Microbial Bioinformatics Hackathon and Workshop

Virtual event, 11-15 October 2021

JPIAMR Secretariat
Swedish Research Council
Box 1035
SE-101 38 Stockholm

+46 8 546 44 000
www.jpiamr.eu
twitter.com/JPIAMR
facebook.com/JPIAMR

jpiamr
Joint Programming Initiative
on Antimicrobial Resistance

Contents

Microbial Bioinformatics Hackathon 2021

Executive Summary

As part of the JPIAMR Roadmap of Actions 2019-2024, JPIAMR proposed holding and sponsoring hackathon events in order to engage the wider scientific community in the challenges of antimicrobial resistance (AMR). This event was proposed in order to encourage novel solutions, to improve human, animal and environmental health and raise awareness of the challenges of Antimicrobial Resistance.

The Cloud Infrastructure for Microbial Bioinformatics (CLIMB-BIG-DATA) in collaboration with the Public Health Alliance for Genomic Epidemiology (PHA4GE) and the Joint Programming Initiative on Antimicrobial Resistance (JPIAMR) organized the 7th Microbial Bioinformatics Hackathon with a special focus on Antimicrobial Resistance. The event brought together key internationally renowned bioinformaticians to address targeted challenges in bioinformatics and AMR.

The event ran over three afternoons, from the 11th to the 13th of October. Due to high demand for places on the Hackathon, the organizers also ran a workshop event on the afternoon of the 15th of October.

As a result of the Hackathon, several new collaborations have been formed and new bioinformatics tools have been made available. The outputs included a program to take a HGVS variant and create a layman's sentence explaining what is occurring and CUDA application to count k-mers in reads using a Graphics Processing Unit. Outputs from the Hackathon will be uploaded to GitHub (https://github.com/AMR-Hackathon-2021).

Background

A hackathon is usually described as a ‘coding marathon’. Programmers, bioinformaticians and specific subject matter experts meet up to solve a problem (or a set or problems). Hackathons involve creating solutions that can be used or further developed as software solutions.

Hackathons promote collaboration, interdisciplinarity, and often raise awareness of the focus topics. The aim of a hackathon is to generate tools to solve common problems in a collaborative fashion. The events often lead to the development of novel collaborations, expand professional networks and offer professional development opportunities.

Steering committee

The Steering Committee for the Hackathon was composed of:

- Andrew Page, Quadram Institute, UK, representing CLIMB-BD and PHA4GE
- Emma Griffiths, Simon Fraser University, Canada, representing PHA4GE
- Mark Pallen, Quadram Institute, UK, representing CLIMB-BD
- Lisa Marchioretto, Quadram Institute, UK, representing CLIMB-BD
- Finlay Maguire, Dalhousie University, Canada, representing PHA4GE
- Jessica Boname, Medical Research Council, UK, representing JPIAMR and MRC

The Microbial Bioinformatics Hackathon was funded by the JPIAMR. The UKRI funded platform, CLIMB-BIG-DATA hosted the Hackathon, coordinated the event and provided the virtual platform used by participants, with additional input from PHA4GE.

Public Health Alliance for Genomic Epidemiology
jpiamr
Joint Programming Initiative on Antimicrobial Resistance
CLIMB
BIG DATA

Figure 1. Organizers and Sponsors of the Hackathon

The Cloud Infrastructure for Microbial Bioinformatics (CLIMB-BIG-DATA) is a collaboration between Warwick, Birmingham, Cardiff, Swansea, Bath and Leicester Universities, the MRC Unit the Gambia at the London School of Hygiene and Tropical Medicine and the Quadram Institute Bioscience.

The concept underpinning CLIMB-BIG-DATA is to bring bioinformatics capability (computing power, storage and analysis tools) to microbiologists working in UK academia. With improvements in next generation sequencing technologies generating genomic data sets has become easier. Many academics don’t have the access to the resources that they need to perform the required bioinformatics analysis. CLIMB-BIG-DATA provides this resource. CLIMB-BIG-DATA provide cloud-based compute, storage, and analysis tools for microbiologists across the UK, accompanied by a wide range of bioinformatics training activities.

The Public Health Alliance for Genomic Epidemiology (PHA4GE) is a global coalition created with the aim of achieving a rapid global genomic-driven public health response to disease outbreaks. PHA4GE is in part funded through the Bill & Melinda Gates Foundation. PHA4GE is working to establish consensus standards; document and share best practices; improve the availability of critical bioinformatic tools and resources; and advocate for greater openness, interoperability, accessibility and reproducibility in public health microbial bioinformatics.

Participants

The Hackathon was aimed at individuals with prior bioinformatics experience. Participants were expected to know in advance how to use the Linux command line, have a good working knowledge of AMR, and an understanding of genomics.

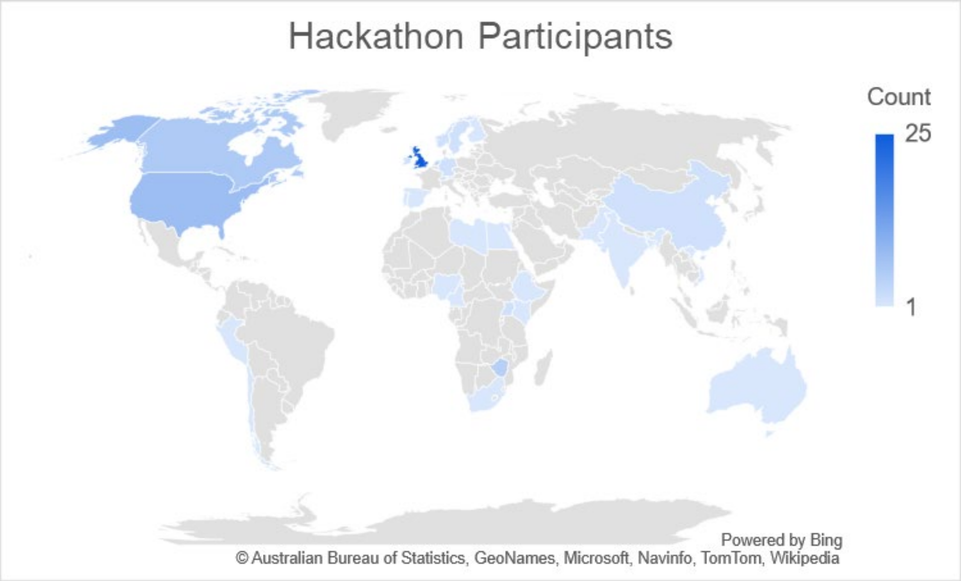

Hackathon Participants
Count
25
1
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Figure 2. Geographic localization of Hackathon participants.

More than 140 researchers applied to take part in the Microbial Bioinformatics Hackathon. A total of 70 individuals from more than 26 countries were then invited to take part in the event (see Figure 2 for the geographical distribution of participants) in addition to invited experts on the steering committee. A total of 65 participants registered on the Discord site on Day 1 of the event, and additional participants registered overnight.

Over half of the applicants stated that this was their first ever Hackathon, and there was a good mix of career stages represented, including:

- 17 Research Scientists
- 13 Post-Docs
- 16 PhD students
- 12 Principal Investigators
- 1 Software Developer

Due to the high number of applications, the Steering Committee made the decision to run an Antimicrobial Resistance Workshop following the Hackathon, to reach out to those applicants who could not be offered a space on the Hackathon.

Aims of the Hackathon

Antimicrobial resistance is a critical universal issue and scientists need reliable, fast, reproducible tools for their research. This hackathon therefore had a special focus on antimicrobial resistance in bacteria. The aim of this hackathon was to improve upon/build and/or extend bioinformatics tools and methods for the AMR research community. The

Hackathon brought together international bioinformatics researchers, scientists and clinicians to collaborate and solve common problems that impact the bacterial AMR community.

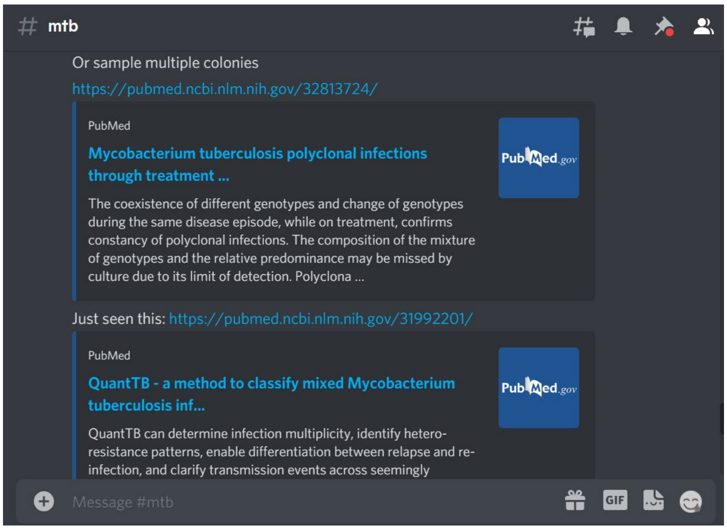
The Microbial Bioinformatics Hackathon Steering Committee suggested several topics of focus for the event. These included:

- SNV (single nucleotide variants) detection standard for AMR
- Alignment of AMR databases (programmatically merging and deduplicating).
- Creation of standardized benchmarking datasets (genomic, metagenomic, assembled, unassembled)
- Integrate hAMRonization (a tool to parse multiple AMR analysis reports into a common data structure) into BioPython (a set of freely available tools for biological computation written in Python) to ease widespread adoption
- Extend open-source AMR bioinformatics tools to improve consistency and interoperability (when the software says 'gene name' it should mean the same thing everywhere)
- GPU (Graphics Processing Unit) enabled AMR calling and analysis

Format of the Hackathon

The Hackathon formally ran over three afternoons. The event began with an introduction to the event, from Emma Griffiths, followed by brief introductions to the sponsors and organizers (Emma Griffiths for PHA4GE, Carolyn Johnson for JPIAMR and Mark Pallen for CLIMB-BIG-DATA), followed by a guide to the format of the event given by Andrew Page and Finlay Maguire. Participants were encouraged to think big but maintain a focus on manageable tasks.

Andrew Page and Finlay Maguire introduced a series of potential topics for groups to focus on, which were replicated in the Discord platform (see Figure 3) used by participants to work collaboratively. Following the introductory session, participants joined the groups that most interested them on Discord and discussed how to create solutions to aid the AMR research community, including useful standardization, ontologies and tools. Groups met up on Zoom periodically throughout the event, to discuss progress and new ideas.



Or sample multiple colonies
https://pubmed.ncbi.nlm.nih.gov/32813724/

PubMed
Mycobacterium tuberculosis polyclonal infections through treatment ...
The coexistence of different genotypes and change of genotypes during the same disease episode, while on treatment, confirms constancy of polyclonal infections. The composition of the mixture of genotypes and the relative predominance may be missed by culture due to its limit of detection. Polyclona ...

Just seen this: https://pubmed.ncbi.nlm.nih.gov/31992201/

PubMed
QuantTB - a method to classify mixed Mycobacterium tuberculosis inf...
QuantTB can determine infection multiplicity, identify hetero-resistance patterns, enable differentiation between relapse and re-infection, and clarify transmission events across seemingly

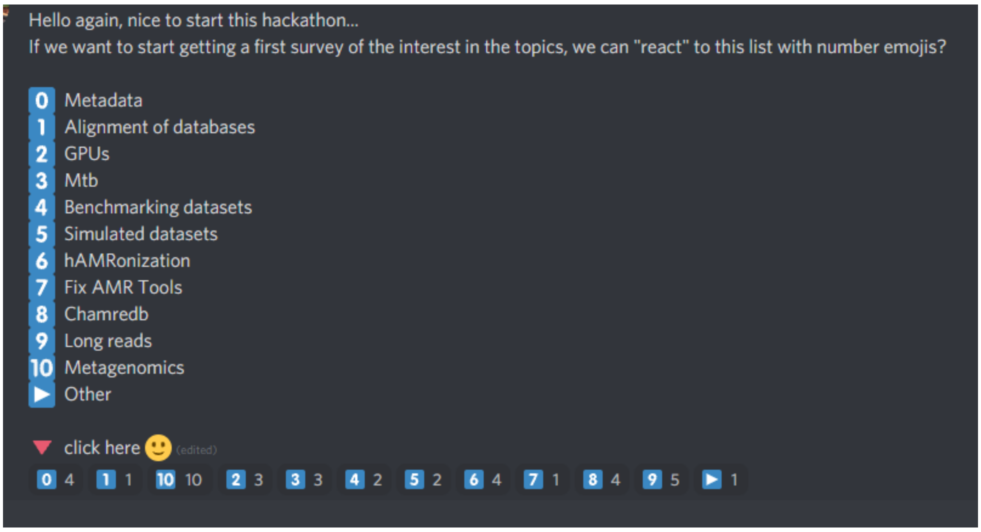
Figure 3. Sharing information over Discord.

The flexibility of the platform allowed participants to continue working on their projects outside of the formal hours of the Hackathon, leading to a more inclusive, accessible event.

CLIMB-BIG-DATA provided two virtual machines, and 5TB of storage, for a three-week period, to allow for projects to be finalized after the end of the event. CLIMB-BIG-DATA also provided two dedicated Cloud Computing Managers to ensure the operability of the Virtual Machines over the three-week period.

Hackathon Topics

An initial list of groups was set out on the Discord platform (see Figure 4). The participants were free to move between groups, participate in multiple groups, suggest merging groups, and suggest their own topics for groups.



Hello again, nice to start this hackathon...
If we want to start getting a first survey of the interest in the topics, we can "react" to this list with number emojis?

0 Metadata
1 Alignment of databases
2 GPUs
3 Mtb
4 Benchmarking datasets
5 Simulated datasets
6 hAMRonization
7 Fix AMR Tools
8 Chamredb
9 Long reads
10 Metagenomics
▶ Other

🔻 click here 🙂 (edited)

0 4 · 1 1 · 10 10 · 2 3 · 3 3 · 4 2 · 5 2 · 6 4 · 7 1 · 8 4 · 9 5 · ▶ 1

Figure 4. Hackathon Topics

Hackathon groups

#GPUs

This grouping discussed the usage of Graphics Processing Units (GPUs) for AMR bioinformatics. Compute Unified Device Architecture (CUDA) is a parallel computing platform and application programming interface (API) that allows software to use graphics processing units for general purpose processing. Each server for the hackathon had access to two Nvidia T4 GPUs, enabling this group to test the tool. The aim of this group was to design a speedy k-mer (a nucleotide sequence of a certain length) counter working from short reads, with a python interface, which can then hopefully be used for several downstream purposes.

#MTB

This pathogen specific grouping (focused on *Mycobacterium tuberculosis*) aimed to produce an SNV variant detection standard for AMR, with a special focus on Mtb; this group decided to work closely with the hAMRonization group, and try to link two existing tools (TB-Profiler and Mykrobe) to the hAMRonization platform (see Figure 5 for more detail on hAMRonization).

#hAMRonization

The focus of this grouping was to enable extended support for hAMRonization, which is a tool that combines the outputs of disparate antimicrobial resistance gene detection tools into a single unified format (see Figure 5 for more detail). The aims of this group were relevant to many topics, including the Long reads topic, Benchmarking and MTB.

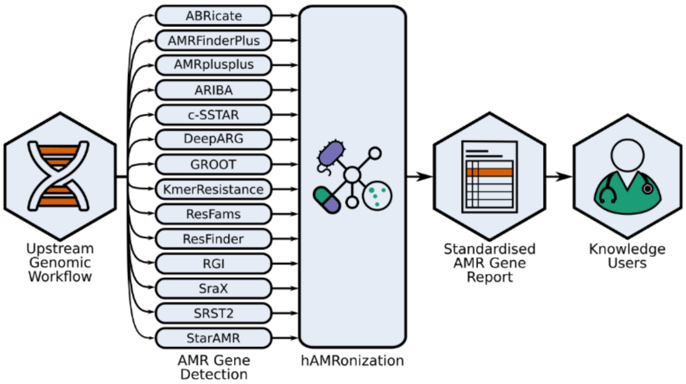


Figure 1. hAMRonization workflow

#Benchmarking datasets and Simulated Datasets

This project was focused on the creation of standardized benchmarking datasets to evaluate the performance of AMR gene detection tools, using assembled and unassembled data. This would allow for the fair evaluation and measurement of bioinformatics pipelines and methods.

This group aimed to create both standard and simulated genomic/metagenomics data (Illumina, PacBio, Nanopore) containing identical genomes of bacteria, but with/without various antimicrobial resistance genes or key SNPs (single nucleotide polymorphisms) in assembled and unassembled genomes. A repository could then be built that could be used to benchmark users own assembly and identification pipelines.

#ChAMReDb

The aim of this group was to extend the ChAMReDb to compare AMR gene's annotation and metadata between CARD, NCBI and ResFinder databases and try to resolve nomenclature differences and data distribution.

#Long reads

Long read sequencing holds the promise of capturing large chunks of a pathogen in single reads, including things like entire mobile genetic elements which are linked to resistance. The only problem is that long reads are more error prone at the base level. This group aimed to build tools that take the error model of long reads into account for faster, more accurate AMR detection (or more).

#Metagenomic

Metagenomics sequencing is being used for AMR surveillance; however, it is particularly difficult to understand which pathogen is linked to which mobile genetic element. This group aimed to look at this problem in more detail and see if bioinformatics methods could help in any way, or at least reduce the size of the challenge. The group focused on training a metagenomic species classifier on the basis of AMR gene distribution using Scagaire (Scagaire allows you to take in gene predictions from a metagenomic sample and filter them by bacterial/pathogenic species) and CARD.

#Metadata

Sequencing combined with high quality metadata makes for exceptionally powerful datasets such as linking phenotypes to genotypes, and source attribution. To support this, well defined (and used) ontologies and ways to share more detailed sets of restricted metadata are required. This group discussed these challenges.

Outcomes of the event

The focus groups came up with six topics to concentrate on- some of these topics were 'cross-groupings'.

- Creation of a GPU-accelerated k-mer counter and FASTA parser for GPU-accelerated AMR genomics
- Extended support for the harmonization of AMR prediction (https://github.com/pha4ge/hAMRonization) and creation of a tool to make HGVS variants simple to understand
- Creation of benchmarking datasets to evaluate the performance of AMR gene detection tools
- Extension of the ChAMRedb to compare AMR gene's annotation and metadata between CARD, NCBI and ResFinder databases
- AMR, MGE and plasmid detection using existing tools on uncorrected long-read data
- Training a metagenomic species classifier on the basis of AMR gene distribution using Scagaire (https://github.com/quadram-institute-bioscience/scagaire) and CARD

Outputs from the Hackathon will be uploaded to GitHub (https://github.com/AMR-Hackathon-2021). GitHub, Inc. is one of the largest providers of Internet hosting for software development, most commonly used to host open-source projects. These included:

- The MTB grouping developed a program to take a HGVS variant and create a layman's sentence explaining what is occurring.
  Link to the tool: https://github.com/conmeehan/laymansHGVS.
- The GPU grouping produced a CUDA application to count k-mers in reads using a GPU. The group developed a user interface for the GPU programme (i.e. a pop up window for people with non-bioinformatic knowledge). The group also wrote out the command line and benchmarked the application against a non-GPU k-mer programme.
  Link to the tool: https://github.com/AMR-Hackathon-2021/gpu_kmer_counter
- The Benchmarking group identified a set of reference genomes as source to generate data sets in order to generate a reference tool to help identify AMR genes and aid identification of the most appropriate antibiotic for treatment.
  Resources: https://github.com/AMR-Hackathon-2021/benchmarking_datasets
- The ChAMRedb group worked on an extension of the ChAMRedb to compare AMR gene's annotation and metadata between CARD, NCBI and ResFinder databases. Resources: https://pypi.org/project/chAMReDb/
- The Long reads group tested three tools to identify AMR genes on Nanopore uncorrected reads.
  Link to outcomes: https://github.com/AMR-Hackathon-2021/long-reads

Participants' Views

The Hackathon was a great success; feedback from participants was overall extremely positive, as can be seen in Figure 6. The majority of applicants noted that the Hackathon was relevant to them, and 95% noted that they would be recommending such events to their colleagues.

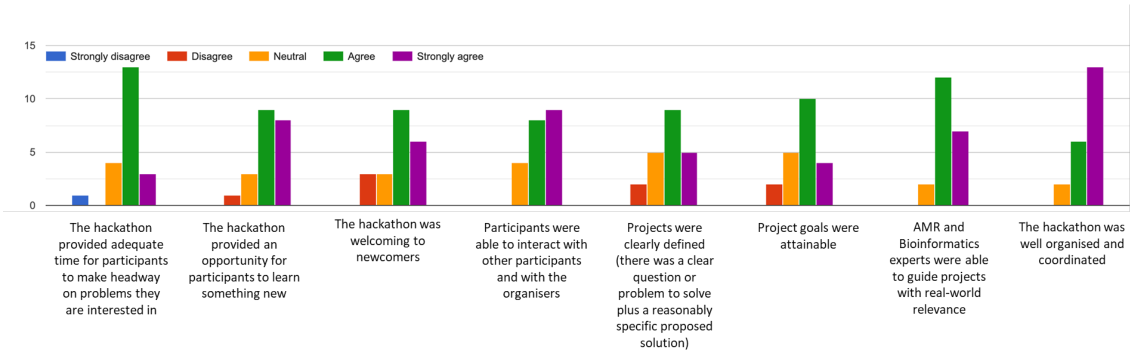


Figure 6. Feedback from the Hackathon

Participants remarked on the usefulness of the Discord platform for collaboration, teamwork and sharing of ideas, however almost all participants suggested that an in-person event would have been even better.

All senior researchers (Independent Researchers) who took part in the event noted that they would be sharing outcomes with colleagues and all were keen to take part in a future event. Most junior researchers (Masters and PhD students) held more divided opinions, but 78% noted that they would be interested in taking part in a similar event, and 93% noted that they would be sharing outcomes with colleagues.

Conclusions

The Hackathon provided a virtual space for groups to tackle important challenges in AMR research, to form new collaborations, and learn new skills. The online format was overall a success and several new tools have been made available as a result of the event. Many of the groups are still working together refining the tools they developed and preparing publications based on the outcomes of the event.

Workshop on Antimicrobial Resistance

Summary

Due to high demand, in addition to the Hackathon, the Steering committee organized a webinar style workshop on Antimicrobial Resistance. The aim of the workshop was to provide a training opportunity for those who were not able to take part in the Hackathon.

The workshop sessions were recorded and the recordings are available online.

The workshop focused on four major themes:

- Use of existing AMR-related databases and resources (including CARD, NCBI, and PATRIC)
- Theory and use of bioinformatics tools to detect AMR genes from genomes (e.g., AMRFinderPlus)
- How to compare and systematically report results from AMR genomics using hAMRonization
- A practical introduction to bioinformatics workflows for AMR genomics

Steering committee

The Steering Committee for the Workshop was composed of:

- Andrew Page, Quadram Institute, UK, representing CLIMB-BD and PHA4GE
- Emma Griffiths, Simon Fraser University, Canada, representing PHA4GE
- Mark Pallen, Quadram Institute, UK, representing CLIMB-BD
- Lisa Marchioretto, Quadram Institute, UK, representing CLIMB-BD
- Finlay Maguire, Dalhousie University, Canada, representing PHA4GE
- Jessica Boname, Medical Research Council, UK, representing JPIAMR and MRC
- Carolyn Johnson, Medical Research Council, UK, representing JPIAMR and MRC

The Microbial Bioinformatics Workshop was hosted by the UKRI funded platform, CLIMB-BIG-DATA. Finlay Maguire hosted the event.

Workshop Participants

The workshop received more than 500 registrations; on the day, 180 people took part; a further 100 have already watched the recordings online. Participants came from as far afield as Sudan and Thailand (see Figure 7 for more detail).

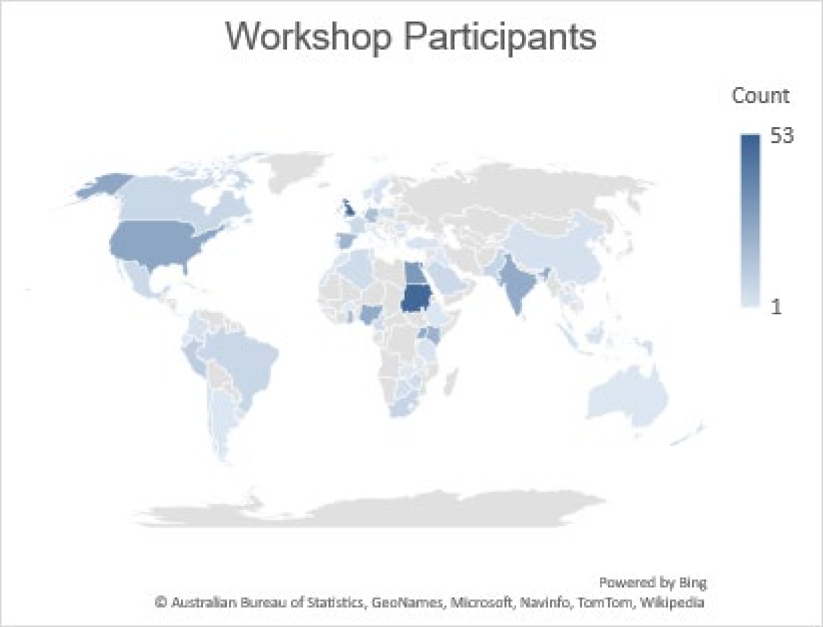

Workshop Participants
Count
53
1
Powered by Bing
© Australian Bureau of Statistics, GeoNames, Microsoft, Navinfo, TomTom, Wikipedia

Figure 7. Geographic localization of participants

Workshop Speakers

The workshop speakers included:

Kara Tsang (London School of Hygiene & Tropical Medicine, UK), who gave an introduction to databases and resources for AMR genomics. This was followed by a talk from Mike Feldgarden (National Center for Biotechnology Information, USA), who gave an overview of AMR gene prediction tools (including detection of variant-related AMR). Ines Mendes (Instituto de Medicina Molecular, PT) then discussed how to compare and report AMR results using hAMRonization.

Finlay Maguire (Dalhousie University, CA) then gave a practical demonstration of how to move from bacterial genomic reads to AMR gene reports.

Panel Discussion

The workshop ended with a panel discussion in which Kara Tsang (London School of Hygiene & Tropical Medicine, UK), Mark Pallen (Quadram Institute, UK), Andrew Page (Quadram Institute, UK), Duncan MacCannell (Center for Disease Control, USA), Ines Mendes (Instituto de Medicina Molecular, PT), Emma Griffiths (Simon Fraser University, CA), Henk den Bakker (University of Georgia, USA), Anthony Underwood (The Centre for Genomic Pathogen Surveillance, UK), Michael Feldgarden (National Center for Biotechnology Information, USA), Lisa Marchioretto (Quadram Institute, UK), Carolyn Johnson (Medical Research Council, UK) and Finlay Maguire (Dalhousie University, CA).

Participants' Views

The majority of participants felt that they had gained both knowledge and skills from the workshop. Ninety five percent of participants felts that the event was extremely relevant. The majority felt that the panellists were well suited to the task and that the event was well organized (see Figure 8).

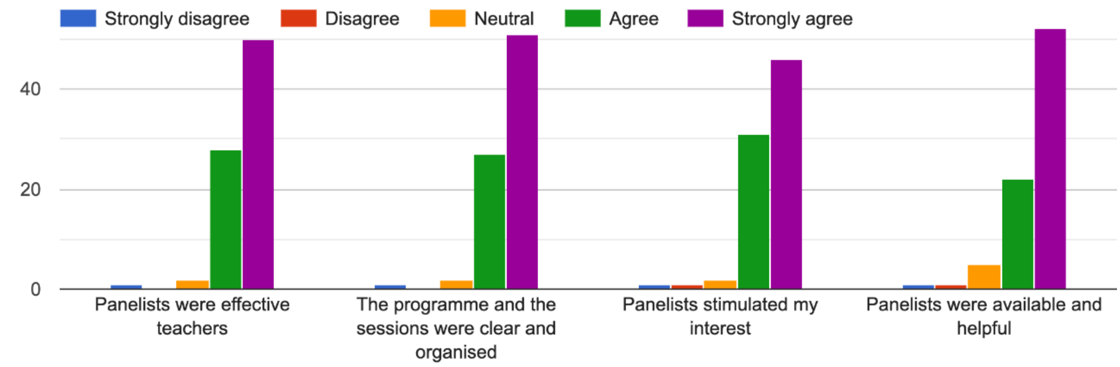


Figure 8: Participants' feedback on the workshop

Participants commented on the usefulness of the practical information provided and several comments that the workshop increased their understanding of the challenges in AMR bioinformatics.

Conclusions

The video recordings of the workshop are now available on the CLIMB-Big-Data Website, and have already received more than 100 views. The Webinar provided attendees with specific examples of the utility of bioinformatics tools for AMR and an opportunity to ask specific questions of an expert panel.

Abbreviations and useful links

Abbreviations

AMR: Antimicrobial Resistance
API: application programming interface
CARD: Comprehensive Antibiotic Resistance Database
CLIMB-BIG-DATA: Cloud Infrastructure for Microbial Bioinformatics
CUDA: Compute Unified Device Architecture
GPU: Graphics Processing Unit
HGVs: Human Genome Variations
JPIAMR: Joint Programming initiative on antimicrobial resistance
MRC: Medical Research Council
MTB: Mycobacterium tuberculosis
NCBI: National Centre for Biotechnology Information
PATRIC: Pathosystems Resource Integration Center
SNP: Single Nucleotide Polymorphism
SNV: Single Nucleotide Variant
PHA4GE: Public Health Alliance for Genomic Epidemiology

Useful links

https://card.mcmaster.ca/ontology/40648
https://cge.cbs.dtu.dk/services/ResFinder/
https://charmed.cgps.group/
https://discord.com/
https://github.com/alexmanuele/arete
https://github.com/ncbi/amr
https://github.com/AMR-Hackathon-2021/gpu_kmer_counter
https://github.com/bactopia/bactopia
https://github.com/pha4ge/hamronization
https://github.com/conmeehan/laymansHGVS
https://github.com/Mykrobe-tools/mykrobe
https://github.com/quadram-institute-bioscience/scagaire
https://github.com/lcerdeira/Spyder
https://github.com/jodyphelan/TBProfiler
https://github.com/tseemann/nullarbor
https://www.jpiamr.eu/
https://www.jpiamr.eu/projects/seq4amr/
https://mrc.ukri.org/
https://www.ncbi.nlm.nih.gov/
https://www.patricbrc.org
https://pha4ge.org/
https://www.protocols.io/workspaces/genometrakr1/publications?sort=views&page_id=2
https://quadram.ac.uk/

Appendix 1: Steering Committee and Workshop Speakers

Steering Committee

Mark Pallen, Quadram Institute, UK,
representing CLIMB-BD

Andrew Page, Quadram Institute, UK,
representing CLIMB-BD and PHA4GE

Emma Griffiths, Simon Fraser University, CA
representing PHA4GE

Carolyn Johnson, Medical research Council, UK
representing JPIAMR and MRC

Lisa Marchioretto, Quadram Institute, UK
representing CLIMB-BD

Finlay Maguire, Dalhousie University, CA
representing PHA4GE

Jessica Boname, Medical research Council, UK
representing JPIAMR and MRC

Workshop Speakers

Kara Tsang, London School of Hygiene & Tropical Medicine, UK

Ines Mendes, Instituto de Medicina Molecular, PT

Mike Feldgarden, National Center for Biotechnology Information, USA

Finlay Maguire, Dalhousie University, CA

With thanks to the wonderful organizing team: Mark Pallen, Finlay Maguire, Andrew Page, Emma Griffiths, Lisa Marchioretto and Jessica Boname.