

# TP Initiation à Galaxy

## Exemple d'analyse différentielle d'expression de gènes



|   |          |
|---|----------|
| <b>Introduction</b>                     | <b>2</b> |
| Récupération des données                | 2        |
| <b>Analyse depuis les fichiers .cel</b> | <b>3</b> |
| Upload des données                      | 3        |
| Contrôle qualité et normalisation       | 3        |
| Analyse différentielle                  | 5        |
| <b>Exercice (mode difficile)</b>        | <b>7</b> |
| <b>Exercice (mode progressif)</b>       | <b>8</b> |

# Introduction

L'objectif de ce TP est de réaliser une analyse différentielle d'expression de gènes, à partir de données de puces à ADN en utilisant la suite d'outil SMAGEXP sous Galaxy. SMAGEXP est une suite d'outil d'analyses et de meta-analyses de données d'expression qui repose sur les packages R limma, metaMA et metaRNAseq.

Ce TP aborde uniquement la partie analyse et laisse de côté la partie méta-analyse.

## Récupération des données

Les données sont accessibles à l'emplacement suivant :


<https://drive.google.com/open?id=0B9hFq8t-2MSITjd4OUJha19xeDQ>

Décompressez le fichier .zip. Il contient 8 fichiers .cel et 1 fichier .cond.

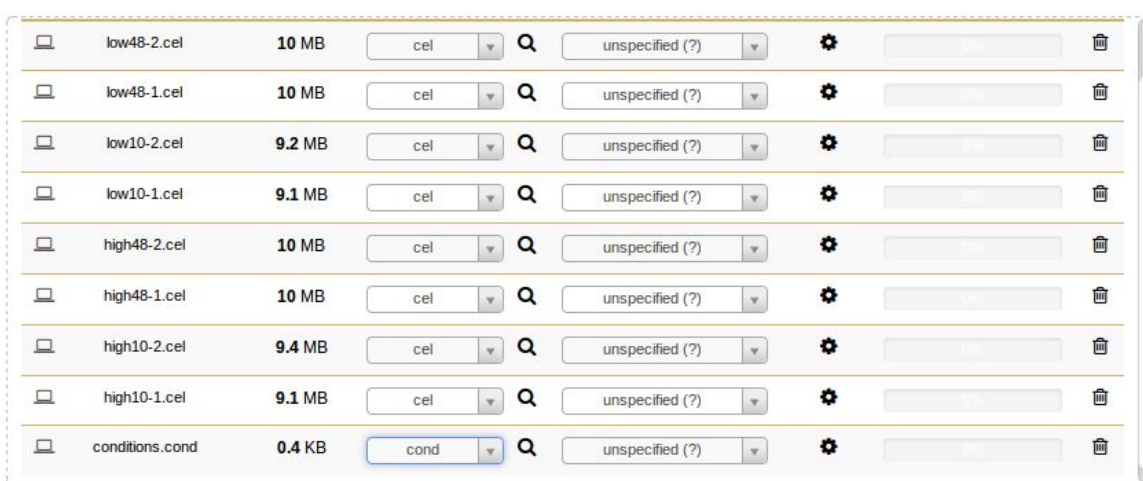
Ces données proviennent d'une expérience liée au cancer du sein réalisée à l'aide de puces à ADN Affymetrix HGU95av2. Les facteurs de cette expérience sont la présence ou non d'oestrogènes (présent ou absent) et la longueur de l'exposition (10h ou 48h).





































# Analyse depuis les fichiers .cel

## Upload des données

- Créer un nouvel historique en cliquant sur l'icône  en haut à droite du panel History
- Uploader les fichiers .cel et .cond sur Galaxy en prenant soin de bien de préciser le type de chaque fichier uploadé (8 fichiers cel ou 1 fichier cond)

Download data directly from web or upload files from your disk



|   |                 |        |      |   |                 |   |                               |   |
|---|-----------------|--------|------|---|-----------------|---|-------------------------------|---|
|    | low48-2.cel     | 10 MB  | cel  |    | unspecified (?) |    | <input type="text" value=""/> |    |
|    | low48-1.cel     | 10 MB  | cel  |    | unspecified (?) |    | <input type="text" value=""/> |    |
|    | low10-2.cel     | 9.2 MB | cel  |    | unspecified (?) |    | <input type="text" value=""/> |    |
|    | low10-1.cel     | 9.1 MB | cel  |    | unspecified (?) |    | <input type="text" value=""/> |    |
|   | high48-2.cel    | 10 MB  | cel  |   | unspecified (?) |   | <input type="text" value=""/> |   |
|  | high48-1.cel    | 10 MB  | cel  |  | unspecified (?) |  | <input type="text" value=""/> |  |
|  | high10-2.cel    | 9.4 MB | cel  |  | unspecified (?) |  | <input type="text" value=""/> |  |
|  | high10-1.cel    | 9.1 MB | cel  |  | unspecified (?) |  | <input type="text" value=""/> |  |
|  | conditions.cond | 0.4 KB | cond |  | unspecified (?) |  | <input type="text" value=""/> |  |

You added 9 file(s) to the queue. Add more files or click 'Start' to proceed.

## Contrôle qualité et normalisation

Objectif : Réaliser une normalisation et s'assurer que les données normalisées sont de bonne qualité.

- Cliquer sur SMAGEXP -> QCnormalization
- Choisir la normalisation RMA

QCnormalization (version 0.1.0)

**.CEL files:**

5: high48-2.cel  
 6: high48-1.cel  
 7: high10-2.cel  
 8: high10-1.cel

.CEL files to be used

**Preprocessing/normalization:**

rma (background correction + quantile normalization + log2)

Execute

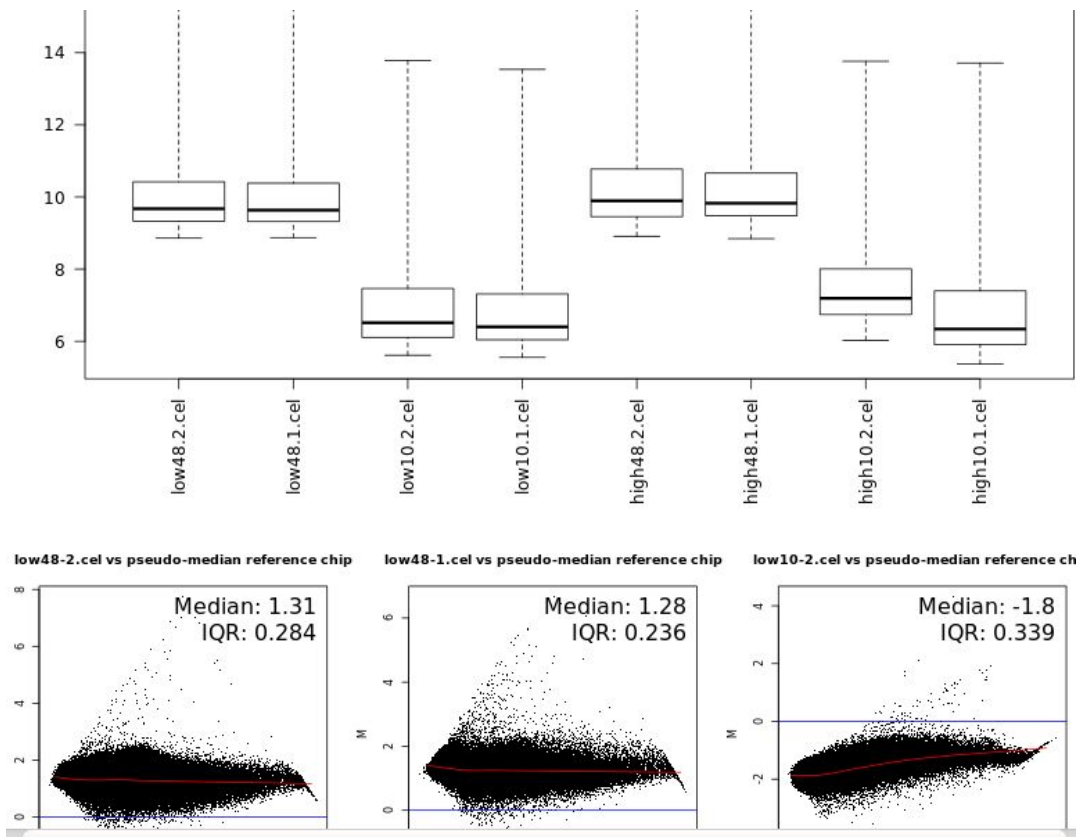
- Cliquer sur Execute

L'outil génère 2 fichiers en sortie :

- Un fichier de Type Rdata (export normalized expressionSet)
- Un rapport html, résumant les figures liées au contrôle qualité et à la normalisation (QC\_result)

En cliquant sur le dataset "QC\_result" on retrouve 3 type de visualisation :

- Les images des puces
- les boxplots
- les MA-plots



On remarquera que la normalisation a bien eu pour effet de normaliser la distribution des données (boxplots devenus quasiment identiques). De plus les MA-plots sont désormais centrés et symétriques autour de 0 et la ligne rouge s'est rapproché de la ligne bleue.

## Analyse différentielle

L'analyse différentielle permet, à partir des données normalisées, de déterminer quels sont les gènes différentiellement exprimés entre 2 conditions. Par exemple dans le jeu de données normalisé précédemment, regardons quels sont les gènes différentiellement exprimés entre la condition low10 et high10.

- Dans le panel "Tools", cliquer sur SMAGEXP -> limma analysis
- Sélectionner le fichier "export normalized expressionSet"
- Sélectionner le fichier .cond et les conditions associées
- Renvoyer les 1000 premiers gènes

**Limma analysis Performs gene expression analysis thanks to limma (Galaxy Version 1.0.0)** Options

**RData**  
17: export normalized expressionSet  
RData to be used

**Conditions**  
1: conditions.cond  
conditions associated with the rData file

**condition 1**  
absent\_10

**condition 2**  
present\_10

**number of top genes**  
1000  
Number of genes to be displayed in result datatable

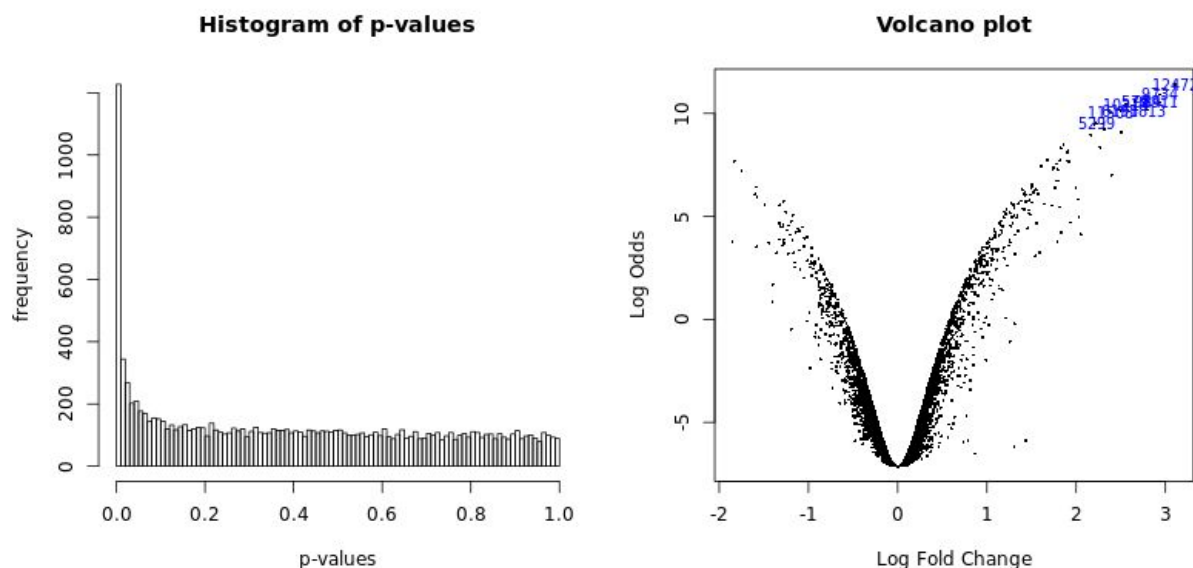
Execute

L'outil renvoie 2 fichiers :

- 1 fichier RData qui peut servir pour une meta-analyse future
- 1 rapport html présentant les résultats de l'analyse.


Le rapport html contient quelques figures : boxplots, histogramme de p-values et volcano plot.

## P-value histogram and Volcano plot













Un tableau présente les valeurs statistiques ainsi que les annotations des 2000 “meilleures” (au sens de la p-value) sondes de la puce.

Les colonnes du tableau sont triables et le champ search permet de réaliser des requêtes.

De plus chaque ligne peut-être déployée à l’aide de l’icône  donnant ainsi accès aux liens vers les annotations du site ncbi et gene ontology.

Enfin, il est possible d’exporter ces données au format csv ou au format Excel.

Copy CSV Excel Search:

|   | ID        | adj_P_Val | P_Value | t     | B      | logFC | Gene_symbol | Gene         |
|---|-----------|-----------|---------|-------|--------|-------|-------------|--------------|
|    | 910_at    | 2.1e-05   | 2.5e-09 | 29.49 | 11.410 | 3.11  | TK1         | thymidine k  |
| Gene Symbol: <a href="#">TK1</a>  |           |           |         |       |        |       |             |              |
| Gene Title: thymidine kinase 1  |           |           |         |       |        |       |             |              |
| GO Function ID: <a href="#">GO:0005524</a> , <a href="#">GO:0042802</a> , <a href="#">GO:0019206</a> , <a href="#">GO:0005515</a> , <a href="#">GO:0004797</a> , <a href="#">GO:0005524</a> |           |           |         |       |        |       |             |              |
|    | 39642_at  | 2.1e-05   | 5.0e-09 | 26.96 | 10.933 | 2.94  | ELOVL2      | ELOVL fatt   |
|    | 1884_s_at | 2.1e-05   | 8.3e-09 | 25.28 | 10.570 | 2.80  | PCNA        | proliferatin |
|    | 1536_at   | 2.1e-05   | 8.3e-09 | 25.27 | 10.569 | 2.66  | CDC6        | cell divisio |
|    | 38827_at  | 2.1e-05   | 8.5e-09 | 25.20 | 10.551 | 2.93  | AGR2        | anterior gr  |
|    | 40117_at  | 2.1e-05   | 9.9e-09 | 24.72 | 10.440 | 2.56  | MCM6        | minichrom    |
|    | 31798_at  | 2.3e-05   | 1.4e-08 | 23.57 | 10.158 | 2.80  | TFF1        | trefoil fact |
|    | 36134_at  | 2.3e-05   | 1.4e-08 | 23.53 | 10.149 | 2.49  | OLFM1       | olfactomed   |
|    | 41400_at  | 2.4e-05   | 1.7e-08 | 23.06 | 10.027 | 2.38  | TK1         | thymidine k  |
|    | 35249_at  | 3.5e-05   | 3.2e-08 | 21.28 | 9.527  | 2.22  | CCNE2       | cyclin E2    |

Show  entries

Showing 1 to 10 of 997 entries Previous  2 3 4 5 ... 100 Next

## Exercice (mode difficile)

- Quels sont les gènes différentiellement exprimés ( $p$ -value ajustée  $< 0.05$  et  $\text{abs}(\log\text{FC}) > 2$ ) communs entre les analyses “absent\_10 vs present\_10” et “absent\_48 vs present\_48”



## Exercice (mode progressif)

- Quels sont les gènes différentiellement exprimés ( $p$ -value ajustée  $< 0.05$  et  $\text{abs}(\log\text{FC}) > 2$ ) communs entre les analyses “absent\_10 vs present\_10” et “absent\_48 vs present\_48”
- 1) Sur le même modèle que précédemment, réaliser l'analyse différentielle entre les conditions absent\_48 et present\_48
- 2) Pour les 2 analyses, sélectionner les gènes différentiellement exprimés ( $p$ -value ajustée  $< 0.05$   $\text{abs}(\log\text{FC}) > 2$ )
- 3) Comparer les 2 listes de gènes ainsi obtenus pour retrouver les gènes communs