Analyse différentielle des régions methylées

Revue de littérature

Jakobi Milan

TIMC-Imag

Vendredi 15 mars 2019

Table des matières

Données

- 450k microarray avec 25% des probes sur des zones intercalaires.
- On utilise les M-values ($= log(\beta)$)

Données

- 450k microarray avec 25% des probes sur des zones intercalaires.
- On utilise les M-values ($= log(\beta)$)
- Méthode extensible à toutes données génomiques : RRBS, WGBS...

Présentation de la méthode

3 types d'analyse proposées :

Présentation de la méthode

3 types d'analyse proposées :

1. Analyse entre deux groupes (Traitement vs Contrôle)
2. Analyse de contraste.
3. Analyse de variabilité (On identifie alors les VMRs).

Pseudo-algorithme

Après avoir choisi le type d'analyse, les étapes de la procédure seront les suivantes :

Pseudo-algorithme

Après avoir choisi le type d'analyse, les étapes de la procédure seront les suivantes :

1. On calcule Y_i nos statistiques de test.

Pseudo-algorithme

Après avoir choisi le type d'analyse, les étapes de la procédure seront les suivantes :

1. On calcule Y_i nos statistiques de test.
2. On estime la distribution de nos statistiques de test Y_i par noyau gaussien (par zones de taille λ).

Pseudo-algorithme

Après avoir choisi le type d'analyse, les étapes de la procédure seront les suivantes :

1. On calcule Y_i nos statistiques de test.
2. On estime la distribution de nos statistiques de test Y_i par noyau gaussien (par zones de taille λ).
3. On modélise, par méthode de Satterthwaite, nos statistiques de test "smoothées".

Pseudo-algorithme

Après avoir choisi le type d'analyse, les étapes de la procédure seront les suivantes :

1. On calcule Y_i nos statistiques de test.
2. On estime la distribution de nos statistiques de test Y_i par noyau gaussien (par zones de taille λ).
3. On modélise, par méthode de Satterthwaite, nos statistiques de test "smoothées".
4. On calcule les pvaleurs du modèle.

Pseudo-algorithme

Après avoir choisi le type d'analyse, les étapes de la procédure seront les suivantes :

1. On calcule Y_i nos statistiques de test.
2. On estime la distribution de nos statistiques de test Y_i par noyau gaussien (par zones de taille λ).
3. On modélise, par méthode de Satterthwaite, nos statistiques de test "smoothées".
4. On calcule les pvaleurs du modèle.
5. On fixe un seuil à partir duquel on exclue les variables dont la pvaleur est trop forte

Pseudo-algorithme

Après avoir choisi le type d'analyse, les étapes de la procédure seront les suivantes :

1. On calcule Y_i nos statistiques de test.
2. On estime la distribution de nos statistiques de test Y_i par noyau gaussien (par zones de taille λ).
3. On modélise, par méthode de Satterthwaite, nos statistiques de test "smoothées".
4. On calcule les pvaleurs du modèle.
5. On fixe un seuil à partir duquel on exclue les variables dont la pvaleur est trop forte
6. On construit nos DMRs/ZMRs finales en regroupant les CpG sites qui sont au plus à λ nucléotides

Construction des Y_i

Selon le type d'analyse, nos statistiques de test Y_i sont différentes : Pour les analyses entre deux groupes ou analyse de contraste, on a :

$$Y_i = \hat{t}^2$$

avec t^2 la statistique du test de Fisher modéré (ratio de la M-value sur son écart-type). Pour l'analyse de variabilité, on a :

$$Y_i = \frac{V_i}{V}$$

avec V_i la variance des M-values de l'échantillon i et V la moyenne de cette variance sur tous les échantillons. (Asymptotiquement équivalent à $F_{n-1,\infty}$).

Kernel smoothing

Pour l'estimation par noyau, chaque noyau est construit sur une longueur de λ nucléotides, et le paramètre d'échelle σ est déterminé par la relation suivante :

$$\sigma = \frac{\lambda}{C}$$

Avec C l'unique hyperparamètre de la méthode (C^* déterminé par CV). On pose :

$$\begin{cases} S_{KY}(i) = \sum_{j=1}^{n} K_{ij} Y_j \\ S_K(i) = \sum_{j=1}^{n} K_{ij} \\ S_{KK} = \sum_{j=1}^{n} K_{ij}^2 \end{cases} \quad (1)$$

Sélection des CpG sites

On pose (Satterwhaite) :

$$\begin{cases} a_i = \frac{S_{KK}(i)}{(\mu S_K(i)} \\ b_i = \frac{\mu S_K^2}{S_{KK}(i)} \end{cases} \tag{2}$$

et on teste $\frac{S_{KY}(i)}{a_i} \sim \chi^2_{b(i)}$
Les pvaleurs de ce test sont corrigées par procédure Bonjemini-Hochberg. On retient les sondes dont la pvaleur est inférieure au seuil choisi (l'auteur conseille 0.05).
On finit par construire nos DMRs (ou VMRs) en regroupant les CpG sites qui sont à au plus λ nucléotides de distance.

Pros et Cons

- Peu d'hyperparamètres : permet de s'affranchir des artefacts du jeu de données
- 95 % de recall sur jeu de données simulés.
- Les auteurs ont pu retrouvé plusieurs ZMRs qui revenaient à chaque fois selon le type de tissu utilisé.
- En comparant à d'autres méthodes utilissant aussi *limma*, les résultats étaient en moyenne meilleure et au moins aussi rapide.
- A l'étape de construction des zones (après sélection des CpG sites, pourquoi ne pas faire du ML non supervisé ?)
- ?
- coût calculatoire élevé (estimation par noyau).
- Taille des DMRs non contrôlable puisqu'adossée au paramètre d'échelle d'estimation par noyau + tailles virtuellement similaires (max 2λ.
- ?