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]

Alexandre Angers-Loustau1, Mauro Petrillo1, Johan Bengtsson-Palme2,3, Thomas Berendonk4, Burton Blais5, Kok-Gan Chan6,7, Teresa M. Coque8, Paul Hammer9, Stefanie Heß4, Dafni M. Kagkli1, Carsten Krumbiegel9, Val F. Lanza8, Jean-Yves Madec10, Thierry Naas11, Justin O'Grady12, Valentina Paracchini1, John W.A. Rossen13, Etienne Ruppé14, Jessica Vamathevan15, Vittorio Venturi16, Guy Van den Eede17

1European Commission Joint Research Centre, Ispra, 21027, Italy
2Department of Infectious Diseases, Institute of Biomedicine, The Sahlgrenska Academy, University of Gothenburg, Gothenburg, SE-413 46, Sweden
3Centre for Antibiotic Resistance research (CARe), University of Gothenburg, SE-413 46, Gothenburg, Sweden
4Institute for Hydrobiology, Technische Universität Dresden, Dresden, 01307, Germany
5Canadian Food Inspection Agency, Ottawa Laboratory (Carling), Ottawa, ON, K1A 0Y9, Canada
6International Genome Centre, Jiangsu University, Zhenjiang, China
7Division of Genetics and Molecular Biology, Institute of Biological Sciences, University of Malaya, Kuala Lumpur, 50603, Malaysia
8Departamento de Microbiología, Hospital Universitario Ramón y Cajal, Instituto Ramón y Cajal de Investigación Sanitaria (IRYCIS), Madrid, 28034, Spain
9BIOMES.world, c/o Technische Hochschule Wildau, Wildau, 15745, Germany
10Unité Antibiorésistance et Virulence Bactériennes, ANSES Site de Lyon, Lyon, F-69364, France
11Service de Bactériologie-Hygiène, Hôpital de Bicêtre, Le Kremlin-Bicêtre, F-94275, France
12Norwich Medical School, University of East Anglia, Norwich, NR4 7TJ, UK
13Department of Medical Microbiology, University Medical Center Groningen, University of Groningen, Groningen, 9713 GZ, The Netherlands
14Laboratoire de Bactériologie, Hôpital Bichat, INSERM, IAM, UMR 1137, Université Paris Diderot, Paris, F-75018, France
15European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, CB10 1SD, UK
16International Centre for Genetic Engineering and Biotechnology (ICGB), Trieste, 34149, Italy
17European Commission Joint Research Centre, Geel, B-2440, Belgium

First published: 13 Apr 2018, 7:459
https://doi.org/10.12688/f1000research.14509.1

Latest published: 07 Dec 2018, 7:459
https://doi.org/10.12688/f1000research.14509.2

Abstract
Next-Generation Sequencing (NGS) technologies are expected to play a crucial role in the surveillance of infectious diseases, with their unprecedented capabilities for the characterisation of genetic information underlying the virulence and antimicrobial resistance (AMR) properties of microorganisms. In the implementation of any
novel technology for regulatory purposes, important considerations such as harmonisation, validation and quality assurance need to be addressed. NGS technologies pose unique challenges in these regards, in part due to their reliance on bioinformatics for the processing and proper interpretation of the data produced. Well-designed benchmark resources are thus needed to evaluate, validate and ensure continued quality control over the bioinformatics component of the process. This concept was explored as part of a workshop on "Next-generation sequencing technologies and antimicrobial resistance" held October 4-5 2017. Challenges involved in the development of such a benchmark resource, with a specific focus on identifying the molecular determinants of AMR, were identified. For each of the challenges, sets of unsolved questions that will need to be tackled for them to be properly addressed were compiled. These take into consideration the requirement for monitoring of AMR bacteria in humans, animals, food and the environment, which is aligned with the principles of a “One Health" approach.

Keywords
Antimicrobial resistance, bioinformatics, next-generation sequencing, benchmarking

This article is included in the International Society for Computational Biology Community Journal gateway.

This article is included in the Antimicrobial Resistance collection.
1. Introduction

Next-Generation Sequencing (NGS) technologies are increasingly regarded as an essential tool in modern regulatory frameworks. Monitoring schemes that rely on the characterisation of genetic information will gain considerably by utilising these technologies. Their importance for infectious diseases surveillance was highlighted by “The Review on Antimicrobial Resistance” in 2014, which stated that “advances in genetics, genomics and computer science will likely change the way that infections and new types of resistance are diagnosed, detected and reported worldwide, so that we can fight back faster when bacteria evolve to resist drugs”1.

This interest can be observed in the rapid expansion in recent years of whole-genome sequencing capacities in national public health infectious diseases surveillance laboratories, as recently reported in a European survey by the European Centre for Disease Prevention and Control (ECDC)2. Antimicrobial resistance (AMR), i.e. the ability of a microorganism to resist the action of an antimicrobial agent, is of particular importance in this surveillance program. Its observed rise places heavy burdens on healthcare systems, leading to prolonged treatment times, higher mortality and high economic impacts (see 3). In March 2017, the Joint Research Centre organised a meeting in order to better understand the state-of-the-art of the application of NGS technologies in the fight against AMR4. Although it is clear that the uses of NGS vary according to the specific need (e.g. to guide clinical intervention or to evaluate the environmental and human health risks of AMR genetic determinants), these discussions highlighted overlaps in the needs and the challenges of implementing NGS for the monitoring of AMR in humans, animals, food and the environment. Some of these were also highlighted in previous workshops organized by the European Food Safety Authority (EFSA) and the ECDC5,6.

A full regulatory implementation of NGS technologies to monitor AMR will need to address many standardisation challenges throughout the process, which broadly includes sample preparation and DNA extraction, library preparation for sequencing, the use of an NGS instrument for generating the sequences, the bioinformatics analysis, and interpretation and reporting of results (see Figure 1). Focusing on the bioinformatics step, an important shared challenge is the need to correctly and reliably identify the known genomic determinants of AMR from a set of NGS reads produced from sequencing a sample. The ECDC study reported the requirement for sufficient bioinformatics expertise as one of the important hurdles to a more general implementation of NGS for routine testing2. This observation has also been made in recent case studies and reviews7–11.

By contrast, within the scientific research community the recent literature reflects widespread enthusiasm for the application of NGS approaches to the determination of AMR.

---

**Figure 1.** Overview of the different steps involved in the use of Next-Generation Sequencing technologies for the detection and monitoring of antimicrobial resistance. The benchmark strategy discussed in the current article focuses on the bioinformatics steps, the pipeline converting the output of the sequencing experiment into a list of identified antimicrobial resistance genetic determinants (dashed rectangle).
characteristics in bacteria. For the bioinformatics steps, many useful strategies have been published. These are, however, very varied in the approaches and resources they use. Some start with sequencing reads produced by the Illumina1,2, Ion Torrent4, PacBio5 or Nanopore6 platforms, just to give a few examples. To predict the resistance profile, interesting results were reported with very different strategies, including k-mer analysis of the reads7, sequence comparisons of individual reads to databases8,9, first assembling the reads into contigs using various software packages10,11 and building and comparing de Bruijn graphs of the sequenced sample reads and the reference database12. The reference set of genetic determinants of AMR used by the bioinformatics pipelines also varied, including databases such as ARGANNOT20, CARD16, ResFinder1, Resqu12, ARDB13, custom-generated from Genbank sequences18,22 or combinations of these14. Interestingly, the choice of the database was shown to greatly influence the interpretation of risk associated with AMR in public health12,24. Even individual steps, such as mapping sequenced reads to a reference, can be done with different tools, each carrying their own compromises (see 25–28).

This complex - and dynamic - reality poses a challenge for the implementation of bioinformatics pipelines in regulatory settings, where the demonstration of reliability and reproducibility is crucial (see also 11,29). Harmonisation approaches must face the variability described above in terms of technologies, strategies, and software used, each with their demonstrated success, limitations and caveats. A further factor influencing the complexity of applying a given bioinformatics pipeline is that new versions of the individual tools that perform tasks such as quality-checking, trimming or assembling the reads, are constantly being released, which may have unanticipated impacts on pipeline performance. Ready-made and/or commercially available solutions that aim to facilitate the implementation of a NGS-based pipeline by lowering the technical skill required (see, for example, 30,31) face the attendant “black-box” issues when proposed for regulatory purposes.

In response to this complex state-of-the-art and the fast-moving environment in which these technologies are developing, efforts for the standardisation and development of best practices have avoided the prescription of restrictive guidelines, methods or technologies in favour of a more flexible approach emphasising quality metrics and fitness-for-purpose12,13. For bioinformatics pipelines, the development of benchmark resources would play an important role in validating specific bioinformatics strategies and workflows, testing any update to the software underlying an established pipeline or allowing proficiency testing of individual laboratories13–15. These resources would need to include a set of inputs for the bioinformatics pipelines (“in silico reference materials”) linked to a “correct” expected output, as well as consideration for the minimum performance requirements to be met by the pipelines. Different initiatives are ongoing to develop these benchmarking resources including, for example, the Critical Assessment of Metagenome Interpretation (CAMI) project for the evaluation methods for metagenome analysis16,17.

On the 5th of October 2017, the Joint Research Centre invited experts in the field of AMR monitoring in order to discuss the challenges involved in the development of such a benchmark strategy, for the specific purpose of evaluating the bioinformatics pipelines that transform a set of NGS reads to a characterised AMR profile. The experts were invited based on their recent publications on the use of NGS to detect the genetic determinants of AMR in diverse fields: human and veterinary health, the food chain and the environment. The conclusions of these discussions are summarised in Table 1, and discussed in this document.

2. The challenges
Although some of the challenges considered reflect the reality of NGS technologies in general, efforts were made to highlight the issues that are specific to the identification of AMR determinants. Broadly, the challenges can be grouped in different, often overlapping categories.

2.1. Nature of the benchmark datasets
How should a benchmark strategy handle the current and expanding universe of NGS platforms? What should be the quality profile (in terms of read length, error rate, etc.) of “in silico reference materials”? Should different sets of reference materials be produced for each platform? In that case, how to ensure no bias is introduced in the process?

As described in the Introduction, different NGS technology platforms exist for the generation of sequence data serving as inputs for the bioinformatics processes used in the analysis of AMR determinants. Moreover, the technology continues to evolve rapidly with the advent of what is now termed “third generation sequencing” methods that can read the nucleotide sequences at the level of single molecules17. Focusing on validating the technology or the instrument itself is therefore not a useful approach to ensure the reliability of the bioinformatics steps, since it can reasonably be expected that sequencing technologies and protocols will undergo many changes over the coming years. Section 862.2265 of the FDA’s Code of Federal Regulations Title 2113 regulates the general use of NGS instruments for clinical use; even when, in this context, devices are cleared as Class II exempt1, laboratories using these instruments must still establish a bioinformatics pipeline for their intended use19. Thus, an effective benchmark strategy will be independent of existing and upcoming NGS technologies, while avoiding any bias that would favour one technology to the detriment of others.

The proprietary nature of the different raw data outputs produced by the various technologies may not be a primary consideration for present purposes since standard file formats exist that can store raw reads and the associated metadata (e.g. QC metrics) produced by the different sequencers. These include FASTQ18 and BAM19, and they have been successfully used

1See, for example, https://www.accessdata.fda.gov/scripts/cdrh/cfdocs/cfRL/rl.cfm?ld=427645
Table 1. Summary of the challenges identified in the generation of benchmark datasets for the purpose of evaluating the bioinformatics pipelines that process a set of NGS reads into a characterised AMR profile. See text for details.

<table>
<thead>
<tr>
<th>Section</th>
<th>Challenges</th>
<th>Questions to be addressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>Nature of the benchmark datasets - NGS platforms</td>
<td>How should a benchmark strategy handle the current and expanding universe of NGS platforms?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What should be the quality profile (in terms of read length, error rate, etc.) of in silico reference materials?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Should different sets of reference materials be produced for each platform? In that case, how to ensure no bias is introduced in the process?</td>
</tr>
<tr>
<td></td>
<td>Nature of the benchmark datasets - datasets origin</td>
<td>Should in silico reference material be composed of the output of real experiments, or simulated read sets?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>If a combination is used, what is the optimal ratio?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How is it possible to ensure that the simulated output has been simulated “correctly”? For real experiments datasets, how to avoid the presence of sensitive information?</td>
</tr>
<tr>
<td></td>
<td>Nature of the benchmark datasets - quality metrics</td>
<td>Regarding the quality metrics in the benchmark datasets (e.g. error rate, read quality), should these values be fixed for all datasets, or fall within specific ranges?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How wide can/should these ranges be?</td>
</tr>
<tr>
<td>2.2</td>
<td>Samples composition - resistance mechanisms</td>
<td>How should the benchmark manage the different mechanisms by which bacteria acquire resistance?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the set of resistance genes/mechanisms that need to be included in the benchmark?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How should this set be agreed upon?</td>
</tr>
<tr>
<td></td>
<td>Samples composition - bacterial species</td>
<td>Should different sample types (isolated clones, environmental samples, …) be included in the same benchmark?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Is a correct representation of different bacterial species (host genomes) important?</td>
</tr>
<tr>
<td>2.3</td>
<td>Evaluation of pipeline performance - dataset characterisation</td>
<td>How can the “true” value of the samples, against which the pipelines will be evaluated, be guaranteed?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is needed to demonstrate that the original sample has been correctly characterised, in case real experiments are used?</td>
</tr>
<tr>
<td></td>
<td>Evaluation of pipeline performance - performance thresholds</td>
<td>How should the target performance thresholds (e.g. specificity, sensitivity, accuracy, …) for the benchmark suite be set?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>What is the impact of these targets on the required size of the sample set?</td>
</tr>
<tr>
<td>2.4</td>
<td>Generation, distribution and update of the benchmark - future proofing</td>
<td>How can the benchmark stay relevant when new resistance mechanisms are regularly characterised?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How is the continued quality of the benchmark dataset ensured?</td>
</tr>
<tr>
<td></td>
<td>Generation, distribution and update of the benchmark - ownership</td>
<td>Who should generate the benchmark resource?</td>
</tr>
<tr>
<td></td>
<td></td>
<td>How can it be efficiently shared?</td>
</tr>
</tbody>
</table>

in laboratory proficiency testing\(^{38,35,42}\). More recent platforms produce outputs using the HDF5 standard or variants of it; conversion into FASTQ would require an additional computational step, using one of the available tools. However, all platforms (as well as sequencer models and versions within each platform) have differences in the profile and amount of raw reads produced, with variations in their number, length, error rates, error types, etc.\(^{35,44}\). Attempting to create a single set of in silico reference materials would either introduce a bias towards a specific platform and/or create a dataset which is not representative. Creating individual sets of reads would increase the work (with no end in sight as platforms appear or evolve) and require careful consideration to avoid, once again, bias.

All this highlights a clear challenge, which is how to address both the evolution of the platforms, differences amongst instruments and run-to-run variabilities, in view of the need for benchmark datasets serving as the basis for the validation and harmonisation of NGS approaches in clinical and/or regulatory frameworks.

Should in silico reference material be composed of the output of real experiments, or simulated read sets? If a combination is used, what is the optimal ratio? How is it possible to ensure that the simulated output has been simulated “correctly”? For real experiment datasets, how to avoid the presence of sensitive information?

The core component of a benchmark resource is, by definition, a set of inputs representative of what the benchmarked bioinformatics pipeline is expected to receive in normal, real-life use. A logical source for this dataset, then, is the actual
output of laboratory sequencing experiments\textsuperscript{17,34}. However, using data generated by real experiments assumes a high level of quality that will need to somehow be assessed and demonstrated. These experiments will need to be properly characterised in terms of the “true” conclusions the benchmarked pipeline is expected to reach. In addition, although there can be actions taken to ensure that most of the host DNA is filtered from the dataset, real metagenomics experiments from a human source could lead to privacy problems, while samples from food should ensure the absence of information on patented genetically modified food potentially present in the sample\textsuperscript{8,43}. Careful filtering against a standard “exclusion database”, or other adequate strategies, may be necessary to solve this issue. For example, de-identification processes for human DNA sequences have been proposed for clinical datasets. However, the impact of applying tools that modify in any way the reads in a metagenomics dataset should be well understood, as there is a risk that the filtered dataset is no longer representative of a real experiment. Experimental data could also be generated using pure cultures of bacteria present as well-characterised strains in biorepositories (see, for example,\textsuperscript{46})

These concerns could be addressed by \textit{in silico}-generated datasets, where the exact quantity of reads and genes from each source in the composite dataset can be better controlled. Many tools have been developed for this purpose, simulating reads from the different available platforms (see, for example\textsuperscript{47–51}). Once again, it will be important to properly understand these tools, agree on their applicability for the purpose of generating the desired benchmark datasets, and correctly set their parameters so that the resulting simulations are a correct representation of the “real” samples.

\textbf{Regarding the quality metrics in the benchmark datasets (e.g. error rate, read quality), should these values be fixed for all datasets, or fall within specific ranges? How wide can/should these ranges be?}

Available published studies of benchmarking NGS bioinformatics pipelines tend to focus on the performance of specific steps at various levels of input quality and/or complexity (SNP rate, GC content, error rate, quality of the reference sequences, contamination, etc.)\textsuperscript{28,32,53}. This is different from a fit-for-purpose evaluation of a complete pipeline under conditions where the quality of the input is guaranteed through the application of best practices and quality control of the laboratory component of the procedure. An important consideration is the extent to which the benchmark should challenge the pipeline robustness by including varying levels of, for example, error rates or reads quality. It is likely that a pipeline that works best under optimal conditions would be sensitive to variation of the sequencing run quality. The extent of desired variation should be agreed upon and captured in the \textit{in silico} reference material included in the benchmark.

\textbf{2.2. Samples composition}

\textit{How should the benchmark manage the different mechanisms by which bacteria acquire resistance? What is the set of resistance genes/mechanisms that need to be included in the benchmark? How should this set be agreed upon?}

Several mechanisms for the development of resistance to antimicrobials have been characterised\textsuperscript{58}, including: 1) production of an enzyme that digests/metabolizes/escapes the antimicrobial; 2) production of efflux pumps that remove the drug from within the cell; 3) modification, through mutations or biochemical reactions, of the intracellular target of the antimicrobial so that their interaction is lost; 4) activation/upregulation/acquisition of alternate pathways that allow survival by bypassing the pathway disrupted by the antimicrobial; and 5) downregulation of the expression of the pores through which the drug enters the bacteria.

Mechanisms 1), 2) and 4), often involve the acquisition of novel genes by the bacteria from its environment (horizontal transfer) and may be detected, for example, by mapping reads to reference sequence databases that compile such genes. The genetic determinants of mechanisms 3) and 5), however, vary on a case-by-case basis, and may require the detection of Single Nucleotide Polymorphisms (SNPs), insertions/deletions (indels) or variations of copy numbers. These represent different types of bioinformatics determinations which a comprehensive pipeline must be able to resolve, and the benchmark needs to reflect this reality by ensuring that the various types of AMR determinants are correctly represented in the dataset.

Many recent evaluations of the use of NGS for the determination of AMR have emphasised the difficulty of establishing a curated knowledge base of drug resistance genetic determinants to be used as a reference database in NGS data analysis\textsuperscript{21,33}. The same problem is mirrored in the design of a benchmark that would ensure all determinants are correctly detected. It is also of foremost relevance to consider that certain genetic determinants such as efflux pumps (mechanism 2 above) are notorious for giving false positive results, as they perform a variety of export functions not necessarily related to antibiotic resistance (see, for example,\textsuperscript{56}). Eliminating these from the search parameters of bioinformatics pipelines was shown to improve positive predictive value\textsuperscript{57}. The results of testing a pipeline using a benchmark dataset involving all mechanisms must be interpreted with the aim of the pipeline in mind, and this should be taken into account when/if criteria are set (see also section 2.3).

Alternatively, choosing to focus a benchmark dataset on specific resistance mechanisms could simplify the task, but these choices would need to be agreed upon, justified and the limitations clearly stated. This reflection is to be linked to ongoing extensive discussions on the generation of appropriate databases of resistance genes and correct interpretation of resistome profiles (see\textsuperscript{24,58}). An \textit{a priori} statement can be made that the benchmark dataset should focus on mechanisms of \textit{acquired} bacterial resistance. Similarly, for lack of being exhaustive in terms of the AMR genetic determinants it includes, a set of \textit{in silico} reference materials can be composed of the resistance mechanisms most relevant for public and environmental safety, for example, focusing on certain specific plasmids and AMR
genes which have been identified as being important in clinical infections.

The decisions through which specific resistance mechanisms are included/excluded from the benchmark should be clear, transparent, agreed upon and justified in order to ensure that the benchmark is relevant to the types of risks considered. These will vary depending on the purpose of the experiment (e.g. environmental risk, clinical decision making), and will be important to evaluate whether the same resistance factors can be incorporated into a single benchmark dataset or if different resources will be needed. Transparency is important since these decisions will also guide the inclusion/exclusion of novel resistance mechanisms according to the changing epidemiology over time (see also section 2.4).

Should datasets representing different sample types (e.g. isolated clones, environmental samples) be included in the same benchmark? Is a correct representation of different bacterial species (host genomes) important?

The preceding section focused on the nature of the genetic determinants to be included in the in silico datasets. These sequences (i.e. AMR genes), however, represent a very small fraction of the overall totality of the sequence data generated from biological materials (e.g. bacterial genomes) in a given experiment. The nature of these majority “background” reads (bacterial host genomes, other contaminants in the sample etc.) in the components of a proper benchmark dataset thus needs to be carefully considered, as they can influence the accuracy of the pipelines.

The detection of drug resistance in clinical settings is often performed by sequencing pure cultured isolates. Pathogens of particular concern in the context of nosocomial infections will, accordingly, need to be properly represented in the in silico datasets. Lists of AMR pathogens presenting significant risks are maintained (see 61) and include the ESKAPE pathogens (Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumanii, Pseudomonas aeruginosa and Enterobacter sp.) and Escherichia coli, among others.

Culture-dependent methods cannot be systematically applied to environmental samples for various reasons, including the fact that most environment bacteria are not recovered under standard culture conditions. Culture-independent approaches (metagenomics) can then be used to analyse the human and environmental resistomes within complex bacterial populations. These approaches have also been proposed for clinical purposes, greatly reducing the time necessary for characterisation. For these samples, agreeing on a realistic genetic diversity within a benchmark - a set of communities which can be considered “representative” - is a significant challenge as there is tremendous variability in the species composing the microbiomes of different communities.

2.3. Evaluation of pipeline performance

How can the “true” value of the samples, against which the pipelines will be evaluated, be guaranteed? What is needed to demonstrate that the original sample has been correctly characterised, in case real experiments are used?

One of the objectives of validating a bioinformatics pipeline is to demonstrate that its accuracy is above an acceptable value, with low instances of false negative and false positive results. Antimicrobial susceptibility testing using traditional methods is, in itself, a complex procedure subject to differences in methodologies and interpretations; hence they have required (and will require) validation and standardisation. There have been reports where discrepancies between NGS-based predictions and susceptibility testing were caused by isolates with inhibition zones close to the susceptibility breakpoint. It was suggested that the results could have been concordant if the susceptibility testing had been performed under different culture conditions, for example, with a different culture medium. The extent to which these “borderline” cases should be included in the benchmark or not, and the final “correct” prediction that will be attached to them will need to be carefully considered. It should also be discussed what the most relevant endpoint in this context is, between, for example, the Minimum Inhibitory Concentration (MIC) prediction and resistance levels above wildtype/type strain.

The realities of veterinary medicine, with specific modalities of antimicrobial administration, mean that susceptibility MIC breakpoints may differ between humans and animals. Thus, the definition of science-based clinical MIC-breakpoints (CBPs) is relevant to interpret results and to harmonise the results of antimicrobial susceptibility testing of veterinary pathogens. Currently, this issue is being discussed in different working groups led by VETCAST. This may cause difficulties in assigning a universal “correct” label to some datasets that would apply to both humans and animals.

Reference samples of metagenomics experiments are even more complex in this regard, with each sample containing numerous instances of genetic AMR determinants. Metagenomics analyses can detect genes (genotype), which are not necessarily translated into resistance (phenotype); expression of the protein(s), which is not directly revealed by DNA sequencing, is important in this context. Assigning accurate profiles to components of a reference dataset will be challenging, as there is no existing pipeline recognised as the ‘gold standard’ to do so. Spiked samples or simulated reads may be a necessary initial step in this context.

Ultimately, the “true” values to be assigned to the samples in the dataset, and the challenges this will pose, will depend on what the validated pipelines will be required to achieve. For example, the benchmark has to be adaptable to whether the aim is detecting the genetic determinants of AMR, predicting AMR or (for human and veterinary health) predicting antimicrobial susceptibility and thus treatment outcomes.

How should the target performance thresholds (e.g. specificity, sensitivity, accuracy) for the benchmark suite be set? What is the impact of these targets on the required size of the sample set?
Validation of a process involves the determination of various performance parameters, such as specificity, sensitivity, accuracy, etc. When used specifically for the detection of antimicrobial resistance the benchmark resources need to include strict performance thresholds, and whether these should be set a priori along with the levels of these thresholds are subjects for consideration. One also needs to clarify how the process can cope with cases where more than one type of resistance needs to be identified in a single sample, in particular for metagenomics studies.

These performance parameters will be important, not only as information to be included in the benchmark, but also because they generally have a significant influence on the size of the in silico dataset needed (see, for example, 76,77). Understanding the target performance characteristics of a valid pipeline will be necessary to guide decisions as to how many samples will be needed in the in silico dataset, with respect to the presence or absence of AMR genetic determinants. Finally, not all parameters are equally important for all samples - for example, considerations of sensitivity are generally not relevant in the case of cultured isolates as the bacteria are present in high numbers, but may be crucial for metagenomics experiments where the proportion of the target(s) relative to the background is variable and unknown. Targeted metagenomics seem promising approaches for the accurate detection of minority genes in complex samples1, and challenging the sensitivity of bioinformatics pipelines with a benchmark dataset would be of added value in this context.

2.4. Generation, distribution and update of the benchmark

How can the benchmark stay relevant when new resistance mechanisms are regularly characterised? How is the continued quality of the benchmark dataset ensured?

An important fact concerning antimicrobial resistance - and one of the reasons it represents a global health emergency - is that novel mechanisms of resistance are constantly being reported and new genes and/or vectors of transmission regularly emerge86,78. Assuming that a benchmark resource can be produced covering the existing complexity of AMR determinants (section 2.2), adapting this resource to new information is a challenge that will need to be addressed in order to ensure that its utility does not diminish with time. Criteria for inclusion of new in silico datasets, and the mechanisms by which these decisions should be taken, need to be discussed and agreed upon when developing the resource.

Newly identified genetic determinants can also impact the information linked to existing datasets in the benchmark resources. These datasets will need to be re-evaluated in view of new information to ensure that their AMR determinants are properly characterised. As an example, this issue was evidenced in 2015 with the identification of mcr-1 as a plasmid-borne colistin resistance gene9; re-analysis of existing NGS data from E.coli isolates from food, feed and hospitalised patients for the previous years in Denmark revealed previously characterised samples containing this gene30,31.

Who should generate the benchmark resource? How can it be efficiently shared?

Current guidelines and recommendations place the responsibility of validating the bioinformatics pipelines (and ensuring reliability after update of any of its components) with the operator/quality manager of the test facility72,83,84. In fact, thus far, many different sets of benchmark materials and resources have been produced for local use or within collaborative endeavours (see 34). Benchmark datasets have also been used to compare different methods or tools12,22,33. The extent to which these datasets address the concerns described in this document is the subject of a case-by-case evaluation that may become crucial for a wide implementation of NGS technology for routine and regulatory use. An open and inclusive discussion on the different issues (described here or arising upon more detailed considerations) will be important for the development of a resource that can gain wide acceptance and use.

Conclusions

The aim of this document is to summarise a list of challenges that were identified at the meeting organised by the Joint Research Centre on the 4th and 5th of October 2017 for the creation of a benchmark resource. The specific objective of this benchmark would be to challenge the bioinformatics step of a workflow to identify antimicrobial resistance in samples, using NGS technologies. It is clear that this covers only a fraction of the work necessary to fully implement this technology in a regulatory context, which will also need to cover additional steps such as the sampling, library preparation, sequencing run, and interpretation of the AMR profiles (see Figure 1). However, this resource would facilitate the implementation of the NGS technology in routine laboratory analyses by:

- Ensuring confidence in the implementation of the bioinformatics component of the procedure, a step currently identified as limiting in the field22,44,45.
- Allowing evaluation and comparisons of new/existing bioinformatics strategies, resources and tools.
- Contributing to the validation of specific pipelines and the proficiency testing of testing facilities.
- “Future-proofing” bioinformatics pipelines to updates and replacement of the tools and resources used in their different steps.

Some of the challenges in building such a resource are common to all NGS-based methods. Many reports on standardisation, quality management and good laboratory practice have focused on clinical testing and the detection of germline sequence variants linked to cancer or other diseases and could guide some of the decisions to be taken. In this context, reference materials were highlighted as necessary for test validation, QC procedures and proficiency testing80. However, many of the challenges also reflect the reality of antimicrobial resistance monitoring and are specific to this framework. How much of the available resources can be directly applied or used to guide future
efforts in this field will need evaluation and, eventually, complementation.

While the present discussion focuses on the monitoring of bacteria, most of the challenges described herein and the means by which they will be approached should apply to AMR monitoring in other organisms, such as viruses, parasites, and fungi. The differences will be in the final solutions proposed for the composition of the benchmark dataset, due, of course, to the different biology in the mode of action of the antimicrobials and their associated resistance mechanisms.

As it was made apparent in the previous sections, many of the challenges are due to the large heterogeneity behind the reality of detecting AMR using NGS. Some of this heterogeneity will require the development of separate benchmark datasets (e.g. the different sequencing platforms) while some will obviously gain by being combined into a single resource (e.g. human and veterinary medicine). Other cases will require more discussions and evaluations of feasibility/added value in being considered together vs separately (e.g. samples composed of isolates vs metagenomics).

Whatever the final composition and number of the benchmark resource(s), the proper path will ensure a holistic view of the problem that also reflects current public health data. This decision-making process should include expertise in AMR characterisation in humans, animals, food and the environment, in order to maximise its impact on the establishment of an AMR surveillance framework that is in line with the principles of a “One Health” approach.

Disclaimer
The contents of this article are the views of the authors and do not necessarily represent an official position of the European Commission.

Data availability
No data is associated with this article.

Grant information
The “Next-generation sequencing technologies and antimicrobial resistance - Working groups kick-off” meeting (4–5 October 2017) was funded by the European Commission’s Joint Research Centre (JRC), Ispra, Italy.

Acknowledgments
We would like to thank Marc Struelens (European Centre for Disease Prevention and Control, ECDC), Ernesto Liebana Criado and Valentina Rizzi (European Food Safety Authority, EFSA) for their participation to the workshop discussions. We are also grateful to Maddalena Querci and Alex Patak (Joint Research Centre) for their help during the workshop and reviewing the manuscript.

References

5. European Food Safety Authority (EFSA): EFSA’s 20th Scientific Colloquium on Whole Genome Sequencing of food-borne pathogens for public health protection. EFSA Support Publ. 2015; 12(2). Publisher Full Text
The opinion article by Alexandre Angers-Loustau and colleagues addresses an increasingly encountered question by clinical and public health microbiology laboratories - what are the considerations in developing a benchmark for validating methods to identify antimicrobial resistance determinants from high-throughput sequencing data? This is an important step in using sequencing data for this purpose, and this article reports back on the discussions by a panel of experts in the field of antimicrobial resistance surveillance.

In general, the article is well-suited to publication in F1000Research, offering insight into an issue of global importance, and I thank the authors for making the discussions from the October 2017 meeting organised by the Joint Research Centre public. It is well written and explores each of the discussion questions with supporting literature. My suggestions for improvement are minor and addressing them is left up to the authors' discretion.

My overall comment is that while many of the relevant considerations are raised, as a clinician and researcher, I am left with the burden of all these concerns, but no clear direction in how to address the challenges (other than a few comments in the conclusion). I appreciate the purpose of the paper was to summarise the challenges identified at the meeting, but I think this (opinion) article could be enhanced and have a greater impact if more expert opinions and ideas on how to proceed were included.

Minor comments:
Introduction (page 4)
It would be nice to know a little more about the panel of experts - for example, what mix of background or fields did the experts come from e.g. clinical or public health microbiology, bioinformaticians, epidemiologists, veterinary microbiology etc.?

Section 2.1: real experiments vs simulated read sets (page 5)
This section highlights some important concerns (e.g. filtering host DNA or patented genetically modified food DNA), but these are more relevant to identification of antimicrobial resistance from metagenomic sequencing data - this is alluded to in the text, but not explicitly stated. Given the other technical challenges with this, I wonder if the discussion should focus more on the "real experimental" data from cultured bacterial organisms – the more likely initial validation before validation proceeds to metagenomic datasets.

Section 2.2: how should the set of resistance genes/mechanisms be agreed upon (page 6)
The considerations here are well explored, but I wonder if there could be some mention of variation in benchmark datasets depending on the purpose of resistance gene/mechanism detection. For example, detection of AMR genes from environmental or livestock samples may have a different focus, and thus different benchmark targets to detection of AMR genes in organisms from human samples. Even among human samples, there may be differences in the mechanisms prioritised for detection between testing for clinical treatment and testing for epidemiological surveillance (mcr-1 is a good example).

Section 2.3: how can the "true" value of the samples be guaranteed (page 7)
Great section and discussion. Along the lines of the discussion, I am interested to hear the authors' thoughts on whether it is sufficient to validate NGS for predicting AMR against standard susceptibility testing (e.g. MIC), or whether clinical studies evaluating treatment outcomes based on NGS detection are required (i.e. treating based on genotype vs phenotype). My only comment for this section is that perhaps some of the commentary is more directed at validation of NGS for "identifying antimicrobial resistance" rather than "identifying antimicrobial resistance determinants" as stated in the title - two subtly different targets. Perhaps some of the discussion could be clarified here.

“HTF5 format” (page 4) should actually be “HDF5 format”.

MIC should be spelt out at the first usage - minimum inhibitory concentration.

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: Clinical infectious diseases, microbial pathogen genomics, antimicrobial agents and resistance, clinical metagenomics
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.

Reviewer Report 11 May 2018

https://doi.org/10.5256/f1000research.15794.r33450

© 2018 Palù G et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Enrico Lavezzo
University of Padova, Padova, Italy

Giorgio Palù
Department of Molecular Medicine, University of Padova, Padova, Italy

In this opinion article, Alexandre Angers-Loustau and colleagues address a hot topic in Diagnostic Microbiology regarding the implementation of next generation sequencing platforms for the identification of antimicrobial resistances. In particular, they focus on a specific step of the long pipeline which goes from sample collection to data interpretation and management, namely the bioinformatics analysis of sequencing data.

It's been more than 10 years since the introduction of the first NGS instrument, but an adequate standardization and harmonization level of the many different platforms is not yet accomplished.

The article is very well written; it is concise but covers all the important issues that need to be considered when willing to implement such technologies for the purpose of AMR detection.

I only have some minor points for the authors:

1. In section 2.1, page 5, regarding the potential presence of sensitive information (e.g. reads of human origin) when including real experiment datasets in the benchmark. I don't see the problem here. How the availability of anonymous DNA sequence information could compromise the patient's privacy? Of course, patient's identity must be separated from sample's ID, but the de-identification process is a standard procedure and others methods have also been developed for this purpose (see for example this paper from Qamar et al. (2014)¹). Please discuss about this.

2. Since this is an opinion article, I would like the authors to 'expose' themselves a little bit more on some points. The challenges are presented by explaining the state-of-the-art, but it is not easy to me to understand authors' thinking about the possible solutions. Obviously they can't have answers for all the challenges, but they could express some personal positions (for example, what about "Who should generate the benchmark resource"? To me, there should be a centralized action aimed, as far as possible, at the definition of a global standard (similar for example to the WHO HPV LabNet Proficiency Studies for Evaluating HPV DNA Typing Methods). What do the authors think about this?
3. Finally, no reference is made to microorganisms other than bacteria, but AMR may be related also to viruses and eukaryotic microorganisms. Are there any difference or peculiarity regarding the bioinformatics analysis of sequencing data coming from them?

References

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?
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:** Microbiology, bioinformatics, NGS data analysis

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.