

# Analyses statistiques et épigénétiques

Mickaël Canouil  
[mickael.canouil@cnrs.fr](mailto:mickael.canouil@cnrs.fr)

Journée Thématique

15 novembre 2016



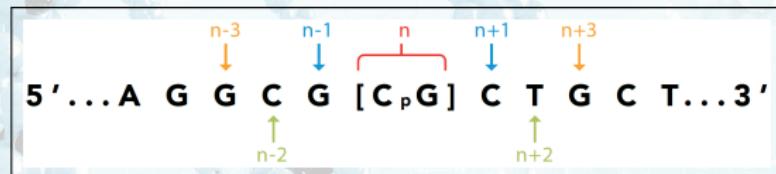
# Introduction I

La méthylation de l'ADN joue un rôle important dans la régulation de l'expression des gènes.

Une hyper- ou hypo-méthylation importante de l'ADN peut avoir un impact sur l'expression des gènes et être impliqués dans la survenue d'une pathologie et/ou ses complications.

# Introduction II

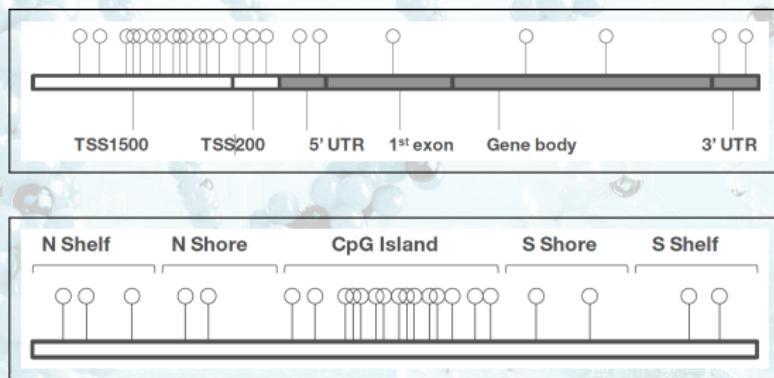
Afin d'étudier les marques de méthylation, notamment celle des sites CpG, des puces existent et fournissent une mesure quantitative (pourcentage) de la méthylation de ces sites.



Ces puces couvrent 99% des gènes identifiés et référencés sur RefSeq.

# Introduction III

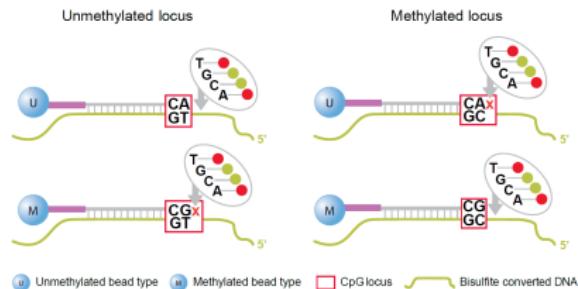
Les sites CpG sont répartis en différentes régions sur l'ADN : géniques et non géniques.



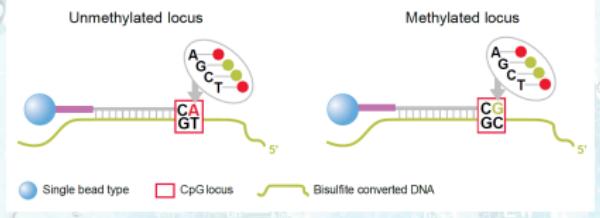
# Méthodes & Analyses I

Au sein de la puce HumanMethylation450, deux technologies différentes sont utilisées pour mesurer le niveau de méthylation d'un site.

## Infinium I



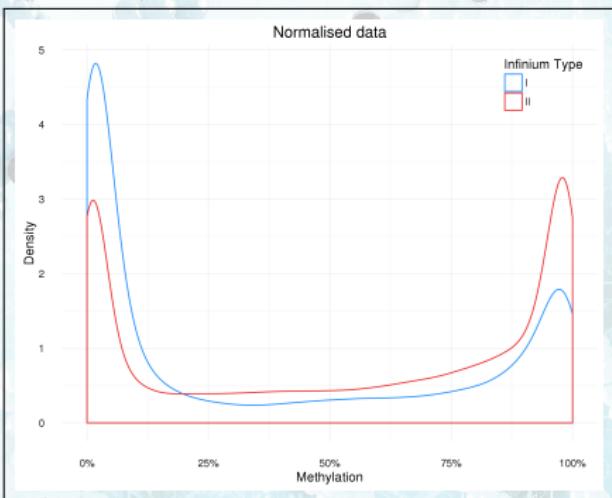
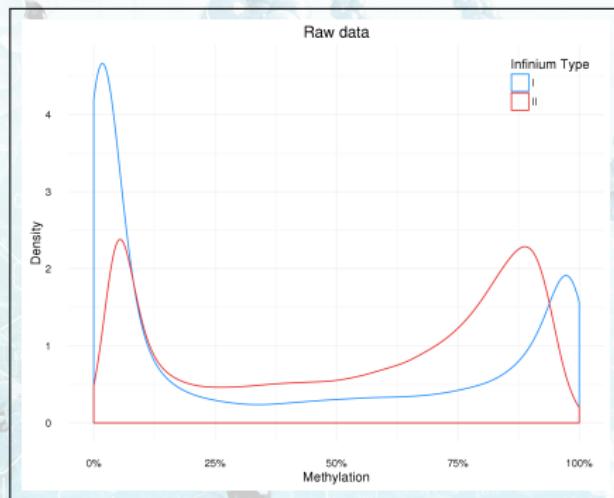
## Infinium II



# Méthodes & Analyses II

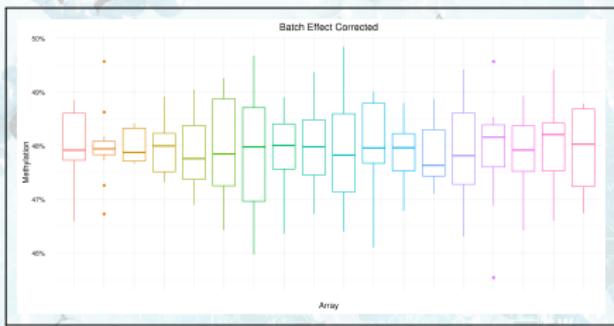
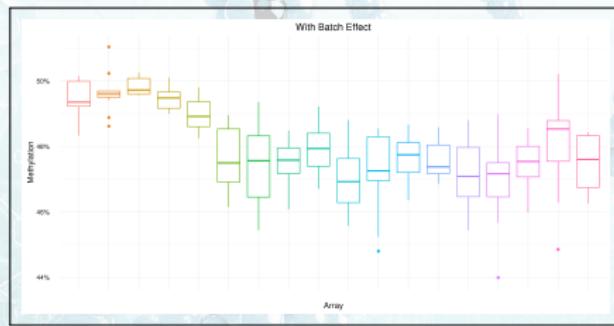
Deux technologies ? Deux mesures différentes...

Une étape de normalisation est donc nécessaire pour prendre en compte ce paramètre.



# Méthodes & Analyses III

Un autre paramètre technique pouvant induire un biais est :  
l'effet plaque, le fameux "Batch Effect".



# Méthodes & Analyses IV

Une fois les données "propres", nous pouvons répondre à la problématique :  
ici un cas/contrôle.

Faisons simple, utilisons une régression linéaire de la forme :

$$Y = \theta_0 + \theta_1 \times X + \gamma \times Z$$

Avec :

$Y$  le niveau de notre site de méthylation compris entre 0 et 1,

$X$  le statut de nos individus (0 : contrôle ; 1 : cas),

$Z$  des covariables comme l'âge, le sexe ou l'IMC,

$\theta_0, \theta_1, \gamma$  les effets estimés.

# Méthodes & Analyses V

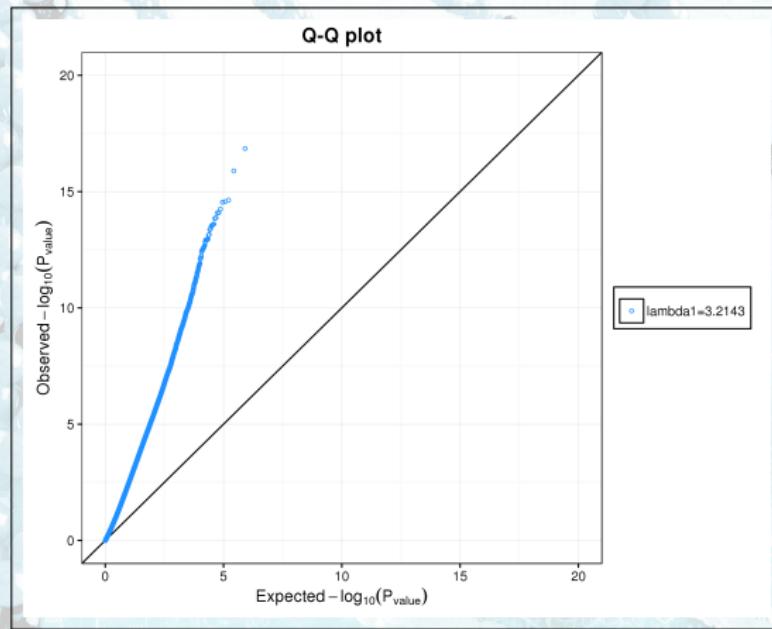
Après avoir effectué pas moins de 500 000 régressions linéaires, nous obtenons un tableau du même nombre de lignes...

Quelques mesures de précautions sont alors de rigueur :

- Regarder le QQplot (quantile-quantile) => Avons-nous quelque-chose de systématique dans les résultats ?
- Corriger les valeurs p ("p-value") pour les tests multiples => Le hasard n'est-il pas le seul responsable ?

# Méthodes & Analyses VI

Le QQ-plot, ça ressemble à quoi ? Que nous dit ce graphique ?



# Méthodes & Analyses VII

Comment pouvons-nous corriger et savoir si le hasard est la cause de tous les maux ?

La méthode la plus répandue et aussi la plus simple à mettre en oeuvre est la correction dite de Bonferroni.

# Méthodes & Analyses VIII

Prenons un exemple, avec un seuil de significativité  $\alpha = 0.05$  :

CpG	Estimée	Valeur p	Significatif
CpG1	0.002315	0.1619	Non
CpG2	0.006276	0.00363	Oui
CpG3	0.004945	0.09727	Non
CpG4	-0.04647	6.212e-06	Oui
CpG5	0.01847	0.04925	Oui

# Méthodes & Analyses IX

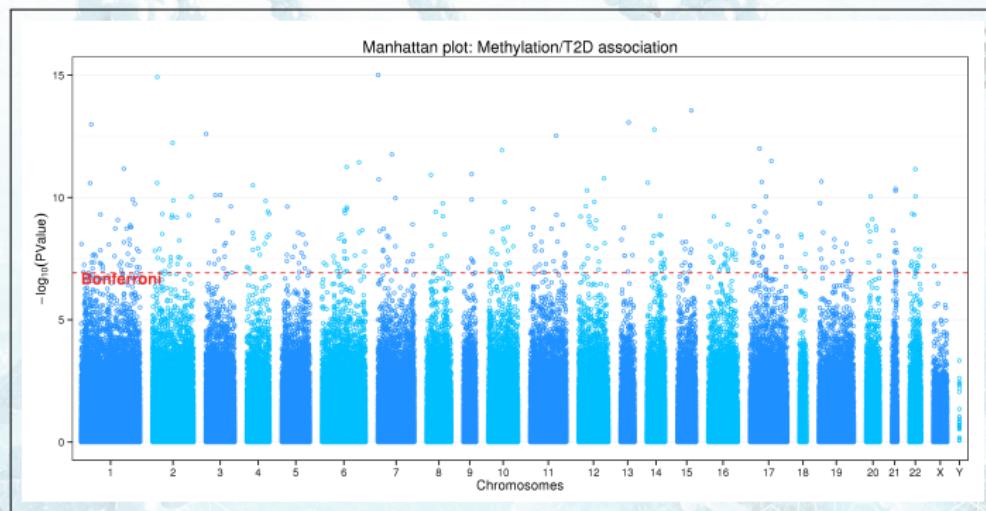
La correction de Bonferroni consiste à diviser notre seuil  $\alpha = 0.05$ , par le nombre de tests effectuer ( $n=5$ ), le nouveau seuil est donc

$$\alpha = 0.05/5 = 0.01 :$$

CpG	Estimée	Valeur p	Significatif
CpG1	0.002315	0.1619	Non
CpG2	0.006276	0.00363	Oui
CpG3	0.004945	0.09727	Non
CpG4	-0.04647	6.212e-06	Oui
CpG5	0.01847	0.04925	Non

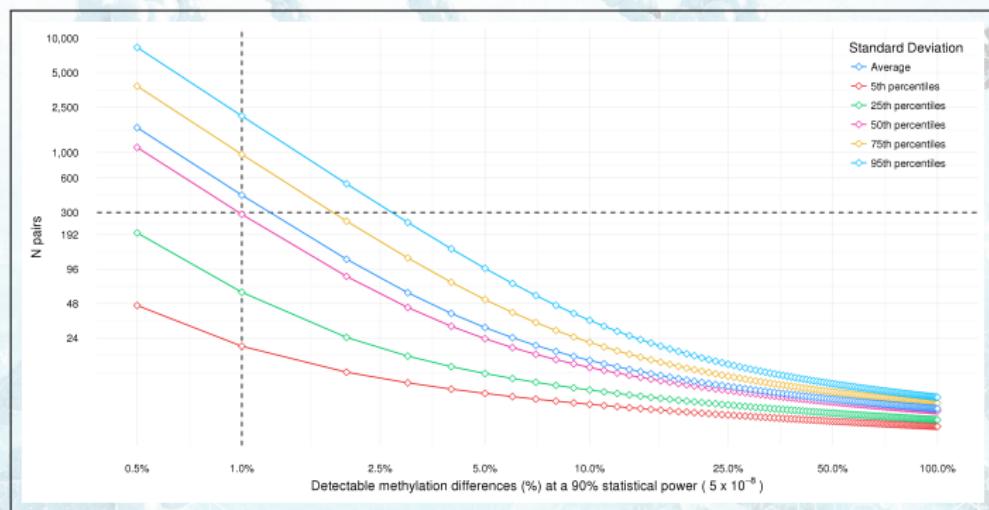
# Méthodes & Analyses X

Une représentation de ces résultats serait le "Manhattan plot".



# Méthodes & Analyses XI

## La puissance statistique (ou du statisticien) !





“The best thing about being a statistician is that you get to play in everyone’s backyard”

— John Tukey