Identification de gènes diagnostic chez les *Rhizobium leguminosarum* à l'aide de la grille informatique sous l'environnement WISDOM

Sébastien Guizard (LPC Clermont-Ferrand), Xavier Bailly (INRA Clermont-Ferrand), Matthieu Reichstadt (INRA Clermont-Ferrand)

Introduction

Rhizobium leguminosarum :
- Espèce bactérienne capable de fixer l'azote atmosphérique
- Différentes souches sont capables de s'associer avec des plantes légumineuses différentes

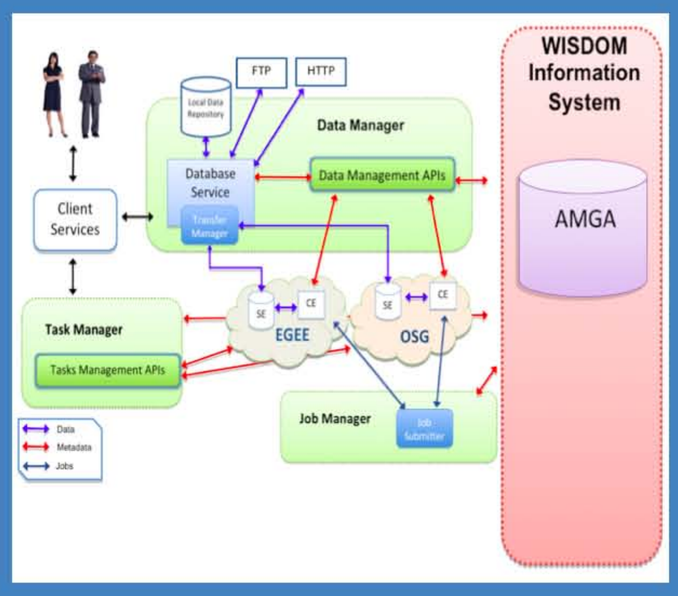
Question:
- Comment évolue la différentiation des génomes de R. leguminosarum isolées de la vesce et du trèfle

Comment ?
- Mesurer la différenciation
- Evaluer la distance génétique moyenne entre souches issues de plantes hôtes différentes par rapport à la distance génétique moyenne entre deux souches de R. leguminosarum

Quels moyens ?
- Mise en place d'un workflow
- Utiliasation de la grille de calcul avec l'environnement de production WISDOM

Le workflow :
- Données de départ : contigs issus de l'assemblage de novo des séquences provenant de 36 souches de R. leguminosarum bv trifolii et de 36 souches de R. leguminosarum bv viciae
- Extraction des séquences des souches étudiées pour les gènes annotés sur les génomes de R. leguminosarum (GenBank)
- Réalisation d'un alignement multiple des séquences extraites pour chaque gène
- Obtention d'un arbre phylogénique pour chaque alignement
- Calcul pour chaque gène du niveau de différenciation

L'environnement



Un environnement en 3 parties

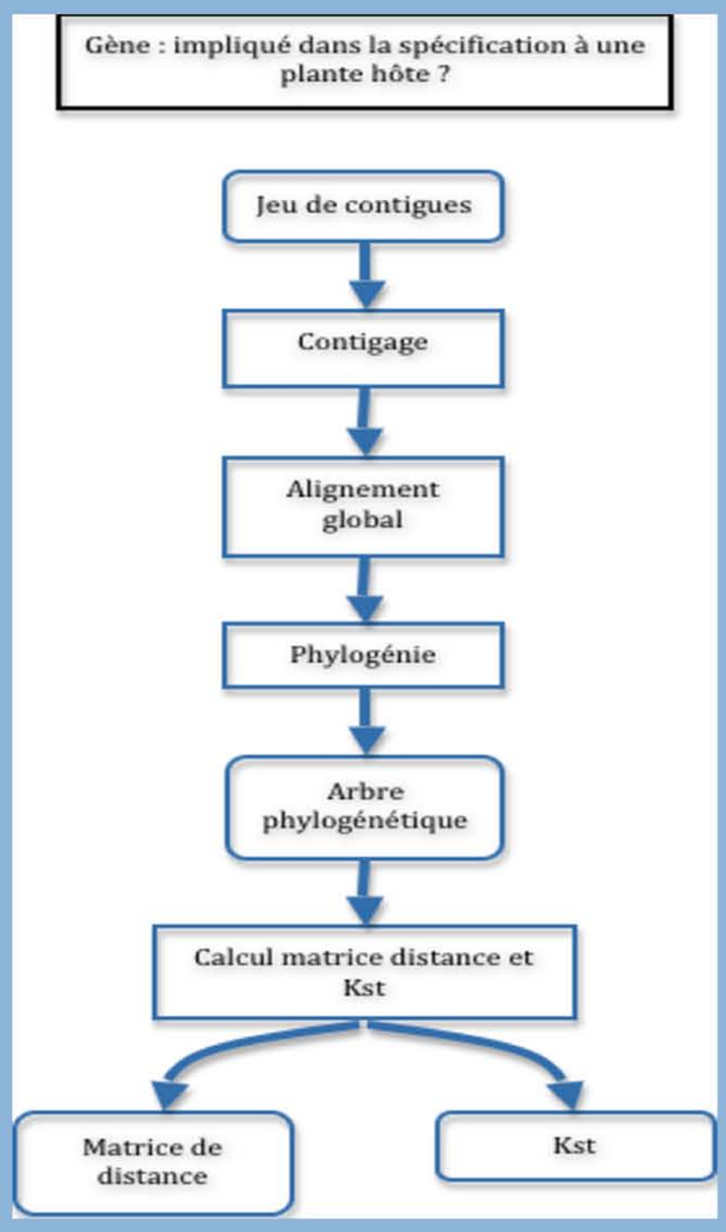
- Le job manager : responsable de la soumission des jobs, il a pour rôle de s'assurer d'un nombre de ressources suffisant pour traiter l'ensemble des tâches qui sont soumises.

- Le data manager : responsable de la copie, de la réplication et de la récupération de toutes les données envoyées sur la grille, notamment tout ce qui concerne les bases de données biologiques.

- Le task manager : responsable du traitement des tâches (calculs) soumis par les utilisateurs de l'environnement Chaque tâche est basée sur 3 éléments :

  - un service
  - des données en entrée
  - des paramètres de calcul

Le workflow



7 étapes:
- sélection des gènes à utiliser
- mapping des contigs sur les séquences retenues
- regroupement des séquences
- alignement global pour chaque gène
- création d'un arbre phylognétique
- calcul des Kst

Résultats
Fusion des listes de gènes :
nombre de gènes : 50% (20559 à 10368)
Mapping et contigage :
648 fichiers créés, 417Mo, 24h →1h30 sur grille
DispatchGenes :
8486 fichiers créés, nombre de gènes étudiés passe de 10368 à 8486
Alignement global et création de l'arbre phylogénique :
42430 fichiers (5/gène), 20 jours → 12h sur grille
total :75368 fichiers