rMAP: Rapid Microbial Analysis Pipeline for ESKAPE bacterial group whole-genome sequence data

Ivan Sserwadda[1,2], Gerald Mboowa[1,3*]

Affiliations

[1]Department of Immunology and Molecular Biology, College of Health Sciences, School of Biomedical Sciences, Makerere University, P.O Box 7072, Kampala, Uganda

[2]Department of Biochemistry and Bioinformatics, School of Pure and Applied Sciences, Pwani University, P.O Box 195-80108, Kilifi, Kenya

[3]The African Center of Excellence in Bioinformatics and Data Intensive Sciences, the Infectious Diseases Institute, College of Health Sciences, Makerere University, P.O Box 22418, Kampala, Uganda

*To whom correspondence should be addressed. gerald.mboowa@chs.mak.ac.ug

Abstract

The recent re-emergence of multidrug resistant pathogens through persistent misuse antibiotics has exacerbated their threat to worldwide human public health and well-being. The evolution of the genomics era has led to generation of huge volumes of sequencing data at an unprecedented rate due to the ever-reducing costs of whole-genome sequencing (WGS). The considerable bioinformatics skills needed to analyze the large volume of genomic data from these platforms and subsequent downstream analysis offer constraints in the implementation of WGS as a routine laboratory technique. We have developed rMAP, a user-friendly pipeline capable of profiling the resistomes of ESKAPE pathogens (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa*, and *Enterobacter* species) from any source using WGS data generated from Illumina's sequencing platforms whilst hiding the technical impediments from inexperienced users. rMAP is designed for beginners and people with little bioinformatics expertise, and automates the steps required for WGS analysis directly from the raw genomic sequence data including: adapter and low quality sequence read trimming, *de-novo* genome assembly, genome annotation, SNP-variant calling, phylogenetic inference by maximum-likelihood, antimicrobial resistance profiling, plasmid profiling, virulence factor determination, multi-locus sequence typing (MLST), pangenome analysis, and insertion sequence characterization (IS). Once the analysis is finished, rMAP generates interactive web-like html report that can be visualized using any web browser, shared and reviewed in a simpler manner. rMAP installation is very simple, it can be run using very simple commands. It is a representation of a rapid and easy way to perform comprehensive bacterial WGS analysis using a personal laptop in low-income settings where High-performance computing infrastructure is limited. This pipeline can be implemented as a surveillance tool for tracking the trends of clinically significant pathogens that pose a threat to public health.

Availability and implementation: rMAP source codes and operation manual freely available at https://github.com/GunzIvan28/rMAP

Introduction

The recent re-emergence of multidrug resistant pathogens through persistent misuse antibiotics has exacerbated their threat to worldwide human public health and well-being. Such organisms consisting of: *Staphylococcus aureus, Pseudomonas aeruginosa,* and *Klebsiella species* belonging to the ESKAPE pathogen group flagged among the top most notorious bugs expressing tremendously high levels of antimicrobial resistance by the WHO have been reported by many studies to contribute to the high frequency of nosocomial infections which have led to high morbidity and mortality rates all over the world[1-3].

In the same spirit, rapid advances in diagnostic science and personalized medicine have ushered in the invention of high-throughput 'next-generation' sequencing technologies to replace the conventional microbiology laboratories which have greatly reduced diagnostic costs and result generation turn-around time for infectious pathogens as a way of keeping stride with their emerging multidrug resistant pathogens. Next-generation processes generally involve parallel sequencing, producing vast quantities of genomic data that require intensive modern computation infrastructure to make sense of the sequencing data in downstream analysis. Furthermore, another bottleneck in the deployment of high-throughput sequencing (HTS) technologies is the ability to analyze the increasing amount of data produced in a fit-for-purpose manner[4]. The field of microbial bioinformatics is thriving and quickly adapting to technological changes, which creates difficulties for clinical microbiologists with little or no bioinformatics background in following the complexity and increasingly obscure jargon of this field[4].
The routine application of WGS requires cheap, user-friendly techniques that can be used on-site by personnel not specialized in big data management[5, 6]. The bioinformaticists' ability to analyze, compare, interpret, and visualize the vast increase in bacterial genomes is valiantly trying to keep up with these developments[7]. Many biologists are drowning in too much data, and in desperate need for the ultimate tool capable of deciphering this complex information, and it is predicted that these trends will continue on in the foreseeable future as generation of genome data becomes cheaper and abundant[7].

Therefore, we introduce rMAP, the rapid microbial analysis pipeline; a one-stop shop toolbox that uses WGS illumina data to characterize the resistomes of bacteria of ESKAPE origin. This is an open-source, user-friendly, command-line, automated and scalable pipeline for conducting analysis of HTS data produced by Illumina platforms.

rMAP takes raw sequencing data as input and performs bacterial bioinformatic analysis steps including: adapter and low-quality sequence trimming, *De-novo* genome assembly, genome annotation, SNP-variant calling, phylogenetic inference by maximum-likelihood, antimicrobial resistance profiling, plasmid profiling, virulence factor determination, multi-locus sequence typing (MLST), pangenome analysis, and insertion sequence characterization (IS).

Pipeline architecture

rMAP is a tool implemented in four programming languages namely Shell script, Python, Perl and R. It was precompiled and supports Linux-64-bit architecture MacOS Catalina. It was originally built using WSL Ubuntu 20.04.1 LTS (Focal Fossa) and Ubuntu 18.04.4 LTS (Bionic Beaver) and the binaries are compatible with noarch-unix style operating systems.

The foundation of rMAP is built using a collection of published reputable tools like FASTQC[8], MultiQC[9], Trimmomatic[10], Shovill, Megahit[11], Prokka[12], Freebayes, SnpEff[13], IQtree[14], BWA[15], Samtools[16], Roary[17], and ISMapper[18] just to mention a few. All the tools and third-party dependencies required by rMAP are resolved and containerized within a conda environment as a single package so as not to interfere with already existing programs. The programs in the conda environment are built on top of Python version 3.7.8[19] and are compatible with R-statistical package version 4.0.2[20] A full list of the packages used by rMAP are shown in **Table 1.**

Table 1: Comprehensive list of third-party tools and algorithms at the base of rMAP

Software	Version	Sypnosis
Abricate	1.0.1	Detection of antimicrobial resistance genes, plasmids virulence factors
AMRfinder	3.8.4	Detection of antimicrobial resistance genes from assembled contigs
Any2fasta	0.4.2	Converts any genomic data format to fasta format
Assembly-stats	1.0.1	Summarizes quality assembly metrics from contigs
Biopython.convert	1.0.3	Conversion and manipulation of different genomic data formats
BMGE	1.12	Block Mapping and Gathering with Entropy for removal of ambiguously aligned reads from multiple sequence alignments
BWA	0.7.17	Burrow Wheeler algorithm for fast alignment of short sequence reads
Cairosvg	2.4.2	Converts SVG to PDF and PNG formats
Fastqc	0.11.9	Quality control and visualization of HTS data
Fasttree	2.1.10	Ultra-fast inference of phylogeny using maximum-likelihood method
Freebayes	1.3.2	Bayesian-based haplotype prediction of
ISMapper	2.0.1	Detection of insertion sequences within genomes
IQtree	2.0.3	Inference of phylogeny using maximum-likelihood method
Kleborate	1.0.0	Screening for AMR genes and MLSTs from genome assemblies
Lxml	4.5.2	Parsing of XML and HTML using python
Mafft	7.471	Algorithm for performing multiple sequence alignments
Multiqc	1.9	Aggregates numerous HTML quality reports into a single file
Megahit	1.2.9	Ultra-fast genome assembly algorithm
Mlst	2.19.0	Characterization and detection of clones within a population of pathogenic isolates
Nextflow	20.07.1	Portable next-generation workflow language that enables reproducibility and development of pipelines
Parallel	20200722	Executes jobs in parallel
Prinseq	0.20.4	Trims, filters and reformats genomic sequence data
Prodigal	2.6.3	Prediction of protein-coding genes in prokaryotic genomes

Prokka	1.14.6	Fast and efficient annotation of prokaryotic assembled genomes
Quast	5.0.2	Quality assembly assessment tool
Roary	3.13.0	Large-scale pangenome analysis
R-base	4.0.2	Statistical data computing and graphical software
Samclip	0.4.0	Filters SAM file for soft and hard clipped alignments
Samtools	1.9	Tools for manipulation of Next-generation sequence data
Shovill	1.0.9	Illumina short read assembler for bacterial genomes
Snippy	4.3.6	Rapid haploid bacterial variant caller
Snpeff	4.5Covid19	Functional effect and variant predictor suite
SRA-tools	2.10.8	Toolbox for acquisition and manipulation of sequences from NCBI
Trimmomatic	0.39	Illumina short-read adapter trimming algorithm
Unicycler	0.4.8	A hybrid assembly pipeline for illumina and long read sequence data
Vt	2015.11.10	A tool for normalizing variants in genomic sequence data

Overview of rMAP workflow

rMAP can be used with an unlimited number of samples of different species and origins. However, it was built to target pathogens of public health exhibiting unrelentless levels of AMR and nosocomial infections. It can be applied to isolates of human and animal origin to give insights of the transmission dynamics of AMR genes at the human-animal interface.

Datasets

The pipeline was tested on numerous bacterial pathogens from ESKAPE group isolated from different origins (clinical, fecal, animal and sewage) sequenced on Illumina platforms obtained from the publicly available repositories of Sequence Read Archive (SRA) and the European Nucleotide Archive (ENA) under the following accessions: *Enterococcus* species(SRR8948878, SRR8948879, SRR8948880, SRR8948881, SRR8948882, SRR8948883, SRR8948884, SRR8948885, SRR8948886, SRR8948887, SRR8948888, SRR8948889, SRR8948890, SRR8948891), *Acinetobacter baumannii(*ERR1989084, ERR1989100, ERR1989115, ERR3197698, SRR3666962, SRR5739056, SRR6037664, SRR8289559, SRR8291681), *Klebsiella species*(SRR8753739, SRR8753737, SRR8291573, SRR11816972, SRR9703249, SRR9029107, SRR9029108, SRR8610335, SRR8610353, SRR8610357, SRR8610351, SRR8610354, SRR9964283, SRR9044171, SRR5687278, SRR5514226, SRR5514224, SRR5514223) and *Staphylococcus aureus*(ERR1794900, ERR1794901, ERR1794902, ERR1794903, ERR1794904, ERR1794905, ERR1794906, ERR1794907, ERR1794908, ERR1794909, ERR1794910, ERR1794911, ERR1794912, ERR1794913, ERR1794914). The genbank references used include; *Acinetobacter baumannii* strain 36-1512 Accession: CP059386.1 GI: 1880620189, *Enterococcus faecalis strain KB1* Accession: CP015410.2 GI: 1173533644, and *Staphylococcus aureus* subsp. aureus strain MRSA252 Accession: BX571856.1 GI: 49240382.

Core pipeline features

rMAP requires three mandatory parameters; the input directory that contains sequence reads in either fastq or fastq.gz formats, an output user-defined directory and a reference genome in either genbank or fasta format. A full genbank reference genome file is recommended to the –reference option to obtain

an annotated VCF files. The raw fastq files are directly submitted to rMAP, with no prior bioinformatics treatment, as follows:

rMAP -t 8 --reference --input dir_name--output dir_name --quality --assembly megahit --amr --varcall --phylogeny --pangenome --gen-ele

The pipeline's features can be summarized in the order of: SRA sequence download, quality control, adapter trimming, *De novo* assembly, resistomes profiling, variant calling, phylogenetic inference, pangenome analysis, insertion sequence mapping and report generation as shown in figure1.

Sequence Read archive download
rMAP is able to retrieve sequences from NCBI-SRA using fastq-dump[21]. A user simply creates a list containing the sample accession numbers to be downloaded saved at the home directory. The downloaded sequences are saved in a default directory called SRA-READS created by rMAP.

Quality assessment and filtering
The pipeline autodetects any non-zipped fastq reads and parses them to fastq.gz format for optimization purposes during downstream analysis. Fastqc[8] generates sequence quality reports and statistics from each individual sample which are then aggregated into a single graphically interactive html report using MultiQC[9].

Adapter and low sequence read trimming
Trimmomatic[10] is used to trim off adapters using a set of pre-defined Illumina library preparation adapters saved in fasta format and low sequence regions from the raw input sequence reads. The pipeline's default parameters for quality and minimum sequence length are set at a phred quality score of 27 and 80 base pairs respectively to accommodate sequencing data that may not be of very high required recommended quality of 33.

***De novo* assembly and annotation**
Two assemblers are selected for this purpose for a user to choose; Shovill[22] and Megahit[11] each demonstrating an advantage over the other. Both algorithms take the trimmed reads as their input and perform k-mer based assembly to produce contigs. Megahit exhibited very fast computational speeds almost half its counterpart but with slightly lower quality assembly metrics. Assembly with Shovill involves guided mapping of the contigs to a reference and numerous rounds of genome polishing using pilon to remove gaps which lasts a longer time but produces good quality assembly metrics (N50, L50, Genome Length). Prodigal[23] is used to predict open reading frames from the assembled contigs which are then functionally annotated using Prokka[12].

Variant calling
The trimmed reads are aligned against a an indexed reference in fasta format using Burrows-Wheeler aligner[15] to produce SAM files. Soft and hard clipped alignments are removed from the Sequence Alignment Map (SAM) files using Samclip(https://github.com/tseemann/samclip). Samtools[16] then sorts, marks duplicates and indexes the resultant Binary Alignment Map (BAM) files. Freebayes[24] calls variants using Bayesian models to produce variant call format (VCF) files containing SNP information which is filtered using bcftools(https://github.com/samtools/bcftools) and normalized of biallelic regions using Vt[25]. The filtered VCF files are annotated using snpEff[13]. Raw, tab-separated, annotated and filtered VCF files are available for the users to manipulate.

Resistome profiling
The conceptualization of rMAP was aimed at exhaustively exploit the resistome of pathogenic bacteria. AMRfinder plus[26] predicts resistance genes using its database. Mass screening for antimicrobial resistance genes against CARD[27], ARG-ANNOT[28], NCBI, ResFinder and MEGARES[29] databases; Plasmids, and virulence factors are type against the PlasmidFinder[30] and Virulence Factor Database(VFDB)[31] respectively using Abricate(https://github.com/tseemann/abricate) from the assembled genomes. Multi-locus sequence typing is performed using Mlst(https://github.com/tseemann/mlst).

Phylogenetic inference
Because of the computationally demanding requirements of algorithms in terms of RAM and core threads to phylogenetic analysis, rMAP incorporates the use of SNP-based analysis which has

successfully proven to be faster than using sequencing data to infer phylogeny. A single VCF file containing all the samples and their SNPs is generated towards the end the variant calling stage which is transposed by vcf2phylip[32] into a multi-alignment fasta file. Multi-sequence alignment is performed using Mafft[33] with the removal of ambiguously aligned reads and selection of informative regions to infer phylogeny using BMGE[34]. IQtree[14] tests various substitution models and constructs trees from the alignments using maximum-likelihood method in 1000 bootstraps. The resulting trees are visualized in form of rectangular (phylogram), circular (phylogram), and circular (cladogram).

Pangenome analysis

Roary[17] is employed by rMAP pipeline to perform core and accessory pangenome analysis across the input samples using general feature format (.gff) files generated from the annotation step. Fasttree is used to convert the core genome alignment to newick format. The scalable vector graphic (SVG) file obtained from the pangenome analysis is converted to a portable network graphic (PNG) file format by cairosvg (https://cairosvg.org/). The resulting trees are visualized in form of rectangular (phylogram), circular (phylogram), and circular (cladogram).

Insertion sequence analysis

rMAP interrogates for the presence of mobile genetic elements in particular insertion sequences using ISMapper[18] which basically spans the lengths of the entire genome of a sequence searching for homology against a set of well-known insertion sequence families commonly found in ESKAPE isolates reported by a study comprehensive study of mobile genetic elements[35] and ISfinder database (https://www-is.biotoul.fr/index.php) shown in Table 2.

Reporting and visualization of the reports

rMAP stores and formats reports from each stage of the pipeline under one directory called 'reports' and uses R-base[20] with a set of R-packages including: ggtree[36], RcolorBrewer, ggplot2[37], knitr[38], rmarkdown[39], plotly[40], reshape2[41] and treeio[42] to generate a web-like html interactive report with expounded explanations at every stage of analysis that can easily be shared and interpreted by inexperienced Bioinformatics individuals. An example of such a report can be accessed via https://gunzivan28.github.io/rMAP/. The reporting format for rMAP was mainly adapted from Tormes[6] pipeline. The results from a successful run can be found under the user-defined output directory and consist of files from: assembly, annotation, insertion sequences, mlsts, pangenomes, phylogeny, plasmids, quality reports, quast assembly stats, reports, resistance genes, trimmed reads, variant calling and virulence factors for further analysis. rMAP retains all the intermediate files generated after a successful run to be interrogated further by experienced Bioinformatics users. The contents extracted from the intermediate files and summarized in the html report with a short description are briefly described in table 3. Examples of visuals generated by the pipeline are illustrated in figure 2.

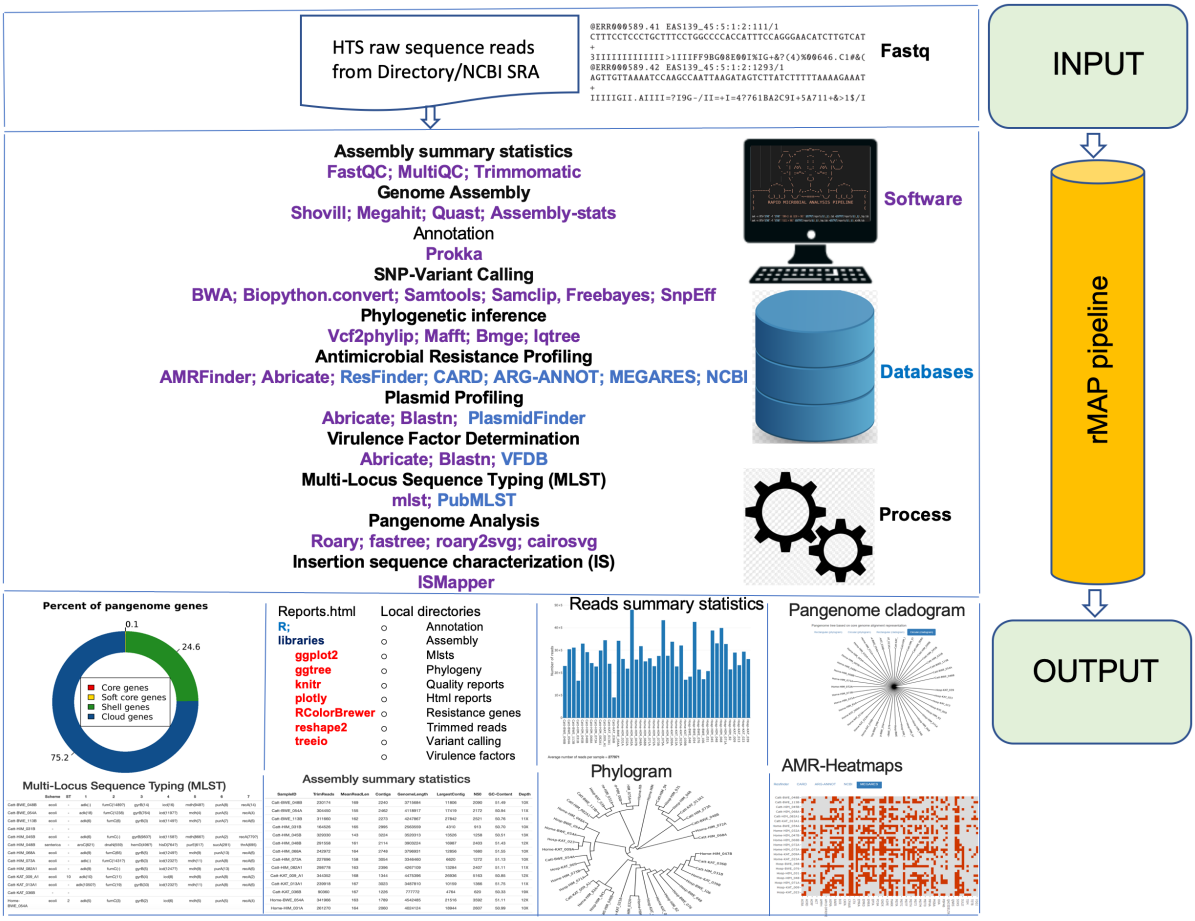


Figure. 1. Schematic graphical representation of rMAP pipeline workflow and associated tools

Table 2: ESKAPE group insertion sequence families (both Gram positive and Gram negative) used by rMAP

Sequence name	Determinant genes	Conferred resistance
IS903	aphA1	Kanamycin
ISApl1	mcr-1	Colistin
ISEc69	mcr-2	Colistin
ISAba14	aphA6	Kanamycin
ISAba1	blaOXA-23	Carbapenems, Beta-lactams
IS16	VanB1	Vancomycin
IS256	cfr	Phenicols, lincosamides, oxazolidinones, pleuromutilins and streptogramin A
IS257-2	aadD, ble, fosB5, fusB, tetL, tetK, aacA-aphD, vatA, dfrK	Kanamycin, bleomycin, fosfomycin, fusidic acid, tetracycline, gentamicin, streptogramin A, trimethoprim
IS1182	aadE, aphaA-3, sat4	Streptomycin, kanamycin, neomycin, streptothricin
IS1216	cfr, str	Phenicols, lincosamides, oxazolidinones, pleuromutilins, streptomycin, streptogramin A
IS1272	*mecR*	Methicillin
IS1182	aphA-3andaadEgenes	Aminoglycoside
ISEnfa4	cfr	Phenicols, lincosamides, oxazolidinones, pleuromutilins and streptogramin A
ISEcp1	CTX-M	Cefotaxime, ceftriaxone, and aztreonam
ISSau1	SCCmec	Methicillin
ISKpn23	blaBKC-1	Carbapenems, cephalosporins, monobactams

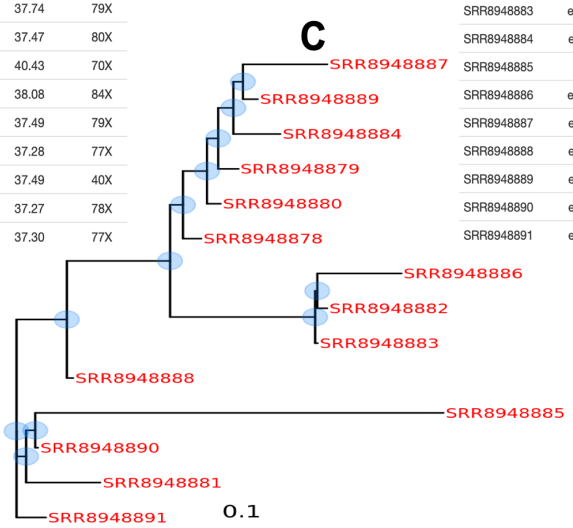
A

SampleID	TrimReads	MeanReadLen	Contigs	GenomeLength	LargestContig	N50	GC-Content	Depth
SRR8948878	1595020	149	68	3101402	545479	273671	37.21	76X
SRR8948879	1588168	149	79	3010836	603711	179463	37.44	78X
SRR8948880	1594518	149	102	3166205	417860	239597	37.18	75X
SRR8948881	1591214	149	44	2924312	516731	202453	37.46	81X
SRR8948882	1586702	149	170	2925331	107207	46219	37.69	80X
SRR8948883	1579276	149	180	2964061	113751	44888	37.74	79X
SRR8948884	1591394	149	57	2934130	644112	179342	37.47	80X
SRR8948885	1584504	149	59	3363692	424585	243114	40.43	70X
SRR8948886	1584116	149	56	2786637	344853	172313	38.08	84X
SRR8948887	1585296	149	89	2984341	603711	185199	37.49	79X
SRR8948888	1576542	149	57	3024061	544043	207386	37.28	77X
SRR8948889	1590020	149	290	5897762	544519	66637	37.49	40X
SRR8948890	1588164	149	52	3016695	544186	207384	37.27	78X
SRR8948891	1589798	149	69	3037920	544195	219893	37.30	77X

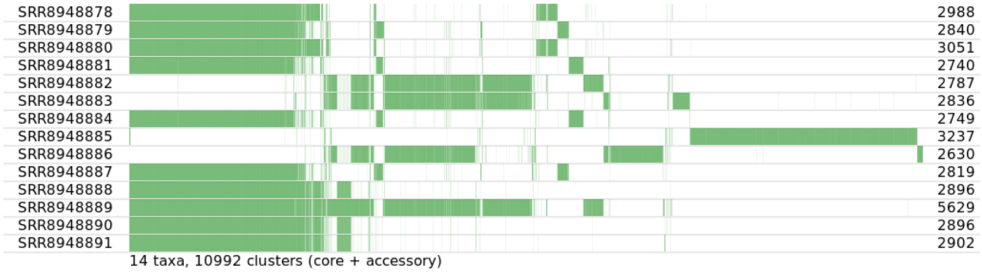
B

	Scheme	ST	1	2	3	4	5	6	7
SRR8948878	efaecalis	16	gdh(5)	gyd(1)	pstS(1)	gki(3)	aroE(7)	xpt(7)	yqiL(6)
SRR8948879	efaecalis	40	gdh(3)	gyd(6)	pstS(23)	gki(12)	aroE(9)	xpt(10)	yqiL(7)
SRR8948880	efaecalis	16	gdh(5)	gyd(1)	pstS(1)	gki(3)	aroE(7)	xpt(7)	yqiL(6)
SRR8948881	efaecalis	40	gdh(3)	gyd(6)	pstS(23)	gki(12)	aroE(9)	xpt(10)	yqiL(7)
SRR8948882	efaecium	80	atpA(9)	ddl(1)	gdh(1)	purK(1)	gyd(12)	pstS(1)	adk(1)
SRR8948883	efaecium	203	atpA(15)	ddl(1)	gdh(1)	purK(1)	gyd(1)	pstS(20)	adk(1)
SRR8948884	efaecalis	40	gdh(3)	gyd(6)	pstS(23)	gki(12)	aroE(9)	xpt(10)	yqiL(7)
SRR8948885	-	-							
SRR8948886	efaecium	1446	atpA(11)	ddl(13)	gdh(18)	purK(15)	gyd(10)	pstS(19)	adk(6)
SRR8948887	efaecalis	40	gdh(3)	gyd(6)	pstS(23)	gki(12)	aroE(9)	xpt(10)	yqiL(7)
SRR8948888	efaecalis	179	gdh(5)	gyd(1)	pstS(1)	gki(3)	aroE(7)	xpt(1)	yqiL(6)
SRR8948889	efaecalis	-	gdh(5,92)	gyd(1)	pstS(1)	gki(3)	aroE(7)	xpt(1)	yqiL(6)
SRR8948890	efaecalis	179	gdh(5)	gyd(1)	pstS(1)	gki(3)	aroE(7)	xpt(1)	yqiL(6)
SRR8948891	efaecalis	179	gdh(5)	gyd(1)	pstS(1)	gki(3)	aroE(7)	xpt(1)	yqiL(6)

C



D



E

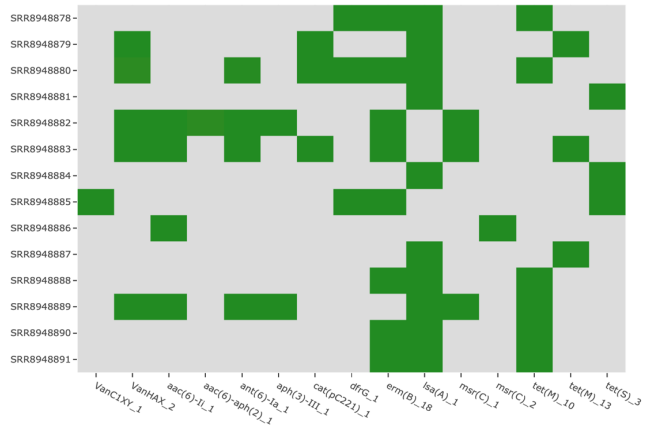


Fig 2. Selected interactive *Enterococcus species* HTML reports. (A) Genome assembly summary statistics for the different *Enterococcus species* isolates. These include common genome analysis key metrics for checking assembly quality. (B) Table of Multi-Locus Sequence Typing (MLST) distribution. (C) SNP-based approximately-maximum-likelihood phylogenetic tree. Three different formats are available i.e Circular (phylogram), Circular (cladogram) and Rectangular (phylogram). An approximately-maximum-likelihood phylogenetic tree is computed based on SNPs detected via read-mapping against a reference genome and stored in standard Newick file format. (D) Pangenome analysis including a schematic representation of genes presence (color) or absence (blank) between samples. (E) Antibiotic resistance profile. Presence/Absence of antibiotic resistance genes (coverage and identity > 90%) on each sample. An antibiotic resistance profile is computed based on Resfinder, CARD, ARG-ANNOT, NCBI and MEGARES annotations for each isolate and transformed into an overview that allows a rapid resistome comparison of all analyzed isolates.

Table 3: Summary description for some stages of intermediate files generated from rMAP

Analysis	Metrics	Description
Assembly	Genome length, Average genome length, N50, GC content and sequencing Depth	• Genome length – An estimate of the draft genome assembly length • Average genome length – Average read length of genomes • N50 – length of smallest contig covering 50% of genome • GC content – Guanine-Cytosine content of draft genome • Depth - Number of times each nucleotide position in the draft genome has a read that aligns to that position
Phylogeny		• SNPs are used to infer phylogenetic relationships between samples
Variant calling	SNPs	• SNP – a single nucleotide base change different from the reference genome that occurs anywhere within the genome
Antimicrobial resistance profiling	Contig, gene, identity, product	• Contig – continuous consensus nucleotide sequences without gaps • Gene – Antibiotic resistance gene identified within the assembly • Identity – Percentage proportion representing exact nucleotide matches • Product – Artefact produced from antibiotic resistance gene
Pangenome analysis	Core genes, soft core genes, shell genes, cloud genes	• The genes are compared against each other across samples to predict genome plasticity and to detect how much of the accessory genome has been taken up by organisms over the course of time

Computational infrastructure and benchmarking

The original philosophy of creating rMAP was to create a tool that can be easily installed and run on a personal computer having descent standard bioinformatics hardware and computational infrastructure. The pipeline was successfully compiled on two personal computers with the following specifications: Dell Inspiron 5570 8th Gen Intel(R) Core(TM) i7-8550U CPU @ 1.80GHz (8 CPUs), ~2.0GHz with 12GB of RAM and 1 Terabyte of hard disk space running Windows subsystem Linux (WSL) Ubuntu 20.04.1

LTS (Focal Fossa) and Ubuntu 18.04.4 LTS (Bionic Beaver) and a MacBook pro Intel(R) Core i7 CPU @ 3.0GHz, 16GB of RAM and 2TB of SSD space running on MacOS Mojave. When provided with the same samples, the MacBook performed relatively better because of its hardware compared to Ubuntu. Depending on the number of samples provided in the input, rMAP generates intermediate files ranging between 10 and 30GB. The wall clock runtimes and benchmarking statistics for each bacterial species on different platforms is summarized in Table 3

Table 3. rMAP's Wall clock runtimes for different bacterial species across different operating system platforms.

Genomes	Genome size	Ubuntu	macOS X
15 Staphylococcus aureus	~ 2.9 Mbp	22 hrs	18 hrs
9 *Acinetobacter baumannii*	~ 3.9 Mbp	22 hrs	19 hrs
14 *Enterococcus spp*	~ 2.9 Mbp	21 hrs	17 hrs

Discussion

Although other pipelines developed under the same philosophy and functionality as rMAP like Tormes[6], ASA3P[43] and the recently published Bactopia[44] exist, we noticed that each of these software had a short-coming that we aimed at addressing. In terms of useability, Tormes[6] was the most friendly pipeline with one major predicament where it could never be launched without a tab-separated metadata file complying with a set criteria. It was also more oriented to bacterial species-specific analyses namely *Escherichia coli* and *Salmonella species.* ASA3P[43] and Bactopia[44] required a Bioinformatics competent user to operate since they are written in complex languages that is Groovy and Nextflow respectively. Other similar pipelines like Nullarbor(https://github.com/tseemann/nullarbor) was extremely difficult to compile and use compared to the to its counterparts, requiring a metadata file conforming with set criterion. In cases where metadata files are required, the different software flagged errors or halted task executions as the correct conforming metadata files were essential for the downstream analyses.

rMAP on the other hand comes with features aimed overcoming the limitations of its counterparts. It requires no prior pre-processing of the sequences or metadata files. The user only provides three essential requirements namely: an input directory, an output directory and a reference genome to run the pipeline. The pipeline is written in basic programming languages that don't require advanced expertise or troubleshooting to be launched. rMAP is highly portable and capable of running on descent personal computers running on either Ubuntu or MacOS. Installation is pretty easy and straight forward from the GitHub repository (https://github.com/GunzIvan28/rMAP) with the binaries and dependencies built within conda environment packages. Most of all, rMAP shows a high sensitivity towards analysis and not limited to a wide range of public health clinically significant ESKAPE group pathogens, but also to other Enterobacteria like *Escherichia coli* and *Salmonella species*.

As a significant limitation, rMAP is coded exclusively in Bash and not implemented within a modern workflow language manager like Snakemake or Nextflow. The ultimate consequence of this is that a user will have to either restart the whole run or manually check which steps completed successfully and resume the run by selecting only options which were not performed while excluding the computed steps from the main command script. This ushers in implementation of the pipeline within a modern workflow language which will be available in the next release of the software.

Conclusion

rMAP is a robust, scalable, user user-friendly, automated Bioinformatics analysis workflow for Illumina WGS reads that has demonstrated efficiency in the analysis of public health significant pathogens. Therefore, we recommend it as a tool for continuous monitoring and surveillance suitable for antimicrobial resistance gene trends especially in low-income countries that are limited by computational Bioinformatics infrastructure.

Availability and future directions

The source code is available on GitHub under GPL3 license at https://github.com/GunzIvan28/rMAP. Questions and issues can be sent to “ivangunz23@gmail.com”, bug reports can be filed as GitHub issues. Although rMAP itself is published and distributed under a GPL3 license, some of its dependencies bundled within the rMAP volume are published under different license models.

Future directions of rMAP comprise of implementation within Singularity, Docker and Nextflow platform containers as well as the integration of further enhancements in terms of scalability and usability.

Acknowledgments

Ivan Sserwadda is a current trainee of the Makerere University/Uganda Virus Research Institute (UVRI) Centre of Excellence for Infection & Immunity Research and Training (MUII) program. He is also supported by MUII. MUII is supported through the DELTAS Africa Initiative (Grant no. 107743). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS), Alliance for Accelerating Excellence in Science in Africa (AESA) and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (Grant no. 107743) and the UK Government.

We are equally grateful for the support of NIH Common Fund, through the OD/Office of Strategic Coordination (OSC) and the Fogarty International Center (FIC), NIH award number U2R TW 010672. Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the supporting offices”.

We also acknowledge all the authors for the various open source code for inspiring the conception and design of this pipeline. Special thanks go to Narciso Quijada, Torsten Seeman, Oliver Schwengers and Robert Pettit. Gratitude to NCBI and ENA for providing the dataset resources used in development of this tool

Author Contributions

Both authors contributed equally in Conceptualization; Formal analysis; Funding acquisition; Methodology; Resources; Software; Validation; Visualization; Writing – original draft; Writing – review & editing

Data Availability Statement

All source code is accessible at GitHub (https://github.com/GunzIvan28/rMAP). Genomes from the exemplary data sets are publicly stored in the SRA database; accession IDs are provided.

Funding

Grant information: This work was supported through the Grand Challenges Africa program (GCA/AMR/rnd2/058). Grand Challenges Africa is a program of the African Academy of Sciences (AAS) implemented through the Alliance for Accelerating Excellence in Science in Africa (AESA) platform, an initiative of the AAS and the African Union Development Agency (AUDA-NEPAD). GC Africa is supported by the Bill & Melinda Gates Foundation (BMGF) and The African Academy of Sciences and partners. The views expressed herein are those of the author(s) and not necessarily those of the AAS and its partners. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests

The authors have declared that no competing interests exist.

References

1. Sserwadda I, Lukenge M, Mwambi B, Mboowa G, Walusimbi A, Segujja F. Microbial contaminants isolated from items and work surfaces in the post- operative ward at Kawolo general hospital, Uganda. BMC Infect Dis. 2018;18(1):68-. doi: 10.1186/s12879-018-2980-5. PubMed PMID: 29409447.
2. Mulani MS, Kamble EE, Kumkar SN, Tawre MS, Pardesi KR. Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. 2019;10(539). doi: 10.3389/fmicb.2019.00539.
3. Ma Y-X, Wang C-Y, Li Y-Y, Li J, Wan Q-Q, Chen J-H, et al. Considerations and Caveats in Combating ESKAPE Pathogens against Nosocomial Infections. Adv Sci (Weinh). 2019;7(1):1901872-. doi: 10.1002/advs.201901872. PubMed PMID: 31921562.
4. Carriço JA, Rossi M, Moran-Gilad J, Van Domselaar G, Ramirez M. A primer on microbial bioinformatics for nonbioinformaticians. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases. 2018;24(4):342-9. Epub 2018/01/09. doi: 10.1016/j.cmi.2017.12.015. PubMed PMID: 29309933.
5. Hyeon J-Y, Li S, Mann DA, Zhang S, Li Z, Chen Y, et al. Quasimetagenomics-Based and Real-Time-Sequencing-Aided Detection and Subtyping of Salmonella enterica from Food Samples. Applied and Environmental Microbiology. 2018;84(4):e02340-17. doi: 10.1128/AEM.02340-17.
6. Quijada NM, Rodríguez-Lázaro D, Eiros JM, Hernández M. TORMES: an automated pipeline for whole bacterial genome analysis. Bioinformatics (Oxford, England). 2019;35(21):4207-12. Epub 2019/04/09. doi: 10.1093/bioinformatics/btz220. PubMed PMID: 30957837.
7. Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. Functional & integrative genomics. 2015;15(2):141-61. Epub 2015/02/28. doi: 10.1007/s10142-015-0433-4. PubMed PMID: 25722247; PubMed Central PMCID: PMCPMC4361730.
8. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
9. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple tools and samples in a single report. Bioinformatics (Oxford, England). 2016;32(19):3047-8. Epub 2016/06/18. doi: 10.1093/bioinformatics/btw354. PubMed PMID: 27312411; PubMed Central PMCID: PMCPMC5039924.
10. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England). 2014;30(15):2114-20. Epub 2014/04/04. doi: 10.1093/bioinformatics/btu170. PubMed PMID: 24695404; PubMed Central PMCID: PMCPMC4103590.
11. Li D, Liu CM, Luo R, Sadakane K, Lam TW. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics (Oxford, England). 2015;31(10):1674-6. Epub 2015/01/23. doi: 10.1093/bioinformatics/btv033. PubMed PMID: 25609793.
12. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England). 2014;30(14):2068-9. Epub 2014/03/20. doi: 10.1093/bioinformatics/btu153. PubMed PMID: 24642063.
13. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. Fly. 2012;6(2):80-92. doi: 10.4161/fly.19695.
14. Nguyen LT, Schmidt HA, von Haeseler A, Minh BQ. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Molecular biology and evolution. 2015;32(1):268-74. Epub 2014/11/06. doi: 10.1093/molbev/msu300. PubMed PMID: 25371430; PubMed Central PMCID: PMCPMC4271533.
15. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2009;25(14):1754-60. doi: 10.1093/bioinformatics/btp324.
16. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics (Oxford, England). 2009;25(16):2078-9. doi: 10.1093/bioinformatics/btp352 %J Bioinformatics.
17. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics (Oxford, England). 2015;31(22):3691-3. Epub 2015/07/23. doi: 10.1093/bioinformatics/btv421. PubMed PMID: 26198102; PubMed Central PMCID: PMCPMC4817141.

18. Hawkey J, Hamidian M, Wick RR, Edwards DJ, Billman-Jacobe H, Hall RM, et al. ISMapper: identifying transposase insertion sites in bacterial genomes from short read sequence data. BMC genomics. 2015;16(1):667. Epub 2015/09/04. doi: 10.1186/s12864-015-1860-2. PubMed PMID: 26336060; PubMed Central PMCID: PMCPMC4558774.
19. Rossum G. Python reference manual. 1995.
20. Ihaka R, Gentleman RJJoc, statistics g. R: a language for data analysis and graphics. 1996;5(3):299-314.
21. ncbi/sra-tools. NCBI - National Center for Biotechnology Information/NLM/NIH; 2020.
22. Seemann T. tseemann/shovill. 2020.
23. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJJBb. Prodigal: prokaryotic gene recognition and translation initiation site identification. 2010;11(1):119.
24. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv:12073907 [q-bio]. 2012.
25. Tan A, Abecasis GR, Kang HMJB. Unified representation of genetic variants. 2015;31(13):2202-4.
26. Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. 2019;63(11):e00483-19.
27. McArthur AG, Waglechner N, Nizam F, Yan A, Azad MA, Baylay AJ, et al. The comprehensive antibiotic resistance database. 2013;57(7):3348-57.
28. Gupta SK, Padmanabhan BR, Diene SM, Lopez-Rojas R, Kempf M, Landraud L, et al. ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. 2014;58(1):212-20.
29. Doster E, Lakin SM, Dean CJ, Wolfe C, Young JG, Boucher C, et al. MEGARes 2.0: a database for classification of antimicrobial drug, biocide and metal resistance determinants in metagenomic sequence data. Nucleic Acids Res. 2019;48(D1):D561-D9. doi: 10.1093/nar/gkz1010 %J Nucleic Acids Research.
30. Carattoli A, Zankari E, García-Fernández A, Larsen MV, Lund O, Villa L, et al. In silico detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. 2014;58(7):3895-903.
31. Liu B, Zheng D, Jin Q, Chen L, Yang JJNar. VFDB 2019: a comparative pathogenomic platform with an interactive web interface. 2019;47(D1):D687-D92.
32. Ortiz E. vcf2phylip v1. 5: convert a VCF matrix into several matrix formats for phylogenetic analysis; 2018.
33. Katoh K, Standley DMJMb, evolution. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. 2013;30(4):772-80.
34. Criscuolo A, Gribaldo S. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence alignments. BMC Evolutionary Biology. 2010;10(1):210. doi: 10.1186/1471-2148-10-210.
35. Partridge SR, Kwong SM, Firth N, Jensen SOJCmr. Mobile genetic elements associated with antimicrobial resistance. 2018;31(4).
36. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods in Ecology and Evolution. 2017;8(1):28-36. doi: 10.1111/2041-210X.12628.
37. Wickham H. ggplot2: Elegant Graphics for Data Analysis: Springer; 2016 2016/06/08/. 266 p.
38. Xie Y. Dynamic Documents with R and knitr.294.
39. Grolemund YXJJAG. R Markdown: The Definitive Guide.
40. Plotly R Graphing Library.
41. Wickham H. Reshaping Data with the reshape Package. Journal of Statistical Software. 2007;21(1):1-20. doi: 10.18637/jss.v021.i12.
42. Yu G. treeio.
43. Schwengers O, Hoek A, Fritzenwanker M, Falgenhauer L, Hain T, Chakraborty T, et al. ASA3P: An automatic and scalable pipeline for the assembly, annotation and higher level analysis of closely related bacterial isolates. 2020.
44. Petit RA, Read TD. Bactopia: a flexible pipeline for complete analysis of bacterial genomes. bioRxiv. 2020:2020.02.28.969394. doi: 10.1101/2020.02.28.969394.