FWO DMP Template - Flemish Standard Data Management Plan

Version KU Leuven

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	**Atilgan Yilmaz** 0000-0003-0309-1588
Contributor name(s) (+ ORCID) & roles	**Eslam Mohammed, postdoc** **Margaux van Puyvelde, PhD student**
Project number [1] & title	G0DCO23N - Essential gene circuits regulating cell state transitions and diseases of skeletal muscle
Funder(s) GrantID [2]	G0DCO23N
Affiliation(s)	X KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☐ Other: ROR identifier KU Leuven: 05f950310

[1] "Project number" refers to the institutional project number. This question is optional. 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.

Please provide a short project description	Skeletal muscle makes up the largest tissue mass in the human body. It is also prone to a vast array of genetic and metabolic disorders, as well as physical trauma. In addition, age-related muscle wasting has become an increasingly common health problem in today's aging societies. Therefore, enhacing the regenerative capacity of skeletal muscle is a particularly crucial aim to treat these conditions. A key step towards this goal is to understand how muscle cell identity is established in human and which genes govern this process. Despite the previous efforts showing a handful of key regulators, a comprehensive characterization of gene networks that are essential for muscle cell identity remains elusive. Thanks to the advances in stem cell biology and gene editing technologies, we recently devised stem cell-based experimental tools that we used to identify essential genes for different cell types in early embryonic stages in human. This research proposal aims to utilize this unique and robust experimental platform to identify the essential genes for muscle cell identity and also the genes that modify the severity of the two most common genetic muscle disorders, namely the Duchenne Muscular Dystrophy (DMD) and Myotonic Dystrophy Type 1 (DM1). These studies will provide a comprehensive understanding of how muscle cells are made in human, giving us a better control over theregenerative capacity of this tissue, while also potentially revealing novel drug targets for DMD and DM1.

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 [3].

				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
Digital images	Phase contrast and fluorescence microscope images, time-lapse movies, gel scans, graphs, schematic illustrations	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	.png, .tiff, .jpg, .mp4, .avi, .pdf, .ai, .pptx,	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ > 5 TB ☐ NA	
Sequencing data	Bulk and single cell RNA sequencing (next-generation sequencing (NGS)) data, NGS data for custom-made DNA libraries, Sanger sequencing data	☒ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☒ Other:	.sam, .bam, .tsv (.gz), .mtx, .loom, .rds (.gz), .bcl, .fastq (.gz), .csv, .txt, .xlsx, .R, .py	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ > 5 TB ☐ NA	

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

Cytometry data	Flow cytometry and fluorescence-activated cell sorting (FACS) data	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☐ Numerical ☐ Textual ☐ Model ☐ Software ☐ Other:	.fcs, .ai, .ppt, .pdf	☐ < 1 GB ☐ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
Algorithms and scripts	Codes written and reused from existing pipelines for NGS analyses	☒ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☐ Images ☐ Sound ☒ Numerical ☐ Textual ☐ Model ☒ Software ☐ Other:	R and python scripts	☐ < 1 GB ☐ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ > 5 TB ☐ NA	
Nucleic acid and cell culture samples	Nucleic acid libraries for NGS, plasmid libraries encoding gRNAs, newly cloned plasmids for gRNAs or homology-directed recombination, frozen cell pellets and frozen cell vials	☒ Generate new data ☒ Reuse existing data	☐ Digital ☒ Physical	NA	NA	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ > 5 TB ☒ NA	Tubes containing the frozen cell vials are stored at -150 ^{0}C freezers, frozen cell pellets are stored at -80 ^{0}C freezers, NGS libraries and plasmids are kept in 20 ^{0}C freezers. An electronic database is used to keep track of the physical storage

							places of the samples.

GUIDANCE:
The data description forms the basis of your entire DMP, so make sure it is detailed and complete. It includes digital and physical data and encompasses the whole spectrum ranging from raw data to processed and analysed data including analysis scripts and code. Physical data are all materials that need proper management because they are valuable, difficult to replace and/or ethical issues are associated. Materials that are not considered data in an RDM context include your own manuscripts, theses and presentations; documentation is an integral part of your datasets and should described under documentation/metadata.
RDM Guidance on data

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 following published datasets will be reused: https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE87365 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE111163 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE147457 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE121154 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE129505 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE212366 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE165075 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE46633 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE161025 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE234616 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE221912 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE178784 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE105211 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE93263 https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA610985 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE158216 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE214495 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE236120 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1527072 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE128844 https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE98509

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.	☒ Yes, human subject data; provide SMEC or EC approval number: EC approval S66794 ☐ Yes, animal data; provide ECD reference number: ☐ Yes, dual use; provide approval number: ☐ No Additional information:
Will you process personal data[4]? 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).	☐ Yes (provide PRET G-number or EC S-number below) ☒ No Additional information:
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: We do not exclude the possibility that the research project could lead to findings with potential for commercial valorization. This applies to the results of the genetic screens that will be performed throughout the project. These include the sequencing data, in particular the NGS datasets for custom-made DNA libraries. In case the commercial valorisation becomes relevant, the invention will be assessed in the context of KU Leuven policies and with the help of KU Leuven Research and Development office and will be IP protected without jeopardizing the publication of the results of the project.
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:

[4] See Glossary Flemish Standard Data Management Plan

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:

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). *RDM guidance on documentation and metadata*.	To keep data understandable and usable, we will document and store the data generated throughout the project in an electronic laboratory notebook using our Benchling account. For previously published NGS datasets, we will generate a shared folder in our KU Leuven OneDrive storage space, where we organize the URLs of the source data, the names, dates, authors of the associated manuscripts, the type of NGS datasets and the species of the samples used in the study. We will keep a record of experimental protocols, the digital map of the freezers where experimental materials are kept in numbered boxes, data files (raw and processed), analysis scripts (R and Python -as Jupyter Notebooks), observations during the experiments. NGS samples will be stored for up to 5 years after the end of the project.

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: Next Generation Sequencing data is required to be deposited to public repository such as GEO, SRA, ArrayExpress or ENA at the time of the publication. The data that has not yet been submitted to these databases will be standardized using the metadata schemes such as Dublin Core, DataCite or Genome Metadata. This standardization will include a title, the name and the affiliation of the creator, date and time references, the subject and a text describing the contents of the dataset and its analysis pipeline, the format of the file, the category of the data such as NGS datasets, images or audio/video files, an identifier and access rights. In any data deposition case, a readme.txt document will be added to the initial directory to include all of the above-mentioned information to allow easy sharing, reusing and interpretation of the data in future. If no, please specify (where appropriate per dataset or data type) which metadata will be created:

4. Data Storage & Back-up during the Research Project	
Where will the data be stored? *Consult the interactive KU Leuven storage guide to find the most suitable storage solution for your data.*	☒ Shared network drive (J-drive) ☐ Personal network drive (I-drive) ☒ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☐ Large Volume Storage ☐ Digital Vault ☐ Other:

How will the data be backed up? *WHAT STORAGE AND BACKUP PROCEDURES WILL BE IN PLACE TO PREVENT DATA LOSS?*	☒ Standard back-up provided by KU Leuven ICTS for my storage solution ☒ Personal back-ups I make (specify) ☐ Other (specify)
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 KU Leuven ICTS has sufficient storage options, which are scalable and variable. The Staging and Archive on VSC are also scalable. 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.* *Guidance on security for research data*	The J drive will only be available to the members of the laboratory. Researchers working on the project have the control over who can access the files in their personal or shared OneDrive folders. The VSC storage is reached by personal accounts in a space accessible to group members. The ICTS data center at KU Leuven ensures a mirror network storage as a back-up strategy and easy recovery of the data.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	The current cost of J Drive is 51.9 Euros/100Gb/Year. OneDrive accounts are free of charge to KU Leuven personnel. VSC staging costs 130 Euros/TB/Year. The total estimated cost for data storage throughout the project will be ~6500 Euros. These costs are budgeted on this grant and in the grants of related projects, whenever relevant and possible.

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...). *Guidance on data preservation*	☐ 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) NGS data will be made publicly available upon the publication of the manuscripts. Unpublished NGS data will be made available upon request 5 years after the end of the project. Since the sample quality may drop in time, physical data including the NGS samples and cell pellets will be stored until 5 years after the end of the project.
Where will these data be archived (stored and curated for the long-term)? *Dedicated data repositories are often the best place to preserve your data. Data not suitable for preservation in a repository can be stored using a KU Leuven storage solution, consult the interactive KU Leuven storage guide.*	☐ KU Leuven RDR ☐ Large Volume Storage (longterm for large volumes) ☒ Shared network drive (J-drive) ☒ Other (specifiy): VSC archive and public repositories All published NGS datasets will be deposited in dedicated publicly available data repositories (i.e. GEO and EBI ArrayExpress). In addition, we will store these sequencing files in VSC archive. All digital data will be stored in our KU Leuven OneDrive folders and J-drive, while the experimental procedures and notes will also be kept in our electronic lab notebook, together with physical notebooks. Algorithms that deviate from the standard analysis pipelines and developed within our group will be stored on Github.
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	For a total of ~10TB over 5 years, we anticipate a cost of around 3500 Euros. These costs will be covered by this grant and other grants.

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: NGS data for transcriptomics studies as well as genetic screens will be deposited in public databases at the time of submission of the manuscripts for publication. The data will be embargoed until the publication of the manuscript and will be made fully available to the public thereafter.
If access is restricted, please specify who will be able to access the data and under what conditions.	As soon as the embargo is lifted on the data, it will be open to public for re-use, conforming the community norms and the rules listed in the relevant public repository. In such cases of re-use, the article associated with the datasets generated under this project will be cited by the third parties re-using the data. For datas shared directly by the PI upon request from third parties, a material transfer agreement will be made to describe the extent and the types of the re-use. Data will also be shared under a CC-BY 4.0 reuse license.
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:

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) NGS datasets will be deposited in open access repositories such as the NCBI Gene Expression Omnibus (GEO) and the EBI ArrayExpress databases. Experimental protocols will be documented in the published manuscripts. Manuscripts submitted to scientific journals for publication will be made publicly available in pre-print servers such as BioRxiv. Publications will be added to KU Leuven institutional repository, Lirias.
When will the data be made available?	☒ 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.* *Check the RDR guidance on licences for data and software sources code or consult the License selector tool to help you choose.*	☒ 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
What are the expected costs for data sharing? How will these costs be covered?	We will use free of charge data repositories, as much as possible. Depending on the repository of choice, the costs will be covered by the project funding.

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	Th PI, Atilgan Yilmaz, together with the research staff (currently Dr. Eslam Mohammed and Margaux Van Puyvelde) and technical staff (currently Sandra Jansen) involved in the data collection and analysis.
Who will manage data storage and backup during the research project?	The PI, together with the research and technical staff, under the guidance or with the help of ICTS and gbiomed-IT staff.
Who will manage data preservation and sharing?	The PI will be responsible and will receive support from the ICTS and gbiomed-IT staff, if needed.
Who will update and implement this DMP?	The PI has the ultimate responsibility to update and implement this DMP.