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	**Steven Van Belleghem 0000-0001-9399-1007**
Contributor name(s) (+ ORCID) & roles	**Frederik Hendrickx 0000-0002-1176-0318 co-Promotor**
Project number [1] & title	Unravelling structural genome variation evolution by exploring functional pan-genomics
Funder(s) GrantID [2]	G020025N
Affiliation(s)	☒ KU Leuven ☐ Universiteit Antwerpen ☐ Universiteit Gent ☐ Universiteit Hasselt ☐ Vrije Universiteit Brussel ☒ Other: RBINS 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	Structural variations (SVs) in genomes, such as insertions, deletions, inversions, duplications, and translocations, are prevalent yet often overlooked in population genetic studies. Three critical questions remain largely unaddressed: 1) “How much SVs are present within and between populations?”, 2) “What is the effect of SVs on genome functioning and phenotypic development?”, and 3) “What is the evolutionary significance of SVs?”. Recent technological advances have revolutionized our ability to cost-effectively and confidently identify all types of SVs. Here, we propose to apply these technologies to Pogonus (Europe) and Calosoma (Galápagos) beetles, which each possess a diversity of elaborate structural rearrangements linked to habitat-specific adaptation. By collecting chromosomal genome assemblies across their geographic ranges, we aim to initially delineate the comprehensive spectrum of SVs in both species. Subsequently, we will collect gene expression, chromatin accessibility, chromatin conformation, and DNA methylation data for key developmental stages and integrate these data into a composite pan-genome assembly. This approach will allow us to assess how SVs result in changes to gene content, gene regulation and development. Finally, we will reconstruct the evolution of these SVs across the unique evolutionary history of each species. Ultimately, this project is expected to enhance our understanding of the rol and evolution of SVs in natural populations.

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
		☒ Generate new data ☒ Reuse existing data	☒ Digital ☐ Physical	☐ Audiovisual ☒ Images ☐ Sound ☒ Numerical ☒ Textual ☐ Model ☐ Software ☐ Other:	DNA/RNA sequences, gene expression profiles, population genetics images & metrics	☐ < 1 GB ☐ < 100 GB ☐ < 1 TB ☐ < 5 TB ☒ > 5 TB ☐ NA	
DNA-seq short	Short-read DNA sequences of beetles	Generate new data	Digital	Observational	.fastq	<5Tb	
DNA-seq long	Long-read DNA sequences of beetles	Generate new data	Digital	Observational	.fastq	>5Tb	
RAD-tag	RAD-tag sequencing for association mapping	Generate new data	Digital	Observational	.fastq	<1Tb	
ATAC-seq	Chromatin	Generate new data	Digital	Observational	.fastq	<1Tb	

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

	accessibility profiles of beetles						
RNA-seq	Gene expression profiles of beetles, developmental changes	Generate new data	Digital	Observational	.fastq	<1Tb	
Iso-seq	Long read RNA sequencing	Generate new data	Digital	Observational	.fastq	<1Tb	
Phenotypic measurements	Phenotypic measurements of beetles	Generate new data	Digital	Observational	.txt	<1Gb	
Processed sequenced data	Sequence data mapped to reference genome	Generate new data	Digital	Observational	.bam	<5Tb	
Genotype calls	Genotypes from sequence data relative to reference genome	Generate new data	Digital	Observational	.vcf	<5Tb	
Omni-C data	Chromatin contacts in the genome	Generate new data	Digital	Observational	.fastq	<1Tb	
DNA methylation	Methylation status obtained from Pacbio	Generate new data	Digital	Observational	.fastq	<1Tb	

	data						
Pan genome	Pan genome assembly with SVs	Generate new data	Digital	Observational	.fasta .xmfa .maf	<1Tb	
Bioinformatic analyses pipeline	Data processing scripts	Generate new data	Digital	Observational	.py .sh .R	<1Gb	

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.	**Pogonus chalceus original draft genome: NCBI BioProject accession code: PRJNA381601** **Genome assembly provided by Sanger institute: Tree of Life QC: Species Report - Pogonus chalceus** **Calosoma resequencing data: NCBI BioProject accession code: PRJNA706924** **Calosoma draft genome: NVBI: JAGJTL000000000**
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: ☐ 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:
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:

3. Documentation and Metadata

[4] See Glossary Flemish Standard Data Management Plan

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*.	**For genomic sequence data (including short- and long-read DNA, RNA, and ATAC), NCBI metadata standards for data submission will be used. This includes:** **- Sample ID (NCBI accession and lab ID)** **- Sample origin** **- Date of sampling** **- Location of sampling** **- Tissue** **- Life stage** **- Sex** **- Sequence type and sequence platform** **Data processing pipelines stored on GitHub will be accompanied with readme.txt files.** **Phenotypic measurements of beetles will be stored in excel sheets including clone ID, clone origin, phenotypic value, experimenter and date of experiment.** **An electronic lab book in which all metadata is shared and curated to common standards will be kept through a lab group on Microsoft teams.** **I will also keep printed version of my protocols for the fieldwork, the molecular biology experiments and bioinformatic pipelines.**

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: **1. Raw sequencing reads for: NCBI metadata standards for data submission ('SRA_metadata.xlsx' and 'invertebrate.1.0.xlsx' file)** **2. Reference genome assemblies: NCBI metadata standards for data submission** If no, please specify (where appropriate per dataset or data type) which metadata will be created: **3. Processed sequenced data: external harddrives (in duplicates) and KU Leuven VSC Tier-1 Data Storage for at least 10 years** **4. Bioinformatic analyses pipeline: readme.txt files with descriptions of the pipelines. Stored on Github, external harddrive and Microsoft cloud service.** **5. Phenotypic measurements of beetles: excel sheets including clone ID and phenotypic value. Stored on Github, external harddrive and Microsoft cloud service.**

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) **At Ku Leuven** **Personal computer** **External harddrives** **KU Leuven VSC Tier-1 Data Storage** **Online:** **Public submission of sequence read data to NCBI (with embargo until publication)** **Github**
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 **We have already secured 32 TB of storage on the iRODs server in Tier 1 of KU Leuven for our lab and purchased a 5TB personal external hard drives (both currently in use).** 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 KU Leuven VSC Tier-1 Data Storage has a strict permissions system (read/write/modify), which will be limited to the primary collectors of the data.** **External hardddrives will be stored safely lockers in the lab space or in the office of Steven Van Belleghem.** **Personal computers will be protected password protected.**
What are the expected costs for data storage and backup during the research project? How will these costs be covered?	**Costs for data storage are included in the funded project and allow purchasing external harddrives and KU Leuven VSC Tier-1 Data Storage until the end of the project.**

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) **All raw sequencing will be publicly stored >25 years.** **1. Raw sequencing reads for:** **- Short-read DNA sequences (.fastQ)** **- Long-read DNA sequences (.fastQ)** **- ATAC-seq (.fastQ)** **- RNA-seq (.fastQ)** **- Hi-C (.fastQ)** **2. Reference genome assemblies (.fasta)** **Processed sequencing data will be stored for at least 10 years using the KU Leuven VSC Tier-1 Data Storage and external harddrives** **3. Processed sequence data:** **- read alignments to genomes (.bam)** **- gene expression counts (.counts)** **- open chromatin regions (.bw .bed)** **- PAN genome alignments (.maf)** **Protocols, pipelines and phenotypic measurements will be stored indefinitely on Github, onedrive, and external harddrives, and as supplementary material with published studies.** **4. Bioinformatic analyses pipeline** **5. Phenotypic measurements of beetles**

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): **1. Raw sequencing reads for: NCBI (National Center for Biotechnology Information)** **2. Reference genome assemblies: NCBI (National Center for Biotechnology Information)** **3. Processed sequence data: external harddrives and KU Leuven VSC Tier-1 Data Storage** **4. Bioinformatic analyses pipeline: GitHub** **5. Phenotypic measurements of beetles: Github and publication supplements**
What are the expected costs for data preservation during the expected retention period? How will these costs be covered?	**NCBI: Free** **GitHub: Free (up to 2gb, which is sufficient for storing pipelines and scripts)** **KU Leuven VSC Tier-1 Data Storage: ~35 euro per Tb. 32Tb is supported by the FWO Data component of the Flemish Tier-1 supercomputing platform until 2027.**

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:
If access is restricted, please specify who will be able to access the data and under what conditions.	Processed sequence data will be stored in at least two location using external harddrives and KU Leuven VSC Tier-1 Data Storage. These data can be reproduced from raw reads and the available processing pipelines (shared on GitHub). These processed data types are not typically uploaded to NCBI for storage and sharing, but after publication, these data will also be freely available upon request and shared through web transfers.
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) **1. Raw sequencing reads for: NCBI (National Center for Biotechnology Information)** **2. Reference genome assemblies: NCBI (National Center for Biotechnology Information)** **3. Processed sequenced data: external harddrives and KU Leuven VSC Tier-1 Data Storage** **4. Bioinformatic analyses pipeline: GitHub** **5. Phenotypic measurements of beetles: Github and publication supplements**
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?	**NCBI: Free** **GitHub: Free (up to 2gb, which is sufficient for storing pipelines and scripts)** **KU Leuven VSC Tier-1 Data Storage: ~35 euro per Tb. 32Tb is supported by the FWO Data component of the Flemish Tier-1 supercomputing platform until 2027.**

7. Responsibilities	
Who will manage data documentation and metadata during the research project?	**Steven Van Belleghem / Maria Madrid**
Who will manage data storage and backup during the research project?	**Steven Van Belleghem / Maria Madrid / Frederik Hendrickx**
Who will manage data preservation and sharing?	**Steven Van Belleghem / Maria Madrid / Frederik Hendrickx**
Who will update and implement this DMP?	**Steven Van Belleghem / Maria Madrid / Frederik Hendrickx**