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	**Joris Vermeesch (0000-0002-3071-1191)**
Contributor name(s) (+ ORCID) & roles	
Project number[1] & title	Mapping the role of the low copy repeats in the phenotypic variability of the 22q11 Deletion Syndrome
Funder(s) GrantID[2]	GOA2622
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	The 22q11 deletion syndrome (22q11DS) is the most common genomic disorder, with a prevalence of 1 in 3000 births. The reason for this high incidence remains an enigma. The presence and degrees of severity of most phenotypic features are highly variable across patients and it remains unknown why some patients acquire neuropsychiatric features and others do not. The deletion is caused by non-allelic homologous recombination, typically causing a 3MB deletion in 90% of patients. We demonstrated human specific expansion and hypervariability of the low copy repeats (LCR) causing the rearrangement with sizes ranging from 200kb to over 2.5Mb. Since duplications in the genome are drivers of evolution and genes in other LCRs have been shown to modulate brain development, we hypothesize this variation could be an important determinant for the 22q11DS phenotypic variability and especially the neuropsychiatric features. Using CRISPR/Cas9 editing, we will engineer human embryonic stem cell lines to remove individual LCRs, determine the effect on gene expression and map their differentiation potential into neurons. We will map the haplotype structure, determine the rearrangement breakpoints and map the effect on gene expression in 22q11DS patients to unravel the role of the LCRs in both the mechanism causing and the phenotypic variation affecting 22q11DS.

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
Long read sequencing data	Genome wide long read sequencing data from 22q11.2DS parents and patients to map the structural rearrangements and the 22q11.2 LCRs	☒ 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 Sequencing data: .fastq.gz Reference genomes: .fasta Aligned reads: .bam, .bai,	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☐ < 10 TB ☒ < 50 TB ☐ > 50 TB ☐ NA	Biobanking

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

					10x Cell Ranger output files: .bam, .mtx, .tsv .csv,.htlm, Analysis with Seurat package: .R, .Rdata, .rds .csv, .xls/.xlsx .jpeg □ < 1 TB Array data Illumina array data on gDNA ☒ Generate new data ☒ Digital ☒ Experimental .idat, .csv., .xls/.xlsx, flat text files		

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.	
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: Human genome data is sensitive data as it cannot be anonymized. We store those data under secure environment. The use of clinical data and samples included in this study is approved by the Ethical Review Committee of the University Hospitals UZ/KU Leuven

[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:
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:

[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).	Wet lab protocols are described in detail and recorded in Word files and PDF files, stored in appropriately labelled folders on project-specific KU Leuven OneDrive or UZ Leuven M drive. For some wet lab procedures SOPs from the UZ diagnostic unit will be followed. Where applicable, final bioinformatic scripts will be tracked in Jypiter notebooks and for reproducibility and data analysis will be upload on GitHub platform or e.g. Figshare, which will be accompanied by a README.txt file. Sequencing data will be collected and stored either on KU Leuven Large Volume Storage (L: Drive) and mainly at VSC Flemish Super Computer. A metadata file will be provided with the clear description of the raw data and how they were generated; the metadata file will be kept together with the sequencing data. Patient inclusions will be kept in an Excel file, stored in the lab TEAMS KULeuven environment.
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: Sequencing data will be stored on VSC, accompanied by a metadata file, containing the necessary information to find and re-use specific files (sample key, technical parameters). Sequencing data require specific metadata when submitted to access-controlled repositories (e.g., EGA). Data documentation will be tailored to their ultimate deposition in public repositories. When depositing data in a repository, the final dataset will be accompanied by detailed information regarding technical and analytical methods used to generate and analyze the data, to allow for independent reproduction; bioinformatics scripts will be provided in repositories like Figshare or GitHub.

4. Data Storage & Back-up during the Research Project	
Where will the data be stored?	☒ Personal network drive (I-drive) ☒ OneDrive (KU Leuven) ☐ Sharepoint online ☐ Sharepoint on-premis ☒ Large Volume Storage ☒ Other: Vlaamse Super Computer (VSC) and UZ Leuven serve
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.*	Data is stored on KU/UZ Leuven and VSC servers with back-up capacities. Eventual upload in EGA will secure long term storage.
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: If no, please specify:

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

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]	Data are stored on KU Leuven IT infrastructure (KU Leuven Large Volume Storage, KU Leuven One Drive, UZ Leuven Server and VSC Flemish Super Computer), requiring for the access a Multifactor Authentication. Also, initial access is defined by the corresponding PI research group, so it will be only available to authorized personnel.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	VSC Staging storage: € 30 / TB / year. The costs for data storage for this project are foreseen and allocated within the project budget.

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

Where will these data be archived (stored and curated for the long-term)?	☒ KU Leuven RDR ☒ Other (specifiy): VSC archive for raw digital files and after publication sequencing data will be deposited to European Genome-phenome Archive/GEO data repositories with controlled access meaning that a third party can obtain access to the data only following approval by the KU Leuven/UZ Leuven Data Access Committee (DAC).
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	VSC archive storage: € 30 / TB / year. The costs for data storage for this project are allocated within the project budget and within future projects.

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: Pseudonymized (coded) data will not be shared, unless a proper Data Transfer Agreement (DTA) or Material Transfer Agreement (MTA) is in place. This implies that pseudonymized data will not be made public, also not after the end of the project, but deposited to deposited to European Genome-phenome Archive/GEO data repositories with controlled access meaning that a third party can obtain access to the data only following approval by the KU Leuven/UZ Leuven Data Access Committee (DAC). Anonymized aggregated datasets could be made available after the publication. Scripts, algorithms and software tools will be described in manuscripts as supplementary files and/or on GitHub (https://github.com), or Figshare repositories. Research results will be published as preprints and as Open Access in peer reviewed journals.
If access is restricted, please specify who will be able to access the data and under what conditions.	Access to human data will be granted by the data access committee to bonafide researchers affiliated with recognized research institutions upon a proper Data Transfer Agreement (DTA) is in place between UZ/KU Leuven DAC and other research institution.
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: Due to nature of the data and also potential intellectual property, data access to human data will restricted according to the specified clauses in the informed consent forms for the different studies or due to associated intellectual property rights.

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) EGA/GEO Algorithms, scripts and software: The relevant algorithms, scripts and software tools driving the project will be described in manuscripts and/or on GitHub (https://github.com) or figshare, if no novel intellectual property rights are associated. (Pre-print) publications will also be automatically added to our institutional repository, Lirias 2.0, based on the authors name and ORCID ID. Research results will be published as BioRxiv preprints or/and as Open Access in peer reviewed journal
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 the research results
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]	Data Transfer Agreement (restricted data) GNU GPL-3.0 (code)

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

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: 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?	We don't expect any costs for datasharing

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	The Ph.D students and PI of the project is responsible for data documentation
Who will manage data storage and backup during the research project?	Yes, a PID will be added upon deposit in a data repository
Who will manage data preservation and sharing?	Yes, a PID will be added upon deposit in a data repository
Who will update and implement this DMP?	Yes, a PID will be added upon deposit in a data repository