Simultaneous detection of bacterial, fungal and oomycete plant pathogens by Nanopore sequencing of multiple genetic markers (Nanotect)

A Data Management Plan created using DMPonline.be

Creator: Marc Venbrux

Affiliation: KU Leuven (KUL)

Funder: Fonds voor Wetenschappelijk Onderzoek - Research Foundation Flanders (FWO)

Template: FWO DMP (Flemish Standard DMP)

Grant number / URL: 1SHFY24N

ID: 206545

Start date: 01-11-2023

End date: 31-10-2027

Project abstract:

Microbial plant pathogens cause considerable yield losses in many economically important crops worldwide. To effectively manage these plant pathogens, their timely detection and accurate identification is of crucial importance. Although detection technologies have improved over the years, both in accuracy and sensitivity, detecting multiple plant pathogens simultaneously remains a challenge. Up until now, a diagnostic tool that combines high speed with high-throughput, multiplexing, and accurate pathogen identification is still lacking. In response to the market demands, a Nanopore-based amplicon sequencing approach (Nanotect) is used in this project. This platform allows simultaneous detection and accurate identification of bacterial, fungal and oomycete plant pathogens in a single assay. To attain this goal, a novel approach is used in which multiple and large genetic markers are sequenced simultaneously on the Oxford Nanopore Technologies MinION device. In this project, optimal genetic markers are selected and amplified. Next, a data analysis pipeline is developed and optimized, which includes a machine learning approach that improves accuracy of taxonomic identification. The Nanotect platform is further evaluated by analysis of spiked water and plant samples. Finally, this novel detection platform is benchmarked to currently used detection techniques regarding several parameters, including accuracy, sensitivity, cost-effectiveness, and analysis time.

Last modified: 29-04-2024

Simultaneous detection of bacterial, fungal and oomycete plant pathogens by Nanopore sequencing of multiple genetic markers (Nanotect)

DPIA

DPIA

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

- Not applicable

Simultaneous detection of bacterial, fungal and oomycete plant pathogens by Nanopore sequencing of multiple genetic markers (Nanotect)

GDPR

GDPR

Have you registered personal data processing activities for this project?

- Not applicable

Simultaneous detection of bacterial, fungal and oomycete plant pathogens by Nanopore sequencing of multiple genetic markers (Nanotect)

Application DMP

Questionnaire

Describe the datatypes (surveys, sequences, manuscripts, objects …) the research will collect and/or generate and /or (re)use. (use up to 700 characters)

Experimental data from wet lab experiments is stored in .xlsx format. Separate tabs will describe the experimental set-up, resulting data, processed data, and analysis. The sequencing results are stored in the following formats: raw sequencing data (.fast5), the basecalled sequences (.fastq), results metadata (.csv). Processed sequences are stored as mapping data (.bam), reference databases (.fasta) and identification and read counts (.csv). Scripts used will be stored in their respective format (.py, .sh, etc.), with a ReadMe.txt file containing instructions. DNA libraries for the sequencing experiments will be coded and stored at -20°C, along with an .xlsx file containing metadata

Specify in which way the following provisions are in place in order to preserve the data during and at least 5 years after the end of the research? Motivate your answer. (use up to 700 characters)

All provisions are in place to store and backup the data as requested by FWO. All data generated will be uniquely coded and stored on a local computer with cloud based incremental backups made every day (OneDrive & internal server, responsible: Marc Venbrux). Promotor prof. Rediers will be the final responsible for long-term data storage and management. For this, the data will be saved for at least 5 years on the internal and secured server of PME&BIM, with automatic daily backups. At the start of the project a data management plan will be made using DMPonline. Published datasets and genetic data will be stored in the following repositories and databases (RDR, DRYAD, GenBank, and ENA).

What's the reason why you wish to deviate from the principle of preservation of data and of the minimum preservation term of 5 years? (max. 700 characters)

NA

Are there issues concerning research data indicated in the ethics questionnaire of this application form? Which specific security measures do those data require? (use up to 700 characters)

NA

Which other issues related to the data management are relevant to mention? (use up to 700 characters)

NA

Simultaneous detection of bacterial, fungal and oomycete plant pathogens by Nanopore sequencing of multiple genetic markers (Nanotect)

FWO DMP (Flemish Standard DMP)

---

1. 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.

see table on next page (appendix 1)

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:

For the purpose of designing primers, as well as creating reference sequence databases for sequence identification, we will make use of the genomes submitted to the NCBI assembly database (https://www.ncbi.nlm.nih.gov/assembly).

Are there any ethical issues concerning the creation and/or use of the data (e.g. experiments on humans or animals, dual use)? Describe these issues in the comment section. Please refer to specific datasets or data types when appropriate.

- No

Will you process personal data? If so, briefly describe the kind of personal data you will use in the comment section. Please refer to specific datasets or data types when appropriate.

- No

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

if succesful, there is a possibility of some data being restricted to retain the intellectual property. More specifically, we will try to protect IP related to a machine learning algorithm for sequence identification, as well as any newly generated pathogen sequences to append to our database.

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 in the comment section to what data they relate and what restrictions are in place.

- No

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 in the comment section to what data they relate and which restrictions will be asserted.

Appendix 1

					Only for digital data	Only for digital data	Only for digital data	Only for physical data
Data Chapter	Dataset Name	Description	New or reused	Digital or Physical	Digital Data Type	Digital Data format	Digital data volume (MB/GB/TB)	Physical volume
WP1: Selection and evaluation of genetic markers	1.1 Obtain genomes and extracting candidate genetic markers for universal primer design of large amplicons	For the purpose of generating novel, larger than conventional, universal primers, a collection of genomes obtained from NCBI assemblies (only representative and reference genomes) is downloaded. This will be in various file formats depending on their availability. Next, candidate genetic markers are extracted from these genome assemblies, preferably by annotation, using an in-house developed script. The sequences of the selected genes are then compiled into a single fasta file.	Reuse existing data	digital	compiled/aggregated data	.fasta .gbff .py	<1 TB	
	1.2 Designing primers	The resulting sequences of the genetic markers are subsequently aligned using MAFFT. The alignment is necessary for the evaluation of the candidate genes, as (1) it allows for a visual confirmation of the presence of highly variable (for identification) and highly conserved (possible primer binding sites) through the generation of a Shannon diversity plot. (2) The alignment is then used to generate a consensus sequence, which can subsequently be used for the creation of possible primers. These primers are selected based on parameters estimated by the primer design software, such as length, degeneracy, melting temperature, etc.	reuse existing data generate new data	digital	compiled/aggregated data	.fasta .py .png .xslx	<1GB	
	1.3 In silico evaluation of designed primers	Compatible primer pairs will be evaluated in silico before continuing with PCR tests. These in silico evaluations will primarily concern the assessment of the broad targetting potential of the universal primers. For this purpose, compatible primer pairs will be used to extract their targetted regions on the large panel of genomes (see 1.1). Subsequently, it will be assessed whether the primers are specific (only the intented target is amplified). And lastly, the amount of mismatches the primers have with their intended targets. For this, a combination of publicly available software tools, as well as in-house developed scripts will be used. The 3 best performing primer pairs for each genetic marker will be used for further analysis.	generate new data	digital	compiled/aggregated data experimental	.tsv .xlsx .gff	<1GB	
	1.4 Obtaining genomic material of microorganisms	A representative panel of microorganisms, obtained from the host-lab's culture collection, will be used to evaluate the performance of the primers. In order to achieve this, genomic material of a taxonomically diverse group of microorganisms is used. The respective microorganisms are cultured, and their DNA is extracted. The microorganisms will be stored in a -80 °C freezer, whilst the genomic DNA will be stored in a -20 °C freezer for the remainder of the project.	generate new data	physical				vials of genomic DNA, labelled and stored in boxes at -20°C

	1.5 Evaluating the primers on a representative panel of microorganisms	The primers are evaluated against the representative panel of microorganisms with PCR reactions using both DNA templates consisting of singular and a combination of genomes (i.e. mock communities). They will be evaluated on factors such as specificity, sensitivity, amplification efficiency, universality (all strains show amplification), etc. In addition, PCR optimizations occur simultaneously by performing temperature gradient PCRs, increasing polymerase concentrations, increasing extension times, etc. The resulting best primer pairs, and their optimized PCR conditions will be used to generate the DNA amplicon libraries for the subsequent Nanopore sequencing.	generate new data	digital	experimental	.xlsx .tiff .jpg		
	1.6 Nanopore sequencing of the genetic markers	For each genetic marker, a sequencing run will be initiated for (1) seperately barcoded amplicons obtained from a single strain, as well as from (2) a series of complex mock communities. The resulting data will be used to evaluate various sequencing result analyses, to select the best performing targets. Furthermore, the discriminative capacities of the selected genetic markers can be evaluated, together with the efficiency and performance of the primers (PCR bias, sequencing bias, etc.).	generate new data	digital	experimental	.fast5 .pod5 .fastq .fasta .xlsx	<5TB	
WP2: Construction of sequencing data analysis pipelines	2.1 Creating a suitable reference sequence database for the selected genetic markers	The selected primers are used to extract gene sequences from publicly available genomes from the NCBI assembly database. Next to the sequence information itself, their taxonomies are linked to their respective sequences.	generate new data reuse existing data	digital	compiled/aggregated data	.fasta .tsv	<100GB	
	2.2 Assigning taxonomic identities to the reads of each genetic marker, using conventional methods	the resulting Nanopore sequences of the genetic markers will be identified by conventional sequence identification methods. Two main routes are explored here: (1) aligning the sequences to the reference sequence database directly and (2) clustering the sequences into OTUs, after which the sequences are identified. Before the sequences are used, data curation occurs by filtering for qualities of the sequences, their length, as well as the sequences being their intented targets.	generate new data	digital	experimental	.tsv .xlsx	<1GB	
	2.3 Exploring alternative sequence identification strategies	Next to already established sequence identification strategies, the implementation of machine learning techniques will be studied. For this, custom written scripts, primarily by utilizing Python, will be used. The sequences previously obtained in 1.6 will be used.	generate new data reuse existing data	digital	experimental	.py .tsv .xlsx	<1GB	
WP3: Initial evaluation and validation on spiked samples	3.1 Creating spiked water and plant samples	In order to further evaluate and assess the performance of the genetic markers, as well as Nanopore sequencing, suitable spiked samples will be created, by either (1) mixing in genomic material with these samples, or (2) using whole cells to spike the samples. A wide range of samples is constructed, and the genomic material obtained from their DNA extraction will be used as template DNA for the PCR reaction, obtaining a DNA library for the Nanopore sequencing runs.	generate new data reuse existing data	physical	experimental			vials of genomic DNA, labelled and stored in boxes at -20°C
	3.2 Sequencing the DNA libraries from the spiked water and plant samples	The resulting amplicons from 3.1 are sequenced, to be used for further analysis of the suitability of the selected genetic markers for pathogen detection.	generate new data	digital	experimental	.fast5 .pod5 .fastq .fasta .xlsx	<10TB	
	3.3 Analysis of the sequencing results of the spiked samples	The optimal sequence identification techniques will be used to evaluate our methods performance on a wide range of samples. This allows us to assess the suitability of our method for real world conditions.	generate new data	digital	experimental	.tsv .xlsx	<1GB	
WP4: Comparison of our method with current state of the art methods	4.1 Obtaining data from other methods	A range of alternative tehcniques will be used to compare our method to. This will primarily consist of techniques such as conventional PCRs, qPCRs, immunological methods, culture based methods, illumina sequencing, etc.	generate new data reuse existing data	physical digital	experimental	.tsv .jpg .xlsx	<1GB	

- No

2. 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).

All protocols used in this project will be explicitly described according to the format we use at our lab. This format comprises the following information:

- purpose and application
- principle
- equipment and reagents
- quality control
- safety precaution
- procedure
- results
- recording data
- method limitation
- reporting results
- reference procedure
- references
- appendices

Daily work will be monitored in an electronic lab notebook on OneNote, with the location of the data resulting from the experiment on the computer. Data will be sorted according to the work package of the project, and each folder will comprise:

- raw data (e.g.: .fast5, .pod5, .fasta, gel pictures, etc)
- compiled data (e.g.: excel file, .csv files)
- scripts together with a readme file explaining its functioning(e.g. .py, .sh, etc [if applicable])
- reports of the obtained results, in a more comprehensible format containing the steps undertaken (.ppt, .docx)
- a read me text file explaining the structure of the respective folder

Will a metadata standard be used to make it easier to find and reuse the data? If so, please specify (where appropriate per dataset or data type) which metadata standard will be used. If not, please specify (where appropriate per dataset or data type) which metadata will be created to make the data easier to find and reuse.

- No

3. Data storage & back-up during the research project

Where will the data be stored?

The data will be stored in three environments:

- locally on the computer (C:\ environment) for daily work
- on the KU Leuven network drive associated with the staff number (:I environment)
- on the cloud storage used at KU (One drive)
- Long-term storage is provided on the KULeuven network (K-)drive managed by the supervisor (Prof. H. Rediers)
- The KU Leuven network drives are automatically backed up every day

For material data (DNA extracts, primers, etc.) each box will be labelled and will contain an information sheet summarizing the experiment and the dates, in order to be able to link it with the lab notebook on OneNote.

How will the data be backed up?

Every month, the data stored on the local environment of the computer (:C environment) will be backed-up on the network environment of KUL (:I environment) and on OneDrive (automatically backed up to the cloud).
Every 6 months, the data will be manually exported to a hard drive for the supervisor of the project to upload on KU Leuven servers (K:\ environment).

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

How will you ensure that the data are securely stored and not accessed or modified by unauthorized persons?

The local environment (C:\ as well as the network (I:\) and OneDrive environment are only accessible with personal credential. The physical storage of experiments results are only accessible for colleagues and other staff members. Lasty, all data backed up on the K-drive is only accessible to the PIs.

What are the expected costs for data storage and backup during the research project? How will these costs be covered?

The cost for storage are the following:

- external hard drive: covered by PME&BIM lab
- OneDrive subscription: covered by KU Leuven
- K-drive: covered by PME&BIM
- physical storage: covered by PME&BIM lab (-80°C freezer and -20°C freezer)

4. 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 the digital data will be saved for at least 10 years according to KU Leuven policy. Plant materials will be stored for the time needed to process the samples. DNA samples obtained during the project will be stored for the duration of the project. All the strains that are used are either currently stored in the strain collection of PME&BIM, or will be added to the strain collection, and will be stored indefinitely.

Where will these data be archived (stored and curated for the long-term)?

Data will be stored on the KUL RDR (Research Data Repositories).

What are the expected costs for data preservation during the expected retention period? How will these costs be covered?

Costs associated to long term data storage on KU Leuven servers is covered by PME&BIM

5. Data sharing and reuse

Will the data (or part of the data) be made available for reuse after/during the project? In the comment section please explain per dataset or data type which data will be made available.

- Other, please specify:

The data will be shared on Open Access Repositories if possible. For example, DNA sequences will be shared on NCBI. Should there be any reason, such as due to intellectual property rights issues, select data will not be made publicly available.

If access is restricted, please specify who will be able to access the data and under what conditions.

In case results, or scripts developed during this PhD, are deemed as intellectual property to be used for further valorization (in consultation with Leuven Research & Development (LRD)), the access will be restriced. The R&D team, supervisor and the PhD student working on the project will have access to the data.

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 in the comment section per dataset or data type where appropriate.

- Yes, Intellectual Property Rights

As mentioned above, should the outcome of this PhD be subject to valorization, certain parts of the obtained data will not be made publicly available.

Where will the data be made available? If already known, please provide a repository per dataset or data type.

The sequencing data will be available for public usage on NCBI
Publication will be done in Open Access journal as much as possible, following KUL policy.

When will the data be made available?

upon publication of research results

Which data usage licenses are you going to provide? If none, please explain why.

Creative Commons Attribution, for utilization of any publicly disclosed data such as those contained within publications.

Do you intend to add a PID/DOI/accession number to your dataset(s)? If already available, you have the option to provide it in the comment section.

- Yes

Not available yet

What are the expected costs for data sharing? How will these costs be covered?

None, the RDR is free of charge. Data sharing will also happen through publication in scientific and professional oriented journals, and the costs associated will be covered by the PME&BIM lab.

6. Responsibilities

Who will manage data documentation and metadata during the research project?

Marc Venbrux

Who will manage data storage and backup during the research project?

Hans Rediers

Who will manage data preservation and sharing?

Marc Venbrux and Hans Rediers

Who will update and implement this DMP?

Marc Venbrux