Rapport d’analyses Bulk-RNA seq

Informations demandeurs

Demandeurs :

- Pauline Garcia (INMG PGNM – Voies de signalisation et muscles striés : julie.sitolle@univ-lyon1.fr
- Fabien LeGrand (INMG PGNM – Voies de signalisation et muscles striés) : fabien.le-grand@cnrs.fr

Projet :

Analyse classique sans filtration des features

sur des échantillons d’origine murine de Bulk RNA-seq

avec recherche d’éléments transposables

L’analyse comprend : le traitement upstream et l’analyse downstream.

- La partie upstream est assurée par le pipeline nf-core/rnaseq.
- La partie downstream est assurée par le pipeline mise en place au sein de MNA2.

Design expérimental

Le design expérimental a été conçu par les demandeurs. Une étape de Data Management des informations a été effectuée sans compromission des comparaisons demandées. Le design utilisé peut être consultable : *results_downstream_nfcorernaseq_3.11.08072024/design.tsv*

Comparaison demandée

Les demandeurs ont établi 1 comparaisons à effectuer. Une étape de Data Management des informations a été réalisée sans compromission des informations.

Les comparaisons peuvent être consultables :
results_downstream_nfcorernaseq_3.11.08072024/contrastlist.tsv

Analyses effectuées

1. Partie upstream pour le niveau Gene

Le détail du pipeline est disponible ici : https://nf-co.re/rnaseq/3.11.2

Les traitements et analyses effectuées au cours du pipeline sont consultables :
results_upstream_nfcorernaseq_3.11.08072024/multiqc_report.html

2. Partie upstream pour le niveau d’élément transposable (TE)

Le détail du pipeline est disponible ici :
https://academic.oup.com/bioinformatics/article/38/13/3361/6591201

3. Partie downstream

Les analyses suivantes ont été réalisées :

- Sélection des features*
- Contrôles-Qualité des échantillons par PCA/UMAP/Clustering non supervisé
- Analyse différentielle
- Analyse d'enrichissement pour le niveau gène
- Analyse fonctionnelle pour le niveau gène

*** Une feature est une annotation génomique (gène, pseudogène, microRNA, etc.)**

Rendu de résultats au niveau Gène

Les résultats sont consultables dans le répertoire *results_downstream_3.11.08072024*

Le répertoire comprend un fichier html réunissant un visuel des sélections, QC et analyses différentielles. Il comprend un dossier propre à la comparaison comprenant l'analyse d'enrichissement et fonctionnelle.

Il y a 3 séries de fichiers csv (séparateur « , ») :

- *comparaison*_gosea.csv :

Il contient le résultat de l'analyse différentielle, les colonnes sont délimitées selon :

- l'identification : les lignes portent le nom d'une feature d'après la référence GRCm39.111. Chaque ligne est associée à un identifiant Ensembl unique (*ensembl_gene_id*) et la localisation chromosomique (uniquement le chromosome), la catégorie RNA auquel appartient la feature
- les métriques de l'analyse différentielle : *baseMean, log2FoldChange, lfcSE, stat, pvalue, padj.*
- Les variables catégorielles de l'analyse différentielle : *LFC, FDR, signed*
  - Une feature est « up-regulated » si et seulement si log2FoldChange ≥ 1.2
  - Une feature est « down-regulated » si et seulement si log2FoldChange ≤ -1.2

Tout autre situation est considérée comme « unchanged »

  - Une feature est « Significant » si et seulement si padj < 0.05 sinon elle est « NoSignificant »

Dans certains cas, où le nombre de features dépasse le millier d'éléments, un filtre plus fort est appliqué, dans ce cas-là, les features retenues pour les analyses suivantes est padj < 0.001. Les valeurs du tableau ne sont pas corrigées dans ce cas de situation.

- Les rangs : *ranks_signed_padj, ranks_signed_pval, ranks_LFC*

- gseaGO_clusterprofiler_results2_*indexcomparaison*.csv

Il contient le résultat de l'analyse d'enrichissement des Goterm assuré par clusterprofiler sur les ontologies provenant de MsigDB.

La colonne *geneSets* correspond à la liste de gènes testée et appartenant à une ontologie donnée (colonne *ONTOLOGY*). Les valeurs associées correspondent au Normalized Enrichment Score (NES), à la confiance dans le test hypergéométrique réalisé (*p.adjust* et *qvalue*). Les autres colonnes servent dans le processus de représentation disponible dans chaque répertoire de comparaison.

- gsea_gprofiler2_results1*indexcomparaison*.csv

Il contient le résultat de l'analyse fonctionnelle assurée par gprofiler2 sur différentes bases de données :

Le pipeline réalise 2 requêtes : une pour les features down-regulated et une autre pour les features up-regulated. Chaque requête est indépendante. Cette information est disponible dans la colonne « query ». Pour le détail des autres colonnes, veuillez lire : https://biit.cs.ut.ee/gprofiler/page/apis (partie g:GOSt)

Les graphiques de chaque comparaison sont les suivants :

- BarPlot_GSEA

A partir des résultats de gprofiler2, on représente l'enrichissement des genesets selon la base de données et associé à chaque requête effectuée.

- NESplot_GOTERM

A partir des résultats de clusterprofiler, on représente l'enrichissement des genesets selon l'ontologie.

- UpsetPlot

On associe l'analyse d'enrichissement de clusterprofiler avec les résultats de l'analyse différentielle pour représenter le chevauchement des gènes entre les genesets et leur caractère différentielle dans la comparaison.

- GeneConceptNetwork

A partir des résultats de clusterprofiler, on a représenté la dérégulation des gènes et leurs associations de parentés avec des Goterm parents.

- Similarity

A partir des résultats de clusterprofiler, on a représenté la similarité sémantique entre chaque Goterm calculé dérégulé. C'est-à-dire la proportion entre chaque goterm à être similaire par rapport aux gènes qui les constituent.

- Unsupervised_analysis

Il s'agit des QC.

- Z-Score_CountNormalized_on_deregulated_genes

On a représenté les features par biotype différentiellement dérégulées au moyen d'un Z-score sur les comptages normalisé et on les associe à leur p-adj et leur log2FoldChange.

Pour plus de détails, veuillez consulter : https://yulab-smu.top/biomedical-knowledge-mining-book/enrichment-overview.html#gsea-algorithm

Les résultats simples sont donc des graphiques, des csv et des html. L'ensemble « ligth » fait : ~18Mb. L'ensemble complet des données traitées (hors fastq.gz) fait : ~40Gb.

Le présent envoie (10 juillet 2024) ne contient pas archives R utilisés.

Les counts/abondances/tpm selon différentes méthodes de comptages peuvent être envoyé à la demande.

Des fichiers de counts brutes et sélectionnée mais non normalisés ont été envoyé.

Rendu de résultats au niveau TE

Les résultats sont consultables dans le répertoire *results_downstream_3.11.08072024/TELevel*

Le répertoire comprend des fichiers pdf et deux fichiers csv (séparateur « , »)

Les fichiers csv contiennent :

- les counts non filtré (*countsTE_unormalized.csv*) et non normalisé
- les counts filtré (*countsTE_selected_unormalized.csv*) et non normalisé

Il contient le résultat de l'analyse différentielle, les colonnes sont délimitées selon :

- l'identification : les lignes portent le nom d'une feature ainsi que son loci d'après la référence GRCm39.111 de RepeatMasker. Chaque ligne est associée à un identifiant élément transposable unique (*te_id*) et la localisation chromosomique (uniquement le chromosome)
- les métriques de l'analyse différentielle : *baseMean, log2FoldChange, lfcSE, stat, pvalue, padj.*
- Les variables catégorielles de l'analyse différentielle : *LFC, FDR, signed*
  - Une feature est « up-regulated » si et seulement si log2FoldChange ≥ 1.2
  - Une feature est « down-regulated » si et seulement si log2FoldChange ≤ -1.2

Tout autre situation est considérée comme « unchanged »

  - Une feature est « Significant » si et seulement si padj < 0.05 sinon elle est « NoSignificant »

Dans certains cas, où le nombre de features dépasse le millier d'éléments, un filtre plus fort est appliqué, dans ce cas-là, les features retenues pour les analyses suivantes est padj < 0.001. Les valeurs du tableau ne sont pas corrigées dans ce cas de situation.

- Z-Score_CountNormalized_on_deregulated_genes

On a représenté les features par famille de TE différentiellement dérégulées au moyen d'un Z-score sur les comptages normalisé et on les associe à leur p-adj et leur log2FoldChange.

Matériels & Méthodes

A la date de rédaction de ce rapport, un matériel & méthode n’est pas disponible.