VISCREEN: Engineering viral vectors for high-throughput CRISPR genetic screens in plants (MSCA-PF)

A Data Management Plan created using DMPonline.be

Creator: Mireia Uranga Ruiz De Eguino https://orcid.org/0000-0001-6343-4031

Affiliation: KU Leuven (KUL)

Funder: European Commission (Horizon)

Template: Horizon 2020 FAIR DMP +

Principal Investigator: Mireia Uranga Ruiz De Eguino https://orcid.org/0000-0001-6343-4031

Project Administrator: Nico De Storme

Grant number / URL: HORIZON-MSCA-2022-PF-01-101110621

ID: 202837

Start date: 01-04-2023

End date: 31-03-2025

Project abstract:

Tools based on CRISPR-Cas systems hold great potential both for fundamental plant genetics and precision crop breeding. Unlike traditional mutagenic approaches, CRISPR-Cas systems allow for the simultaneous delivery of thousands of unique guide RNAs, enabling the generation of collections of higher-order mutants for rapid functional screens and the discovery of genetic interactions. However, CRISPR technology suffers from a 'transformation bottleneck'; relatively few crop genotypes can be transformed and often only a handful of mutant lines can be produced. In virus-induced genome editing (VIGE), plant viruses are engineered to deliver gRNAs into cells. Such a capability is highly attractive for CRISPR knock-out (KO) screens as it could simplify the delivery and generate large populations of mutant lines in a broader range of crop genotypes. However, the application of VIGE for CRISPR KO screens is hampered by the lack of a proper viral-based multiplex gRNA delivery vector. My host lab has pioneered the development of CRISPR KO screens in plants, and in VISCREEN I will combine this knowledge with my expertise on VIGE to scale up CRISPR-Cas-mediated manipulation of plant genomes. The main objectives of my project are: (i) to develop Barley stripe mosaic virus (BSMV) for highthroughput CRISPR screens in wheat and maize; and (ii) to engineer and validate a set of viral vectors for VIGE in soybean. The viral vector-based, high-throughput editing systems developed in VISCREEN will expand CRISPR screens in a wider variety of crops and allow to test fundamental biological questions that are limited by functional redundancy. Overall, this innovative approach will make me a well-trained academic researcher with the capabilities to develop a professional career within the plant genome editing field.

Last modified: 17-11-2023

VISCREEN: Engineering viral vectors for high-throughput CRISPR genetic screens in plants (MSCA-PF)
Horizon 2020 FAIR DMP +

Version information

Version number

v1.0

Description

Initial DMP (submitted in month 6) for the first periodic evaluation of the MSCA-PF project VISCREEN.

Date of first version

17/11/2023

Date of last update

17/11/2023

1. Data summary

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Data collection and generation provides the evidentiary basis for the scientific findings of the project. It also provides content for scientific dissemination (i.e. publications, conferences) as well as public-facing outreach and engagement activities.

What types and formats of data will the project generate/collect?

WP1 - Development of a high-throughput genome editing system in wheat and maize

- **Task 1.1.** *Construction of recombinant BSMV vectors and optimization of CRISPR screen.* sgRNA sequences for targeting yield- and meiosis related genes will be stored in .fasta format and also in an Excel file (.csv). The set of BSMVγ-derived vectors generated for sgRNA delivery will be stored in .gb format.
- **Task 1.2.** *Plant inoculation, phenotyping, and analysis of gene editing.* Physical samples (i.e. virus-infected leaves) collected from the inoculation experiments will be labelled, appropriately stored and included in database with a reference code. DNA and RNA extracted from these samples will be accordingly encoded. An Excel file (.csv) will be created for each sequencing delivery, where the PCR-amplified target genes will have unique reference codes. Sequencing data (.abi, .seq) of the target genetic regions will be used as input data for the editing analysis (ICE Synthego). ICE numeric data will provide us detailed information on the editing efficiency of the different BSMV-based multiplexing approaches. Results from the RT-PCR analysis of virus infectivity will be stored as electrophoresis images (.tif, .png).
- **Task 1.3.** *Inheritance of targeted modifications.* M1 seeds from inoculated plants will be labelled, appropriately stored and included in database with a reference code. Data from gene editing and RT-PCR analysis will be processed as detailed in task 1.2.

Dataset name	Description	Data type	File format	Estimated volume
Task 1.1	Target genes and sgRNAs	Genetic sequences	.fasta .csv	< 100 MB
Recombinant plasmids	DNA sequences	.gb	< 100 MB	
Task 1.2	Virus-infected leaf tissue	Physical (-80C frozen samples)	.csv	< 200 plants
DNA and RNA extractions	Physical (-20C/-80C frozen samples)	.csv	< 200 plants	
Sequencing of target genes	DNA sequences	.csv .seq .abi	< 1GB	
Gene editing analysis (ICE)	Numerical Image	.csv .png	< 1GB	
Virus infectivity analysis (RT-PCR)	Image	.tif .png	< 200 MB	
Task 1.3	M1 progeny	Physical (-80C frozen samples)	.csv	< 200 plants
DNA and RNA extractions	Physical (-20C/-80C frozen samples)	.csv	< 200 plants	
Sequencing of target genes	DNA sequences	.csv .seq .abi	< 1GB	
Gene editing analysis (ICE)	Numerical Image	.csv .png	< 1GB	
Virus infectivity analysis (RT-PCR)	Image	.tif .png	< 200 MB	

WP2 - Engineering of viral vectors for legume functional genomics

- **Task 2.1.** *Construction recombinant vectors for a mobile editing system.* Plasmid vectors used for soybean hairy root transformation will be stored in .gb format.
- **Task 2.2.** *Validation of soybean hairy root transformation.* A GFP reported will be co-delivered with the mobile CRISPR-Cas9 plasmid in the transformation experiments. Confocal microscopy images (.tif) for GFP fluorescence in roots will provide us with the qualitative information to estimate the transformation efficiency and select the transformed plants for ongoing steps.
- **Task 2.3.** *Validation of gene editing in soybean.* Physical samples (i.e. leaves) collected from the transformation experiments will be labelled, appropriately stored and included in database with a reference code. DNA extracted from these samples will be accordingly encoded. Sequencing data (.abi, .seq) of the target genetic regions will be used as input data for the editing analysis (ICE Synthego). ICE numeric data will provide us detailed information on the performance of the mobile CRISPR-Cas system.
- **Task 2.4.** *Inheritance of targeted modifications.* M1 seeds from inoculated plants will be labelled, appropriately stored and included in database with a reference code. Data from gene editing and RT-PCR analysis will be processed as detailed in task 1.2.

Dataset name	Description	Data type	File format	Estimated volume
Task 2.1	Target genes and sgRNAs	Genetic sequences	.fasta .csv	< 100 MB
Recombinant plasmids	DNA sequences	.gb	< 100 MB	
Task 2.2	Confocal microscopy (GFP)	Image	.tif	< 500 MB
Task 2.3	Transformed leaf tissue	Physical (-80C frozen samples)	.csv	< 50 plants
DNA and RNA extractions	Physical (-20C/-80C frozen samples)	.csv	< 50 plants	
Sequencing of target genes	DNA sequences	.csv .seq .abi	< 1GB	
Gene editing analysis (ICE)	Numerical Image	.csv .png	< 1GB	
Task 2.4	M1 progeny	Physical (-80C frozen samples)	.csv	< 200 plants
DNA and RNA extractions	Physical (-20C/-80C frozen samples)	.csv	< 200 plants	
Sequencing of target genes	DNA sequences	.csv .seq .abi	< 1GB	
Gene editing analysis (ICE)	Numerical Image	.csv .png	< 1GB	

Will you re-use any existing data and, if so, how?

Authors from a recent research paper on BSMV editing (doi: 10.1111/pbi.13910) kindly provided us with BSMV vectors for virus induced gene silencing. These will be used in WP1 for preliminary inoculation experiments, which will serve to: (i) study the apearance of appearance of viral infection symptoms; and (ii) compared the performance of different inoculation methods (in vitro transcription vs. mechanical inoculation).
Additionally, I will use data of gene sequences from diverse plant species (i.e. wheat, maize, soybean) from databanks such as NCBI, PLAZA 5.0 and Ensembl Plants.

What is the origin of the data?

Since the principal methodology of the project is experimental, all the information related to the construction of recombinant vectors, experimental designs, sample management, data adquisition and analysis will be recorded in the researcher's lab notebook and in the cloud-based research platform Benchling. For short to mid-term storage, data will be stored on the KU Leuven OneDrive server, which will also enable to share files with team members and collaborators.

What is the expected size of the data (if known)?

The expected size of the data is less than 10 GB.

To whom might the data be useful (‘data utility’)?

The data will be useful to:
(i) The plant science research community, especially those with expertise in CRISPR-Cas9 and in virus biotechnology.
(ii) Agronomic companies interested in applying novel genome editing methodologies for precise crop breeding.

2.1 FAIR data: Making data findable, including provisions for metadata

Are the data produced and/or used in the project discoverable with metadata?

Yes. We will use the RDR data repository of KU Leuven, which automatically applies a metadata standard upon depositing the data. The metadata model will include fields that are required, recommended and optional. Using the RDR repository, all the
data generated as part of VISCREEN will follow the FAIR principles (Findable, Accessible, Interoperable and Reusable) and will be open to other researchers.

Are the data produced and/or used in the project identifiable and locatable by means of a standard identification mechanism?

Yes, the RDR repository system uses Digital Object Identifiers (DOIs) as persistent and unique identifiers.

What naming conventions do you follow?

Within the head folder 'research', I have sub-folders for the types of document: primary sources, oral histories, published materials etc. Archival material: The naming conventions follow the organizational classification of the archive itself, to assist with accurate citation and retrieval in the case that I need to go back to the original document. In the case of the main Amnesty International archive, for example, I use the following: International Secretariat > Folder xxx > name of document. In the case of the Amnesty International USA archive, their own folder structure is more complex, and so I reproduce that: AIUSA > Series II Executive Directors Files > Box 11.1 1 > Folder 10 > File name Oral histories: Within the head research folder, oral histories are arranged in the following way: Research > oral histories > interviewee name > name_date_place.wav Secondary sources: Zotero has its own file structure, and also names the .pdf file based on document metadata.
Within the head folder 'Experiments', two sub-folders will be included for each WP as "WP1'' and "WP2''. A reference document (e.g. Readme.txt file) will be included in each sub-folder to describe the purpose of the different files in them.
File names will include the following info:
- Project acronym (i.e. VISCREEN)
- WP number (i.e. WP1 or WP2)
- Type of data (e.g. samples, sequencing, ICE, photos)
- Date (YYYYMMDD) - only when necessary
- Version number of the file (vXX, being X the number) - only when necessary
- Different terms will be separated using underscores (_)

Will search keywords be provided that optimize possibilities for re-use?

Keywords are provided through the RDR repository system.

What is your approach for clear versioning?

N/A.

What metadata will be created?

I will follow the metadata standards for Life Sciences, as outlined inGenome Metadata and MIBBI (Minimum Information for Biological and Biomedical Investigations)

2.2. FAIR data: Making data openly accessible

Which data produced and/or used in the project will be made openly available as the default? If some data is kept closed provide a rationale for doing so.

We will publish our findings via BioRXiv upon manuscript submission to peer-reviewed open-access journals to ensure a rapid dissemination of the results. Upon publication, sequences of all recombinant plasmids will be available to the scientific community by uploading them to GenBank (https://www.ncbi.nlm.nih.gov/genbank/). Sharing all experimental data is not possible because of contractual restrictions. Alternatively, a subset of data related to sequencing and gene editing analysis will be openly shared.

How will the data be made accessible?

Data will be deposited in the KU Leuven'sResearch Data Repository (RDR). This means that it will remain accessible after the life of the project, even after my own association with KU Leuven ends.

What methods or software tools are needed to access the data? Is documentation about the software needed to access the data included? Is it possible to include the relevant software (e.g. in open source code)?

- A molecular biology software platform (paid, downloadable: Geneious, SnapGene; or free, web-based: Benchling) will be needed to access recombinant plasmids, genetic sequences, alignments and sequencing data.
- An image visualizer (e.g. Adobe Photoshop, Pixlr) will be needed to access to RT-PCR images and photos of plant phenotyping.
- Excel will be needed to access to experiment and sample data files.

Where will the data and associated metadata, documentation, and code be deposited? Have you explored appropriate arrangements with the identified repository?

Data will be deposited in the KU Leuven'sResearch Data Repository (RDR). All documentation and published results will be available throughLirias, which is provided by Ku Leuven to facilitate sharing research output. Lirias also provides a gateway to materials stored on the RDR.
I confirm that the condition for use of the KU Leuven RDR fulfill the requirements of this project.

If there are restrictions on use, how will access be provided?

No restrictions.

2.3. FAIR data: Making data interoperable

Are the data produced in the project interoperable? What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Other researchers and users will easily be able to access and re-use this project's data:

- Molecular biology data files (i.e. plasmid maps, genetic sequences, aligments...) will be saved in the following formats:
  - .fasta: easily opened with NotePad
  - .gb: easily opened with a subscription software (e.g. Geneious, SnapGene) or the free, web-based Benchling software.
- Image files are easily opened with an image visualizator (e.g. Adobe Photoshop, Pixlr).
- Excel will be needed to access to experiment and sample data files.
- Other data files in .txt, .cvs or .pdf formats are easily opened with the MS Office pack and Adobe Acrobat Reader, which are installed standardly on most digital devices.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability? In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

All descriptors will be in given in plain language that can be understood across multiple disciplines to ensure accessibility and use by the scientific community (i.e. MIBBI, Minimum Information for Biological and Biomedical Investigations). If uncommon or project specific ontologies or vocabularies are used, mapping will be provided.

2.4. FAIR data: Increase data re-use (through clarifying licenses)

How will the data be licensed to permit the widest re-use possible?

Readme files will be provided on the methodology, if not stated in the document itself. This will ensure that the data will be easy to re-use and understood. Additionally, research data will be licensed under standard re-use licenses (EUDAT B2SHARE, Creative Commons Attribution International Public Licence - CC BY-) to allow other to use it for non-commercial purposes, as long as they acknowledge the source.

When will the data be made available for re-use? If applicable, specify why and for what period a data embargo is needed.

Research data will be made available once all research has been concluded. This will be before the end of the project, or at the fellowship end at the latest.

Are the data produced and/or used in the project usable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

CRISPR-Cas is a rapidly evolving research area and expected results expected of VISCREEN will contribute to expanding the use of viral vectors for novel editing purposes and host species. Only a limited number of research groups have enough expertise in engineering plant viruses, so our findings will be of great value to the broad scientific community.
Additionally, developing genome editing tools in crops will result in technology transfer. Many aggrotech companies are located in Belgium and all over Europe, so the CRISPR screen systems developed in the project will be appealing for them to set up collaborations.

How long is it intended that the data remains re-usable?

According to KU Leuven RDM policy, all relevant data will be preserved for at least 10 years after the end of the project, in a safe, secure and sustainable way for purposes of reproducibility, verification, and potential re-use.

Are data quality assurance processes described?

Regular meetings with the project supervisor will ensure quality control over the data collection process. These will include sample replicates (either biological or experimental), standardized data recording, data entry validation, peer review of data and representation with controlled vocabularies.

3. Allocation of resources

What are the costs for making data FAIR in your project? How will these costs be covered?

There are no costs associated with making data FAIR for this project, as repositories and other resources are provided by KU Leuven free of charge.

Who will be responsible for data management in your project?

As the sole researcher in VISCREEN, I am responsible for data capture, metadata production, data quality and storage and backup, archiving and sharing. Regular meetings
with the project supervisor ensure that this is being done correctly. Reporting to the EU will act as a second safeguard to ensure it is being done correctly.

What are the costs and potential value of long term preservation?

According to KU Leuven RDM policy, all data is preserved free of charge for 10 years after the project ends. I will discuss with the KU Leuven about any time limits they impose for open data storage and work with them to ensure the longest possible access to the data.

4. Data security

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

To allow secure storage and allow data recovery in case of loss or damage, all data and documentation related to the project will be stored in two separate locations: computer hard drive and related KU services.
During the project's data collection and generation phase (i.e. before data is made open) this will be stored on the secure, two-factor authentication protected OneDrive server provided by KU Leuven. Back-up copies will be automatically generated on the KU Leuven internal server, which is also password-protected and synched with OneDrive.
KU Leuven has IT specifications for data storage and management, including standard back-ups. The IT Office provides tailored solution to ensure that data is securely stored, and cannot be altered by an unauthorized entity.

5. Ethical aspects

Are there any ethical or legal issues that can have an impact on data sharing?

No.

6. Other issues

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

I follow the RDM policy at KU Leuven.

VISCREEN: Engineering viral vectors for high-throughput CRISPR genetic screens in plants (MSCA-PF)
GDPR Record

GDPR record

Have you registered personal data processing activities for this project?

- Not applicable

VISCREEN: Engineering viral vectors for high-throughput CRISPR genetic screens in plants (MSCA-PF)
DPIA

DPIA

Have you performed a DPIA for the personal data processing activities for this project?

- Not applicable