M2 Génomique et Post-Génomique GBI

Université d'Evry Val d'Essonne – Année 2014-2015

TP Transcriptome : Analyse de données de puces à ADN

Problématique Biologique : la différenciation des kératinocytes

Les kératinocytes subissent en permanence une évolution morphologique témoignant de leur kératinisation sous-tendant le rôle de barrière protectrice (mécanique et chimique) de l'épiderme. Cette évolution se fait de la profondeur vers la superficie et permet de distinguer sur une coupe d'épiderme quatre couches superposées de la profondeur vers la superficie : la couche germinative (ou basale), la couche à épines (ou épineuse), la couche granuleuse et la couche cornée (compacte, puis desquamante) : **Figure 1.**

Couche cornée

Couche granuleuse

Couche épineuse

Couche basale

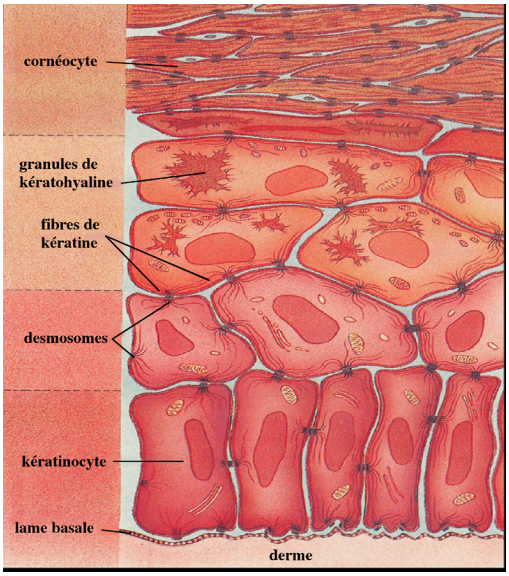

Figure 1 : structure de l'épiderme

La couche germinative ou basale assure par les mitoses de ses cellules le renouvellement de l'épiderme ; ses cellules, cubiques ou prismatiques, contiennent de nombreux grains de mélanine phagocytés qui permettent à l'épiderme d'assurer son rôle de protection de la lumière et qui sous-tendent le rôle de régulation de la pigmentation cutanée qu'ont les kératinocytes.

Dans la couche épineuse, les cellules commencent à s'aplatir, mais le noyau et les organites cytoplasmiques sont intacts, les filaments intermédiaires de kératine groupés en faisceaux denses, les desmosomes normaux.

Dans la couche granuleuse, la cellule est très aplatie, le noyau commence à dégénérer et surtout apparaissent au sein des trousseaux de filaments de kératine de nombreux **grains de kératohyaline.**

Enfin, dans la couche cornée, le kératinocyte (qui prend maintenant le nom de cornéocyte) est complètement aplati, le noyau et les organites cytoplasmiques ont totalement disparu et le cytoplasme est rempli de trousseaux fibrillaires formés à partir des filaments de kératine et des grains de kératohyaline. En superficie de la couche cornée, les cornéocytes, se détachent de l'épiderme (desquamation).

Le kératinocyte passe donc d'un état prolifératif dans la couche basale à un état de différenciation terminale dans la couche cornée avant sa mort cellulaire et sa desquamation. Dans la peau, ce cycle de différenciation dure une vingtaine de jours.
Ce processus de différenciation peut-être reproduit *in vitro*. Notamment, en culture, les kératinocytes se différencient naturellement à partir du moment où la confluence est atteinte, cette technique a été utilisée pour générer les données que nous allons analyser.

L'objectif du TP est d'analyser la modulation de l'expression des gènes au cours de la différenciation *in vitro* de kératinocytes humains. Des expériences d'analyse du transcriptome ont été réalisées en utilisant des puces à ADN sur lesquelles ont été déposées des oligonucléotides longs.
Au total les lames contiennent 26495 spots.
Les cellules ont été cultivées *in vitro* dans des conditions de prolifération (noté P dans le nom de l'échantillon) ou de différenciation (noté D dans le nom de l'échantillon).
Pour chaque état P ou D, une extraction d'ARN a été faites pour 3 individus différents (I1, I2 ou I3). Deux inversions de marquage ont ensuite été réalisées pour chaque échantillon en utilisant une référence commune (le numéro de l'inversion de fluorochrome est noté _1 ou_2 dans le nom de l'échantillon et le fluorochrome de l'ARN test est noté _3 pour Cy3 et _5 pour Cy5).

A. Quantification des images

Dans un premier temps, vous allez réaliser la quantification des images obtenues pour une puce donnée.
Cette analyse est effectuée en salle informatique sur une version de démonstration du logiciel GenePix 6.0 (Axon Instruments).
ARN test (kératinocytes différenciés) marqué en Cy3 : 532 nm # GreenAWS018.tif
ARN de référence marqué en Cy5 : 635 nm # RedAWS018.tif

- Ouvrir le logiciel GenePix 6.0
- Charger les 2 images (RedAWS018.tif et GreenAWS018.tif) en même temps (icône de droite : open /save → open images) en faisant attention à indiquer la longueur d'onde de chaque fluorochrome et la couleur qui servira à représenter les valeurs de ce fluorochrome.
- L'image de la lame entière apparaît
- Zoomer sur la zone des dépôts avec la loupe (icônes de gauche)

1ère possibilité : Création d'une grille manuellement

- Créer une grille (icônes de gauche)
- Une fenêtre « new blocks » apparaît, la remplir avec les paramètres suivants : regarder les paramètres que vous devez spécifier pour créer manuellement une grille.
- Si vous créez une grille de cette manière il faut ensuite la centrer, puis centrer les spots, etc… c'est trop long quand on a des milliers de spots comme sur la lame qu'on utilise ici.

2ème possibilité : Chargement de la liste d'annotation de la puce
On sait exactement ce qui est déposé sur la lame (ordre des dépôts), informations contenues dans le programme de pilotage du robot.
A chaque lot de lames produit on récupère un fichier de sortie qui indique comment le dépôt a été réalisé : fichier .gal.

- Charger le fichier ASW.gal (icône de droite : open /save → Load array list)
- L'image de la lame entière apparaît ainsi que les grilles
- Zoomer la zone spottée avec la loupe (icônes de gauche)
- Centrer l'ensemble des grilles sur les spots à l'aide de la souris, puis cliquer sur F8 : permet de lancer automatiquement toute une série de commandes : Positionning Array, Finding blocks, Finding features)
- Vérifier que tous les spots soient bien alignées dans le cas contraire repositionner la et utiliser la touche F5 pour rechercher les spots (Finding features)
- Se mettre en mode features (spot) en utilisant le bouton droit de la souris ou les icones de gauche
- Replacer les cercles et/ou les agrandir pour que tous les spots soient bien encadrés par un cercle (flèches + bouton Ctrl)
- Eliminer certains spots si nécessaires (présence de poussière / rayure, spot trop petits, …) en utilisant la touche A ou bien en cliquant sur le bouton droit de la souris et en sélectionnant « flag bad », ce qui revient à éliminer le spot. Vous pouvez regarder le « pixel plot » en cliquant droit sur la souris pour regarder la distribution des intensités de chaque pixel correspondant à un spot.
- Extraire les valeurs de chaque spot : analyse (icône de droite)

  Aller sur l'onglet Scatter Plot et regarder la distribution des données F635 median-B635vsF532 median-B532. Vous pouvez si vous le souhaitez, vous mettre en échelle log. Regarder également le plot MvsA généré à partir des données en allant sur l'onglet report et en sélectionnant MvsA plot puis S start.

  Question 1 : *Qu'en pensez-vous?*

-

  Question 2 *: Si vous observiez de nombreux spots localisés tout à droite du plot formant une ligne verticale, comment interpréteriez-vous ce résultat (Sur votre graphique vous devriez observer 2 de ces spots)?*
  Regarder l'onglet résultats et le scatter plot et déterminer s'il y a autant de gènes de chaque coté de la médiane. Qu'est ce que ce résultat indique ?
  Rechercher dans l'onglet résultats les contrôles positifs « ctr1_ACTB » (actine) et « ctr2_GAPDH » et les ratios associés. Vous pouvez également regarder les spots sélectionnés sur l'onglet image ou le scatter plot. Qu'attendiez-vous en théorie pour ces spots ?

- Normaliser les données (Ratio of medians) en utilisant une normalisation par la moyenne sur tous les spots.
-

  Question 3 : *Comparer la distribution des données à celles obtenues avant normalisation sur le Scatter plot ou la distribution MvsA. Quel est le gène le plus surexprimé et le plus réprimé dans les kératinocytes par rapport à la référence. Indiquez leur position sur la distribution des données et justifier votre choix.*

B. Analyse des données

Nous allons analyser la modulation de l'expression des gènes au cours de la différenciation *in vitro* de kératinocytes humains. Des expériences d'analyse du transcriptome ont été réalisées en utilisant des puces à ADN sur lesquelles ont été déposées des oligonucléotides longs.
Au total les lames contiennent 26495 spots.
Les cellules ont été cultivées *in vitro* dans des conditions de prolifération (noté P dans le nom de l'échantillon) ou de différenciation (noté D dans le nom de l'échantillon).
Pour chaque état P ou D, une extraction d'ARN a été faites pour 3 individus différents (I1, I2 ou I3). Deux inversions de marquage ont ensuite été réalisées pour chaque échantillon en utilisant une référence commune (le numéro de l'inversion de fluorochrome est noté _1 ou_2 dans le nom de l'échantillon et le fluorochrome de l'ARN test est noté _3 pour Cy3 et _5 pour Cy5).

1- Design expérimental

Préalablement au démarrage d'un projet de transcriptome, on peut déterminer le nombre de puces à ADN qu'il faut hybrider par échantillon afin de s'assurer d'avoir un taux de résultats faux-positifs relativement faible. C'est ce qu'on appelle le design expérimental.

Sous R, suivez la procédure d'installation suivante

```
source("http://bioconductor.org/biocLite.R")
biocLite("OCplus")
library(OCplus)
```

> ***Question 4 :*** *Déterminer le FDR théorique que l'on aura dans ce projet si on estime que 5 % des gènes sont différentiellement exprimés avec un ratio >2 et que l'on réalise l'analyse différentielle avec un seuil de significativité de 1/1000, en utilisant la fonction 'sample size' du package R OCplus.*
> *Faire la même chose en considérant cette fois que 10% des gènes sont différentiellement exprimés. Qu'en pensez-vous ?*

2- Créer un projet

Copier sur le bureau le dossier contenant l'ensemble des lames que vous allez analyser.
Aller dans TPstuds /Debily/Transcriptome: il s'agit d'un dossier "data_output genepix"

Ouvrir Excel.
Aller dans Fichiers puis Option puis Options avancées. Vérifier que le séparateur de décimales correspond à un point sinon modifiez-le.

1- Ouvrir un des fichiers .gpr et regarder la structure de ce fichier.

1- Récupérer la description des échantillons
Ouvrir le fichier Experimental_descriptor.xls
Faites un copier coller dans un nouveau fichier et sauvegarder le.
Fermer les fichiers excel ouverts.

2- Télécharger les données
Dans la barre outils « Complément », sélectionner "Arraytools" : **ATTENTION NE PAS UPDATER ARRAYTOOLS !!!!**
Dans la barre outils « Complément », sélectionner "Arraytools" > import data > Data Import Wizard

Choisissez DataType « Genepix Dual-Channel Data » File Type « The expression data are in separate files stored in one folder » et allez sélectionner le dossier contenant les fichiers de sorties de Genepix pro (fichiers .gpr).
Cliquer sur OK.
Cliquer sur Next.
Sélectionner le fichier « Experimental_descriptor » que vous avez sauvegardé préalablement. Cliquer sur next.
Enregistrer le projet → Projet1

Attention à bien définir les paramètres dans les 3 fenêtres
1 Spot filters : Sélectionner "Use a common reference design" puis "change current setting"
Was the common reference RNA always labelled Cy3 → No
Column indicating arrays with Cy5→ Dys swap indicator
Cy5 label → 1
Cliquer sur OK
Désélectionner les filtres pré-sélectionnés

Sélectionner Apply Background adjustment

2 Normalization: désélectionner les filtres pré-sélectionnés
*3*Gene filters : désélectionner les filtres pré-sélectionnés

Cliquer sur OK
Proceed to Annotation → sélectionner "Annotate data later through "Utilities" menu, cliquer sur OK.

3- Normalisation

Analyser la répartition des données avant normalisation sur un plot MvsA

Sélectionner Arraytools → graphics → MvsA plot,
column of exper descriptor sheet for array names → Experiment Names
Sélectionner 'used normalized log ratio' & 'Intensity filter'
Cliquer submit
Regarder les images obtenues qui sont dans le dossier plugins/plotMvA de votre projet 1.

Question 5 : *Analyser la répartition des données avant normalisation sur un plot MvsA (fichiers sauvegardés dans le dossier de votre projet dans le sous-dossier plugins)*

Créer un deuxième projet (Projet 2) identique sauf que vous ajoutez une étape de normalisation
/ ou modifier le 1^{ER} en sélectionnant filter and subset the data

1 Spot filters : désélectionner les filtres pré-sélectionnés
Use a common reference design puis change current setting
Was the common reference RNA always labelled Cy3 → No
Column indicating arrays with Cy5→ Dys swap indicator
Cy5 label → 1
Désélectionner les filtres pré-sélectionnés
Sélectionner Apply Background adjustment
2 Normalization
Using lowess smoother
*3*Gene filters : désélectionner les filtres pré-sélectionnés

Enregistrer le projet → Projet 2
Sélectionner Arraytools → Graphics → MvsA plot et cocher 'used normalized log ratio" et 'intensity filter', Regarder les images obtenues qui sont dans le dossier plugins/plotMvsA de votre projet2.

Question 6: *Comparer la répartition des données à celles obtenues avant normalisation sur un plot MvsA. Que pensez-vous de la qualité de la normalisation effectuée? Justifier.*

4- Clustering & Expression différentielle

4-1 Clustering des échantillons
Faites un clustering sur les échantillons, en prenant en compte l'ensemble des gènes présents sur la puce à ADN, de manière à voir comment ils se regroupent en utilisant le projet 2.

Sélectionner Arraytools→ clustering→ samples alone→ Hierarchical clustering

Utiliser comme méthode de calcul des distances « centered » (équivalent à la corrélation de Pearson), « average linkage » et centrer les gènes.

Question 7 : *Analyser les résultats*

4-2 Expression différentielle avec les 3 individus

Sélectionner Class Comparison → Between groups of arrays
Faire une comparaison PvsD comprenant tous les individus avec un seuil de 1/1000 en gardant le modèle de variance par défaut. (Vous pouvez changer le nom du fichier de sortie en cliquant sur option avant de lancer l'analyse).

Question 8: *Combien de gènes différentiellement exprimés obtenez-vous? Que vous apporte le Volcano plot comme information ?*

Recommencez en ajoutant 1000 permutations (cliquer dans option). Copier & coller le tableau obtenu dans excel et supprimer les gènes pour lesquels la « permutation p-value » est >0,001).

> ***Question 9 :*** *Combien de gènes différentiellement exprimés avec une permutation p-value <0,001 obtenez-vous? Que signifie ce résultat par rapport à l'analyse précédente (question 8) ?*

Copiez le tableau obtenu lors de l'analyse statistique sur tous les échantillons avec un seuil de 1/1000 et 1000 permutations dans un nouveau fichier excel et sauvegarder cette liste "liste_gènes_diff"

> ***Question 10 :*** *Quel est le gène le plus surexprimé et le plus réprimé au cours de la différenciation?*
> *Rechercher à quoi correspondent ces gènes? Sont-ils déjà connus comme étant impliqués dans la différenciation?*

Filtrer les données de la liste "liste_gènes_diff" pour ne conserver que les gènes présentant un différentiel d'expression > ±1.3 et une valeur de « parametric p-value » < 1e -07, copier la liste des numéros d'identifiant ("unique_id") des sondes correspondant à ces critères en copiant également l'intitulé de la colonne, coller cette liste dans une autre feuille excel et sauvegarder là au format .txt. sous le nom par exemple "liste_genes_diff_sousgroupe"

Vous devez ensuite déplacer ce fichier.txt dans le dossier Projet2> Genelist > Analysis Results.

4-3 Analyse différentielle avec un seul individu

Faire une comparaison PvsD comprenant uniquement l'individu 3 avec un seuil de 1/1000 en gardant le modèle de variance par défaut et en ajoutant 1000 permutations (cliquer dans option)

> ***Question 11 :*** *Combien de gènes différentiellement exprimés obtenez-vous? Comment expliquez-vous la différence par rapport à l'analyse réalisée sur les 3 individus différents?*

4-4 Clustering des échantillons

Faites à nouveau un clustering sur les échantillons seulement mais en utilisant cette fois uniquement une partie des gènes identifiés comme différentiellement exprimés (Projet2) : au moment de lancer le clustering sélectionner l'option gene subset, user gene list et sélectionner le fichier "liste_genes_diff_sousgroupe" que vous avez obtenu précédemment.

> ***Question 12 :*** *Comparer le clustering obtenu à celui obtenu sur l'ensemble des gènes.*

Faire un clustering sur les échantillons et les gènes
Sélectionner Clustering → Genes (and samples)

> ***Question 13 :*** *Que constatez-vous? Y'a-t-il plus de gènes induits ou réprimés au cours de la différenciation d'après vos données?*

Sélectionner la feuille Cluster view et cliquer sur previous pour visualiser le dendrogramme des gènes.
Couper l'arbre de manière à ne visualiser que 2 clusters. Une nouvelle feuille a été crée avec la liste des gènes dans chaque cluster.

> ***Question 14 :*** *A quoi correspondent ces deux clusters? Combien y-a-t-il précisément de gènes dans chaque cluster?*
> ***NB :*** *ne comptabilisez pas les contrôles actine « ctr1_ACTB »*

Sauvegarder ces listes indépendamment l'une de l'autre: liste cluster 1 et cluster 2 après avoir éliminé les contrôles actine.
Voir avec l'enseignant comment les reformater pour les utiliser dans l'annotation GO → enlever « _position oligo ».

5-Annotation GO

Connectez-vous à TOPPGENE
https://toppgene.cchmc.org/
Sélectionner "TOPPFUN"
Charger la liste de gènes du cluster 1, sélectionner le numéro d'accession correspondant Entry Type « RefSeq»
Soumettez la requête.
Lancer l'analyse en conservant les paramètres par défaut.
Visualiser les résultats en choisissant Functional annotation Chart
Refaites la même chose avec la liste de gène du cluster 2.

> ***Question 15 :*** *Quels sont les 3 processus biologiques GO et voies de signalisation (pathway) les plus enrichis en termes de significativité pour chacun des clusters. Qu'en pensez-vous?*