SOFTWARE TOOL ARTICLE

NastyBugs: A simple method for extracting antimicrobial resistance information from metagenomes [version 1; peer review: 2 approved with reservations]

Hsinyi Tsang¹, Matthew Moss², Greg Fedewa ³, Sharif Farag⁴, Daniel Quang⁵, Alexey V. Rakov ⁶, Ben Busby ⁷

¹Center for Biomedical Informatics and Information Technology, National Cancer Institute, National Institutes of Health, Gaithersburg, MD, 20850, USA

²Cold Spring Harbor Laboratory, Cold Spring Harbor, NY, 11724, USA

³Bioinformatics, University of California, San Francisco, San Francisco, CA, 94158, USA

⁴Bioinformatics and Computational Biology, Eshelman School of Pharmacy, University of North Carolina at Chapel Hill, Chapel Hill, NC, 27599, USA

⁵Department of Computer Science, Donald Bren School of Information and Computer Sciences, University of California, Irvine, Irvine, CA, 92617, USA

⁶Department of Pathobiology, School of Veterinary Medicine, University of Pennsylvania, Philadelphia, PA, 19104, USA

⁷National Center for Biotechnology Information, National Library of Medicine, Bethesda, MD, 20894, USA

V1 First published: 08 Nov 2017, 6:1971
<https://doi.org/10.12688/f1000research.12781.1>
 Latest published: 08 Nov 2017, 6:1971
<https://doi.org/10.12688/f1000research.12781.1>

Abstract

Multidrug resistant bacteria are becoming a major threat to global public health. While there are many possible causes for this, there have so far been few adequate solutions to this problem. One of the major causes is a lack of clinical tools for efficient selection of an antibiotic in a reliable way. NastyBugs is a new program that can identify what type of antimicrobial resistance is most likely present in a metagenomic sample, which will allow for both smarter drug selection by clinicians and faster research in an academic environment.

Keywords

Metagenome, microbiome, antibiotic, resistance, bacterial, genomic signature, MDR, workflow

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

Open Peer Review

Approval Status ? ?

	1	2
version 1
08 Nov 2017	view	view

1. **Torsten Seemann** , University of Melbourne, Melbourne, Australia

2. **Tom J. B. de Man**, Centers for Disease Control and Prevention, Atlanta, USA

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

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

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

Corresponding authors: Hsinyi Tsang (tsang@mail.nih.gov), Ben Busby (ben.busby@nih.gov)

Author roles: **Tsang H:** Conceptualization, Methodology, Project Administration, Software, Writing – Review & Editing; **Moss M:** Methodology, Software, Writing – Review & Editing; **Fedewa G:** Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Farag S:** Methodology, Software, Writing – Review & Editing; **Quang D:** Methodology, Software, Writing – Review & Editing; **Rakov AV:** Methodology, Software, Writing – Original Draft Preparation, Writing – Review & Editing; **Busby B:** Conceptualization, Methodology, Resources, Supervision, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: Ben Busby was funded by the Intramural Research Program of the NLM. Hsinyi Tsang was funded by NCI Contract #D14PD00826.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Copyright: © 2017 Tsang H *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: Tsang H, Moss M, Fedewa G *et al.* **NastyBugs: A simple method for extracting antimicrobial resistance information from metagenomes [version 1; peer review: 2 approved with reservations]** F1000Research 2017, **6**:1971 <https://doi.org/10.12688/f1000research.12781.1>

First published: 08 Nov 2017, **6**:1971 <https://doi.org/10.12688/f1000research.12781.1>

Introduction

Antimicrobial resistance (AMR) of bacterial pathogens is a growing public health threat around the world. Most concerning are multidrug resistant (MDR) bacteria, which have become more prevalent in recent decades¹. Well known examples of these pathogens include methicillin-resistant *Staphylococcus aureus* (MRSA), vancomycin-resistant *S. aureus*, extended spectrum beta-lactamase, and vancomycin-resistant *Enterococcus*. MRSA is prevalent in surgical and maternity hospitals and nursing homes, where it is often associated with hospital-acquired infection with high morbidity and mortality². The current method of determining if a patient has a MDR infection is based on being able to culture a patient-derived sample in the presence of different antibiotic drugs³. This is a slow process that can take days to weeks, which can put the patient in danger of not receiving the correct antibiotic in time.

There are several known mechanisms by which bacteria acquire AMR. One common mechanism involves acquiring resistance through horizontal genetic transfer (HGT), which can include plasmid-, phage-, transposon-, and integron-mediated resistance⁴. The second major mechanism involves SNPs in chromosomal genes that can result in a change in antibiotic binding sites⁵.

High-throughput whole genome sequencing (WGS) of microbiomes is a state-of-the-art method for studying complex microbial communities, such as the human gut. WGS creates large raw data sets, which must be processed quickly and efficiently to guide clinicians for the best and most efficient treatment strategy for the given patient. However, simple, clinically applicable bioinformatics methods that can provide a fast, reusable, reproducible, scalable pipeline to locate AMR genomic signatures in large metagenomics datasets from the NCBI Sequence Read Archive (SRA) and other public datasets are still lacking. Such pipelines could also be used for crowdsourcing of this analysis, such as with undergraduate students. The problem of determining efficient strategies for antibiotic usage is the keystone of the modern antibacterial therapy and prevention⁶.

In the last few years, a variety of different papers and tools have been developed that exploit AMR detection for both complete genomes and metagenomes. Some of the existing detection methods for AMR genomic signatures include: ResFinder⁷, PointFinder⁸, SSTAR⁹, DeepARG¹⁰, ARIBA¹¹ and ResCap¹². Another approach is Galaxy-based pipeline Amr++¹³. All detection methods depend on the availability of collections of known AMR genomic signatures. These signatures are then directly searched for, or models are generated for the detection of novel AMR genes/loci. One of the most updated manually curated databases is the Comprehensive Antibiotic Resistance Database (CARD)¹³; others include ResFinder⁷, ARG-ANNOT¹⁴, and MegaRES¹⁵. Although some of these tools provide user-friendly web interfaces and use both FASTA and FASTQ files as input, they do not use the power of command line. Moreover, these solutions are not universal, e.g. ResFinder searches for only HGT-mediated resistance, whereas its successor PointFinder only looks for AMR caused by chromosomal point mutations. Other disadvantages of the existing solutions include an inability to work with big datasets or multiple

raw sequence files, slow speed, and poor handling of metagenomic data.

The primary objective of this project is designing a reliable system for rapid diagnostics and prompt treatment of patients with MDR infections. At the heart of the system is a reusable, reproducible, scalable, and interoperable workflow to locate AMR genomic signatures in SRA shotgun sequencing (including metagenomics) datasets. To ease this task we used only RefSeq reference genomes for the bacterial pathogens important for public health, but the pipeline can be scalable to include databases for other microbes, viruses, and fungi. The result, NastyBugs, is a new program that can identify what type of antimicrobial resistance is most likely present in a metagenomic sample, which will allow for both smarter drug selection by clinicians and faster research done in an academic environment. NastyBugs is a framework created during the National Center of Biotechnology Information Hackathon in August 2017.

Methods

The detailed workflow to extract antimicrobial resistance gene signatures is described in [Figure 1](#).

Three BLAST databases for downstream analysis were created using the latest versions of: 1) RefSeq human genome assembly GRCh37/UCSC hg19 (RefSeq accession no. GCF_000001405.37); 2) [RefSeq bacterial genomes](#); 3) [CARD](#)¹³. For comparison purposes, CARD was used as two databases: one consisted of genes and another of SNPs.

Further analysis consists of three steps: 1) Host (human) reads removal; 2) Antimicrobial resistance signature identification; 3) Bacterial identification and characterization. Steps 2 and 3 were performed in parallel. Input data is SRA accession numbers (ERR or SRR) of the metagenome of interest. Another option is using FASTQ files from local storage.

Host reads removal

Using [STAR](#)¹⁶ or [Magic-BLAST](#), all reads mapped to human genome GRCh37/hg19 were removed, and unmapped non-human reads (considered as bacterial) were collected using [SAMtools](#) for further analysis.

Antimicrobial resistance signature identification

To remove adapters/linkers/barcodes we used [FASTX Clipper and Trimmer](#). The non-human reads are mapped again using Magic-BLAST; however this time they are mapped to bacterial genes/SNPs from the CARD. This allows for the identification of genes and SNPs that can lead to antimicrobial resistance in the bacterial population. Obtained reads were sorted and the sum of read abundance was calculated.

Bacterial identification and characterization

The identification of bacterial species and abundance is carried out in parallel. For that we again used Magic-BLAST and the database of NCBI RefSeq reference bacterial genomes. The resulting list of species was visualized using [Krona](#)¹⁷.

Figure 1. An overview of the NastyBugs pipeline.

Output format

The output TAB-delimited formatted file contains data in five columns: 1) RefSeq accession number; 2) genus; 3) resistance gene; 4) ARO (Antibiotic Resistance Ontology) accession number; 5) score (number of mapped reads per 1kb). The data can be used for constructing a scatter plot showing relative abundance of antimicrobial resistance and the corresponding bacterial species in metagenomic sample analysed.

Dependencies

The documented workflow contains the script with containerized tools in Docker.

We used the following dependencies: 1) Magic-BLAST v. 1.3, a novel tool allowing mapping of large next-generation RNA or DNA sequencing runs against a whole genome; 2) SAMtools v. 1.3.1, a popular suite of programs for interacting with HTS data; 3) FASTX-Toolkit v. 0.0.13, a collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing; 4) STAR v. 2.5.3a, RNA-seq aligner; and 5) Krona v. 2.7, a tool for metagenomic pie chart visualization.

Use case

To validate our pipeline we used two human gut metagenomic datasets, SRA acc. no. [ERR1600439](#) and [SRR5239736](#). The SRR5239736 sample was used for comparison of our results with the results obtained by ResCap¹². For metagenome sample SRR5239736, 24% of reads were mapped to the gene database of CARD and 1.6% reads were mapped to the SNP database of CARD.

Magic-BLAST, a novel program from the BLAST family, can provide a faster and more accurate way to align reads of interest with reference sequences. A quick comparison of STAR and Magic-BLAST showed at least a 10-fold difference in speed increase with Magic-BLAST for mapping of SRA reads to human genome compared to STAR. For that reason, we chose not to use STAR in the pipeline.

Conclusion and next steps

Obtained results showed high efficiency of identification of antibiotic signatures in the studied samples. However, the presented workflow may be improved. Planned improvements will include: 1) optimization of the pipeline; 2) additional large-scale validation for different metagenomic samples; 3) representation of results with more information; 4) adding information about proteins participating in AMR; and 5) prediction of novel resistance genes using Hidden Markov Model.

Moreover, implementation of machine learning analysis may provide additional capabilities.

The pipeline can be used to efficiently identify the presence of antimicrobial resistant genes, which in turn can be used as features for further downstream machine learning analysis. One useful application of machine learning in antimicrobial resistance is the prediction of the appropriate antimicrobial therapy to apply to a critically ill patient. For these patients, the time taken to administer an appropriate antibiotic agent inversely correlates with improved patient outcomes¹⁸. Whole genome sequencing of microbial isolates, followed by antimicrobial

resistance genes identification and machine learning prediction provides an attractive solution to this problem. A previous application in this regard applied a simple rules-based approach and a logistic regression model¹⁹. More sophisticated, non-linear supervised machine learning methods, such as random forests, gradient boosting, and artificial neural networks may play a key role in producing accurate predictions for clinical use. Artificial neural networks, such as convolutional and recurrent neural networks, are particularly interesting as they may be able to extract novel features.

Data and software availability

The code for the pipeline is publically available on GitHub: <https://github.com/NCBI-Hackathons/MetagenomicAntibioticResistance>.

Archived source code as at time of publication: <http://doi.org/10.5281/zenodo.1020266>²⁰

License: MIT

Competing interests

No competing interests were disclosed.

Grant information

Ben Busby was funded by the Intramural Research Program of the NLM. Hsinyi Tsang was funded by NCI Contract #D14PD00826.

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgements

The authors would like to acknowledge NCBI for hosting the hackathon, Grzegorz Boratyn, Sean Davis and Lisa Federer for technical discussions.

References

- Friedman ND, Temkin E, Carmeli Y: **The negative impact of antibiotic resistance.** *Clin Microbiol Infect.* 2016; **22**(5): 416–22.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Boucher HW, Corey GR: **Epidemiology of Methicillin-Resistant Staphylococcus aureus.** *Clin Infect Dis.* 2008; **46** Suppl 5: S344–49.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Bonev B, Hooper J, Parisot J: **Principles of assessing bacterial susceptibility to antibiotics using the agar diffusion method.** *J Antimicrob Chemother.* 2008; **61**(6): 1295–301.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Gyles C, Boerlin P: **Horizontally transferred genetic elements and their role in pathogenesis of bacterial disease.** *Vet Pathol.* 2014; **51**(2): 328–40.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Woodford N, Ellington MJ: **The emergence of antibiotic resistance by mutation.** *Clin. Microbiol. Infect.* 2007; **13**(1): 5–18.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Lee CR, Cho IH, Jeong BC, *et al.*: **Strategies to minimize antibiotic resistance.** *Int J Environ Res Public Health.* 2013; **10**(9): 4274–305.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zankari E, Hasman H, Cosentino S, *et al.*: **Identification of acquired antimicrobial resistance genes.** *J Antimicrob Chemother.* 2012; **67**(11): 2640–4.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Zankari E, Allesøe R, Joensen KG: **PointFinder: a novel web tool for WGS-based detection of antimicrobial resistance associated with chromosomal point mutations in bacterial pathogens.** *J Antimicrob Chemother.* 2017; **72**(10): 2764–8.
[Publisher Full Text](#)
- de Man TJ, Limbago BM: **SSTAR, a Stand-Alone Easy-To-Use Antimicrobial Resistance Gene Predictor.** *mSphere.* 2016; **1**(1): pii: e00050-15.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Arango-Argoty GA, Garner E, Pruden A, *et al.*: **DeepARG: A deep learning approach for predicting antibiotic resistance genes from metagenomic data.** *bioRxiv.* 2017; 149328.
[Publisher Full Text](#)
- Hunt M, Mather AE, Sánchez-Busó L, *et al.*: **ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads.** *Microbial Genomics.* 2017.
[Publisher Full Text](#)
- Lanza VF, Baquero F, Martinez JL, *et al.*: **In-depth resistome analysis by targeted metagenomics.** *bioRxiv.* 2017; 104224.
[Publisher Full Text](#)
- Jia B, Raphenya AR, Alcock B, *et al.*: **CARD 2017: expansion and model-centric curation of the comprehensive antibiotic resistance database.** *Nucleic Acids Res.* 2017; **45**(D1): D566–73.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Gupta SK, Padmanabhan BR, Diene SM, *et al.*: **ARG-ANNOT, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes.** *Antimicrob Agents Chemother.* 2014; **58**(1): 212–20.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Lakin SM, Dean C, Noyes NR, *et al.*: **MEGARes: an antimicrobial resistance database for high throughput sequencing.** *Nucleic Acids Res.* 2017; **45**(D1): D574–80.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Dobin A, Davis CA, Schlesinger F, *et al.*: **STAR: ultrafast universal RNA-seq aligner.** *Bioinformatics.* 2013; **29**(1): 15–21.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Ondov BD, Bergman NH, Phillippy AM: **Interactive metagenomic visualization in a Web browser.** *BMC Bioinformatics.* 2011; **12**(1): 385.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Kumar A, Roberts D, Wood KE, *et al.*: **Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock.** *Crit Care Med.* 2006; **34**(6): 1589–96.
[PubMed Abstract](#) | [Publisher Full Text](#)
- Peseky MW, Hussain T, Wallace M, *et al.*: **Evaluation of Machine Learning and Rules-Based Approaches for Predicting Antimicrobial Resistance Profiles in Gram-negative Bacilli from Whole Genome Sequence Data.** *Front Microbiol.* 2016; **7**: 1887.
[PubMed Abstract](#) | [Publisher Full Text](#) | [Free Full Text](#)
- Rakov A, harper357, DCGenomics, Tsang S, *et al.*: **NCBI-Hackathons/ MetagenomicAntibioticResistance: Nastybugs.** *Zenodo.* 2017.
[Data Source](#)

Open Peer Review

Current Peer Review Status: ? ?

Version 1

Reviewer Report 08 December 2017

<https://doi.org/10.5256/f1000research.13848.r27751>

© 2017 de Man T. 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.

Tom J. B. de Man

Centers for Disease Control and Prevention, Atlanta, GA, USA

The authors describe a new tool that is able to detect antimicrobial resistance (AMR) genes and identify bacterial species composition directly from metagenomic datasets. This pipeline removes the need to culture bacterial specimens and perform PCR for AMR gene detection and could potentially guide patient antimicrobial therapy.

Introduction comments:

Extended-spectrum beta-lactamase (ESBL) is associated with what bacteria? (e.g., Enterobacteriaceae?)

Please define the acronym 'SNP' (Single Nucleotide Polymorphism) the first time it is used within the manuscript.

PointFinder is not the successor of ResFinder, they serve different purposes.

Methods comments:

It is not immediately clear when STAR or Magic-Blast will be selected for read mapping. Or is STAR not used in the current pipeline? There seems to be conflicting comments about the use of STAR in the manuscript ("Methods" versus "Use Case").

The pipeline screens out human reads prior to further analysis; however, the authors might also consider removing PhiX reads in order to get cleaner output in your Krona plots.

Bacterial species composition is determined using Bacterial RefSeq and Magic-Blast. How does this compare to tools like Kraken? Did the authors consider using a different tool than Magic-Blast for determining bacterial species composition? Is there a way to link bacterial species to identified AMR genes?

The paper does not mention what cut-offs are being used for reporting AMR genes in terms of coverage and percentage sequence similarity. Is this something that can be adjusted by the user?

Output format comments:

It would be useful for the reader to have a figure that illustrates and explains the output generated by this pipeline.

GitHub repository comments:

KRONA is missing from the dependency list on the authors GitHub repo

The “workflow method” section on the GitHub page does not contain any STAR aligner commands, only Magic-Blast. Users will be interested in the STAR parameters as well. Or is STAR not part of the pipeline anymore?

Overall comments:

Tried to run the pipeline but it does not work. The pipeline, by default, is not able to find RefSeq database files. It would be a good solution to let the user choose the location of databases using a parameter option.

Pipeline has no help function or usage string when executed without any input parameters.

Disclaimer:

The findings and conclusions in this review are those of the author and do not necessarily represent the views of the Centers for Disease Control and Prevention.

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

Yes

Is the description of the software tool technically sound?

No

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

No

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

No

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

No

Competing Interests: No competing interests were disclosed.

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

Reviewer Report 20 November 2017

<https://doi.org/10.5256/f1000research.13848.r27750>

© 2017 Seemann T. 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.

Torsten Seemann

University of Melbourne, Melbourne, Vic, Australia

"A Software Tool Article should include the rationale for the development of the tool and details of the code used for its construction. The article should provide examples of suitable input data sets and include an example of the output that can be expected from the tool and how this output should be interpreted."

The authors present a software tool which claims to detect AMR genes from raw metagenomic FASTQ reads. However, there are several problems with the manuscript and the source code.

The discussion of existing tools, methods and databases is confusing. The term "AMR genomic signature" is not defined. ResCap is mentioned, but that is a wet lab set of probes, not a bioinformatics method per se. CARD does have a command line tool, it is called "RGI". This section would benefit from a table of methods which explain if they are a tool, a database, or both; whether they work with contigs or FASTQ reads; and whether they work on bacterial isolates or metagenomes.

The algorithm is poorly explained. Why was Magic-BLAST chosen as an aligner? Did you turn splicing off (as bacteria don't have introns). It is stated that reads were removed of adaptors etc using Fastx_Clipper but no details of which adaptor files were used. The database used was CARD, but not explanation of how it was split into HGT and SNP/Variant parts. Alignment to Bacterial RefSeq is also done to get a species distribution, but it is unclear why or how, and how it links to the genes found. No examples of output results are provided, nor is it described how you identify the SNPs.

I was unable to install and run the software. The github site <https://github.com/NCBI-Hackathons/MetagenomicAntibioticResistance> is very confusing also. It seems I have to use Docker to just get MagicBlast, but it doesn't include all the other dependencies. The primary script is called "main.h" but it appears to have lots of hard-coded parameters (eg. -threads 12) and refers to reference database files that do not exist. It also has bash syntax errors like "var= wc -l file" which should probably be "var=`wc -l file`" or "var=\$(wc -l file)" - it appears something has lost characters in the file? The docs say "The pipeline use three databases that should be downloaded with the script". But there is no such script. I tried hard to find more info, but I was unable to.

In summary, more detail is needed, and the software needs to be installable and have some basic tests.

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

Partly

Is the description of the software tool technically sound?

No

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

No

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

No

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

No

Competing Interests: No competing interests were disclosed.

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I 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