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

Tuteur : M. Le TUTEUR
Chef d'équipe : M. Chef
D'ÉQUIPE

12 Avril 2017 — 20 Juin 2017

Table des matières

1 INTRODUCTION :

2 ACQUISITION DES DONNÉES :

2.1 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 des ADNc plutôt que directement les ARN. 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.

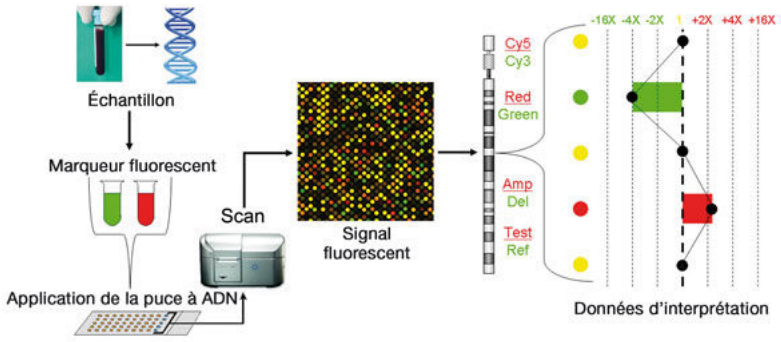

Figure 1 – Principe général d'utilisation des puces ADN

2.2 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, la puce "BeadArray technologie".

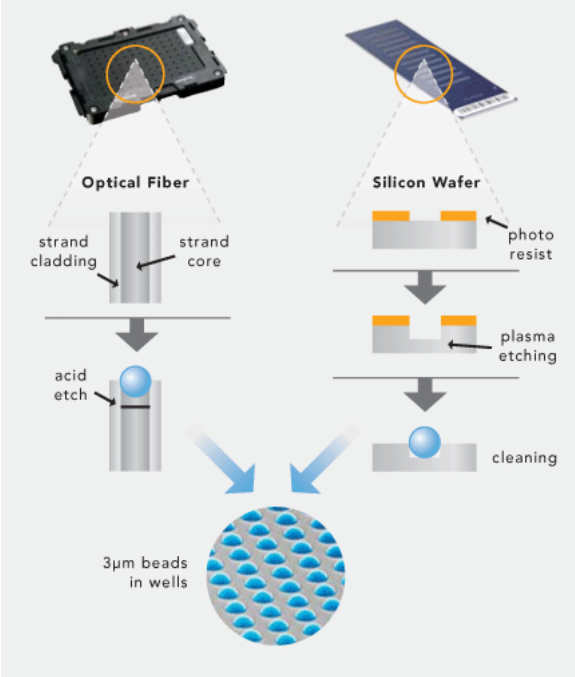

Figure 2 – Illumina BeadArray Technologie

Dans l'analyse des expressions des genes, Illumina utilise deux approches différentes : l'hybridation directe(Direct Hybridization assay) 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 par spot. Cette séquence monocaténaire est censée s'hybrider avec la séquence cible étiquetée dans l'échantillon. La quantité du signal fluorescente produit détermine la quantité de l'ARN cible dans l'échantillon.

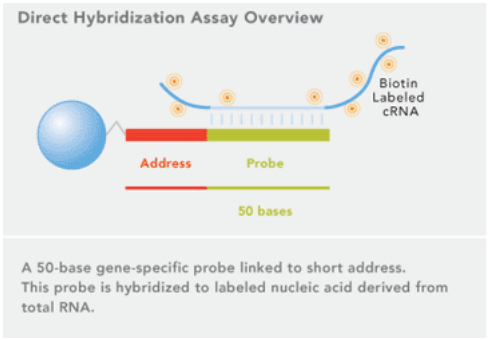

Figure 3 – An Illumina Direct Hybridization probe

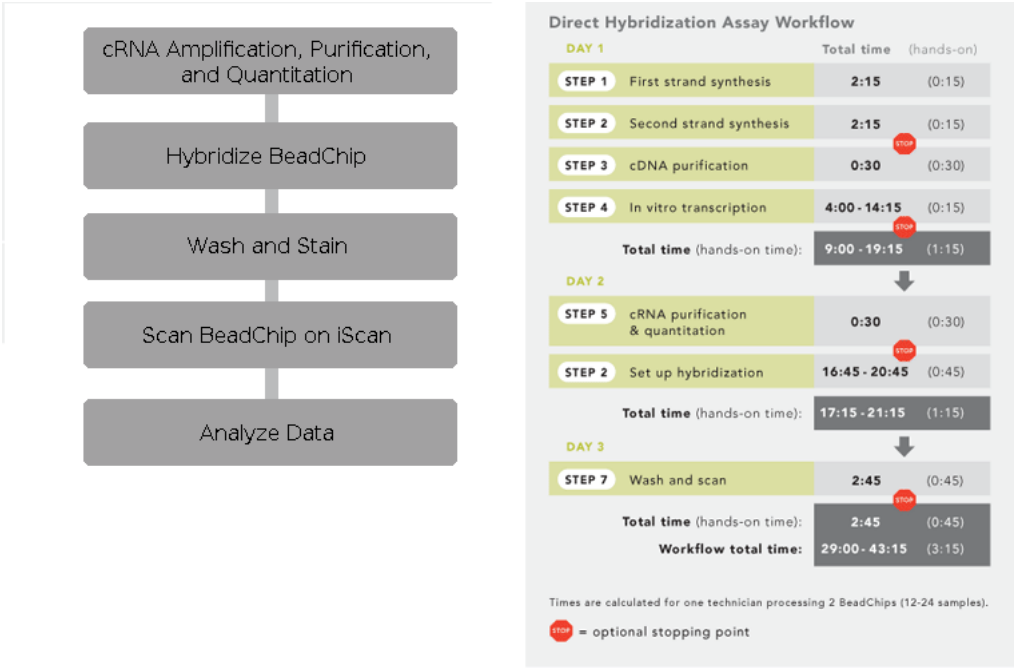

Figure 4 – Direct Hybridization assay workflow

2.3 Plan expérimental :

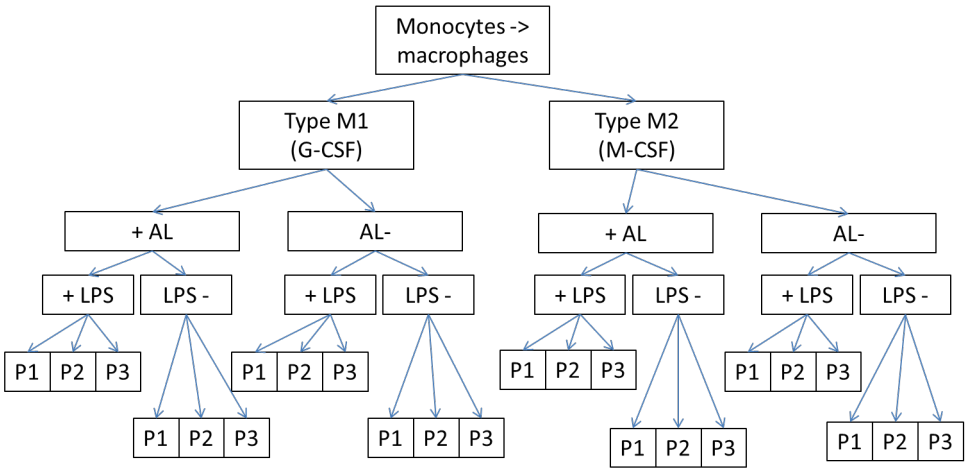

Figure 5 – Plan éxpérimental

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 dimmension :

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

2.4 Les biais possibles :

Il peut exister des biais systématiques dus à d'autres facteurs, tels que l'affinité des séquences ou l'efficacité du marquage.

3 DESCRIPTION DES DONNÉES :

3.1 Decryptage et lecture :

Après avoir scanné la puce, le scanner iScan d'Illumina exporte et produit des fichiers de sortie selon les paramètres définis. Les fichiers qui contiennent les intensités des spots (.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 spots, 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 spots

Table 1 – 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'un spot (regular, negative, biotin,...) Dans la suite logique des choses, Illumina fourni un logiciel payant (GenomeStudio Software) qui aide sur le traitement et l'analyse des puces ADN (Genotyping Module, Gene Expression Module, Methylation Module).

3.2 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 est un des packages 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. On obtient après un objet limma de type EListRaw qui contient les objets suivants :

E	matrice des intensité brutes
other$NumBeads	matrice de mêmes dimensions que E donnant les nombre de spots (bead) utilisées pour chaque valeur d'intensité.
other$STDEV	matrice de mêmes dimensions que E donnant un écart type au niveau des spots 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 2 – Contenu de l'objet EListRaw retourné par la fonction read.idat() de limma

— Barcode
— Section
— ChipType
— RunInfo
— Quants

MeanBinData

TrimmedMeanBinData

DevBinData

MedianBinData

BackgroundBinData

BackgroundDevBinData

CodesBinData

NumBeadsBinData

NumGoodBeadsBinData

IllumicodeBinData

3.3 Contrôle et qualité :

4 PRÉTRAITEMENT DES DONNÉES :

4.1 Transformation :

4.2 Normalisation :

4.3 Filtrage :

5 ANALYSE DES DONNÉES DE TRANSCRIPTOME :

5.1 Gènes différentiellement exprimés :

5.2 Gènes co-exprimés :

6 INTERPÉTATION :

Caractérisation d'un ensemble de gènes

Résumé