REPHACTOR - PHAGE-BASED BIOSYNTHETIC GENE CLUSTER REFACTORING FOR THE DISCOVERY AND YIELD IMPROVEMENT OF NOVEL BIOACTIVE NATURAL PRODUCTS

A Data Management Plan created using DMPonline.be

Creators: Jorien Poppeliers https://orcid.org/0000-0002-9015-8658, n.n. n.n., Rob Lavigne

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Principal Investigator: n.n. n.n., Rob Lavigne

Grant number / URL: G044624N

ID: 208250

Start date: 01-01-2024

End date: 31-12-2027

Project abstract:

Microorganisms produce a wealth of structurally diverse bioactive natural products with important applications in medicine and agriculture. Unfortunately, the majority of natural product biosynthetic gene clusters (BGCs) are not, or only minimally, expressed under laboratory conditions. This poses a major challenge to the discovery and functional characterization of novel natural products, especially in non-model bacteria. Current methods for activating and enhancing BGC expression, such as heterologous expression in model organisms, are costly, inefficient and require extensive screening capacity. To overcome these limitations, this project aims to draw inspiration from Nature and exploit the ability of phages to hijack and control bacterial metabolic pathways. Specifically, it aims to activate and enhance the expression of natural product BGCs in non-model bacteria by integrating orthogonal, phage-derived genetic elements while taking into account the underlying regulatory mechanisms that modulate these pathways. As a proof concept, this innovative phage-based refactoring pipeline will be applied to the identification of cryptic metabolites in five diverse Pseudomonas species and to the yield enhancement of a natural product anticancer agent with therapeutic potential. Together, these innovations will have broad translational applications and will provide a much more efficient, sustainable and cost-effective means of generating industrially valuable natural products.

Last modified: 20-06-2024

REPHACTOR - PHAGE-BASED BIOSYNTHETIC GENE CLUSTER REFACTORING FOR THE DISCOVERY AND YIELD IMPROVEMENT OF NOVEL BIOACTIVE NATURAL PRODUCTS
DPIA

DPIA

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

- Not applicable

REPHACTOR - PHAGE-BASED BIOSYNTHETIC GENE CLUSTER REFACTORING FOR THE DISCOVERY AND YIELD IMPROVEMENT OF NOVEL BIOACTIVE NATURAL PRODUCTS
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

REPHACTOR - PHAGE-BASED BIOSYNTHETIC GENE CLUSTER REFACTORING FOR THE DISCOVERY AND YIELD IMPROVEMENT OF NOVEL BIOACTIVE NATURAL PRODUCTS

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)

Experimental approaches will result in newly generated data, consisting of small datasets like digital images (e.g., .jpg, .tiff, .png) and numerical data (e.g., .xls and .csv), as well as large datasets such as high-throughput RNA seq (e.g., .fastq) and mass spectrometry (e.g., .d) data. These experiments will also have associated data such as methods, protocols and lab books (e.g., .docx, .txt, .pdf). Furthermore, dissemination of the research will occur through manuscripts and presentations (e.g., .pptx). Data (e.g., DNA sequences and publications in scientific journals) from public databases (e.g., PubMed, NCBI) will be (re)used for analysis. Biological data will be reused (e.g., production strains) and new biological data will be generated (e.g., engineered production strains, expression vectors). Experimental data (e.g., RNA-seq, mass spectrometry data) will also be reused. We will not be working with personal data in this project.

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 persons
    - Joleen Masschelein (joleen.masschelein@kuleuven.be)
    - Rob Lavigne (rob.lavigne@kuleuven.be)
2. Storage capacity/repository
    - During the research, data collection and a summary of the data will be described in digital lab books. They are saved to a secure shared network drive from KU Leuven to share the obtained data with colleagues in the lab. A back-up of the drive is made automatically and previous files can be retrieved when current files are damaged. Since the used network drive is suitable to store large datasets at minimal cost, all relevant data will be archived here for a minimum of 10 years (following the RDM policy of KU Leuven). After this time, irrelevant versions of digital DNA sequences, digital lab notebooks and experimental datasets will be scrutinized for prolonged storage or disposal. Upon publication, the data will be shared in public data bases (e.g., Gene Expression Omnibus (NCBI))

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)

Not applicable

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)

Yes, synthetic biology in general aims to re-write biological systems to design new parts and devices for applications with high industrial and societal benefits. At the same time it raises concerns in terms of biosecurity, and can be considered as a dual use research. Researchers involved in this project will be trained to create awareness of any potential ethical or misuse implications of this research. This responsibility lies with the PIs (Joleen Masschelein & Rob Lavigne) as head of research within this project and will be achieved through discussions, making available relevant literature and by general adherence and scrutiny by the scientific community. Through Leuven Research and development of the KU Leuven, material transfer agreements will be set up (data sharing agreement) that
ensure a clear communication of the designated uses and plans by groups receiving information/materials. In research publications and dissemination of results, all members performing the research will be mindful towards the release of sensitive
information in terms of dual use.

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

Not applicable

REPHACTOR - PHAGE-BASED BIOSYNTHETIC GENE CLUSTER REFACTORING FOR THE DISCOVERY AND YIELD IMPROVEMENT OF NOVEL BIOACTIVE NATURAL PRODUCTS

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, .csv,.pdf, .txt, .rtf, .dwg, .gml, … • NA	*Please choose from the following options:* • <100MB • <1GB • <100GB • <1TB • <5TB • <10TB • <50TB • >50TB • NA	
ONT-cappable-seq	RNA-seq	New/ reuse	Digital	Experimental	.fastq	< 5 TB	
MS	Mass spectrometry	New	Digital	Experimental	.d	< 2 TB	
Lab books	Summary of experimental research, images, numerical, data, graphs	New	Digital	Experimental/compiled data	.docx, .pdf	< 100 MB	
Biological samples	Engineered bacterial strain	New	Physical	Experimental	NA	NA	Glycerol stocks stored in duplicate (> 10 strains)
Production hosts	Original (non-engineered) bacterial strains	Reuse	Physical	Experimental	NA	NA	Glycerol stocks stored in duplicate (6 strains)
Publications	Publications	New	Digital	Other	.pdf	< 100 MB	
Inventory	Inventory and description of the bacterial strains	New	Digital	Compiled data	.xlsx	< 100 MB	
Plasmids	Plasmids for engineering	New/reuse	Physical	Experimental	NA	NA	DNA stored in duplicate
HPLC	Chromatograms from the HPLC purifications	New	Digital	Experimental	.pdf	< 25 Gb	
NMR	Data from NMR spectroscopic analysis	New	Digital	Experimental	.FID	< 50 Gb	

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:

- Production hosts
  - *P. fluorescens NCIMB 10586*
    - resource: https://www.ncimb.com/
  - *P. chlororaphis NRRL-B-30761*
    - resource: https://nrrl.ncaur.usda.gov/
  - *P. syringae pv. porri CFBP1770*
    - resource: https://www.frontiersin.org/journals/microbiology/articles/10.3389/fmicb.2016.00279/full
  - *P. baetica*
    - resource: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5909648/
  - *P. wadenswilerensis*
    - resource: https://pubmed.ncbi.nlm.nih.gov/28820109/
- ONT-cappable-seq
  - ONT-cappable-seq data generated as part of an SB fellowship of FWO (1S18723N)
- Plasmids
  - Generated as part of previous research:
    - CRISPR Cas3: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10715078/
    - pPUT system: generated as part of an SB project of FWO (G096519N)

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, dual use

Risk assessment strategy: Synthetic biology in general aims at re-writing biological systems to design new parts and devices for applications with high industrial and societal benefits. At the same time it raises concerns in terms of biosecurity, and can be considered as a “dual use research”. Considering this research program, which is of high scientific (and potential commercial) value, we remain mindful for potential abuse of this research.

Research results will include the identification of new synthetic biology tools, sourced from nature, which can be optimized/engineered for introduction into (laboratory strain) bacteria. If this information or materials would fall into the wrong hands, the research and its tools would still require a highly advanced level of expertise and skill to convert the expected outcomes of this research into something which could harm people, animals or the environment.

Risk mitigation strategy:

- Researchers involved in this program will be trained to create awareness of any potential ethical or misuse implications of this research. This responsibility lies with my promoter as head of research within this project and will be achieved through discussions, making available relevant literature and by general adherence and scrutiny by the scientific community.

- Our lab adheres to the standards required in terms of safety, according to the guidelines mentioned above. More specifically, our L2 environment has limited access to authorized personnel.

-Through Leuven Research and development of the KU Leuven, material transfer agreements will be set up (‘data sharing agreement’) that ensure a clear communication of the designated uses and plans by groups receiving information/materials.

-In research publications and dissemination of results, all members performing the research will be mindful towards the release of sensitive information in terms of dual use.

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.

- No

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.

- Yes

The final engineered *P. baetica* strain can be used for commercial valorization. It can be licensed out to industrial partners for the optimized production of the secondary metabolite. In addition, potential novel natural products with industrially relevant biological activities resulting

from this project will be protected through patent applications which will enable future exploitation and commercial development.

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.

- No

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).

A detailed lab book is maintained containing information on the methods used to generate and analyze data:

- protocols to describe how the data is obtained
- code used to analyze the data
- graphs/tables to interpret the data
- Inventory of the bacterial strains and plasmids

All data will be timestamped, enabling easy lookup in the lab books how it was acquired.

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

Upon publication, data will be made available through online databases following their metadata standards:

- RNA sequencing data
  - Gene Expression Omnibus (NCBI): https://www.ncbi.nlm.nih.gov/geo/
- Metabolomics data
  - Metabolomics workbench (NIH): https://www.metabolomicsworkbench.org/
- Other metadata for which no specific tools is available can be stored on the KU Leuven RDR platform

All data will be collected in lab books, where a time indication will enable us to trace back results.

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

Where will the data be stored?

During the Project:

- shared network drive (J:) from KU Leuven to store lab books, protocols, etc.

- large volume storage (K:) from KU Leuven to store sequencing and MS data
- biological data will be stored at -80°C/-20°C

After PhD

- All data will be transferred to large volume storage (K:) for at least 10 years
- biological data will be stored in duplicate at -80°C/-20°C freezer for long term storage

How will the data be backed up?

Data will be backed-up according to the standard back-up procedures provided by KU Leuven ICTS. This involves automatic version management of the files. Version management is done using "snapshot" technology, where the previous versions of the changed files are kept online in a snapshot on the same storage system.

- by default, 1 snapshot is taken daily and is kept for 14 days. So you can go back to previous versions of the file up to 14 days.
- end users can restore older files themselves from within their Windows PC via the "previous versions | previous versions" functionality.

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.

- Yes

Currently, the large storage volume (K:) has a storage capacity of 6TB. If more space is needed in the future, extra storage space can be bought per TB.

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

Both labs adhere to the standards required in terms of safety, according to KULeuven guidelines. More specifically, our L2 environment has limited access to authorized personnel.
The network drives can only be accessed by selected members in the lab.
Freezers with biological data are located in the labs which have restricted access for unauthorized personnel (by means of a badge-system).

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

The costs for storing large datasets equals €104,42/TB each year. Upon publication, data will be deposited in online repositories as mentioned above, reducing the data storage costs of the lab. Costs will be covered by the research group.

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...).

Computational data (e.g., sequencing data, MS data, lab books) and biological data

- stored for at least 10 years according to the KU Leuven RDM policy

For both datatypes a selection will be made for data that is either hard/costly to replicate (such as engineered strains and MS data) or has potential future applications (RNA seq data)

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

- Computational data
  - large volume storage drive (K:-drive) of KU Leuven
- Biological samples
  - In duplicate in 2 different -80°C/-20°C freezers

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

Most of the data will be made available through online repositories, reducing the storage costs to below 1TB/year (<104,42/TB/year). The costs will be covered by the project budget.

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

- Sequencing data will be made available through (GEO (NCBI))
- Metabolomics data will be made available through (metabolomics workbench (NIH))
- Upon request, other data can be shared with interested parties

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

Not applicable

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, Intellectual Property Rights
- Yes, Aspects of dual-use

The final engineered *P. baetica* strain can be used for commercial valorization. It can be licensed out to industrial partners for the optimized production of the secondary metabolite. In addition, potential novel natural products with industrially relevant biological activities resulting from this project will be protected through patent applications which will enable future exploitation and commercial development.
Concering dual use, material transfer agreements ('data sharing agreement') will be set up through Leuven Research and development of the KU Leuven, that ensure a clear communication of the designated uses and plans by groups receiving information/materials. Before sending the samples, we want clear communication of the designated uses and plans by groups receiving information/materials. In addition, background check on the receiver will be performed to see if there are any red flags regarding misuse of the material. If in doubt, material will not be send.

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

- RNA sequencing data
  - Gene Expression Omnibus (NCBI): https://www.ncbi.nlm.nih.gov/geo/
- Metabolomics data
  - Metabolomics workbench (NIH): https://www.metabolomicsworkbench.org/
- Other metadata for which no specific tools is available can be stored on the KU Leuven RDR platform

When will the data be made available?

Upon publication of an associated research article.

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

Data from the project that can be shared will be made available under a creative commons attribution license (cc-by 4.0), so that users have to give credit to the original data creators.

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

Standard, a DOI will be created.

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

The proposed repositories don't charge a fee for uploading data. if other repositories are used, the project budget will cover the costs.

6. Responsibilities

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

Joleen Masschelein (PI) and Rob Lavigne (PI)

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

Joleen Masschelein (PI) and Rob Lavigne (PI)

Who will manage data preservation and sharing?

Joleen Masschelein (PI) and Rob Lavigne (PI)

Who will update and implement this DMP?

Jorien Poppeliers (PhD student), under supervision of Joleen Masschelein (PI) and Rob Lavigne (PI).