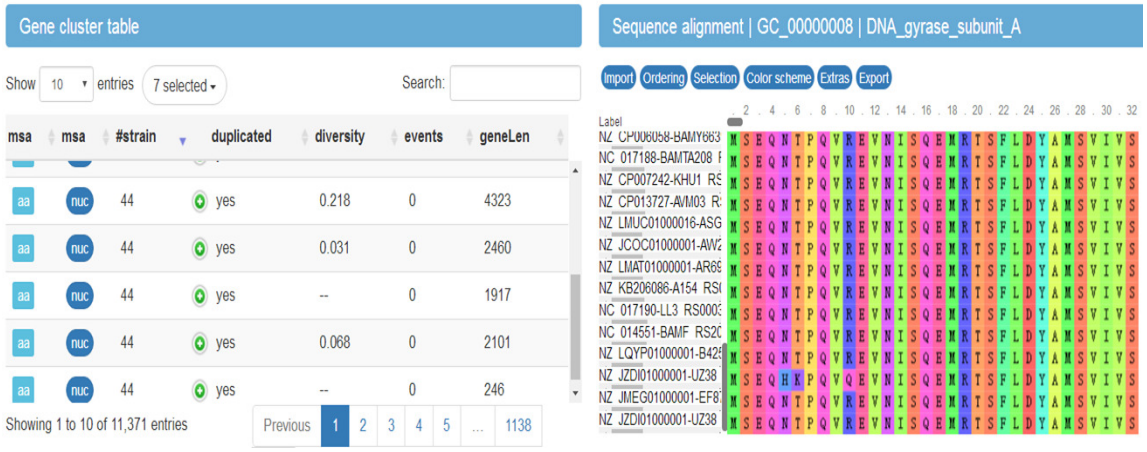

Fig. 3 Screenshot of website(Gene cluster table an Sequence algnment of 44 *B. amyloliquefaciens* strains).

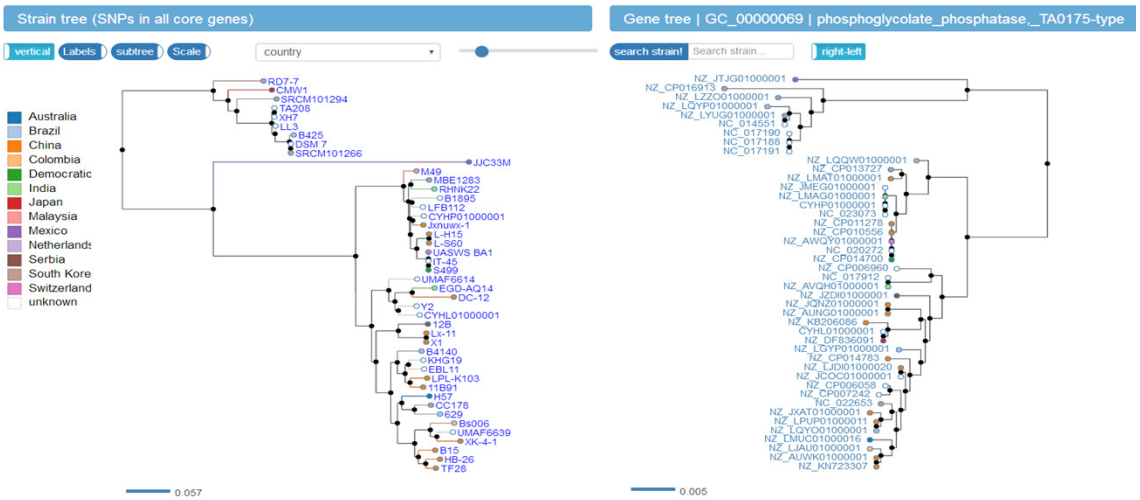

Fig. 4 Screenshot of website (Phylogenetic tree of 44 B. amyloliquefaciens strains and gene. The scale bar represents genetic distance).

Datamining and visualization biosynthetic gene clusters of *B. amyloliquefaciens*

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 can be downloaded without registration.

Here, we provide brief summary statistics of the biosynthetic gene clusters of all 44 strains (Table 2), as well as providing an example of the types 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(http://mibig.secondarymetabolites.org/), such as surfactin, mersacidin, and fengycin, while the other 10 gene clusters are unknown.

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

Accession	Strains	Biosynthesis gene clusters				Genome of strains		
		Total	known	Unknown	Type	Size(Mb)	Gene	Protein
CYHL01000001	JRS5	38	27	11	12	4.03148	3870	3863
CYHP01000001	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	3.9392	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	LFB112	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_AVQH01000001	EGD-AQ141	36	26	10	12	4.22259	4121	3995
NZ_AWQY01000001	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
NZ_CP007242	KHG19	32	20	12	10	3.95336	3816	3658
NZ_CP010556	L-H15	32	19	13	10	3.90597	3769	3615
NZ_CP011278	L-S60	32	19	13	10	3.90302	3773	3611
NZ_CP013727	MBE1283	35	22	13	12	3.97993	3856	3681
NZ_CP014700	S499	33	19	14	11	3.93593	3819	3671
NZ_CP014783	B15	29	19	10	10	4.00675	3875	3704
NZ_CP016913	RD7-7	31	17	14	8	3.68821	3656	3483
NZ_DF836091	CMW1	30	20	10	11	3.90857	3901	3706
NZ_JCOC01000001	EBL11	35	23	12	11	3.92932	3773	3682
NZ_JMEG01000001	B1895	38	24	14	12	4.10728	4026	3623
NZ_JQNZ01000001	X1	40	28	12	10	3.9211	3766	3619
NZ_JTJG01000001	JJC33M	36	25	11	12	3.96166	3952	3796
NZ_JXAT01000001	LPL-K103	36	23	13	9	3.87327	3743	3637
NZ_JZDI01000001	12B	69	49	20	11	7.59676	8194	7985
NZ_KB206086	DC-12	28	19	9	11	4.01656	3984	3842
NZ_KN723307	TF281	31	20	11	11	3.98764	3782	3571
NZ_LGYP01000001	629	31	18	13	10	3.90337	3785	3427
NZ_LJAU01000001	Bs006	45	30	15	10	4.17309	4074	3969
NZ_LJDI01000020	XK-4-1	37	24	13	12	3.94181	3821	3701
NZ_LMAG01000001	RHNK22	38	27	11	12	3.97818	3837	3698
NZ_LMAT01000001	Jxnuwx-1	40	27	13	10	4.08932	4008	3870
NZ_LMUC01000016	H57	34	23	11	11	3.95883	3859	3732
NZ_LPUP01000011	11B91	33	20	13	10	4.02366	3892	3702
NZ_LQQW01000001	M49	41	30	11	11	3.88665	3741	3617

NZ_LQYO01000001	B4140	39	25	14	11	4.01425	3847	3713
NZ_LQYP01000001	B425	29	20	9	9	3.9682	4034	3844
NZ_LYUG01000001	SRCM101266	31	19	12	11	3.76536	3781	3628
NZ_LZZO01000001	SRCM101294	32	20	12	10	3.96275	3982	3850

"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). "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 using this probabilistic algorithm to detect putative gene clusters of unknown types. "-" of host is unrecorded.

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
Cluster 7	Cf_putative	-	-
Cluster 8	Otherks	-	-
Cluster 9	Cf_fatty_acid	-	-
Cluster 10	Cf_putative	-	-
Cluster 11	Terpene	-	-
Cluster 12	Cf_fatty_acid	-	-
Cluster 13	Cf_putative	-	-
Cluster 14	Cf_putative	-	-
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)	BGC0001089_c1
Cluster 17	Nrps-Transatpks	Fengycin_biosynthetic_gene_cluster (93% of genes show similarity)	BGC0001095_c1
Cluster 18	Terpene	-	-
Cluster 19	Cf_saccharide-T3pks	-	-
Cluster 20	Transatpks	Difficidin_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0000176_c1
Cluster 21	Cf_putative	-	-
Cluster 22	Nrps-Bacteriocin	Bacillibactin_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0000309_c1
Cluster 23	Cf_saccharide	-	-
Cluster 24	Nrps	-	-
Cluster 25	Cf_saccharide	Teichuronic_acid_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0000868_c1
Cluster 26	Cf_putative	-	-
Cluster 27	Cf_saccharide	Bacilysin_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0001184_c1
Cluster 28	Cf_putative	-	-
Cluster 29	Lantipeptide	Mersacidin_biosynthetic_gene_cluster (90% of genes show similarity)	BGC0000527_c1
Cluster 30	Cf_saccharide	-	-
Cluster 31	Cf_putative	-	-

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

Select Gene Cluster:

Overview 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31

Identified secondary metabolite clusters

Cluster	Type	From	To	Most similar known cluster	MIBiG BGC-ID
The following clusters are from record NC_017912.1:					
Cluster 1	Nrps	338057	358115	Surfactin_biosynthetic_gene_cluster (43% of genes show similarity)	BGC0000433_c1
Cluster 2	Cf_putative	382926	400100	-	-
Cluster 3	Cf_putative	457505	463879	-	-
Cluster 4	Cf_fatty_acid	553834	571410	-	-
Cluster 5	Phosphonate	611959	652849	Pactamycin_biosynthetic_gene_cluster (3% of genes show similarity)	BGC0000119_c1
Cluster 6	Cf_saccharide	707404	724559	Plantathiazolicin_/_plantazolicin_biosynthetic_gene_cluster (33% of genes show similarity)	BGC0000569_c1
Cluster 7	Cf_putative	815908	837072	-	-
Cluster 8	Otherks	928506	969750	-	-
Cluster 9	Cf_fatty_acid	997814	1018791	-	-
Cluster 10	Cf_putative	1019821	1028475	-	-
Cluster 11	Terpene	1052066	1072806	-	-
Cluster 12	Cf_fatty_acid	1099917	1105735	-	-
Cluster 13	Cf_putative	1132407	1140612	-	-
Cluster 14	Cf_putative	1217728	1234302	-	-
Cluster 15	Transatpks	1414044	1469984	Macrolactin_biosynthetic_gene_cluster (90% of genes show similarity)	BGC0000181_c1
Cluster 16	Nrps-Transatpks	1766710	1842979	Bacillaene_biosynthetic_gene_cluster (85% of genes show similarity)	BGC0001089_c1
Cluster 17	Nrps-Transatpks	2075376	2178857	Fengycin_biosynthetic_gene_cluster (93% of genes show similarity)	BGC0001095_c1
Cluster 18	Terpene	2224424	2246307	-	-
Cluster 19	Cf_saccharide-T3pks	2296276	2344896	-	-
Cluster 20	Transatpks	2506980	2581912	Difficidin_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0000176_c1
Cluster 21	Cf_putative	3239477	3252309	-	-
Cluster 22	Nrps-Bacteriocin	3278880	3345673	Bacillibactin_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0000309_c1
Cluster 23	Cf_saccharide	3533963	3550378	-	-
Cluster 24	Nrps	3624364	3661226	-	-
Cluster 25	Cf_saccharide	3683737	3738552	Teichuronic_acid_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0000868_c1
Cluster 26	Cf_putative	3771770	3782210	-	-
Cluster 27	Cf_saccharide	3882481	3944920	Bacilysin_biosynthetic_gene_cluster (100% of genes show similarity)	BGC0001184_c1
Cluster 28	Cf_putative	4020100	4026182	-	-
Cluster 29	Lantipeptide	4078161	4102145	Mersacidin_biosynthetic_gene_cluster (90% of genes show similarity)	BGC0000527_c1
Cluster 30	Cf_saccharide	4170395	4195268	-	-
Cluster 31	Cf_putative	4203706	4219124	-	-

Fig. 5 Screenshot of visualization website of biosynthetic gene clusters total numbers of Y2

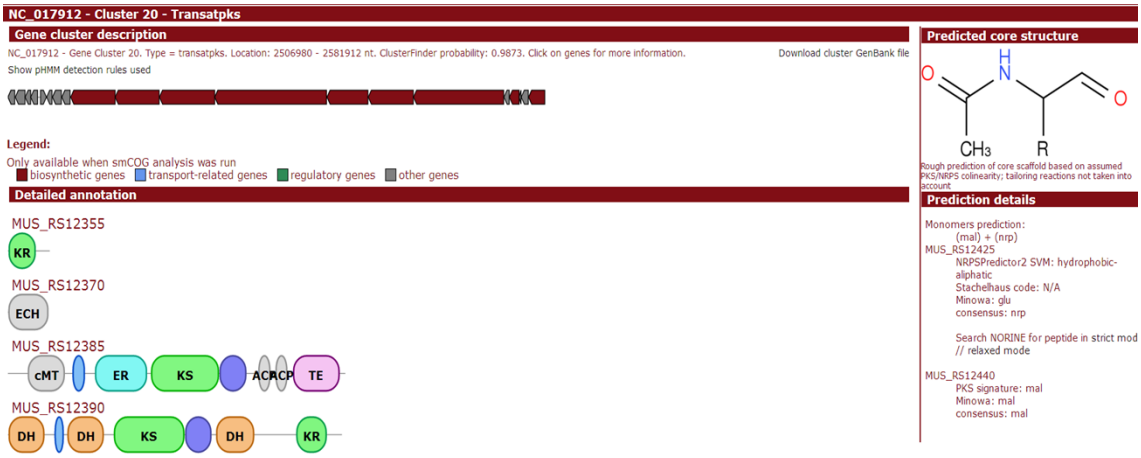

Fig. 6 Screenshot of visualization website of Transatpks type cluster of Y2 (Most similar known cluster Difficidin_biosynthetic_gene_cluster)

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 the best flexibility application environments. 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 can run 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 can automatically and effectively analyze biological problems of interest. The BGDMdocker Container enables application isolation with high efficiency and flexibility. Applications can run Container independently with Docker technology, and each management command (start, stop, boot, etc.) can be executed in seconds or milliseconds. Hundreds or thousands of Containers could be run on a single host at same time (Ali *et al*., 2016), thus ensuring that the failure of one task does not cause disruption of the entire process. Instead, new Containers can be initialized quickly to continue the task until the completion of the entire process, 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 can be complicated, and researchers are therefore not able to directly focus on their analyses of interest(Table S5). Although the BGDMdocker workflow includes several tools, installing and running the software is quite simple. Biologists can automatically install, configure, and test the scripts, making these processes faster and the results repeatable.

Table S5. Function of BGDMdocker workflow compared with others pan-genome tools

Tools	Automatic installation	Cross platform	Resulet visualization	Genome annotation	Gene Cluster mining
Roary	×	×	×	√	×
PGPA	×	×	×	×	×
SplitMEM	×	×	×	×	×
PanGP	×	×	×	×	×
PanTools	×	×	√	×	×
BGDMdocker workflow	√	√	√	√	√

√ is provided with the function, × is not provided with the function

CONCLUSIONS

Here, we presented a BGDMdocker workflow to achieve bacterial and viral genome annotation, pan-genome analysis, mining of biosynthetic gene clusters, and result visualization on a local host or online. This allows researchers to browse information for every gene, including duplication, diversity, indel events, and sequence alignments, and 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 that were 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 for the analysis and visualization of pan-genomes and biosynthetic gene clusters can be fully reused immediately across different computing platforms (Linux, Windows, Mac, and cloud-based systems), with the flexible and rapid deployment of integrated software packages across various platforms. This workflow could also be used for other pan-genome analysis and visualization of other species. The visual display of data provided in this paper can be completely duplicated as well. All resulting data and relevant tools and files can 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

Funding

This work was supported by the National Key Research and Development Project (SQ2017YFNC030122), Fundamental Research Fund for Central Non-profit Scientific Institution (Y2016PT38), and CAAS-ASTIP.

Competing Interests

The authors declare there are no competing interests.

Data Availability

The following information is supplied regarding data and code availability:

GitHub: https://github.com/cgwyx/debian_prokka_panx_antismash_biodocker

Website: http://pangenome.zggskj.com/home

FigShare: https://figshare.com/s/5e7b44f4bb8aba4a65f1

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.1791#supplemental-information.

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.

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.

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.

Tommaso PD, Palumbo E, Chatzou M, Prieto P, Heuer ML, Notredame C. 2015. AntiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. Nucleic Acids Research, 43(W1), W237–W243.