DMP title

Project Name My plan (FWO DMP) - DMP title
Grant Title 1S43022N
Principal Investigator / Researcher David Bahling
Institution KU Leuven

1. General Information

Name applicant
David Bahling

FWO Project Number & Title
1S43022N

Heterologous Expression of Calvin Cycle Enzymes in S. cerevisiae for Hemi-autotrophic Growth and Bioethanol Production in a Novel Yeast-Driven Bionic Leaf

Affiliation

- KU Leuven

2. Data description

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

- Generate new 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).

The researcher will generate, collect, process, analyze and store the data listed below, as detailed in the project description.
The following datasets will be generated:

1. **Experimental data**

Dataset 1.1. – Digital images (WP 1,2,3) – (~1-2 GB)

Microscopy pictures, gel scans, graphs, illustrations, figures
Digital images in raster formats: uncompressed TIFF (.tif/.tiff), JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf)

Dataset 1.2. – Electrophysiology data (WP 3) – (~1-5 GB)

Data generated from potentiostat
Electrophysiology data: Applied Biosystems Sequence Tracer Sequence Trace (.abf)

Dataset 1.3. – Omics data (WP 1,2,3) – (~3-5 GB)

Genomics, transcriptomics, metabolomics

Dataset 1.4. – Plasmids (WP 1,2,3)

The researcher will generate approximately 300 bacterial plasmids containing gene constructs for gene integration. Gene constructs will consist, for example, of Calvin cycle enzymes (RuBisCO, phosphoribulokinase) and autotrophic enzymes (hydrogenase, carbonic anhydrase).
Plasmid-vectors: glycerol stocks frozen at -20°C

Dataset 1.5. – Strains (WP 1,2,3)

Bacterial strains (containing plasmids) and yeast strains (approximately 200, containing heterologous genes)
Bacterial and yeast strains generated: glycerol stocks frozen at -80°C

2. **Derived and compiled data**

Dataset 2.1 – Research documentation (WP 1,2,3) – (~0.5 GB)

Research documentation generated by the researcher or collected from online sources and from collaborators, including laboratory notes protocols, and experimental data.

MS Word (.doc/.docx), Adobe Portable Document Format (.pdf), Survey data: MS Excel (.xls)

Dataset 2.2 – Manuscripts (WP 1,2,3) – (~0.5 GB)

Progress/journal reports describing experiments, interpretation and conclusions derived thereof.

MS Word (.doc/.docx), MS PowerPoint (.ppt), Survey data: MS Excel (.xls), Adobe Portable Document Format (.pdf), Digital images in vector formats: scalable vector graphics (.svg), encapsulated postscript (.eps), Scalable Vector Graphics (.svg), Adobe Illustrator (.ai), Chemical structures: .sdf or .csv format

3. **Canonical data**

Dataset 3.1 – Nucleic acid sequences (WP 1,2,3) – (~1-2 GB)

Nucleotide sequences: raw sequence data trace (.ab1), text-based format (.fasta/.fa)

Dataset 3.2 – Protein sequences (WP 1,2,3) – (~1-2 GB)

Protein sequences: raw sequence data trace (.ab1), text-based format (.fasta/.fa)

Protein structures: Protein Data Bank format (.pdb / .pdbx)

**All relevant generated data will be incurred through the duration of this project and derived from the host lab unless stated otherwise*

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

Ownership of the generated data belongs to KU Leuven & VIB in accordance with the framework agreement between these institutes; copyright of the data belongs to prof. Kevin Verstrepen.

We identify in an early phase the valorization potential of research lines and has a vast network of industrial contacts to efficiently start the route to commercialization. The lab is supported in this matter by Dr. Stijn Spaepen, IOF innovation manager responsible for research valorization. For research with valorization potential, the host lab actively protects its IP by filling patent applications with support from the IP department of VIB. Type of data with potential for tech transfer and valorization include yeast strains isolated and generated during the timeframe of this project (including phenotypic data), sequencing information generated during the timeframe of this project and (analysis) data and models hereof derived. Valorization potential includes a) licensing of (improved) strains or information on linking a specific sequence variant to a phenotype and b) creation or participation in start-up companies.

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

No third-party agreement restricts the dissemination or exploitation of the data or strains generated from this project. In particular, existing agreements between VIB and KU Leuven do not restrict the publication of data. There is no IP on the generated strains that would prevent us

from storing the strains, performing the anticipated experiments, or publishing the results.

4. Documentation and metadata

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

Data will be generated following standardized protocols. Data will be documented by the researcher at the time of data collection and analysis, by taking careful notes in their laboratory notebook. Cryotubes of biological samples (bacterial and yeast strains) stored at -80°C will be labeled with a reference number that links to an entry in our strain database.

Digital data files will be stored on KU Leuven servers and will be accompanied by a read-me text file that contains relevant metadata for understanding and re-use of data.

The read me file will be structured as follows:

PROJECT INFORMATION

P01. Project name/nickname, including the name of funder, type of grant, grant number

P02. Name and contact of the Principal Investigator and scientists (the whole team) involved in the project with an indication of the lead scientist

P03. Small description of the project

P04. (Archive) Dropbox folder link associated with this project if any

GENERAL INFORMATION

G01. Names of the file(s) or dataset(s) that this README file describes

G02. Date of creation/last update of the README file.

For each update/change made to the files, please add an extra line here with the date and main changes made.

G03. Created by (name)

G04. Description of the dataset

FILE OVERVIEW

F01. Software used to generate the data, including the software version used

F02. Software necessary to open the file

F03. Relationship between the files

For example, for scripts used, please also include a list of files generated by each script.

METHODOLOGICAL INFORMATION

M01. Date (beginning-end) and place of data collection

M02. Data collecting method

M03. Information about data processing methods and scripts used.

M04. Information about the instrument, calibration, settings used

M05. Information about limitations of the dataset (missing values, …), information that ensures correct interpretation of the dataset

M06. People involved in the creation or processing of the dataset

DATA ACCESS AND SHARING

A01. Confidentiality information/restrictions on the use of data

DATA SPECIFIC INFORMATION (ABOUT THE DATA THEMSELVES)

D01. Full names and definitions for columns and rows.

This can be provided in a .csv file

D02. Explanation of abbreviations

D03. Explanation of strain codes and sample names

Crosslink to lab strain list if possible.

RELATIONSHIPS
R01. Publications based on this dataset
R02. This dataset derives from… (other datasets)
R03. This dataset is related to… (documents, dataset)

Cryotubes and plates with biological strains are labeled with a reference number that links to an entry in our strain database (stored on both a professional Dropbox account as well as KU Leuven servers, see also below). All relevant information on the specific strains (strain ID, genetic information, origin of strain) is included in this database.

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

Since there is no formally acknowledged metadata standard specific to our discipline, Dublin Core Metadata will be used. Moreover, we will closely monitor MIBBI (Minimum Information for Biological and Biomedical Investigations) for metadata standards that are more specific to our data.

5. Data storage and backup during the FWO project

Where will the data be stored?

Biological samples: Cryotubes with strains will be stored in a -80° freezer present in the Verstrepen lab at KU Leuven (Gaston Geenslaan 1, Heverlee). Costs are covered by general lab expenses. Unauthorized people do not have access to strains.

Data (digital files) generated in this project will be stored in a Dropbox Business Advanced account for processing and analyses; following secure data transfer, modern data encryption standards, and encrypted block storage (256-bit AES and SSL/TLS encryption. For more details see: https://www.dropbox.com/business/trust

Sequencing data will be stored on an internal lab server (present in the host lab) as well as on a secure Dropbox Business account for processing and analyses.

All the relevant algorithms, scripts, and software code driving the project will be stored on a secure Dropbox account. Scripts used for analysis will also be stored in Jupyter notebook (jupyter.org - an open-source web application to store and share scripts), in github or in the GitLab service of KU Leuven.

Upon publication, all sequences supporting a manuscript will be made publicly available via repositories such as the GenBank database or the European Nucleotide Archive (nucleotide sequences from primers / new genes / new genomes), NCBI Gene Expression Omnibus (microarray data / RNA-seq data / CHIPseq data), the Protein Database (for protein sequences), the EBI European Genome-phenome Archive (EGA) for personally identifiable (epi)genome and transcriptome sequences.

How is backup of the data provided?

Strains are backed up in a compressed form (96-well plates) in the Verstrepen lab and on a 2nd location (Kasteelpark Arenberg 20, Heverlee) at KU Leuven. Data (digital files) are automatically backed up by the secure Dropbox Business Advanced account cloud backup services. Additionally, project data and sequencing data will be backed up to KU Leuven servers.

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

Dropbox Business offers unlimited storage and backup capacity in their clouds.
There is sufficient storage and backup capacity on all KU Leuven servers:

- the “L-drive” is an easily scalable system, built from General Parallel File System (GPFS) cluster with NetApp e-series storage systems, and a CTDB samba cluster in the front-end.
- the “J-drive” is based on a cluster of NetApp FAS8040 controllers with an Ontap 9.1P9 operating system.
- KU Leuven also offers a 2TB OneDrive already included in our group’s Office 365 subscription

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

The total estimated cost of data storage during the project is 250 EUR. This estimation is based on the following costs:

- Yeast/bacteria strains are easily kept alive for several weeks. This costs on average 5 EUR. When no experiments are planned with a specific strain, cryopreservation will thus be used to safeguard strains and prevent genetic drift, loss of transgene, and potential contaminations. -80°C freezers are present in the lab of prof. Verstrepen and costs are included in general lab costs.
- The costs associated with a Dropbox Business account have been negotiated by the lab to 10 USD/month/user.
- The costs of digital data storage are as follows: 173,78€/TB/Year for the “L-drive” and 519€/TB/Year for the “J-drive”.

Data security: how will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

For biological samples:

Unauthorized people do not have access to the strain collections.

For the digital data:

Access to data stored on the Dropbox Business Advanced cloud is granted based on role-based access control and all access requires layers of authentication that include strong passwords, SSH keys, 2-factor authentication, and one-time passcodes. Dropbox safeguards data with document watermarking, granular content permissions and policies, document watermarking, and legal holds.

Both the “L-drive” and “J-drive” KU Leuven servers are accessible only by laboratory members and are mirrored in the second ICTS data center for business continuity and disaster recovery so that a copy of the data can be recovered within an hour.

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, ...).

The minimum preservation term of 5 years after the end of the project will be applied to all datasets All datasets will be stored on the university’s central servers with automatic backup procedures for at least 10 years, conform the KU Leuven RDM policy. The costs (€113,84 per TB per year for “Large volume storage”) will be covered by the general lab costs of the Verstrepen Lab.

Where will the data be archived (= stored for the longer term)?

As a general rule, datasets will be made openly accessible, whenever possible via existing platforms that support FAIR data sharing (www.fairsharing.org), at the latest at the time of publication.

For all other datasets, long term storage will be ensured as follows:

- Biological data: yeast and bacterial strains will be stored locally in the laboratory (-80°C). Other biological and chemical samples: storage at 4°C and/or as frozen samples as appropriate
- Digital datasets: files will be stored on a Dropbox account and the KU Leuven “L-drive”.

What are the expected costs for data preservation during the retention period of 5

years? How will the costs be covered?

The total estimated cost of data storage during the 5 years after the end of the project is 1250 EUR. These costs will be covered by the Verstrepen lab general budget and have been included in the project's budget.

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)?

- Yes. Specify:

The researcher and host lab aim to communicate our results in top journals that require full disclosure of all included data. Biological material will be shared upon simple request following publication, unless we identify valuable IP, in which case we will first protect commercial exploitation, either through patenting or via an MTA that restricts the material from commercial use.

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

The researcher and host lab are committed to publishing research results in order to communicate them to peers and to a wide audience. All research outputs supporting publications will be made openly accessible, unless we identify valuable IP, in which case we will first protect commercial exploitation, either through patenting or via an MTA that restricts the material from commercial use.

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

- In an Open Access repository
- Upon request by mail
- Other (specify):

Open-access publications in peer-reviewed journals, including supplemental information.

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 datasets will be provided in the corresponding publications.

As a general rule, datasets will be made openly accessible via existing platforms that support FAIR data sharing (www.fairsharing.org). Sharing policies for specific research outputs are detailed below:

- Biological data: Bacteria and yeast strains will be shared upon simple request following publication unless we identify valuable IP. In this case, we will first protect commercial exploitation, either through patenting or via an MTA that restricts the material from commercial use.
- Datasets will be deposited in open access repositories. Research documentation: All protocols used to generate published data will be described in the corresponding manuscript(s), and the related documentation will be included as supplementary information. These data and all other documents deposited in the lab notebook are accessible to the PI and the research staff involved in the project and will be made available upon request.
- Manuscripts: We opt for open access publications where possible. Publications will be automatically listed in our institutional repository, Lirias 2.0, based on the author's name and ORCID ID.
- Algorithms, scripts, and software: As soon as a manuscript is publicly available, algorithms, scripts, and software code will be deposited in a github repository.
- Nucleic acid and protein sequences: Upon publication, all sequences supporting a manuscript will be made publicly available via repositories such as the GenBank database or the European Nucleotide Archive (nucleotide sequences from primers / new genes / new genomes), NCBI Gene Expression Omnibus (microarray data / RNA-seq data / CHIPseq data), the Protein Database (for protein sequences).

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

As a general rule all research outputs will be made openly accessible at the latest at the time of publication. No embargo will be foreseen unless imposed e.g. by pending publications, potential IP requirements – note that patent application filing will be planned so that publications need not be delayed – or ongoing projects requiring confidential data. In those cases, datasets will be made publicly available as soon as the embargo date is reached.

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

Whenever possible, datasets and the appropriate metadata will be made publicly available through repositories that support FAIR data sharing. As detailed above, metadata will contain sufficient information to support data interpretation and reuse and will conform to community norms. These repositories clearly describe their conditions of use (typically under a Creative Commons CC0 1.0 Universal (CC0 1.0) Public Domain Dedication, a Creative Commons Attribution (CC-BY), or an ODC Public Domain Dedication and License, with a material transfer agreement when applicable). Interested parties will thereby be allowed to access data directly, and they will give credit to the authors for the data used by citing the corresponding DOI. For data shared directly by the PI, a material transfer agreement (and a non-disclosure agreement if applicable) will be concluded with the beneficiaries in order to clearly describe the types of reuse that are permitted.

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

It is the intention to minimize data management costs by implementing standard procedures e.g. for metadata collection and file storage and organization from the start of the project, and by using free-to-use data repositories and dissemination facilities whenever possible. Data management costs will be covered by the laboratory budget.

8. Responsibilities

Who will be responsible for data documentation & metadata?

Metadata will be documented by the researcher and technical staff involved in the project at the time of data collection and analysis, by taking careful notes in their laboratory notebook that refer to specific datasets and by ensuring detailed README files and other documentation and metadata. Ms. Veerle Saels (senior lab technician) and Ms. Eef Lemmens (personal assistant to Prof. Kevin Verstrepen) will follow this up.

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

The researcher and technical staff involved in the project will ensure data storage and backup. Mr. Nico Vangoethem (IT responsible) will follow this up.

Who will be responsible for ensuring data preservation and reuse ?

The PI (Prof. Kevin Verstrepen) is responsible for data preservation and sharing, with support from the research and technical staff involved in the project. He will also be assisted by Mr. Nico Vangoethem and Ms. Eef Lemmens in these aspects.

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

The PI (Prof. Kevin Verstrepen) is ultimately responsible for all data management during and after data collection, including implementing and updating the DMP.