Sujet de stage ingénieur bioinfo biostat M1

Analyse multi-omique de la régulation de la transcription par la méthylation de l’ADN

Groupe EpiMed, Laboratoire IAB, Grenoble, France

Sujet : Conception et mise en oeuvre d’un outil pour l’analyse conjointe du méthylome et du transcriptome.

Encadrement : florent.chuffart@univ-grenoble-alpes.fr

Mots-clefs : regression linéaire, methylome, transcriptome,

Compétences : calcul scientifique, R, python, ssh, linux/unix

# Contexte scientifique

La plateforme EpiMed se donne pour objectif de développer des outils permettant, à partir des données multi-omiques, d’obtenir des molécules candidates pour le dépistage, le diagnostic et/ou l’évaluation pronostique adaptés au type ou sous-type de tumeurs étudiées.

# Contexte biologique

La méthylation de l’ADN est une marque épigénétique qui régule l’activité transcriptionnelle des cellules. Ce contrôle épigénétique est impliquée dans le processus de différenciation cellulaire, (développement embryonnaire, régulation du système immunitaire, oncogenèse).

Si dans la plupart des tissus somatiques, le niveau de methylation du génome présente peu de variations au sein de la population de cellules, dans les tumeurs, on observe une augmentation de variabilité de cette marque épigénétique [Morales Torres et al. Science 2016].

Dès lors, les tumeurs deviennent des modèles biologiques permettant d’étudier la régulation épigénétique de la transcription.

# Objectif du stage

L'objectif du stage est de développer un outil capable indépendamment i) d’interroger le modèle biologique *a priori* sur la régulation épigénétique de la transcription, ii) d’identifier et de caractériser des régions du génome dont le statut épigénétique contrôlent la transcription.

# Organisation du stage

Le/la stagiaire aura à sa charge :

* d’adapter le pipeline existant [Rousseaux et al. 2020] pour le jeu de données TCGA LUAD et LUSC (conception)
* de le déployer sur les infrastructures de calcul (mise en oeuvre)
* optimiser les paramètres des méthodes et évaluer la pertinence des résultats obtenus
* publier un package (code + documentation)
* renforcer l’activité methylome de la plateforme EpiMed (présentation, animation de la communauté)

# Profil recherché :

* Niveau master en statistique.
* Compétences avancées en programmation scientifique (R/python).
* Travail à distance sur cluster de calcul (bash/ssh/linux).
* Aptitudes à gérer et utiliser des données massives (big data).
* Fort intérêt pour les applications biologiques ou médicales.
* Capacité à travailler en équipe dans un contexte interdisciplinaire.
* Maîtrise courante de l’anglais.

# Biblio

[Morales Torres et al. Science 2016] "The linker histone H1.0 generates epigenetic and functional intratumor heterogeneity", figure 2G page 20

[Rousseaux et al. 2020] https://www.biorxiv.org/content/10.1101/852186v1

[2] RefFreeEWAS, https://cran.r-project.org/web/packages/RefFreeEWAS/index.html

[3] rlm, https://cran.r-project.org/web/packages/rlm/index.html

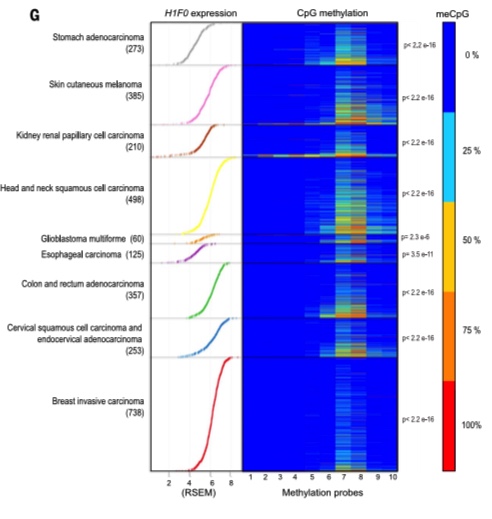
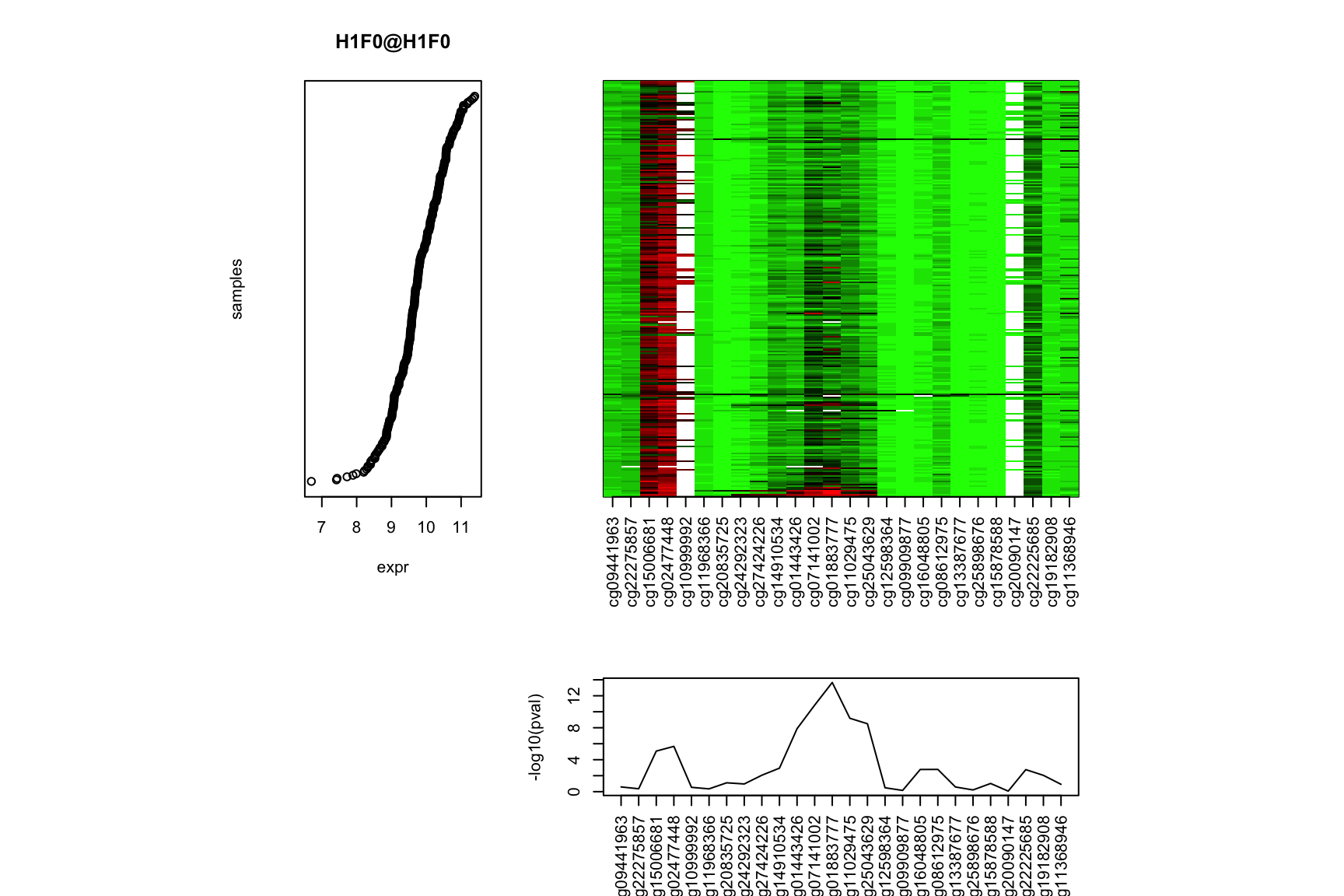
[4] comb-p, https://github.com/brentp/combined-pvalues

# Les données

Cohorte de test : TCGA LUAD et LUSC. Le jeu de données est composé de 1200 patients (500 LUAD, 500 LUSC, 200 NT)

Pour chaque patient nous disposons du methylomes (450k) et du transcriptome et CNV (2500 gènes), des profils mutationnels et des small RNA...

# Résultats préliminaires



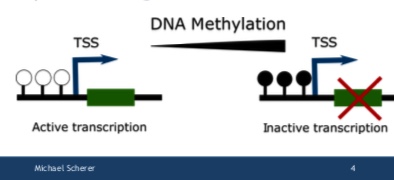
# Pipeline

Le pipeline d'analyse envisagé se fonde sur celui utilisé dans cette article [Rousseaux et al. 2020]. Il se décompose en 3 étapes principales :

* déconvolution du methylome à la l'aide de refFreeEwas [2]
* epigenome-wide association studies (EWAS) à l'aide d'une régression robuste [3]
* recherche de "differentially methylated region" (DMR) à l'aide de comb-p [4]
* Analyse fonctionnelle (biologie intégrative) et facteur pronostic (survie) des DMR trouvées.

**Généralité dsur la méthylationd e l’ADN**

La méthylation de l’ADN du promoteur d’un gène contrôle sont niveau d’expression. Un haut niveau de méthylation du promoteur d'un gène est souvent associé à une répression de ce gène (pas de transcription). A l'inverse, un faible niveau de méthylation du promoteur d'un gène est souvent associé à une expression de ce gène (transcription).



Le génome humain contient autour de  25000 gènes, 2/5 de ces gènes (15000) sont dits CpG riches, c’est à dire qu’ils contiennent un îlot CpG dans leur promoteur ou encore que la séquence d’ADN située en amont de ces gènes est beaucoup plus riches en CG qu’une distribution aléatoire. Leur expression est donc potentiellement associée à l’état de méthylation de leur promoteur.

Dans la plupart des tissus sains, 2/3 des gènes CpG riches (10000) seront constamment déméthylés, 1/3 des gènes CpG riches (5000) seront constamment méthylés sauf pour quelques tissus (lire Saxonov 2006). On parle alors de gènes tissus spécifiques et c’est la méthylatioin de l’ADN qui va contrôler cette spécificité (contrôle de l'expression par la méthylation). Dans les faits, les promoteurs sont méthylés ou démethylés et cet état est stable dans un tissue donné.

Dans les cancer, cette stabilité est remise en cause. Par exemple cette article [Morales Torres et al. Science 2016] met en évidence, à l’intérieur d’une même cohorte, différents niveaux de méthylation du promoteur du gène H1F0. De plus, les auteurs corrèlent négativement le niveau d’expression de ce gène et le niveau de méthylation de son promoteur. Ainsi les cancers deviennent des modèles biologiques capables de mettre en évidence des régions du génomes plastiques du point de vue de la méthylation (hotspot épigénétiques). L’identification de ces régions est un enjeu majeur pour comprendre les mécanismes de l'oncogenèse (diagnostic), relier ces mécanismes à l'agressivité des cancers (pronostic) et identifier des cibles potentielles pour des futurs traitements.

**Objectif du stage**

L'objectif du stage développer un outil capable de d’identifier et de cataloguer des régions du génome affectées par la tumeur et l’environnement tumoral.

Cette objectif s’organise autour de deux tâches : i) reproduire le pattern, ii) étendre l’approche.

Reproduire le pattern promoteur/gene décrit dans la figure 2G page 20 du papier [1] dans une jeu de données de cancers du poumon.

Etendre l’approche pour trouver d’autre régions du génome sensible à l’environnement tumoral.

Cela se concrétise par la recherche systématique et/ou a priori de régions du génome qui régulent l’expression des gènes.

**Biblio**

[13] Saxonov S, Berg P, Brutlag DL. A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proceedings of the National Academy of Sciences of the United States of America. 2006;103:1412-7.

Un bon préalable est de comprendre ce qu’est la methylation :

Tu peux lire ça concernant la méthylation, c’est en français et c’est très clair.

https://www.medecinesciences.org/en/articles/medsci/full\_html/2008/10/medsci2008248-9p731/medsci2008248-9p731.html

Ils utilisent une autre technologie mais les enjeux sont les mêmes.

Ceci est interessant aussi (pour info)

https://www.medecinesciences.org/en/articles/medsci/full\_html/2008/10/medsci2008248-9p747/medsci2008248-9p747.html

Ensuite concernant la techno :

EPIC, le chapitre 1 page 2 à 9 de ce document devrait t’aider, commence pas là :

https://support.illumina.com/content/dam/illumina-support/documents/documentation/chemistry\_documentation/infinium\_assays/infinium\_hd\_methylation/infinium-hd-methylation-guide-15019519-01.pdf

RRBS :

Le protocole est dispo ici :

https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1258174/

Mais ca va etre difficile à lire.

Peut ceci est plus facile :

https://en.wikipedia.org/wiki/Reduced\_representation\_bisulfite\_sequencing

ceci aussi:

https://www.illumina.com/science/sequencing-method-explorer/kits-and-arrays/rrbs-seq-scrrbs.html?langsel=/us/