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	**Yourae Hong / 0000-0001-6683-433X**
Contributor name(s) (+ ORCID) & roles	
Project number[1] & title	12D5823N
Funder(s) GrantID[2]	D-2023-1809
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	To understand the characteristics of colorectal cancer (CRC), diverse molecular classifiers have been generated. But these only capture the characteristics of the tumor at a specific stage or in a specific section. CRC has continuously evolving entities, with a complex adenoma-carcinoma sequence and later carcinoma-metastasis sequence, in a continuous coevolutionary crosstalk with the tumor microenvironment. Our past work has shown representative features in the carcinoma, and the current work will elaborate these in detail, integrating different data inputs needed to characterize the dynamic information in these lesions. The adenoma-carcinoma stages of CRC offer a unique opportunity, if exploited well, to understand evolutionary trajectories of the epithelium, the co- evolution of the TME and the possible impact on the targetable immune states. With this project, I will characterize the tumor cells' sub-characteristics within classified epithelial molecular groups with integration approaches of bulk, single-cell and spatial technologies. In the preliminary data, the heterogeneous pattern within molecular subtypes can clearly be found. Especially, the understudied non adenomatous lesions, non-canonical Wnt driven tumors show very interesting heterogeneity and novel cancer stem cell dynamics. Collectively, signatures combined with bulk-, single-cell RNA sequencing and spatial techniques will identify the optimal stratification for CRC prognosis and therapy.

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
Single-cell RNA sequencing data (basecall)	Basecall reads data of single-cell RNA sequencing	☒ 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: FASTQ ☐ NA	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ < 10 TB ☐ < 50 TB ☐ > 50 TB ☐ NA	4TB
Single-cell RNA sequencing data (aligned)	Aligned reads data of single-cell RNA sequencing from FASTQ	☒ Generate new data ☐ Reuse existing data	☒ Digital ☐ Physical	☐ Observational ☒ Experimental ☐ Compiled/ aggregated data ☐ Simulation data	☐ .por ☐ .xml ☐ .tab ☐ .csv ☐ .pdf ☐ .txt	☐ < 100 MB ☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☒ < 5 TB ☐ < 10 TB	3TB

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

				☐ Software ☐ Other ☐ NA	☐ .rtf ☐ .dwg ☐ .tab ☐ .gml ☒ other: BAM ☐ NA	☐ < 50 TB ☐ > 50 TB ☐ NA	
Quantified single-cell RNA sequencing data	After alignment, quantified data for each gene levels for each single cells	☒ 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	100GB

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.	We also reuse data, for processed scRNA-seq (#3), deposited by Synapse (https://www.synapse.org/#!Synapse:syn26844071/wiki/615389).
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: Usage of human subject data is a part of this project. For that, we got the approval by the Ethical committee (S66460).

[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).	**Data containing raw and processed data will be organized based on the source of the data, and date to data generation. It stored by seperated document with specific date. Scripts for data analysis will be tracked using each R script / R Markdown files, with explanation. Script files for data analysis will be stored in workstation in KU Leuven, and periodically stored using One Drive (managed by KU Leuven IT department)**
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: Document for each experiment will be stored in L-Drive (managed by KU Leuven), including sample keys, experimental condition, to ensure the reusability of the data. Also, computational analysis pipelines after generation of processed data (finishing anonymous step), will be stored using OneDrive, managed by KU Leuven IT department.

4. Data Storage & Back-up during the Research Project

Where will the data be stored?	**We stored two systems. First, we generated FASTQ data from Genomic Core, and uploaded FASTQ & processed data on cloud server, only accessible to restricted users. Also, we copied FASTQ data in hard drive in KU Leuven, with safety lock.**
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.*	**To avoid data loss, we backed up FASTQ data on cloud server, managed by Genomic Core.**
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: We have ~10TB hard drive to store data. 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]	We only used cloud system managed by Genomic Core, under the permission from PI and only used specific account generated by Genomic Core. So only restricted user can access data, but only possible to download. Also, we downloaded and backed up data and it is already locked to use that.
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	For the store of data, about 245euro per year, for 1 TB. We expected 8TB sufficiently, to store all data generated as part of this project. These costs will be covered by the budget of the principal investigator (Prof. Sabine Tejpar).

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 stored the minimum preservation term of 5 years, after the end of the project.**

Where will these data be archived (stored and curated for the long-term)?	**After publication, raw sequencing data will be deposited into an access-controlled public depository, for example, European Genome-Phenome Archive (EGA). Also for processed data (after anonymous) will be deposited other depository, Synapse or Gene Expression Omnibus (GEO).**
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	**That data is not actively worked after end of project, so cost will be reduced because it will be moved into archival storage, not standard storage. So cost will be down and will be covered by the budget of Prof. Sabine Tejpar.**

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: - Prior to publication, data sharing is restricted to members of the IMMUCAN consortium, already concluded a treaty. - After publication, For raw sequencing data, we will upload public depository as controlled access like EGA. It only allows restrict user after approval from host institution. - After publication, script for downstream analysis will be stored on public repository like GitHub.
If access is restricted, please specify who will be able to access the data and under what conditions.	Prior to publication, data sharing is only available to member of IMMUCAN consortium (concluded a treaty), and after publication, only allow user after getting approval from host 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: As above, raw and processed data is restricted to member of IMMUCAN consortium.
Where will the data be made available? If already known, please provide a repository per dataset or data type.	**In Controlled access repository.**

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 from the project will be made available under a creative commons attribution license (CC-BY 4.0).**
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:
What are the expected costs for data sharing? How will these costs be covered?	**After publication, all data sharing is via public depository, so the cost will be free. Before publication, the cost will be covered the budget of Prof. Sabine Tejpar.**

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

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	**The applicant (Dr. Yourae Hong) and the principal investigator (Prof. Sabine Tejpar) will share the responsibility for data documentation.**
Who will manage data storage and backup during the research project?	**The applicant (Dr. Yourae Hong) will be primarily responsible to store and back up data during the project.**
Who will manage data preservation and sharing?	**The applicant (Dr. Yourae Hong) and PI (Prof. Sabine Tejpar) will be sharing the responsibility for data preservation and sharing.**
Who will update and implement this DMP?	**The applicant (Dr. Yourae Hong) and PI (Prof. Sabine Tejpar) will be sharing the responsitiliby for the updating and implementing this DMP.**