Stirrer

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateurs du PGD : mathieu rousseau-gueutin, Gautier Richard

Affiliation du créateur principal : INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement

Modèle du PGD : ANR - DMP template (english)

Dernière modification du PGD : 07/09/2020

Financeur : ANR

Numéro de subvention : ANR-19-CE20-0013

Résumé du projet :

Generating diversity using meiotic recombination is the main mechanism through which agro-economic players can reduce the environmental footprint of production while maintaining the yield and quality. However, recombination is strictly controlled, with 1 to 3 crossovers (COs) per homologous chromosomes. In Brassica, we showed that we can increase COs number (3.4x) and create COs in normally cold regions (pericentromeres) by using allotriploid hybrids. The mechanisms involved in this modification of meiosis regulations remain unknown while it opens new avenues for B. napus (oilseed rape) breeding. Several questions will be tackled in the project: (1) Are epigenetic modifications and transcriptomic regulations responsible of changing the recombination rules? (2) How far can we break the linkage disequilibrium by successive cycles at the allotriploid level? Are the modifications reversible? These results will have major applied impacts by providing a novel, effective breeding strategy.

Chercheur Principal : mathieu rousseau-gueutin

Identifiant ORCID : https://orcid.org/0000-0002-1130-1090

Contact pour les Données : mathieu rousseau-gueutin

Stirrer

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

Le stockage de données NGS (BS ChIP et RNA seq) dans Genouest pendant l'utilisation des données, et sur Agrodataring en duplicata. Genouest est une plateforme bio-informatique ouverte et nationale proposant des services de stockage et de calcul. Agrodataring quant à elle correspond au serveurs de stockage longue durée de l'INRAE. Notre UMR y dispose d'un espace disque de 100 To. Recherche de financement en cours pour une 3eme plateforme de stockage.

Après publication, des fichiers README seront générés et stockés dans le espaces décrits ci-dessus prévus pour le "stockage froid" afin de renvoyer vers les repository publiques contenant les données publiées et accessibles à l'ensemble de la communauté scientifique.

En effet, les données NGS seront déposées sur des bases de données internationales via SRA. Un BioProject sera ainsi défini où de nombreuses metadonnées seront associées aux données NGS: matériel végétal (génotypes utilisés, rang foliaire utitilisés, conditions de culture et homogénéité des réplicas biologiques), protocoles wet-lab, protocoles d'analyse des données (outils, versions, workflows, options utilisées).

Un repository github sera de plus créé. Les données de génotypage, brutes et analysées, y seront stockées en plus de metadonnées telles que les scripts d'analyse, la provenance du matériel végétal, et les conditions de cultures.

Avant publication l'ensemble des données seront utilisées de manière privée. Après publication il n'y aura aucune restriction de l'utilisation des données, via SRA, le github, ainsi que via le stockage longue durée des données sur Agrodataring. Dans ces différents repositories, la provenance des données ainsi que les metadata seront précisés.

* Données NGS:

- Brutes: fastq.gz ; compression due à la taille importante des fichiers

- Processées : bed et bedGraph pour BS-seq et ChIP-seq, csv ou txt pour RNA-seq.

* Données de Génotypage:

- Brutes : Données de TraitGenetics, plusieurs dossiers avec differents formats de fichiers.

Les fichiers important sont les .idat.

- Processées : Projet GenomeStudio, format .bsc. cluster_file en .egt. fichier brut de sortie en .xlsx.

* Données de Cartographie:

Données sous forme de fichiers .xlsx, txt et docx.

c'est un deuxième process des données de génotypage (avec en plus des résultats de blast pour les mks SNP sur les génomes de référence)

* Données Cytogénétiques:

- Brutes : images au format ? .czi

- Processé : images et fichiers tabulés. Pas d'images processées, uniquement des résultats en fichiers .xlsx, et des scripts (Zen, Fiji, Python et R)

L'ensemble des ces formats ont été choisis car étant adaptés aux entrepôts de données et facilement réutilisables sans conversion.

A l'heure actuelle 1 To de données ont été produites, et environ 1 à 2 To devrait être produits au cours de ce projet.

2. Documentation and data quality

Les métadonnées seront renseignées selon les réglementations imposées par les repositories. Elles correspondront au matériel végétal (génotypes utilisés, rang foliaire utitilisés, conditions de culture et homogénéité des réplicas biologiques), protocoles wet-lab (kits, conditions expérimentales, composition des réactifs), protocoles d'analyse des données (outils, versions, workflows, options utilisées).

Au cours du projet les données brutes type NGS seront déposés et utilisés pour processing et analyse en aval via le serveur Genouest. Les fichiers processés finaux seront mis à disposition de l'ensemble des collaborateurs sur un serveur partage: CeSGO, géré par la plateforme Genouest.

Le premier niveau d'organisation des données sera lié à leur origine (BS-seq, RNA-seq, ChIP-seq, données de génotypage, microscopie), puis au sein de chaque sous-dossier, les données seront séparées selon

Stockage CeSGO - Genouest

|

|__ Données_NGS
| |__ Processées
| | |_ Matériel_végétal n
| |__ Analyses
| |__ Analyse_comparative n
| |__ Input
| |__ Scripts
| |__ Output (figures, tableaux)
|___ Cartographie
|___ Génotypage
|___ Microscopie

Autant que possible des README seront générés à chaque niveau de l'arborescence afin de décrire au mieux l'origine des données, quelque soit leur niveau d'analyse (brutes, processées, finales).

NGS: 3 réplicats biologiques correspondant à des feuilles provenant de boutures du même génotype. L'ensemble des génotypes utilisés par expérimentation ont été cultivés au même moment dans la même chambre de culture en milieu contrôlé. Tous les échantillons ont été prélevé le même jour au même stade de développement, sur le même rang foliaire.

Avant l'envoi au séquençage, tous les échantillons sont controlés par des contrôles qualités:

Durant le séquençage, des contrôles qualités sont effectués par la plateforme de séquençage.

A la réception des données, des contrôles qualités bio-informatique sont effectués par nos soins: qualité de l'appel des bases, qualité de mapping (quantité de reads mappés, fraction de reads mappant de manière unique), vérification de l'origine des banques (homogénéité des réplicats via ACP après mapping).

3. Storage and backup during the research process

Après réception des données brutes sur disque dur externe via la plateforme de séquençage, celles-ci sont stockées en copie sur Genouest (Rennes) et sur Agrodataring (Paris). Les données en cours d'analyse seront stockées sur Genouest.

Concernant la sécurité des données, elles ne seront accessibles qu'aux participant au projet grâce à des dossier protégés sur Genouest. Il y aura un droit de lecture pour l'ensemble des collaborateurs, et un droit d'écriture uniquement pour les membres participant à l'analyse. Les données brutes seront quant à elles uniquement accessibles en lecture.

Les données importantes (brutes et processées finales) seront stockées de manière securisée sur un serveur de l'institut (Agrodataring).

4. Legal and ethical requirements, code of conduct

Aucune donnée personelle ne sera utilisée dans de projet et l'ensemble des données générées sera la propriété de l'INRAE.

L'ensemble des données sont la propriété de l'INRAE, et seul des membres de l'INRAE sont impliqués dans ce projet. Toutefois, suite à la publication des articles, les données seront accessibles à l'ensemble de la communauté scientifique. Donc aucune restriction pour la réutilisation des données.

Aucun problème éthique ou déontologique ne sera rencontré au cours de ce projet.

5. Data sharing and long-term preservation

Les données seront partagées entre les différents membres du projet durant leur analyse via le service CeSGO de Genouest, mais non accessibles publiquement.

Par la suite, après publication… SRA… github

Les données seront archivées sur le long terme sur le serveur Agrodataring pour une durée indéterminée.

Les travaux sont prévus d'être publiés fin 2021.

Aucune données ne doit détruite pour des raisons contractuelles, légales ou réglementaires.

Les données brutes, finales d'analyses et les scripts seront conservés sur le long terme. Les données intermédiaires seront effacées.

Les données brutes (NGS et génotypage principalement) et scripts seront utiles à la communauté scientifique dans le cadre d'autres projets de recherche.

Les données NGS seront déposées sur les bases de données du NBCI (SRA et GEO). Les données de génotypage seront accessibles à la communauté scientifique en supplémentaire de la publication via le github qui sera créé. Les fichiers brutes de génotypage seront disponibles à la demande (réflexion actuellement en cours par la communauté Brassica au niveau international pour la mise en commun de ce type de données). Ces données seront stockées sur le long terme sur les serveurs INRAE Agrodataring.

Via SSH ou l'interface CesGo pour Genouest, et Agrodataring.

A l'heure actuelle seule les données de génotypage nécessite un logiciel spécifique pour l'analyse des données brutes. Néanmoins les données processées finales seront disponible sur le long terme.

Les données pas encore publiées seront disponibles aux membres du projet de manière directe et illimité. Il en sera de même pour les données publiée via SRA ou github.

Les données en cours d'analyse ne disposerons pas d'un DOI. Néanmoins des noms explicites seront donnés aux fichiers, qui, accompagnés des fichiers README, permettront d'assurer leur tracabilité et réutilisation.

Les données publiées sur SRA disposerons de multiplies identifiants tels que des numéros BioProject, BioSamples, etc. Le repository github comportera un DOI attribué par Zenodo, qui suivra le contrôle de version permit par github, si de nouvelles données doivent être ajoutées.

6. Data management responsibilities and resources

- Mathieu Rousseau-Gueutin (Chargé de Recherche INRAE): gestion, établissement de la structure et de l'archivage des données de l'ensemble du projet, tout au long de ce dernier et au-delà.
- Gautier Richard (Post-doc INRAE): gestion des données NGS: dépôt les différents entrepot de données, privés (Genouest, Agrodataring) et publiques (SRA), gestion de la qualité des données. Gestion du github des données de génotypage.
- Franz Boideau (Doctorant INRAE): gestion des données de génotypage, de cartographie et de microscopie , suivi du matériel végétal, gestion de la qualité des données et metadonnées associées.
- Cyril Falentin (Ingénieur d'Étude INRAE): gestion de l'ensemble des metadonnées: protocoles expérimentaux.

Ces quatre personnes participe à la mise en oeuvre, au suivi et à la mise à jour du plan de gestion des données.

Cette mise à jour aura lieu tout au long du projet, notamment à l'arrivée de nouvelles données et dans le cadre de la prise de prise décision de la communauté Brassica concernant la mise en commun des données de génotypage.

Une semaine de travail a été prévue pour la mise en ligne des données brutes et processées dans les formats demandés par la communauté scientifique (SRA, github…).

A l'heure actuelle, il n'y a pas de frais de stockage des données. Néanmoins, il sera nécessaire de participer à l'achat d'une nouvelle baie de stockage (évaluée à 20.000e) afin d'assurer la duplication des données à l'INRAE pour stockage long terme ; le stockage Genouest n'ayant pas vocation à être pérenne.