Documentation

# Description

Ce travail a été consacré à la simulation des données RNA-seq et microarrays pour appliquer les normalisations et les standardisations à ces données simulées afin de pouvoir les combiner.

Il contient 4 modules:

* **microarray\_simul** : Permet de simuler les données microarrays à partir d’un modèle prédéfini. Les données simulées ont un comportement similaire aux données microarrays produites par la plateforme « Affimetrix ».
* **rnaseq\_simul** : Permet de simuler les données de comptages RNAseq puis les normaliser. Nous proposons trois méthodes de normalisation des données RNAseq : DESeq2, edgeR , et VOOM.
* **normalisation.rna\_seq.r:** Permet de normaliser les données RNA-seq. Nous proposons également trois méthodes de normalisation des données RNAseq : DESeq2, edgeR , et VOOM.
* **naseq\_microarray\_fusion** : Permet de standardiser puis fusionner les deux matrices des données simulées. Nous proposont également trois méthodes de standardisation : Zscore, Zscore Robuste et la QN.

# Usage in R scripts

## “microarray\_simul.r”

Cette fonction permet de simuler des données *microarrays* à partir d’un modèle prédéfini. Les données simulées ont un comportement similaire aux données *microarrays* produites par la plateforme *« Affimetrix* ».

Pour cela, l’utilisateur doit fournir un ensemble des paramètres, ou utiliser ceux disponibles par défaut.

### Description des arguments

* **"-gn" ou "--gene\_number":** Un nombre entier naturel indiquant le nombre de gènes dans les données simulées. La valeur par défaut est : --gene\_number=10,000
* **"-sn1" ou "--samples\_n1":** Un nombre entier naturel indiquant le nombre d’échantillons du phénotype 1 (condition 1). La valeur par défaut est : --samples\_n1=75
* **"-sn2" ou "--samples\_n2":** Un nombre entier naturel indiquant le nombre d’échantillons du phénotype 1 (condition 1). La valeur par défaut est : --samples\_n2=75
* **"-diff "ou "--diff\_genes\_ratio":** Un nombre décimal indiquant le the pourcentage des gènes différentiellement exprimés. Sa valeur par défaut est : --diff\_genes\_ratio=0.1
* **"-up", ou "--up\_ratio":** Un nombre décimal indiquant le pourcentage de gènes surexprimés. Sa valeur par défaut est : --up\_ratio=0.5
* **"-m1" ou "--m1":** Un nombre décimal correspondant à la différence moyenne entre la moyenne totale et la moyenne des gènes différentiellement exprimé avec des valeurs élevées. Sa valeur par défaut est : --m1=1.4
* **"-m2" ou "--m2":** Un nombre décimal correspondant à la différence moyenne entre la moyenne totale et la moyenne des gènes différentiellement exprimé avec des valeurs peu élevées. Sa valeur par défaut est : --m1=0.8
* "-s" ou "--seed": Un entier utilisé pour générer un nombre aléatoire par l'ordinateur dans le but de rendre la simulation reproductible

### Plus de détails

Si l’utilisateur fournit un nombre décimal au lieu d’un nombre entier pour les trois premiers paramètres, la valeur sera arrondie.

La fonction ne sera pas exécutée et retournera un message d’erreur dans les cas suivant :

* Si un des paramètres numérique ne l’est pas
* Si un des paramètres numériques est négatif
* Si le nombre de gènes à simuler est nul
* Si le nombre d’échantillons est nul
* Si les deux paramètres de proportions ne sont pas compris entre 0 et 1

### Sortie

La fonction renvoie une matrice de données avec respectivement le nombre de lignes et de colonnes spécifié par les paramètres d'entrée "--gene\_number et "--samples\_n1 + "--samples\_n2.

Les données sont supposées etre semblables aux données microarrays produites par la plateforme « Affimetrix » log2 intensité.

### Usage in R scripts

**" microarray\_simul.r" --gene\_number 1000 --samples\_n1 20 --samples\_n2 20 --up\_ratio 0.5 --diff\_genes\_ratio 0.1 –m1 1.4 --m2 0.8**

## “ rnaseq\_simul.r”

-“rnaseq\_simul.r" --gene\_number 1000 --samples\_n1 20 --samples\_n2 20 --up\_ratio 0.5 --diff\_genes\_ratio 0.1

### Description des arguments

* **"-gn" ou "--gene\_number**": Un nombre entier naturel indiquant le nombre de gènes dans les données simulées. La valeur par défaut est : --gene\_number=10,000
* **"-sn1" ou "--samples\_n1"** : Un nombre entier naturel indiquant le nombre d’échantillons du phénotype 1 (condition 1). La valeur par défaut est : --samples\_n1=75
* **"-sn2" ou "--samples\_n2"** : Un nombre entier naturel indiquant le nombre d’échantillons du phénotype 1 (condition 1). La valeur par défaut est : --samples\_n2=75
* **"-diff "ou " --diff\_genes\_ratio"** : Un nombre décimal indiquant le the pourcentage des gènes différentiellement exprimés. Sa valeur par défaut est : --diff\_genes\_ratio=0.1
* **"-up", ou "--up\_ratio"** : Un nombre décimal indiquant le pourcentage de gènes surexprimés. Sa valeur par défaut est : --up\_ratio=0.5
* **"-fc" ou "--fc\_file"** : Un fichier ‘txt’ contenant un vecteur FC (fc= "FC.txt"). Si aucun fichier n’est fourni,Le vecteur des fold-change par défaut sera récupéré, soit tel quel, soit avec un échantiollnage pour avoir les bons nombres des génes DE et up
* **"-rseq\_n", "--rnaseq\_norm"** : Un caractère indiquant la méthode de normalisation des données RNA-seq souhaitée. --rnaseq\_norm=’’DESeq2 ‘’ est la valeur par défaut, les alternatives à passer sont ‘’edgeR ‘’ et ‘’VOOM’’.
* **"-s", "--seed"** : Un entier utilisé pour générer un nombre aléatoire par l'ordinateur dans le but de rendre la simulation reproductible

### Plus de détails :

A l'instar de la fonction "microarray\_simul.r", des vérifications seront faites.

### Sortie

La fonction renvoie une matrice de données RNA-seq normalisée avec respectivement le nombre de lignes et de colonnes spécifié par les paramètres d'entrée "--gene\_number et "--samples\_n1 + "--samples\_n2.

## ‘’ normalisation.rna\_seq.r’’

#### Description des arguments

• "-count" ou "--count\_file": matrice de comptage des données RNA-seq

## Deployment and usage in Galaxy workflows

...

## Software dependencies