Check for updates

SOFTWARE TOOL ARTICLE

Lessons learned in virulence factor identification and data management from a hackathon on microbial virulence

[version 1; peer review: 1 approved with reservations]

Brett E. Pickett^{1,2}, Ryan Connor³, Tamiru Berhanu-Denka⁴, Sherry Bhalla⁵, Vyacheslav Brover³, Michael James Chambers⁶, Kumardeep Chaudhary⁵, Ousmane H. Cissé⁷, Allissa Dillman³, Moamen M. Elmassry⁸, Michael Feldgarden³, Eric Holloway³, Xin Huang⁹, William Klimke³, Catarina Inês Mendes^{10,11}, S. Elizabeth Norred¹², Jonathan Parkinson¹³, Samantha Sevilla^{14,15}, Monica Garcia Solache¹⁶, Defne Surujon¹⁷, Udana Torian¹⁸, Vadim Zalunin³, Ben Busby^{3,19}

¹Booz Allen Hamilton, Rockville, MD, 20852, USA

²Brigham Young University, Provo, UT, 84602, USA

³National Center for Biotechnology Information, National Library of Medicine, National Institutes of Health, Bethesda, MD, 20894, USA

⁴City Paws Veterinary Hospital, Washington, DC, 20009, USA

⁵Icahn School of Medicine at Mount Sinai, New York, NY, 20019, USA

⁶Systems Biology and Genome Engineering Section, NIH - NHGRI, Bethesda, MD, 20894, USA

⁷Critical Care Medicine Department, NIH Clinical Center, National Institutes of Health, Bethesda, MD, 20814, USA

⁸Department of Biological Sciences, Texas Tech University, Lubbock, TX, USA

⁹National Human Genome Research Institute, National Institutes of Health, Bethesda, MD, 20851, USA

¹⁰Universidade de Lisboa, Lisbon, Portugal

¹¹Department of Medical Microbiology and Infection Prevention, University of Groningen, Groningen, The Netherlands

¹²Bredesen Center for Interdisciplinary Research and Education, University of Tennessee, Knoxville, TN, 37922, USA

¹³Qpex Biopharma, Inc., San Diego, CA, 92121, USA

¹⁴Cancer Genomics Research Laboratory, Leidos Biomedical Research, Inc., Frederick National Laboratory for Cancer Research, Frederick, MD, 21702, USA

¹⁵Division of Cancer Epidemiology and Genetics, National Cancer Institute (NCI), National Institutes of Health (NIH), Bethesda, MD, 20814, USA

¹⁶Department of Medicine, Brown University, Providence, RI, 02903, USA

¹⁷Boston College, Chestnut Hill, MA, 02467, USA

¹⁸Laboratories of Infectious Disease, National Institutes of Allergy and Infectious Disease, National Institutes of Health, Bethesda, MD, 20892, USA

¹⁹DNAexus, Mountain View, CA, 94040, USA

V1 First published: 09 Feb 2022, 11:160
<https://doi.org/10.12688/f1000research.26452.1>

Latest published: 09 Feb 2022, 11:160
<https://doi.org/10.12688/f1000research.26452.1>

Open Peer Review**Approval Status** ?**Abstract**

Virulence is a complex mix of microbial traits and host susceptibility that could ultimately lead to disease. The increased prevalence of

multidrug resistant infections complicates treatment options, augmenting the need for developing robust computational methods and pipelines that enable researchers and clinicians to rapidly identify the underlying mechanism(s) of virulence in any given sample/isolate. Consequently, the National Center for Biotechnology and Information at the National Institutes of Health hosted an in-person hackathon in Bethesda, Maryland during July 2019 to assist with developing cloud-based methods to reduce reliance on local computational infrastructure. Groups of attendees were assigned tasks that are relevant to identifying relevant tools, constructing pipelines capable of identifying microbial virulence factors, and managing the associated data and metadata. Specifically, the assigned tasks consisted of the following: data indexing, metabolic functions, virulence factors, antimicrobial resistance, mobile elements in enterococci, and metatranscriptomics. The cloud-based framework established by this hackathon can be augmented and built upon by the research community to aid in the rapid identification of microbial virulence factors.

Keywords

enterococcus, transposons, plasmids, antimicrobial resistance, blast, data indexing, bioinformatics, data mining, metadata, metagenomics, algorithmic information

This article is included in the **Pathogens** gateway.

This article is included in the **Bioinformatics** gateway.

This article is included in the **Hackathons** collection.

This article is included in the **Antimicrobial Resistance** collection.

1

version 1

09 Feb 2022

[view](#)

1. **Rick Stevens**, Argonne National Laboratory,
Lemont, USA

James Davis, Argonne National Laboratory,
Lemont, USA

Any reports and responses or comments on the article can be found at the end of the article.

Corresponding author: Ben Busby (ben.busby@gmail.com)

Author roles: **Pickett BE:** Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Connor R:** Formal Analysis, Investigation, Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Berhanu-Denka T:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Bhalla S:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Brover V:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Chambers MJ:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Chaudhary K:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Cissé OH:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Dillman A:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Elmassry MM:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Feldgarden M:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Holloway E:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Huang X:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Klimke W:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Inês Mendes C:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Norred SE:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Parkinson J:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Sevilla S:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Garcia Solache M:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Surujon D:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Torian U:** Formal Analysis, Investigation, Methodology, Software, Writing – Review & Editing; **Zalunin V:** Data Curation, Formal Analysis, Investigation, Methodology, Resources, Software, Writing – Review & Editing; **Busby B:** Conceptualization, Funding Acquisition, Project Administration, Resources, Supervision, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: This work was partially supported by the Intramural Research Program of the National Library of Medicine. C.I.M. was supported by the Fundação para a Ciência e Tecnologia (grant SFRH/BD/129483/2017) and BacGenTrack [TUBITAK/0004/2014]; Fundação para a Ciência e a Tecnologia (FCT)/Scientific and Technological Research Council of Turkey (TÜBİTAK). *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.*

Copyright: © 2022 Pickett BE *et al.* This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The author(s) is/are employees of the US Government and therefore domestic copyright protection in USA does not apply to this work. The work may be protected under the copyright laws of other jurisdictions when used in those jurisdictions.

How to cite this article: Pickett BE, Connor R, Berhanu-Denka T *et al.* **Lessons learned in virulence factor identification and data management from a hackathon on microbial virulence [version 1; peer review: 1 approved with reservations]** F1000Research 2022, 11:160 <https://doi.org/10.12688/f1000research.26452.1>

First published: 09 Feb 2022, 11:160 <https://doi.org/10.12688/f1000research.26452.1>

Introduction

A variety of publications have reported the diverse mechanisms of microbial virulence (Leschem et al. 2020; Tarsillo and Priefer 2020; Nogueira et al. 2019; Saeki et al. 2020; Lange et al. 2019; Saiardi et al. 2018; Russo and Marr 2019; Feßler et al. 2018; Geisinger and Isberg 2017; Schroeder et al. 2017; Diard and Hardt 2017). In the vast majority of cases, one or more features/biomarkers can be identified and compared against databases of known virulence factors. Examples of such features can include genetic (e.g. pathogenicity islands, antimicrobial resistance (AMR) genes, mobile elements) (Kaushik et al. 2018; Vestergaard et al. 2019; Madec et al. 2017) and metabolomic (e.g. carbohydrates, lipids, small molecules) (Jia et al. 2019; Lloyd-Price et al. 2019) features.

The National Center for Biotechnology and Information (NCBI) at the National Institutes of Health hosts in-person hackathons to crowd-source the rapid development and prototyping of various methods (Connor et al. 2019). Recently (July 2019), a hackathon was convened to develop methods to detect various aspects of microbial virulence. The pipelines and methods that were generated during this hackathon were made publicly available to rapidly disseminate these tools and to enable the rapid identification of various microbial virulence factors.

The aim of this study was to develop computational pipelines that enabled cloud-based analyses to identify markers of microbial virulence in a modular and publicly available way. Attendees focused on implementing proof-of-concept strategies and methods to address “high-priority” needs. These tasks included data indexing, metabolic functions, virulence factors, AMR, mobile element detection in enterococci, and metatranscriptomics. Participants were grouped into teams, with each team developing at least one automated pipeline for their assigned task. These pipelines and workflows can be adapted by the research community to enable improved detection of various microbial virulence factors in a variety of data types.

Methods

Data indexing

All data and metadata from NCBI Sequence Read Archive (SRA) were assembled into contigs using SKESA version 2.3.0 (Souvorov et al. 2018). Virulence factors, AMR genes, and mobile elements were then identified and annotated in these contigs with AMRfinder (Feldgarden et al. 2019). These data and metadata were then uploaded to storage buckets and Big Query (respectively) in the Google Cloud Platform. A table containing pointers to the above data and metadata was then generated.

A script was then used to: 1) concatenate the metadata for each SRR in a subset of SRA experiments, 2) concatenate all assembled contigs for each SRR in the same subset of SRA experiments, and 3) apply a previously reported quantitative clustering by compression method (Cilibrasi and Vitanyi 2005), to identify any relationship between the complexity in contig sequence data and the associated metadata.

Metabolic functions

BioProjects within the NCBI-hosted Sequence Read Archive were selected for analysis based upon the presence of shotgun sequencing metagenomic samples from the gut microbiome and paired labels identifying the sample with a health status. Disease statuses encompass individuals diagnosed with colorectal cancer, diabetes, obesity, Parkinson's disease, and ulcerative colitis. Metagenomic gut samples from healthy individuals were also selected. A subset of 100 representative SRA accessions (Table 1) were selected for an initial characterization to prototype the metabolic functions workflow. Sequence read files for the selected SRA accessions were downloaded using SRA Toolkit version 2.9.6, followed by a read quality filter using PRINSEQ++ version 1.2 (Schmieder and Edwards 2011). All filtered reads contained paired ends, which were concatenated into a single read file using a custom script before being passed as input into HUMAnN2 version 2.8.1 using the UniRef90 full database to optimize the identification of uncharacterized proteins (Franzosa et al. 2018). Following the HUMAnN2 workflow, output files were normalized based on relative abundance followed by joining all sample tables to create a single table for gene families, path coverage, and path abundance. All three tables were preprocessed using a custom R script before being passed through a random forest classifier. A total of 100 trees were generated for the binary classification using the random forest algorithm implemented in the Python scikit-learn library. Owing to the low sample size with 18 healthy and 41 diseased samples (belonging to Parkinson's disease, obesity etc.), leave-one out cross validation was used.

Virulence factors

Experimentally validated virulence factor (VF) genes from the Virulence Factor Database (VFDB) were used to represent virulence-associated genes (Chen et al. 2016). Example metagenomes used for testing were drawn from public datasets listed on NCBI SRA and included healthy and disease-state human skin metagenomic samples (Table 1). Specifically, diseased metagenomes were drawn from the Diabetic Foot Ulcer metagenome study (BioProject: [PRJNA506988](#)) and

Table 1. Data used for the pipelines and workflows were obtained from publicly accessible repositories, including NCBI.

Pipeline/workflows	Accession numbers
<i>Metabolic functions</i> (Sequence Read Archive, NCBI)	SRR5903325; SRR5903328; SRR5903329; SRR5903330; SRR5903336; SRR5903338; ERR1018190; ERR1018226; ERR1018307; ERR1018293; SRR1778450; SRR413673; SRR413674; SRR413683; SRR413684; SRR413685; SRR413713; SRR413714; SRR413715; SRR413769; SRR413770; SRR413758; SRR413759; SRR413760; SRR413761; SRR1804586; SRR541350; SRR062467; SRR528218; SRR061181; SRR514185; SRR1803339; SRR541162; SRR062388; SRR1571144; SRR1565333; SRR540442; SRR541169; SRR528002; SRR1804737; SRR059993; SRR1031437; SRR1564357; SRR1804738; SRR341635; SRR341636; SRR341643; SRR341708; SRR341709; SRR341710; SRR341711; ERR1018258; ERR1018268; ERR688551; ERR688516; SRR2155174; SRR2155175; SRR2155176; SRR2155177; SRR2155178; SRR2155179; SRR2155306; SRR2155319; SRR2155333; SRR2155345; SRR2155356; SRR2155370; SRR2155382; SRR2155394; SRR2155482; ERR1912997; ERR1913127; ERR1912950; ERR1912975; ERR1913096; ERR1913053; ERR1913022; ERR1913083; ERR1913012; ERR1913104; ERR1913032; ERR1913019; ERR1913094; ERR1913071; ERR1912999; SRR5950679; SRR5935741; SRR5947016; SRR5946862; SRR5946961; SRR5950495; SRR5950608; SRR5935859; SRR5946838; SRR5950712; SRR5936199; SRR5950505; SRR5946688; SRR5950591; SRR5950789
<i>Virulence factors</i> (BioProject, NCBI)	PRJNA506988; PRJEB30094
AMR	ERR1600439
<i>Mobile Elements</i>	All 15,000 available <i>E. faecalis</i> and <i>E. faecium</i> sequence runs available as of July 2019.
<i>Metatranscriptomics</i>	PRJNA398089; PRJNA188481; PRJNA4068

healthy foot skin metagenomes were drawn from BioProject: [PRJEB30094](#). Metagenomes were assembled using MetaSPADES version 3.13.1 ([Nurk et al. 2017](#)).

A Hidden Markov Model (HMM) was applied to the VFDB genes to create virulence profiles. Genes were selected for which at least five different bacterial species were available. Multiple sequence alignments were generated using MUSCLE ([Edgar 2004](#)) and HMMs using HMMER3 ([Eddy 2011](#)). Genomes and/or corresponding protein coding sequences were screened with HMMSEARCH using pre-computed significance scores. Scores were calculated as 80% of the envelope alignment score of a representative sequence corresponding to its HMM. Alignments were filtered by custom scripts to extract putative virulence factors' loci. VF sequences were concatenated, aligned and used as input for phylogenetic analyses. Phylogenetic trees were constructed using RAXML-ng ([Kozlov et al. 2019](#)) and analyzed using the R package Ape ([Paradis and Schliep 2019](#)) and Newick Utilities ([Junier and Zdobnov 2010](#)). Virulence tags were assigned based on the number of virulence loci found and phylogenetic classification. All analyses are described within a Snakemake workflow ([Köster and Rahmann 2018](#)).

A support vector machine (SVM) model was also developed to classify virulent and non-virulent gene segments by training on a reference set of labelled pathogen and commensal genomes. The pathogen genomes were acquired from an NCBI Assembly search and included the species identified in the VFDB dataset. Commensal genomes were also acquired from an NCBI Assembly search, and included species selected from the [NHSN Common Commensals List](#) and from prior work ([Busby et al. 2013](#)). To characterize the novel virulence factors, the known virulence factors were masked from pathogenic genomes and non-pathogenic genomes.

AMR

Utilizing FASTQ files, either by providing them directly or by a list of SRA run accession numbers, this Nextflow pipeline creates a curated list of AMR genes present in a metagenomic sample and includes their context by performing a guided assembly with the detected genes. In this case, the sample was from a preterm infant ([ERR1600439](#)) with information of the context of the detected gene. After quality control, three modes to detect AMR genes are available: "magicblast", "hmm" and "mash", all utilizing the AMRFinder AMR database ([Feldgarden et al. 2019](#)). The first mode utilizes magicBLAST to align the sequencing data to the reference database and, after filtering by coverage breadth and depth, outputs the selected AMR genes found in the sample. The "hmm" mode takes the sequencing data and translates it into the six possible reading frames, and breaks each into open reading frames (ORFs). These ORFs are used as input to

the HMMER algorithm to search against the profile HMMs from AMRFinder and selects AMR sequences based on a threshold bit score. The final mode, “mash”, builds a MASH sketch with the same AMR database and screens the reads against the sketch. This produces k-mer distances between the read set and each AMR, which then is used to extract only the AMRs that are close to the readset. The resulting AMRs, from any mode of the pipeline, are used as bait to perform a guided assembly with SKESA version 2.3.0.

Mobile elements/enterococcus

All 15,000 available *E. faecalis* and *E. faecium* sequence runs were downloaded from the SRA database. Short reads were assembled using SKESA version 2.3.0 (Souvorov et al. 2018), with AMRFinder used to identify resistance genes (Feldgarden et al. 2019). Genomes were annotated using the Prokaryotic Genome Annotation Pipeline version 4.8 (Tatusova et al. 2016). BLAST (Altschul et al. 1990) was used to compare the assembled sequences against a database of transposon and plasmid sequences. Annotations were also searched for terms associated with mobile elements, including transposase, recombinase, integrase. After performing the described BLAST analyses, a report that combined and summarized these results was constructed.

Metatranscriptomics

Processed data was downloaded from the Inflammatory Bowel Disease Multi'omics Database (IBDMDB; PRJNA398089). Patient metadata, including inflammatory bowel disease information, was also obtained from the same source. Metagenomics and metatranscriptomics data were obtained as taxon and pathway abundance tables, respectively. Comparison of metagenomic and metatranscriptomic abundances were done using custom code.

To see whether metatranscriptomic data is predictive of disease state of an individual, seven machine learning classification models were trained and subsequently combined in a consensus model that classifies disease outcome. A random selection of a third of the individuals were placed in a test set to evaluate the performance of the consensus model and these individuals were not used in any training step. The input features consisted of the relative abundance of each metabolic pathway, which were normalized to have 0 mean and equal variance. The models trained include Decision Tree, K-Nearest Neighbor, Support Vector Machine, Logistic Regression, Random Forest, and Multi-Layered Perceptron. Each model was trained on the training data, using hyper-parameter tuning with five-fold cross-validation. The seven models were then combined using a majority vote scheme into a consensus model. The performance of the consensus was evaluated on the training and previously unseen test datasets. All machine learning analysis was done using the Python library scikit-learn.

Implementation

The Github code repository for each of the described projects contains sufficient documentation to install all of the tools required for each of the workflows (see *Software availability*).

Operation

All workflows were designed to run on modern cloud-based hardware, such as the Google Cloud Platform. This approach can maximize data processing efforts, minimize data transfer, and efficiently complete the analyses. Overviews of each workflow are provided in Figures 1, 3, 5, and 6.

Results

Before beginning the hackathon, we assembled the relevant sequence data from SRA into contigs prior to annotating them with known virulence factors, AMR genes, and mobile elements (for Enterococcus-related samples). We then added these processed data, together with the original SRA metadata and the pre-computed virulence metadata into the Google Cloud Platform. These data and metadata formed the basis for the downstream analyses described below.

Data indexing

The data indexing group focused on the underlying requirement by funding agencies to submit the experimental data, together with the associated metadata, to public data repositories. These large repositories include an extremely large corpus from which to perform data mining activities on existing data. Such activities are enhanced when the submissions conform to the findable, accessible, interoperable, and reusable (FAIR) data principles (Reiser et al. 2018). While the repositories provide a search interface that enables filtering by user-specified criteria, searching on relationships between datasets and supporting downstream data mining activities are often not supported. As such, the data included in this hackathon were all hosted in the cloud, using Big Query within the Google Cloud Platform. This approach facilitates the rapid identification of relevant datasets based on not only SRA metadata, but also using preprocessed data and metadata including assembled contigs, presence of AMR genes, and other elements in the contigs. Providing community access to this, and other relevant information in the cloud will improve the pace of developing solutions for human health.

Modular Approach for Expanding and Extending Metadata

Figure 1. Relevant metadata that could be indexed. A graphical representation showing the important relationship between data generation, data analysis, and data indexing.

The aim of this effort was to measure the feasibility of storing the assembled contigs, original SRA metadata, and additional metadata in a cloud-based platform to facilitate downstream analytical workflows (Figure 1). The benefit of this design is that both the tools and the data are co-hosted in the same cloud environment. This approach minimizes time required for file transfer and ensures that workflows are consistently accessible to all users.

In addition to the storage of data, we hypothesized that the complexity, or sequence diversity, of contigs could correlate with the complexity of the associated metadata (e.g. disease state, year, host species, AMR gene presence). To quantitatively compare metadata and sequence data we used a technique based on algorithmic information theory. Briefly, this approach takes advantage of the fact that two related sequences show a high degree of sequence similarity (i.e. low genetic distance), and the files containing this information could therefore be compressed more efficiently. Similarly, files containing metadata that is highly repetitive would also compress much more efficiently. As such, we applied a previously published metric of mutual information that compares how well sequences and metadata compress individually and how well they compress when concatenated together (Cilibrasi and Vitanyi 2005) (Figure 2). This algorithm was shown to be agnostic to the input data and consequently robust in a variety of applications including DNA sequences. Their previous study showed that known biological taxonomies could be replicated without having to adjust the metric. A related paper demonstrated that comparing the dictionary sequences compared to the logarithmic size of the dictionary, it is possible to identify statistically significant sequences by length above a threshold (Milosavljević and Jurka 1993). Applying this method to pairs of SRR identifiers from SRA and their associated metadata revealed better compression of the metadata overall than the sequence data. We expect that the combination of more detailed and descriptive metadata for future SRA submissions, together with generating additional metadata from the relevant sequence data will improve the performance of this analytical method.

Metabolic functions

The metabolic functions group focused on identifying metabolic signatures of relevant virulence factors from a subset of the approximately one million metagenomic datasets that have been deposited at NCBI in the past two decades. Tens of thousands of these sequences were human-related. Due to the diversity in the metabolic capacity of the human microbiome, the role of the gut microbiome has been extensively studied in metabolic disorders and chronic diseases. At the same time, the value of machine learning has been recognized and applied to study complex datasets such as next-generation sequencing datasets. The goal of this effort was to uncover the missing metabolic functions that could exacerbate disease or disorder conditions. Finally, the established model would help predict the gut microbiome metabolic profile that may be associated with dysbiosis in the human body.

The goal of this project was to identify unique metabolic functions in gut metagenomes for disease states such as colorectal cancer, diabetic, obese, Parkinson's disease, and ulcerative colitis populations. We began by retrieving a variety of publicly available human samples from the SRA. These samples were derived from patients with a variety of disease states and were subjected to a novel computational workflow that transforms metagenomics sequencing reads into metabolic profiles (Figure 3). Specifically, we generated 100 trees in the random forest to select the features with the depth of 2 for pruning and classification. We then calculated the useful performance metrics for this approach including

Figure 2. Data compression workflow. The computational workflow that was used to test whether sequence complexity in contigs relates to metadata complexity.

Figure 3. Metabolic function workflow. A graphical description of the computational workflow used to identify metabolic functions from metagenomics data.

precision (1.00, 0.72), recall (0.11, 1.00), and F1-score (0.20, 0.84) for non-disease vs disease samples respectively, along with the area under the curve value of 0.86 (Figure 4). Overall, we observe that this cloud-based workflow shows a high amount of promise for differentiating between disease and non-disease samples using only metagenomic data.

Virulence factors

The virulence factors group developed methods to better understand how disease progression could be impacted by: metagenomic identification of microbial virulence, community relationships, and environmental context. Maintaining an appropriate scope for the identification of new virulence factors can often be challenging. Especially since certain contexts, such as the introduction of at least one virulence factor in a population can significantly alter the landscape of the

Figure 4. ROC curve for distinguishing patients with or without disease. AUC quantifies the ability of the machine-learning based process to correctly assign healthy and unhealthy patients based on metagenomics data.

surrounding microbiome and the pathological potential of individual microbes, including those normally considered commensal. Crucially, few tools exist for probing microbial virulence in metagenomes, which constrains the development of culture-free methods in public health and epidemiology.

We found it difficult to manually identify and acquire datasets with sufficiently detailed metadata for this task. After identifying an appropriate metagenomic study, the data from both healthy and diseased human skin microbiomes was subjected to the HMM portion of the workflow (Figure 5). We found that assembling the sequencing reads was relatively

Figure 5. Virulence factors workflow. A graphical representation of the workflow to predict the presence of virulence factors from metagenomic datasets.

fast but was still the limiting factor in the overall speed of the analysis. Although the workflow that we implemented seems logical, additional software improvements are necessary to improve tractability and obtain the desired results in a reasonable amount of time.

AMR

AMR has been identified as a global concern, as microorganisms worldwide gain the ability to resist drugs that have been used to treat common infections. Identifying AMR genes has been the focus of several organizations, leading to the curation of large bacteria-focused databases such as the Bacterial Antimicrobial Resistance Reference Gene Database, and a call for the characterization of these genes (Buckner et al. 2018). Plasmids, which are self-replicating elements, can contain AMR genes and as they are often transmissible through horizontal gene transfer, have become a concern for the global spread of antibiotic resistance (San Millan 2018). Given that a human microbiome sample includes thousands of species, surveillance and the detection of AMR genes is critical. Developing tools that can not only identify which AMR genes are present but that can also classify which bacterial species the AMR genes are likely from, and whether

Figure 6. AMR detection workflow. A depiction of the computational workflow used to identify antimicrobial resistance markers in metagenomic datasets.

these AMR genes were likely derived from a plasmid, is extremely informative. The intersection of this information is of utmost importance as plasmids have been identified in the spread of AMR in specific bacterial families, such as *Enterobacteriaceae* and *Enterococcaceae*.

We designed each of the pipelines such that they generated a list of AMR gene “hits” from the sequencing reads used as input. In this case we used the publicly available human fecal gut SRA sample [ERR1600439](#), available from NCBI. We observed that the number of hits varied depending on the algorithm that was used. Specifically, we found that magicBLAST identified 17 AMR candidates and MASH identified 45 AMR candidates. After guided assembly, each pipeline created a varied number of contigs, with magicblast creating seven, which was the most generated by any of the three algorithms.

Mobile elements/enterococci

Infections caused by *Enterococcus* spp. present a major threat to public health as a leading cause of nosocomial infections. One major contributor to the severity of infections caused by enterococci is its frequent resistance to antimicrobial therapy. The association of resistance with mobile elements is of particular concern, as this can result in the rapid spread of AMR to previously susceptible isolates. Thus, the development of tools to identify mobile elements and their association with resistance genes will fill an important gap in current knowledge and provide a useful resource for public health and research applications.

Our mobile element (“mobilome”) analysis was implemented as a computational workflow capable of identifying antimicrobial resistance genes that are flanked by active mobile elements ([Figure 7](#)). The tabular report that was generated by the workflow consisted of rows that contained information about each isolate-feature combination. The location, feature type, analysis method, name, and relevant BLAST statistics for each isolate-feature combination are reported in a tab-delimited text format. BLAST was used to compare each individual genome against each reference plasmid to determine if a particular reference plasmid is present in the sequence of a given isolate. The global coverage of each plasmid by each isolate is reported in a tab-delimited text format.

Use cases

We wanted to construct a pilot workflow within this cloud-based platform to rapidly and effectively preprocess and analyze metatranscriptomic data. This workflow could be used to support future analyses to distinguish between active and inactive/dormant bacterial taxa in a given sample. The aim for this use case was to determine whether we could detect such differences in bacterial populations using metatranscriptomics, and whether any biological insight could be derived from the results. While many interesting questions can be pursued by analyzing metatranscriptomic data, we decided to

Figure 7. Quantified RNA and DNA values in two bacterial species. Mean RNA and DNA abundance values for *Escherichia coli* (left) and *Bacteroides vulgatus* (right). Each point represents a patient sample. Most patients have lower (relative) amounts of RNA from *B. vulgatus* compared to DNA; and the converse is true for *E. coli*.

Figure 8. Quantified RNA and DNA abundance values in two bacterial metabolic pathways. Arginine biosynthesis (left) and Fucose-rhamalose catabolism (right) pathways. Each point is a patient sample. For the arginine synthesis pathway, RNA and DNA abundance correlate well. However, for the fucose-rhamalose catabolism pathway is an example of low variability in terms of DNA abundance, but high variability in RNA abundance.

quantify differences between levels of DNA and RNA for a given gene--to distinguish between persistence and active metabolism.

Persisters are often defined as metabolically inactive bacteria, characterized by arrested growth, low ATP levels, and low mRNA expression. Bacteria can enter a persistent state when exposed to supra-lethal concentrations of antimicrobial compounds, and can resume growth after the compound is removed from the environment (Fisher et al. 2017). Thus, persistence is a form of transient antibiotic tolerance, and can result in treatment failure and relapse of infection in the clinic (even in the absence of antibiotic resistance). It is therefore crucial to be able to detect persisters in clinical samples. One possible method of detection of persisters is genome-wide absence of RNA transcripts, but a presence of genomic DNA. This can be accomplished by considering the mean or median RNA vs DNA abundances across all pathways of each individual species (Figure 7).

Similar to how a human cell with one genome can differentiate into vastly diverse cell types, similar metagenomic compositions in microbial communities can have vastly different transcriptional/functional profiles (Franzosa et al. 2014). Many genes/pathways demonstrate a strong correlation between RNA and DNA abundance. However, there are also examples of pathways where DNA abundance does not vary much across samples, but RNA abundance does (Figure 8).

Recent studies classified healthy vs diseased patients based on NGS data (Bang et al. 2019) (Pasolli et al. 2016). Early analysis of metatranscriptomic data has identified differentially-expressed pathways in periodontal disease when compared to healthy oral microbial communities (Jorth et al. 2014). Metatranscriptomic profiling data potentially has richer and more informative features than metagenomic or 16S data; since it provides information on which genes and pathways are being actively expressed, which may contribute to a host phenotype. We therefore selected the IBDMDB dataset for training ML models since it has metatranscriptomics data, and disease metadata.

As a pilot experiment, we trained seven machine learning models (decision tree, logistic regression, SVM, Naive Bayes, K-nearest neighbor, random forest and multi-layer perceptron) using the pathway abundance data extracted from the metatranscriptomics data from the HMP2 cohort in IBDMDB. A majority voting consensus model was also generated using the outputs of the 7 individual trained models (Figure 9). This workflow demonstrates the advantage of using metatranscriptomic data to stratify patients. We expect that this technology could be improved in the future to enable the accurate identification of persister/dormant bacteria that may contain virulence factors.

Discussion

This work describes our efforts to demonstrate the usefulness of using existing cloud-based platforms to not only perform analytical tasks related to identifying markers of microbial virulence, but also to use these platforms to store the relevant

Figure 9. Confusion matrix of the consensus model for detecting disease. The training set (left) and test set (right). nonIBD: no Inflammatory Bowel Disease. UC: Ulcerative colitis. CD: Crohn's Disease.

data and metadata. Specifically, we generated proof-of-concept computational workflows for data storage and indexing, metabolic functions, virulence factors, antimicrobial resistance, and mobile element detection in enterococci.

As metadata and derived metadata become more descriptive and comprehensive in the future, we expect that data indexing and storage will continue to play an important role in data mining activities. The subtask that involved comparing the complexity of compressed sequence data and metadata provided a foundation upon which to build and improve. We anticipate that the degree of metadata compression will be reduced as the metadata becomes more expansive and less repetitive overall. The increased metadata complexity could approach the complexity of the sequencing data in the future. The increased amounts of metadata would therefore help with both identifying relevant datasets by applying filtering criteria as well as with predicting additional related datasets. Overall, efforts for continued data indexing, together with the association of relevant and informative metadata are imperative to facilitate the future use and re-use of public datasets. In addition, future meta-analyses that both increase sample size and augment knowledge will benefit from enhancing existing minimum metadata reporting standards.

The ability to accurately identify metabolic functions associated with the disease or non-disease state from metagenomic data will be useful as sequencing-based assays continue to gain traction in the clinic. When sufficient datasets and metadata are used to train the machine learning model, this proof of concept shows that the presence of a disease state, and potentially even the identification of the specific disease state, can be possible. Similarly, identifying gene patterns that are specific to particular disease states and their contribution to disease progression would be invaluable as we seek to better understand the underlying mechanism(s) of disease.

We envision the computational identification of potential virulence factors to be useful not only in current data mining applications, but also to identify potential microbial factors that could be impacting human health once sequencing becomes more routine in the clinical setting. The results from a similar workflow could be used to inform the best treatment strategy for a given individual.

The computational AMR gene detection workflow that we implemented can be useful in clinical contexts where time from sample collection to prescribed treatment is precious. For example, rather than waiting for minimum inhibitory concentration (MIC) assay results to be performed and reported by the clinical lab, sequencing a sample and running an AMR-detection pipeline could inform the clinician which AMR genes are present in the sample and could nicely complement the MIC results.

The mobilome analysis will be used in cases where users are interested in the transmissibility of genomic material, particularly related to antimicrobial resistance. Association of resistance genes with transposons, plasmids, or other mobile elements can yield valuable information on the transmission of resistance through bacterial populations. This can help track the emergence and spread of antimicrobial resistance on particular genetic elements.

Overall, we found the NCBI hackathon for Microbial Virulence in the Cloud was extremely productive as measured by the various workflows, analyses, and results that were generated. These tools and results can be built upon by the scientific research community to improve existing methods and initiate important discussions about roadblocks and successes associated with our approach. For example, combining the various virulence detection methods could enable the rapid identification of emerging pathogens, while linking metabolism and signaling results with expression analyses could be used to identify novel indicator genes and therapeutic targets.

Conclusions and next steps

We are pleased with the publicly available workflows that were generated during this hackathon to build tools that can be applied to augment our understanding of microbial virulence and antimicrobial resistance. Potential next steps for this effort could include one or more of the following tasks: making the existing versions of pipelines and algorithms more robust, supporting the analysis of multiple input file types, applying and comparing the results from other algorithms, import workflow(s) into the [NCBI Pathogen Detection Pipeline](#), dockerize tools and/or port them to external analysis platforms, improve data normalization methods, formalize the various database structure to facilitate data transfer across platforms, improve ease-of-use by developing web-based graphical user interfaces for a subset of tools, develop a database with the ability to take tab-delimited files containing metadata or other values and generate a table with standardized column names, and ensure the compatibility of tools, data structures, and containers across multiple cloud-based platforms.

Each of the workflows implemented during this hackathon greatly benefited from the availability of public data in the cloud. The analyses that were performed incorporated statistical analysis to minimize bias and quantify observed variability. Beginning each analysis from data and algorithms hosted in the cloud shows the relevance of cloud-based resources such as Google Big Query for storing, retrieving, and mining large datasets and repositories. This approach and the underlying cloud platform are both scalable in a way that should allow the support of other organisms and additional metadata. Throughout this process the extreme importance of accurate, descriptive, and available metadata became apparent. In addition, we found that augmenting the current metadata for each experiment by identifying features such as the presence of AMR or other virulence factors could improve downstream reuse and analysis. In summary, this hackathon quickly generated a number of publicly available workflows that can contribute to future research efforts related to AMR, metagenomics, and metatranscriptomics.

Data availability

Underlying data

The data used for these projects were obtained from publicly accessible repositories, including NCBI, and are available in [Table 1](#).

Software availability

Pipelines and workflows available from:

- Virulence AMR Data Index Resource: https://github.com/NCBI-Hackathons/Virulence_AMR_Data_Index_Resource/
- Nasty Metagenomes: https://github.com/NCBI-Hackathons/Nasty_Metagenomes
- MetaClaMP-ML: <https://github.com/NCBI-Hackathons/MetaClaMP-ML>
- Virulence Factor Characterization: https://github.com/NCBI-Hackathons/Virulence_Factor_Characterization
- Enterococcus Mobile Elements: https://github.com/NCBI-Hackathons/Enterococcus_Mobile_Elements
- Enterococcus Mobile Elements: https://github.com/NCBI-Hackathons/Metatranscriptomics_Pilot

Archived pipelines and workflows as at time of publication:

- Virulence AMR Data Index Resource: <https://doi.org/10.5281/zenodo.5227710> (DrCapt et al. 2021)
- Nasty Metagenomes: <https://doi.org/10.5281/zenodo.5361759> (Mendes et al. 2021)
- MetaClaMP-ML: <https://doi.org/10.5281/zenodo.5236567> (Elmassry et al. 2021)
- Virulence Factor Characterization: <https://doi.org/10.5281/zenodo.5338778> (Payne et al. 2021)
- Enterococcus Mobile Elements: <https://doi.org/10.5281/zenodo.5565736> (Ghtyson et al. 2021)
- Metatranscriptomics Pilot: <https://doi.org/10.5281/zenodo.5227692> (Defne & Busby 2021)

License: MIT

Acknowledgments

We thank Gregory Tyson at the U.S. Food and Drug Administration, Center for Veterinary Medicine, Office of Research and Shurjo Sen at the Division of Genome Sciences, National Human Genome Research Institute. We also gratefully acknowledge NCBI and the U.S. National Institutes of Health for hosting this hackathon. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other federal agencies. These agencies had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

- Altschul SF, Gish W, Miller W, *et al.*: **Basic local alignment search tool.** *J. Mol. Biol.* 1990; **215**(3): 403–410.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bang S, Yoo D, Kim S-J, *et al.*: **Establishment and evaluation of prediction model for multiple disease classification based on gut microbial data.** *Sci. Rep.* 2019; **9**(1): 10189.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Buckner MMC, Ciusa ML, Piddock LJV: **Strategies to combat antimicrobial resistance: anti-plasmid and plasmid curing.** *FEMS Microbiol. Rev.* 2018; **42**(6): 781–804.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Busby B, Kristensen DM, Koonin EV: **Contribution of phage-derived genomic islands to the virulence of facultative bacterial pathogens.** *Environ. Microbiol.* 2013; **15**(2): 307–312.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Chen L, Zheng D, Liu B, *et al.*: **VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on.** *Nucleic Acids Res.* 2016; **44**(D1): D694–D697.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Cilibrasi R, Vitanyi PMB: **Clustering by Compression.** *IEEE Trans. Inf. Theory.* 2005 Apr; **51**(4): 1523–1545.
[Publisher Full Text](#)
- Connor R, Brister R, Buchmann JP, *et al.*: **Ncibi's virus discovery hackathon: engaging research communities to identify cloud infrastructure requirements.** *Genes.* 2019; **10**(9).
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Defne AAD, Busby B: **NCBI-Hackathons/Metatranscriptomics_Pilot: First release (v0.9).** *Zenodo.* 2021.
[Publisher Full Text](#)
- Diard M, Hardt W-D: **Evolution of bacterial virulence.** *FEMS Microbiol. Rev.* 2017; **41**(5): 679–697.
[PubMed Abstract](#) | [Publisher Full Text](#)
- DrCapt, Yters, Bpickett, *et al.*: **NCBI-Hackathons/Virulence_AMR_Data_Index_Resource: First release of the Data Index Resources (v0.9).** *Zenodo.* 2021.
[Publisher Full Text](#)
- Eddy SR: **Accelerated profile HMM searches.** *PLoS Comput. Biol.* 2011; **7**(10): e1002195.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Edgar RC: **MUSCLE: multiple sequence alignment with high accuracy and high throughput.** *Nucleic Acids Res.* 2004; **32**(5): 1792–1797.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Elmassry M, Chambers M, *et al.*: **NCBI-Hackathons/MetaClAMP-ML: Current release - MetaClAMP-ML (v1.0.0).** *Zenodo.* 2021.
[Publisher Full Text](#)
- Feldgarden M, Brover V, Haft DH, *et al.*: **Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates.** *Antimicrob. Agents Chemother.* 2019; **63**(11).
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Feßler AT, Wang Y, Wu C, *et al.*: **Mobile macrolide resistance genes in staphylococci.** *Plasmid.* 2018; **99**: 2–10.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Fisher RA, Gollan B, Helaine S: **Persistent bacterial infections and persister cells.** *Nat. Rev. Microbiol.* 2017; **15**(8): 453–464.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Franzosa EA, McIver LJ, Rahnavard G, *et al.*: **Species-level functional profiling of metagenomes and metatranscriptomes.** *Nat. Methods.* 2018; **15**(11): 962–968.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Franzosa EA, Morgan XC, Segata N, *et al.*: **Relating the metatranscriptome and metagenome of the human gut.** *Proc. Natl. Acad. Sci. U. S. A.* 2014; **111**(22): E2329–E2338.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Geisinger E, Isberg RR: **Interplay Between Antibiotic Resistance and Virulence During Disease Promoted by Multidrug-Resistant Bacteria.** *J. Infect. Dis.* 2017; **215**(suppl_1): S9–S17.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ghtyson ES, Feldgard VB, WarriorScientist M, Busby B: **NCBI-Hackathons/Enterococcus_Mobile_Elements: v0.9.0 (v0.9.0).** *Zenodo.* 2021.
[Publisher Full Text](#)
- Jia Q, Li H, Zhou H, *et al.*: **Role and effective therapeutic target of gut microbiota in heart failure.** *Cardiovasc. Ther.* 2019; **2019**: 5164298.
- Jorth P, Turner KH, Gumus P, *et al.*: **Metatranscriptomics of the human oral microbiome during health and disease.** *MBio.* 2014; **5**(2): e01012–e01014.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Junier T, Zdobnov EM: **The Newick utilities: high-throughput phylogenetic tree processing in the UNIX shell.** *Bioinformatics.* 2010; **26**(13): 1669–1670.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kaushik M, Kumar S, Kapoor RK, *et al.*: **Integrins in Enterobacteriaceae: diversity, distribution and epidemiology.** *Int. J. Antimicrob. Agents.* 2018; **51**(2): 167–176.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Köster J, Rahmann S: **Snakemake—a scalable bioinformatics workflow engine.** *Bioinformatics.* 2018; **34**(20): 3600.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Kozlov AM, Darriba D, Flouri T, *et al.*: **RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference.** *Bioinformatics.* 2019; **35**(21): 4453–4455.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lange PT, Lagunoff M, Tarakanova VL: **Chewing the fat: the conserved ability of DNA viruses to hijack cellular lipid metabolism.** *Viruses.* 2019; **11**(2).
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Leshem A, Liwinski T, Elinav E: **Immune-Microbiota Interplay and Colonization Resistance in Infection.** *Mol. Cell.* 2020; **78**: 597–613.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lloyd-Price J, Arze C, Ananthakrishnan AN, *et al.*: **Multi-omics of the gut microbial ecosystem in inflammatory bowel diseases.** *Nature.* 2019; **569**(7758): 655–662.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Madec JY, Haenni M, Nordmann P, *et al.*: **Extended-spectrum β -lactamase/AmpC- and carbapenemase-producing Enterobacteriaceae in animals: a threat for humans?** *Clin. Microbiol. Infect.* 2017; **23**(11): 826–833.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Mendes SI, Huang X, Zalunin V, *et al.*: **NCBI-Hackathons/Nasty_Metagenomes: First release (v0.9.0).** *Zenodo.* 2021.
[Publisher Full Text](#)
- Milosavljević A, Jurka J: **Discovering simple DNA sequences by the algorithmic significance method.** *Computer applications in the biosciences : CABIOS.* 1993; **9**(4): 407–411.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Nogueira T, David PHC, Pothier J: **Antibiotics as both friends and foes of the human gut microbiome: The microbial community approach.** *Drug Dev. Res.* 2019; **80**(1): 86–97.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Nurk S, Meleshko D, Korobeynikov A, *et al.*: **metaSPAdes: a new versatile metagenomic assembler.** *Genome Res.* 2017; **27**(5): 824–834.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Paradis E, Schliep K: **ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R**. *Bioinformatics*. 2019; **35**(3): 526–528.
[PubMed Abstract](#) | [Publisher Full Text](#)

Pasolli E, Truong DT, Malik F, *et al.*: **Machine Learning Meta-analysis of Large Metagenomic Datasets: Tools and Biological Insights**. *PLoS Comput. Biol.* 2016; **12**(7): e1004977.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Payne J, Norred N, Cisse OH, Busby B: **NCBI-Hackathons/Virulence_Factor_Characterization: First release of this software (v0.9.0)**. *Zenodo*. 2021.
[Publisher Full Text](#)

Reiser L, Harper L, Freeling M, *et al.*: **FAIR: A call to make published data more findable, accessible, interoperable, and reusable**. *Mol. Plant*. 2018; **11**(9): 1105–1108.
[PubMed Abstract](#) | [Publisher Full Text](#)

Russo TA, Marr CM: **Hypervirulent *Klebsiella pneumoniae***. *Clin. Microbiol. Rev.* 2019; **32**(3).
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Saeki EK, Kobayashi RKT, Nakazato G: **Quorum sensing system: Target to control the spread of bacterial infections**. *Microb. Pathog.* 2020; **142**: 104068.
[PubMed Abstract](#) | [Publisher Full Text](#)

Saiardi A, Azevedo C, Desfougères Y, *et al.*: **Microbial inositol polyphosphate metabolic pathway as drug development target**. *Advances in biological regulation*. 2018; **67**: 74–83.
[PubMed Abstract](#) | [Publisher Full Text](#)

Advances in biological regulation. 2018; **67**: 74–83.

[PubMed Abstract](#) | [Publisher Full Text](#)

San Millan A: **Evolution of Plasmid-Mediated Antibiotic Resistance in the Clinical Context**. *Trends Microbiol.* 2018; **26**(12): 978–985.
[PubMed Abstract](#) | [Publisher Full Text](#)

Schmieder R, Edwards R: **Quality control and preprocessing of metagenomic datasets**. *Bioinformatics*. 2011; **27**(6): 863–864.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Schroeder M, Brooks BD, Brooks AE: **The Complex Relationship between Virulence and Antibiotic Resistance**. *Genes*. 2017; **8**(1).
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Souvorov A, Agarwala R, Lipman DJ: **SKESA: strategic k-mer extension for scrupulous assemblies**. *Genome Biol.* 2018; **19**(1): 153.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Tarsillo B, Priefer R: **Proteobiotics as a new antimicrobial therapy**. *Microb. Pathog.* 2020; **142**: 104093.
[PubMed Abstract](#) | [Publisher Full Text](#)

Tatusova T, DiCuccio M, Badretdin A, *et al.*: **NCBI prokaryotic genome annotation pipeline**. *Nucleic Acids Res.* 2016; **44**(14): 6614–6624.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)

Vestergaard M, Frees D, Ingmer H: **Antibiotic resistance and the MRSA problem**. *Microbiology spectrum*. 2019; **7**(2).
[PubMed Abstract](#) | [Publisher Full Text](#)

Open Peer Review

Current Peer Review Status: ?

Version 1

Reviewer Report 03 March 2022

<https://doi.org/10.5256/f1000research.29206.r123011>

© 2022 Stevens R et al. This is an open access peer review report distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Rick Stevens

Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL, USA

James Davis

Computing, Environment and Life Sciences, Argonne National Laboratory, Lemont, IL, USA

The report summarizes a hackathon that was held at NCBI in July of 2019. The participants in the group conducted pilot studies to explore building pipelines for identifying metabolic functions, virulence factors, AMR, mobile elements, and analyzing metatranscriptomic data. The workflows and software are listed in the article and are available for public use.

Overall, while none of the workflows appear to be at a point of development where they stand alone as publishable entities, I thought that this was a fairly interesting article since it summarized the results of a seemingly productive hackathon. The paper is also a useful reference for various methods relating to these topics. In my opinion, there is a place for a publication of this type and that it should be accepted on its merit, assuming that it meets the requirements of this journal. My comments are fairly minor.

- I didn't think the title fully matched what had been done in the hackathon.
- I found the justification of the need for the cloud to be tenuous, all of the other sections of the report and workflows have some biological justification, and could in theory stand alone. If that part is to be kept, it seems like it should go last with an explanation of how to use the data in the cloud (assuming that was built).
- We know what AMR stands for, but please define "AMRs". Is this genes, proteins, etc? "AMRs" also appears at the bottom of figure 6.
- In the mobile elements section, "Figure 7" has the wrong figure number.
- Figure 9 would be easier to read with the raw counts instead of scientific.

- Accuracy or F1 scores of the transcriptomic disease model(s) and other models could be stated in the results

Is the rationale for developing the new software tool clearly explained?

Partly

Is the description of the software tool technically sound?

Partly

Are sufficient details of the code, methods and analysis (if applicable) provided to allow replication of the software development and its use by others?

Yes

Is sufficient information provided to allow interpretation of the expected output datasets and any results generated using the tool?

Yes

Are the conclusions about the tool and its performance adequately supported by the findings presented in the article?

Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Bioinformatics, AI, large-scale biological databases, machine learning, HPC

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however we have significant reservations, as outlined above.

The benefits of publishing with F1000Research:

- Your article is published within days, with no editorial bias
- You can publish traditional articles, null/negative results, case reports, data notes and more
- The peer review process is transparent and collaborative
- Your article is indexed in PubMed after passing peer review
- Dedicated customer support at every stage

For pre-submission enquiries, contact research@f1000.com

F1000Research