RARe-BRC4Env

Données de séquençage numérique

L'exemple des données « sols »

Samuel Mondy[1], Christian Mougin[2]

[1]UMR Agroécologie Dijon, [2]UMR ECOSYS Versailles

Les thèmes abordés

- Le contexte
- Les infrastructures de recherche
- La nature et la gestion des données
- Les écueils et difficultés rencontrés

Le contexte

Les données de séquence numérique de ressources génétiques, des enjeux pour les recherches en agriculture, sciences de l'environnement et de l'alimentation

- **De nombreuses applications : biodiversité, étude d'impact…**
- **Une évolution rapide des outils d'analyse et d'interprétation**
- **Le contexte de l'Open Science : impact sur les relations, obligations**

Les infrastructures de recherche

Ressources Agronomiques pour la Recherche (RARe, https://www.agrobrc-rare.org/) et son pilier Environnement (BRC4ENv, https://www.brc4env.fr/)

- Fédèrent plusieurs CRBs stockant des ressources et échantillons environnementaux (sols, eau, sédiments)

Analyses et Expérimentations sur les Ecosystèmes – France (AnaEE-F, https://www.anaee-france.fr/)

- Fédèrent plusieurs plateformes de génomiques impliquées dans le séquençage d'échantillons environnementaux
- GenoSol est impliquée dans ces IR

Les règles pour l'utilisation des données

Analyses et Expérimentations sur les Ecosystèmes – France (AnaEE-F, https://www.anaee-france.fr/)

- **Charte d'accès**

« 6.1.5. Accès aux données
La plateforme est en accord avec les principes de l'Open Science et la politique de l'Infrastructure de Recherche AnaEE-France.
Il respecte également les exigences émanant des agences de financement nationales et internationales concernant l'accessibilité des données. »

- **Plateforme en génomique moléculaire**

GenOuest
E-DNA

Les règles pour l'utilisation des données

Ressources Agronomiques pour la Recherche (RARe, https://www.agrobrc-rare.org/) et son pilier Environnement (BRC4ENv, https://www.brc4env.fr/)

- **Accès à la ressource**
- **Connaissance des métadonnées associées**
- **Obligation d'un statut de l'échantillon**

Contexte et objectifs

La plateforme GenoSol de l'UMR 1347 Agroécologie est une structure unique en Europe, centrée sur la biologie des sols.

- un Conservatoire des Ressources Génétiques Microbiennes (labellisé ISO9001-2008
- une Plateforme Technique de caractérisation moléculaire du microbiote du sol
- une expertise bioinformatique et statistique sur la diversité microbienne des sols et de l'environnement.

Objectif : mesure de l'impact de l'homme (agriculture, réchauffement..) sur la capacité des sols à rendre un service écosystémique

Plat. Strat. Nat. (2013)

C.R.B. (2013)

Investissements d'Avenir (2014)

Certification ISO 9001 pour le conservatoire (Septembre 2015)

De nouvelles données à acquérir

De nouveaux types de données

GESTAD
Gestion Durable des Adventices

Amplicon : 16S, 18S, gènes de fonction
Stock de graines des adventices

GEAPSI
Déterminismes Génétiques et Environnementaux de l'Adaptation des Plantes a des Systèmes de Cultures Innovants

Génomes
Transcriptomes
Amplicon : 16S, 18S, gènes de fonction

GenoSol PLATEFORME

BIOmE
Biologie et fonctions écosystémiques des sols

Amplicon : 16S, 18S, gènes de fonction
Métagénome
Métatranscriptome

IPM
Mécanismes et gestions des Interactions Plantes Microorganismes

Amplicon : 16S, 18S, gènes de fonction
Génomes
Transcriptomes
Métagénome
Métatranscriptome

Système d'Informations Environnementales

Mise en place d'une Système d'information environnemental (2009-2014)

Serveur de calcul permettant les analyses bioinformatiques et statistiques
Base de données *MicroSol* stockant les informations relatives aux échantillons de sols

Site Web permettant l'affichage de la plateforme sur l'extérieur

Stockage des indices et métriques mais pas d'accès aux données brutes

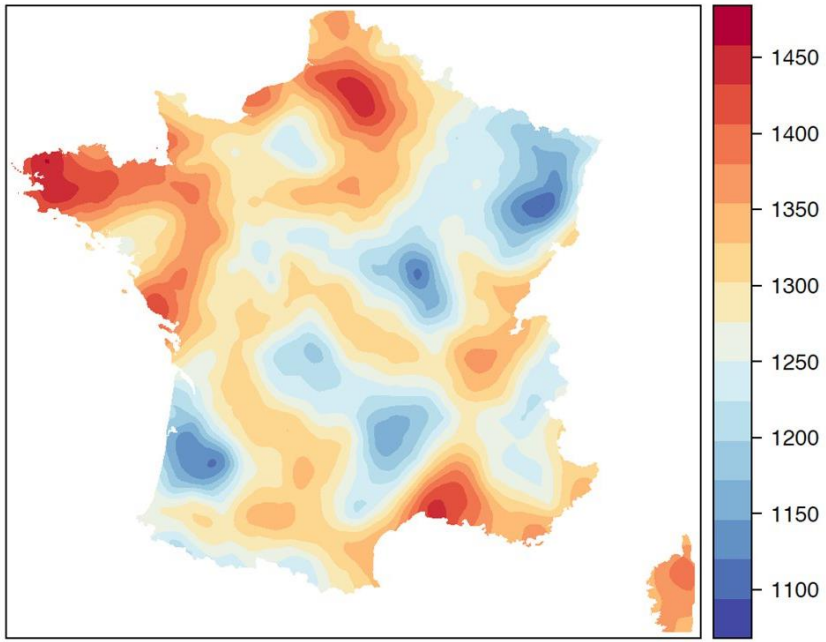
Gestion des droits d'accès : données non accessibles directement

Exemple de productions

Évolution de la biodiversité bactérienne des sols

http://indicateurs-biodiversite.naturefrance.fr/indicateurs/evolution-de-la-biodiversite-bacterienne-des-sols

Cartographie nationale de la diversité bactérienne des sols de France
(nombre de taxons bactériens par gramme de sol)



Source : © GIS Sol, UMR Agroécologie – équipe BIOCOM, plateforme GenoSol

Données :
Amplification du gène 16S ribosomiques (V3-V4) sur ADN extraits des sols du RMQS

Données publiées et accessibles
https://www.ebi.ac.uk/ena/data/view/PRJEB21351

Devenir des données de séquençages issues de la plateforme

Données de séquençage : Données brutes ou nettoyées, assemblées
PI : en fonction de la nature de la convention

Charte AnaEE : Embargo (5 ans pour AnaEE-F et 6 mois pour AnaEE-Europe) après publication des données

Editeurs scientifiques : Données accessibles au moment de la publication

SRA ou EBI : stockage des données brutes ou assemblées (fastq, sff, gff, fasta)

Problème : certains scientifiques vont mettre uniquement les séquences assemblées, ou oublier de préciser à SRA ou EBI la fin de l'embargo / Dépendance pour la fourniture du service

Qui met en ligne : la plateforme ou les partenaires scientifiques ?

Utilisation des métadonnées : restriction d'utilisation, anonymisation, RGPD

Réutilisation des données issues de la plateforme GenoSol

Les données de séquençage deviennent des méta-données associées aux échantillons de sols

AnaEE-F/RARe – BRC4Env : Mise à disposition des ressources génétiques des sols

Données brutes dans un DataVerse institutionnel?

Affichage des DOI dans moteur de recherche RARe?

Quel statut juridique des données de séquençage issues du sol ?

Intégrer l'APA, Le protocole de Nagoya, l'Open Science

MERCI POUR VOTRE ATTENTION