A roadmap for the generation of benchmarking resources for antimicrobial resistance detection using next generation sequencing

Mauro Petrillo, Marco Fabbri, Dafni Maria Kagkli, Maddalena Querci, Guy van den Eede, Erik Alm, Derya Aytan-Aktug, Salvador Capella-Gutierrez, Catherine Carrillo, Alessandro Cestaro, et al.

OPINION ARTICLE

REVISED A roadmap for the generation of benchmarking resources for antimicrobial resistance detection using next generation sequencing [version 2; peer review: 1 approved, 2 approved with reservations]

Mauro Petrillo[1*], Marco Fabbri[1*], Dafni Maria Kagkli[1*], Maddalena Querci[1], Guy Van den Eede[1,2], Erik Alm[3], Derya Aytan-Aktug[4], Salvador Capella-Gutierrez[5], Catherine Carrillo[6], Alessandro Cestaro[7], Kok-Gan Chan[8,9], Teresa Coque[10,11], Christoph Endrullat[12], Ivo Gut[13,14], Paul Hammer[15], Gemma L. Kay[16], Jean-Yves Madec[17], Alison E. Mather[16,18], Alice Carolyn McHardy[19], Thierry Naas[20], Valentina Paracchini[1], Silke Peter[21], Arthur Pightling[22], Barbara Raffael[1], John Rossen[23], Etienne Ruppé[24], Robert Schlaberg[25], Kevin Vanneste[26], Lukas M. Weber[27-29], Henrik Westh[30], Alexandre Angers-Loustau[31*]

[1]European Commission Joint Research Centre, Ispra, Italy
[2]European Commission Joint Research Centre, Geel, Belgium
[3]The European Centre for Disease Prevention and Control, Stockholm, Sweden
[4]National Food Institute, Technical University of Denmark, Lyngby, Denmark
[5]Barcelona Supercomputing Centre (BSC), Barcelona, Spain
[6]Ottawa Laboratory – Carling, Canadian Food Inspection Agency, Ottawa, Ontario, Canada
[7]Fondazione Edmund Mach, San Michele all'Adige (TN), Italy
[8]Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur, Malaysia
[9]International Genome Centre, Jiangsu University, Zhenjiang, China
[10]Spanish Consortium for Research on Epidemiology and Public Health (CIBERESP), Carlos III Health Institute, Madrid, Spain
[11]Servicio de Microbiología, Hospital Universitario Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, Spain
[12]MSD SHARP & DOHME GMBH, Haar, Germany
[13]Universitat Pompeu Fabra, Barcelona, Spain
[14]Centro Nacional de Análisis Genómico, Centre for Genomic Regulation (CNAG-CRG), Barcelona Institute of Technology, Barcelona, Spain
[15]BIOMES. NGS GmbH c/o Technische Hochschule Wildau, Wildau, Germany
[16]Quadram Institute Bioscience, Norwich Research Park, Norwich, UK
[17]Unité Antibiorésistance et Virulence Bactériennes, ANSES Site de Lyon, Lyon, France
[18]University of East Anglia, Norwich, UK
[19]Helmholtz Centre for Infection Research, Braunschweig, Germany
[20]French-NRC for CPEs, Service de Bactériologie-Hygiène, Hôpital de Bicêtre, Le Kremlin-Bicêtre, France
[21]Institute of Medical Microbiology and Hygiene, University of Tübingen, Tübingen, Germany
[22]Center for Food Safety and Applied Nutrition, US Food and Drug Administration, College Park, MD, USA
[23]Department of Medical Microbiology, University Medical Center Groningen, University of Groningen, Groningen, The Netherlands
[24]IAME, Université de Paris, Paris, France
[25]Department of Pathology, University of Utah, Salt Lake City, UT, USA

[26]Transversal activities in Applied Genomics, Sciensano, Brussels, Belgium
[27]Present address: Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
[28]SIB Swiss Institute of Bioinformatics, University of Zurich, Zurich, Switzerland
[29]Institute of Molecular Life Sciences, University of Zurich, Zurich, Switzerland
[30]Hvidovre University Hospital, Hvidovre, Denmark
[31]European Commission Publications Office, Luxembourg, Luxembourg

* Equal contributors

v2 **First published:** 08 Feb 2021, **10**:80
https://doi.org/10.12688/f1000research.39214.1
Latest published: 16 Mar 2022, **10**:80
https://doi.org/10.12688/f1000research.39214.2

Abstract

Next Generation Sequencing technologies significantly impact the field of Antimicrobial Resistance (AMR) detection and monitoring, with immediate uses in diagnosis and risk assessment. For this application and in general, considerable challenges remain in demonstrating sufficient trust to act upon the meaningful information produced from raw data, partly because of the reliance on bioinformatics pipelines, which can produce different results and therefore lead to different interpretations. With the constant evolution of the field, it is difficult to identify, harmonise and recommend specific methods for large-scale implementations over time. In this article, we propose to address this challenge through establishing a transparent, performance-based, evaluation approach to provide flexibility in the bioinformatics tools of choice, while demonstrating proficiency in meeting common performance standards. The approach is two-fold: first, a community-driven effort to establish and maintain "live" (dynamic) benchmarking platforms to provide relevant performance metrics, based on different use-cases, that would evolve together with the AMR field; second, agreed and defined datasets to allow the pipelines' implementation, validation, and quality-control over time. Following previous discussions on the main challenges linked to this approach, we provide concrete recommendations and future steps, related to different aspects of the design of benchmarks, such as the selection and the characteristics of the datasets (quality, choice of pathogens and resistances, etc.), the evaluation criteria of the pipelines, and the way these resources should be deployed in the community.

Keywords

Antimicrobial resistance, bioinformatics, next-generation sequencing, benchmarking

This article is included in the Bioinformatics gateway.

Open Peer Review

Approval Status ? ✓ ?

	1	2	3
version 2 (revision) 16 Mar 2022		✓ view	? view
version 1 08 Feb 2021	? view	? view	

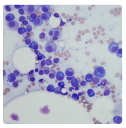
1. **Rene Hendriksen**, Technical University of Denmark, Bygning, Denmark

2. **Anna Abramova**, University of Gothenburg, Gothenburg, Sweden
Marcus Wenne, University of Gothenburg, Gothenburg, Sweden

3. **Enrico Lavezzo**, University of Padova, Padua, Italy
Emilio Ispano, University of Padova, Padua, Italy

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

This article is included in the Pathogens gateway.



This article is included in the Cell & Molecular Biology gateway.

This article is included in the Antimicrobial Resistance collection.

Corresponding authors: Mauro Petrillo (mauro.petrillo@ext.ec.europa.eu), Marco Fabbri (Marco.FABBRI@ec.europa.eu)

Author roles: Petrillo M: Conceptualization, Methodology, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Fabbri M**: Conceptualization, Methodology, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing; **Kagkli DM**: Conceptualization, Methodology, Writing – Original Draft Preparation, Writing – Review & Editing; **Querci M**: Conceptualization, Funding Acquisition, Methodology, Project Administration, Supervision, Writing – Review & Editing; **Van den Eede G**: Conceptualization, Funding Acquisition, Project Administration, Supervision, Writing – Review & Editing; **Alm E**: Conceptualization, Methodology, Writing – Review & Editing; **Aytan-Aktug D**: Conceptualization, Methodology, Writing – Review & Editing; **Capella-Gutierrez S**: Conceptualization, Methodology, Writing – Review & Editing; **Carrillo C**: Conceptualization, Methodology, Writing – Review & Editing; **Cestaro A**: Conceptualization, Methodology, Writing – Review & Editing; **Chan KG**: Conceptualization, Methodology, Writing – Review & Editing; **Coque T**: Conceptualization, Methodology, Writing – Review & Editing; **Endrullat C**: Conceptualization, Methodology, Writing – Review & Editing; **Gut I**: Conceptualization, Methodology, Writing – Review & Editing; **Hammer P**: Conceptualization, Methodology, Writing – Review & Editing; **Kay GL**: Conceptualization, Methodology, Writing – Review & Editing; **Madec JY**: Conceptualization, Methodology, Writing – Review & Editing; **Mather AE**: Conceptualization, Methodology, Writing – Review & Editing; **McHardy AC**: Conceptualization, Methodology, Writing – Review & Editing; **Naas T**: Conceptualization, Methodology, Writing – Review & Editing; **Paracchini V**: Conceptualization, Methodology, Writing – Review & Editing; **Peter S**: Conceptualization, Methodology, Writing – Review & Editing; **Pightling A**: Conceptualization, Methodology, Writing – Review & Editing; **Raffael B**: Conceptualization, Methodology, Writing – Review & Editing; **Rossen J**: Methodology, Writing – Review & Editing; **Ruppé E**: Conceptualization, Methodology, Writing – Review & Editing; **Schlaberg R**: Conceptualization, Methodology, Writing – Review & Editing; **Vanneste K**: Conceptualization, Methodology, Writing – Review & Editing; **Weber LM**: Conceptualization, Methodology, Writing – Review & Editing; **Westh H**: Conceptualization, Methodology, Writing – Review & Editing; **Angers-Loustau A**: Conceptualization, Methodology, Supervision, Visualization, Writing – Original Draft Preparation, Writing – Review & Editing

Competing interests: No competing interests were disclosed.

Grant information: The "Benchmarking of NGS bioinformatics pipelines for AMR" meeting (27-28 of May 2019) was funded by the European Commission's Joint Research Centre (JRC), Ispra, Italy. AEM is supported by the Biotechnology and Biological Sciences Research Council Institute Strategic Programme Grant Microbes in the Food Chain BB/R012504/1 and its constituent project BBS/E/F/000PR10348 (Theme 1, Epidemiology and Evolution of Pathogens in the Food Chain).
The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

How to cite this article: Petrillo M, Fabbri M, Kagkli DM *et al.* **A roadmap for the generation of benchmarking resources for antimicrobial resistance detection using next generation sequencing [version 2; peer review: 1 approved, 2 approved with reservations]** F1000Research 2022, **10**:80 https://doi.org/10.12688/f1000research.39214.2

First published: 08 Feb 2021, **10**:80 https://doi.org/10.12688/f1000research.39214.1

REVISED **Amendments from Version 1**

This version contains text additions following the suggestions and comments from the reviewers in their referee reports. These additions include:

- A revised version of sections: *General considerations*, *Section 3*, and *Conclusions*.

- Correction of wrong numbering of sections of the manuscript.

- Update of references and insertion of the suggested ones

Any further responses from the reviewers can be found at the end of the article

1. Introduction

The technological advances in Whole Genome Sequencing (WGS) and the increasing integration of Next Generation Sequencing (NGS) platforms in the arsenal of testing laboratories is having a profound impact on health sciences. Affordable human genome sequencing is bringing about an era of improved diagnostics and personalised healthcare. For microorganisms, the reliable characterisation of their genetic material allows improved insights in their identity and physiology. For example, once sequenced, the genome of a microorganism can also be used to (re-)identify the species and infer important phenotypic properties, such as virulence, resistance to antibiotics, typing and other adaptive traits. In addition, novel strategies for the implementation and analysis of NGS data are being developed and improved, and they can be used, for instance, to reconstruct the timeline and relationships between the cases of an infectious disease outbreak, which is something difficult to achieve with classical microbiological techniques.

An important aspect of the implementation of NGS technologies are considerations related to quality and consistency (see 1), in particular if the result of the method is to be used in a regulatory context (for example, in a monitoring framework) or, more importantly, in a clinical setting linked to decisions on medical treatments[2–4], veterinary, agricultural or environmental interventions and food safety[5,6] which may be linked under One Health initiatives.

Methods for predicting antimicrobial resistance (AMR) based on genetic determinants from NGS data rely on complex bioinformatics algorithms and procedures to transform the large output produced by the sequencing technologies into relevant information. Traditionally, regulatory implementation of analytical methods focuses on harmonisation of the protocol and the subsequent steps of analysis, i.e. ensuring the implementation of specific methods previously validated according to a set of criteria. For methods with important bioinformatics components, this is often not optimal, due to both the large variability in the developed strategies, variations in the particular computational resources available and the speed at which technologies and analytical approaches evolve. For the prediction of AMR determinants, very different strategies have been proposed, processing the sequencing data either as a set of reads or as pre-processed assemblies[7,8], even using neural networks[9]; sometimes, the system itself is proprietary and operates as a "black box" from the point of view of the user. In such cases like this, it has been proposed to approach the quality assurance challenge through performance-based evaluations, i.e. ensuring that the implemented methods, although different, perform at a similar (acceptable) level in this context[10]. The same performance-based evaluation can then be applied whenever a component of the pipeline, or its environment, is replaced or updated.

An important component for a performance-based evaluation scheme is the availability of resources (in particular, datasets) that enable these evaluations[11–13]. In 2017, the Joint Research Centre (JRC) initiated a reflection on the subject by inviting experts in the field of AMR detection with NGS from the four compartments of a "One Health" perspective, i.e. clinics, food, animals and the environment[14,15]. These discussions led to a compilation of the challenges involved in the development of a benchmark strategy for bioinformatics pipelines, both for NGS-based approaches in general and in this specific field of application[16]. These challenges were grouped into often overlapping categories, including the nature of the samples in the dataset (e.g. their origin, quality and associated metadata), their composition (e.g. the determinants and species to include), their use (e.g. expected results and performance thresholds) and their sustainability (e.g. their development, release and update).

On the 27th and 28th of May 2019, the JRC held a follow-up meeting, including most of the authors of the original article and additional experts that expressed interest, to discuss and propose solutions to the identified challenges for AMR detection using next generation sequencing. The present article represents a summary of these discussions and the conclusions reached. We propose this document as a baseline for a roadmap and guidelines to harmonise and standardise for the generation of the benchmark resources in the field of AMR.

2. Framing the aims and purposes of the benchmarking resources

An important observation that arose from the two-day discussions is that the concept of benchmarking, even when focusing on a single component of the method (i.e. the bioinformatics pipeline), may refer to different activities that can vary in their scope and objectives (see also 17–19). Clarifying these scopes is crucial when proposing recommendations, as these (and the final datasets) will be influenced by the scope of the evaluation.

In the conclusions of Angers *et al.* article, the use of the benchmark resources was reported as follows: *"(1) Ensuring confidence in the implementation of the bioinformatics component of the procedure, a step currently identified as limiting in the field. (2) Allowing evaluation and comparison of new/existing bioinformatics strategies, resources and tools. (3) Contributing to the validation of specific pipelines and the proficiency testing of testing laboratories and (4) "Future-proofing" bioinformatics pipelines to updates and replacement of tools and resources used in their different steps*."[14].

These four summarising points made above, in practice, cover two different questions: 1, 3 and 4 (implementation, validation,

proficiency testing and future proofing) ask whether the bioinformatics pipeline performs as expected, while 2 (evaluation/comparison) focuses on identifying gold standard pipelines and resources for implementation. The first scope of a benchmark resource would thus address the question: "Which pipeline performs best and at least by the agreed minimum standards?" A second scope addresses the question: "What is the quality of the information produced by the implemented bioinformatics pipeline?"

The latter question requires further refinement, based on the "what" the pipeline is "required" to achieve. Although there may be different contexts to the use of the methods (e.g. guide clinical intervention, contribute data to a monitoring framework, outbreak management, monitor the spread of AMR genes (ARGs) in or between different settings/environments, etc.), for a benchmark resource, these can be split in three main use cases:

- to predict the resistance of a pathogen of interest based on the identification of ARGs (either acquired ARGs and/or chromosomal point mutations conferring AMR);
- to identify the link of ARGs to a specific taxon in a metagenomic sample (i.e. taxon-binning approaches like described by Sangwan *et al.*[20]);
- to identify the complete repertoire of the AMR determinants (i.e. the resistome) in a complex bacterial community from which total genomic DNA has been extracted and sequenced (i.e. the metagenome).

Finally, another important scope for a benchmark resource was identified, having, once again, an impact on the decisions regarding the benchmark dataset: "How **robust** is the bioinformatics pipeline?" Studies addressing this question focus on identifying how the pipelines can tolerate variation in characteristics of the input data, most often related to the quality of the sample or sequencing steps: robustness against contamination or low number/poor quality reads, for example. Robustness, in certain contexts, could also be seen as the effect (or lack of) of swapping a tool (or the reference database) at a certain step in the pipeline for a different one that is functionally equivalent (see, for example,[21]).

In summary, it is important to be specific about the purpose and scope of the benchmark resource in the decisions taken when generating the datasets. We propose that the scope of a benchmark has three major parts, summarised in Figure 1.

General considerations

When discussing the different challenges described in 16, rarely can an absolute "best" answer be identified for a given question; recommendations thus need to be made, taking into account the specific purpose of the benchmark resource and the fact that they may evolve with the state-of-the-art in the field.

Still, some general observations and conclusions were proposed regarding difficulties like what type of AMR mechanisms to include, which pathogens to consider, lack of expertise and of harmonisation, rapid evolution of the fields of informatics and bioinformatics in parallel to progress of scientific knowledge on AMR. They are summarised in this section and, by discussing two main use cases for benchmarking resources (single isolates and mixed samples), represent proposals for a way through these and other challenges with common, transparent, performance-based evaluation approaches, to be applied in different fields in a "One Health" perspective.

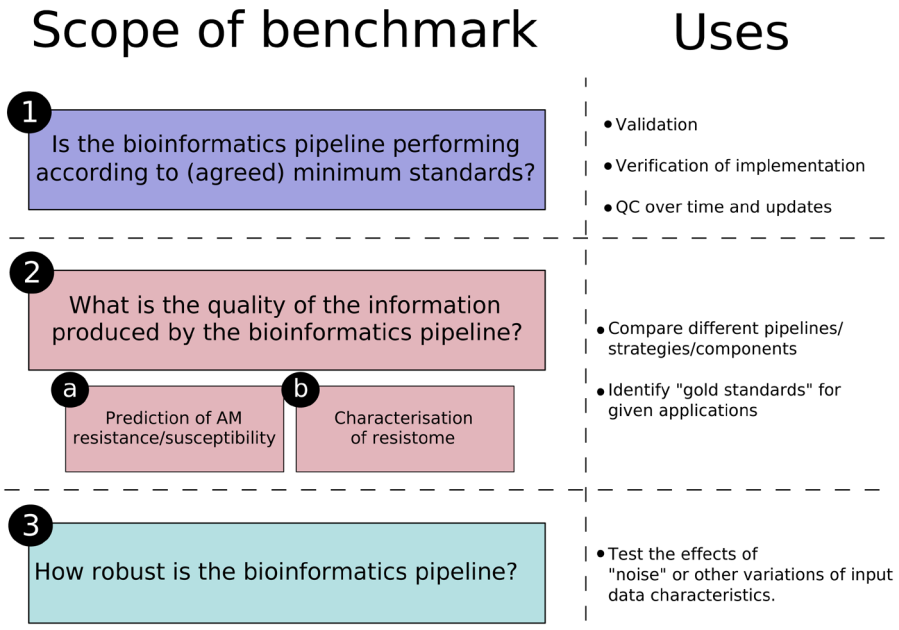


Figure 1. Summary of the different "scopes" for the benchmark resources for AMR detection using next generation sequencing discussed in the current document, with an indication of the uses for each.

2.1. NGS platforms

A quick analysis of the different NGS platforms currently available and in development makes it obvious that the set of reads that they produce have very different characteristics. In addition, each platform has its strengths and weaknesses. Both the error rate (about 0.1% for Illumina (RRID:SCR_010233), 1–10% for newer technologies like from Pacific Biosciences and Oxford Nanopore Technologies (RRID:SCR_003756)) and the types of errors (miscalls, insertions or deletions, or problems of particular motifs such as homopolymer sequences) vary according to the platform used. The average length of the reads can vary from hundreds (Illumina, Ion Torrent) to thousands (PacBio, Nanopore) of base pairs[22–24].

Bioinformatics pipelines are thus usually designed to handle the output of a specific platform, often in a certain configuration. Although exceptions exist (e.g. 25,26), in the context of a benchmark resource (and independently of the question asked), we thus believe that different datasets are needed for each of the different NGS platforms, each containing reads that have a profile that matches closely with the normal output of the respective technologies. It is important, in this case, to ensure that the datasets produced for the different platforms are developed so that they do not inadvertently introduce any bias resulting in apparent difference in performance among the different platforms due to composition of the benchmark dataset (e.g. when bioinformatics pipelines analysing the output of different platforms are compared). Although the absence of bias may be hard to demonstrate *a posteriori*, efforts should be made to ensure that the datasets derive from strategies that are as similar as possible, for example by containing reads generated from the same input samples.

The platforms for which benchmark datasets are produced should be selected based on pragmatic considerations. Ideally, equivalent resources should be available for all technologies; however, it may be necessary to prioritize the most common platforms if resource limitations are an issue. Recent surveys have shown a clear preference for the Illumina platform in this context[27,28]. The same trend can be observed when counting the number of published articles in a scientific literature database (Figure 2). That being said, sequencing technologies are constantly evolving, and benchmark datasets should be extended to new technologies with improved performance as these are increasingly adopted by testing laboratories.

The so-called "Third Generation Sequencing" technologies that sequence single DNA molecules and, importantly, produce long reads, have been shown to provide substantial benefits in the context of AMR detection. First, many resistance genes are located on plasmids, which are challenging to assemble using short-read sequencing technologies, because the short read lengths do not allow spanning of repetitive regions[29]. The presence of an AMR determinant on a plasmid is also important for its transfer and eventual spread, and thus their correct assembly using long-read technologies represent a substantial advantage[30–34]. In addition, the proper and timely treatment of a pathogen infection is critical for successful prevention and control of diseases in clinical settings as well as in the community. In line with this, the Nanopore sequencing technology has shown the promise of providing accurate antibiotic resistance gene identification within six hours of sample acquisition[35–37] We thus propose to include DNA Nanopore sequencing as an additional priority platform to develop benchmark resources.

The choice of formats for the different components of the datasets is also important. Each instrument produces a raw data output in a specific format (for example, the Illumina platforms generate raw data files in binary base call (BCL) format,

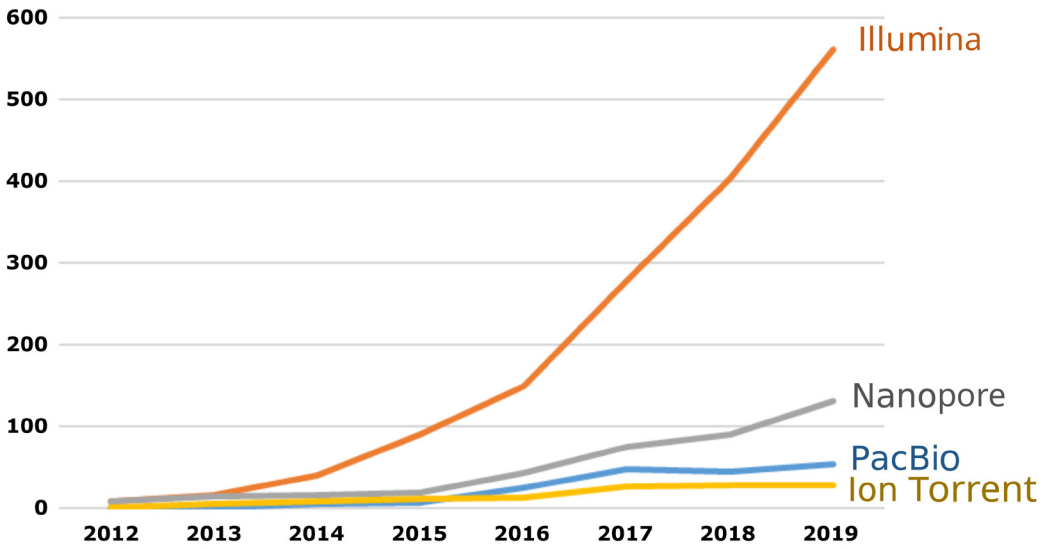


Figure 2. Number of articles published each year in the scientific literature mentioning the selected platform. Source: Scopus, using the search: ALL ("X" AND "antimicrobial resistance").

while the Nanopore platforms produce FAST5 (HDF5) files). However, the entry point of most bioinformatics pipelines in this context is the description of the sequence of the produced reads, with an indication of the quality (or confidence) score for each of the base positions. The FASTQ format is a standard format in this context[38], which should be used in the benchmark resources; many tools exist to convert the raw data output files into this format in case of different platform outputs (see, for example,[39,40]) although, it should be noted, different tools may produce different results and this step should be carefully planned.

Other standard formats exist to describe intermediate states of processing, for example for the description of assembled contigs or variant calling[41]. However, using these formats would make an *a priori* assumption about the strategy of the bioinformatics pipeline that may not be universal; indeed, not all reported solutions involve assembling reads, or mapping them to reference genomes or databases (see, for example,[42,43]).

2.2. Datasets origin

Three main sources of data for creating a benchmark dataset were identified. The first is to simulate the output (reads) *in silico* using an input sequence of a resistant pathogen and a specialised software. The second is to use the archived output of previously performed experiments that are available in different repositories. The third is to perform NGS experiments on biological samples.

Although the disadvantage of simulating *in silico* data is obvious (it is not 'real'), there are some substantial advantages: it is a lot cheaper than performing sequencing runs, a lot faster, and can be applied to any genome previously sequenced. Thus, many more potential scenarios can be tested, for which the ground truth is well-established (i.e., the annotation of the genome reference that is used: different species, different classes of AMR, different localization of AMR), which usually cannot be done by actually sequencing them. Finally, it is also potentially 'safer' to do this for pathogenic bacteria for which high biosafety levels would be required to sequence in a laboratory. However, a major drawback is that simulating variation the way nature evolves is very challenging – genetic variation happens in places in the genome where it is hardest to find.

Many methods and programs have been developed to simulate genetic data. Their use in this context is, in itself, an exercise of Open Science and mechanisms should be used to guarantee quality and reproducibility (see 44) In 2013, Peng *et al.*[45] developed the catalogue "Genetic Simulation Resources" (GSR, available at https://popmodels.cancercontrol.cancer.gov/gsr/) to help researchers compare and choose the appropriate simulation tools for their studies. However, after reviewing the software listed in the GSR catalogue, the authors realised that the quality and usefulness of published simulation tools varied greatly due to inaccessible source code, lack of or incomplete documentation, difficulties in installation and execution, lack of support from authors and lack of program maintenance[45]. For these reasons, a defined checklist of features that may benefit end users was defined[46]; the "GSR Certification Program" was developed and recently implemented into the GSR in order to assess simulation tools based on these criteria[47]. Established criteria are grouped to attribute four "certificates" (https://popmodels.cancercontrol.cancer.gov/gsr/certification/):

- Accessibility: it ensures that the simulator is openly available to all interested users and is easy to install and use.
- Documentation: it ensures that the simulator is well documented so that users can quickly determine if the simulator provides needed features and can learn how to use it.
- Application: it ensures that the software simulator is peer-reviewed, is reasonably user-friendly to be useful to peer researchers, and has been used by researchers in the scientific community.
- Support: it ensures that the authors of the simulator are actively maintaining the simulator, addressing users' questions, bug reports and feature requests.

As of December 2019, the GSR catalogue lists 148 simulators and many of them have been assessed for their compliance with the requirements in order to be certified. Obviously, not all of them are for simulation of NGS reads. In 2016 Escalona *et al.*[48] identified and compared 23 computational tools for the simulation of NGS data and established a decision tree for the informed selection of an appropriate NGS simulation tool for the specific question at hand.

By browsing the GSR catalogue, 20 out of 23 tools assessed by Escalona *et al.* (45) have been recorded, including only one with the four "GSR certificates" (Table 1), i.e. the ART tool[49]. Other tools not assessed by Escalona are also present in the GSR catalogue with certificates, like NEAT[50] and VISOR[51].

For choice of the simulation methods and programs for NGS reads, the decision tree proposed by Escalona *et al.* is robust. However, it should be complemented by "certification" steps and, in this respect, we encourage the use of the "certification" criteria established by the GSR Certification Program, to tackle the challenge of following agreed principles for rigorous, reproducible, transparent, and systematic benchmarking of omics tools, in line with those proposed by Mangul *et al.*[13].

Using pre-existing experiments, from private or public repositories, ensures that the components of the dataset are representative of a real-life experiment, including the complete panel of real-life variabilities that are difficult to simulate. The main issues then are: a) there is a need to demonstrate that the experiment met the necessary quality criteria (see section 2.3); b) the "correct" value (i.e. the 'ground truth') for the experiment needs to be determined. This can be already described in the metadata associated with the record and/or determined (verified) *a posteriori* – although this requires strict annotation of the experiment; c) it will not be possible (besides rare exceptions) to build datasets for the different platforms using the same initial samples.

Table 1. Analysis of the GSR certifications of the computational tools for the simulation of next-generation sequencing described in 48. See text for details.

Tool	In GSR?	GSR certificate?			
		Accessibility	Documentation	Application	Support
454sim	Yes	not yet evaluated			
ART	Yes	Yes	Yes	Yes	Yes
ArtificialFastqGenerator	Yes	not yet evaluated			
BEAR	No	-			
CuReSim	Yes	No	Yes	No	No
DWGSIM	Yes	Yes	Yes	No	Yes
EAGLE	Yes	not yet evaluated			
FASTQSim	Yes	not yet evaluated			
Flowsim	No	-			
GemSIM	Yes	Yes	Yes	No	No
Grinder	Yes	not yet evaluated			
Mason	Yes	Yes	Yes	No	Yes
MetaSim	Yes	No	Yes	Yes	No
NeSSM	No	-			
pbsim	Yes	not yet evaluated			
pIRS	Yes	Yes	Yes	No	Yes
ReadSim	Yes	not yet evaluated			
simhtsd	Yes	not yet evaluated			
simNGS	Yes	not yet evaluated			
SimSeq	Yes	not yet evaluated			
SInC	Yes	Yes	No	No	Yes
wgsim	Yes	not yet evaluated			
XS	Yes	not yet evaluated			

Generating experiments specifically for the sake of a benchmark dataset has almost the same advantages and disadvantages as using pre-existing data. Additional advantages include a better capacity to determine the "ground truth" of each sample by ensuring access to the original pathogen, as well as the possibility to generate datasets for the different platforms while using the same samples, if the same pathogen/purified DNA is processed through the different protocols and instruments. This also allows better control of the quality aspects of the procedure performed, e.g. through the use of accredited laboratories who have therefore demonstrated by audits that they possess the necessary knowhow and expertise to create high-quality data. However, an additional disadvantage is that this process requires a substantial investment of time and resources (although this investment may be deemed worthwhile given the importance of the topic, and could benefit from the involvement of the instrument vendors).

Because each approach has advantages and disadvantages, the choice must be carefully considered, according to the purpose of the dataset, which will be discussed in section 3.

2.3. Quality metrics

The quality of the original sample and the wet laboratory procedures (e.g. DNA extraction, library preparation and sequencing) have a strong impact on the quality of the reads fed into the bioinformatics pipelines. Contamination, low amounts of reads passing the machine QC, higher error rates than normal, etc. can influence the output of bioinformatics pipelines. Usually, the pipelines are designed to be resilient to some extent to these variations.

Although understanding this resilience is important, we propose, as shown in Figure 1, to separate these considerations from resources meant for quality control and performance evaluation (questions 1, 2a and 2b) for two reasons: first, many of these factors are variable, heterogeneous, technology-specific, and can be implemented at different stages of the bioinformatics pipeline; attempting to incorporate them all in the same resource would be impractical and too costly. Second, pipelines implemented for regulatory or clinical decision-making will be incorporated into a larger quality assurance framework that will ensure the quality of the input until that step[2]. Although examples exist of end-to-end NGS workflow validation (like in the case of WGS) where bioinformatics is one of the components[52], our approach emphasises a process where each step is validated separately (see 53).

It is then crucial to closely follow the proposed quality control schemes, either published or in development, in particular for the upstream steps (DNA isolation, library extraction, sequencing, etc.), for example ISO/TC 34/SC 9/WG 25. From these, both the metrics and the thresholds that can be applied at the level of the reads should be identified (some of which may vary according to the sequencing methodology), such as percent of bases with quality scores over Q30, percent alignment and error rates of the positive control (if present), the number of reads after trimming, quality of the draft assembly (e.g. N50, number of contigs), etc. Tools exist that can provide a panel of quality metrics from FASTQ files, such as FASTQC (RRID: SCR_014583)[54]. It is important to include the quality metrics as metadata in the dataset samples.

For studies evaluating resilience (question 3), a variety of datasets are needed for the "low quality dimensions" to be tested. For example, datasets incorporating high error rates, contaminating reads, reads with lower Q-scores could be used to assess resilience of pipelines to quality issues that may or may not be detectable with standard QC pipelines. For this reason, the establishment of standard datasets for this type of benchmarking is a complex exercise and answering question 3 of Figure 1 should be attempted on a case-by-case basis, and may be better suited to individual studies. One way to harmonise the approach would be to use the datasets produced for questions 1 and 2 as a starting point, as there are tools that can add some extent of "noise" to existing good quality datasets[49].

2.4. Choice of bacteria/resistance to include

In the context of challenging/evaluating a bioinformatics pipeline for the detection of AMR genetic determinants, a very pragmatic approach could be the generation of random DNA sequences, to which particular sequences of interest are added (i.e. fragments of AMR genes). However, the genomic background of the bacteria (i.e. the "non-AMR related" sequences) might have a profound influence on the performance of the pipelines. For example, pipelines that include a contig assembly step will be affected by the frequency and level of repetitive sequences in the background genome, as well as its GC content[55,56]. Some species also have genes that are similar at the sequence level to known AMR determinants that efficient pipelines must be able to distinguish.

In conclusion, the bacterial species included in the benchmark datasets, and the AMR genes they contain, need thus to be carefully selected, with the appropriate justifications. These are specific to the purpose of the dataset (Figure 1) and will be discussed in section 3.1–section 3.3 below.

2.5. Genomic and phenotypic endpoints

A pipeline processing sequencing information for AMR can produce two closely linked but conceptually different outputs: a) they can detect the genetic determinants of AMR (genomic endpoint), and in addition b) some can predict the AMR/susceptibility of the bacteria in the original sample (phenotypic endpoint).

In a clinical context, the phenotypic endpoint is the most relevant, as it provides the information that is most useful for the end users. Studies that evaluated AMR genotype to phenotype relationships have indicated that despite generally high correspondence, this can vary greatly between pathogens / case studies, and even for different antimicrobial agents within the same species[57,58]. There are different reasons for discrepancies between phenotype and genotype, including the extent of the expression of the resistance determinants for the resistance to be conferred, and also relatively complex molecular pathways that can influence the eventual phenotype. In some cases, genes can also confer reduced susceptibility (i.e. increasing the concentration of an antimicrobial necessary for treatment) rather than resistance *per se*. A genotypic endpoint may also be problematic due to the definition of "antibiotic resistance" in different settings[59], which can complicate the interpretation of results.

In practice, however, focusing on a genomic endpoint has many advantages:

- The end-point (determinant; gene or SNP) is better defined: presence or absence.
- The gene copy number can be calculated, this is important even if obtaining gene copy numbers with short read data remains pretty difficult.
- It provides high resolution information that is useful when many genetic determinants confer resistance to the same antimicrobials.
- It offers additional information to contribute to an evaluation of the history of the spread of AMR[60].
- It does not rely on breakpoints such as MICs, which may vary between human and animal bacterial isolates, or may not be available for some animals (or pathogens), or because it may be updated based on phenotypic scientific observations[61,62].
- Even in the cases of AMR determinants not being expressed (thus not leading to a resistance phenotype), this may be important to characterise/record for epidemiological purposes.

2.6. Benchmark datasets metadata

Besides the set of reads themselves, additional information needs to be associated with each sample in the dataset, not for

the benchmarking *per se* but its use for next benchmarking exercises.

Obviously, each sample needs to include a "true" value, i.e. the 'ground truth' to be used for comparison when evaluating the performance of the pipeline. For a genotypic endpoint, this would take the form of the name (with a reference to a database) of the AMR determinants present. If real-life samples are used, the phenotypic information should be included, on top of the genotypic endpoint.

Public resources and repositories are available to host both the data and the metadata, and should be used as appropriate for the sake of transparency, obsolescence and traceability of the datasets of the benchmark resource. In practice, this means:

- The NGS reads data should be hosted in one of the International Nucleotide Sequence Database Collaboration (INSDC, RRID:SCR_011967) sequence reads archives[63], compiling the metadata information as appropriate.
- For simulated reads, this information should include the simulation tool used (source, version, parameters).
- For simulated reads, the "input" sequence(s) should be a closed genome, and any additional genes, that should be available in INSDC sequence archives[64], and the record ID(s) included in the reads metadata information. Optimally, the closed genomes should be linked to a "real" sample in the INSDC BioSample database.
- For real experiments, the originally sequenced sample should be present/submitted in the INSDC BioSample database[65], with all the appropriate metadata information (including identified resistances and the MIC(s) determined according to the standard culture-based evaluation methods).

3. Design of the scope-specific benchmark resources

Two main use cases for benchmarking resources have been discussed, i.e. single isolates and mixed samples, here summarised in sections 3.1-3.2 and 3.3, respectively.

3.1. Is the bioinformatics pipeline performing according to (agreed) minimum standards?

The scope of this benchmark resource is to address the questions of validation, implementation and quality control over time (i.e. following any change in the pipeline or the environment on which it is executed). The dataset required for this should be compiled based on an agreed "minimum" standard, i.e. thresholds for the acceptance values of certain performance metrics for the bioinformatics pipeline in the context of the detection of AMR determinants, no matter the exact final use of the information produced.

This evaluation of performance should be based on challenging the pipeline with input representing a carefully selected set of resistance determinants and bacterial hosts. These sets of NGS reads should be fully characterised regarding their genetic content and serve as (*in silico*) reference materials for the validation and quality control of the bioinformatics component of the methods (see, for other host models, 66,67).

To maintain this necessary control on the genetic content of the reads, the dataset could be composed exclusively of simulated experiments. Synthetic artificial reads can be generated on a large scale in a harmonised manner and most importantly allow full control on the content. Alternatively, real sequencing datasets could be used, which would be extremely relevant for cases where the presence/absence of some AMR determinants has been established using consolidated classical molecular-biology-based methods and/or first generation-sequencing (e.g. PCR and/or Sanger sequencing). Both approaches enable the generation of datasets for which the 'ground truth' is well-established.

For the choice of resistances and bacterial species to be included, it is proposed to select them based on three sources, based on their current public health relevance and regulatory frameworks:

- The WHO's list of antibiotic-resistant "priority pathogens"[68].
- The AMR reporting protocol for the European Antimicrobial Resistance Surveillance Network (EARS-Net)[69].
- The Commission Implementing Decision of 12 November 2013 on the monitoring and reporting of AMR in zoonotic and commensal bacteria[70].

Table 2 shows the combination of these three lists, in terms of both the bacterial species and the antibiotics mentioned.

It is important to highlight that the combination of these three lists still leaves important regulatory gaps, and should be complemented by the World Organisation for Animal Health (OIE, RRID:SCR_012759) list of antimicrobial agents of veterinary importance[71] or others[72]. However, the lists do not mention specific species associated to each antibiotic, and these should be selected by the appropriate experts for the context of this benchmark resource. In practice, the generation of reference sequencing datasets should focus on:

1. Pathogens in Table 2 for which high-quality and complete reference genome sequences are available. See, for example, the FDA-ARGOS database[11] and the NCBI RefSeq Genomes database (RRID:SCR_003496).
2. Known genetic determinants for the resistance against the antibiotics in Table 2, using available resources[7,8,73]. If more than one determinant is associated with a resistance phenotype, one possibility is to collect them all; expert knowledge and empirical evidence on the relative contribution of different genes to the phenotypes, from published large-scale studies (e.g. 74) can also be used to objectively reduce the list of determinants to include for a given antibiotic.

Table 2. Summary of the bacterial species and antibiotics resistances mentioned in the three lists discussed in the text. a: WHO's list of antibiotic-resistant "priority pathogens". b: EARS-Net reporting protocol for 2018. c: Commission Implementing Decision 2013/652/EU.

	Acinetobacter baumannii	*Campylobacter coli*	*Campylobacter jejuni*	*Enterococcus faecalis*	*Enterococcus faecium*	*Escherichia coli*	*Haemophilus influenzae*	*Helicobacter pylori*	*Klebsiella pneumoniae*	*Neisseria gonorrhoeae*	*Pseudomonas aeruginosa*	*Salmonella spp.*	*Shigella spp.*	*Staphylococcus aureus*	*Streptococcus pneumoniae*
Amikacin	b					b			b		b				
Amoxicillin				b	b	b			b						
Ampicillin				b,c	b,c	b,c	a					c			
Azithromycin						c						c			b
Cefepime						b			b		b				
Cefotaxime						b,c			b			c			b
Cefoxitin														b	
Ceftazidime						b,c			b		b	c			
Ceftriaxone						b			b						b
Cephalosporin										a					
Chloramphenicol				c	c	c						c			
Ciprofloxacin	b	a,c	a,c	c	c	b,c			b	a	b	a,c	a	b	
Clarithromycin								a							b
Cloxacillin														b	
Colistin	b					b,c			b		b	c			
Daptomycin				c	c									c	
Dicloxacillin														b	
Ertapenem	a					a,b			a,b		a	a	a		
Erythromycin		c	c	c	c										b
Flucloxacillin														b	
Gentamicin	b	c	c	b,c	b,c	b,c			b		b	c			
Imipenem	a,b					a,b			a,b		a,b	a	a		
Levofloxacin	b	a	a			b			b	a	b	a	a	b	b
Linezolid				b,c	b,c									b	
Meropenem	a,b					a,b,c			a,b		a,b	a,c	a		
Methicillin														a,b	
Moxifloxacin						b			b						b
Nalidixic acid		c	c			c						c			

	Acinetobacter baumannii	*Campylobacter coli*	*Campylobacter jejuni*	*Enterococcus faecalis*	*Enterococcus faecium*	*Escherichia coli*	*Haemophilus influenzae*	*Helicobacter pylori*	*Klebsiella pneumoniae*	*Neisseria gonorrhoeae*	*Pseudomonas aeruginosa*	*Salmonella spp.*	*Shigella spp.*	*Staphylococcus aureus*	*Streptococcus pneumoniae*
Netilmicin	b					b			b		b				
Norfloxacin						b			b					b	b
Ofloxacin		a	a			b			b	a		a	a	b	
Oxacillin														b	b
Penicillin															a,b
Piperacillin						b			b		b				
Polymyxin B	b					b			b		b				
Quinupristin/ Dalfopristin				c	c										
Rifampin														b	
Streptomycin		c	c												
Sulfamethoxazole						c						c			
Teicoplanin				b,c	b,c										
Tetracycline		c	c	c	c	c						c			
Tigecycline				c	c	b,c			b			c			
Tobramycin	b					b			b		b				
Trimethoprim						c						c			
Vancomycin				b,c	a,b,c									a,b	

3. Combinations of (1) and (2) present in at least one of the chosen lists (see cells in Table 2) (see section 2.3).

The availability of reference datasets for which the ground truth is known, i.e. the composition of AMR determinants within the sample (whether SNPs, genes, or more complex features) is well-established, allows to compare the output of bioinformatics pipelines to evaluate their performance. The endpoint considered for this benchmark is thus genotypic (see section 2.5), i.e. the proper detection and identification of genomic AMR determinants is evaluated. The reference datasets for each pathogen should be carefully chosen and characterised to ensure all present AMR determinants are carefully recorded.

Several classical performance metrics used traditionally for method performance evaluation can then be used to describe performance of the pipeline, such as sensitivity, specificity, accuracy, precision, repeatability, and reproducibility[2]. Because of the selection of this subset of bacteria/resistances and their immediate clinical and regulatory importance, an important performance metric to be evaluated is accuracy, i.e. the likelihood that results are correct. Definitions for accuracy and other performance metrics should be carefully considered and tailored to the genomics context[53] (for example, "reproducibility", could be evaluated as the result of running the same bioinformatics pipeline, with the same datasets, implemented in different systems).

For all performance metrics, the minimum acceptable values should be subsequently determined once the outputs of real benchmarking exercises considering all the aspects described in this article are available. This will allow to enforce thresholds

that pipelines should obtain. Such an approach focusing on performance-based evaluation offers a robust framework for pipeline validation that is more flexible than focusing all efforts on single pipelines or sequencing technologies. Scientists should be able to choose the pipeline best suited to their own needs, whether commercial, open-source, web-based, etc.[75], as long as it conforms to minimum agreed upon performance standards, evaluated by analyzing the community AMR benchmark datasets that are created. This also offers flexibility with respect to the currently quickly evolving sequencing technologies for which it is not always clear yet which algorithmic approaches will become community standards, allowing different data analysis methodologies if minimum performance is demonstrated.

When generated, the benchmark should be deployed on a dedicated (and sustainably maintained) platform that includes all the links to the data (see section 2.6) and a description of all the steps/decisions that were taken to generate it. It is also important to implement, from the start, a clear version control system for the benchmark resource, in order to properly document the changes over time, and the exact versions used at the different times that the resource is used. In addition to the unique accession numbers of the individual samples in their respective repositories, the dataset as a whole should have a unique identifier (e.g. a DOI) that changes when any modification is made. The versioning should also allow access to and use of any previous versions of the resource, even after being updated.

This minimal dataset contains, by definition, a limited number of species and may lack pathogens of clinical importance (for example, *Mycobacterium tuberculosis*, for which WGS-based approaches have shown particular advantages, see 76,77). A full validation exercise for a specific pipeline, applied to a specific context, will need additional samples that complement the resource described in this section with the appropriate species/resistances. These datasets may, for example, be taken from the resources described in the following sections, that focus on evaluating the actual performance of methods in broader contexts by gathering the many datasets necessary to do so.

3.2. What is the quality of the information produced by the bioinformatics pipeline (prediction of resistance)?

The scope of this benchmark resource is to identify gold standards for bioinformatics pipelines, in this case linked to the specific use of predicting resistance/susceptibility of a pathogen.

There is a step between identifying the determinants of AMR and predicting resistance, which is not always straightforward as factors such as expression of the AMR gene may affect the prediction[57,74]. For this reason, and because it is conceptually closer to the information that is acted upon, the endpoint for this benchmark should be phenotypic. In addition, the dataset should be composed of real NGS experiments, since artefacts and variations are more complex in real sequencing reads than in simulated reads, a factor crucial to consider for this scope that focuses on accuracy.

To minimise the need of extensive resources to produce these "real" datasets, we propose to focus on re-using experiments previously performed under standardised conditions. A great source of data are the published ring trials; these have the additional advantage of providing an accurate characterisation of the sequenced samples, since the same original samples are sequenced many times by different laboratories. If needed, the data generated by single-site studies can also be evaluated, although in this case the issue of the correct characterisation of the samples (their "true" resistance patterns) should be addressed. One possibility is to use studies performed in a hospital setting, linked to clinical outcome (for example,[78]), or where sufficient information is available to evaluate the way the susceptibility testing was performed.

In practice, this would mean:

1. Performing an extensive review of the published literature to identify studies, ring trials, and proficiency testing that meet the criteria (focused on the detection of AMR using NGS, starting from a "real" sample). Table 3 provides a non-exhaustive list of recent references to be used as a starting point, together with reports from national and international reference laboratories dealing with AMR (see for example, the external quality assessment reports from the EU Reference Laboratory – Antimicrobial Resistance).
2. Assessing whether the raw sequencing output for the projects meet the FAIR principles (Findability, Accessibility, Interoperability, and Reusability)[79], and are retrievable from publicly available repositories – even if they are access controlled. If not fully open, the corresponding authors should be contacted and asked whether the data could be obtained and deposited in long-term archives (e.g. Zenodo (RRID: SCR_004129), EuDat and/or the European Nucleotide Archive (ENA, RRID:SCR_006515) depending on the deposited data).

These datasets would then be used to test and compare the different bioinformatics pipelines in order to calculate the accuracy of their phenotypic predictions. Although not exhaustive, these datasets should cover the most relevant "real-life" cases, as they warranted their inclusion into a ring trial, with the associated resources committed to produce the data. The final size and composition (species, resistances) of the dataset would depend on what is provided by the available projects; *ad hoc* ring trials could be organised to cover eventual important gaps in species and/or resistance.

Although the chosen endpoint is mostly phenotypic, the purpose is to evaluate bioinformatics pipelines that process information at the sequence level, so it was agreed that there was little added value of inserting resistant samples (based on a characterised or inferred phenotype) for which the resistance mechanism is still unknown. In any case, it is improbable that these cases would have been included in ring trials projects.

Table 3. Non-exhaustive list of sample studies to be analysed for the availability of FAIR raw reads data, to include in the benchmark resource.

Pathogens	Study	Year	Study type	Ref
Clostridium (Clostridioides) difficile	Berger *et al.*	2019	ring trial	80
Neisseria meningitidis	Bogaerts *et al.*	2019	validation study	53
Salmonella enterica	Mensah *et al.*	2019	single study	81
Enterobacteriales	Ruppé *et al.*	2019	single study	58
Escherichia coli	Stubberfield *et al.*	2019	single study	82
Staphylococcus aureus	Deplano *et al.*	2018	ring trial	83
Brucella melitensis	Johansen *et al.*	2018	ring trial	84
Salmonella enterica	Neuert *et al.*	2018	single study	85
Salmonella, Campylobacter	Pedersen *et al.*	2018	proficiency testing	86
Escherichia coli	Pietsch *et al.*	2018	single study	87
Enterococcus faecium, Enterococcus faecalis	Tyson *et al.*	2018	single study	88
Actinobacillus pleuropneumoniae	Bossé *et al.*	2017	single study	89
Klebsiella pneumoniae	Brhelova *et al.*	2017	single study	90
Salmonella enterica	Carroll *et al.*	2017	single study	91
Escherichia coli	Day and al.	2016	single study	92
Salmonella spp., Escherichia coli, Staphylococcus aureus	Hendriksen *et al.*	2016	proficiency testing	93
Salmonella	McDermott *et al.*	2016	single study	94
Staphylococcus aureus, Enterococcus faecium, Escherichia coli, Pseudomonas aeruginosa	Mellmann *et al.*	2016	single study	78
Staphylococcus aureus, Mycobacterium tuberculosis	Bradley *et al.*	2015	single study	42
Escherichia coli	Tyson *et al.*	2015	single study	95
Mycobacterium tuberculosis	Walker *et al.*	2015	single study	76
Campylobacter jejuni, Campylobacter coli	Zhao *et al.*	2015	single study	96
Pseudomonas aeruginosa	Koos *et al.*	2014	single study	97
Staphylococcus aureus	Gordon *et al.*	2013	single study	98
Escherichia coli, Klebsiella pneumoniae	Stoesser *et al.*	2013	single study	99
Staphylococcus aureus, Clostridium difficile	Eyre *et al.*	2012	single study	100
Salmonella typhimurium, Escherichia coli, Enterococcus faecalis, Enterococcus faecium	Zankari *et al.*	2012	single study	101
Salmonella	Cooper *et al.*	2020	single study	102

Although the performance metrics described in Section 3.1 apply and are relevant in this case, the main performance metric for this benchmark is the accuracy. Because of the difficulty of predicting the link between the presence of AMR determinants and their impact on the pathogen susceptibility to the antimicrobial agents, the target accuracy is expected to be lower than for a genotypic endpoint. Both false positives and false negatives can be an issue when the information is used for clinical intervention, so a sufficient amount of "borderline" cases should be included, and both sensitivity and specificity evaluated. It is also possible to consider attaching different relative costs for false positives and false negatives when evaluating the accuracy metrics.

Once selected and combined, the data should be separated by NGS platform, and by species and antibiotic. Because this benchmarking aims at evaluating and comparing performance of methods, which are continuously developed and optimised, against a large and constantly expanding dataset, it is crucial to define an environment where the AMR community can establish a continuous benchmarking effort. Within this platform, pipelines would be compared simultaneously based on up-to-date datasets, under the same conditions, and over time. Constantly updating and adding to the reference datasets is important both to keep up with the evolution of the knowledge/reality in the field, and to avoid that pipelines are developed that are optimised to specific datasets only.

One option is OpenEBench (Open ELIXIR Benchmarking and Technical Monitoring platform), which is developed under the ELIXIR-EXCELERATE umbrella in order to provide such a platform[18]. In this framework, in addition to compiling the data resources to be included (as described above), and whatever the platform chosen, there will be the need for efforts to:

- Establish guidelines for input and output formats (and, in the case of the phenotypic endpoint, an agreed ontology for the conclusions).
- Encouraging a "FAIR" implementation of the pipelines themselves, to increase the number of pipelines accessible for the benchmarking platform, and for interested end users to retrieve and implement in house.

Provisions should be included to allow the possibility to evaluate, in this context, pipelines that cannot be made "FAIR" based on intellectual property rights, institutional policies or available resources.

A final step will be to communicate these efforts within the scientific community and the potential end users, as well as to demonstrate the added value of this "live" benchmark resource to ensure that future studies (in particular, their pipelines and the datasets they generate) are efficiently integrated in the platform.

3.3. What is the quality of the information produced by the bioinformatics pipeline (mixed samples)?

Many gaps exist in the scientific understanding of antibiotic resistance development and transmission, making it difficult to properly advise policy makers on how to manage this risk. There is strong evidence that a multitude of resistance genes in the environment have not yet made it into pathogens[103,104]; understanding the relative importance of different transmission and exposure routes for bacteria is thus crucial[59,105–107].

Establishing a baseline for resistance determinants in the environment, and linking this to a surveillance scheme, requires a good understanding of the relative performance of methods that are and have been developed to characterise the resistome in a complex sample. There would be, also for this use case, a great value in the establishment of a community-driven "live" benchmarking using a platform such as OpenEBench, and many of the concepts that were discussed in section 3.2 apply here as well, with the following differences:

- As, by definition, the resistome refers to the genetic determinants (and not directly the associated phenotypes)[108,109], the endpoint for this benchmark should be genotypic.
- Culture-dependent methods established for clinical samples cannot always be readily applied to environmental samples[110], so establishing "true" values for real samples, to compare the output of the evaluated pipelines, will be difficult, so the benchmark should be performed, at this stage, with simulated reads.

The resistome is usually derived from a sample containing a complex microbial community (see 111–113 for recent examples). For this reason, the approaches[114] and tools[115] from the ongoing Critical Assessment of Metagenome Interpretation (CAMI) could be considered when organising the community around this challenge.

In practice, this means an effort to engage and coordinate the community of bioinformatics pipelines designed to predict the resistome of a sample in order to:

1. Design the scope of the challenge, including the relevant metrics for performance evaluation. For this, "accuracy", the main metrics for the previous two benchmarks, may not be the most appropriate, and the focus should be placed, e.g., on "recall" and "precision".
2. Describe the microbial communities (i.e. microbial abundance profiles and their typical AMR gene profiles) most relevant for the determination/monitoring of the resistome, in order to generate congruent datasets that accurately represent real-life samples. Of particular interest, for which validation will eventually be a prerequisite, are blood, serum, saliva etc., i.e. the types of samples clinical microbiology laboratories and national reference centres/laboratories typically process.
3. Identify both the microbial genomes and the resistance determinants (as single genetic determinants or plasmids) necessary to generate the profiles identified in (2). As stated in section 3.1, the genomes should be well analysed to ensure no lack of, or an adequate characterisation of, AMR determinants. This is crucial in order to establish a resistome "ground truth" for the generated datasets.
4. Combine these sequences, as appropriate, to generate the benchmark datasets, using appropriate tools (such as CAMISIM, developed as part of the CAMI challenge[115]).

The community should decide whether (or at what stage) the use of real data can also be considered in the challenge. As for purified bacteria (see Table 3), many studies have been published as potential sources of raw data. These studies can also be used as a source of information to define the relevant profiles (point 2 above). Recent studies include resistome

determination in samples from drinking water[116,117], wastewater plants[118,119], hospital wastewater[120,121], human gut[111,122], sewage[123], to name a few. In the benchmarking platform, the datasets (and the calculated pipeline performances) should be separated by the type of source they originate from or simulate. Another important point for this scope is the detection of minority populations and the correct normalisation of the samples to be analysed[124].

Conclusions

The scientific community quickly adopted the new NGS technologies to develop methods that can efficiently detect, identify and characterise genetic determinants of AMR. In parallel with these research uses, NGS technologies can have immediate impacts on how AMR is diagnosed, detected and reported worldwide, complementing etiologic agent diagnosis, clinical decision making, risk assessment and established monitoring frameworks[3,125–127].

For this application and in general, there are great challenges in the implementation of NGS-based methods for public decision-making. Capacity building and its cost is of course a factor, but recent surveys show that capacity development is ongoing in many countries[28]. A greater concern is the interpretation of the produced genomic data into meaningful information that can be acted upon or used for regulatory monitoring, in great part because of the important bioinformatics component of these methods.

The difficulties posed by this reliance on bioinformatics processes are many, and include:

- The specific expertise needed for their implementation and maintenance, which is still limited compared to the needs of routine testing environments.
- The lack of harmonisation in their design, as the same sequencer output can be processed to produce the same target information by pipelines that either follow the same general strategy, with different tools for the individual steps, or completely different strategies entirely (see 128).
- The constant, rapid evolution of the fields of informatics and bioinformatics, which makes uneasy (or even unwise) to "freeze" a harmonised, validated, implemented pipeline with the same components in the same environment over long periods of time.
- For AMR, as for other fields, the pipelines (and their performance metrics) are built based on *a priori* scientific knowledge, in this case the genetics of resistance, which is constantly progressing.

In this document, we propose a way through these difficulties with a transparent, performance-based evaluation approach to assess and demonstrate that pipelines are fit-for-purpose and to ensure quality control. The discussions, initiated in 2017[16], have involved experts in different fields: human health, animal health, food and environmental monitoring, and general bioinformatics.

The approach is two-fold: first, an agreed-upon, limited dataset to contribute to performance-based control of the pipeline implementation and their integration in quality systems. We propose selection criteria for this common dataset based on bacterial species and resistances relevant to current public health priorities (see section 3.1).

Second, a community-driven effort to establish a "live" benchmarking platform where both the datasets and the bioinformatics workflows are available to the community according to the FAIR principles. After an initial investment of resources to establish the rules and integrate the existing resources, a proper engagement of the community will be needed to ensure that both the datasets and the workflows will constantly be updated, with live monitoring of the resulting comparative performance parameters. For this, two main use cases were identified, each necessitating its own platform: the analysis of isolates (with a focus on the prediction of resistance, see section 3.2), and the analysis of mixed samples (with a focus on the interpretation of the resistome, see section 3.3).

To ensure acceptance of this approach by regulators and policy-makers, the conclusions and the roadmap proposed in this document should be complemented (and, if necessary, revised) with the continuous involvement of all relevant actors in the field, including (but not limited to) the scientific community (see 13 and 114 as excellent examples of defining principles for benchmarking, and a roadmap for software selection, respectively, to guide researchers), the collaborative organisation and platforms active in the field (e.g. the European Committee on Antimicrobial Susceptibility Testing (EUCAST), the Joint Programming Initiative on Antimicrobial Resistance (JPIAMR), the Global Microbial Identifier (GMI), the European Society of Clinical Microbiology and Infectious Diseases and its Study Groups (ESCMID)), regulatory agencies (e.g. the European Food Safety Authority (EFSA, RRID:SCR_000963), the European Centre for Disease Prevention and Control (ECDC)), European Union reference laboratories and their networks (e.g. the EURL AR and the EURLs for the different pathogens) and the existing bioinformatics infrastructures (e.g. the European Bioinformatics Institute (EMBL/EBI), ELIXIR).

Such an approach would be a way to facilitate the integration of NGS-based methods in the field of AMR, and may be a case study on how to approach the overlapping challenges in other potential fields of applications, including some at high level in policy agendas (food fraud, genetically modified organism detection, biothreats monitoring for biodefense purposes, etc.).

Disclaimer

The contents of this article are the views of the authors and do not necessarily represent an official position of the European Commission or the U.S. Food and Drug Administration.

Data availability

No data is associated with this article.

Acknowledgments

We would like to thank Valentina Rizzi (EFSA) and Alberto Orgiazzi (JRC) for their participation to the workshop discussions. The authors are also grateful to the following colleagues for their comments on the manuscript: Sigrid De Keersmaecker and Nancy Roosens (Sciensano) and Tewodros Debebe (Biomes). We are also grateful to Laura Oliva for her invaluable help during the workshop.

References

1. Doyle RM, O'Sullivan DM, Aller SD, *et al.*: **Discordant bioinformatic predictions of antimicrobial resistance from whole-genome sequencing data of bacterial isolates: an inter-laboratory study.** *Microb Genom.* 2020; **6**(2): e000335.
PubMed Abstract | Publisher Full Text | Free Full Text
2. Kozyreva VK, Truong CL, Greninger AL, *et al.*: **Validation and Implementation of Clinical Laboratory Improvements Act-Compliant Whole-Genome Sequencing in the Public Health Microbiology Laboratory.** *J Clin Microbiol.* 2017; **55**(8): 2502–2520.
PubMed Abstract | Publisher Full Text | Free Full Text
3. Ellington MJ, Ekelund O, Aarestrup FM, *et al.*: **The role of whole genome sequencing in antimicrobial susceptibility testing of bacteria: report from the EUCAST Subcommittee.** *Clin Microbiol Infect.* 2017; **23**(1): 2–22.
PubMed Abstract | Publisher Full Text
4. Rossen JWA, Friedrich AW, Moran-Gilad J, *et al.*: **Practical issues in implementing whole-genome-sequencing in routine diagnostic microbiology.** *Clin Microbiol Infect.* 2018; **24**(4): 355–360.
PubMed Abstract | Publisher Full Text
5. Collineau L, Boerlin P, Carson CA, *et al.*: **Integrating Whole-Genome Sequencing Data Into Quantitative Risk Assessment of Foodborne Antimicrobial Resistance: A Review of Opportunities and Challenges.** *Front Microbiol.* 2019; **10**: 1107.
PubMed Abstract | Publisher Full Text | Free Full Text
6. EFSA Panel on Biological Hazards (EFSA BIOHAZ Panel), Koutsoumanis K, Allende A, *et al.*: **Whole genome sequencing and metagenomics for outbreak investigation, source attribution and risk assessment of food-borne microorganisms.** *EFSA J.* 2019; **17**(12): e05898.
PubMed Abstract | Publisher Full Text | Free Full Text
7. Hendriksen RS, Bortolaia V, Tate H, *et al.*: **Using Genomics to Track Global Antimicrobial Resistance.** *Front Public Health.* 2019; **7**: 242.
PubMed Abstract | Publisher Full Text | Free Full Text
8. Boolchandani M, D'Souza AW, Dantas G: **Sequencing-based methods and resources to study antimicrobial resistance.** *Nat Rev Genet.* 2019; **20**(6): 356–370.
PubMed Abstract | Publisher Full Text | Free Full Text
9. Aytan-Aktug D, Clausen PTLC, Bortolaia V, *et al.*: **Prediction of Acquired Antimicrobial Resistance for Multiple Bacterial Species Using Neural Networks.** *mSystems.* 2020; **5**(1): e00774–19.
PubMed Abstract | Publisher Full Text | Free Full Text
10. Lambert D, Pightling A, Griffiths E, *et al.*: **Baseline Practices for the Application of Genomic Data Supporting Regulatory Food Safety.** *J AOAC Int.* 2017; **100**(3): 721–731.
PubMed Abstract | Publisher Full Text
11. Sichtig H, Minogue T, Yan Y, *et al.*: **FDA-ARGOS is a database with public quality-controlled reference genomes for diagnostic use and regulatory science.** *Nat Commun.* 2019; **10**(1): 3313.
PubMed Abstract | Publisher Full Text | Free Full Text
12. Hardwick SA, Deveson IW, Mercer TR: **Reference standards for next-generation sequencing.** *Nat Rev Genet.* 2017; **18**(8): 473–484.
PubMed Abstract | Publisher Full Text
13. Mangul S, Martin LS, Hill BL, *et al.*: **Systematic benchmarking of omics computational tools.** *Nat Commun.* 2019; **10**(1): 1393.
PubMed Abstract | Publisher Full Text | Free Full Text
14. Angers A, Petrillo M, Patak A, *et al.*: **The role and implementation of next-generation sequencing technologies in the coordinated action plan against antimicrobial resistance.** Publications Office of the European Union. 2017.
Publisher Full Text
15. Hernando-Amado S, Coque TM, Baquero F, *et al.*: **Defining and combating antibiotic resistance from One Health and Global Health perspectives.** *Nat Microbiol.* 2019; **4**(9): 1432–1442.
PubMed Abstract | Publisher Full Text
16. Angers-Loustau A, Petrillo M, Bengtsson-Palme J, *et al.*: **The challenges of designing a benchmark strategy for bioinformatics pipelines in the identification of antimicrobial resistance determinants using next generation sequencing technologies. [version 2; peer review: 2 approved].** *F1000Res.* 2018; **7**: ISCB Comm J–459.
PubMed Abstract | Publisher Full Text | Free Full Text
17. Weber LM, Saelens W, Cannoodt R, *et al.*: **Essential guidelines for computational method benchmarking.** *Genome Biol.* 2019; **20**(1): 125.
PubMed Abstract | Publisher Full Text | Free Full Text
18. Capella-Gutierrez S, de la Iglesia D, Haas J, *et al.*: **Lessons Learned: Recommendations for Establishing Critical Periodic Scientific Benchmarking.** *bioRxiv.* 2017.
Publisher Full Text
19. Belmann P, Dröge J, Bremges A, *et al.*: **Bioboxes: standardised containers for interchangeable bioinformatics software.** *Gigascience.* 2015; **4**(1): 47.
PubMed Abstract | Publisher Full Text | Free Full Text
20. Sangwan N, Xia F, Gilbert JA: **Recovering complete and draft population genomes from metagenome datasets.** *Microbiome.* 2016; **4**: 8.
PubMed Abstract | Publisher Full Text | Free Full Text
21. Del Fabbro C, Scalabrin S, Morgante M, *et al.*: **An Extensive Evaluation of Read Trimming Effects on Illumina NGS Data Analysis.** *PLoS One.* 2013; **8**(12): e85024.
PubMed Abstract | Publisher Full Text | Free Full Text
22. Fox EJ, Reid-Bayliss KS, Emond MJ, *et al.*: **Accuracy of Next Generation Sequencing Platforms.** *Next Gener Seq Appl.* 2014; **1**: 1000106.
PubMed Abstract | Free Full Text
23. Ip CLC, Loose M, Tyson JR, *et al.*: **MinION Analysis and Reference Consortium: Phase 1 data release and analysis. [version 1; peer review: 2 approved].** *F1000Res.* 2015; **4**: 1075.
PubMed Abstract | Publisher Full Text | Free Full Text
24. Ardui S, Ameur A, Vermeesch JR, *et al.*: **Single molecule real-time (SMRT) sequencing comes of age: applications and utilities for medical diagnostics.** *Nucleic Acids Res.* 2018; **46**(5): 2159–2168.
PubMed Abstract | Publisher Full Text | Free Full Text
25. Giordano F, Aigrain L, Quail MA, *et al.*: ***De novo* yeast genome assemblies from MinION, PacBio and MiSeq platforms.** *Sci Rep.* 2017; **7**(1): 3935.
PubMed Abstract | Publisher Full Text | Free Full Text
26. Kaas RS, Leekitcharoenphon P, Aarestrup FM, *et al.*: **Solving the Problem of Comparing Whole Bacterial Genomes across Different Sequencing Platforms.** *PLoS One.* 2014; **9**(8): e104984.
PubMed Abstract | Publisher Full Text | Free Full Text
27. European Food Safety Authority (EFSA), Fierro RG, Thomas-Lopez D, *et al.*: **Outcome of EC/EFSA questionnaire (2016) on use of Whole Genome Sequencing (WGS) for food- and waterborne pathogens isolated from animals, food, feed and related environmental samples in EU/EFTA countries.** *EFSA Support Publ.* 2018; **15**(6): 1432E.
Publisher Full Text
28. Revez J, Espinosa L, Albiger B, *et al.*: **Survey on the use of Whole-Genome Sequencing for infectious diseases surveillance: rapid expansion of European national capacities, 2015–2016.** *Front Public Health.* 2017; **5**: 347.
PubMed Abstract | Publisher Full Text | Free Full Text
29. Arredondo-Alonso S, Willems RJ, van Schaik W, *et al.*: **On the (im)possibility of reconstructing plasmids from whole-genome short-read sequencing data.** *Microb Genom.* 2017; **3**(10): e000128.
PubMed Abstract | Publisher Full Text | Free Full Text
30. Lemon JK, Khil PP, Frank KM, *et al.*: **Rapid Nanopore Sequencing of Plasmids and Resistance Gene Detection in Clinical Isolates.** *J Clin Microbiol.* 2017; **55**(12): 3530–3543.
PubMed Abstract | Publisher Full Text | Free Full Text
31. Greig DR, Dallman TJ, Hopkins KL, *et al.*: **MinION nanopore sequencing identifies the position and structure of bacterial antibiotic resistance determinants in a multidrug-resistant strain of enteroaggregative *Escherichia coli*.** *Microb Genom.* 2018; **4**(10): e000213.
PubMed Abstract | Publisher Full Text | Free Full Text
32. Cao MD, Nguyen SH, Ganesamoorthy D, *et al.*: **Scaffolding and completing genome assemblies in real-time with nanopore sequencing.** *Nat Commun.* 2017; **8**(1): 14515.
PubMed Abstract | Publisher Full Text | Free Full Text
33. Ashton PM, Nair S, Dallman T, *et al.*: **MinION nanopore sequencing identifies the position and structure of a bacterial antibiotic resistance island.** *Nat Biotechnol.* 2015; **33**(3): 296–300.
PubMed Abstract | Publisher Full Text
34. Xie H, Yang C, Sun Y, *et al.*: **PacBio Long Reads Improve Metagenomic Assemblies, Gene Catalogs, and Genome Binning.** *Front Genet.* 2020; **11**: 516269.
PubMed Abstract | Publisher Full Text | Free Full Text

35. Schmidt K, Mwaigwisya S, Crossman LC, *et al.*: **Identification of bacterial pathogens and antimicrobial resistance directly from clinical urines by nanopore-based metagenomic sequencing.** *J Antimicrob Chemother.* 2017; **72**(1): 104–114.
PubMed Abstract | Publisher Full Text

36. Arredondo-Alonso S, Top J, McNally A, *et al.*: **Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen *Enterococcus faecium*.** *mBio.* 2020; **11**(1): e03284–19.
PubMed Abstract | Publisher Full Text | Free Full Text

37. Charalampous T, Kay GL, Richardson H, *et al.*: **Nanopore metagenomics enables rapid clinical diagnosis of bacterial lower respiratory infection.** *Nat Biotechnol.* 2019; **37**(7): 783–792.
PubMed Abstract | Publisher Full Text

38. Cock PJA, Fields CJ, Goto N, *et al.*: **The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants.** *Nucleic Acids Res.* 2010; **38**(6): 1767–1771.
PubMed Abstract | Publisher Full Text | Free Full Text

39. Holtgrewe M, Messerschmidt C, Nieminen M, *et al.*: **Digestiflow: from BCL to FASTQ with ease.** *Bioinformatics.* 2019; **7**: e27717v4.
Publisher Full Text

40. Loman NJ, Quinlan AR: **Poretools: a toolkit for analyzing nanopore sequence data.** *Bioinformatics.* 2014; **30**(23): 3399–3401.
PubMed Abstract | Publisher Full Text | Free Full Text

41. Zhang H: **Overview of Sequence Data Formats.** In: *Statistical Genomics.* E. Mathé and S. Davis, Eds. New York, NY: Springer New York, 2016; **1418**: 3–17.
Publisher Full Text

42. Bradley P, Gordon NC, Walker TM, *et al.*: **Rapid antibiotic-resistance predictions from genome sequence data for *Staphylococcus aureus* and *Mycobacterium tuberculosis*.** *Nat Commun.* 2015; **6**: 10063.
PubMed Abstract | Publisher Full Text | Free Full Text

43. Clausen PTLC, Aarestrup FM, Lund O: **Rapid and precise alignment of raw reads against redundant databases with KMA.** *BMC Bioinformatics.* 2018; **19**(1): 307.
PubMed Abstract | Publisher Full Text | Free Full Text

44. Jiménez RC, Kuzak M, Alhamdoosh M, *et al.*: **Four simple recommendations to encourage best practices in research software [version 1; peer review: 3 approved].** *F1000Res.* 2017; **6**: ELIXIR-876.
PubMed Abstract | Publisher Full Text | Free Full Text

45. Peng B, Chen HS, Mechanic LE, *et al.*: **Genetic Simulation Resources: a website for the registration and discovery of genetic data simulators.** *Bioinformatics.* 2013; **29**(8): 1101–1102.
PubMed Abstract | Publisher Full Text | Free Full Text

46. Chen HS, Hutter CM, Mechanic LE, *et al.*: **Genetic Simulation Tools for Post-Genome Wide Association Studies of Complex Diseases.** *Genet Epidemiol.* 2015; **39**(1): 11–19.
PubMed Abstract | Publisher Full Text | Free Full Text

47. Peng B, Leong MC, Chen HS, *et al.*: **Genetic Simulation Resources and the GSR Certification Program.** *Bioinformatics.* 2019; **35**(4): 709–710.
PubMed Abstract | Publisher Full Text | Free Full Text

48. Escalona M, Rocha S, Posada D: **A comparison of tools for the simulation of genomic next-generation sequencing data.** *Nat Rev Genet.* 2016; **17**(8): 459–469.
PubMed Abstract | Publisher Full Text | Free Full Text

49. Huang W, Li L, Myers JR, *et al.*: **ART: a next-generation sequencing read simulator.** *Bioinformatics.* 2012; **28**(4): 593–594.
PubMed Abstract | Publisher Full Text | Free Full Text

50. Stephens ZD, Hudson ME, Mainzer LS, *et al.*: **Simulating Next-Generation Sequencing Datasets from Empirical Mutation and Sequencing Models.** *PLoS One.* 2016; **11**(11): e0167047.
PubMed Abstract | Publisher Full Text | Free Full Text

51. Bolognini D, Sanders A, Korbel JO, *et al.*: **VISOR: a versatile haplotype-aware structural variant simulator for short- and long-read sequencing.** *Bioinformatics.* 2020; **36**(4): 1267–1269.
PubMed Abstract | Publisher Full Text

52. Portmann AC, Fournier C, Gimonet J, *et al.*: **A Validation Approach of an End-to-End Whole Genome Sequencing Workflow for Source Tracking of *Listeria monocytogenes* and *Salmonella enterica*.** *Front Microbiol.* 2018; **9**: 446.
PubMed Abstract | Publisher Full Text | Free Full Text

53. Bogaerts B, Winand R, Fu Q, *et al.*: **Validation of a Bioinformatics Workflow for Routine Analysis of Whole-Genome Sequencing Data and Related Challenges for Pathogen Typing in a European National Reference Center: *Neisseria meningitidis* as a Proof-of-Concept.** *Front Microbiol.* 2019; **10**: 362.
PubMed Abstract | Publisher Full Text | Free Full Text

54. Andrews S: **FastQC: a quality control tool for high throughput sequence data.** *Babraham Bioinformatics.* Babraham Institute, Cambridge, United Kingdom, 2010.
Reference Source

55. Chen YC, Liu T, Yu CH, *et al.*: **Effects of GC Bias in Next-Generation-Sequencing Data on *De Novo* Genome Assembly.** *PLoS One.* 2013; **8**(4): e62856.
PubMed Abstract | Publisher Full Text | Free Full Text

56. Phillippy AM, Schatz MC, Pop M: **Genome assembly forensics: finding the elusive mis-assembly.** *Genome Biol.* 2008; **9**(3): R55.
PubMed Abstract | Publisher Full Text | Free Full Text

57. Su M, Satola SW, Read TD: **Genome-Based Prediction of Bacterial Antibiotic Resistance.** *J Clin Microbiol.* 2019; **57**(3): e01405–18.
PubMed Abstract | Publisher Full Text | Free Full Text

58. Ruppé E, Cherkaoui A, Charretier Y, *et al.*: **From genotype to antibiotic susceptibility phenotype in the order Enterobacterales: a clinical perspective.** *Clin Microbiol Infect.* 2020; **26**(5): 643.e1–643.e7.
PubMed Abstract | Publisher Full Text

59. Martínez JL, Coque TM, Baquero F: **What is a resistance gene? Ranking risk in resistomes.** *Nat Rev Microbiol.* 2015; **13**(2): 116–123.
PubMed Abstract | Publisher Full Text

60. Baker S, Thomson N, Weill FX, *et al.*: **Genomic insights into the emergence and spread of antimicrobial-resistant bacterial pathogens.** *Science.* 2018; **360**(6390): 733–738.
PubMed Abstract | Publisher Full Text | Free Full Text

61. European Food Safety Authority, European Centre for Disease Prevention and Control: **The European Union Summary Report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food in 2011.** EU summary report on antimicrobial resistance in zoonotic and indicator bacteria from humans, animals and food 2011. *EFSA J.* 2013; **11**(5): 3196.
Publisher Full Text

62. Toutain PL, Bousquet-Mélou A, Damborg P, *et al.*: **En Route towards European Clinical Breakpoints for Veterinary Antimicrobial Susceptibility Testing: a position paper explaining the VetCAST Approach.** *Front Microbiol.* 2017; **8**: 2344.
PubMed Abstract | Publisher Full Text | Free Full Text

63. Leinonen R, Sugawara H, Shumway M, *et al.*: **The Sequence Read Archive.** *Nucleic Acids Res.* 2011; **39**(Database issue): D19–D21.
PubMed Abstract | Publisher Full Text | Free Full Text

64. Stoesser G, Baker W, van den Broek A, *et al.*: **The EMBL Nucleotide Sequence Database.** *Nucleic Acids Res.* 2002; **30**(1): 21–26.
PubMed Abstract | Publisher Full Text | Free Full Text

65. Gostev M, Faulconbridge A, Brandizi M, *et al.*: **The BioSample database (BioSD) at the european bioinformatics institute.** *Nucleic Acids Res.* 2012; **40**(Database issue): D64–D70.
PubMed Abstract | Publisher Full Text | Free Full Text

66. Zakaria O, Rezali MF: **Reference Materials as a Crucial Tools for Validation and Verification of the Analytical Process.** *Procedia Soc Behav Sci.* 2014; **121**: 204–213.
Publisher Full Text

67. Zook JM, Catoe D, McDaniel J, *et al.*: **Extensive sequencing of seven human genomes to characterize benchmark reference materials.** *Sci Data.* 2016; **3**: 160025.
PubMed Abstract | Publisher Full Text | Free Full Text

68. WHO: **WHO publishes list of bacteria for which new antibiotics are urgently needed**.
Reference Source

69. European Centre for disease prevention and control (ECDC): **Antimicrobial resistance (AMR) reporting protocol 2018**.
Reference Source

70. European Commission: **2013/652/EU: Commission Implementing Decision of 12 November 2013 on the monitoring and reporting of antimicrobial resistance in zoonotic and commensal bacteria**. 2013.
Reference Source

71. W. OIE: **OIE list of antimicrobial agents of veterinary importance**. 2015.
Reference Source

72. Scott HM, Acuff G, Bergeron G, *et al.*: **Critically important antibiotics: criteria and approaches for measuring and reducing their use in food animal agriculture.** *Ann N Y Acad Sci.* 2019; **1441**(1): 8–16.
PubMed Abstract | Publisher Full Text | Free Full Text

73. Xavier BB, Das AJ, Cochrane G, *et al.*: **Consolidating and Exploring Antibiotic Resistance Gene Data Resources.** *J Clin Microbiol.* 2016; **54**(4): 851–859.
PubMed Abstract | Publisher Full Text | Free Full Text

74. Sadouki Z, Day MR, Doumith M, *et al.*: **Comparison of phenotypic and WGS-derived antimicrobial resistance profiles of *Shigella sonnei* isolated from cases of diarrhoeal disease in England and Wales, 2015.** *J Antimicrob Chemother.* 2017; **72**(9): 2496–2502.
PubMed Abstract | Publisher Full Text

75. Bogaerts B, Delcourt T, Soetaert K, *et al.*: **A Bioinformatics Whole-Genome Sequencing Workflow for Clinical *Mycobacterium tuberculosis* Complex Isolate Analysis, Validated Using a Reference Collection Extensively Characterized with Conventional Methods and *In Silico* Approaches.** *J Clin Microbiol.* 2021; **59**(6): e00202–21.
PubMed Abstract | Publisher Full Text | Free Full Text

76. Walker TM, Kohl TA, Omar SV, *et al.*: **Whole-genome sequencing for prediction of *Mycobacterium tuberculosis* drug susceptibility and resistance: a retrospective cohort study.** *Lancet Infect Dis.* 2015; **15**(10): 1193–1202.
PubMed Abstract | Publisher Full Text | Free Full Text

77. Votintseva AA, Bradley P, Pankhurst L, *et al.*: **Same-day diagnostic and surveillance data for tuberculosis via whole-genome sequencing of direct respiratory samples.** *J Clin Microbiol.* 2017; **55**(5): 1285–1298.
PubMed Abstract | Publisher Full Text | Free Full Text

78. Mellmann A, Bletz S, Böking T, *et al.*: **Real-Time Genome Sequencing of Resistant Bacteria Provides Precision Infection Control in an Institutional

Setting. *J Clin Microbiol.* 2016; **54**(12): 2874–2881.
PubMed Abstract | Publisher Full Text | Free Full Text

79. Wilkinson MD, Dumontier M, Aalbersberg IJ, *et al.*: **The FAIR Guiding Principles for scientific data management and stewardship.** *Sci Data.* 2016; **3**(1): 160018.
PubMed Abstract | Publisher Full Text | Free Full Text

80. Berger FK, Mellmann A, von Müller L, *et al.*: **Quality assurance for genotyping and resistance testing of *Clostridium (Clostridioides) difficile* isolates - Experiences from the first inter-laboratory ring trial in four German speaking countries.** *Anaerobe.* 2020; **61**: 102093.
PubMed Abstract | Publisher Full Text

81. Mensah N, Tang Y, Cawthraw S, *et al.*: **Determining antimicrobial susceptibility in *Salmonella enterica* serovar Typhimurium through whole genome sequencing: a comparison against multiple phenotypic susceptibility testing methods.** *BMC Microbiol.* 2019; **19**(1): 148.
PubMed Abstract | Publisher Full Text | Free Full Text

82. Stubberfield E, AbuOun M, Sayers E, *et al.*: **Use of whole genome sequencing of commensal *Escherichia coli* in pigs for antimicrobial resistance surveillance, United Kingdom, 2018.** *Euro Surveill.* 2019; **24**(50): 1900136.
PubMed Abstract | Publisher Full Text | Free Full Text

83. Deplano A, Dodémont M, Denis O, *et al.*: **European external quality assessments for identification, molecular typing and characterization of *Staphylococcus aureus*.** *J Antimicrob Chemother.* 2018; **73**(10): 2662–2666.
PubMed Abstract | Publisher Full Text

84. Johansen TB, Scheffer L, Jensen VK, *et al.*: **Whole-genome sequencing and antimicrobial resistance in *Brucella melitensis* from a Norwegian perspective.** *Sci Rep.* 2018; **8**(1): 8538.
PubMed Abstract | Publisher Full Text | Free Full Text

85. Neuert S, Nair S, Day MR, *et al.*: **Prediction of Phenotypic Antimicrobial Resistance Profiles From Whole Genome Sequences of Non-typhoidal *Salmonella enterica*.** *Front Microbiol.* 2018; **9**: 592.
PubMed Abstract | Publisher Full Text | Free Full Text

86. Karlsmose PS, Hendriksen RS, Bortolaia V: **The 23rd EURL-AR Proficiency TestSalmonella, Campylobacter and genotypic characterisation 2017**.
Reference Source

87. Pietsch M, Irrgang A, Roschanski N, *et al.*: **Whole genome analyses of CMY-2-producing *Escherichia coli* isolates from humans, animals and food in Germany.** *BMC Genomics.* 2018; **19**(1): 601.
PubMed Abstract | Publisher Full Text | Free Full Text

88. Tyson GH, Sabo JL, Rice-Trujillo C, *et al.*: **Whole-genome sequencing based characterization of antimicrobial resistance in *Enterococcus*.** *Pathog Dis.* 2018; **76**(2).
PubMed Abstract | Publisher Full Text

89. Bossé JT, Li Y, Rogers J, *et al.*: **Whole Genome Sequencing for Surveillance of Antimicrobial Resistance in *Actinobacillus pleuropneumoniae*.** *Front Microbiol.* 2017; **8**: 311.
PubMed Abstract | Publisher Full Text | Free Full Text

90. Brhelova E, Antonova M, Pardy F, *et al.*: **Investigation of next-generation sequencing data of *Klebsiella pneumoniae* using web-based tools.** *J Med Microbiol.* 2017; **66**(11): 1673–1683.
PubMed Abstract | Publisher Full Text

91. Carroll LM, Wiedmann M, den Bakker H, *et al.*: **Whole-Genome Sequencing of Drug-Resistant *Salmonella enterica* Isolates from Dairy Cattle and Humans in New York and Washington States Reveals Source and Geographic Associations.** *Appl Environ Microbiol.* 2017; **83**(12): e00140–17.
PubMed Abstract | Publisher Full Text | Free Full Text

92. Day M, Doumith M, Jenkins C, *et al.*: **Antimicrobial resistance in Shiga toxin-producing *Escherichia coli.* serogroups O157 and O26 isolated from human cases of diarrhoeal disease in England, 2015.** *J Antimicrob Chemother.* 2017; **72**(1): 145–152.
PubMed Abstract | Publisher Full Text

93. Global Microbial Identifier initiative's Working Group 4: **The proficiency test (pilot) report of the global microbial identifier (GMI) initiative, year 2014**.
Reference Source

94. McDermott PF, Tyson GH, Kabera C, *et al.*: **Whole-Genome Sequencing for Detecting Antimicrobial Resistance in Nontyphoidal *Salmonella*.** *Antimicrob Agents Chemother.* 2016; **60**(9): 5515–5520.
PubMed Abstract | Publisher Full Text | Free Full Text

95. Tyson GH, McDermott PK, Li C, *et al.*: **WGS accurately predicts antimicrobial resistance in *Escherichia coli*.** *J Antimicrob Chemother.* 2015; **70**(10): 2763–2769.
PubMed Abstract | Publisher Full Text

96. Zhao S, Tyson GH, Chen Y, *et al.*: **Whole-Genome Sequencing Analysis Accurately Predicts Antimicrobial Resistance Phenotypes in Campylobacter spp.** *Appl Environ Microbiol.* 2015; **82**(2): 459–466.
PubMed Abstract | Publisher Full Text | Free Full Text

97. Kos VN, Déraspe M, McLaughlin RE, *et al.*: **The Resistome of *Pseudomonas aeruginosa* in Relationship to Phenotypic Susceptibility.** *Antimicrob Agents Chemother.* 2015; **59**(1): 427–436.
PubMed Abstract | Publisher Full Text | Free Full Text

98. Gordon NC, Price JR, Cole K, *et al.*: **Prediction of *Staphylococcus aureus* antimicrobial resistance by whole-genome sequencing.** *J Clin Microbiol.* 2014; **52**(4): 1182–1191.
PubMed Abstract | Publisher Full Text | Free Full Text

99. Stoesser N, Batty EM, Eyre DW, *et al.*: **Predicting antimicrobial susceptibilities for *Escherichia coli* and *Klebsiella pneumoniae* isolates using whole genomic sequence data.** *J Antimicrob Chemother.* 2013; **68**(10): 2234–2244.
PubMed Abstract | Publisher Full Text | Free Full Text

100. Eyre DW, Golubchik T, Gordon NC, *et al.*: **A pilot study of rapid benchtop sequencing of *Staphylococcus aureus.* and *Clostridium difficile.* for outbreak detection and surveillance.** *BMJ Open.* 2012; **2**(3): e001124.
PubMed Abstract | Publisher Full Text | Free Full Text

101. Zankari E, Hasman H, Kaas RS, *et al.*: **Genotyping using whole-genome sequencing is a realistic alternative to surveillance based on phenotypic antimicrobial susceptibility testing.** *J Antimicrob Chemother.* 2013; **68**(4): 771–777.
PubMed Abstract | Publisher Full Text

102. Cooper AL, Low AJ, Koziol AG, *et al.*: **Systematic Evaluation of Whole Genome Sequence-Based Predictions of *Salmonella* Serotype and Antimicrobial Resistance.** *Front Microbiol.* 2020; **11**: 549.
PubMed Abstract | Publisher Full Text | Free Full Text

103. Berglund F, Marathe NP, Österlund T, *et al.*: **Identification of 76 novel B1 metallo-β-lactamases through large-scale screening of genomic and metagenomic data.** *Microbiome.* 2017; **5**(1): 134.
PubMed Abstract | Publisher Full Text | Free Full Text

104. Hatosy SM, Martiny AC: **The ocean as a global reservoir of antibiotic resistance genes.** *Appl Environ Microbiol.* 2015; **81**(21): 7593–7599.
PubMed Abstract | Publisher Full Text | Free Full Text

105. Bengtsson-Palme J, Kristiansson E, Larsson DGJ: **Environmental factors influencing the development and spread of antibiotic resistance.** *FEMS Microbiol Rev.* 2018; **42**(1): fux053.
PubMed Abstract | Publisher Full Text | Free Full Text

106. Ashbolt NJ, Amézquita A, Backhaus T, *et al.*: **Human Health Risk Assessment (HHRA) for Environmental Development and Transfer of Antibiotic Resistance.** *Environ Health Perspect.* 2013; **121**(9): 993–1001.
PubMed Abstract | Publisher Full Text | Free Full Text

107. Martínez JL, Coque TM, Baquero F: **Prioritizing risks of antibiotic resistance genes in all metagenomes.** *Nat Rev Microbiol.* 2015; **13**(6): 396–396.
PubMed Abstract | Publisher Full Text

108. Wright GD: **The antibiotic resistome: the nexus of chemical and genetic diversity.** *Nat Rev Microbiol.* 2007; **5**(3): 175–186.
PubMed Abstract | Publisher Full Text

109. Perry JA, Westman EL, Wright GD: **The antibiotic resistome: what's new?** *Curr Opin Microbiol.* 2014; **21**: 45–50.
PubMed Abstract | Publisher Full Text

110. Berendonk TU, Manaia CM, Merlin C, *et al.*: **Tackling antibiotic resistance: the environmental framework.** *Nat Rev Microbiol.* 2015; **13**(5): 310–317.
PubMed Abstract | Publisher Full Text

111. Sinha T, Vila AV, Garmaeva S, *et al.*: **Analysis of 1135 gut metagenomes identifies sex-specific resistome profiles.** *Gut Microbes.* 2019; **10**(3): 358–366.
PubMed Abstract | Publisher Full Text | Free Full Text

112. Liu J, Taft DH, Maldonado-Gomez MX, *et al.*: **The fecal resistome of dairy cattle is associated with diet during nursing.** *Nat Commun.* 2019; **10**(1): 4406.
PubMed Abstract | Publisher Full Text | Free Full Text

113. Ruppé E, Ghozlane A, Tap J, *et al.*: **Prediction of the intestinal resistome by a three-dimensional structure-based method.** *Nat Microbiol.* 2019; **4**(1): 112–123.
PubMed Abstract | Publisher Full Text

114. Sczyrba A, Hofmann P, Belmann P, *et al.*: **Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.** *Nat Methods.* 2017; **14**(11): 1063–1071.
PubMed Abstract | Publisher Full Text | Free Full Text

115. Fritz A, Hofmann P, Majda S, *et al.*: **CAMISIM: simulating metagenomes and microbial communities.** *Microbiome.* 2019; **7**(1): 17.
PubMed Abstract | Publisher Full Text | Free Full Text

116. Ma L, Li B, Zhang T: **New insights into antibiotic resistome in drinking water and management perspectives: A metagenomic based study of small-sized microbes.** *Water Res.* 2019; **152**: 191–201.
PubMed Abstract | Publisher Full Text

117. Bai Y, Ruan X, Xie X, *et al.*: **Antibiotic resistome profile based on metagenomics in raw surface drinking water source and the influence of environmental factor: A case study in Huaihe River Basin, China.** *Environ Pollut.* 2019; **248**: 438–447.
PubMed Abstract | Publisher Full Text

118. Ju F, Beck K, Yin X, *et al.*: **Wastewater treatment plant resistomes are shaped by bacterial composition, genetic exchange, and upregulated expression in the effluent microbiomes.** *ISME J.* 2019; **13**(2): 346–360.
PubMed Abstract | Publisher Full Text | Free Full Text

119. Ng C, Tan B, Jiang XT, *et al.*: **Metagenomic and resistome analysis of a full-scale municipal wastewater treatment plant in Singapore containing membrane bioreactors.** *Front Microbiol.* 2019; **10**: 172.
PubMed Abstract | Publisher Full Text | Free Full Text

120. Makowska N, Philips A, Dabert M, *et al.*: **Metagenomic analysis of β-lactamase and carbapenemase genes in the wastewater resistome.** *Water Res.* 2020; **170**: 115277.
PubMed Abstract | Publisher Full Text

121. Buelow E, Rico A, Gaschet M, *et al.*: **Classification of hospital and urban wastewater resistome and microbiota over time and their relationship to the eco-exposome.** *bioRxiv.* 2019.
Publisher Full Text

122. Feng J, Li B, Jiang X, *et al.*: **Antibiotic resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses.** *Environ Microbiol.* 2018; **20**(1): 355–368.
PubMed Abstract | Publisher Full Text

123. Aarestrup FM, Woolhouse MEJ: **Using sewage for surveillance of antimicrobial resistance.** *Science.* 2020; **367**(6478): 630–632.
PubMed Abstract | Publisher Full Text

124. Lanza VF, Add-Baquero F, Luís Martínez J, *et al.*: **In-depth resistome analysis by targeted metagenomics.** *Microbiome.* 2018; **6**(1): 11.
PubMed Abstract | Publisher Full Text | Free Full Text

125. EFSA Panel on Additives and Products or Substances used in Animal Feed (FEEDAP); Rychen G, Aquilina G, *et al.*: **Guidance on the characterisation of microorganisms used as feed additives or as production organisms.** *EFSA J.* 2018; **16**(3).
PubMed Abstract | Publisher Full Text | Free Full Text

126. European Centre for Disease Prevention and Control: **Expert opinion on whole genome sequencing for public health surveillance**. 2016.
Reference Source

127. European Centre for Disease Control (ECDC), European Food Safety Authority (EFSA), Van Walle I, *et al.*: **EFSA and ECDC technical report on the collection and analysis of whole genome sequencing data from food-borne pathogens and other relevant microorganisms isolated from human, animal, food, feed and food/feed environmental samples in the joint ECDC-EFSA molecular typing database.** *EFSA Support Publ.* 2019; **16**(5).
Publisher Full Text

128. Mason A, Add-Foster D, Bradley P, *et al.*: **Accuracy of Different Bioinformatics Methods in Detecting Antibiotic Resistance and Virulence Factors from *Staphylococcus aureus*. Whole-Genome Sequences.** *J Clin Microbiol.* 2018; **56**(9).
PubMed Abstract | Publisher Full Text | Free Full Text

Open Peer Review

Current Peer Review Status: ? ✓ ?

Version 2

Reviewer Report 29 June 2022

https://doi.org/10.5256/f1000research.122296.r135212

?

Enrico Lavezzo
Department of Molecular Medicine, University of Padova, Padua, Italy
Emilio Ispano
Department of Molecular Medicine, University of Padova, Padua, Italy

The manuscript “A roadmap for the generation of benchmarking resources for antimicrobial resistance detection using next generation sequencing” by Petrillo and coauthors addresses a crucial issue, which is the need for a standard benchmark dataset to assess bioinformatics pipeline performing AMR detection/prediction. The logic followed by the authors to produce guidelines to build such benchmark dataset(s) is sound and clear, and so are the discussion and conclusion. Although the paper is well-written and holds solid evidence, and surely meets the need of the scientific community to have a gold standard for this topic, there are some paragraphs that could be adjusted for clarity and they will be listed below. In addition, it is not clear why past attempts to create such databases are not mentioned, since it could provide a much more advanced starting point than building it from scratch. I am referring to McArthur, A.G *et al.* (2013)[1] (which introduces also an ontology to represent AMR, could it serve the purpose addressed in the last bullet of section 3.2?), Zankari, *et al.* (2012)[2].

Abstract

- "For this application and in general, considerable challenges remain in demonstrating sufficient trust to act upon the meaningful information produced from raw data...", meaningful but convoluted sentence, could benefit from rephrasing.

Introduction

- First paragraph: A reference could be added to support the stated implementation of strategies using NGS to track the timeline and relationships between cases of an outbreak.

- Third paragraph: "For methods with important bioinformatics components...", what do you mean by important? That the bioinformatic component is crucial for the method? Or that the component is by itself complex and structured?

- Third paragraph: "in such cases like this..." is a repetition. I suggest using just "In such cases".

Section 2.3

- Second paragraph: from "Although understanding..." to "...too costly" is a sentence longer than needed, I suggest a full stop instead of the colon.

Section 2.4

- First paragraph: "Some species also have genes that are similar at the sequence level to known AMR determinants that efficient pipelines must be able to distinguish.", this could benefit from some references pointing to such species.

- Second paragraph: The authors imply the need to cut several bacterial species from the database, which is nonsense in the scope of an ideally global benchmark database. I would rather include as much information as possible and add metadata to characterise every database entry to discern whether to use it or not for a certain evaluation.

Section 2.6

- First paragraph: "Besides the set of reads themselves, additional information needs to be associated with each sample in the dataset, not for the benchmarking per se but its use for next benchmarking exercises", this sentence is hard to understand, please rephrase.

- Third paragraph: "and should be used as appropriate for the sake of transparency", what do the authors mean?

Section 3.1

- Fourth paragraph: Check also Murray (2022)[3] as it provides comprehensive data and estimations regarding many AMR cases worldwide.

Section 3.2

- Third paragraph: "A great source of data are the published ring trials...". Explain what a ring trial is since it is also a key step for the understanding of Table 3.

- Last bullet: I am again pointing to McArthur, A.G *et al.* (2013)[1] for such ontology.

References

1. McArthur AG, Waglechner N, Nizam F, Yan A, et al.: The comprehensive antibiotic resistance database.*Antimicrob Agents Chemother*. 2013; **57** (7): 3348-57 PubMed Abstract | Publisher Full Text
2. Zankari E, Hasman H, Cosentino S, Vestergaard M, et al.: Identification of acquired antimicrobial resistance genes.*J Antimicrob Chemother*. 2012; **67** (11): 2640-4 PubMed Abstract | Publisher Full Text
3. Murray C, Ikuta K, Sharara F, Swetschinski L, et al.: Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet*. 2022; **399** (10325): 629-655 Publisher Full Text

Is the topic of the opinion article discussed accurately in the context of the current literature?

Yes

Are all factual statements correct and adequately supported by citations?
Partly

Are arguments sufficiently supported by evidence from the published literature?
Partly

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbiology, bioinformatics, genetics

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.

Author Response 20 Jul 2022

Mauro Petrillo

Dear Dr Lavezzo and Dr Ispano,

Thanks a lot for your valuable comments and suggestions. It is our intention to address all of them as part of an updated version of the published paper.

Best regards,

Mauro Petrillo, on behalf of the authors.

Competing Interests: No competing interests were disclosed.

Reviewer Report 13 May 2022

https://doi.org/10.5256/f1000research.122296.r127727

Anna Abramova
Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden
Marcus Wenne

Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden

We thank the authors for taking into consideration our comments and suggestions.

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Yes

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: AMR, NGS data analysis, bioinformatics

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.

Author Response 13 May 2022

Mauro Petrillo

Dear Dr Abramova and Dr Wenne,

We would like to thank you for your valuable comments and suggestions which definitively improved the quality of our manuscript.

Best regards,

Mauro Petrillo, on behalf of the authors.

Competing Interests: No competing interests were disclosed.

Version 1

Reviewer Report 14 September 2021

https://doi.org/10.5256/f1000research.42277.r91069

?

Anna Abramova
[1] Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden
[2] Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden
Marcus Wenne
[1] Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden
[2] Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden

Petrillo and colleagues discuss approaches and associated challenges for establishing a performance-based benchmarking resource for antimicrobial resistance detection. The authors first describe general considerations and then provide scope-specific use cases. The paper represents a summary of the discussions held by experts in AMR and NGS-related fields during the JRC meeting. It is a relevant and important paper, and the initiative will be widely appreciated by the scientific community working with antimicrobial resistance. Overall, the paper is well-written, however, it would benefit from some additional clarifications and adjustments, which are explained in more detail in the comments below.

Introduction
- Since the main focus of the paper is benchmarking, it would be beneficial to provide a short background on the previous benchmarking resources/initiatives in the introduction (e.g. Mangul *et al*., 2019[1], Sczyrba *et al*., 2017[2], etc).

Section 2
- Paragraph 2: “In the conclusions of the previous article…”, it is confusing which article the authors refer to since reference 19 (Bellman *et al*., 2015) provided at the end of the sentence does not contain the cited text.

General considerations
- Several important questions such as which tools to include in the benchmarking, should they be run with optimised or default parameters (e.g. default or customised database), and what performance metrics are to be used for evaluation - could be added to the "General discussion" to clarify and benefit in the understanding of the subsequent sections.

Section 2.1
- This section discusses very important considerations when it comes to different sequencing technologies. Since sequencing technologies are constantly evolving perhaps it would be relevant to add some future perspectives, e.g. how to deal with emerging outperforming technologies.
- Paragraph 2: From the sentence starting with “It is important, in this case…”, it is not clear whether the authors mean the bias among different platforms’ outputs or outputs from the same platform.

Section 2.2
- Apart from being in silico generated or obtained from a real-life sample, the data can come from different types of samples. In the case of metagenomics, it would be an important consideration for the evaluation of bioinformatics tools (i.e. a human gut sample or a soil sample would be characterised by a very different complexity).

- Paragraph 7: “The main issues then are: a) there is a need to demonstrate that the experiment met the necessary quality criteria (see section 3.3)” - section 3.3 does not contain any information on the quality criteria.

- Paragraph 11: “Because each approach has advantages and disadvantages, the choice must be carefully considered, according to the purpose of the dataset, which will be discussed in section 4.” - section 4 is missing.

Section 2.4

- This is perhaps the most important section considering the focus of the paper on AMR detection, however, it is very concise and does not provide a good overview of the challenges (e.g. what type of AMR mechanisms to include, which pathogens to consider). These topics are described later in section 3.1 in the example of a particular use case. However, it would be beneficial to outline them in the General considerations section.

- Paragraph 2: “These are specific to the purpose of the dataset (Figure 1) and will be discussed in section 4.1–section 4.3 below” - sections 4.1-4.3 are missing.

Section 2.5

- To aid understanding it would be helpful to clarify genomic and phenotypic endpoint, perhaps by adding to the first sentence, e,g, “..a) they can detect the genetic determinants of AMR (genomic endpoint), and in addition b) some can predict the AMR/susceptibility of the bacteria in the original sample (phenotypic endpoint).”

Section 3.1

- Paragraph 4: “3. Combinations of (1) and (2) present in at least one of the chosen lists (see cells in Table 2), the sequences are combined and used as the input to simulate the reads using the appropriate tools (see section 3.2).” - maybe the authors meant section 2.3 instead of 3.2.

- Paragraph 6: “The endpoint considered for this benchmark is thus genotypic (see section 3.5),” - section 3.5 is missing.

- Paragraph 10: “When generated, the benchmark should be deployed on a dedicated (and sustainably maintained) platform that includes all the links to the data (see section 3.6) ” - section 3.6 is missing.

Conclusions

- The authors mention in the conclusion that they identified two main use cases for this benchmarking resource, each necessitating its own platform: single isolates and mixed samples. This could be expanded upon in the General considerations section to give the reader an understanding of the different challenges and approaches for the two use cases. It is also implied, but not specifically stated, in sections 3.1 and 3.2 that they focus on single isolates. This only becomes clear when reading section 3.3.

- Sections 4.1-4.3 are missing.

References

1. Mangul S, Martin L, Hill B, Lam A, et al.: Systematic benchmarking of omics computational tools. *Nature Communications*. 2019; **10** (1). Publisher Full Text

2. Sczyrba A, Hofmann P, Belmann P, Koslicki D, et al.: Critical Assessment of Metagenome Interpretation-a benchmark of metagenomics software.*Nat Methods*. 2017; **14** (11): 1063-1071
PubMed Abstract | Publisher Full Text

Is the topic of the opinion article discussed accurately in the context of the current literature?
Yes

Are all factual statements correct and adequately supported by citations?
Yes

Are arguments sufficiently supported by evidence from the published literature?
Yes

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Yes

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: AMR, NGS data analysis, bioinformatics

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.

Author Response 16 Sep 2021

Mauro Petrillo

Dear Dr Abramova and Dr Wenne,

Thanks a lot for your valuable comments and suggestions that you have provided in the report.

We will address all of them, together with those of other reviewers, in order to provide a fully revised version of the manuscript.

Best regards,

Mauro Petrillo, on behalf of the authors.

Competing Interests: No competing interests were disclosed.

Author Response 09 Mar 2022

Mauro Petrillo

Replies to:

Anna Abramova, Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden
Marcus Wenne, Department of Infectious Diseases, University of Gothenburg, Gothenburg, Sweden

REPLY: Thanks for your comments, we have revised the whole manuscript and we hope it is now in line with your expectations.

Introduction

- Since the main focus of the paper is benchmarking, it would be beneficial to provide a short background on the previous benchmarking resources/initiatives in the introduction (e.g. Mangul et al., 20191, Sczyrba et al., 20172, etc).

REPLY: Thanks the suggestion. We added them in the Conclusions section.

Section 2

Paragraph 2: "In the conclusions of the previous article...", it is confusing which article the authors refer to since reference 19 (Bellman et al., 2015) provided at the end of the sentence does not contain the cited text.

REPLY: Thanks for spotting this error, which has been corrected.

General considerations

- Several important questions such as which tools to include in the benchmarking, should they be run with optimised or default parameters (e.g. default or customised database), and what performance metrics are to be used for evaluation - could be added to the "General discussion" to clarify and benefit in the understanding of the subsequent sections.

REPLY: Section 3.1 was substantially rewritten to provide more information on the performance-based evaluation. Concretely, both tools and their parameters, should be open to the preferences and requirements of the end users. Certain groups might have a preference for certain tools because it is easier to implement for them, they have developed the tool themselves in-house, it's freely available as open-source solutions or alternatively commercially available but with full customer support etc. As long as said tools are demonstrated to provide a certain minimum performance, they can be used since the benchmarking should focus on performance rather than enforcing a single pipeline to be used by all (which at any rate, would be unlikely, given the plethora of different sequencing technologies, platforms, and chemistries). The same logic applies to settings, for which certain groups may wish to extensively finetune certain settings but others may prefer to keep them at defaults. With respect to the actual performance metrics to be used, more information was provided in the revised manuscript but we prefer redirecting to reference publications such as Kozyreva et al. (https://doi.org/10.1128/JCM.00361-17) for a more detailed explanation and example of how these could be implemented, since providing all that information would substantially enlarge the size of this paper.

Section 2.1

- This section discusses very important considerations when it comes to different sequencing technologies. Since sequencing technologies are constantly evolving perhaps it would be relevant to add some future perspectives, e.g. how to deal with emerging outperforming technologies.

REPLY: A sentence has been added to the end of the section indicating that implementation of new benchmark datasets should be prioritized if new and better technologies emerge.

- Paragraph 2: From the sentence starting with "It is important, in this case…", it is not clear whether the authors mean the bias among different platforms' outputs or outputs from the same platform.

REPLY: Text has been modified to try to clarify that this means bias among the different platforms.

Section 2.2

- Apart from being in silico generated or obtained from a real-life sample, the data can come from different types of samples. In the case of metagenomics, it would be an important consideration for the evaluation of bioinformatics tools (i.e. a human gut sample or a soil sample would be characterised by a very different complexity).

REPLY: We have addressed this point in section 3, which was revised.

- Paragraph 7: "The main issues then are: a) there is a need to demonstrate that the experiment met the necessary quality criteria (see section 3.3)" - section 3.3 does not contain any information on the quality criteria.

REPLY: Should be 2.3 Text has been modified accordingly.

- Paragraph 11: "Because each approach has advantages and disadvantages, the choice must be carefully considered, according to the purpose of the dataset, which will be discussed in section 4." - section 4 is missing.

REPLY: Should be 3. Text has been modified accordingly.

Section 2.4

- This is perhaps the most important section considering the focus of the paper on AMR detection, however, it is very concise and does not provide a good overview of the challenges (e.g. what type of AMR mechanisms to include, which pathogens to consider). These topics are described later in section 3.1 in the example of a particular use case. However, it would be beneficial to outline them in the General considerations section.

REPLY: The "General considerations" section has been expanded accordingly.

- Paragraph 2: "These are specific to the purpose of the dataset (Figure 1) and will be discussed in section 4.1–section 4.3 below" - sections 4.1-4.3 are missing.

REPLY: Should be 3.1-3.3. Text has been modified accordingly.

Section 2.5

- To aid understanding it would be helpful to clarify genomic and phenotypic endpoint, perhaps by adding to the first sentence, e,g, "..a) they can detect the genetic determinants of AMR (genomic endpoint), and in addition b) some can predict the AMR/susceptibility of the bacteria in the original sample (phenotypic endpoint)."

REPLY: Added as suggested.

Section 3.1

- Paragraph 4: "3. Combinations of (1) and (2) present in at least one of the chosen lists

(see cells in Table 2), the sequences are combined and used as the input to simulate the reads using the appropriate tools (see section 3.2).” - maybe the authors meant section 2.3 instead of 3.2.

REPLY: Should be 2.3, text has been modified.

- Paragraph 6: “The endpoint considered for this benchmark is thus genotypic (see section 3.5),” - section 3.5 is missing.

REPLY: Should be 2.5- Text has been modified accordingly

- Paragraph 10: “When generated, the benchmark should be deployed on a dedicated (and sustainably maintained) platform that includes all the links to the data (see section 3.6) ” - section 3.6 is missing.

REPLY: Should be 2.6. Text has been modified accordingly

Conclusions

- The authors mention in the conclusion that they identified two main use cases for this benchmarking resource, each necessitating its own platform: single isolates and mixed samples. This could be expanded upon in the General considerations section to give the reader an understanding of the different challenges and approaches for the two use cases. It is also implied, but not specifically stated, in sections 3.1 and 3.2 that they focus on single isolates. This only becomes clear when reading section 3.3.

REPLY: A short introduction in section 3 has been added. The “General considerations” section has been expanded accordingly, too.

- Sections 4.1-4.3 are missing.

REPLY: Should be section 3. Text has been modified accordingly.

Competing Interests: No competing interests were disclosed.

Reviewer Report 21 May 2021

https://doi.org/10.5256/f1000research.42277.r84352

?

Rene Hendriksen
[1] National Food Institute, Technical University of Denmark, Bygning, Lyngby, Denmark
[2] National Food Institute, Technical University of Denmark, Bygning, Lyngby, Denmark

This manuscript describes a road map how to set up and conduct benchmarking to assess bioinformatics pipelines to detect AMR genes in three levels.

Overall, comments:
The manuscript is well-written but I find it in several paragraphs hard to comprehend the sentences as the authors contradict themselves. This needs to be address as the topic is important

but it needs to be all clear.
I suggest to focus the paper on single isolate genomes rather than but this and metagenomics. It is really two separate technologies and will need different approaches. It is trying to explain all but fails really to in depth address metagenomics.

Specific comments:

- Page 4 1st paragraph: Describe other epidemiological traits alongside with the timeline and relatedness by merging the paragraphs "....such as virulence, resistance to antibiotics, typing and other adaptive traits... to the same paragraph. I was missing the characterization of AMR genes in the initial lines as this is the focus and title of the paper.

- Page 4 introduction: I miss an explanation about "complex microbial communities". I know this is metagenomics but the term needs to be introduced.

- Page 4 introduction: ".... antimicrobial resistance (AMR) genetic determinants from NGS data.... This needs to be clarified if this include acquired antimicrobial resistance genes AND/ OR chromosomal point mutations. This is not clear what the approach includes.

- Page 5 2nd bullet: Clarify what is mean by resistome. I suggest to make it clear that this bullet deals with metagenomics.

- Page 5 last paragraph: "Bioinformatics pipelines are thus usually designed to handle the output of a specific platform, often in a certain configuration". This is true but also the achilles heel of the further description where this is being contradicted.

- Fig 1: I miss the word "concordance to phenotype" as well as curation for scope 1.

- Page 6 2nd paragraph: "in practice, a prioritisation exercise should be made based on the capacity building efforts in testing laboratories". Clarify why this is needed!

- Fig 2: The data could easily be explained in text – omit fig 2 or reference already published similar figures.

- Page 6 4th paragraph: I don't understand the concept. It was earlier explained that "Bioinformatics pipelines are thus usually designed to handle the output of a specific platform, often in a certain configuration" so, why compare the output of a certain pipeline from data generated from different platforms knowing that the result for certain platforms will be biased due to the low comparability of a certain pipeline.

- Page 6 5th paragraph: "The FASTQ format is a standard format in this context, which should be used in the benchmark resources; many tools exist to convert the raw data output files into this format in case of different platform outputs (see, for example,39,40) although, it should be noted, different tools may produce different results and this step should be carefully planned." I find this a source for bring in bias to the benchmarking. I find it hard to see how one can trust the analysis when bringing in variation which might not even be controlled.

- Page 7 1st, 2nd, 7th paragraph: I did it contradicting that its phrased that " Although the

disadvantage of simulating in silico data is obvious (it is not 'real'), there are some substantial advantages: it is a lot cheaper than performing sequencing runs, a lot faster, and can be applied to any genome previously sequenced." and "However, a major drawback is that simulating variation the way nature evolves is very challenging – genetic variation happens in places in the genome where it is hardest to find." And "although this requires strict annotation of the experiment; c) it will not be possible (besides rare exceptions) to build datasets for the different platforms using the same initial samples." First of all, its not all that can prepare simulated datasets and secondly, its correct that it will never mimic nature, Thus, I don't see why this is so heavily recommended.

- Page 9 1st paragraph: I miss N50, no of contigs etc.. to be mentioned as QC metrics.
- Page 9 2nd paragraph: Not easy to understand.
- Page 9 2.4 1st paragraph: "a very pragmatic approach could be the generation of random DNA sequences, to which particular sequences of interest are added (i.e. fragments of AMR genes). However, there is sufficient evidence that the genomic background of the bacteria (i.e. the "non-AMR related" sequences) can have a profound influence on the performance of the pipelines". Contradiction – see Page 7 1st, 2nd , 7th paragraph.
- Page 9 6th paragraph: define "phenotypic endpoint".
- Page 9 6th paragraph: "Studies that evaluated AMR genotype to phenotype relationships have indicated that despite generally high correspondence, this can vary greatly between pathogens / case studies, and even for different antimicrobial agents within the same species 57,58." I need to be further elaborated – what do one trust the phenotypic data or detected genes.
- Page 9 intro to bullets: define "genomic endpoint".
- Page 9 section 2.6: Agree that metadata is needed but it needs to be explained that its not needed for the benchmarking itself but for others to use the dataset for future exercises.
- Page 10 1st paragraph of section 3.1: explain what is meant by "agreed minimum standards"- what performance metrics.
- Page 10 3rd paragraph of section 3.1: I find it contradicting to Page 7 1st, 2nd, 7th paragraph.
- Page 10: WHO CIA has been updated – provide ref.
- Page 10: The decision has been updated in 2021 – provide ref.
- Page 13: Tabel 3 greatly lack a million PT/ EQA reports from EURLs e.g. https://www.eurl-ar.eu/reports.aspx https://antimicrobialresistance.dk/eqas.aspx
- Page 14 section 3.3: I would omit this part as it add more confusion to bring in also metagenomics to the concept – a completely different approach with complex samples. Past studies has also show that benchmarking metagenomics is not a trivial discipline.

Is the topic of the opinion article discussed accurately in the context of the current literature?
Partly

Are all factual statements correct and adequately supported by citations?
Partly

Are arguments sufficiently supported by evidence from the published literature?
No

Are the conclusions drawn balanced and justified on the basis of the presented arguments?
Partly

Competing Interests: No competing interests were disclosed.

Reviewer Expertise: Microbiologist focusing on NGS and EQA/ benchmarking

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.

Author Response 16 Sep 2021

Mauro Petrillo

Dear Dr Hendriksen,

Thanks a lot for your valuable comments and suggestions that you have provided in the report.

We will address all of them, together with those of other reviewers, in order to provide a fully revised version of the manuscript.

Best regards,

Mauro Petrillo, on behalf of the authors.

Competing Interests: No competing interests were disclosed.

Author Response 09 Mar 2022

Mauro Petrillo

Replies to Rene Hendriksen, National Food Institute, Technical University of Denmark, Bygning, Lyngby, Denmark.
Overall, comments
The manuscript is well-written but I find it in several paragraphs hard to comprehend the

sentences as the authors contradict themselves. This needs to be address as the topic is important but it needs to be all clear.
I suggest to focus the paper on single isolate genomes rather than but this and metagenomics. It is really two separate technologies and will need different approaches. It is trying to explain all but fails really to in depth address metagenomics.
REPLY: Thanks for your comments, we have revised the whole manuscript and we hope it is now in line with your expectations.

Specific comments

- Page 4 1st paragraph: Describe other epidemiological traits alongside with the timeline and relatedness by merging the paragraphs "....such as virulence, resistance to antibiotics, typing and other adaptive traits... to the same paragraph. I was missing the characterization of AMR genes in the initial lines as this is the focus and title of the paper.

REPLY: Paragraphs 1 and 2 have been merged, and reference to the use of this data for inferring AMR has been moved to early in paragraph 1 as suggested.

- Page 4 introduction: I miss an explanation about "complex microbial communities". I know this is metagenomics but the term needs to be introduced.

REPLY: As requested, bullets on this page have been rewritten to introduce the "metagenome" terminology

- Page 4 introduction: ".... antimicrobial resistance (AMR) genetic determinants from NGS data.... This needs to be clarified if this include acquired antimicrobial resistance genes AND/ OR chromosomal point mutations. This is not clear what the approach includes.

REPLY: The bullets on page 4 have been modified to indicate that phenotype prediction relies on detection of both acquired ARGs and point mutations. We felt that this was a better place to define the types of genetic determinants associated with AMR

- Page 5 2nd bullet: Clarify what is mean by resistome. I suggest to make it clear that this bullet deals with metagenomics.

REPLY: This bullet has been rewritten to address clarity.

- Page 5 last paragraph: "Bioinformatics pipelines are thus usually designed to handle the output of a specific platform, often in a certain configuration". This is true but also the achilles heel of the further description where this is being contradicted.

REPLY: We have rewritten section 3.1 to make clearer that validation of different pipelines should be performance-based, i.e. focus on their performance rather than enforcing a single pipeline to be adopted by the community. Different sequencing technologies exist, but also for the same technology there exist differences in instruments and chemistries that potentially require specifically adopted pipelines to accommodate specific configurations. By building community benchmarking datasets for which the ground truth is well-established for those configurations, prioritizing dominant technologies, it becomes possible that different pipelines are used by different groups (even when both groups otherwise use exactly the same configuration), as long as those pipelines meet minimum agreed upon acceptance values for their performance.

- Fig 1: I miss the word "concordance to phenotype" as well as curation for scope 1.

REPLY: Sorry, we did not understand this point. Do you mean these concepts should be added to Figure 1? Or that what is reported in the figure 1 is inconsistent with what mentioned in the text?

- Page 6 2nd paragraph: "in practice, a prioritisation exercise should be made based on the capacity building efforts in testing laboratories". Clarify why this is needed!

REPLY: This has been reworded to indicate that it may be necessary if resource limitations are an issue.

- Fig 2: The data could easily be explained in text – omit fig 2 or reference already published similar figures.

REPLY: We added additional text to justify the presence of Figure 2.

- Page 6 4th paragraph: I don't understand the concept. It was earlier explained that "Bioinformatics pipelines are thus usually designed to handle the output of a specific platform, often in a certain configuration" so, why compare the output of a certain pipeline from data generated from different platforms knowing that the result for certain platforms will be biased due to the low comparability of a certain pipeline.

REPLY: We provide here a clarification. Even though pipelines can be drastically different, e.g. made for different sequencing technologies or use completely different algorithmic approaches for a certain configuration (e.g. assembly-based or read-mapping based strategies), their final aim in the context of this manuscript considers the correct detection and identification of AMR determinants (whether SNPs, genes or more complex features), as quantified by validating their performance. Consequently, it is not the output of those pipelines themselves that will be directly compared, but rather the ability of said pipelines to correctly detect and identify AMR determinants. Whatever sequencing technology or bioinformatics methodology is applied, if a pipeline designed to detect AMR genes cannot properly detect and identify those genes, then said pipeline cannot be considered validated for AMR gene detection in a clinical or regulatory application. Contrarily, if said pipeline can correctly detect and identify those genes, then it could be applied in such contexts.

- Page 6 5th paragraph: "The FASTQ format is a standard format in this context, which should be used in the benchmark resources; many tools exist to convert the raw data output files into this format in case of different platform outputs (see, for example,39,40) although, it should be noted, different tools may produce different results and this step should be carefully planned." I find this a source for bring in bias to the benchmarking. I find it hard to see how one can trust the analysis when bringing in variation which might not even be controlled.

REPLY: This is exactly what the generation of community benchmark datasets circumvents by providing reference datasets for which the ground truth is well-established. By starting from such samples, whether using reference genomes with known AMR determinants for which reads are simulated with error profiles modelled to mimic specific sequencing technologies, or alternatively using real samples where certain AMR determinants have been shown to be present/absent with traditional methods (e.g. PCR and/or Sanger sequencing) subjected to sequencing by specific sequencing technologies, the resulting datasets can be analyzed with pipelines for which it is known which AMR determinants should be detected. If such an approach uncovers that certain variation and biases specific to certain sequencing technologies and/or configurations cannot be controlled for bioinformatically and negatively affect pipeline performance leading to missed detection (or alternatively false positive detections), said pipeline cannot be considered validated for AMR characterization in a clinical or regulatory application. We hope this clarifies.

- Page 7 1st, 2nd, 7th paragraph: I did it contradicting that its phrased that " Although the disadvantage of simulating in silico data is obvious (it is not 'real'), there are some substantial advantages: it is a lot cheaper than performing sequencing runs, a lot faster, and can be applied to any genome previously sequenced." and "However, a

major drawback is that simulating variation the way nature evolves is very challenging – genetic variation happens in places in the genome where it is hardest to find." And "although this requires strict annotation of the experiment; c) it will not be possible (besides rare exceptions) to build datasets for the different platforms using the same initial samples." First of all, its not all that can prepare simulated datasets and secondly, its correct that it will never mimic nature, Thus, I don't see why this is so heavily recommended.

REPLY: We have rewritten section 3.1 to take the suggestion of the reviewer into account by rendering the second approach for generating benchmark data, i.e. sequencing of real samples on the condition that their AMR determinants are well-established by other approaches (such as conventional PCR and/or Sanger sequencing), more prominent.

- Page 9 1st paragraph: I miss N50, no of contigs etc.. to be mentioned as QC metrics.

REPLY: We have now included quality of the assembly as an example of an important QC metric.

- Page 9 2nd paragraph: Not easy to understand.

REPLY: This paragraph has been rewritten for clarity.

- Page 9 2.4 1st paragraph: "a very pragmatic approach could be the generation of random DNA sequences, to which particular sequences of interest are added (i.e. fragments of AMR genes). However, there is sufficient evidence that the genomic background of the bacteria (i.e. the "non-AMR related" sequences) can have a profound influence on the performance of the pipelines". Contradiction – see Page 7 1st, 2nd , 7th paragraph.

REPLY: Thanks for spotting this contradiction. It has been clarified by rephrasing.

- page 9 6th paragraph: define "phenotypic endpoint".

REPLY: This has been defined in the first paragraph of section 2.5

- Page 9 6th paragraph: "Studies that evaluated AMR genotype to phenotype relationships have indicated that despite generally high correspondence, this can vary greatly between pathogens / case studies, and even for different antimicrobial agents within the same species 57,58." I need to be further elaborated – what do one trust the phenotypic data or detected genes.

REPLY: The subsequent paragraphs explain that focusing on a genomic endpoint has advantages. The name of the section has been change, as its aim is not to assess what is better between the two described endpoints.

- Page 9 intro to bullets: define "genomic endpoint".

REPLY: This has been defined in the first paragraph of section 2.5

- Page 9 section 2.6: Agree that metadata is needed but it needs to be explained that its not needed for the benchmarking itself but for others to use the dataset for future exercises.

REPLY: Added a sentence to clarify this.

- Page 10 1st paragraph of section 3.1: explain what is meant by "agreed minimum standards"- what performance metrics.

REPLY: We have substantially revised section 3.1 to render the "agreed minimum standards" clearer.

- Page 10 3rd paragraph of section 3.1: I find it contradicting to Page 7 1st, 2nd, 7th paragraph.
- Page 10: WHO CIA has been updated – provide ref.
  REPLY: the WHO web page states the a new update will be issued in 2022. WHO release:

2020 annual review of the clinical and preclinical antibacterial pipelines evaluates the potential of antibacterial candidates (https://www.who.int/publications/i/item/9789240021303)

- Page 10: The decision has been updated in 2021 – provide ref.
  REPLY: changed according to request.
- Page 13: Tabel 3 greatly lack a million PT/ EQA reports from EURLs e.g. https://www.eurl-ar.eu/reports.aspx https://antimicrobialresistance.dk/eqas.aspx

REPLY: We agree that a lot of PT/ EQA reports from EURLs exist. However, Table 3 is proposed as "a non-exhaustive list of recent references to be used as a starting point" and, among them, reference 85 is a report from the EURL-AR. Anyway, we modified the text to highlight the relevance of these reports.

Page 14 section 3.3: I would omit this part as it add more confusion to bring in also metagenomics to the concept – a completely different approach with complex samples. Past studies has also show that benchmarking metagenomics is not a trivial discipline.
REPLY: We did not omit this part as we believe it is linked to the resistome concept that we addressed, thanks to your suggestions.

Competing Interests: No competing interests were disclosed.