Development of a CRISPR-based tool for localized random mutagenesis of biosynthetic pathways

A Data Management Plan created using DMPonline.be

Creator: Karin Voordeckers

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

ID: 199756

Start date: 01-01-2023

End date: 31-12-2026

Project abstract:

Yeast is one of the most frequently used microbes in today's biotechnology and fermentation industry. But, despite its domestication and adaptation to man-made environments, new applications (such as the production of bioethanol or high-value chemicals) confront it with new challenges and unfamiliar, harsh environments. This leaves many opportunities to generate new, superior variants optimized for specific applications. Current approaches based on classic genetic modification often only partly succeed in meeting these goals and are very time-consuming, especially for multigenic phenotypes. Therefore, in this project, we address these limitations by developing a novel, powerful strategy for industrial strain engineering. This strategy, referred to as 'confined mutagenesis', is founded on the most recent developments in CRISPR-based gene editing tools, such as Type-I CRISPR systems and 'base editors', which we will repurpose for microbe engineering. It will allow for a rapid and inducible introduction of mutations within a user-defined part of the genome. We will develop such systems for a wide array of industrially-relevant yeasts. As a proof-of-concept, we will use this technology for the optimization of 1) xylose conversion (for the production of biofuels) in *Kluyveromyces marxianus* and 2) astaxanthin production (a carotenoid widely used as food colorant and dietary supplement) in *Saccharomyces cerevisiae*. The proposed tool will address an unmet need in synthetic biology.

Last modified: 06-06-2023

Development of a CRISPR-based tool for localized random mutagenesis of biosynthetic pathways
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
Dataset 1.1. – Digital images	gel scans (from e.g. check PCRs),	Generate new data	Digital	Experimental	.tiff, .jpeg	< 100GB	
Dataset 1.1. – Digital images	graphs, illustrations, figures.	Generate new data	Digital	Experimental	.tiff, .jpeg,.svg, .pdf	< 100 GB	
Dataset 1.2. – Cytometry data.	flow cytometry data	Generate new data	Digital	Experimental	.fsc	< 100 GB	
Dataset 1.3. – Omics data.	sequencing data of variants and strains	Generate new data	Digital	Experimental	.bcl, .gz, bed, .bg, .bedGraph, .bw, .bigwig.	< 10 TB	
Dataset 1.4. – Plasmids	plasmids with varying base editor constructs	Generate new data	Physical	Experimental			~30 plasmids
Dataset 1.5. – Strains.	Bacterial strains for cloning and propagating plasmids, yeast strains	Generate new data	Physical	Experimental			around 1000 bacterial and yeast strains, most in 96 well-plate format.
Dataset 1.6 – Spectrometry and chromatography data.	HPLC, GC, GC-MS measurements of compounds produced during fermentation in WP4-5.	Generate new data	Digital	Experimental	.csv, .msp	< 100 GB	
Dataset 2.1 – Research documentation.	Lab protocols	Generate new data	Digital	Compiled/aggregated	.doc, .pdf	< 1GB	
Dataset 2.1 – Research documentation.	Lab notebooks	Generate new data	Physical	Compiled/aggregated			~ 4 lab notebooks
Dataset 2.2 – Manuscripts	Manuscripts	Generate new data	Digital	Compiled/aggregated	.doc, .pdf	< 1GB	
Dataset 2.3 – Scripts.	Standard scripts, using existing packages, written in R to analyze and plot obtained data.	Generate new data	Digital	Software	.r	< 1 GB	
Dataset 3.1 – Nucleic acid sequences	DNA sequences	Generate new data	Digital	Compiled/aggregated	.sam, bam, .ab1, .fasta/fa, .qual, gb/gbk, .dna	< 100 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:

Not applicable

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.

- No

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 project has potential for tech transfer and valorization, since it aims at further developing synthetic biology tools for strain optimization. These tools and the resulting strains have industrial application potential.
Specifically, the following datasets have the potential for commercial valorization:

- Dataset 1.3-Omics data (eg. identification of mutations responsible for specific phenotype)
- Dataset 1.4- Plasmids (eg plasmids expressing the developed base-editing tools)
- Dataset 1.5-Strains (eg. strains with optimized industrially relevant phenotype)
- Dataset 2.1-Research documentation (eg detailed protocols on strain optimization and mimicking of industrial conditions on labscale)
- Dataset 3.1 Nucleic acid sequences

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).

Data will be generated following standardized protocols. Clear and detailed descriptions of these protocols are present in the lab and will be made publicly available upon publication.
Data will be documented at the time of data collection and analysis, by taking careful notes in the laboratory notebook. 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.

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. Additionally, project data and sequencing data will be backed up to KU Leuven servers.

Digital data files will be accompanied with a read me text file that contains relevant metadata for understanding and re-use of data.
All the relevant scripts 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.

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.

- No

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.

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

Where will the data be stored?

Biological material: Strains are stored in a -80°C freezer in the Verstrepen lab.
Experimental results: Data will be stored in a Dropbox Business account (see above).

How will the data be backed up?

Biological material: 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. Experimental results: 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

Biological material: The lab has sufficient space in the different -80°C freezers
Experimental results: Dropbox Business offers unlimited storage and back-up capacity in their clouds. There is sufficient storage and back-up capacity on all KU Leuven servers:
- the “L-drive” is an easily scalable system, built from General Parallel File System (GPFS) cluster with NetApp eseries 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

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 includes 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 datacenter for business continuity and disaster recovery so that a copy of the data can be recovered within an hour.

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

The total estimated cost of data storage during the project is 1300 euros. This estimation is based on the following costs:
Yeast/bacteria strains are easily kept alive for several weeks. This costs on average 5 euro. When no experiments are planned with a specific strain, cryopreservation will thus be used to safeguard strains, 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 has been negotiated by the lab to 10 USD/month/user; and costs are covered by the lab.

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 the biological material will be saved as well as all the experimental results that have potential for commercial valorization and/or required as a back-up for the research papers developed during the project.

Where will these data be archived (stored and curated for the long-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 Dropbox account and the KU Leuven “L-drive”.

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

The costs (€105 per TB per year for “Large volume-storage” at KU Leuven) will be covered by general lab budgets.

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, ...)
- Other, please specify:

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. Plasmids will be made available via Addgene (non-profit plasmid repository). If we identify valuable IP, we will first protect commercial exploitation, either through patenting or via an MTA that restricts the biological material (strains/plasmids) 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 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 authors name and ORCID ID.
- **scripts**: As soon as a manuscript is publicly available, algorithms, scripts and software code will be deposited in a github repository.
- **Nucleic acid 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).

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

For restricted access, this can be institutional only. 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.

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

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, 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.
Datasets for which IP rights might be applicable include:
Dataset 1.3. – Omics data.
Dataset 1.4. – Plasmids
Dataset 1.5. – Strains.
Dataset 3.1 – Nucleic acid sequences

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

- **Biological data**: Bacteria and yeast strains will be shared upon simple request following publication. Plasmids will be made available via Addgene (non-profit plasmid repository). If we identify valuable IP, we will first protect commercial exploitation, either through patenting or via an MTA that restricts the biological material (strains/plasmids) 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 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 authors name and ORCID ID.
- **scripts**: As soon as a manuscript is publicly available, algorithms, scripts and software code will be deposited in a github repository.
- **Nucleic acid 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).

When will the data be made available?

Upon publication of research results. Importantly, publication will be done after analyzing the IP of the project and filling the respective patents.

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

In principle:
Creative Commons Attribution-NonCommercial-NoDerivs (CC-BY-NC-ND)

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

Not available yet.

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

The KU Leuven RDR allow researchers to store 50 GB per year for free. If more data is required to be stored, then contact and negotiations with RDM helpdesk will be mandatory. GitHub, the price is 0.008 USD per GB/day and 0.5 USD per GB of data transfer. Also, depending on the journal where the paper is published will have different costs for data sharing and saving. Depositing plasmids at Addgene is free.

Costs will be covered from general lab budget.

6. Responsibilities

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

Metadata will be documented by the research 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 read me files and other documentation and metadata. Mr. Jeroen Cortebeeck (lab manager) will follow this up.

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

Dries De Vadder, our lab's IT responsable.

Who will manage data preservation and sharing?

Jeroen Cortebeeck (lab manager) and Karin Voordeckers (staff scientist)

Who will update and implement this DMP?

The PI (Kevin Verstrepen) bears the end responsibility of updating & implementing this DMP.

Development of a CRISPR-based tool for localized random mutagenesis of biosynthetic pathways
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)

1. **Experimental data**

Dataset 1.1. – Digital images. gel scans (from e.g. check PCRs), graphs, illustrations, figures.
Dataset 1.2. – Cytometry data. Flow Cytometry and fluorescence-activated cell sorting (FACS) data
Dataset 1.3. – Omics data. Genomics data.
Dataset 1.4. – Plasmids
Dataset 1.5. – Strains. Bacterial strains for cloning and propagating plasmids, yeast strains (*S. cerevisiae*, *K.marxianus*, *Y. lipolytica* and *K. pastoris*). For yeast strains, this includes lab strains, natural/clinical/industrial isolates, variant libraries, optimized clones (i.e. site-specific mutants constructed through genome engineering).
Dataset 1.6 – Spectrometry and chromatography data. HPLC, GC, GC-MS measurements of compounds produced during fermentation in WP4-5.

1. **Derived and compiled data**

Dataset 2.1 – Research documentation. Research documentation generated by the research and technical staff or collected from online sources and from collaborators, including laboratory notes, protocols.
Dataset 2.2 – Manuscripts
Dataset 2.3 – Scripts. Standard scripts, using existing packages, written in R to analyze and plot obtained data.

1. **Canonical data**

Dataset 3.1 – Nucleic acid sequences
Data will be stored in the following formats:

- **Text files**: MS word (.doc/.docx), Adobe Portable Document Format (.pdf)
- **Quantitative tabular data**: comma separated value files (.csv), MS Excel (.xls/xlsx).
- **Digital images**: uncompressed TIFF (.tif/.tiff),JPEG (.jpg), Adobe Portable Document Format (.pdf); scalable vector graphics (.svg).
- **Flow cytometry data**: Flow Cytometry Standard (.fcs).
- **Nucleotide and protein sequences**: raw sequence data trace (.ab1), text-based format (.fasta/.fa) and accompanying QUAL file (.qual), Genbank format (.gb/.gbk).
- **Next generation sequencing raw data**: binary base call format (.bcl), .fastq(.gz).
- **Spectrometry and chromatography data**: computable document format (.cdf), comma-separated value files (.csv), mass spectral files (.msp);
- **Bacterial and yeast strains generated**: glycerol stocks frozen at -80°C.

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)

RESPONSIBLE PERSON: Kevin Verstrepen, assisted by Jeroen Cortebeeck (senior labtech) and Dries de Vadder (lab IT responsible).

BIOLOGICAL MATERIAL: Yeast and bacterial strains are stored in a -80°C freezer and backed up in a compressed form (96-well plates) in the Verstrepen lab and on a 2nd location at KU Leuven. Costs are covered by general lab expenses. Unauthorized people do not have access to strains. Strain will be stored for at least 5 years after the project ends.

EXPERIMENTAL RESULTS: Data will be stored in a Dropbox Business account managed by the research group; following secure data transfer, modern data encryption standards, encrypted block storage. All data will be stored for at least 10 years, conform KU Leuven RDM 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)

Not applicable to this project.

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)

Not applicable to this project.

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

IP: This project could result in research data with potential for tech transfer and valorization. The lab identifies in an early phase the valorization potential 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 and the Business Development team of VIB-HQ.

DOCUMENTATION: Data will be generated following standardized protocols. Clear and detailed descriptions of these protocols are present in the lab and will be made publicly available upon publication.

SHARING: All data will be made publicly available upon publication, as associated files with the publication and/or by posting the data in relevant data repositories.

Development of a CRISPR-based tool for localized random mutagenesis of biosynthetic pathways
DPIA

DPIA

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

- Not applicable

Development of a CRISPR-based tool for localized random mutagenesis of biosynthetic pathways
GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable