Mutational robustness and evolvability in model microbes

A Data Management Plan created using DMPonline.be

Creators: Jan Michiels, Kevin Verstrepen, Sander Govers, Piet van den Berg, Natalie Verstraeten, Karin Voordeckers

Affiliation: KU Leuven (KUL)

Template: KU Leuven BOF-IOF

Principal Investigator: Jan Michiels, Kevin Verstrepen, Sander Govers, Piet van den Berg

Grant number / **URL:** C16/23/007

ID: 202713

Start date: 01-10-2023

End date: 30-09-2029

Project abstract:

Organisms need to balance the need for mutations to evolve with mitigating these mutations' potential negative effects. One hypothesis is that some cellular mechanisms help minimize the phenotypic consequences of mutations and hence confer mutational robustness. However, whereas the concept of mutational robustness is central to our understanding of evolution and genetics, our knowledge about it remains surprisingly limited. This is because studying robustness is extremely challenging: it requires interrogating the role of a large number of genes on the effect of a large number of (de novo or standing) mutations on several traits and in several environments. In this project, we combine the power of two model organisms with an evolutionary modelling framework to identify genes and mechanisms associated with reducing mutational effects, assess the type of mutations that can be mitigated and model and test if and how mutational robustness influences evolvability.

Last modified: 14-03-2024

Mutational robustness and evolvability in model microbes

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.

Dataset name / ID	Description	New or reuse	Digital or Physical data	Data Type	File format	Data volume	Physical volume
		*Indicate: **N**(ew data) or **E**(xisting data)*	Indicate: **D**(igital) or **P**(hysical)	Indicate: **A**udiovisual **I**mages **S**ound **N**umerical **T**extual **M**odel **SO**ftware Other (specify)		Indicate: <1GB <100GB <1TB <5TB >5TB NA	
WP1. Material - EMS yeast strains	EMS mutagenized yeast deletion collection	N	P	/	/	/	~4500 strains, stored in 96 well plates at -80°C
WP1. Material - re-engineered yeast mutants	re-engineered deletion strains + EMS mutagenized variants	N	P	/	/	/	20 cryovials stored at -80°C
WP1. Dataset - growth EMS yeast strains	microscopy images of growth EMS mutagenized yeast strains	N	D	I	.mov, .jpeg, .tiff	< 100GB	/
WP1. Dataset - cell morphology EMS yeast strains	microscopy images of stained, EMS mutagenized yeast cells	N	D	I	.tiff	<5TB	/
WP2. Material - EMS *E. coli* strains	EMS mutagenized *E. coli* Keio deletion collection	N	P	/	/	/	~4000 strains, stored in 96 well plates at -80°C
WP2. Material - *E. coli* clones with known mutational load	*E. coli* clones created using DinB error-prone DNA polymerase	N	P	/	/	/	20 cryovials stored at -80°C
WP2. Material - sgRNA library in *E. coli* clones with known mutational load (including WT)	clones for CRISPRi screening, created by transferring sgRNA library to dCas9 strains	N	P	/	/	/	30 cryovials stored at -80°C
WP2. Material - in-frame deletions and CRISPRi mutants of potential buffer genes	validation of identified buffer genes in 10 randomly mutagenized backgrounds	N	P	/	/	/	~200 strains, stored in 96 well plates at -80°C
WP2. Dataset - growth EMS *E. coli* strains	microscopy images of growth EMS mutagenized *E. coli* strains	N	D	I	.mov, .jpeg, .tiff	< 100GB	/

WP2. Dataset - cell morphology EMS *E. coli* strains	microscopy images of stained, EMS mutagenized *E. coli* cells	N	D	I	.tiff	<5TB	/
WP2. Dataset - CRISPRi screening data	sgRNAs counts following growth of clones in selective conditions	N	D	/	sam, bam, .ab1, .fasta/fa, .qual, gb/gbk, .dna	<1G	/
WP3&7. Dataset - digital oligos	List of 10.000 variants and associated oligos for editing	N	D	T	.txt	<1G	/
WP3&7. Material - oligopools	oligopools to create specific mutations (linked to barcodes)	N	P	/	/	/	12 eppendorf tubes
WP3&7. Material - plasmid libraries	plasmid libraries (stored in *E. coli*) created from oligopools, containing guide, donor, barcode	N	P	/	/	/	12 cryovials stored at -80°C
WP3&7. Material - yeast strains	pools of yeast strains containing mutations introduced with MAGESTIC	N	P	/	/	/	~600 cryovials stored at -80°C
WP3&7. Data - sequencing data	sequencing data to determine fitness of different variants	N	D	/	sam, bam, .ab1, .fasta/fa, .qual, gb/gbk, .dna	<5TB	/
WP4. Material - yeast strains	natural yeast strains with specific robustness gene deleted	N	P	/	/	/	480 strains, stored in 5 deepwell plates at -80°C
WP4. Data - plate images	images of agar plates with yeast colonies	N	D	I	.jpeg, .tiff	<1GB	/
WP4. Dataset - cell morphology of natural yeast strains with specific robustness gene deleted	microscopy images of stained yeast cells	N	D	I	.tiff	<5TB	/
WP5&8. Dataset - digital oligos	List of 10.000 variants and associated oligos for editing	N	D	T	.txt	<1G	/
WP5&8. Material - *E. coli* strains	*E. coli* libraries created using CREATE technology: 10.000 variants in WT and 5 mutants of buffer genes	N	P	/	/	/	~100 cryovials stored at -80°C
WP5&8. Data - sequencing data	sequencing data to determine fitness of different variants	N	D	/	sam, bam, .ab1, .fasta/fa, .qual, gb/gbk, .dna	<5TB	/
WP6. Material - *E. coli* strains	natural *E. coli* strains with specific robustness gene deleted	N	P	/	/	/	480 strains, stored in 5 deepwell plates at -80°C

WP6. Data - plate images	images of agar plates with *E. coli* colonies	N	D	I	.jpeg, .tiff	<1GB	/
WP6. Dataset - cell morphology of natural *E. coli* strains with specific robustness gene deleted	microscopy images of stained *E. coli* cells	N	D	I	.tiff	<5TB	/
WP9&10. Scripts - evolutionary modelling	Python scripts of the evolutionary simulation work	N	D	SO	.py	<1GB	/
WP9&10. Dataset - evolutionary modelling	output file of evolutionary simulation work	N	D	N	.txt, .csv	<5TB	/
WP10. Material - yeast strains	yeast strains with different expression levels of robustness genes	N	P	/	/	/	15 strains, stored in cryovials at -80°C
WP10. Material - *E. coli* strains	*E. coli* strains with different expression levels of robustness genes	N	P	/	/	/	15 strains, stored in cryovials at -80°C
WP10. Data - growth curves	growth curves of yeast populations during and after experimental evolution	N	D	/	.csv	<5GB	/
All WP- Digital images	gel scans (from e.g. PCRs), figures, graphs, illustrations ...	N	P	I	.tiff, .jpeg,.svg, .pdf	<10GB	/
All WP - Research documentation	lab protocols and lab notebooks	N	D&P	/	.doc, .pdf	<1GB	/
All WP - Manuscripts	manuscripts	N	D	/	.doc, .pdf	<1GB	/
All WP - Scripts	new and existing scripts using existing packages, written in R to analyze and plot obtained data	N&E	D	/	.r	<1GB	/

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:

- We will perform EMS mutagenesis on the haploid yeast gene knockout collection (Brachmann *et al.* (1998) *Yeast*, 14: 115-132). Similarly, EMS mutagenesis will be performed on the *E. coli* Keio knockout collection (Baba *et al.* (2006). *Molecular Systems Biology*, 2: 2006.0008). Both collections are available in-house.
- A genome-scale CRISPRi library targeting each protein- and RNA-coding gene of *E. coli* (~15 sgRNA targets/gene) is also available in the Michiels lab (Wang *et al.* (2018) *Nature Communications* 9: 2475).
- To analyze and plot obtained data we will use new and existing scripts, using existing packages written in R.

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, refer to specific datasets or data types when appropriate and provide the relevant ethical approval number.

- No

Will you process personal data? If so, please refer to specific datasets or data types when appropriate and provide the KU Leuven or UZ Leuven privacy register number (G or S number).

- 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

Potential tech transfer will be discussed with the KU Leuven Research & Development - Tech Transfer Office. Valorization potential includes licensing of (improved) strains or information on linking a specific sequence variant to a phenotype.

Do existing 3rd party agreements restrict exploitation or dissemination of the data you (re)use (e.g. Material or 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

Existing agreements between VIB and KU Leuven do not restrict 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.

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.

- Yes

Materials requested from other labs (e.g. reporters needed to perform the phenotypic characterization described in T6.1) might be subject to MTAs. This will be done in consultation with our host institution's legal departments to minimize restrictions on the use of these materials.

Documentation and Metadata

Clearly describe what approach will be followed to capture the accompanying information necessary to keepdata 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:** Cryotubes and multi-well plates 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, 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).
- **Scripts:** All scripts for producing evolutionary simulations and figures will be properly annotated so that the code is understandable and can be used to re-generate the results. After publication or upon submission of manuscripts for publication, all scripts will be uploaded in dedicated data repositories (see below), including a readme file that explains what each script is exactly used for.

Will a metadata standard be used to make it easier tofind 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.

- 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… 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 the various data files and give a broad overview of the analyses 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.

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
- Large Volume Storage
- Other (specify below)

- **Biological material:** Cryotubes and multi-well plates will be stored in -80° freezers with restricted access.
- **Experimental and evolutionary simulation results:** An E-Notebook will be used to collect data. Low-volume data, protocols and analyses will subsequently be stored in secure and internally shared folders on university servers with built-in backup and versioning (SharePoint). Although built-in backup systems are in place, password-protected hard drives equipped with anti-virus programs will be used as backup. A network drive will also be used for large-scale data (e.g. NGS data and microscopy data). A copy of these datasets will be made to desktop PCs with large computational power (or to a computing cluster of our host institution) whenever data analyses will be performed. For final datasets (including evolutionary simulation results) 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 or the KU Leuven GitLab server (https://gitlab.kuleuven.be/).

How will the data be backed up?

- Standard back-up provided by KU Leuven ICTS for my storage solution
- Other (specify below)

- **Biological material:** A backup of critical strains will be stored in the different host labs involved in the project.
- **Experimental and evolutionary simulation results:** Data will be stored on KU Leuven’s central servers with automatic daily back-up and version control procedures.

Is there currently sufficient storage & backup capacity during the project?

If no or insufficient storage or backup capacities are available, explain how this will be taken care of.

- Yes

- **Biological material:** Sufficient storage is available.
- **Experimental and evolutionary simulation results:** OneDrive at KU Leuven 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.

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 and evolutionary simulation results:** E-notebooks are password protected and data stored in KU Leuven's secure environments are secured by a two factor authorization and frequently changed passwords. External HDD are password-protected and stored in the safety of the labs.

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

- **Biological material:** -80° freezers are currently present in the host labs (costs are covered by general lab expenses).
- **Experimental and evolutionary simulation results:** The costs for large volume storage are limited and covered by general lab expenses.

Data Preservation after the end of the Research Project

Which data will be retained for 10 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

Where will these data be archived (stored and curated for the long-term)?

- KU Leuven RDR
- Other (specify below)

- **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 different host labs. Relevant strains will also be deposited in a public repository (e.g. the Belgian Coordinated Collections of Micro-organisms (BCCM)).
- **Experimental and evolutionary simulation results:** Data will in first instance be stored on KU Leuven central servers, and, after publication, data will additionally indefinitely be stored in open access repositories (e.g. Zenodo, Mendeley Data, KU Leuven's RDR). Dedicated repositories will be used for specific datatypes e.g. SRA for NGS data.

What are the expected costs for data preservation during the expected retention period? How will these 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 and evolutionary simulation results:** The costs will be covered by general lab budgets.

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.

- Other (specify below)

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 KU Leuven LRD.
Unpublished, essential data will be available to (future) lab members via internal IT provisions.

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

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 KU Leuven LRD.
Unpublished, essential data will be available to (future) lab members via internal IT provisions.

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.

- No

Where will the data be made available?

If already known, please provide a repository per dataset or data type.

- Other (specify below)

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.

- **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.
- **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.
- **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 notebooks are accessible to the PI and the research staff involved in the project, and will be made available upon request.
- **Algorithms and scripts:** As soon as a manuscript is publicly available, algorithms and scripts will be deposited in a GitHub repository.
- **Datasets:** Datasets (including those of evolutionary simulations studies) will be deposited in open access repositories.
- **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) and the Protein Database (for protein sequences).

When will the data be made available?

- Upon publication of research results
- Other (specify below)

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.

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

Which data usage licenses are you going to provide?

If none, please explain why.

- CC-BY 4.0 (data)

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 be conform 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).

Do you intend to add a persistent identifier (PID) to your dataset(s), e.g. a DOI or accession number? If already available, please provide it here.

- Yes, a PID will be added upon deposit in a data repository

What are the expected costs for data sharing? How will these 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. A budget for publication costs has been requested in this project.

Responsibilities

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

All researchers will regularly upload their data, protocols and metadata on the E-notebook and university servers, and regular back-ups will be additionally made on external hard disks (password protected and encrypted). The final responsibility for data documentation & metadata will be with the WP leaders:

- WP1. Verstrepen, Govers
- WP2. Michiels, Govers
- WP3. Verstrepen, van den Berg
- WP4. Verstrepen, Govers
- WP5. Michiels, van den Berg
- WP6. Michiels, Govers
- WP7. Verstrepen, Govers
- WP8. Michiels, Govers
- WP9. van den Berg

- WP10. van den Berg, Michiels

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

WP leaders will be responsible for data storage & back up during the project:

- WP1. Verstrepen, Govers
- WP2. Michiels, Govers
- WP3. Verstrepen, van den Berg
- WP4. Verstrepen, Govers
- WP5. Michiels, van den Berg
- WP6. Michiels, Govers
- WP7. Verstrepen, Govers
- WP8. Michiels, Govers
- WP9. van den Berg
- WP10. van den Berg, Michiels

Who will manage data preservation and sharing?

WP leaders will be responsible for data preservation and reuse:

- WP1. Verstrepen, Govers
- WP2. Michiels, Govers
- WP3. Verstrepen, van den Berg
- WP4. Verstrepen, Govers
- WP5. Michiels, van den Berg
- WP6. Michiels, Govers
- WP7. Verstrepen, Govers
- WP8. Michiels, Govers
- WP9. van den Berg
- WP10. van den Berg, Michiels

Who will update and implement this DMP?

Jan Michiels bears the end responsibility of updating & implementing this DMP.