HavePhAlth: Human phage therapy against the ESKAPE pathogens using AI

A Data Management Plan created using DMPonline.be

Creators: Cédric Lood, Rob Lavigne

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: Cédric Lood, Rob Lavigne

Data Manager: Cédric Lood, Rob Lavigne

Project Administrator: Cédric Lood, Rob Lavigne

Grant number / **URL:** 12D8623N

ID: 198873

Start date: 01-11-2022

End date: 31-10-2025

Project abstract:

Human Phage Therapy (PT) is a promising route for the treatment of drug-resistant bacterial infections. Belgium is currently leading the implementation of PT in Europe, and the technique is currently in operations at the Queen Astrid Military Hospital (QAMH) and at UZ Leuven. A Multidisciplinary Phage Task Force (MPTF) has been set up within UZ Leuven to provide a PT framework for patients with difficult-to-treat infections. However, the current design strategies of phage cocktails are a black box, relying on empirical rules that fail to leverage the rapidly expanding omics data to predict bacteria-phage interactions. This in turn limits the inclusion criteria for patients to receive PT treatment. In my PhD research, I developed machine learning models of phage infectivity in P. aeruginosa that predict which phages from a collection can infect given bacterial strains based on their genomic content. As a member of the MPTF and collaborator of QAMH, two entities that will generate big datasets of omics/clinical data on PT in vivo, I will translate these modeling approaches to the ESKAPE pathogens in patients. This will put me in a unique position to assess the dynamics of bacteria-phage co-evolution in vivo, by inspecting longitudinal isolates from given patients undergoing treatment. Importantly, these analyses will enable us to extract design rules for phage products, while productively translating our research towards "sur-mesure" phage treatment of individual patients.

Last modified: 29-04-2023

HavePhAlth: Human phage therapy against the ESKAPE pathogens using AI Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
UMC host-range experiment	Dataset of 503 sequenced P. aeruginosa isolates, of which 452 have been tested against 8 phages	Reuse existing data	Digital	Experimental	.fastq .fasta .gbk .faa .gff .xlsx .txt	<100 GB	
Pirnay host-range experiment	Dataset of 300 isolates, with sequencing ongoing, which have been tested against 20 phages	Reuse existing data	Digital	Experimental	fastq .fasta .gbk .faa .gff .xlsx .txt	<100 GB	
Pirnay collection	Collection of P. aeruginosa isolates	New	Physical	Experimental			300 eppendorf tubes of 2 ml in the ultrafreezer
Hospital isolates	Collection of pathogens relevant to therapeutic efforts	New	Physical	Experimental			100 eppendorf tubes of 2 ml in the ultrafreezer

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

1. Designation of responsible person (If already designated, please fill in his/her name.)
Rob Lavigne - Head of the Laboratory of Gene Technology
2. Storage capacity/repository
2.1 During the research:
The data will be stored on the KU Leuven network drive available to Rob Lavigne and Cédric Lood.
2.2 After the research:
The data will remain available on the KU Leuven network drive and accessible by Rob Lavigne. All relevant data will be archived there for a minimum of 10 years (RDM policy of KU Leuven). Raw sequencing data that has been published and made available in perpetuity via the INSDC databases may be removed from the drives. (https://www.insdc.org/policy/)

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

We do not wish to deviate from the principle of preservation of data.

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

Patient information will be treated confidentially as we described in the phageforce protocol publication (doi: https://doi.org/10.3390/v13081543 - section 2.7)

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

Not applicable.

HavePhAlth: Human phage therapy against the ESKAPE pathogens using AI
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable

HavePhAlth: Human phage therapy against the ESKAPE pathogens using AI
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

HavePhAlth: Human phage therapy against the ESKAPE pathogens using AI
FWO DMP (Flemish Standard DMP)

1. Research Data Summary

List and describe all datasets or research materials that you plan to generate/collect or reuse during your research project. For each dataset or data type (observational, experimental etc.), provide a short name & description (sufficient for yourself to know what data it is about), indicate whether the data are newly generated/collected or reused, digital or physical, also indicate the type of the data (the kind of content), its technical format (file extension), and an estimate of the upper limit of the volume of the data.

				Only for digital data	Only for digital data	Only for digital data	Only for physical data
Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
		Please choose from the following options: • Generate new data • Reuse existing data	*Please choose from the following options:* • Digital • Physical	*Please choose from the following options:* • Observational • Experimental • Compiled/aggregated data • Simulation data • Software • Other • NA	*Please choose from the following options:* • .por, .xml, .tab, .cvs,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
UMC host-range experiment	Dataset of 503 sequenced P. aeruginosa isolates, of which 452 have been tested against 8 phages	Reuse existing data	Digital	Experimental	.fastq .fasta .gbk .faa .gff .xlsx .txt	<100 GB	
Pirnay host-range experiment	Dataset of 300 isolates, with sequencing ongoing, which have been tested against 20 phages	Reuse existing data	Digital	Experimental	fastq .fasta .gbk .faa .gff .xlsx .txt	<100 GB	
Pirnay collection	Collection of P. aeruginosa isolates	New	Physical	Experimental			300 eppendorf tubes of 2 ml in the ultrafreezer
Hospital isolates	Collection of pathogens relevant to therapeutic efforts	New	Physical	Experimental			100 eppendorf tubes of 2 ml in the ultrafreezer

If you reuse existing data, please specify the source, preferably by using a persistent identifier (e.g. DOI, Handle, URL etc.) per dataset or data type:

UMC host-range experiment: data published as part of the dissertation of Cédric Lood (https://kuleuven.limo.libis.be/discovery/fulldisplay?docid=lirias3630794&context=SearchWebhook&vid=32KUL_KUL:Lirias&search_scope=lirias_profile&tab=LIRIAS&adaptor=SearchWebhook&lang=nl)
Pirnay host-range experiment: data published as part of the dissertation of Cédric Lood (https://kuleuven.limo.libis.be/discovery/fulldisplay?docid=lirias3630794&context=SearchWebhook&vid=32KUL_KUL:Lirias&search_scope=lirias_profile&tab=LIRIAS&adaptor=SearchWebhook&lang=nl)

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes, human subject data

Hospital isolates dataset is linked to the implementation of phage therapy in UZ Leuven. The collection of these isolates for analysis of host-range and genomics purpose is covered under the Ethics Committee Research UZ/KU Leuven (S64854).

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- Yes

For the hospital isolates dataset, I will know in which disease the bacteria were involved, as well as being in a position to tell which isolates were sampled from the same patient (longitudinal series). However, the data will be treated confidentially as described in our consortium publication (doi: https://doi.org/10.3390/v13081543 - section 2.7)

Does your work have potential for commercial valorization (e.g. tech transfer, for example spin-offs, commercial exploitation, …)? If so, please comment per dataset or data type where appropriate.

- No

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material/Data transfer agreements/ research collaboration agreements)? If so, please explain in the comment section to what data they relate and what restrictions are in place.

- Yes

Both the UMC and Pirnay host-range datasets involved collaborations with external groups. In the first case, the group of Pieter-Jan Haas (Utrecht, NL), and in the second case, the groups of Jean-Paul Pirnay at the Queen Astrid Military Hospital (Brussels, BE) and Pieter-Jan Ceyssens at Sciensano (Brussels, BE). These collaborations are scientific in nature and do not prevent publications of the data once the publication stage has been reached.
Finally, the Hospital isolates dataset is being collected as part of an interdisciplinary network project from KU Leuven (IDN/20/024).

Are there any other legal issues, such as intellectual property rights and ownership, to be managed related to the data you (re)use? If so, please explain in the comment section to what data they relate and which restrictions will be asserted.

- No

2. Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keep data understandable and usable, for yourself and others, now and in the future (e.g., in terms of documentation levels and types required, procedures used, Electronic Lab Notebooks, README.txt files, Codebook.tsv etc. where this information is recorded).

Project-related folders are created and their scope is documented through a top-level README file. Each sub-project folders contains its own README file and accompanying office documents describing the protocol followed when appropriate, or the scripts necessary for the data analysis. The programming scripts used to process the sequencing data and proceed with the data analysis are maintained in the folders and documented inline with the code.
Published code will be made available via the git repository of the lab:
https://github.com/LoGT-KULeuven

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- Yes

Genomics data will be made available via the INSDC databases, which have a set of requirements in terms of metadata. We routinely make use of the following NCBI database (member of the INSDC consortium):
- Sequence Read Archive (SRA): https://submit.ncbi.nlm.nih.gov/about/sra/
- Genbank: https://submit.ncbi.nlm.nih.gov/about/genome/
- Gene Expression Omnibus (GEO): https://www.ncbi.nlm.nih.gov/geo/info/submission.html
- BioSample & BioProject: https://submit.ncbi.nlm.nih.gov/about/bioproject-biosample/
Each dataset then is allocated one or multiple accession numbers, which will be recorded in the documentation of the project and related publications.

3. Data storage & back-up during the research project

Where will the data be stored?

The raw sequencing data is stored on the KU Leuven large volume storage drive of the laboratory (K: drive), accessible to all lab members. The data will be duplicated and stored on the workstation of the researcher during their processing, before being transferred to the KU Leuven drive (J: drive) of the laboratory.
Metadata and scripts are created on the workstation of the researcher, and stored on the researcher's folder in the KU Leuven network drive.
Processed sequencing data will be sent to the relevant NCBI database where accession numbers will be available and made public.
Biological data will be stored at -80°C/-20°C

How will the data be backed up?

The KU Leuven network drives are maintained by the central IT service of the university and they ensure proper backup. The researcher is responsible for the backup of the data onto an external drive on his workstation.

Is there currently sufficient storage & backup capacity during the project? If yes, specify concisely.
If no or insufficient storage or backup capacities are available, then explain how this will be taken care of.

- No

Extended storage will be necessary to store the data on the KU Leuven network drives. This is covered by the bench fee of the researcher where we specifically requested and extra 500 Eur/year for 5 years to obtain an extra 1TB of storage on the network drive. Sequencing data will be transferred to the INSDC databases throughout the project and maintained there in perpetuity (https://www.insdc.org/policy/)

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The workstation of the researcher is maintained by the central IT and accessible via central login. The list of people that can access the computer is maintained by IT, and includes the researcher

and the head of the laboratory (Rob Lavigne).
The KU Leuven network drives include shared folders available to all lab members (raw sequencing data), and private folders accessible to the researcher and to the head of the laboratory.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

We have budgeted in the bench fee a total of 500 Eur/Year for 1TB of network drive storage for 6 years.

4. Data preservation after the end of the research project

Which data will be retained for at least five years (or longer, in agreement with other retention policies that are applicable) after the end of the project? In case some data cannot be preserved, clearly state the reasons for this (e.g. legal or contractual restrictions, storage/budget issues, institutional policies...).

All genomics data will be maintained in perpetuity on the INSDC databases. The scripts for data analysis will be maintained in perpetuity on the git repository of the laboratory: https://github.com/LoGT-KULeuven. All datasets and scripts will be additionally available through the RDR repository of KU Leuven and accessible via DOI.

Where will these data be archived (stored and curated for the long-term)?

All sources mentioned above (section 2.2) maintain data in perpetuity.

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

No additional costs are foreseen for the long-term retention of the data.

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Yes, in a restricted access repository (after approval, institutional access only, …)
- Yes, in an Open Access repository

Any published digital dataset will be made available in perpetuity in the KU Leuven Research Data Repository (RDR) with a corresponding DOI accession number.
Digital data awaiting publication will be available through the KU Leuven network drives through which lab members have access (K: drive) or to which Rob Lavigne and Cédric Lood have access (J: drive).

If access is restricted, please specify who will be able to access the data and under what conditions.

Cédric Lood and Rob Lavigne have unrestricted access to the datasets.

Are there any factors that restrict or prevent the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)? Please explain in the comment section per dataset or data type where appropriate.

- Yes, Ethical aspects

For the hospital isolates dataset, we will know in which disease the bacteria were involved, as well as being in a position to tell which isolates were sampled from the same patient (longitudinal series). However, the data will be treated confidentially as described in our consortium publication (doi: https://doi.org/10.3390/v13081543 - section 2.7)

Where will the data be made available? If already known, please provide a repository per dataset or data type.

All genomics data will be made available in perpetuity via the INSCD databases. Programs for the analysis of the data will be made available via the git platform of the lab: https://github.com/LoGT-KULeuven

When will the data be made available?

Upon publication of research results.

Which data usage licenses are you going to provide? If none, please explain why.

Concerning the genomics datasets, as stated in the mission of the INSDC:
"The INSDC has a uniform policy of free and unrestricted access to all of the data records their databases contain. Scientists worldwide can access these records to plan experiments or publish any analysis or critique. Appropriate credit is given by citing the original submission, following the practices of scientists utilising published scientific literature."

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

Any published digital dataset will be made available in perpetuity in the KU Leuven Research Data Repository (RDR) with a corresponding DOI accession number.

What are the expected costs for data sharing? How will these costs be covered?

Not applicable.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Cédric Lood

Who will manage data storage and backup during the research project?

Cédric Lood

Who will manage data preservation and sharing?

Cédric Lood and Rob Lavigne

Who will update and implement this DMP?

Cédric Lood and Rob Lavigne