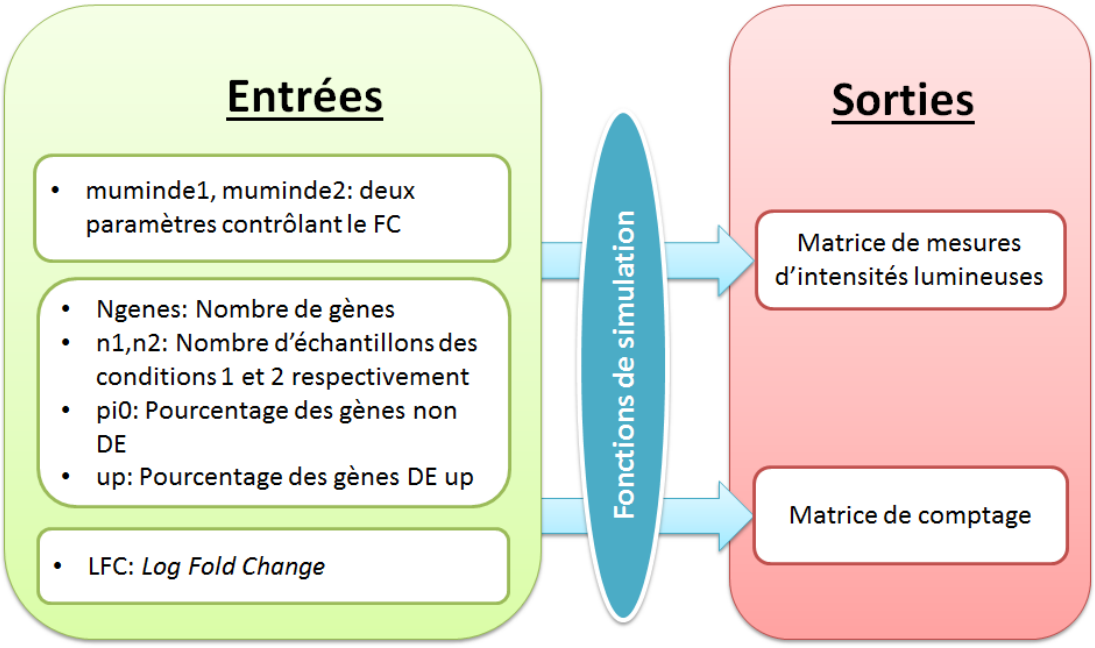
Descriptions

On se place dans le cas où le jeu de données est partagé en deux classes, représentant deux conditions (phénomènes) biologiques différents. Cela est valable pour les deux technologies (microarrays/ RNA-seq).

Nous générons une matrice de taille N*m. Nous fixons le nombre de gènes à nGenes=10.000, le nombre total des échantillons à m=150: n1=135 et n2=15 avec n1 et n2 sont le nombre d'échantillons pour la première condition et la deuxième condition, le pourcentage des gènes non différentiellement exprimés (NDE) à pi0=90% et le pourcentage des gènes surexprimés (gènes up) à up=50%. Ces paramètres seront communs aux deux fonctions de simulation.

Pour la fonction de simulation des données *microarrays*, on ajoute deux autres paramètres permettant de contrôler la variation du vecteur des fold-changes (FC). Concernant la fonction de simulation des données de comptage, elle prend 1 paramètres d'entrée en plus sous forme de vecteur représentant FC. Voir le schéma en dessous pour plus de clarté.



Apres avoir généré les matrices des données de simulations, nous pouvons appliquer la méthode de normalisation choisie par l'utilisateur aux données de comptage (les données *microarrays* simulées sont préporocessées et normalisées), ensuite nous effectuons une standardisation aux données normalisées (*microarrays*/RNA-seq) . Enfin nous fusionnons les deux matrices d'une manière horizontale, puisque les deux matrices comportent exactement le même nombre de gènes et sont structurées de la même façon.

I. Le fichier ***counts_data_simulator.R*** contient le scripte R qui permet de simuler les données de comptages. Il prend comme 6 arguments :

- Arg1 : Nombre de gènes total (entier naturel : nGenes=10.000)
- Arg2: Nombre d'échantillons pour la condition 1 (entier naturel n1=135)
- Arg3 : Nombre d'échantillons pour la condition 1 (entier naturel n2=15)
- Arg4 : Pourcentage de gènes non DE (nombre réel pi0=0.9)
- Arg5 : Pourcentage des gens up (nombre réel up=0.5)
- Arg6 : Fichier 'txt' contenant un vecteur (matrice d'une colonne) FC (fc= "FC.txt")

La sortie est sous forme d'une matrice de comptage sauvegardée dans le dossier « Simulation » sous format 'txt'

II. Le fichier ***Array_simulator.R*** contient le scripte R qui permet de simuler les données *microarrays* (Affy). Il prend comme 7 arguments :

- Arg1 : Nombre de gènes total (entier naturel : nGenes=10.000)
- Arg2 : Nombre d'échantillons pour la condition 1 (entier naturel n1=135)
- Arg3 : Nombre d'échantillons pour la condition 1 (entier naturel n2=15)
- Arg4 : Pourcentage de gènes non DE (nombre réel pi0=0.9)
- Arg5 : Pourcentage des gens up (nombre réel up=0.5)
- Arg6 : muminde1 (Nombre réel compris c entre 1.1 et 1.9) (muminde1=1.4)
- Arg6 : muminde2 (Nombre réel compris c entre 0.2 et 0.9) (muminde2=0.8)

La sortie est sous forme d'une matrice de mesures d'intensités sauvegardée dans un fichier 'txt' dans le dossier « Simulation »

III. Le fichier ***Cross_platform.R*** contient le scripte R qui permet de normaliser les données RNA-seq puis effectuer la fusion. Il prend comme 7 arguments :

- Args: préfixe ('simulation')
- Args2: Méthode de normalisation ("DESeq")
- Args3 : Méthode de standardisation (''Zscor'')
- Args4: Nombre d'échantillons des données *microarrays* pour la condition 1 (entier naturel Mn1=135)
- Args5: Nombre d'échantillons des données *microarrays* pour la condition 2 (entier naturel Mn2=15)

- Args6 : Nombre d'échantillons des données RNA-seq pour la condition 1 (entier naturel Rn1=135)
- args7 : Nombre d'échantillons des données RNA-seq pour la condition 2 (entier naturel Rn2=15)

La sortie est une matrice de fusion sauvegardée dans le dossier « Fusion » dans un fichier 'txt '