UNIVERSITE LIBRE DE BRUXELLES

Faculté des Sciences

Institut de Recherche Interdisciplinaire en Biologie humaine et moléculaire (IRIBHM)

Promoteur : Pr. Vincent Detours

Etude de l'édition de l'ARN dans onze types de cancers

Mémoire présenté en vue de l'obtention du grade de master en Bioinformatique et Modélisation

Alizée Vercauteren Drubbel

Année académique 2015-2016

Vercauteren Drubbel Alizée
Année académique 2015-2016

Vercauteren Drubbel Alizée
Année académique 2015-2016

Vercauteren Drubbel Alizée
Année académique 2015-2016

Remerciements

Je tiens à remercier le Pr. Vincent Detours pour l'opportunité qu'il m'a offerte de faire mon mémoire au sein de son laboratoire. Cette expérience enrichissante a développé au mieux mes connaissances en bio-informatiques. Elle m'a également permis d'acquérir les compétences recherchée sur un sujet riche et peu étudié jusqu'à présent.

Je remercie également les membres du laboratoire David & Danai pour leurs gentillesses et leur aide ainsi que Joël et Benoit pour leur bonne humeur et conseils éclairés

Merci à ma sœur Maëlle pour son écoute et ses encouragements.

Merci à papa pour son soutien.

Merci à maman de croire en moi depuis toujours et de me soutenir coûte que coûte.

Je tiens également à remercier mes amis pour leur motivation. Tout d'abord Audrey pour ses 23 ans d'amitié inaltérable, pleine de souvenirs pittoresques et Charlotte pour son enthousiasme enrichissant. Bien sûr, un grand merci à Aline Wuidart, binôme et amie depuis 8 ans qui a su faire preuve d'écoute, de disponibilité et d'excellents conseils. Merci également aux pool'et matata qui, avec leur humour et brin de folie, m'ont apporté la distraction dont j'avais si souvent besoin. En particulier, merci à Manu pour sa relecture.

C'est avec une affection particulière que je remercie Valentin pour son écoute, sa compréhension, sa tolérance, son humour et son soutien.

Merci également à toutes les personnes que je n'ai pu citer mais qui ont contribué à la réalisation de mon mémoire et à mon épanouissement.

Résumé

L'édition de l'ARN adénosine-inosine (A-I) est un mécanisme post-transcriptionnel, qui a peu été étudié dans les cancers jusqu'à présent. Celui-ci est catalysé par l'enzyme ADAR, qui remplace une adénosine par une inosine dans une séquence d'ARN double brin. Ces inosines sont interprétées comme des guanosines par la cellule. Ce phénomène dynamique et flexible modifiant directement la séquence nucléotidique de l'ARN pourrait donc, au même titre que les mutations génétiques, jouer un rôle clé dans la tumorigénèse.

Dans notre étude nous caractérisons l'édition et les mécanismes gouvernant l'édition dans plus de 4000 échantillons regroupant 11 types de cancers provenant de *The Cancer Genome Atlas.* Nous confirmons que l'édition est augmentée dans la plupart des cancers comparés à leurs tissus sains correspondants. Nous montrons également que dans la plupart des cancers, cette édition concorde avec la surexpression de la protéine ADAR, dont l'expression est positivement corrélée à l'édition. Dans chacun des cancers, nous identifions le pourcentage de variation de l'édition qui peut être expliqué par la réponse à l'interféron et le nombre de copies génomiques d'ADAR. Nous démontrons que parmi toutes les voies biologiques, la voie de l'interféron semble la plus associée à l'édition dans la majorité des cancers. L'identification des gènes différentiellement exprimés avec l'édition nous montre que l'édition est associée à l'expression de milliers de gènes et serait donc associée à de nombreuses modifications cellulaires. En conséquence, nous confirmons que l'utilisation d'une signature d'édition, même cancer spécifique, n'est pas suffisante pour prédire de manière précise l'édition dans les échantillons. Finalement, nous investiguons l'association entre l'édition et la progression tumorale via une analyse de survie des patients. Nous démontrons que dans deux types de cancers l'édition semble significativement affecter la survie.

Ces découvertes démontrent que différents mécanismes intensifiés dans les cancers sont associés à l'augmentation de l'édition de l'ARN. Suggérant que celle-ci apporterait potentiellement un avantage prolifératif aux cellules cancéreuses dans certains types et/ou sous-types tumoraux et jouerait un rôle dans la tumorigénèse. Cependant, étudier l'édition et identifier ses conséquences précises dans la tumorigénèse semble plus complexe que prévu, entre autres, à cause de la grande hétérogénéité retrouvée entre les types tumoraux mais également au sein d'un même type tumoral.

Abstract

Adenosine-to-Inosine (A-to-I) RNA editing is a post-transcriptional mechanism that has been poorly investigated in cancer until now. It is catalyzed by the ADAR enzyme that turns adenosines into inosines in double-stranded RNA. Subsequently, the cellular machinery interprets these inosines as guanosines. This flexible and dynamic mechanism alters directly the nucleotide sequence of RNA and could consequently, as genetic mutations, play a key in role in tumor development.

In our study, we characterised RNA editing and its mechanisms in 4000 samples from 11 tumor types from The Cancer Genome Atlas. We confirmed that more RNA editing is found in tumor samples compared to normal matched tissue. We observed that in most cancers, ADAR is overexpressed and we showed a positive correlation between ADAR expression and RNA editing. In each cancer type, we identified the proportion of editing that can be explained both by the interferon response and ADAR genomic copy number gains. We revealed that considering all metabolic pathways, interferon response is the most associated with editing in most of the cancers. Analysis of differentially expressed genes with editing showed that this mechanism is associated with the expression of thousand of genes and likely to numerous cellular mechanisms. Consequently, we confirmed that using a gene expression signature does not predict accurately editing frequency in samples. Finally we investigated the link between editing and survival rate of the patient, showing a significant effect of editing on survival in two cancer types.

These results show that several mechanisms enhanced in tumors are associated with increasing RNA editing frequency. This suggests that editing could give a proliferation gain to tumor cells in some cancer types or subtypes and could play a role in tumor growth. However, studying editing and identifying its exact consequences in tumorigenesis will have to tackle, amongst other things, with tumor heterogeneity between tumor types but also within the same tumor type.

Liste des abréviations

A: adénosine

ADAR: Adenosine Deaminase, RNA-Specific

ADARB1: Adenosine Deaminase, RNA-Specific, B1

ADN: Acide désoxyribonucléique

ADNc: Acide désoxyribonucléique complémentaire

ARN: Acide ribonucléique

ARNm: Acide ribonucléique messager

BCR: Biospecimen Core Resource

BLCA: *Bladder urothelial carcinoma* (carcinome urothélial de la vessie)

BRCA: *Breast invasive carcinoma* (carcinome invasif du sein)

C: cytidine

CAMERA: *Correlation Adjusted Mean Rank*

CESC: *Cervical and endocervical cancers* (cancers cervicaux et endocervicaux)

CNV: *Copy Number Variation* (variation du nombre de copies d'un gène)

FDR: *False Discovery Rate* (Taux de faux positifs)

G: guanosine

GDACs: The Genome Data Analysis Centers

HNSC: *Head and Neck squamous cell carcinoma* (carcinome spinocellulaire tête et cou)

I: inosine

IFIH1: *Interferon Induced With Helicase C Domain 1*

IFIT: *Interferon-Induced Protein With Tetratricopeptide Repeats*

INF: interféron

KO: *knock out* (gène désactivé)

LFC: log fold change

LIHC: *Liver hepatocellular carcinoma* (carnicome hépatocellulaire du foie)

log2: logarithme en base 2

LUAD: *Lung adenocarcinoma* (adenocarcinome du poumon)

LUSC: *Lung squamous cell carcinoma* (carcinome spinocellulaire du poumon)

miRNA: microARN

NCI: National Cancer Institute

NHGRI: National Human Genome Research Institute

OAS: *Oligoadenylate Synthetase-Like*

PARP: Poly (ADP-Ribose) Polymerase 1

PCA: analyse en composantes principales

PRAD: *Prostate adenocarcinoma* (adénocarcinome de la protsate)

read: représente une séquence de transcrit alignée sur le génome

RNA-seq: séquençage d'ARN

RSEM: *RNASeq by Expectation Maximization*

SINE: *short interspersed element* (petits éléments nulcéaires intercalés)

STAD: *Stomach adenocarcinoma* (adénocarcinome de l'estomac)

STAT1: *Signal Transducer And Activator Of Transcription 1*

TCGA: The Genome Cancer Atlas (L'atlas génomique du cancer)

THCA: *Thyroid carcinoma* (carcinome de la thyroïde)

UCEC: *Uterine Corpus Endometrial Carcinoma* (carcinome de l'endomètre (corps utérin))

UTR: *untranslated region* (région non traduite)

VOOM: *variance modeling at the observational level*

Vercauteren Drubbel Alizée
Année académique 2015-2016

Table des matières

Vercauteren Drubbel Alizée
Année académique 2015-2016

Introduction

1. Les cancers

Un cancer est une pathologie qui résulte de la prolifération anormale de cellules ne répondant plus aux signaux régulateurs. Dans le cas des tumeurs solides, cette prolifération anarchique entraine la formation d'une masse appelée tumeur qui peut être agressive, envahir les tissus, former des métastases et entrainer la mort. Dans le cas des leucémies, les cellules sanguines anormales envahissent la moelle osseuse pour se répandre dans la circulation sanguine et le système lymphatique et peuvent également envahir les organes vitaux.

En 2012 à travers le monde, 14.1 millions de nouveaux cas de cancers ont été diagnostiqués (sans prendre en comptre les cancers de la peau non mélanomiens). On estime que 8.2 millions de personnes sont décédées du cancer et 32.6 millions en étaient atteintes à travers le monde en 2012. Plus de la moitié des cas nouvellement diagnostiqués et des morts répertoriés le sont dans les régions moins développées (Globocan 2012). Le cancer est donc un problème majeur de santé publique dans le monde entier et l'étude des mécanismes associés à son apparition et/ou sa progression est fondamentale.

Il existe plus de 200 formes de cancers et bien plus de sous-types tumoraux. Chacun d'eux est causé par plusieurs mutations dans l'ADN entrainant des gains et/ou des pertes de fonctions de gènes. De nombreuses mutations peuvent mener au cancer, celles-ci sont très hétérogènes. Les identifier et comprendre comment elles interagissent pour mener à la maladie est donc un objectif majeur dans le but d'améliorer la prévention contre le cancer, d'établir un diagnostic précoce et d'élaborer des traitements. L'apparition récente des technologies de séquençage haut débit (NGS) a révolutionné l'étude génomique des tumeurs. Depuis leur apparition en 2005, le coût, le débit et la qualité de ces technologies s'améliorent à un rythme soutenu. À l'heure actuelle, elles permettent ainsi d'étudier, rapidement et à moindre coût, les séquences nucléotidiques de l'ADN afin d'identifier les modifications génomiques retrouvées dans les cancers. Cependant, l'ADN séquencé n'est pas une mesure exhaustive de toute l'information contenue dans une cellule. En effet, l'expression des gènes codés par l'ADN diffère selon le type cellulaire et les stimuli environnants. Il est donc tout aussi important d'étudier les gènes transcrits et les modifications post-transcriptionnelles que ces transcrits subissent. De nombreux mécanismes post-transcriptionnels sont à prendre en considération, parmi ceux-ci on retrouve l'édition de l'ARN (Bass, 2002).

2. L'édition de l'ARN:

L'édition de l'ARN est un mécanisme qui entraine la conversion d'un nucléotide en un autre dans une séquence d'ARN. Elle cible les ARN messagers (ARNm) et prémessagers, les ARN de transfert et les ARN ribosomiques. Dans certains ARNm, l'édition transforme seulement un à deux nucléotides, dans d'autres plus de 50% des résidus sont modifiés (Gott and Emeson, 2000). De ce fait, elle crée ou altère un site d'épissage, déstabilise la structure de l'ARN affectant ainsi ses interactions avec d'autres protéines, modifie la localisation ou la traduction de l'ARNm entrainant la modification de la fonction de la protéine (Bass, 2002).

Ce faisant, l'édition modifie l'information encodée initialement par le génome et diversifie l'ensemble des transcrits (Bass, 2002). Mais elle a également un impact important sur les fonctions et régulations cellulaires.

2.1. Mécanismes moléculaires impliqués dans l'édition

Jusqu'à ce jour, deux types d'édition ont été décrits pour les ARNm, chacun d'eux implique la désamination de nucléotides. L'un transforme la cytidine (C) en uridine (U) tandis que l'autre transforme l'adénosine (A) en inosine (I) (Bass, 2002). La forme d'édition la plus fréquente chez l'humain est de type adénosine-inosine (A-I). Ce type d'édition est catalysée par la famille d'enzymes adénosine désaminase qui agissent directement sur l'ARN (ADAR). L'enzyme se lie aux ARN double brin et transforme une adénosine en inosine qui sera interprétée comme une guanosine (G) par la machinerie cellulaire (Bass, 2002; Bass et al., 1997).

Ce phénomène peut avoir lieu dans le noyau, cibler les ARN prémessagers et précéder l'épissage de l'ARN. Il peut également se passer dans le cytoplasme et agir directement sur les ARNm (Gott and Emeson, 2000). Cependant, seulement 0.4% des événements d'édition affectent des séquences codant pour des protéines (Mannion et al., 2014; Peng et al., 2012). Plus de 99% de ces phénomènes ont lieu dans les séquences Alu situées dans les introns et les régions 3' UTR non traduites des transcrits. Ces séquences non codantes font partie de la famille des SINE (petits éléments nucléaires intercalés) et représentent 10% du génome. Les séquences Alu sont donc fréquentes (plus d'un million de séquences présentes dans le génome humain), répétées tout le long du génome et en particulier dans les régions riches en gènes. Il est donc probable que deux séquences Alu en palindromes localisées sur le même ARN fassent une structure double brin et deviennent ainsi la cible d'un phénomène d'édition (Levanon et al., 2004). L'étude réalisée par Bazak et al. a détecté plus d'un million de sites

d'édition dans le génome humain et présente un modèle selon lequel toutes les adénosines dans les séquences Alu répétées formant de l'ARN double brin subissent une édition de type A-I bien que la plupart des sites montrent un niveau d'édition inférieur à 1% (Bazak et al., 2014). La fréquence d'édition d'un site particulier représente le nombre de transcrits provenant d'une séquence d'ADN qui sont édités par rapport au nombre de transcrits totaux issus de cette même séquence. La plupart de ces sites édités se situant dans des régions non codantes ou des régions d'éléments répétitifs, la fonction biologique de l'édition dans ces sites reste encore inconnue. Même si de plus en plus de données tendent à montrer qu'elle pourrait avoir un impact sur le traitement des ARN, l'expression des gènes et la déstabilisation de l'ARN double brin (Galeano et al., 2012). Il semblerait également que l'édition régulerait négativement l'expression et l'activité de certains microARN (miRNA) (Nishikura, 2016; Ota et al., 2013). De manière intéressante, ce type d'édition a également été retrouvé dans les transcrits de certains virus (Gott and Emeson, 2000). Il a été proposé que l'édition affecte la façon dont les virus interagissent avec l'hôte, renforçant ou réduisant la croissance virale selon le virus concerné (Samuel, 2011). Les processus biologiques modulés par cette édition sont donc encore peu connus. Cependant, comme mentionné plus tôt les mécanismes moléculaires responsables de l'édition sont un peu mieux caractérisés et reposent sur la famille d'enzymes ADAR.

2.2. Rôle des enzymes ADAR dans l'édition

La famille d'enzyme ADAR est composée de trois protéines paralogues. ADAR (ADAR1) qui est ubiquitaire et possède deux isoformes, l'une est constitutivement active, l'autre est inductible par l'interféron. Les deux isoformes voyagent entre le noyau et le cytoplasme (Nishikura, 2016). ADARB1 (ADAR2) est principalement exprimée dans le cerveau. ADARB2 (ADAR3) n'a pas d'activité d'édition (Gott and Emeson, 2000). Le gène codant pour ADAR est situé sur le bras q du chromosome 1, les deux isoformes qu'il encode lient l'ARN via la reconnaissance d'un motif typiquement présent dans l'ARN double brin et non dans l'ARN simple brin (Patterson and Samuel, 1995). Les enzymes de la famille ADAR sont indispensables à la vie des mammifères. En effet les souris Knock Out (KO) pour ces enzymes se développent de manière anormale. Elles sont victimes de crises convulsives et meurent jeunes (KO pour ADARB1) (Higuchi et al., 2000), ou ne dépassent pas le stade embryonnaire et présentent, entres autres, une induction aberrante de la voie de l'interféron (KO pour ADAR) (Wang et al., 2000). Chez l'humain, les mutations d'ADAR causent le syndrome d'Aicardi-Goutières (AGS). Celui-ci est un désordre inflammatoire, affectant

principalement le cerveau et la peau, caractérisé par une suractivation de la réponse à l'interféron (Rice et al., 2012). Ceci suggère qu'ADAR jouerait également un rôle dans le contrôle de l'expression d'interféron, ce qui a été confirmé par une étude in vitro sur des cellules humaines (Yang et al., 2014). Différentes études ont souligné le fait qu'ADAR était responsable des phénomènes d'édition (Patterson and Samuel, 1995) mais Fumagalli et al. ont été les premiers à montrer, in vitro, qu'il existait un lien direct entre la quantité d'ADAR et la fréquence d'édition de tous les sites potentiellement éditables dans le transcriptome des cellules cancéreuses (Fumagalli et al., 2015).

3. L'édition de l'ARN dans les cancers

Avant 2015 peu d'études sur l'édition de l'ARN dans les cancers avaient été publiées et celles-ci rapportaient des données divergentes sur l'implication de l'édition dans les cancers. Une étude bio-informatique réalisée en 2007 a rapporté une hypoédition des séquences Alu dans le cancer du cerveau, de la prostate, du rein, du poumon et du testicule, comparée aux tissus sains. Cette diminution était couplée à une diminution de l'expression des enzymes de la famille ADAR suggérant qu'une diminution de l'édition de l'ARN dans ces types de cancer jouerait un rôle central dans la tumorigénèse (Paz et al., 2007). En revanche l'édition et ADAR sont augmentées dans les leucémies myéloïdes chroniques suggérant que l'augmentation de l'édition par ADAR joue un rôle dans la progression tumorale en favorisant le développement des progéniteurs myéloïdes malins (Jiang et al., 2013). Pareillement dans les cancers colorectaux l'édition est augmentée dans un site qui pourrait jouer un rôle dans l'invasion (Han et al., 2014). Cependant différents facteurs limitants affectent la qualité de ces études, que ce soit la méthode utilisée pour mesurer les sites d'édition (via les séquences EST) (Paz et al., 2007), le nombre de sites d'édition étudiés (un seul site étudié pour le cancer colorectal) (Han et al., 2014), ou le faible nombre d'échantillons (une dizaine) (Jiang et al., 2013).

Récemment trois études se sont intéressées à l'édition de l'ARN dans les cancers et ont mené une analyse plus systématique, sur plusieurs centaines d'échantillons et sur de nombreux sites d'édition.

Avec la progression des technologies NGS et des outils bio-informatiques, à partir du séquençage d'exome et du séquençage d'ARN il est possible de détecter les sites d'édition à grande échelle. En effet, en alignant les transcrits sur un génome de référence on peut

identifier des mutations A-G (I lu comme un G). Si celles-ci ne sont pas retrouvées dans le séquençage d'exome aligné sur le même génome de référence, cela signifie que la mutation a eu lieu après la transcription et que celle-ci est probablement un site d'édition. La fréquence d'édition d'un site représente la fraction de transcrits édités alignés sur ce site par rapport au nombre de transcrits totaux alignés sur ce même site. Une méthode moins lourde mais moins précise est d'utiliser les sites d'édition déjà connus répertoriés dans les bases de données. Une étape de filtration, à l'aide de bases de données répertoriant des mutations et polymorphismes connus, permet d'éliminer les sites qui sont modifiés par des mutations dans l'ADN plutôt que par l'édition. Cette méthode est moins précise puisqu'elle n'identifie pas de nouveaux sites d'édition et pourrait ne pas identifier des mutations génétiques déjà présentes dans l'ADN dont le transcrit est issu.

Han et al. ont investigué les sites différentiellement édités dans les tumeurs par rapport aux tissus sains dans différents types de cancers répertoriés dans le TCGA. Ils démontrent que dans le cancer du sein, le carcinome tête et cou, le cancer de la thyroïde et l'adénocarcinome pulmonaire, l'édition des sites différentiellement édités dans ces cancers est augmentée dans les prélèvements tumoraux comparés à leur tissu sain (Han et al., 2015).

Paz-Yacoov et al. étudient l'édition de manière plus globale en calculant une édition moyenne des régions Alu afin d'attribuer un degré d'édition à chaque échantillon. Ils démontrent que l'édition est altérée dans les séquences Alu des tumeurs comparées au tissu sain provenant du même patient (Paz-Yaacov et al., 2015).

Fumagalli et al. ont analysé des tumeurs mammaires provenant de l'Institut Bordet. Ils démontrent que l'édition moyenne est positivement corrélée à l'expression d'ADAR. Mais également que l'enzyme agit de façon globale et édite uniformément les mêmes loci à travers les tissus qu'ils soient cancéreux ou sains. Malgré que les sites édités soient identiques dans les tumeurs et les tissus sains, ils identifient une différence dans la fréquence d'édition de ces sites. Ainsi, ils démontrent que la fréquence d'édition pour ces sites dans les tumeurs est augmentée par rapport à celle des tissus sains. Ils démontrent également que l'expression d'ADAR est augmentée dans ces tumeurs (Fumagalli et al., 2015).

Ces trois études réalisées de manière indépendante convergent toutes vers une même idée qui remet en cause les études précédentes : l'édition de l'ARN semble plus fréquente dans les tumeurs que dans les tissus sains. Les mécanismes régulant l'expression d'ADAR dans les tissus sains et les cancers sont méconnus. Cependant, Fumagalli et al. ont investigué les principes gouvernant les processus d'édition principalement dans le cancer du sein. Ils

montrent que 53% des variations de l'expression d'ADAR peuvent être expliquées par le nombre de copies du gène ADAR présent dans le génome et par la voie de l'interféron (Fumagalli et al., 2015).

4. Mécanismes régulant l'expression d'ADAR

4.1. La réponse immune à l'interféron

Un des facteurs activant la voie de l'interféron est la présence d'ARN double brin dans la cellule interprété comme de l'ARN viral par l'organisme. Suite à la reconnaissance du duplex ARN une série de protéines vont être activées en chaîne pour déclencher *in fine* l'expression des gènes codant pour l'interféron de type I et les cytokines pro-inflammatoires (Mannion et al., 2014).

4.1.1. Mécanismes moléculaires de la voie de l'interféron

Les molécules interférons de type 1 (INF) sont une famille de cytokines jouant un rôle médiateur clé dans la réponse de l'organisme contre les pathogènes. L'interféron induit un ensemble de gènes qui déclenchent les effecteurs antiviraux. Parmi ceux-ci on retrouve le gène codant pour la protéine ADAR qui va altérer directement l'ARN viral (Mannion et al., 2014; Samuel, 2011). La réponse à l'interféron est très rapide et a un rétrofeedback positif très fort. En particulier la synthèse de plus d'INF assure qu'une réponse totale soit activée malgré une faible expression d'INF initiale. L'INF est donc un composant clé de l'immunité intrinsèque des cellules et une cellule peut produire, à elle seule, différents types d'interféron (Mostafavi et al., 2016). Les interférons de type I comprennent plusieurs formes, parmi celles-ci on retrouve l'INF α et β qui sont sécrétés par de nombreux types cellulaires infectés par un virus. Les souris déficientes pour le récepteur à l'INF succombent rapidement suite à une variété d'infections virales (Sy et al., 1995). Outre son activité antivirale, l'interféron joue aussi un rôle anti prolifératif et immunorégulateur intervenant dans la surexpression des MHC de classe I, la promotion de la cytotoxicité des cellules *Natural Killer* et le recrutement et la maturation des cellules myéloïdes. Rappelons que les MHC de classe I sont des molécules du complexe majeur d'histocompatibilité. Elles présentent les peptides viraux aux lymphocytes T.

Il existe également l'interféron de type II comprenant l'interféron γ. Celui-ci est principalement produit par les cellules T activées et les *Natural Killer* et joue un rôle majeur dans l'activation de la réponse immunitaire durant les infections avec des pathogènes

intracellulaires telles les infections bactériennes et parasitaires (Stifter and Feng, 2015). La déficience de production d'INF γ entraîne une sensibilité accrue aux mycobactéries.

Lorsque l'INF est sécrété par les cellules infectées, il se lie au récepteur de l'INF des cellules adjacentes saines. Le récepteur de l'INF est hétérodimérique et ses sous-unités sont associées aux kinases Janus JAK1. L'activation de ces kinases entraine la phosphorylation de STAT1 et STAT2 (*Signal Transducer And Activator Of Transcription*). Ces dernières activent la transcription d'une panoplie de gènes capables de déclencher les voies de signalisation inhibant la réplication virale et détruisant le génome viral des cellules infectées (Stark and Darnell Jr., 2012).

Les interférons de type I activent donc STAT1 qui lui même active les processus d'inflammation. L'expression de STAT1 est donc un bon indicateur de la réponse à l'interféron.

4.1.2. La protéine STAT1

STAT1 est un facteur de transcription appartenant à la famille STAT. Son activité dépend de différentes cytokines incluant l'interféron de type I et III. Elle module différents processus cellulaires comme la prolifération, la différentiation et la mort cellulaire et joue un rôle central dans l'immunité adaptative en protégeant l'organisme des infections (Meissl et al., 2015).

L'activation de STAT1 est transitoire et finement contrôlée. Elle est activée et transloquée dans le noyau par les kinases Janus (JAKs) et une fois activée elle régule elle-même sa propre transcription. Son inactivation est contrôlée à différents niveaux, par l'inhibition de JAKs, sa déphosphorylation ou son export nucléaire.

Dans les tumeurs, STAT1 joue un rôle dans la réponse immune innée et adaptative contre les cellules transformées, cependant les conséquences de son expression dans la progression de la maladie divergent selon les cancers. En effet, de manière générale STAT1 est considéré comme un suppresseur de tumeur, cependant de plus en plus d'études démontrent que la fonction de STAT1 semble favorisée par les cellules tumorales. Dans de nombreux cancers, l'expression de STAT1 est corrélée avec un bon pronostic, dans d'autres, c'est le cas par exemple pour le cancer du sein, cette association est plus ambiguë (Meissl et al., 2015).

4.2. L'amplification du bras q du chromosome 1

Comme mentionné précédemment, le nombre de copies d'ADAR présent dans le génome est un autre phénomène pouvant expliquer les variations d'expression d'ADAR.

Chaque chromosome humain est composé d'un bras court nommé p et d'un bras long nommé q. ADAR est un gène situé sur le bras long du chromosome 1. Ce chromosome est le plus grand du génome humain, il contient plus de 2000 gènes codant pour des protéines (EnsEMBL). Il représente à lui seul 6% de l'ADN des cellules humaines. Les réarrangements de ce chromosome sont particulièrement fréquents dans différents types de malignité (Povey and Parrington, 1986).

Les réarrangements chromosomiques consistent entre autres en l'amplification ou la perte d'un segment chromosomique. La taille de ces régions réarrangées est variable et va de quelques paires de bases jusqu'à un chromosome entier. Lorsque ce réarrangement affecte une région englobant des gènes, il entraine des anomalies concernant le nombre d'exemplaires de ces gènes. Dans les tumeurs ce phénomène est fréquent et est mesuré comme la variation du nombre de copies d'un gène (CNV) retrouvée dans les cellules tumorales comparée au nombre de copies trouvé dans les échantillons de cellules saines (Zhang et al., 2014). D'un point de vue fonctionnel une étude sur le cancer du sein a démontré qu'à la fois les amplifications et les délétions du nombre de copies d'un gène avaient un impact sur son expression (Hyman et al., 2002).

Dans plus d'une quinzaine de variétés de cancers chez l'adulte, une région ou l'entièreté du bras q du chromosome 1 est amplifiée (Knuutila et al., 1998a, 2000). Ces types d'amplifications ont également été identifiés dans 9 types de cancers pédiatriques (Puri and Saba, 2014). Cette aberration est plus souvent retrouvée dans les tumeurs récurrentes que dans les tumeurs primaires suggérant que cette amplification est associée à la progression plutôt qu'à l'initiation tumorale (Weith et al., 1996).

Le chromosome 1q comprend plus de 900 gènes et parmi ceux-ci on retrouve des gènes potentiellement impliqués dans la tumorigénèse, et dans les processus cellulaires tels que la différenciation et la prolifération (Puri and Saba, 2014). Dans la plupart des cas, les amplifications 1q retrouvées dans les cancers sont associées à un mauvais pronostique et à des rechutes de la maladie. ADAR, situé sur 1q, est donc fréquemment amplifié dans les cancers. L'analyse du *The Cancer Genome Atlas* (TCGA) montre que sur 20 cancers analysés seuls le cancer du rein et les tumeurs de la thyroïde ont moins de 10% des échantillons présentant une amplification d'ADAR (Fumagalli et al., 2015).

5. *The Cancer Genome Atlas*

Nous allons nous intéresser à l'édition et aux mécanismes la gouvernant dans 11 types de cancers répertoriés dans le TCGA.

La base de données du TCGA (*The Cancer Genome Atlas*) est un outil répertoriant des données concernant le génome de plus de 10 000 échantillons tumoraux humains. Celle-ci a été créée afin d'accélérer la compréhension des mécanismes moléculaires tumoraux à travers l'analyse génomique, dans le but ultime d'améliorer le diagnostic, le traitement et de prévenir le cancer. Elle reprend 11 000 patients pour lesquels un échantillon de tissu tumoral et un échantillon de tissu sain (souvent un prélèvement sanguin, mais parfois également un échantillon de tissu sain adjacent à la tumeur) ont été prélevés. Ces deux échantillons peuvent donc être appariés pour un même patient.

Le *National Cancer Institute* (NCI) en collaboration avec le *National Human Genome Research Institute* (NHGRI) a développé cet outil pour explorer de manière systématique l'ensemble des changements génomiques clés affectant plus de 33 types et sous-types de cancers humains. Le TCGA comprend plusieurs plates-formes « genome wide » qui traitent différents aspects moléculaires des cellules cancéreuses. Ainsi pour chaque échantillon :

- D'un point de vue génétique on retrouve les données concernant:
  - La séquence de l'exome et donc les informations concernant les mutations somatiques
  - La méthylation de l'ADN
  - Les polymorphismes de l'ADN
  - Les variations de nombre de copies de gènes
- Du point de vue de l'expression des gènes on trouve les données concernant :
  - L'expression des ARNm évaluée par microarrays et/ou RNAseq
  - L'expression des micro ARN
  - L'expression de différentes protéines oncologiques

De plus on retrouve également un certain nombre d'informations cliniques concernant les patients. Dans notre étude nous utilisons les données concernant les variations du nombre de copies de gène, les données reprenant l'expression des gènes ainsi que les données cliniques.

Pour l'expression des gènes, nous utilisons les données de RNAseq plutôt que celles de micro-array. Le RNAseq est une méthode récente qui utilise les technologies de séquençage haut débit pour profiler le transcriptôme. Elle fournit une mesure très précise de la quantité de

transcrits et de leurs isoformes présents dans l'échantillon. Celle-ci est nettement plus précise que les micro-arrays et a permis la caractérisation de nombreux transcriptômes (Wang et al., 2009).

L'utilisation de nouvelles technologies et l'immense quantité de données générées ont mené à l'expansion du réseau de recherche du TCGA comprenant de nouveaux centres consacrés à l'analyse de données. Parmi ceux-ci on retrouve The Genome Data Analysis Centers (GDACs) développé par le Broad institute de Harvard. Celui-ci s'attelle à élaborer des outils qui aident les chercheurs à traiter les données à travers l'entièreté du génome et donc à faciliter une large utilisation des données du TCGA. Ainsi les données que nous utilisons dans notre étude proviennent de GDACs firehose version *2015_02_04*.

L'avantage du TCGA est qu'il fournit une caractérisation uniforme d'un grand nombre de cancers (33). Dans notre étude nous ne nous intéressons qu'à 11 types de cancers pour lesquels les données d'édition de l'ARN ont été calculées par Han et al. (Han et al., 2014).

L'identification des sites d'édition est un calcul lourd, et nous avons jugé que réaliser celui-ci pour le reste des cancers répertoriés dans le TCGA n'entrait pas dans le cadre d'un mémoire. En effet, comme mentionné plus tôt, identifier les sites d'édition à partir de données de RNAseq implique à la fois un alignement des données de RNAseq et des données de séquençage d'exome sur le génome de référence. Ceci nécessite0 donc la manipulation de centaines de gigabits de données. Les détails concernant l'identification des sites d'édition par Han et al. peuvent être retrouvés dans la section Matériels&Méthodes.

6. Objectifs de l'étude

Les objectifs poursuivis dans notre recherche sont les suivants :

- Premièrement nous nous sommes attelés à reproduire les résultats concernant la différence d'édition retrouvée entre les tissus normaux et tumoraux établis par Han et al. et Paz-Yaacov dans les données provenant du TCGA. Nous avons ainsi validé la méthode utilisée pour calculer le ratio d'édition global par échantillon.
- Deuxièmement nous avons étendu à 11 types de cancers les résultats de Fumagalli et al. concernant les mécanismes gouvernant l'édition dans le cancer du sein.
    - Ainsi nous investiguons le rôle joué par la voie de l'interféron et l'amplification du nombre de copies d'ADAR sur les variations d'édition observées.

- Nous recherchons également d'autres mécanismes biologiques pouvant être associés à l'édition de l'ARN dans tous les cancers.
- Nous évaluons la possibilité de dériver un métagène dont l'expression permettrait de prédire l'édition de manière précise.

➢ Finalement nous investiguons les conséquences de l'édition sur la progression tumorale en évaluant son impact sur la survie des patients.

Matériel et Méthodes

Toutes les analyses ont été réalisées avec le logiciel libre R version 3.1.2 (2014) développé par *« R Foundation for Statistical Computing »*. R est un langage de programmation qui fournit un environnement pour les calculs statistiques et les analyses graphiques.

Les packages utilisés proviennent du projet Bioconductor. Celui-ci procure des outils spécifiques pour l'analyse et la compréhension de données génomiques en se basant sur le langage de programmation de R.

1. Descriptions des données utilisées

1.1. The Cancer Genome Atlas

La base de données du TCGA (*The Cancer Genome Atlas*) met à disposition gratuitement des données à la fois cliniques et génétiques, concernant plus de 10 000 patients atteins de cancers.

Parmi les données disponibles sur le TCGA nous utilisons les données de séquençage d'ARN normalisées et non normalisées, les données concernant les variations du nombre de copies de gène et les données cliniques. Un résumé du nombre d'échantillons disponibles pour chaque type de données en fonction du type tumoral est disponible table 9.

Ces données sont téléchargées à partir du serveur *The Genome Data Analysis Centers* (GDACs) firehose version *2015_02_04* (Broad Institute TCGA Genome Data Analysis Center (2015): Firehose stddata__2015_02_04 run. Broad Institute of MIT and Harvard. doi:10.7908/C19P30S6) développé par le Broad Institute de Harvard. Plus de détails concernant le TCGA et GDACs peuvent être trouvés dans l'introduction. Les descriptions techniques concernant la plateforme TCGA ont été trouvées sur le site web du National Cancer Intitute Wiki (https://wiki.nci.nih.gov/display/TCGA/TCGA+Encyclopedia).

Pour toutes les données nous n'avons utilisé que les échantillons provenant de tumeurs solides et de tissus sains solides. Nous avons éliminé les échantillons de métastases, de tumeurs récurrentes ainsi que les échantillons sanguins. En effet, comparer nos données aux échantillons sanguins comme tissu sain ne nous semblait pas pertinent car la transcription des gènes est spécifique à chaque tissu.

1.1.1. Description des données de RNA-seq

La procédure de séquençage comprend plusieurs étapes. Tout d'abord l'ARNm est isolé des cellules et rétrotranscrit en ADN complémentaire (ADNc), plus stable, qui est ensuite flanqué d'adaptateurs. Chaque molécule est ensuite séquencée à haut débit. Le séquençage détermine la séquence exacte de l'ADNc. Les séquences obtenues ainsi sont ensuite alignées sur le génome et le nombre de transcrits alignés sur une même région peut être quantifié et utilisé pour évaluer l'abondance d'un transcrit.

Les données de RNAseq provenant du TCGA ont été alignées sur le génome avec le pipeline MapSplice. Les transcrits ont été quantifiés avec RSEM (RNASeq by Expectation Maximization). MapSplice est un algorithme développé pour aligner les reads de RNAseq sur les jonctions d'épissages du génome. Il reconnait de nouvelles jonctions d'épissage indépendamment des sites d'épissage déjà connus. Ainsi, les séquences issues d'un épissage alternatif, déjà répertoriées ou nouvelles peuvent être identifiées (Wang et al., 2010). RSEM permet d'estimer l'abondance relative des transcrits de gènes et de leurs isoformes provenant de données de RNA-seq. RSEM ne nécessite pas l'entièreté du génome de référence. Il suffit de lui fournir des séquences de transcrits de références pour lesquels il quantifie le nombre de reads qui y sont alignés reflétant ainsi l'abondance de ce transcrit et la confiance qu'on peut lui accorder. Un autre avantage de RSEM est qu'il peut traiter les reads qui s'alignent de manière ambigüe entre les gènes et les isoformes, augmentant la précision de l'estimation de l'abondance (Li and Dewey, 2011). Ces données ont ensuite été annotées en utilisant la version du génome et les transcrits GRCh37/hg19 d'ensEMBL.

Dans notre étude nous analysons les décomptes RNA-seq : on assigne à chaque gène le nombre de reads alignés sur ce gène. Nous avons également utilisé les données normalisées par le TCGA, celles-ci représentent la division de chaque décompte par le percentile 75 des décomptes RNAseq pour les +/- 20 000 gènes de cet échantillon. Cette valeur est ensuite multipliée par 1000. La normalisation rend les valeurs plus comparables entre les expériences. Nous avons ensuite transformé ces valeurs de manière systématique en $\log_2$ car la distribution des valeurs d'expression décroit quasi exponentiellement, et dans un souci de cohérence avec la plupart des études RNAseq. Les données de RNAseq non normalisées et normalisées présentaient 30 probes de gènes non identifiés qui ont été systématiquement éliminés.

1.1.2. Description des données de variation du nombre de copies

Les données de CNV répertoriées dans le TCGA ont été obtenues par SNP array Affymetrix (Santa Clara, CA, USA) Genome-Wide Human 6.0 qui contient plus 946 000 sondes pour détecter les variations de nombres de copies (Zhang et al., 2014).

Les anomalies de nombre de copies (CNA) des tumeurs ont été calculées par rapport à des tissus sains, en général des prélèvements sanguins. Un pipeline nommé « CopyNumberInferencePipeline » implémenté par « Cancer Genomics Computation Analysis group » (Broad Institute) transforme les mesures brutes des Affymetrix SNP6.0 en un nombre, le *"segment mean"*. Le *segment mean* est une estimation par échantillon du nombre de fragments, et donc de copies, identifiés à chaque position du génome étudiée. L'algorithme procède comme suit. Premièrement, les intensités brutes sont calibrées afin de pouvoir être comparées entre elles, ensuite une *base line* est créée à partir des échantillons sains diploïdes afin de déterminer la valeur de l'intensité pour aucune copie, une copie (échantillon hétérozygote) et deux copies (échantillon homozygote pour le gène étudié). Les valeurs d'intensités peuvent alors être converties en nombre de copies pour l'entièreté des échantillons étudiés. Il est ensuite possible de calculer le bruit de fond et de l'éliminer du nombre de copies calculé. Un algorithme de segmentation est ensuite utilisé afin d'identifier les régions chromosomiques qui, malgré le bruit, ont un nombre de copies uniforme. L'algorithme segmente les chromosomes de manière à ce que le nombre de copies identifié pour un segment soit constant. En effet, la limite du segment est identifiée lorsque la distribution des sondes du segment est différente de la distribution des sondes voisines. Ces valeurs provenant de sondes situées dans un même segment dans le génome sont moyennées, transformées en $\log_2$ et seront appelées *segment mean*. Les intervalles générés ainsi ne se chevauchent jamais, mais sont contigus. Le nombre de copies est ainsi défini pour chaque position du génome.

Lorsqu'on s'intéresse, comme dans notre cas, à des altérations somatiques il est préférable d'utiliser les données de CNV qui excluent les CNV contenues dans l'ADN germinal. Dans cette étude nous avons donc utilisé les *segment means* des tumeurs corrigées pour les *segment means* des lignées germinales provenant du GDAC firehose (Broad Institue). Une valeur négative indique une perte d'ADN tandis qu'une valeur positive indique un gain. Si le ratio est égal à zéro (ou très proche de zéro), l'échantillon tumoral a le même nombre de copies de cette région génomique que l'échantillon sain.

Afin d'identifier les gènes correspondant aux coordonnées des différentes CNV, nous avons utilisé le site de référence UCSC Genome browser. Sur celui-ci, nous pouvons définir la version d'annotation du génome qui nous intéresse (dans ce cas GRCh37/hg19 d'ensEMBL qui est la version qui a été utilisée pour l'annotation des CNV) et rechercher les coordonnées génomiques du gène d'intérêt. Pour certains gènes il existe plusieurs versions de transcrits possibles. En comparant les corrélations retrouvées entre l'expression du gène et leurs CNV définis selon les coordonnées correspondant à différents isoformes nous n'avons pas trouvé de différence entre les résultats obtenus (données non montrées). Nous avons donc arbitrairement choisi le transcrit d'étendue génomique maximale pour les gènes auxquels nous nous sommes intéressés.

Les coordonnées utilisées pour ADAR sont les suivantes : chromosome 1, début du transcrit=154554534, fin du transcrit =154600456.

1.2. Données d'édition provenant de Han et al.

Les données d'édition que nous utilisons proviennent de l'étude réalisée sur les données du TCGA par le groupe de Liang (Han et al., 2015). Après avoir identifié et réannoté dans les données du TCGA plus d'un million de sites d'édition provenant de *RNA Editing Database* (RADAR), ils ont éliminé environ 4000 sites répertoriés comme des mutations somatiques dans différentes bases de données. De plus, à l'aide de données de séquençage complet (*whole genome sequencing*), provenant du *International Cancer Genome Consortium,* et des données de séquençage d'exomes du TCGA pour les types de cancers étudiés, ils ont évalué s'il y avait des signaux mutationnels situés à des sites d'édition identifiés précédemment. Ainsi ils ont éliminé 0.28% des sites identifiés à travers les cancers étudiés. En se basant sur les données de RNA-seq alignées sur le génome de référence, la fraction de reads édités pour chaque site spécifique dans chaque échantillon donné a été calculée. Cette fraction est égale au nombre de reads alignés sur le site et ayant un 'G', divisé par la somme du nombre de reads ayant un 'G' et du nombre de reads ayant le nucléotide de référence, 'A' à cette même position. Suite à cela, plusieurs conditions devaient être satisfaites pour que le site soit considéré comme un site réel d'édition, plus de détails peuvent être trouvés dans les procédures expérimentales de Han et al. (Han et al., 2015). Cette approche implique que pour chaque type tumoral, les mêmes sites d'édition ont été évalués dans chacun des échantillons. Le nombre de sites d'édition évalués pour chaque cancer peut être retrouvé table 9.

Dans cette étude, nous avons attribué à chaque échantillon un ratio d'édition global calculé comme moyenne des fractions de reads édités à travers tous les sites analysés pour cet échantillon. Les sites sur lesquels aucun read n'est aligné, et donc pour lesquels la fraction de sites édités n'est pas définie, sont bien sûr omis dans ce calcul.

2. Méthodes

2.1. Comparaison entre les tissus sains et tumoraux

Pour un certain nombre de patients du TCGA, du tissu sain adjacent à la tumeur primaire a été prélevé en plus du tissu tumoral. Apparier le tissu tumoral et le tissu sain d'un même patient permet de distinguer les modifications spécifiques à la tumorigénèse de celles qui sont présentes à l'origine chez le patient.

Pour chaque type de cancer étudié nous avons apparié les données d'édition des échantillons sains à leur échantillon tumoral correspondant. De même, les échantillons sains et tumoraux ont été appariés pour les données de RNA-seq afin d'étudier la différence d'expression d'ADAR. Le nombre d'échantillons tumoraux et sains pour chaque type de cancer est disponible table 9.

Le test de Wilcoxon pour échantillons pairés a été utilisé dans chacun des cas afin de calculer la confiance que l'on peut accorder à la différence observée entre les échantillons sains et les échantillons tumoraux. Nous estimons qu'une p-valeur inférieure à 0.05 permet de rejeter l'hypothèse nulle selon laquelle la distribution des données comparées est la même. Finalement le test a été implémenté dans R via la fonction `wilcox.test`.

2.2. Comparaison des données d'expression

L'expression du gène d'intérêt, p.ex. ADAR, STAT1, etc. a été extraite des données RNA-seq normalisées du TCGA, tandis que les données d'édition ont été extraites comme décrit plus haut.

Les corrélations de Spearman ont été calculées via la fonction R, `cor.test`, retournant le coefficient de corrélation, ρ, qui indique l'intensité et la direction de l'association qui existe entre les deux variables. La p-valeur retournée indique la confiance que l'on peut accorder au fait que ces deux variables ne sont pas indépendantes et donc que le coefficient calculé est différent de 0. La corrélation de Spearman permet, par rapport au coefficient de Pearson, de diminuer l'effet massif des outliers, et ne suppose pas une distribution normale des données.

En effet cette corrélation calcul si les variables tendent à changer ensemble mais pas de manière constante (taux constant). Celle-ci se base sur les rangs des valeurs plutôt que sur les valeurs elles-mêmes. Les coefficients de corrélation et les p-valeurs ont été arrondis au chiffre significatif le plus proche via la fonction R `signif`.

La régression linéaire a été utilisée de différentes manières. La régression linéaire estime l'équation d'une droite qui approxime les valeurs observées en minimisant au maximum les écarts entre les valeurs de la droite et celles observées. La fonction R `lm` nous retourne ainsi les paramètres de cette droite, et diverses métriques diagnostiques : la pente c'est à dire comment la variable d'intérêt varie lorsque la variable explicative augmente de une unité ; la confiance que l'on peut accorder au fait que cette pente est différente de 0 et donc que les variables sont dépendantes l'une de l'autre. Elle retourne également le R^2, qui n'est autre que le coefficient de corrélation de Pearson au carré et représente la proportion de variation qui peut être attribuée au modèle calculé.

Nous pouvons également extraire de `lm` les résidus de la droite de régression. Ceux-ci représentent les différences entre les valeurs observées et celles estimées à partir de la droite de régression linéaire. De cette manière, l'expression de l'édition peut être corrigée pour les CNV d'ADAR pour chaque échantillon. En effet, en calculant la régression linéaire entre l'édition et les CNV d'ADAR, et en extrayant les valeurs des résidus, nous obtenons la variabilité qui est indépendante du modèle expliquant la variabilité de l'édition en fonction des CNV. Les ratios d'édition corrigés pour les CNV de ADAR ont donc été considérés comme la somme entre le ratio moyen à travers les échantillons d'un type de cancers et le résidu du modèle de régression linéaire pour l'échantillon.

Le modèle de régression linéaire est généralisable aux analyses multivariées. Ceci est utile pour estimer comment une variable dépend de plusieurs autres variables explicatives. La fonction `lm` permet d'inclure plusieurs variables dans le modèle. Ainsi nous pouvons étudier individuellement la proportion de la variation observée due à chacune des variables explicatives indépendamment, mais également la proportion de la variation observée due à ces variables ensemble.

Finalement, une variante de la fonction `lm` permet d'inclure les interactions entre les variables explicatives dans le modèle si elles existent. Elle va également calculer une p-valeur pour déterminer la confiance que l'on peut accorder au fait que ces covariables interagissent. Ici nous utiliserons cette fonction uniquement pour vérifier que les deux variables étudiées

sont indépendantes. Si la p-valeur retournée par le test de dépendance est supérieure à 0,05 on peut rejeter l'hypothèse nulle selon laquelle ces variables interagissent et assurer qu'elles sont indépendantes. Pour éviter de dissimuler certaines informations étant donné la diversité intertumorale, nous avons évité les représentations abstraites telles que les *barplots* et *boxplots*. Nos graphiques visualisent chaque échantillon individuel, représentant intuitivement la tendance globale, la distribution des données et la présence d'éventuels *outliers*, ainsi que le nombre de données analysées.

2.3. Stabilisation de la variance via Voom

Les mesures de RNA-seq, contrairement aux données de microarrays, produisent des nombres entiers, puisqu'elles correspondent à des décomptes plutôt que des valeurs continues d'intensité. En conséquence, les méthodes développées pour analyser l'expression différentielle des gènes dans les microarrays sont inapplicables directement aux données de RNA-seq pour lesquelles l'estimation de la variance biologique pour des expériences avec un petit nombre de réplicats pose problème. En effet, les décomptes ont souvent une variabilité importante et un nombre de comptes élevé entraine une déviation standard plus grande qu'un nombre de comptes plus petit. À l'inverse lorsque les comptes sont transformés en logarithmes cette variabilité est "trop" atténuée et les décomptes élevés ont une déviation standard plus petite que les décomptes petits. La méthode Voom (Law et al., 2014) a été développée pour pallier ce problème. Elle estime la relation entre la moyenne et la variance de chacune des observations (nombre de reads par gène) du RNA-seq pour lui attribuer un poids précis, de cette manière elle permet de stabiliser la variance (Law et al., 2014). Pour ce faire, elle va estimer la tendance variance-moyenne au niveau des gènes qu'elle va interpoler afin de prédire la variance des observations individuelles. Cette méthode reste performante malgré les différences de profondeur de séquençage entre les échantillons. Voom transformera donc les données brutes non normalisées de RNAseq en log-cpm (log_2-counts per million) auxquelles elle attribuera un poids. Ces résultats sont ensuite traitables par des algorithmes de détection de l'expression différentiels classiques comme Limma (Law et al., 2014), conçus au départ pour les microarrays. Dans une comparaison de dix algorithmes, le traitement par Voom suivi de Limma donne le meilleur contrôle des erreurs de type I pour l'identification de gènes différentiellement exprimés (Law et al., 2014).

La méthode Voom est implémentée à travers la fonction `voom` du package limma de Bioconductor.

2.4. Analyse de l'expression différentielle des gènes par Limma

L'analyse de l'expression différentielle des gènes en fonction d'une variable a été réalisée avec la fonction `limma` du package limma de Bioconductor, dans R (Smyth, 2005). Ce package est conçu de manière à analyser simultanément des comparaisons entre un grand nombre d'ARN cibles dans des expériences au design complexe. En effet, `limma` va comparer la différence d'expression d'un gène à une variable donnée et va répéter cette analyse pour chacun des gènes provenant du RNAseq, ensuite il va appliquer une correction pour test multiple et fournir un « False Discovery Rate » (FDR) ou p-valeur ajustée. Attribuant ainsi un degré de confiance à la différence de *fold change* trouvé en fonction de la variable analysée. L'idée est donc d'utiliser un modèle linéaire pour analyser l'entièreté des données comme un ensemble plutôt que de faire des comparaisons par paire. Ceci a pour effet le partage d'informations entre les échantillons. En effet il va attribuer un poids afin de prendre en compte les variations de précisions entre les différentes observations, et va utiliser le modèle empirique de Bayes pour partager les informations entre les gènes. Ce partage va permettre de fournir des résultats fiables malgré un faible nombre d'échantillons et une variation spécifique à chaque gène. L'approche nécessite de définir une matrice de design qui répertorie les valeurs de la variable continue à laquelle on s'intéresse pour chaque échantillon, dans ce cas-ci l'édition. Une covariable peut éventuellement être définie lorsqu'on veut éliminer son effet sur notre variable dans ce cas-ci les CNV d'ADAR, pour corriger l'effet de l'amplification du chromosome 1q. Dans cette étude nous avons réalisé les deux analyses.

Nos analyses limma sont réalisées sur tous les échantillons disponibles pour chacun des 11 organes couverts par Han et al..

Seuls les gènes significativement différentiellement exprimés ont été gardés (FDR < 0.05). Les tables ont été triées sur la valeur de FDR la plus petite et ensuite sur le LFC le plus grand. Le LFC est le $\log_2$ du Fold Change c'est-à-dire le $\log_2$ de la différence d'expression mesurée entre les échantillons plus et moins édités, s'il est négatif le gène est sous-exprimé, s'il est positif le gène est surexprimé quand la variable augmente.

2.5. Analyse de l'expression différentielle de groupes de gènes via CAMERA

Un test sur un groupe de gènes permet d'analyser l'expression différentielle globale de plusieurs gènes faisant partie d'une même catégorie fonctionnelle. La fonction CAMERA est un test qui met en compétition les groupes de gènes d'intérêt par rapport à tous les autres

gènes (Wu and Smyth, 2012). Il permet donc de détecter des processus biologiques significativement associés à la variable d'intérêt. La procédure CAMERA 'non directionnelle' va tester l'hypothèse nulle selon laquelle la valeur absolue du logFC moyen des gènes se trouvant dans le groupe de gènes d'intérêt est égale à celle des gènes ne se trouvant pas dans le groupe d'intérêt. Si celle-ci est vraie, cela signifie que les gènes dans le groupe d'intérêt ne sont en moyenne pas plus différentiellement exprimés en fonction de la variable que tous les autres gènes ne se situant pas dans le groupe. L'hypothèse alternative étant que la valeur absolue du logFC moyen est plus grande pour les gènes faisant partie du groupe d'intérêt que celle pour les autres gènes. L'hypothèse est émise directement sur les fold changes et non sur les corrélations ou les variances. L'algorithme CAMERA est jusqu'à présent le meilleur pour contrôler l'erreur de type I (qui prédira à tort des groupes de gènes comme différentiellement exprimés) malgré une forte corrélation entre les gènes appartenant au même groupe d'intérêt. En effet le package ajuste le test effectué en fonction de la corrélation calculé entre les gènes d'un même groupe (Wu and Smyth, 2012).

Les résultats retournés par CAMERA nous indiquent les groupes de gènes les plus différentiellement exprimés par rapport à notre variable, le nombre de gènes appartenant à ce groupe, la corrélation entre les gènes de ce groupe, la direction du changement (dans le cas des voies métaboliques si la voie est activée par exemple cette direction sera « Up »), les FDR par groupe. Les données ont été triées en fonction du FDR le plus petit.

Les ensembles de gènes utilisés pour cette étude proviennent de la base de données MSigDB (http://bioinf.wehi.edu.au/software/MSigDB/), d'où les données peuvent être téléchargées directement sous forme d'objet R. Dans cette étude nous ne nous sommes intéressés qu'aux groupes de gènes impliqués dans les voies métaboliques répertoriées dans les bases de données Biocarta, Reactome et Kegg.

Une covariable peut éventuellement être définie lorsque l'on souhaite éliminer son effet sur notre variable. Dans ce cas-ci, afin de corriger l'effet de l'amplification du chromosome 1q nous avons utilisé les CNV d'ADAR comme covariable.

La fonction `Camera` est disponible à partir du logiciel Limma de Bioconductor.

2.6. Identification d'une signature d'édition

Afin de tester s'il est possible d'établir un métagène de prédiction d'édition, les échantillons disponibles pour chaque type de cancers ont été séparés de manière aléatoire en 2 groupes de taille égale. Un groupe d'entrainement « *trainset* » et un groupe de test « *testset* ».

Le « trainset » servira à dériver un métagène lié à l'édition. Le « testset » permettra d'évaluer l'efficacité du métagène trouvé pour prédire l'édition.

La fonction `sample` dans R permet de tirer de manière aléatoire un nombre d'échantillons précisé par l'utilisateur dans un groupe d'échantillons. Pour chacun des cancers séparément nous avons ainsi récupéré les noms d'échantillons appartenant à notre « trainset », les échantillons restants ont été attribués au « testset ».

Les données d'édition et de RNAseq ont été récupérées pour notre « trainset » et les fonctions `voom` et `limma` ont été appliquées comme décrit plus haut, afin d'identifier les gènes différentiellement exprimés en fonction de l'édition dans nos « trainsets ».

Les gènes communs détectés par `limma` ont été extraits via la fonction `intersect` dans R, ceux-ci formaient notre métagène d'édition commun à tous les cancers.

La signature d'édition cancer spécifique a été établie en récupérant les 50 gènes les plus significativement surexprimés avec l'édition dans chacun des « trainsets » des types de cancers. Ceux-ci formaient le métagène d'édition cancer spécifique.

Pour chaque cancer, l'expression du métagène a été mesurée par échantillon appartenant au « trainset ». Pour ce faire, pour chaque échantillon, nous avons récupéré les valeurs de RNAseq normalisées pour chacun des gènes appartenant à la signature, nous avons calculé la valeur médiane et considéré celle-ci comme la valeur d'expression du métagène.

Les valeurs d'éditions pour ces échantillons ont été extraites et corrélées à l'expression calculée pour le métagène.

2.7. Analyse de survie

Une analyse de survie caractérise l'association de variables avec un événement temporel, dans notre cas, la survie du patient. Une spécificité de ce type d'analyse est que l'apparition des événements est mesurée durant un laps de temps limité, dans notre étude, le temps de suivi des patients. Il est donc possible que l'événement n'ait pas été observé durant ce laps de temps mais cela n'implique pas nécessairement que l'événement ne peut se produire hors de la période d'observation. Dans ce cas les patients sont considérés comme sortis de l'étude et l'événement est dit censuré. Lors de l'analyse de survie il est important de prendre en compte à la fois le temps jusqu'à la mort du patient des suites de la maladie (un événement) et celui jusqu'à la fin du suivi du patient (un événement censuré) (Zwiener et al., 2011).

Les données cliniques du TCGA ont été collectées par le *Biospecimen Core Resource* pour chaque participant dont les échantillons tissulaires ont été prélevés. Pour chaque type de cancer, nous avons téléchargé ces données cliniques à partir de firehose GDAC version *2015_02_04*. Pour réaliser l'analyse de survie nous utilisons les données « jours jusqu'à la mort » et « jours jusqu'au dernier suivi » ainsi que le « statut vital du patient ».

Le test du log-rank compare les temps de survie entre les différents groupes tout au long de la période d'observation et indique si une différence entre ceux-ci est identifiée. L'analyse de Cox permet d'analyser l'effet d'une variable continue mais également l'effet simultané de plusieurs variables sur le temps de survie. La régression de Cox se base sur le principe que le *"hazard ratio"* reste constant au cours du temps. Ceci signifie que le ratio du taux de mortalité entre les groupes reste constant et donc que le risque qu'un évènement ait lieu dans un groupe est proportionnel au risque dans l'autre groupe. L'analyse de Cox nous fournit la valeur du *hazard ratio* mesuré selon la variable étudiée. Il mesure à quel point le risque d'un évènement dans un groupe est différent du risque dans l'autre groupe. Ainsi lorsqu'on mesure le ratio du groupe 2 / groupe 1 si ce ratio est plus grand que 1, le taux de mortalité dans le groupe 2 est plus élevé que le taux de mortalité dans le groupe 1 (Zwiener et al., 2011).

L'analyse a été implémentée en R via la fonction `coxph` du package *Suvival* afin d'identifier l'impact de l'édition sur la survie des patients. Seuls les échantillons tumoraux ont été utilisés. L'analyse de Cox a été réalisée en utilisant soit l'édition comme variable continue, soit l'édition comme variable continue et l'expression de STAT1 comme covariable afin de corriger l'effet potentiel de l'inflammation sur la survie.

Résultats

1. Consistance de nos résultats avec les études précédentes

L'édition de l'ARN est un phénomène physiologique qui convertit un nucléotide en un autre dans les ARNs double brin. La séquence de l'ARN est ainsi différente de l'ADN dont elle provient. L'étude du génome tumoral sans prendre en compte ces modifications post-transcriptionnelles, peut donc ne pas être suffisante pour identifier les mécanismes responsables des changements moléculaires observés lors de la tumorigénèse.

1.1. L'édition globale de l'ARN est plus élevée dans les tumeurs comparées au tissu sain correspondant

Pour un certain nombre de patients répertoriés dans le TCGA, du tissu tumoral ainsi que du tissu sain ont pu être prélevés chez un même patient. Le tissu sain est considéré comme le tissu adjacent à la tumeur, provenant donc du même organe que la tumeur étudiée. Ceci permet d'apparier le tissu tumoral et le tissu sain d'un même patient afin d'identifier les différences réellement spécifiques aux tumeurs et non à une différence entre les patients déjà présente dans le génome d'origine.

En octobre 2015 trois études majeures sur l'édition dans les cancers sont publiées.

Han et al. démontrent que l'édition de certains sites bien particuliers est augmentée dans différents cancers comparé à leur tissu sain, incluant BRCA, HNSC, THCA, LUAD (Han et al., 2015). L'étude réalisée par Fumagalli et al. sur des prélèvements tumoraux mammaires provenant de patients de l'Institut Bordet démontre que l'édition globale de l'ARN est augmentée dans ces tumeurs (Fumagalli et al., 2015). Paz-Yacoov et al. étendent cette étude à 9 cancers dont les données proviennent du TCGA, et démontrent qu'il existe une édition altérée dans les sites de séquence Alu retrouvés dans les cancers comparés au tissu sain du même patient (Paz-Yaacov et al., 2015).

Dans notre étude nous utilisons les sites d'édition identifiés par l'étude de Han et al.. La plupart des sites sont situés dans des régions introniques ou 3' UTR. Cependant dans quelques rares cas ces sites d'édition sont retrouvés dans des régions codantes (Han et al., 2015), suggérant que l'édition pourrait avoir une conséquence directe sur la protéine produite. Pour chacun des sites étudiés, un ratio d'édition a été calculé. Celui-ci représente la quantité de transcrits édités alignés sur une position génomique par rapport à la quantité de transcrits

totaux alignés sur ce même site. En calculant la moyenne d'édition de tous ces sites par échantillon nous pouvons avoir une idée globale de la fréquence d'édition du tissu prélevé.

Lorsqu'on compare le degré d'édition moyen retrouvé dans les tissus sains appariés à ces tumeurs, on observe une tendance assez nette. La plupart des tumeurs ont un taux d'édition de l'ARN plus élevé que leur tissu sain correspondant (Figure 1).

Un test des rangs de wilcoxon pour échantillons pairés affirme qu'il existe une différence significative (p-valeur <0.05) entre les valeurs d'édition pour les groupes de tissus tumoraux par rapport aux tissus sains. Les valeurs exactes de ces tests peuvent être retrouvées sur chaque graphique.

Le CESC, le LUSC et l'UCEC ne donnent pas de p-valeurs significatives par manque de données. Cependant à vue d'œil on observe également cette tendance, la majorité des échantillons se retrouve à gauche de la diagonale x=y. Ceci signifie que les taux d'éditions tumoraux situés sur l'axe y, sont supérieurs aux taux d'éditions des tissus sains appariés situés sur l'axe des x.

L'adénocarcinome de la prostate (PRAD), contrairement aux autres cancers, ne montre pas de différence significative entre les valeurs d'édition pour les tissus tumoraux comparés aux tissus sains. De manière cohérente, ce résultat avait également été mentionné par Paz-Yaacov (Paz-Yaacov et al., 2015).

Nos résultats réalisés sur 11 types de cancers reproduisent donc assez fidèlement ceux retrouvés par les trois études réalisées en 2015, validant ainsi la méthode utilisée pour mesurer l'édition globale de l'ARN dans un échantillon.

Dans notre étude nous quantifions l'édition globale de l'ARN d'un échantillon en prenant en compte entre 14 000 et plus de 75 000 sites d'édition d'ARN selon le type de cancer étudié. Nous démontrons également qu'une méthode relativement simple permet de comparer ce taux d'édition global de l'ARN entre les tissus tumoraux et sains de manière reproductible dans différents types de cancers. Finalement nous prouvons que l'édition est globalement plus élevée dans plusieurs types de cancers suggérant que celle-ci pourrait être considérée comme un phénomène analogue aux mutations génomiques.

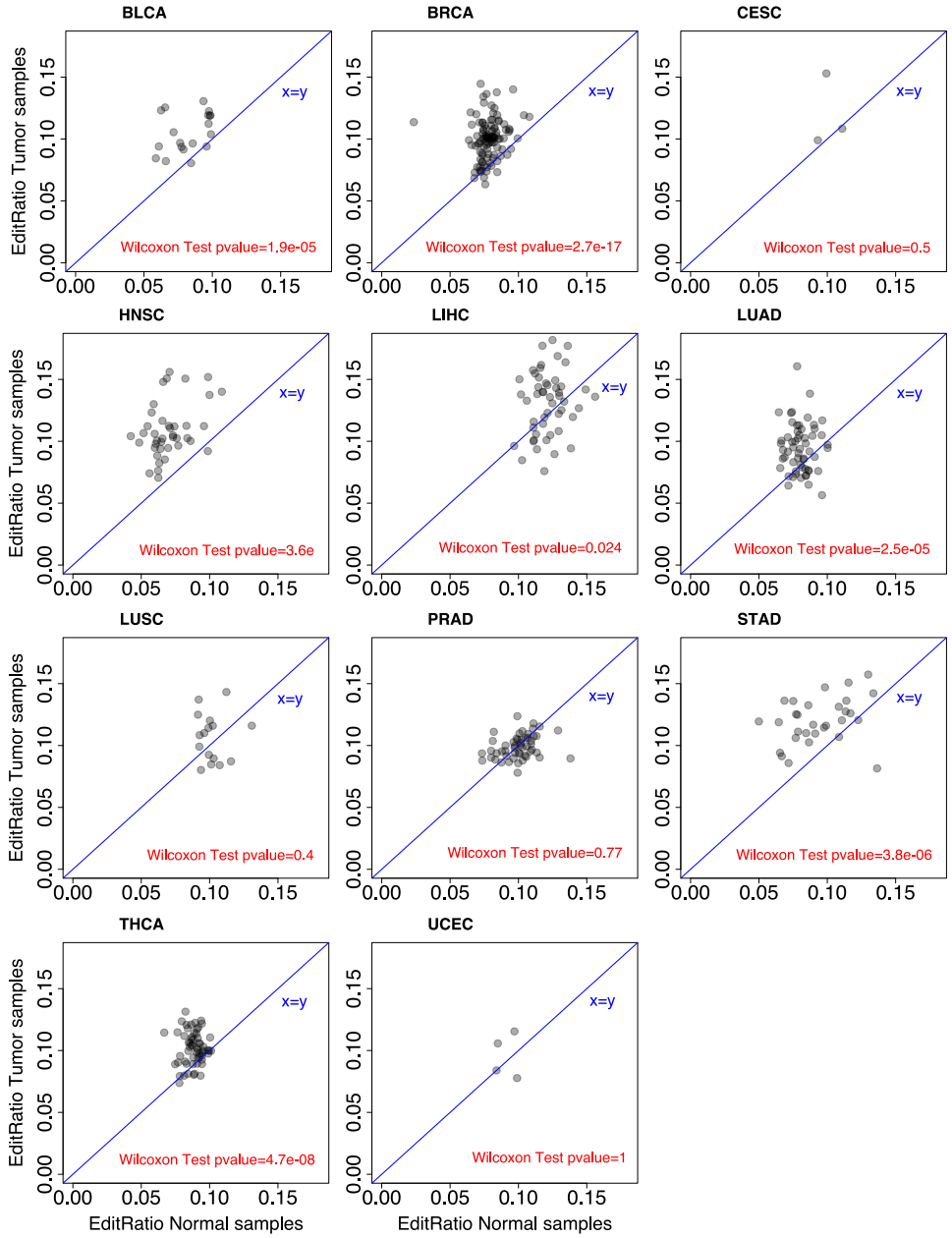


Figure 1: Ratio d'édition dans les tumeurs comparées à leurs tissus sains. Chaque point représente un patient. En abscisse la moyenne du ratio d'édition de son tissu normal, en ordonnée la moyenne du ratio d'édition de son tissu tumoral.

1.2. L'expression d'ADAR est augmentée dans les tissus tumoraux par rapport à leurs tissus sains correspondants

Une raison évidente qui pourrait expliquer l'augmentation de l'édition retrouvée dans les cancers serait une augmentation de l'expression d'ADAR. Comme mentionné précédemment ADAR est responsable de la déamination des A en I des ARN double brin, phénomène définissant l'édition. Une augmentation de l'expression d'ADAR avait été démontrée par Paz-Yaacov dans 8 types de tumeurs. Afin de démontrer que, l'augmentation de l'édition de l'ARN que nous observons dans les cancers concorde avec un changement dans l'expression d'ADAR, nous étudions l'augmentation de l'expression d'ADAR dans les cancers comparé à leur tissu sain apparié.

Ainsi nous confirmons qu'il y a également une tendance bien nette pour la plupart des cancers (Figure 2). L'expression d'ADAR est augmentée et le test des rangs de wilcoxon pour échantillons pairés affirme qu'il y a bien une différence entre les valeurs mesurées pour les tumeurs et les tissus sains. Les résultats des tests peuvent être retrouvés sur les graphiques.

Pour PRAD et LUSC la tendance est moins nette à visualiser mais nous constatons qu'il y a plus d'échantillons tumoraux surexprimant ADAR par rapport à leurs tissus tumoraux correspondants. De manière intéressante les données de RNAseq, comprenant plus d'échantillons, font ressortir une différence significative (p-valeur < 0.05) entre les échantillons tumoraux et sains de LUSC et UCEC. Cancers pour lesquelles, la différence ne pouvait être affirmée pour les données d'édition. Néanmoins CESC a toujours trop peu d'échantillons sains ce qui nous empêche de tirer des conclusions significatives.

Cette étude démontre donc qu'une augmentation de l'expression d'ADAR est observée dans la plupart des tissus tumoraux comparés à leurs tissus sains.

1.3. Il existe une corrélation positive entre l'expression d'ADAR et l'édition globale

Malgré ces observations rien ne prouve que la variation d'ADAR observée est réellement associée à la variation de l'édition observée. Le groupe de Fumagalli et al. a démontré que dans les tissus mammaires sains et tumoraux, l'expression d'ADAR était fortement corrélée à l'édition (ρ=0.7) (Fumagalli et al., 2015). Ici, nous étendons ces résultats à 11 types de cancers afin de confirmer l'association entre ADAR et l'édition de l'ARN.

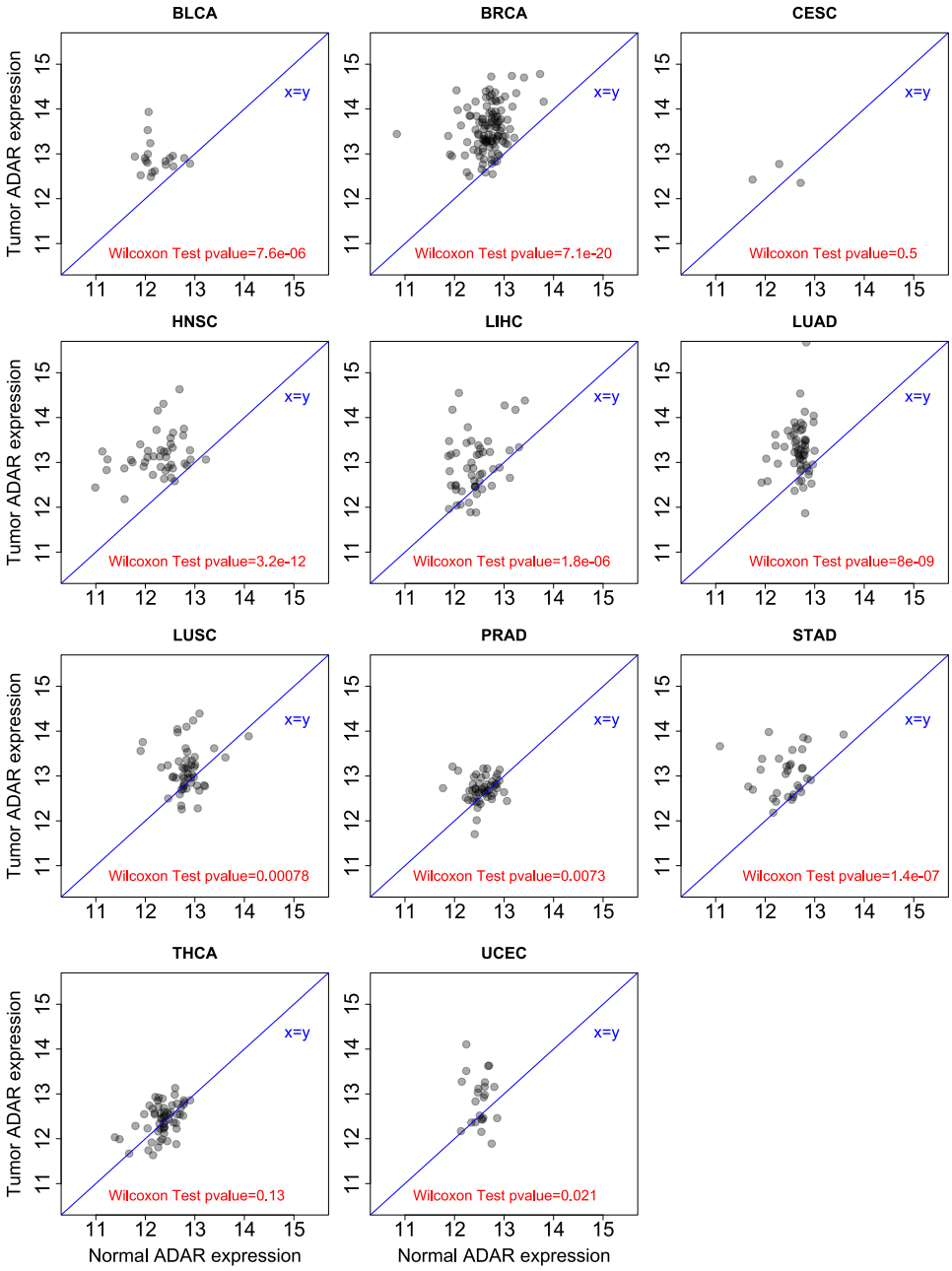


Figure 2: Expression d'ADAR dans les tumeurs comparés à leurs tissus sains. Chaque point représente un patient. En abscisse l'expression d'ARNm d'ADAR pour son tissu sain, en ordonnée l'expression d'ARNm d'ADAR pour son tissu tumoral. Les expressions sont mesurées par RNAseq normalisées et transformées en log2.

L'analyse de l'expression des gènes via RNA-seq nous permet d'étudier la corrélation potentielle existant entre l'expression d'ADAR et le taux d'édition global des échantillons provenant des tissus étudiés.

Nous traitons tous les échantillons sans faire de différence entre le tissu sain ou tumoral puisque nous voulons étudier de manière générale la corrélation entre l'édition de l'ARN et ADAR. Ceci a pour avantage, l'étude d'un plus grand nombre d'échantillons et ainsi l'augmentation de notre puissance statistique. Les données de RNA-seq proviennent de la base de données du TCGA tandis que les données d'édition proviennent, comme mentionnées précédemment, de Han et al. (Han et al., 2015).

La majorité des corrélations de Spearman trouvées sont hautement significatives $<2x10^{-16}$, et les corrélations sont fortes, variant entre 0.8 pour le cancer du sein et 0.5 pour les cancers de la prostate et de la thyroïde (Figure 3).

Seul l'adénocarcinome de l'estomac (STAD) présente un coefficient de corrélation de 0.2, nettement inférieur aux autres, sa p-valeur de $5x10^{-04}$, certes significative, l'est moins que les autres. On pourrait imaginer que dans ce type de cancer l'édition est assurée par une autre enzyme qu'ADAR. ADARB1 étant également connue comme une enzyme ayant une activité d'édition nous avons calculé la corrélation entre son expression et l'édition globale de l'ARN dans STAD. Sa corrélation tout juste significative (p-valeur =0.02) est encore plus faible (ρ=0.1) que celle trouvée précédemment, nous empêchant de conclure qu'ADARB1 a un lien avec l'édition plus fort qu'ADAR.

Cette étude démontre donc que dans 10 cancers sur 11, l'expression d'ADAR est fortement corrélée au taux d'édition global retrouvé dans les tissus. Cependant cette corrélation varie d'un tissu étudié à l'autre.

Les résultats obtenus jusqu'à présent démontrent que l'édition globale est augmentée dans la plupart des cancers et que cette augmentation est très probablement liée à la surexpression d'ADAR dans les cancers.

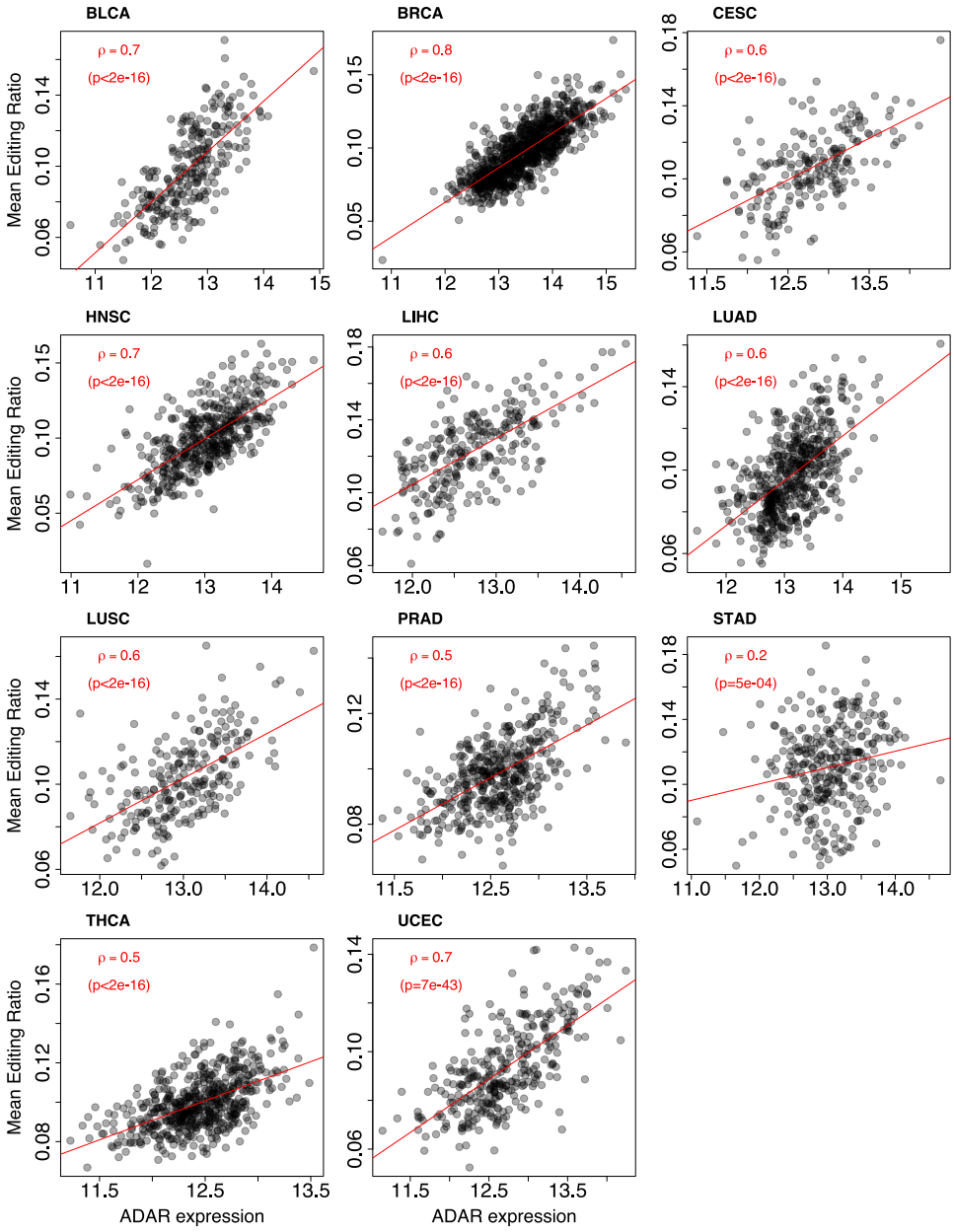


Figure 3: Corrélation entre le ratio d'édition moyen et l'expression d'ADAR. Chaque point représente un échantillon. En abscisse l'expression d'ARNm d'ADAR, en ordonnée la moyenne du ratio d'édition globale par échantillon. Les expressions sont mesurées par RNAseq normalisées et transformées en log2. La droite de régression linéaire en rouge indique la tendance de la distribution. ρ est le

2. La variation du nombre de copies d'ADAR et l'expression de STAT1 expliquent une partie des variations d'éditions de l'ARN

L'édition de l'ARN est connue pour être associée à la réponse antivirale en participant à l'élimination de l'ARN double brin toxiques pour la cellule (Mannion et al., 2014). La voie principale associée à la réponse antivirale est la voie de l'interféron. Celle-ci active systématiquement la protéine STAT1 nécessaire pour déclencher ensuite l'expression des gènes responsables de la signature de la réponse à l'interféron. L'inflammation est un phénomène retrouvé dans tous les cancers et expliquerait en partie l'édition augmentée dans les cancers.

Dans les cancers une mutation très fréquemment retrouvée est l'amplification du bras q du chromosome 1 comprenant plus de 900 gènes, dont ADAR. Cette amplification aurait donc des conséquences sur la variation du nombre de copies d'ADAR et par extension sur son expression.

Les mécanismes qui gouvernent cette édition dans les cancers sont peu connus et ont été principalement étudiés dans le cancer du sein jusqu'à présent (Fumagalli et al., 2015). Fumagalli et al. ont démontré que dans les tissus mammaires étudiés, 53% de la variation d'expression d'ADAR était expliquée par la combinaison de la variation du nombre de copies (CNV) d'ADAR et de la réponse à l'interféron qui a été mesurée par l'expression de STAT1. Puisque nous venons de démontrer que l'expression d'ADAR était significativement corrélée à l'édition globale nous avons étendu leur étude à l'édition de l'ARN et à un plus grand nombre de tissus.

Nous trouvons ainsi que, selon le tissu, 25 à presque 60% de la variation de l'édition peut être expliquée à la fois par les nombres de copies d'ADAR et l'expression de STAT1 (Table 1). Excepté pour l'adénocarcinome de l'estomac pour lequel STAT1 et les CNVs d'ADAR n'expliquent que 10% de la variation de l'édition. Ceci reste en accord avec les données précédentes soulignant que l'expression d'ADAR était peu corrélée à l'édition (ρ=0.2).

Il est important de noter que seulement 1% et 7% de la variation d'édition peut être expliquée par STAT1 respectivement dans LIHC et dans STAD.

De manière intéressante, dans la plupart des cancers, la régression linéaire modélisant à la fois l'effet de la CNV d'ADAR et de STAT1 explique une plus grande proportion de la

Cancer	CNV	STAT1	CNV + STAT1	p-valeur de dépendance
BLCA	0,23	0,29	0,59	0,71
BRCA	0,29	0,2	0,51	0,22
CESC	0,15	0,29	0,43	0,49
HNSC	0,05	0,3	0,44	0,51
LIHC	0,23	0,01	0,25	0,79
LUAD	0,2	0,13	0,39	0,34
LUSC	0,15	0,2	0,34	0,08
PRAD	0,16	0,22	0,37	0,96
STAD	0,03	0,07	0,11	0,58
THCA	0,1	0,2	0,3	0,002
UCEC	0,16	0,2	0,42	0,09

Table 1: proportion de la variation d'édition expliquée par la variation du nombre de copies d'ADAR (CNV), l'expression du gène STAT1(STAT1), la combinaison de la CNV et de STAT1 dans chacun des cancers. Les p-valeurs étaient toutes significatives <0.05 pour l'analyse éffectuée. Les p-valeurs des coefficients de Spearman pour l'association entre STAT1 et l'édition peuvent être retrouvées Figures 4 et 5. "p-valeur de dépendance" est la confiance que l'on peut accorder au fait que la CNV et STAT1 interagissent. Une p-valeur >0.05 nous fait rejeter l'hypothèse nulle selon laquelle les variables dépendent l'une de l'autre.

variation d'édition par rapport à la somme des proportions observées séparément par les deux modèles. Ceci signifie que lorsqu'on modélise séparément l'effet de la CNV et l'effet de STAT1 sur l'édition, on sous-estime l'effet réel de ces deux variables explicatives.

De plus une étude statistique calcule la confiance que l'on peut accorder à la dépendance de ces deux variables. Une p-valeur supérieure à 0.05 nous permet de rejeter l'hypothèse nulle selon laquelle ces deux variables dépendent l'une de l'autre. Ainsi on voit que dans tous les cancers, excepté l'adénocarcinome de la thyroïde, les CNVs et STAT1 varient indépendamment l'un de l'autre (Table 1). Il faut cependant prendre en compte que dans le cancer de la thyroïde les variations du nombre de copies d'ADAR sont très peu nombreuses (Fumagalli et al., 2015). Les variations de l'édition dans ce type de cancer ne peuvent donc être expliquées par celles-ci.

En conclusion, dans la plupart des cancers étudiés, entre un quart et plus de la moitié des variations d'édition observées peuvent être expliquées par l'activation de la voie de l'interféron et par les variations du nombre de copies du gène ADAR qui agissent indépendamment.

3. La correction de l'édition pour la variation du nombre de copies d'ADAR renforce la corrélation avec l'expression de STAT1

Afin d'avoir une idée plus précise des variations individuelles de l'édition et de STAT1 l'une par rapport à l'autre dans chaque échantillon, nous réalisons un graphique du ratio d'édition en fonction de l'expression de STAT1.

Le calcul de la corrélation entre le ratio d'édition et l'expression de STAT1 montre qu'il existe une corrélation significative entre ces deux variables dans tous les cancers étudiés excepté LIHC (Figure 4). Afin d'éliminer l'influence de la CNV d'ADAR dans la mesure de la corrélation entre le ratio d'édition et l'expression de STAT1, nous corrigeons le ratio d'édition via un modèle de régression linéaire. En effet en récupérant les résidus du modèle linéaire expliquant la variation de l'édition par la CNV d'ADAR, nous obtenons les ratios d'édition qui ne sont pas expliqués par les CNVs. Plus de détails sur cet ajustement peuvent être trouvés dans la section Matériel&Méthodes.

Suite à cet ajustement, cette relation ne change pas, voir est renforcée dans tous les cancers. En particulier pour LIHC où la confiance que l'on peut accorder au fait que le

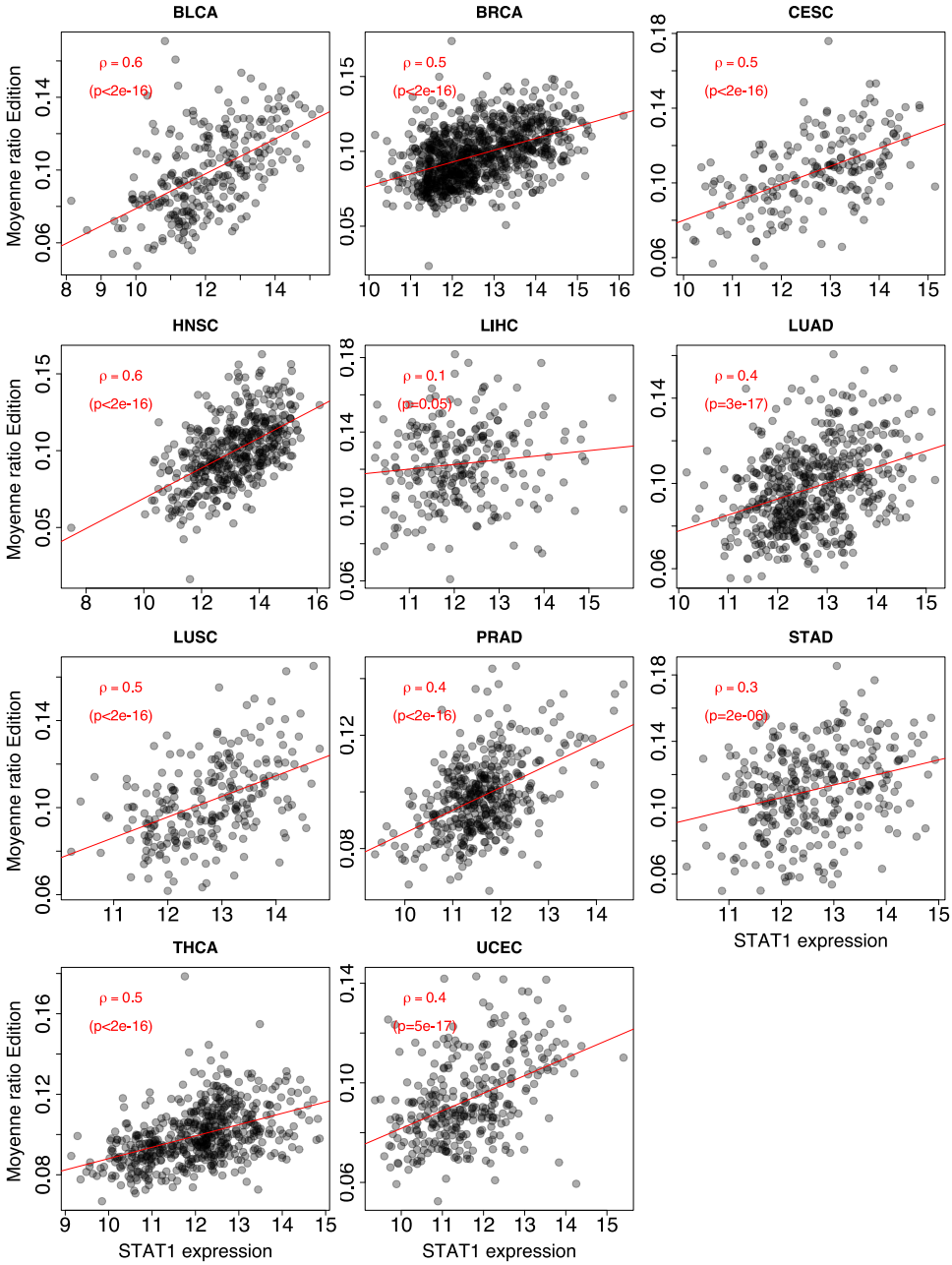


Figure 4: Corrélation entre le ratio d'édition moyen et l'expression de STAT1. Chaque point représente un échantillon. En abscisse l'expression d'ARNm STAT1, en ordonnée la moyenne du ratio d'édition globale par échantillon. Les expressions sont mesurées par RNAseq normalisées et transformées en log2. La droite de régression linéaire en rouge indique la tendance de la distribution. ρ est le coefficient de corrélation de Spearman, p sa p-valeur.

coefficient de corrélation soit différent de 0 augmente et devient significative (Figure 5). Néanmoins le coefficient de corrélation entre l'édition et STAT1 est seulement de 0.2. Ceci reste en accord avec le résultat trouvé précédemment pour LIHC, où STAT1 n'expliquait que 1% de la variance d'édition observée. Ceci suggère que d'autres mécanismes seraient responsables des variations d'édition observées, et que ceux-ci confondent l'effet de l'interféron sur l'édition dans ce type de cancer.

La variation du nombre de copies d'ADAR est uniquement retrouvée dans les cancers pour lesquelles le bras q du chromosome 1 est amplifié, et la réponse à l'interféron est fortement présente dans les cancers. Ces deux phénomènes semblent être, du moins en partie, responsables des variations d'édition observées dans la majorité des cancers. L'édition pourrait donc faire partie d'un phénomène plus large jouant un rôle dans les changements cellulaires observés lors de la tumorigénèse.

4. La voie de la réponse à l'interféron est la plus associée à l'édition dans la plupart des cancers

Nous investiguons maintenant les processus biologiques liés à l'édition de l'ARN de manière plus globale. Ceci nous permettrait de confirmer le rôle de l'inflammation et de découvrir éventuellement d'autres processus biologiques liés à l'édition.

Afin d'identifier les processus biologiques potentiellement associés à l'édition nous utilisons l'algorithme CAMERA (Correlation Adjusted Mean Rank) qui fait une analyse pour des groupes de gènes, plus de détails peuvent être trouvés sur cette méthode dans le matériel & méthodes.

Dans cette étude nous ne nous intéressons qu'aux groupes de gènes jouant un rôle dans les voies métaboliques répertoriées dans les bases de données Biocarta, KEGG et Reactome.

Nous trouvons de manière étonnante que très peu de cancers retournent des voies significativement liées à l'édition. À nouveau il est important de prendre en considération l'amplification 1q qui contient des centaines de gènes (900) jouant un rôle dans plusieurs voies métaboliques. Cette amplification fait donc ressortir certaines voies qui sont en réalité simplement liées à cette amplification et non réellement à l'édition. Cet ajustement retourne de manière générale plus de voies significatives.

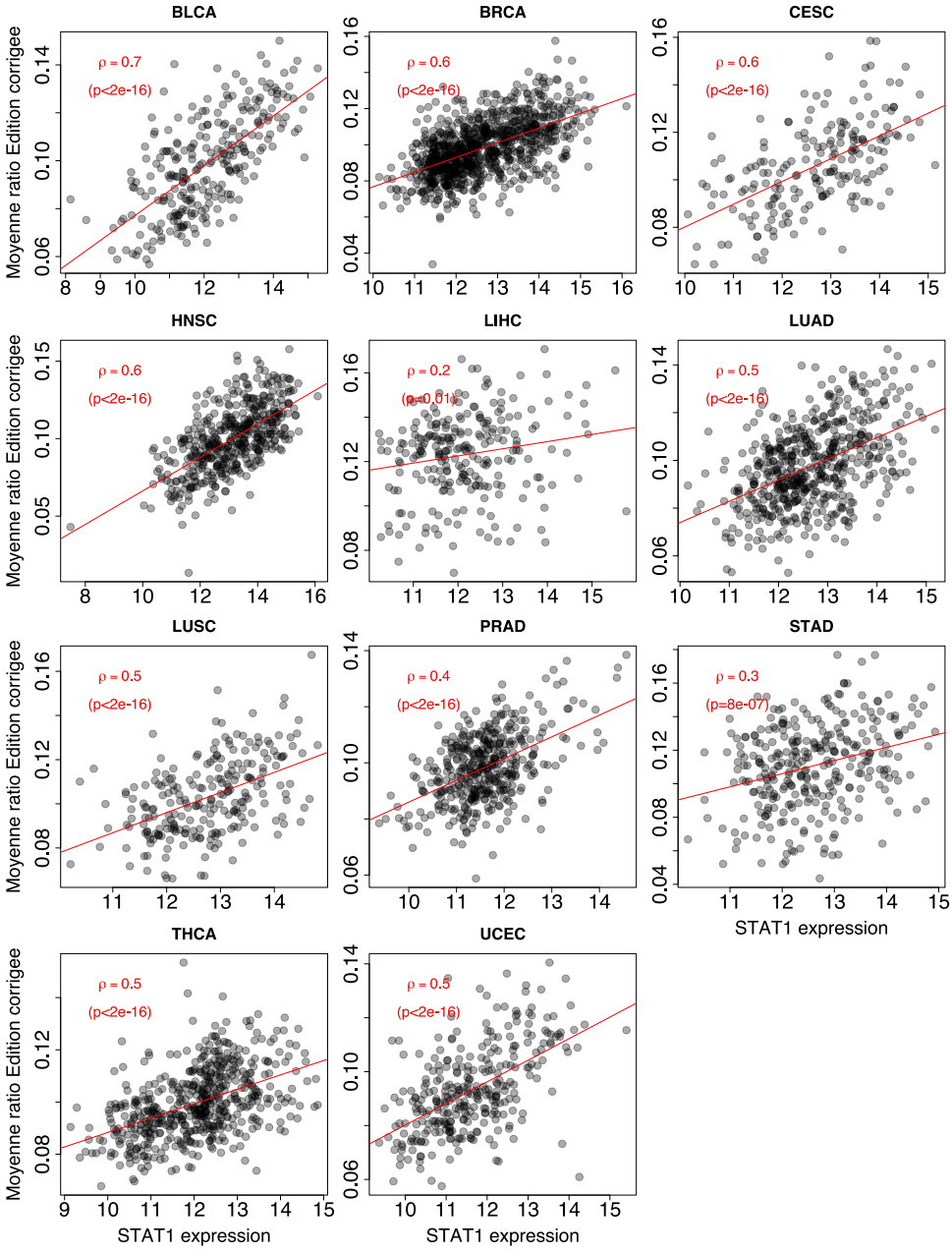


Figure 5: Corrélation entre le ratio d'édition moyen ajusté et l'expression de STAT1. Chaque point représente un échantillon. En abscisse l'expression d'ARNm de STAT1, en ordonnée le ratio d'édition global ajusté pour la CNV d'ADAR. Les expressions sont mesurées par RNAseq normalisées et transformées en log2. La droite de régression linéaire en rouge indique la tendance de la distribution. ρ est le coefficient de corrélation de Spearman, p sa p-valeur.

En accord avec nos résultats précédents, en utilisant un modèle multivarié afin de corriger les expressions pour les CNVs d'ADAR, la voie qui ressort en général comme la plus significativement liée à l'édition est la voie de l'interféron. C'est le cas pour BRCA, BLCA, CESC, HNSC, LUSC, UCEC.

De plus, en accord avec nos résultats précédents cette voie ressort à un rang plus élevé dans LIHC (en 15^{e} position sachant que le FDR le plus bas est en position 1) avec un FDR de 0,22.

La voie de l'interféron n'est pas dans les premiers rangs également pour LUAD (avec un FDR de 0,35 et en 43^{e} position), PRAD (FDR de 0,33 et 46^{e} position), STAD (FDR de 0,98 et 161^{e} position), THCA (FDR de 0,39 et 21^{e} position). Cependant aucune voie ne ressort significativement pour ces cancers en fonction de l'édition. Signifiant sans doute une grande hétérogénéité entre les échantillons étudiés.

Un exemple des 20 premières voies retournées suite à l'analyse CAMERA dans le BRCA peut être trouvé Table 2 pour l'analyse sans correction, Table 3 pour l'analyse corrigée pour la CNV d'ADAR. Les tableaux complets pour tous les cancers étudiés peuvent être retrouvés à l'adresse suivante : https://www.dropbox.com/sh/ee3i9ujkw8d8v1h/AACG_MIZBHn7lA9I7-mws1RDa?dl=0

En conclusion, toutes les voies significativement liées à l'édition de l'ARN dans les différents cancers étudiés sont en lien avec la réponse immunitaire et aucune voie biologique retrouvée ne semble jouer un rôle significatif dans de nouveaux mécanismes totalement indépendants de la réponse immune.

5. Analyse des gènes différentiellement exprimés en fonction de l'édition avec l'algorithme limma

5.1. Des milliers de gènes sont différentiellement exprimés en fonction de l'édition

Étant donné que la recherche de groupes de gènes potentiellement associés à l'édition de l'ARN nous a donné peu de résultats significatifs et qu'aucune nouvelle voie biologique n'a été découverte comme significativement liée à l'édition, nous concentrons notre analyse sur l'étude des gènes individuels.

Afin d'investiguer les gènes potentiellement associés à l'édition nous avons utilisé la fonction `limma` dans R. Celle-ci détecte les gènes dont l'expression différentielle est associée avec l'édition globale de l'ARN.

	NGenes	Direction	FDR
REACTOME_BETA_DEFENSINS	34	Down	0,32699609
REACTOME_OLFACTORY_SIGNALING_PATHWAY	316	Down	0,32699609
REACTOME_DEFENSINS	40	Down	0,32699609
REACTOME_E2F_ENABLED_INHIBITION_OF_PRE_REPLICATION_COMPLEX_FORMATION	10	Up	0,32699609
KEGG_OLFACTORY_TRANSDUCTION	381	Down	0,32699609
REACTOME_POL_SWITCHING	13	Up	0,32699609
REACTOME_CDC6_ASSOCIATION_WITH_THE_ORC_ORIGIN_COMPLEX	11	Up	0,32699609
REACTOME_ACTIVATION_OF_ATR_IN_RESPONSE_TO_REPLICATION_STRESS	35	Up	0,32699609
REACTOME_G2_M_CHECKPOINTS	41	Up	0,32699609
REACTOME_CYCLIN_A_B1_ASSOCIATED_EVENTS_DURING_G2_M_TRANSITION	15	Up	0,32699609
REACTOME_COPI_MEDIATED_TRANSPORT	10	Up	0,32699609
REACTOME_PROCESSIVE_SYNTHESIS_ON_THE_LAGGING_STRAND	15	Up	0,32699609
REACTOME_SYNTHESIS_OF_DNA	90	Up	0,32699609
REACTOME_E2F_MEDIATED_REGULATION_OF_DNA_REPLICATION	33	Up	0,32699609
REACTOME_DEPOSITION_OF_NEW_CENPA_CONTAINING_NUCLEOSOMES_AT_THE_CENTROMERE	60	Up	0,32699609
REACTOME_DNA_REPLICATION	188	Up	0,32699609
REACTOME_MITOTIC_M_M_G1_PHASES	168	Up	0,32699609
REACTOME_S_PHASE	106	Up	0,32699609
REACTOME_REGULATION_OF_MITOTIC_CELL_CYCLE	77	Up	0,32699609
REACTOME_G1_S_TRANSITION	106	Up	0,32699609

Table 2: 20 premières voies retournées suite à l'analyse CAMERA des voies métaboliques associées à l'édition dans le BRCA.

	NGenes	Direction	FDR
REACTOME_INTERFERON_SIGNALING	153	Up	0,02050095
REACTOME_NEGATIVE_REGULATORS_OF_RIG_I_MDA5_SIGNALING	30	Up	0,02050095
REACTOME_INTERFERON_ALPHA_BETA_SIGNALING	62	Up	0,06306016
REACTOME_TRAF3_DEPENDENT_IRF_ACTIVATION_PATHWAY	14	Up	0,07671006
REACTOME_CYTOKINE_SIGNALING_IN_IMMUNE_SYSTEM	263	Up	0,11023133
REACTOME_ANTIVIRAL_MECHANISM_BY_IFN_STIMULATED_GENES	65	Up	0,17309333
REACTOME_INTERFERON_GAMMA_SIGNALING	59	Up	0,17796957
REACTOME_ANTIGEN_PRESENTATION_FOLDING_ASSEMBLY_AND_PEPTIDE_LOADING_OF_CLASS_I_MHC	20	Up	0,17796957
REACTOME_NFKB_ACTIVATION_THROUGH_FADD_RIP1_PATHWAY_MEDIATED_BY_CASPASE_8_AND10	12	Up	0,17796957
REACTOME_ADAPTIVE_IMMUNE_SYSTEM	511	Up	0,17796957
REACTOME_CLASS_I_MHC_MEDIATED_ANTIGEN_PROCESSING_PRESENTATION	236	Up	0,17796957
KEGG_ENDOCYTOSIS	181	Up	0,18884905
KEGG_OLFACTORY_TRANSDUCTION	381	Down	0,38655895
REACTOME_IMMUNE_SYSTEM	886	Up	0,38655895
REACTOME_OLFACTORY_SIGNALING_PATHWAY	316	Down	0,38731428
REACTOME_ENDOSOMAL_VACUOLAR_PATHWAY	8	Up	0,45893633
REACTOME_ANTIGEN_PROCESSING_UBIQUITINATION_PROTEASOME_DEGRADATION	199	Up	0,4662636
REACTOME_BETA_DEFENSINS	34	Down	0,5257747
REACTOME_DEFENSINS	40	Down	0,54305914
REACTOME_INTRINSIC_PATHWAY_FOR_APOPTOSIS	29	Up	0,54305914

Table 3: 20 premières voies retournées suite à l'analyse CAMERA des voies métaboliques associées à l'édition dans le BRCA, en tenant compte de la covariable CNV d'ADAR

En accord avec Fumagalli et al. une bonne partie des gènes les plus significativement différentiellement exprimés trouvés sont situés sur le chromosome 1q, tout comme ADAR (Fumagalli et al., 2015) . Puisque l'amplification 1q est retrouvée dans une large fraction des cancers humains (Knuutila et al., 1998b, 2000) nous détecterons logiquement une association entre l'édition induite par la CNV d'ADAR et ces gènes amplifiés de la même manière qu'ADAR. Nous corrigeons donc le modèle avec la covariable du nombre de copies d'ADAR.

Suite à la correction du modèle pour les CNVs, nous trouvons qu'ADAR et STAT1 sont systématiquement associés à l'édition de manière fortement significative pour tous les cancers, excepté STAD. En effet, la p-valeur étayant qu'il existe réellement une association entre ADAR et l'édition est de maximum 10^{-11} pour CESC, toutes les autres sont inférieures à cela.

En revanche, ADAR n'est pas détectée par le crible de la limma dans l'adénocarcinome de l'estomac. De manière étonnante, nous retrouvons ici, que ADARB1 est associée à l'édition suite à l'analyse limma. Ceci n'est pas en accord avec les données évoquées dans le point 3.2 pour lequel on trouvait un lien moins significatif entre ADARB1 et l'édition qu'avec ADAR. Une étude plus poussée de nos données retournées par la limma, identifie ADAR avec un FDR de 0,053 et un Log Fold Change de 2,6. Ceci ne veut pas dire que le gène n'est en aucun cas lié à l'édition. Cela veut simplement dire que la probabilité de trouver un lien entre ADAR et l'édition dans l'adénocarcinome de l'estomac, par hasard dans un crible du génome, est légèrement supérieure à 5%.

Le fait que, ni ADAR ni ADARB1 ne ressortent comme fortement associés à l'édition, reste intrigant et nous avons donc évalué si nos données pouvaient être biaisées par un effet de batch. L'effet de batch est un biais technique qui apparaît lorsqu'on fusionne plusieurs données obtenues par des techniques de RNA-seq différentes. Celui-ci peut être détecté par une analyse en composante principale (PCA) qui classe les variables responsables de la variation observée dans nos données. Ainsi nous pouvons évaluer s'il existe 2 variables principales, qui seraient ici probablement nos deux méthodes de RNAseq, qui divisent les données de manière évidente. L'analyse PCA sur les données de RNAseq de STAD ne montre aucun effet de batch évident (Annexe Figure 1). Les mécanismes liés à l'édition dans ce cancer semblent donc différents de ceux retrouvés dans les autres cancers étudiés.

Un exemple des 20 premiers gènes retournés suite à l'analyse limma effectuée sur les données du BRCA peut être trouvé Table 4 pour la limma classique, Table 5 pour la limma corrigée pour la CNV. La totalité des tables comprenant tous les gènes significativement

Nom de gènes	log Fold Change	Expression Moyenne	FDR
ADAR	26,10365914	9,235980839	4,15E-176
OAS3	40,08378502	6,871326551	1,58E-80
DTX3L	21,88181876	6,922142494	3,37E-75
PARP9	25,57763198	6,781550541	3,38E-69
UBAP2L	14,47087033	7,849975093	8,15E-65
OAS2	43,40381181	6,413880558	8,15E-65
SLC25A44	13,35414621	5,887559471	3,83E-64
SCAMP3	16,48800094	6,524118615	3,09E-60
OAS1	40,53603572	5,442613396	4,69E-60
DDX60	34,34028582	5,707760779	5,27E-60
DDX58	27,86506456	5,881628584	4,05E-59
PIP5K1A	14,43429075	6,553694584	1,19E-58
IFI6	52,78934246	7,178105431	2,17E-57
PYGO2	14,62827381	6,678206787	1,10E-55
MX1	46,75537274	6,742360739	2,59E-55
UBE2Q1	12,52838673	6,779531738	3,01E-54
YY1AP1	12,36714775	6,963775014	2,62E-53
IFIT1	47,79830781	5,714553431	4,52E-53
EIF2AK2	22,23692218	4,961286645	4,52E-53
TARS2	17,34463787	5,352968319	1,03E-51

Table 4: 20 premiers gènes les plus significativements différentiellements exprimés avec l'édition suite à l'analyse Limma dans le BRCA

Nom de gènes	log Fold Change	Expression Moyenne	FDR
OAS3	54,30833192	6,871326551	8,41E-114
ADAR	21,38294545	9,235980839	8,72E-114
DDX60	51,50192527	5,707760779	5,01E-110
OAS1	58,10942625	5,442613396	2,60E-95
PARP9	34,31146944	6,781550541	4,03E-95
OAS2	59,70482029	6,413880558	3,14E-93
IFIT1	70,652567	5,714553431	2,23E-89
IFIT5	31,13312852	5,01011051	4,53E-86
IFIT3	55,64472832	5,638076482	1,17E-85
DDX58	38,24233388	5,881628584	2,36E-84
DTX3L	26,64731744	6,922142494	3,32E-83
XAF1	47,55728236	5,405197052	1,51E-81
CMPK2	52,35622872	3,709185182	2,12E-81
OASL	64,62765406	3,387517218	5,52E-80
SAMD9	44,75519838	5,230868725	1,66E-79
PARP14	34,40458003	6,956417562	9,84E-79
IFI6	71,17176876	7,178105431	1,05E-77
MX1	63,43522349	6,742360739	3,25E-76
IFI44	52,50343328	4,672086737	9,87E-76
STAT1	39,166253	8,374237628	4,81E-74

Table 5: 20 premiers gènes les plus significativements différentiellements exprimés avec l'édition suite à l'analyse Limma dans le BRCA en tenant compte de la covariable CNV d'ADAR

différentiellement exprimés en fonction de l'édition par cancer peut être trouvée à l'adresse suivante : https://www.dropbox.com/sh/ee3i9ujkw8d8v1h/AACG_MIZBHn7lA9I7-mws1RDa?dl=0

Malgré la correction pour l'amplification 1q, un grand nombre de gènes sont significativement associés à l'édition dans tous les cancers. On retrouve dans le BRCA par exemple ~ 12 400 gènes, c'est à dire plus de la moitié des gènes humains, et le minimum retrouvé est de ~ 2700 gènes pour CESC. Étant donné la quantité de gènes retournés, il est assez difficile d'identifier des gènes communs retrouvés systématiquement dans les cancers comme liés à l'édition. À vue d'œil on retrouve beaucoup de gènes appartenant à la famille IFIT (Interferon-Induced Protein With Tetratricopeptide Repeats). Ceci n'est pas étonnant puisque ce sont des protéines, dont l'expression est induite par l'interféron, qui inhiberaient la réplication et l'initiation de la traduction virale. On retrouve également fréquemment des protéines de la famille OAS (Oligoadenylate Synthetase-Like) également induite par l'interféron qui jouerait un rôle dans la dégradation de l'ARN viral. Les gènes de la famille PARP sont également retrouvés, ceux-ci sont associés à la chromatine et jouent un rôle dans de nombreux processus cellulaires telles la différenciation, la prolifération et la transformation tumorale mais également dans les processus ayant lieu après l'altération de l'ADN (GeneCards).

5.2. Septante-et-un gènes différentiellement exprimés sont communs aux 11 types de cancers

La fonction `intersect` dans R nous identifie de manière plus rigoureuse les gènes communs aux différents cancers en vue d'avoir un signal plus spécifique à l'édition de l'ARN et pas à tous les mécanismes qui sont modifiés en concordance avec celle-ci.

Nous avons ainsi identifié 71 gènes communs à tous les cancers dont l'expression est systématiquement modifiée avec l'édition. ADAR n'en fait pas partie puisque comme expliqué précédemment il n'est pas significatif dans STAD. Ceux-ci sont tous surexprimés (Table 6) sauf 3: KIAA0114, LEUTX, GPR 119.

Parmi ceux-ci on retrouve STAT1, des gènes des familles IFIT, OAS et PARP mentionnées précédemment. Mais on retrouve également IFIH1 (Interferon Induced With Helicase C Domain 1) participant aux processus qui altèrent l'ARN, des gènes de la famille IRF (Interferon Regulatory Factor 1) qui servent comme des activateurs de la transcription de l'interféron.

STAT1	CMPK2	HLA-F	TRIM21
UBE2L6	GBP5	IFIT5	BIRC3
PARP9	IFI44L	MX1	ETV7
PARP14	IFIT2	SP110	BTN3A2
TAP2	RSAD2	SAMD9	ISG15
IFIH1	DDX58	OAS2	TRANK1
IFIT3	IFI6	PSMB8	ERAP1
DTX3L	IRF1	USP18	ZNFX1
DDX60	CXCL11	CSF1	RTP4
NLRC5	APOL6	BATF2	DHX58
NMI	OAS3	DDX60L	LOC100133669
PSMB9	BTN3A1	PRIC285	GBP3
IFI44	PLSCR1	IRF9	C19orf66
EIF2AK2	IFIT1	SP100	OAS1
EPSTI1	GBP4	STAT2	
SAMD9L	HLA-E	RNF213	

Table 6: Signature pan-cancer de l'édition suite à l'analyse Limma des gènes surexprimés avec l'édition dans 11 types de cancers

PARP9
PARP14
IFIH1
DTX3L
PSMB9
CMPK2
OAS3
APOL6
MX1
PRIC285
IRF9
ETV7
ISG15

Table 7: Signature pan-cancer de l'édition suite à l'analyse Limma des gènes surexprimés avec l'édition dans les trainsets des 11 types de cancers

En conclusion l'analyse limma nous a retourné des milliers de gènes dont les modifications étaient significativement associées aux variations d'édition d'ARN. L'édition de l'ARN semble donc faire partie d'un phénomène très large se mêlant à de nombreuses modifications dans l'expression des gènes. Il a été possible d'identifier des gènes liés à l'édition retrouvés dans les 11 cancers mais il reste néanmoins à évaluer à quel point ces gènes sont réellement associés spécifiquement à l'édition.

6. Recherche d'un métagène d'édition : existe-t-il un moyen simple de prédire l'édition en se basant sur l'expression d'une signature d'édition?

Nous avons identifié des gènes communs liés à l'édition dans tous les types de cancers. Nous n'avons, en revanche, pas prouver qu'il était possible de trouver une corrélation entre l'expression de ces gènes et l'édition se distinguant de celle trouvée pour l'expression des autres gènes sortis en réponse à la limma. Démontrer qu'il existe une corrélation spécifique entre ces gènes et l'édition pourrait nous permettre de dériver un métagène, c'est-à-dire une liste de gènes, dont la mesure d'expression prédirait l'édition de l'ARN de manière précise dans les échantillons. Le rôle de l'édition dans les cancers est encore peu connu mais in vitro la diminution de l'expression d'ADAR mène à une diminution de la prolifération et à une augmentation de l'apoptose (Fumagalli et al., 2015) ce qui suggère un effet sur la croissance tumorale. S'il existait une manière simple de prédire le niveau d'édition à partir de l'expression de certains gènes cela simplifierait l'analyse à deux niveaux. Premièrement, lors d'analyse de RNAseq, on éviterait une étape assez lourde de recherche des sites d'édition par alignement du transcriptome que l'on doit ensuite comparer au génome des patients dont il provient. Deuxièmement, on pourrait étendre cette analyse d'édition pour des échantillons analysés par microarray. En effet, l'analyse de l'expression des gènes obtenue par microarray ne fournit pas les séquences des ARN identifiés, ce qui est nécessaire pour la recherche des sites d'édition.

Afin de prédire un métagène lié à l'édition et tester ensuite son efficacité réelle nous avons partitionné nos données.

D'une part, la moitié des échantillons de chaque cancer sont récupérés et appelés « *trainset* ». Celui-ci nous sert de base pour dériver un métagène lié à l'édition.

D'autre part, les données restantes sont appelées « *testset* » et serviront à tester notre métagène trouvé. Comme nous possédons également les données d'édition pour notre *testset* nous évaluerons l'efficacité de notre métagène en analysant le lien entre l'expression du métagène et l'édition. La taille des *trainsets* et par conséquent des *testsets* varient selon le type de cancers en fonction du nombre d'échantillons initialement disponibles (Table 9).

6.1. Intersection des gènes différentiellement exprimés dans tous les cancers

Une fois les données partitionnées de manière aléatoire, nous évaluons les gènes différentiellement exprimés avec l'édition dans chacun des trainsets des cancers via le package limma.

Ensuite, les gènes communs à tous les cancers sont récupérés. Ainsi nous trouvons 13 gènes surexprimés (Table 7), que nous définirons comme le métagène d'édition. La quantité de gènes communs retrouvée ici est nettement inférieure à celle trouvée précédemment, ce qui n'est pas étonnant puisque nous ne travaillons qu'avec la moitié des échantillons.

De manière intéressante, STAT1 et ADAR ne font pas partie de ce métagène. Comme précédemment, ADAR n'est pas significativement lié à l'édition dans STAD tandis que STAT1 n'est pas significativement lié à l'édition dans le trainset LIHC. Ceci reste en accord avec les données trouvées précédemment selon lesquelles STAT1 n'expliquerait que 1% de la variation observée de l'édition dans ce cancer.

6.2. Corrélations entre l'expression de la signature d'édition par rapport à l'édition

Dans un premier temps nous avons testé le métagène prédit sur les données dont il provient, c'est-à-dire les différents trainset spécifiques à chaque cancers.

Pour chacun des échantillons nous avons calculé la médiane d'expression de ce métagène et nous l'avons comparé au taux d'édition moyen de cet échantillon.

Toutes les corrélations de Spearman trouvées entre ce métagène et l'édition sont significatives (< 0.05) pour chacun des types de cancers et aucun coefficient de corrélation ne dépasse 0,6 (Figure 6a).

Détaillant l'analyse du graphique représentant la corrélation la plus élevée retrouvée pour HNSC (Annexe Figure2a), on observe que l'expression du métagène n'est pas un bon prédicteur d'édition de l'ARN. En effet, si l'on prend un seuil d'expression pour lequel on estimerait pouvoir prédire un intervalle de ratio d'édition, on voit que ce taux d'édition varie

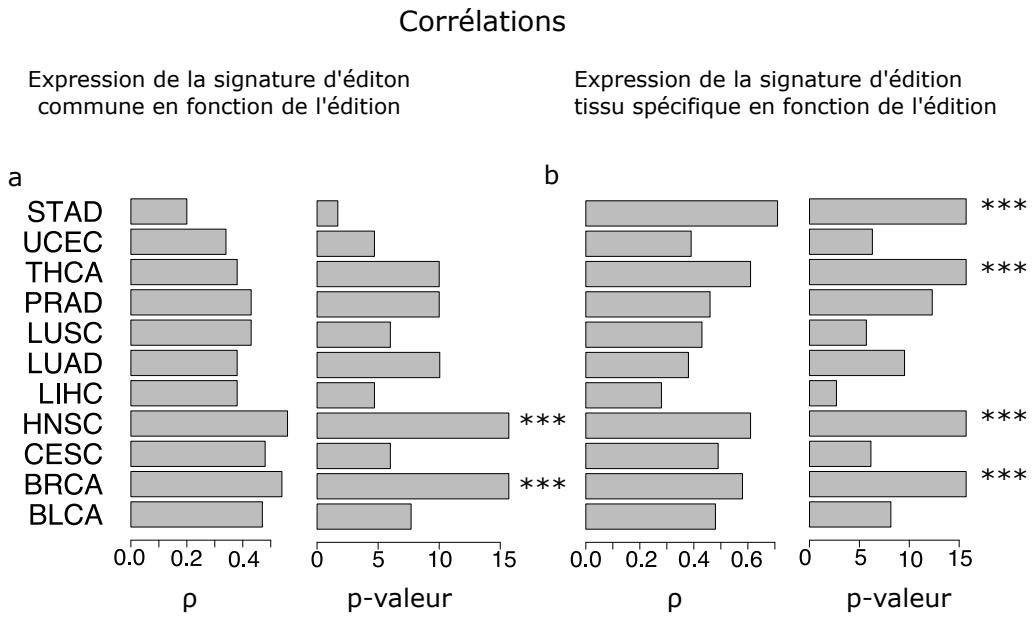


Figure 6: Résultats des corrélations entre l'expression de la signature d'édition et le ratio d'édition moyen. a. résultat de la corrélation de la médiane d'expression des gènes surexprimés avec l'édition communs à tous les cancers en fonction de l'édition dans chacun des cancers b. résultat de la corrélation de la médiane d'expression des gènes surexprimés avec l'édition dans chaque cancer en fonction de l'édition dans ce cancer. ρ est le coefficient de corrélation de Spearman. Les p-valeurs sont exprimées en -log10 et les *** sont inférieures à 2e-16.

énormément d'un échantillon à l'autre. Par exemple, en fixant la médiane de l'expression du métagène à 11, on pourrait prédire un intervalle d'édition d'ARN minimal retrouvé dans l'échantillon. Cependant on observe que les ratios d'édition retrouvés pour cette valeur varient entre ~ 0.06 et ~ 0.16. Ceci représente un intervalle très peu précis étant donné l'intervalle total de variation de l'édition dans tous les échantillons de ce cancer (le ratio d'édition global variant entre moins de 0.06 et plus de 0.16).

Le test de notre métagène sur le trainset est en réalité une estimation optimiste de la qualité du classifieur. Puisque cette estimation optimiste de la capacité du métagène à prédire l'édition est insuffisante, il n'est pas utile d'aller plus loin en testant la qualité de notre classifieur sur les testsets.

6.3. Étude de la corrélation de la médiane d'expression d'une signature d'édition cancer spécifique par rapport à l'édition

La recherche d'un métagène commun a échoué. Cependant, chaque type de cancer étant bien différent, prédire un métagène pan-cancers est certes désirable, mais peut-être trop ambitieux.

Afin d'investiguer l'idée qu'un métagène spécifique par type de cancer pourrait être utilisé pour prédire l'édition, nous avons testé une variation de la procédure précédemment appliquée. Pour chaque type de cancers nous avons dérivé un métagène comprenant les 50 gènes significatifs (FDR < 0.05) les plus surexprimés avec l'édition à partir de son trainset. Chaque métagène a ensuite été testé sur son propre trainset.

Comme précédemment, les corrélations trouvées entre le métagène cancer spécifique et l'édition sont toutes significatives. La corrélation maximale de 0.7 est retrouvée pour STAD (Figure 6b). Pour rappel ADAR ne fait pas partie de cette signature. Cependant, cela n'exclut pas que, les gènes trouvés comme différentiellement exprimés avec l'édition soient fortement corrélés à celle-ci.

De même que précédemment, lorsqu'on fixe un seuil d'expression pour le métagène par exemple de 4, on observe que les variations d'édition retrouvées au-delà de ce seuil varient entre moins de 0.1 et 0.18 (Annexe Figure 2b). De plus, lorsqu'on observe les variations d'édition retrouvées sous ce seuil d'expression de 4 on trouve également un large intervalle allant de moins de 0.06 à plus de 0.14. Sachant que l'intervalle total de variation de l'édition va de moins de 0.06 à 0.18 il sera compliqué de choisir un seuil qui prédira le degré d'édition de manière précise dans les échantillons.

La prédiction de l'édition par un moyen simple semble donc compromise.

Nous pourrions appliquer des méthodes de « machine learning » plus sophistiquées mais nous avons jugé cet investissement de temps superflu en regard du bien possible que la prédiction apporterait. En effet il est toujours possible d'évaluer directement l'édition à partir des données de RNAseq, et celui-ci, de moins en moins couteux et de plus en plus accessible, commence à être utilisé en routine.

7. Les modifications apportées à la survie en fonction de l'édition sont significatives dans les carcinomes spinocellulaire tête et cou

Le rôle de l'édition n'a pas été étudié in vivo dans les cancers et nous ne pouvons donc affirmer que prédire l'édition de manière très précise sur base de l'expression des gènes sera un jour essentiel. Afin d'investiguer l'influence pronostique de l'édition dans les cancers nous avons réalisé une analyse de survie en fonction de l'édition. L'analyse de survie vise à établir une association entre des variables et des évènements survenant durant un intervalle d'observation limité dans le temps. Ici nous utilisons cette analyse pour évaluer l'existence d'une association entre l'édition et la progression du cancer influençant la survie des patients.

L'étude statistique de Cox mesure l'évolution potentielle de la survie en fonction d'une variable donnée qui peut être continue ou discrète. L'utilisation d'une variable continue pour l'édition est plus précise, car elle ne nécessite pas de choix arbitraire d'un seuil d'édition à partir duquel l'échantillon est considéré comme édité.

L'analyse de Cox nous indique que l'édition ne semble pas influencer la survie dans les cancers que nous avons étudiés, excepté pour HNSC. Seul celui-ci a une p-valeur significative de 0.04 et un coefficient positif, suggérant que dans ce type de cancer un taux d'édition élevé mènerait à une mort plus rapide (Table 8).

7.1. Lorsque l'édition est corrigée pour STAT1 l'impact de l'édition sur la survie augmenterait dans l'adénocarcinome thyroïdien

Comme nous l'avons vu précédemment l'édition est fortement liée à la réponse à l'interféron et donc à l'inflammation dans les cancers. La voie de l'interféron présente des conséquences différentes au niveau du pronostic selon les cancers. En effet, dans la plupart des cancers l'augmentation de STAT1 est considérée comme un facteur de bon pronostic, STAT1 étant même considéré comme un gène suppresseur de tumeurs. Cependant dans le

	L'édition		L'édition corrigée pour l'influence de l'inflammation	
	exp(coef)	Pr(>|z|)	exp(coef)	Pr(>|z|)
BLCA	0,057058305	0,48706687	0,023247163	0,44570975
BRCA	0,06587249	0,60002108	0,226546776	0,79517893
CESC	0,445809448	0,91821061	0,032164513	0,72163887
HNSC	1519,1475	0,04490604	29777,633	0,01428071
LIHC	98,80315737	0,36032739	109,4176922	0,35464415
LUAD	1249,958572	0,07517806	164,0343447	0,24076126
LUSC	1,599746392	0,92692984	15,27533993	0,63531128
PRAD	1,716E+15	0,21765083	4,90449E+15	0,24490433
STAD	0,001497827	0,12797556	2,23653E+16	0,25633852
THCA	2162270360	0,20344883	1,14E+20	0,00363787
UCEC	152108,1175	0,13731791	469,448609	0,49541683

Table 8: Etude de l'influence de l'édition sur la survie des patients via l'analyse de Cox. exp(coef) représente le "hazard ratio" qui mesure à quel point le risque d'un évènement dans un groupe est différent du risque dans l'autre groupe. Pr(>z) représente la probabilité de trouver une telle valeur par hasard sous l'hypothèse nulle.

Types de cancers	Nombre de sites édités étudiés	Nombre d'échantillons tumoraux dans les données d'édition	Nombre d'échantillons normaux dans les données d'édition	Nombre d'échantillons tumoraux dans les données de RNAseq	Nombre d'échantillons normaux dans les données de RNAseq	Nombre d'échantillons tumoraux dans les données de CNV d'ADAR	Nombre d'échantillons normaux dans les données de CNV d'ADAR
BLCA	39270	252	19	408	19	407	34
BRCA	76555	837	105	1093	112	1079	133
CESC	32797	196	3	304	3	294	7
HNSC	35510	426	42	520	44	522	74
LIHC	23540	200	50	371	50	367	83
LUAD	54362	488	58	515	59	515	176
LUSC	36822	220	17	501	51	500	238
PRAD	43078	374	52	497	52	491	113
THCA	52701	498	59	501	58	499	95
UCEC	14217	316	4	545	35	536	49
STAD	26389	285	33	415	35	440	95

Table 9: nombre d'échantillons retrouvés dans les différents groupes de données utilisés pour nos analyses

cancer du sein par exemple, son rôle n'est pas aussi net et un des isoformes de STAT1 semblerait même associé à un mauvais pronostic (Meissl et al., 2015). Il y a donc possibilité que l'inflammation confonde l'analyse de survie. En conséquence, nous allons utiliser le modèle multivarié de Cox pour ajuster l'édition pour STAT1.

Suite à cette correction nous obtenons peu de valeurs significatives également. La p-valeur retrouvée pour HNSC égale à 0.01 est plus significative que précédemment.

De plus nous obtenons également une p-valeur significative de 0.004 pour le THCA.

Nous pouvons donc en conclure que dans la plupart des cancers le ratio d'édition global ne semble pas avoir d'influence sur la survie des patients, excepté pour le cancer de la thyroïde et le carcinome tête et cou où la survie semble associée à l'édition (Table 8). Cependant les valeurs trouvées pour ces cancers ne sont pas hautement significatives et une analyse plus poussée avec un meilleur suivi des patients devrait être envisagée avant de pouvoir tirer des conclusions sur de tels résultats.

Ces derniers résultats nous font supposer que prédire l'édition via l'expression des gènes de manière rapide et précise ne devrait pas être au cœur des recherches actuelles, mieux comprendre les rôles que pourraient avoir l'édition dans les cancers devrait rester une priorité.

Discussion & Perspectives

En 2015, trois études sur l'édition de l'ARN dans les cancers ont été réalisées sur les échantillons répertoriés dans la base de données du TCGA et sur des échantillons de tumeurs mammaires provenant de l'Institut Bordet. Ces trois études, utilisant des techniques rigoureuses pour l'identification des sites d'éditions, sont réalisées sur plusieurs centaines d'échantillons et étudient des centaines de sites d'édition. Elles tendent à montrer que l'édition de l'ARN est plus élevée dans la plupart des cancers étudiés suggérant que l'édition de l'ARN dans les cancers est un mécanisme qui pourrait avoir des conséquences importantes dans la tumorigénèse.

À l'heure actuelle le débat sur la reproductibilité des résultats et des recherches précliniques concernant les cancers est ouvert. Les études tentant de reproduire les résultats obtenus lors d'études oncogéniques démontrent que seuls 11% des expériences semblent reproductibles sans nécessairement remettre en cause la confiance que l'on peut accorder aux résultats rapportés (Begley and Ellis, 2012).

Dans notre étude nous reproduisons une partie des résultats trouvés par les trois études menées en 2015. Nous observons, que l'édition, ainsi que l'expression d'ADAR, sont significativement augmentées dans le cancer de la vessie, le cancer du sein, le carcinome spinocellulaire tête et cou, l'adénocarcinome du poumon et l'adénocarcinome de l'estomac. De manière non significative, sans doute par manque d'échantillons normaux, elles sont très probablement augmentées également dans les cancers cervicaux et endocervicaux, et dans le carcinome de l'endomètre (corps utérin). De plus, dans de nombreux cancers, nous confirmons une association entre la voie de l'interféron et l'édition. Et nous démontrons qu'une partie des variations de l'édition observées dans différents cancers peuvent être expliquées par cette voie et par la variation du nombre de copies d'ADAR, comme démontré précédemment dans les cancers mammaires.

La plupart de nos résultats confirment les résultats trouvés par les études réalisées en 2015, validant ainsi notre méthode d'évaluation de l'édition globale de l'ARN dans chaque échantillon. Mais on remarque cependant que l'association entre l'augmentation d'ADAR et l'édition n'est pas systématique. C'est sur ces déviations de la norme que nous allons nous pencher dans la section suivante.

1. Diversité du pattern d'édition dans les cancers

Nos résultats tendent à montrer que la corrélation entre l'expression d'ADAR, l'édition de l'ARN et le pronostic de la maladie dépend fortement du tissu d'origine et donc du type cellulaire.

1.1. L'édition dans le carcinome hépatocellulaire

Dans le carcinome hépatocellulaire (LIHC), l'édition est significativement augmentée (Figure 1). Cependant, l'augmentation de l'édition est retrouvée dans bien moins d'échantillons tumoraux (n= ~27/50) (Figure 1) que l'augmentation de l'expression d'ADAR (n= ~40/50) (Figure 2).

Concrètement, l'édition semble cependant déjà assez élevée dans les tissus hépatiques sains. Nous confirmons cette observation avec un test de Wilcoxon démontrant que l'édition dans les tissus sains hépatiques est plus élevée que l'édition dans les autres tissus sains étudiés (Table Annexe). Fumagalli et al. démontrent que la relation entre ADAR et l'édition peut être modélisée par une fonction logistique et est donc limitée par un seuil de saturation. Lorsqu'on atteint ce seuil de saturation, la surexpression d'ADAR n'augmente plus l'édition (Fumagalli et al., 2015). Il est donc probable que l'édition étant déjà élevée dans les tissus sains hépatiques, ce seuil de saturation serait atteint dans certains échantillons, en conséquence, l'expression d'ADAR n'augmenterait que faiblement l'édition.

L'étude de Fumagalli et al. démontrent qu'il y a une corrélation significativement supérieure à 0.5 entre l'expression d'ADAR corrigée pour la CNV d'ADAR et STAT1 dans ce cancer. Celle-ci est bien plus élevée que la corrélation de 0.2 que nous avons calculée entre l'édition corrigée et STAT1. Ceci étaye notre hypothèse mentionnée précédemment, en démontrant qu'une surexpression de STAT1 est plus associée à une surexpression d'ADAR qu'à une augmentation de l'édition. Ces résultats ont été fidèlement reproduits par notre analyse (Annexe Figure 3a). Un des facteurs de risque les plus importants pour le cancer hépatocellulaire est le virus de l'hépatite C et B. Il est donc probable que la réponse immunitaire soit activée dans les tissus sains afin de combattre ce virus, entrainant une augmentation de l'édition dans ces tissus.

1.2. L'édition dans le cancer de la prostate

Dans le cancer de la prostate (PRAD), ni l'édition de l'ARN ni ADAR ne sont significativement augmentés dans les cancers (Figures 1,2). L'édition et ADAR sont associés

l'un à l'autre comme démontré par leur corrélation hautement significative de 0.5 (Figure 3). Cependant, en analysant chaque patient une édition élevée ne semble pas particulièrement retrouvée dans les cancers plutôt que dans les tissus sains prostatiques.

Ces résultats sont en accord avec les études précédentes. Le groupe de Paz-Yaacov montre qu'il n'y a pas d'augmentation significative de l'index d'édition dans le cancer de la prostate. Le groupe de Han et al. démontre que la proportion de sites différentiellement édités dans ce cancer est quasi nulle et qu'il semblerait y avoir autant de sites dont l'édition est augmentée que de sites dont l'édition est diminuée. Le rôle de l'édition dans la tumorigénèse de ce type de cancer peut donc être remis en question.

1.3. L'édition dans l'adénocarcinome de l'estomac

De manière intéressante, l'édition est augmentée dans STAD comparé à son tissu sain correspondant (Figure 1). Cependant, cette augmentation ne semble que peu associée à la surexpression d'ADAR (ρ=0.2) (Figure 3) malgré que celle-ci soit présente dans les tumeurs (Figure 2). Les variations d'édition sont peu expliquées par l'expression de STAT1 et les CNV d'ADAR. En revanche, on retrouve une corrélation hautement significative de 0.6 lorsqu'on calcule la corrélation entre l'expression d'ADAR, corrigée pour la CNV, et l'expression de STAT1 (Annexe Figure 3b). Finalement, lors de l'analyse limma, ADAR n'est pas détecté comme significativement différentiellement exprimé avec l'édition. Par ailleurs, STAT1 est détecté mais est loin derrière les gènes les plus significativement différentiellement exprimés avec l'édition.

La voie de l'interféron semble donc liée à l'expression d'ADAR dans ce cancer mais ADAR semble avoir peu d'impact sur l'édition. L'édition augmentée dans ce cancer serait donc expliquée par un autre phénomène n'impliquant que peu les enzymes de la famille ADAR (faible corrélation pour ADARB1 également) ou par un phénomène modulant l'activité des enzymes ADAR sans en moduler l'expression (pour revue voir Deffit and Hundley, 2016). Par ailleurs, l'expression d'ADAR pourrait jouer un autre rôle dans ce cancer. Par exemple, il a été démontré qu'ADAR pouvait former un complexe avec Dicer afin d'interagir avec la machinerie responsable des mécanismes de l'interférence d'ARN. Le mécanisme d'interférence de l'ARN utilise les microARNs (miRNA) pour réguler la traduction de certains ARNm en protéine. ADAR favoriserait, entre autres, le développement de miRNA et participerait ainsi à la dégradation d'ARNs cibles (Ota et al., 2013).

1.4. L'édition dans le carcinome thyroïdien

Les résultats pour le carcinome thyroïdien (THCA) démontrent que l'édition est augmentée dans les tumeurs comparées aux tissus normaux correspondants (Figure 1), et que celle-ci est significativement corrélée à ADAR (Figure 3). Étonnamment, l'expression d'ADAR n'est pas augmentée significativement dans les cancers comparés aux tissus normaux (Figure 2).

Ceci suggère qu'ADAR joue un rôle direct dans l'édition, mais qu'il existe probablement un autre mécanisme expliquant une partie de cette augmentation d'édition spécifiquement observée dans ces cancers. Il est également possible que dans certains échantillons l'effet d'ADAR soit renforcé en aval de son expression, entrainant ainsi une augmentation de l'édition sans nécessairement impliquer une forte augmentation de l'expression d'ADAR. Plusieurs études suggèrent que, la régulation de l'accès à la protéine cible et les interactions protéine-protéine altérant directement l'activité d'ADAR, pourraient jouer un rôle dans le contrôle de l'édition (pour revue voir Deffit and Hundley, 2016).

1.5. L'édition dans le carcinome spinocellulaire du poumon

Dans LUSC, l'édition n'est pas augmentée significativement dans les échantillons tumoraux comparés à leurs tissus sains (Figure 1). En revanche, l'expression d'ADAR est significativement augmentée dans la majorité des échantillons tumoraux (Figure 2) et l'édition est significativement associée à l'expression d'ADAR (ρ=0.6) (Figure 3).

Il faut cependant noter que le nombre d'échantillons disponibles pour ces comparaisons est assez faible. En particulier, il y a moins de données disponibles pour l'édition dans les tissus sains (n=17) (Figure 1) que pour les données d'expression de gènes (n=51) (Figure 2). Ceci pourrait être une des raisons qui expliquerait la petite divergence observée entre ADAR et l'édition. Cependant, en accord avec nos résultats, Han et al. avaient démontré que dans ce type de cancer, la proportion de sites différentiellement édités était extrêmement faible (Han et al., 2015). L'analyse graphique nous montre que les échantillons de carcinome spinocellulaire pulmonaire sont assez hétérogènes (Figure 1,3). Ceci est probablement une des raisons rendant une tendance globale, difficile à identifier.

2. Impact potentiel de l'hétérogénéité tumorale sur l'association ADAR / édition dans les cancers

Nos résultats démontrent donc qu'il existe plusieurs cas de figure. Selon le type de cancer, l'édition et les mécanismes la gouvernant ne sont pas toujours les mêmes. Nous retrouvons ainsi des types de cancers où l'édition et ADAR sont augmentées dans la grande majorité des échantillons et fortement corrélés (ρ entre 0.6 et 0.8) c'est le cas pour BLCA, BRCA, CESC, HNSC, LUAD et UCEC. Ces augmentations semblent entre autres gouvernées par la voie de l'interféron et la CNV d'ADAR (Tableau 1), comme prédit par Fumagalli et al. dans le cancer du sein. Dans le cas de PRAD, l'édition et ADAR semblent rester assez constants dans tous les échantillons, les mécanismes gouvernant l'édition sont donc plus compliqués à investiguer. Nous retrouvons également le cas de LIHC pour lequel ADAR est augmenté, cependant l'édition, déjà élevée dans les cellules normales, ne montre pas d'augmentation flagrante dans les cellules cancéreuses. Enfin, nous retrouvons quelques cas de figure pour lesquels les mécanismes régissant l'édition semblent assez hétérogènes d'un échantillon à l'autre (LUSC, THCA). Ces résultats pris ensemble démontrent donc qu'il existe une hétérogénéité entre les différents types de cancers. Mais également qu'une hétérogénéité est présente au sein même d'un type de tumeurs. Cette hétérogénéité est observée graphiquement par la dispersion des échantillons dans l'espace de représentation autour de la droite de régression (Figure 3) par exemple pour PRAD, STAD, LUSC, THCA...

L'hétérogénéité tumorale peut-être la conséquence de la combinaison de différents mécanismes intrinsèques et extrinsèques. Les facteurs extrinsèques englobent le microenvironnement tumoral comprenant les cellules endothéliales, les fibroblastes, les cellules immunitaires et la matrice extracellulaire. Les mécanismes intrinsèques englobent le profil génétique et épigénétique ainsi que potentiellement la cellule à l'origine du cancer. D'une part, de plus en plus d'études démontrent qu'une même mutation dans deux types cellulaires différents provenant du même organe mène à des sous types de cancers différents (pour revue voir Sutherland and Visvader, 2015). D'autre part, l'association entre les profils de mutation et le sous-type tumoral a été démontrée par de nombreuses études utilisant le séquençage d'exomes. Les mutations, dans un même type cellulaire, déterminent initialement le phénotype tumoral (pour revue voir dans Cusnir and Cavalcante, 2012). Cette hétérogénéité peut donc se manifester de différentes manières. Elle se caractérise par des morphologies différentes, par une diversité dans les marqueurs histopathologiques exprimés, et par une divergence de comportement concernant la croissance, l'agressivité tumorale et la réponse aux

thérapies. En fonction du type cellulaire, de la mutation et du microenvironnement, l'interaction complexe, entre les mécanismes moléculaires et cellulaires, entraine une dérégulation des signaux qui favorise l'oncogénèse et influence le pronostic tumoral (pour revue voir dans Sutherland and Visvader, 2015).

Dans le TCGA l'hétérogénéité au sein des mêmes types tumoraux a été démontrée à deux niveaux.

Han et al. analysent le niveau d'édition de 8 sites d'édition engendrant un changement d'acide aminé. Ils montrent que, selon le type de cancer, les niveaux d'édition de ces sites varient entre autres selon le sous-type tumoral et les stades tumoraux.

Cette hétérogénéité est également retrouvée dans l'analyse de l'amplification d'ADAR dans chacun des cancers du TCGA réalisée par Fumagalli et al.. Ils identifient une certaine proportion d'échantillons pour lesquels une amplification est trouvée. Cette proportion varie, selon le type de cancers, de ~ 80% à moins de 10% des échantillons. Par exemple, lorsqu'on s'intéresse au carcinome spinocellulaire du poumon, ~60% des échantillons ont une amplification d'ADAR, les 40% restant ne l'ont pas (Fumagalli et al., 2015). Un gain d'ADAR implique l'amplification 1q qui affecte potentiellement l'expression de 900 gènes dans 60% des tumeurs.

Nous mentionnons ici, des hétérogénéités que nous avons pu mesurer mais il existe probablement bien d'autres phénomènes qui pourraient être concernés. Des études récentes sur le séquençage de génomes tumoraux indiquent que les mutations oncogéniques (*driver mutations*) menant à un même type de cancer sont en réalité très variées (Ciriello et al., 2013). Ainsi, dans la plupart des types de cancers, une mutation oncogénique spécifique n'est retrouvée que dans une certaine proportion des tumeurs. Ceci peut amener de grandes variabilités au niveau des mécanismes moléculaires qui régissent le comportement cellulaire. On sait que la variété des mutations oncogéniques influence la réponse aux thérapies, mais elle pourrait influencer bien d'autres mécanismes tels que l'édition. Il est donc probable que certains des cancers que nous avons étudiés suivent un autre schéma à cause de cette hétérogénéité retrouvée entre les mutations oncogéniques (Greenman et al., 2007).

De nombreuses études démontrent qu'une tumeur n'est pas uniquement composée de cellules cancéreuses. On y retrouve aussi des cellules appartenant au microenvironnement tumoral et influençant grandement la tumorigénèse (pour revue voir Balkwill et al., 2012). Ces cellules sont les fibroblastes, les cellules immunitaires (macrophages, lymphocytes…),

les cellules endothéliales pour la vascularisation tumorale et les cellules de la matrice extracellulaire (pour revue Balkwill et al., 2012). Lors d'une ponction tumorale, on prélève à la fois les cellules du microenvironnement et les cellules tumorales sans distinction. Ces cellules peuvent avoir des génomes différents, exprimer d'autres gènes, avoir des voies métaboliques activées différemment, jouer un rôle dans des mécanismes cellulaires variés. Les analyses retrouvées dans le TCGA sont faites sur la totalité de ces cellules sans distinction. On ne peut donc exclure que certains mécanismes spécifiques aient été noyés dans ces informations variées provenant de différents types cellulaires. Il est même envisageable que la différence d'édition mesurée entre les cellules normales et tumorales provienne d'une augmentation de l'édition dans les cellules composant le microenvironnement plutôt que dans les cellules cancéreuses. Cependant, l'immunohistochimie d'un prélèvement tumoral mammaire démontre que l'expression d'ADAR dans les cellules tumorales luminales est plus élevée que dans l'épithélium sain et dans les lymphocytes (Fumagalli et al., 2015). Les autres cancers étudiés mériteraient plus d'investigation.

3. Rôle de l'édition dans les cancers

Pour certains types de tumeurs la tendance observée est bien nette et peu d'hétérogénéité d'un point de vue de l'édition est présente. On observe peu de dispersion autour des droites de régression linéaire calculées (Figure 3,5) pour BRCA, BLCA, HNSC… Nous pourrions imaginer que pour ce type de tumeurs, l'édition procure un avantage sélectif. Au cours de la tumorigénèse, au sein d'une même tumeur on retrouve des cellules cancéreuses différentes. Il est possible que pour certain type cellulaire, l'édition procure un avantage sélectif entrainant par exemple une prolifération plus rapide de la cellule, lui permettant de générer un clone de cellules filles prédominant, suivant le modèle d'évolution clonale (Greaves and Maley, 2012). La croissance de ces tumeurs serait donc grandement favorisée par les cellules ayant une édition augmentée. Les résultats obtenus, in vitro, sur l'effet antiapoptotique d'ADAR et son effet sur la prolifération dans les cellules mammaires semblent aller dans ce sens (Fumagalli et al., 2015). Han et al. ont également démontré que certaines protéines éditées augmentaient la survie des cellules épithéliales mammaires in vitro (Han et al., 2015).

À ce jour, peu d'études fonctionnelles concernant l'impact de l'édition sur la progression tumorale ont été menées. Une étude concernant le carcinome hépatocellulaire, s'est intéressée à l'édition particulière d'un site entrainant la modification d'un acide aminé et donc de la protéine produite. Ils ont démontré que greffer des cellules exprimant cette protéine modifiée

augmentait significativement l'incidence d'apparition de tumeurs comparées à des cellules exprimant la protéine sauvage (Chen et al., 2013). Suggérant ainsi un rôle de cette édition dans la tumorigénèse des carcinomes hépatocellulaires.

L'impact de l'édition globale sur la progression tumorale a été investigué de manière rétrospective par l'analyse de données cliniques. Paz-Yaacov et al. rapporte qu'une diminution de la survie des patients est significativement associée à une augmentation de l'édition globale dans le cancer hépatocellulaire (LIHC) et le carcinome tête et cou (HNSC). Dans notre étude, nous démontrons que l'édition joue un rôle néfaste sur la survie des patients atteins de HNSC et de THCA. Dans les autres cancers, dont LIHC, nous ne trouvons pas de lien significatif entre la survie et l'édition de l'ARN. Nos données divergent donc en partie avec celles présentées par Paz-Yaacov. Il faut noter que nous avons utilisé l'édition comme une variable explicative continue tandis que Paz-Yaacov et al. ont dichotomisé leurs données en choisissant un seuil d'édition pour lequel l'échantillon était considéré comme édité ou non. Ce seuil, différent selon le cancer, a été choisi de manière à maximiser le contraste de survie entre les deux groupes. En l'absence de données de validation pour établir le seuil, cette méthode ne peut conduire qu'à de l'overfitting. De plus, nous observons que très peu d'évènements ont été pris en considération, en particulier pour HNSC où l'étude pour les échantillons "sous-édités" implique seulement 3 évènements. Finalement nous n'observons aucun évènement censuré comme si ceux-ci n'avaient pas été pris en compte lors de l'étude (Paz-Yaacov et al., 2015). Notre étude est donc manifestement plus robuste et ne permet de reproduire qu'une partie des résultats trouvés précédemment. Malgré cela, il faut prendre en considération, le fait que les données cliniques provenant du TCGA sont parfois assez pauvres. En effet, le suivi des patients du TCGA n'est pas terminé et est donc mis à jour régulièrement. Nos résultats pourraient donc changer si on répète cette étude dans quelques années. De plus, les courbes de Kaplan-Meier sont très variables d'un cancer à l'autre (beaucoup d'évènements censurés dans un mais pas dans l'autre…) (Annexe Figure 4) et celles-ci mériteraient une étude plus approfondie pour prendre en compte tous les éléments pouvant influencer cette étude de survie.

Une meilleure compréhension de l'influence de l'édition dans les cancers nécessiterait donc de caractériser plus précisément les processus d'édition (d'un point de vue génétique et cellulaire) dans chaque type de cancer.

4. Perspectives

Nous avons observé une grande hétérogénéité entre les types de cancers mais également au sein d'un même type de cancer. Celle-ci pourrait être due à la fois aux mutations spécifiques retrouvées dans les cellules mais également à la mixité des cellules retrouvées dans les prélèvements tumoraux.

De futures analyses bio-informatiques permettraient d'évaluer l'influence des mutations génétiques sur l'édition. Nous pourrions, dans un premier temps, uniquement nous concentrer sur des mutations bien connues comme jouant un rôle dans la tumorigénèse (PI3K, P53, KRAS,…). Nous évaluerions alors l'existence potentielle d'une association entre une édition altérée et certaines mutations retrouvées dans un type de cancer. Ceci permettrait, non seulement d'étudier la raison pour laquelle l'édition serait favorisée par certaines mutations mais également d'investiguer les mécanismes associés à l'édition uniquement pour ces sous-types tumoraux.

Pour analyser l'expression génique malgré une hétérogénéité au sein même d'un prélèvement tumoral nous pourrions faire des analyses histologiques. Celle-ci démontrerait que la surexpression d'ADAR mesurée est bien spécifique aux cellules cancéreuses et non à leur micro-environnement, comme cela été montré dans le cancer du sein (Fumagalli et al., 2015). Dans un second temps, pour investiguer les mécanismes liés à l'édition de l'ARN spécifiquement dans les cellules cancéreuses nous pourrions faire du « single cell sequencing ». Cette technique utilise les nouvelles technologies de séquençage pour étudier le génome d'une seule cellule isolée. Nous isolerions ainsi par cytométrie en flux différentes cellules formant la tumeur et les analyserions individuellement. Les données de séquençage d'exome et de RNA-seq nous permettraient d'étudier l'édition dans ces cellules individuelles et d'investiguer l'expression des gènes et par conséquent les mécanismes impliqués.

Finalement, afin d'investiguer le rôle potentiel de l'édition dans la tumorigénèse, des expériences de xenotransplantations sous-cutanées de cellules tumorales humaines dans des souris immunodéprimées seraient réalisées. Ceci permettrait d'étudier, in vivo, la croissance et la propagation tumorale. Pour ce faire, nous grefferions des cellules tumorales humaines pour lesquelles ADAR et/ou le niveau d'édition seraient augmentés. Nous comparerions le développement de ces tumeurs au développement de tumeurs provenant de greffes contrôles. Les greffes contrôle seraient par exemple des cellules tumorales qui auraient un niveau d'ADAR et/ou d'édition plus faible. Ceci démontrerait un éventuel lien entre l'édition et/ou

ADAR et la compétence des cellules à former des tumeurs secondaires, à favoriser la croissance tumorale et potentiellement à former des métastases à distance du site d'injection. Nous pourrions analyser ces résultats en parallèle avec les données cliniques des patients dont proviennent les cellules greffées afin de comprendre de quelle façon ADAR et/ou l'édition influencent potentiellement le pronostic clinique.

Nous pourrions également utiliser des souris transgéniques avec un KO (Knock Out) inductible d'ADAR qui permettrait de choisir le moment où l'on inactiverait le gène codant la protéine ADAR. Le KO induit avant l'initiation tumorale mettrait en évidence le rôle d'ADAR dans l'initiation tumorale. Le KO induit après l'initiation tumorale, c'est-à-dire une fois que les tumeurs sont formées, évaluerait le rôle d'ADAR dans la progression tumorale.

5. Conclusion

En conclusion, l'édition de l'ARN est un mécanisme fréquemment augmenté dans les cancers. Les mécanismes qui en sont responsables doivent encore être approfondis malgré un rôle évident joué par la réponse à l'interféron et l'amplification 1q, tous deux augmentés dans de nombreux cancers. L'édition de l'ARN joue donc très probablement un rôle dans la tumorigénèse. De futures études prenant en considération l'hétérogénéité intratumorale via le single cell sequencing et l'étude des mutations génomiques permettraient de déterminer plus précisément le rôle d'ADAR et de l'édition dans la régulation de l'expression des gènes et de la tumorigénèse.

Bibliographiques

Balkwill, F.R., Capasso, M., and Hagemann, T. (2012). The tumor microenvironment at a glance. J. Cell Sci. *125*, 5591–5596.

Bass, B.L. (2002). RNA Editing by Adenosine Deaminases That Act on RNA. Annu. Rev. Biochem. *71*, 817–846.

Bass, B.L., Nishikura, K., Keller, W., Seeburg, P.H., Emeson, R.B., O'Connell, M.A., Samuel, C.E., and Herbert, A. (1997). A standardized nomenclature for adenosine deaminases that act on RNA. RNA *3*, 947–949.

Bazak, L., Haviv, A., Barak, M., Jacob-Hirsch, J., Deng, P., Zhang, R., Isaacs, F.J., Rechavi, G., Li, J.B., Eisenberg, E., et al. (2014). A-to-I RNA editing occurs at over a hundred million genomic sites, located in a majority of human genes. Genome Res. *24*, 365–376.

Begley, C.G., and Ellis, L.M. (2012). Drug development: Raise standards for preclinical cancer research. Nature *483*, 531–533.

Chen, L., Li, Y., Lin, C.H., Chan, T.H.M., Chow, R.K.K., Song, Y., Liu, M., Yuan, Y.-F., Fu, L., Kong, K.L., et al. (2013). Recoding RNA editing of AZIN1 predisposes to hepatocellular carcinoma. Nat. Med. *19*, 209–216.

Ciriello, G., Miller, M.L., Aksoy, B.A., Senbabaoglu, Y., Schultz, N., and Sander, C. (2013). Emerging landscape of oncogenic signatures across human cancers. Nat. Genet. *45*, 1127–1133.

Cusnir, M., and Cavalcante, L. (2012). Inter-tumor heterogeneity. Hum. Vaccines Immunother. Hum. Vaccines Immunother. *8*, *8*, 1143, 1143–1145.

Deffit, S.N., and Hundley, H.A. (2016). To edit or not to edit: regulation of ADAR editing specificity and efficiency. Wiley Interdiscip. Rev. RNA *7*, 113–127.

Fumagalli, D., Gacquer, D., Rothé, F., Lefort, A., Libert, F., Brown, D., Kheddoumi, N., Shlien, A., Konopka, T., Salgado, R., et al. (2015). Principles Governing A-to-I RNA Editing in the Breast Cancer Transcriptome. Cell Rep. *13*, 277–289.

Galeano, F., Tomaselli, S., Locatelli, F., and Gallo, A. (2012). A-to-I RNA editing: The "ADAR" side of human cancer. Semin. Cell Dev. Biol. *23*, 244–250.

Gott, J.M., and Emeson, R.B. (2000). Functions and Mechanisms of RNA Editing. Annu. Rev. Genet. *34*, 499–531.

Greaves, M., and Maley, C.C. (2012). Clonal evolution in cancer. Nature *481*, 306–313.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G.L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., et al. (2007). Patterns of somatic mutation in human cancer genomes. Nature *446*, 153–158.

Han, L., Diao, L., Yu, S., Xu, X., Li, J., Zhang, R., Yang, Y., Werner, H.M.J., Eterovic, A.K., Yuan, Y., et al. (2015). The Genomic Landscape and Clinical Relevance of A-to-I RNA Editing in Human Cancers. Cancer Cell *28*, 515–528.

Han, S.-W., Kim, H.-P., Shin, J.-Y., Jeong, E.-G., Lee, W.-C., Kim, K.Y., Park, S.Y., Lee, D.-W., Won, J.-K., Jeong, S.-Y., et al. (2014). RNA editing in RHOQ promotes invasion potential in colorectal cancer. J. Exp. Med. *211*, 613–621.

Higuchi, M., Maas, S., Single, F.N., Hartner, J., Rozov, A., Burnashev, N., Feldmeyer, D., Sprengel, R., and Seeburg, P.H. (2000). Point mutation in an AMPA receptor gene rescues lethality in mice deficient in the RNA-editing enzyme ADAR2. Nature *406*, 78.

Hyman, E., Kauraniemi, P., Hautaniemi, S., Wolf, M., Mousses, S., Rozenblum, E., Ringnér, M., Sauter, G., Monni, O., Elkahloun, A., et al. (2002). Impact of DNA Amplification on Gene Expression Patterns in Breast Cancer. Cancer Res. *62*, 6240–6245.

Jiang, Q., Crews, L.A., Barrett, C.L., Chun, H.-J., Court, A.C., Isquith, J.M., Zipeto, M.A., Goff, D.J., Minden, M., Sadarangani, A., et al. (2013). ADAR1 promotes malignant progenitor reprogramming in chronic myeloid leukemia. Proc. Natl. Acad. Sci. U. S. A. *110*, 1041–1046.

Knuutila, S., Björkqvist, A.M., Autio, K., Tarkkanen, M., Wolf, M., Monni, O., Szymanska, J., Larramendy, M.L., Tapper, J., Pere, H., et al. (1998a). DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. Am. J. Pathol. *152*, 1107–1123.

Knuutila, S., Björkqvist, A.M., Autio, K., Tarkkanen, M., Wolf, M., Monni, O., Szymanska, J., Larramendy, M.L., Tapper, J., Pere, H., et al. (1998b). DNA copy number amplifications in human neoplasms: review of comparative genomic hybridization studies. Am. J. Pathol. *152*, 1107–1123.

Knuutila, S., Autio, K., and Aalto, Y. (2000). Online Access to CGH Data of DNA Sequence Copy Number Changes. Am. J. Pathol. *157*, 689–690.

Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29.

Levanon, E.Y., Eisenberg, E., Yelin, R., Nemzer, S., Hallegger, M., Shemesh, R., Fligelman, Z.Y., Shoshan, A., Pollock, S.R., Sztybel, D., et al. (2004). Systematic identification of abundant A-to-I editing sites in the human transcriptome. Nat. Biotechnol. *22*, 1001–1005.

Li, B., and Dewey, C.N. (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics *12*, 323.

Mannion, N.M., Greenwood, S.M., Young, R., Cox, S., Brindle, J., Read, D., Nellåker, C., Vesely, C., Ponting, C.P., McLaughlin, P.J., et al. (2014). The RNA-Editing Enzyme ADAR1 Controls Innate Immune Responses to RNA. Cell Rep. *9*, 1482–1494.

Meissl, K., Macho-Maschler, S., Müller, M., and Strobl, B. (2015). The good and the bad faces of STAT1 in solid tumours. Cytokine.

Mostafavi, S., Yoshida, H., Moodley, D., LeBoité, H., Rothamel, K., Raj, T., Ye, C.J., Chevrier, N., Zhang, S.-Y., Feng, T., et al. (2016). Parsing the Interferon Transcriptional Network and Its Disease Associations. Cell *164*, 564–578.

Nishikura, K. (2016). A-to-I editing of coding and non-coding RNAs by ADARs. Nat. Rev. Mol. Cell Biol. *17*, 83–96.

Ota, H., Sakurai, M., Gupta, R., Valente, L., Wulff, B.-E., Ariyoshi, K., Iizasa, H., Davuluri, R.V., and Nishikura, K. (2013). ADAR1 Forms a Complex with Dicer to Promote MicroRNA Processing and RNA-Induced Gene Silencing. Cell *153*, 575–589.

Patterson, J.B., and Samuel, C.E. (1995). Expression and regulation by interferon of a double-stranded-RNA-specific adenosine deaminase from human cells: evidence for two forms of the deaminase. Mol. Cell. Biol. *15*, 5376–5388.

Paz, N., Levanon, E.Y., Amariglio, N., Heimberger, A.B., Ram, Z., Constantini, S., Barbash, Z.S., Adamsky, K., Safran, M., Hirschberg, A., et al. (2007). Altered adenosine-to-inosine RNA editing in human cancer. Genome Res. *17*, 1586–1595.

Paz-Yaacov, N., Bazak, L., Buchumenski, I., Porath, H.T., Danan-Gotthold, M., Knisbacher, B.A., Eisenberg, E., and Levanon, E.Y. (2015). Elevated RNA Editing Activity Is a Major Contributor to Transcriptomic Diversity in Tumors. Cell Rep. *13*, 267–276.

Peng, Z., Cheng, Y., Tan, B.C.-M., Kang, L., Tian, Z., Zhu, Y., Zhang, W., Liang, Y., Hu, X., Tan, X., et al. (2012). Comprehensive analysis of RNA-Seq data reveals extensive RNA editing in a human transcriptome. Nat. Biotechnol. *30*, 253–260.

Povey, S., and Parrington, J.M. (1986). Chromosome 1 in relation to human disease. J. Med. Genet. *23*, 107–115.

Puri, L., and Saba, J. (2014). Getting a Clue from 1q: Gain of Chromosome 1q in Cancer. J. Cancer Biol. Res. *2*, 1053.

Rice, G.I., Kasher, P.R., Forte, G.M.A., Mannion, N.M., Greenwood, S.M., Szynkiewicz, M., Dickerson, J.E., Bhaskar, S.S., Zampini, M., Briggs, T.A., et al. (2012). Mutations in ADAR1 cause Aicardi-Goutieres syndrome associated with a type I interferon signature. Nat. Genet. *44*, 1243–1248.

Samuel, C.E. (2011). Adenosine deaminases acting on RNA (ADARs) are both antiviral and proviral. Virology *411*, 180–193.

Smyth, G.K. (2005). limma: Linear Models for Microarray Data. In Bioinformatics and Computational Biology Solutions Using R and Bioconductor, R. Gentleman, V.J. Carey, W. Huber, R.A. Irizarry, and S. Dudoit, eds. (Springer New York), pp. 397–420.

Stark, G.R., and Darnell Jr., J.E. (2012). The JAK-STAT Pathway at Twenty. Immunity *36*, 503–514.

Stifter, S.A., and Feng, C.G. (2015). Interfering with Immunity: Detrimental Role of Type I IFNs during Infection. J. Immunol. *194*, 2455–2465.

Sutherland, K.D., and Visvader, J.E. (2015). Cellular Mechanisms Underlying Intertumoral Heterogeneity. Trends Cancer *1*, 15–23.

Sy, H., Pj, H., Ka, H., Sh, S., Mj, T., Ja, H., G, W., I, B., and I, K. (1995). A null mutation in the gene encoding a type I interferon receptor component eliminates antiproliferative and antiviral responses to interferons alpha and beta and alters macrophage responses. Proc. Natl. Acad. Sci. U. S. A. *92*, 11284–11288.

Wang, K., Singh, D., Zeng, Z., Coleman, S.J., Huang, Y., Savich, G.L., He, X., Mieczkowski, P., Grimm, S.A., Perou, C.M., et al. (2010). MapSplice: Accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. *38*, e178–e178.

Wang, Q., Khillan, J., Gadue, P., and Nishikura, K. (2000). Requirement of the RNA Editing Deaminase ADAR1 Gene for Embryonic Erythropoiesis. Science *290*, 1765–1768.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nat. Rev. Genet. *10*, 57–63.

Weith, A., Brodeur, G.M., Bruns, G.A., Matise, T.C., Mischke, D., Nizetic, D., Seldin, M.F., van Roy, N., and Vance, J. (1996). Report of the second international workshop on human chromosome 1 mapping 1995. Cytogenet. Cell Genet. *72*, 114–144.

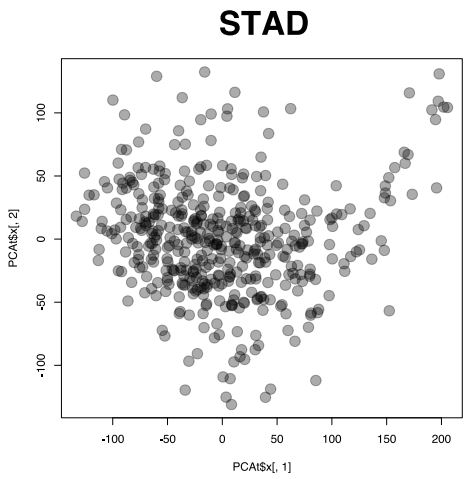
Wu, D., and Smyth, G.K. (2012). Camera: a competitive gene set test accounting for inter-gene correlation. Nucleic Acids Res. *40*, e133–e133.

Yang, S., Deng, P., Zhu, Z., Zhu, J., Wang, G., Zhang, L., Chen, A.F., Wang, T., Sarkar, S.N., Billiar, T.R., et al. (2014). ADAR1 Limits RIG-I RNA Detection and Suppresses IFN Production Responding to Viral and Endogenous RNAs. J. Immunol. Baltim. Md 1950 *193*, 3436–3445.

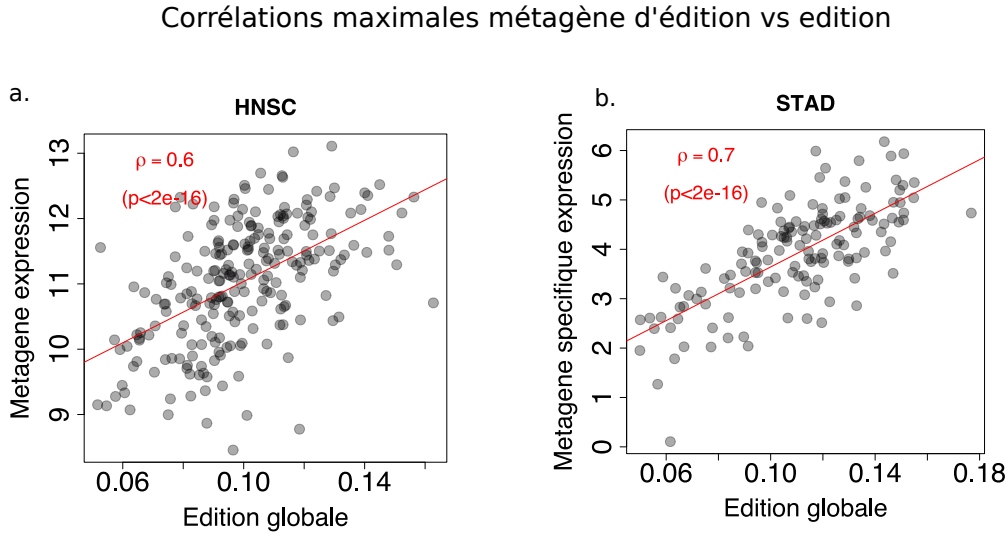
Zhang, X., Du, R., Li, S., Zhang, F., Jin, L., and Wang, H. (2014). Evaluation of copy number variation detection for a SNP array platform. BMC Bioinformatics *15*, 50.

Zwiener, I., Blettner, M., and Hommel, G. (2011). Survival analysis: part 15 of a series on evaluation of scientific publications. Dtsch. Ärztebl. Int. *108*, 163–169.

Annexes



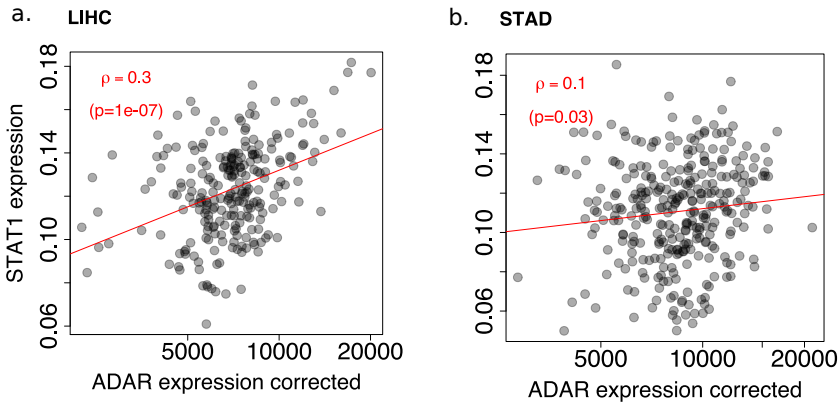
Annexe Figure 1: Analyse en composante principale (PCA) pour STAD. Distribution des données en fonction des deux premières composantes principales. Pas d'effet de batch visible. Les composantes 1 et 2 expliquent respectivement 14% et 7% de la variance observée.



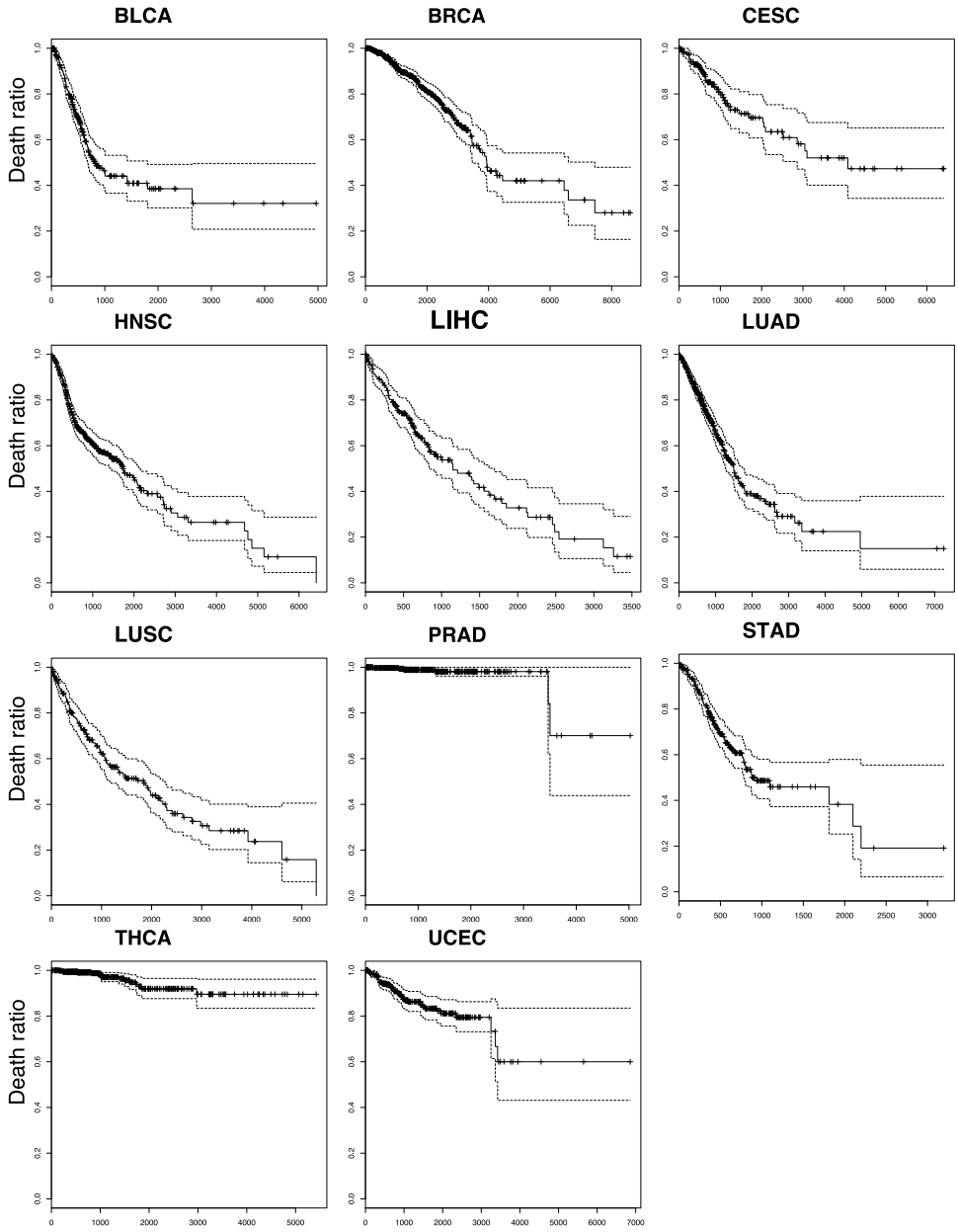
Annexe Figure 2: Représentation de la corrélation maximale retrouvée entre l'expression de la signature d'édition pan-cancer (a) ou cancer spécifique (b) et l'édition. Chaque point représente un échantillon. En abscisse la mesure de l'édition globale par échantillon. En ordonnées la mesure de la médiane d'expression du métagène par échantillon. La droite de régression linéaire en rouge indique la tendance de la distribution. ρ est le coefficient de Spearman et p sa p-valeur.

Cancers	P-valeur Test Wilcoxon	Ratio d'édition moyen des tissus sains
LIHC	/	0,121752829
BLCA	3,10E-10	0,082359933
BRCA	1,40E-23	0,078587102
CESC	0,0094	0,101097714
HNSC	3,20E-16	0,070158111
LUAD	4,70E-19	0,080573678
LUSC	2,70E-06	0,102611136
PRAD	1,00E-12	0,100082998
STAD	4,10E-08	0,092153095
THCA	4,30E-19	0,088191819
UCEC	0,0013	0,091219159

Table annexe: test de Wilcoxon pour évaluer la différence significative existant entre les ratios d'édition moyen des échantillons provenant des tissus sains LIHC par rapport aux ratios d'édition dans les tissus sains des autres cancers. La p-valeur indique la confiance que l'on peut accorder au fait que la distribution dans ces cancers est différente de celle de LIHC. "moyenne" indique la moyenne du ratio d'édition global retrouvée dans les tissus sains de chacun des types de cancers.



Annexe Figure 3: chaque point représente un échantillon. En abscisse l'expression d'ARNm d'ADAR ajustée pour la CNV d'ADAR, en ordonnée l'expression d'ARNm de STAT1. Les expressions sont mesurées par RNAseq normalisées et transformées en log2. La droite de régression linéaire en rouge indique la tendance de la distribution. ρ est le coefficient de corrélation de Spearman et p sa p-valeur. a. graphique pour le carcinome hépatocellulaire (LIHC). b. graphique pour l'adénocarcinome de l'estomac (STAD).



Annexe Figure 4: Graphique de Kaplan-Meier représentant la survie des patients dans chacun des types de cancers en fonction du temps en jours. Les barres verticales sont des évènements censurés.