FWO DMP Template - Flemish Standard Data Management Plan

Project supervisors (from application round 2018 onwards) and fellows (from application round 2020 onwards) will, upon being awarded their project or fellowship, be invited to develop their answers to the data management related questions into a DMP. The FWO expects a **completed DMP no later than 6 months after the official start date** of the project or fellowship. The DMP should not be submitted to FWO but to the research co-ordination office of the host institute; FWO may request the DMP in a random check.

At the end of the project, the **final version of the DMP** has to be added to the final report of the project; this should be submitted to FWO by the supervisor-spokesperson through FWO's e-portal. This DMP may of course have been updated since its first version. The DMP is an element in the final evaluation of the project by the relevant expert panel. Both the DMP submitted within the first 6 months after the start date and the final DMP may use this template.

The DMP template used by the Research Foundation Flanders (FWO) corresponds with the Flemish Standard Data Management Plan. This Flemish Standard DMP was developed by the Flemish Research Data Network (FRDN) Task Force DMP which comprises representatives of all Flemish funders and research institutions. This is a standardized DMP template based on the previous FWO template that contains the core requirements for data management planning. To increase understanding and facilitate completion of the DMP, a standardized **glossary** of definitions and abbreviations is available via the following link.

1. General Project Information	
Name Grant Holder & ORCID	Sara Baco & https://orcid.org/0000-0001-8850-5174
Contributor name(s) (+ ORCID) & roles	Sibylle Vonesch (https://orcid.org/0000-0003-2485-1048) & Supervisor
Project number[1] & title	3E221094 & Dissecting the functional consequences of synonymous mutations using massively parallel precision genome editing
Funder(s) GrantID[2]	11PS824N
Affiliation(s)	X KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: Provide ROR[3] identifier when possible:

[1] "Project number" refers to the institutional project number. This question is optional since not every institution has an internal project number different from the GrantID. Applicants can only provide one project number.

[2] Funder(s) GrantID refers to the number of the DMP at the funder(s), here one can specify multiple GrantIDs if multiple funding sources were used.

[3] Research Organization Registry Community. https://ror.org/

Please provide a short project description	In this project, I will investigate the mechanisms with which synonymous codon substitutions can affect different layers of gene expression and drive phenotypic change. Combining genome editing, functional genomics and bioinformatics, I will engineer naturally occurring and artificial synonymous mutations in the eukaryotic model S. cerevisiae and measure their effect at the RNA and protein level. To do this, I will use a CRISPR-based tool that can engineer and measure the impact of thousands of mutations in parallel recently developed by my host lab. By measuring different phenotypes at the RNA and protein level in high-throughput, this comprehensive study aims to dissect the mechanism underlying the phenotypic impact of synonymous mutations and will shed light on the importance of synonymous codon usage as a determinant of protein expression and function.

2. 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[4].

				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
BIOLOGICAL MATERIAL	Yeast libraries (*S. cerevisiae*) constructed through genome engineering, bacterial	☒ Generate new data ☐ Reuse existing data	☐ Digital ☒ Physical				Stored in Eppendorfs at -80°C as yeast and plasmid library pool in the lab

[4] Add rows for each dataset you want to describe.

	plasmids and plasmid libraries						
EXPERIMENTAL RESULTS	Digital images, FACS data, sequencing data raw and processed, analysis scripts, software	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☒ Software ☐ Other:	☐ .por ☐ .xml ☐ .tab ☒ .csv ☒ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: - gel scans, colony plate pictures, plots - sorting/ analysis plots - FASTQ, BAM, VCF, textfile ☐ NA	☒ < 100 MB ☒ < 1 GB ☒ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
DATA REUSE	For analysis purposes we will use data from published datasets	☒ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☒ Software ☐ Other:	☐ .por ☐ .xml ☐ .tab ☒ .csv ☐ .pdf ☒ .txt ☐ .rtf ☐ .dwg ☐ .tab	☒ < 100 MB ☒ < 1 GB ☒ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	

					☐ .gml ☒ other: published datasets ☐ NA		

Guidance:

Data can be digital or physical (for example biobank, biological samples, …). Data type: Data are often grouped by type (observational, experimental etc.), format and/or collection/generation method.

Examples of data types: observational (e.g. survey results, sensor readings, sensory observations); experimental (e.g. microscopy, spectroscopy, chromatograms, gene sequences); compiled/aggregated data[5] (e.g. text & data mining, derived variables, 3D modelling); simulation data (e.g. climate models); software, etc.

Examples of data formats: tabular data (.por,. spss, structured text or mark-up file XML, .tab, .csv), textual data (.rtf, .xml, .txt), geospatial data (.dwg,. GML, ..), image data, audio data, video data, documentation & computational script.

digital data volume: Please estimate the upper limit of the volume of the data per dataset or data type.

physical volume: Please estimate the physical volume of the research materials (for example the number of relevant biological samples that need to be stored and preserved during the project and/or after).

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 yet applicable.
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, please describe these issues further and refer to specific datasets or data types when appropriate.	☐ Yes, human subject data ☐ Yes, animal data ☐ Yes, dual use ☒ No If yes, please describe:

[5] These data are generated by combining multiple existing datasets.

Will you process personal data[6]? If so, briefly describe the kind of personal data you will use. Please refer to specific datasets or data types when appropriate. If available, add the reference to your file in your host institution's privacy register.	☐ Yes ☒ No If yes: - Short description of the kind of personal data that will be used: - Privacy Registry Reference:
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 ☐ No If yes, please comment: Results obtained in this thesis are of interest for potential industrial biotechnological applications and for the development of therapies in the context of human disease, therefore may result in intellectual properties. This will be decided in cooperation with VIB's IP Management team.
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 to what data they relate and what restrictions are in place.	☐ Yes ☒ No If yes, please explain:
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 to what data they relate and which restrictions will be asserted.	☒ Yes ☐ No If yes, please explain: Results obtained in this thesis are of interest for potential industrial biotechnological applications and for the development of therapies in the context of human disease, therefore may result in intellectual properties. This will be decided in cooperation with VIB's IP Management team.

[6] See Glossary Flemish Standard Data Management Plan

3. 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).	BIOLOGICAL MATERIAL: Yeast strains are stored in a -80°C freezer and as yeast and plasmid library pool in the lab, for at least 10 years after the project ends. Costs are covered by general lab expenses. Unauthorized people do not have access to strains. EXPERIMENTAL RESULTS: (Meta)data will be documented in lab notebooks and digital files will be stored in a Dropbox Business account, 256-bit AES and SSL/TLS encryption. Raw and processed sequencing data and any end values derived from these data will be stored on a server in an ordered structure, and a separate hard drive as third backup. All data will be stored for at least 10 years, conform KU Leuven RDM policy.
Will a metadata standard be used to make it easier to **find and reuse the data**? If so, please specify which metadata standard will be used. If not, please specify which metadata will be created to make the data easier to find and reuse. *Repositories could ask to deliver metadata in a certain format, with specified ontologies and vocabularies, i.e. standard lists with unique identifiers.*	☐ Yes ☒ No If yes, please specify (where appropriate per dataset or data type) which metadata standard will be used: If no, please specify (where appropriate per dataset or data type) which metadata will be created: Text documents and Excel files stored within each experiment folder will respectively contain guidelines describing data collection/analysis methods and all relevant metadata (including experimental conditions, quality control metrics, computational analysis pipelines and their parameters) to ensure the reusability of the data and the reproducibility of any further data generation.

4. Data Storage & Back-up during the Research Project

Where will the data be stored?	☐ Shared network drive (J-drive) ☐ Personal network drive (I-drive) ☒ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☐ Large Volume Storage ☐ Digital Vault ☒ Other: - Dropbox Business account, 256-bit AES and SSL/TLS encryption. Raw and processed sequencing data and any end values derived from these data will be stored on a server in an ordered structure, and a separate hard drive as third backup. All data will be stored for at least 10 years, conform KU Leuven RDM policy. Yeast strains are stored in a -80°C freezer and as yeast and plasmid library pool in the lab, for at least 10 years after the project ends. Costs are covered by general lab expenses. Unauthorized people do not have access to strains.
How will the data be backed up? *What storage and backup procedures will be in place to prevent data loss? Describe the locations, storage media and procedures that will be used for storing and backing up digital and non-digital data during research.*[7] *Refer to institution-specific policies regarding backup procedures when appropriate.*	☐ Standard back-up provided by KU Leuven ICTS for my storage solution ☐ Personal back-ups I make (specify) ☒ Other (specify) - Dropbox Business account, 256-bit AES and SSL/TLS encryption. Raw and processed sequencing data and any end values derived from these data will be stored on a server in an ordered structure, and a separate hard drive as third backup. All data will be stored for at least 10 years, conform KU Leuven RDM policy. Yeast strains are stored in a -80°C freezer and as yeast and plasmid library pool in the lab, for at least 10 years after the project ends. Costs are covered by general lab expenses. Unauthorized people do not have access to strains.

[7] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

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 ☐ No If yes, please specify concisely: - Dropbox Business account, 256-bit AES and SSL/TLS encryption. Raw and processed sequencing data and any end values derived from these data will be stored on a server in an ordered structure, and a separate hard drive as third backup. All data will be stored for at least 10 years, conform KU Leuven RDM policy. - Yeast strains are stored in a -80°C freezer and as yeast and plasmid library pool in the lab, for at least 10 years after the project ends. Costs are covered by general lab expenses. Unauthorized people do not have access to strains. If no, please specify:
How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons? *Clearly describe the measures (in terms of physical security, network security, and security of computer systems and files) that will be taken to ensure that stored and transferred data are safe.* [7]	- Dropbox Business account, 256-bit AES and SSL/TLS encryption. Raw and processed sequencing data and any end values derived from these data will be stored on a server in an ordered structure, and a separate hard drive as third backup. All data will be stored for at least 10 years, conform KU Leuven RDM policy. - Yeast strains are stored in a -80°C freezer and as yeast and plasmid library pool in the lab, for at least 10 years after the project ends. Costs are covered by general lab expenses. Unauthorized people do not have access to strains.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	Costs are covered by general lab expenses.

5. 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 data will be preserved for 10 years according to KU Leuven RDM policy ☐ All data will be preserved for 25 years according to CTC recommendations for clinical trials with medicinal products for human use and for clinical experiments on humans ☐ Certain data cannot be kept for 10 years (explain)
Where will these data be archived (stored and curated for the long-term)?	☐ KU Leuven RDR ☐ Large Volume Storage (longterm for large volumes) ☐ Shared network drive (J-drive) ☒ Other (specifiy): - Dropbox Business account, 256-bit AES and SSL/TLS encryption. Raw and processed sequencing data and any end values derived from these data will be stored on a server in an ordered structure, and a separate hard drive as third backup. All data will be stored for at least 10 years, conform KU Leuven RDM policy. - Yeast strains are stored in a -80°C freezer and as yeast and plasmid library pool in the lab, for at least 10 years after the project ends. Costs are covered by general lab expenses. Unauthorized people do not have access to strains.
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	Costs are covered by general lab expenses.

6. Data Sharing and Reuse	
Will the data (or part of the data) be made available for reuse after/during the project? Please explain per dataset or data type which data will be made available. *Note that 'available' does not necessarily mean that the data set becomes openly available, conditions for access and use may apply. Availability in this question thus entails both open & restricted access. For more information: https://wiki.surfnet.nl/display/standards/info-eu-repo/#infoeurepo-AccessRights*	☒ Yes, as open data ☒ Yes, as embargoed data (temporary restriction) ☐ Yes, as restricted data (upon approval, or institutional access only) ☒ No (closed access) ☒ Other, please specify: I aim for 3 high impact manuscripts with the data generated in this project. 1) Dissecting the mechanism underlying the impact of synonymous mutations at scale. 2) Analysis of the impact of synonymous codon usage and amino acid identity on protein expression in PGK1. 3) Building a machine learning model to predict the effect of synonymous mutations on protein expression. Conform the Open Access publication requirement for FWO, data used in published manuscripts will be openly available.
If access is restricted, please specify who will be able to access the data and under what conditions.	Only lab members can access the data.
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 per dataset or data type where appropriate.	☐ Yes, privacy aspects ☒ Yes, intellectual property rights ☐ Yes, ethical aspects ☐ Yes, aspects of dual use ☐ Yes, other ☒ No If yes, please specify: Results obtained in this thesis are of interest for potential industrial biotechnological applications and for the development of therapies in the context of human disease, therefore may result in intellectual properties. This will be decided in cooperation with VIB's IP Management team.

Where will the data be made available? If already known, please provide a repository per dataset or data type.	☐ KU Leuven RDR ☐ Other data repository (specify) ☐ Other (specify)
When will the data be made available? *This could be a specific date (dd/mm/yyyy) or an indication such as 'upon publication of research results'.*	☒ Upon publication of research results ☐ Specific date (specify) ☐ Other (specify)
Which data usage licenses are you going to provide? If none, please explain why. *A data usage license indicates whether the data can be reused or not and under what conditions. If no licence is granted, the data are in a grey zone and cannot be legally reused. Do note that you may only release data under a licence chosen by yourself if it does not already fall under another licence that might prohibit that.* *Example Answer: E.g. "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."* [8]	☐ CC-BY 4.0 (data) ☐ Data Transfer Agreement (restricted data) ☐ MIT licence (code) ☐ GNU GPL-3.0 (code) ☐ Other (specify)
Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, please provide it here. *Indicate whether you intend to add a persistent and unique identifier in order to identify and retrieve the data.*	☒ Yes, a PID will be added upon deposit in a data repository ☐ My dataset already has a PID ☐ No

[8] Source: Ghent University Generic DMP Evaluation Rubric: https://osf.io/2z5g3/

What are the expected costs for data sharing? How will these costs be covered?	No extra costs

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Sara Baco & Sibylle Vonesch, assisted by Célie Cokelaere (lab manager)
Who will manage data storage and backup during the research project?	Sara Baco & Sibylle Vonesch, assisted by Célie Cokelaere (lab manager)
Who will manage data preservation and sharing?	Sara Baco & Sibylle Vonesch, assisted by Célie Cokelaere (lab manager)
Who will update and implement this DMP?	Sara Baco