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...

Méthode

3 types d'analyse proposées :

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 meilleurs et la méthode au moins aussi rapide.
- ?
- 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.
- ?
- A l'étape de construction des zones (après sélection des CpG sites, pourquoi ne pas faire du ML non supervisé ?)

Discovering high-resolution patterns of differential DNA methylation that correlate with gene expression changes.
VanderKraats et al. 2013

Données

Données d'expression des gênes et données de methylation mises côte à côte.

- WGBS et Methyl-MAPS sur données cancéreuses et saines sur cellules mammaires principalement. 17 datasets différents dont l'intersection représente 24.9% des CpG sites du génôme humain.
- N'importe quelles données d'expression différentielles (eg : RNA-Seq, expression array analysis).

Présentation générale

Cette méthode propose une mise en relation des données de methylation (construction de DMRs) et du niveau d'expression des gênes.
Permet l'identification de patterns de relation entre la méthylation, l'expression du gêne, et la position du TSS.

Pseudo-algorithme

2 échantillons A et B, on dispose de leur données de methylation et des niveaux d'expression différentielle des gênes correspondants :

1. Différence de methylation : B-A

Pseudo-algorithme

2 échantillons A et B, on dispose de leur données de methylation et des niveaux d'expression différentielle des gênes correspondants :

1. Différence de methylation : B-A
2. Interpolation autour de chaque TSS : création de signatures

Pseudo-algorithme

2 échantillons A et B, on dispose de leur données de methylation et des niveaux d'expression différentielle des gênes correspondants :

1. Différence de methylation : B-A
2. Interpolation autour de chaque TSS : création de signatures
3. Calcul de la distance de Fréchet discrète entre chaque paire de signature.

Pseudo-algorithme

2 échantillons A et B, on dispose de leur données de methylation et des niveaux d'expression différentielle des gênes correspondants :

1. Différence de methylation : B-A
2. Interpolation autour de chaque TSS : création de signatures
3. Calcul de la distance de Fréchet discrète entre chaque paire de signature.
4. Clustering hiérarchique ascendant en utilisant la distance comme critère d'aggrégation (Complete linkage).

Pseudo-algorithme

2 échantillons A et B, on dispose de leur données de methylation et des niveaux d'expression différentielle des gênes correspondants :

1. Différence de methylation : B-A
2. Interpolation autour de chaque TSS : création de signatures
3. Calcul de la distance de Fréchet discrète entre chaque paire de signature.
4. Clustering hiérarchique ascendant en utilisant la distance comme critère d'aggrégation (Complete linkage).
5. Sélection des clusters correlés à l'expression (KS test).

Pseudo-algorithme

2 échantillons A et B, on dispose de leur données de methylation et des niveaux d'expression différentielle des gênes correspondants :

1. Différence de methylation : B-A
2. Interpolation autour de chaque TSS : création de signatures
3. Calcul de la distance de Fréchet discrète entre chaque paire de signature.
4. Clustering hiérarchique ascendant en utilisant la distance comme critère d'aggrégation (Complete linkage).
5. Sélection des clusters correlés à l'expression (KS test).
6. Aggrégation des clusters jugés similaires dans une liste de gênes ayant un schéma de relation similaire entre expression-methylation-position de la variation par rapport au TSS.

Methylation Signatures

- On commence par soustraire pour chaque site le niveau de methylation entre B et A, la valeur obtenue pour chaque site est donc bornée en $[-1, 1]$.
- On interpole ensuite la courbe de methylation (spline cubique) en prenant pour chaque CpG les 1 à 5 plus proches CpG sites.
- Les régions ainsi crées avec moins de 5 CpG sites, les régions ne disposant pas d'au moins un CpG voisin et les régions ne contenant aucune différence de methylation >0.2.
- On lisse avec une estimation par noyau gaussien ($\sigma = 50bp$)

Methylation Signatures

- On commence par soustraire pour chaque site le niveau de methylation entre B et A, la valeur obtenue pour chaque site est donc bornée en $[-1, 1]$.
- On interpole ensuite la courbe de methylation (spline cubique) en prenant pour chaque CpG les 1 à 5 plus proches CpG sites.
- Les régions ainsi crées avec moins de 5 CpG sites, les régions ne disposant pas d'au moins un CpG voisin et les régions ne contenant aucune différence de methylation >0.2.
- On lisse avec une estimation par noyau gaussien ($\sigma = 50bp$)
- **On obtient donc de nombreuses signatures qui se chevauchent**

Clustering

On aggrège les régions en suivant la procédure suivante :

- On convertit chaque signature en courbe polynomiale (on moyenne tous les 10bp pour améliorer le temps de calcul) ce qui permet d'approximer efficacement la distance de Fréchet : on obtient notre critère de similarité/dissimilarité.
- On peut suite effectuer notre classification hiérarchique ascendante.
- On introduit ensuite l'expression des gènes : on test pour chaque cluster m l'expression des gènes face à toutes les autres données. On test par Kolmogorov-Smirnov et on contrôle l'inflation du risque α par procédure BH.

Clustering (suite) : création de la liste de gènes retenus

On cherche suite à dégager une liste de gènes de nos cluster. Les auteurs ont construit un critère de pureté défini comme la fraction des gênes qui ont une différence d'expression de même polarité. On peut ajouter à la liste des gènes remarquables les clusters suivant ce critère (le seuil proposé est 0.85 de pureté), avec comme contraintes :

- Si le cluster c' considéré ne chevauche aucun autre cluster déjà présent dans la liste, il est ajouté.

Clustering (suite) : création de la liste de gènes retenus

On cherche suite à dégager une liste de gènes de nos cluster. Les auteurs ont construit un critère de pureté défini comme la fraction des gênes qui ont une différence d'expression de même polarité. On peut ajouter à la liste des gènes remarquables les clusters suivant ce critère (le seuil proposé est 0.85 de pureté), avec comme contraintes :

- Si le cluster c' considéré ne chevauche aucun autre cluster déjà présent dans la liste, il est ajouté.
- Si c' est descendant d'un autre cluster qui est déjà sur la liste, il n'est pas ajouté

Clustering (suite) : création de la liste de gènes retenus

On cherche suite à dégager une liste de gènes de nos cluster. Les auteurs ont construit un critère de pureté défini comme la fraction des gênes qui ont une différence d'expression de même polarité. On peut ajouter à la liste des gènes remarquables les clusters suivant ce critère (le seuil proposé est 0.85 de pureté), avec comme contraintes :

- Si le cluster c' considéré ne chevauche aucun autre cluster déjà présent dans la liste, il est ajouté.
- Si c' est descendant d'un autre cluster qui est déjà sur la liste, il n'est pas ajouté
- Si c' est le parent d'un ou plusieurs clusters déjà présents sur la liste, on compare leur pureté : Si les descendants ont une pureté représentant moins de 30 % de la pureté de c', on ajoute c' a notre liste de gène et on retire les descendants, sinon on n'ajoute pas c' et on garde les descendants.

Avantages, inconvénients, perspectives...

- Forte capacité à distinguer des patterns de relation methylation/niveau d'expression **en prenant en compte la position du TSS**
- L'interpolation permet d'inférer efficacement les données manquantes
- Algorithme stable et robuste au bruit
- Enormément d'heuristiques : sélection des régions d'intérêt, paramètre de lissage, seuil et contraintes dans le critère de pureté construit
- Kolmogorov-Smirnov...
- La méthode gagnerait à ne pas utiliser Kolmogorov-Smirnov mais un autre critère de significativité (éventuellement mathématique). On pourrait imaginer d'autres distances pour la classification (JSD ?) et pour la pureté (negentropie ?).

Modeling complex patterns of differential DNA methylation that associate with gene expression changes. Schlosberg et al. 2017

Présentation générale