DMP du projet "Characterizing Network Wirings of Maize Water Deficit Tolerance for Genomic Selection Enhancement"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - DMP template (english)" fourni par Agence nationale de la recherche (ANR).

Plan Details

Plan title	DMP du projet "Characterizing Network Wirings of Maize Water Deficit Tolerance for Genomic Selection Enhancement"
Deliverable	D0.1
Version	First version
Plan purpose/scope	Here will be detailed where all the data, whether generated or re-used in the context of NETWITS, will be stored and made publicly available; which scripts, pipelines and softwares will be used for data analysis, and where the scripts/pipelines/softwares that will be developped during this project will be made available.
Fields of science and technology (from OECD classification)	Biological sciences (Natural sciences)
Language	eng
Creation date	2023-10-20
Last modification date	2023-10-24
License	**Name**: Creative Commons Attribution Non Commercial Share Alike 4.0 International **URL**: http://spdx.org/licenses/CC-BY-NC-SA-4.0.json
Associated documents (publications,reports, patents, experimental plan...), website	• Website : https://anr.fr/Projet-ANR-23-CE20-0024

Project Details

Project title Characterizing Network Wirings of Maize Water Deficit Tolerance for Genomic Selection Enhancement

Acronym NETWITS

Abstract

Climate change is already responsible for massive loss of crop yield worldwide. Adapting varieties to future climatic conditions is thus a timely challenge. However, models used in genomic selection are not yet accurate enough to identify the most adapted genotypes to specific environment, because they lack prior biological knowledge on the molecular bases of crops response to environment. At the molecular level, crops' tolerance to an environmental constraint is highly polygenic, and is determined by complex gene regulatory networks regulating the expression of several biological processes. It is thus likely that tolerance to a constraint can be reached by leveraging distinct networks depending on the genotype. Distal *cis*-regulatory elements (dCREs) are likely to play a key role in articulating these networks. But the diversity of dCRE-articulated gene regulatory networks underlying tolerance to a constraint, and its impact on the selection process, remains an open question. Using maize response to water deficit as a model, I will aim to (i) assess the diversity of gene regulatory network wiring underlying maize WD-tolerance; (ii) determine how network topology and its variations may constrain or facilitate selection; and (iii) integrate these novel information into a predictive model of maize response to WD.

In NETWITS, I will uncover biological information on regulatory networks underlying WD tolerance in maize by integrating genomic and transcriptomic data already available in the lab using cutting-edge approaches in plant systems biology and population genetics. I will then use this biological knowledge to develop a more accurate model to predict maize phenotype in response to WD. NETWITS will bring an important breakthrough in our understanding of the role of dCRE variations and those of the gene regulatory networks they articulate in determining WD tolerance. This will be key to preserve yield while retaining some genetic diversity among cultivated crops.

Funding

- Agence Nationale de la Recherche : ANR-23-CE20-0024-01

Start date 2024-03-01

End date 2028-02-28

Partners

- Génétique quantitative et Evolution - Le Moulon https://ror.org/012fqzm33

Research outputs :

1. Software for detecting signatures of positive seletion (Software)
2. Model of polygenic selection targeting quanttative phenotype with pleiotropy (Model)
3. Model for predicting phenotype in response to water deficit (Model)
4. Software for inferring gene regulatory networks including genetic variation effects

(Software)
5. Software implementing the DROUGHTPREDICT model (Software)
6. Data set (RNA-Seq Gene ATLAS) generated by the AMAIZING project (Dataset)
7. Dataset (RNA-seq, genotyping and phenotyping data from the drought panel) generated by the AMAIZING project (Dataset)
8. Dataset (RNA-seq, genotyping and phenotyping data from the drought panel) generated by the SeqOccIn project (Dataset)

Contributors

Name	Affiliation	Roles
Fagny Maud - https://orcid.org/0000-0002-7740-2521/	Institut national de recherche pour l'agriculture, l'alimentation et l'environnement - 198017849T	• DMP manager • Personne contact pour les données (POLYGENDETECT, DROUGHTPREDICT, POLYGENSIM, GENEATLAS AMAIZING, EGRET, DROUGHTPREDICTSOFT, SEQOCCIN PANEL, AMAIZING PANEL) • Project coordinator

DMP du projet "Characterizing Network Wirings of Maize Water Deficit Tolerance for Genomic Selection Enhancement"

1. Data description and collection or re-use of existing data

Software for detecting signatures of positive seletion

1a. How will new data be collected or produced and/or how will existing data be re-used?

- A software will be developped to implement the POLYGENSIM model and compute summary statistics.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

- Scripts in R, python and bash (a few Ko).
- A docker allowing users to install and use the software easily (a few Ko).

Model of polygenic selection targeting quanttative phenotype with pleiotropy

1a. How will new data be collected or produced and/or how will existing data be re-used?

A new forward-in-time model of population evolution simulating polygenic selection with epistasis and pleiotropy will be developped.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

A TXT file describing the model (a few Ko).

Model for predicting phenotype in response to water deficit

1a. How will new data be collected or produced and/or how will existing data be re-used?

A new model to predict yield in response to drought, that will include biological information (network topology, selection scores) will be developped.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

A TXT file describing the model (a few Ko).

Software for inferring gene regulatory networks including genetic variation effects

1a. How will new data be collected or produced and/or how will existing data be re-used?

Question sans réponse.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

- Scripts in python (a few Ko).
- The new version will be pushed on the netZooPy github.

Software implementing the DROUGHTPREDICT model

1a. How will new data be collected or produced and/or how will existing data be re-used?

- An implementation of the DROUGHTPREDICT model will be provided.

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

- Scripts in R, python and bash (a few Ko).
- A docker allowing users to install and use the software easily (a few Ko).

Data set (RNA-Seq Gene ATLAS) generated by the AMAIZING project

1a. How will new data be collected or produced and/or how will existing data be re-used?

- Existing data (DNA methylation, RNA-Seq, assembled genomes and phenotyping data) will be re-analysed (distal cis-regulatory elements detection, in-silico TFBS predictiong, gene regulatory network inference).
- Potential distal cis-regulatory elements will be detected
- Gene regulatory networks will be inferred

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

- List of potential distal cis-regulatory elements for 6 maize lines. Zipped TXT files (~12Mo each, 84Mo)
- List of predicted TFBS for 6 maize lines, generated using the FIMO software (~70Mo each, ~7-8Go total)
- Gene regulatory networks for each tissue, cell line, and watering condition (17 x 6 = 102). Zipped TXT files (~80Mo each, ~8-9Go total).

Dataset (RNA-seq, genotyping and phenotyping data from the drought panel) generated by the AMAIZING project

1a. How will new data be collected or produced and/or how will existing data be re-used?

- Existing data (RNA-Seq, genotyping and phenotyping data) will be re-analysed.
- eQTL bipartite networks will be inferred

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

- eQTL networks for each watering condition (150Mo x2 =~300Mo)

Dataset (RNA-seq, genotyping and phenotyping data from the drought panel) generated by the SeqOccIn project

1a. How will new data be collected or produced and/or how will existing data be re-used?

- Existing data (RNA-Seq, genotyping and phenotyping data) will be re-analysed.
- eQTL bipartite networks will be inferred

1b. What data (for example the kind, formats, and volumes), will be collected or produced?

- eQTL networks for each watering condition (150Mo x2 =~300Mo)

2. Documentation and data quality

2a. What metadata and documentation (for example the methodology of data collection and way of organising data) will accompany the data?

- For generated data, metadata will include how data will be produced (raw data used and where to find them, softwares - including their version - and pipeline used and where to find them).
- For scripts, pipelines and softwares, a README file will be included in the repository to explain how to install and use the products. A test dataset will be provided for each software.

2b. What data quality control measures will be used?

Not Applicable

---

3. Storage and backup during the research process

3a. How will data and metadata be stored and backed up during the research?

- All generated data and softwares/scripts/pipeliens will be stored on a computer space dedicated to this project on the GQE computer servers, which is automatically backed-up.
- The scripts/pipeline/softwares will be stored on a private space on the INRAE ForgeMIA to keep track of the modifications and allow access of all participants to this projects

---

3b. How will data security and protection of sensitive data be taken care during the research

Data security is ensured by the IT team of GQE. Processes have already been implemented to ensure data security, and protection according to the European and French legislation.

---

4. Legal and ethical requirements, code of conduct

4a. If personal data are processed, how will compliance with legislation on personal data and on security be ensured?

NA

4b. How will other legal issues, such as intellectual property rights and ownership, be managed? What legislation is applicable?

The software and data will be published on ForgeMIA, in OpenAccess, under **CC-by-nc-sa** licences. INRAE has legal intellectual property rights and ownership. The intellectual property rights and ownership will be managed according to

what is described in the project's leader and participants contracts with their employers.

4c. What ethical issues and codes of conduct are there, and how will they be taken into account?

Generated data and softwares/pipelines/scripts will be stored and made available, and results will be published according to the FAIR principles, and following the ANR rules (open access for publications, open source for softwares/pipeline/code).

---

5. Data sharing and long-term preservation

5a. How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

- Re-used data (genomes, RNA-seq, DNA methylation, phenotypes) are already or will be made available on appropriate public databases before publication of our results. Embargo are in effect for some of the data, but will expire before the end of the NETWITS project.
- Generated data will be made available on publication of the papers describing the results, or before the end of NETWITS. They will be referenced and stored on https://entrepot.recherche.data.gouv.fr/dataverse/inrae. No restriction further than those imposed by the licence, no embargo.
- The software will be published on the INRAE ForgeMIA and on the project leader github account. No restriction further than those imposed by the licence, no embargo. Publication will be made at the latest before the end of the NETWITS project.

5b. How will data for preservation be selected, and where data will be preserved long-term (for example a data repository or archive)?

- All data, code and software generated during this project will be stored on the GQE server and a copy will be generated on band.
- Generated data will be deposited on the INRAE dataverse (https://entrepot.recherche.data.gouv.fr/dataverse/inrae).
- Software will be stored on the ForgeMIA.

5c. What methods or software tools are needed to access and use data?

R
python (pyslim, tskit)
bash

5d. How will the application of a unique and persistent identifier (such as a Digital Object Identifier (DOI)) to each data set be ensured?

Maud Fagny and the [IR] will be responsible to ensure that all data and softwares have a proper DOI. This will be done through the data INRAE portal.

---

6. Data management responsibilities and resources

6a. Who (for example role, position, and institution) will be responsible for data management (i.e. the data steward)?

Maud Fagny (Researcher, INRAE) and [IR] (Engineer, INRAE) will be responsible for data management.

6b. What resources (for example financial and time) will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

A Research Engineer will be hired to help the project manager with data management.