Biomarqueurs Pronostiques dans le Cancer du Côlon

Lancelot Ravier, Jade Bordet, Manal Belouarda, Sara Mekkaoui

Université Grenoble Alpes (UGA), IM2AG, Master SSD

.

Encadrantes : Ekaterina Bourova-Flin, Séverine Valmary-Degano
Encadrement : IAB Grenoble

.

Professeurs : Adeline Leclerq-Samson, Jean-François Coeurjolly

Abstract

Le cancer du côlon est l'un des cancers les plus fréquents dans le monde. Récemment, l'équipe "Régulations épigénétiques" de l'IAB a mis en évidence une signature transcriptomique basée sur l'activation aberrante de gènes spécifiques de tissus, appelée GEC (Gene Expression Classifier), capable de prédire la survie des patients. Dans ce projet, nous avons évalué la robustesse de cette signature à travers trois jeux de données transcriptomiques (TCGA-COAD, GSE39582 et GSE17536) et testé une version simplifiée à trois gènes (3-GEC) utilisable en immunohistochimie. Par des analyses de survie (Kaplan-Meier, test du log-rank, modèle de Cox), nous avons confirmé l'impact pronostique du score GEC, même dans sa version simplifiée. En parallèle, nous avons reconstruit le statut MSI/MSS des tumeurs à partir de données multi-omiques, puis développé des modèles de machine learning (XGBoost, SVM, et autres) pour le prédire. Enfin, nous avons montré que le GEC reste un marqueur pronostique significatif, indépendamment du statut MSI/MSS. Ces résultats soulignent l'intérêt d'intégrer le GEC dans la stratification clinique des patients atteints de cancer du côlon. Ce projet propose ainsi un outil utile, accessible et potentiellement transposable à la pratique hospitalière.

Introduction

L'amélioration de la prise en charge des cancers passe aujourd'hui par une compréhension précise des mécanismes moléculaires impliqués dans le développement des tumeurs. Dans ce contexte, la recherche de biomarqueurs pronostiques fiables est essentielle pour adapter les traitements aux profils biologiques de chaque patient. Le cancer du côlon, en particulier aux stades II et IV soulèvent des questions cliniques quant à la nécessité - ou non - d'un traitement post-chirurgie. Or, les critères actuellement utilisés (fondés sur des paramètres cliniques et histologiques) présentent des limites et ne permettent pas toujours de prédire avec précision l'évolution de la tumeur cancéreuse.

Les avancées récentes en épigénétique ont permis de découvrir de nouveaux mécanismes impliqués dans le développement de la tumeur, comme l'activation ectopique de gènes spécifiques de tissus. Ces gènes, normalement silencieux dans les cellules somatiques, peuvent être activées de manière anormale dans les cellules tumorales, contribuant ainsi à leur prolifération et à leur agressivité. Ce phénomène, observé dans plusieurs types de cancer dont le cancer colorectal, est porteur d'un fort potentiel dans les outils de classification pronostique.

Parmi les signatures développées, le Gene Expression Classifier (GEC), élaboré récemment par l'équipe "Régulations épigénétiques" de l'IAB, repose sur l'expression aberrante de quatre gènes identifiés comme fortement corrélés au pronostic des patients atteints de cancer du côlon. L'intérêt de cette signature réside dans sa capacité à prédire la survie à partir d'un simple score, permettant ainsi une utilisation clinique. Cependant, pour permettre son utilisation par les services hospitaliers, une version simplifiée de ce score doit être envisagée, en tenant compte de la disponibilité des outils de détection (notamment les anticorps pour l'immunohistochimie pour le gène ERFE).

Parallèlement, un autre critère pronostique moléculaire important dans le cancer colorectal est le statut MSI/MSS (Microsatellite Instable/Stabilité des Microsatellites), déjà utilisé pour guider certaines décisions thérapeutiques, notamment en immunothérapie. Un enjeu majeur est donc de déterminer si le score GEC apporte une information complémentaire ou redondante par rapport au statut MSI/MSS, afin de préciser son intérêt clinique.

Ce projet s'inscrit dans cette double problématique : valider l'utilité pronostique de la signature GEC dans le cancer du côlon et évaluer son indépendance vis-à-vis du statut MSI/MSS, en s'appuyant sur des méthodes de biostatistique et de machine learning appliquées à des données transcriptomiques et multi-omiques.

Présentation de l'IAB

L'Institut pour l'Avancée des Biosciences (IAB) est un institut de recherche biomédicale fondamentale et translationnelle situé à Grenoble employant environ 300 personnes. L'IAB mène des projets de recherche pour mieux comprendre les mécanismes épigénétiques responsables du développement et de la progression des cancers pour découvrir de nouvelles approches thérapeutiques ciblées. Il est constitué de 3 départements dans lesquels évoluent 19 équipes. L'Institut dépend de trois tutelles de la fonction publique : l'Institut National de la Santé et de la Recherche Médicale (INSERM), le Centre National de la Recherche Scientifique (CNRS) et l'Université Grenoble Alpes (UGA). Il a également développé une forte interaction avec le Centre Hospitalier Universitaire Grenoble Alpes (CHUGA) et le CLARA (Cancéropôle Lyon Rhône-Alpes Auvergne). L'IAB est dirigé par le Pr. Christophe Arnoult. L'IAB développe plusieurs programmes de recherche, fondamentale et translationnelle, dans les domaines suivants : l'épigénétique, l'environnement cellulaire et la signalisation, la reproduction, l'immunologie, l'épidémiologie environnementale et le cancer. La recherche à l'IAB s'organise autour de trois axes principaux : la signalisation par la chromatine, la plasticité cellulaire, le microenvironnement et la signalisation et la prévention et le traitement des maladies chroniques.

Notre projet tutoré a été réalisé en collaboration avec l'équipe de recherche "Régulations épigénétiques" de l'IAB et le Centre Hospitalier Universitaire Grenoble Alpes (CHUGA). Il a été co-encadré par Ekaterina Flin, ingénieur de recherche de l'équipe "Régulations épigénétiques" de l'IAB et par le Pr. Séverine Valmary-Degano, pathologiste dans le service d'anatomie et cytologie pathologiques de l'institut de Biologie et de Pathologie du CHUGA.

Présentation du projet

Contexte scientifique

Le cancer du colon Le cancer colorectal (regroupant les cancers du côlon et du rectum) est l'un des cancers les plus fréquents en France et dans le monde. En 2023, on estime à 47582 le nombre de nouveaux cas diagnostiqués, avec une répartition de 26212 cas chez les hommes et 21370 cas chez les femmes. L'âge médian au moment du diagnostic est de 71 ans chez les hommes et 72 ans chez les femmes. Concernant la mortalité, 17100 décès ont été enregistrés en 2018, répartis entre 9200 hommes et 7900 femmes, avec un âge médian de décès de 77 ans chez les hommes et 81 ans chez les femmes. Entre 2010 et 2018, une diminution annuelle de moyenne du taux de mortalité a été observé : -1.8% chez les hommes et -1.6% chez les femmes [**cancer_colorectal**].

Le cancer du côlon peut présenter des anomalies du système de réparation de l'ADN - ou système MMR (pour MisMatch Repair) - qui sont alors responsables d'une instabilité des microsatellites (séquences répétitives de courtes unités de bases d'ADN, généralement 1 à 6 paires de bases entre 2 gènes), par absence de réparation des erreurs de réplication de l'ADN à cet endroit. Les tumeurs peuvent alors alors définies selon le statut stable ou instable des microsatellites reflétant ainsi l'intégrité ou non du système de réparation. On distingue deux catégories :

- Stabilité des microsatellites (MSS) ou statut pMMR (pour proficient MMR) : La majorité des cancers colorectaux, soit environ 85%, sont caractérisés par une stabilité des microsatellites. Ainsi, ces tumeurs sont généralement moins sensibles aux immunothérapies [**RubyGupta2018**] et peuvent nécessiter des approches thérapeutiques différentes [**KYamaguchi2025**].
- Instabilité élevée des microsatellites (MSI) ou statut dMMR (pour deficient MMR) : Environ 15% des tumeurs colorectales présentent une instabilité élevée des microsatellites. Ce phénomène est alors associé à une meilleure réponse à certaines immunothérapies et à un pronostic généralement plus favorable. [**EduardoVilar2010**]

Ainsi, la distinction entre MSI et MSS est cruciale pour orienter les choix des médecins dans la prise en charge des patients atteints du cancer colorectal. Le statut MSI est accordé lorsque un des 3 cas suivant est avéré :

- Méthylation du gène MLH1 : Si la région promotrice du gène MLH1 est méthylée et son expression est faible La méthylation de l'ADN est un processus normal impliqué dans le contrôle de l'expression des gènes. Cependant, dans de nombreux cancers, une hyperméthylation anormale des îlots CpG situés dans les promoteurs de certains gènes peut entraîner leur silençage transcriptionnel. C'est le cas du gène MLH1, un acteur essentiel du système de réparation des mésappariements de l'ADN [**Kane2008**]. Son inactivation par hyperméthylation du promoteur est fréquemment observée dans les cancers colorectaux sporadiques présentant une instabilité des microsatellites (MSI)[**Xia2013**]. L'absence d'expression fonctionnelle de MLH1 conduit à une accumulation d'erreurs de réplication au

niveau des séquences microsatellites, entraînant une instabilité génomique caractéristique des tumeurs MSI [**Kane2008**]. Ce phénomène favorise la mutation d'autres gènes clés impliqués dans le contrôle du cycle cellulaire et l'apoptose, facilitant ainsi la progression tumorale [**Xia2013**]. Contrairement aux formes héréditaires du syndrome de Lynch, où des mutations germinales des gènes de réparation sont impliquées, l'inactivation de hMLH1 dans les cancers sporadiques est principalement d'origine épigénétique [**Kane2008**]. L'identification des échantillons tumoraux présentant une forte méthylation et une faible expression des gènes de réparation est essentielle pour classifier les cancers du colon selon leur statut MSI. Cette classification permet non seulement d'affiner le diagnostic, mais aussi d'orienter les décisions thérapeutiques, car les tumeurs MSI sont connues pour répondre différemment aux traitements chimiothérapeutiques et sont associées à un meilleur pronostic dans certains contextes.

- Mutations des gènes MLH1, MSH2, MSH6, PMS2 : Si l'un de ces gènes présente une mutation de type "perte de fonction" (LOF - Loss of Function). Les mutations des gènes MLH1, MSH2, MSH6 et PMS2, qui font partie du système de réparation des erreurs de l'ADN (réparation par mésappariement de l'ADN), sont des facteurs importants notamment dans le cadre du syndrome de Lynch (ou cancer colorectal héréditaire non polyposique). Lorsque l'un de ces gènes subit une mutation de type "perte de fonction" (LOF), la capacité de réparation des erreurs de l'ADN est compromise, conduisant à une accumulation de mutations dans le génome des cellules. Ces mutations perturbent le mécanisme de correction des erreurs dans la séquence de l'ADN lors de la réplication cellulaire, entraînant une instabilité microsatellite (MSI. Détecter ces mutations peutt aider à identifier les individus à risque élevé de développer un cancer du côlon, offrant ainsi la possibilité de surveiller et de traiter ces patients de manière précoce et plus ciblée.
- Perte de gènes par altération du nombre de copies (CNA) : Si une région entière de l'ADN contenant l'un de ces gènes est perdue. La perte de gènes par altération du nombre de copies (CNA - Copy Number Alterations) est un phénomène où une région entière de l'ADN, contenant un ou plusieurs gènes, est perdue ou dupliquée. Dans le contexte du cancer du côlon, la perte d'une région génétique peut entraîner l'élimination de gènes importants pour la régulation de la croissance cellulaire, la réparation de l'ADN ou la suppression tumorale, ce qui contribue à la progression tumorale. Lorsqu'une région contenant des gènes tumor-suppresseurs ou des gènes impliqués dans la réparation de l'ADN est perdue, cela peut altérer les mécanismes de contrôle de la cellule, favorisant l'apparition de mutations et la croissance incontrôlée des cellules cancéreuses.

Activation aberrante des gènes spécifiques de tissus dans le cancer Lors du processus oncogénique la cellule subit des dérégulations majeures au niveau du génome et de l'épigénome qui perturbent les mécanismes d'expression des gènes, c'est-à-dire la manière dont les gènes sont activés ou désactivés dans la cellule. Certaines dérégulations génétiques, par exemple, des mutations des séquences d'ADN, peuvent activer des oncogènes (gènes favorisant la division cellulaire) et d'autres peuvent désactiver des gènes suppresseurs de tumeurs (gènes inhibant la croissance cellulaire). Ces événements contribuent à l'apparition d'un cancer. Les dérégulations épigénétiques, quant à elles, incluent des altérations des différentes marques chimiques qui sont capables de modifier l'activité des gènes sans affecter la séquence d'ADN. Ces modifications dites épigénétiques jouent également un rôle important dans la régulation d'expression des gènes et peuvent aussi contribuer au développement du cancer.

Les travaux de recherche menés depuis plus d'une décennie dans l'équipe "Régulations épigénétiques" à l'Institut pour l'Avancées des Biosciences (IAB) ont démontré que les dérégulations épigénétiques dans les cellules tumorales aboutissent à une activation aberrante d'un grand nombre de gènes qui doivent normalement rester silencieux dans une cellule saine. Il s'agit des gènes spécifiques de tissus, en particulier, des gènes spécifiques de la lignée germinale mâle, du placenta et des cellules embryonnaires souches. Ces gènes sont normalement exprimés de la façon prédominante dans un tissu particulier et ne sont pas exprimés ou faiblement exprimés dans d'autres tissus somatiques adultes. Dans les cancers, les perturbations épigénétiques déclenchent une activation aberrante de ces gènes normalement silencieux. Ce phénomène est appelé expression ectopique ou expression hors-contexte des gènes. Il survient massivement dans tous les types de cancers.

L'activation ectopique de certains gènes dans les tumeurs est significativement associée à un pronostic vital défavorable des patients. Ainsi, ces gènes représentent des biomarqueurs pronostiques potentiels dont le statut d'activation (OFF/ON) de la tumeur permet de prédire la probabilité de survie de chaque patient individuellement. L'équipe "Régu-

lations épigénétiques" de l'IAB a développé une approche bioinformatique originale, nommée "ectopy", pour découvrir systématiquement de nouveaux biomarqueurs pronostiques à partir des données d'expression de gènes, en se basant sur le phénomène d'expressions ectopiques. Cette stratégie a conduit à la découverte de biomarqueurs pronostiques dans plusieurs pathologies malignes, en particulier, dans le cancer du poumon [**Rousseaux2013**] [**LeBescont2015**], dans le lymphome [**Emadali2013**], dans des leucémies aiguës lymphoblastiques [**Wang2015**] [**Peng2022**], dans le cancer oropharyngé [**BourovaFlin2021**] et dans le cancer du sein [**Jacquet2023**].

Signature pronostique GEC basée sur une activation aberrante de gènes dans le cancer du côlon Récemment, la méthode "ectopy" a été appliquée aux données d'expression de gènes du cancer du côlon. Les chercheurs ont identifié un panel de quatre gènes, ERFE, HOXC6, LAMP et ULBP2, dont l'activation aberrante est significativement associée à un pronostic défavorable des patients [**Spinelli2024**]. Cette signature est appelée Gene Expression Classifier ou GEC.

Le classifier GEC stratifie les patients selon le nombre de gènes activés dans le panel, de 0 à 4. Chaque gène du panel contribue à la prédiction du classifier avec le même poids. Les patients dont la tumeur n'exprime aucun gène du panel GEC ont un pronostic favorable. Les patients dont la tumeur exprime un ou plusieurs gènes du panel ont un pronostic plus sombre. Plus grand est le nombre de gènes activés (statut ON), moins longue sera la probabilité de survie des patients. A titre d'exemple, les courbes de survie obtenues en fonction du score GEC dans trois jeux de données du cancer du côlon sont présentées dans la Figure 1 1.

Les chercheurs de l'IAB et du CHUGA souhaitent réaliser une validation expérimentale sur une nouvelle cohorte rétrospective de 140 patients du CHUGA à l'aide de la technologie RT-qPCR ou en immunohistochimie. Cette validation permettra d'utiliser l'outil pronostique GEC en routine clinique hospitalière pour mieux guider la prise en charge des patients atteints du cancer du côlon. Dans ce but, des analyses statistiques supplémentaires sont nécessaires et feront l'objet principal de ce projet tutoré.

Objectifs du projet

Le panel initial GEC est composé de quatre gènes (4-GEC) : ERFE, HOXC6, LAMP5 et ULBP2. Cependant, dans le but de développer un test pronostique simplifié utilisable en routine hospitalière, il est nécessaire de s'assurer que les anticorps spécifiques aux gènes de la signature GEC sont disponibles pour une utilisation en immunohistochimie. Or, les anticorps industriels existent uniquement pour trois des quatre gènes : HOXC6, ULBP2 et LAMP5.

La question posée est donc la suivante : peut-on réduire la signature pronostique 4-GEC à 3-GEC tout en conservant un impact significatif sur la survie des patients ? Ainsi, le premier objectif du projet consiste à comparer la signature GEC à 4 gènes avec celle à 3 gènes et à évaluer leurs impacts pronostiques respectifs.

Le deuxième objectif du projet vise à étudier la relation entre la signature GEC et le statut MSI ou MSS du cancer du côlon. Nous souhaitons notamment savoir si le test pronostique GEC apporte une information indépendante par rapport au statut MSI/MSS des tumeurs. Cette information est essentielle pour décider si le futur test clinique GEC pourrait être applicable à tous les patients atteints du cancer du côlon indépendamment de leur statut MSI/MSS, être utile uniquement dans un de ces groupes (MSI ou MSS) ou n'a pas d'intérêt particulier si le statut MSI/MSS est déjà connu.

Il se trouve que le statut MSI/MSS n'est pas directement disponible dans les annotations bioclniques des jeux de données du cancer du côlon dont nous disposons. En revanche, pour une partie d'échantillons tumoraux, le statut MSI/MSS peut être établi à partir d'une analyse croisée des données moléculaires disponibles, en particulier, des données d'expression de gènes, de mutation, de méthylation et d'altération de nombre copies (en anglais, copy number alterations ou CNA). En utilisant ces échantillons, nous pouvons reconstruire les annotations MSI/MSI et ensuite créer un classifier en machine learning pour prédire les annotations MSI/MSS manquantes pour les autres échantillons.

Les étapes concrètes de cette approche seront les suivantes :

- Identifier le statut MSI/MSS du cancer du côlon dans le jeu de données TCGA-COAD à partir des données d'expression de gènes, de mutation, de méthylation et de CNA, dans les échantillons où ces données sont simultanément disponibles.
- En se basant sur les annotations MSI/MSS identifiées, développer des modèles de machine learning pour prédire le statut MSI/MSS des autres échantillons, à partir des données d'expression de gènes uniquement. Évaluer les performances de différents modèles.
- Réaliser une analyse pronostique de la signature GEC par rapport au statut MSI/MSS des échantillons tumoraux. Répondre à la question principale du deuxième objectif du projet : est-ce

que la signature GEC apporte une information pronostique indépendante du statut MSI/MSS ?

Organisation, planification et outils de travail

Le projet a été réalisé par un groupe de quatre étudiants, ayant contribué de manière équitable à l'ensemble des tâches. Afin de garantir une bonne coordination, nous avons mis en place une organisation rigoureuse dès le début du projet. Des réunions régulières de suivi ont été tenues après avoir accompli au moins une tâche principale du cahier des charges qui nous a été confiée, pour faire le point sur l'avancement, ajuster les objectifs si nécessaire, et répartir les tâches pour les semaines suivantes.

Pour la gestion des tâches et le suivi collaboratif, nous avons utilisé l'outil en ligne Trello, en organisant notre travail selon les grandes étapes du projet : préparation des données, analyses de survie, modélisation MSI/MSS, interprétation des résultats et rédaction du rapport.

Le développement informatique a été réalisé principalement avec Visual Studio Code (VSCode) en utilisant le langage Python. Parmi les principaux packages utilisés, on retrouve :

- Pandas et NumPy pour la manipulation de données,
- Matplotlib et Seaborn pour la visualisation,
- Scikit-learn pour les modèles de machine learning (SVM, Random Forest, XGBoost, etc.),
- Lifelines pour l'analyse de survie (Kaplan-Meier, Cox),
- SciPy et statsmodels pour les tests statistiques.

Cette organisation nous a permis d'avancer efficacement tout au long du projet et de maintenir une bonne cohérence dans les analyses et la rédaction finale.

Matériels et methodes

Jeux de données

Dans ce projet, nous avons utilisé trois jeux de données du cancer du côlon, TCGA-COAD, GSE39582 et GSE17536, issus des bases de données publiques GDC Portal (https://portal.gdc.cancer.gov) et NCBI GEO (https://www.ncbi.nlm.nih.gov/geo). L'aperçu des différentes types de données disponibles pour chaque dataset est présenté dans le tableau ci-dessous :

Types de données	TCGA-COAD	GSE39582	GSE17536
Expression de gènes	Oui (RNA-seq*, 398)	Oui (puces**, 566)	Oui (puces**, 177)
Méthylation d'ADN	Oui	Non	Non
Mutations de gènes	Oui	Non	Non
CNA	Oui	Non	Non
Survie	Oui	Oui	Oui
Score GEC	Oui	Oui	Oui

Table 1: *Résumé des types de données disponibles*

* Technologie de séquençage de l'ARN
** Technologie de puces à ADN

Les données brutes issues des bases de données publiques ont été préalablement prétraitées par les membres de la plateforme EpiMed.

Les données d'expression de gènes ont été alignées sur un génome de référence, normalisées et ensuite log-transformées. Les données de méthylation sont présentées en valeurs numériques de 0 à 1, correspondant aux taux de méthylation dans les régions promotrices de gènes (0: absence de méthylation, 1: méthylation totale).

Les données de mutations contiennent le nombre de mutations des gènes MLH1, MLH2, MSH6 et PMS2 dans chaque échantillons du cancer du côlon. De façon similaire, les données de CNA indiquent les échantillons du cancer du côlon dans lesquels il existe une perte d'une région génique entière pour les gènes MLH1, MLH2, MSH6 et PMS2.

Les données de survie sont codées sur deux valeurs numériques :

- la durée de la survie globale (en mois) de chaque patient après le diagnostic du cancer jusqu'à l'événement (le décès du patient) ou jusqu'à la date de la dernière nouvelle (habituellement, sa dernière consultation médicale).
- la "censure" : la valeur binaire indiquant si l'événement a été observé après la durée indiquée.

Enfin, le statut d'activation de chaque gène du panel GEC (ERFE, HOXC6, LAMP et ULBP2) était disponible individuellement pour tous les patients.

Les données de méthylation, de mutations et de CNA étaient présentes uniquement dans le jeu de données TCGA-COAD alors que les données d'expression de gènes, de survie et le score GEC étaient disponibles dans les trois jeux de données.

Préparation des données

Les fichiers bruts des annotations biocliniques n'étaient pas nettoyés. Ils contenaient divers types d'échantillons, incluant des tissus normaux et tumoraux, ainsi que des échantillons ne correspondant pas au côlon ou dépourvus de données d'expression. En appliquant des filtres sur le statut d'échantillon

(normal ou cancer), le type de tissu et la présence de l'échantillon dans les données d'expression de gènes, nous avons identifié les échantillons du cancer du côlon dans les trois jeux de données. Nous avons obtenu 398 échantillons du cancer du côlon dans le dataset TCGA-COAD, 566 échantillons dans le dataset GSE39582 et 177 échantillons dans le dataset GSE17536.

Dans le jeu de données TCGA-COAD, les données de méthylation étaient présentes pour 34 échantillons sur 398, les données de mutations pour 20 échantillons et les données de CNA pour 12 échantillons.

Modèles d'analyse de survie

L'analyse de survie est une méthode statistique utilisée pour étudier le temps qu'il faut avant qu'un événement d'intérêt se produise, tel que le décès, la récidive d'une maladie ou d'autres événements cliniques pertinents. Dans le contexte du cancer du côlon, ces méthodes permettent d'évaluer l'impact de différents facteurs cliniques et biologiques sur la survie des patients. Les modèles de survie sont essentiels pour identifier des sous-groupes de patients ayant des pronostics différents et pour guider les décisions thérapeutiques.

Cette section présente trois approches statistiques pour l'analyse de survie que nous avons utilisées au cours de ce projet : la courbe de Kaplan-Meier, le test de log-rank et le modèle de régression de Cox. Chacune de ces méthodes offre une manière différente d'explorer les données de survie et d'évaluer les facteurs influençant la survie des patients.

Courbe de Kaplan-Meier La courbe de Kaplan-Meier est une méthode non paramétrique utilisée pour estimer la fonction de survie à partir de données censurées. Elle permet de visualiser l'évolution de la probabilité de survie d'un groupe de patients au fil du temps, en tenant compte des individus qui n'ont pas encore connu l'événement d'intérêt (par exemple, un décès ou une récidive) au moment de l'analyse. Ce type de courbe est particulièrement utile dans les études où certains patients n'ont pas vécu suffisamment longtemps pour que l'événement se produise, ce qui entraîne des données censurées.

La courbe de Kaplan-Meier est tracée en représentant le temps sur l'axe des abscisses et la probabilité de survie sur l'axe des ordonnées. Chaque "saut" dans la courbe correspond à un événement observé (comme un décès), et l'amplitude de chaque saut est inversement proportionnelle au nombre d'individus à risque au moment de l'événement. Cette courbe permet de comparer facilement la survie entre plusieurs groupes (dans notre cas, nous pouvons comparer la survie entre les différents scores GEC) en observant les différences dans la forme des courbes.

Afin d'évaluer l'influence du niveau d'expression des gènes sur la probabilité de survie des patients, nous avons recours à deux méthodes statistiques en analyse de survie : le test de logrank et le modèle de Cox à risques proportionnels.

Modèle de Cox Nous avons également utilisé le modèle de Cox proportionnel pour estimer l'impact du niveau d'expression d'un gène (ou d'une covariable) sur la probabilité de survie. Le modèle s'écrit sous la forme :

$$h(t \mid Z^i) = h_0(t) \exp\left(\sum_{k=1}^{p} \beta_k Z_k^i\right) = h_0(t) e^{\theta^T Z}$$

où :

- $h(t \mid Z^i)$ est le taux de risque instantané pour l'individu i,
- $h_0(t)$ est la fonction de risque de base, indépendante des covariables,
- Z^i est le vecteur des covariables pour l'individu i,
- $\theta = (\beta_1, \ldots, \beta_p)$ est le vecteur des coefficients de régression inconnus.

Le hazard ratio (HR) associé à une covariable Z_j est défini par :

$$HR = e^{\beta_j}, \quad \forall j \in \{1, \ldots, p\}$$

L'interprétation est la suivante :

- Si $HR > 1$, la variable augmente le risque de décès (effet défavorable sur la survie),
- Si $HR < 1$, elle a un effet protecteur,
- Si $HR = 1$, elle n'a pas d'effet significatif.

Par exemple, un $HR > 1$ pour la variable *statut MSI/MSS* indique que ce statut a un effet significatif sur la survie.

Pour tester la significativité des coefficients β_j, on utilise le test de Wald. On teste l'hypothèse :

$$H_0 : \beta_j = 0 \quad \text{vs} \quad H_1 : \beta_j \neq 0$$

La statistique de test est :

$$W = \frac{\hat{\beta}_j^2}{\text{Var}(\hat{\beta}_j)}$$

Sous H_0, cette statistique suit asymptotiquement une loi du chi-deux à 1 degré de liberté :

$$W \sim \chi^2(1)$$

On rejette l'hypothèse nulle H_0 au seuil α si :

$$\mathbb{P}(\chi_1^2 \geq W) \leq \alpha$$

Test de Log-rank On s'intéresse à la comparaison des fonctions de survie des deux groupes A et B, notées respectivement $S_A(t)$ et $S_B(t)$. L'objectif est de tester l'hypothèse nulle suivante :

$$H_0 : S_A(t) = S_B(t) \quad \text{pour tout } t$$

$$H_1 : S_A(t) \neq S_B(t) \quad \text{pour au moins un } t$$

Dans ce contexte, où des données censurées peuvent être présentes, on utilise le test du log-rank pour comparer les deux groupes.

Sous H_0, la statistique de test est définie par :

$$T_n = \frac{U^2}{V(U)}$$

où

$$U = \sum_{i=1}^{k} d_{B,i} - \frac{R_{B,i}}{R_{A,i} + R_{B,i}}(d_{A,i} + d_{B,i})$$

avec :

- k : le nombre total d'instants de décès observés (événements),
- $d_{A,i}, d_{B,i}$: le nombre d'événements (décès) aux temps t_i dans les groupes A et B,
- $R_{A,i}, R_{B,i}$: le nombre de patients à risque juste avant t_i dans les groupes A et B.

Sous l'hypothèse nulle H_0, lorsque la taille de l'échantillon est grande, la statistique T_n suit asymptotiquement une loi du chi-deux à un degré de liberté :

$$T_n \xrightarrow{\mathcal{L}} \chi^2(1) \quad \text{quand } n \to \infty$$

On rejette l'hypothèse nulle H_0 au seuil α si :

$$\mathbb{P}_{H_0}(T_n > t_{n,\text{obs}}) \leq \alpha$$

Méthode d'identification du statut MSI/MSS du cancer du côlon à partir des données multi-omiques

L'objectif de cette analyse est d'identifier le statut MSI (Microsatellite Instability) ou MSS (Microsatellite Stable) de chaque échantillon tumoral du dataset TCGA-COAD-FPKM. Le statut MSI est déterminé par l'une des trois causes suivantes :

- Méthylation du gène MLH1 : si la région promotrice du gène MLH1 est méthylée et son expression est faible.
- Mutations des gènes MLH1, MSH2, MSH6, PMS2 : si l'un de ces gènes présente une mutation de type "perte de fonction" (LOF - Loss of Function).
- Perte de gènes par altération du nombre de copies (CNA) : si une région entière de l'ADN contenant l'un de ces gènes est perdue.

En l'absence de ces trois conditions, l'échantillon est classé comme MSS. Pour identifier le statut MSI/MSS d'un échantillon du cancer du côlon, il est nécessaire d'utiliser quatre types de données omiques simultanément qui doivent être disponibles pour cet échantillon : les données d'expression de gènes, les données de méthylation, les données de mutations et les données de CNA .

Résultats

Impact pronostic de la signature GEC

L'objectif de cette partie est d'analyser l'impact du Gene Expression Classifier (GEC) sur la survie des patients atteints du cancer du côlon dans trois jeux de données publics: TCGA-COAD, GSE39582 et GSE17536.

Pour limiter des biais éventuels liés aux décès survenus après 10 ans pour des causes autres que le cancer, la durée maximale de suivi de patients a été plafonnée à 120 mois. L'analyse statistique a été réalisée en traçant les courbes de Kaplan-Meier pour chaque groupe GEC afin d'estimer la probabilité de survie en fonction du temps. La comparaison des courbes a été effectuée à l'aide du test du log-rank, tandis que l'impact du score GEC sur la survie a été évalué à l'aide d'un modèle de Cox.

Les Figures 1, 2 et 3 présentent les courbes de survie obtenues pour chacun des trois jeux de données TCGA-COAD, GSE39582 et GSE17536., mettant en évidence les différences de survie entre les groupes GEC.

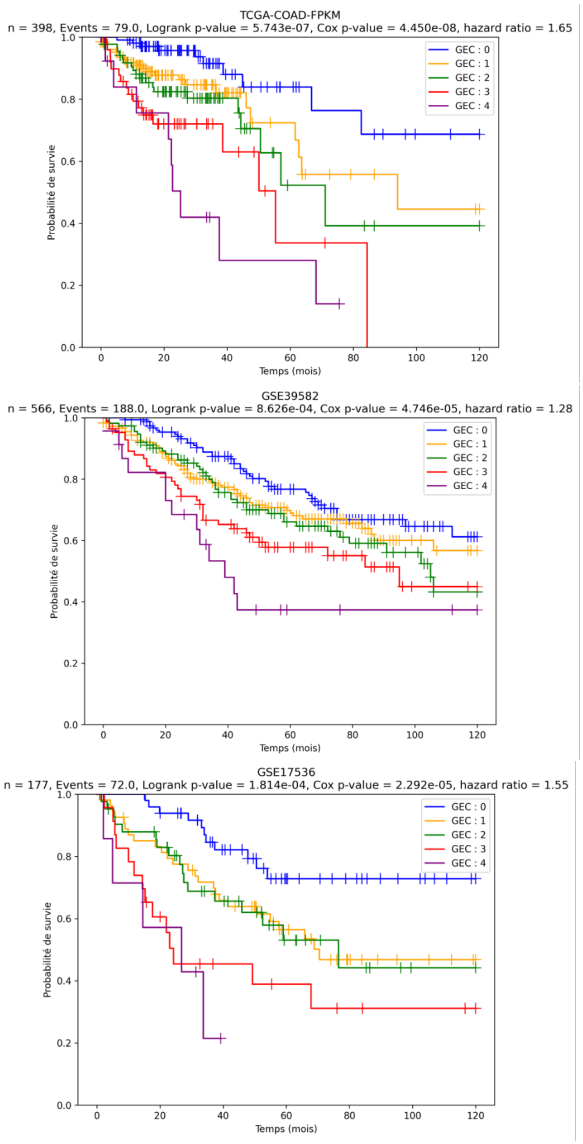


Figure 1: *Courbes de survie de Kaplan-Meier dans trois jeux de données du cancer du côlon en fonction du nombre de gènes activés du panel GEC à 4 gènes (ERFE, HOXC6, LAMP et ULBP2).*

Dans le but de savoir s'il est possible de réduire la signature pronostique 4-GEC à 3-GEC tout en conservant un impact significatif sur la survie des patients, nous avons constitué le protocole suivant :

- Calcul du score GEC basé uniquement sur 4 gènes (ERFE, HOXC6, LAMP5, ULBP2) et sur trois gènes pour lesquels un anticorps est utilisable en immunohistochimie (HOXC6, LAMP5, ULBP2).
- Création des courbes de survie Kaplan-Meier pour les signatures 4-GEC et 3-GEC sur les trois jeux de données (TCGA-COAD, GSE39582 et GSE17536).
- Évaluation de la significativité statistique des différences de survie à l'aide du test du log-rank et d'un modèle de Cox.

A la suite du protocole, les courbes de survies ci-dessous ont été obtenues :

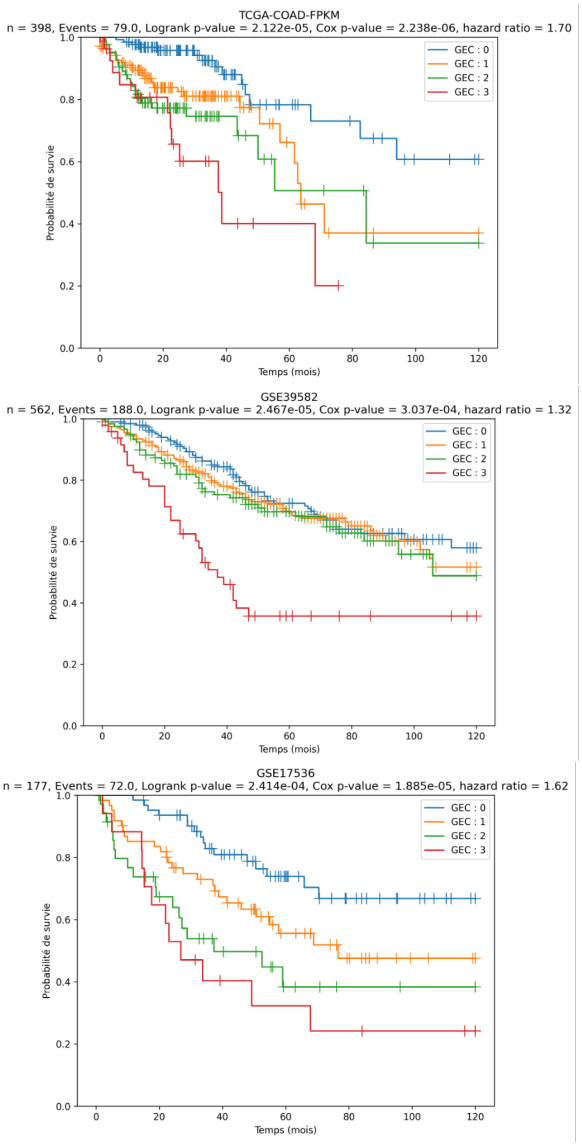


Figure 2: *Courbes de survie de Kaplan-Meier dans trois jeux de données du cancer du côlon en fonction du nombre de gènes activés du panel GEC à 3 gènes (HOXC6, LAMP et ULBP2)*

Nos résultats montrent que la signature 3-GEC reste un bon facteur pronostique avec un impact significatif sur la survie. Toutefois, une légère perte de performance est observée par rapport à la signature complète 4-GEC.

Ainsi, les cliniciens pourraient envisager d'utiliser la signature 3-GEC en approche d'immunohistochimie, car elle repose sur des anticorps déjà disponibles et conserve un bon pouvoir pronostique.

Identification du statut MSI/MSS du cancer du côlon dans le jeu de données TCGA-COAD

Le second objectif principal de notre projet tutoré est d'identifier le statut MSI/MSS des échantillons du cancer du côlon et d'évaluer l'impact pronostique du score GEC dans les sous-groupes du statut MSI et MSS séparément. Le statut MSI/MSS peut être déterminé directement à partir des données multi-omiques d'expression de gènes, de méthylation, de mutations et de CNA si ces données sont simultanément disponibles dans les échantillons du cancer du côlon concernés. L'algorithme d'identification du statut MSI/MSS est décrit ici . Il se base sur le niveau de méthylation de la région promotrice du gène MLH1, les statuts de mutation des gènes MLH1, MSH2, MSH6 et PMS2, ainsi que sur des CNA possibles concernant ces gènes.

Les données multi-omiques nécessaires à l'identification du statut MSI/MSS sont disponibles uniquement dans le jeu de données TCGA-COAD pour une partie d'échantillons. Pour les autres échantillons, seulement les données d'expression de gènes sont présentes. Compte tenu de cette limitation, nous avons procédé de manière suivante. Tout d'abord, dans les échantillons du dataset TCGA-COAD où les données nécessaires étaient disponibles, nous avons analysé séparément les données de méthylation, de mutation et de CNA, en identifiant les différentes conditions liées aux statuts MSI/MSS. Ensuite, nous avons regroupé ces résultats afin de déterminer le statut MSI/MSS final de ces échantillons.

Finalement, nous avons utilisé le statut MSI/MSS ainsi déterminé pour développer des algorithmes d'apprentissage supervisé en machine learning avec l'objectif de créer un modèle de prédiction du statut MSI/MSS à partir des données d'expression de gènes uniquement. Grâce à ce modèle nous avons réussi à prédire les statut MSI/MSS des échantillons pour lesquels les données multi-omiques n'étaient pas disponibles.

Analyse de méthylation dans la région promotrice du gène MLH1 Sur les données analysées, nous avions 398 échantillons de colon tumoral de personnes atteintes de cancer du côlon, ainsi que les niveaux de méthylation et d'expression du gène MLH1. Le gène MLH1 possédant plusieurs positions dans sa région promotrice pouvant être méthylés, l'analyse a été effectuée sur la moyenne de méthylation sur les différentes régions. Ainsi, en croisant les données d'expression et de méthylation, nous avons effectué un clustering afin d'identifier les échantillons ayant une faible expression et une forte méthylation. Pour ce faire, nous utilisons les seuils suivants :

- Expression < 1.5
- Methylation moyenne > 0.3

Aussi, nous avons calculé un coefficient de corrélation de pearson entre expression et méthylation. Nos résultats sont les suivants :

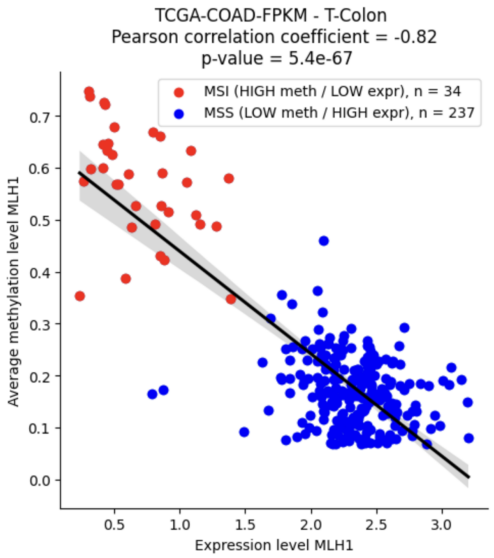


Figure 3: *Graphique des samples par methylation moyenne et expression. En bleu : samples MSI / En orange : samples MSS (courbe de regression linéaire en bleu avec intervalles de confiance à 95%*

Le coefficient de corrélation est significatif et négatif : ainsi, pour nos données, le niveau de méthylation moyen est est décroissant en fonction des données d'expression, c'est logique (source). Sur ce graphique, on voit bien la séparation entre les échantillons ayant une faible expression et forte méthylation (n = 34, points en bleu) des autres echantillons (en orange).

Analyse de mutations des gènes MLH1, MSH2, MSH6 et PMS2 Le statut de mutation des gènes MLH1, MSH2, MSH6 et PMS2 est disponible dans 362 échantillons sur 398 (91.5%) dans le jeu de données TCGA-COAD. Dans 20 de ces échantillons, nous avons trouvé au moins une mutation de type "perte de fonction" des gènes MLH1, MSH2, MSH6 ou PMS2. Par conséquent, ces 20 échantillons ont le statut MSI. Dans les 342 restants aucune mutation de ces gènes n'a été détectée. Pour connaître le statut de ces échantillons, il sera nécessaire de croiser les résultats obtenus avec d'autres types de données (méthylation et CNA).

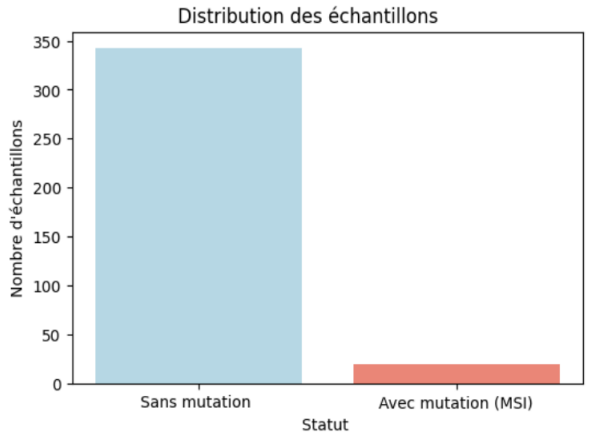


Figure 4: *Répartition des échantillons de l'analyse de mutation*

Analyse d'altérations du nombre de copies (CNA) des gènes MLH1, MSH2, MSH6 et PMS2 Dans cette sous-partie, on cherche à identifier les échantillons présentant une altération du nombre de copies (CNA) des gènes MLH1, MSH2, MSH6 et PMS2. Les données CNA sont disponibles dans 395 échantillons sur 398 dans le jeu de données TCGA-COAD.

Pour ces 395 échantillons, nous avons obtenu les résultats suivants :

- 12 échantillons possèdent une altération de CNA avec une perte de copie. Ces échantillons ont le statut MSI, soit 96,96% de la cohorte.
- 383 échantillons n'ont pas d'altérations de nombre de copies.

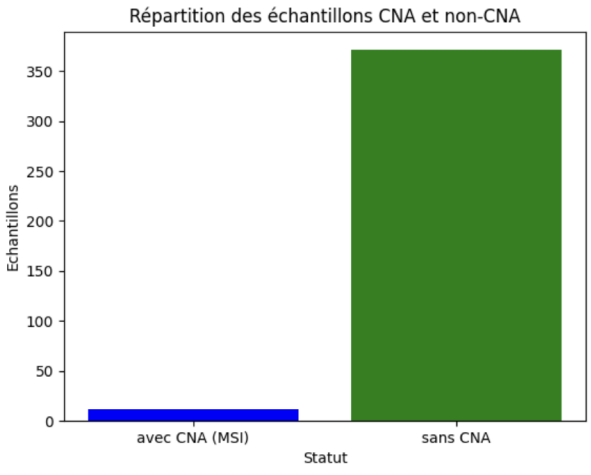


Figure 5: *Répartition des échantillons CNA MSI/MSS*

Ces résultats devront en plus être analysés avec ceux des données de méthylation et de mutation pour pouvoir ensuite conclure sur leur statut MSI/MSS.

Synthèse des trois approches pour la classification MSI/MSS dans les données TCGA/COAD Pour déterminer le statut MSI (Microsatellite Instable) ou MSS (Microsatellite Stable) des échantillons tumoraux du côlon (T-Colon), nous avons intégré trois types d'approches complémentaires :

- Mutations de type perte de fonction (LOF) dans les gènes du système MMR (MLH1, MSH2, MSH6, PMS2).
- Altérations du nombre de copies (CNA) délétères.
- Hyperméthylation du promoteur du gène MLH1 associée à une down-régulation de son expression.

Après croisement de ces données, le statut MSI/MSS des échantillons a pu être défini comme suit :

Statut	Nombre d'échantillons
MSI	56
MSS	212
Inconnu	130
Total	398

Table 2: *Répartition des échantillons selon le statut MSI/MSS*

Sur les 398 échantillons tumoraux du côlon (T-Colon), l'information sur les mutations était disponible pour 362 d'entre eux. Vingt présentaient au moins une mutation de type LOF . Par ailleurs, 12 échantillons montraient une délétion CNA et 34 une hyperméthylation du promoteur de MLH1 avec une expression faible. Certains échantillons répondaient à plusieurs critères. L'union de ces trois approches permet d'identifier au total 56 échantillons MSI.

Les 212 échantillons MSS sont ceux pour lesquels les quatre types de données étaient disponibles et aucun critère d'instabilité microsatellitaire n'était présent. Le statut de 130 échantillons reste inconnu, faute de données complètes.

Algorithme de machine learning pour prédire le statut MSI/MSS

L'objectif de cette analyse est de développer des modèles d'apprentissage supervisé capables de prédire le statut MSI (Microsatellite Instability) / MSS (Microsatellite Stability) des échantillons du jeu de données TCGA-COAD à partir de données d'expression de gènes.

Suite aux analyses précédentes, nous avons identifié le statut MSI/MSS pour 268 échantillons du cancer du côlon sur 398 disponibles au total dans le jeu de données TCGA-COAD. Parmi ces 268 échantillons, 56 (20.9%) avaient le statut MSI et 212 (79.1%) le statut MSS. Pour 130 échantillons restants, il n'était pas possible d'identifier le statut MSI/MSS à cause de

certaines données multi-omiques manquantes. En revanche, le statut MSI/MSS a un impact biologique important sur les profils d'expression de gènes. Puisque nous disposons des données d'expression de gènes pour la totalité des 130 échantillons, nous pouvons estimer le statut MSI/MSS dans ces échantillons avec une approche de machine learning supervisée.

Sélection de variables Le jeu de données que nous utilisons pour créer un modèle de prédiction du statut MSI/MSI contient 268 échantillons et 42696 paramètres (aussi appelés features ou variables explicatives). Ces paramètres correspondent à la totalité des transcrits, c'est-à-dire, la totalité des molécules ARN, présents dans chaque échantillon. Les valeurs numériques d'expression représentent les niveaux d'abondance normalisés de ces transcrits. Ainsi, le nombre de paramètres dans notre jeu de données est plus de 150 fois supérieur au nombre d'échantillons. Un tel déséquilibre peut favoriser un surapprentissage lors de l'entraînement d'un modèle de machine learning.

Il existe plusieurs approches pour résoudre ce problème. Dans notre projet, nous avons choisi de réaliser une sélection de variables, en se basant sur deux considérations. Premièrement, nous avons utilisé une liste de gènes, déjà décrite dans la littérature scientifique, dont l'expression dépend significativement du statut MSI/MSS. En particulier, Watanabe et collègues [**Watanabe2006**] ont défini une liste de 69 gènes réprimés et de 30 gènes surexprimés dans le cancer du côlon avec le statut MSI par rapport au statut MSS. La liste de ces gènes est disponible dans la base de données publique de signatures moléculaires MSigDB (https://www.gsea-msigdb.org).

Deuxièmement, pour compléter la liste de Watanabe, nous avons réalisé une analyse différentielle entre le transcriptome d'échantillons MSI versus le transcriptome d'échantillons MSS dans le jeu de données TCGA-COAD et avons sélectionné les gènes dont l'expression varie significativement entre ces deux conditions. Concrètement, pour chaque gène disponible dans TCGA-COAD, nous avons calculé le fold change entre les niveaux moyens d'expression dans le groupe MSI versus le groupe MSS. De plus, nous avons appliqué le test statistique non paramétrique de Wilcoxon-Mann-Whitney entre les niveaux d'expression du groupe MSI versus le groupe MSS et avons obtenu les p-valeurs correspondantes. Puisque le nombre de gènes est très élevé et que le test a été répété plusieurs fois, nous avons ajusté les p-valeurs obtenues selon la procédure de Benjamini-Hochberg pour contrôler le taux de fausses découvertes (False Discovery Rate, FDR). En appliquant les critères de sélection sur le fold change > 2 et le FDR < 0.05, nous avons trouvé 102 gènes significativement réprimés et 47 gènes significativement surexprimés dans le groupe MSI versus le groupe MSS dans TCGA-COAD. Les résultats sont présentés sur un volcano plot de la figure XX. Certains gènes identifiés étaient communs avec la signature de Watanabe.

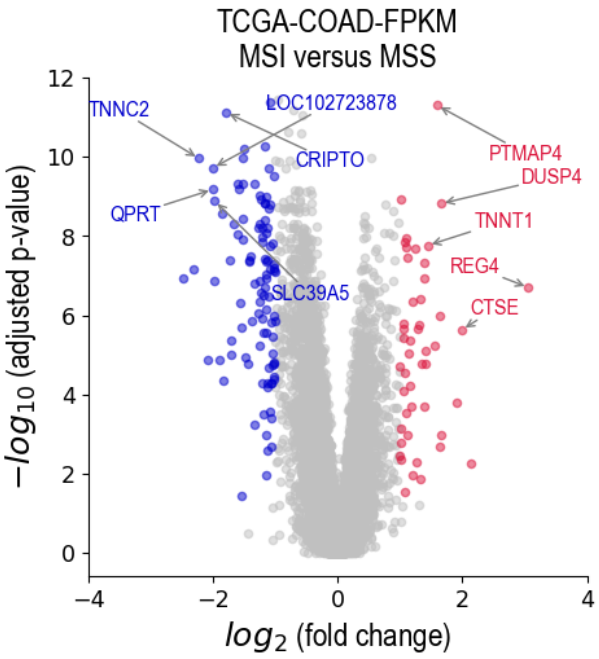


Figure 6: *Résultats de l'analyse différentielle entre les échantillons du cancer du côlon avec le statut MSI versus le statut MSS dans le jeu de données TCGA-COAD. Les points bleus indiquent les gènes significativement réprimés dans le groupe MSI par rapport au groupe MSS. Les points rouges représentent les gènes significativement surexprimés. Les points gris montrent les autres gènes pour lesquels la différence de niveaux d'expression entre les deux conditions n'était pas significative.*

Finalement, nous avons fusionné les listes de gènes identifiés avec les deux approches et avons obtenu 217 gènes au total. Ils ont été utilisés en tant que variables explicatives sélectionnés pour développer des modèles de machine learning.

Développement de modèles de machine learning Le jeu de données TCGA-COAD, initialement composé de 398 échantillons du cancer du côlon, a été prétraité pour l'apprentissage. Seuls les échantillons avec un statut MSI/MSS ont été conservés, et 217 gènes d'intérêt ont été sélectionnés. Les valeurs manquantes ont été éliminées, et la variable cible a été encodée en deux catégories (MSI/MSS). Au final, la matrice d'apprentissage comprenait 268 échantillons et 217 variables explicatives.

Plusieurs modèles d'apprentissage automatique ont été évalués en utilisant une validation croisée stratifiée à 3 folds sur l'ensemble de formation. Les modèles testés sont les suivants :

- Régression Logistique (simple, Ridge L2, Lasso L1, Elastic Net)
- Machines à Vecteurs de Support (SVM)
- Arbres de Décision et Forêts Aléatoires (Random Forest, Extra Trees, Decision Tree)
- Gradient Boosting (XGBoost, XGBoost-Tuned)
- Naïve Bayes
- Réseaux de Neurones (MLPClassifier)

Chaque modèle a été évalué à l'aide de la validation croisée et les performances ont été mesurées en utilisant l'accuracy moyenne obtenue sur les folds. Les résultats des performances des modèles sont les suivants :

Modèle	Accuracy	Precision	F1-score
XGBoost_Tuned	0.925373	0.973684	0.787234
XGBoost	0.910448	0.833333	0.769231
Extra Trees	0.899254	0.891892	0.709677
Neural Network	0.895522	0.780000	0.735849
Random Forest	0.888060	0.809524	0.693878
SVM	0.884328	0.735849	0.715596
Decision Tree	0.873134	0.683333	0.706897
Lasso reg.	0.869403	0.677966	0.695652
Linear reg.	0.828358	0.562500	0.661765
Ridge reg.	0.828358	0.562500	0.661765
Naïve Bayes	0.828358	0.564103	0.656716
Elastic Net reg.	0.820896	0.546512	0.661972

Table 3: *Comparaison des performances des modèles de classification selon plusieurs métriques. Les meilleures valeurs par colonne sont surlignées.*

Le modèle XGBoost Tuné a été sélectionné comme le meilleur modèle après une optimisation des hyperparamètres via une recherche sur grille (GridSearchCV). L'optimisation a été réalisée en testant différentes combinaisons de paramètres pour un Gradient Boosting Classifier, notamment :

- Nombre d'arbres : 100, 200, 500
- Profondeur maximale des arbres : 3, 5, 7
- Taux d'apprentissage : 0.01, 0.05, 0.1
- Proportion d'échantillons utilisés pour chaque arbre : 0.7, 0.8, 1.0

La validation croisée (5 folds) a été utilisée pour évaluer la performance de chaque combinaison en fonction de l'accuracy. La meilleure combinaison d'hyperparamètres a ensuite été sélectionnée et appliquée au modèle final.

Une fois le modèle entraîné, nous avons tenté de prédire le statut MSI/MSS pour les échantillons "Inconnu" dans TCGA. Cependant, nous avons rencontré certaines limites :

- L'absence d'annotations de statut MSI/MSS dans les autres jeux de données (GSE39582 et GSE17536) empêche une évaluation rigoureuse des performances sur ces datasets.
- Les mauvaises performances obtenues lors des tests suggèrent des différences d'acquisition des données et d'harmonisation entre TCGA et les autres bases.

Ainsi, aucune prédiction fiable n'a pu être réalisée sur les autres jeux de données.

En conclusion, nous avons développé plusieurs modèles de classification pour prédire le statut MSI/MSS des échantillons TCGA en utilisant l'expression des 217 gènes sélectionnés.Le modèle XGBoost Tuné s'est révélé le plus performant. Toutefois, des défis persistent pour généraliser ce modèle aux autres bases de données en raison du manque d'annotations et de possibles variations dans les méthodologies d'acquisition des données.

Analyse pronostique de la signature GEC dans les cancers MSI et MSS L'analyse vise à évaluer l'impact du GEC sur la survie des patients, séparément pour les groupes MSI (Microsatellite Instable) et MSS (Microsatellite Stable), en utilisant plusieurs modèles de classification (XGBoost, SVM,etc.). Les métriques clés incluent :

- Logrank p-value : Teste la différence significative entre les courbes de survie (GEC 0 vs GEC 1-4).
- Hazard Ratio (HR) : Mesure le risque relatif associé au GEC.

Il serait mieux de calculer si possible non seulement les p-valeurs et les hazard ratios mais aussi le C-index (index d'Harell ou index de concordance) pour pouvoir comparer les performances de ces modèles. En analysant ces métriques, répondez clairement aux questions des objectifs. Pour le moment, vous donnez uniquement des résultats MSS. Cela ne suffit pas pour conclure.

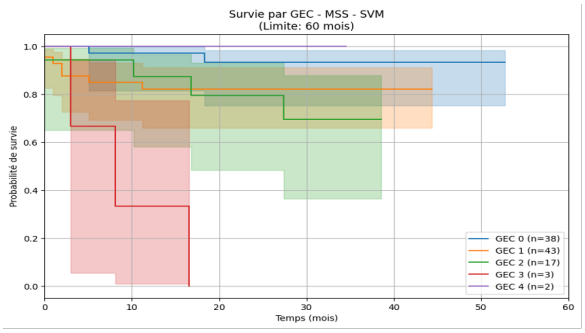
Ainsi, pour le groupe MSI, tous les modèles montrent une différence significative avec une p-valeur inférieure à 0,05 entre les groupes GEC 0 et GEC 1-4. Le modèle SVM est cependant le plus performant avec la p-valeur la plus faible (0.025) et le meilleur C-index (0.683). Un HR supérieur à 1 indique que les patients avec un score GEC compris entre 1 et 4 ont un risque de décès accru (75 à 88% plus élevé que GEC 0).

Pour le groupe MSS, les résultats sont similaires au groupe MSI, avec des p-values significatives et des Hazard Ratio cohérents.

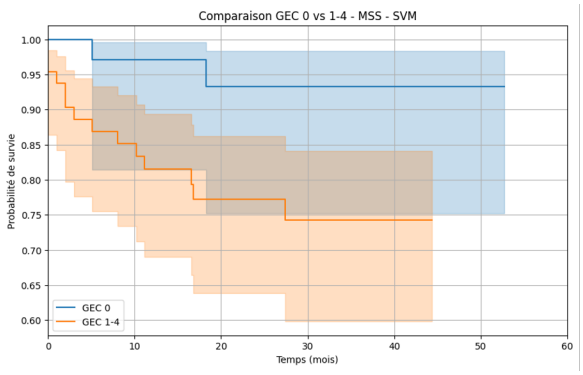
Le SVM (Support vector machine) reste, ici aussi, le modèle le plus robuste pour discriminer les groupes GEC.

Afin de visualiser les résultats obtenus, nous utilisons les courbes de Kaplan-Meïer avec un tracé pour chaque valeur du score GEC et ce dans chaque groupe. Les courbes Kaplan-Meier confirment une séparation nette entre GEC 0 (meilleure probabilité de survie) et GEC 1-4 (probabilité de survie réduite).

Exemple pour le SVM (MSI) : Probabilité de survie à 60 mois : 80% (GEC 0) vs 50% (GEC 1-4).



(a) Survie par GEC à 60 mois pour les echantillons MSS (modèle SVM)



(b) comparaison entre GEC 0 et GEC 1-4 pour les echantillons MSS (modèle SVM)

L'analyse démontre que le GEC est un marqueur prédictif significatif de la survie, aussi bien dans les groupes MSI (Microsatellite Instable) que MSS (Microsatellite Stable). Les résultats des tests statistiques (Logrank p-value < 0.05, Cox p-value < 0.05) confirment une différence significative entre les courbes de survie des patients GEC 0 (meilleur pronostic) et GEC 1-4 (survie réduite). Parmi les modèles évalués, le SVM s'est révélé le plus performant, affichant une p-valeur de 0.025 (MSI) et 0.00411 (MSS), ainsi qu'un C-index élevé (0.68), indiquant une bonne capacité discriminative. Les Hazard Ratios (HR) compris entre 1.75 et 1.88 suggèrent une augmentation du risque de décès de 75 à 88% pour les patients avec un GEC ≥ 1 par rapport à ceux avec un GEC 0, renforçant ainsi la valeur pronostique du score. Sur le plan clinique, ces résultats impliquent que le GEC pourrait servir d'outil de stratification pour identifier les patients à risque accru dans les deux sous-types moléculaires (MSI/MSS), permettant une prise en charge personnalisée. Cependant, une limite importante réside dans les effectifs réduits des groupes GEC 3-4 (n < 5), ce qui nécessite une validation sur des cohortes plus larges pour confirmer ces observations. En résumé, cette étude confirme l'utilité du GEC comme marqueur pronostique indépendant, avec des performances robustes pour les modèles SVM et XGBoost dans les deux contextes MSI et MSS, ouvrant des perspectives pour son intégration dans les analyses pronostiques en oncologie.

Dans un second temps, nous avons réalisé une analyse de survie multivariée en combinant le score GEC avec le statut MSI/MSS. Cette analyse porte uniquement sur le sous-ensemble de 268 patients pour lesquels le statut MSI/MSS a été déterminé à partir des données multi-omiques (56 MSI, 212 MSS).

Deux modèles ont été construits : l'un avec une variable binaire GEC3 (GEC ≥ 3), l'autre avec GEC4 (GEC ≥ 4). L'objectif est d'évaluer si le score GEC et le statut MSI apportent une information pronostique complémentaire sur la survie globale des patients.

Les résultats montrent que le GEC est un facteur pronostique majeur de la survie. Un GEC ≥ 3 est associé à un risque de décès 3,8 fois plus élevé (HR = 3.80, $p < 0.005$), et ce risque augmente encore pour un GEC ≥ 4 (HR = 4.42, $p < 0.005$). Ces effets sont observés indépendamment du statut MSI ou MSS. En revanche, le statut MSI/MSS n'est pas significativement associé à la survie ($p = 0.47$ pour MSI, $p = 0.54$ pour MSS), avec par exemple un HR pour MSI de 0.79 (IC 95% : 0.41–1.52), indiquant un effet neutre. Les modèles présentent une concordance modeste (0.57 à 0.63), soulignant que le GEC est la variable prédominante. Le test du rapport de vraisemblance est significatif ($p < 0.005$), ce qui confirme que le modèle incluant GEC et MSI/MSS est supérieur à un modèle nul.

En somme, le GEC (seuils ≥ 3 ou ≥ 4) constitue un marqueur indépendant et puissant de mauvais pronostic, alors que le statut MSI/MSS n'apporte pas d'information pronostique notable. D'un point de vue clinique, le GEC pourrait donc suffire à identifier les patients à haut risque, sans qu'il soit nécessaire de le combiner au statut MSI/MSS.

Comparaison MSI vs MSS sans prise en compte du GEC Le graphique compare la survie entre les patients MSI et MSS sans tenir compte du score GEC. Les courbes sont très proches, parfois superposées, ce qui indique que le statut MSI/MSS seul n'a pas d'effet pronostique significatif dans cette cohorte. Cela confirme les résultats multivariés où MSI/MSS n'était pas associé à la survie ($p > 0.05$). Bien que l'étude conclut que le GEC influence la survie, ce graphique ne le montre pas. L'analyse visait sans doute à évaluer si le statut MSI/MSS apportait une information indépendante. Les résultats confirment que non : seul le GEC impacte significativement le pronostic.

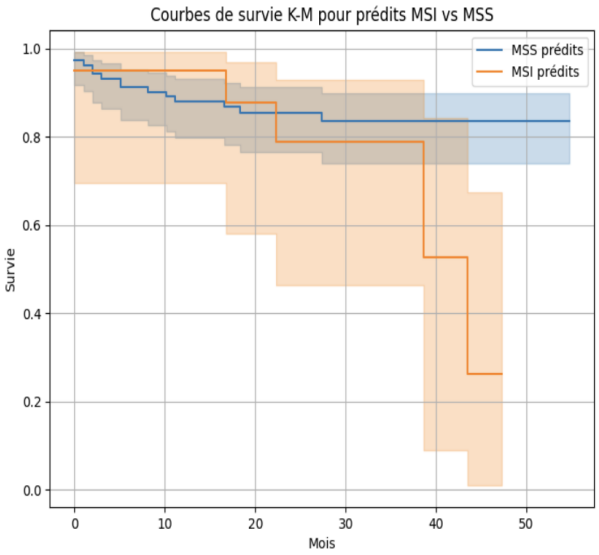


Figure 7: *Courbes de survie des prédictions MSI / MSS*

Résultats finaux Le GEC ressort comme le marqueur pronostique central, avec un risque de décès accru de 75 à 88% dès un score ≥ 1, indépendamment du statut MSI ou MSS. Le statut MSI/MSS n'apporte pas d'information pronostique significative, ni seul ni en interaction avec le GEC. Le graphique Kaplan-Meier MSI vs MSS illustre cette neutralité, et son rôle est méthodologique : écarter un effet confondant. Malgré l'absence de stratification par GEC, la cohérence des résultats entre analyses univariées et multivariées confirme la solidité du GEC. Ce marqueur pourrait suffire, à lui seul, à guider la stratification clinique du risque. Une analyse visuelle complémentaire stratifiée par GEC et MSI/MSS renforcerait encore cette conclusion.

Impact sociétal et environnemental du projet

Le numérique nous offre la possibilité d'approfondir notre compréhension du monde et, dans le cadre de ce projet, d'améliorer l'efficacité des traitements médicaux, notamment pour le cancer du côlon. Toutefois, avec près de cinquante ZettaOctets de données générées en 2020, il est devenu essentiel que le secteur numérique s'interroge sur ses usages et participe activement à la réduction de ses impacts environnementaux. Si ce projet contribue à la recherche contre le cancer du côlon, il n'échappe pas aux réalités de l'empreinte écologique du numérique. En effet, au-delà des algorithmes utilisés et des calculs informatiques nécessaires, des activités comme les déplacements, les visioconférences et les besoins en infrastructures matérielles et logicielles génèrent une consommation énergétique importante.

Dans la recherche scientifique en général, et dans ce projet également, l'accumulation et la consommation croissante de données, résultant de calculs complexes, de modélisations ou de traitements d'images par exemple, engendrent une forte demande en matériel informatique et en énergie. Bien que des progrès aient été réalisés en matière d'efficacité énergétique, ces avancées peinent à compenser la croissance exponentielle des besoins en ressources, particulièrement dans le contexte des technologies de l'intelligence artificielle, des big data et du machine learning.

Cette réalité soulève une question cruciale : comment conjuguer progrès technologique et durabilité environnementale ? Le numérique, en dépit de ses bienfaits indéniables dans la recherche, a un impact direct sur l'environnement, contribuant aux dérèglements climatiques par la consommation d'énergie, la production de déchets électroniques, et la pollution numérique. Ces défis soulèvent de nombreuses questions quant à la justice sociale et aux inégalités. Les actions pour limiter ces impacts sont multiples, allant de l'optimisation des algorithmes pour réduire leur consommation énergétique à l'adoption de pratiques plus responsables en, par exemple, limitant au maximum l'utilisation d'intelligence artificielle. Toutefois, ces réponses restent complexes et nécessitent une réflexion globale, prenant en compte les aspects sociaux, sociétaux et humains, en plus des considérations écologiques. Dans le cadre de ce projet, il est donc essentiel d'intégrer ces enjeux dans la conception et la mise en œuvre des solutions, en adoptant une approche éco-responsable et en envisageant des stratégies pour réduire l'empreinte environnementale du traitement et de l'analyse des données, tout en maximisant leur utilité pour la recherche et le progrès médical.

Dans cette optique, et bien que le projet ait engendré un certain impact carbone, l'un des objectifs de l'analyse 3-GEC est de développer un test immunohistochimique présentant un coût environnemental inférieur à celui des RT-qPCR, tout en conservant, voire en améliorant, son efficacité sur le plan sociétal.

Ce projet s'inscrit dans une démarche active visant à limiter l'empreinte environnementale des outils de diagnostic, tout en maintenant une exigence élevée en matière d'efficacité médicalel.

Estimation du Bilan Carbone du projet :

- Déplacements : 2,88 kg CO2
- Conversation entre les membres du groupe ainsi que les échanges par mail avec notre tutrice : l'empreinte carbone totale varie entre 23,6 g CO2 et 54 g CO2
- Visio conférence : 15,39 kg CO2
- Code : 11,28 kg CO2 (estimation pour des script

Python d'environ 500 lignes lancés une trentaine de fois dans leur intégralité sur VSCode)
- Les stockages des bases de données et des autres documents étant peu volumineux et de courte période (seulement les quelques mois de ce projet tutoré) ont un impact négligeable sur le calcul.

Conclusion

Ce projet tutoré, réalisé en collaboration avec l'équipe « Régulations épigénétiques » de l'IAB, s'est inscrit dans une démarche de recherche translationnelle visant à améliorer la stratification pronostique des patients atteints de cancer du côlon. Dans un premier temps, nous avons rappelé le rôle fondamental de l'épigénétique dans l'activation aberrante de gènes normalement silencieux, un phénomène récurrent dans divers types de cancers, dont le côlon. Ces observations ont conduit au développement de la signature GEC (Gene Expression Classifier), basée sur l'expression de quatre gènes (ERFE, HOXC6, LAMP5, ULBP2), identifiés comme marqueurs pronostiques défavorables.

Notre première analyse a démontré que la signature GEC à quatre gènes présentait un pouvoir discriminant significatif pour la survie des patients. Afin de rendre ce test plus facilement applicable en milieu hospitalier, nous avons évalué une version simplifiée à trois gènes (3-GEC), basée uniquement sur les gènes pour lesquels des anticorps commerciaux sont disponibles. Les résultats ont montré une légère perte de performance, mais une robustesse suffisante pour envisager une utilisation clinique.

Le second volet de notre travail a porté sur le statut MSI/MSS des tumeurs colorectales. Grâce à une méthodologie intégrant données de méthylation, mutations et CNA, nous avons pu classifier les échantillons TCGA-COAD et construire des modèles d'apprentissage supervisé pour prédire ce statut à partir des seules données d'expression. Le modèle XGBoost optimisé s'est avéré le plus performant, bien que sa généralisation à d'autres bases de données reste limitée par des divergences techniques.

Enfin, l'analyse croisée entre le score GEC et le statut MSI/MSS a confirmé que le GEC constitue un marqueur pronostique indépendant, aussi bien pour les cancers MSI que MSS. Les modèles SVM et XGBoost ont confirmé la valeur prédictive significative du GEC avec des Hazard Ratios élevés et des indices de concordance satisfaisants.

Au-delà des résultats biologiques et statistiques, ce projet s'est également inscrit dans une réflexion plus large sur les impacts sociétaux et environnementaux de la recherche numérique. Conscients des coûts énergétiques associés à l'utilisation de ressources computationnelles, nous avons proposé une estimation du bilan carbone et souligné l'importance de démarches écoresponsables.

En conclusion, notre étude confirme la pertinence de la signature GEC comme outil de prédiction de la survie dans le cancer du côlon, avec un potentiel d'intégration en routine clinique, y compris dans une version simplifiée. Elle ouvre également la voie à de futures recherches sur la complémentarité entre signatures transcriptomiques et autres biomarqueurs moléculaires.

References

[1] https://www.cancer.fr/professionnels-de-sante/statistiques-et-chiffres-sur-les-cancers/epidemiologie-des-cancers/cancer-colorectal.

[Eduardo Villar et al., 2010] Vilar E, Gruber SB. Microsatellite instability in colorectal cancer-the stable evidence. Nat Rev Clin Oncol. 2010 Mar;7(3):153-62. doi: 10.1038/nrclinonc.2009.237. Epub 2010 Feb 9. PMID: 20142816; PMCID: PMC3427139.

[Ruby Gupta et al., 2018] Gupta R, Sinha S, Paul RN. The impact of microsatellite stability status in colorectal cancer. Curr Probl Cancer. 2018 Nov;42(6):548-559. doi: 10.1016/j.currproblcancer.2018.06.010. Epub 2018 Jul 18. PMID: 30119911.

[K. Yamaguchi et al., 2025] Yamaguchi K, Tsuchihashi K, Ueno S, Uehara K, Taguchi R, Ito M, Isobe T, Imajima T, Kitazono T, Tanoue K, Ohmura H, Akashi K, Baba E. Efficacy of pembrolizumab in microsatellite-stable, tumor mutational burden-high metastatic colorectal cancer: genomic signatures and clinical outcomes. ESMO Open. 2025 Jan;10(1):104108. doi: 10.1016/j.esmoop.2024.104108. Epub 2025 Jan 6. PMID: 39765187; PMCID: PMC11758824.

[Rousseaux et al., 2013] Rousseaux S, Debernardi A, Jacquiau B, Vitte AL, Vesin A, Nagy-Mignotte H, Moro-Sibilot D, Brichon PY, Lantuejoul S, Hainaut P, Laffaire J, de Reyniès A, Beer DG, Timsit JF, Brambilla C, Brambilla E, Khochbin S. Ectopic activation of germline and placental genes identifies aggressive metastasis-prone lung cancers. Sci Transl Med. 2013 May 22;5(186):186ra66. doi: 10.1126/scitranslmed.3005723. PMID: 23698379; PMCID: PMC4818008.

[Le Bescont et al., 2015] Le Bescont A, Vitte AL, Debernardi A, Curtet S, Buchou T, Vayr J, de Reyniès A, Ito A, Guardiola P, Brambilla C, Yoshida M, Brambilla E, Rousseaux S, Khochbin S. Receptor-Independent Ectopic Activity of Prolactin Predicts Aggressive Lung Tumors and Indicates HDACi-Based Therapeutic Strategies. Antioxid Redox Signal. 2015 Jul 1;23(1):1-14. doi: 10.1089/ars.2013.5581. Epub 2014 Mar 6. PMID: 24512221; PMCID: PMC4492736.

[Emadali et al., 2013] Emadali A, Rousseaux S, Bruder-Costa J, Rome C, Duley S, Hamaidia S, Betton P, Debernardi A, Leroux D, Bernay B, Kieffer-Jaquinod S, Combes F, Ferri E, McKenna CE, Petosa C, Bruley C, Garin J, Ferro M, Gressin R, Callanan MB, Khochbin S. Identification of a novel BET bromodomain inhibitor-sensitive, gene regulatory circuit that controls Rituximab response and tumour growth in aggressive lymphoid cancers. EMBO Mol Med. 2013 Aug;5(8):1180-95. doi: 10.1002/emmm.201202034. Epub 2013 Jul 4. PMID: 23828858; PMCID: PMC3944460.

[Wang et al., 2015] Wang Y, Xiao M, Chen X, Chen L, Xu Y, Lv L, Wang P, Yang H, Ma S, Lin H, Jiao B, Ren R, Ye D, Guan KL, Xiong Y. WT1 recruits TET2 to regulate its target gene expression and suppress leukemia cell proliferation. Mol Cell. 2015 Feb 19;57(4):662-673. doi: 10.1016/j.molcel.2014.12.023. Epub 2015 Jan 15. PMID: 25601757; PMCID: PMC4336627.

[Peng et al., 2022] Peng LJ, Zhou YB, Geng M, Bourova-Flin E, Chuffart F, Zhang WN, Wang T, Gao MQ, Xi MP, Cheng ZY, Zhang JJ, Liu YF, Chen B, Khochbin S, Wang J, Rousseaux S, Mi JQ. Ectopic expression of a combination of 5 genes detects high risk forms of T-cell acute lymphoblastic leukemia. BMC Genomics. 2022 Jun 24;23(1):467. doi: 10.1186/s12864-022-08688-1. PMID: 35751016; PMCID: PMC9233359.

[Bourova Flin et al., 2021] Ekaterina Bourova-Flin, Samira Derakhshan, Afsaneh Goudarzi, Tao Wang, Anne-Laure Vitte, et al.. The combined detection of Amphiregulin, Cyclin A1 and DDX20/Gemin3 expression predicts aggressive forms of oral squamous cell carcinoma. British Journal of Cancer, 2021, ⟨10.1038/s41416- 021-01491-x⟩. ⟨hal-03359619⟩

[Jacquet et al., 2023] Emmanuelle Jacquet, Florent Chuffart, Anne-Laure Vitte, Eleni Nika, Mireille Mousseau, et al.. Aberrant activation of five embryonic stem cell-specific genes robustly predicts a high risk of relapse in breast cancers. BMC Genomics, 2023, 24, pp.463. ⟨10.1186/s12864-023-09571-3⟩. ⟨hal-04216151⟩

[Spinelli, 2024] Aurélien Spinelli. Signature pronostique dans le cancer du côlon par l'étude de l'expression ectopique de gènes spécifiques de tissus. Médecine humaine et pathologie. 2024. ⟨dumas-04738592⟩

[Kane, 2008] Poynter JN, Siegmund KD, Weisenberger DJ, Long TI, Thibodeau SN, Lindor N, Young J, Jenkins MA, Hopper JL, Baron JA, Buchanan D, Casey G, Levine AJ, Le Marchand L, Gallinger S, Bapat B, Potter JD, Newcomb PA, Haile RW, Laird PW; Colon Cancer Family Registry Investigators. Molecular characterization of MSI-H colorectal cancer by MLHI promoter methylation, immunohistochemistry, and mismatch repair germline mutation screening. Cancer Epidemiol Biomarkers Prev. 2008 Nov;17(11):3208-15. doi: 10.1158/1055-9965.EPI-08-0512. PMID: 18990764; PMCID: PMC2628332.

[Xia Li, 2013] Li X, Yao X, Wang Y, Hu F, Wang F, Jiang L, Liu Y, Wang D, Sun G, Zhao Y. MLH1 promoter methylation frequency in colorectal cancer patients and related clinicopathological and molecular features. PLoS One. 2013;8(3):e59064. doi: 10.1371/journal.pone.0059064. Epub 2013 Mar 29. PMID: 23555617; PMCID: PMC3612054.