Importance de la variabilité de la méthylation dans le cancer

RAPPORT BIBLIOGRAPHIQUE

Fabien KON-SUN-TACK

MASTER Sciences et Numériques pour la Santé
PARCOURS Bioinformatique Connaissances et Données

ENCADRANTS : Magali RICHARD, Florent CHUFFART
ANNÉE UNIVERSITAIRE : 2020-2021

26 février 2021

Remerciements

Je souhaiterais remercier l'ensemble des personnes qui m'ont accompagné lors de la rédaction de ce projet de recherche bibliographique.

Je tiens tout d'abord à remercier Mme. Richard et M. Chuffart, qui ont accepté ma collaboration pour ce travail de recherche bibliographique, et qui ont accepté ma candidature pour le stage du second semestre dans leur équipe de recherche.

Je remercie également l'ensemble de l'équipe pédagogique du master Sciences et numériques pour la Santé, en particulier Mme. Lautier et M. Mancheron pour leur présence et leurs conseils concernant l'organisation de ce travail de recherche.

Je tiens aussi à remercier la responsable pédagogique de ce Master, Mme.Fiston-Lavier, qui nous a apporté tant de soutien nécessaire en cette période troublée.

Enfin, je souhaiterais remercier mon tuteur pédagogique, M. Boureux, pour les précieux conseils de rédaction qu'il m'a apportés.

Table des matières

INTRODUCTION

Des études récentes sur l'épigénome dans le cancer ont montré le rôle central joué par les modifications épigénétiques dans le processus tumoral. En particulier, des variations dans la méthylation de l'ADN dans le cancer modifient significativement l'expression des gènes et impactent l'évolution de la tumeur. Depuis l'avènement du séquençage haut-débit, de grands jeux de données multiomiques ont été généré sur des échantillons tumoraux. Un des enjeux majeurs pour la recherche en cancérologie est maintenant d'intégrer ces nouvelles données dans un modèle de dérégulation génétique tenant compte des variations de la méthylation. Un tel modèle offrira une meilleure compréhension des mécanismes biologiques en jeu lors de l'apparition du cancer.

L'analyse de la méthylation de l'ADN s'accompagne de problèmes méthodologiques lié à la nature de ces données : grande dimension, données manquantes, intégration des facteurs de confusion, méconnaissances des déterminants biologiques sous-jacents, variabilité temporelle et spatiale. Nous proposons ici une approche bioinformatique originale permettant une meilleure caractérisation de la variance de la méthylation sur des lignées cellulaires. Ce projet devrait permettre de développer de nouveaux outils contribuant à l'interprétation de la dérégulation épigénétique dans le cadre du cancer.

Dans ce rapport, nous allons présenter : i) la nature de la méthylation de l'ADN, ii) l'état de l'art sur la méthylation de l'ADN dans le cadre du cancer et iii) l'approche méthodologique proposée pour étudier la variabilité de la méthylation de l'ADN sur des lignées cellulaires.

1 La méthylation de l'ADN

1.1 Définition de la méthylation

La méthylation au niveau de l'ADN correspond à l'implantation d'un groupement méthyle (ou CH3) sur des résidus cytosine. Cette modification a été initialement décrite comme un marqueur de répression de la transcription : les gènes présents dans les régions méthylées du génome ne sont plus transcrits, et ne peuvent plus coder pour leurs protéines.
Cette méthylation s'opère majoritairement dans des régions où on observe une grande concentration en cytosine et guanine, comparé au reste du génome. Ces régions sont appelées des îlots CpG (abréviation pour Cytosine-phosphate-Guanine).

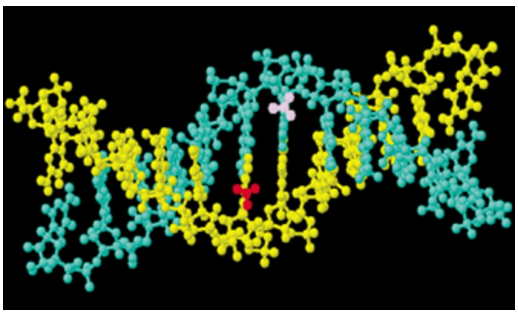

FIGURE 1 – **Structure de l'ADN au niveau d'un dinucléotide CpG méthylé.** Les groupements méthyle des deux chaînes (en violet et en rouge) se situent dans le grand sillon, en position accessible. Après la réplication de l'ADN, les sites hémiméthylés sont complétés par une méthyltransférase de maintien, ce qui assure la transmission de ce motif [1]

1.2 Dynamique de la méthylation

La fonction de la méthylation de l'ADN est intrinsèquement associée aux mécanismes permettant l'instauration, le maintien et la suppression des groupements méthyl. Cette méthylation est établie de novo par l'activité catalytique des DNA méthyltransférases DNMT3a, DNMT3b et DNMT3c et est maintenue par DNMT1 au cours des divisions cellulaires.

La méthylation de l'ADN est une marque stable dans la plupart des cas, mais cette méthylation doit également être régulée de manière active ou passive, afin d'établir un état permissif pour l'expression des gènes en aval. Des études récentes ont mis en évidence l'implication d'enzymes spécifiques dans la déméthylation active, notamment lors de la reprogrammation des cellules vers un état pluripotent [2].

Ce processus fait intervenir un mécanisme au cours de la division cellulaire ou de la réparation de l'ADN, qui engendrerait une excision des bases plutôt que la suppression des motifs méthyl à partir des 5MC. Ces enzymes comprennent entre autres les enzymes dites TET (ten-eleven translocation) et les AID (activation-induced cytidine deaminase) ([3]).

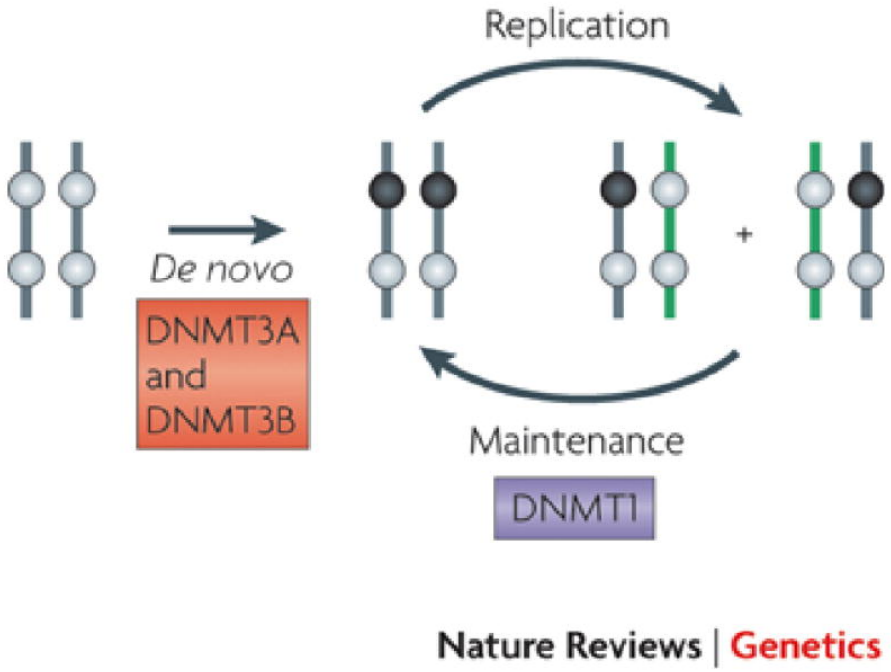

FIGURE 2 – Modèle actuel pour l'établissement et l'héritage de la méthylation de l'ADN [4]

1. Filion, Guillaume, et Pierre-Antoine Defossez. « Les protéines se liant à l'ADN méthylé : interprètes du code épigénétique ». médecine/sciences 20, n° 1 (janvier 2004) : 7-8.

3. Parry, Aled, Steffen Rulands, et Wolf Reik. « Active Turnover of DNA Methylation during Cell Fate Decisions ». Nature Reviews Genetics, 6 octobre 2020.

3. Bhutani, Nidhi, Jennifer J. Brady, Mara Damian, Alessandra Sacco, Stéphane Y. Corbel, et Helen M. Blau. « Reprogramming towards Pluripotency Requires AID-Dependent DNA Demethylation ». Nature 463, n° 7284 (février 2010) : 1042-47.

4. Jones, Peter A., et Gangning Liang. « Rethinking How DNA Methylation Patterns Are Maintained ». Nature Reviews Genetics 10, n° 11 (novembre 2009) : 805-11.

1.3 Méthylation et expression génétique

La méthylation de l'ADN a un effet variable et complexe sur l'expression des gènes selon les régions concernées (Figure 7).
Par exemple, elle est considérée comme une marque répressive de l'expression lorsqu'elle est située dans le promoteur d'un gène. La méthylation des îlots CpG est généralement associée avec une absence d'activité de transcription [5]. La méthylation de l'ADN au sein des régions transcrites (séquences codantes) est à l'inverse une marque d'activation de la transcription. En effet, elle conduit à l'inactivation des îlots CpG dans les régions intragéniques, ce qui favoriserait l'élongation de la transcription [6].

La méthylation des sites dits isolateurs (Insulator) a également un effet sur l'expression des gènes. Ces sites sont positionnés à distance des séquences géniques qu'ils régulent. Dans ces régions, la méthylation de l'ADN inhiberait la liaison du facteur de transcription CTCF [7]. La liaison de CTCF contribue à définir la délimitation entre chromatine active et hétérochromatine (structure tridimensionnelle de la chromatine). Ainsi, l'inhibition de liaison de CTCF par la méthylation servirait de marque de répression des gènes.

Au niveau des sites dits amplificateurs (ou Enhancers), la densité des îlots CpG y est bien plus faible. Situés à des distances variables des sites promoteurs, il s'agit de sites clés pour la régulation de l'expression génique lors du développement embryonnaire. Dans ces régions amplificatrices, la méthylation de l'ADN y présente des niveaux variables [8], et sont alors dénommées « Low-Methylated Regions ».

Une explication de ce statut variable de la méthylation de l'ADN serait liée à un dynamisme des quelques îlots CpG dans ces régions, couplé à un héritage inefficace de la méthylation lors de la division cellulaire (Jones 2012).

En conclusion, le rôle exact de la méthylation dans le contrôle de l'expression des gènes n'est pas encore résolu, et fait toujours l'objet de nombreuses études scientifiques à l'heure actuelle [9].

---

5. Weber, Michaël. « Profils de méthylation de l'ADN dans les cellules normales et cancéreuses ». médecine/sciences 24, n° 8-9 (1 août 2008) : 731-34.

6. Larsen, Frank, Jorun Solheim, et Hans Prydz. « A Methylated CpG Island 3' in the Apolipoprotein-E Gene Does Not Repress Its Transcription ». Human Molecular Genetics 2, n° 6 (1993) : 775-80.

7. Bell, Adam C., et Gary Felsenfeld. « Methylation of a CTCF-Dependent Boundary Controls Imprinted Expression of the Igf2 Gene ». Nature 405, n° 6785 (mai 2000) : 482-85.

9. Weber, Michaël. « Profils de méthylation de l'ADN dans les cellules normales et cancéreuses ». médecine/sciences 24, n° 8-9 (1 août 2008) : 731-34.

9. (Stadler et al, 2011 : DNA-binding factors shape the mouse methylome at distal regulatory regions)

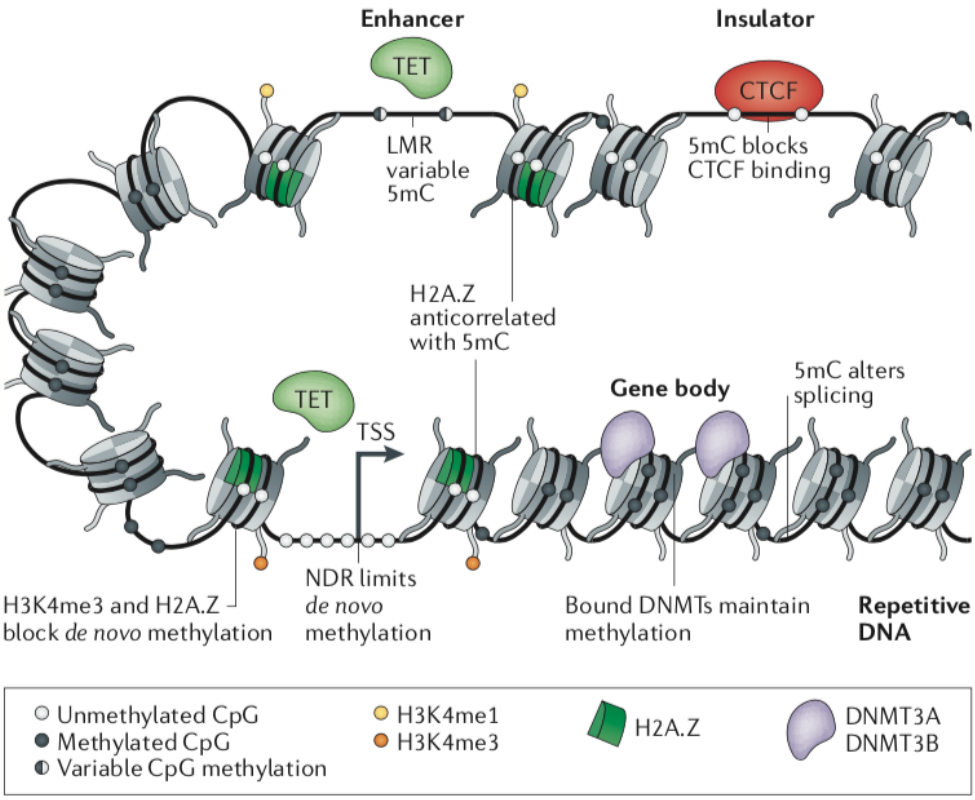

FIGURE 3 – Rôle des îlots CpG dans la régulation de l'expression des gènes (JONES 2012)

1.4 Mesure de la méthylation : approches bioinformatiques

Plusieurs approches se sont développées ces dernières années pour détecter les motifs des cytosines méthylées (5mc) à l'échelle du génome. Ces méthodes font intervenir la digestion des cytosines par des enzymes de restriction sensibles à la méthylation, et la capture des 5Mc par des protéines méthylées liant l'ADN, suivies par un séquençage de nouvelle génération (NGS).

Une autre approche est basée sur l'immunoprécipitation de l'ADN méthylé, dans laquelle l'ADN extrait est clivé, dénaturé et précipité en utilisant un anticorps ciblant la 5MC, pour aboutir à un séquençage des fragments précipités.

La méthode de mesure de la méthylation la plus répandue est basée sur le traitement de l'ADN par les bisulfites.

Ce traitement convertit les cytosines non méthylées en Uraciles, qui seront par la suite amplifiées par méthode de PCR, et seront reconnues en tant que Thymines. Les cytosines méthylées seront quant à elles inchangées, ce qui permet de mesurer de manière efficace le taux de méthylation dans l'ADN étudié.

L'utilisation de micropuces, telles que les puces Illumina 450k, s'est largement répandue pour l'analyse de l'ADN traité par bisulfites.

Le traitement de l'ADN par les bisulfites est accompagné par le développement d'une méthode de séquençage intitulée *Reduced Representation Bisulfite Sequencing* (RRBS), qui consiste à cliver l'ADN par les enzymes DNMT avant le traitement par les bisulfites.

La couverture de la méthylation de l'ADN peut s'obtenir par différentes méthode, mais la plus compréhensive à l'échelle d'une seule base est obtenue par le séquençage en "shotgun" de l'ADN traité par bisulfites.

Cependant, l'utilisation de ces composés soulèvent le problème de l'ambiguïté entre la 5MC et la 5-hydroxyméthylcytosine (5-hmC), un composé qui a récemment été identifié dans l'ADN.

En effet, le traitement des 5-hmC par les bisulfites ne permet pas la conversion de ces cytosines en thymines, ce qui correspond au même comportement que les 5mC. Ainsi, la technique des bisulfites ne différencie pas les deux composés, ce qui peut engendrer des faux positifs dans l'analyse statistique [10].

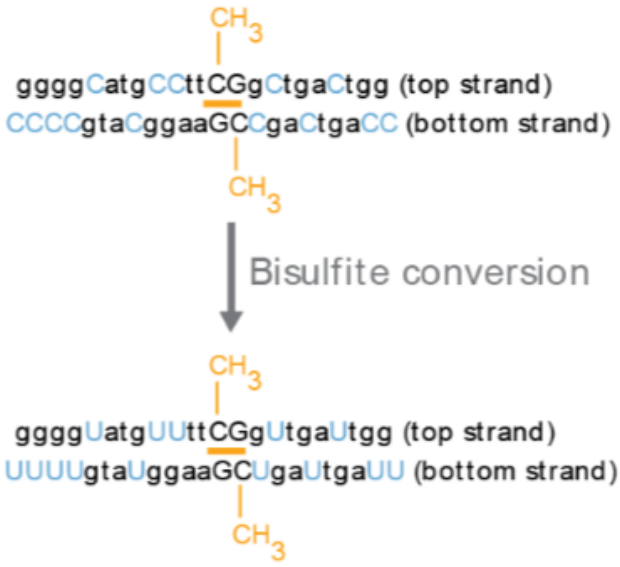

FIGURE 4 – **Principe de la méthode des bisulfites** [11]

2 La méthylation dans le cancer

Les marqueurs du cancer

Le cancer est considérée comme la maladie caractéristique du XXIè siècle. Il constitue la 2e cause mondiale de décès, avec plus de 8,8 millions de morts en 2015, soit 1 personne sur 6 dans le monde. Les mécanismes du cancer sont très complexes de par la variabilité des types de cellules dans les différents types de tumeurs. Par conséquent, le nombre accru d'études scientifiques à ce sujet n'a pas encore permis de résoudre l'ensemble des problématiques associées à cette variabilité. Cependant, plusieurs caractéristiques sont communes à l'ensemble des types de cancers [12] :

10. Huang, Yun, William A. Pastor, Yinghua Shen, Mamta Tahiliani, David R. Liu, et Anjana Rao. « The Behaviour of 5-Hydroxymethylcytosine in Bisulfite Sequencing ». Édité par Jun Liu. PLoS ONE 5, n° 1 (26 janvier 2010) : e8888.

11. « Infinium HD Methylation Assay Protocol Guide (15019519) », s. d., 244.

12. Hanahan, Douglas, et Robert A. Weinberg. « Hallmarks of Cancer : The Next Generation » . Cell 144, n° 5 (mars 2011) : 646-74.

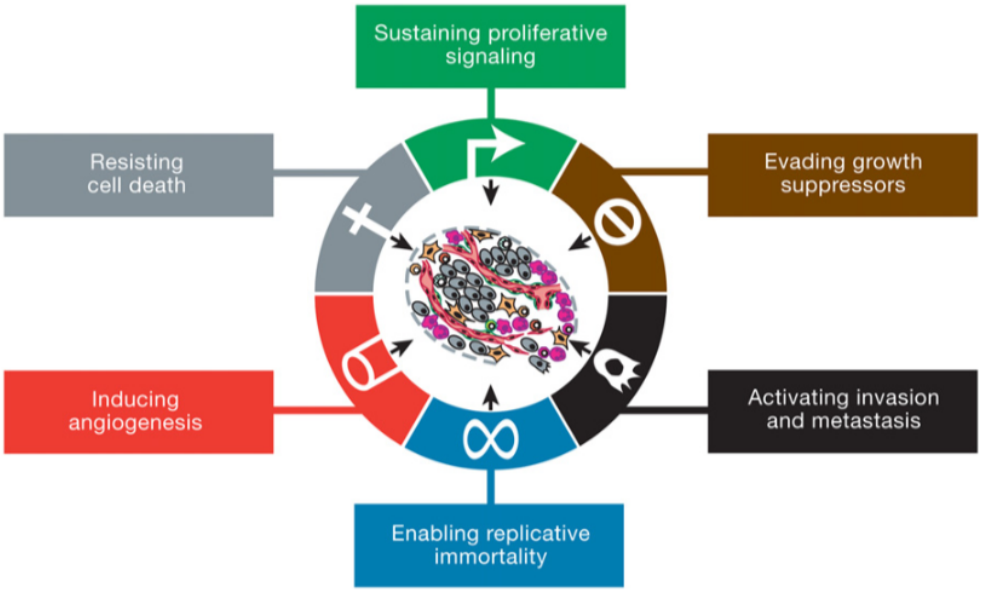

FIGURE 5 – Les caractéristiques du cancer

1. Ainsi, les cellules cancéreuses détournent la machinerie cellulaire pour maintenir une signalisation cellulaire propice à leur prolifération, notamment en activant des voies de signalisation de manière constitutive.
2. Elles peuvent également contourner l'action des gènes dits « suppresseurs de tumeurs » (dont celui codant pour la protéine p53), induisant par la suite leur prolifération permanente.
3. L'inactivation du gène TP53 traduit en outre la capacité des cellules tumorales à contourner ou à limiter les mécanismes de l'apoptose (ou mort cellulaire programmée). Cette capacité peut également se retrouver dans la modulation de l'expression des signaux de vie ou de mort.
4. La résistance des cellules cancéreuses face à leur « destin » se retrouve aussi avec le phénomène d'immortalisation : elles manifestent une réplication quasiment infinie, et ne rentrent pas en phase de vieillissement (ou sénescence). Cela est dû principalement à l'expression d'une enzyme, la télomérase, qui empêche l'érosion des télomères aux extrémités des chromosomes.
5. Une autre caractéristique répandue dans les cancers implique la capacité des cellules tumorales à induire l'angiogénèse, c'est-à-dire le développement d'une vascularisation destinée à alimenter ces cellules en nutriments ou en oxygène.
6. Enfin, la majorité des types de cancers ont pour caractéristique commune le développement de métastases, c'est-à-dire des croissances de cellules tumorales à distance de la tumeur primaire.

Malgré la mise en évidence de caractéristiques communes à l'ensemble des cancers, il existe quelques spécificités selon les types de cancer :

1. Plusieurs mutations génétiques différentes peuvent être à l'origine de plusieurs types de cancers différents : les mutations des gènes BRCA sont impliquées dans le cancer du sein, tandis que celle du gène de la polypose adénomateuse colique (APC) est associée avec le cancer colorectal.

2. Des aberrations chromosomiques distinctes peuvent induire la survenue de différents cancers, avec un effet spécifique sur la fonction des gènes : une délétion peut être associée avec une perte de fonction (ex : inactivation du gène Rb dans le rétinoblastome), tandis qu'une amplification de gène conduit à une surexpression (ex : le gène HER2 dans le cancer du sein).
3. Des variations dans les marques épigénétiques peuvent aussi être associées avec la survenue de cancers : ainsi la méthylation de certains gènes est préférentiellement associée avec un type de tumeur (ex : la méthylation des gènes BRCA, et le cancer du sein ou de l'ovaire).

2.1 La méthylation dans le cancer

Parmi les différents marqueurs épigénétiques étudiés, la méthylation de l'ADN est le plus répandu parmi les modifications aberrantes dans les cancers humains. La méthylation aberrante de l'ADN est impliquée dans de nombreux processus cellulaires : elle est retrouvée dans les composants majeurs de la division cellulaire, la réparation de l'ADN, ainsi que dans les voies de signalisation associées au cancer, telles que Wnt, TGF-β et NF-κB.

De la même manière, la déméthylation de l'ADN peut également être associée avec l'induction de certains cancers : ainsi, la suppression de la marque épigénétique dans l'ADN peut induire l'expression de certains gènes spécifiques aux lignées germinales, ou certains gènes présents dans le placenta, de façon ectopique, c'est-à-dire à une position anormale ou inhabituelle [13].

2.2 Importance de l'étude de la variabilité : aspects biostatistiques

Le monde du vivant est régi par le jeu des lois déterministes (en lien avec le principe de causalité) et de la variabilité de l'aléatoire. Ainsi, l'ensemble des processus biologiques sont soumis aux effets dits stochastiques (ou aléatoires). Il en résulte une variabilité dans les produits de ces processus biologiques, que l'on peut nommer bruit biologique. À l'échelle génomique, les niveaux d'expression des gènes peuvent alors varier entre différentes cellules, même lorsqu'elles possèdent le même matériel génétique, et sont soumises aux mêmes conditions environnementales [14].

En résumé, le lien entre génotype et phénotype n'est pas strictement déterministe, et des petites variations stochastiques à l'échelle moléculaire peuvent avoir un impact sur la probabilité d'apparition d'une phénotype et le devenir d'une cellule (par exemple l'engagement dans un processus tumoral).

Nous émettons alors l'hypothèse suivante : des régions présentant une forte variabilité du niveau de méthylation de l'ADN seraient plus facilement susceptibles d'être modifiées de façon significative dans le cadre du cancer. Jusqu'à présent, la variabilité du niveau de méthylation d'un CpG donné, en environnement fixé et stable, n'a pas réellement fait l'objet d'études scientifiques. Nous proposons dans ce projet de développer une méthode bioinformatique permettant de caractériser la variabilité de la méthylation dans le contexte de lignées cellulaires en culture. Cette méthode est basée sur une approche biostatistique de l'analyse des données de méthylation, qui permet de résoudre la problématique de l'auto-corrélation dans les jeux de données sur l'ensemble du génome.

13. Rousseaux, Sophie, Alexandra Debernardi, Baptiste Jacquiau, Anne-Laure Vitte, Aurélien Vesin, Hélène Nagy-Mignotte, Denis Moro-Sibilot, et al. « Ectopic Activation of Germline and Placental Genes Identifies Aggressive Metastasis-Prone Lung Cancers ». Science translational medicine 5, n° 186 (22 mai 2013) : 186ra66.
14. Tsimring, Lev S. « Noise in Biology ». Reports on progress in physics. Physical Society (Great Britain) 77, n° 2 (février 2014) : 026601.

3 Méthodologie pour l'étude de la variabilité de la méthylation de l'ADN

3.1 Stratégies à mettre en œuvre

Pour apporter une réponse à cette problématique, Magali Richard, en collaboration avec Florent Chuffart, propose d'utiliser un « pipeline » précédemment exploité pour analyser l'association entre la méthylation différentielle de l'ADN placentaire et la consommation de tabac des mères lors de la grossesse (Rousseaux et al. 2020). Cette analyse est décomposée en 3 étapes :

1. En premier lieu, on évalue l'hétérogénéité des cellules tumorales au niveau de la méthylation
2. Ensuite, on réalise un étude d'association à l'échelle de l'épigénome, ou EWAS (Epigenome-Wide Association Study)
3. Enfin, on détermine les régions considérées comme Différentiellement Méthylées (ou DMR : Differentially Methylated Regions).

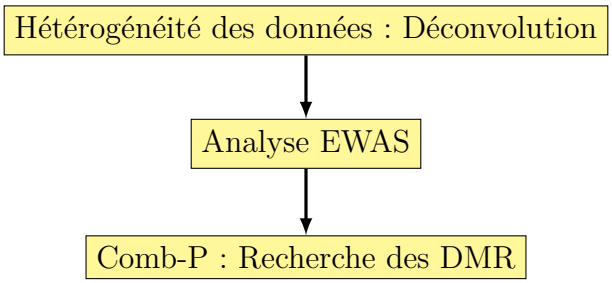

Figure 6 – Figure test

3.1.1 Hétérogénéité cellulaire

Pour cette première étape, les données de méthylation issues des puces sont traitées en utilisant une méthode ne nécessitant pas de profil de méthylation de référence, intitulée RefFreeEWAS, appliquée dans le cadre des EWAS.

Cette méthode est notamment utilisable en tant que module dans le langage R. Il s'agit d'une méthode de déconvolution des données qui se base sur l'identification de variables latentes en tant que substituts du mélange des types cellulaires.

Cette étape est particulièrement importante dans l'étude de la méthylation de l'ADN : cette marque épigénétique joue un rôle essentiel dans la différentiation des cellules pendant le développement. En effet, la méthylation de l'ADN dans les cellules différenciées permet entre autres d'inhiber l'expression de gènes qui ne sont pas nécessaires à la spécification de ces cellules : on peut ainsi dire que la méthylation de l'ADN est très spécifique du type de cellule et de tissu.

Cette caractéristique pose cependant un problème dans l'analyse biologique : en comparant différents phénotypes, les variations dans la composition des types de cellules peuvent fausser les résultats. Ce biais se traduit alors par un fort taux de faux positifs dans les données de méthylation : certaines variables sont supposées importantes dans la survenue d'un phénotype, alors qu'elles n'y contribuent pas ; à l'inverse, certaines variations dans les types de cellules, censées être sans rapport avec le phénotype, peuvent masquer des associations potentielles [15].

15. Teschendorff, Andrew E, et Shijie C Zheng. « Cell-Type Deconvolution in Epigenome-Wide Association Studies : A Review and Recommendations ». Epigenomics 9, n° 5 (mai 2017) : 757-68

Une des limitations de l'analyse de l'hétérogénéité cellulaire repose sur le manque de profils de méthylation de référence pour certains types de tissus.

Pour y remédier, RefFreeEWAS part du principe suivant, illustré dans la figure 7.

On définit la matrice Y pour les données de méthylation observées ; la matrice M représente l'ensemble des profils de méthylation connus pour un ensemble de types de cellules ; Ω^T est définie comme la matrice des proportions des types de cellules propre à chaque individu ; et X représente l'ensemble des métadonnées des phénotypes pour chaque individu.

Si les associations entre les matrices Y et X peuvent être expliquées par la décomposition suivante $Y = M\Omega^T$, on peut alors déduire que ces associations s'expliquent totalement ou en partie par les variations dans la distribution des types cellulaires.

Cette relation peut également être renforcée si les méthylomes répertoriés dans la matrice M peuvent correspondre à des processus biologiques qui permettent à leur tour de discerner des populations de cellules distinctes.

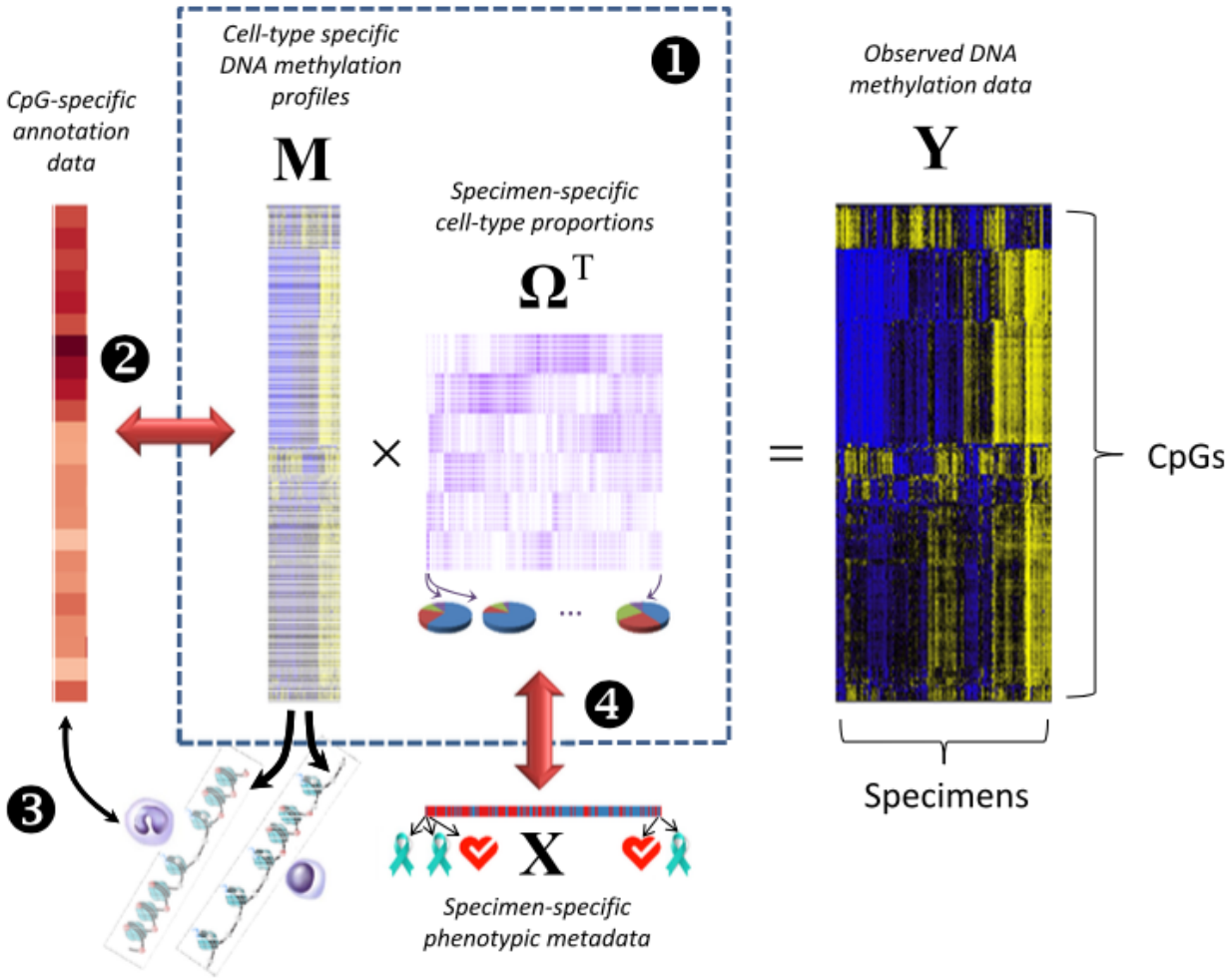

FIGURE 7 – Principe de la déconvolution des données de méthylation sans référence. Concernant les étapes de l'analyse : 1) Déconvolution des données. 2) Détermination des loci dits discriminants. 3) Analyse des ensembles de gènes. 4) Analyse des associations avec les phénotypes [16].

3.1.2 Modèle linéaire : EWAS

Dans un second temps, une analyse statistique est réalisée pour étudier l'association à l'échelle de l'ensemble de l'épigénome. Ce type d'étude, nommée EWAS, consiste à déterminer un lien entre les variations de niveau dans les marques épigénétiques (en l'occurrence ici la méthylation de l'ADN) et l'apparition de certains phénotypes. Dans le cadre de ce projet de recherche, il s'agit de trouver quelles variations dans la méthylation de l'ADN impactent la survenue de cancer dans la population. Cette analyse se fait notamment en comparant deux groupes d'individus : ceux ayant le phénotype d'intérêt (donc les patients atteints de cancer) et les cas témoins.

Pour procéder à l'analyse EWAS, chaque sonde mesurant le taux de méthylation de l'ADN est traitée de façon indépendante. Une première analyse statistique des données se fait à l'aide de tests d'association sur une seule variable (ou univariés), pour détecter les sites des îlots CpG où la méthylation de l'ADN varie avec le phénotype. Ensuite, la quantification de la méthylation est associée avec une correction sur des tests statistiques multiples, afin de limiter les facteurs dits confondants.

L'objectif de l'analyse EWAS est ainsi de définir un modèle statistique qui quantifie l'association entre chaque variable expliquée et le niveau de la méthylation de l'ADN dans chaque sonde.

Détermination des régions DMR : Adaptation de Comb-P

Enfin, à partir des résultats de l'analyse EWAS, une recherche des régions différentiellement méthylées (ou DMR) est réalisée, notamment grâce au programme Comb-P Pedersen et al. 2012. Il s'agit d'un outil en ligne de commande, associé à une libraire utilisable en langage Python, qui est utilisé dans le cadre de données qui sont auto-corrélées localement, comme par exemple dans les données de méthylation.

Le programme permet entre autres de prendre en compte la dépendance (ou corrélation) des sondes de méthylation le long du génome. Il ne traite donc pas chaque sonde de manière indépendante, comme dans le cadre de l'analyse EWAS précédente, mais il quantifie le degré de corrélation des sondes.

Le logiciel Comb-P prend ainsi en paramètre l'ensemble des P-values calculées, et corrige les corrélations entre ces probabilités en plusieurs étapes : Après avoir calculé l'auto-corrélation des P-values, ces dernières sont ajustées en fonction des valeurs adjacentes. Ensuite, un score de q-value est évalué par la méthode des fausses découvertes (ou FDR = False Discovery Rate), notamment par la correction de Benjamini-Hochberg. Enfin, le logiciel permet d'identifier les régions du génome considérées comme enrichies, à savoir les séries de faibles P-values adjacentes, et réattribue une P-value à ces régions pour leur apporter une significativité.

3.2 Détails concernant le logiciel Comb-P

La mise en oeuvre de Comb-P est composée de 4 étapes :

1. En premier lieu, la corrélation entre les valeurs P récupérées pour chaque sonde de méthylation est corrigée, en utilisant la méthode de l'auto-corrélation ;
2. Ensuite, les valeurs de probabilité pour chaque sonde sont ajustées en fonction des sondes voisines, en utilisant la méthode de Stouffer-Liptak-Kechris ;
3. Entre temps, on effectue une autre correction sur les valeurs P de base, pour déterminer par la suite des scores de probabilité ajustés par la méthode du taux de fausse découverte (ou FDR) ;

4. Enfin, les valeurs P ajustées (ou valeurs Q) permettent de déterminer les régions dites différentiellement méthylées. Puis, une nouvelle correction de Stouffer-Liptak permet de quantifier la significativité de ces régions, en attribuant une valeur P combinée à chaque région identifiée.

3.2.1 Auto-corrélation

La première étape dans le déroulement du programme Comb-P consiste à corriger la corrélation entre les différentes P-values, qui sont espacées de manière irrégulière, en raison de la répartition de la méthylation de l'ADN en îlots CpG.

Pour cela, Comb-P calcule l'auto-corrélation des P-values, par la méthode ACF. Cette méthode consiste à corréler une variable avec elle-même en utilisant un décalage de cette même variable au niveau du génome.

Alors que la plupart des méthodes d'implémentation de l'ACF se basent sur des décalages fixes pour les sondes adjacentes, le logiciel a pour paramètres une distance maximale et un seul décalage qui segmente cette distance en intervalles. Ainsi, si une sonde donnée est éloignée de 2 ou plusieurs sondes à 40 bases de distance, alors elle sera visible plusieurs fois dans l'ensemble des paires de sondes séparées entre elles de moins de 40 bases. La figure 8 décrit le principe de la méthode ACF.

Cette méthode s'applique bien aux données issues des puces de méthylation puisque les sondes ne sont pas distribuées de manière uniforme le long du génome.

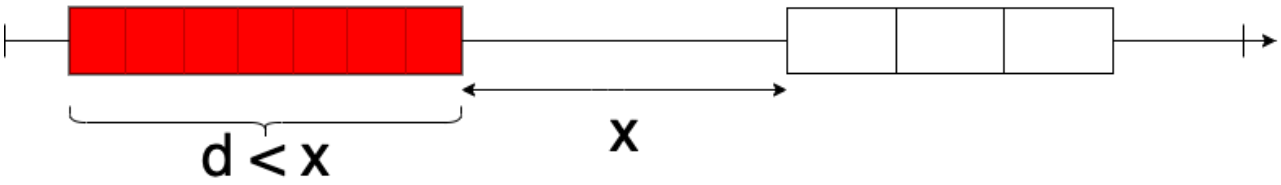

FIGURE 8 – **Schéma expliquant le principe de l'auto-corrélation**. Chaque section représente une sonde de méthylation le long du génome. La région colorée en rouge représente un ensemble de sondes auto-corrélées entre elles : la distance entre les différentes sondes, notée d est inférieure à la distance que sépare la dernière sonde avec un autre ensemble de sondes (notée x).

Il est à noter ici qu'étant donné les paramètres utilisés pour le calcul de l'ACF, cette auto-corrélation est limitée : la distance maximale pour la correction est spécifiée par l'utilisateur, mais cette distance peut ne pas être suffisante pour couvrir des régions de méthylation plus grandes.

3.2.2 Correction des P-values

Une fois l'ACF calculée, une correction des P-values est réalisée par la méthode de Stouffer-Liptak-Kechris (ou *slk*). Cette méthode s'opère en deux étapes principales.

1. Tout d'abord, on applique la correction de Stouffer-Liptak sur l'ensemble des valeurs P calculées. Cette première étape consiste à convertir ces probabilités en scores standards (ou Scores Z), puis à les combiner pour aboutir à un seul score de déviation standard. La conversion des valeurs P se base sur l'hypothèse qu'elles suivent une distribution selon la loi normale.

   Cependant, la méthode est utilisée dans le cas où les variables étudiées font partie des variables indépendantes et identiquement distribuées.

2. Dans les données des sondes de méthylation, chaque valeur est dépendante des valeurs voisines. Ainsi, les tests statistiques doivent être ajustés en prenant en compte la corrélation entre les différentes valeurs.

   Pour cela, l'ensemble des scores de déviation standard corrélés (noté Q) est transformé en un ensemble de scores Z indépendants (noté Q*), pour pouvoir être à nouveau recalculé par la correction de Stouffer-Liptak.

   Cette transformation se base sur le processus suivant : les valeurs P sont supposées être corrélées selon une matrice de corrélation nommée Σ. En supposant que cette matrice soit définie positive, la factorisation de Cholesky définit alors une matrice triangulaire C telle que Σ= CC' (où C' est la matrice transposée de C).

   En appliquant la transformation suivante : $Q* = C^{-1}Q$ (où C^{-1} représente la matrice inverse de C), l'ensemble des scores Z indépendants suit une distribution normale, et peut donc être traité par la méthode de Stouffer-Liptak [17].

Il s'agit alors d'une méthode dite « de moyennes mobiles » (ou *moving averages*), qui consiste à calculer une fenêtre coulissante où chaque P-value est ajustée en rapport avec les valeurs adjacentes. Ainsi, une sonde donnée peut voir sa P-value être ajustée vers le niveau inférieur si les sondes voisines ont également une P-value faible ; cependant, cette P-value reste insignifiante si les P-values adjacentes sont élevées.

Cette méthode permet ainsi une correction des P-valeurs indépendamment du test statistique utilisé pour générer ces valeurs.

3.2.3 Q-value : méthode de FDR

Après correction des P-values pour chaque sonde, une autre méthode de correction est réalisée sur les P-values d'origine, et permet de calculer un score appelé q-value, basé sur la correction de type FDR (*False Discovery Rate*), qui s'applique notamment dans le cas des tests de comparaison multiples.

Cette correction peut être réalisée selon la méthode traditionnelle de Benjamini-Hochberg, où l'hypothèse nulle posée est une distribution uniforme des p-values, ou alors en spécifiant une distribution dite nulle, notamment en réalisant un brassage des données cliniques issues des données des sondes avant de générer les p-values à traiter.

Elle permet de recalculer les p-values en mettant en relation le taux de faux positifs possibles dans les tests statistiques.

17. Kechris, Katerina J, Brian Biehs, et Thomas B Kornberg. « Generalizing Moving Averages for Tiling Arrays Using Combined P-Value Statistics ». Statistical Applications in Genetics and Molecular Biology 9, n° 1 (6 janvier 2010)

3.2.4 Détermination des régions enrichies

A partir des différents scores de q-value obtenus par la correction de FDR, un algorithme permet de déterminer des régions dites d'enrichissement. Elles sont définies par des regroupements de sondes dont l'ensemble des q-values calculées par FDR sont faibles.

Par la suite, la significativité de ces régions est déterminée par la correction de Stouffer-Liptak, qui assigne à chaque région une nouvelle P-value dite combinée.

Cette étape ne se base pas sur l'auto-corrélation calculée précédemment. En effet, l'ACF de la première étape est réalisée sur une distance maximale entre les sondes déterminée par l'utilisateur. Or cette distance peut ne pas couvrir certaines régions du génome (car elles sont plus grandes).

À l'inverse, une nouvelle auto-corrélation est calculée par ACF à partir d'une distance équivalente à la longueur de la région la plus longue, d'où la nécessité de réaliser une correction par FDR des p-values, afin de déterminer les régions différentiellement méthylées.

Ainsi, la p-value combinée est déterminée pour chaque région en utilisant les P-values d'origine qui se situeraient entre les régions (ou *peaks*) et l'ACF totale. Cela permet ainsi d'éviter les problèmes liés à la modification de la distribution des valeurs lors des étapes de correction par slk ou par FDR.

Pour résumer le déroulé du programme, la correction des p-values par FDR permet de définir les limites des régions DMR, puis on définit l'importance et la significativité des régions en utilisant la correction slk sur les p-values d'origine.

Conclusion

La recherche bibliographique que j'ai effectuée portait sur la méthylation de l'ADN, et plus particulièrement sur la variabilité de cette marque épigénétique. Elle s'est notamment focalisée sur l'impact de cette variabilité, ou bruit biologique, dans le cadre du cancer.

La problématique soulevée par ce projet de recherche concernait l'élaboration de méthodes d'analyse de données de méthylation capables de prendre en compte l'aspect stochastique expliquant les régions exprimant de manière différentielle cette méthylation.

Pour apporter une solution à cette problématique, de nouvelles méthodes biostatistiques peuvent être imaginées, compte tenu d'une part de la présence de grands jeux de données disponibles publiquement (*e.g* TCGA), et dont dispose l'équipe de recherche de M. Richard (Meth Epic, RRBS), et d'autre part du développement de nouveaux outils d'analyse bioinformatique capables d'extraire différentes informations à partir de ces jeux de données.

Le pipeline proposé par l'équipe de recherche permettrait d'apporter une réponse, notamment en modifiant les analyses statistiques sur les données de méthylation, qui génèrent des valeurs de probabilité auto-corrélées le long du génome, mais de manière irrégulière.

Références bibliographiques

[1] Peter A. Jones. “Functions of DNA methylation : islands, start sites, gene bodies and beyond”. en. In : *Nat Rev Genet* 13.7 (juil. 2012), p. 484-492. issn : 1471-0056, 1471-0064. doi : 10.1038/nrg3230. url : http://www.nature.com/articles/nrg3230 (visité le 06/11/2020).

[2] B. S. Pedersen et al. “Comb-p : software for combining, analyzing, grouping and correcting spatially correlated P-values”. en. In : *Bioinformatics* 28.22 (nov. 2012), p. 2986-2988. issn : 1367-4803, 1460-2059. doi : 10.1093/bioinformatics/bts545. url : https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts545 (visité le 14/12/2020).

[3] Sophie Rousseaux et al. “Immediate and durable effects of maternal tobacco consumption alter placental DNA methylation in enhancer and imprinted gene-containing regions”. In : *BMC Medicine* 18.1 (oct. 2020), p. 306. issn : 1741-7015. doi : 10.1186/s12916-020-01736-1. url : https://doi.org/10.1186/s12916-020-01736-1 (visité le 09/10/2020).