SOFTWARE

Open Access

AMRViz enables seamless genomics analysis and visualization of antimicrobial resistance

Duc Quang Le[1,2*], Son Hoang Nguyen[1*], Tam Thi Nguyen[3], Canh Hao Nguyen[4], Tho Huu Ho[5,6], Nam S. Vo[7], Trang Nguyen[1], Hoang Anh Nguyen[1] and Minh Duc Cao[1*]

*Correspondence:
quangld@huce.edu.vn;
dr.sonhoangnguyen@gmail.com;
minhduc.cao@gmail.com

[1] AMROMICS JSC, Nghe An, Vietnam
[2] Faculty of IT, Hanoi University of Civil Engineering, Hanoi, Vietnam
[3] Oxford University Clinical Research Unit, Hanoi, Vietnam
[4] Bioinformatics Center, Institute for Chemical Research, Kyoto University, Kyoto, Japan
[5] Department of Medical Microbiology, The 103 Military Hospital, Vietnam Military Medical University, Hanoi, Vietnam
[6] Department of Genomics and Cytogenetics, Institute of Biomedicine and Pharmacy, Vietnam Military Medical University, Hanoi, Vietnam
[7] Center for Biomedical Informatics, Vingroup Big Data Institute, Hanoi, Vietnam

Abstract

We have developed AMRViz, a toolkit for analyzing, visualizing, and managing bacterial genomics samples. The toolkit is bundled with the current best practice analysis pipeline allowing researchers to perform comprehensive analysis of a collection of samples directly from raw sequencing data with a single command line. The analysis results in a report showing the genome structure, genome annotations, antibiotic resistance and virulence profile for each sample. The pan-genome of all samples of the collection is analyzed to identify core- and accessory-genes. Phylogenies of the whole genome as well as all gene clusters are also generated. The toolkit provides a web-based visualization dashboard allowing researchers to interactively examine various aspects of the analysis results. Availability: AMRViz is implemented in Python and NodeJS, and is publicly available under open source MIT license at https://github.com/amromics/amrviz.

Keywords: Microbial genomics, Visualization, Antimicrobial resistance

Introduction

With the advances of high-throughput sequencing, the number of bacterial genomes sequenced has increased exponentially over the last decade. Studying the genomics of pathogenic bacteria, especially those with antibiotic resistance can provide insights into the development and transmission of antibiotic resistance. Pan-genome analysis [1] has increasingly become the method of choice for population genomics. Here, the pan-genome refers to the entire set of gene families of a given clade; core genes, the set of gene families presenting in the majority of isolates in the clade, are considered to characterize the clade whereas accessory genes, those present in a small number of organisms, provide a view of the variation and evolutionary trajectories [2]. Pan-genome analysis is particularly relevant to microbial genomics because of the plasticity of their genomes; they can rapidly acquire genes from other organisms, especially in an environment with evolutionary pressure such as with antibiotics [3]. The ability to study the complete genetic information of a collection of large number of bacterial genomes also provides the potential to generate insights into the pathogenic genotype/phenotype relationships [4, 5], pathogenic virulence transmissibility [6] and antibiotic resistance tracking [7].

While there are numerous software tools for analyzing bacterial genomics data, for example [8–11], they are often designed for a specific task of the large analysis pipeline. These tools often require bioinformatics expertise to set up, and to connect them together to make up an end-to-end pipeline. Many tools require users to resort to multiple external websites such as [12–14] for the analysis. As a result, bacterial genomics analysis, especially for a collection of hundreds of samples, involves many manual steps that can be time-consuming and error-prone. Recently, several computational pipelines have been developed by stitching the analysis tools into a workflow to facilitate the analysis of microbial genomic data, notably Nullarbor [15], Bactopia [16], and ASA^{3}P [17]. However, these software pipelines usually require high-end computational infrastructures and are not scalable for large collections of samples [18]. In addition, there is a lack of integrated tools for visualization and interpretation of analysis results.

To address these shortcomings, we developed AMRViz, a software package to manage the analysis of drug-resistant bacteria collections. AMRViz is bundled with an end-to-end analysis pipeline consisting of the considered best practice methodologies for microbial genomics to streamline data processing. AMRViz also provides a comprehensive visualization of important aspects of the analysis and a dashboard to manage samples and collections. The analysis of a collection of bacterial samples can be initiated with a single command line. The analyses and visualization will be generated automatically. AMRViz is cross-platform, and can be conveniently installed, managed and operated with modest hardware requirements and minimum bioinformatics support.

Methods and implementation

AMRViz toolkit consists of three components, namely (1) a pipeline for genomics analysis of a bacterial collection, (2) a platform for visualization of the analysis results, and (3) a system to manage samples and collections.

Analysis pipeline

The analysis pipeline takes in as input a list of samples for a collection, and performs genomics analysis on each single sample as well as pan-genome analysis on the collection. The samples are listed in a tsv file, one sample per line. Metadata associated with the samples such as phenotypic information can be included in the tsv file. Data input for each sample can be in the form of sequencing data (fastq format), assembly sequence (fasta format) or annotation (gff3 format). AMRViz supports the most common sequencing technologies including Illumina, Pacbio and Oxford Nanopore.

Under the hood, AMRViz utilizes AMRomics [18], an efficient workflow to generate analysis results. Briefly, AMRomics performs single sample analysis for each sample, and then performs collection analysis for the collection (Fig. 1a). If the input data for a sample are sequencing data, the pipeline applies the appropriate assembly algorithm for the sequencing technology: SKESA [8] or SPAdes [9] for Illumina data and Flye [19] for Pacbio and Oxford Nanopore data (Fig. 1b). The assembly of the sample is then annotated with Prokka [10] to identify protein coding genes and transfer RNA sequences and their functions. The sample assembly is also subject to antibiotic resistance and virulence gene analysis using BLAST [20] against major relevant gene databases, and to multi-locus strain typing analysis with pubMLST database [12].

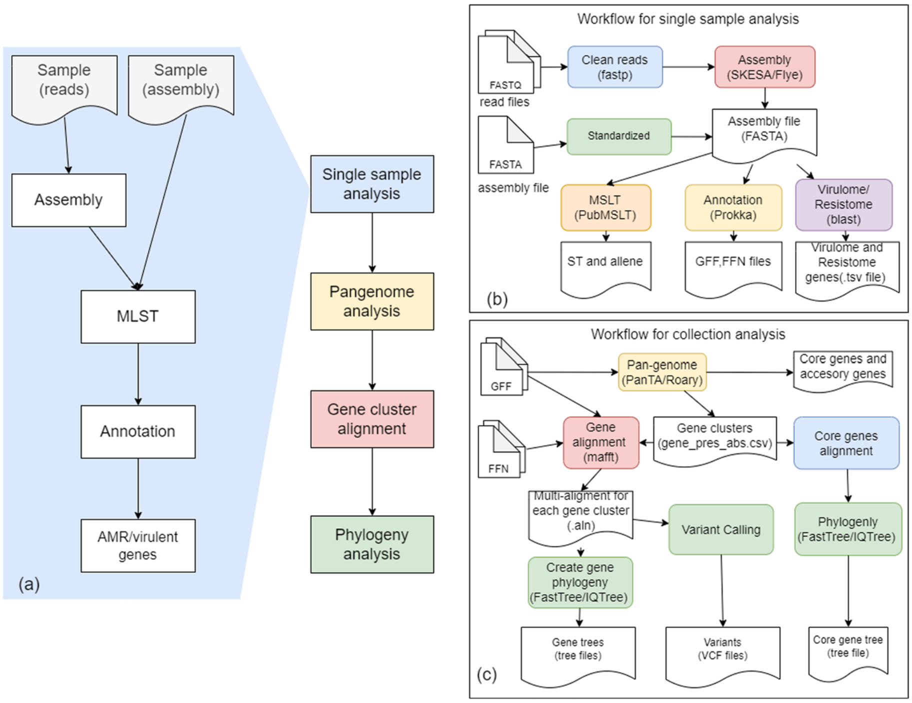
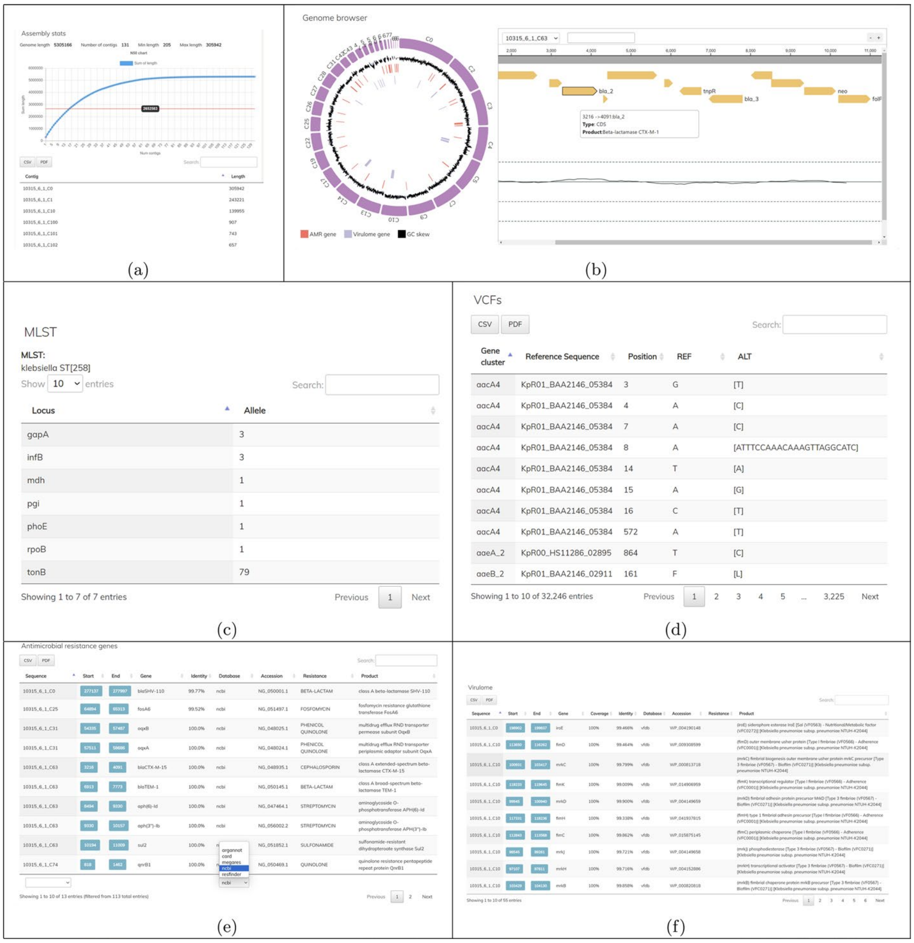


Fig. 1 Schematic description of the AMRViz analysis pipeline. **a** Overall workflow; **b** workflow for single sample analysis, and **c** workflow for collection analysis

After all samples in the collection have undergone single sample analysis, the pipeline performs genomics analysis of the collection (Fig. 1c). Specifically, PanTA [21] or Roary [11] is used for pan-genome analysis of the collection to identify core- and accessory-genes. For each gene cluster, the gene sequences are translated to amino acid sequences which are used to generate the gene multiple sequence alignment (MSA) using MAFFT [22]. The alignment is performed on protein sequences to make use of the functional information encoded in protein sequences. Subsequently, the nucleotide MSA of the gene is inferred from the protein alignment and is used for phylogenetic analysis with FastTree 2 [23] or IQ-TREE2 [24] to utilize the information in the synonymous mutations. In addition, AMRViz builds the phylogeny of the collection from the concatenation of the MSAs of all core genes. Finally, AMRViz calls variants of all the samples in the collection against the reference pangenome, termed pan-SNPs, a novel concept introduced in AMRomics [18]. The variant profile of a sample is reported in a VCF file.

Visualization platform

The platform exports all analysis results to a web server for visualization via a web browser. For each sample, AMRViz displays and visualizes detailed genomic information allowing users to interactively investigate the genomics analysis results (Fig. 2). Specially, the assembly statistics and a chart showing the length distribution of the contigs are presented (Fig. 2a). AMRViz also presents the genome structure of the sample via a circos chart and a genome browser, allowing users to have both an overview and detailed views of the genome (Fig. 2b). From the circos chart, users can view the genome structure, the contigs and their CG content, and the high-level locations of resistance genes and



Assembly stats

Genome length 5305166 Number of contigs 131 Min length 205 Max length 305942

CSV PDF Search:

Contig	Length
10315_6_1_C0	305942
10315_6_1_C1	243221
10315_6_1_C10	139955
10315_6_1_C100	907
10315_6_1_C101	743
10315_6_1_C102	657

(a)

Genome browser

3216 ->4091:bla_2
Type: CDS
Product:Beta-lactamase CTX-M-1

(b)

MLST

MLST:
klebsiella ST[258]

Show 10 entries Search:

Locus	Allele
gapA	3
infB	3
mdh	1
pgi	1
phoE	1
rpoB	1
tonB	79

Showing 1 to 7 of 7 entries Previous 1 Next

(c)

VCFs

CSV PDF Search:

Gene cluster	Reference Sequence	Position	REF	ALT
aacA4	KpR01_BAA2146_05384	3	G	[T]
aacA4	KpR01_BAA2146_05384	4	A	[C]
aacA4	KpR01_BAA2146_05384	7	A	[C]
aacA4	KpR01_BAA2146_05384	8	A	[ATTTCCAAACAAAGTTAGGCATC]
aacA4	KpR01_BAA2146_05384	14	T	[A]
aacA4	KpR01_BAA2146_05384	15	A	[G]
aacA4	KpR01_BAA2146_05384	16	C	[T]
aacA4	KpR01_BAA2146_05384	572	A	[T]
aaeA_2	KpR00_HS11286_02895	864	T	[C]
aaeB_2	KpR01_BAA2146_02911	161	F	[L]

Showing 1 to 10 of 32,246 entries Previous 1 2 3 4 5 … 3,225 Next

(d)

Antimicrobial resistance genes

CSV PDF Search:

Sequence	Start	End	Gene	Identity	Database	Accession	Resistance	Product
10315_6_1_C0	277137	277997	blaSHV-110	99.77%	ncbi	NG_050001.1	BETA-LACTAM	class A beta-lactamase SHV-110
10315_6_1_C25	64894	65313	fosA6	99.52%	ncbi	NG_051497.1	FOSFOMYCIN	fosfomycin resistance glutathione transferase FosA6
10315_6_1_C31	54335	57487	oqxB	100.0%	ncbi	NG_048025.1	PHENICOL QUINOLONE	multidrug efflux RND transporter permease subunit OqxB
10315_6_1_C31	57511	58686	oqxA	100.0%	ncbi	NG_048024.1	PHENICOL QUINOLONE	multidrug efflux RND transporter periplasmic adaptor subunit OqxA
10315_6_1_C63	3216	4091	blaCTX-M-15	100.0%	ncbi	NG_048935.1	CEPHALOSPORIN	class A extended-spectrum beta-lactamase CTX-M-15
10315_6_1_C63	6913	7773	blaTEM-1	100.0%	ncbi	NG_050145.1	BETA-LACTAM	class A broad-spectrum beta-lactamase TEM-1
10315_6_1_C63	8494	9330	aph(6)-Id	100.0%	ncbi	NG_047464.1	STREPTOMYCIN	aminoglycoside O-phosphotransferase APH(6)-Id
10315_6_1_C63	9330	10157	aph(3")-Ib	100.0%	ncbi	NG_056002.2	STREPTOMYCIN	aminoglycoside O-phosphotransferase APH(3")-Ib
10315_6_1_C63	10194	11009	sul2	100.0%	n	NG_051852.1	SULFONAMIDE	sulfonamide-resistant dihydropteroate synthase Sul2
10315_6_1_C74	818	1462	qnrB1	100.0%	n	NG_050469.1	QUINOLONE	quinolone resistance pentapeptide repeat protein QnrB1

Showing 1 to 10 of 13 entries (filtered from 113 total entries) Previous 1 2 Next

(e)

Virulome

CSV PDF Search:

Sequence	Start	End	Gene	Coverage	Identity	Database	Accession	Resistance	Product
10315_6_1_C0	198902	199837	iroE	100%	99.466%	vfdb	WP_004190148		(iroE) siderophore esterase IroE [Sal (VF0563) - Nutritional/Metabolic factor (VFC0272)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	113650	116262	fimD	100%	99.464%	vfdb	WP_009308599		(fimD) outer membrane usher protein [Type I fimbriae (VF0566) - Adherence (VFC0001)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	100931	103417	mrkC	100%	99.799%	vfdb	WP_000813718		(mrkC) fimbrial biogenesis outer membrane usher protein mrkC precursor [Type 3 fimbriae (VF0567) - Biofilm (VFC0271)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	118233	119645	fimK	100%	99.009%	vfdb	WP_014906959		(fimK) transcriptional regulator [Type I fimbriae (VF0566) - Adherence (VFC0001)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	99945	100940	mrkD	100%	99.900%	vfdb	WP_004149659		(mrkD) fimbrial adhesin protein precursor MrkD [Type 3 fimbriae (VF0567) - Biofilm (VFC0271)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	117331	118236	fimH	100%	99.338%	vfdb	WP_041937815		(fimH) type 1 fimbrial adhesin precursor [Type I fimbriae (VF0566) - Adherence (VFC0001)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	112843	113568	fimC	100%	99.862%	vfdb	WP_015875145		(fimC) periplasmic chaperone [Type I fimbriae (VF0566) - Adherence (VFC0001)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	98545	99261	mrkJ	100%	99.721%	vfdb	WP_004149658		(mrkJ) phosphodiesterase [Type 3 fimbriae (VF0567) - Biofilm (VFC0271)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	97107	97811	mrkH	100%	99.716%	vfdb	WP_004152886		(mrkH) transcriptional activator [Type 3 fimbriae (VF0567) - Biofilm (VFC0271)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]
10315_6_1_C10	103429	104130	mrkB	100%	99.858%	vfdb	WP_000820818		(mrkB) fimbrial chaperone protein mrkB precursor [Type 3 fimbriae (VF0567) - Biofilm (VFC0271)] [Klebsiella pneumoniae subsp. pneumoniae NTUH-K2044]

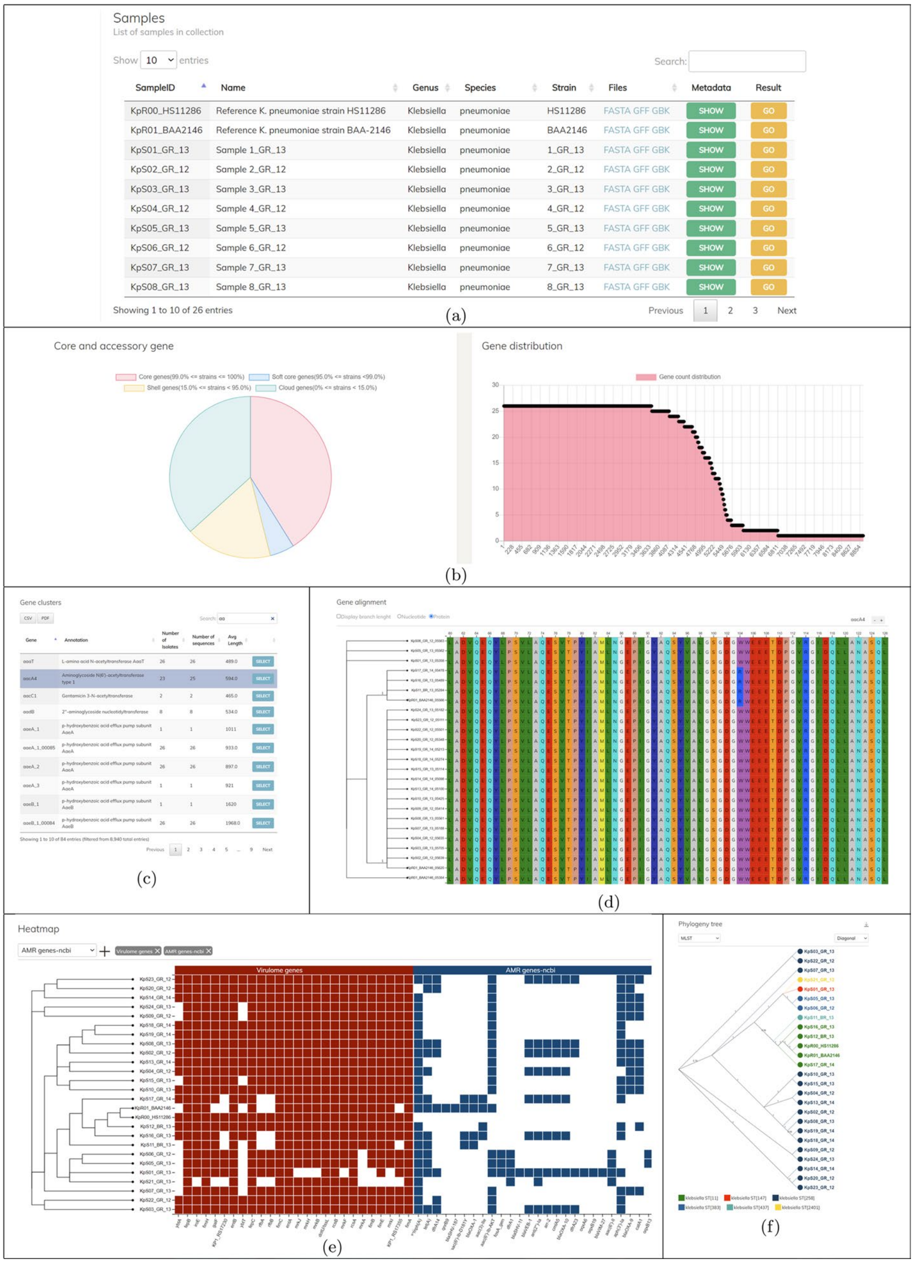
Showing 1 to 10 of 55 entries Previous 1 2 3 4 5 6 Next

(f)

Fig. 2 Visualization of the single analysis of a sample. **a** Showing of the assembly statistics. **b** Depiction of the genome as a circos and the genome browser. **c** MLST profile of the sample. **d** Listing of antimicrobial resistance genes, users can select the source of resistance gene database to show. **e** Listing of virulence genes

virulence genes on the genome. Users can select a contig or a gene of interest presented in the genome browser. AMRViz also shows the sample's metadata, gene annotation, MLST profile (Fig. 2c) and variant against the pan-genome (Fig. 2d). Resistance genes and virulence genes are displayed in tables with paging and sorting for easy navigation (Fig. 2e, f). As AMRViz uses several databases of resistance genes for resistome analysis, users can select which source of resistance genes to display. Users can also search for resistance genes and virulence genes with keywords via the search boxes. When users select one of the resistance/virulence genes, the genome browser is zoomed to the location of the selected gene and its contig, enabling visual inspection of the gene and the vicinity of its position.

The pan-genome analysis of the whole collection is also be visualized, allowing interactive detailed investigation (Fig. 3). At the top of the page, AMRViz lists all samples in the collection (Fig. 3a). In this section, users can sort and search the list to find the sample of interest, view its metadata as well as access its single sample analysis result. AMRViz summarizes statistics of the core- and pan-genes in a pie chart and a bar graph (Fig. 3b).



Samples

List of samples in collection

Show 10 entries Search:

SampleID	Name	Genus	Species	Strain	Files	Metadata	Result
KpR00_HS11286	Reference K. pneumoniae strain HS11286	Klebsiella	pneumoniae	HS11286	FASTA GFF GBK	SHOW	GO
KpR01_BAA2146	Reference K. pneumoniae strain BAA-2146	Klebsiella	pneumoniae	BAA2146	FASTA GFF GBK	SHOW	GO
KpS01_GR_13	Sample 1_GR_13	Klebsiella	pneumoniae	1_GR_13	FASTA GFF GBK	SHOW	GO
KpS02_GR_12	Sample 2_GR_12	Klebsiella	pneumoniae	2_GR_12	FASTA GFF GBK	SHOW	GO
KpS03_GR_13	Sample 3_GR_13	Klebsiella	pneumoniae	3_GR_13	FASTA GFF GBK	SHOW	GO
KpS04_GR_12	Sample 4_GR_12	Klebsiella	pneumoniae	4_GR_12	FASTA GFF GBK	SHOW	GO
KpS05_GR_13	Sample 5_GR_13	Klebsiella	pneumoniae	5_GR_13	FASTA GFF GBK	SHOW	GO
KpS06_GR_12	Sample 6_GR_12	Klebsiella	pneumoniae	6_GR_12	FASTA GFF GBK	SHOW	GO
KpS07_GR_13	Sample 7_GR_13	Klebsiella	pneumoniae	7_GR_13	FASTA GFF GBK	SHOW	GO
KpS08_GR_13	Sample 8_GR_13	Klebsiella	pneumoniae	8_GR_13	FASTA GFF GBK	SHOW	GO

Showing 1 to 10 of 26 entries

Previous 1 2 3 Next

(a)

Core and accessory gene
Core genes(99.0% <= strains <= 100%)
Soft core genes(95.0% <= strains <99.0%)
Shell genes(15.0% <= strains < 95.0%)
Cloud genes(0% <= strains < 15.0%)
Gene distribution
Gene count distribution
30
25
20
15
10
5
0
(b)

Gene clusters

CSV PDF Search: aa

Gene	Annotation	Number of Isolates	Number of sequences	Avg Length	
aaaT	L-amino acid N-acetyltransferase AaaT	26	26	489.0	SELECT
aacA4	Aminoglycoside N(6')-acetyltransferase type 1	23	25	594.0	SELECT
aacC1	Gentamicin 3-N-acetyltransferase	2	2	465.0	SELECT
aadB	2"-aminoglycoside nucleotidyltransferase	8	8	534.0	SELECT
aaeA_1	p-hydroxybenzoic acid efflux pump subunit AaeA	1	1	1011	SELECT
aaeA_1_00085	p-hydroxybenzoic acid efflux pump subunit AaeA	26	26	933.0	SELECT
aaeA_2	p-hydroxybenzoic acid efflux pump subunit AaeA	26	26	897.0	SELECT
aaeA_3	p-hydroxybenzoic acid efflux pump subunit AaeA	1	1	921	SELECT
aaeB_1	p-hydroxybenzoic acid efflux pump subunit AaeB	1	1	1620	SELECT
aaeB_1_00084	p-hydroxybenzoic acid efflux pump subunit AaeB	26	26	1968.0	SELECT

Showing 1 to 10 of 84 entries (filtered from 8,940 total entries)

Previous 1 2 3 4 5 … 9 Next

(c)

Fig. 3 Visualization of the analyses for a collection. **a** Visualization of the statistics of core- and accessory genes and the gene distribution. **b** Listing of the gene families in the collection. **c** The multiple alignment of a gene family is displayed when a gene family is selected. **d** The heatmap showing the presence and absence of resistance and virulence genes. **e** The phylogeny of the samples in the collection

It also lists all gene families in a table (Fig. 3) with sorting and searching functions. When a gene family is selected, the multiple alignment of the gene is displayed (Fig. 3d). Users can toggle between the nucleotide and protein modes for the multiple alignment and zoom into the columns of interest for a closer inspection. A gene phylogeny is presented on the side of the alignment. The phylogeny of the gene can also be visually compared to the core-gene phylogeny. User can observe the difference between the two trees if there is a crossover between the inheritance links in the gene tree (data not shown).

AMRViz also produces a heatmap showing the presence of resistance genes and virulence genes together with the phenotype and metadata provided in the input, allowing visual inspection of any associations between gene presence and phenotypes (Fig. 3e). The phylogeny tree from the core-gene alignment of all samples in the collection is also generated (Fig. 3f). This tool displays a full range of tree information in various styles and bootstrap values. Furthermore, users can highlight samples with different colors based on metadata. When a node is selected on the tree, its child-node samples will be highlighted on the heatmap and alignment tool.

The web application is developed on the VUE framework (https://vuejs.org) with well-modularized visualization tools and is easy to integrate with other systems. The visualization tools are built on top of the following graphics libraries and tools: chart.js (https://www.chartjs.org/), d3js (https://d3js.org), phylocanvas (http://phylocanvas.org), circosJS (https://github.com/nicgirault/circosJS).

Sample and collection management system

AMRViz provides a web-interface dashboard to manage dataset collections. Users can view and navigate all collections that have been analyzed. From the dashboard menu, users can also view the list of samples in the system. Both the collection and sample list views allow sorting by various fields and searching by keywords. They also provides access to analysis results and visualization for each collection/sample. Collections and samples can be added and updated with pipeline runs as described in the Analysis pipeline section.

Case study

A collection of clinical *Klebsiella pneumoniae*

We demonstrate AMRViz's ability to perform an end-to-end automatic analysis and to visualize analysis results of a genome collection with two case studies. In the first case study, we used the tool to re-analyze a dataset including 89 samples of *Klebsiella pneumoniae* collected from major hospitals in Kathmandu [25]. We included in the collection two additional reference genomes with Genbank accessions GCF_000240185 and GCF_000364385 in fasta format. While the pipeline did not require a reference genome as part of the analysis, including one or several reference genomes enables the use of known genes in the construction of the pan-reference genome [18]. We also included 11 samples that were sequenced by Oxford Nanopore and a sample by Pacbio technologies to show AMRViz's capability to handle the heterogeneity of data. Altogether, the collection contained 103 samples in various input types including sequencing assembly, short read and long read sequencing data. The details of the data of the case study are provided in *Availability of data and materials.*

Both the web server and the analysis were run on a laptop computer with a 20 hyper-thread CPU (Intel Core i7-1280P) and 32Gb of memory, running Ubuntu Linux 22.0. With one command line, we were able to process the whole dataset through the pipeline. The analysis took in total 12.54 h and consumed 10.01 Gb of memory to perform all analysis steps and import results into the web portal.

AMRViz was able to replicate the analysis performed by [25] as described in the original paper. We note that the MLST types were identical to the original paper, except that the cluster of ST1559 was identified as ST3395. We manually checked the sequences of the housekeeping genes for ST3395 in the MLST database and found they are identical to the genes in the cluster, thus confirming the accuracy of AMRViz. To ascertain that the discrepancy is not due to the error during assembly, we downloaded the assembly of one the samples in the cluster from Refseq database (Refseq accession GCF_000943155.1) and ran MLST typing using both our system and MLST server [26]. Again, both systems return ST3395 with 100% identity.

From the dashboard of the analysis results, AMRViz showed on the phylogeny diagram the formation of two major genetically distinct lineages, ST15 and ST3395 (Additional file 1: Figure S7). According to the results from [25] the ST15 isolates have S83F and D87A mutations in the gyrA gene and the S80I mutation in the parC gene. Using the alignment view tool (Additional file 1: Figures S3 and S4), these mutations can be easily seen when comparing the amino acid sequences of the isolates with the reference MGH78578. Furthermore, the gene tree comparing tool (Additional file 1: Figure S5) allowed us to see that among the ST15 isolates, 10315_6#28 has the gyrA gene closest to the reference. As shown in alignment view tool, 10315_6#28 has one less SNP than the rest of the ST15 isolates. For finding resistance genes, we used the AMRViz scanned across resistance gene databases such as ncbi, card, megares, argannot and found most of the resistance genes mentioned in [25]. The heatmap showed the relationship between resistance genes and drug resistance of isolates, by displaying both the presence of resistance genes and the resistance phenotype of each isolate. As an example, the relationship between *carbapenem* resistance and the bla-NDM1 gene was displayed (Additional file 1: Figure S6). Most of the isolates in the ST15 group with the bla-NDM1 gene were resistant to Imipenem and Meropenem.

For detailed analysis of the results of a single isolate, AMRViz provided a set of tools to visualize the result data in the form of interactive tables and graphs. As an example, the statistics of the assembly for sample 10315_6#17 was displayed (Additional file 1: Figure S1). The statistics included the number of contigs, the length of each contig and N50 statistics. The results of MLST analysis showed that this sample has ST 15. The list of resistance genes and virulence genes were presented in two interactive tables showing the sample harbored multiple resistance genes such as blaSHV, bla-NDM1,blaCTX-M-15, qnRB1, mrSE,armA, sul1,aadA2, strA, strB. We also found additional resistance genes that have not been reported previously in [25] such as fosA and tet(D). We could see the elements around this gene when looking for the location of gene bla-NDM-1 in this isolate using the genome browser (Additional file 1: Figure S1B).

A set of *Streptococcus pyogenes* from Victoria, Australia

The gram-positive cocci *S. pyogenes*, or group A streptococcus (GAS), linked with seasonal scarlet fever, has increasingly caused invasive infection (iGAS) in many Australian states over the past decades. There are ongoing efforts in genomic surveillance from national public health labs to understand the epidemiology and upregulated virulence factors of iGAS samples around the country [27, 28]. For demonstration purposes, we extract a small set of 72 samples from the project PRJNA872282—sequencing of *Streptococcus pyogenes emm1* from Victoria and Queensland between 2005 and 2020. Together with 4 other *S. pyogenes* assemblies, we ran the pipeline on this dataset and the results were added to the collection dashboard of the visualization interface for browsing. The analysis took 2.16 h on the aforementioned computer, requiring 9.4 Gb of memory. The *emm* type for each sample can be identified by searching the gene from the Samples view. The phylogeny tree, on the other hand, can show the relationships between collected samples, highlighted by the MLST to aid the epidemiological investigation. In addition, the heatmap can depict the presence/absence of the virulence genes, such as the superantigens specA/B/C/G/J, for each and every sample in the set.

Discussion

We have presented AMRViz, a toolkit for analyzing, visualizing, and managing large collections of bacterial genomes. The toolkit is bundled with the current best practice analysis pipeline for bacterial genomics, providing a comprehensive analysis of every single sample and the collection. The analysis results can be visualized via a web-based platform, allowing investigators to interactively examine various aspects of the collection. The complete software package can be easily installed on any common computer system. Both the underlying analysis pipeline and the web application are highly extendable; new analysis tools can be easily incorporated into the pipeline as they become established. We will also include additional visualization tools to support new analyses as required. We anticipate the software to be of general importance in the field of bacterial genomics. The method presented will be a utility for scientists to rapidly investigate bacterial antibiotic resistance.

Availability of source code and requirements

- Project name: AMRViz.
- Project home page: https://github.com/amromics/amrviz
- Operating system(s): Platform independent
- Programming language: Python and JavaScript
- License: MIT

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05792-9.

Additional file 1. Supplementary Figures S1–S7.

Author contributions

MDC, HAN and TN conceptualized the work. DQL and SHN developed the pipeline and the visualization framework. DQL, TTN and SHN ran the case study and analyzed the results. CHN, HTH and NSV contributed to the analysis of the results. DQL, TN, MDC and SHN wrote the main manuscript. All authors reviewed the manuscript.

Funding

This work has been supported by Vingroup Innovation Foundation (VINIF) in project code VINIF.2019.DA11.

Availability of data and materials

Data used in the manuscript are made available on the project home page https://github.com/amromics/amrviz/tree/main/examples. Scripts to reproduce the analysis are also provided on the GitHub repository.The data results from the case study are available on figshare at https://doi.org/10.6084/m9.figshare.25560645.v1.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interest

The author(s) declare that they have no conflict of interest.

Received: 26 September 2023 Accepted: 18 April 2024
Published online: 16 May 2024

References

1. Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, DeBoy RT, Davidsen TM, Mora M, Scarselli M, Margarit y Ros I, Peterson JD, Hauser CR, Sundaram JP, Nelson WC, Madupu R, Brinkac LM, Dodson RJ, Rosovitz MJ, Sullivan SA, Daugherty SC, Haft DH, Selengut J, Gwinn M.L, Zhou L, Zafar N, Khouri H, Radune D, Dimitrov G, Watkins K, O'Connor KJB, Smith S, Utterback TR, White O, Rubens CE, Grandi G, Madoff LC, Kasper DL, Telford JL, Wessels MR, Rappuoli R, Fraser CM. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". Proc Natl Acad Sci. 2005;102:(39)13950–13955. https://doi.org/10.1073/pnas.0506758102.
2. Cummins EA, Hall RJ, Connor C, McInerney JO, McNally A. Distinct evolutionary trajectories in the *Escherichia coli* pangenome occur within sequence types. Microb Genom. 2022;8(11):1–13. https://doi.org/10.1099/mgen.0.000903.
3. McInerney JO, McNally A, O'Connell MJ. Why prokaryotes have pangenomes. Nat Microbiol. 2017;2(March):1–5. https://doi.org/10.1038/nmicrobiol.2017.40.
4. Karlsen ST, Rau MH, Sánchez BJ, Jensen K, Zeidan AA. From genotype to phenotype: computational approaches for inferring microbial traits relevant to the food industry. FEMS Microbiol Rev. 2023;47:030.
5. Do VH, Nguyen SH, Le DQ, Nguyen TT, Nguyen CH, Ho TH, Vo NS, Nguyen T, Nguyen HA, Cao MD. Pasa: leveraging population pangenome graph to scaffold prokaryote genome assemblies. Nucleic Acids Res. 2024;52(3):15–15. https://doi.org/10.1093/nar/gkad1170.
6. De la Fuente J, Diez-Delgado I, Contreras M, Vicente J, Cabezas-Cruz A, Tobes R, Manrique M, Lopez V, Romero B, Bezos J, et al. Comparative genomics of field isolates of *Mycobacterium bovis* and *M. caprae* provides evidence for possible correlates with bacterial viability and virulence. PLoS Negl Trop Dis. 2015;9(11):0004232.
7. Hendriksen RS, Bortolaia V, Tate H, Tyson GH, Aarestrup FM, McDermott PF. Using genomics to track global antimicrobial resistance. Front Public Health. 2019;7:242.
8. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol. 2018;19(1):153. https://doi.org/10.1186/s13059-018-1540-z.
9. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012;19(5):455–77. https://doi.org/10.1089/cmb.2012.0021.
10. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30(14):2068–9. https://doi.org/10.1093/bioinformatics/btu153.
11. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush D, Keane JA, Parkhill J. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691–3. https://doi.org/10.1093/bioinformatics/btv421.
12. Jolley KA, Maiden MC. BIGSdb: scalable analysis of bacterial genome variation at the population level. BMC Bioinform. 2010;11:1–11.
13. Carattoli A, Zankari E, Garcia-Fernandez A, Voldby Larsen M, Lund O, Villa L, Moller Aarestrup F, Hasman H. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. Antimicrob Agents Chemother. 2014;58(7):3895–903. https://doi.org/10.1128/AAC.02412-14.
14. Feldgarden M, Brover V, Gonzalez-Escalona N, Frye JG, Haendiges J, Haft DH, Hoffmann M, Pettengill JB, Prasad AB, Tillman GE, et al. AMRFinderPlus and the reference gene catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Sci Rep. 2021;11(1):1–9.
15. Seemann T, G.d.S.A. Github https://github.com/tseemann/nullarbor (2018).
16. Petit RA III, Read TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. Msystems. 2020;5(4):10–1128.
17. Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, Goesmann A. ASA3P: an automatic and scalable pipeline for the assembly, annotation and higher-level analysis of closely related bacterial isolates. PLoS Comput Biol. 2020;16(3):1007134.
18. Le DQ, Nguyen TT, Nguyen CH, Ho TH, Vo NS, Nguyen T, Nguyen HA, Cao MD, Nguyen SH. AMRomics: a scalable workflow to analyze large microbial genome collection. bioRxiv. 2024. https://doi.org/10.1101/2024.04.02.587817.
19. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat graphs. Nat Biotechnol. 2019;37(5):540–6. https://doi.org/10.1038/s41587-019-0072-8.
20. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL. BLAST+: architecture and applications. BMC Bioinform. 2009;10(1):421. https://doi.org/10.1186/1471-2105-10-421.

21. Le DQ, Nguyen TA, Nguyen TT, Nguyen SH, Do VH, Nguyen CH, Phung HT, Ho TH, Nam VS, Nguyen T, Nguyen HA, Cao MD. Efficient inference of large pangenomes with PanTA. Bioarxiv. 2023. https://doi.org/10.1101/2023.07.03.547471.
22. Nakamura T, Yamada KD, Tomii K, Katoh K. Parallelization of MAFFT for large-scale multiple sequence alignments. Bioinformatics. 2018;34(14):2490–2. https://doi.org/10.1093/bioinformatics/bty121.
23. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5(3):9490. https://doi.org/10.1371/journal.pone.0009490.
24. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37(5):1530–4. https://doi.org/10.1093/molbev/msaa015.
25. ...The HC, Karkey A, Pham Thanh D, Boinett CJ, Cain AK, Ellington M, Baker KS, Dongol S, Thompson C, Harris SR, Jombart T, LeThiPhuong T, Tran Do Hoang N, Ha Thanh T, Shretha S, Joshi S, Basnyat B, Thwaites G, Thomson NR, Rabaa MA, Baker S. A high-resolution genomic analysis of multidrug-resistant hospital outbreaks of *Klebsiella pneumoniae*. EMBO Mol Med. 2015;7(3):227–39. https://doi.org/10.15252/emmm.201404767.
26. Larsen MV, Cosentino S, Rasmussen S, Friis C, Hasman H, Marvig RL, Jelsbak L, Sicheritz-Pontén T, Ussery DW, Aarestrup FM, Lund O. Multilocus sequence typing of total-genome-sequenced bacteria. J Clin Microbiol. 2012;50(4):1355–61. https://doi.org/10.1128/JCM.06094-11.
27. Davies MR, Keller N, Brouwer S, Jespersen MG, Cork AJ, Hayes AJ, Pitt ME, De Oliveira DM, Harbison-Price N, Bertolla OM, et al. Detection of *Streptococcus pyogenes* m1uk in Australia and characterization of the mutation driving enhanced expression of superantigen SpeA. Nat Commun. 2023;14(1):1051.
28. Butler TA, Story C, Green E, Williamson KM, Newton P, Jenkins F, Varadhan H, van Hal S. Insights gained from sequencing Australian non-invasive and invasive *Streptococcus pyogenes* isolates. Microb Genom. 2024;10(1):001152.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.