Microarray Data Analysis

Reports

Coppin Georges

31 août 2015

Introduction

En biologie, les microarray sont des technologies qui ont évoluée à partir de la méthode d'analyse Southern Blot. Les microarray permettent d'analyser le niveau d'expression des gènes d'un échantillon. En effet, les microarrays sont composées de nombreux fragments d'oligonucléotides (ADN ou ARN) de l'échantillon que l'on souhaite analyser. Ces oligonucléotides sont déposés sur un support solide (en verre, en plastique, etc.) et chaque fragment correspond à une sonde (probe) qui est spécifique à un gène d'intéret. On étudie l'expression des gènes en hybridant les sondes avec des oligonucléotides complémentaires issus de précédents échantillons expérimentaux ou d'échantillons cliniques. Ces méthodes d'analyses d'expression génétiques sont beaucoup utilisées par les scientifiques grâce à leur efficacité et leur modulabilité (choix des sondes et de leur localisation sur le substrat formant la puce). Les oligonucléotides déposés sur les plaques microarrays ont généralement une taille d'environ 25 à 60 nucléotides. La taille augmente la spécificité des sondes mais augmente également le prix de production.

Il existe plusieurs façons de détecter l'expression d'un gène suivant le type de micrarray que l'on utilise. Il y a notamment la détection *two-colors*, ou *one-color*. Dans le cas de la détection deux couleurs, on prépare généralement une plaque microarray issue de deux échantillons que l'on souhaite comparer. Les sondes obtenues à partir des deux échantillons sont labellisées via deux fluorochromes (marqueur émettant une fluorescence après excitation lumineuse) différents pour chacun des deux échantillons, généralement Cy3 (vert) et Cy5 (rouge). Après excitation des sondes sur plusieurs longueurs d'onde, on peut alors comparer l'intensité relative des deux marqueurs de la biopuce pour identifier quels gènes sont exprimés. Ces microarrays portent généralement des sondes contrôles permettant de normaliser les intensités mesurées entre échantillons. Ce type de microarray est généralement utilisé pour comparer l'expression relatives entre deux échantillons plutôt que pour mesurer l'expression absolue d'une échantillon.

Pour estimer le niveau absolu d'expression génétique d'un échantillon, on priviligera plutôt une microarray *one-color*. Pour ce genre d'analyse, on utilise qu'une couleur et on mesure les intensités lumineuses relatives des sondes excités après hybridation. Il est possible de comparer ces mesures sur bases de données précédentes normalisées. Une des force de cette méthode réside dans le fait qu'un échantillon aberrant ne peut pas altérer les données brutes d'autres échantillons puisqu'il n'y a qu'un échantillon par plaques, contrairement à la méthode *two-colors*. Un autre avantage avec les microarrays *one-color*, c'est qu'il est plus aisé de comparer les données d'autres échantillons. Un des inconvénient de cette méthode est qu'elle requiert deux fois plus de microarray que les microarrays *two-colors*.

Dans le cadre des travaux pratiques du cours de génomique, protéomique et évolution, il nous est demandé de réaliser une analyse de données microarrays. Les données à analyser sont issues de microarray développées par Affymetrix. Le modèle de microarray sur lequel nous travaillons sont des HGU133 plus 2 (format développé par Affymetrix). Les données que j'ai choisies d'analyser traitent de la différence d'expression génétique de cellules périphériques sanguines sur des patients atteints d'insuffisance rénale chronique (IRC). Les analyses d'expression génétique des cellules périphériques sanguines par microarray ont été effectuées sur trois types de patients : des patients traités avec hémodialyse, des

patients atteints d'IRC, et des patients atteins d'hypertension. Les résultats de ces expériences sont disponibles depuis le 7 juillet 2015 et les chercheurs n'ont pas encore dévoilé d'articles analysant ceux-ci. A l'aide de l'outil R qui permet de réaliser de nombreux tests statistiques, j'essayerai d'analyser ces données et de tirer des interprétations biologiques de mes résultats. A l'issus des analyses, le but est d'identifier les gènes dont l'expression évolue similairement et de les associer au regard des phénotypes observés. Ici, en comparant les individus atteint d'IRC avec les deux autres types de patients.

L'insuffisance rénale chronique se caractérise par une altération irréversible du système de filtration glomérulaire, de la fonction tubulaire et endocrine des reins. Cette maladie est irréversible et touche de plus en plus de monde (vieillissement de la population, diabète menant à l'IRC). Il n'existe pas encore de traitement pour guérir d'une IRC, on ne peut que ralentir la progression de la maladie (wikipedia).

Matériel et Méthode

Les données utilisées ont été téléchargées sur le site `https://www.ebi.ac.uk/arrayexpress/`. Il s'agit des données référencées E-GEOD-70528. L'analyse de ces données est basée sur le code R fourni au TP. Celle-ci se déroule en plusieurs étapes.

1. Chargement et Prétraitement des Données

Les informations brutes obtenues après exposition des microarray à de la lumière ne peuvent pas directement être utilisées (voir igure 1). Elles sont au format .CEL. Il est nécessaire de normaliser les données avant d'analyser celles-ci, afin de calibrer les erreurs systématiques entre plaques microarrays et rendre la comparaison de ces plaques possible. L'idée est donc d'enlever toutes les informations "non biologiques". Plusieurs types de méthode existent pour transformer nos données. Il est souvent utile de tester plusieurs de ces méthodes, car d'un jeux à un autre, une méthode ou l'autre peut s'avèrer meilleure. Ici, nous travaillerons avec GCRMA (normalisation basée sur plusieurs microarray) et MAS5 (normalisation individuelle basée sur les PM/MM). Avec MAS5, une microarray est choisie comme modèle pour les autres de sorte que celles-ci aient la même intensité moyenne. Avec GCRMA, la normalisation s'effectue avec les quantiles, qui imposent une distribution identique. Il est également possible de customiser les opérations effectuées sur les données à l'aide de la fonction espresso.

2. Contrôle Qualité

Pour analyser l'effet de la normalisation des données, on peut porter sur graphique la différence d'intensité d'échantillons en fonction de la moyenne géométrique des intensité. C'est ce qu'on appelle un MA-plot. On peut également utiliser des boxplots.

Cette étape est importante car tous les échantillons ne sont pas toujours bon à anaylser et faussent les résultats. Il faut donc les identifier pour les exclure. Une fois que les données passent le contrôle qualité, celles-ci peuvent être analysées. Il faut néanmoins faire attention à ne pas supprimer trop de données (perte de degrés de libertés).

4. Analyses et Illustration des Données

Sur R, on utilise le package Limma pour identifier les gènes dont l'expression diffère (gènes DE). Limma effectue des tests student ou ANOVA pour les identifier. Plusieurs statistiques sont donc effectuées sur les données pour déterminer les gènes qui seront choisis pour illustrer les données. Pour sélectionner les données, on peut utiliser un volcano plot. Un volcano plot est un type de graphique en nuage de point qui porte la significativité des données (p-value) en fonction du fold-change (niveau relatif d'intensité d'expression). En jouant sur ces deux paramètres, on peut isoler les gènes potentiellement intéressant à étudier. Une fois les gènes d'intérets déterminés, on peut comparer leurs expressions relatives sur un Heatmap, leurs relations à l'aide des dendrogrammes associés aux heatmaps. On associe souvent les heatmaps et les dendrogrammes ensemble a fin de réarranger la matrice en fonction des distances ou des similarités entre les gènes et/ou les échantillons.

Résultats

Sur la figure 1, on peut observer à quoi ressemble les données brutes issus d'une microarray Affymetrix au format .CEL d'un des échantillons. On constate le nombre important de données contenues dans une seule plaque.

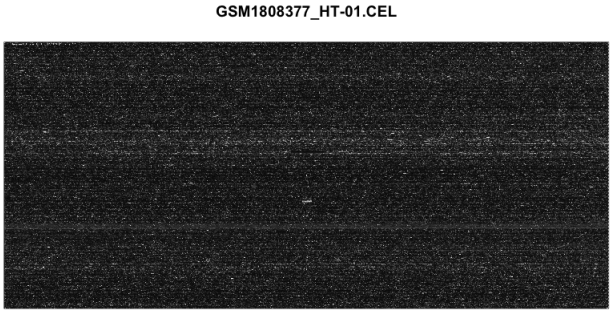

FIGURE 1 – Image .CEL brute de l'échantillon GSM1808377

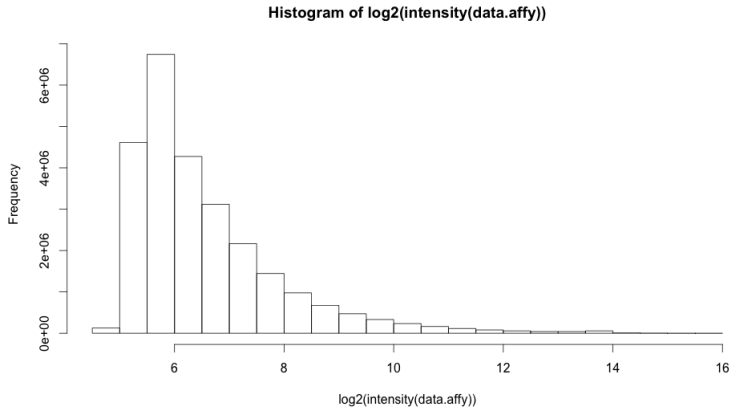

FIGURE 2 – Histogramme des distributions des données brutes des 19 échantillons

Sur la figure 3, on peut observer l'effet de la normalisation sur la distribution des résidus. Ceux-ci semblent mieux distribués.

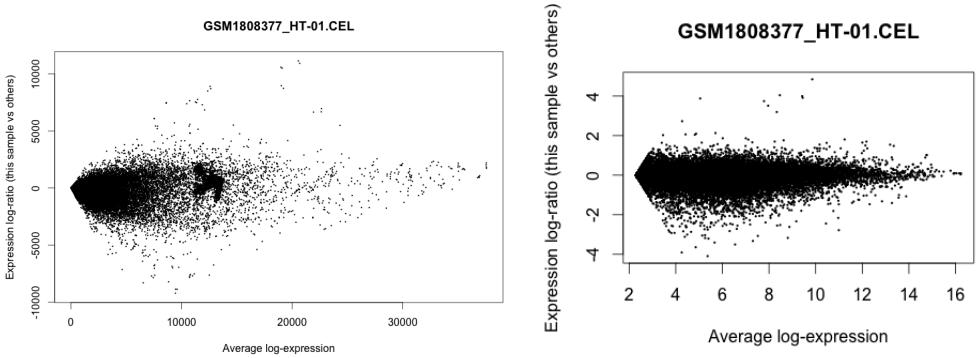

FIGURE 3 – Expression log-ratio de l'échantillon GSM1808377 en fonction du log-expression moyen avant normalisation (gauche) et après normalisation GCRMA (droite)

Sur la figure 4, on peut observer l'effet de la normalisation MAS5 sur la distribution des signaux enregistrés par

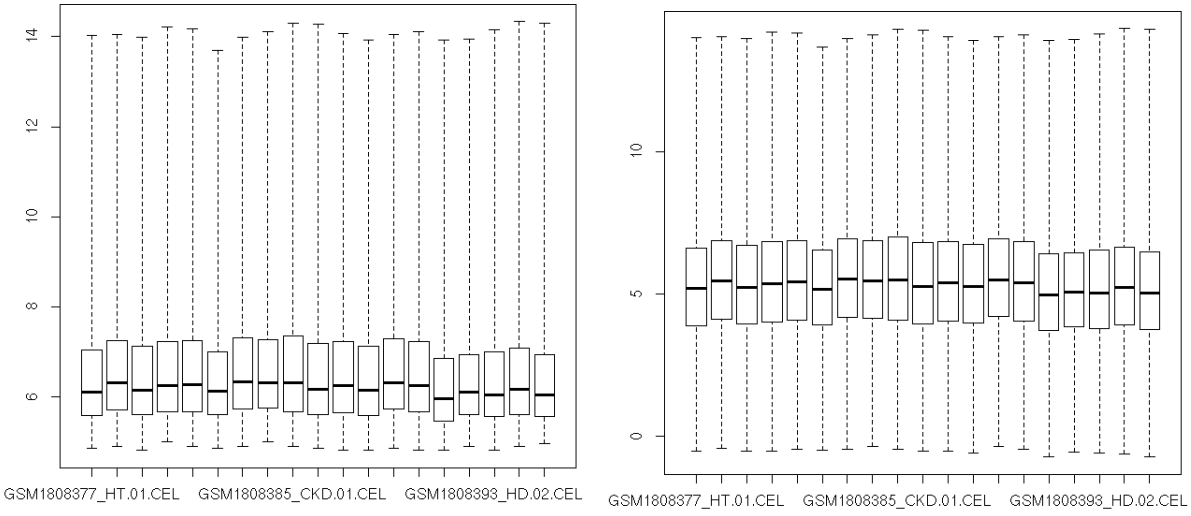

FIGURE 4 – Boxplots des 19 microarrays avec données brutes (gauche) et données traités avec MAS5 (droite)

Sur les graphiques de la figure 5, on peut observer les différences de distribution des volcano plots obtenus avec Limma pour les deux algorithmes utilisés, MAS5 et GCRMA. On remarque que la distribution du nuage de point n'est pas la même pour les données traitées avec l'un ou l'autre algorithme. En rouge, ce sont les données sélectionnées en fonction des seuils imposés par la p-value et le fold-change (seuils relativement bas ici par soucis de clarté (FC > 2, p-value < 0.01).

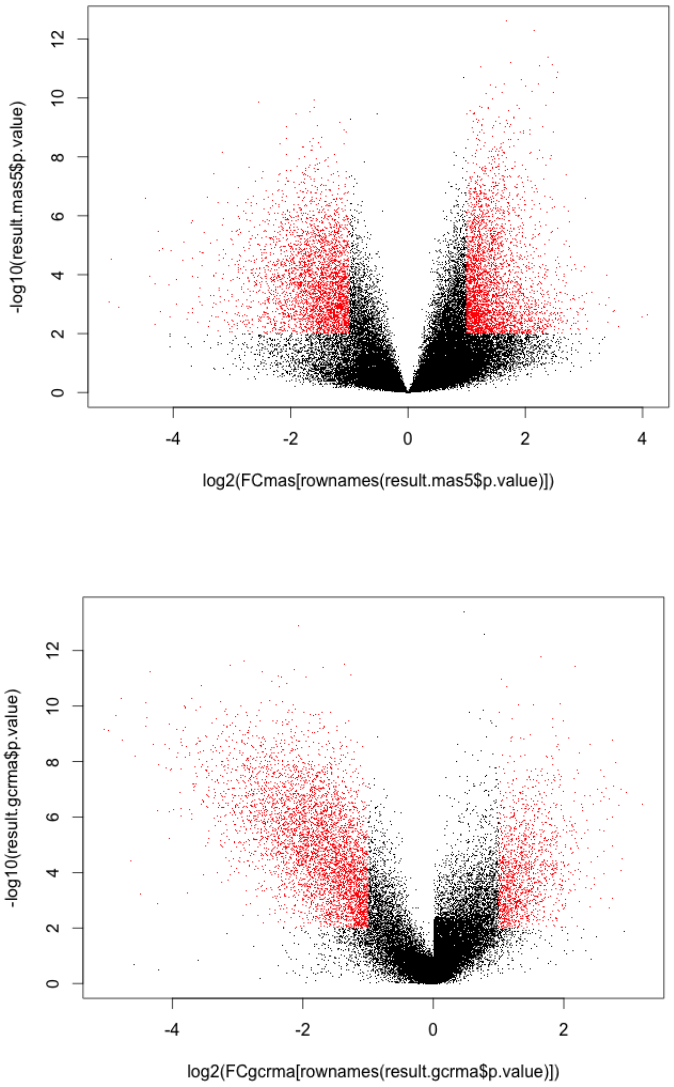

FIGURE 5 – VolcanoPlot des données traitées par MAS5 (gauche) et GCRMA (droite), données sélectionnées en rouge

Le diagramme de Venn de la figure 6 montre les différentes relations entre les gènes des différents échantillons. Cond1 = patients sous hémodialyse, cond2 = patients atteints d'IRC et cond3 = patients atteints d'hyper tension.

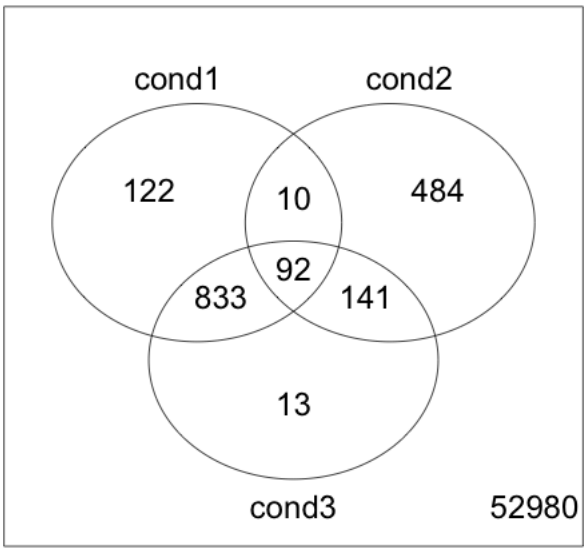

FIGURE 6 – Diagramme de Venn des trois types d'échantillons (cond1, cond2, cond3)

Les figures 7 et 8 sont deux heatmaps obtenus avec MAS5 et GCRMA. Ces heatmaps sont combinés avec des clusters hierarchisés.

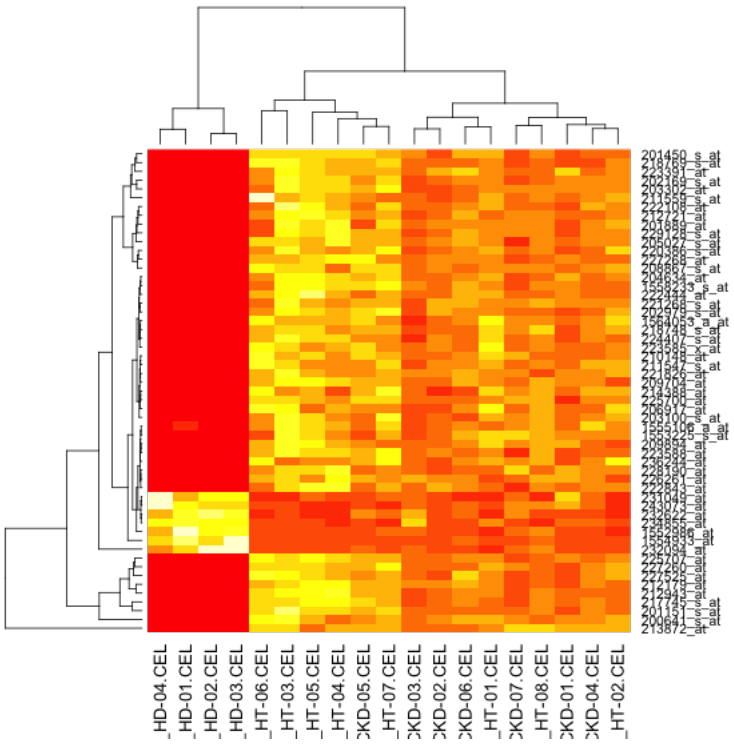

FIGURE 7 – Heatmaps obtenus à l'aide de l'algorithme MAS5

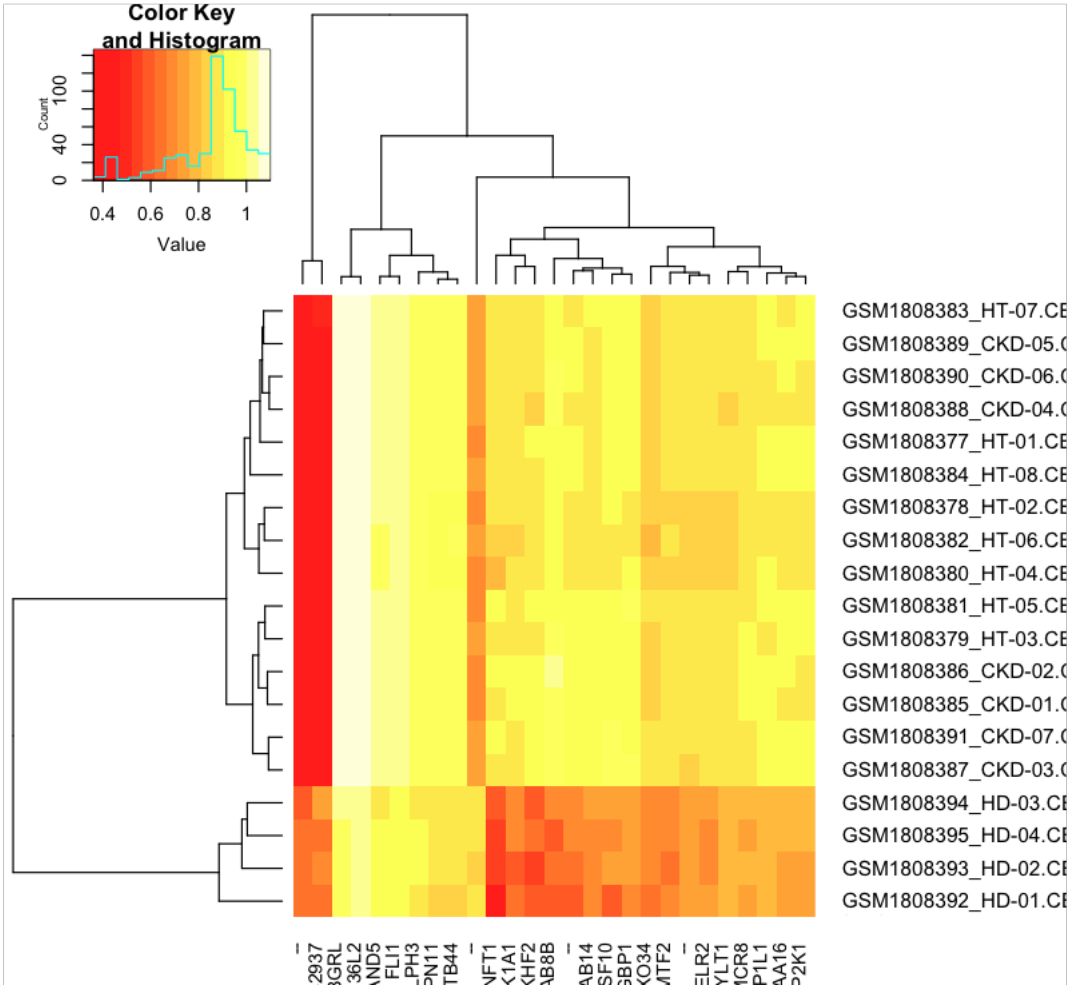

FIGURE 8 – Heatmaps obtenus à l'aide de l'algorithme GCRMA

Le tableau ci-dessous répertorie les gènes qui ont été sélectionnés. La colonne GOBPID fourni un ID permettant de regrouper la littérature respective de chaque gène. Sur ce (lien), des informations relatives sont disponibles pour chacun des gènes à l'aide des ID.

GOBPID	Pvalue	OddsRatio	ExpCount	Count	Size	Term
GO:0035855	0.000	143.676	0	2	14	megakaryocyte development
GO:0016482	0.000	7.927	1	6	813	cytoplasmic transport
GO:0044060	0.001	57.400	0	2	32	regulation of endocrine process
GO:0060986	0.001	52.171	0	2	35	endocrine hormone secretion
GO:0051463	0.001	Inf	0	1	1	negative regulation of cortisol secretion
GO:0060125	0.001	Inf	0	1	1	negative regulation of growth hormone secretion
GO:0060324	0.001	41.968	0	2	43	face development
GO:0008543	0.002	15.091	0	3	183	fibroblast growth factor receptor signaling pathway
GO:0071774	0.002	13.494	0	3	204	response to fibroblast growth factor
GO:0007030	0.002	31.837	0	2	56	Golgi organization
GO:0033277	0.003	814.778	0	1	2	abortive mitotic cell cycle
GO:0045053	0.003	814.778	0	1	2	protein retention in Golgi apparatus
GO:2000847	0.003	814.778	0	1	2	negative regulation of corticosteroid hormone secretion
GO:0051254	0.003	5.280	2	6	1185	positive regulation of RNA metabolic process
GO:1903311	0.004	24.890	0	2	71	regulation of mRNA metabolic process
GO:0061086	0.004	407.361	0	1	3	negative regulation of histone H3-K27 methylation
GO:0003056	0.005	271.556	0	1	4	regulation of vascular smooth muscle contraction
GO:0051461	0.005	271.556	0	1	4	positive regulation of corticotropin secretion
GO:0060352	0.005	271.556	0	1	4	cell adhesion molecule production
GO:0090170	0.005	271.556	0	1	4	regulation of Golgi inheritance
GO:2001275	0.005	271.556	0	1	4	positive regulation of glucose import in response to insulin stimulus
GO:0015031	0.005	4.663	2	6	1327	protein transport
GO:0046887	0.006	19.947	0	2	88	positive regulation of hormone secretion
GO:0051173	0.006	4.531	2	6	1362	positive regulation of nitrogen compound metabolic process
GO:0006892	0.006	19.055	0	2	92	post-Golgi vesicle-mediated transport
GO:0034645	0.007	3.482	5	11	4163	cellular macromolecule biosynthetic process
GO:0000244	0.008	162.911	0	1	6	spliceosomal tri-snRNP complex assembly
GO:0033629	0.008	162.911	0	1	6	negative regulation of cell adhesion mediated by integrin
GO:0036302	0.008	162.911	0	1	6	atrioventricular canal development
GO:0045046	0.008	162.911	0	1	6	protein import into peroxisome membrane
GO:0061087	0.008	162.911	0	1	6	positive regulation of histone H3-K27 methylation
GO:0048308	0.009	135.750	0	1	7	organelle inheritance
GO:0006355	0.009	3.393	4	9	3084	regulation of transcription, DNA-templated
GO:0051651	0.010	15.019	0	2	116	maintenance of location in cell

Discussion

On a vu que MAS5 et GCRMA sont des algorithmes différents mais qui remplissent la même fonction. On observe que la normalisation s'est correctement effectuée dans les deux cas. Une première différence entre ces deux algorithmes est observée lors de la sélection des gènes d'intérets (Differentially Expressed Genes). En effet, on observe une distribution différente des points sur les volcano plots de la figure 5. Pour de mêmes p-value et de mêmes FC, la sélection ne sera pprobablement pas identique. Ceci est tout à fait attendu puisque les intensités reprises dans les données brutes sont utilisées de façons complètement différentes. Le nombre de gènes sélectionnés, determiné à l'aide de diagrammes de Venn, a été fixé à environ 30-40 gènes dans un soucis visuel, notamment pour les heatmaps. Sur ceux-ci, on constate que les gènes sélectionnés par les deux algorithmes sont différents. Il est possible d'associer l'expression de différents groupes de gènes à l'aide de ces graphiques. On remarque néanmoins des similitudes entre ces deux figures. Par exemple, les quatres patients sous traitement d'hémodialyse (HD-01, HD-02, HD-03, HD-04) semblent présenter une expression génétique différentielle des autres patients dans les deux figures. Sur le heatmap obtenu à partir de l'algorithme MAS5 (figure 7), il est plus

aisé d'identifier les différences d'expression génétique entre les patients. On remarque que le dendrogramme associé aux patients de cette figure dégage trois groupes dans lesquels les patients n'ont pas nécéssairement les mêmes symptomes.

Les chercheurs à l'origine de ces recherches n'ont pas publié d'analyses de leurs résultats. Il aurait été intéressant de pouvoir comparer les résultats obtenus lors de la réalisation du projet avec les leurs. Il serait intéressant d'étudier les différents gènes associés dans les heatmaps/dendrogrammes avec les gènes répertoriés dans la table. En effet, des liens entre ces gènes permettraient peut-être de mieux comprendre l'insuffisance rénale chronique en vue de trouver un traitement efficace.

Conclusion

Cette première analyse de données sur biopuce microarray est une bonne introduction aux outils d'analyses statistiques qui existent. Il subsiste néanmoins de nombreuses lacunes dans le traitement et l'interprétation des données. D'autres projets du même type ainsi que d'autres cours de statistiques s'avèrent nécessaires pour comprendre pleinement l'analyse de données volumineuses comme celles obtenues par les biopuces Affymetrix.

Références

http://www3.nd.edu/~steve/Rcourse/Lecture10v1.pdf
http://arrayanalysis.org/main.html
http://www.transcriptome.ens.fr/sgdb/contact/download/200611_LeCrom_pretraitementAnaDiff.pdf
http://lectures.molgen.mpg.de/swp13/affy_diffexp_clustering_exercise-1.pdf
https://www.ebi.ac.uk/arrayexpress/
http://arrayanalysis.org/main.html
https://fr.wikipedia.org/wiki/Insuffisance_rénale_chronique http://wiki.bits.vib.be/index.php/Analyze_your_own_microarray_data_in_R/Bioconductor#Three_groups_of_samples