EXCELLGEN - FWO

A Data Management Plan created using DMPonline.be

Creators: Kristina Nesporova, Sander Govers

Affiliation: KU Leuven (KUL)

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

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1251624N

ID: 205792

Start date: 01-11-2023

End date: 31-10-2026

Project abstract:
Extraintestinal pathogenic *E. coli* (ExPEC) are one of the greatest challenges for the healthcare system, with a frightening prognosis for the future. Despite being heavily studied in recent years, no unifying principle explaining their success has been identified. Only limited attention has been paid to the additional genomic properties of ExPEC strains that are not virulence factors or antimicrobial resistance genes. I aim to demonstrate that these additional genomic features lead to alterations in

Last modified: 24-04-2024

EXCELLGEN - FWO

DPIA

DPIA

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

- Not applicable

EXCELLGEN - FWO

GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

EXCELLGEN - FWO

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 project will generate various types of data/material involving:
• Wild-type strains and deletions strains
• NGS data and results from their analysis
• Raw data from experimental assays (e.g., growth curves)
• Imaging data coming from light microscopy
• Scientific manuscript submitted to preprint servers and international peer-reviewed journals (open access)
• Potentially varolizable results (e.g., new therapeutic targets) will be submitted for priority filing through the KU Leuven Technology transfer office.

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)

Prof Govers will be responsible for data management as the supervisor of the project.
The data and materials will be stored for at least 5 years following the ending of the project:
• Bacterial strains will be stored at -80C. They will be available upon request and stored also as a part of the bacterial collections of other labs.
• Digitalised data will be stored in triplicate: personal computer, backup disc, and a shared lab repository. Handwritten notes will be kept in registered lab books that are stored for several years.
• The accessibility of the data will be ensured by depositing the produced results in the Research Data Repository (RDR) of KU Leuven.

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)

NA

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)

The project takes advantage of genomes and metadata stored in public databases. These include patient-related data such as gender or type of infection. The project will also use 100 ExPEC strains from collaborating laboratories together with their metadata. However, in both cases, the data are already anonymized so no ethical issue in that regard is connected to the project as designed

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

NA

EXCELLGEN - FWO

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
ExPEC isolates, laboratory control strains, manipulated strains and relevant metadata

☒ Generate new data
☒ Reuse existing data
☒ Digital
☒ Physical
☒ Observational
☒ Experimental
□ Compiled/ aggregated data
□ Simulation data
□ Software
☒ Other
□ NA
□ .por
□ .xml
☒ .tab
☒ .csv
☒ .pdf
☒ .txt
□ .rtf
□ .dwg
□ .tab
□ .gml
□ other:
□ NA
□ < 100 MB
☒ < 1 GB
□ < 100 GB
□ < 1 TB

☐ < 5 TB
☐ < 10 TB
☐ < 50 TB
☐ > 50 TB
2 L
Genome sequencing data
Raw and assembled data .fasta format
☒ Generate new data
☒ Reuse existing data
☒ Digital
☐ Physical
☐ Observational
☒ Experimental
☐ Compiled/ aggregated data
☐ Simulation data
☐ Software
☐ Other
☐ NA
☐ .por
☐ .xml
☐ .tab
☐ .csv
☐ .pdf
☐ .txt
☐ .rtf
☐ .dwg
☐ .tab
☐ .gml
☒ other: fasta, fastq, fast5
☐ NA
☐ < 100 MB
☐ < 1 GB
☐ < 100 GB
☐ < 1 TB
☐ < 5 TB
☒ < 10 TB
☐ < 50 TB
☐ > 50 TB
☐ NA

Microscopy images and photographs
Digital images
Microscopy images, gel scans, photographs from the experimental setting
☒ Generate new data
☐ Reuse existing data
☒ Digital
☐ Physical

☐ Observational
☒ Experimental
☐ Compiled/ aggregated data
☐ Simulation data
☐ Software
☐ Other
☐ NA

☐ .por
☐ .xml
☐ .tab
☐ .csv

☐ .pdf
☐ .txt
☐ .rtf
☐ .dwg
☐ .por
☐ .xml
☐ .tab
☐ .csv
☐ .pdf
☐ .txt
☐ .rtf
☐ .dwg
☐ .tab
☐ .gml
☒ other: tif
☐ NA

☐ < 100 MB
☐ < 1 GB
☐ < 100 GB
☐ < 1 TB
☐ < 5 TB
☐ < 10 TB
☒ < 50 TB
☐ > 50 TB
☐ NA

Manuscripts, tables and figures
Derived and compiled data
Manuscripts

☒ Generate new data
☐ Reuse existing data

☒ Digital
☐ Physical

☒ Observational
☒ Experimental
☒ Compiled/ aggregated data
☐ Simulation data
☐ Software
☐ Other
☐ NA

☐ .por
☐ .xml
☒ .tab
☒ .csv
☒ .pdf
☒ .txt
☐ .rtf
☐ .dwg
☐ .tab
☐ .gml
☐ other:

□ NA

□ < 100 MB
☒ < 1 GB
□ < 100 GB
□ < 1 TB
□ < 5 TB
□ < 10 TB
□ < 50 TB
□ > 50 TB
□ NA

Raw data from experimental assays (e.g., growth curves measurements)

☒ Generate new data
□ Reuse existing data
☒ Digital
□ Physical
☒ Observational
☒ Experimental
☒ Compiled/ aggregated data
□ Simulation data
□ Software
□ Other
□ NA

□ .por
□ .xml
☒ .tab
☒ .csv
☒ .pdf
☒ .txt
□ .rtf
□ .dwg
□ .tab
□ .gml
□ other:
□ NA

□ < 100 MB
□ < 1 GB
☒ < 100 GB
□ < 1 TB
□ < 5 TB
□ < 10 TB
□ < 50 TB
□ > 50 TB
□ NA

Algorithms and scripts:
R and bash scripts used for data analysis.

☒ Generate new data
☒ Reuse existing data

☒ Digital
☐ Physical

☐ Observational
☐ Experimental
☐ Compiled/ aggregated data
☐ Simulation data
☒ Software
☐ Other
☐ NA

☐ .por
☐ .xml
☒ .tab
☒ .csv
☒ .pdf
☒ .txt
☐ .rtf
☐ .dwg
☐ .tab
☐ .gml
☒ other:
☐ NA

☐ < 100 MB
☒ < 1 GB
☐ < 100 GB
☐ < 1 TB
☐ < 5 TB
☐ < 10 TB
☐ < 50 TB
☐ > 50 TB
☐ NA

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:

The physical isolates will originate from our collaborators with already anonymized metadata.
The EnteroBase collection of the E. coli genomes and related metadata is publicly accessible:
https://enterobase.warwick.ac.uk/

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

The patient data related to the used strains were already anonymized before the start of the project.

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

There is a potential for valorization of results produced in WP4 (identifying a prominent drug target).

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).

We will accompany our datasets with metadata and documentation whenever it is necessary/appropriate. We will update metadata from EnteroBase (specifying the "pathogenic" status of E. coli genomes based mainly on the source of isolation - which is information present in existing EnteroBase metadata). We will create documentation and metadata describing the process of used analyses (readme .txt files) and conditions used for experiments/analyses. We will provide documentation (readme files) describing the code produced within the Excellgen project. The metadata will follow standard vocabulary and structure following the repository where we will submit them (e.g., Research Data Repository – RDR of KU Leuven, or NCBI and EnteroBase for genomic data).

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.

- Yes

Fastq files store metadata for genomic data.
EnteroBase temples provide a unified structure for bacterial genomic metadata.

Format of .nd2 files stores metadata of microscopy images based on Nikon metadata standards.
Data without specific metadata standards will follow the Dublin Core Metadata codex.

3. DATA STORAGE & BACK-UP DURING THE RESEARCH PROJECT

Where will the data be stored?

The digital data will be stored on the working laptop. Analysis of genomic data produced on The Flemish Supercomputer Centre (VSC) will be initially stored there and later transferred to the working laptop for further processing. All relevant scripts and software code used in the project will be additionally deposited in GitHub. All genomic data will also be deposited in NCBI or EnteroBase. When appropriate all the data, metadata, and documentation relevant to a specific manuscript will be deposited at RDR to be accessible at the time of publication release.
Bacterial strains will be stored as a glycerol stock in freezers (-80°C).

How will the data be backed up?

The digital data stored on the working laptop are automatically backed up by the university's built-in backup system.
The data will also be regularly (every month) backed up on a hard drive.
The strains will be frozen in a minimum of two copies and stored in two different freezers.

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

There is sufficient storage capacity within the personal drive (5 TB) on the KU Leuven working laptop which is automatically backed up. In case more space is needed, additional hard drives will be purchased to have all the data backed up in a minimum of two copies.

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The data access will require authentication via KU Leuven login credentials and password and extra authentication can be activated for sensitive datasets. The hard drives with backups will be stored within a space requiring a minimum of two keys for access (e.g. locked office and locked drovers).

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

The costs to purchase the hardware to produce, store, and back up data properly, to purchase additional space on RDR, and to purchase computational time on HPC from Vlaams Supercomputer Centrum will be covered by the bench fee related to the Excellgen project.

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...).

The relevant datasets will be stored for 5 years minimally. KU Leuven policy requires the storage of some types of data for 10 years, therefore, most of the datasets will be stored and accessible for this period.

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

The data will be mostly stored in RDR. RDR is an open-access repository (unless the access is restricted for specific cases).
The genomic data (fasta or fastq, together with metadata) from 100 ExPEC strains will be reposited and made openly accessible in well-established genomic repositories such as NCBI or EnteroBase in case they have not been reposited before this project (we are reusing strains from multiple projects of our collaborators and for some strains, the NGS data has been already made publicly accessible).

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

We expect to exceed the free storage limit for RDR (50 GB per year), yet the pricing for additional data should not exceed 200 euros/year. This cost will be covered by the bench fee of the Excellgen project during the project duration and later from funds accessible to Prof. Govers (the project´s supervisor).

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

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

Not applicable.

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.

- No

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

The final data will be reposited on RDR. The RDR enables data accessibility and provides different options for accessibility limitations in case of sensitive data (e.g. if potentially valorizable results would be created during the project).

When will the data be made available?

The data will be made available once the relevant manuscript is published.

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

Data from the project that can be shared openly will be made available under an appropriate type of Creative Commons Attribution license (e.g., CC-BY 4.0).

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

The DOI will be provided for the relevant datasets.

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

The main cost is of data sharing will be related to covering the open-access option for publishing the manuscripts. The estimation for the cost is 3500 USD (based on Elsevier Open Access charge, 2024) per manuscript, therefore 7000 USD for two expected manuscripts. These costs will be covered by the Excellgen bench fee budget.

6. RESPONSIBILITIES

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

The metadata and documentation will be managed by the fellow and the technical staff at the time of data collection.

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

The storage and the backup of data will be managed by the fellow.

Who will manage data preservation and sharing?

The supervisor of the project will be responsible for data preservation and sharing.

Who will update and implement this DMP?

The DMP will be updated and implemented by both, the fellow and the supervisor of this project, with the supervisor being the responsible person.