MICROBIAL GENOMICS

SHORT COMMUNICATION

Matamoros *et al.*, *Microbial Genomics* 2020;6

DOI 10.1099/mgen.0.000342

MICROBIOLOGY SOCIETY

OPEN DATA OPEN MICROBIOLOGY

Accelerating surveillance and research of antimicrobial resistance – an online repository for sharing of antimicrobial susceptibility data associated with whole-genome sequences

Sébastien Matamoros[1,*], Rene S. Hendriksen[2], Bálint Ármin Pataki[3,4], Nima Pakseresht[5], Marc Rossello[5], Nicole Silvester[5], Clara Amid[5], Frank M. Aarestrup[2], Marion Koopmans[6], Guy Cochrane[5], Istvan Csabai[3,4], Ole Lund[2], Constance Schultsz[1,7] and COMPARE ML-AMR Group[8]

Abstract

Antimicrobial resistance (AMR) is an emerging threat to modern medicine. Improved diagnostics and surveillance of resistant bacteria require the development of next-generation analysis tools and collaboration between international partners. Here, we present the 'AMR Data Hub', an online infrastructure for storage and sharing of structured phenotypic AMR data linked to bacterial whole-genome sequences. Leveraging infrastructure built by the European COMPARE Consortium and structured around the European Nucleotide Archive (ENA), the AMR Data Hub already provides an extensive data collection of more than 2500 isolates with linked genome and AMR data. Representing these data in standardized formats, we provide tools for the validation and submission of new data and services supporting search, browse and retrieval. The current collection was created through a collaboration by several partners from the European COMPARE Consortium, demonstrating the capacities and utility of the AMR Data Hub and its associated tools. We anticipate growth of content and offer the hub as a basis for future research into methods to explore and predict AMR.

DATA SUMMARY

1. All scripts, antibiogram templates and tools (including validation and upload) created for this project are deposited in a public GitHub repository for easy access: https://github.com/EBI-COMMUNITY/compare-amr

2. No new sequences were generated specifically for this project. All new antibiograms created for this project have been deposited in the AMR Data Hub. All data used in the AMR Data Hub is publicly available after the completion of the project and can be accessed via the Project-specific pages in the ENA browser. The AMR data containing projects are listed under 'dcc_schubert' here: https://www.ebi.ac.uk/ena/pathogens/datahubs

3. The European Committee on Antimicrobial Susceptibility testing (EUCAST) maintains a database of MIC and zone diameter distributions with their respective ecological cut-off (ECOFF) values for a wide range of micro-organisms and antimicrobial agents. The *E. coli* ciprofloxacin resistance values were used for comparison with values deposited in the AMR Data Hub: https://mic.eucast.org/Eucast2/

4. Project PRJEB14981 conducted by the National Food Institute, Technical University of Denmark describes the genomes

Received 10 September 2019; Accepted 31 January 2020; Published 07 April 2020

Author affiliations: [1]Amsterdam UMC, University of Amsterdam, Department of Medical Microbiology, Amsterdam, The Netherlands; [2]National Food Institute, Technical University of Denmark, Lyngby, Denmark; [3]Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary; [4]Department of Computational Sciences, Wigner Research Centre for Physics of the HAS, Budapest, Hungary; [5]European Molecular Biology Laboratory (EMBL), European Bioinformatics Institute (EBI), Wellcome Genome Campus, Hinxton, Cambridge, UK; [6]Department of Viroscience, Erasmus University Medical Center, Rotterdam, The Netherlands; [7]Amsterdam UMC, University of Amsterdam, Department of Global Health, Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands; [8]See the full list of the COMPARE ML-AMR group members, in acknowledgements.
***Correspondence:** Sébastien Matamoros, sebastien.matamoros@gmail.com

Keywords: antimicrobial resistance; whole-genome sequencing; database; data sharing; surveillance.

Abbreviations: AMR, Antimicrobial resistance; API, application programming interface; AST, antimicrobial susceptibility testing; ECOFF, ecological cut-off; EMBL-EBI, European Molecular Biology Laboratory - European Bioinformatics Institute; ENA, European Nucleotide Archive; EUCAST, European Committee on Antimicrobial Susceptibility testing; INSDC, International Nucleotide Sequence Database Collaboration; MIC, Minimum inhibitory concentration; ML, machine learning; NCBI, National Center for Biotechnology Information; NGS, next-generation sequencing; SNPs, single nucleotide polymorphisms (SNPs); SRA, sequence repository archive.

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table is available with the online version of this article.

and associated antibiograms of two cases of extremely drug-resistant *Salmonella enterica* serovar Senftenberg isolated from patients in Zambia. This project is used as an example of association of bacterial genomes and antibiogram using the AMR Data Hub architecture described in this report: https://www.ebi.ac.uk/ena/browser/view/PRJEB14981

INTRODUCTION

Antimicrobials are widely regarded as one of the major advances in modern medicine [1]. The global emergence, however, of antimicrobial resistance (AMR) threatens the very core of modern medicine with the potential to turn the global population back in time to the pre-antibiotic era in which simple surgical procedures and common infections could have deadly consequences [2].

The decreased costs of next-generation sequencing (NGS) combined with the progress made in big data analysis such as machine learning (ML) represent innovative opportunities to tackle the AMR crisis [3]. Many bacterial phenotypic traits, including AMR, can be directly linked to the presence of genomic determinants such as genes, single nucleotide polymorphisms (SNPs) or transcription promoters, which can be identified using functional genomics approaches on large databases of genomes. AMR mechanisms are often multifactorial, which is a complicating factor for prediction. In order to circumvent the limitations of the traditional AMR gene presence/absence detection, methods using large amounts of data and unguided analysis of the whole bacterial genome are required. In this vein, recent studies have used computational approaches such as ML to predict antimicrobial susceptibility from genomic data or to discover previously unidentified antibiotic resistance determinants [4–6]. Today, the major limitation for such approaches is not the lack of advanced computational methods or hardware resources but the lack of large enough well curated, annotated data sets where phenotypic AMR data and genomes are linked.

Academic research initiatives and public health organisations could benefit from the implementation of online repositories capable of storing large amounts of genome sequence and antimicrobial susceptibility testing (AST) data [7]. For example, the PATRIC database (https://patricbrc.org/ [6]) has been used for the development of AMR prediction algorithms, but is a closed system which does not support direct access and sharing with other major online genome repositories. The NCBI is offering a similar service, linking antibiograms with genomes deposited in the Sequence Repository Archive (SRA), but lacks a pre-publication data sharing mechanism for stakeholders within a given project (https://www.ncbi.nlm.nih.gov/pathogens/isolates#/search/).

Different stakeholders in the AMR field have different requirements regarding the accessibility of AST data. Making optimal use of the opportunities described above requires enabling the global sharing of data, but some institutions are reluctant to immediately make their AST data publicly accessible for privacy, legal or other reasons [8]. National public health institutes would be encouraged to create supranational networks using a standard format to share data for AST result analysis, while academics would find a solution to share post-publication data, encouraging reproducibility and cross-validation experiments, if such a database structure would become available. Thus, we have identified a clear need for a database structure that can support public AST data as well as those data that are to be shared privately for a period of time until publication.

The Horizon 2020 funded EU consortium COMPARE aims at bringing NGS to public health and clinical practice (http://www.compare-europe.eu/). European experts in AMR working within the COMPARE consortium, including the European Bioinformatics Institute (EMBL-EBI), which is part of the International Nucleotide Sequence Database Collaboration (INSDC: http://www.insdc.org/), have deployed the 'Data Hub' system to allow pre-publication sharing of isolate NGS and linked phenotypic AMR data. The Data Hub system provides a platform that allows data providers, such as public health and clinical laboratories, food safety agencies and veterinary institutes, consented sharing and download of genomic and related data sets. Data can either be kept private

Impact Statement

Antimicrobial resistance is recognized as a major threat to modern medicine and needs new and improved methods for detection and surveillance. Whole-genome sequencing and subsequent bioinformatics analysis of resistant isolates are highly effective methods for identification of AMR markers and surveillance of their spread worldwide. However, further development in this field requires large amounts of genomic and phenotypical data, which will need to be stored and shared using reliable tools and infrastructure. In order to solve this issue the EU-H2020 funded COMPARE consortium has initiated the 'AMR Data Hub', an online infrastructure built around the European Nucleotide Archive for storage and sharing of structured phenotypic AMR data linked to bacterial genome sequences. We describe the structure of this repository, its integration within the European Nucleotide Archive, the standardized antibiogram form used to provide AMR metadata to accompany genome sequences, and a collection of tools for preparation, retrieval and visualization of the data. An example of the data contained in the Data Hub is presented and its functionalities are compared with other existing platforms such as the NCBI BioSample database. In particular, this new architecture allows the creation of private Data Hubs that can be used for sharing data among members of a specific project. This is critical when working with sensitive data (patients, public health or surveillance records). The data is finally publicly released after publication of the results.

for a limited time (pre-publication) or released as open-access immediately at the discretion of the data providers [9], but regardless of the data status (pre-publication private or public) all data can be linked to a Data Hub where it can be accessed by consumers who have been given consent by data providers. The data can be downloaded and analysed by the stakeholders or run through analysis pipelines that are available through the COMPARE platform [9] that is linked to the Data Hubs, and finally analysis results can also be fed back into the Data Hubs and be made available to users of the Data Hub of interest. Novel software was developed for use in the AMR Data Hub that validates the conformity of submitted datasets. The system supports both qualitative and quantitative AST data such as those resulting from disk diffusion and micro-broth dilution tests. A system of Jupyter Notebooks allows for fast browsing and overview of the Data Hubs content.

Here, we present the AMR Data Hub, which permits sharing of large amounts of information that could be used for ML and other data analysis approaches, eventually resulting in accurate and quantitative, hence clinically relevant, predictions of AMR phenotypes based on NGS data.

THE AMR DATA HUB

The Data Hub system

The COMPARE Data Hub system has been built as a broad infrastructure for the sharing and analysis of pathogen NGS data and related data types. Here is presented a brief outline of the relevant parts of the system while a full description is provided in Amid *et al.*[9]. The Data Hubs are provided upon the foundation of the European Nucleotide Archive (ENA: https://www.ebi.ac.uk/ena), an open repository for sequence and related data [10]. The concept was developed and introduced as a model for rapid sharing of data and analysis outputs in public and pre-publication confidential status within the COMPARE consortium. Data providers can choose to use an existing Data Hub or request the creation of a new one from the ENA, in which case they will also decide who are the authorized members of the project. Each Data Hub is configured for a specific project and its access is restricted (by login and password) to the authorized members of the project. These members can link data (sequence and metadata) previously deposited in the ENA database to the Data Hub. The data can be pre-publication confidential but accessed through a Data Hub by authorized users only and/or be made publicly available (i.e. accessible by any user of the ENA). In case of private data linked to a Data Hub, it will be shared with all members of the project with authorized access but will remain hidden from other ENA users. Ultimately, all data archived in Data Hubs are released into the public domain (through the ENA) after a period defined by data owners, and at the latest at the time of publication in a manuscript. Data and metadata reported by data providers are submitted to the hubs through systematic processes supported by a number of tools. Subsequently, structured and accessioned data/metadata are available for sharing between data consumers who have received consent from data providers. The ENA system then serves at the same time as a database for sequence data and metadata and a repository for storage and browsing of results which can be queried through the Pathogen portal (https://www.ebi.ac.uk/ena/pathogens/home).

The antibiogram

In order to represent AST data, a new analysis type, 'antibiogram', has been added to cater for this new ENA database object. This data type leverages the extensible 'analysis object' system, with the addition of a new class specifically for the storage of phenotypic AMR data, designated 'AMR_ANTIBIOGRAM' and is linked to sequence data as a metadata object. Antibiograms are treated as data objects within the system and are supported in data submission and access services. They can easily be associated with a Data Hub for specific projects. As a new data type, building this support has required the development of open software that is distributed publicly (https://github.com/EBI-COMMUNITY/compare-amr) and used internally at EMBL-EBI for the validation and submission of incoming AST data.

We have aimed with the antibiogram for a format that is flexible, complete and compatible with the NCBI equivalent, to allow for a future comprehensive antibiogram data exchange across INSDC. Minimum requirements include, for each combination of isolate/ antibiotic provided, INSDC Sample accession (SAM*); species; antibiotic name; antibiotic susceptibility testing standard; breakpoint version; antibiotic susceptibility test method; measurement; measurement units; measurement sign; susceptibility phenotype; and test platform. Any combination of bacterial species and antimicrobial is supported. Reported antimicrobial susceptibility can be measured by micro-broth dilution or zone diameter, all major testing platforms (Sensititre, VITEK and Phoenix) and standards (EUCAST or CLSI) are accepted. More uncommon methods can be added by using the free-text format of these sections. To help data providers, a detailed protocol explaining the preparation of the metadata form is presented on the GitHub repository of the project (https://github.com/EBI-COMMUNITY/compare-amr) along with tools for batch creation of antibiograms from Excel files. An interactive web page allowing the manual creation of antibiograms is in preparation. It will provide an easy alternative for data providers registering a limited number of samples. Finally, a tutorial explaining the steps required to retrieve data (genomes and antibiograms) from the datahub is available on the GitHub repository.

Each antibiogram is linked to a bacterial genome within ENA. The association is asserted by linking the analysis object, i.e. the antibiogram, with the corresponding study, example: https://www.ebi.ac.uk/ena/browser/view/PRJEB14981. As with other data deposited in the ENA, antibiograms can be kept confidential for a provider-defined period but must ultimately be released at the same time of publishing the associated genomic data into public view. Antibiograms can be queried and retrieved through the AMR Data Hub

	measurement_units	mg.l⁻¹																							Totals
	measurement	0.01	0.012	0.015	0.016	0.023	0.03	0.06	0.12	0.125	0.19	0.25	0.38	0.5	0.75	1.0	2.0	4.0	8.0	12.0	16.0	24.0	32.0	64.0	
antibiotic_name	country																								
ciprofloxacin	Denmark			121			12	1				6													140
	Italy			13																					13
	Netherlands											2				1		1							4
	USA			49			3	1	2			11		2		2		6							76
	United Kingdom			92			4					3							1						100
	Viet Nam	2	1		2	1				6	10	16	5	11	1	5	1	1	2	1		1	45		111
	nan	9		43			6	7	11			22		6		5	3	2	30		26		76	31	277
	Totals	11	1	318	2	1	25	9	13	6	10	60	5	19	1	13	4	10	33	1	26	1	121	31	721

Fig. 1. Visualization of the AST data deposited in the database on 03/12/2019. Data were filtered for: ciprofloxacin (antibiotic_name); *E. coli* (scientific_name) and mg l^{-1} (measurement_units). The numbers in each table cell represent the number of isolates from the given country with the given measurement value. Whole-genome-sequencing data are available for all isolates.

(while confidential) as well as (when made public) through other alternatives, such as the Pathogen Portal (https://www.ebi.ac.uk/ena/pathogens/home), a Discovery Application Programming Interface (API: https://www.ebi.ac.uk/ena/portal/api/), the ENA browser (https://www.ebi.ac.uk/ena) and services providing high-volume data access such as the ENA File Downloader (https://github.com/enasequence/ena-ftp-downloader/). Using the Pathogen Portal or the Portal API, various filters can be used to refine the query such as bacterial species, country of origin, host, and more. An authenticated user can access confidential and public data comprehensively across all authorized Data Hubs and public data not connected to Data Hubs.

Visualization tools

To visualize the contents of the AMR Data Hub, a Notebook was configured (Fig. 1) that has several options for comparison of a defined set of parameters from the database, such as the distribution of the MIC of different antimicrobials as a function of the country of origin, or the comparison of MIC distributions between different antimicrobials. As such, this functionality can be used for surveillance purposes, providing a rapid overview of MIC distribution for a specific collection of isolates and how this compares to isolates from the same host or from different geographical regions.

The Notebook is integrated into the Pathogen Portal and can be accessed from https://www.ebi.ac.uk/ena/pathogens/home. While analysing the data hosted in the AMR Data Hub, access to private data sets as well access to the Notebook results requires authentication via login and password with authorization to the corresponding project. In the meantime, all genomic data, associated AST and visualization results are publicly available via the Pathogen Portal under the 'Explore' tab. While the Notebook report to dcc_schubert (name of the AMR Data Hub), with views such as Fig. 1 and more, is directly available from the above link, the genomic data and associated AST can be accessed either via the respective project pages in ENA (see dcc_schubert: https://www.ebi.ac.uk/ena/pathogens/datahubs), or by using the 'Search' option in the Pathogen Portal (https://www.ebi.ac.uk/ena/pathogens/search), selecting 'Data type: analysis' and tailoring the search to find all analysis records of type 'AMR_ANTIBIOGRAM'.

CURRENT CONTENT

As of 03-12-2019 the AMR Data Hub contains 824 *Escherichia coli* genomes with attached AST data, 1839 *Salmonella enterica* and 29 *Enterococcus faecium* originating from ten different countries. Data on susceptibility against 59 different antibiotics (or combinations) have been entered in the database so far. As an example, 824 *E. coli* isolates originating from seven different countries (Bangladesh; Denmark; Italy; Netherlands; UK; USA; Japan; and Vietnam) have been tested for ciprofloxacin susceptibility, 721 by various dilution-based methods and 103 by disk diffusion. All the data listed here were shared privately by COMPARE partners or were publicly available in the ENA.

A total of 471 *E. coli* antibiograms were submitted directly by COMPARE partners, 247 by an external partner (Toho University School of Medicine, Japan) while 106 were imported from the US CDC database (n=30) or previous publications (n=76) [11, 12].

In order to accurately predict MICs from genomic data, the database used for machine-learning training should represent as accurately as possible the distribution of susceptibility levels expected in clinically relevant bacterial isolates. The EUCAST database (https://mic.eucast.org/Eucast2/) comprises more than 15000 ciprofloxacin MIC values for *E. coli*. We compared the distribution of MICs in the EUCAST database and AMR Data Hub (Fig. 2, Table S1, available in the online version of this article).

In the AMR Data Hub set three distinct peaks can be observed at 0.016, 0.25 and 32 mg l^{-1}. The distribution of MICs in the EUCAST database is slightly different, with a less pronounced peak at 0.25 mg l^{-1} (3.6% of isolates at 0.25 mg l^{-1}, 3.4% at 0.125 mg l^{-1} 1.2% at 0.5 mg l^{-1}). Additionally high MIC values seem to be more abundant in the AMR Data Hub and low MICs in the EUCAST database. In particular no values below 0.016 mg l^{-1} are recorded in the AMR Data Hub, probably reflecting the fact that the genomes of highly susceptible

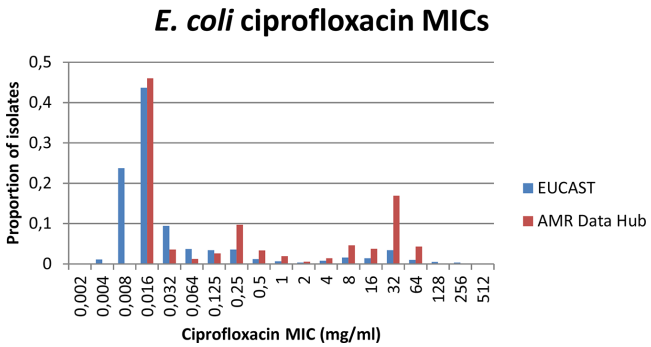

Fig. 2. Proportion of MICs in mg l^{-1} for ciprofloxacin in *E. coli* in EUCAST and AMR Data Hub databases. See Table S1 for the absolute number of isolates.

isolates are rarely sequenced. Similarly, no values above 64 mg l^{-1} are recorded. A Kolmogorov–Smirnov test showed that the distributions are indeed not similar (P<0.01). More effort should be directed towards increasing the availability of whole-genome sequencing of highly susceptible and to a lesser extent highly resistant *E. coli* isolates.

DISCUSSION

We have built the AMR Data Hub with the aim to provide a system for public health, food and veterinary institutes, clinical laboratories and researchers to share their genomic and related AST data. It can be used for standardized open-access data sharing, for example for published data, thus creating an ever-growing source of AST metadata available to researchers worldwide. The large volume of data made available will make it easier to use advanced statistical methods such as machine-learning to predict AMR phenotypes from genomic data and discover new AMR determinants as was recently demonstrated [13].

It has been recently underlined that application of the Nagoya Protocol, which regulates material and data sharing, to genetic information might threaten the timely sharing of data in times of public health emergencies [14]. In allowing the organization and sharing of linked genomics and AST data, the AMR Data Hub promotes openness and accessibility for these important data types while at the same time meeting the privacy concerns for pre-publication data. Considering the exponential rise in the number of bacterial genomes available, and the threat to modern medicine represented by the rise of AMR, the establishment of an AMR Data Hub that can address these privacy concerns, represents a timely effort to improve collaboration in this field.

The design of a standard data submission format benefited from the expertise of the COMPARE consortium, a group of international experts in bacterial genomics and AMR surveillance and research. It is designed to be as exhaustive, and at the same time as flexible as possible to ensure easy sharing of AST data. The database is hosted at the EMBL-EBI, ensuring its connection to the world's largest online repository of bacterial genomes. This system is also part of the ELIXIR infrastructure for sustainable maintenance in the future [15].

As members of the INSDC, EMBL-EBI and NCBI are part of a joined effort for standardization and sharing of genomic data. The NCBI can also host antibiograms in a similar format to that presented here, and efforts are currently ongoing to allow automated synchronization of content from both sides. This will greatly increase the flexibility and the reach of the AMR Data Hub. At the time of writing (3 December 2019) the NCBI BioSample collection contains 12234 genomes with associated antibiograms compared to 2695 in the EBI AMR Data Hub. However, while NCBI does not provide a platform for pre-publication data sharing, the COMPARE Data Hubs [9] hosted at the EMBL-EBI support this model as a mean for rapid data sharing and analysis prior to publication. Participating institutions can thus choose whether they want their data to be open-access immediately or whether they prefer sharing it with selected members of a consortium before public release. This service is unique and could help data providers from public health and clinical laboratories, food safety agencies and veterinary institutes to overcome barriers of data sharing

The view is that the AMR Data Hub will soon become an essential resource for functional genomic studies of AMR. By encouraging data providers from different fields and geographical origins to share their data, this collection can greatly improve our ability to answer questions related to the current AMR crisis.

Funding information
This study was supported by the COMPARE Consortium, which has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 643476.

Acknowledgements
COMPARE contributing institutions for AST data: Amsterdam UMC(1), APHA(11), DTU(2), FM(7), OUCRU(13), SSH(12), UNIBO(10), Toho University(16). The COMPARE ML-AMR group: S. Matamoros(1); V. Janes(1); R. S. Hendriksen(2); O. Lund(2); P. Clausen(2); Frank M. Aarestrup(2); Marion Koopmans(6); B. Pataki(3,4); D. Visontai(3,4); J. Stéger(3,4); JM. Szalai-Gindl(3,4); I. Csabai(3,4); N. Pakseresht(5); M. Rossello(5); N. Silvester(5); C. Amid(5); G. Cochrane(5); Marion Koopmans(6); C. Schultsz(1,7), F. Pradel(8); E. Westeel(8); S. Fuchs(9); S. Malhotra Kumar(10); B. Britto Xavier(10); M. Nguyen Ngoc(10); D. Remondini(11); E. Giampieri(11); F. Pasquali(12); L. Petrovska(13); D. Ajayi(13); E. M. Nielsen(14); N. V. Trung(15); N. T. Hoa(15); Yoshikazu Ishii(16); Kotaro Aoki(16); P. McDermott(17). (1) Amsterdam UMC, University of Amsterdam, Department of Medical Microbiology, Amsterdam, The Netherlands. (2) National Food Institute, Technical University of Denmark, Lyngby, Denmark. (3) Department of Physics of Complex Systems, ELTE Eötvös Loránd University, Budapest, Hungary. (4) Department of Computational Sciences, Wigner Research Centre for Physics of the HAS, Budapest, Hungary. (5) European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK. (6) Department of Viroscience, Erasmus University Medical Center, Rotterdam, The Netherlands. (7) Amsterdam UMC, University of Amsterdam, Department of Global Health, Amsterdam Institute for Global Health and Development, Amsterdam, The Netherlands. (8) Fondation Mérieux, Lyon, France. (9) Department of Infectious Diseases, Robert Koch Institut, Berlin, Germany. (10) Department of Medical Microbiology, Vaccine and Infectious Disease Institute, Antwerp University, University Hospital Antwerp, Antwerp, Belgium. (11) Department of Physics and Astronomy (DIFA), University of Bologna, Bologna, Italy.

(12) Department of Agricultural and Food Sciences (DISTAL), University of Bologna, Bologna, Italy. (13) Animal and Plant Health Agency, Addlestone, Surrey,United Kingdom. (14) Statens Serum Institut, Denmark. (15) Oxford University Clinical Research Unit, Centre for Tropical Medicine, Ho Chi Minh City, Vietnam (16) Department of Microbiology and Infectious Diseases, Faculty of Medicine, Toho University School of Medicine, 5-21-16 Omorinishi,Ota-ku, Tokyo 143–8540, Japan (17) Food and Drug Administration, Center for Veterinary Medicine, Office of Research, Laurel, MD, USA.

Author contributions

S. Matamoros, R.S. Hendriksen, O. Lund, C. Schultsz, G. Cochrane and I. Csabai designed the project. S. Matamoros and C. Schultsz supervised the execution of the project. B. Pataki, N. Pakseresht, M. Rossello, N. Silvester and C. Amid were involved in development of the Notebooks, writing of the validation software, the creation of the Pathogen Portal and the coordination of the COMPARE Data Hub activities, respectively. R. S. Hendriksen and S. Matamoros coordinated the data sharing. New (unpublished) data was provided by the COMPARE ML- AMR group. M. Koopmans and F. Aarestrup provided supervision through the COMPARE consortium. The draft manuscript was written by S. Matamoros, R. S. Hendriksen, C. Amid and B. Pataki. All authors critically reviewed the manuscript and provided feedback.

Conflicts of interest

The authors declare that there are no conflicts of interest.

Data Bibliography

1. AMR Data Hub antibiogram validation and import tools: https://github.com/EBI-COMMUNITY/compare-amr

2. https://www.ebi.ac.uk/ena/data/view/PRJEB14981

3. European Committee on Antimicrobial Susceptibility testing (EUCAST) database for Antimicrobial wild type distributions of microorganisms; https://mic.eucast.org/Eucast2/

References

1. Aminov RI. A brief history of the antibiotic era: lessons learned and challenges for the future. *Front Microbiol* 2010;1:134.
2. WHO. 2014. Antimicrobial resistance: global report on surveillance [Internet]. http://apps.who.int/iris/handle/10665/112642
3. van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. *Trends Genet* 2014;30:418–426.
4. Kavvas ES, Catoiu E, Mih N, Yurkovich JT, Seif Y *et al*. Machine learning and structural analysis of Mycobacterium tuberculosis pan-genome identifies genetic signatures of antibiotic resistance. *Nat Commun* 2018;9:4306.
5. Niehaus KE, Walker TM, Crook DW, Peto TEA, Clifton DA. Machine learning for the prediction of antibacterial susceptibility in Mycobacterium tuberculosis. *2014 IEEE-EMBS Int Conf Biomed Heal Informatics, BHI 2014* 2014:618–621.
6. Davis JJ, Boisvert S, Brettin T, Kenyon RW, Mao C *et al*. Antimicrobial resistance prediction in PATRIC and RAST. *Sci Rep* 2016;6:1–12.
7. Otto M. Next-Generation sequencing to monitor the spread of antimicrobial resistance. *Genome Med* 2017;9:68.
8. van Panhuis WG, Paul P, Emerson C, Grefenstette J, Wilder R *et al*. A systematic review of barriers to data sharing in public health. *BMC Public Health* 2014;14:1144.
9. Amid C, Pakseresht N, Silvester N, Jayathilaka S, Lund O *et al*. The compare data hubs. *Database* 2019;2019 [Epub ahead of print 01 01 2019].
10. Amid C, Alako BTF, Balavenkataraman Kadhirvelu V, Burdett T, Burgin J *et al*. The European nucleotide Archive in 2019. *Nucleic Acids Res* 2020;48:D70–D76.
11. Tyson GH, McDermott PF, Li C, Chen Y, Tadesse DA *et al*. WGS accurately predicts antimicrobial resistance in Escherichia coli. *J Antimicrob Chemother* 2015;70:2763–2769.
12. Tyson GH, Zhao S, Li C, Ayers S, Sabo JL *et al*. Establishing genotypic cutoff values to measure antimicrobial resistance in Salmonella. *Antimicrob Agents Chemother* 2017;61 [Epub ahead of print 23 02 2017].
13. Kim J, Greenberg DE, Pifer R, Jiang S, Xiao G *et al*. VAMPr: variant mapping and prediction of antibiotic resistance via explainable features and machine learning. *bioRxiv* 2019:537381.
14. Dos S Ribeiro C, Koopmans MP, Haringhuizen GB. Threats to timely sharing of pathogen sequence data. *Science* 2018;362:404–406.
15. Durinx C, McEntyre J, Appel R, Apweiler R, Barlow M *et al*. Identifying ELIXIR core data resources. *F1000Res* 2017;5:2422.