# Development of an innovative CRISPR-based tool for industrial yeast strain improvement

*A Data Management Plan created using DMPonline.be*

**Creator:** Julian Prieto Vivas

**Affiliation:** KU Leuven (KUL)

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

**Template:** FWO DMP (Flemish Standard DMP)

**Project Administrator:** n.n. n.n.

**Grant number** / **URL:** 1S25923N

**ID:** 195631

**Start date:** 25-12-2022

**End date:** 01-12-2025

**Project abstract:**

The generation of genetic diversity via mutagenesis is a routinely used strategy for protein engineering and pathway optimization. Current technologies for random mutagenesis often target either the whole genome or relatively narrow windows. To bridge this gap, we developed a new tool, CoMuTER (**Co**nfined **Mu**tagenesis using a **T**ype I-**E** CRISP**R**-Cas system), that allows inducible and targetable, *in-vivo* mutagenesis of genomic loci of around 55 kilobases. CoMuTER employs a fusion protein of the targetable helicase Cas3 - signature enzyme of the class 1 type I-E CRISPR-Cas system- and nucleotide deaminases. As a result, Cas3 targets and unwinds large stretches of DNA while the deaminases introduce random point mutations. This strategy can be used to increase the mutation rate of complete metabolic pathways.

For the project I will compare the on-target (using amplicon sequencing and whole genome sequencing) vs off-target (using whole genome sequencing) of different CoMuTER variants when targeting it to a specific locus in the genome.
Then I will test the tool by optimizing a metabolic pathway, for example the production of riboflavin. For measuring riboflavin content I will use the flowcytometer. Maybe other pathways are going to be selected, so perhaps and HPLC will be used.

**Last modified:** 21-04-2023

# Development of an innovative CRISPR-based tool for industrial yeast strain improvement 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)**

The Work packages (WP) are divided as follows:

- WP1: Development of a 'Cas3 base-editor' with adenosine deaminases
- WP2: Improving versatility of CoMuTER– combining cytidine and adenosine base editing.
- WP3: Increasing throughput of CoMuTER: Multiplexing
- WP4: Proof-of-Concept: Improvement of riboflavin production

The following data will be generated:

1. **Experimental data**
   **Dataset 1.1-Digital Images:** Gel scans (to check PCR producst), graphs, illustrations and figures.
   **Dataset 1.2-Cytometry data:** Flow cytometry and fluorescence-activated cell sorting (FACS) data.
   **Dataset 1.3-Omics data:** Genomics data.
   Datasets 1.2 and 1.3 will be obtained from:
   WP1+WP2: For selecting the best constructs.
   WP3: For evaluating multiplexing
   WP4: Screening for high riboflavin producers.
   WP5: Evaluating the construct in non-conventional yeast strain.
   **Dataset 1.4 Strains:** Bacteria strains and yeast strains. this includes lab strains, variant libraries and optimized clones.
2. **Derived and compiled data**
   **Dataset 2.1- Research documentation:** Generated by the research and technical staff or collected from online sources and from collaborators, including laboratory notes, protocolsand summaries.
   **Dataset 2.2- Manuscripts**
   **Dataset 2.3-Algorithms and scripts**
3. **Canonical data**
   **Dataset 3.1- Nucleic acid sequences:** Primers, oligos, ultramers, genes and constructs developed throught the entire project.
   **Dataset 3.2- Protein sequences**

These dataset represent an important source of information for Kevin Verstrepen's lab (including future staff) and also for scientists, journalists and higher education teachers working in the field of metabolic engineering and genome evolution.
Data will be stored in the following formats:

- **Text files:** MS word (.doc/.docx), Adobe Portable Document Format (.pdf), LaTex (.tex) format.
- **Quantitative tabular date:** comma separated value files (.csv), MS Excel (.xls/xlsx).
- **Digital images in raster formats** : uncompressed TIFF (.tif/.tiff),JPEG (.jpg), JPEG 2000 (.jp2), Adobe Portable Document Format (.pdf).
- **Digital images in vector formats:** scalable vector graphics (.svg).
- **Flow cytometry data:** Flow Cytometry Standard (.fcs).
- **Theoretical model data code:** R (.r), python (.py).
- **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).
- **Sequence alignment data:** (.sam), .bam.
- **Protein and DNA sequences:** .dna
- **Coverage data:** .bed, .bg, .bedGraph, .bw, .bigwig.
- **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 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; following secure data transfer, modern data encryption standards, encrypted block storage. All data will be stored for at least 5 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.

# Development of an innovative CRISPR-based tool for industrial yeast strain improvement DPIA

## DPIA

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

- Not applicable

# Development of an innovative CRISPR-based tool for industrial yeast strain improvement GDPR

## GDPR

**Have you registered personal data processing activities for this project?**

- Not applicable

# Development of an innovative CRISPR-based tool for industrial yeast strain improvement 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 |
| 1.1.I-gel pictures | Contains all the PCR figures taken during the study | Generate new data | Digital | Experimental | .jpg, .jp2 | <100GB | |
| 1.1.II-Images | Contains all the illustrations and figures | Generate new data | Digital | Experimental | .jpg, .jp2, .tif/.tiff, .pdf, .svg | <100GB | |
| 1.2-Cytometry | Contains all the flow cytometry data | Generate new data | Digital | Experimental | .fcs | <100GB | |
| 1.3-WGS | Contains the whole genome sequencing data | Generate new data | Digital | Experimental | .bcl, .gz, bed, .bg, .bedGraph, .bw, .bigwig. | <5TB | |
| 1.4-Strains | Bacteria and yeast strains | Generate new data | Physical | Experimental | | | around 1000 bacterial and yeast strains, most in 96 well-plate format. |
| 2.1.I-Protocols | Lab protocols | Generate new data | Digital | Compiled/aggregated data | .doc/docx, xls/xlsx. .csv | <1GB | |
| 2.1.II-Notes | Lab notes | Generate new data | Physical | Compiled/aggregated data | | | around 4 lab notebooks |
| 2.2-Manuscripts | | Generate new data | Digital | Compiled/aggregated data | .doc/docx, .tex | <1GB | |
| 2.3-Algorithms | Codes | Generate new data | Digital | Compiled/aggregated data | .r , .py | <1GB | |
| 3.1-Sequences | DNA and protein sequences | Generate new data | Digital | Compiled/aggregated data | .sam, bam, .ab1, .fasta/fa, .qual, gb/gbk, .dna | <100GB | |

**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 aplicable for this project

**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. Particularly for these cases:
Dataset 1.3-Omics data
Dataset 1.4-Strains
Dataset 2.1-Research documentation
Dataset 2.3-Algorithms and scripts
Dataset 3.1 Nucleic acid sequences
Dataset 3.2 Protein sequences

But, in general, the other datasets can be used as well for a patent application.

**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
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 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.

**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

## 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

**How will the data be backed up?**

Biological material: Strains will be backed up in a compressed form (96-well plates) in glycerol on a 2nd location at KU Leuven.

Experimental results: Data will be also saved physically in hard drives in case cloud storage will suffer from losses.

**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 counts with enough -80°C freezers
Experimental results: There is sufficient space in the Dropbox business account.

**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 strains because they are securely stored inside the lab and there is a lock for the freezer.

Experimental results: Data will be saved following secure data transfer, modern data encryption standards and encrypted block storage.

**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)?**

Same place as mentioned before hand:
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

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

All datasets will be stored on the university's central servers with automatic back-up procedures for at least 10 years, conform the KU Leuven RDM policy. The costs (€105 per TB per year for "Large volume-storage" ) 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, …)

Dataset 1.1-Pictures and images: Open
Dataset 1.2-Cytometry: Open
Dataset 1.3-Omics data: both Open and Restricted
Dataset 1.4-Strains: both Open and Restricted
Dataset 2.1-Research documentation: both Open and Restricted
Dataset 2.2 Manuscripts: Open
Dataset 2.3-Algorithms and scripts: Open
Dataset 3.1 Nucleic acid sequences: both Open and Restricted
Dataset 3.2 Protein sequences: both Open and Restricted

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

Institutional access only, but it can be different, depending on each case (after approval).

**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

For datasets 1.3 Omics data and 1.4-Strains: Several of the strains created and/or used during the project are industrially used (which cannot be freely distributed) and/or have potential or future industrial aplications.

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

Electronical data: 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).
Physical data: All published vectors and the associated sequences will be sent to the non-profit plasmid repository Addgene, which will take care of vector storage and shipping upon request.

**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)

But it will be case specific, depending the type of data, the intended usage and who asks for it.

**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 already available

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

The KU Leuven RDM 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. For 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.

## 6. Responsibilities

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

The researcher, that is Julian Ernesto Prieto vivas

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

Dries de Vadder

**Who will manage data preservation and sharing?**

Jeroen Cortebeeck and Karin Voordeckers

**Who will update and implement this DMP?**

The researcher: Julian Ernesto Prieto vivas