DMP title

Project Name FWO research project 2021 - DMP title

Project Identifier G0C4322N

Grant Title G0C4322N

Principal Investigator / Researcher Jan Michiels

Project Data Contact Natalie Verstraeten, Tel +32 16 376643, natalie.verstraeten@kuleuven.be

Description There is a pressing need for novel antibiotics in face of increasing drug resistance. Most antibiotics in clinical use are derived from natural products. It is generally accepted that there is still an enormous untapped potential left, especially in soil. However, many promising strains remain out of reach of functional exploration because of their low abundance, compounded by limited culturability under standard lab conditions. In this proposal, we will harness the power of culture-independent, genotype-based selection, while dramatically improving the lower limit of detection - aiming for single-clone sensitivity. We will investigate the feasibility of achieving this aim by adapting a high-throughput droplet microfluidics platform to perform digital PCR for eDNA cosmid clone selection. To demonstrate proof of concept, we will focus on calcium-dependent lipopeptide antibiotics, with the aim of discovering structurally novel building blocks of this promising class of antibiotics. Next, we will explore whether the capabilities of the resulting platform can be expanded to allow genotype-based selection of single cells, based on taxonomic markers informed by publicly available metagenomics studies. This would enable the enrichment of specific, rare, high-interest taxonomic groups for targeted whole-genome sequencing. This analysis is expected to yield a wealth of novel antibiotic candidates, thus opening the door to truly transformative discovery of novel natural products.

Institution KU Leuven

1. General Information

Name applicant

Jan Michiels (KU Leuven, VIB)

Liesbet Lagae (KU Leuven, IMEC)

FWO Project Number & Title

Project number: G0C4322N

Title: Digging deep into the arsenal of calcium-dependent antibiotics through highly targeted genotype-based selection

Affiliation

- KU Leuven
- Other

VIB, IMEC

2. Data description

Will you generate/collect new data and/or make use of existing data?

- Generate new data
- Reuse existing data

Describe in detail the origin, type and format of the data (per dataset) and its (estimated) volume. This may be easiest in a table (see example) or as a data flow and per WP or objective of the project. If you reuse existing data, specify the source of these data. Distinguish data types (the kind of content) from data formats (the technical format).

Type of data	Format	Volume	How created	Reuse/new data

Sequence analysis and primer design for CDA biosynthetic gene cluster detection	.fasta	500 MB	Multiple sequence alignments performed on A domains of characterized biosynthetic gene clusters of CDAs and manually curated; additional tools and databases can be used, e.g. antiSMASH and eSNaPD	Generate new data
Primers	Vials stored at -20°C	2 mL cryotubes	Primer ordered at IDT	Generate new data
PCR efficiency and analytical sensitivity for primer validation	.xslx	500 MB	qPCR experiments on dilution series of different DNA templates	Generate new data
Soil samples	Bags with soil	30 bags	Soil samples from geographically diverse environments in Belgium (meadow, cropland, mixed woodland, forest, river valley, sandy soil, alkaline soil, loamy soil)	Reuse our own data
eDNA libraries for soil samples	Glycerol stocks frozen at -80°C	2 mL cryotubes	High-molecular-weight DNA extracted and used to construct a saturating eDNA fosmid/BAC library for each soil sample	Generate new data
Genomic DNA libraries for strain collections	Glycerol stocks frozen at -80°C	2 mL cryotubes	High-molecular-weight DNA extracted and used to construct a saturating genomic DNA fosmid/BAC library for each strain collection	Generate new data
Designs for microfluidic chips	.gds	1 GB	Drawings in mask design software e.g. Cadence	Generate new data
Silicon-glass droplet microfluidics high-throughput screening platform	.csv, .txt, .tif, .jpeg	2 TB	Implemented according to the state-of-the-art, i.e. featuring picoliter droplets, droplet splitting and merging, and dielectrophoresis-based sorting capabilities	Reuse our own data

Digital PCR assay	.xslx	5 GB	PCR assay mix compartmentalized in a water-in-oil emulsion using droplet generator chips; droplets are collected in PCR tubes, subjected to temperature cycling using a benchtop tool, and analyzed using droplet detection chips and microscope-based fluorimetry to quantify the number of positive versus negative droplets	Generate new data
Digital PCR results on eDNA libraries	.fasta	500 MB	Use of the optimized digital PCR assay to select eDNA fosmid/BAC clones containing CDA biosynthesis clusters; results confirmed by DNA sequencing, compared to sequences present in databases, and subjected to hierarchical clustering to identify subclades that do not contain genes involved in the synthesis of previously characterized compounds	Generate new data
E. coli, S. lividans and *S. albus* cells harboring eDNA of interest	Glycerol stocks frozen at -80°C	2 mL cryotubes	Clone DNA is first amplified by a replication-cycle reaction and subsequently transferred to hosts of interest by in-flow electroporation using custom microchip technology (developed as part of the current project)	Generate new data
Antibiotic activity on ESKAPE pathogens of *Streptomyces* strains harboring eDNA of interest	.tif, .jpeg	10 GB	Plate assay with overlay of a reporter strain; effect of extracts and supernatant from producer strains	Generate new data

Shortlist of taxons to target by taxonomic enrichment	.docx	1 MB	Literature search	Generate new data
PCR primers for demonstrating taxonomic enrichment	Digitally preserved sequences and vials stored at -20°C	20 KB	DNA sequences will be retrieved from public databases; multiple sequence alignments will be performed and manually curated; degenerate primers will be designed to amplify these regions using e.g. CLC Main Workbench (Qiagen)	Generate new data
Validated PCR assays for demonstrating taxonomic enrichment	.xslx	5 GB	PCR assay mix compartmentalized in a water-in-oil emulsion using droplet generator chips; droplets are collected in PCR tubes, subjected to temperature cycling using a benchtop tool, and analyzed using droplet detection chips and microscope-based fluorimetry to quantify the number of positive versus negative droplets	Generate new data
Validated protocol to enable direct single-cell droplet PCR	.xslx	5 GB	Cell lysis within droplets by heating; validation by plating and qPCR	Generate new data
Validated platform for single-cell sensitivity taxon selection	.csv, .txt, .tif, .jpeg	2 TB	Validation of single-cell taxon selection using mock samples representative for bacterial soil communities	Generate new data

Proof of concept for single-cell sensitivity taxon selection from soil samples	.fasta	2 TB	Filtered cell suspension from a soil sample will be diluted to reach single-cell encapsulation in droplets; droplets will be subjected to cell lysis and bulk PCR, targeting taxons of interest; positive droplets will be sorted and genomic DNA purified and subjected to sequencing (CoolMPS)	Generate new data
Taxon-based genotypic CDA markers for selection of eDNA clones	.fasta	500 MB	Biosynthetic gene clusters identified using publicly available sequence analysis tools e.g. antiSMASH and eSNaPD; PCR assays designed to target CDA gene clusters in taxons of interest; fosmid/BAC clone selection using our available eDNA libraries	Generate new data

3. Legal and ethical issues

Will you use personal data? If so, shortly describe the kind of personal data you will use. Add the reference to your file in KU Leuven's Register of Data Processing for Research and Public Service Purposes (PRET application). Be aware that registering the fact that you process personal data is a legal obligation.

- No

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? If so, add the reference to the formal approval by the relevant ethical review committee(s)

- No

Does your work possibly result in research data with potential for tech transfer and valorisation? Will IP restrictions be claimed for the data you created? If so, for what data and which restrictions will be asserted?

- Yes

Research data with potential for tech transfer and valorisation includes chip designs, DNA sequences and chemical structures of compounds of medical or industrial relevance. Potential tech transfer will be discussed with the research and development offices of KU Leuven, IMEC and VIB. Ownership of the generated data has been stipulated in a Collaboration Agreement.

Do existing 3rd party agreements restrict dissemination or exploitation of the data you (re)use? If so, to what data do they relate and what restrictions are in place?

- No

4. Documentation and metadata

What documentation will be provided to enable reuse of the data collected/generated in this project?

BIOLOGICAL MATERIAL
Cryotubes will be labeled with a reference number that links to an entry in our Microsoft Access Database which is hosted on a central server and accessible to all people involved in the project. All relevant information on the specific strains will be included in this database. This includes strain identifier, a clear description of how the mutants were constructed and a link to whole genome sequence if applicable.

EXPERIMENTAL RESULTS
Data will be generated following standardized protocols which are stored in a central OneNote notebook. Furthermore, specifically by the Michiels group, an E-notebook will be used to register day-by-day activities. Raw data, history and context of experiments, protocols and analyzed data will be uploaded to this E-Notebook and backed up in the cloud. After publication or upon submission of manuscripts for publication, all datasets described in the publication will be deposited in dedicated data repositories (see below).

DESIGNS AND SIMULATION RESULTS
Simulation results including a description of the used parameters and designs will be stored on secure servers with restricted access, maintained by IMEC ICT. Aims, results and conclusions will be described in Powerpoint files.

Will a metadata standard be used? If so, describe in detail which standard will be used. If no, state in detail which metadata will be created to make the data easy/easier to find and reuse.

- Yes

Various data types come with their own metadata containing technical information about settings, machine types, pixel density, resolution, channels... Examples of these include .fastq NGS files containing standard metadata on sequencing technique, or .nd2 following the Nikon metadata standards. Throughout the project, these data files will be preserved with their original metadata. For .txt, .csv, .xlsx files containing tabular information, extra tabs or a head text section will be used to explain the data, the meaning of the columns etc. For others lacking a formally acknowledged metadata standard, Dublin Core Metadata will be used and a readme file will be saved in the same directory of the datafiles to explain all the various data files and give a broad overview of the analysis steps. Moreover, we will closely monitor MIBBI (Minimum Information for Biological and Biomedical Investigations) for metadata standards that are more specific to our data.

After publication or upon submission of manuscripts for publication, all datasets described in the publication will be deposited in data repositories (see below). Depending on the repository that is used, the metadata standard used by that specific repository will be filled in.

5. Data storage and backup during the FWO project

Where will the data be stored?

BIOLOGICAL MATERIAL
Cryotubes will be stored in -80° freezers with restricted access.

EXPERIMENTAL RESULTS
An E-Notebook and/or OneNote will be used to collect data. Low-volume data, protocols and analyses will subsequently be stored in secure and internally shared folders on university/IMEC servers with built-in backup and versioning. Although built-in backup systems are in place, passwordprotected hard drives equipped with anti-virus programs will be used as backup for selected data sets. A network drive will also be used for large-scale data. A copy of these datasets will be made to desktop PCs with large computational power (or to computing cluster of our host institution) whenever data analyses will be performed. For final datasets that are part of publications or manuscripts posted on preprint servers, datasets will be deposited in publicly available repositories. Depending on the data type, this could be the SRA depository (for NGS data), KU Leuven's own data repository (RDR), Mendeley Data... and, whenever possible or required, data will also be fully shared via the publisher's website. Scripts and code will be stored (and shared after reaching a finality) via Github. Chips and lab setups at IMEC are accessible only by people with the required security clearing and proper training, controlled by badge programming.

DESIGNS AND SIMULATION RESULTS
Designs and simulation results will be stored on secure servers with restricted access, maintained by IMEC ICT.

How is backup of the data provided?
BIOLOGICAL MATERIAL
A backup of critical strains will be stored in a different physical location (Kevin Verstrepen lab, Bioincubator Leuven).

EXPERIMENTAL RESULTS
Data will be stored on KU Leuven / IMEC central servers with automatic daily back-up and version control procedures.

DESIGNS AND SIMULATION RESULTS
Data will be stored on secure servers with restricted access and automatic daily back-up and version control procedures, maintained by IMEC ICT.

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

BIOLOGICAL MATERIAL
Sufficient storage is available.

EXPERIMENTAL RESULTS
OneDrive at KU Leuven already offers 5TB of data per user. Network storage is purchased on a group level and increased whenever needed. Github space is currently free of charge and only requires small volumes. External hard drives are cheap for large volumes and are readily available in the labs. IMEC ICT ensures sufficient digital data storage space at IMEC.

What are the expected costs for data storage and back up during the project? How will these costs be covered?
BIOLOGICAL MATERIAL
-80° freezers are present in our lab (costs are covered by general lab expenses).

EXPERIMENTAL RESULTS
The costs for large volume storage are covered by the project (€600 per year).

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?
BIOLOGICAL MATERIAL
Unauthorized people do not have access to the strains.

EXPERIMENTAL RESULTS
E-notebooks are password protected and data stored in KU Leuven's / IMEC's secure environment are secured by a two factor authorization and frequently changed passwords. External HDD are password-protected and stored in the safety of the lab. At IMEC, chips and lab setups can only be accessed by people with required security clearance and safety training, controlled through badge programming.

6. Data preservation after the FWO project
Which data will be retained for the expected 5 year period after the end of the project? In case only a selection of the data can/will be preserved, clearly state the reasons for this (legal or contractual restrictions, physical preservation issues, ...).
All data will be retained at least for 10 years, as required by KU Leuven's RDM policy.

Where will the data be archived (= stored for the longer term)?
BIOLOGICAL MATERIAL
All strains will be stored for at least 10 more years after the end of the project. For this purpose, -

80° freezers are available in the Michiels lab. Relevant strains will also be deposited in a public repository (e.g. the Belgian Coordinated Collections of Micro-organisms (BCCM)).

EXPERIMENTAL RESULTS
Data and designs will in first instance be stored on KU Leuven / IMEC central servers, and, after publication, data will additionally indefinitely be stored in open access repositories (e.g. Zenodo, Mendeley Data, KU Leuven's RDR).

What are the expected costs for data preservation during the retention period of 5 years? How will the costs be covered?

BIOLOGICAL MATERIAL
-80° freezers are present (included in general lab costs). Deposit of biological material in public repositories is generally without a fee.

EXPERIMENTAL RESULTS
The costs (€99,55 per TB per year) will be covered by general lab budgets.

7. Data sharing and reuse

Are there any factors restricting or preventing the sharing of (some of) the data (e.g. as defined in an agreement with a 3rd party, legal restrictions)?

- No

All published data will be made available at the time of publication. However, in case we identify valuable IP, we will first protect commercial exploitation, either through patenting or via an MTA that restricts the material from commercial use. This will be done after consulting with the research and development offices of KU Leuven, IMEC and VIB.

Which data will be made available after the end of the project?

BIOLOGICAL MATERIAL
Microbial strains generated in this project will be made available upon simple request.

EXPERIMENTAL RESULTS
Published data are likely to include chip designs, droplet microfluidics data, strain descriptions, sequencing data, and platform optimization procedures. These datasets will be made available via dedicated public databases or via open access repositories (e.g. Zenodo) (see above).

Where/how will the data be made available for reuse?

- Other (specify):

BIOLOGICAL MATERIAL
We aim at communicating our results in top journals that require full disclosure of all included data. Biological material will be shared upon simple request following publication.

EXPERIMENTAL RESULTS
We aim at communicating our results in top journals that require full disclosure upon publication of all included data, either in the main text, in supplementary material or in a data repository if requested by the journal and following deposit advice given by the journal. Depending on the journal, accessibility restrictions may apply. Proper links to these data sets will be provided in the corresponding publications.

When will the data be made available?

- After an embargo period. Specify the length of the embargo and why this is necessary
- Upon publication of the research results

Unpublished data will be embargoed for public access for another 5 years to allow the research groups to publish research findings.

Who will be able to access the data and under what conditions?

BIOLOGICAL MATERIAL
Biological material will be distributed to other parties if requested.

EXPERIMENTAL RESULTS
Depending on the journal: open access or standard subscription-based publication.

What are the expected costs for data sharing? How will the costs be covered?
BIOLOGICAL MATERIAL
Generally, shipment is paid by requesting parties.

EXPERIMENTAL RESULTS
A budget for publication costs has been requested in this project (€2500 per year).

8. Responsibilities

Who will be responsible for data documentation & metadata?

Both applicants (i.e. Jan Michiels and Liesbet Lagae) will share responsibility for data documentation & metadata. At IMEC, this responsibility may be shared with a Principal Scientist acting as project lead (currently Maarten Fauvart).

Who will be responsible for data storage & back up during the project?

Both applicants (i.e. Jan Michiels and Liesbet Lagae) will share responsibility for data storage & back up during the project. At IMEC, this responsibility may be shared with a Principal Scientist acting as project lead (currently Maarten Fauvart).

Who will be responsible for ensuring data preservation and reuse ?

Both applicants (i.e. Jan Michiels and Liesbet Lagae) will share responsibility for ensuring data preservation and reuse. At IMEC, this responsibility may be shared with a Principal Scientist acting as project lead (currently Maarten Fauvart).

Who bears the end responsibility for updating & implementing this DMP?

The PI (Jan Michiels) bears the end responsibility of updating & implementing this DMP.