microorganisms

MDPI

Article

BacSeq: A User-Friendly Automated Pipeline for Whole-Genome Sequence Analysis of Bacterial Genomes

Arnon Chukamnerd [1], Kongpop Jeenkeawpiam [2], Sarunyou Chusri [1], Rattanaruji Pomwised [3], Kamonnut Singkhamanan [2,*] and Komwit Surachat [2,4,5,*]

[1] Division of Infectious Diseases, Department of Internal Medicine, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand; arnonchukamnerd@hotmail.com (A.C.); sarunyouchusri@hotmail.com (S.C.)
[2] Department of Biomedical Sciences and Biomedical Engineering, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand; kongpop.je@gmail.com
[3] Division of Biological Science, Faculty of Science, Prince of Songkla University, Songkhla 90110, Thailand; rattanaruji.p@psu.ac.th
[4] Translational Medicine Research Center, Faculty of Medicine, Prince of Songkla University, Songkhla 90110, Thailand
[5] Division of Computational Science, Faculty of Science, Prince of Songkla University, Songkhla 90110, Thailand
* Correspondence: skamonnu@medicine.psu.ac.th (K.S.); komwit.s@psu.ac.th (K.S.)

Abstract: Whole-genome sequencing (WGS) of bacterial pathogens is widely conducted in microbiological, medical, and clinical research to explore genetic insights that could impact clinical treatment and molecular epidemiology. However, analyzing WGS data of bacteria can pose challenges for microbiologists, clinicians, and researchers, as it requires the application of several bioinformatics pipelines to extract genetic information from raw data. In this paper, we present BacSeq, an automated bioinformatic pipeline for the analysis of next-generation sequencing data of bacterial genomes. BacSeq enables the assembly, annotation, and identification of crucial genes responsible for multidrug resistance, virulence factors, and plasmids. Additionally, the pipeline integrates comparative analysis among isolates, offering phylogenetic tree analysis and identification of single-nucleotide polymorphisms (SNPs). To facilitate easy analysis in a single step and support the processing of multiple isolates, BacSeq provides a graphical user interface (GUI) based on the JAVA platform. It is designed to cater to users without extensive bioinformatics skills.

Keywords: whole-genome sequencing; BacSeq; assembly; annotation; bioinformatics

Citation: Chukamnerd, A.; Jeenkeawpiam, K.; Chusri, S.; Pomwised, R.; Singkhamanan, K.; Surachat, K. BacSeq: A User-Friendly Automated Pipeline for Whole-Genome Sequence Analysis of Bacterial Genomes. *Microorganisms* **2023**, *11*, 1769. https://doi.org/10.3390/microorganisms11071769

Academic Editors: Pufeng Du, Bing Niu and Suren Rao Sooranna

Received: 1 June 2023
Revised: 4 July 2023
Accepted: 4 July 2023
Published: 6 July 2023

1. Introduction

High-throughput sequencing (HTS) technologies have revolutionized the field of genomics by allowing researchers to analyze large quantities of genetic material in a relatively short amount of time [1,2]. Short-read sequencing (SRS) and long-read sequencing (LRS) are powerful tools to study the entire genomes of bacteria [2]. The sequence reads from these technologies are generated as a fastq file, which needs bioinformatics tools for further analysis. Command-line, web-based, and program-based tools are currently used for sequence analyses [3]. Among them, command-line tools provide maximum flexibility and are highly customizable, but require a higher level of technical expertise and may be more time-consuming for certain tasks. Web-based tools, on the other hand, are generally more user-friendly and accessible to users without extensive bioinformatics training, but may have limitations in terms of customization and flexibility. Program-based tools provide a balance between the two, offering a graphical user interface that is more accessible than command-line tools while still providing a high degree of flexibility.

In a previous study, Quijada et al., (2019) developed automated pipelines called TORMES for analyzing whole-genome sequencing (WGS) data of bacteria generated by Illumina platforms [4]. TORMES automates the bioinformatic analysis steps, including

sequence quality filtering, de novo assembly, draft genome ordering against a reference, genome annotation, multi-locus sequence typing (MLST), searching for antibiotic resistance and virulence genes, and pan-genome comparisons. The pipeline can be used for any set of bacteria from any species and origin, and more extensive analyses for *Escherichia* and *Salmonella* can be enabled using the −g/− genera option. Once the analysis is finished, TORMES generates an interactive web-like report that can be opened in any web browser, and shared and revised by researchers in a simple manner. However, it should be noted that TORMES may not be suitable for all types of WGS data, and researchers probably consider using other inputs, such as short-read sequences and long-read sequences from different platforms, to obtain a more comprehensive understanding of their bacterial genomes. Additionally, many researchers may not have the necessary bioinformaticians to fully utilize TORMES or other sequencing analysis tools. We then aimed to generate and improve an easy-to-use automated pipeline for WGS and bioinformatics analyses of bacterial genomes, which is beneficial for non-bioinformatician users.

2. Materials and Methods

2.1. Bioinformatics Pipeline

BacSeq integrates several frequently used open-source bioinformatics tools to perform a single-step analysis including assembly, assembly quality evaluation, gene prediction, functional annotation, specific gene identification, and pan-genome analysis. The pipeline begins with loading compressed raw read files (.fastq.gz; accessed on 5 May 2023) and checking the quality via FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/; accessed on 5 May 2023), and the results are imported into MultiQC [5] to generate summary reports. The trimming step is then performed using fastp [6] to remove the adapter sequence, cut low-quality bases, and trim all reads at the 5′ and 3′ ends. Next, SPAdes [7] is used for assembling the filtered sequences into contigs and scaffolds using various k-mer lengths.

Next, genome assembly assessment and completeness evaluation are performed using QUAST [8] and BUSCO [9], respectively. For the annotation process, Prokka [10] is called to identify genomic features of interest in the assembled genome. Functional annotation is then performed with eggNOG-mapper [11] which combines HMMER [12], DIAMOND [13], MMSEQS2 [14], and PRODIGAL [15] to search against several databases including Clusters of Orthologous Groups of proteins (COGs) [16], Gene Ontology (GO) [17], Protein family (PFAM) [18], and Kyoto Encyclopedia of Genes and Genomes (KEGG) [19]. Next, the downstream analysis to identify pathogenic-related genes starts by running the ABRicate pipeline (https://github.com/tseemann/abricate/; accessed on 5 May 2023) to search against several databases including the Comprehensive Antibiotic Resistance Database (CARD) [20], ResFinder [21], Antibiotic Resistance Gene-ANNOTation (ARG-ANNOT) [22], MEGARes [23], Virulence Factor Database (VFDB) [24], PlasmidFinder [25], and ISfinder [26]. Carbohydrate-active enzymes (CAZyme) and Clustered Regularly Interspaced Short Palindromic Repeats and CRISPR-associated proteins (CRISPR-Cas) are then searched via automated CAZyme annotation [27,28] and CRISPRCasFinder [29], respectively.

A pan-genome analysis was then performed by Roary [30] to identify the core and accessory genes from a collection of assembled genomes. Single-nucleotide polymorphisms (SNPs) of core genes are called by SNP-sites [31] and constructed the phylogenetic tree using FastTree [32]. All analysis reports are finally generated by combining all results into web format and Comma-Separated Values (CSV) files. The overall bioinformatics workflow is presented in Figure 1.

2.2. Requirements

BacSeq is a JAVA-based application for analyzing WGS data using paired-end reads and supports either single or multiple genomes in one analysis. BacSeq can automatically complete assembling, annotating, identifying target genes, and analyzing comparative genomes. To start analysis using BacSeq, raw paired-end reads of the sample(s) are required.

To run this tool, Conda (https://docs.anaconda.com/; accessed on 5 May 2023), an open-source package management system, is required to install BacSeq version 1.0 and all prerequisite software. BacSeq only supports Linux systems and requires a minimum space capacity of 1 Gb for installation.

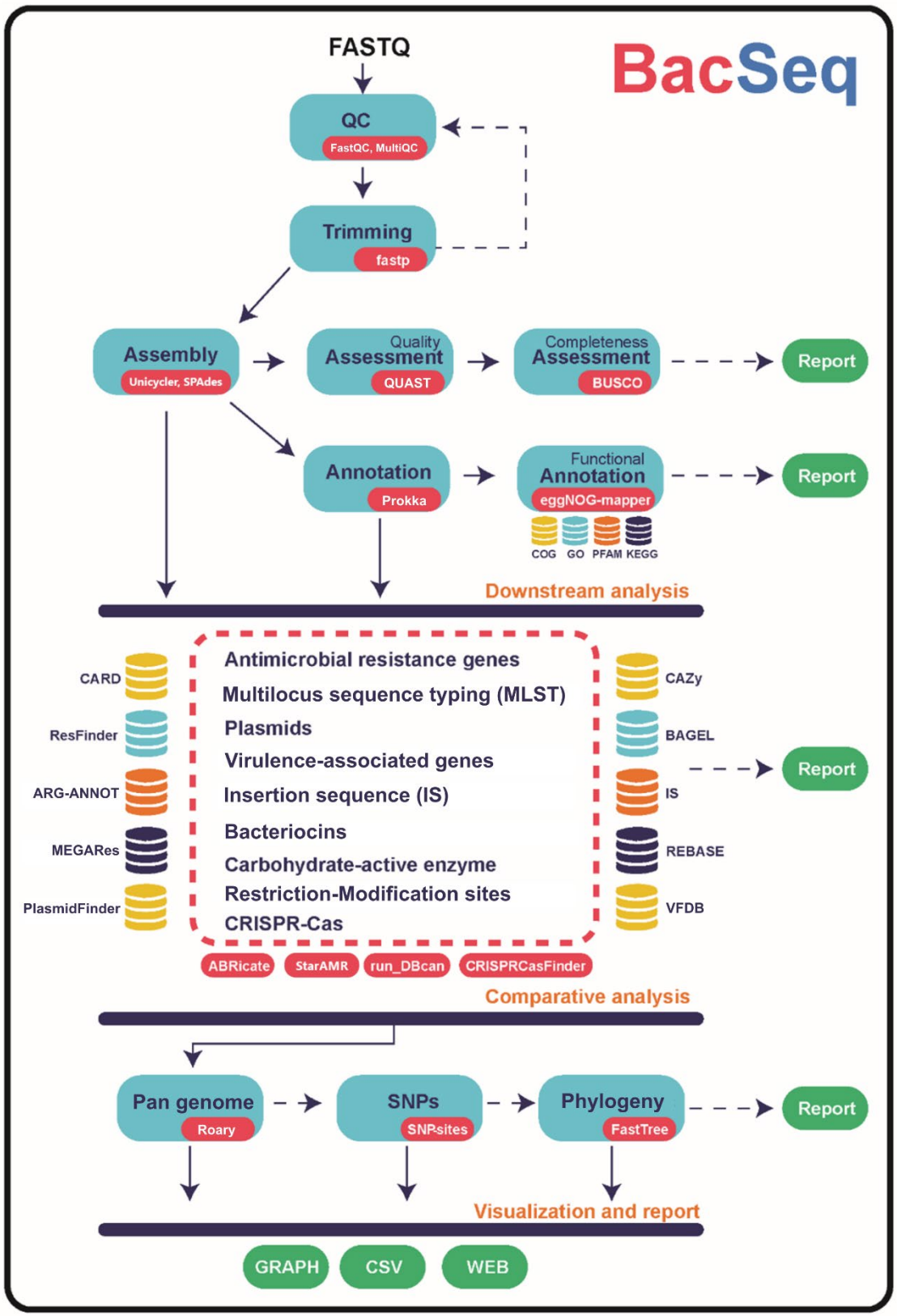

Figure 1. Bioinformatics workflow of BacSeq analysis steps.

2.3. Pipeline Customization

BacSeq offers two modes i.e., quick and advanced modes for novice and expert users, respectively. In quick mode, all results can be easily analyzed with a single click without further configuration. Default parameters are set in all tools in this mode for convenience. However, manual configuration can be used in the advanced mode. All bioinformatics tools can be configured parameters using a graphical user interface (GUI). In addition, expert users can optionally use any bioinformatics tools integrated into BacSeq via a command-line interface.

3. Results and Discussion

3.1. Graphical User Interface (GUI)

The BacSeq pipeline was deployed as a JAVA-based application, enabling users to interact directly with the graphical user interface (GUI) for performing bioinformatics analysis, as shown in Figure 2. To start using BacSeq, users can simply select the directory containing the genome data and execute the program to complete the analysis in a single step. Users only need to use the command-line interface once for program installation. The pipeline supports both single files and multiple files in one analysis by just providing the absolute path of the directory of the file. However, all files must be prepared and renamed to an allowed pattern for the program to recognize the files and import them into the pipeline.

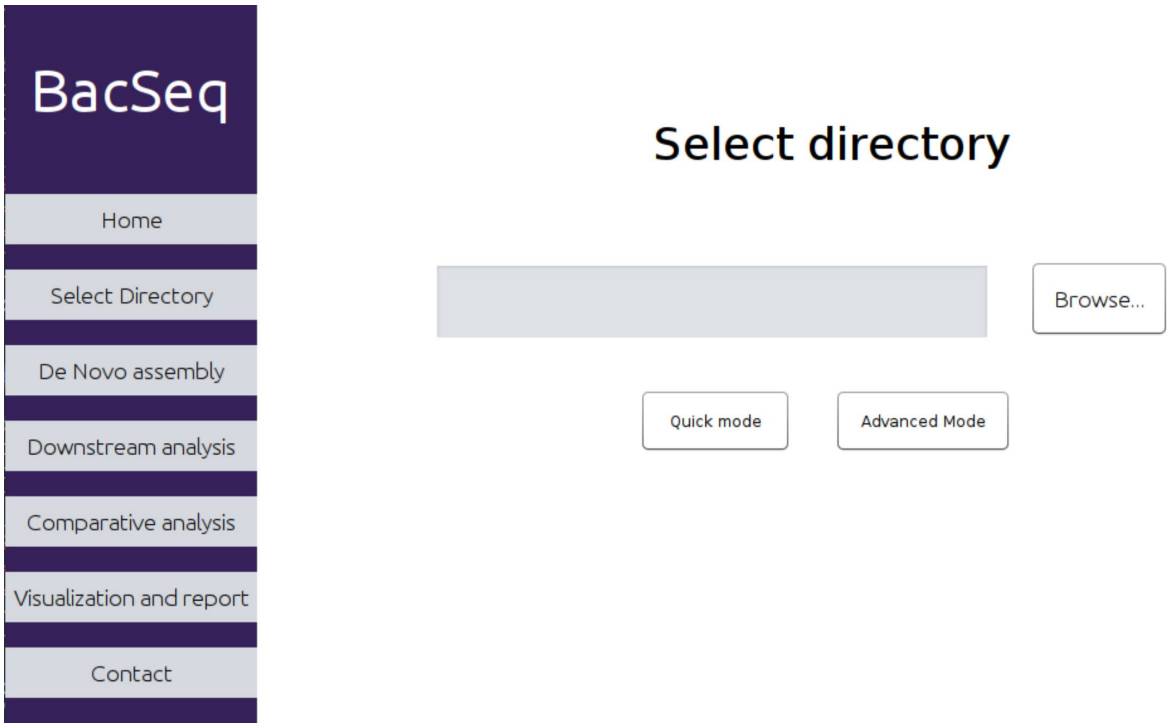

Figure 2. Graphical user interface (GUI) of BacSeq program.

3.2. Use Case: Draft Genome Sequences of Acinetobacter baumannii Isolates

We used BacSeq to analyze the short-read sequencing data of 13 carbapenem-resistant *Acinetobacter baumannii* (CRAB) isolates, including PA020, PA025 (JAIGYO000000000), ST001 (JAIGSU000000000), ST009 (JAIGST000000000), ST010 (JAIGSS000000000), ST024 (JAIGSR000000000), ST028, ST032 (JAIGSQ000000000), ST034 (JAIGSP000000000), ST035 (JAIGSO000000000), ST036, YL005 (JAIGQD000000000), and YL006 (JAIGQC000000000). The qualified genomes of 10 isolates were deposited into the NCBI GenBank, except for the PA020, ST028, and ST036 isolates. Although the unqualified genomes existed in these 3 isolates, they were still included here as examples. The isolates were approved by the Human Research Ethics Committee (HREC) from Prince of Songkla University, Thailand

(reference number: 64–284-14–1, date of approval: 9 June 2021). *A. baumannii* is a Gram-negative, rod-shaped, and aerobic bacterium that commonly causes hospital-acquired infection, particularly in intensive care units (ICUs) and among critically ill patients [33]. It has gained notoriety for its remarkable ability to acquire resistance to multiple antibiotics, especially carbapenem, through several mechanisms [33]. Moreover, their genetic materials, such as antibiotic resistance genes (ARGs) and virulence-associated genes (VAGs), could be transferred between the genus and other Gram-negative bacteria [34]. The Centers for Disease Control and Prevention (CD) have classified carbapenem-resistant *Acinetobacter* spp. as an urgent threat level [35]. Thus, the entire genome of this pathogen is necessary to be sequenced, which may provide more understanding of the genetic features related to its molecular evolution.

3.2.1. Quality Control

According to the analysis workflow (Figure 1), quality control was initially performed to verify the raw reads. The reports exhibited total sequences, sequences flagged as poor quality, sequence length, %GC, total deduplicated percentage, average sequence length, basic statistics, Per base sequence quality, Per sequence quality scores, Per base sequence content, Per sequence GC content, Per base N content, sequence length distribution, sequence duplication levels, overrepresented sequences, and adapter content (Table S1). The results were reported as quantity or quality (pass, warn, and fail). As shown in Table S1 and Figure 3, the quality of most isolates was acceptable, while Per sequence GC content of the ST028 genome failed. This failure occurs when the cumulative deviations from the normal distribution of GC content in the reads exceed 30% [36].

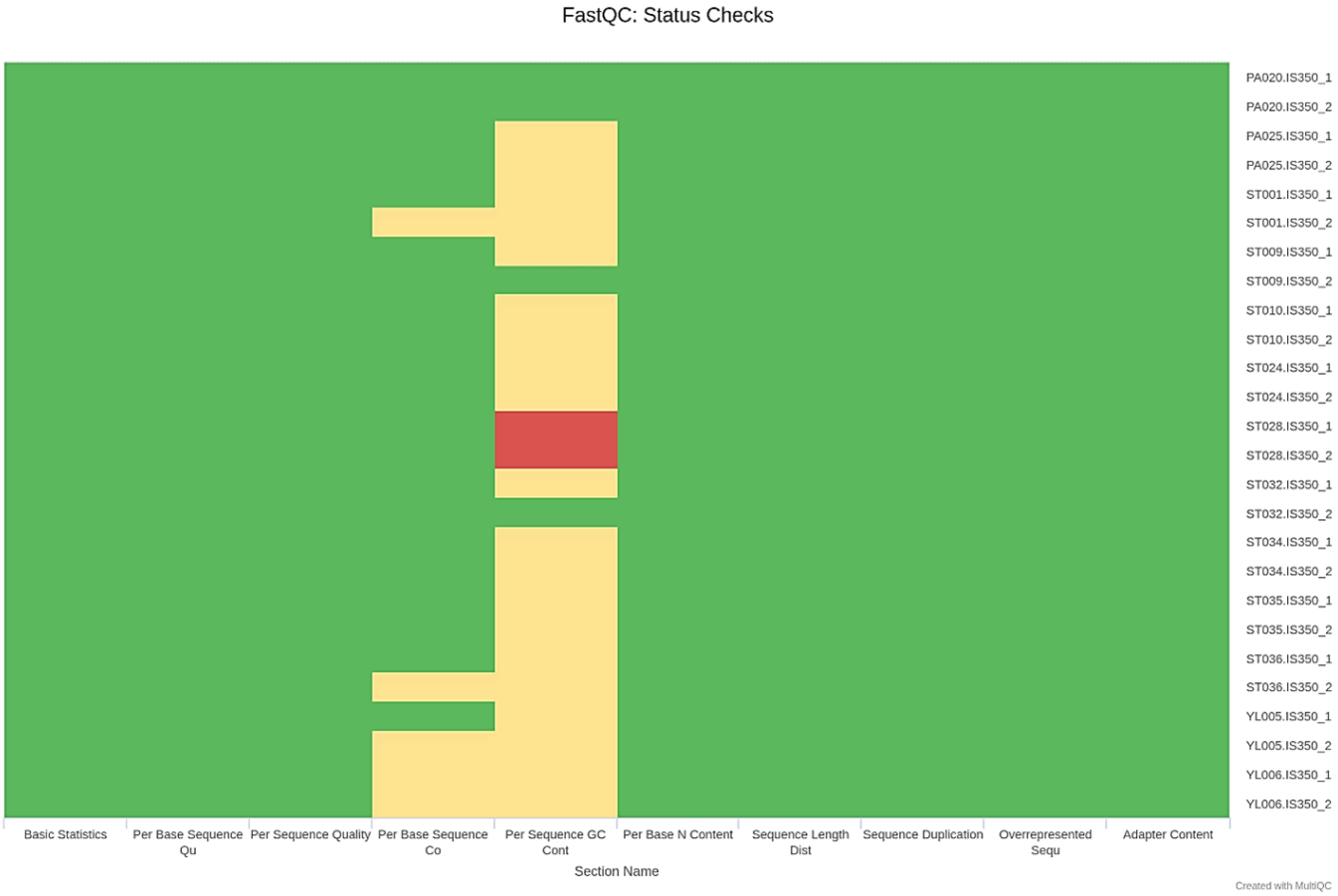

Figure 3. A report of status check by FastQC. Green, yellow, and red boxes represent pass, warn, and fail qualities, respectively.

3.2.2. Genome Assembly, Assembly Quality Assessments, and Genome Annotation

The assembled sequences were subjected to quality assessments using QUAST, which reported the number of contigs, total sequence lengths, %GC, N50, N90, L50, and L90 (Table 1). We found that assembly of the ST036 genome provided a high number (n = 818) of total contigs, which generally indicates a more fragmented assembly. It means that the genome was not fully reconstructed into large, contiguous sequences but rather fragmented into numerous smaller pieces. This may occur due to various reasons, including repetitive or complex regions in the genome, sequencing errors, low coverage depth, or difficulties in resolving repetitive sequences [37]. However, a high number of contigs might be acceptable for some applications, such as comparative genomic analysis or identification of gene families; it is often desirable to have fewer contigs for a more complete and accurate representation of the genome. We also found that the PA020 and ST028 genomes consist of 9,243,789 bp and 9,539,281 bp, which is over the common length (approximately 3.7–4.4 bp) of the *A. baumannii* genome [38]. The reason may be the contamination of other bacterial genomes. However, these contaminant genomes were still included in further analyses, which could be used to compare it to other clean genomes. In addition, the completeness of assembled sequences was also assessed by BUSCO, which reported the percentages of complete, single-copy, duplicated, fragmented, and missing sequences (Figure 4). The result demonstrated that the duplicated sequences were observed in the PA020 and ST028 genomes, while the high percentages of fragmented and missing sequences were detected in the ST036 genome. For genome annotation, Prokka reports the amounts of tmRNA, tRNA, rRNA, miscRNA, gene, and coding sequence (CDS), as shown in Figure 5. Unusually high numbers of tRNA, genes, and CDS were observed in the PA020 and ST028 genomes due to their duplicated sequences. Additionally, Prokka also provided a GFF (General Feature Format) file that can be used as an input file in Roary for pan-genome analysis.

Table 1. A report of quality assessments by QUAST.

Isolate Code	Number of Contigs	Total Length	%GC	N50	N90	L50	L90
PA020	92	9,243,789	36.80	569,197	59,780	6	29
PA025	66	3,906,279	38.89	152,139	41,611	6	24
ST001	68	4,111,741	39.02	128,875	39,322	8	28
ST009	69	3,844,585	39.01	113,438	40,316	11	32
ST010	68	3,872,017	38.93	123,627	42,752	10	30
ST024	62	4,294,911	38.82	250,119	64,045	6	18
ST028	187	9,539,281	49.62	187,251	35,730	17	58
ST032	67	3,844,381	39.01	122,061	42,602	11	31
ST034	58	4,225,388	38.88	250,219	71,864	6	16
ST035	70	4,035,126	38.99	176,611	43,801	7	25
ST036	818	4,329,065	38.79	65,441	1278	15	255
YL005	109	3,910,735	38.91	76,044	195,99	18	55
YL006	53	3,894,856	38.99	190,977	61,096	6	19

3.2.3. Antibiotic Resistance, Also including Plasmids and Virulence Factors

For the identification of acquired ARGs, we provided an analysis against various databases in the ABRicate pipeline, including National Center for Biotechnology Information (NCBI), CARD, ResFinder, ARG-ANNOT, and MEGARes. Plasmid makers, VAGs, and sequence type (ST) could be also investigated, and their results were reported together with antibiotic resistance on the Hypertext Markup Language (HTML) page. In our case study, we reported the results of ARGs, plasmids, and VAGs, as illustrated in Figures 6–8. We found that all clinical isolates of CRAB carried the genes that encoded for antibiotic efflux pumps conferring resistance to fluoroquinolone (*abaQ* and *abeM*), macrolide (*amvA* and *abeS*), and tetracycline (*adeA*, *adeB*, *adeL*, *adeR*, and *adeS*) (Figure 6). They also carried the genes that encoded for resistance-nodulation-cell division (RND) antibiotic efflux pump conferring multidrug resistance to tetracycline and fluoroquinolone (*adeF*, *adeG*, and

adeH) and to rifamycin, diaminopyrimidine, penem, carbapenem, phenicol, tetracycline, macrolide, lincosamide, fluoroquinolone, and cephalosporin (*adeI*, *adeJ*, *adeK*, and *adeN*). In addition, antibiotic target alteration conferring aminoglycoside resistance (*armA*) and antibiotic inactivation (e.g., $bla_{\text{OXA-23}}$, $bla_{\text{NDM-1}}$, $bla_{\text{CARB-16}}$, *aph(3″)-Ib*, *aph(6′)-Id*, *mphE*, *msrE*, *fosA6*, and so on) were also detected in a high number of these CRAB isolates. Plasmid identification revealed that three plasmids, including IncFIA(HI1)_1_HI1, IncFIB(K)_1_Kpn3, and IncFII_1_pKP91, were only observed in the ST028 isolate (Figure 7). These plasmids, which have been classified as multidrug-resistant (MDR) plasmids, are commonly found in the Enterobacterales family, particularly *Salmonella* spp. and *Klebsiella* spp. [39–42], implying that the ST028 isolate may be contaminated with *Salmonella* spp. and/or *Klebsiella* spp. In virulence factor detection, we found that thiol-activated cytolysin gene (*BAS3109*), cytotoxin K (*cytk*), immune inhibitor A (*inhA*), hemolytic enterotoxin HBL complex genes (*hblA*, *hblC*, and *hblD*), and non-hemolytic enterotoxin genes (*nheA*, *nheB*, and *nheC*) were only harbored by the PA020 isolate (Figure 8). Enterotoxin genes (*entA*, *entB*, and *fepC*), outer membrane protein A gene (*ompA*), and adhesive virulence genes (*yagV/ecpE*, *yagW/ecpD*, *yagX/ecpC*, *yagY/ecpB*, *yagZ/ecpA*, and *ykgK/ecpR*) were only harbored by the ST028 isolate.

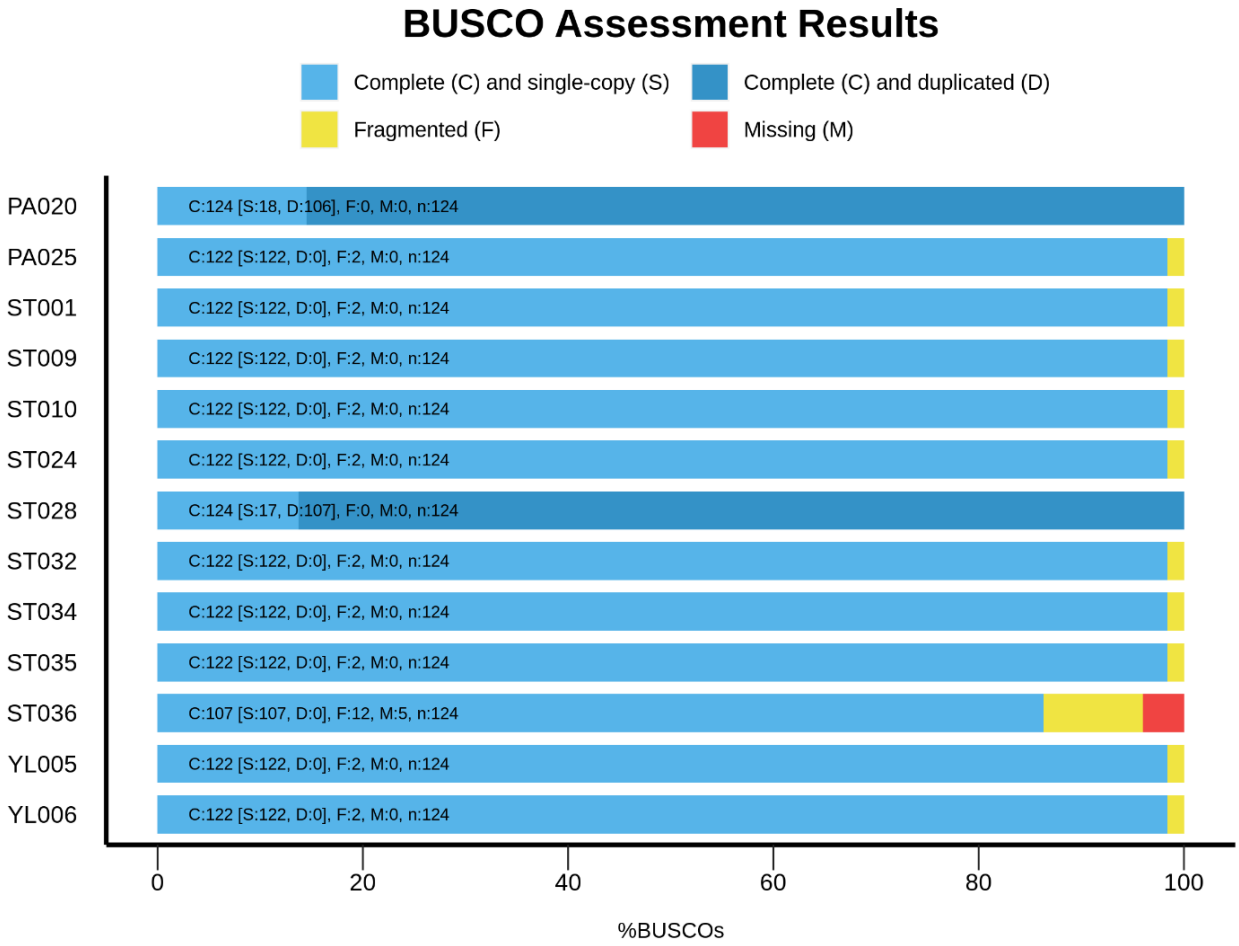

Figure 4. A report of genome completeness by BUSCO.

3.2.4. Comparative Analysis

In comparative genomic analysis, we provided Roary for analyzing pan-genome profiles among the studied genomes, which provide valuable insights into the genetic complexity and adaptability of species, helping us better understand their biology and evolution. For our case study, we found that 2509 (14.22%) core genes and 15,135 (85.78%) accessory genes were observed from 17,644 pan genes (Figure 9). Contaminant sequences in the PA020 and ST028 genomes resulted in the presence of high-level accessory genes existing in the pan-genome profile. The uncommon presence of these accessory genes

could also be observed from the results of genome annotation and downstream analysis. In general, a large proportion of accessory genes in a pan-genome analysis can be indicative of an open pan genome, suggesting substantial genetic diversity within the studied population or species [43]. This observation further suggests the occurrence of horizontal gene transfer (HGT) events, facilitating the acquisition of novel genes from different isolates, species, or organisms [44,45]. Such pan-genome analysis provides valuable insights into the evolutionary history of the species, including ongoing speciation processes and the existence of subpopulations harboring distinct gene sets [44,45].

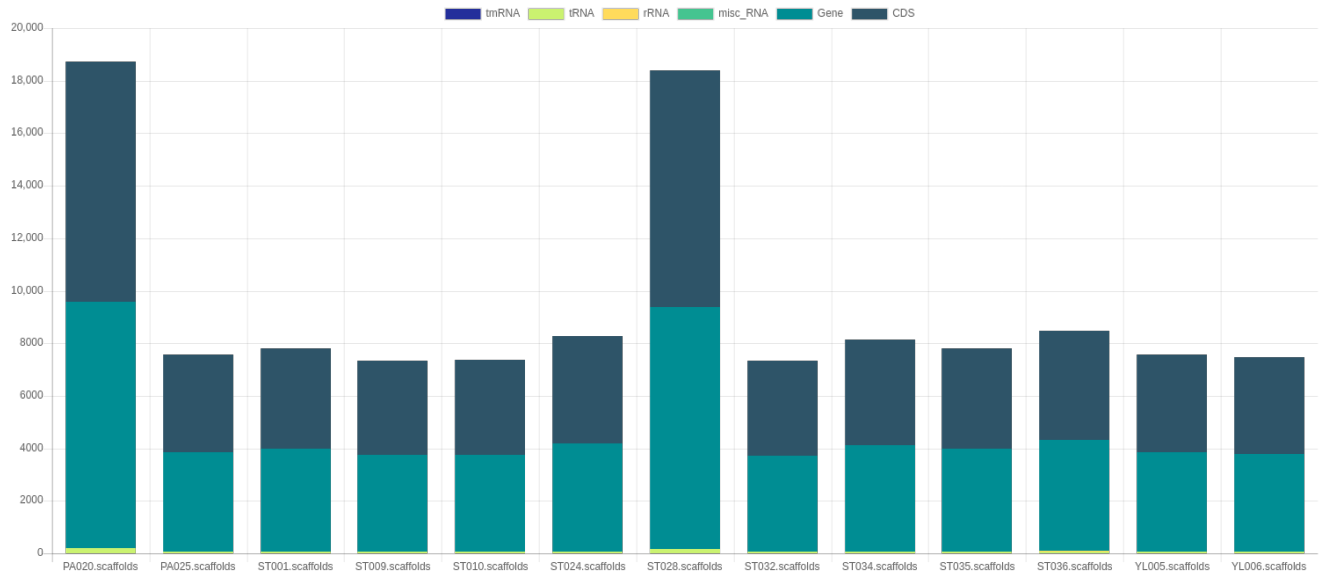

Figure 5. A report of genome annotation by Prokka.

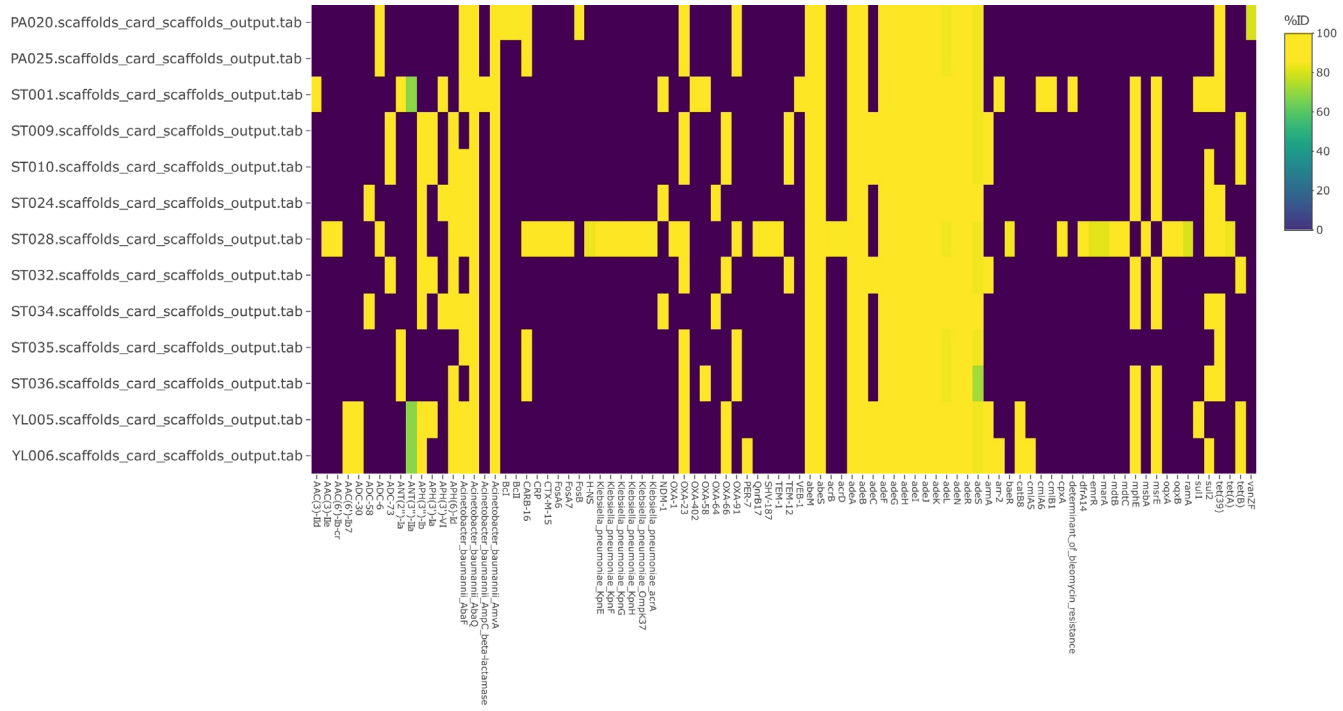

Figure 6. A report of antibiotic resistance genes (ARGs) against comprehensive antibiotic resistance database (CARD). ID, identity.

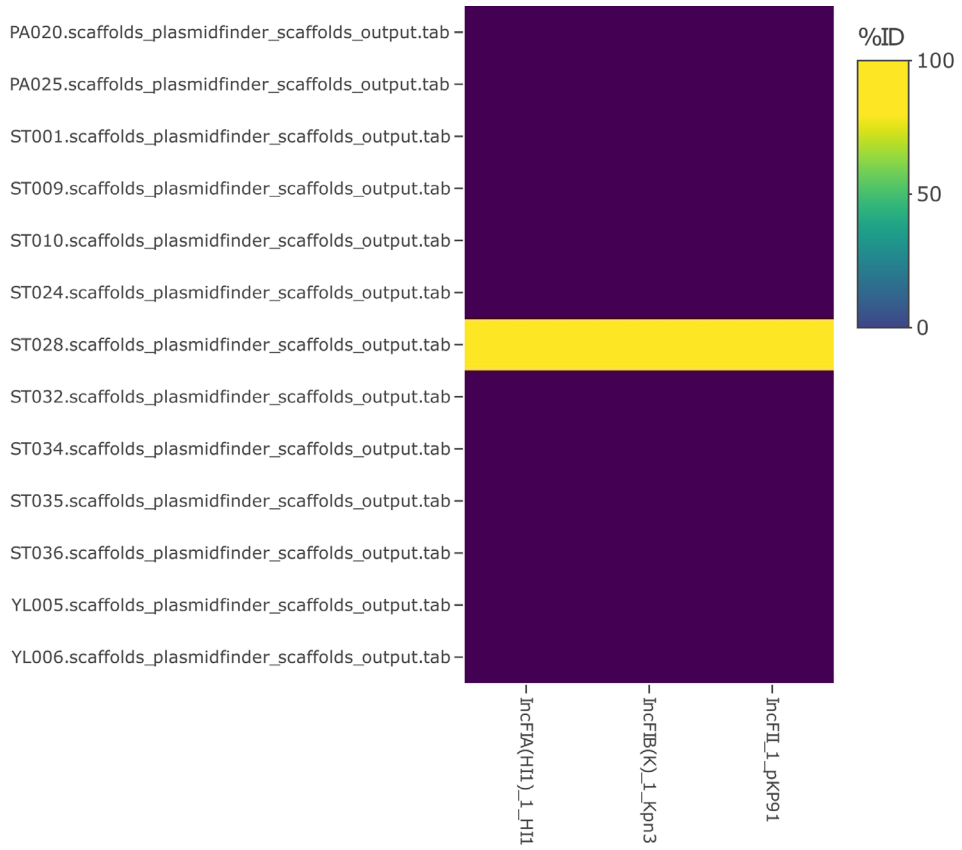

Figure 7. A report of plasmid types against PlasmidFinder database. ID, identity.

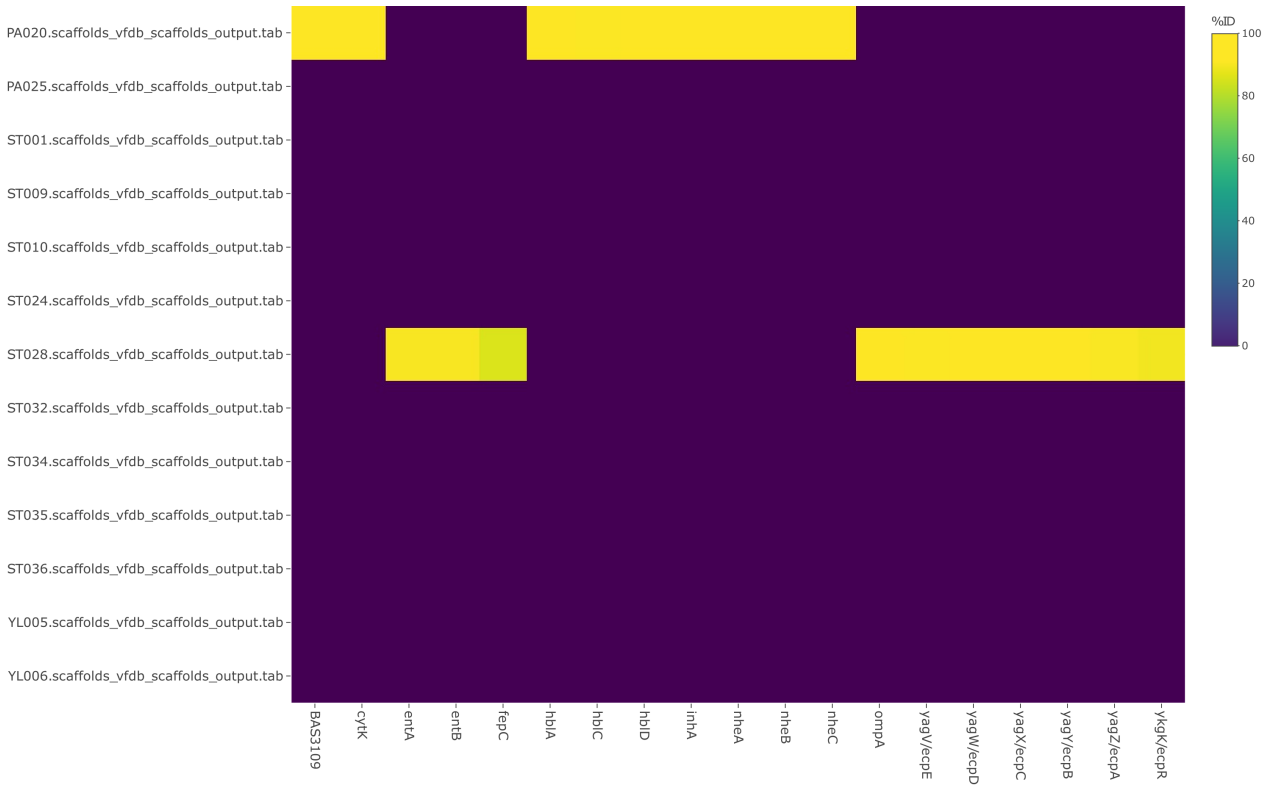

Figure 8. A report of virulence-associated genes (VAGs) against virulence factor database (VFDB). ID, identity.

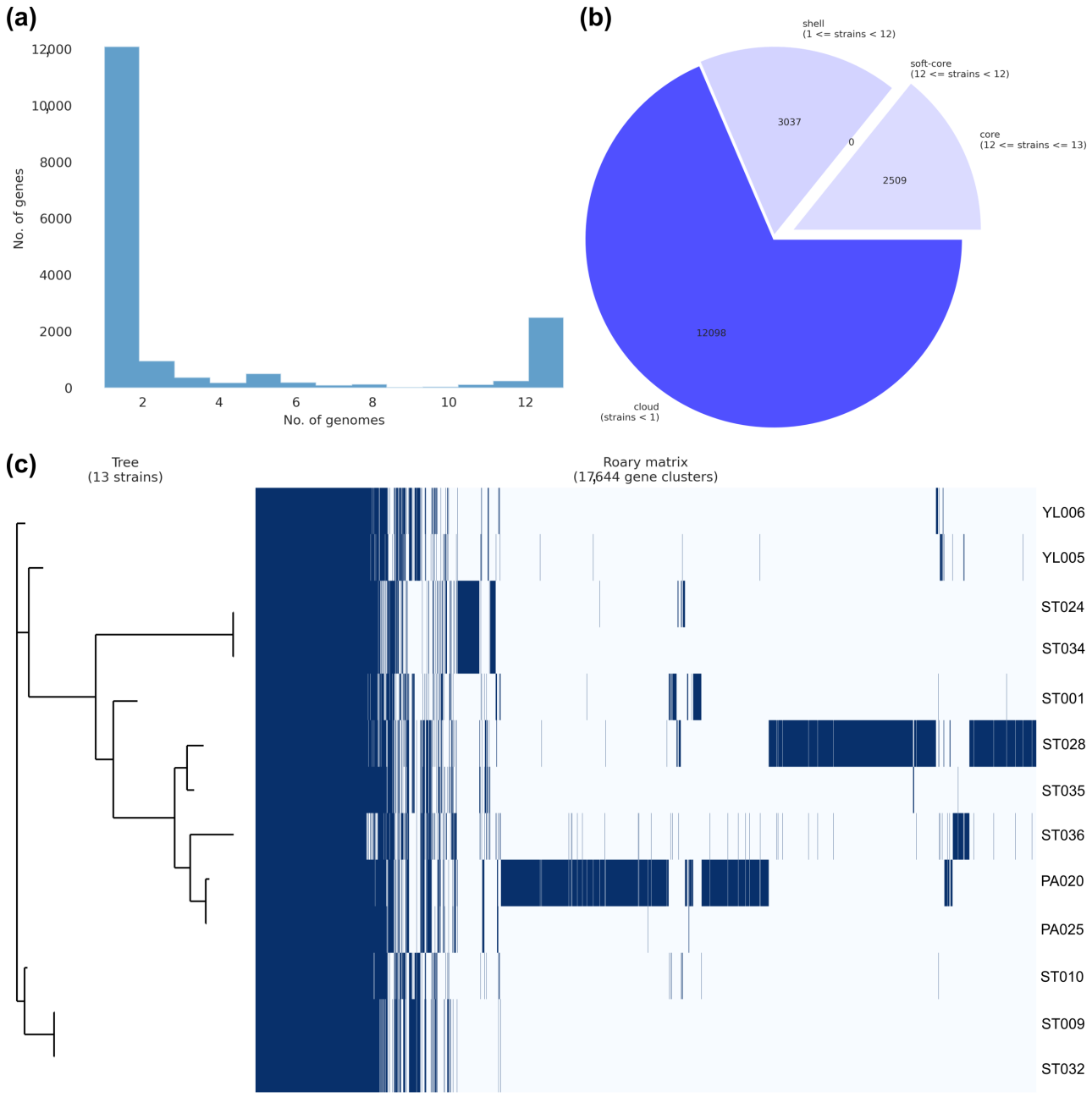

Figure 9. A report of pan-genome analysis by Roary. The frequency of genes versus the number of genomes (**a**), the breakdown of genes and the number of gene-presented isolates (**b**), and the phylogenetic tree against a matrix of present (blue) and absent (light blue) genes among core and accessory genomes (**c**) were exhibited.

3.2.5. Other Analysis

In BacSeq, we also provide bioinformatics tools for identifying CAZyme and CRISPR-Cas systems, and the results are reported in the Other Analysis button. The CAZyme reported annotated genes that encoded for the families of carbohydrate-active enzymes. Carbohydrate-active enzymes are enzymes involved in the breakdown, biosynthesis, or modification of carbohydrates, and they play a crucial role in various biological processes, including digestion, microbial metabolism, and the degradation of complex carbohydrates [46]. This tool is beneficial for analyzing WGS data of potential bacteria, especially probiotic strains. It can provide insights into their ability to metabolize and interact with different carbohydrates. This information is valuable for understanding the potential health benefits of probiotics, as carbo-

hydrates are a significant component of the diet and can influence various aspects of human health, including the gut microbiota composition and metabolic activities [47,48]. Meanwhile, CRISPRCasFinder reported the location of CRISPR-Cas systems (bacterial adaptive immune system), which included the sequences of direct repeats and spacers as well as the types of Cas gene groups. The presence of this system could imply the adaptive evolution of the studied genomes to foreign genetic elements, especially bacteriophage genomes [49]. This tool is good for bacterial evolution study as well, as it could be the model of genetic engineering [50,51]. Furthermore, the identification of bacteriocin-encoding genes and restriction–modification (R-M) sites will be implemented in the future versions of BacSeq.

3.3. Hybrid Library Assembly for Complete Genome Analysis

Long-read WGS is commonly beneficial for studying the complete genome of particular bacteria because it can be used to distinguish bacterial chromosome(s) and plasmids. The analytical steps in BacSeq were almost similar, except for genome assembly. Here, we used Unicycler for assembling the bacterial genome using both SRS and LRS data with a hybrid assembly method. The complete genome probably includes the chromosome and plasmids, which can be separated into distinct contigs or fragments. The separation of chromosome and plasmid sequences within a complete genome assembly enhances the understanding of the bacterial genomic structures, facilitates comparative genomics studies, enables functional analysis of plasmid-borne elements, and supports various applications in genetic engineering and biotechnology.

3.4. Limitation of BacSeq

In this study, we provide BacSeq as an automated pipeline for analyzing WGS data of bacterial genomes. However, the main limitation of BacSeq is that when the user runs the program and contaminant sequences occur in the analysis, the program cannot exclude contaminating sequences from the studied genome. We suggest using Kraken [52] or other tools to identify the contaminated sequences and remove them from the analysis, as demonstrated in our previous study [38]. Nevertheless, we recommend that the researcher avoid the contamination in the bacterial culture and pick a single isolated pure colony for genomic DNA extraction before performing WGS and bioinformatics analysis. This limitation will be addressed in future versions of BacSeq.

4. Conclusions

BacSeq is an open-source comprehensive pipeline integrating various bioinformatics tools for analyzing WGS data of bacterial genomes that research communities can easily install and implement on laptops and high-performance computers. BacSeq provided an automated bioinformatics workflow starting from genome assembly, annotation, and antimicrobial resistance gene identification to comparative genome analysis. Furthermore, BacSeq can generate comprehensive reports and plots in a web form which could help users simply explore and extract interesting information from the analysis.

Supplementary Materials: The following supporting information can be downloaded at: https://www.mdpi.com/article/10.3390/microorganisms11071769/s1, Table S1: a report of FastQC results.

Author Contributions: Conceptualization, K.S. (Komwit Surachat); methodology, K.S. (Komwit Surachat), A.C., K.S. (Kamonnut Singkhamanan)., and K.J.; software, K.S. (Komwit Surachat) and K.J.; validation, K.S. (Komwit Surachat), A.C., K.S. (Kamonnut Singkhamanan), and K.J.; formal analysis, K.S. (Komwit Surachat), A.C., K.S. (Kamonnut Singkhamanan), and K.J.; investigation, K.S. (Komwit Surachat), A.C., K.S. (Kamonnut Singkhamanan), and K.J.; resources, K.S. (Komwit Surachat); data curation, K.S. (Komwit Surachat), A.C., and K.S. (Kamonnut Singkhamanan); writing—original draft preparation, A.C. and K.S. (Komwit Surachat); writing—review and editing, K.S. (Komwit Surachat), A.C., K.S. (Kamonnut Singkhamanan), K.J., S.C., and R.P.; visualization, K.S. (Komwit Surachat), A.C., K.S. (Kamonnut Singkhamanan), and K.J.; supervision, K.S. (Komwit Surachat); project administration, K.S.; funding acquisition, K.S. (Komwit Surachat). All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Faculty of Science, Prince of Songkla University, Thailand (grant number SCI64040135) and the National Science, Research and Innovation Fund (NSRF) and Prince of Songkla University, Thailand (grant number MED6505096b). In addition, this work was also supported by the Postdoctoral Fellowship from Prince of Songkla University, Thailand.

Data Availability Statement: The BacSeq pipeline with its detailed user manual is publicly available at https://github.com/mecobpsu/bacseq (accessed on 5 May 2023).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Mardis, E.R. The impact of next-generation sequencing technology on genetics. *Trends Genet.* **2008**, *24*, 133–141. [CrossRef] [PubMed]
2. Mardis, E.R. DNA sequencing technologies: 2006–2016. *Nat. Protoc.* **2017**, *12*, 213–218. [CrossRef] [PubMed]
3. Pevsner, J. *Bioinformatics and Functional Genomics*; John Wiley & Sons: Hoboken, NJ, USA, 2015.
4. Quijada, N.M.; Rodríguez-Lázaro, D.; Eiros, J.M.; Hernández, M. TORMES: An automated pipeline for whole bacterial genome analysis. *Bioinformatics* **2019**, *35*, 4207–4212. [CrossRef]
5. Ewels, P.; Magnusson, M.; Lundin, S.; Kaller, M. MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics* **2016**, *32*, 3047–3048. [CrossRef]
6. Chen, S.; Zhou, Y.; Chen, Y.; Gu, J. fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **2018**, *34*, i884–i890. [CrossRef]
7. Prjibelski, A.; Antipov, D.; Meleshko, D.; Lapidus, A.; Korobeynikov, A. Using SPAdes De Novo Assembler. *Curr. Protoc. Bioinform.* **2020**, *70*, e102. [CrossRef]
8. Gurevich, A.; Saveliev, V.; Vyahhi, N.; Tesler, G. QUAST: Quality assessment tool for genome assemblies. *Bioinformatics* **2013**, *29*, 1072–1075. [CrossRef]
9. Waterhouse, R.M.; Seppey, M.; Simao, F.A.; Manni, M.; Ioannidis, P.; Klioutchnikov, G.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Mol. Biol. Evol.* **2018**, *35*, 543–548. [CrossRef]
10. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **2014**, *30*, 2068–2069. [CrossRef] [PubMed]
11. Cantalapiedra, C.P.; Hernandez-Plaza, A.; Letunic, I.; Bork, P.; Huerta-Cepas, J. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. *Mol. Biol. Evol.* **2021**, *38*, 5825–5829. [CrossRef] [PubMed]
12. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput. Biol.* **2011**, *7*, e1002195. [CrossRef]
13. Buchfink, B.; Reuter, K.; Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* **2021**, *18*, 366–368. [CrossRef]
14. Steinegger, M.; Soding, J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.* **2017**, *35*, 1026–1028. [CrossRef]
15. Hyatt, D.; Chen, G.L.; Locascio, P.F.; Land, M.L.; Larimer, F.W.; Hauser, L.J. Prodigal: Prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* **2010**, *11*, 119. [CrossRef] [PubMed]
16. Tatusov, R.L.; Galperin, M.Y.; Natale, D.A.; Koonin, E.V. The COG database: A tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.* **2000**, *28*, 33–36. [CrossRef]
17. Harris, M.A.; Clark, J.; Ireland, A.; Lomax, J.; Ashburner, M.; Foulger, R.; Eilbeck, K.; Lewis, S.; Marshall, B.; Mungall, C.; et al. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.* **2004**, *32*, D258–D261.
18. Finn, R.D.; Bateman, A.; Clements, J.; Coggill, P.; Eberhardt, R.Y.; Eddy, S.R.; Heger, A.; Hetherington, K.; Holm, L.; Mistry, J.; et al. Pfam: The protein families database. *Nucleic Acids Res.* **2014**, *42*, D222–D230. [CrossRef]
19. Kanehisa, M.; Goto, S. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* **2000**, *28*, 27–30. [CrossRef]
20. Alcock, B.P.; Raphenya, A.R.; Lau, T.T.Y.; Tsang, K.K.; Bouchard, M.; Edalatmand, A.; Huynh, W.; Nguyen, A.V.; Cheng, A.A.; Liu, S.; et al. CARD 2020: Antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **2020**, *48*, D517–D525. [CrossRef]
21. Bortolaia, V.; Kaas, R.S.; Ruppe, E.; Roberts, M.C.; Schwarz, S.; Cattoir, V.; Philippon, A.; Allesoe, R.L.; Rebelo, A.R.; Florensa, A.F.; et al. ResFinder 4.0 for predictions of phenotypes from genotypes. *J. Antimicrob. Chemother.* **2020**, *75*, 3491–3500. [CrossRef]
22. Gupta, S.K.; Padmanabhan, B.R.; Diene, S.M.; Lopez-Rojas, R.; Kempf, M.; Landraud, L.; Rolain, J.-M. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **2014**, *58*, 212–220. [CrossRef]
23. Doster, E.; Lakin, S.M.; Dean, C.J.; Wolfe, C.; Young, J.G.; Boucher, C.; Belk, K.E.; Noyes, N.R.; Morley, P.S. MEGARes 2.0: A database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. *Nucleic Acids Res.* **2020**, *48*, D561–D569.
24. Liu, B.; Zheng, D.D.; Jin, Q.; Chen, L.H.; Yang, J. VFDB 2019: A comparative pathogenomic platform with an interactive web interface. *Nucleic Acids Res.* **2019**, *47*, D687–D692. [CrossRef]

25. Carattoli, A.; Hasman, H. PlasmidFinder and In Silico pMLST: Identification and Typing of Plasmid Replicons in Whole-Genome Sequencing (WGS). *Methods Mol. Biol.* **2020**, *2075*, 285–294.
26. Siguier, P.; Perochon, J.; Lestrade, L.; Mahillon, J.; Chandler, M. ISfinder: The reference centre for bacterial insertion sequences. *Nucleic Acids Res.* **2006**, *34*, D32–D36. [CrossRef]
27. Yin, Y.; Mao, X.; Yang, J.; Chen, X.; Mao, F.; Xu, Y. dbCAN: A web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **2012**, *40*, W445–W451. [CrossRef]
28. Zhang, H.; Yohe, T.; Huang, L.; Entwistle, S.; Wu, P.; Yang, Z.; Busk, P.K.; Xu, Y.; Yin, Y. dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **2018**, *46*, W95–W101. [CrossRef]
29. Grissa, I.; Vergnaud, G.; Pourcel, C. CRISPRFinder: A web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res.* **2007**, *35*, W52–W57. [CrossRef] [PubMed]
30. Page, A.J.; Cummins, C.A.; Hunt, M.; Wong, V.K.; Reuter, S.; Holden, M.T.G.; Fookes, M.; Falush, D.; Keane, J.A.; Parkhill, J. Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **2015**, *31*, 3691–3693. [CrossRef]
31. Page, A.J.; Taylor, B.; Delaney, A.J.; Soares, J.; Seemann, T.; Keane, J.A.; Harris, S.R. SNP-sites: Rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2016**, *2*, e000056. [CrossRef] [PubMed]
32. Price, M.N.; Dehal, P.S.; Arkin, A.P. FastTree: Computing large minimum evolution trees with profiles instead of a distance matrix. *Mol. Biol. Evol.* **2009**, *26*, 1641–1650. [CrossRef]
33. Antunes, L.; Visca, P.; Towner, K.J. Acinetobacter baumannii: Evolution of a global pathogen. *Pathog. Dis.* **2014**, *71*, 292–301. [CrossRef]
34. Partridge, S.R.; Kwong, S.M.; Firth, N.; Jensen, S.O. Mobile genetic elements associated with antimicrobial resistance. *Clin. Microbiol. Rev.* **2018**, *31*, e00088-17. [CrossRef] [PubMed]
35. CDC. Antibiotic Resistance Threats in the United States, 2019 (2019 AR Threats Report), Centers for Disease Control and Prevention (CDC), Atlanta, GA. 2019. Available online: https://www.cdc.gov/drugresistance/Biggest-Threats.html (accessed on 4 January 2020).
36. Andrews, S. *FastQC: A Quality Control Tool for High Throughput Sequence Data*; Babraham Bioinformatics, Babraham Institute: Cambridge, UK, 2010.
37. Treangen, T.J.; Salzberg, S.L. Repetitive DNA and next-generation sequencing: Computational challenges and solutions. *Nat. Rev. Genet.* **2012**, *13*, 36–46. [CrossRef] [PubMed]
38. Chukamnerd, A.; Singkhamanan, K.; Chongsuvivatwong, V.; Palittapongarnpim, P.; Doi, Y.; Pomwised, R.; Sakunrang, C.; Jeenkeawpiam, K.; Yingkajorn, M.; Chusri, S. Whole-genome analysis of carbapenem-resistant Acinetobacter baumannii from clinical isolates in Southern Thailand. *Comput. Struct. Biotechnol. J.* **2022**, *20*, 545–558. [CrossRef]
39. Hernández-Díaz, E.A.; Vázquez-Garcidueñas, M.S.; Negrete-Paz, A.M.; Vázquez-Marrufo, G. Comparative Genomic Analysis Discloses Differential Distribution of Antibiotic Resistance Determinants between Worldwide Strains of the Emergent ST213 Genotype of Salmonella Typhimurium. *Antibiotics* **2022**, *11*, 925. [CrossRef] [PubMed]
40. Tsui, C.K.; Abid, F.B.; McElheny, C.L.; Almuslamani, M.; Omrani, A.S.; Doi, Y. Genomic epidemiology revealed the emergence and worldwide dissemination of ST383 carbapenem-resistant hypervirulent Klebsiella pneumoniae and hospital acquired infections of ST196 Klebsiella quasipneumoniae in Qatar. *bioRxiv* **2022**. [CrossRef]
41. Alzahrani, K.O.; Al-Reshoodi, F.M.; Alshdokhi, E.A.; Alhamed, A.S.; Al Hadlaq, M.A.; Mujallad, M.I.; Mukhtar, L.E.; Alsufyani, A.T.; Alajlan, A.A.; Al Rashidy, M.S. Antimicrobial resistance and genomic characterization of Salmonella enterica isolates from chicken meat. *Front. Microbiol.* **2023**, *14*, 1104164. [CrossRef]
42. Bloomfield, S.; Duong, V.T.; Tuyen, H.T.; Campbell, J.I.; Thomson, N.R.; Parkhill, J.; Le Phuc, H.; Chau, T.T.H.; Maskell, D.J.; Perron, G.G. Mobility of antimicrobial resistance across serovars and disease presentations in non-typhoidal Salmonella from animals and humans in Vietnam. *Microb. Genom.* **2022**, *8*, 000798. [CrossRef]
43. Mira, A.; Martín-Cuadrado, A.B.; D'Auria, G.; Rodríguez-Valera, F. The bacterial pan-genome: A new paradigm in microbiology. *Int. Microbiol.* **2010**, *13*, 45–57.
44. Polz, M.F.; Alm, E.J.; Hanage, W.P. Horizontal gene transfer and the evolution of bacterial and archaeal population structure. *Trends Genet.* **2013**, *29*, 170–175. [CrossRef]
45. Palmer, M.; Venter, S.N.; Coetzee, M.P.; Steenkamp, E.T. Prokaryotic species are sui generis evolutionary units. *Syst. Appl. Microbiol.* **2019**, *42*, 145–158. [CrossRef]
46. Cerqueira, F.M.; Photenhauer, A.L.; Pollet, R.M.; Brown, H.A.; Koropatkin, N.M. Starch digestion by gut bacteria: Crowdsourcing for carbs. *Trends Microbiol.* **2020**, *28*, 95–108. [CrossRef]
47. Surachat, K.; Kantachote, D.; Deachamag, P.; Wonglapsuwan, M. Genomic insight into Pediococcus acidilactici HN9, a potential probiotic strain isolated from the traditional Thai-style fermented Beef Nhang. *Microorganisms* **2020**, *9*, 50. [CrossRef]
48. Rowland, I.; Gibson, G.; Heinken, A.; Scott, K.; Swann, J.; Thiele, I.; Tuohy, K. Gut microbiota functions: Metabolism of nutrients and other food components. *Eur. J. Nutr.* **2018**, *57*, 1–24. [CrossRef] [PubMed]
49. Amitai, G.; Sorek, R. CRISPR–Cas adaptation: Insights into the mechanism of action. *Nat. Rev. Microbiol.* **2016**, *14*, 67. [CrossRef]
50. Chevallereau, A.; Meaden, S.; van Houte, S.; Westra, E.R.; Rollie, C. The effect of bacterial mutation rate on the evolution of CRISPR-Cas adaptive immunity. *Philos. Trans. R. Soc. B* **2019**, *374*, 20180094. [CrossRef]

51. De la Fuente-Núñez, C.; Lu, T.K. CRISPR-Cas9 technology: Applications in genome engineering, development of sequence-specific antimicrobials, and future prospects. *Integr. Biol.* **2017**, *9*, 109–122. [CrossRef] [PubMed]
52. Wood, D.E.; Salzberg, S.L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **2014**, *15*, 1–12. [CrossRef] [PubMed]