UNIVERSITÉ D'ANGERS
UFR INFORMATIQUE

RAPPORT DE STAGE
MASTER 1 2016-2017

UMR INSERM 1232
-EQUIPE IMMUNITÉ INNÉE ET IMMUNOTHÉRAPIE

ANALYSE TRANSCRIPTOMIQUE

Présenté par : RASOLONIAINA MARLINO

Maître de stage : Dr. VALÉRIE SEEGERS

Responsable d'équipe : Dr. YVES DELNESTE

Laboratoire d'accueil : INSERM U1232-Equipe 7
Bâtiment IRIS
CHU-4,rue Larrey
49933 ANGERS

12 Avril 2017 — 20 Juin 2017

Table des matières

Table des figures

Liste des tableaux

1 INTRODUCTION

2 PRÉSENTATION DE L'ORGANISME D'ACCUEIL

2.1 Organisme d'accueil : Centre de Recherche en Cancérologie et Immunologie Nantes Angers (CRCINA)

Le CRCINA est une structure de recherche intégrée aux universités de Nantes et d'Angers et labellisée par l'Institut National de la Santé et de la Recherche Médicale (Inserm). Le CRCINA regroupe environ 400 personnes dont environ 150 chercheurs et enseignants/chercheurs [1].

Le CRCINA développe des programmes multidisciplinaires alliant recherche fondamentale et clinique, principalement dans le domaine de l'oncologie. Il est organisé autour de trois départements :

- le département « Immunologie et Immunothérapie » regroupe 8 équipes dont les travaux se focalisent sur l'étude de l'immunité cellulaire humaine antitumorale et antivirale et sous un angle plus appliqué sur la mise en oeuvre de protocoles d'immunothérapie passive ou active
- le département « Oncogénèse et Thérapies Ciblées »
- le département « Ciblage immunospécifique des radionucléides et radiobiologie »

Les équipes de recherche sont installées sur différents sites : CHU de Nantes, CHU d'Angers et Institut de Cancérologie de l'Ouest (site Gauducheau à Nantes et site Papin à Angers). Les activités de recherche sont adossées à des plateformes technologiques et plateaux techniques localisés dans les différents sites ; ces laboratoires sont, pour la plupart, intégrés dans les structures fédératives de recherche des sites Santé des Universités de Nantes et d'Angers (SFR Bonamy à Nantes ; SFR ICAT à Angers).

2.2 Service du rattaché : Équipe 7 de l'U1232 Immunité innée et immunothérapie

Le système immunitaire est organisé autour de deux composantes, le système immunitaire inné et le système immunitaire adaptatif. Le système immunitaire inné est spécialisé dans la détection des signaux de danger, qu'ils soient d'origine endogène (le soi modifié), représenté par les cellules mortes, ou d'origine exogène (le non soi), à savoir les microbes. Toute altération du système immunitaire inné peut avoir des conséquences pathologiques sévères, pouvant aller du déficit immunitaire aux maladies inflammatoires chroniques. Ainsi, les cellules myéloïdes (monocytes, macrophages. . .) jouent un rôle essentiel dans la réponse immunitaire. Elles sont impliquées dans l'élimination des microbes et des cellules mortes ainsi que dans l'initiation et la polarisation des réponses immunitaires adaptatives.

1. `http://www.crcina.org/`

Un des objectifs de l'unité est de comprendre les mécanismes cellulaires et moléculaires impliqués dans la polarisation fonctionnelle des macrophages et identifier des stratégies immunothérapeutiques ciblant les macrophages associés aux tumeurs (cf projet d'équipe au point suivant). Pour explorer la polarisation des macrophages, l'équipe a utilisé la technologie des puces à ADN pour comprendre les mécanismes d'expression géniques mobilisés par les macrophages en conditions physiologiques particulières (Acide Lactique) par analogie avec un milieu tumoral.

3 PRÉSENTATION DU SUJET DE STAGE

3.1 Projet d'équipe où s'inscrit le stage :

Les macrophages sont des cellules d'origine myéloïde présentes dans tous les tissus. Principalement connus pour leur activité de sentinelles immunitaires impliquées dans l'élimination des microbes, les macrophages sont également impliqués dans le maintien de l'homéostasie et le métabolisme tissulaire. Différentes sous-classes de macrophages ont été définies pour décrire cette diversité fonctionnelle.

- Les macrophages de type M1 sont impliqués dans la réponse anti-microbienne et sont caractérisés par un profil pro-inflammatoire
- Les macrophages de type M2 sont impliqués dans la réparation tissulaire et sont caractérisés par un profil anti-inflammatoire.
- Il est actuellement admis que ces deux sous-types de macrophages représentent les extrêmes d'un continuum de cellules [5].

Une des caractéristiques essentielles des macrophages est leur plasticité, c'est-à-dire leur capacité à adopter différents phénotypes en fonction de la nature des signaux reçus localement.

Les macrophages jouent un rôle essentiel dans le développement des tumeurs et leur accumulation est, dans la majorité des tumeurs solides, de mauvais pronostic. De nombreuses études ont montré que le statut métabolique de la tumeur est altéré comparativement aux tissus sains, avec notamment une importante glycolyse. Le laboratoire s'est donc intéressé à l'analyse de l'impact de l'acide lactique, métabolite de la glycolyse qui s'accumule en grande quantité dans les tumeurs, et en particulier le cancer de l'ovaire, sur la polarisation fonctionnelle (M1 versus M2) des macrophages humains. Une étude transcriptomique a donc été réalisée pour analyser l'impact de l'acide lactique sur la différentiation des monocytes en macrophages.

3.2 Objectif et missions du stage :

L'étudiant a pu choisir en arrivant entre 3 sujets concernant la question biologique « Impact de l'acide lactique sur la polarisation et l'expression génique des macrophages » (actuellement dans l'unité par 3 doctorants) avec des puces ADN. Chacun de ces sujets faisait appel à des compétences plus particulières en bioinformatique (informatique/statistiques/biologie) :

— Identification de la structure des données brutes issues des puces ADN et étapes de prétraitement des données pour l'analyse différentielle (sujet bio-informatique / informatique)
— Analyse différentielle des données des puces (sujet bioinformatique/biostatistiques)
— Comment interpréter des résultats de l'analyse différentielle d'une puce ADN (bio-informatique/ biologie)

L'objectif est de permettre à l'étudiant de se familiariser avec l'une des étapes du traitement des données d'expression génique.

3.3 Missions réellement réalisées :

La première partie du projet, donc le premier sujet, a été choisi et réellement réalisée par l'étudiant à partir des données brutes issues de la technologie illumina. Il a identifié la structure des données contenues dans ce fichier, et toutes les étapes de contrôle et de prétraitement des données nécessaires aux analyses statistiques ultérieures.

4 ACQUISITION DES DONNÉES

4.1 Plan expérimental :

Le sujet du stage s'inscrit dans la thématique principale du laboratoire qui porte sur l'adaptation du macrophage aux modifications environnementales et plus précisément : l'impact de la présence d'acide lactique (AL), qui s'accumule au sein des tumeurs, sur l'expression génique des macrophages de type M1 (pro-inflammatoires et initiateurs de la réponse immunitaire) et de type M2 (immunotolérants et qui s'accumulent dans les tissus cancéreux).

Pour répondre à cette question, le laboratoire modélise in vitro la polarisation des macrophages en modifiant les conditions de culture des cellules envisageant ainsi 8 conditions expérimentales. (Graphique) Comme il y avait 3 donneurs différents, 24 prélèvements ont été préparés et analysés sur les puces ADN.

Les questions biologiques à explorer sont les suivantes : Quels gènes sont différentiellement exprimés par les macrophages G (M1) et par les macrophages M (M2) en présence d'acide lactique (AL), en dehors de toute stimulation par LPS ? Avec une stimulation par LPS ? Quels sont les processus biologiques concernés par l'exposition des macrophages à l'acide lactique ? Quelles sont les voies de signalisation concernées ?

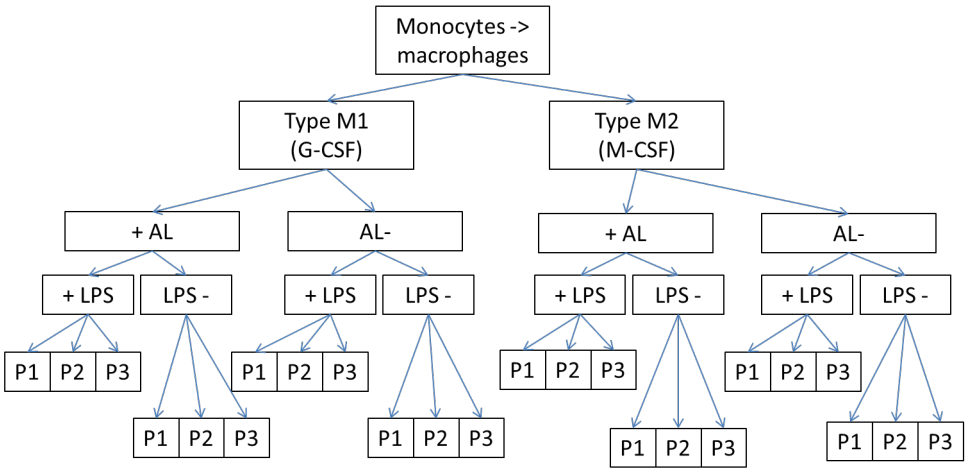

Figure 1 – Plan éxpérimental

L'expérience a été faite avec la puce ADN Illumina HumanHT-12 v4.0 BeadChip 12x1 avec 48210 sondes pour chaque prélèvement. 887 de ses sondes sont classés comme des sondes de contrôles. On se trouve donc avec 47323 individus sur 24 variables. Ce qui nous donne une matrice de données de dimension :

$$\overbrace{\phantom{\begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix}}}^{24 \text{ columns}} \left\{ \begin{pmatrix} a_{11} & \cdots & a_{1m} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_{nm} \end{pmatrix} \right. 47323 \text{ rows}$$

4.2 Les puces à ADN :

Une puce à ADN est constituée d'un support physique (le plus souvent une lame de verre) sur lequel sont déposées des molécules d'ADN correspondant à de petits fragments du génome (jusqu'à 40 000 dépôts différents par puce). On recouvre la puce de la solution contenant la population d'ARN à étudier. Les ARN s'hybrident sur les fragments d'ADN complémentaires. La quantité d'ARN fixée reflète la concentration de cet ARN dans la solution.
Pour des raisons pratiques, on utilise de l'ADN complémentaire plutôt que directement l'acide ribonucléique car l'ADN est plus stable. Les ADNc sont marqués par un nucléotide radioactif ou un fluorochrome. Il est possible d'étudier simultanément plusieurs populations d'ADNc sur une même puce en utilisant des fluorochromes différents. La meilleure façon d'utiliser cette possibilité est de marquer l'ADN génomique avec un fluorochrome, toujours le même. On obtient ainsi une référence stable au cours des années qui permet de mettre toutes les puces à la même échelle, quelle que soit leur origine.

Un scanner mesure l'intensité du signal émis par l'ADNc hybridé au niveau de chaque dépôt. Parmi les valeurs que proposent les logiciels pour cette intensité, la plus fiable est la médiane de l'intensité des pixels car elle est moins sensible aux défauts de l'image (pixels sur-brillants par exemple).
Les puces comportent généralement plusieurs dépôts identiques pour chaque gène. Cela simplifie le travail lorsqu'il faut repérer les aberrations dans la lecture des intensités puisqu'il suffit d'examiner les cas où les valeurs diffèrent beaucoup d'un dépôt à l'autre. Il s'agit le plus souvent d'un défaut physique sur la puce et il est facile d'éliminer la valeur aberrante. Dans le doute, on conserve la médiane des différentes mesures.
Plusieurs types de puces à ADN existent selon le support, la nature des fragments fixés à la surface, le mode de fabrication, la densité, le mode de marquage des cibles et les méthodes d'hybridation. On sait que toutes les technologies des puces ADN se basent sur le principe fondamental de l'hybridation complémentaire des brins d'acide nucléique même si leurs techniques se diffèrent largement entre-elles, par exemple sur la longueur et la type de la sonde utilisée (cDNA arrays, oligonucleotide array), l'étiquetage et le protocole d'hybridation... La vraie différence entre ces approches résident sur la précision, la spécificité, la sensibilité et la robustesse de chaque plateforme.

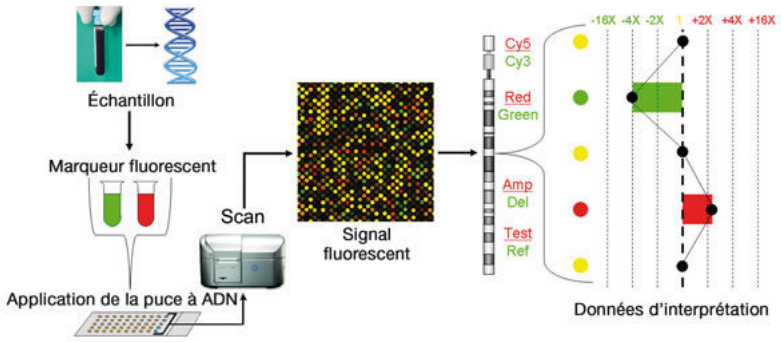

Figure 2 – Principe général d'utilisation des puces ADN

4.3 La technologie Illumina :

Illumina, Inc. est une société américaine qui fabrique et commercialise des systèmes intégrés pour l'analyse de la variation génétique et la fonction biologique, notamment des gammes de produits et services qui servent les marchés du séquençage, génotypage et expression génétique.
Une de ces récentes fabrications est la puce "BeadArray technologie"[2].

2. https://www.illumina.com/technology/beadarray-technology.html

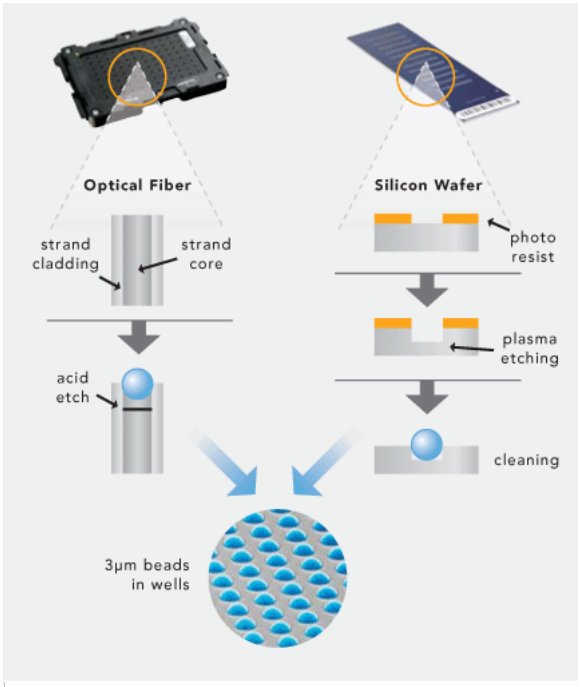

Figure 3 – Illumina BeadArray Technologie

Dans l'analyse des expressions des gènes, Illumina utilise deux approches différentes : l'hybridation directe (Direct Hybridization assay) [3] et le DASL (cDNA-mediated Annealing Selection Extension and Ligation). L'expérience est faite avec la première approche qui consiste à utiliser un simple brin de la séquence d'ADN (pour chaque sonde). Cette séquence monocaténaire s'hybride avec la séquence cible étiquetée dans l'échantillon. La quantité du signal fluorescent produit détermine la quantité de l'ARN cible dans l'échantillon.

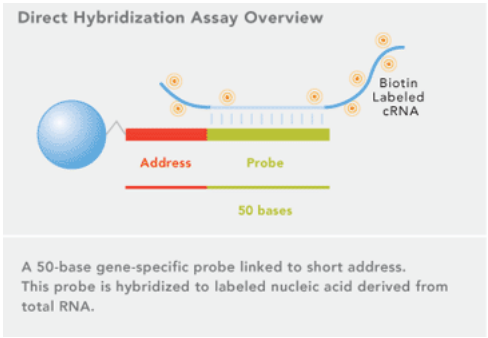

Figure 4 – An Illumina Direct Hybridization probe

3. https://support.illumina.com/array/array_kits/humanht-12_v4_expression_beadchip_kit/training.html

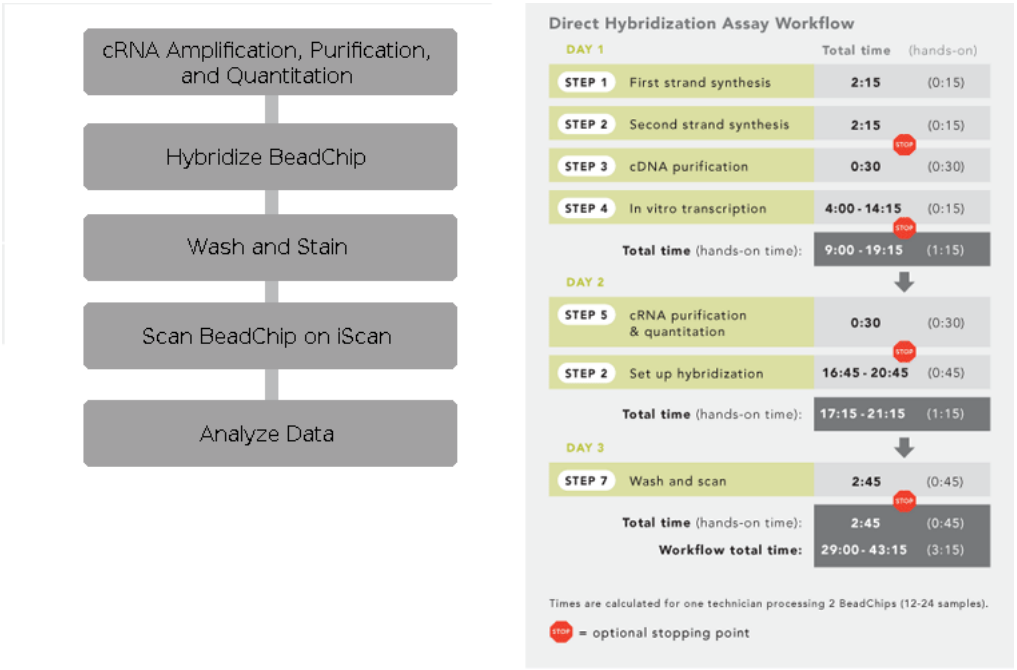

Figure 5 – Direct Hybridization assay workflow

4.4 Les sources de variations et les défis sur l'utilisation des puces ADN :

En Biologie,si on effectue à plusieurs reprises une même expérience, on peut se heurter à des valeurs d'expérience légèrement différentes à chaque exécution. Il est de ce fait très intéressant de voir de plus près les grandes étapes et les effets des processus biologiques qui sont derrières ces sources de quantité de variabilité dans l'étude des expressions génomique avec les puces ADN. De ce point de vue, ces variations sont considérées comme des bruits dans la phase d'analyse des expressions.
Est-ce que la variation d'un gène particulier est due au bruit de fond de la puce ou c'est réellement une différence entre les conditions expérimentales ? C'est là le vrai defi. Si on prend un gène spécifique, combien de quantité de sa valeur représente la mesure de la variance due à la régulation des gènes et due à la quantité de bruit ? Ces sources de variation (Tableau : 1[2, 9]) nous mènent aux problèmes de fiabilité et de reproductibilité dans les mesures des puces ADN. Néanmoins, une grande partie de la variabilité induite par la puce elle-même peut être déterminée à l'aide des techniques de réplications ou d'autres techniques de séparations des bruits (Exemples : conception d'expérience statistique, normalisation des données). Plusieurs efforts ont été menés pour évaluer la fiabilité, la précision et la reproductibilité des puces ADN, inclus des projets comme MAQC(MicroArray Quality Control[3]).

Facteur	*Commentaires*
Préparation des mRNA	Tissus, les kits et les procedures variantes
Transcription	Les variations inhérentes dans la réaction, le type de l'enzymes utilisé
L'étiquetage (Labeling)	Depend du type,des procédures et l'âge de l'étiquette
Amplification (PCR)	Il est difficille de quantifier le rendement de la PCR
Variations géométriques des broches	différentes surfaces et propriétés dues à des erreurs aléatoires de production
Volume de l'échantillon	fluctue stochastiquement même pour la même broche (pin)
Fixation de l'échantillon	La fraction de l'ADNc cible (une gouttelette) qui est chimiquement liée à la surface de la diapositive n'est pas prise en compte
Paramètre d'hybridation	influencé par plusieurs facteurs comme la temperature,le temps,le buffering
Hybridation non-spécifique	un ADNc s'hybride avec une sequence qui n'est pas exactement son complémentaire
Réglages de gain	déplace la répartition des intensités de pixels
Limitation de la plage dynamique	Variabilité de la saturation au bas de gamme ou au haut de gamme
Alignement d'image	Les images d'un même BeadArray à diverses longueurs d'onde correspondant à des canaux différents ne sont pas alignées ; différents pixels sont considérés pour le même emplacement
Placement de la grille	le centre de la sonde (spot) n'est pas bien localisé
Bruit de fond non-spécifique	Élévation erronée de la moyenne de l'intensité du bruit de fond
Forme de la sonde (spot)	L'intensité des sondes irréguliers sont difficille à segmenter en bruit de fond
Segmentation	Des contaminants lumineux peuvent ressembler comme un signal(ex : poussière)
Quantification	la moyenne des pixels, la médiane,...

Table 1 – Sources typiques de fluctuations dans une expérience avec les puces à ADN

5 DESCRIPTION DES DONNÉES

5.1 Décryptage et lecture :

Après avoir scanné la puce, le scanner iScan d'Illumina exporte et produit des fichiers de sortie (Tableau : 2) . Les fichiers qui contiennent les intensités de chaque sonde (.idat) sont encryptés et d'autres fichiers sont fournis à titre indicatif et de mesure pour l'analyse.

File	*Description*
(Serial Number).txt	un fichier qui stocke la positions et l'identité de chaque sonde, qui contient quelques informations sur les paramètres du scanner
Metrics.txt	un pour chaque BeadChip et contient des informations récapitulatives sur l'intensité des signals, la quantité de saturation, la mise au point et l'enregistrement sur l'image (s) de chaque section
Effective.cfg	fichier de configuration des paramètres du scanner
(Serial Number).sdf	fichier de description des échantillons d'Illumina utilisé pour déterminer les propriétes (positions) physique d'une section et savoir les sections liées sur chaque échantillons
*.idat	contiennent la moyenne des intensités du signal de chaque sonde

Table 2 – Description des fichiers de sortie

Pour la lecture des fichiers .idat, l'utilisation d'un fichier manifeste qui contient l'ensemble de tout les informations nécessaires concernant la puce est indispensable pour le décryptage : le nom et l'identifiant des gènes (Probe_id, Array_Address_Id, Symbol, Barcode), le statut d'une sonde (regular, negative, biotin,...) Dans la suite logique des choses, Illumina fourni un logiciel payant (GenomeStudio Software [4]) qui aide sur le traitement et l'analyse des puces ADN (Genotyping Module, Gene Expression Module, Methylation Module).

5.2 Données brutes via Bioconductor :

Bioconductor est un projet de développement et un ensemble de package (1380 packages en 2017) gratuit et open source dans l'analyse et la compréhension des données génomiques basé principalement en langage de programmation statistique R.
Limma [5] est un des packages (de choix) dans Bioconductor pour l'analyse des expressions génomique des puces ADN. La function read.idat() de package Limma permet de lire les fichiers idat d'Illumina BeadArray en fournissant en paramètre le fichier manifeste .bgx correspondant à la plateforme d'expression de gène à étudier. En fait, read.idat () améliore la fonction readIDAT() du package Illuminaio[7] en se basant sur les statuts des sondes (régulier, négatif) parce qu'actuellement, il est le seul package R qui est conçu de décrypter et extraire toutes les informations possibles d'un fichier binaire IDAT (encodé en base64) de plateforme BeadArrayd'Illumina.
Après la lecture, on obtient un objet de type ElistRaw de limma (Tableau : 3) qui contient les informations qu'on a besoin. Il est pratique par la suite de mettre en

4. https://support.illumina.com/content/dam/illumina-support/documents/documentation/software_documentation/genomestudio/genomestudio-2011-1/genomestudio-gx-module-v1-0-user-guide-11319121-a.pdf

5. https://bioconductor.org/packages/release/bioc/vignettes/limma/inst/doc/usersguide.pdf

correspondance les informations liées aux échantillons avec l'objet nouvellement crée, par exemple changer les noms de la colonne de la matrice des intensités.

E	matrice des intensité brutes
other$NumBeads	matrice de mêmes dimensions que E donnant les nombre de la sonde (bead) utilisées pour chaque valeur d'intensité.
other$STDEV	matrice de mêmes dimensions que E donnant un écart type ou une erreur standard pour chaque valeur d'intensité.
genes	un data.frame des annotations des sondes qui contient des informations extraites du fichier manifeste relatif au type de puce utilisé : Probe_Id, Array_Address_Id, Status

Table 3 – Contenu de l'objet EListRaw retourné par la fonction read.idat() de limma

5.3 Contrôle et qualité :

Par approximation, on peut considérer qu'un signal émis par la puce soit :
- la vraie intensité produit par le gène cible
- un signal d'une hybridation non-spécifique
- un bruit de fond non-spécifique

La puce d'Illumina introduit alors des sondes appelées sonde de contrôle pour pouvoir mesurer et quantifier la qualité des données obtenues. Avec ces sondes de contrôles, on peut quantifier les bruits et la qualité du signal, vérifier la qualité de la mesure d'expression de l'ensemble des sondes de la puce. Une valeur anormale produit par un seul BeadArray peut compromettre le résultat d'une analyse sur l'ensemble des données. On ne peut pas donc être assuré d'avoir un bon résultat en phase d'analyse si la qualité des données obtenues n'est pas acceptable (Tableau : 5) [6] [7].

6. https://www.bioconductor.org/packages/release/data/experiment/vignettes/BeadArrayUseCases/inst/doc/BeadArrayUseCases.pdf
7. http://dnatech.genomecenter.ucdavis.edu/wp-content/uploads/2013/11/technote_gene_expression_data_quality_control.pdf

Contrôle de spécimen biologique	ces sont des gènes appelés 'housekeeping genes' qui doivent être exprimés dans tous les échantillons
Contrôle de l'étiquetage des échantillons (Labeling)	des ARN spécifiques(lysA,pheA,thrB,trpF) sont introduits dans les échantillons juste avant la transcription inverse (cDNA) et l'étiquetage. Des faibles signaux provenant de ses sondes indiquent des éventuelles problèmes lors de la réaction
Contrôle de l'hybridation	— Cy3-labeled hyb : ce contrôle se compose de 6 sondes d'oligonucléotides marqué par le fluorochrome Cy3 avec trois concentrations (low, medium, high) et doit produire des signaux progressivement croissants. — Low-stringency hyb : ce contrôle de stringence d'hybridation se compose de 8 sondes (medium, high) avec exception que chaque sonde contient deux bases mésappariés (Perfect Match & Mismatch) — High-stringency hyb
Contrôle de génération des signaux	des ARN sont marqués par de la biotin. On attend un signal d'hybridation positif provenant de ces sondes
Contrôle de sonde négative	des centaines de sondes de séquences aléatoires sans cibles dans le génome sont intégrées dans la puce reflétant les signaux de bruit de fond du système d'imagerie,d'hybridation croisée et autres.On s'attend à des faibles signaux provenant de ces sondes.

Table 4 – Liste des contrôles des données d'Illumina BeadArray

Contrôle	*Valeurs attendues*
Hybridization Controls*	High > Medium > Low
Low Stringency*	PM > MM2
Biotin and High Stringency*	valeurs élevées
Negative Controls (Background and Noise)	valeurs faibles
Gene Intensity (Housekeeping and All Genes)	Plus élevée que les bruits de fond (Housekeeping > All Genes)
Labeling and Background	Labeling >= Background

Table 5 – Contrôle d'hybridation direct de la technologie BeadArray d'Illumina
Les contrôles marqués par (*) sont fortement recommandé par Illumina.

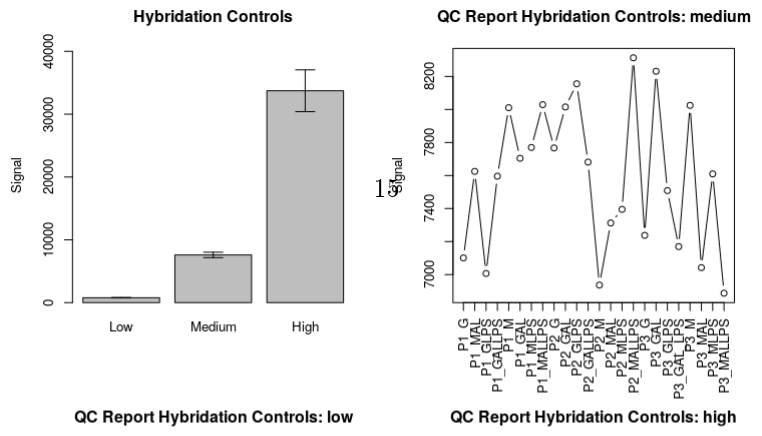

Figure 6 – Contôle d'hybridation

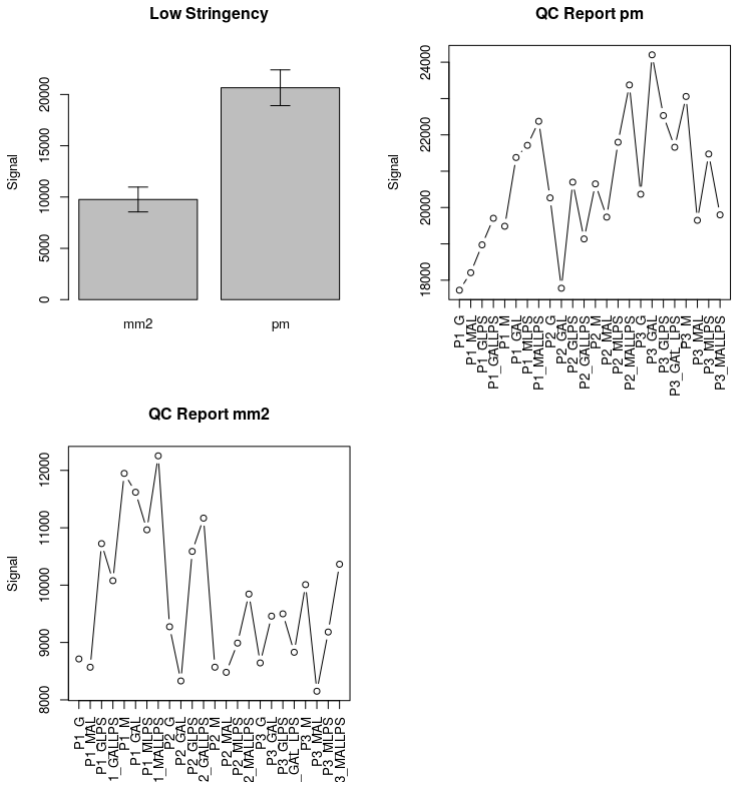

Figure 7 – Low Stringency control

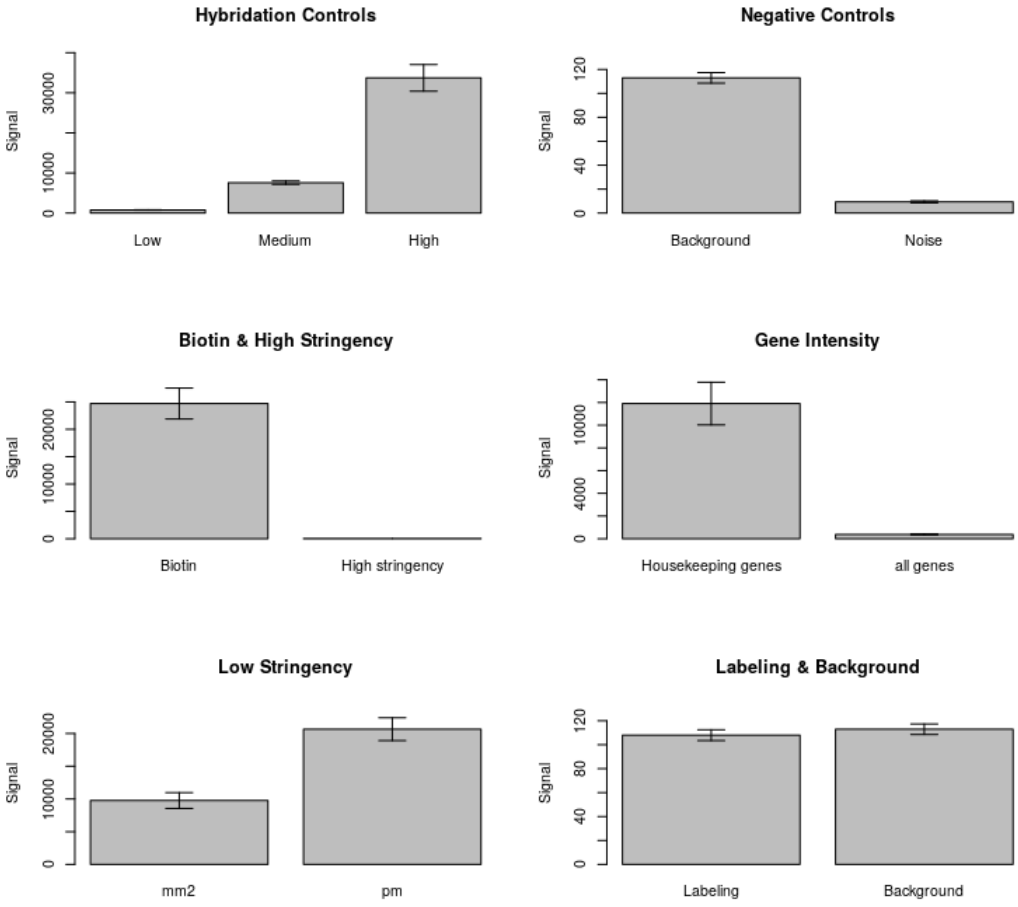

Figure 8 – Résume de l'ensemble de contôle d'hybridation direct de la technologie BeadArray d'Illumina

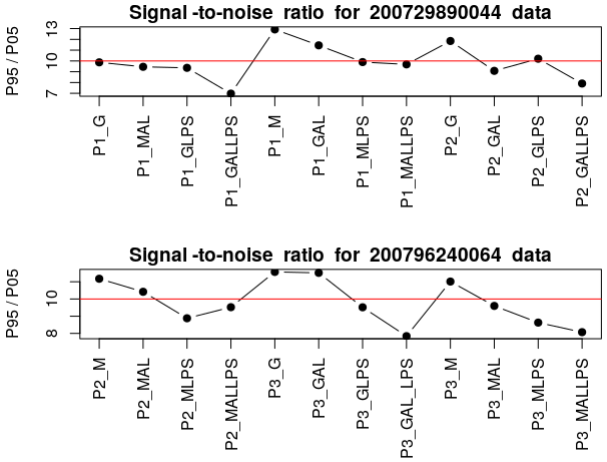

Figure 9 – Rapport signal-bruit sur les deux puces

Un rapport signal/bruit (SNR) peut être calculé en utilisant les mesures fournies par le scanner (dans le fichier metrics.txt) incluant les 95(P95) et 5(P05) quantiles de toutes les intensités de pixels de l'image de chaque section. Ces informations de mesures dépendent du paramètre de scanner et sont tout aussi utile pour l'évaluation de la qualité des données des échantillons ou bien pour évaluer si des échantillons semblent être des valeurs aberrantes. Illumina recommande que le ratio SNR soit supérieur à 10 pour les puces HT-12.

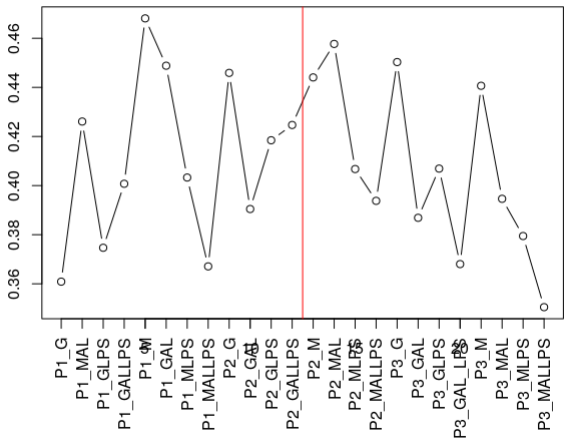

Figure 10 – Estimation de proportion des sondes exprimées sur les deux puces avec la fonction propexpr() de limma

Ces valeurs ne sont pas vraiment des probabilités, elles estiment la proportion globale de sondes sur chaque section de la puce Illumina BeadChip qui correspondent à des gènes exprimés selon la méthode de Shi et al (2010)[4].

$pi1 = (pb - p)/(pb - p1)$ avec $(pi1[pi1 > 1] \leftarrow 1$ et $pi1[pi1 < 0] \leftarrow 0)$

La fonction compare la distribution d'intensité empirique des sondes de contrôle négatif avec celle des sondes régulières. Un modèle de mélange est adapté aux données de chaque échantillon de la puce pour inférer la distribution d'intensité des sondes exprimées et estimer la proportion exprimée.

5.4 Commentaire des résultats :

Un premier préavis sur la qualité de notre donnée est indiqué par le rapport signal/bruit(Figure 9).

```
> summary(ht12snr)#Signal-to-noise ratio for 200729890044 data
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   6.968 9.290 9.775 9.886 10.520 12.920
> summary(ht12snrB)#Signal-to-noise ratio for 200796240064 data
   Min. 1st Qu. Median Mean 3rd Qu. Max.
   7.841 8.817 9.561 9.813 11.050 11.570
```

Les valeurs SNR minimumes respectivement pour la puce 44 et 64 sont 6.968 et 7.871. Ces chiffres sont inférieurs à la valeur préconisée par Illumina mais en moyenne, on voit que la moitie des échantillons sur chaque puce a un SNR

aux alentours de 10. Les valeurs ne sont pas très critiques. En examinant la figure 6, on trouve pas des anomalies particuliers sur l'ensemble des contrôles. On voit que les housekeeping genes produisent des signaux presque 100 fois plus fort que les sondes négatives ($12000 > 120$). L'écart type de l'estimation de la proportion des sondes exprimés sur l'ensemble des échantillons n'est pas énorme ($sd(propexpr(obj)) = 0.03361209$) par rapport à la moyenne($mean(propexpr(obj)) = 0.4087041$), donc la variation de la dite proportion entre les échantillons n'est pas signifiante.

6 PRÉTRAITEMENT DES DONNÉES

6.1 Correction de bruit de fond :

6.2 Transformation :

La distribution du niveau d'expression des gènes est très asymétrique avec un petit nombre de valeurs élevées. C'est une source de problèmes, car de nombreuses méthodes statistiques supposent implicitement une distribution gaussienne. Un simple calcul d'écart type ne satisfait donc pas de donner une interprétation habituelle de la distribution. La transformation logarithmique est la plus utilisée. Il y a plusieurs avantages d'utiliser la transformation logarithmique. Les données transformées sont plus faciles à interpréter (avec la variation du niveau d'expression des gènes plus réaliste) et aussi plus signifiantes de point de vue biologique (les intensités sont généralement comprises entre 0 et 65 535). L'asymétrie est fortement diminuée et la transformation rend la distribution du niveau d'expression des gènes presque normale(distribution gaussienne). Après transformation, les conditions d'application des méthodes et des tests statistiques sont mieux satisfaites.

6.3 Normalisation :

Il n'est pas très judicieux de se lancer tout de suite à la comparaison et l'analyse des expressions des gènes à partir des échantillons multiples, car des sources parasites de variations des expressions peuvent fausser le résultat, exemples :quantité d'ARN différentes dans les échantillons, efficacité de la détection de fluorescence, biais systématiques, artefacts, conditions d'hybridation des échantillons. Parmi ces sources, on cible plus précisément :

- *l'hétérogénéité du bruit de fond* : Si le bruit de fond présente des variations très différentes d'une puce à une autre, ou très structurées spatialement sur une puce, alors on peut être amené à corriger le signal par soustraction du bruit de fond.
- *l'hétérogénéité du signal* : De la même manière, le principe d'invariance d'une très grande majorité des expressions géniques d'une puce à une autre doit se traduire par une répartition comparable des valeurs des signaux entre les différentes puces. Si des différences marquées existent, il est judicieux de ramener les signaux moyens de chaque puce à la même valeur.

Avant de s'approcher des hypothèses favorables pour l'analyse différentielle, la normalisation est nécessaire afin de s'assurer que les données des différentes puces sont exploitables et comparable entre elles, que les différences d'intensité sont en effet dues à l'expression différentielle et non aux artefacts et les biais techniques expérimentaux. Il y a plusieurs méthodes de normalisation souvent classé en deux catégories :

— *méthodes qui utilisent des données de référence (baseline array)* : scaling methods and non-linear methods

— *méthodes qui combinent l'information de toutes les sections de la puce dans un ensemble de données donné (méthode complet)* : Lowess,normalisation par quantile,RMA(Robust Multi-Array Analysis)

6.4 La fonction neqc() du package Limma :

Le package Limma (écrit par Gordon Smyth,Matthew Ritchie et autres) contient pas mal de fonction de normalisation de puce à ADN que ce soit à une ou double couleur. Mais la fonction qui nous intéresse est la fonction neqc() spécialement personnalisée pour les puces Illumina BeadChips. Cette fonction R effectue avant la transformation logarithmique des données une correction de bruit de fond utilisant des sondes de contrôle négatif suivie après par la normalisation par quantile utilisant à la fois les sondes de contrôle positif et négatif. L'algorithme utilise le modèle « normexp »[6] pour la correction de bruit de fond qui consiste à modéliser les intensités de pixels observées en tant que somme de deux variables aléatoires, une normalement distribuée et l'autre répartie exponentiellement, représentant respectivement le bruit et le signal de fond. La moyenne (*mu*) et l'écart-type (*sigma*) du bruit de fond normalement distribuée du modèle normexp sont estimés avec les valeurs des sondes de contrôle négatif et la moyenne (*alpha*) du signal répartie exponentiellement est estimée comme la différence entre la moyenne du signal et la moyenne des sondes de contrôle négatif.

$mu \leftarrow colMeans(xn, na.rm = TRUE)$

$sigma \leftarrow sqrt(rowSums((t(xn) - mu)^2, na.rm = TRUE)/(nrow(xn) - 1))$

$alpha \leftarrow pmax(colMeans(xr, na.rm = TRUE) - mu, 10)$

$mu.sf \leftarrow x - mu - sigma^2/alpha$

$signal \leftarrow mu.sf + sigma^2 * exp(dnorm(0, mean = mu.sf, sd = sigma, log = TRUE) - pnorm(0, mean = mu.sf, sd = sigma, lower.tail = FALSE, log.p = TRUE))$

Après la correction, un petit décalage (offset) est ajouté (par défaut 16) aux intensités corrigées pour améliorer la performance dans la phase d'analyse d'expression différentielle et on applique la normalisation par quantile(*normalizeBetweenArrays*). Le but de la normalisation par quantile est de mettre la distribution, médiane et la moyenne des intensités des sondes de chaque puce sur le même niveau pour toutes les échantillons. Ceci est fait de façon suivante :

1. Donner la matrice des intensités X de dimensions $p * n$ avec les colonnes représentent les échantillons et les lignes représentent les sondes
2. Trier chaque colonne de X par ordre croissant pour construire X_{sort}
3. Calculer la moyenne par ligne de X_{sort} et affecter cette moyenne sur chaque élément dans la ligne pour avoir X_{sm}
4. Construire $X_{normalized}$ en réarrangeant les éléments de chaque colonne de Xsm dans l'ordre de la matrice original X

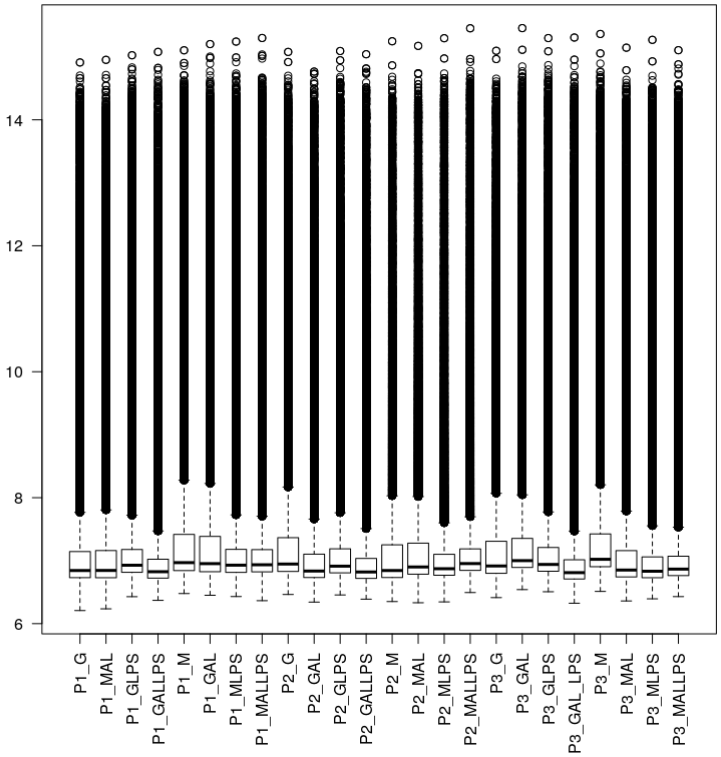

Figure 11 – Box plot des signaux avant normalisation

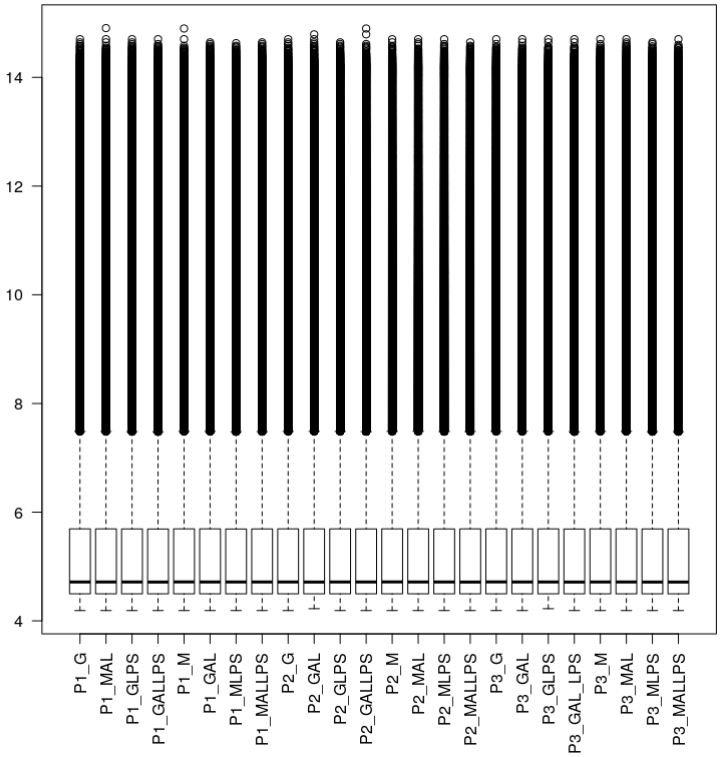

Figure 12 – Box plot des signaux après normalisation

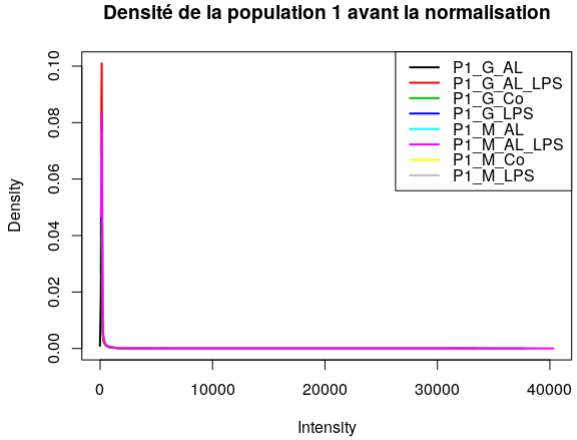

Figure 13 – Densité de la population 1 avant la normalisation

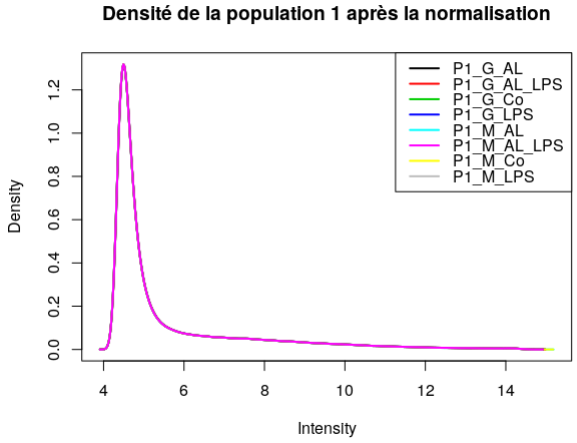

Figure 14 – Densité de la population 1 après la normalisation

7 ANALYSE DES DONNÉES DE TRANSCRIPTOME :

=> En cours

8 INTERPÉTATION :

=> Pas de temps

9 ANNEXE : Scripts R développés et quelques fonctions de Limma

9.1 Lecture des fichiers IDAT :

9.2 Contrôle des données :

9.3 Prétraitement des données :

9.4 Analyse des données :

10 CONCLUSION :

Références

[1] Sorin Drăghici. *Statistics and data analysis for microarrays using R and bioconductor*. CRC Press, 2016.

[2] Lance D Miller, Philip M Long, Limsoon Wong, Sayan Mukherjee, Lisa M McShane, and Edison T Liu. Optimal gene expression analysis by microarrays. *Cancer cell*, 2(5) :353–361, 2002.

[3] Leming Shi, Laura H Reid, Wendell D Jones, Richard Shippy, Janet A Warrington, Shawn C Baker, Patrick J Collins, Francoise De Longueville, Ernest S Kawasaki, Kathleen Y Lee, et al. The microarray quality control (maqc) project shows inter-and intraplatform reproducibility of gene expression measurements. *Nature biotechnology*, 24(9) :1151–1161, 2006.

[4] Wei Shi, Carolyn A de Graaf, Sarah A Kinkel, Ariel H Achtman, Tracey Baldwin, Louis Schofield, Hamish S Scott, Douglas J Hilton, and Gordon K Smyth. Estimating the proportion of microarray probes expressed in an rna sample. *Nucleic acids research*, 38(7) :2168–2176, 2010.

[5] Antonio Sica and Alberto Mantovani. Macrophage plasticity and polarization : in vivo veritas. *The Journal of clinical investigation*, 122(3) :787–795, 2012.

[6] Jeremy D Silver, Matthew E Ritchie, and Gordon K Smyth. Microarray background correction : maximum likelihood estimation for the normal–exponential convolution. *Biostatistics*, page kxn042, 2009.

[7] Mike L Smith, Keith A Baggerly, Henrik Bengtsson, Matthew E Ritchie, and Kasper D Hansen. illuminaio : An open source idat parsing tool for illumina microarrays. *F1000Research*, 2, 2013.

[8] Gordon Smyth. Limma : linear models for microarray data. *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420, 2005.

[9] SE Wildsmith, GE Archer, AJ Winkley, PW Lane, and PJ Bugelski. Research report maximization of signal derived from cdna microarrays. *Biotechniques*, 30(1) :202–208, 2001.