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	**Kato Ramaekers (https://orcid.org/0000-0003-0023-7940)**
Contributor name(s) (+ ORCID) & roles	**Kathleen Freson (promotor)** **Veerle Labarque (copromotor)**
Project number[1] & title	1126723N Improved disease modelling and genetics for inherited platelet disorders
Funder(s) GrantID[2]	FWO 1126723N
Affiliation(s)	☒ 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	Inherited platelet disorders (IPD) are highly heterogeneous with genetic defects in 63 genes and characterized by a high susceptibility to bleeding and often associated with syndromic features. Many genes have been discovered by whole exome/genome sequencing (WES/WGS) but their exact role in platelet formation and function remains unknown due to the lack of animal or efficient cell-based models. Today, high throughput-sequencing multi-gene panel tests comprising these genes can diagnose 48 to 26% of patients with platelet formation and function disorders, respectively. Many patients still remain undiagnosed and even with WGS data available, it remains extremely difficult to pinpoint the causative genetic defect because of the hundreds of rare coding variants present. This project therefore aims to deliver a rapid stem cell-based model to study known (but functionally undefined) and novel genes using CRISPR/Cas technology, quantitative in vivo imaging and genetic studies. Novel genes will be obtained from WGS data obtained for undiagnosed patients with very well-known platelet disorders, supplemented with platelet transcriptomes. This will allow WGS data analysis of differentially expressed genes. Validated novel candidate genes will be used to expand our existing diagnostic multi-gene panel test to increase detection of patients with the same IPD. This project will provide novel insights in platelet biology that is of use for therapeutic discoveries.

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
Whole genome sequencing	WGS data of patients mentioned in the project application	☐ 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: .bam ☐ NA	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☒ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
RNAseq	RNAseq data of platelets, monocytes, neutrophils and T-cell for patients with an	☒ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☐ Experimental ☐ Compiled/ aggregated data ☐ Simulation	☐ .por ☒ .xml ☐ .tab ☒ .csv ☐ .pdf	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☒ < 1 TB ☐ < 5 TB	

[4]

	inherited platelet disorder and healthy controls			data ☐ Software ☒ Other ☐ NA	☒ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: .fastq, .bam, .count ☐ NA	☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
Flow cytometry data	Flow Cytometry and FACS sort files (FlowJo and equipment specific files)	☒ 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: .fcs .jpg .tiff ☐ NA	☐ < 100 MB ☒ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
Cytation5 imager	Imaging and quantification files from BioTek Cytation5 imager (KU	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☒ Experimental ☐ Compiled/ aggregated data ☐ Simulation	☐ .por ☒ .xml ☐ .tab ☐ .csv ☐ .pdf	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB	

	Leuven servers).			data ☐ Software ☐ Other ☐ NA	☐ .txt ☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: .jpg .tiff .mp4 .exp ☐ NA	☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
Confocal microscopy	Imaging files from confocal microscope	☒ 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: .jpg .tiff ☐ NA	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	
Data analysis and manuscript preparation		☒ Generate new data ☐ Reuse existing data	☒ Digital ☒ Physical	☐ Observational ☒ Experimental ☒ Compiled/ aggregated data ☐ Simulation data ☐ Software ☐ Other	☐ .por ☒ .xml ☐ .tab ☐ .csv ☒ .pdf ☒ .txt ☒ .rtf ☐ .dwg	☐ < 100 MB ☐ < 1 GB ☒ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB	

				☐ NA	☐ .tab ☐ .gml ☒ other: .docx .jpg .tiff ☐ NA	☐ 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.	WGS data: This data is stored on secured servers of University of Cambridge-UCAM High Performance Computing facility, Freson K (promotor) has access to this server. RNAseq: part of this data (20 control platelet samples) has been generated during the PhD of Fabienne Ver Donck and is stored on the EGAS000106339 archive.

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

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: Reference to ethical committee approval: S63666
Will you process personal data? 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: Whole genome sequencing, RNA sequencing and personal metadata (age, gender, clinical condition, laboratory data (hemostatic tests results), blood counts) will be collected - Privacy Registry Reference: not applicable
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: If data will be obtained of interest for valorization, IP restriction will be claimed. It is not clear from the start what novel genetic targets relevant for megakaryopoeisis and platelet formation/function can be identified.

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: The megakaryocyte model itself "imMKCL" cannot be protected by IP restrictions as this model was developed by our collaborators and we have signed an MTA (with dr K Eto, Kyoto university). If this model is needed for the overall IP restriction, a joined application is a possibility (as stated in the MTA). The flow chamber to enhance platelet production from imMKCL cells from the Bristol group (PI A. Poole) cannot be protected by IP restrictions as we have signed an MTA to use these slides in this project.
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).	Generated sequencing data will be uploaded to EGA in combination with related metadata (e.g. age, gender, case/control status, sequencing platform/library... etc.) to be accessible to the public. Flow cytometry and sorting: information on gating strategy for cell identification and sorting will be saved in electronic files with details on antibody concentrations and protocols for cell preparation and staining will be described in detail in lab books. Imaging (confocal and cytation5 imager): images and settings will be saved in electronic files. Details on staining techniques and antibody or dye concentrations and protocols for cell preparation will be described in detail in lab books.

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: The metadata standards of EGA will be used for submission of sequencing data, as can be consulted on https://ega-archive.org/submission/sequence/unaligned 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?	PC owned by the group
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.*[6] *Refer to institution-specific policies regarding backup procedures when appropriate.*	Double backup on: KU Leuven OneDrive, external SSD drive

[6] 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: external SSD capacity of 2 TB, KU Leuven OneDrive capacity of 2 TB (can be increased) 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]	The KU Leuven OneDrives are protected from unauthorized persons. External SSD will be stored in a locked cabinet.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	OneDrive is included for all employees of KU Leuven (no additional storage cost), J drive storage (add. Backup) costs 52 euro per year

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...).	Data that will no longer be used for analyses will be transfer from our active KULeuven servers to the archive KULeuven-servers after 10 years. No data will be disposed.

Where will these data be archived (stored and curated for the long-term)?	The data will be stored on the university's central archive servers (with automatic backup procedures) for at least 10 years, conform the KU Leuven RDM policy.
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	Permanent storage after the project for at least 10 years after the end of the project: (Archive storage K-drive) €5,69 per 100GB per year.

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, in an Open Access repository ☒ Yes, in a restricted access repository (after approval, institutional access only, …) ☐ No (closed access) ☐ Other, please specify: DNA and RNA datasets will be shared on EGA servers after publication (open access repository). Other data can be obtained by researchers after request and approval by the PI (Kathleen Freson).
If access is restricted, please specify who will be able to access the data and under what conditions.	Researchers can access that data after request.
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: DNA samples from patients can't be shared because of privacy reasons.
Where will the data be made available? If already known, please provide a repository per dataset or data type.	The WGS and RNAseq data will be made available on the EGA, where access to data is granted based on applications to a data access committee that oversees the dataset.

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'.*	After publication of the research results in a peer-reviewed journal
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."* [7]	Genetic data can be reused after submission to EGA.
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 ☐ No If yes: EGA submission is accompanied by a specific accession number that will be mentioned in the publications
What are the expected costs for data sharing? How will these costs be covered?	Data will be released after publication of a manuscript and data sharing costs will be covered by the publication cost.

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

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	PI Kathleen Freson and PhD student Kato Ramaekers
Who will manage data storage and backup during the research project?	PIs Kathleen Freson, Veerle Labarque PhD student Kato Ramaekers Technician Chantal Thys
Who will manage data preservation and sharing?	PI Kathleen Freson
Who will update and implement this DMP?	The PI bears the end responsibility of updating & implementing this DMP.