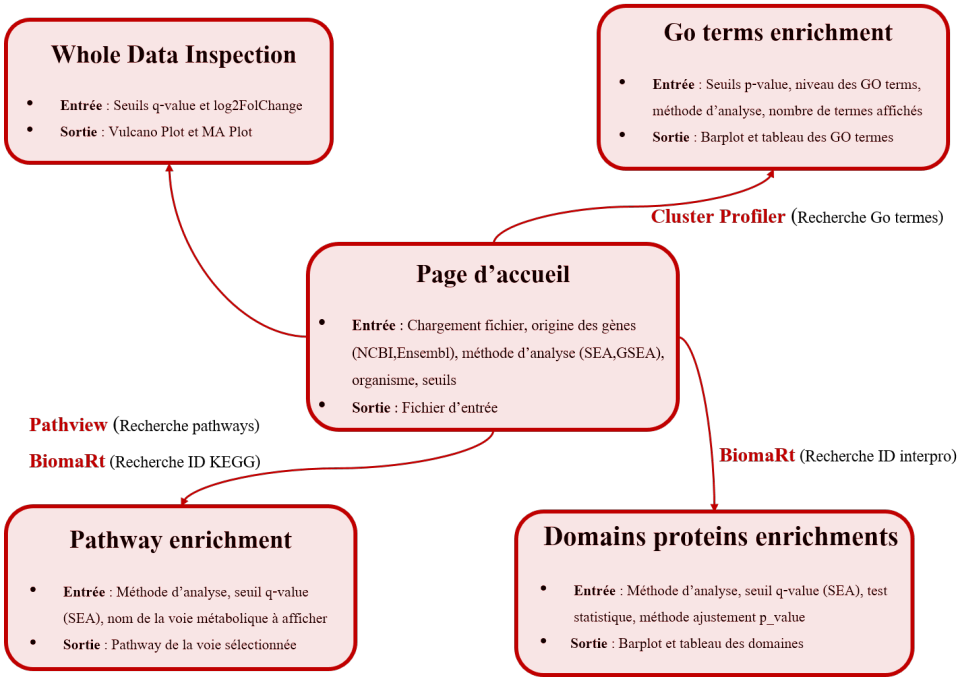
Documention de l'interface Shiny d'enrichissement fonctionnel

Pauline Barbet
Benjamin Bourgeois
Thomas Gabilly
Lysiane Hauguel

15 Mars 2017

1 Schéma général de l'application



2 Explications fondamentales des modèles SEA et GSEA

2.1 Méthode SEA

Cette méthode d'analyse d'enrichissement fonctionnel se base sur la présélection d'un ensemble de gènes par l'utilisateur présentant une différence d'expression par rapport à la normale. Ces gènes ont une p-value inférieure ou égale 0.05 et un fold change supérieur ou égal à 1.5. Des tests d'enrichissement pour chaque type d'annotation sont alors réalisés. Les termes enrichis qui franchissent le seuil d'enrichissement sont présentés dans un tableau avec les termes et leur p-value d'enrichissement. Le calcul de la p-value d'enrichissement peut être effectué à l'aide de différentes méthodes statistiques : les test du χ^2, de Fisher et les loi binomiale et hypergéométrique.

Il s'agit d'une méthode très efficace pour extraire une signification biologique majeure à partir d'une grande liste de gènes à partir d'études génomique et transcriptomique.

2.2 Méthode GSEA

Le principe de cette méthode d'analyse d'enrichissement fonctionnel est le même que pour la méthode SEA sauf qu'il n'y a pas de sélection de gènes différentiellement exprimés par l'utilisateur. Tous les gènes de l'expérience sont utilisés.

Cette stratégie est bénéfique à l'analyse d'enrichissement sous deux aspects :

— Les facteurs arbitraires dans l'étape de sélection des gènes de la méthode SEA qui pourraient avoir une incidence sur l'analyse d'enrichissement sont réduis

— Toutes les informations obtenues à partir des expériences sont utilisées en permetant aux gènes qui n'ont pas passé le seuil de sélection d'expression différentielle de contribuer à l'analyse.

Les 2 précédents avantages de la GSEA sont aussi un limite dans certaines études biologiques. En effet la méthode GSEA nécessite de résumer chaque gène en une valeur biologique (ex : fold change). Il est parfois difficile de résumer de nombreux aspects biologiques d'un gène en une seule valeur lorsque l'étude biologique et la plate-forme génomique sont complexes.

Plus d'informations sur ces méthodes sont disponible sur cet article de Da Wei Huang.al

3 Problématique des génomes peu ou mal annotés

Le fait que le génome étudié soit peu ou mal annoté est susceptible de fausser les résultats des analyse d'enrichissement. En effet si certaines informations sur les gènes ne sont pas disponible tous les gènes de l'ensemble d'étude ne contribueront pas à l'analyse.

4 Le niveau des GO terms

Le niveau d'un GO terms correspond à son degré de précision du terme. Le niveau de précision va de l'organisme à la molécule. Plus le niveau est élevé, plus la précision est importante.

5 Choix des seuillages

Les seuillages du risque p-values de l'application sont définies par défaut à 5% car c'est la valeur utilisée en biologie pour les tests statistiques. Il peut être augmenté ou réduit en fonction de l'expérience et du risque accepté.