Retour d'expérience sur l'analyse de données de méthylation

Lijiao NING (lijiao.ning@cnrs.fr)

2018-10-12

UMR 8199
GÉNOMIQUE INTÉGRATIVE ET MODÉLISATION DES MALADIES MÉTABOLIQUES

Egid European Genomic Institute for Diabetes

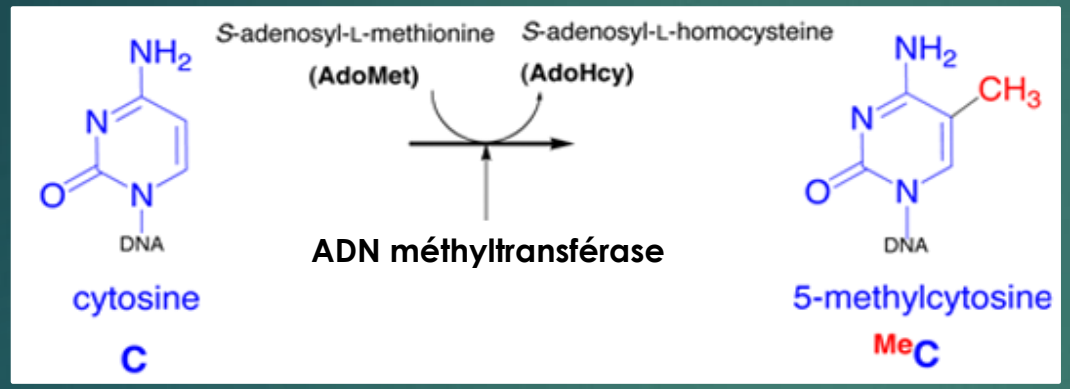
cnrs

Université de Lille

Institut Pasteur de Lille
Vivre mieux plus longtemps

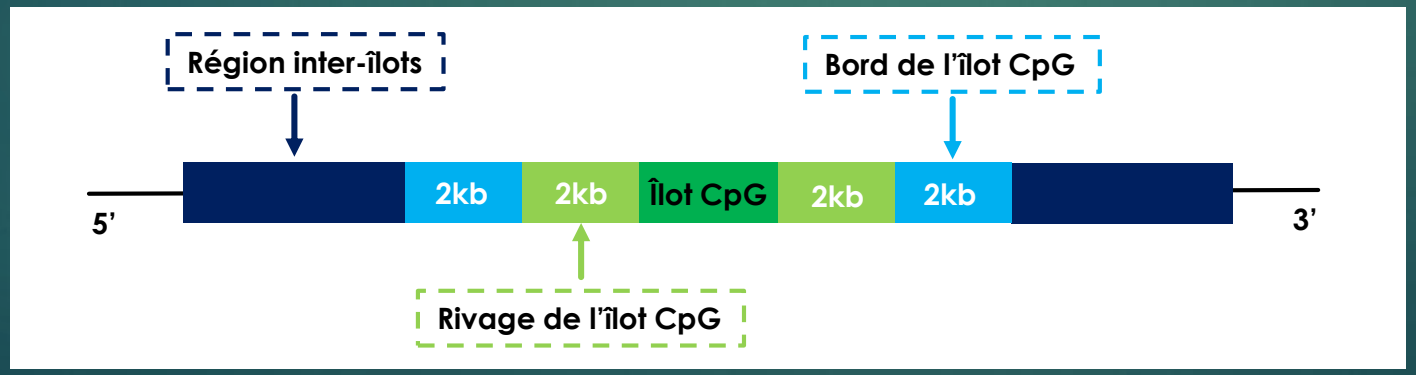
Méthylation d'ADN

- Ajout d'un groupe méthyle (CH3) sur la cytosine



(Source atlasgeneticsoncology.org)

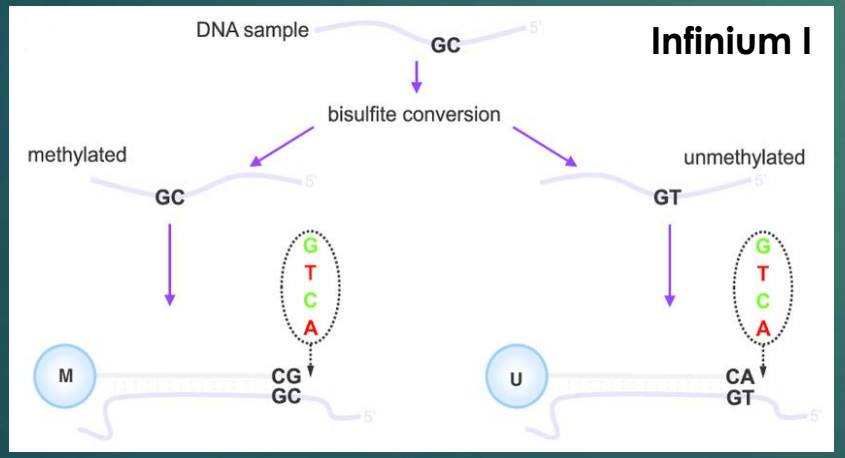
- Régions de sites CpG

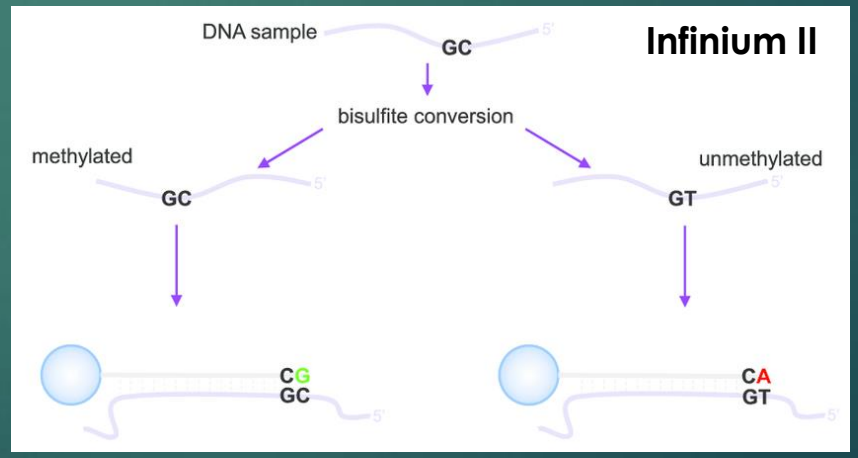


(Figure reproduite d'annotatr)

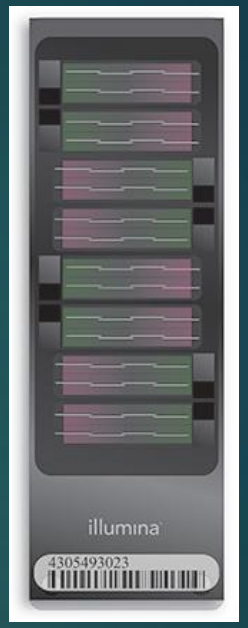
Conception d'EPIC Array

- > 850 000 sites CpG
- 8 échantillons/puce
- 2 types de design
  - Infinium I : une paire de sondes méthylée (M) et non méthylée (U)
  - Infinium II : un seul type de sonde





(Figures de Maksimovic, Gordon, and Oshlack , 2012)



Infinium® MethylationEPIC BeadChip

Valeur de méthylation

- Valeur bêta

$$\beta = \frac{Meth}{Meth + nonMeth + 100}, avec\ \beta \in [0, 1]$$

- Valeur M

$$M = \log_2 \frac{Meth + 1}{nonMeth + 1}, avec\ M \in (-\infty, +\infty)$$

- Relation

$$M = \log_2 \frac{\beta}{1 - \beta}; \qquad \beta = \frac{2^M}{2^M + 1}$$

Outils d'analyse

- Projet « Bioconductor » :
  - Collection de packages R pour l'analyse de données génomiques à haut débit
  - Libre d'accès

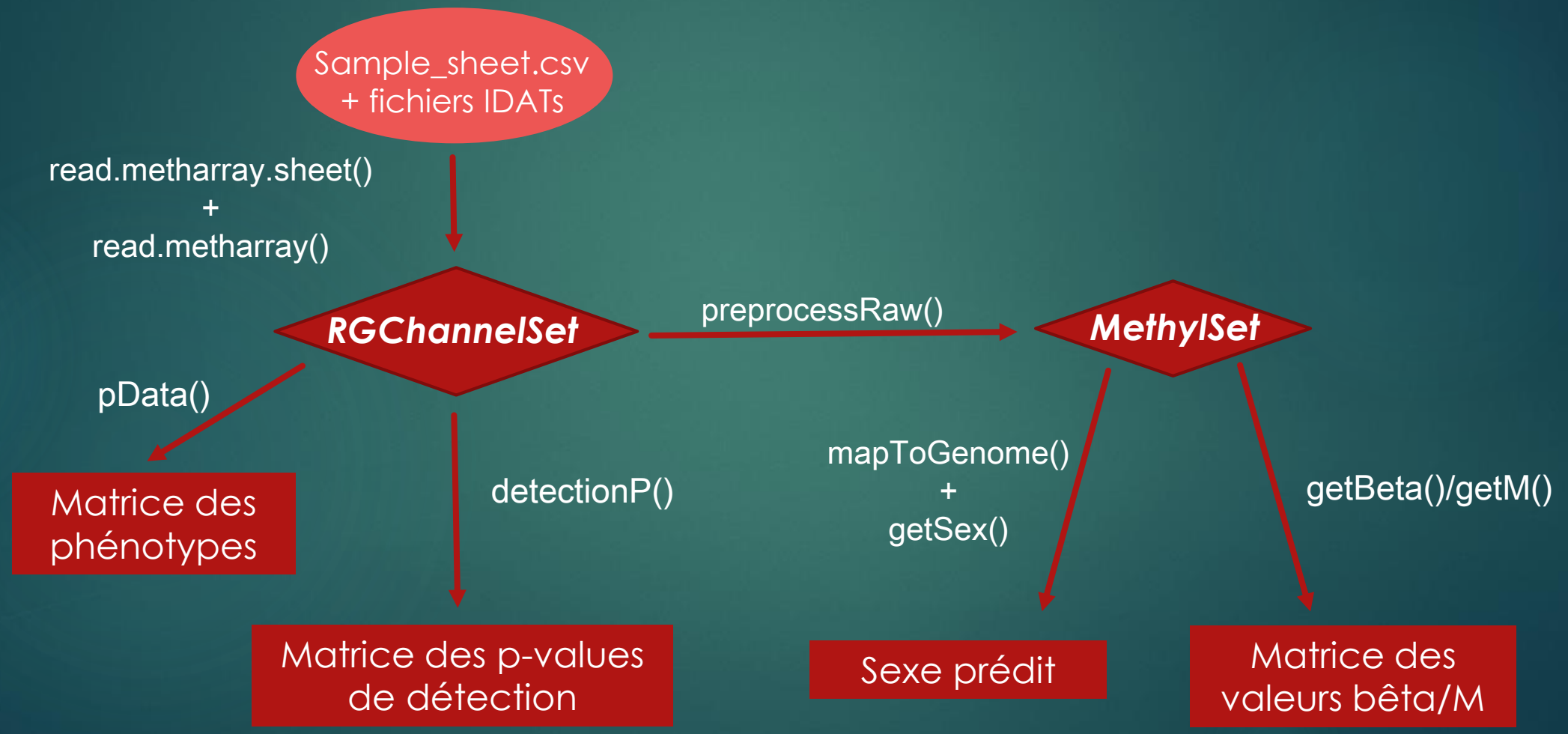
- Packages pour l'analyse des données de méthylation, p. ex. :
  - minfi *(objets « RGChannelSet » et « MethylSet »)*
  - methylumi *(objets « MethyLumiSet »)*
  - ChAMP
  - wateRmelon
  - missMethyl
  - limma

Pipeline d'analyse

*EWAS: epigenome-wide association study

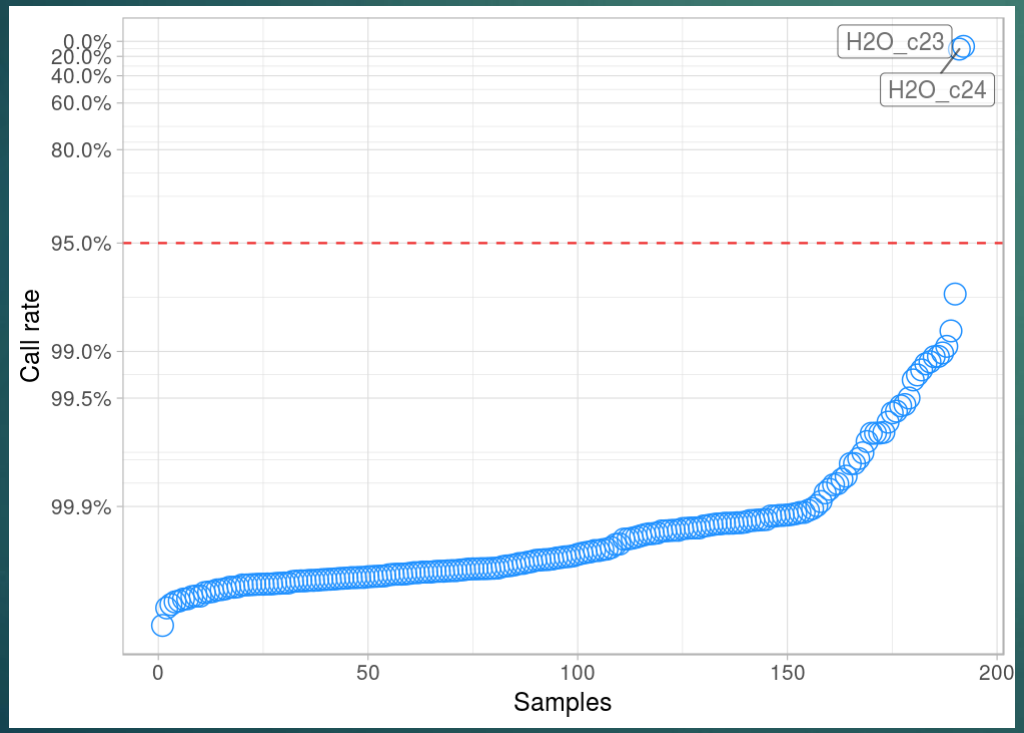
Importer les données – *minfi* (Aryee *et al.*, 2014)

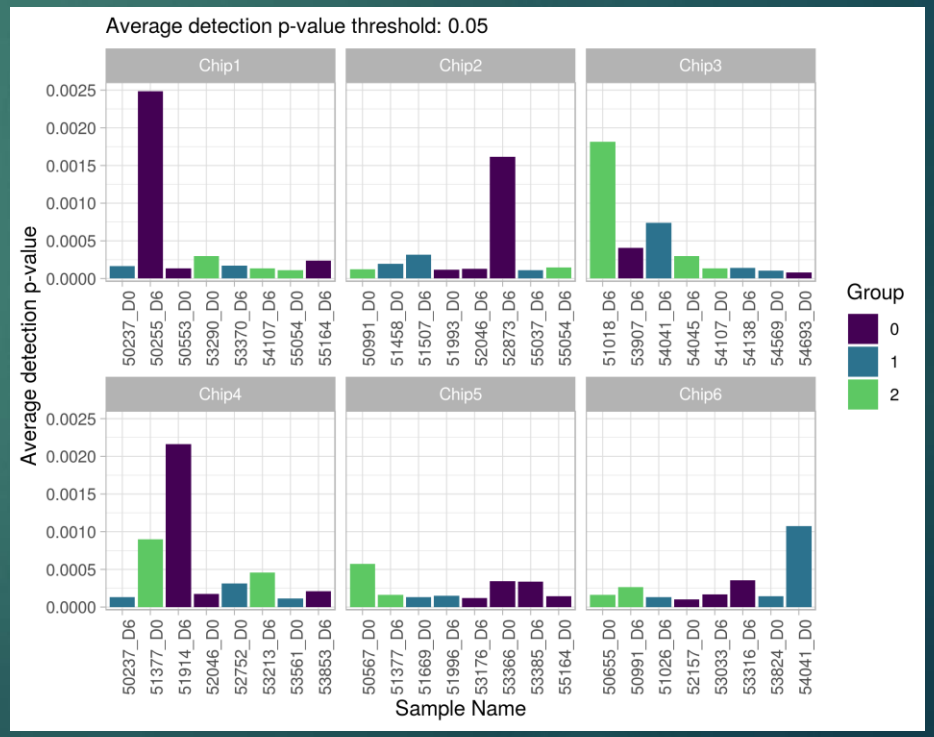
- Structure des données :



Contrôle de qualité

- Niveau des échantillons
  - P-value de détection
    - % de sondes fiables (p. ex. : 95% de $sondes\ ayant\ la\ p < 0{,}01$)
    - La moyenne des p-values (p. ex. : $\bar{p} < 0{,}05$)

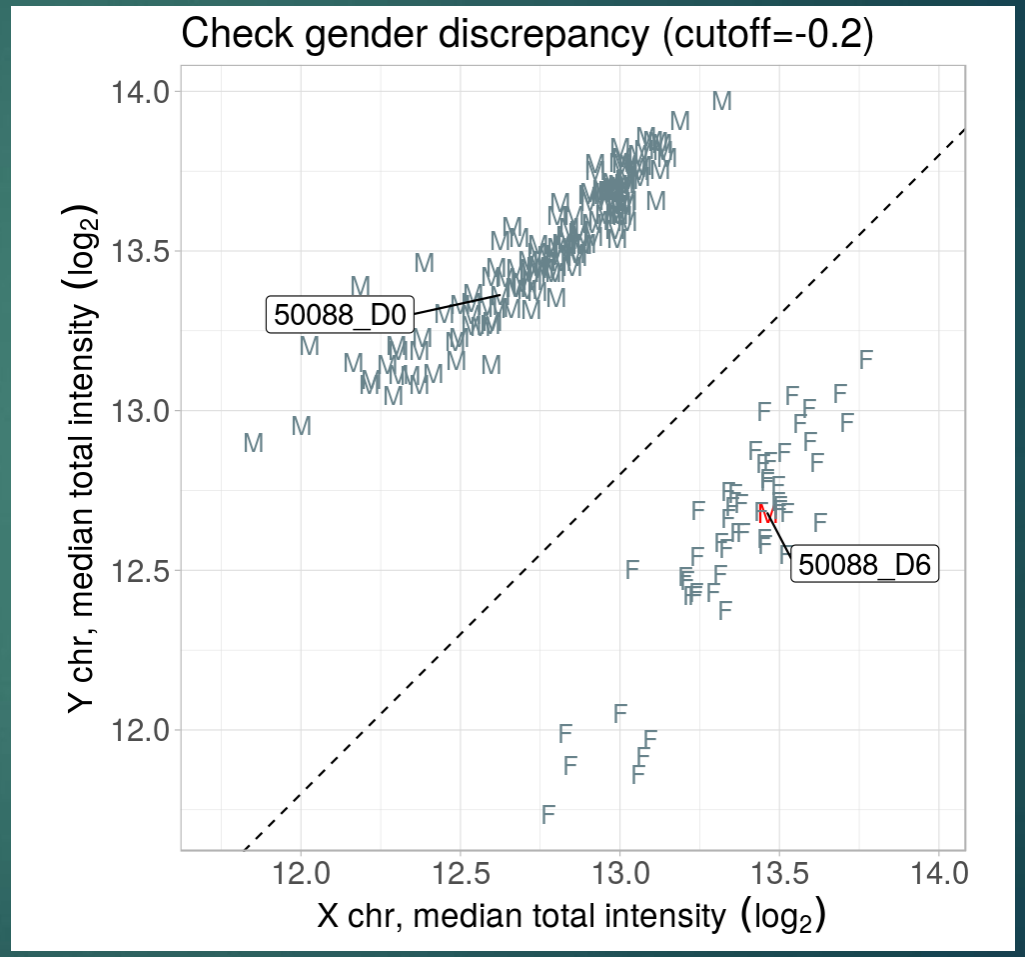




Contrôle de qualité

- Niveau des échantillons
  - P-value de détection
    - % de sondes fiables
    - La moyenne des p-values
  - Genre prédit *vs.* observé

    Liée à une maladie ?
  - Corrélation entre les réplicats



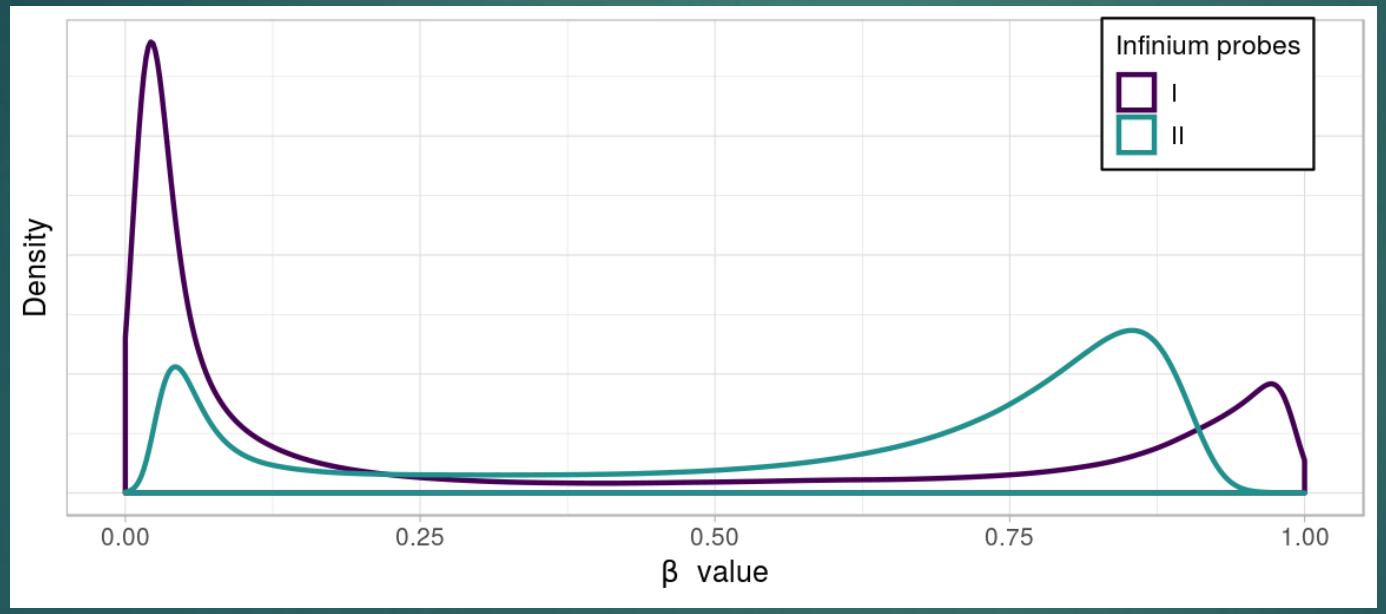
Contrôle de qualité

- Niveau des sondes
  - Sondes avec une $p > 0{,}01$ dans 1% d'échantillons
  - Sondes avec le nombre de billes < 3 dans 5% d'échantillons
  - Sondes non spécifiques
    - ⇒ s'alignent à plusieurs endroits sur le génome
  - Sondes avec des SNPs sur le site CpG
    - ⇒ possible de détecter une différence artificielle
  - Sondes sur les chromosomes sexuels
    - ⇒ méthylation très variée liée au genre
  - Sondes sur sites non-CpG

Éliminables selon besoins

Normalisation

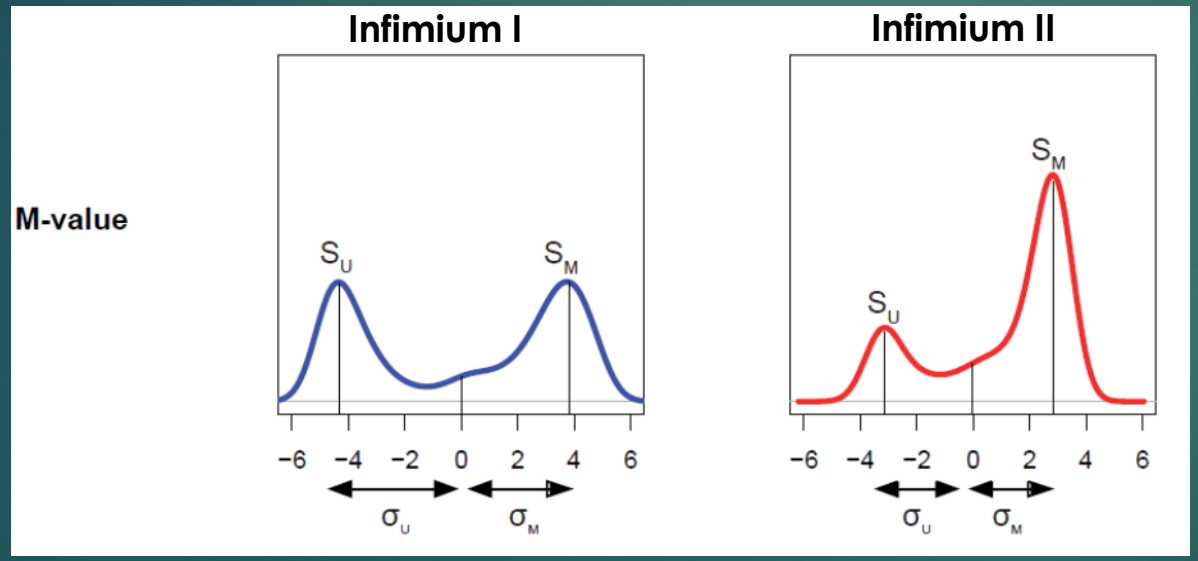
- But : ↓ le biais technique
- Source : sondes mixtes
  - **distributions différentes** des valeurs de méthylation



  - densité de sites CpG différente

Normalisation

- But : ↓ le biais technique
- Source : sondes mixtes
- Quelques méthodes
  - **PBC** : peak-based correction (Dedeurwaerder, S. *et al.*, 2011)



(Figure de Dedeurwaerder, S. *et al.*)

Pour toutes les valeurs M de sondes du type Infinium II :

$$Mval'_S = \frac{Mval_S}{\sigma_S^{InfII}} \times \sigma_S^{InfI}$$

$Mval'$: *valeur M corrigée*
S: *état de la sonde*, $S \in \{M, U\}$
σ: *distance entre l'abscisse du sommet et 0*

Normalisation

- But : ↓ le biais technique
- Source : sondes mixtes
- Quelques méthodes
  - **BMIQ** : beta-mixture quantile normalization (Teschendorff, A. E. *et al.*, 2013)
    - Étape 1 : adapter deux modèles de mélange de loi Bêta

$$p(\beta^t) = \pi_U^t B(\beta|a_U^t, b_U^t) + \pi_H^t B(\beta|a_H^t, b_H^t) + \pi_M^t B(\beta|a_M^t, b_M^t)$$

t : *type de sondes*, t ∈ $\{I, II\}$
a, b : *paramètres de loi Bêta*
π : *probabilité de la distribution bêta*

U : *état non méthylé*
H : *état hémiméthylé*
M : *état méthylé*

    - Étape 2 : normalisation de sondes du type II
      - Sonde U ou M : transformer séparément leurs probabilités en quantiles selon les paramètres de loi Bêta estimés à partir de sondes U ou M du type I
      - Sondes H : transformer les valeurs d'après les valeurs normalisées de sondes M et U du type II.

Normalisation

- But : ↓ le biais technique
- Source : sondes mixtes
- Quelques méthodes

Méthode	Description	Propriétés	Packages R
PBC[3]	Peak-based correction	Intra-échantillon	ChAMP wateRmelon
BMIQ[4]	Beta-mixture quantile normalization	Intra-échantillon Normalisation des quantiles	ChAMP wateRmelon
SWAN[5]	Subset-quantile within array normalization		minfi ChAMP
Funnorm[6]	Functional normalization		minfi ChAMP
Dasen[7]	Data-driven separate normalization	Inter-échantillon Normalisation des quantiles Ajustement du bruit de fond	wateRmelon

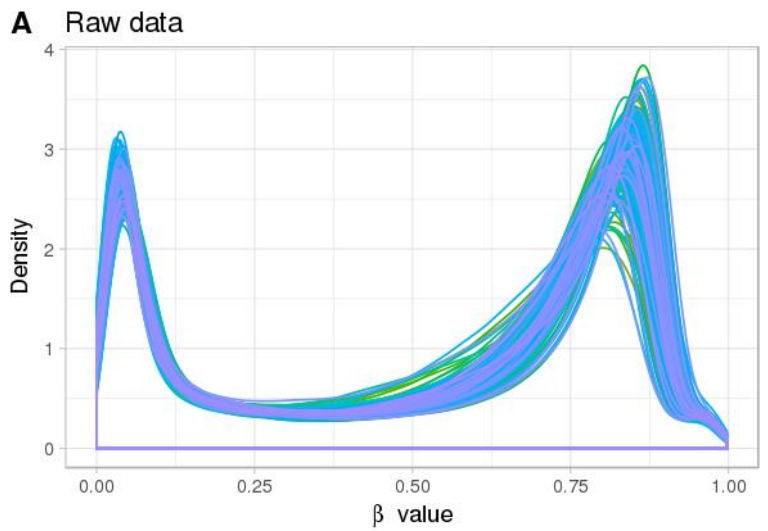
Correction de l'effet de lot

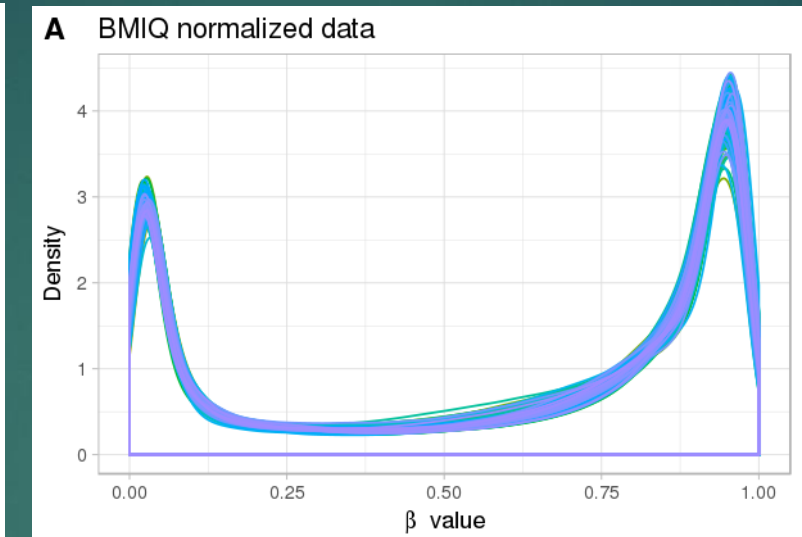
- But : éliminer la variation non biologique

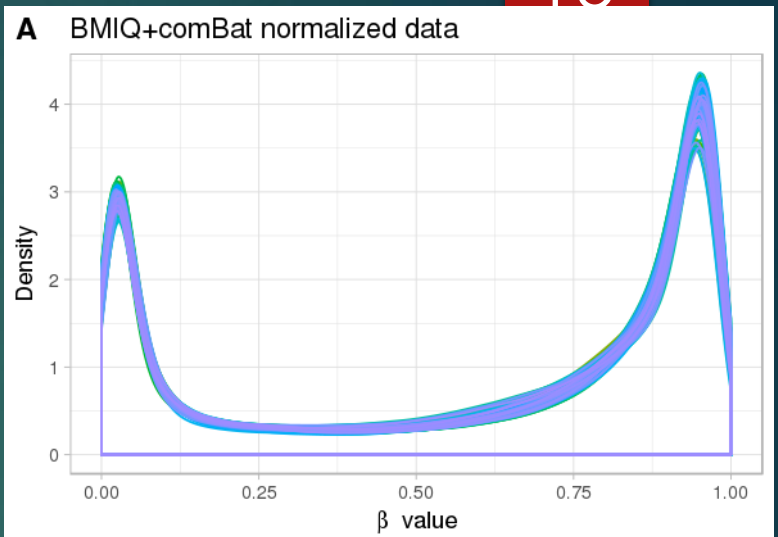
- Vérification
  - L'analyse en composantes principales (PCA)
  - La décomposition en valeurs singulières (SVD)

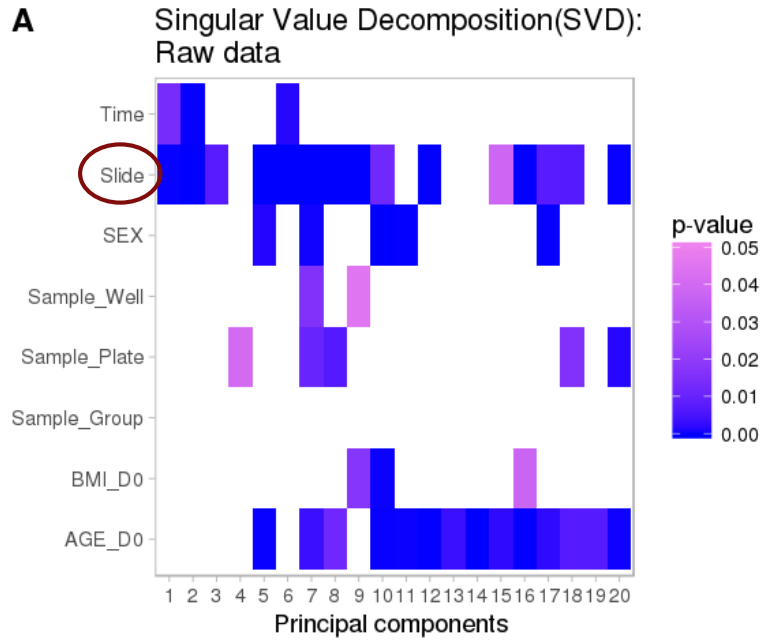
- Correction
  - Ajouter les composantes principales dans le modèle statistique
  - ComBat (méthode bayésienne empirique)

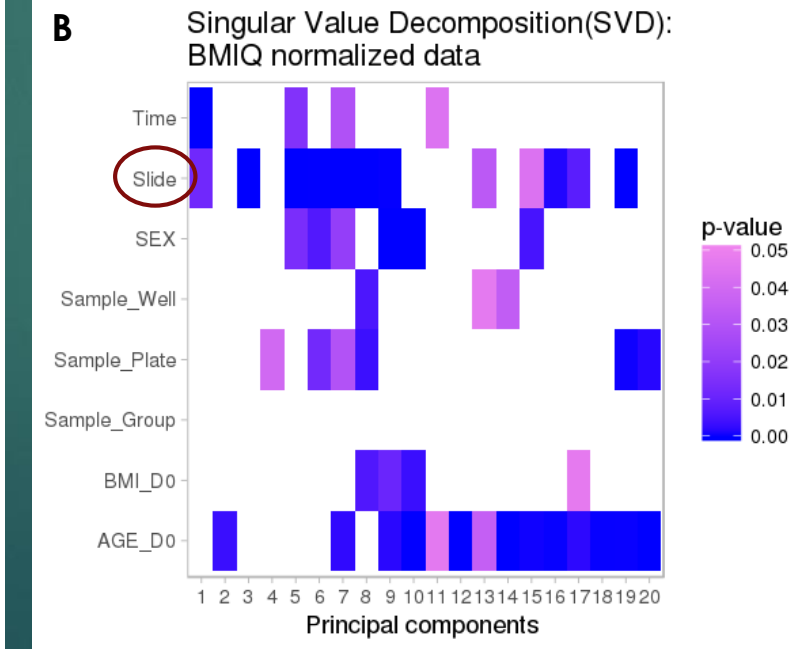
    ajuster l'effet de lot connu, corriger sur les valeurs bêta

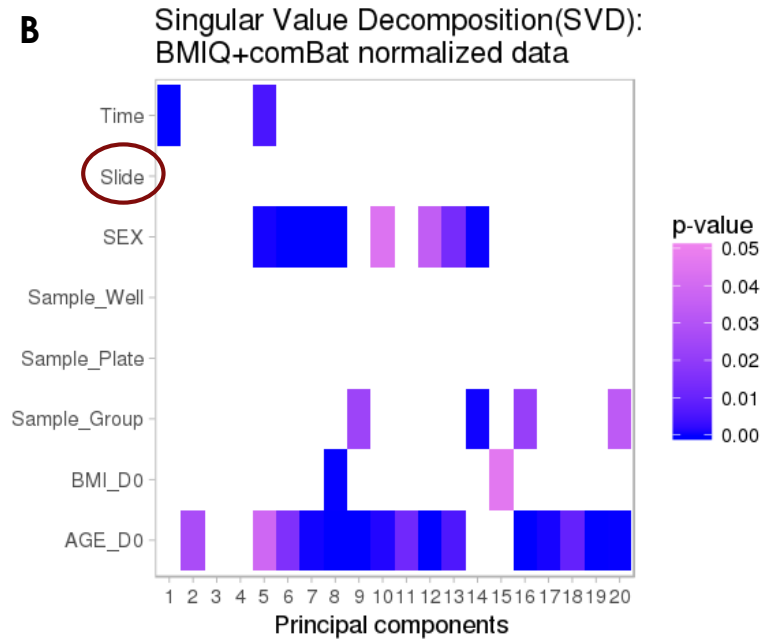

A
Raw data
Density
β value


A
BMIQ normalized data
Density
β value


A
BMIQ+comBat normalized data
Density
β value

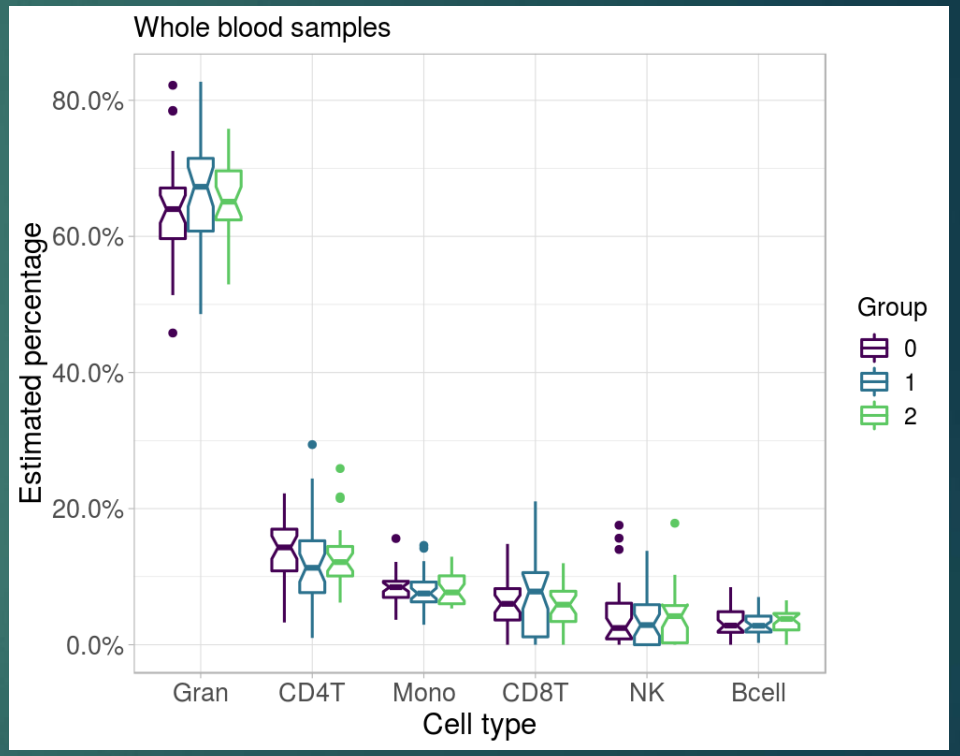

A
Singular Value Decomposition(SVD):
Raw data
Time
Slide
SEX
Sample_Well
Sample_Plate
Sample_Group
BMI_D0
AGE_D0
Principal components
p-value


B
Singular Value Decomposition(SVD):
BMIQ normalized data
Time
Slide
SEX
Sample_Well
Sample_Plate
Sample_Group
BMI_D0
AGE_D0
Principal components
p-value


B
Singular Value Decomposition(SVD):
BMIQ+comBat normalized data
Time
Slide
SEX
Sample_Well
Sample_Plate
Sample_Group
BMI_D0
AGE_D0
Principal components
p-value

Compositions cellulaires

- Cause : hétérogénéité de type cellulaire
  - Type d'échantillon : sang ? tissu ?
- Méthode d'estimation
  - Basée sur une référence
    - Types cellulaires connus
    - Sites CpG discriminants
    - ✓ Sang total, sang de cordon, cortex frontal
  - Sans référence
    - Analyse des variables de substitution (SVA): inférer les variables non modélisées depuis la matrice des valeurs bêta
- Compositions cellulaires liées au phénotype?



Analyse

- Méthylation différentielle au niveau des :
  - sites CpG
    - p. ex. : modèle linéaire site par site :

$$CpG_i = \alpha_0 + \alpha_1 Group + \alpha_2 Age + \alpha_3 BMI + \alpha_4 Gender + \sum_{j=1}^{k} (\gamma_j cellType_j) + \epsilon$$

  - régions
    - Les sites proches sont-ils méthylés différemment de même manière ?

      p. ex. : minfi::bumphunter, ChAMP::DMR
- Analyse d'enrichissement
  - p. ex. : analyse d'ontologie génétique
- Analyse combinée avec d'autres données omiques

Références

1. Moore, L. D., Le, T. & Fan, G. DNA Methylation and Its Basic Function. *Neuropsychopharmacology* **38,** 23–38 (2013).
2. Du, P. *et al.* Comparison of Beta-value and M-value methods for quantifying methylation levels by microarray analysis. *BMC Bioinformatics* **11,** 587 (2010).
3. Dedeurwaerder, S. *et al.* Evaluation of the Infinium Methylation 450K technology. *Epigenomics* **3,** 771–784 (2011).
4. Teschendorff, A. E. *et al.* A beta-mixture quantile normalization method for correcting probe design bias in Illumina Infinium 450 k DNA methylation data. *Bioinformatics* **29,** 189–196 (2013).
5. Maksimovic, J., Gordon, L. & Oshlack, A. SWAN: Subset-quantile within array normalization for illumina infinium HumanMethylation450 BeadChips. *Genome Biol.* **13,** R44 (2012).
6. Fortin, J.-P. *et al.* Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome Biol.* **15,** 503 (2014).
7. Pidsley, R. *et al.* A data-driven approach to preprocessing Illumina 450K methylation array data. *BMC Genomics* **14,** 293 (2013).
8. Tian, Y. *et al.* ChAMP: updated methylation analysis pipeline for Illumina BeadChips. *Bioinformatics* **33,** 3982–3984 (2017).
9. Aryee, M. J. *et al.* Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics* **30,** 1363–1369 (2014).
10. Teschendorff, A. E. & Relton, C. L. Statistical and integrative system-level analysis of DNA methylation data. *Nat. Rev. Genet.* (2017). doi:10.1038/nrg.2017.86
11. Maksimovic, J., Phipson, B. & Oshlack, A. A cross-package Bioconductor workflow for analysing methylation array data. *F1000Research* **5,** 1281 (2016).

Merci pour votre attention, Questions ?