Analyse de signatures transcriptomiques et épigénétiques des carcinomes hépatocellulaires

Léa Meunier

Université de Paris

Ecole doctorale Hématologie Oncogénèse et Biothérapies (ED561)

INSERM UMRS-1138, Centre de Recherche des Cordeliers (CRC)
Équipe 28 : Génomique Fonctionnelle des Tumeurs Solides

Analyse de signatures transcriptomiques et épigénétiques des carcinomes hépatocellulaires

Léa Meunier

Thèse de doctorat d'Oncogénèse

Dirigée par Éric Letouzé

Présentée et soutenue publiquement le 26 Février 2020

Devant un jury composé de :

Dr. François Radvanyi, Directeur de recherche, Institut Curie, Rapporteur
Dr. Michael Weber, Directeur de recherche, Université de Strasbourg, Rapporteur
Dr. Anne Biton, Ingénieure de Recherche, Institut Pasteur, Examinatrice
Dr. Pierre-Antoine Defossez, Directeur de recherche, Université de Paris, Examinateur
Dr. Judith Favier, Directeur de recherche, Université de Paris, Examinatrice

Titre : Analyse de signatures transcriptomiques et épigénétiques des carcinomes hépatocellulaires

Résumé :

Élucider les processus transcriptionnels et épigénétiques dérégulés dans les cancers est fondamental pour mieux comprendre les voies biologiques impliquées et proposer une thérapie adaptée au phénotype moléculaire de chaque tumeur. Les approches classiques de classification non supervisée définissent des groupes moléculaires principaux pour chaque type tumoral. Cependant, ces méthodes, appliquées à des tumeurs complexes comme le carcinome hépatocellulaire (CHC), le 3ème cancer le plus mortel au monde, définissent des groupes qui restent relativement hétérogènes et ne reflètent qu'imparfaitement la diversité des mécanismes biologiques à l'œuvre dans ces tumeurs. Au cours de ma thèse, j'ai développé une stratégie d'analyses innovante, basée sur l'analyse en composantes indépendantes (ACI), pour extraire des signatures de processus biologiques précis à partir de grands jeux de données transcriptomiques et épigénétiques. Grace à cette nouvelle approche, j'ai identifié des groupes de gènes co-régulés, associés à des phénotypes ou altérations moléculaires précises. De même, l'analyse en composantes indépendantes du méthylome de 738 CHC m'a permis d'isoler 13 signatures épigénétiques stables, préférentiellement actives dans certaines tumeurs et certains sites CpG. Ces signatures incluent des signatures de méthylation précédemment associées au vieillissement et au cancer, mais aussi de nouvelles signatures d'hyper- et d'hypométhylation liées à des événements « drivers » et sous-groupes moléculaires spécifiques. Ces résultats nous éclairent sur la diversité des mécanismes moléculaires impliqués dans la carcinogenèse hépatique. Les outils d'analyse biostatistique innovants que j'ai développés ont été incorporés dans un package R librement utilisable par la communauté scientifique.

Mots clefs :

Carcinomes hépatocellulaire, Analyse en composantes indépendantes, RNAseq, Méthylation, Signatures épigénétiques

Title : Transcriptomic and epigenetic signature analysis of hepatocellular carcinomas

Abstract :

Elucidating deregulated transcriptional and epigenetic processes in cancers is fundamental to better understand the biological pathways involved and to propose a therapy adapted to the molecular phenotype of each tumor. Classical unsupervised classification approaches define, for each tumor type, the main molecular groups. However, these methods, applied to complex tumors such as hepatocellular carcinoma (HCC), the 3rd cause of cancer-associated mortality worldwide, define groups that remain relatively heterogeneous and only imperfectly reflect the diversity of biological mechanisms at work in these tumors. During my PhD, I developed a, innovative strategy involving independent component analysis (ICA) to extract signatures of precise biological processes in large transcriptomic and epigenetic tumor data sets. This new approach allowed me to identify groups of co-regulated genes associated with specific phenotypes or molecular alterations. Similarly, independent component analysis of the methylomes of 738 HCC revealed 13 stable epigenetic signatures preferentially active in specific tumors and CpG sites. These signatures include signatures previously associated with ageing and cancer, but also new hyper- and hypomethylation signatures related to specific driver events and molecular subgroups. The work presented in this thesis sheds light on the diversity of molecular processes remodeling liver cancer transcriptomes and methylomes, improve the understanding of the molecular mechanisms involved in hepatic carcinogenesis and provides a statistical framework to unravel the signatures of these processes.

Keywords :

Hepatocellular Carcinoma, Independent Component Analysis, RNAseq, Methylation, Epigenetic Signatures

Table des matières

Introduction

1. Les carcinomes hépatocellulaires

Le cancer est une des principales causes de mortalité dans le monde. Il est à l'origine de 9,5 millions de décès en 2018 (Ferlay et al., 2019). Il s'agit d'une maladie due à l'accumulation d'altérations génétiques et épigénétiques qui procurent un avantage sélectif aux cellules touchées et permettent l'expansion clonale et la transformation maligne (Hanahan and Weinberg, 2000). Les altérations sont variées et peuvent être causées par des mécanismes intrinsèques à la cellule ou liées aux expositions environnementales (Stratton et al., 2009). La diversité des facteurs de risque et des altérations fait du cancer une maladie hétérogène multifactorielle. L'un des organes les plus touchés par l'apparition de tumeurs est le foie, avec environ 800 000 nouveaux cas en 2018 (Rawla et al., 2018). Il existe plusieurs types de tumeurs hépatiques, de formes et de localisations différentes. On distingue les tumeurs primaires, qui prennent naissance dans le foie, des tumeurs secondaires ou métastatiques, originaires d'un autre organe et s'étant propagées au tissu hépatique.

Parmi les tumeurs malignes, on retrouve le carcinome hépatocellulaire, le cholangiocarcinome, l'hépatoblastome et l'angiosarcome. Le carcinome hépatocellulaire (CHC) est le plus fréquent des cancers primitifs du foie (Yang and Roberts, 2010). Il se développe à partir d'hépatocytes, le type cellulaire majoritaire dans le foie qui a pour fonction d'assurer le bon déroulement des fonctions métaboliques et de détoxification. Le carcinome hépatocellulaire est responsable de plus de 500 000 décès par an, ce qui en fait le 3ème cancer le plus mortel au monde (Bray et al., 2018) ainsi qu'un problème de santé publique. Les autres formes de tumeurs malignes du foie sont plus rares et ont pour origine des types cellulaires différents : les cholangiocarcinomes proviennent de cellules épithéliales intra hépatiques des voies biliaires et représentent 6 % des cancers du foie (Sia et al., 2017) ; les hépatoblastomes, des tumeurs pédiatriques rares, se développent à partir de tissu embryonnaire ; et les angiosarcomes, des tumeurs mésenchymateuses, ont pour origine les cellules des vaisseaux sanguins (Chaudhary et al., 2015).

Il existe aussi des tumeurs hépatiques bénignes. Les principales, classées par fréquence, sont : l'hémangiome, une malformation vasculaire ; l'hyperplasie nodulaire focale, une prolifération d'hépatocytes polyclonale qui semble survenir suite à une perturbation locale du flux sanguin ; et l'adénome hépatocellulaire (HCA) dérivé de la prolifération monoclonale d'hépatocytes matures (Grazioli et al., 2017).

1.1. Facteurs de risque, histologie et traitement

1.1.1. Épidémiologie et facteurs de risque

Mon travail se focalise sur le carcinome hépatocellulaire (CHC), la forme la plus fréquente de cancer primitif du foie. Il se développe principalement sur un foie ayant subi des lésions sévères, cirrhotique dans 70 % à 90 % des cas. La cirrhose apparaît après de nombreuses années pendant lesquelles l'organe a connu une inflammation chronique pouvant être due à l'exposition à des composés toxiques, à des agents infectieux, à la présence de syndrome métabolique, ou à des maladies rares (El–Serag and Rudolph, 2007). Dans le cas de l'affection du patient par une maladie chronique du foie, l'inflammation peut entraîner l'apparition d'une fibrose hépatique plus ou moins importante, ce qui correspond à un tissu cicatriciel. Plusieurs stades de fibrose existent, définis en fonction de la sévérité des altérations dans le foie. Le plus sévère, correspondant à la cirrhose. Cette dernière est un état dans lequel l'architecture du foie est détruite par l'apparition de nodules structurellement anormaux, présentant une accumulation de mutations somatiques, qui s'accompagnent d'une diminution de la fonction hépatique (Brunner et al., 2019; Ginès et al., 2016). La cirrhose rend le foie incapable de remplir ses fonctions habituelles et constitue le terrain favorable au développement d'un cancer. Son origine et les paramètres personnels comme l'âge, le sexe et l'hygiène de vie, influent sur le risque individuel des patients de développer un cancer.

L'incidence du CHC n'est pas la même en fonction des régions, on le retrouve de manière plus fréquente en Asie de l'Est et du Sud-Est, en Afrique centrale et en Afrique de l'Est (cf. Figure 1). On remarque que la disparité de l'incidence dans les différentes régions du monde correspond à la répartition des facteurs étiologiques (McGlynn et al., 2001). En France, le développement d'un CHC est dans 50 % des cas, dû à la surconsommation d'alcool (Rosa et al., 2010). Aux États-Unis, c'est l'obésité qui représente le facteur de risque le plus important avec l'infection par l'hépatite C. En effet, l'accumulation de graisses dans le foie peut

provoquer une inflammation locale chronique (Ascha et al., 2010). Dans les pays d'Afrique et d'Asie, les carcinomes hépatocellulaires sont plus nombreux et sont en grande majorité associés à une infection chronique aux virus de l'hépatite B (Evans et al., 1998) et C (Lok et al., 2009). On retrouve également une exposition à l'aflatoxine B1 (Hsia et al., 1992) qui est produite par un champignon de type Aspergillus et qui prolifère dans le maïs, les cacahuètes ou encore les graines de coton, cultivés dans les pays chauds et humides. Des maladies, comme l'hémochromatose qui entraîne des dépôts de fer dans le foie ou des pathologies métaboliques peuvent également provoquer l'apparition d'une cirrhose.

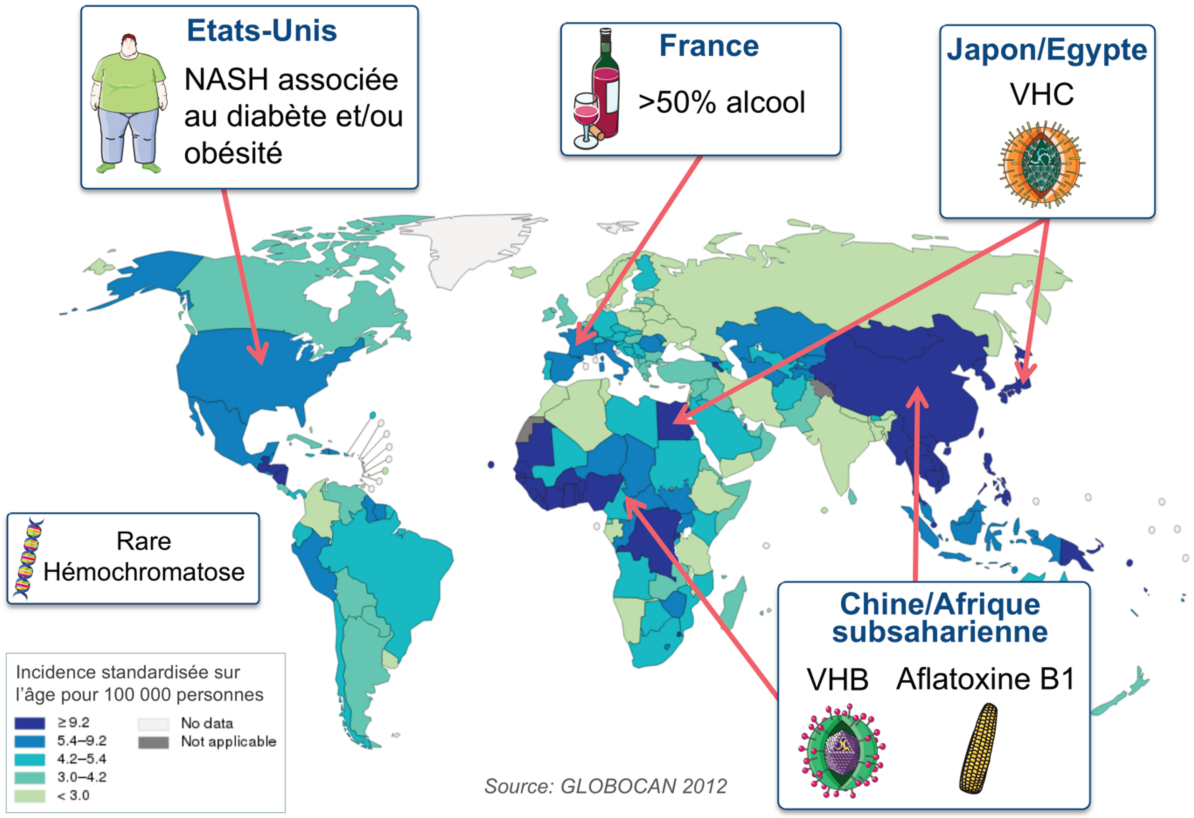


Figure 1 ***: Incidence globale et principales éthologies du CHC dans le monde*** *(Figure provenant de l'Habilitation à Diriger les Recherches de Sandra Rebouissou).*

Le CHC peut aussi se développer sur un foie non cirrhotique comme dans les rares cas de CHC résultant de la transformation maligne d'un adénome hépatocellulaire (Tokoro et al., 2014) (cf. Figure 2). De plus récemment l'exposition à l'acide aristolochique (Ng et al., 2017), une substance fortement mutagène contenue dans certaines plantes traditionnellement utilisées pour leurs vertus médicinales dans les pays asiatiques a été identifiée comme associée au développement de CHC en particulier à Taiwan.

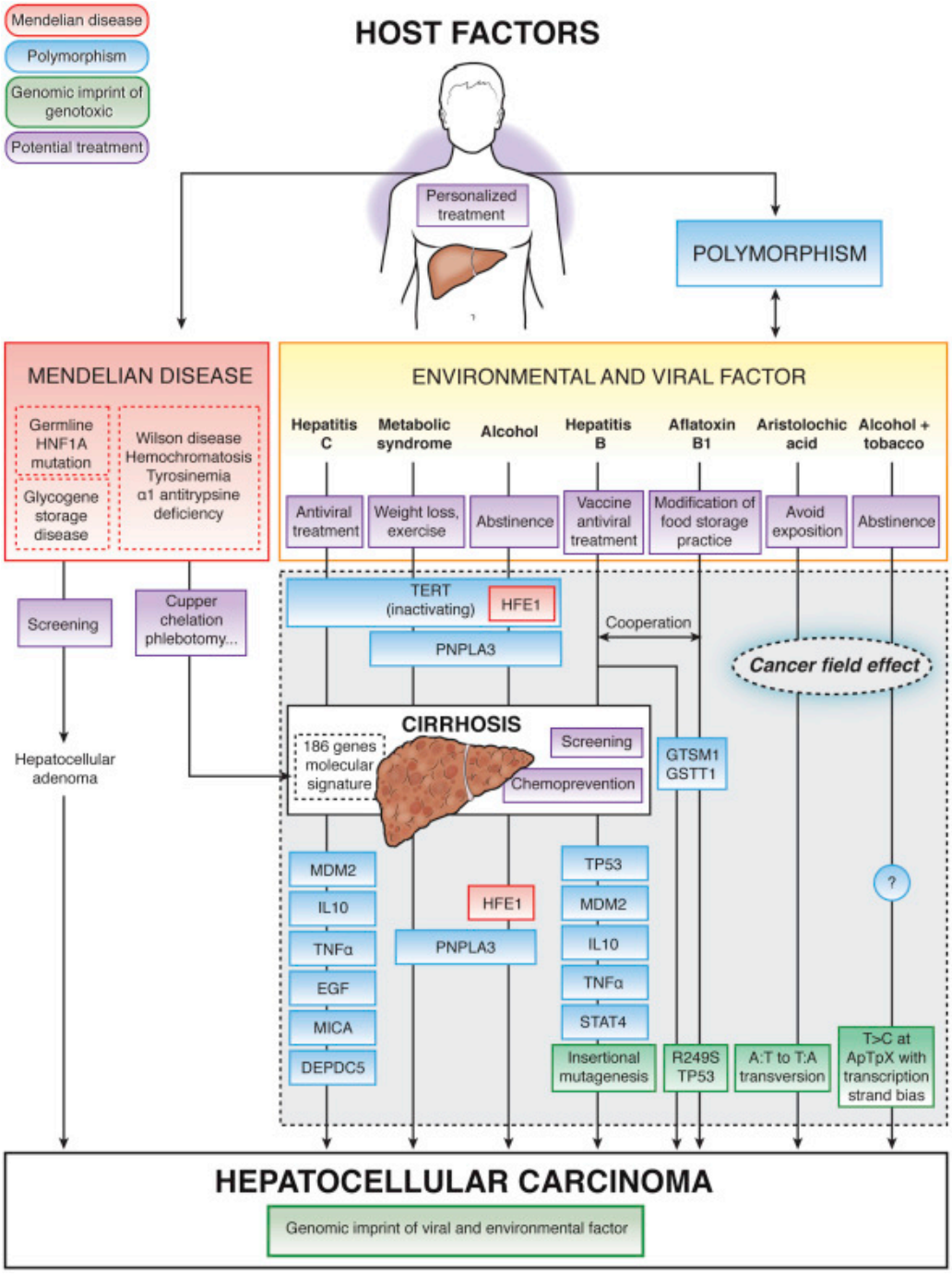


Figure 2 : ***Interaction entre la prédisposition génétique, les facteurs environnementaux et la survenue du CHC.*** *Rouge, modèle héréditaire mendélien, bleu, polymorphismes impliqués dans l'apparition du CHC, vert, empreinte génétique du CHC des expositions génotoxiques, et violet, les traitements potentiels utilisés pour réduire l'incidence du CHC (Zucman-Rossi et al., 2015).*

1.1.2. Caractéristiques et classification morphologique des tissus

Le carcinome hépatocellulaire peut avoir un aspect macroscopique très variable et son diamètre peut aller de quelques millimètres à 20 cm. Pour caractériser le foie du patient à partir duquel il s'est développé, on évalue la sévérité de la cirrhose grâce au score METAVIR (Poynard et al., 1997). Ce dernier définit 5 stades allant du foie sans fibrose F0 au stade

cirrhotique F4. Des classifications ont également été établies pour quantifier la différenciation du CHC. Le grade d'Edmondson-Steiner sépare les CHC en 4 groupes, allant du grade I, de CHC très bien différenciés parfois difficiles à distinguer du tissu non tumoral, au grade IV, peu différenciés et plus agressifs (Edmondson and Steiner, 1954). Le pronostic du carcinome hépatocellulaire dépend du stade de découverte et de la gravité de la cirrhose du foie non-tumoral. Il existe plusieurs classifications visant à orienter le traitement et notamment sélectionner les patients pouvant bénéficier d'un traitement à visée curative susceptible d'améliorer leur survie (Levy and Sherman, 2002). Actuellement, la classification pronostic recommandée par l'EASL (European Association for the Study of the Liver) et l'AASLD (American Association for the Study of Liver Diseases) est la classification du Barcelona Clinic Liver Cancer (BCLC), comprenant les stades BCLC 0, A, B, C et D, allant du plus précoce au stade terminal. Elle prend en compte le nombre et la taille des nodules, la présence d'invasion vasculaire, de lésions extra-hépatiques, la fonction hépatique, et l'état général de l'individu pour orienter les patients vers les thérapies adaptées (Sia et al., 2017).

1.1.3. Pronostic et traitement

Dans l'ensemble, le pronostic des CHC est mauvais, puisque seuls 30 % des patients, ceux dont le CHC est détecté à un stade précoce (BCLC 0 et A), peuvent bénéficier d'un traitement curatif comme la résection chirurgicale, la transplantation ou l'ablation percutanée (par alcoolisation ou radiofréquence). La transplantation hépatique a l'avantage de guérir à la fois la tumeur et la cirrhose sous-jacente du foie, elle est donc un traitement incontournable pour les tumeurs détectées à un stade précoce. Cependant le manque de donneurs de foie limite le nombre de transplantations, ce qui entraîne un temps d'attente de plus en plus long, pendant lequel l'état du patient peut s'aggraver (Bruix and Llovet, 2002).

Pour les patients avec un CHC de stade intermédiaire (BCLC B), non éligibles à ces traitements curatifs, les possibilités sont : la chimio-embolisation artérielle, qui permet de bloquer l'apport sanguin de la tumeur, et la chimiothérapie intra-artérielle, qui consiste à injecter des molécules de chimiothérapie dans l'artère hépatique grâce à un cathéter. Ces traitements visent à améliorer la survie, que ce soit en durée ou en qualité de vie, sans aggraver l'évolution de la cirrhose, mais peuvent présenter des effets secondaires (Galuppo et al., 2014).

Si la tumeur est à un stade avancé (BCLC C), ou à un stade intermédiaire (BCLC B), et qu'elle ne répond pas aux traitements décrits ci-dessus ou récidive, un traitement médicamenteux peut-être administré. Pendant longtemps, le seul traitement proposé était le Sorafenib. Il s'agit d'un inhibiteur qui cible les récepteurs VEGFR-1/2/3 (vascular endothelial growth factors receptors), les kinases sérine/thréonine BRaf et CRaF, et les PDGFR-α/β (platelet-derived growth factor receptors). L'essai SHARP qui a comparé les effets du Sorafenib par rapport au placebo a montré une amélioration de la survie globale des patients de 3 mois en moyenne et une diminution de la progression radiologique (Llovet et al., 2008). Ces dernières années de nouveaux traitements de première ligne, le Lenvatinib, un inhibiteur de la tyrosine-kinase à cibles multiples (VEGFR 1-3, FGFR 1-4, PDGFRα, RET), et de deuxième ligne tels que le Regorafenib, le Cabozantinib et le Nivolumab, ont été approuvés. Le Regorafenib, un inhibiteur multikinase plus efficace contre les kinases VEGFR et dont l'activité est plus large (contre le TIE2, le KIT et le RET) que le Sorafenib, a montré des résultats positifs chez les patients dont les tumeurs progressent sous Sorafenib (Bruix et al., 2017). Une amélioration similaire de la survie est obtenue avec le Cabozantinib, un inhibiteur de tyrosine kinases (VEGFR 1-3, MET et AXL) (Abou-Alfa et al., 2018). Le Nivolumab est pour le moment la seule immunothérapie autorisée dans le traitement des CHC. Il s'agit d'un anticorps monoclonal ciblant le point de contrôle immunitaire inhibiteur de la molecule PD-1 (Programmed cell death 1). Cependant, le Ramucirumab, un anticorps monoclonal qui se lie au VEGFR-2 et bloque son activation, a montré des résultats prometteurs en test de phase III (Zhu et al., 2018). Plus récemment, l'association de l'Atezolizumab (inhibiteur de PD-L1) avec le Bevacizumab (inhibiteur du VEGF) a montré des résultats exceptionnels dans un essai de phase 3 avec environ 30 % de réponses objectives et une amélioration significative de la survie globale des patients.

Au cours des dernières années, malgré la caractérisation des altérations dans les CHC et le développement de thérapies moléculaires ciblant les voies de signalisation altérées, de nombreux essais cliniques ont échoué. Les raisons de leur échec sont hétérogènes et comprennent le manque de compréhension des facteurs critiques de la progression/dissémination tumorale, la toxicité hépatique ou les défauts de conception des essais (Llovet and Hernandez-Gea, 2014). De nouveaux essais sont également conçus pour tester des médicaments sur des sous-populations de patients atteints de CHC à partir de

biomarqueurs. On retrouve par exemple le test d'inhibiteurs de MEK chez les patients présentant une activation de RAS (Das, 2018). Ces stratégies ont été jugées efficaces dans les cancers du sein, du mélanome et du poumon, et pourraient changer le paysage de la conception des essais cliniques du CHC. Le profilage moléculaire et la réponse au traitement sont étroitement liés ; par conséquent, l'amélioration de la compréhension des dérégulations dans les CHC permettra de les cibler efficacement avec une thérapie spécifique.

1.2. Altérations moléculaires *driver*

Les étiologies et les processus moléculaires menant à la transformation maligne des hépatocytes en CHC sont multiples, induisant une grande hétérogénéité dans ces tumeurs. En effet, de nombreuses altérations génétiques s'accumulent dans les tissus durant la vie du patient, particulièrement pendant la prolifération des nodules cirrhotiques (Brunner et al., 2019) et le développement du carcinome hépatocellulaire. Seules certaines altérations ont des effets sur les mécanismes de régulation cellulaire, ce qui peut leur conférer un avantage sélectif dans la transformation maligne de la cellule. On les appelle altérations *driver*. On estime que le nombre d'altérations dites *driver* est limité, généralement moins de dix (Hanahan and Weinberg, 2000), et confère à la cellule une ou des capacités permettant le développement de la tumeur. Les autres altérations sans effet dans le développement tumorale sont dites *passager* (Pon and Marra, 2015). Il est donc important d'identifier et caractériser les altérations qui contribuent à la transformation et aux dysfonctionnements cellulaires afin de comprendre la biologie et la diversité des tumeurs dans l'optique du développement de thérapies ciblées adaptées au phénotype tumoral.

1.2.1. Différents types d'altérations *driver*

Il existe différents types d'altérations somatiques dans le génome d'une cellule tumorale qui modifient la séquence et parfois le nombre de copies de l'ADN : les substitutions nucléotidiques, les petites insertions et délétions (indels) ou les réarrangements structuraux de plus grande taille (Lengauer et al., 1998). D'autres processus moins fréquents peuvent aussi participer à la carcinogénèse hépatique.

1.2.1.1. Mutations ponctuelles

Les mutations ponctuelles sont les substitutions nucléotidiques et les indels dont la taille est inférieure à 10 nucléotides. Parmi les processus responsables de la survenue de mutations

dans l'ADN d'une cellule, on peut trouver des mutations dues à des erreurs dans la réplication de l'ADN, qui arrive tout au long de la vie, ou des mutations dues à différentes expositions à des agents environnementaux mutagènes (substances chimiques, rayonnements…). En général, ces altérations ponctuelles sont corrigées par les mécanismes de réparation de l'ADN. Cependant, ces mécanismes ont un taux d'erreur intrinsèque qui fait que des mutations, non corrigées, sont intégrées au génome et transmises aux générations cellulaires suivantes (cf. Figure 3) (Stratton et al., 2009).

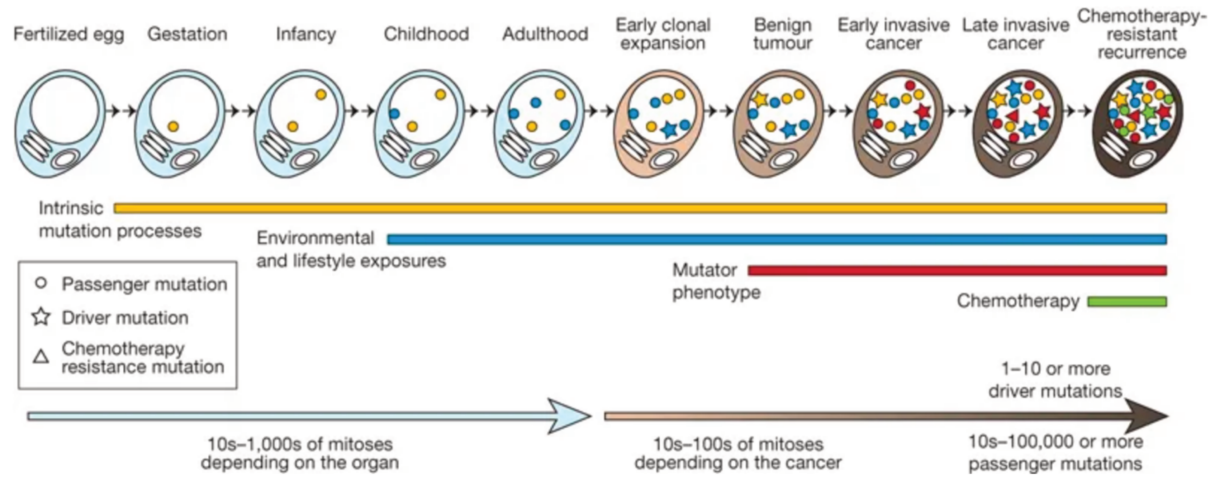


Figure 3 : ***Schéma de l'acquisition de mutations au cours de la vie par différents mécanismes pouvant aboutir à la formation et au développement de tumeurs*** *(Stratton et al., 2009).*

Les mutations ponctuelles peuvent survenir tout le long du génome, et si la plupart sont des mutations *passager* n'ayant aucun effet sur les cellules, certaines entraînent des conséquences fonctionnelles qui participent à la dérégulation de la machinerie cellulaire observée dans les tumeurs. Par exemple, dans les gènes codants exprimés, les mutations dans un exon peuvent induire un changement de codon au moment de la traduction et activer ou inactiver la fonction de la protéine (Hanna et al., 2005). Les petites insertions et délétions, en plus de modifier la traduction d'un codon, peuvent, si elles ne touchent pas des nucléotides par multiple de 3, décaler le cadre de lecture. Ce décalage entraîne la formation d'une protéine tronquée dans laquelle il manque des domaines. On retrouve également ces altérations *driver* dans les séquences non codantes de l'ADN. Par exemple, des mutations qui touchent les éléments de régulation de la transcription des gènes ont été particulièrement étudiées. En effet, la mutation dans un promoteur peut entraîner la surexpression d'un gène comparé au tissu normal, comme c'est le cas pour le gène de la télomérase *TERT* (telomerase transcriptase inverse) dans les CHC (Nault and Zucman-Rossi, 2016).

1.2.1.2.Aberrations chromosomiques

La plupart des CHC présente une instabilité chromosomique importante, avec de nombreux gains, pertes, ou réarrangements chromosomiques, dus à différents mécanismes, ou erreurs lors de la réplication et division cellulaire (Daughtry and Chavez, 2016). L'altération du nombre de copies de l'ADN peut engendrer un changement d'expression. Les délétions hétérozygotes et homozygotes peuvent inactiver partiellement ou totalement le fonctionnement de gènes suppresseurs de tumeurs, comme *CDKNA2*, *TP53* et *AXIN1* dans les CHC. Les gains et amplifications focales peuvent augmenter l'expression des oncogènes tel que *MYC*, *CCND1* et *MET* (cf. Figure 4).

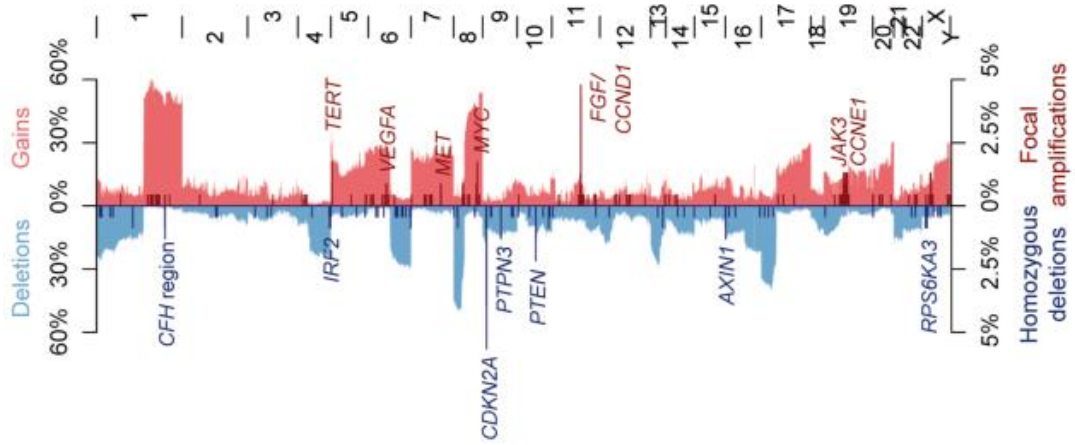


Figure 4 : ***Fréquence des anomalies du nombre de copies le long du génome.*** *L'axe de droite indique la fréquence des changements de faible amplitude (gains et pertes) ; l'axe de gauche indique la fréquence des changements de forte amplitude (amplifications focales et délétions homozygotes). Les gènes cibles des amplifications récurrentes et des délétions homozygotes sont indiqués (Schulze et al., 2015).*

Les anomalies chromosomiques macroscopiques, touchant une grande portion ou un bras de chromosome, sont fréquentes dans les CHC, comme les gains des bras 1q, 5, 22 6p, 7, 8q, 17q et 20 et les pertes des bras 1p, 4q, 6q, 8p, 13q, 16, 17p et 21 (Schulze et al., 2015; Zucman-Rossi et al., 2015). L'accumulation de gains chromosomiques multiples a été identifiée comme un événement tardif dans les CHC faisant suite à une première altération driver favorisant l'instabilité chromosomique (Letouzé et al., 2017). Par ailleurs, il a été montré que les CHC les plus instables étaient globalement associés à des tumeurs peu différentiées, plus prolifératives et avec un mauvais pronostic (Boyault et al., 2007).

1.2.1.3. Réarrangements structuraux

Les réarrangements structuraux sont composés de 4 types d'altérations simples : duplications, délétions, inversions et translocations qui touchent la séquence d'ADN (Feuk et al., 2006). On observe également des évènements complexes qui sont la combinaison synchrone de ces altérations élémentaires, comme le chromothripsis, qui se caractérise par la pulvérisation d'une séquence chromosomique en plusieurs fragments et leurs réassemblages aléatoires (Pellestor et al., 2014). Les réarrangements structuraux sont de taille variable et leurs points de cassure peuvent se trouver, comme pour les mutations, dans des séquences codantes pour des protéines, ou non-codantes.

Un point de cassure dans la séquence codante d'un gène ou dans ses introns entraîne directement la fusion d'une partie du gène avec une séquence distante, induisant une molécule d'ADN chimérique. Quand la région distante correspond à un autre gène, cela engendre une fusion de gènes, qui, si elle est exprimée induit l'apparition d'un transcrit chimérique. Le transcrit peut-être non fonctionnel, ou au contraire combiner différents domaines qui vont augmenter et modifier l'activité de la protéine qu'il code (Honeyman et al., 2014). Par exemple, l'inversion de la séquence d'ADN comprenant le domaine 3'UTR de IL6 identifiée dans le foie (cf. Figure 5), entraîne la formation d'un transcrit ne contenant pas les éléments de régulation d'*IL6*, ce qui empêche sa dégradation et aboutit à une accumulation dans les cellules (Calderaro et al., 2018).

Quand la région remaniée est intergénique, l'altération peut changer le contexte chromatinien dans lequel se trouve le gène ce qui peut rapprocher des éléments de régulation de la transcription et modifier son expression. Par exemple, la délétion de la séquence contenant les exons du gène *INHBE* identifiés dans certains adénomes hépatocellulaires, détourne son promoteur, qui active alors la transcription du gène *GLI1*, et entraîne la formation de tumeurs *(Nault et al., 2017)*. La translocation d'une séquence contenant des enhancers actifs, ou la délétion d'insulators, peut augmenter l'expression d'un oncogène et participer à la transformation tumorale (Haller et al., 2019; Hnisz et al., 2016) (cf. Figure 5).

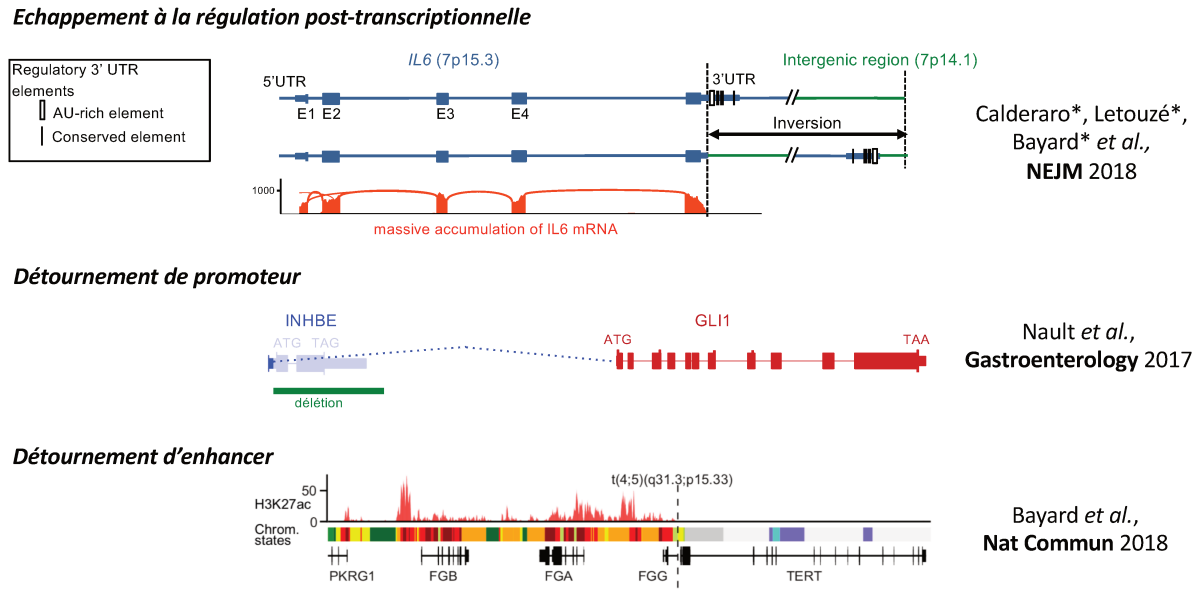


Figure 5 : ***Exemples de réarrangements structuraux identifiés dans les tumeurs du foie****. Le détournement d'un enhancer placé en amont de TERT et l'association du promoteur d'INHBE à son gène voisin GLI1 suite à une délétion entraînent respectivement la surexpression de TERT et GLI1 (Bayard et al., 2018; Nault et al., 2017). La perte du domaine 3'UTR de IL6 modifie sa régulation post-transcriptionnelle (Calderaro et al., 2018).*

1.2.1.4. Insertions virales

L'infection des patients par le virus de l'hépatite B (HBV) augmente le risque de développer des CHC (Trépo et al., 2014), principalement en favorisant l'inflammation chronique du foie, et le développement d'une cirrhose (Neuveut et al., 2010). En plus de son rôle dans l'établissement des lésions du foie, HBV joue un rôle direct dans la carcinogenèse hépatique. Il est retrouvé dans les cellules sous forme épisomale (circulaire) mais aussi sous forme linéaire intégrée dans le génome humain (Locarnini and Zoulim, 2010). L'insertion du virus modifie la séquence d'ADN ce qui peut altérer la structure des transcrits et l'expression des gènes et participer à la transformation maligne des cellules du foie. Le virus peut s'insérer partout dans le génome humain, mais on le retrouve plus fréquemment dans les régions répétées et des séquences transcrites. HBV a notamment été retrouvé inséré et caractérisé comme altération driver pour le développement des tumeurs dans le promoteur de TERT ainsi que dans d'autres oncogènes connus comme KMT2B, CCNE1 et SENP5 (Zhao et al., 2016).

Un deuxième virus, inséré dans de rares cas de CHC, a été récemment identifié au laboratoire : le virus adénome-associé de type 2 -AAV2 (Nault et al., 2015). Initialement, un fragment de séquence de ce virus a été trouvé par hasard dans le promoteur de TERT. Par la suite,

l'utilisation des données de séquençage exome, génome et RNAseq, alignées sur une référence de la séquence virale d'AAV2 ont permis d'identifier d'autres insertions. Les reads s'alignant en partie sur le génome humain et en partie sur le génome viral permettent d'identifier le point d'insertion du virus. De cette manière, on retrouve des insertions de ce virus, caractérisées comme driver dans la transformation maligne des cellules du foie, dans des oncogènes typiques comme dans le promoteur de TERT, les gènes du cycle cellulaire CCNA2 et CCNE1, TNFSF10 et KMT2B (Bella et al., 2019; Nault et al., 2015).

1.2.2. Gènes et pathways *driver*

Le CHC a fait l'objet de plusieurs études, de notre laboratoire (Guichard et al., 2012; Schulze et al., 2015) et d'autres (Ahn et al., 2014; Fujimoto et al., 2012), visant à identifier les voies et processus affectés de manière récurrente par des altérations génétiques, qui conduisent à la carcinogénèse hépatique. On retrouve 6 voies de signalisation principales (cf. Figure 6).

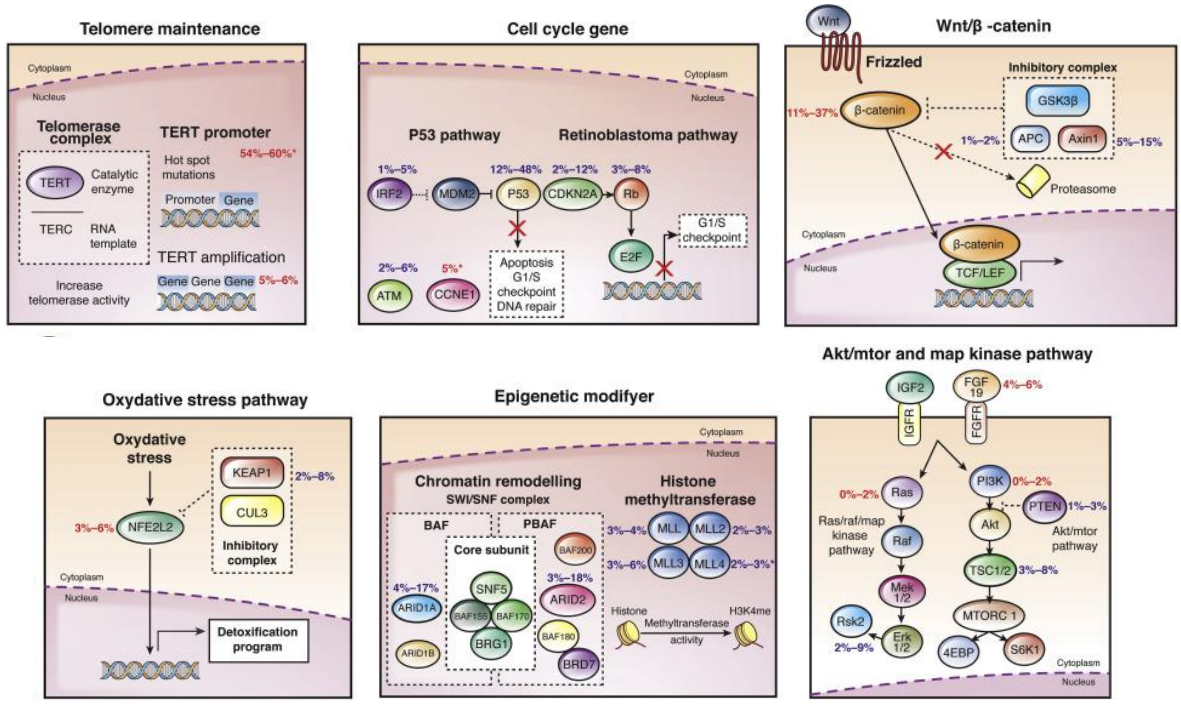


Figure 6 *: **Paysage génétique du CHC.** Rouge, activant les mutations des oncogènes, bleu, inactivant les mutations des suppresseurs de tumeurs. ∗Gène ciblé de façon récurrente par l'intégration du virus HBV. (Zucman-Rossi et al., 2015).*

Le gène le plus fréquemment altéré dans les CHC est *TERT* qui code pour la sous-unité catalytique du complexe de télomérase. Ce complexe est essentiel pour l'allongement et l'entretien des télomères, évitant ainsi l'érosion chromosomique dans les cellules. Le gène *TERT* n'est pas exprimé dans le foie sain, mais il est réactivé au début de la carcinogenèse

hépatique, entraînant une altération de la longueur des télomères. On retrouve une surexpression de la télomérase dans 90 % des CHC (Farazi et al., 2003; Plentz et al., 2007) pouvant être induite par différents mécanismes : les mutations du promoteur (54 % à 60 %) (Nault et al., 2013), l'amplification du gène (5 % à 6 %) (Totoki et al., 2014) ; l'insertion des virus HBV ou AAV2 (10 % à 15 %) (Chiang et al., 2008; Nault et al., 2015; Paterlini-Bréchot et al., 2003) ou des réarrangements structuraux du promoteur. On retrouve ces altérations associées de manière fréquente à une instabilité génomique accrue.

Après la maintenance des télomères, la deuxième voie la plus altérée dans le CHC est la voie Wnt/B-caténine, essentielle pour la zonation du foie et la différenciation des cellules hépatiques. On la retrouve activée dans 17 % à 44 % des CHC, avec une surexpression des gènes cibles comme *GLUL* et *LGR5*, particulièrement chez les patients avec des tumeurs bien différenciées (Boyault et al., 2007). Cette activation est en grande majorité due à la présence de mutations activatrices du gène *CTNNB1* (Coste et al., 1998). Plusieurs positions récurrentes de mutation dans ce gène ont été caractérisées, induisant différents degrés d'activation de la voie, comme les substitutions ou délétions in-frame touchant la séquence de *CTNNB1* cible du complexe inhibiteur *APC/AXIN1/GSK3B* (Rebouissou Sandra et al., 2016). Des mutations inactivatrices des gènes *AXIN1* (10% des cas) (Satoh et al., 2000), *APC* (1-2% des cas) ou *ZNRF3* (3% des cas) (Basham et al., 2019) affectant la voie Wnt/B-caténine sont également retrouvées dans les CHC, mais de manière moins fréquente.

L'inactivation du gène *TP53* par mutation somatique touche 12 % à 48 % des CHC, le plus souvent dans les tumeurs induites par le virus HBV (Zucman-Rossi et al., 2015). Ces mutations sont retrouvées tout le long du gène, la seule position de mutation récurrente caractérisée est la mutation R249S associée à l'exposition à l'aflatoxine B1 (Guichard et al., 2012). *TP53* est un gène suppresseur de tumeur impliqué dans la régulation du cycle cellulaire, spécifiquement dans la transition G1/S, c'est-à-dire la transition entre la croissance cellulaire pendant laquelle a lieu la plupart de la régulation transcriptionelle et la phase de réplication de l'ADN. Il joue aussi un rôle dans le mécanisme de réparation de l'ADN et dans le contrôle de l'apoptose, la mort cellulaire programmée. En dehors de l'altération du gène *TP53*, des mutations inactivatrices de *RB1* (3 % à 8 % des cas) et *CDKN2A* (2 % à 12 %) sont également

impliquées dans l'altération du cycle cellulaire (Ahn et al., 2014; Totoki et al., 2014). Les mutations des gènes *TP53* et *CTNNB1* sont majoritairement exclusives.

Le stress oxydatif est la troisième voie la plus altérée dans le CHC. L'exposition continue à des facteurs de stress lors d'une inflammation ou d'une consommation prolongée d'alcool entraîne une dérégulation de l'équilibre entre le taux de production et d'élimination des espèces réactives de l'oxygène (ROS). La rupture de cette homéostasie favorise la transformation maligne. L'une des voies de signalisation impliquée dans le processus de la réponse au stress est celle régulée par le complexe NFE2L2/KEAP1 (Guichard et al., 2012). Une étude in vitro a montré que l'activation constitutive de la voie favorise la survie et inhibe la mort cellulaire (Sporn and Liby, 2012). Il a été montré que 5 % à 15 % des cas de CHC (Zucman-Rossi et al., 2015) présentent une activation de cette voie, soit par mutation activatrice de *NFE2L2* (nuclear factor erythroid 2-related factor 2), soit par mutation inactivatrice de *KEAP1* (Kelch-like ECH-associated protein 1), ce qui empêche sa liaison à NFE2L2 et la dégradation de ce dernier (Bryan et al., 2013).

La voie PI3K-AKT-mTOR, qui module le métabolisme et la prolifération cellulaire, est activée dans les CHC par mutation inactivatrice de *PTEN* (1 % à 3 % des cas), de *TSC1* ou de *TSC2* (3 % à 8 % des cas) (Totoki et al., 2014). La voie RAS/RAF/MAPK, également impliquée dans la prolifération, la différenciation, la migration et la survie, est altérée dans les CHC. On retrouve dans 2 % à 9 % des mutations de *RPS6KA3* entraînant la perte d'une boucle de rétroaction négative et l'activation constitutive de la voie (Zucman-Rossi et al., 2015). De plus, FGFR est le récepteur en amont dans la signalisation de PI3K-AKT-mTOR et de RAS/RAF/MAPK. L'activation de ces voies pourrait également être liée à l'amplification d'une région qui inclut le *FGF3*, *FGF4* et *FGF19*, décrite dans 5 % des CHC (Babina and Turner, 2017).

Enfin, différents régulateurs épigénétiques, en particulier le complexe de remodelage de la chromatine SWI/SNF, sont fréquemment altérés dans les CHC. Ce complexe est impliqué dans le changement de conformation de l'ADN, permettant ainsi, quand il est recruté, de faire glisser les nucléosome le long de la séquence d'ADN pour rendre possible la transcription de certains gènes (Masliah-Planchon et al., 2015). Des mutations inactivatrices des gènes *ARID1A* et *ARID2* sont identifiées respectivement dans 4 % à 17 % et 3 % à 18 % des CHC (Nakamura et al., 2019). Les effets précis sur changement induit par l'inactivation de ces deux gènes et

les mécanismes par lesquels ils promeuvent la carcinogénèse hépatique restent flous. Cependant, il a été mis en évidence qu'*ARID1A* code pour la sous-unité BAF250a du complexe SWI/SNF, qui se lie à l'ADN et recrute d'autres composants du complexe pour ouvrir la chromatine à la machinerie de transcription (Guan et al., 2011). Il a été décrit comme étant à la fois un oncogène, dont la surexpression intervient dans l'initiation de la transformation tumorale, et un suppresseur de tumeur dans le développement du cancer (Sun et al., 2017). L'inactivation du gène *ARID1A* après l'établissement de la tumeur augmenterait le potentiel métastatique du CHC (Nakamura et al., 2019). Il a également été montré que sa perte dans le foie normal était liée à la dérégulation du métabolisme lipidique ce qui entraîne un stockage aberrant de graisse dans le foie et une prédisposition aux maladies métaboliques non alcooliques (NASH), un terrain favorable au développement des CHC (Qu et al., 2019). Le gène *ARID2* est quant à lui nécessaire à la réparation des dommages de l'ADN induits par les UV et les composés cancérigènes dans le CHC (Oba et al., 2017). Le rétablissement de son expression dans le foie inhibe la croissance cellulaire et la progression tumorale chez la souris, alors que son inhibition entraîne une surexpression des gènes codant pour les protéines du cycle cellulaire comme *CCND1* et *CCNE1* (Duan et al., 2016; Toh et al., 2019). On retrouve aussi dans les CHC, des altérations d'autres régulateurs épigénétiques, notamment les histones méthyltransférases MLL/KMT2 qui jouent un rôle dans la modulation de la chromatine et l'accessibilité de l'ADN, dans 2 % à 6 % des tumeurs (Zucman-Rossi et al., 2015).

1.3. Groupes et signatures moléculaires

Plusieurs approches ont été utilisées pour identifier des groupes de tumeurs hépatiques présentant des profils moléculaires homogènes et identifier des signatures transcriptionnelles de pronostics et de différents phénotypes. Les méthodes statistiques utilisées et les principaux résultats obtenus sont décrits ci-dessous.

1.3.1. Méthodes d'analyse non supervisées

Le clustering englobe les techniques de data mining visant à partitionner un gros volume de données en un ensemble de sous-groupes homogènes (Jain et al., 1999). Il s'agit de méthodes non-supervisées, c'est-à-dire sans hypothèse préalable pour guider l'analyse. Elles sont couramment utilisées pour l'analyse de données transcriptomiques.

Par exemple, Eisen *et al.* sont parmi les premiers à avoir utilisé la technique de clustering hiérarchique pour déterminer des groupes de gènes co-régulés et les visualiser selon leur profil d'expression dans les levures (Eisen et al., 1998). Le clustering hiérarchique est un algorithme qui s'appuie sur une notion de distances entre individus pour regrouper les plus semblables en groupes appelés clusters, tout en maximisant la dissimilarité entre les groupes formés. Deux grandes familles de clustering hiérarchique co-existent : les approches dites ascendantes (qui agglomèrent les individus selon leur similarité) (Guha et al., 1998) et les approches dites descendantes (qui divisent le groupe initial d'individus par rapport à leur dissimilarité).

Il existe plusieurs méthodes permettant de calculer des distances entre les individus (euclidienne, corrélation de Pearson…) et d'estimer le critère de liaison entre les clusters (simple, moyen, complet, centroïdes…) (Szekely and Rizzo, 2005). Les méthodes hiérarchiques regroupent les données sous la forme d'un dendogramme, arbre où les individus sont regroupés de proche en proche (cf. Figure 7). Les méthodes de clustering hiérarchique ont l'avantage de proposer une classification des données suivant plusieurs niveaux (Struyf et al., 1997). Le regroupement des tumeurs permet d'identifier des groupes moléculaires homogènes, et le regroupement des gènes permet d'identifier des modules de gènes co-régulés (Segal et al., 2003).

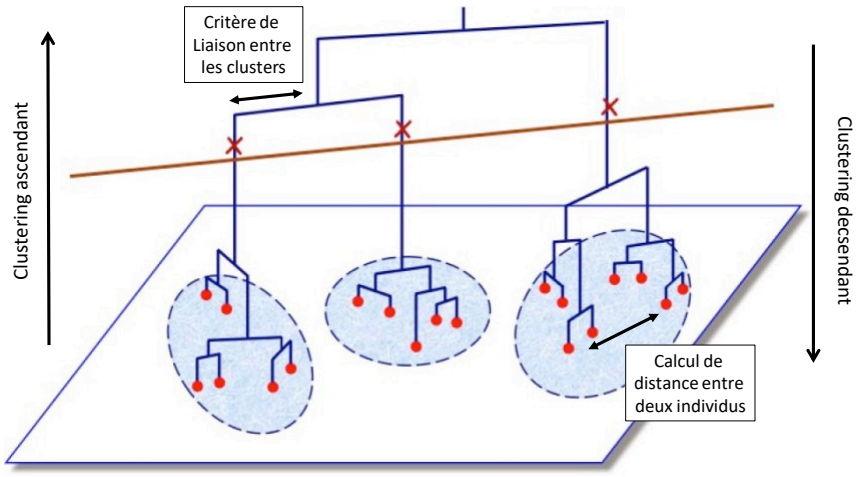


Figure 7 : ***Classification par clustering hiérarchique***. *Les clusters sont formés en utilisant un calcul de la distance entre 2 individus. Ils sont ensuite appariés en fonction de leur similarité grâce à un critère de liaison.*

Les questions à prendre en compte lors de la détermination des clusters de données sont les choix du nombre de clusters et la confiance dans l'assignation de chaque individu à un cluster.

Pour répondre à ces questions, il est possible d'utiliser le technique de rééchantillonnage et de validation croisée, afin d'évaluer la stabilité des résultats des clusters obtenus par rapport à la variabilité de la sélection des échantillons (Ben-Hur et al., 2002; Bhattacharjee et al., 2001; Dudoit and Fridlyand, 2002). Le bootstrapping a notamment été utilisé pour effectuer ce rééchantillonnage en lançant plusieurs classifications par clustering hiérarchique sur un sous-ensemble de gènes et d'individus (Bhattacharjee et al., 2001). Une méthode de visualisation des résultats du rééchantillonage a également été proposée pour appuyer les décisions concernant le nombre de clusters, en choisissant les résultats les plus stables (Monti et al., 2003). Cette technique est appelée clustering consensus.

L'analyse en composantes principales (ACP) est également fréquemment utilisée pour visualiser la similarité d'individus décrits par plusieurs variables quantitatives, comme l'expression des gènes (Ma and Dai, 2011). L'ACP permet l'analyse et la visualisation d'un jeu de données à plus de 3 variables, en synthétisant l'information contenue dans le jeu de données en un nombre réduit de nouvelles variables appelées composantes principales, correspondant à une combinaison linéaire des variables d'origine (Abdi and Williams, 2010). L'ACP permet également d'estimer la variance capturée par chaque composante, et les composantes pour lesquelles la variation est maximale permettent de visualiser aisément la dispersion des échantillons et d'identifier des groupes associés aux annotations cliniques et moléculaires (cf. Figure 8). La contrainte de cette méthode est que 2 composantes principales doivent être orthogonales entre elles (Ma and Dai, 2011).

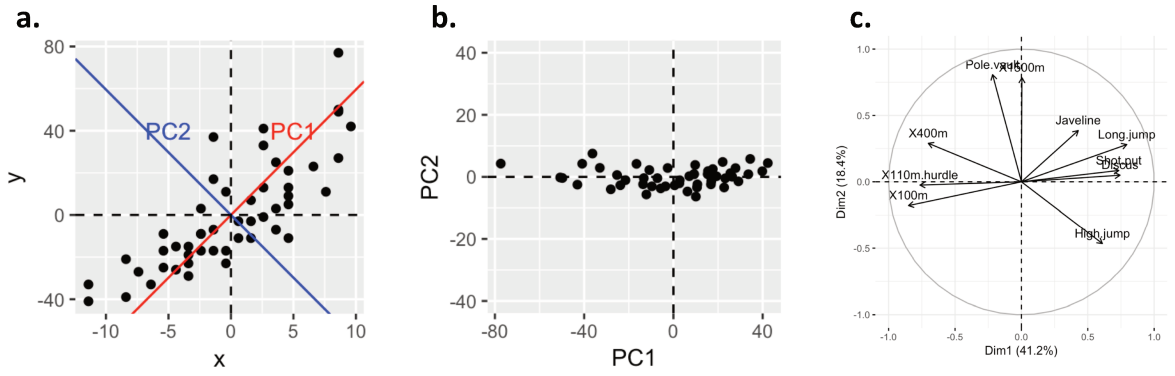


Figure 8 : ***Explication des données par analyse en composantes principales (ACP).*** *(a) Ici, l'axe PC1 est l'axe principal le long duquel les échantillons ont la plus grande variation, PC2 est le second axe avec la plus grande variation respectant la contrainte d'orthogonalité à PC1. (b) Il est ensuite possible de représenter les données en les projetant selon ces deux axes. (c) Les composantes principales 1 et 2 peuvent être décomposées pour déterminer les variables les plus contributrices (Abdi and Williams, 2010).*

1.3.2. Groupes moléculaires de CHC

Les CHC présentent une grande diversité de facteurs de risque, d'évolutions cliniques et de gènes *driver*. En analysant le transcriptome de ces tumeurs sur puce à ADN ou par séquençage RNAseq, plusieurs groupes dont le nôtre ont proposé des classifications moléculaires des CHC (Boyault et al., 2007; Chiang et al., 2008; Hoshida et al., 2009; Wheeler and Roberts, 2017). Dans notre laboratoire, une classification en 6 groupes (G1-G6), associée à des caractéristiques cliniques et moléculaires distinctes a été proposée sur la base d'un clustering hiérarchique de 57 CHC.

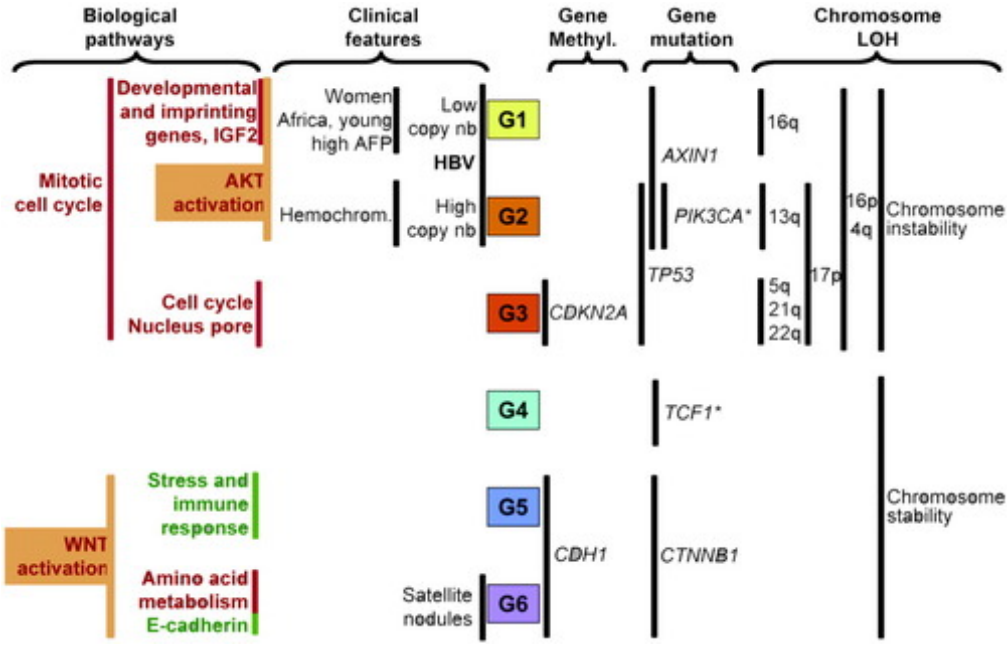


Figure 9 : ***Schématisation des différents sous-groupes de CHC de G1 à G6 définis par l'analyse du transcriptome avec leurs voies cliniques et génétiques associées par l'équipe du laboratoire*** *(Boyault et al., 2007). Les lignes verticales indiquent les caractéristiques associées de façon significative. Le rouge et le vert indiquent respectivement les principales voies de signalisation surexprimées et sous-exprimées. LOH = perte d'hétérozygotie ; hemochrom = hémochromatose ; AFP = alpha-fœtoprotéine ; VHB = virus de l'hépatite B ; *caractéristique rare.*

Les groupes G1 à G3 présentent une forte activation du cycle cellulaire, associée à une instabilité chromosomique, alors que les groupes G4, G5 et G6 sont plus stables. Les groupes G5 et G6 rassemblent les tumeurs présentant des mutations activatrices du gène CTNNB1, qui induit une activation de la voie de signalisation Wnt/B-catenine. Ils se différencient par l'altération d'autres voies. Les tumeurs du groupe G5 montrent une inhibition de la réponse immunitaire, alors que celles du groupe G6 présentent une activation du métabolisme des acides aminés et une inhibition de la E-cadhérine, qui promeut la migration des cellules tumorales. Le groupe G6 est notamment associé à la présence de nodules satellites. Parmi les

3 groupes qui activent les gènes du cycle cellulaire, G1 et G2 sont associés aux mutations du gène *AXIN1*, à l'infection par le virus de l'hépatite B et à l'activation de la voie AKT. On retrouve dans le groupe G1 un enrichissement en patients de sexe féminin, plus jeunes, originaires d'Afrique, avec un taux plus élevé d'alpha-foetoprotéine dans le sang, alors que les patients avec une hémochromatose appartiennent majoritairement dans le groupe G2. La mutation inactivatrice du gène TP53, qui code la protéine p53 régulatrice du cycle cellulaire est retrouvé significativement associée avec les groupes G2 et G3.

Les travaux de Chiang *et al.,* publiés un an après, ont séparé les CHC en 5 groupes en utilisant le clustering hiérarchique de 91 échantillons analysés sur puces à ADN.

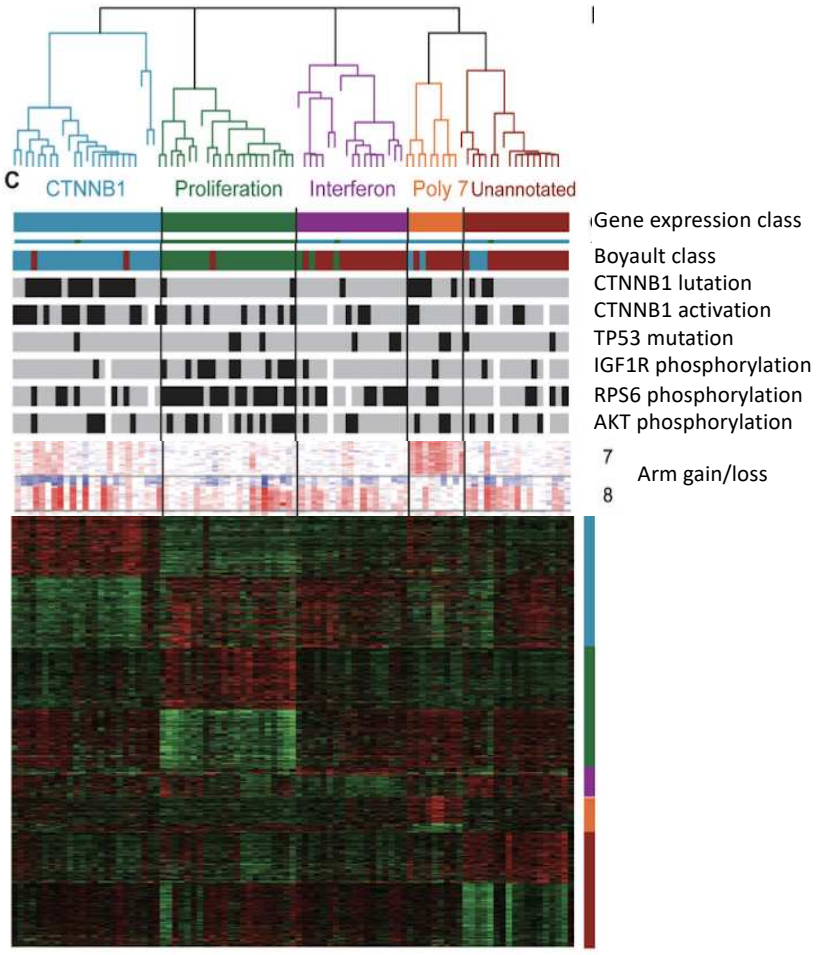


Figure 10 : ***Classification des altérations génomiques et moléculaires dans les carcinomes hépatocellulaires*** *(Chiang et al., 2008). Clustering hiérarchique consensus de 91 tumeurs. Les couleurs dans le dendrogramme des échantillons indiquent les 5 classes d'expression génique déterminées, associées aux modifications du nombre de copies. L'état de mutation (CTNNB1 ou TP53) et la coloration immunohistochimique sont indiqués en niveaux de gris : présent (noir), absent (gris) ou données manquantes (blanc).*

Le groupe « CTNNB1 » identifié dans cette classification présente une activation de la voie Wnt/ß-catenine par mutation activatrice de *CTNNB1* et correspond aux groupes G5-G6 identifiés au laboratoire. Le groupe « Prolifération » est similaire aux groupes G1-G2-G3 et montre un taux élevé de phosphorylation de RPS6. L'élément nouveau de cette classification consiste en la subdivision du groupe G4 en 3 groupes aux caractéristiques distinctes : le groupe « interféron » lié à l'infiltration des leucocytes, le groupe « polysomy 7 » caractérisé par le gain du chromosome 7 et l'absence de gain du bras chromosomique 8q ; et un dernier groupe encore non annoté (Chiang et al., 2008).

Plus tard, une autre classification, basée sur la méta-analyse de 8 études transcriptomiques du foie, a été réalisée dans l'optique de définir une classification universelle des carcinomes hépatocellulaires (Hoshida et al., 2009). Sur un total de 603 patients aux étiologies et origines diverses, 3 groupes de CHC ont été déterminés (S1, S2 et S3).

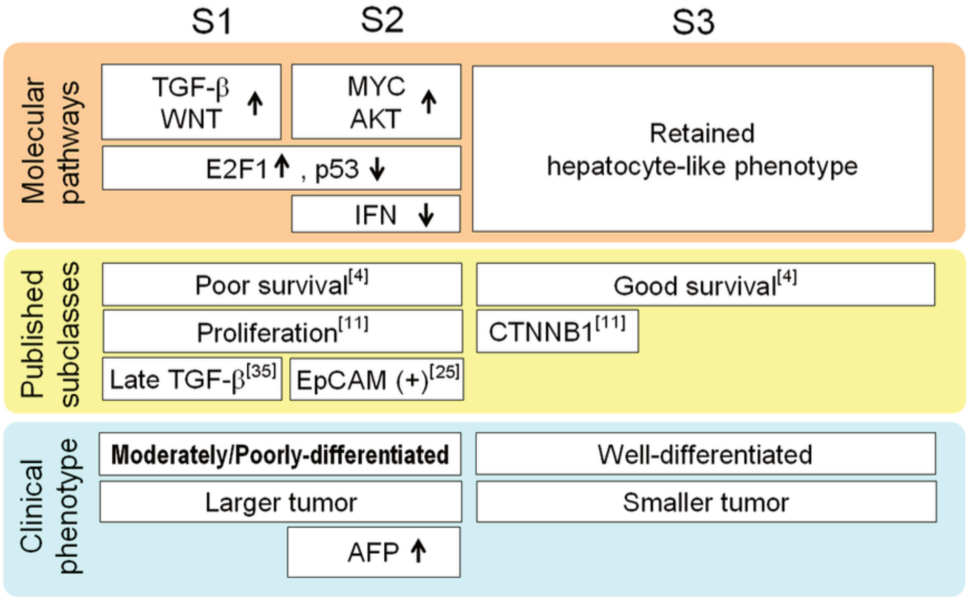


Figure 11 : ***Schématisation des différents sous-groupes de CHC de S1 à S3 définis par l'analyse du transcriptome avec leurs voies cliniques et génétiques associées*** *(Hoshida et al., 2009). AFP = Alpha-foetoproteine.*

Les groupes S1 et S2 présentent une augmentation de l'expression du gène E2F1 et une diminution de TP53. Les tumeurs de ces groupes sont significativement plus grosses et moyennement à peu différenciées. Comparées à la classification G1-G6 décrite précédemment, elles sont classées dans les groupes considérés comme proliférateur avec activation du cycle cellulaire G1, G2 et G3 et présentent souvent une mutation de *TP53*. Le

groupe S1 présente également une activation des voies TGF-B et de la voie Wnt, tandis que le groupe S2 active MYC et la voie AKT et est associé à un taux élevé d'alpha-foetoprotéine dans le sang. Des études fonctionnelles ont indiqué que l'activation de la voie Wnt dans le groupe S1 n'est pas le résultat d'une mutation activatrice de la β-caténine mais de l'activation du TGF-β (transforming growth factor β). Le groupe S3 regroupe les tumeurs qui conservent un phénotype hépatocytaire. Ce sont majoritairement des petites tumeurs bien différenciées associées à une bonne survie. Dans ce dernier groupe, une partie correspond aux tumeurs classées comme G5-G6, avec une activation de la voie Wnt dans la classification du laboratoire (Boyault et al., 2007).

La dernière classification majeure des carcinomes hépatocellulaires est une classification multi-omique de 363 patients (Wheeler and Roberts, 2017). Elle est basée sur le clustering non supervisé de cinq types de données : les données de séquençage d'exome, avec analyse de mutations *driver* et du nombre de copies de l'ADN ; les données de méthylation de l'ADN ; d'expression des gènes, des miARN, et des protéines. De cette manière, trois sous-types ont été identifiés (cf.Figure 12).

Le premier cluster intégré, iClust1, est associé à des patients jeunes, d'origine ethnique asiatique, de sexe féminin et de poids corporel normal. Ces tumeurs sont généralement peu différenciées avec une invasion macrovasculaire et présentent une surexpression de gènes marqueurs de prolifération tels que *MYBL2, PLK1* et *MKI6*. Les deux autres clusters, iClust2 et iClust3 ont une fréquence élevée de répression du gène CDKN2A par hyperméthylation de son promoteur, un fort taux de mutations de CTNNB1 et du promoteur TERT comparé à iClust1. La corrélation avec les variables cliniques a révélé une association d'iClust2 avec des tumeurs très différenciées et moins d'invasion microvasculaire, alors qu'iClust3 est caractérisé par une forte instabilité chromosomique avec perte du 17p, et une fréquence élevée de mutation *TP53*.

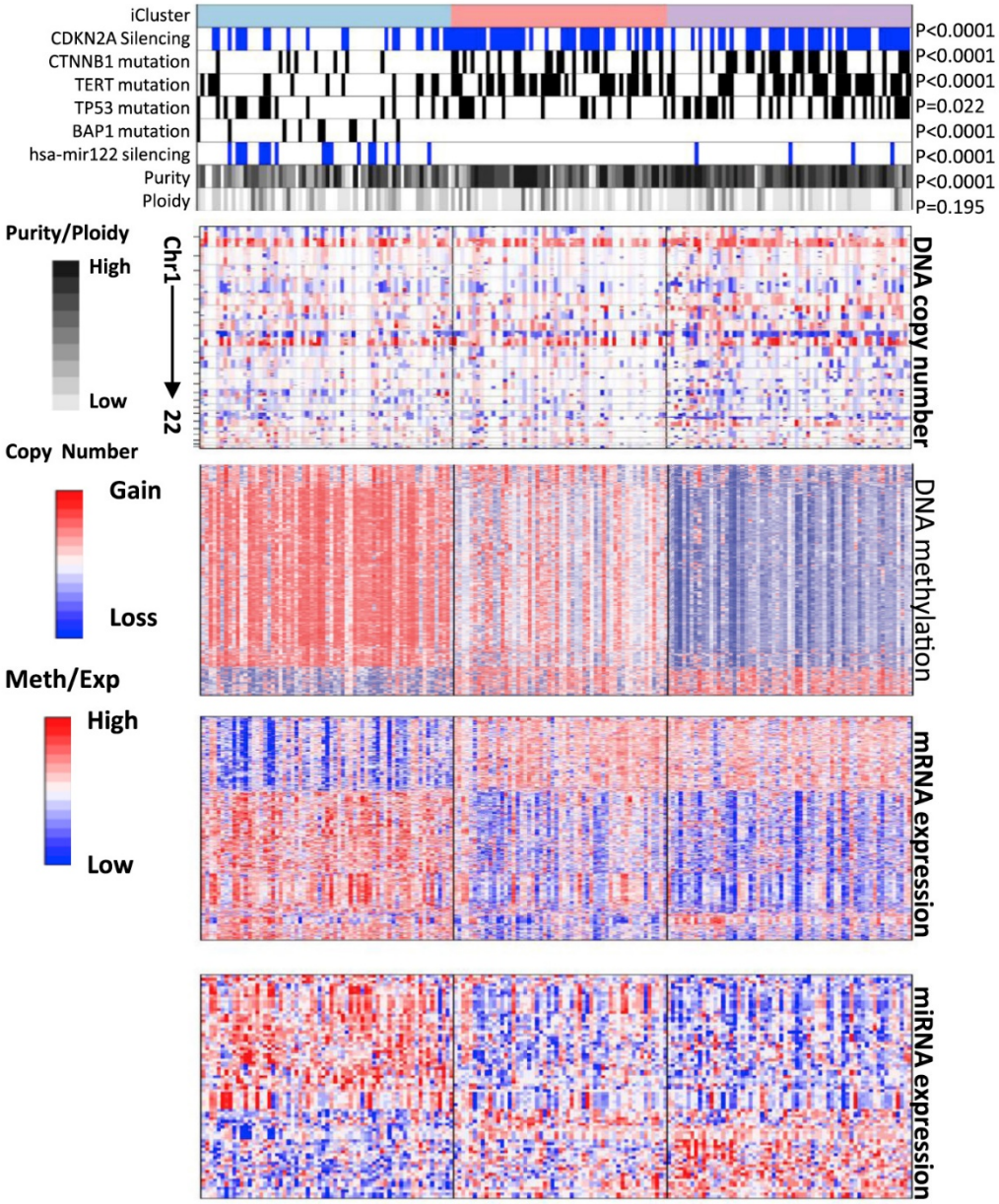


Figure 12 : ***Identification de trois sous-types moléculaires à partir du clustering multi-plateforme de données de cancer du foie*** *(Wheeler and Roberts, 2017). Heatmaps organisées par groupe iCluster en fonction du nombre de copies d'ADN, du statut de méthylation de l'ADN, de l'expression de l'ARNm et de l'expression de l'ARNm et de l'expression de l'ARNm et corrélées avec certaines caractéristiques moléculaires (traces supérieures). Les tumeurs sont en colonnes, regroupées par iCluster.*

Ainsi, plusieurs classifications majeures ont ainsi été établies pour les CHC (Boyault, Hoshida, Chiang, Wheeler), qui présentent à la fois des similarités et des spécificités propres. Cette diversité peut s'expliquer par l'hétérogénéité des séries en termes d'étiologie et d'origine géographique, par des différences méthodologiques et par des frontières relativement poreuses entre certains groupes. Il est néanmoins nécessaire de proposer une classification consensus des CHC, et ce sera l'un des objectifs de ma thèse.

1.3.3. Signatures pronostics et phénotypiques

Au-delà des groupes moléculaires, plusieurs équipes ont proposé des signatures de gènes permettant de caractériser le pronostic ou le phénotype des carcinomes hépatocellulaires. Notre équipe a notamment identifié un score pronostic, basé sur l'expression de 5 gènes (*HN1*, *RAN*, *RAMP3*, *KRT19* et *TAF9*), associé à la survie spécifique dans une série de 189 patients traités par résection. Ce score est associé à la survie spécifique en analyse multivariée, indépendamment des autres caractéristiques cliniques et pathologiques (Nault et al., 2013).

On retrouve également diverses signatures liées aux voies biologiques impliquées dans la carcinogénèse hépatique. Une signature d'activation de la voie TGF-B, un facteur de croissance connu pour ses propriétés oncogéniques, permet d'identifier des tumeurs associées à un phénotype invasif, une récidive tumorale accrue et une réduction significative du temps de survie moyen (Coulouarn et al., 2008). Une autre signature associée à une mauvaise survie est la signature de la voie oncogénique *MET*. Cette signature, corrélée à une augmentation du taux d'invasion vasculaire du foie, est retrouvée dans un sous-ensemble de CHC et toutes les métastases hépatiques (Kaposi-Novak et al., 2006). Le microenvironnement joue également un rôle dans le développement tumoral. Partant de l'hypothèse qu'il pouvait induire une plus grande agressivité tumorale, une équipe a modélisé un environnement d'hypoxie chronique, et déterminé une signature de sept gènes associée à un mauvais pronostic dans les CHC (van Malenstein et al., 2010). L'étude d'un marqueur de cellules souches hépatiques, la molécule d'adhésion des cellules épithéliales (EpCAM), a également permis d'identifier une signature moléculaire distincte avec des caractéristiques des cellules progénitrices hépatiques chez certains CHC. Ceux ne présentant pas cette signature au contraire présentent les caractéristiques des hépatocytes matures (Yamashita et al., 2008).

2. La méthylation de l'ADN

Au niveau moléculaire, l'épigénétique rassemble toutes les modifications influençant l'expression des gènes indépendamment des changements dans la séquence d'ADN elle-même. Cela comprend la méthylation de l'ADN, les modifications des histones, le remodelage de la chromatine et l'action des ARNs non codants comme les microARNs. La chromatine constitue le support de ces modifications épigénétiques (Nicoglou and Merlin, 2017),

nécessaires au développement embryonnaire et à la différenciation des différents types de cellules au sein des organismes multicellulaires (Khavari et al., 2010). Les altérations épigénétiques sont à l'origine de plusieurs pathologies (Ladd-Acosta and Fallin, 2015). Par exemple, le syndrome de Beckwith–Wiedemann, causé par un défaut de méthylation de l'ADN au niveau du locus soumis à empreinte parentale des gènes *IGF2* et *H19*, est caractérisé par des malformations congénitales et une forte prédisposition au cancer (Mussa et al., 2016). La méthylation de l'ADN a été caractérisée dans de nombreuses tumeurs et constitue un biomarqueur de risque intéressant en raison de sa spécificité et de sa stabilité dans les échantillons humains (Werner et al., 2017). Contrairement aux altérations de séquences, certains changements épigénétiques sont réversibles, ce qui en fait des cibles intéressantes pour le développement de nouvelles stratégies thérapeutiques (Kelly and Issa, 2017).

2.1. Régulation et distribution de la méthylation

2.1.1. Mécanismes de méthylation et déméthylation de l'ADN

La méthylation est l'attache ou la substitution d'un groupement méthyle (radical alkyle hydrophobe dérivé du méthane (CH_4)) sur un substrat. La méthylation de l'ADN est une marque épigénétique qui intervient sur les bases nucléiques, le type de méthylation varie selon les espèces. Chez les mammifères, on retrouve majoritairement la fixation covalente d'un groupe méthyle à la position C5 de l'anneau cytosine dans un contexte CpG (Holliday and Pugh, 1975), même si elle a aussi été identifiée de manière beaucoup plus rare dans d'autres contextes (Lister et al., 2009)**.** Les groupements méthyles sont ajoutés par les membres de la famille des enzymes méthyltransférases d'ADN (DNMTs).

L'enzyme DNMT1 est la première méthyltransférase de l'ADN à avoir été caractérisée (Bestor and Ingram, 1983). Il s'agit d'une enzyme qui maintient la méthylation, notamment au moment de la réplication cellulaire, ce qui permet une hérédité des schémas de méthylation des cytosines (Prokhortchouk and Defossez, 2008). Les enzymes DNMT3A, DNMT3B sont des méthyltransférases homologues qui ajoutent de nouvelles marques de méthylation sur l'ADN, ce que l'on nomme méthylation *de novo* (Okano et al., 1999). Elles sont fortement exprimées pendant les premières phases du développement, puis leur expression diminue. DNMT3A est le principal acteur dans le mécanisme des gènes soumis à empreinte, c'est à dire la répression épigénétique de l'allèle transmis par l'un des deux parents. DNMT3B, quant à lui, semble plus

actif dans la méthylation des CpG situés dans les zones centromériques (Bestor, 2000). On retrouve également une autre protéine homologue, DNMT3L, qui ne présente aucune activité catalytique quand elle est seule, mais semble être un cofacteur de DNMT3A (Aapola et al., 2000). L'inactivation des DNMT3 bloque l'établissement de la méthylation *in vitro* et *in vivo* (Chen and Li, 2004). Historiquement, l'enzyme DNMT2 a été identifiée comme méthyltransférase car présentant une séquence similaire au domaine C-terminal de la DNMT1 (Wilkinson et al., 1995), mais des travaux ultérieurs ont montré qu'elle n'était active que sur l'ARN de transfert ; elle a donc été renommée méthyltransférase d'ARN de transfert (TRDMT1) (Schaefer et al., 2010).

Le rôle de ces méthyltransférases de l'ADN est essentiel, mais reste encore mal compris (Bestor, 2000). L'élimination des gènes qui les codent chez la souris cause une mort précoce au cours du développement ou peu après la naissance (Bird, 2002). Les enzymes de la famille des DNMTs sont responsables de la méthylation de l'ADN, mais il existe également des mécanismes impliqués dans la déméthylation des cytosines, qui n'est pas exclusivement passive (Paroush et al., 1990). Des études montrent que les protéines TET (Ten-Eleven Translocation enzymes) 1, 2, ou 3 ont un lien avec la déméthylation active de l'ADN (Bhutani et al., 2010 ; Bhutani et al., 2011). Elles catalysent l'oxydation successive de la 5-méthylcytosine (5mC) en 5-hydroxyméthylcytosine (5hmC), 5-formylcytosine (5fC), et 5-carboxylcytosine (5caC) (He et al., 2011). Ces produits d'oxydation de 5mc sont des intermédiaires dans la conversion de 5mC en cytosines non modifiées, fournissant les premières étapes d'une voie de déméthylation active de l'ADN (Rasmussen and Helin, 2016).

2.1.2. Distribution de la méthylation le long du génome

La distribution de la méthylation le long du génome est très inégale. Ce déséquilibre est dû à la déamination spontanée des cytosines méthylés en thymines, entraînant une sous-représentation de CpG (équivalente à 21 % de celle attendue dans le génome humain), sauf dans les régions non méthylées (Bird, 2002; Irizarry et al., 2009; Klose and Bird, 2006). On observe notamment de petites régions fortement déméthylées et extrêmement riches en CpG, d'environ 1kb, nommées îlots CpG ou CpG Island (CGI) (Illingworth and Bird, 2009). Chez l'homme, on en retrouve environ 45 000 CpG qui chevauchent 60 % à 70 % des régions promotrices des gènes connus (Larsen et al., 1992; Saxonov et al., 2006; Weber et al., 2007).

Le reste de la chromatine est globalement pauvre en CpG, à l'exception des régions de 2kb et 4kb adjacentes aux îlots, appelées respectivement shore et shelf, qui présentent une quantité intermédiaire de CpG (cf. Figure 13).

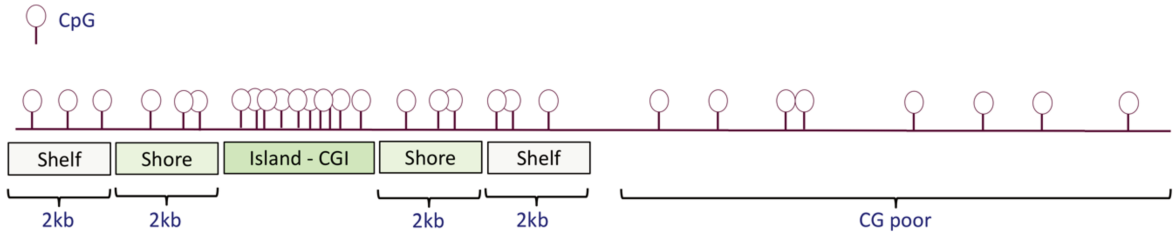


Figure 13 ***: Schéma des domaines de répartition des CpG.*** *Ils sont déterminés en fonction de la densité de CG sur la séquence d'ADN (Island ou îlots CpG = région riche en CG) et de la distance autour de ses régions riches en CG.*

2.1.3. Domaines de méthylation

La première étude de méthylation des CpG à l'échelle du génome, à partir de cellules souches embryonnaires humaines et de fibroblastes fœtaux a mis en évidence des niveaux de méthylation réduits dans les séquences des fibroblastes associés à une faible activité transcriptionnelle (Lister et al., 2009). A partir de ces résultats, plusieurs travaux ont cherché à segmenter le génome des différents tissus en fonction de leur méthylation, notamment en utilisant le modèle de Markov caché (HMM) (Burger et al., 2013). Les résultats obtenus le décomposent en 4 classes distinctes, d'un côté les grands domaines méthylés, allant de 100 kb à 20 Mb, subdivisés en "High Methylated Domain" (HMD - methylation > 80 %) et "Partially Methylated Domain" (PMD méthylation entre 50 % et 80 %), et de l'autre côté des régions courtes de régulation qui sont de deux types : les régions faiblement méthylées "Low Methylation Region" (LMR entre 50 et 10 %) et les régions non méthylées "UnMethylated Region" (UMR <10%).

Les niveaux de méthylation sont associés à des caractéristiques de la chromatine, comme aux domaines de réplication, des segments d'ADN d'environ 400-800 kb répliqués dans un ordre spécifique pendant la phase S du cycle cellulaire (Dileep et al., 2015). Dans la plupart des tissus humains normaux, les "High Methylated Domain" (HMD) constituent la grande majorité du génome, et on les retrouve associés à des domaines de réplication précoce. De leur côté, les "Partially Methylated Domain" (PMD), avec leur méthylation entre 50 % et 80 %, sont associés à des domaines de réplication tardive de l'ADN (Berman et al., 2011), enrichis en modifications d'histones caractéristiques de l'hétérochromatine, comme H3K27me3, pauvres en gènes et

moins actives. Les PMD peuvent représenter un pourcentage variable du génome, jusqu'à 74 %, et sont différents en fonction des tissus et de l'origine des cellules étudiées, ce qui fait d'eux une caractéristique claire pour déterminer le type cellulaire (Salhab et al., 2018).

2.2. Lien avec la régulation des histones

2.2.1. Structure du nucléosome

La méthylation de l'ADN coopère avec d'autres mécanismes épigénétiques, notamment l'organisation du nucléosome. Le nucléosome, unité de base d'organisation de la chromatine, est constitué de 8 histones centraux autour desquels s'enroulent environ 147 bases du brin d'ADN (Luger et al., 1997). Il permet le repliement et la compaction plus ou moins importante de l'ADN. Les histones ne sont pas répartis uniformément, et sont mobiles, ce qui confère une structure dynamique à la chromatine (cf. Figure 14).

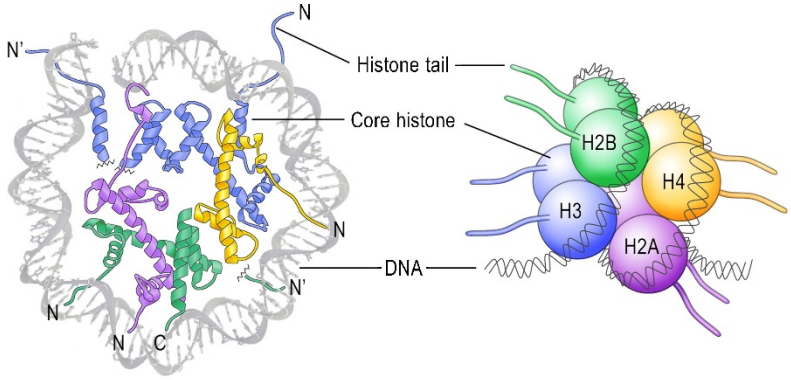


Figure 14 *: **Représentation du nucléosome.** Vue le long de l'axe de l'hélice d'ADN montrant la moitié de la structure du nucléosome et représentation schématique des histones du noyau à quatre nucléosomes, H2A, H2B, H3 et H4 (Gräff and Mansuy, 2008).*

2.2.2. Modifications post-traductionnelles des histones

La mobilité des histones est en partie due à la diversité des modifications post-traductionnelles qui les affectent, essentiellement au niveau de leur queue d'histone, comme la méthylation, l'acétylation, la phosphorylation, l'acétylation et l'ubiquitination (Castillo et al., 2017) (cf. Figure 15). Ces marques épigénétiques sont apposées et retirées par des enzymes différentes selon la position de la lysine et le type de marquage (Hyun et al., 2017). Le type de modification et le résidu de queue d'histone concerné jouent un rôle important dans la régulation transcriptionnelle, la réparation, la réplication de l'ADN, l'épissage alternatif et la condensation chromosomique (Portela and Esteller, 2010), et

peuvent aussi créer des sites de liaison spécifiques pour d'autres protéines ou des complexes enzymatiques (Bannister and Kouzarides, 2011).

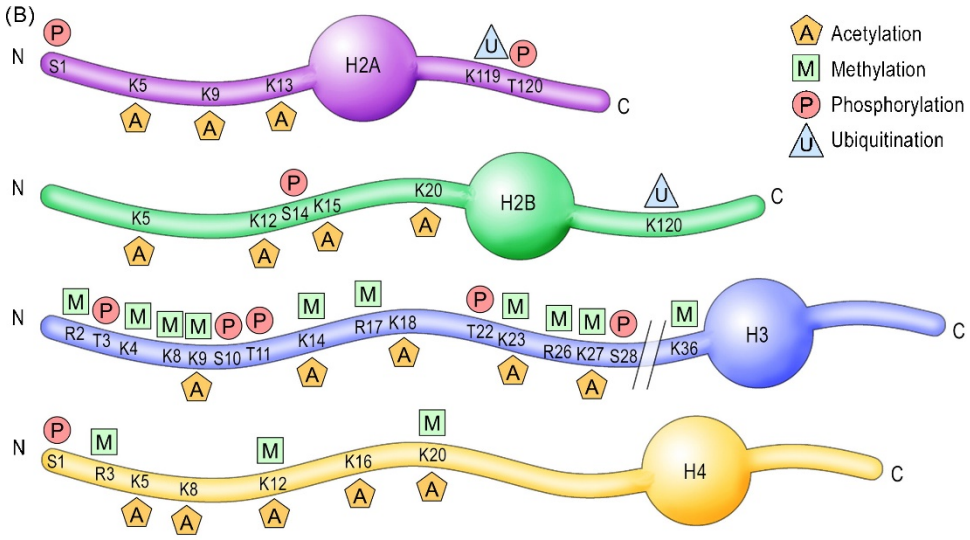


Figure 15 : ***Marques épigénétiques sur les queues d'histone.*** *Représentation schématique des N- et C-terminaux des histones centrales et de leurs modifications épigénétiques spécifiques aux résidus (Gräff and Mansuy, 2008).*

La méthylation de l'ADN et les modifications post-traductionnelles des histones interagissent les uns avec les autres via divers mécanismes et s'influencent mutuellement. La méthylation des histones peut aider à orienter les modèles de méthylation de l'ADN, et la méthylation de l'ADN semble servir de modèle pour reconstruire les modèles de modification des histones après réplication de l'ADN (Cedar and Bergman, 2009). Par exemple, dans leur état méthylé, les CpG sont reconnus par des protéines présentant un Methyl-CpG Binding Domain (MBD). Ces protéines recrutent les enzymes de modifications post-traductionnelles des histones, par exemple les enzymes qui enlèvent la marque d'acétylation de la Lysine (HDAC), ce qui a pour effet de compacter un peu plus l'ADN autour des histones, le rendant alors moins accessible (Nan et al., 1998) (cf. Figure 16). Dans ces zones, on retrouve également un haut niveau de méthylation des résidus H3K9, H3K27 et H4K20. Au contraire, un état non méthylé des CpG, ainsi que la perte d'un nucléosome directement en amont du TSS sont nécessaires à l'activation d'un gène. Si ce n'est pas le cas, la transcription n'a pas lieu (Li et al., 2007). Les domaines chromatiniens ouverts permettant la transcription, sont caractérisés par de hauts niveaux de triméthylation de H3K4, d'acétylation d'H3K4, d'H3K27 et d'H2BK5 de la

méthylation de H4K20. Dans le corps des gènes transcrits, les résidus H3K79 et H4K20 sont fortement méthylés (Peterson and Laniel, 2004).

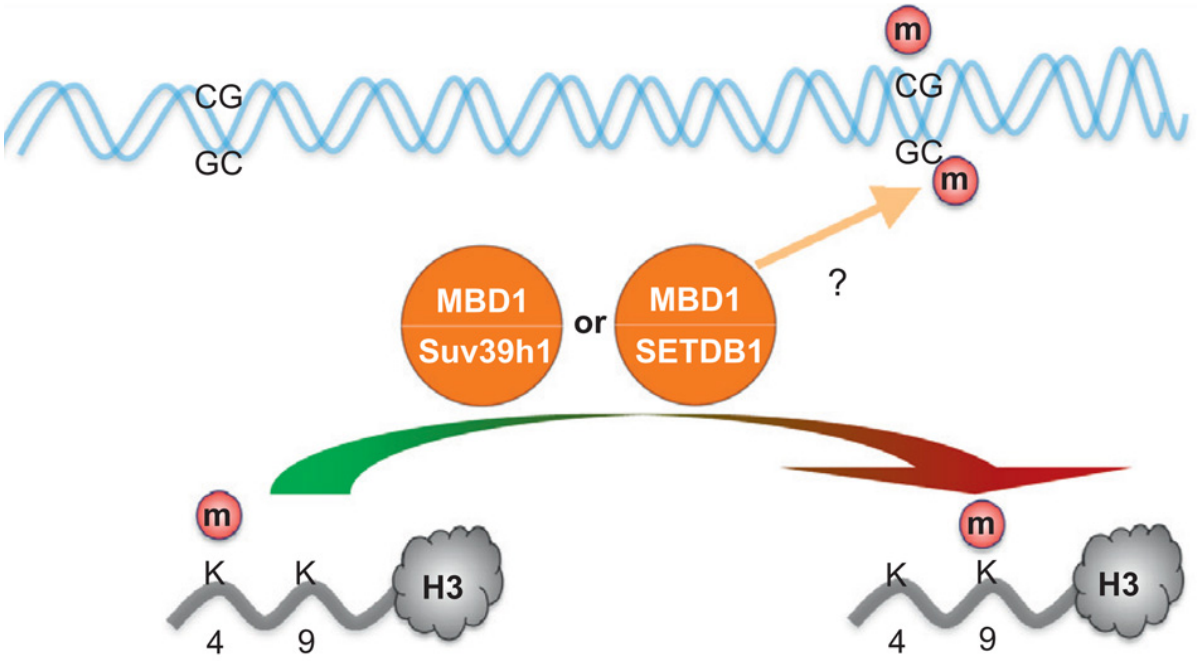


Figure 16 : ***Protéines avec un domaine MBD***. *Modèle des réactions qui régulent la méthylation du H3K9 par SUV39H1 ou SETDB1. La reconnaissance des CpG méthylés par le domaine méthyl-CpG de MBD1 dirige la méthylation du H3K9me3 (Cheng and Blumenthal, 2011).*

2.2.3. Cartographie de la chromatine

La combinaison des différents niveaux de régulation épigénétique définit différents états chromatiniens, fortement associés à la régulation de l'expression des gènes (Kundaje et al., 2015). L'analyse à large échelle des différentes marques épigénétiques est essentielle pour comprendre et modéliser cette régulation. De gros projets de cartographie de la chromatine et de ses caractéristiques ont donc été entrepris dans différents tissus et types de cellules, notamment par le NIH Roadmap Epigenomics Mapping Consortium (Kundaje et al., 2015) et le projet Encyclopedia of DNA Elements (Davis et al., 2018; ENCODE Project Consortium, 2012). Grâce à diverses techniques de séquençage, ces projets ont produit des cartes de haute qualité à l'échelle du génome de plusieurs modifications importantes des histones (ChIP-seq), de l'accessibilité de la chromatine (DNase-seq), de la méthylation de l'ADN (whole genome bisulfite sequencing), de l'expression des gènes (RNAseq) et de la fixation des facteurs de transcription (ChIP-seq) dans des centaines de types de cellules et tissus humains.

Des modèles statistiques ont ensuite été développés pour intégrer différentes marques épigénétiques et définir des états caractéristiques de la chromatine. Ainsi, le programme

ChromHMMM (Ernst and Kellis, 2012), basé sur un modèle de Markov caché multivarié, a permis de définir 18 états chromatiniens à partir de 6 marques d'histones (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3 et H3K27ac) (Kundaje et al., 2015) (cf. Figure 17). Chaque état chromatinien correspond à une combinaison particulière de marques histones, associée à un timing de réplication plus ou moins précoce, et à une expression plus ou moins active des gènes.

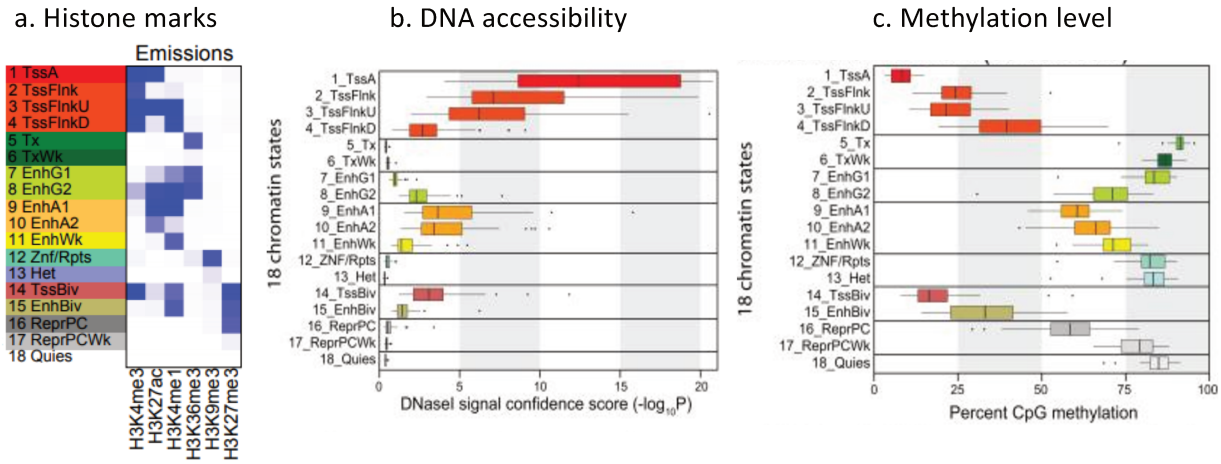


Figure 17 : ***Relation entre les domaines, les marques d'histones, la méthylation de l'ADN et l'accessibilité de l'ADN.*** *a. Marques d'histones caractéristiques de chaque domaine b. Accessibilité de l'ADN de chaque état chromatinien. c. Niveau de méthylation pour tous les états chromatiniens dans le modèle à 18 états - WGBS (Roadmap Epigenomics Consortium et al., 2015).*

On retrouve 11 états plutôt actifs et 7 inactifs. Les états actifs correspondent aux TSS (1) et à leurs zones adjacentes (2-4), au corps des gènes transcrits (5-6) et aux enhancers (7-11). Ces domaines sont associés à des modifications d'histones identifiés comme activatrices de la transcription et caractéristiques des gènes exprimés. Des combinaisons de marques d'histones permettent de distinguer les différents types de régions au sein des TSS et des enhancers. Les états inactifs incluent la chromatine quiescente (18), la chromatine réprimée par des complexes protéiques du groupe des polycomb qui triméthylent le résidu H3K27 (16-17), la chromatine bivalente, (14-15), l'hétérochromatine (13) et la chromatine ZNF/Rpts (12). Cette dernière est caractérisée uniquement par la trimethylation d'H3K9. Les domaines de chromatine bivalente (14-15) présentent à la fois la marque active H3K4me1 et la marque répressive H3K27me3. On la retrouve au niveau des régulateurs proximaux et distaux de gènes importants pour le développement, au niveau des gènes soumis à empreinte parentale, réprimés au niveau de l'allèle d'un des deux parents, et dans les cellules souches

embryonnaires. La presence de marques activatrices et repressives permet de réguler étroitement et d'activer rapidement l'expression génique au cours de différents processus de développement (Kanayama et al., 2019). La quantification de chaque de domaines chromatinien montre que le génome est majoritairement constitué de chromatine inactive quiescente (18), pauvre en CpG. Les caractéristiques associées à chacun des 18 domaines, comme l'accessibilité de l'ADN et le niveau de méthylation ont été étudiées (cf. Figure 17). On peut voir que pour la chromatine active, les domaines correspondants à des TSS sont faiblement méthylés et fortement accessibles, les enhancers sont en majorité moyennement méthylés et faiblement accessibles, alors que les domaines transcrits ont une méthylation extrêmement forte et correspondent à de l'ADN sous forme compactée, non accessible. Au niveau de la chromatine inactive, les domaines 16, 17 et 18 sont fortement compactés, mais présentent des niveaux de méthylation différents : la chromatine quiessante est fortement méthylée alors que les domaines chromatiniens réprimés par les protéines Polycomb ont une méthylation plus faible. Pour la chromatine bivalente, on retrouve une méthylation similaire à celle des TSS mais une plus faible accessibilité.

2.3. Lien avec l'expression

2.3.1. Impact de la méthylation du promoteur des gènes

Majoritairement, les ilôts sont déméthylés. Leur profil de méthylation joue néanmoins un rôle clé dans la différenciation cellulaire. En effet, la méthylation des ilôts CpG (CGI) localisés aux promoteurs des gènes entraîne la répression de leur transcription (Renfree et al., 2013), et la déméthylation spécifique des gènes dans certains tissus est associée à leur activation transcriptionnelle (Meissner et al., 2008). La répression de la transcription par l'hyperméthylation des CpG dans un promoteur résulte de plusieurs mécanismes. D'un côté, elle empêche les facteurs de transcription de se lier à l'ADN (Bird, 2002). De l'autre, elle crée des sites de liaison spécifiquement reconnus par les MBPs (Methyl-Binding Protein), qui se lient à l'ADN méthylé et recrutent des complexes de remodelage de la chromatine. Ces derniers la condensent sous forme d'hétérochromatine, ce qui la rend inaccessible aux complexes de transcription et empêche l'expression des gènes (Sasai and Defossez, 2009) (cf. Figure 18). Il faut noter que la méthylation des CpG dans le promoteur d'un gène réprimé n'est pas systématique, il s'agit d'un mécanisme de régulation d'un nombre restreint de sites qui induit la formation de profils de méthylation tissus et cellules spécifiques (Weber et al., 2007).

Des études à grande échelle menées sur la méthylation dans les tissus normaux de plusieurs organes ont ainsi permis d'identifier des CpG différentiellement méthylés selon les tissus, supposés participer à la différentiation cellulaire. On les retrouve dans certains Îlots CpG, mais surtout dans les *shores*, les régions de 2kb directement adjacentes aux îlots. Ces observations soutiennent un rôle fonctionnel pour les *shores*, dans lesquels la méthylation des CpG est étroitement liée à l'expression des gènes (Irizarry et al., 2009).

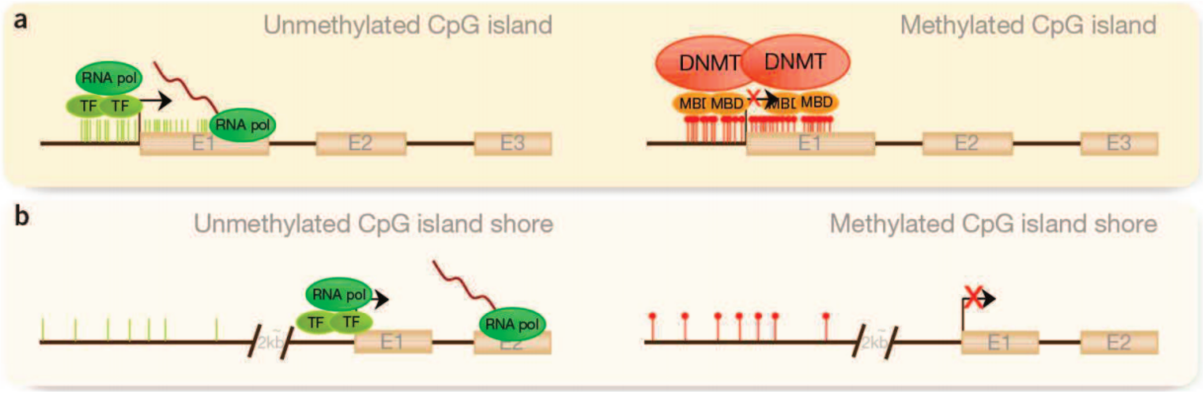


Figure 18 ***: Schéma de l'interaction entre la méthylation et la transcription des gènes****. a. dans les ilôts CpG b. dans les shores.*

2.3.2. La méthylation du corps des gènes

En dehors des promoteurs, on retrouve également des CpG dans le corps des gènes. Ceux-ci sont peu denses et généralement fortement méthylés (80 % à 90 %) (Lister et al., 2009). La méthylation de ces CpG, caractéristiques des gènes transcrits, est positivement corrélée à leur expression. Ainsi, le traitement de cellules en culture par un agent déméthylant induit non seulement la réactivation de gènes dont le promoteur était hyperméthylé, mais diminue aussi l'expression de gènes activements transcrits dont le corps est fortement méthylé (Yang et al., 2014). La fonction de la méthylation du corps génétique n'est pas encore bien comprise, mais un ensemble de preuves suggère qu'elle pourrait être liée à l'efficacité de l'allongement de la transcription et empêcherait l'initiation de transcription intragénique en ne permettant pas à la machinerie de transcription de se fixer à des éléments transposables ou des promoteurs cryptiques (Wolf et al., 1984; Zilberman et al., 2007). La méthylation dans le corps des gènes est dépendante de la méthyltransférase de l'ADN DNMT3b, qui la protège de la fixation de l'ARN polymérase II. Cette fonction de DNMT3b dépend de son activité enzymatique, mais

aussi de son recrutement par H3K36me3, particulièrement enrichi dans le corps des gènes (Neri et al., 2017). Cette marque est catalysée par l'enzyme SETD2 en même temps que l'élongation transcriptionnelle par l'ARN polymérase II (RNA Pol II) (Wagner and Carpenter, 2012) (cf. Figure 19).

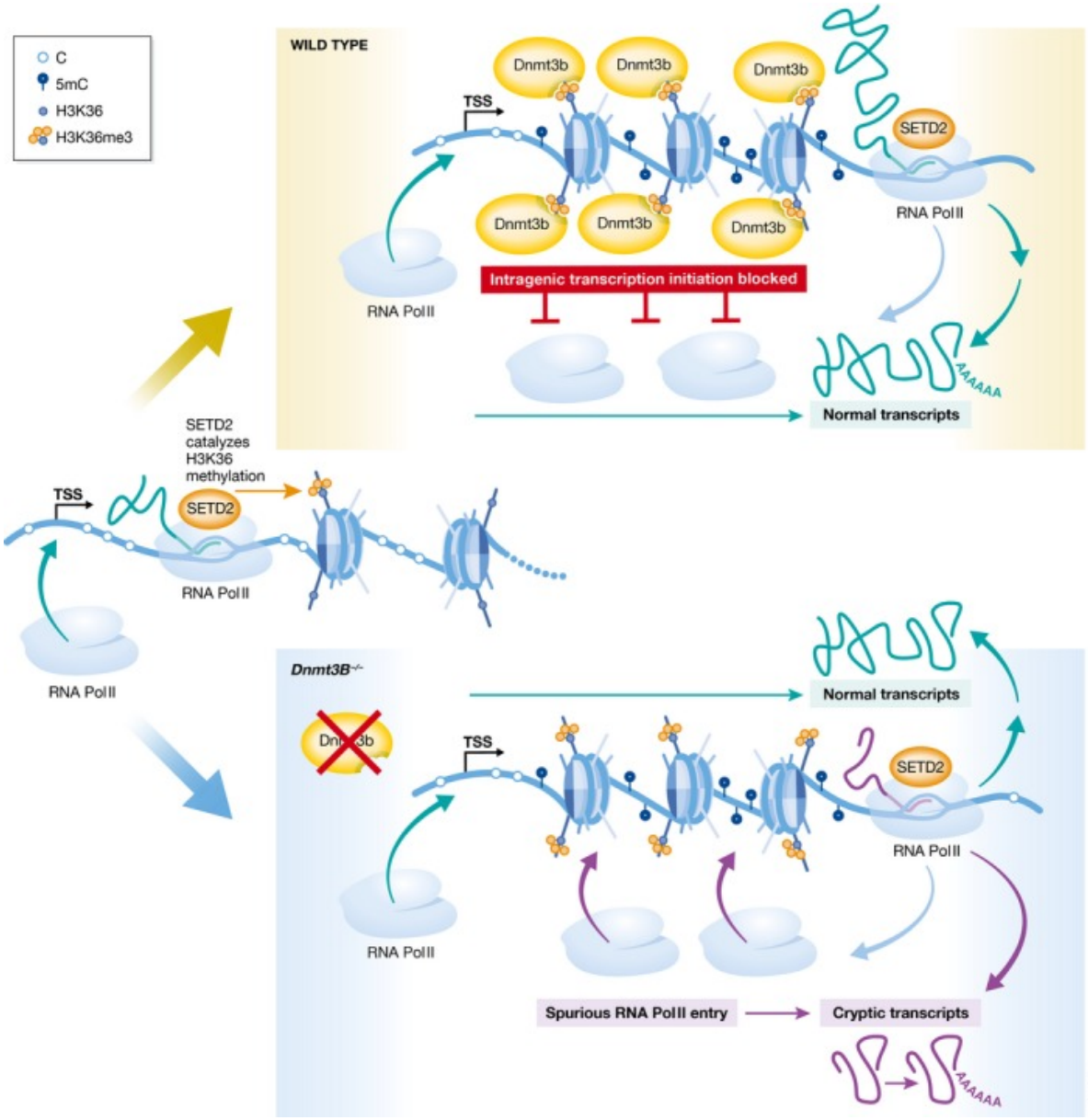


Figure 19 : ***Rôle de la méthylation de l'ADN dans le corps des gènes.*** *La méthylation intragénique restreint l'initiation de la transcription par l'intermédiaire de DNMT3B- et H3K36me3. Dans les cellules DNMT3B KO, les régions intragéniques hypométhylées sont pénétrées par l'ARN polymérase II, ce qui conduit à des initiations de la transcription cryptique et à la génération de faux transcrits. L'ARN Pol II recrute l'histone méthyltransférase SETD2 pendant l'allongement de la transcription pour catalyser les marques H3K36me3. (Teissandier and Bourc'his, 2017).*

2.3.3. La méthylation des régions de régulation distales

En plus des promoteurs, régions régulatrices proximales, la méthylation des régions de contrôle distales comme les enhancers (qui augmentent l'expression des gènes) et les

répresseurs ou insulateurs (qui la diminuent), permet également de moduler l'expression des gènes (Lister et al., 2009). Les changements de méthylation de l'ADN au niveau des éléments activateurs permettraient même de prédire plus efficacement les changements d'expression génétique dans les cancers que les promoteurs (Aran et al., 2013).

Parmi les domaines de méthylation décrits précédemment, les deux régions les moins méthylées, LMR et UMR, beaucoup plus focales que les HMD et PMD, correspondent aux domaines de régulation de la transcription (Salhab et al., 2018). Les UMR regroupent les CpG avec une méthylation < 10 %, c'est à dire les îlots non méthylés ainsi qu'une partie de leurs shores, et sont enrichis aux promoteurs. Ces régions correspondent aux éléments de régulation proximale. Les régions faiblement méthylées (LMR) présentent une méthylation moyenne de 30 % et correspondent aux régions régulatrices distales. Elles ne sont généralement pas des îlots CpG puisqu'elles ont une teneur en CpG plus faible, sont plus courtes et plus éloignées des sites d'initiation de la transcription (TSS). On retrouve généralement les enhancers dans ces régions, leur état de méthylation étant étroitement lié à l'expression. Les LMRs se forment dynamiquement lors de la différenciation, pilotés par des facteurs de transcription spécifiques à chaque type cellulaire (Stadler et al., 2011). Plus de 4 % de tous les CpG sont dans les LMR.

Étudier l'impact transcriptionnel de la méthylation des éléments distaux n'est pas aisé car ceux-ci sont plus difficiles à identifier, et surtout à relier à leur(s) gène(s) cible(s), que les promoteurs. Des méthodes statistiques ont été développées, comme l'outil ELMER, qui corrèlent données de méthylation et d'expression pour (1) identifier les enhancers différentiellement méthylés entre deux conditions, (2) identifier les gènes régulés par chaque enhancer, et (3) caractériser les facteurs de transcription en amont via une analyse de motifs autour des CpG altérés (Yao et al., 2015) (cf. Figure 20). Cette méthode, récemment étendue au-delà des enhancers à l'ensemble des CpG distaux (Silva et al., 2019), a notamment permis d'identifier des réseaux transcriptionnels dérégulés dans les cancers par des facteurs de transription clés comme *GATA3* et *FOXA1* (cancer du sein), ou encore *NFE2L2*, *SOX2* et *TP63* (cancer du poumon à cellules squameuses).

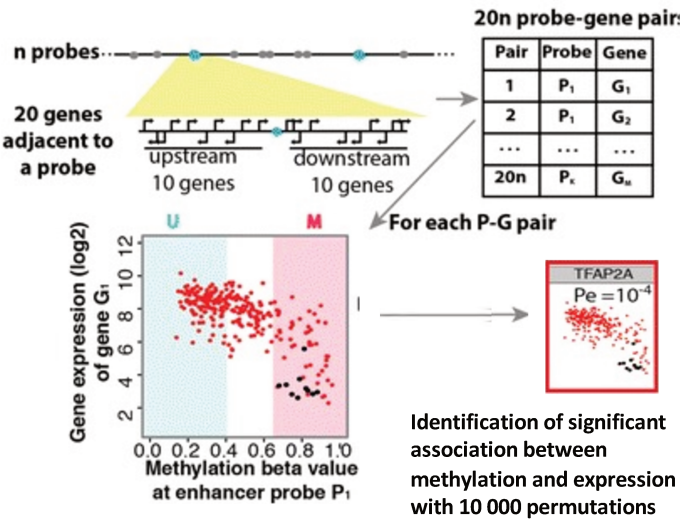


Figure 20 : ***Principe de détection de couple CpG-gène implementé dans le package ELMER***. *(Silva et al., 2019).*

3. Dérégulations épigénétiques et cancers

Les mécanismes épigénétiques sont essentiels au développement normal et au maintien de l'expression tissus spécifique des gènes. Leur altération peut modifier l'activité des gènes et promouvoir la transformation maligne des cellules (Sharma et al., 2010). L'étude à l'échelle du génome de la méthylation dans différents cancers met en évidence une hyperméthylation locale des CpG ainsi qu'une hypométhylation globale (Berman et al., 2011; Eden et al., 2003; Ehrlich, 2009; Esteller et al., 2001).

3.1. Altérations locales de méthylation

3.1.1. Méthylation anormale dans les promoteurs des oncogènes ou gènes suppresseurs de tumeurs

Compte tenu du lien entre méthylation et expression décrit précédemment, la méthylation aberrante des ilots CpG au niveau des promoteurs a été particulièrement étudiée. L'hyperméthylation du promoteur est reconnue par des protéines à domaine MBD (methyl-CpG binding domain) qui recrutent des DNMTs et des protéines impliquées dans la modification des queues d'histone. Ce recrutement permet de maintenir la méthylation de l'ADN, d'enlever les marques d'acétylation des histones et de triméthyler H3K19 (cf. Figure 21). Ces mécanismes contribuent à diminuer l'accessibilité du promoteur aux complexes de transcription, et donc la transcription du gène (Deltour et al., 2005). Ainsi, plusieur gènes de

réparation de l'ADN sont inactivés par l'hyperméthylation de CpG dans leur promoteur, comme *BRCA1* dans le cancer du sein (Dobrovic and Simpfendorfer, 1997), *MGMT* dans les cancers du côlon, du poumon et des lymphoïdes (Esteller et al., 1999) et *hMLH1* associé à de l'instabilité microsatellite dans les cancers du côlon (Herman et al., 1998). A l'inverse, l'hypométhylation de CpG dans un promoteur peut entraîner l'activation d'oncogènes habituellement réprimés. Par exemple, dans le cancer du poumon non à petites cellules (NSCLC), le gène *ELMO3* est surexprimé dans les tumeurs primaires de patients présentant des métastases distantes. L'augmentation de son expression coïncide avec une hypométhylation dans la région promotrice de ELMO3 (Søes et al., 2014).

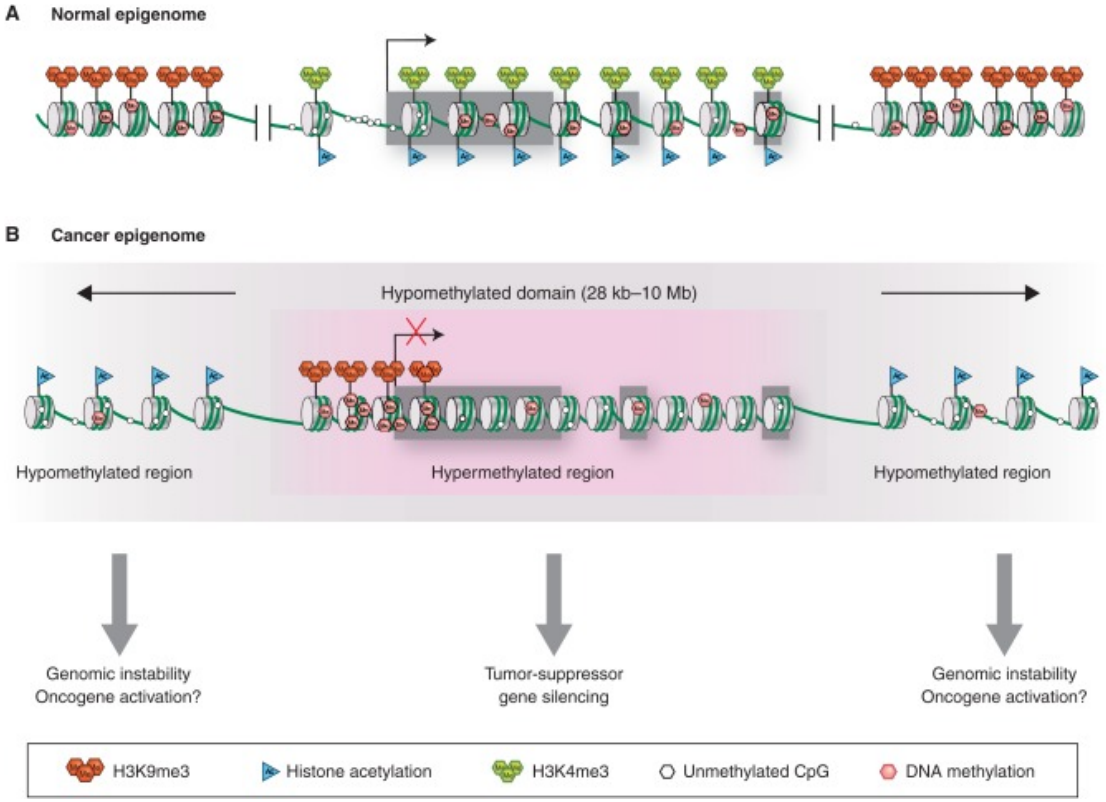


Figure 21 ***: Profil de méthylation de l'ADN et compaction de la chromatine inversés dans les cellules cancéreuses.*** *L'analyse des taux de méthylation entre cellules normales et cellules tumorales, montre un niveau de méthylation plus élevé de certains ilôts CpG, associé à une extinction du gène situé en aval, à une perte des marques H3K4me3, remplacées par H3K9m3. Il est également montré une diminution de méthylation au niveau des séquences répétées, une perte d'H3K9me3 et une acétylation des queues d'histone, ce qui participe à une instabilité du génome.*

Les îlots CpG ne représentent qu'une partie de la variabilité de méthylation entre les différents tissus, un grand nombre des CpG différenciellement méthylés étant retrouvés dans les *shores*, adjacents aux îlots. En effet, dans le cancer du côlon, la plupart des changements de méthylation (à la fois hyper- et hypométhylation) ont lieu dans les shores. De plus, entre 45 % à 65 % de ces changements affectent des régions rétrouvées différenciellement méthylées

entre les différents tissus normaux (Irizarry et al., 2009). Cette observation soutient l'hypothèse que la dérégulation de la différenciation cellulaire est l'un des principaux mécanismes par lequel la méthylation participe au développement de tumeurs (Feinberg et al., 2006).

On peut distinguer deux causes possibles aux altérations de la méthylation de gènes impliqués dans la transformation tumorale. Soit il s'agit d'une survenue aléatoire qui donne un avantage sélectif, soit il s'agit du résultat de l'altération d'un oncogène qui va modifier la méthylation dans les promoteurs d'autres gènes. On sait par exemple que les altérations qui touchent les méthyltransférases ou les protéines qui interagissent avec les DNMTs sont impliquées dans plusieurs types de cancers, dont les leucémies myéloïdes aiguës (Wong et al., 2019). Une altération forte des mécanismes de méthylation, par la perte de *DNMT1*, mène à l'apoptose (Zhou et al., 2017), à des altérations durant la mitose (Chen et al., 2007) et à une importante instabilité chromosomique, ce qui peut entraîner la transformation maligne des cellules (Eden et al., 2003). De même, l'activation de l'oncogène *K-ras* entraîne la transformation de cellules en recrutant plusieurs effecteurs épigénétiques qui méthylent directement les promoteurs des gènes cibles, comme celui du gène pro-apoptotique *Fas* (Gazin et al., 2007). L'étude des gènes altérés dans les cancers a montré que de nombreux gènes touchés étaient impliqués dans la régulation épigénétique par l'établissement de la méthylation, le remodelage de la chromatine, et la modification des histones (Klose and Bird, 2006). Néanmoins, la conséquence exacte de l'altération de ces gènes sur le méthylome des cellules tumorales reste à élucider (Witte et al., 2014).

3.1.2. Perte de l'empreinte parentale

On retrouve également des altérations locales de la méthylation près des gènes soumis à empreinte parentale. En effet, la perte de l'empreinte à l'intérieur de ces régions régulatrices peut perturber l'expression des gènes et conduire à la survenue de maladies ou favoriser le développement de cancer. Par exemple, l'altération du locus 11p15.4-5 peut entraîner la survenue du syndrome de Beckwith–Wiedemann (BWS), un trouble de régulation de la croissance présentant une surcroissance somatique et une prédisposition aux tumeurs embryonnaires, notamment aux hépatoblastomes, des tumeurs pédiatriques du foie (Brioude et al., 2018). Au niveau de ce locus, on retrouve 2 domaines, chacun contenant une région de

contrôle soumise à empreinte parentale. Dans le domaine télomérique, on retrouve notamment l'ARN non codant exprimé par l'allèle maternel H19 et le facteur de croissance fœtale exprimé paternellement IGF2 (cf. Figure 22). La marque épigénétique associée à ces gènes (IC1), située entre les deux, est généralement méthylée sur le chromosome paternel et non méthylée sur le chromosome maternel. Dans 5 % des cas de BWS, IC1 présente un gain de méthylation sur le chromosome maternel, ce qui perturbe l'équilibre d'expression de ces gènes entre les 2 allèles parentaux, entraînant une répression de H19 et une surexpression d'IGF2. L'autre domaine soumis à empreinte parentale et impliqué dans la survenue de BWS est le domaine centromérique. Il contient les gènes KCNQ1, KCNQ1OT1, et l'inhibiteur du cycle cellulaire CDKN1C associé à la marque épigénétiques est IC2, retrouvée dans le gène CDKN1C et le promoteur de KCNQ1OT1 (cf. Figure 22). IC2 est méthylé sur le chromosome maternel, associé à l'expression de KCNQ1 et CDKN1C, est non méthylé sur le chromosome paternel où seulement KCNQ1OT1 est exprimé. 50 % des BWS présente une perte de méthylation d'IC2, entraînant une répression de KCNQ1 et de l'inhibiteur du cycle cellulaire CDKN1C ainsi qu'une surexpression de KCNQ1OT1, un long ARN non codant qui interagit avec la chromatine et régule la transcription de multiples gènes cibles par des modifications épigénétiques. Ce dernier est notamment impliqué dans la carcinogénèse des cellules du colon et associé à un mauvais pronostic (Zhang et al., 2019).

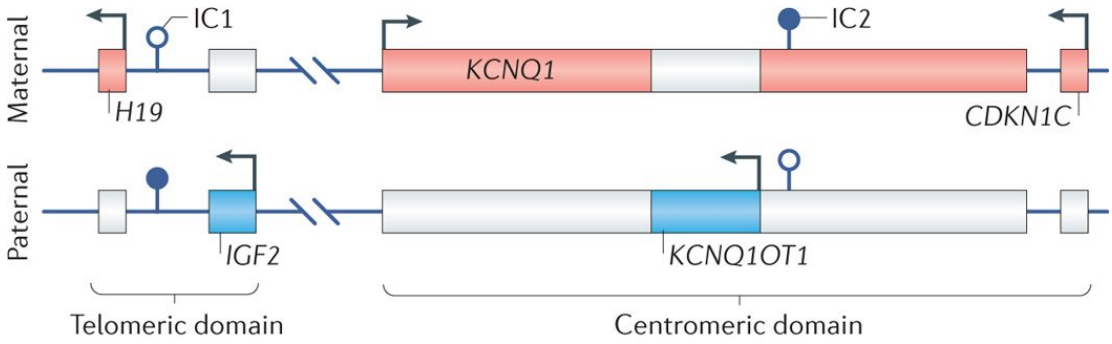


Figure 22 : ***Schéma représentant les gènes soumis à empreinte du locus 11p15****. Les gènes exprimés à partir du chromosome maternel sont représentés en rouge, et les gènes exprimés à partir du chromosome paternel en bleu. Les gènes en gris indiquent que les allèles ne sont pas exprimés. Les cercles indiquent les marques épigénétiques soumises à empreinte parentale, remplies elles sont méthylées, alors que vides elles ne le sont pas. Les flèches indiquent l'orientation de la transcription.*

3.2. Hypométhylation des domaines partiellement méthylés

Dès 1983, il a été mis en évidence que les cellules tumorales présentaient une déméthylation étendue des CpG comparée aux cellules de tissus normaux (Gama-Sosa et al., 1983). Cette hypométhylation globale du génome est une caractéristique commune à tous les types de

cancers (Ehrlich, 2009; Klein Hesselink et al., 2018; Pérez et al., 2018; Shen et al., 2017; Zelic et al., 2015) et peut rendre l'ADN plus sensible aux cassures ou aux remaniements chromosomiques (Baylin and Jones, 2016). Cette hypométhylation affecte principalement les 'partially methylated domains' (PMDs) (Berman et al., 2011). Ces domaines, qui couvrent plusieurs mégabases, présentant une méthylation entre 50 % et 70 % et sont un excellent indicateur du type cellulaire et de l'histoire de la prolifération de la cellule. Plusieurs travaux ont montré que les niveaux de méthylation de l'ADN dans différents organes évoluent avec l'âge (Wilson et al., 1987). Une diminution progressive des niveaux de méthylation de l'ADN a notamment été observée dans les régions pauvres en GpG. Cette perte progressive de méthylation pourrait être causée par des erreurs dans son maintien au cours de la réplication par DNMT1, qui a une fidélité d'environ 95 % par génération cellulaire (Stein et al., 1982; Wigler et al., 1981). Cette perte de méthylation associée à la réplication s'accélère dans les cancers suite à l'augmentation de la prolifération, en particulier au niveau des CpG isolés situés dans les PMDs, sans base C/G adjacente et sans autre CpG dans les 35 pb alentour (Zhou et al., 2018). L'hypométhylation de ces régions s'accumule avec le nombre de divisions cellulaires, en raison du timing de réplication tardif et de l'efficacité réduite des mécanismes de réplication de la méthylation dans les régions à très faible densité de CG (Zhou et al., 2018).

L'hypométhylation globale des tumeurs est un processus continu qui persiste tout au long de la vie des cellules tumorales, impliquant des mécanismes existants dans les cellules normales, mais amplifiés par la dérégulation des mécanismes cellulaires (Bell et al., 2019). L'altération des mécanismes responsables de la méthylation comme les méthyltransférases de l'ADN, leurs partenaires et les déméthylases, peut également participer à l'hypométhylation globale observée. Ainsi, dans les carcinomes hépatocellulaires, la surexpression d'UHRF1, une protéine qui identifie l'ADN hémiméthylé généré pendant la réplication de l'ADN et recrute ensuite des DNMT1 (Bostick et al., 2007), entraîne une hypométhylation globale de l'ADN (Mudbhary et al., 2014).

3.3. Phénotypes hyperméthylateurs

3.3.1. Hyperméthylation liée à l'âge

Les changements épigénétiques, et particulièrement la méthylation, sont très efficaces pour estimer de manière extrêmement précise l'âge chronologique d'un individu (Field et al., 2018;

Horvath and Raj, 2018). Leur changement avec l'âge du patient suggèrent un rôle dans les maladies liées au vieillissement (Jylhävä et al., 2017). L'utilisation de méthodes d'apprentissage supervisées pour l'estimation de l'âge chronologique, met en évidence, dans certains cas, une différence entre l'âge du patient et l'âge estimé avec les marqueurs de méthylation (Bell et al., 2019). Ce phénotype lié à l'âge peut être dû à une maladie, à des mesures cliniques de "fragilité" ou des phénotypes cellulaires, comme l'âge mitotique (le nombre total de divisions cellulaires durant la vie d'un tissu). Il est particulièrement marqué dans les tumeurs présentant une augmentation de la prolifération (Tomasetti and Vogelstein, 2015; Yang et al., 2016). En plus de la déméthylation de CpG située dans les PMD, d'autres changements de méthylation apparaissent avec l'âge, notamment une hyperméthylation des CpG dans les îlots (Day et al., 2013; Unnikrishnan et al., 2019).

Parmi les promoteurs hyperméthylés avec l'âge dans différents types de tissus, les domaines chromatiniens bivalents, présentant à la fois la marque active H3K4me1 et la marque répressive H3K27me3, sont particulièrement enrichis. Cette même catégorie de promoteurs, associée à des gènes clés du développement, présente une hyperméthylation encore plus marquée dans les cancers (Rakyan et al., 2010). Une des hypothèses proposée est que, dans un contexte cancéreux, l'hyperméthylation aberrante des promoteurs de la chromatine bivalente entraîne la répression permanente des gènes nécessaires à la différenciation, poussant ainsi les cellules souches vers un état prolifératif prédisposant au développement du cancer (Ohm and Baylin, 2007).

3.3.2. Phénotypes hyperméthylateurs

Au-delà de l'hyperméthylation classiquement observée dans les cancers, des phénotypes hyperméthylateurs ont été identifiés, caractérisés par l'hyperméthylation d'un grand nombre de CpGs situés dans des îlots et les promoteurs des gènes. Ce phénotype, qui définit un sous-ensemble de tumeurs présentant des aberrations épigénomiques importantes et une biologie distincte, a été initialement décrit dans le cancer colorectal (Toyota et al., 1999). Il a ensuite été trouvé dans beaucoup d'autres tissus dont le carcinome hépatocellulaire (Cheng et al., 2010; Toyota et al., 1999; Zhang et al., 2007), dans lequel il s'est révélé associé à un mauvais pronostic. Même si le terme de phénotype hyperméthylateur est utilisé pour plusieurs types

de tumeurs, les CpG ciblés ne sont pas forcément les mêmes entre les différents cancers et la cause de ce phénotype reste en majorité à identifier (Hughes et al., 2013).

Parmi les causes moléculaires identifiées, on retrouve 4 altérations qui conduisent à un phénotype hyperméthylateur via un mécanisme commun, l'inactivation des déméthylases de la famille TET. La première est la mutation inactivatrice d'un des membres de la famille des TET, par exemple TET 2 dans les leucémies (Hughes et al., 2013). On retrouve également, associées à ce phénotype, les mutations somatiques des isocitrates déshydrogénases-1 et 2 (IDH1/2), fréquemment observées dans les gliomes et les leucémies (Hughes et al., 2013; Turcan et al., 2012). Les gènes mutants IDH1 ou 2 produisent du (R)-2-hydroxyglutarate [(R)-2HG] qui s'accumule dans les cellules (Dang et al., 2010). Ce métabolite présente des similitudes structurales avec le 2-oxoglutarate (2-OG) et entraîne l'inhibition compétitive des déméthylases d'ADN de la famille TET. De plus, les enzymes du cycle de Krebs, la succinate déhydrogénase (SDH) et la fumarate hydratase (FH) sont des suppresseurs de tumeurs dont les mutations de perte de fonction prédisposent aux syndromes familiaux de cancer (Frezza et al., 2011). L'inactivation de la SDH et de la FH entraîne respectivement un blocage du cycle de Krebs, une altération de la respiration et une accumulation anormale de leurs substrats succinate et fumarate. Les deux métabolites sont également des inhibiteurs des dioxygénases 2OG-dépendantes (Yang and Pollard, 2013). On retrouve ce phénotype hyperméthylateur associé aux mutations des gènes codant pour les succinates deshydrogénases dans les parangliomes, des tumeurs neuroendocrines (Letouzé et al., 2013), et associé à la mutation des gènes codant pour les furamates hydratases (FHx) dans les carcinomes du rein (Pollard et al., 2007).

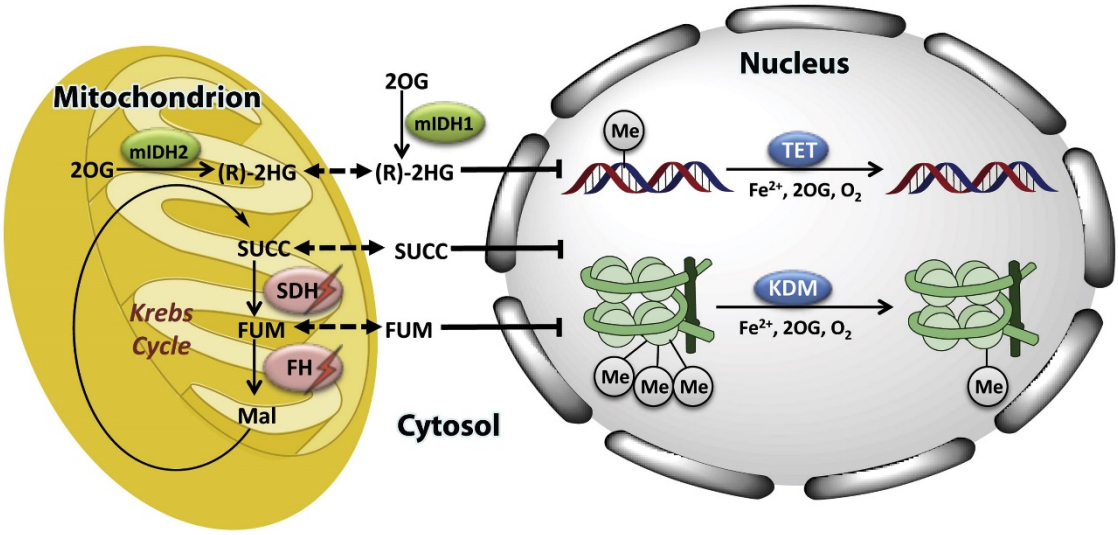


Figure 23 : ***Reprogrammation épigénétique par onco-métabolites****. Les mutations dans les enzymes métaboliques isocitrate déshydrogénase (IDH)-1 et -2, succinate déshydrogénase (SDH) et fumarate hydratase (FH) entraînent une accumulation anormale de (R)-2-hydroxyglutarate ((R)-2HG), succinate et fumarate, respectivement. Leur accumulation inhibe les activités des dioxygénases dépendantes du 2-oxoglutarate (2OG), y compris la famille des enzymes modifiant l'ADN de la TET et les déméthylases d'histone lysine contenant le domaine JmjC (KDMs). SUCC, succinate ; FUM, fumarate ; Mal, malate ; mIDH1/2, mutant IDH1/2 ; Me, groupement méthyle (Yang and Pollard, 2013).*

L'augmentation des onco-métabolites (R)-2HG, et (3)-2HG, succinate et furamate, perturbe le cycle de KREBS à différents niveaux dans les différents cancers. Toutes ces altérations convergent pour altérer les enzymes déméthylases TET (cf. Figure 23). Leur inhibition induit une hyperméthylation étendue de l'ADN. Ce changement et la chromatine, rendue inaccessible à la machinerie de transcription, bloquent la différenciation cellulaire des cellules présentant ce phénotype (Van der Auwera et al., 2009).

3.4. Altération de la méthylation dans les CHC

Les perturbations épigénétiques à l'œuvre dans le carcinome hépatocellulaire ont fait l'objet de plusieurs études pour caractériser ces changements et comprendre comment ils peuvent participer à la transformation maligne des hépatocytes.

3.4.1. Hyperméthylation dans les promoteurs

L'hyperméthylation touchant un grand nombre de CpG, il est intéressant de regarder plus précisément dans les tumeurs, quels sont les CpG fréquemment altérés, et si leur méthylation affecte la régulation de la transcription de gènes clés. Les gènes fréquemment altérés au niveau des promoteurs ne sont pas les mêmes en fonction des types de cancer (Esteller et al., 2001; Shen et al., 2012; Shin et al., 2010; Yang et al., 2003).

Le gène suppresseur de tumeurs *CDKN2A* est fréquemment inactivé par hyperméthylation dans les cancers du foie. Ce gène code pour plusieurs protéines grâce à un épissage alternatif des premiers exons, dont p16 et p19 (inhibiteurs de CDK4, importante pour la progression en phase G1 du cycle cellulaire) et p14 (stabilisateur de p53). Il est inactivé par délétion homozygote dans plusieurs cancers, dont celui du foie (Biden et al., 1997; Herman et al., 1995), entraînant une prolifération des cellules et l'apparition de métastases (Chen et al., 2015). Cependant, l'inactivation épigénétique par hyperméthylation du promoteur est de loin l'altération la plus fréquente de *CDKN2A* dans les CHC (53 % vs. 4 % d'inactivation génomique) (Wheeler and Roberts, 2017).

D'autres gènes sont fréquemment inhibés par hyperméthylation de leur promoteur, notamment *GSTP1*, *CDH1*, *APC*, *RASSF1* et *MGMT*. La fréquence d'altération de ces gènes varie en fonction des séries (Zhang et al., 2016). Une analyse par voie de signalisation montre que les gènes hyperméthylés sont fréquemment impliqués dans le cycle cellulaire, la prolifération, la différenciation, et la mort cellulaire (Song et al., 2013).

Enfin, les ROS (Reactive Oxygen Species) peuvent inhiber l'expression de CDH1, la E-Cadherin, en induisant des changements épigénétiques, dont une hyperméthylation dans son promoteur (Han et al., 2018; Lim et al., 2008). La E-cadhérine participe à des mécanismes régulant l'adhérence des cellules, la mobilité et la prolifération des cellules épithéliales (Meigs et al., 2002). L'inactivation de tous ces gènes suppresseurs de tumeurs, altère des mécanismes importants pour la cellule et joue un rôle dans la carcinogenèse hépatique.

3.4.2. Études globales de la méthylation

L'analyse à large échelle des profils de méthylation de l'ADN, avec une puce de méthylation Illumina ciblant 450k CpG répartis le long du génome, des carcinomes hépatocellulaire a mis en évidence une hypométhylation globale des CpG en dehors des gènes et des promoteurs, ainsi qu'une forte hyperméthylation des CpG dans les tumeurs (Yamada et al., 2016; Zheng et al., 2018). Toutes les tumeurs analysées semblent présenter une hypométhylation, même si elle n'affecte pas les mêmes CpG avec la même intensité (Mah et al., 2014; Zheng et al., 2018).

Plusieurs études ont également décrit un phénotype hyperméthylateur dans les CHC (Cheng et al., 2018; Mah and Lee, 2014). L'une d'elles, menée par le consortium TCGA (The Cancer Genome Atlas) sur 363 CHC, permet de voir que 196 CHC montrent une hyperméthylation, plus ou moins marquée selon les tumeurs (Wheeler and Roberts, 2017). Environ 15 000 sites CpG avec une hyperméthylation significative ont été identifiés, et forment 4 sous-groupes distincts, associés à l'origine géographique et les altérations de gènes *driver* comme *TP53*, *CTNNB1*, du promoteur de *TERT* et de la répression de CDKN2A (cf. Figure 24).

L'origine de ces profils d'hyperméthylation n'est pas bien compris, mais une mutation ou une sous-expression des déméthylases *TET* (Williams et al., 2012) pourraient en être à l'origine (Liu et al., 2019; Thomson et al., 2016). On remarque également que parmi le sous-groupe présentant l'hyperméthylation la plus marquée sont retrouvées les 3 tumeurs avec une altération d'*IDH1*, connue pour être associée à un profil hyperméthylateur dans d'autres cancers.

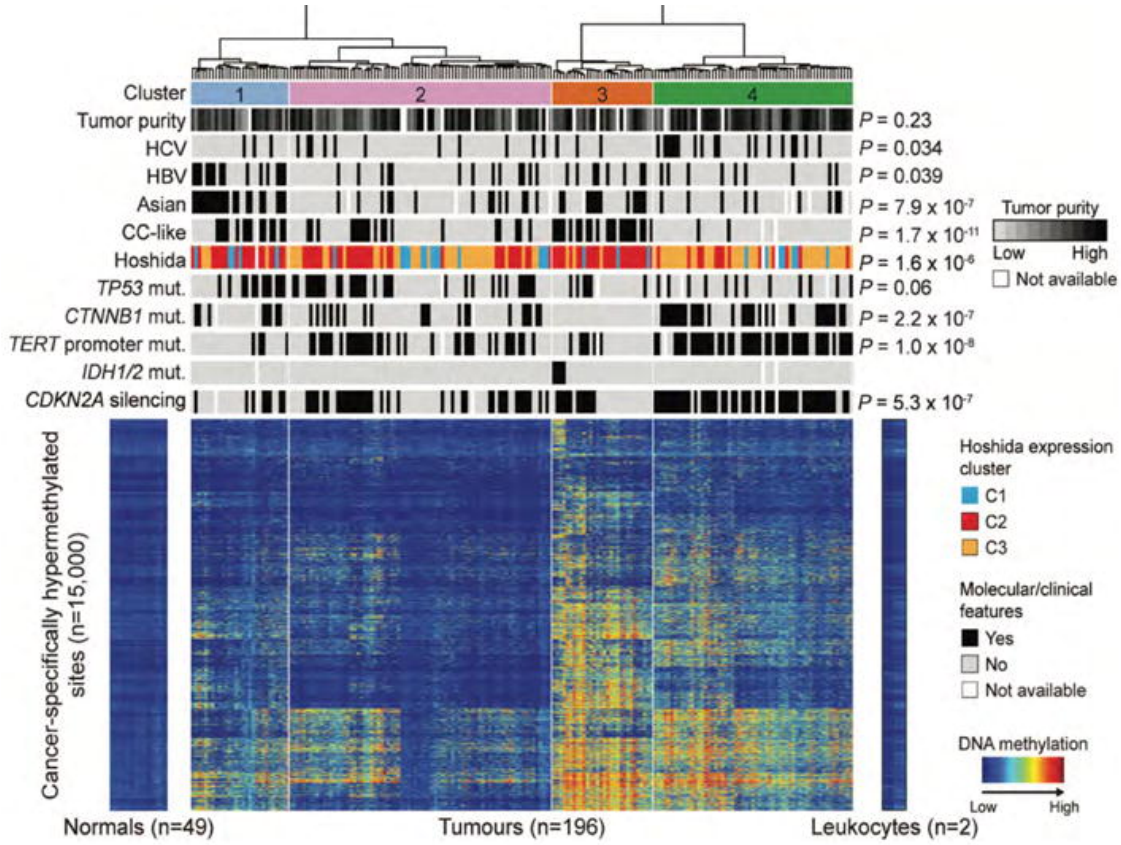


Figure 24 : ***Analyse non supervisée de l'hyperméthylation de 15 000 sites CpG dans les tumeurs comparées aux normaux.*** *Quatre sous-groupes distincts de CHC sont définis. Le heatmap représente la méthylation dans ces 15 000 sondes, et l'association des sous-groupes obtenus avec les annotations cliniques et moléculaires (Wheeler and Roberts, 2017).*

3.4.3. Pertinence clinique de la méthylation de l'ADN dans la prise en charge des carcinomes hépatocellulaires

La méthylation aberrante des promoteurs a été utilisée pour développer une signature pronostic des carcinomes hépatocellulaires, qui prédit avec précision la survie à partir de 36 marqueurs de méthylation de l'ADN (Villanueva et al., 2015). Cette signature permet d'identifier des tumeurs présentant des caractéristiques de cellules progénitrices, plus agressives, associées à un mauvais pronostic.

De plus, l'hyperméthylation de certains gènes est très caractéristique des altérations hépatiques. Le statut de méthylation de 5 gènes suppresseurs de tumeur (*APC, FHIT, CDKN2B, CDKN2A, CDH1*) dans les CHC a également été détecté avec succès dans le plasma (Han et al., 2018; Iyer et al., 2010). L'utilisation de l'hyperméthylation de ces promoteurs, détectable dans les cellules sanguines, en fait des bons marqueurs potentiels pour un diagnostic précoce et sans risque, contrairement à la biopsie qui est une procédure invasive (Hardy et al., 2017).

4. Objectifs de la thèse, stratégie et données

Élucider les processus transcriptionnels et épigénétiques dérégulés dans les cancers est fondamental pour mieux comprendre les voies biologiques impliquées et proposer une thérapie adaptée au phénotype moléculaire de chaque tumeur (Chang et al., 2009). Les approches classiques de classification non supervisée ont permis de définir, sur la base du transcriptome, les principaux groupes moléculaires pour chaque type tumoral, et notamment les tumeurs hépatiques (Boyault et al., 2007; Chiang et al., 2008; Hoshida et al., 2009). Cependant, ces groupes restent relativement hétérogènes et ne reflètent qu'imparfaitement la diversité des mécanismes biologiques à l'œuvre dans ces tumeurs. L'analyse du méthylome a également mis en évidence des groupes de tumeurs présentant des profils similaires dans différents types de cancer (Witte et al., 2014). Les changements de méthylation observés dans les tumeurs incluent une hypométhylation globale des régions partiellement méthylées et pauvres en CpG (Berman et al., 2011), une hyperméthylation des domaines bivalents de la chromatine (Rakyan et al., 2010), un phénotype hyperméthylateur plus ou moins marqué dans certaines tumeurs (Berman et al., 2011; Wheeler and Roberts, 2017) et une altération ciblée de la méthylation d'oncogène et de gènes suppresseur de tumeurs qui influe sur leur

expression. Divers processus sont susceptibles d'affecter le méthylome, comme l'âge (Ashapkin et al., 2017), les expositions environnementales (Ladd-Acosta and Fallin, 2015) ou l'altération de certains *driver* comme les *gènes IDH, SDH, TET* et *H3F3A (Chen et al., 2017; Dang et al., 2010).* Cependant, à quelques exceptions près, les mécanismes biologiques à l'origine des altérations de méthylation observées dans les tumeurs restent largement méconnus. En effet, les approches classiques consistent à définir des groupes de tumeurs puis à étudier les changements d'expression/de méthylation dans ces groupes.

Dans ce contexte, l'objectif de ma thèse était de développer des méthodes d'analyses innovantes pour explorer le transcriptome et le méthylome des tumeurs hépatiques afin d'identifier des signatures moléculaires reflétant plus finement les mécanismes biologiques dérégulés dans ces cancers. Pour cela, j'ai mis en place une méthode de déconvolution du signal (Analyse en Composantes Indépendantes, ACI) afin d'identifier des composantes transcriptionnelles et épigénétique fines dans les cancers du foie. J'ai ensuite exploré en détail ces composantes et leurs interactions afin d'identifier les mécanismes moléculaires sous-jacents.

4.1. Analyses en composantes indépendantes (ACI)

Les méthodes statistiques de déconvolution, telles que l'analyse en composantes indépendantes (ACI) ou la factorisation matricielle non négative (NMF), ont récemment trouvé de nombreuses applications en biologie moléculaire (Zinovyev et al., 2013). En effet, ces approches permettent de séparer différentes sources de variation mélangées dans un signal, ce qui est fréquent en biologie. Ainsi, le transcriptome d'une tumeur reflète à la fois la cellule d'origine, la dérégulation de diverses voies oncogéniques, l'abondance de l'infiltrat inflammatoire etc. Le méthylome reflète lui une superposition de changements liés à l'âge, aux altérations *driver* ou encore aux expositions environnementales. Les méthodes de déconvolution apparaissent donc prometteuses pour extraire les signatures individuelles de ces processus dans les données biologiques. Au cours de ma thèse, j'ai appliqué l'analyse en composantes indépendantes pour isoler les différentes signatures épigénétiques ou transcriptionnelles présentes dans les tumeurs (groupe de CpG altérés conjointement ou de gènes co-régulés) et élucider leur signification biologique.

4.1.1. Principe mathématique de l'ACI

L'analyse en composantes indépendantes est une méthode développée à l'origine pour résoudre les problèmes de séparation aveugle des sources (Jutten and Herault, 1991), c'est à dire, retrouver, uniquement à partir du mélange final, les signaux sources qui ont mené à ce mélange. L'exemple type pour ce problème est "la soirée cocktail". Au cours d'une soirée rassemblant plusieurs convives qui discutent par petits groupes de taille variable, le bruit total des conversations est considéré comme le mélange. L'objectif est d'isoler, à partir des différents micros positionnés dans une pièce, la voix de chaque personne. Pour cela, l'ACI considère que les voix de chaque invité enregistrées à un instant t sont indépendantes les unes des autres (Oja and Nordhausen, 2006) .

D'un point de vu mathématique, on a :

- S = (S1, S2, S3 … Sd) les sources indépendantes inconnues (dans notre exemple : les invités)
- X = (X1, X2, X3 … Xn) les observations (micros) dont les informations contenues sont dépendantes les unes des autres et varient ensemble. Statistiquement, cela signifie que la covariance(X) = 1
- A, la matrice de mélange inconnue de dimension d x n (sources x observations)

Ce qui donne le système X = AS
On cherche la matrice orthogonale A telle que le vecteur $S = A^TX$ soit composé de sources indépendantes.

Pour résoudre ce problème, de nombreux algorithmes ont été développés. Bien qu'ils soient basés sur des méthodes statistiques différentes, chacune de ces méthodes permet de mettre en évidence et de mesurer la propriété d'indépendance recherchée, toutes ces méthodes statistiques aboutissent à des résultats semblables (Cardoso, 1997; Lee et al., 2000). Parmi eux, Hyvärinen et Oja proposent d'utiliser une mesure de "non gaussianité". En effet, d'après le théorème central limite énoncé par Lindeberg-Levy (1920), la somme de variables indépendantes tend vers une distribution normale. Maximiser la non gaussianité entre les composantes S estimées permettrait donc d'assurer leur indépendance (Hyvärinen and Oja, 2000). Pour cela, plusieurs solutions ont été mises en place. L'une d'elle, proposée par

Hyvärinen, est implémentée dans l'algorithme FastICA, basé sur le principe de l'algorithme d'apprentissage itératif de type point fixe (Fixed point algorithm). Cet algorithme cherche à maximiser le coefficient de Kurtosis, un critère permettant de mesurer la répartition d'une variable aléatoire (Hyvärinen, 1999). L'étape préalable à l'extraction des composantes indépendantes est une étape de blanchiment des données. Cette transformation, implémentée dans l'algorithme, permet de décorréler et d'annuler la variance de la matrice initiale, grâce à l'Analyse en Composantes Principales (ACP). FastICA est très utilisé dans de nombreux domaines en raison de sa rapidité et de son absence de paramètres d'optimisation. Cependant, dans le cas où le nombre d'échantillons est grand, il devient très gourmand en temps de calcul et d'espace mémoire. Contrairement à l'ACP, les composantes indépendantes peuvent ne pas être orthogonales par rapport à l'espace de données original, ce qui peut être considéré comme un avantage dans certaines applications (Bugli and Lambert, 2007).

4.1.2. Application aux données biologiques

Cette méthode a été appliquée avec succès dans différentes études de l'expression des gènes (Biton et al., 2014; Lee and Batzoglou, 2003). L'analyse en composantes indépendantes cherche à retrouver des signaux sources sous l'hypothèse d'indépendance statistique. Appliquée aux données transcriptomiques, elle modélise le niveau d'expression de gènes dans plusieurs échantillons donnés comme une somme linéaire pondérée de plusieurs composantes indépendantes, où chaque composante capture l'effet de facteurs ou processus influençant l'expression des gènes (Liebermeister, 2002). Dans ce modèle, la matrice d'expression X qui a pour dimension le nombre d'échantillons sur le nombre de gènes est décomposée comme le produit de 2 matrices (cf. Figure 25). La première matrice correspond aux projections des gènes dans chaque composante et la deuxième correspond aux contributions des échantillons. Les gènes ayant la plus grande projection sur une composante sont les gènes les plus fortement influencés par le processus associé à cette composante. La deuxième matrice, de contribution des échantillons, reflète l'activité de la source de variation représentée par la composante dans les échantillons. En corrélant l'activité des composantes avec les annotations cliniques et moléculaires des échantillons, on peut donc identifier des associations permettant d'interpréter la signification biologique de chacune (Zinovyev et al., 2013). L'interprétation des composantes au niveau des gènes implique l'identification de voies de signalisation ou groupes de gènes particulièrement actifs dans la composante, en utilisant

par exemple la méthode GSEA d'analyse d'enrichissement par rang GSEA (Gene Set Enrichment Analysis) (Anders and Huber, 2010).

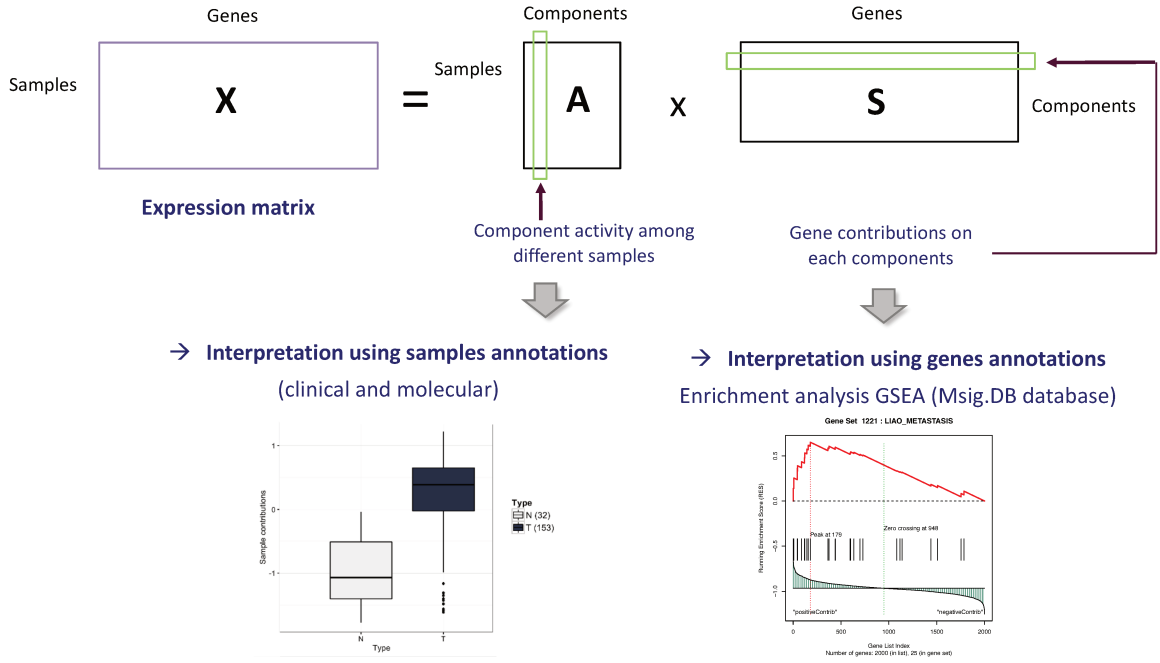


Figure 25 : ***Déconvolution de la matrice d'expression par analyse en composantes indépendantes et interprétation des matrices résultantes optenues.***

Appliquée aux données de méthylation, cette méthode permettrait de séparer les changements de méthylation liés à des processus différents (Freytag et al., 2017). La caractérisation de ces changements, au niveau des échantillons avec les annotations cliniques et altérations génomiques des tumeurs, ainsi qu'au niveau des CpG avec l'étude du contexte épigénétique dans lequel elles interviennent, pourrait nous permettre de mieux comprendre les altérations de méthylation présentes dans les tumeurs, ainsi que leur origine.

4.2. Jeux de données utilisés

Pour ce projet, j'ai analysé différents types de données (épi)génomiques générées au laboratoire sur les cancers du foie. J'ai principalement travaillé sur des données RNAseq et de puces méthylation, que j'ai croisées avec des résultats d'analyses whole exome et whole genome pour interpréter les composantes. Je me suis également appuyée sur des données publiques relatives au cancer du foie ou au paysage épigénétique du foie normal.

4.2.1. Données du laboratoire

L'essentiel de mes analyses portent sur une série de tumeurs du laboratoire (LICA-FR), collectées en collaboration étroite avec plusieurs hôpitaux et analysées par une stratégie multi-omique impliquant le séquençage des ARNs (RNAseq), de l'ADN à l'échelle de l'exome (whole exome sequencing, WES) ou du génome entier (whole genome sequencing, WGS), et l'analyse du méthylome sur puce Illumina 450k.

J'ai ainsi pu analyser les données d'expression RNAseq de 148 tumeurs du foie et 32 échantillons de foie non tumoraux appariés. Les échantillons non tumoraux proviennent de foies normaux ou présentant divers types de cirrhoses (virales, alcooliques).

Les données de méthylation du laboratoire contiennent 239 CHC et 35 foies non tumoraux analysés sur puce Illumina Infinium HumanMethylation450k. Cette puce est basée sur la conversion au bisulfite de sodium (Bibikova, Barnes et al. 2011 ; Sandoval et al. 2011) qui va transformer les cytosines non méthylées en uracile tout en gardant intactes les cytosines méthylées, et mesure la méthylation sur environ 480 000 sites CpG qui couvrent environ 99 % des gènes et ARN non codants connus. À partir des intensités mesurées par les sondes, le niveau de méthylation (Beta-value) compris en 0 et 1 de chaque CpG est calculé. La Beta-value est égale à l'intensité du signal pour la forme méthylée sur l'intensité totale du signal.

A chaque échantillon analysé est associée une sélection pertinente d'annotations cliniques et moléculaires générées au laboratoire. Les annotations cliniques incluent plusieurs informations comme le sexe du patient, le type et le diagnostic de l'échantillon, le grade de la tumeur, la présence de fibrose ou de cirrhose du foie non tumoral et les principaux facteurs de risques liés aux carcinomes. Au niveau moléculaire, j'ai utilisé pour mon projet les résultats de précédentes analyses comme la détection de mutations, aberrations chromosomiques (gains et pertes de fragments chromosomiques) ou réarrangements de gènes *driver*.

4.2.2. Données publiques

En plus des données du laboratoire, j'ai utilisé différents jeux de données publiques afin de comparer et valider les résultats obtenus.

Pour la partie transcriptome, j'ai utilisé les données RNAseq de carcinomes hépatocellulaires et de foies normaux des consortium TCGA (The Cancer Genome Atlas, n= 383) et ICGC (Fujimoto et al., 2016) (International Cancer Genome Consortium, n = 410) (cf. Figure 26). La série TCGA contient des cas Américains et d'Asie du Sud Est et inclut diverses étiologies. La série ICGC a été générée au Japon, où l'étiologie dominante est l'infection par le virus de l'hépatite C.

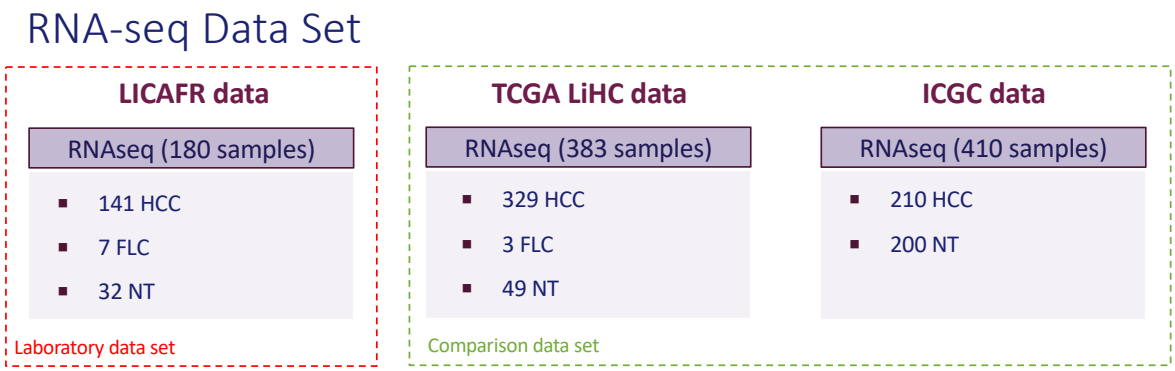


Figure 26 : ***Description des jeux de données RNAseq utilisés pour cette étude****. HCC : carcinome hepatocellulaire, FLC : carcinome fibrollamelaire, NT : foie non tumoral*

Pour la méthylation, j'ai également utilisé la série TCGA (275 CHC et 50 foies non tumoraux) ainsi que la série HEPTROMIC (224 CHC et 19 foies non tumoraux). Ces deux jeux de données indépendants ont été analysés avec la même technologie que les échantillons du laboratoire (puce Illumina Infinium HumanMethylation 450k) permettant de comparer et valider les résultats obtenus (cf. Figure 27).

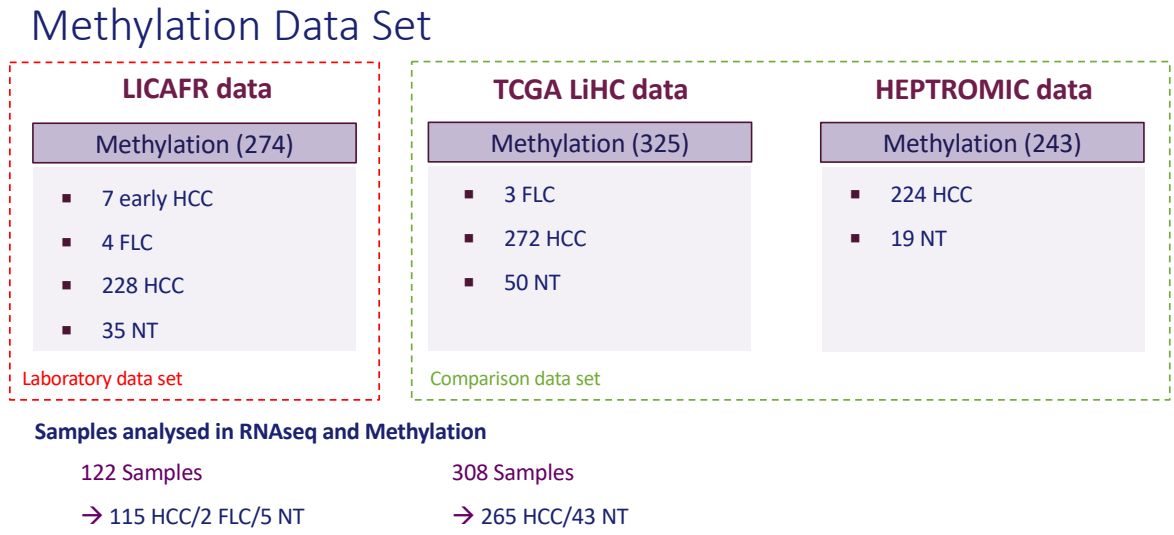


Figure 27 : ***Description des jeux de données de méthylation utilisés pour cette étude, et intersection avec les données RNAseq pour les jeux de données du laboratoire (LICAFR) et du consorcium TCGA****. HCC : carcinome hepatocellulaire, FLC : carcinome fibrollamelaire, NT : foie non tumoral*

Pour chaque série, les informations cliniques classiquement fournies pour les cancers du foie ont été extraites des publications. J'ai également récupéré un maximum d'annotations moléculaires incluant, pour les séries ICGC et TCGA, les profils génomiques obtenus par séquençage de l'exome ou du génome complet.

4.2.3. Annotations épigénétiques et fonctionnelles

Pour interpréter les signatures de méthylation obtenues, j'ai utilisé différentes annotations provenant d'études épigénomiques intégrées afin de mieux comprendre le contexte épigénétique dans lequel les changements de méthylation interviennent. Ont été récupérés :

- les domaines de méthylation du foie normal (HMD/PMD/LMR/UMR) (Salhab et al., 2018) ;
- les données Chip-seq du marquage H3K27ac ainsi que les 18 domaines chromatiniens déterminés sur le foie normal par le consortium Roadmap (Roadmap Epigenomics Consortium et al., 2015) ;
- les données ChiP-seq de fixation de facteur de transcription et les données de timing de réplication déterminées sur la lignée cellulaire de carcinome hépatocellulaire HEPG2 par le consortium ENCODE (Davis et al., 2018; Dunham et al., 2012).

L'objectif principal de ma thèse est de mieux comprendre les relations de cause à effet entre les différents niveaux de régulation dans le cancer du foie, en caractérisant précisément les effets d'une altération moléculaire spécifique et en utilisant des méthodes statistiques capables d'extraire des signatures d'expression des gènes ou de méthylation de l'ADN liées à des mécanismes oncogéniques précis. En croisant ces résultats avec la diversité des altérations génomiques connues dans le cancer du foie, nous pourrons mieux comprendre les mécanismes moléculaires impliqués dans la carcinogenèse hépatique, point de départ au développement de thérapies innovantes.

Résultats

Le profil transcriptionnel et épigénétique de chaque tumeur reflète l'interaction de multiples facteurs, à la fois cliniques (facteurs de risque, cirrhose) moléculaires (mutations, aberrations chromosomiques, dérégulation épigénétique) et phénotypiques (prolifération, différenciation, infiltrat inflammatoire). Ce chapitre présente les résultats des principales études réalisées durant ma thèse visant à mieux comprendre les dérégulations observées dans les tumeurs hétérogènes et fortement remaniées que sont les carcinomes hépatocellulaires. Ces études ont porté sur le niveau de l'expression des gènes et de la méthylation des CpG.

1. Signatures transcriptionnelles des carcinomes hépatocellulaires

1.1. Développement d'un outil pour la caractérisation des groupes et signatures transcriptionnelles existantes

Comme on a pu le voir dans l'introduction, plusieurs équipes ont créé des classifications et scores moléculaires basés sur l'expression des gènes dans les CHC (Boyault et al., 2007; Chiang et al., 2008; Hoshida et al., 2009; Kaposi-Novak et al., 2006; Yamashita et al., 2008). Au cours de ma première année de thèse, j'ai participé au développement de MS.liverK, un prologiciel sous forme de package R qui permet de reproduire et comparer les classifications et scores publiés. MS.liverK permet de déterminer facilement les sous-groupes de CHC et les scores moléculaires de chaque tumeur, aussi bien à partir de données de puces que de séquençage RNAseq. Cet outil a été mis à disposition de la communauté scientifique sur Github, et il est décrit dans un article que nous avons déposé sur bioRxiv. J'ai utilisé MS.liverK dans la suite de ma thèse pour prédire les groupes moléculaires de CHC dans les différentes séries RNAseq que j'ai analysées.

Dans cet article, nous avons également appliqué MS.liverK à la série TCGA afin de comparer les différentes classifications publiées (Fig. 1). Celles-ci se recoupent fortement entre elles ainsi qu'avec les scores fonctionnels. Le groupe A de Lee rassemble l'essentiel les groupes S1-S2 de Hoshida et les groupes G1-G3 de notre laboratoire. Le groupe B de Lee correspond au

groupe S3 de Hoshida, qui englobe dans notre étude le groupe G4 et les groupes G5-G6 associés à l'activation de la voie Wnt ß-caténine. Le groupe « polysomie chromosome 7 » de Chiang correspond à un sous-groupe de notre groupe G4.

Ainsi, chaque classification semble capturer certaines variations biologiques non capturées par les autres. De plus, les frontières entre groupes sont parfois poreuses et certaines tumeurs présentent des caractéristiques pouvant appartenir à plusieurs groupes. Ces observations soulignent les limites des approches classiques de classification non supervisée, qui permettent d'identifier les principaux groupes de tumeurs mais pas de caractériser en détail les différentes sources de variation du transcriptome. Ces limites justifient le développement de méthodes statistiques innovantes, comme l'analyse en composantes indépendantes, pour identifier les gènes co-régulés impliqués dans des processus biologiques précis.

Article 1 : " MS.liverK: an R package for transcriptome-based computation of molecular subtypes and functional signatures in liver cancer" (Available on bioRxiv)

Florent Petitprez, **Léa Meunier**, Eric Letouzé, Yujin Hoshida, Augusto Villanueva, Josep Llovet, Snorri Thorgeirsson, Xin Wei Wang, Wolf H Fridman, Jessica Zucman-Rossi, Aurélien de Reyniès

MS.liverK: an R package for transcriptome-based computation of molecular subtypes and functional signatures in liver cancer

Florent Petitprez[1,2,3,4], Léa Meunier[5], Eric Letouzé[5], Yujin Hoshida[6], Augusto Villanueva[6,7], Josep Llovet[6,8], Snorri Thorgeirsson[9], Xin Wei Wang[9], Wolf H Fridman[2,3,4], Jessica Zucman-Rossi[5,10] and Aurélien de Reyniès[1,*]

[1]Programme Cartes d'Identité des Tumeurs (CIT), Ligue Nationale Contre le Cancer, Paris, France

[2]INSERM UMR_S 1138, Cordeliers Research Centre, Paris, France

[3]Paris Descartes University, Sorbonne Paris Cité, Paris, France

[4]Pierre and Marie Curie University, Paris, France

[5]INSERM UMR-1162, Paris, France; [6]Liver Cancer Program, Division of Liver Diseases, Department of Medicine, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[7]Division of Haematology and Medical Oncology, Department of Medicine, Tisch Cancer Institute, Icahn School of Medicine at Mount Sinai, New York, NY, USA

[8]BCLC, Liver Unit, IDIBAPS, CIBERehd, Hospital Clínic, Universitat de Barcelona, ICREA, Barcelona, Catalonia, Spain

[9]Laboratory of Human Carcinogenesis, CCR, NCI, NIH, Bethesda, MD, USA

[10]AP-HP; HEGP, Department of Oncology, Paris, France.

*To whom correspondence should be addressed: reyniesa@ligue-cancer.net.

Abstract

Abstract

Summary: Liver cancer is a highly heterogeneous disease in terms of etiology, tissue and cellular morphology, tumor molecular characteristics, microenvironment composition and prognosis. Several studies, based on tumor gene-expression profiling (GEP) data, have dissected the molecular heterogeneity of liver cancer. They resulted in various tools, either delineating homogeneous tumor subtypes or calculating molecular scores of prognostic or biological functions. Here, we present MS.liverK, an easy-to-use R package providing a comprehensive implementation of these tools, for research use.

Availability and implementation: The *MS.liverK* R package is available from GitHub (https://github.com/cit-bioinfo/MS.liverK).

1. Introduction

Primary liver cancer is the third most deadly cancer worldwide, with Hepatocellular carcinomas (HCC) accounting for 85%-90% of the cases. The molecular heterogeneity of liver cancers has motivated numerous transcriptome-based studies, yielding gene-signatures and algorithms for molecular subtyping, prognostic prediction and functional scores computation. The use of these tools is of great interest for researchers dedicated to liver cancer study, but is very limited due to their dispersion.

Here, we introduce MS.liverK (Molecular Signatures in Liver Cancer), an R package reimplementing these tools. MS.liverK takes as input a (log2 scale) transcriptome matrix of liver cancer samples, either microarray- or RNA-Seq-derived. It also includes graphical functions, allowing to easily visualize the outputs.

2. MS.liverK

After a careful review of the literature, we identified 5 GEP-based molecular subtyping systems of HCCs published in the last 15 years (Lee *et al.*, 2004; Boyault *et al.*, 2007; Yamashita *et al.*, 2008; Chiang *et al.*, 2008; Hoshida *et al.*, 2009). Lee classification (Lee *et al.*, 2004) was established on a belgo-chinese cohort of 91 HCC samples; it is made of 2 classes (A/B), related to proliferation and survival. Lee subtypes have been further refined in 4 groups using Alpha-fetoprotein (AFP) marker, improving association to survival. Boyault classification (Boyault *et al.*, 2007) was established on a French cohort of 57 HCCs including patients from various geographic origins; it is made of 6 classes (G1 to G6) related to TP53 and CTNNB1 mutations, proliferation, HBV infection and prognosis. Yamashita classification (Yamashita *et al.*, 2008) is based on two markers, EPCAM and AFP, and defines four groups related to differentiation and prognosis. Chiang classification (Chiang *et al.*, 2008) was built using 91 HCV infected HCCs; it identifies 5 classes related to CTNNB1 mutation, proliferation, inflammation and polysomy of the chromosome 7. Hoshida classification (Hoshida *et al.*, 2009) is based on the meta-analysis of 9 cohorts totaling 603 HCC samples; it contains 3 classes (S1/S2/S3) related to proliferation, differentiation and survival. MS.liverK implements all of the above classification systems from transcriptomic data; it also includes a conversion function to allow using different GEP platforms (Fig. S1A).

To ensure classifiers were adequately assigning classes to samples, we ran them on the data from the samples used by the four teams to define their subgroups. We then compared MS.liverK classes to the ones they originally established and found a very good correspondence (Fig. S1B): Chi-squared test p-values for Lee, Boyault, Chiang and Hoshida and Roessler classifications were all < 2.2 e-16, with accuracy rate of respectively 94.5%, 98.2%, 100%, 93.8% and 90.0%.

MS.liverK implements the prognostic prediction algorithm published by (Nault *et al.*, 2013), which is so far the most extensively validated prognostic score for hepatocellular carcinomas (HCC).

Several teams have published gene-signatures related to biological functions/pathways involved in liver cancer oncogenesis. These gene-signatures relate to TGFB1 signalling (Coulouarn *et al.*, 2008), MET pathway (Kaposi-Novak *et al.*, 2006), stemness (Oishi *et al.*, 2012; Yamashita *et al.*, 2008), EPCAM (Yamashita *et al.*, 2008) and hypoxia (van Malenstein *et al.*, 2010). MS.liverK computes all the corresponding scores.

We applied MS.liverK to the TCGA LIHC RNA-seq dataset (n=286). The graphical output shows a striking concordance across the various subtyping systems, the prognostic score and the functional scores (Figure 1). This observation is expected, given that molecular subtypes should represent homogeneous types of tumors with specific biological and prognostic characteristics.

3. Software implementation and use

The package has been entirely written using R language. It includes pre-packaged data from Gene Expression Omnibus (GEO), accession number GSE20238 (Mínguez *et al.*, 2011), to be used as example data. A vignette describing the use of the package is available on GitHub (https://github.com/FPetitprez/MS.liverK/blob/master/vignettes/vignette.pdf). On a standard computer, the computation of all subtypes and scores takes a couple of minutes.

References

Boyault,S. *et al.* (2007) Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. *Hepatology*, **45**, 42–52.

Chiang,D.Y. *et al.* (2008) Focal gains of VEGFA and molecular classification of hepatocellular carcinoma. *Cancer Res.*, **68**, 6779–6788.

Coulouarn,C. *et al.* (2008) Transforming growth factor-beta gene expression signature in mouse hepatocytes predicts clinical outcome in human cancer. *Hepatology*, **47**, 2059–2067.

Hoshida,Y. *et al.* (2009) Integrative transcriptome analysis reveals common molecular subclasses of human hepatocellular carcinoma. *Cancer Res.*, **69**, 7385–7392.

Kaposi-Novak,P. *et al.* (2006) Met-regulated expression signature defines a subset of human hepatocellular carcinomas with poor prognosis and aggressive phenotype. *J. Clin. Invest.*, **116**, 1582–1595.

Lee,J.-S. *et al.* (2004) Classification and prediction of survival in hepatocellular carcinoma by gene expression profiling. *Hepatology*, **40**, 667–676.

van Malenstein,H. *et al.* (2010) A seven-gene set associated with chronic hypoxia of prognostic importance in hepatocellular carcinoma. *Clin. Cancer Res.*, **16**, 4278–4288.

Mínguez,B. *et al.* (2011) Gene-expression signature of vascular invasion in hepatocellular carcinoma. *J. Hepatol.*, **55**, 1325–1331.

Nault,J.-C. *et al.* (2013) A hepatocellular carcinoma 5-gene score associated with survival of patients after liver resection. *Gastroenterology*, **145**, 176–187.

Oishi,N. *et al.* (2012) Transcriptomic profiling reveals hepatic stem-like gene signatures and interplay of miR-200c and epithelial-mesenchymal transition in intrahepatic cholangiocarcinoma. *Hepatology*, **56**, 1792–1803.

Yamashita,T. *et al.* (2008) EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. *Cancer Res.*, **68**, 1451–1461.

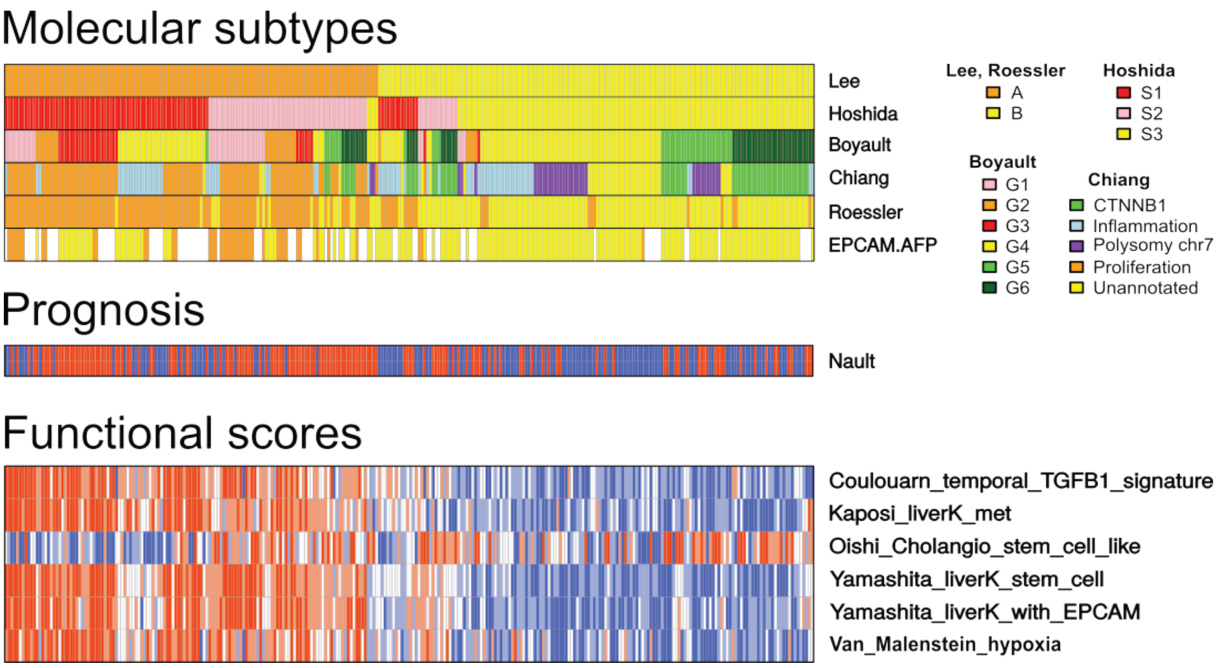


Fig. 1. MS.liverK graphical output. Graphical output of MS.liverK illustrated on the TCGA LIHC transcriptome dataset. Columns represent samples, rows represent molecular subtypes (upper panel), prognostic scores (middle panel) and functional scores (lower panel).

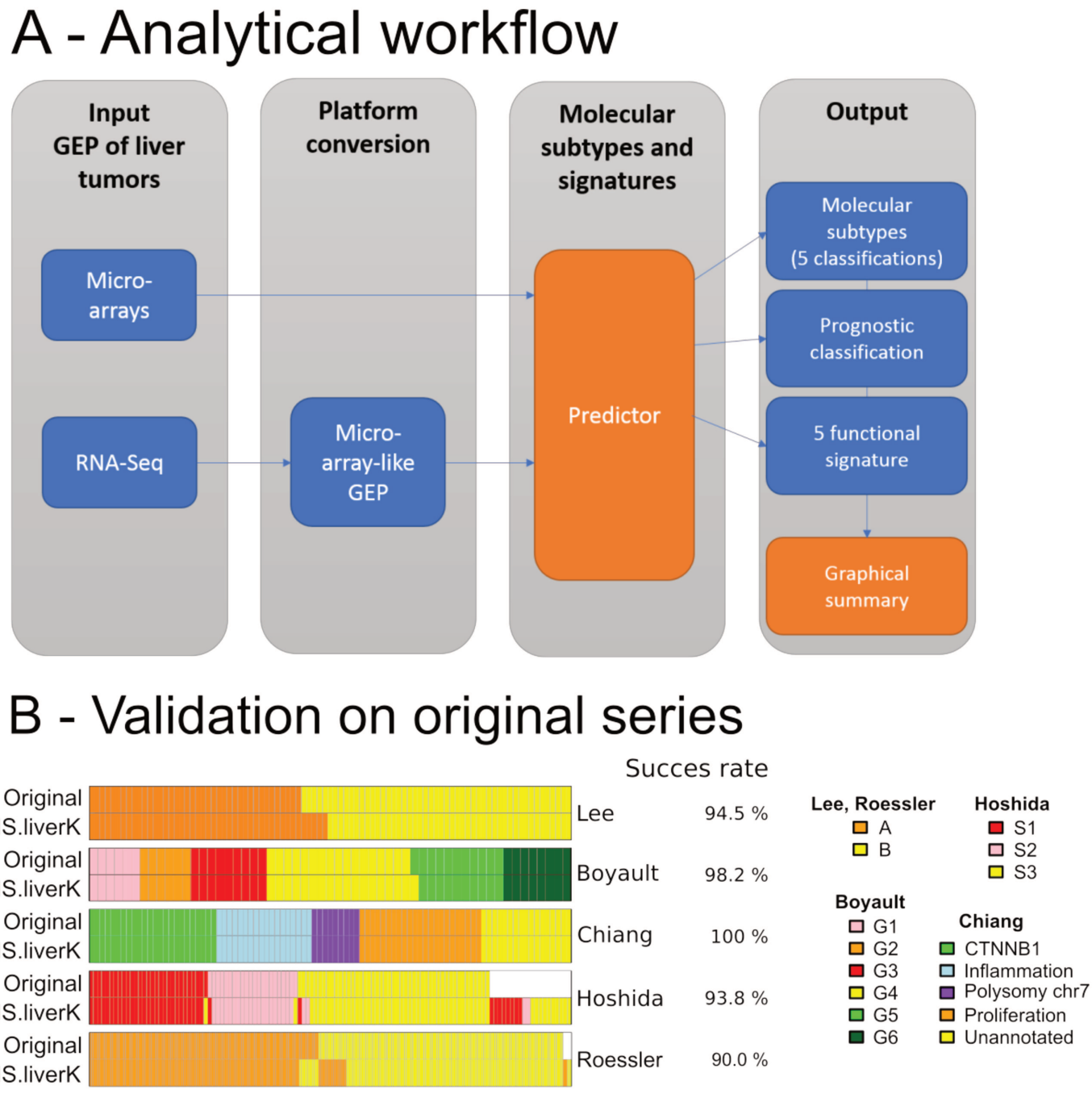


Fig. S1: A. Analytical workflow: (i) MS.liverK takes liver tumors GEP as input -several GEP platforms are supported-, (ii) when needed GEP are converted to microarray-like GEP, (iii) Molecular subtypes, prognostic and functional scores are calculated (iv) textual and graphical outputs are provided. B. Application of MS.liverK to TCGA LIHC cohort and performance of MS.liverK molecular subtyping features with respect to the original data of each classification system.

1.2. Analyse en composantes indépendantes des CHC

Afin d'identifier les gènes co-régulés dans le foie normal et tumoral, et de mieux comprendre la diversité des processus biologiques perturbant le transcriptome des carcinomes hépatocellulaires, j'ai appliqué l'analyse en composantes indépendantes (ACI) aux données RNAseq de tumeurs hépatiques (jeux de données LICA-FR, TCGA-LIHC et ICGC). Les résultats obtenus ont été interprétés en prenant en compte les données additionnelles disponibles pour chaque échantillon comme les données cliniques, anatomo-pathologiques et moléculaires (mutations, aberrations chromosomiques).

1.2.1. Pipeline d'analyse de données RNAseq

Pour obtenir l'expression des gènes à partir des données de séquençage RNAseq, j'ai d'abord mis en place un pipeline d'analyse bioinformatique partant des séquences brutes pour aller jusqu'aux données d'expression normalisées. Les fichiers bruts au format .fastq, contenant les reads avec leurs scores de qualité associés, sont alignés avec le logiciel Tophat2 (Kim et al., 2013). Le génome de référence utilisé pour les alignements est le génome humain hg19 publié en 2007 par Genome Reference Consortium (GRC). Les résultats de l'alignement au format .bam sont filtrés avec le logiciel SamTools (Li et al., 2009) et les reads qui s'alignent sur plusieurs régions du génome sont supprimés. L'outil htseq-count développé à partir du package python HTSeq (Anders et al., 2015) permet alors d'obtenir la matrice de comptage indiquant le nombre de reads chevauchant chaque gène référencé dans la base GENCODE (Harrow et al., 2012), qui regroupe à la fois les gènes codants, mais aussi les longs ARNs non codants (cf. Figure 28).

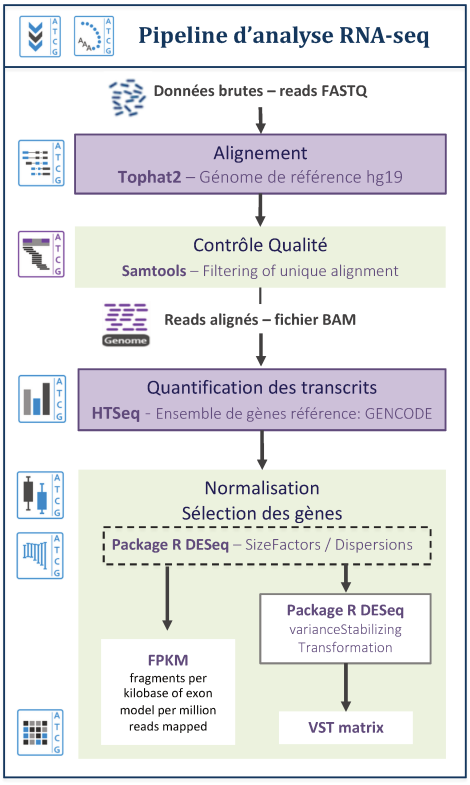


Figure 28 : ***Structure du pipeline d'analyse mis en place au laboratoire pour l'analyse des données RNAseq.***

Deux matrices d'expression normalisées sont alors générées avec le package DESeq2 (Love et al., 2014) :

- la matrice FPKM (Fragments per Kilobase of exon per Million reads) normalisée par rapport à la taille des gènes et au nombre de reads séquencés pour chaque librairie.
- la matrice VST, normalisée par rapport à la taille de la librarie et la longueur codante des gènes, mais aussi en utilisant la fonction '*variance stabilisation transformation*'. Cette méthode estime la variabilité des données grâce à un modèle d'erreur basé sur la distribution binomiale négative (avec une variance et une moyenne liées par régression locale) qui permet de normaliser les données RNAseq.

La même normalisation de l'expression des gènes a été appliquée aux données TCGA et ICGC, pour obtenir la matrice FPKM ainsi que la matrice normalisée avec la fonction variance stabilisation transformation du package DESeq2 (Love et al., 2014). Ce pipeline est maintenant utilisé en routine au laboratoire pour toutes les nouvelles données RNAseq reçues. Pour ce travail, j'ai utilisé le génome de référence hg19, mais d'autres génomes de référence, comme hg38 pour l'homme ou m21 pour la souris, peuvent également être utilisés.

1.2.2. Paramétrage de l'ACI

Plusieurs paramètres sont à choisir pour l'analyse en composantes indépendantes : le nombre de gènes utilisés, le nombre de composantes à extraire et l'algorithme de calcul. Pour ce travail, j'ai sélectionné les 10 000 gènes les plus variants en fonction de leur écart-type, et utilisé l'algorithme de déconvolution FastICA (Hyvärinen, 1999) implémenté dans le package R *fastICA* http://www2.uaem.mx/r-mirror/web/packages/fastICA/index.html. J'ai basé le choix du nombre de composantes sur la stabilité des résultats obtenus. Le but étant d'isoler un maximum de composantes indépendantes stables au cours des différentes itérations (cf. Figure 28).

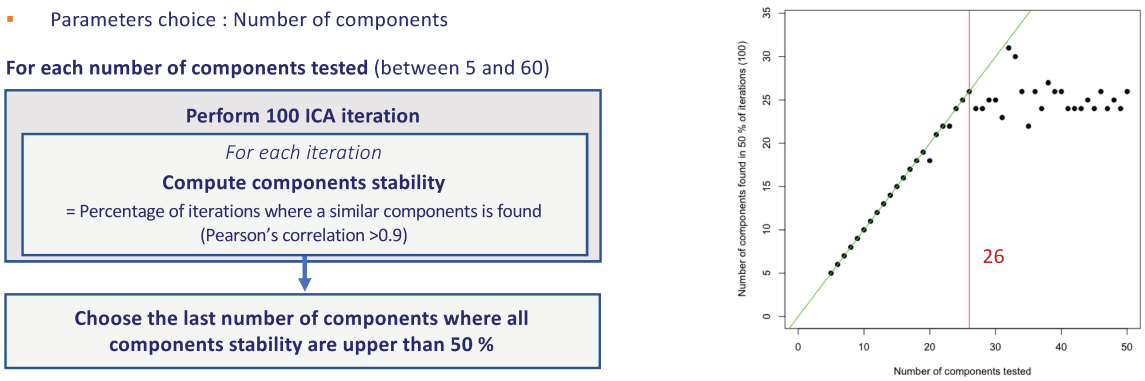


Figure 29 : ***Méthode utilisée pour le choix du nombre de composantes dans l'ACI.***

Pour chaque nombre de composantes entre 5 et 60, j'ai effectué 100 itérations de l'ACI. L'algorithme FastICA étant basé sur une initialisation aléatoire, cela donne 100 résultats différents, qui sont ensuite comparés entre eux avec un calcul de corrélation de Pearson au niveau de la contribution des gènes. La mesure de stabilité d'une composante est le pourcentage d'itérations dans lesquelles une composante similaire (corrélation de Pearson > 0.9) est identifiée. On constate que, pour la cohorte LICA-FR, le nombre de composantes stables plafonne à 26 composantes. Au-delà, les composantes supplémentaires sont instables. J'ai donc choisi ce nombre de 26 composantes, qui correspond au nombre maximal pour lequel toutes les composantes ont une stabilité supérieure à 50 %. La même approche m'a conduit à sélectionner 35 composantes pour la série TCGA-LiHC et 29 pour la série ICGC.

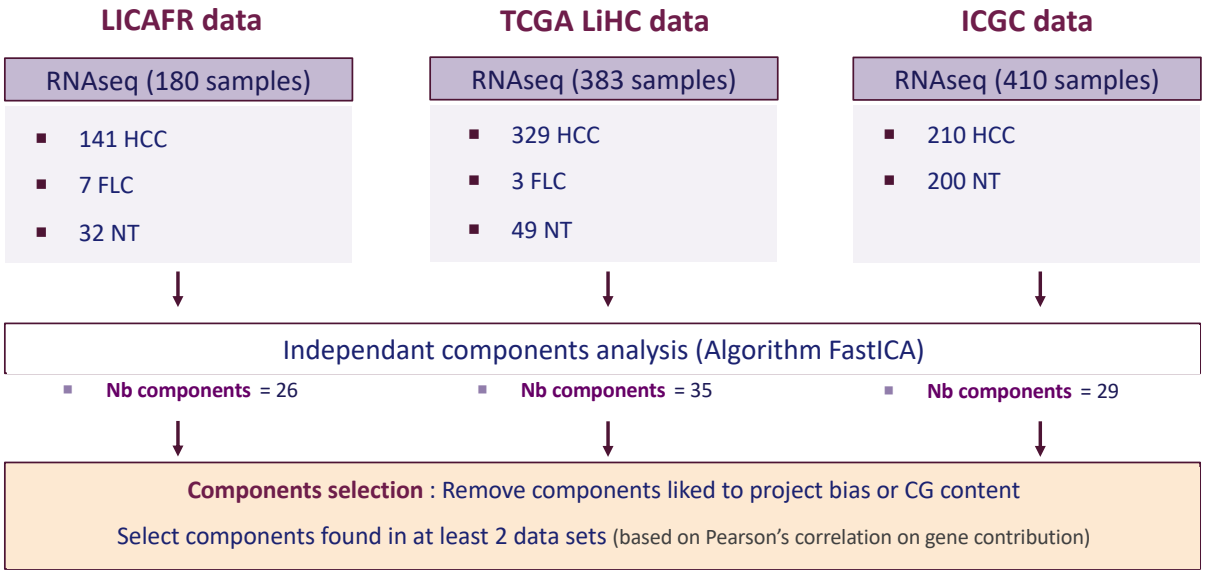


Figure 30 : ***Schéma récapitulatif des jeux de données analysés et des paramètres utilisés pour extraire les compsantes.*** *Chaque jeu de données a été analysé séparément avec un nombre de composantes choisi comme décrit ci-dessus, et la reproductibilité des résultats a été étudiée ensuite. HCC = carcinome hepatocellulaire ; FLC = carcinome fibrolamellaire ; NT = tissu non tumoral*

J'ai ensuite comparé les résultats obtenus pour les 3 jeux de données. La similarité de deux composantes provenant de différents jeux de données est déterminée en calculant, à partir de la contribution de leurs gènes communs, la valeur absolue du coefficient de corrélation de Pearson. Une composante est gardée si elle est retrouvée dans la décomposition d'au moins 2 des 3 jeux de données de CHC (corrélation de Pearson > 0.35). Cette étape permet notamment de supprimer les composantes liées à des biais de projets. Dans les données LICA-FR, passées sur 8 projets de séquençage différents, on retrouve en effet une composante liée

à la taille des reads et une composante liée à un des projets qui a été séquencé avec une librairie « non-stranded », contrairement aux autres (cf. Figure 31).

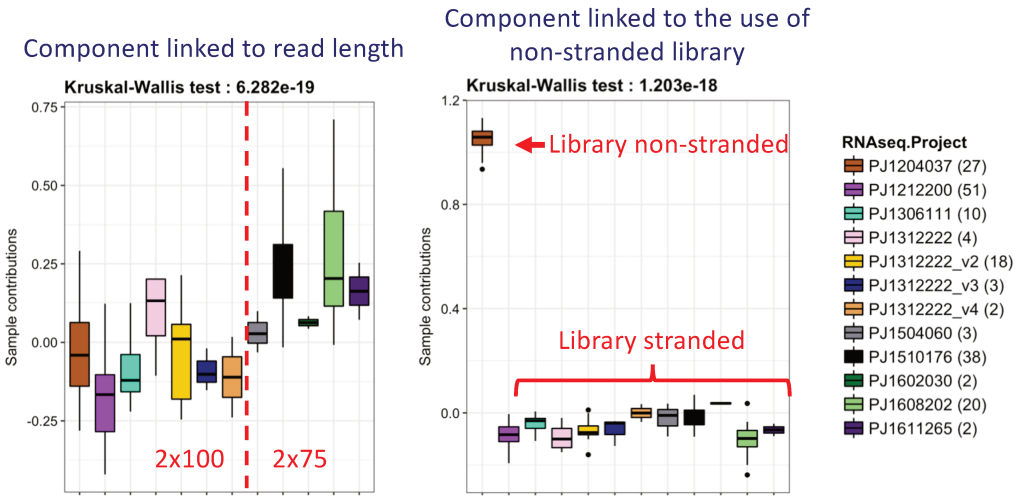


Figure 31 : ***Représentation des associations des composantes liées à des biais de projet.***

Une composante corrélée au contenu en GC des gènes est également identifiée dans les 3 jeux de données. Elle n'est pas gardée pour la suite de l'interprétation.

1.2.3. Interprétation des composantes

Des scripts R spécifiques à l'analyse des résultats de l'ACI ont été utilisés. Ils proviennent du package *MineICA* https://www.bioconductor.org/packages/release/bioc/html/MineICA.html (Biton et al.) ou ont été développés durant ma thèse afin de faciliter la visualisation et l'interprétation des composantes. L'interprétation biologique des composantes est basée sur l'analyse croisée de la contribution des gènes et des échantillons. L'interprétation des composantes au niveau des gènes repose sur une analyse d'enrichissement par rang (Subramanian et al., 2005), implémentée dans le package R *GSEAPreranked* (Gene Set Enrichment Analysis) https://gsea-msigdb.github.io/gseapreranked-gpmodule/v6/index.html, que j'ai modifié pour pouvoir l'appliquer, non pas à une matrice d'expression des gènes, mais directement à un vecteur indiquant la contribution des gènes dans une composante. Cette méthode permet d'identifier des groupes de gènes de la base de données de signatures moléculaires (Molecular Signatures Database – v5.1 MSigDB) (Liberzon et al., 2011, 2015) significativement enrichis parmi les gènes les plus contributeurs d'une composante. MSigDB regroupe à la fois les groupes de gènes des bases de données GO, KEGG et HALLMARK, mais aussi ceux identifiés dans des publications antérieures.

Les échantillons et gènes fortement contributeurs permettent d'interpréter les composantes extraites (cf. Figure 32). Par exemple, la composante 13 est fortement associée à la mutation du gène CTNNB1, codant pour la β-caténine (p-value = 6.14 10-16) ainsi qu'aux échantillons des groupes G5 et G6. De manière cohérente, l'analyse des gènes fortement contributeurs à cette composante révèle un enrichissement en gènes cibles de la voie Wnt/ß-caténine.

GS	SIZE	NOM-pval	FDR-pval	FWER-pval
CHIANG_LIVER_CANCER_SUBCLASS_CTNNB1_UP	164	0	0	0
BOYAULT_LIVER_CANCER_SUBCLASS_G6_UP	49	0	0	0
CHIANG_LIVER_CANCER_SUBCLASS_INTERFERON_DN	44	0	0	0
SABATES_COLORECTAL_ADENOMA_UP	105	0	0.0008066	0.004
CAIRO_HEPATOBLASTOMA_UP	114	0	0.0006453	0.004
SANSOM_WNT_PATHWAY_REQUIRE_MYC	44	0	0.005425	0.04
HALLMARK_WNT_BETA_CATENIN_SIGNALING	26	0	0.0097184	0.08
AMINE_RECEPTOR_ACTIVITY	19	0	0.01658	0.148
V$CDX2_Q5	123	0	0.050716	0.424
NIKOLSKY_BREAST_CANCER_7P15_AMPLICON	11	0	0.047265	0.436

Figure 32 : Représentation des associations au niveau des annotations cliniques et moléculaires et des annotations des gènes de la composante 13, liée à la voie Wnt/B-catenine.

De plus, une étude du laboratoire a montré que les différents types de mutations de la β-caténine entraînent une activation plus ou moins importante de la voie Wnt (Rebouissou Sandra et al., 2016). De manière notable, l'activité de notre composante 13 dans les CHC est

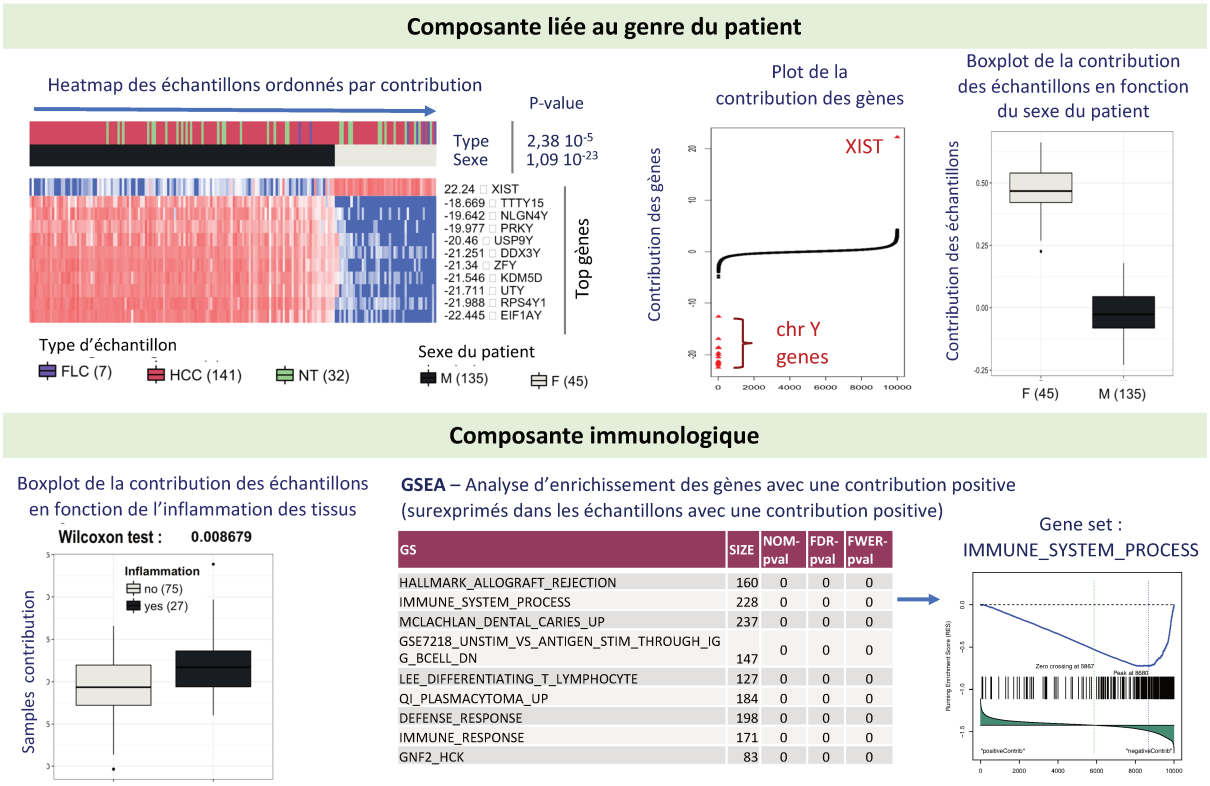
fortement corrélée aux effets connus des différentes mutations de CTNNB1 sur l'activation de la voie Wnt. Elle est également corrélée linéairement à la quantité de glutamine synthase (marqueur d'activation de la voie Wnt) mesurée en RPPA (reverse phase protein array). Toutes ces observations soutiennent l'idée que la composante 13 reflète précisément le niveau d'activation de la voie Wnt/B-catenine dans les tumeurs.

Un travail d'interprétation similaire a été fait pour toutes les composantes retrouvées dans au moins 2 jeux de données (cf.Tableau 1). Les composantes obtenues ont ainsi été attribuées au type d'échantillon, à des sous-groupes de tumeurs ou à des caractéristiques moléculaires comme le gain ou la perte de bras chromosomiques, la présence de certaines mutations et l'activation de voies de signalisation spécifiques.

Association	LiC1162	TCGA	ICGC	n°	
Patient gender	X	X	X	1	Patient characteristic
NT vs T	X	X	X	2	
Immunologic	X	X	X	3	
Immunoglogulin genes	X		X	4	
EMT	X	X	X	5	Biological process
Proliferation	X	X	X	6	
Hepatic differentiation	X	X	X	7	
Fibrollamelar carcinoma	X	X	X	8	HCC groups
Group Chiang polysomy 7	X	X	X	9	
Groupe Chiang proliferation		X	X	10	
AXIN1/RPS6KA3	X	X	X	11	Mutation – Pathway activation
Oxydatif stress	X	X	X	12	
Wnt activation	X	X	X	13	
IL6/JAK/STAT		X	X	14	
NF-kB		X	X	15	
8q gain	X	X	X	16	Chromosomic aberation
1q gain	X	X	X	17	
Methylation	X	X	X	18	
Mitochondrie	X	X	X	19	
Interferon gene		X	X	20	

Tableau 1 : ***Tableau récapitulatif des principales interprétations des 20 composantes étudiées et leur récurrence dans les différents jeux de données.***

On retrouve également une composante liée au sexe des patients (cf. Figure 33), avec une forte contribution des gènes du chromosome Y, et des gènes XIST et TSIX d'inactivation du chromosome X, ainsi qu'une composante liée à la présence d'infiltrats inflammatoires dans les tissus hépatiques (cf. Figure 33), montrant un enrichissement des gènes du système immunitaire et une corrélation au statut inflammatoire des tumeurs déterminé en histologie.



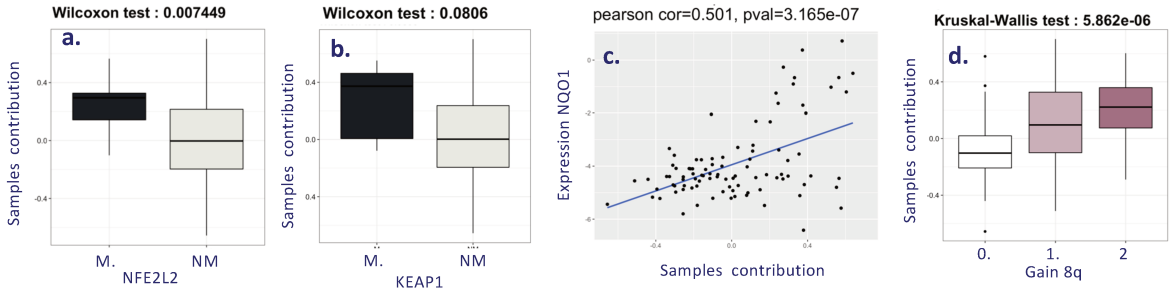
GS	SIZE	NOM-pval	FDR-pval	FWER-pval
HALLMARK_ALLOGRAFT_REJECTION	160	0	0	0
IMMUNE_SYSTEM_PROCESS	228	0	0	0
MCLACHLAN_DENTAL_CARIES_UP	237	0	0	0
GSE7218_UNSTIM_VS_ANTIGEN_STIM_THROUGH_IG G_BCELL_DN	147	0	0	0
LEE_DIFFERENTIATING_T_LYMPHOCYTE	127	0	0	0
QI_PLASMACYTOMA_UP	184	0	0	0
DEFENSE_RESPONSE	198	0	0	0
IMMUNE_RESPONSE	171	0	0	0
GNF2_HCK	83	0	0	0

Figure 33 : ***Représentation des associations au niveau des annotations cliniques et moléculaires et des annotations des gènes de la composante 1 liée au genre du patient et de la composante immunologique 3.***

Outre ces processus biologiques connus, l'ICA m'a permis de mettre en évidence de nouvelles associations entre altérations moléculaires et dérégulations du transcriptome dans les CHC. Une composante particulièrement intéressante est la composante 12 liée à la voie du stress oxydatif. Cette composante montre une association à la mutation activatrice du gène NFE2L2, qui permet d'induire la réponse au stress oxydatif, ainsi qu'à la mutation inactivatrice de KEAP1, un inhibiteur de NFE2L2 (cf. Figure 34) (Bryan et al., 2013). L'activité de la composante dans les échantillons est fortement corrélée à l'expression du gène et la protéine NQO1 (NAD(P)H quinone dehydrogenase 1), un antioxydant activé par NFE2L2. L'analyse des gènes contributeurs montre également un enrichissement en gènes liés à NFE2L2 et au Hallmark « Reactive oxygen species pathway » (cf. Figure 34).

Composante liée au stress oxydatif

Graphe de la contribution des échantillons en fonction des mutations (a) NFE2L2 et (b) KEAP1, de (c) l'expression protéique (RPPA) de NQO1 dans les échantillons et (d) du gain du bras chromosomique 8q



GSEA – Analyse d'enrichissement des gènes avec une contribution positive (surexprimés dans les échantillons avec une contribution positive)

GS	SIZE	NOM-pval	FDR-pval	FWER-pval
NFE2L2.V2	305	0	0	0
REACTOME_PHASE_II_CONJUGATION	53	0	0	0
chr17p11	38	0	0	0
REACTOME_GLUCURONIDATION	14	0	0	0
KEGG_PENTOSE_AND_GLUCURONATE_INTERCONVERSIONS	21	0	0	0
KEGG_ASCORBATE_AND_ALDARATE_METABOLISM	20	0	0.0047665	0.028
KEGG_PORPHYRIN_AND_CHLOROPHYL_METABOLISM	24	0	0.0046648	0.032
HALLMARK_REACTIVE_OXIGEN_SPECIES_PATHWAY	34	0	0.0086526	0.068

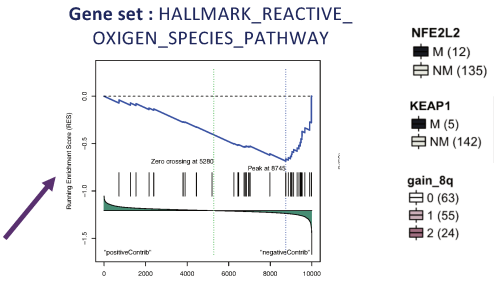


Figure 34 : ***Représentation des associations au niveau des annotations cliniques et moléculaires et des annotations des gènes de la composante 12 liée au stress oxidatif.***

Outre l'association attendue avec les altérations touchant les gènes de la voie du stress oxydatif, la composante 12 est fortement associée au gain du bras chromosomique 8q (P = 5.9×10-6), fréquemment gagné dans les CHC (60 % des tumeurs). Parmi les gènes du bras 8q, l'oncogène MYC (8q24) est l'un des plus fortement contributeurs à la composante 12. Or, il a été montré que la protéine c-Myc, en plus d'être impliquée dans la régulation du cycle cellulaire, la différenciation et l'apoptose, joue un rôle dans l'équilibre d'oxydoréduction cellulaire. En effet, c-Myc active la transcription des sous unités de la gamma-glutamyl-cystéine synthétase (gamma-GCS), l'enzyme limitante dans la biosynthèse du glutathion (GSH). Une étude a montré que le gain d'une seule copie supplémentaire du gène MYC et des gènes voisins incluant PVT1, CCDC26 et GSDMC favorise le développent tumoral. En effet, le gain de l'expression de l'ARN non codant PVT1 bloque la phosphorylation du résidu T58 de c-Myc nécessaire à sa dégradation (Tseng et al., 2014). Notre hypothèse est que le gain à la fois de MYC et PVT1, comme c'est le cas quand le bras 8q complet est gagné, entraînerait une surexpression de c-Myc qui, via son rôle dans la synthèse du glutathion, participerait à l'activation de la voie du stress oxydatif et contribuerait à la résistance des cellules tumorales (cf. Figure 35).

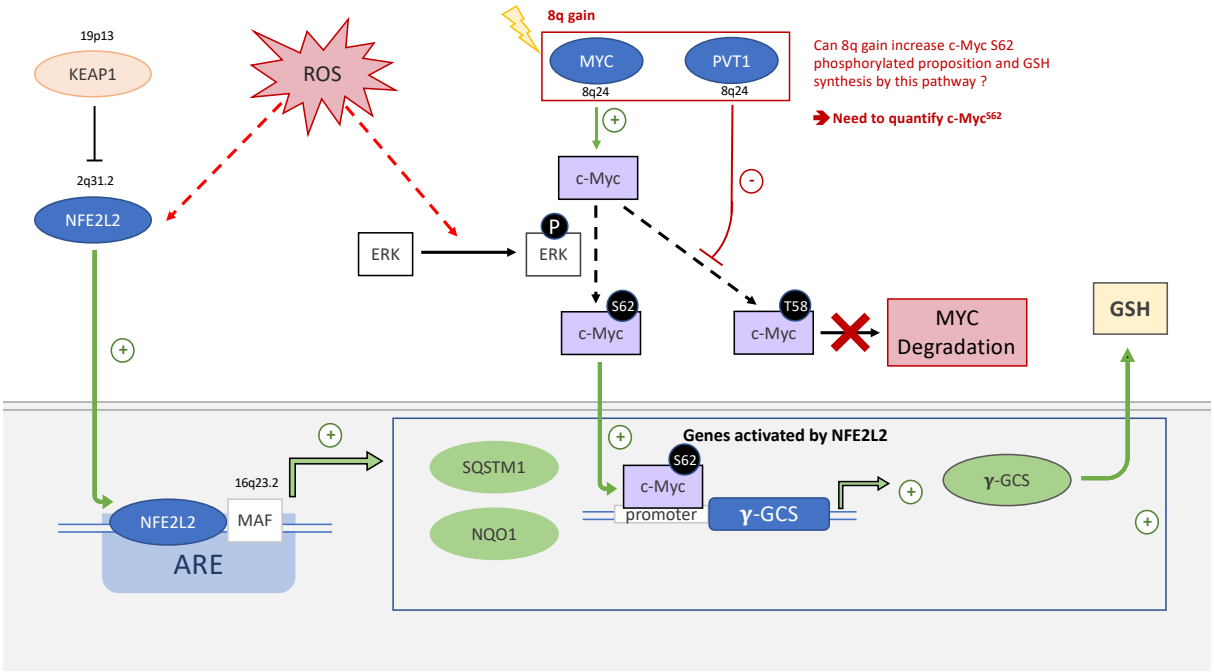


Figure 35 : ***Modèle proposé de l'influence du gain du 8q, et particulièrement de c-Myc et PVT1 sur la voie du stress oxydatif par l'activation de GSH.***

La diversité des associations mises en évidence dans cette étude montre bien que le transcriptome est influencé par un large panel de facteurs biologiques d'origines multiples. La séparation et la caractérisation de ces derniers sont des points essentiels pour la compréhension des effets des multiples altérations aux différents niveaux de régulation du génome survenant dans le développement de tumeurs.

1.2.4. Conclusion et perspectives

J'ai appliqué avec succès l'analyse en composantes indépendantes à 3 jeux de données RNAseq de cancers du foie, ce qui m'a permis d'identifier 20 composantes retrouvées dans au moins 2 des 3 jeux de données étudiés. Les caractéristiques des échantillons et gènes les plus contributeurs m'ont permis d'identifier la signification biologique de la plupart des composantes. Les différentes composantes ont pu être associées à des facteurs techniques, des processus biologiques, des sous-types de CHC ou des caractéristiques moléculaires comme la présence de mutations d'un certain gène. L'ACI a permis de séparer clairement des sources de variations difficilement distinguables avec les méthodes classiquement utilisées (clustering hiérarchique et analyse en composantes principales), comme l'activation de la voie de réponse au stress oxydatif.

La comparaison des composantes obtenues avec les classifications des carcinomes hépatocellulaires existantes a permis de retrouver des composantes correspondant aux variations identifiées dans les sous-groupes (mutation de *CTNNB1*, tumeurs très proliférantes activant le cycle cellulaire, échantillons surexprimant des gènes de l'interféron et groupe de tumeur avec un gain du chromosome 7), mais cette approche permet d'exprimer les dérégulations précédemment observées de manière plus quantitative. Les tumeurs ne sont pas seulement classées selon la dérégulation majeure présente dans leur transcriptome, mais décrites comme la combinaison de différentes altérations transcriptionnelles. Cela résout le problème des tumeurs présentant des caractéristiques de plusieurs groupes et permet de mieux comprendre la diversité des dérégulations à l'œuvre dans les CHC. De nouvelles associations ont également été mises en évidence, comme l'activation de la voie du stress oxydatif liée au gain du bras chromosomique 8q, qui suggère l'implication du gène *MYC* localisé au locus 8q24. Cette observation doit être approfondie, notamment au niveau expérimental, pour comprendre le lien entre les deux altérations et leur possible coopération, mais elle reflète l'utilité de l'ACI pour identifier de nouvelles connexions fonctionnelles à l'aide des données transcriptomiques.

1.3. Caractérisation des dérégulations liées à l'activation des cyclines

En parallèle de mes travaux sur les composantes transcriptionnelles, j'ai participé à la caractérisation d'un nouveau sous-groupe de carcinomes hépatocellulaires, particulièrement agressif et principalement développé chez les patients non-cirrhotiques, défini par les altérations activatrices des cyclines A2 ou E1 (CCH-HCC). Ces deux gènes régulent le cycle cellulaire en favorisant l'entrée et la progression en phase S. Nous avons identifié différents mécanismes d'activation de ces gènes, comme l'insertion du virus de l'hépatite B (HBV) ou du virus adéno-associé de type 2 (AAV2), le détournement d'un enhancer et des fusions récurrentes de *CCNA2*. L'activation de ces gènes induit une prolifération cellulaire avec une entrée prématurée en phase S et un stress réplicatif intense qui induit l'apparition d'une signature particulière de réarrangements structuraux, caractérisée par des centaines de duplications focales et translocations inter-chromosomiques.

Dans ce projet, je suis intervenue à plusieurs étapes. Dans un premier temps, j'ai utilisé le pipeline mis en place dans le laboratoire pour obtenir les données d'expression des gènes analysé en RNAseq et identifier les tumeurs présentant une surexpression de *CCNA2* ou *CCNE1*. J'ai également réalisé une classification non supervisée des tumeurs du laboratoire (série LICA-FR) et du TCGA (TCGA-LiHC) par t-SNE et montré que les tumeurs avec activation de *CCNA2* ou *CCNE1* définissent un groupe moléculaire homogène. J'ai alors analysé les gènes dérégulés dans les CCN-HCC et mis en évidence une forte activation de la voie E2F et de la voie ATR impliquée dans la réponse au stress réplicatif. A l'inverse, les gènes de la phosphorylation oxydative sont sous-exprimés dans ces tumeurs, suggérant un switch vers un métabolisme anaérobie (effet Warburg) permettant de soutenir la prolifération cellulaire. Enfin, j'ai analysé les caractéristiques cliniques et moléculaires des CCN-HCC, qui sont typiquement des tumeurs de grosse taille, développées chez des patients jeunes en l'absence de cirrhose, avec une fréquence élevée d'inactivation des gènes *RB1* et *PTEN*. A l'inverse, ces tumeurs sont mutuellement exclusives des tumeurs mutées *CTNNB1*. Ainsi, l'activation des cyclines A2 ou E1 définit un nouveau groupe de CHC de mauvais pronostic associé à une signature d'instabilité génomique liée au stress réplicatif, qui pourrait conférer une vulnérabilité ciblable à ces tumeurs, par exemple via des inhibiteurs de la voie ATR.

Article 2 : " Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress" (Nature Communications, 2018)

Quentin Bayard, **Léa Meunier**, Camille Peneau, Victor Renault, Jayendra Shinde, Jean-Charles Nault, Iadh Mami, Gabrielle Couchy, Giuliana Amaddeo, Emmanuel Tubacher, Delphine Bacq, Vincent Meyer, Tiziana La Bella, Audrey Debaillon-Vesque, Paulette Bioulac-Sage, Olivier Seror, Jean-Frédéric Blanc, Julien Calderaro, Jean-François Deleuze, Sandrine Imbeaud, Jessica Zucman-Rossi & Eric Letouzé

nature COMMUNICATIONS

ARTICLE

DOI: 10.1038/s41467-018-07552-9 OPEN

Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress

Quentin Bayard[1,2,3,4], Léa Meunier[1,2,3,4], Camille Peneau[1,2,3,4], Victor Renault[5], Jayendra Shinde[1,2,3,4], Jean-Charles Nault[1,2,3,4,6,7], Iadh Mami[1,2,3,4], Gabrielle Couchy[1,2,3,4], Giuliana Amaddeo[8,9], Emmanuel Tubacher[5], Delphine Bacq[10], Vincent Meyer[10], Tiziana La Bella[1,2,3,4], Audrey Debaillon-Vesque[11], Paulette Bioulac-Sage[12,13], Olivier Seror[1,14], Jean-Frédéric Blanc[11,12], Julien Calderaro[8,15], Jean-François Deleuze[5,10], Sandrine Imbeaud[1,2,3,4], Jessica Zucman-Rossi[1,2,3,4,16] & Eric Letouzé[1,2,3,4]

Cyclins A2 and E1 regulate the cell cycle by promoting S phase entry and progression. Here, we identify a hepatocellular carcinoma (HCC) subgroup exhibiting cyclin activation through various mechanisms including hepatitis B virus (HBV) and adeno-associated virus type 2 (AAV2) insertions, enhancer hijacking and recurrent *CCNA2* fusions. Cyclin A2 or E1 alterations define a homogenous entity of aggressive HCC, mostly developed in non-cirrhotic patients, characterized by a transcriptional activation of E2F and ATR pathways and a high frequency of *RB1* and *PTEN* inactivation. Cyclin-driven HCC display a unique signature of structural rearrangements with hundreds of tandem duplications and templated insertions frequently activating *TERT* promoter. These rearrangements, strongly enriched in early-replicated active chromatin regions, are consistent with a break-induced replication mechanism. Pan-cancer analysis reveals a similar signature in *BRCA1*-mutated breast and ovarian cancers. Together, this analysis reveals a new poor prognosis HCC entity and a rearrangement signature related to replication stress.

[1] INSERM, UMR-1162, Génomique Fonctionnelle des Tumeurs Solides, Equipe Labellisée Ligue Contre le Cancer, Institut Universitaire d'Hématologie, Paris 75010, France. [2] Université Paris Descartes, Labex Immuno-Oncology, Sorbonne Paris Cité, Faculté de Médecine, Paris 75006, France. [3] Université Paris 13, Sorbonne Paris Cité, Unité de Formation et de Recherche Santé, Médecine, Biologie Humaine, Bobigny 93017, France. [4] Université Paris Diderot, Sorbonne Paris Cité, Paris 75013, France. [5] Laboratory for Bioinformatics, Fondation Jean Dausset – CEPH, Paris 75010, France. [6] Liver unit, Hôpital Jean Verdier, Hôpitaux Universitaires Paris-Seine-Saint-Denis, Assistance-Publique Hôpitaux de Paris, APHP, Bondy 93140, France. [7] Unité de Formation et de Recherche Santé Médecine et Biologie Humaine, Université Paris 13, Communauté d'Universités et Etablissements Sorbonne Paris Cité, Bobigny 93017, France. [8] Inserm, U955, Team 18, Université Paris-Est Créteil, Faculté de Médecine, Créteil 94010, France. [9] Assistance Publique-Hôpitaux de Paris, Service d'Hépatologie, CHU Henri Mondor, Créteil 94010, France. [10] Centre National de Recherche en Génomique Humaine, CEA, Evry 91000, France. [11] Service Hépato-Gastroentérologie et Oncologie Digestive, Hôpital Haut-Lévêque, Centre Hospitalier Universitaire de Bordeaux, Bordeaux 33076, France. [12] Université Bordeaux, Bordeaux Research in Translational Oncology, Bordeaux 33076, France. [13] Service de Pathologie, Hôpital Pellegrin, Centre Hospitalier Universitaire de Bordeaux, Bordeaux 33000, France. [14] Radiology Department, Jean Verdier Hospital, Hôpitaux Universitaires Paris-Seine-Saint-Denis, APHP, Bondy 93140, France. [15] Assistance Publique-Hôpitaux de Paris, Département de Pathologie, Hôpital Henri Mondor, Créteil 94010, France. [16] Assistance Publique-Hôpitaux de Paris, Hopital Européen Georges Pompidou, 75015 Paris, France. These authors contributed equally: Léa Meunier, Camille Peneau. These authors jointly supervised this work: Jessica Zucman-Rossi, Eric Letouzé. Correspondence and requests for materials should be addressed to J.Z-R. (email: jessica.zucman-rossi@inserm.fr) or to E.Lé. (email: eric.letouze@inserm.fr)

Hepatocellular carcinoma (HCC) is the third leading cause of cancer death worldwide. Only 30% of cases are diagnosed at an early stage and are amenable to curative treatment by tumor resection or liver transplantation[1]. The multikinase inhibitors sorafenib[2] and regorafenib[3] are currently the only drugs approved for advanced HCC cases, but the median life expectancy of patients with HCC on sorafenib is only 1 year. All phase III clinical trials involving targeted molecular therapies have failed so far for various reasons including liver toxicity, lack of antitumoral potency, and the molecular heterogeneity of the disease[4]. Identifying homogeneous HCC subgroups sharing similar driving mechanisms and vulnerabilities is thus crucial to design successful patient-tailored clinical trials.

Most HCC develop in a cirrhotic liver, associated with various etiologies including hepatitis B virus (HBV) and hepatitis C virus (HCV) infections, alcohol abuse, metabolic disease, and exposure to carcinogenic compounds like aflatoxin B1[5]. The natural history of HCC in cirrhosis follows a well-established sequence with the successive development of dysplastic nodules that can transform into early stage and advanced HCC. *TERT* promoter mutations are the initial oncogenic events already detected in dysplastic nodules[6] whereas alterations in other HCC drivers[7–11] involved in cell cycle control (*TP53*, *RB1*, *CCND1*, *CDKN2A*), Wnt/ß-catenin signaling (*CTNNB1*, *AXIN1*), oxidative stress response (*NFE2L2*, *KEAP1*) epigenetic regulation (*ARID1A*, *ARID2*) and the AKT/mTOR and MAP kinase pathway (*RPS6KA3*, *TSC1*, *TSC2*, *PTEN*) only occur in progressed HCC[12].

In 20% of the cases, HCC develops in absence of cirrhosis. These patients usually maintain adequate liver functions and, being less subject to liver toxicity, may be eligible for more treatment options. The etiology of HCC in absence of cirrhosis is largely unknown, but one mechanism of transformation involves insertional mutagenesis by the HBV virus. The first oncogenic HBV insertion was identified in cyclin A2 gene (*CCNA2*)[13]. Since then, recurrent HBV insertions were mapped in several oncogenes including *CCNE1*, *KMT2B* and *TERT*[14,15]. Recently, we identified adeno-associated virus type 2 (AAV2) insertions as a new etiology for HCC developed in absence of cirrhosis, with recurrent insertions in *CCNA2* and *CCNE1* genes[16]. However, the molecular consequences of viral insertions in cyclin genes and their precise role in HCC development remain poorly understood.

Here, we report the systematic screening of *CCNA2* and *CCNE1* alterations in 751 HCC. We identify new mechanisms of cyclin A2/E1 activation, and we explore the clinical and molecular characteristics of this tumor subgroup.

Results

Viral insertions and gene fusions activate cyclin A2. To identify the exhaustive landscape of *CCNA2* and *CCNE1* alterations in HCC, we analyzed 751 HCC comprising an in-house series of 160 tumors (LICA-FR) analyzed by RNA sequencing (RNAseq, $n = 160$), whole exome (WES, $n = 156$) and whole genome sequencing (WGS, $n = 45$) (Supplementary Data 1), the TCGA[17] series (334 HCC with RNA-seq and WES, 48 or which also analyzed by WGS) and the ICGC-JP[11] series (257 HCC with WGS data, Supplementary Data 2).

We first screened the LICA-FR series of 160 tumors to characterize the exhaustive mechanisms activating *CCNA2* and *CCNE1* in HCC. We identified one HBV and 5 AAV2 insertions (four previously described in the ref. [16]) in *CCNA2* gene (Supplementary Data 3), all but one located within *CCNA2* intron 2 (Fig. 1a). Viral insertions were associated with *CCNA2* mRNA over-expression ($P = 8.2 \times 10^{-9}$, fold-change = 5.6, Fig. 1b), but also altered the transcript and protein structure. AAV2 and HBV insertions induced the expression of various abnormal transcripts (Supplementary Fig. 1), predicted to generate a truncated cyclin A2 protein starting at methionine 148 or 158 with occasionally a few amino acids translated from the viral genome (Fig. 1c).

In addition we identified novel gene fusions in 4 tumors (Supplementary Data 4), all involving the C-terminal part of *CCNA2* (exons 3–8) at chromosome 4q27 downstream 3 different partner genes: *GSTCD* at 4q24, *SNX29* at 16p13.13 and *TET2* (×2) at 4q24 (Fig. 1a, d). In the *TET2-CCNA2* and *GSTCD-CCNA2* fusion transcripts, the first untranslated exons of *TET2* and *GSTCD* were linked with *CCNA2* exons 3–8. The *SNX29-CCNA2* fusion revealed an alternative transcription start site (TSS) in *SNX29* intron 14 generating a 448-nucleotide sequence spliced with *CCNA2* exon 3. In all fusions, the predicted translation initiation site of the fused RNA was located at methionine 158 in *CCNA2* exon 3, predicted to generate a truncated cyclin A2 protein of 275 amino acids (32 Kda), lacking the destruction box[18] and ubiquitination targeting sequences[19] but retaining the functional cyclin box, without any protein fragment from the partner genes (Fig. 1e).

Western blot analysis of 9 tumors with viral insertion or gene fusion confirmed the over-expression, as predicted, of a truncated 32 KDa protein (Fig. 1f). Thus, gene fusions and viral insertions in *CCNA2* both lead to the production of a stable protein lacking the N-terminal regulatory domains.

In the TCGA series, we identified 7 *CCNA2* fusions with 5 different partner genes (*FAM160A1*, *KIAA1109* × 3, *LIPC*, *UBA6* and *TDO2*, Fig. 1a, d), all of which involved the first untranslated exon(s) of the partner gene linked with exons 3–8 of *CCNA2*. WGS revealed in another tumor a focal deletion starting in the 5′ UTR region and ending in *CCNA2* intron 2 (Supplementary Fig. 2). All these events were predicted to generate the same 32 KDa truncated cyclin A2 protein lacking N-terminal regulatory domains. We also identified one tumor with HBV insertion and 3 tumors with AAV2 insertions in *CCNA2*. Finally, 6 tumors strongly overexpressed *CCNA2* (FPKM > 15), 3 of which displayed 23–48 Mb intra-chromosomal deletions linking the intergenic region downstream *CCNA2* with the highly expressed *ALB*, *AFP*, and *ADH6* genes (Supplementary Fig. 2). The ICGC-JP cohort comprised one HBV insertion in *CCNA2* intron 2 and one fusion between the first untranslated exon of *ANXA5* and exons 3–8 of *CCNA2* (Fig. 1a, d).

In total, we identified 10 HCC with *CCNA2* activation events in the LICA-FR series (6.2%), 2 in the ICGC-JP series (0.8%) and 18 in the TCGA series (5.4%), associated with a significant increase of *CCNA2* mRNA expression, but also generating a truncated cyclin A2 protein lacking the N-terminal destruction box and the ubiquitination site.

Viral insertions and enhancer hijacking activate cyclin E1. In our series of 160 HCC, we identified 5 AAV2 insertions (three previously described in the ref. [16]) and one HBV insertion in the 5′ region or upstream the transcription start site (TSS) of *CCNE1* (Fig. 2a, Supplementary Data 3). These viral insertions induced a massive overexpression of the full-length *CCNE1* gene (Fig. 2b), confirmed by western-blot analysis (Supplementary Fig. 3). Interestingly, one case with AAV2 insertion (FR2141T) also displayed an amplification of *CCNE1* locus including the viral sequence (Supplementary Fig. 3), suggesting a two-step selection of *CCNE1* activation in the natural history of this tumor. Four other tumors overexpressed *CCNE1* (FPKM > 6), explained by high-level amplification in one case. In the 3 remaining cases, whole genome sequencing revealed interchromosomal translocation breakpoints in the regulatory region of *CCNE1* (Fig. 2a). Tumor FR2048T

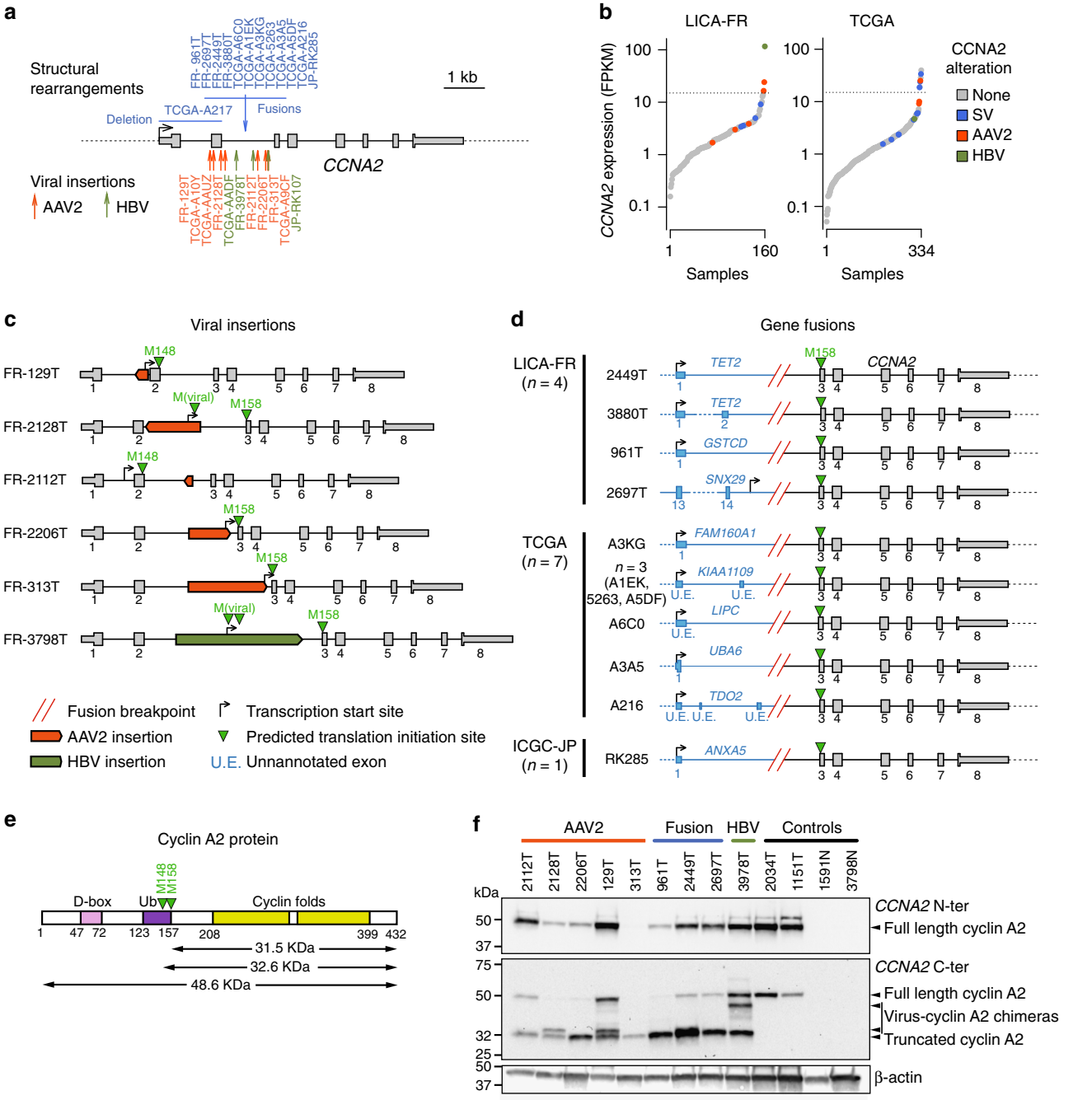


Fig. 1 Diverse mechanisms leading to *CCNA2* activation in HCC. **a** Summary of structural rearrangements (top) and viral insertions (bottom) affecting *CCNA2* gene identified in 751 HCC from the LICA-FR, TCGA and ICGC-JP cohorts. **b** Sorted *CCNA2* expression (log scale) in the LICA-FR and TCGA cohorts. Gene expression was obtained from RNA-seq data and is given in fragments per kilobase of exons per million reads (FPKM). Samples harboring structural variants (SV) or viral insertions are indicated with a color code. **c** Functional consequences of AAV2 and HBV insertions in *CCNA2*. Viral insertions identified in the LICA-FR cohort were precisely mapped using WGS or viral capture data, and RNA-seq reads were aligned on the reconstructed chimeric DNA to identify the transcription start sites and predicted translation initiation sites of abnormal transcripts. **d** *CCNA2* fusions identified in the LICA-FR, TCGA and ICGC-JP cohorts. The transcription start site of the fusion transcript is represented together with the predicted translation initiation site. Fusions with *KIAA1109*, *LIPC* and *TDO2* involve 5′ exons not annotated in transcript databases but expressed in normal liver. **e** Schematic representation of cyclin A2 protein with functional domains. D-box Destruction box; Ub, Ubiquitination targeting sequences. **f** Western blot analysis of cyclin A2 using antibodies targeting the N-terminal (top) or C-terminal (middle) domains. Tumors with viral insertions or gene fusions are compared with tumors without *CCNA2* alteration and non-tumoral liver controls

displayed a translocation placing *CCNE1* downstream the first untranslated exon of the highly expressed *ERRFI1* gene, leading to a highly expressed *ERRFI1-CCNE1* fusion. The two other translocations lead to juxtapose *CCNE1* promoter with enhancer-rich chromatin areas located close to the highly expressed genes *RAPH1* and *CYB5A* (Fig. 2c). Thus, both viral insertions and structural rearrangements can activate *CCNE1* expression by bringing viral or distal human enhancers in the regulatory region of the gene.

In the TCGA series, 10 tumors overexpressed *CCNE1* (Fig. 2b), including 2 cases with HBV insertion, one with HBV insertion

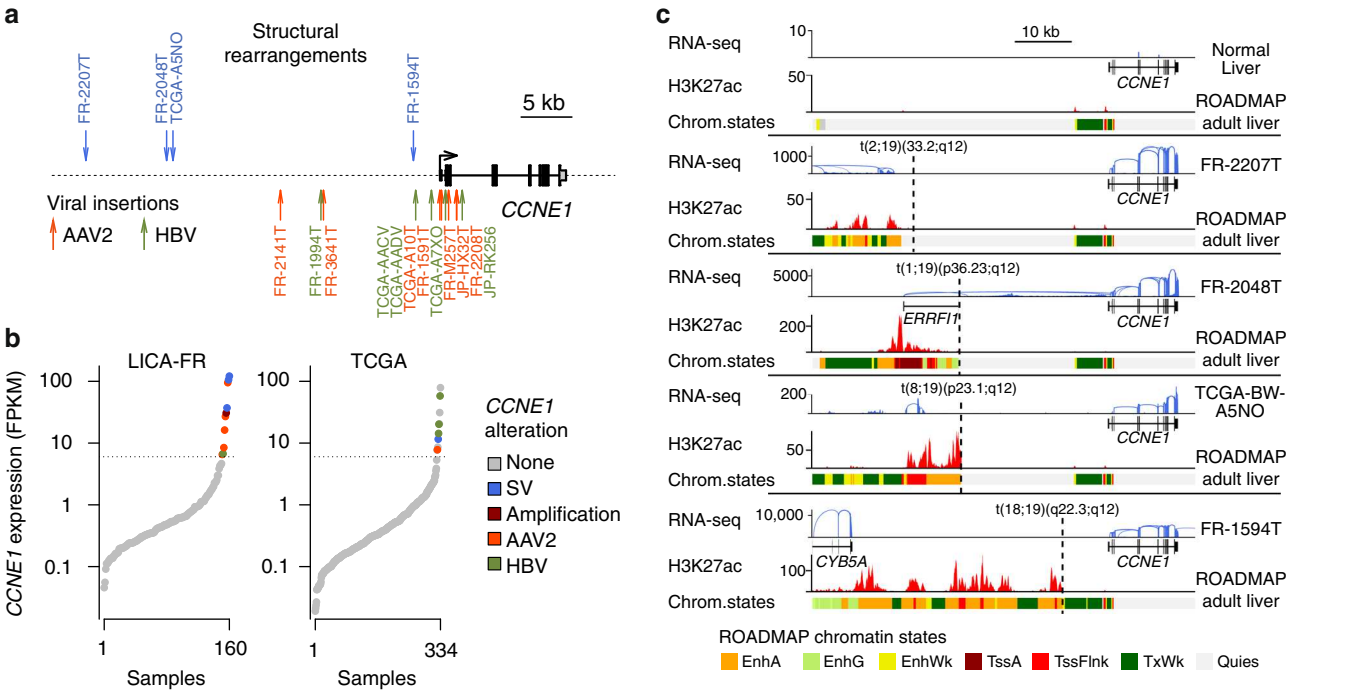


Fig. 2 Viral and non-viral mechanisms of *CCNE1* activation in HCC. **a** Summary of structural rearrangements (top) and viral insertions (bottom) affecting *CCNE1* gene identified in 751 HCC from the LICA-FR, TCGA and ICGC-JP cohorts. **b** Sorted *CCNE1* expression (log scale) in the LICA-FR and TCGA cohorts. Gene expression was obtained from RNA-seq data and is given in fragments per kilobase of exons per million reads (FPKM). Samples harboring structural variants (SV), focal amplifications and viral insertions are indicated with a color code. **c** Functional consequences of structural rearrangements affecting *CCNE1* regulatory region. RNA-seq read counts along *CCNE1* locus are represented in normal liver (top) and in 4 tumors harboring structural rearrangements upstream *CCNE1* trancription start site (TSS). H3K27Ac chromatin immunoprecipitation sequencing (ChIP-seq) signal and chromatin states in adult liver were obtained from the ROADMAP consortium and are depicted below each reconstruted DNA sequence. EnhA: active enhancer; EngG: genic enhancer; EnhWk: weak enhancer; TssA: active TSS; TssFlnk: flanking TSS; TxWk: weak transcription; Quies: quiescent chromatin

plus high-level amplification, one with AAV2 insertion and one with a translocation between *CCNE1* regulatory region and an enhancer-rich region on chromosome 5 (Fig. 2c). In the 5 remaining cases, the mechanism leading to *CCNE1* overexpression remained unexplained in absence of WGS data. In the ICGC-JP cohort, we identified one AAV2 and one HBV insertion associated with *CCNE1* overexpression. In total, we identified 10 HCC with *CCNE1* activation events in the LICA-FR cohort (6.2%), two in the ICGC series (0.8%) and 10 in the TCGA series (3.0%).

Across the three data sets, 52/751 tumors (6.9%) displayed an activation of cyclin A2 ($n = 30$) or E1 ($n = 22$) due to viral insertions or structural rearrangements. These are later referred to as CCN-HCC. The proportion of CCN-HCC varied between the cohorts (12.5% in our series, 8.4% in TCGA and 1.6% in ICGC-JP) due to differences in etiological backgrounds (Supplementary Data 2). It was particularly high in our series enriched in cancers developed in a non-fibrotic liver, and low in the IGGC-Japan series dominated by HCV-related cases.

Cyclin A2 or E1 activation defines a homogenous HCC subgroup. We next explored the molecular and clinical characteristics of CCN-HCC. Gene expression analysis of the LICA-FR and TCGA showed that CCN-HCC defined homogeneous transcriptional clusters (Fig. 3a). They were characterized by an overexpression of cell cycle genes, in particular E2F targets, and an activation of the ATR pathway in response to replication stress (Fig. 3b, Supplementary Data 5). The most significant downregulated pathways were oxidative phosphorylation, suggesting a metabolic switch to aerobic glycolysis (Warburg effect), and *MYC* targets. We also compared the alteration frequencies of known liver cancer driver genes[10] between CCN-HCC and others. *CCNA2* and *CCNE1* activation events were remarkably exclusive from *CTNNB1* and *TERT* promoter mutations, but frequently associated with *PTEN* and *RB1* inactivation in both the LICA-FR and TCGA series (Fig. 3b, Supplementary Data 6). *RB1* inactivation may allow cells to overcome oncogene-induced senescence[20] in these tumors, whereas *PTEN* inactivation might favor the oncogenic metabolic switch that we observed at the transcriptional level[21]. Compared to the other tumors in the LICA-FR series, CCN-HCC were enriched in large tumors (median largest nodule diameter = 115 vs. 60 mm, $P = 0.0033$), of poor prognosis (median overall survival = 21 vs. 69 months, $P = 0.0072$, Fig. 3c), developed in younger patients (median age = 57 vs. 67 years old, $P = 0.050$) with a non-fibrotic liver (fibrosis stage F0-F1 80 vs. 42%, $P = 0.0011$). Thus, CCN-HCC define a homogenous HCC entity with characteristic clinical and molecular features.

CCN-HCC display a unique structural rearrangement signature. To identify mutational signatures associated with CCN-HCC, we analyzed the whole genome sequences of 45 of our 160 HCC (35 were previously published[22], 10 new), including 13 CCN-HCC. With a median of 12,463 mutations, CCN-HCC were rather less mutated than others (median = 16,397 mutations, $P = 0.065$). Mutational signatures 4, 5, and 16 (COSMIC nomenclature), ubiquitous in liver cancers[22], accounted for most mutations in CCN-HCC, with a slight increase of signature 5 (53 vs. 33%, $P = 0.036$) and decrease of signature 16 (23 vs. 32%, $P = 0.05$) as compared with other HCC (Supplementary Fig. 4).

In contrast, CCN-HCC displayed > 3 times more structural variants (median = 415 vs. 126, $P = 1.1 \times 10^{-4}$). We identified 6 rearrangement signatures, termed RS1 to RS6, characterized by

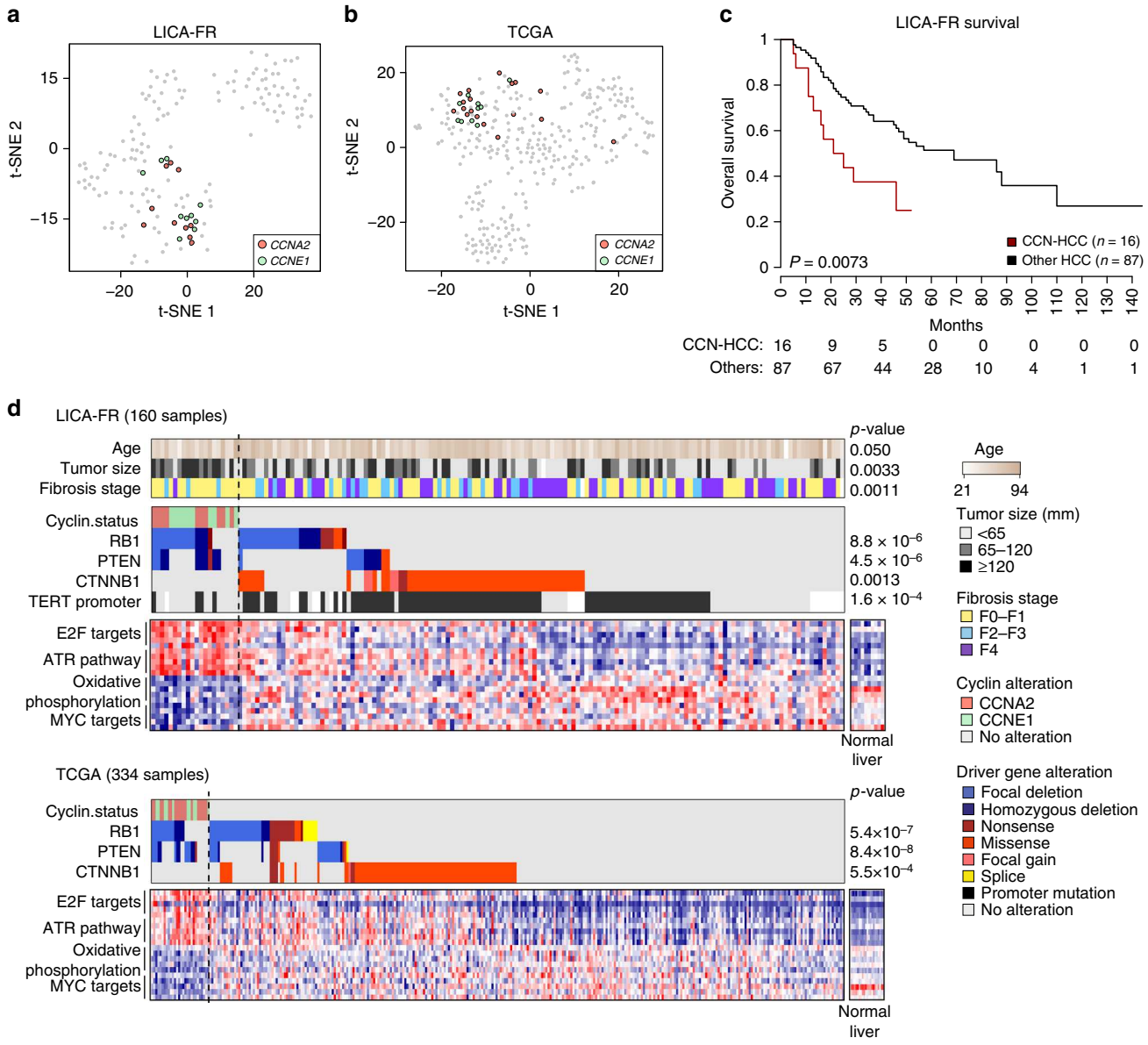


Fig. 3 Clinical and molecular features of cyclin-activated HCC. **a** t-SNE plots depicting the classification of HCC from the LICA-FR and TCGA cohorts based on their transcriptional profiles. Tumors harboring *CCNA2* or *CCNE1* activating alterations are indicated with a color code. **b** Clinical characteristics, driver genes and deregulated pathways associated with CCN-HCC in the LICA-FR (top) and TCGA (bottom) cohorts. **c** Overall survival in CCN-HCC as compared with other HCC in the LICA-FR cohort. Only HCC with curative resection (R0) were included

different combinations of rearrangement categories defined according to the type, size, and clustered nature of rearrangements (Fig. 4a). Strikingly, a high number of rearrangements attributed to signature RS1 (≥50 events) was specifically encountered in a cluster of 13 tumors corresponding exactly to CCN-HCC ($P = 1.4 \times 10^{-11}$, Fig. 4b). We validated this association using WGS data from the ICGC-JP series and a subset of 48 samples from the TCGA series (Fig. 4c, Supplementary Data 7). In absence of WGS data for the rest of the TCGA series, we used SNP array data to estimate the number of focal gains (<200 kb) in each tumor as a surrogate marker of the RS1 signature. With a median of 120 events, CCN-HCC displayed a significant increase of focal gains as compared with other HCC in the TCGA series (median = 6, $P < 2.2 \times 10^{-16}$, Supplementary Fig. 5). Thus, CCN-HCC have a relatively low mutation burden but a large number of structural rearrangements with a specific signature.

RS1 features suggest a replication stress-induced mechanism. Almost all rearrangements in CCN-HCC belonged to signature RS1, characterized by a combination of small tandem duplications (<100 kb) and inter-chromosomal translocations (Fig. 4d). CCN-HCC also displayed a typical copy-number profile showing hundreds of focal gains, usually one copy above surrounding chromosome segments (Supplementary Fig. 6). Surprisingly, overlaying structural rearrangement breakpoints with copy-number profiles revealed that only 68% of these gains were due to tandem duplications, other gains being frequently surrounded by translocation or inversion breakpoints (Fig. 4e, Supplementary Fig. 6). A recurrent feature consisted of several chromosome segments, usually between 10 and 100 kb, stung together and with the same duplication level relative to their source chromosomes. Most of these events involved segments from two (Fig. 4f) or more (Supplementary Fig. 7) different chromosomes, a feature recently described as templated

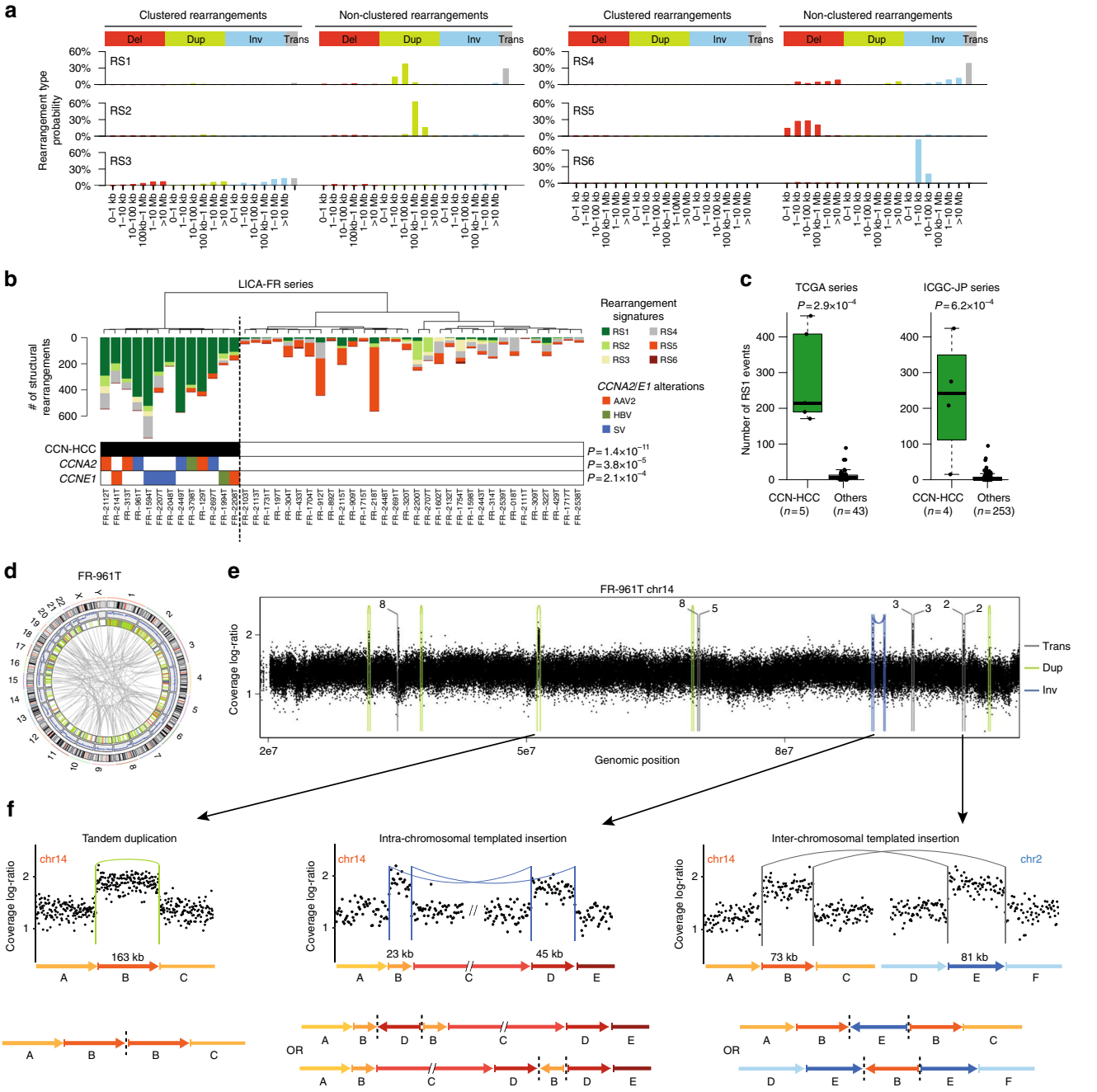


Fig. 4 Cyclin-activated HCC display a specific signature or structural rearrangements. **a** Six rearrangement signatures identified across 350 HCC genomes by non-negative matrix factorization. Structural rearrangements were classified in 38 categories considering their type (del: deletion, dup: tandem duplication, inv: inversion, trans: inter-chromosomal translocation) and size, and distinguishing clustered from non-clustered events. The probability of each rearrangement category in each signature is represented, with rearrangement types indicated above and rearrangement sizes below. **b** Unsupervised classification of 45 HCC from the LICA-FR cohort based on the contribution of rearrangement signatures in each tumor. Significant molecular alterations associated with the cluster of tumors having a high contribution of signature RS1 are represented below. *P*-values were obtained using Fisher's exact tests. **c** Validation of the association between the RS1 signature and CCN-HCC in the TCGA and ICGC-JP series. The middle bar, median; box, interquartile range; bars extend to 1.5 times the interquartile range. **d** CIRCOS plot representing the structural rearrangement profile of a representative CCN-HCC (FR-961T, harboring a *GSTCD-CCNA2* fusion). **e** Copy-number profile showing the accumulation of focal gains along chromosome 14 in tumor FR-961T. Structural rearrangements are overlaid on the copy-number profile with a color code indicating the type of event. trans: inter-chromosomal translocation; dup: tandem duplication; inv: inversion. **f** Three types of rearrangements leading to focal chromosome gains in CCN-HCC. A representative example of each type of event is shown with a copy-number plot above and a schematic representation of the rearranged chromosome below. Structural rearrangements are represented with the same color code as in **e**. Dashed lines on schematic chromosome reconstructions represent the abnormal junctions detected in WGS data

insertion cycle[23]. Inter-chromosomal templated insertions accounted for 11% of focal gains in CCN-HCC. Other events, which we call intra-chromosomal templated insertions, involved distal segments of a same chromosome and appeared as couples of inversions (Fig. 4f) or duplication and deletion (Supplementary Fig. 7), depending on the orientation of the-junctions. Intra-chromosomal templated insertions accounted for 7% of focal gains in CCN-HCC. All these events are consistent with a replication-based mechanism in which a DNA polymerase at a stalled replication fork would switch template, replicate one or more other DNA regions and switch back to the original template strand behind the point of departure, generating a duplication on the host chromosome[23–26]. Such mechanism could be particularly active in CCN-HCC due to replication stress induced by premature S phase entry.

Structural rearrangements activate *TERT* promoter in CCN-HCC. To better understand the functional consequences of the rearrangement phenotype observed in CCN-HCC, we examined the location of 8466 breakpoints attributed to the RS1 signature among the 350 liver cancer genomes from the LICA-FR, TCGA and ICGC cohorts. RS1 breakpoints were not distributed evenly along the genome but formed clusters located almost exclusively within active topologically associated domains (TADs, Fig. 5a) characterized by early replication, high gene expression and active chromatin states in normal liver (Fig. 5b). In particular, RS1 breakpoint hotspots were frequently observed at loci encoding very highly expressed liver enzymes exemplified by the albumin (*ALB*), alcohol dehydrogenase (*ADH*) and hydroxysteroid 17-Beta dehydrogenases (*HSD17B*) loci on chromosome 4 (Fig. 5a, Supplementary Fig. 8). Among the 18 chromatin states defined by the ROADMAP consortium in normal adult liver, active transcription start sites (TSS) and enhancer regions were the most strongly enriched in RS1 breakpoints (fold-change > 3), whereas quiescent and heterochromatin domains were the most depleted (Fig. 5c). TSS and enhancer regions were also enriched, to a lesser extent, in breakpoints related to signature RS2 characterized by large tandem duplications. By contrast, breakpoints related to signature RS6, dominated by inversions < 10 kb, were predominantly observed in heterochromatin and ZNF repeats.

We then used binomial regression[27] to model the density of rearrangement breakpoints along the genome considering an extensive set of genomic features (Supplementary Fig. 9) and to identify hotspots harboring more breakpoints than expected by chance from the background model, which may indicate positive selection in CCN-HCC. We identified a single significant locus corresponding to *TERT* promoter region (q=0.0029, Fig. 5d). Although *TERT* promoter mutations were rare in CCN-HCC (9 vs. 55% in others, $P = 2.4 \times 10^{-5}$), *TERT* promoter rearrangements were highly enriched (82 vs. 7%, $P = 1.8 \times 10^{-15}$, Fig. 5e) and involved regions of active chromatin in normal liver, in the vicinity of highly expressed liver enzymes (*ALB*, *FGG*, *SEP15*, *SLC12A7* and *BAAT*) or transcription factors (*HNF4A*, *CEBPA*, and *CEBPB*) (Supplementary Data 8, Supplementary Fig. 10). *TERT* promoter rearrangements induced an overexpression of *TERT*, stronger than promoter mutations but lower than HBV insertions (Supplementary Fig. 11). Of the 18 *TERT* promoter rearrangements identified in CCN-HCC, 16 could be associated with signature RS1 with a probability ≥ 0.5 (Fig. 5f). By contrast, most *TERT* promoter rearrangements in other HCC were related to signature RS4. Thus, structural rearrangements induced by replication stress are enriched at active chromatin regions and can promote CCN-HCC development by activating oncogenes like *TERT*.

CCN-HCC share a similar signature with *BRCA1*-altered cancers. To investigate the prevalence of the RS1 signature in other cancer types, we applied our method to 2606 tumors from the ICGC PanCancer Analysis of Whole Genomes (PCAWG) dataset[23,28,29]. In this pan-cancer series, we identified 9 rearrangement signatures (Supplementary Fig. 12), including one signature (RS1-pancan) highly similar to the RS1 signature that we identified in liver cancers (cosine similarity = 0.91). The RS1-pancan signature was detected at low frequency in several cancer types (e.g. bladder, lung, esophageal and gastric cancers), and was highly active in breast (18% of samples with ≥ 50 RS1 events) and ovarian (33%) cancers. However, this signature was associated with *CCNA2/E1* rearrangements only in liver cancer (Fig. 6a, Supplementary Data 9). Thus, the relationship between cyclin A2/E1 activation and signature RS1 is specific to liver cancer, and the molecular cause of this signature in other cancer types remains to be elucidated. In ovarian and breast cancer, RS1 signature was not associated with *CCNE1* amplifications but with *BRCA1* inactivation (Fig. 6b, c), consistent with previous reports[30,31]. Despite sharing a common signature of short tandem duplications and templated insertions, *CCNA2*, *CCNE1* and *BRCA1*-altered tumors displayed slightly different characteristics. First, the number of RS1 rearrangements was higher in *CCNA2*-activated HCC (median = 269) than in *CCNE1*-activated HCC (137) and *BRCA1*-altered breast (132) and ovarian (159) cancers (Fig. 6d). Second, tandem duplications were larger in *CCNE1*-activated HCC (median = 39 kb) than in *CCNA2*-activated HCC (22 kb), and smaller in *BRCA1*-altered breast (9 kb) and ovarian (10 kb) cancers (Fig. 6e). Finally, duplication and translocation breakpoints were strongly enriched in early-replicated regions in CCN-HCC as compared with other HCC, but not in *BRCA1*-altered as compared with other breast and ovarian cancers (Fig. 6f). Cyclin E1 activation was recently shown to induce replication stress by firing novel replication origins located within highly transcribed genes and prone to collapse[32]. BRCA1 is implicated in the response to replication stress[33,34] and its inactivation leads to tandem duplication formation at stalled forks by a replication restart-bypass mechanism[35]. Cyclin A2/E1 activation in HCC and *BRCA1* inactivation in breast and ovarian cancers may thus converge towards a similar rearrangement signature, with specificities reflecting the different ways by which these genetic alterations induce replication stress or modulate response to it (Fig. 6g).

Discussion

Here, we report the characterization of a homogeneous HCC subgroup driven by the activation of *CCNA2* or *CCNE1* gene. CCN-HCC represent 7% of HCC across the 3 data sets analyzed here, but up to 14% of HCC developed in a non-fibrotic liver. These patients often have atypical clinical presentation, without any history of primary risk factors, and can be remarkably young, exemplified by tumor FR-3880T developed in a 32 year-old woman without any risk factor, due to a *TET2-CCNA2* fusion. CCN-HCC are usually large tumors of poor prognosis but share molecular characteristics, in particular high proliferation and replication stress, that could provide therapeutic opportunities[36]. First, conventional chemotherapies mainly affect actively dividing cells by generating DNA damage or blocking DNA replication, and the tandem duplicator phenotype was identified as a marker for chemotherapeutic response in breast cancer cell lines and patient-derived xenografts[37]. Transarterial chemoembolization (TACE) with doxorubicin, cisplatin or epirubicin, usually recommended for patients with intermediate HCC not eligible for surgery, may thus be an interesting option for CCN-HCC. Poly (ADP-ribose) polymerase (PARP) inhibitors, the first clinically

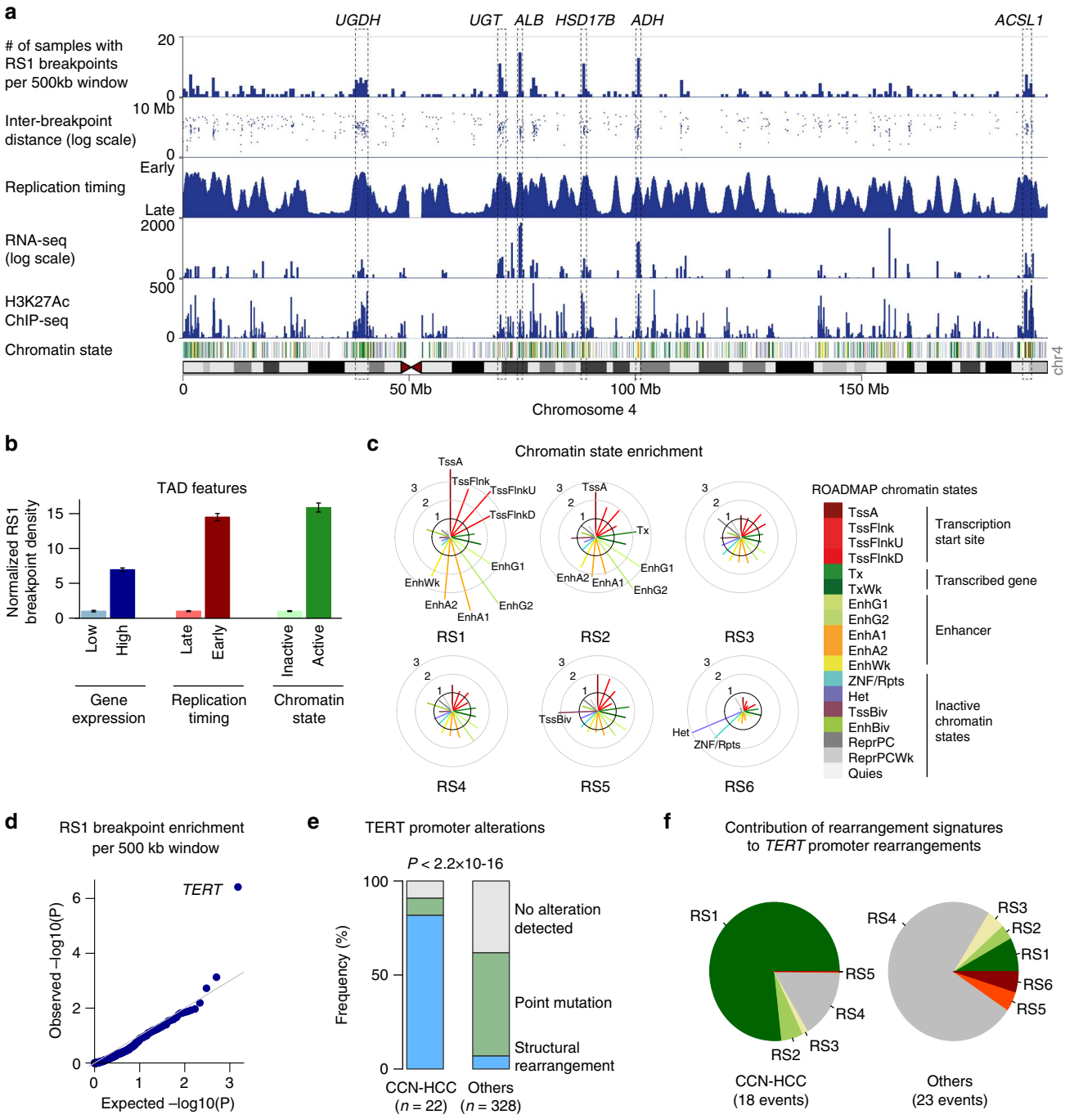
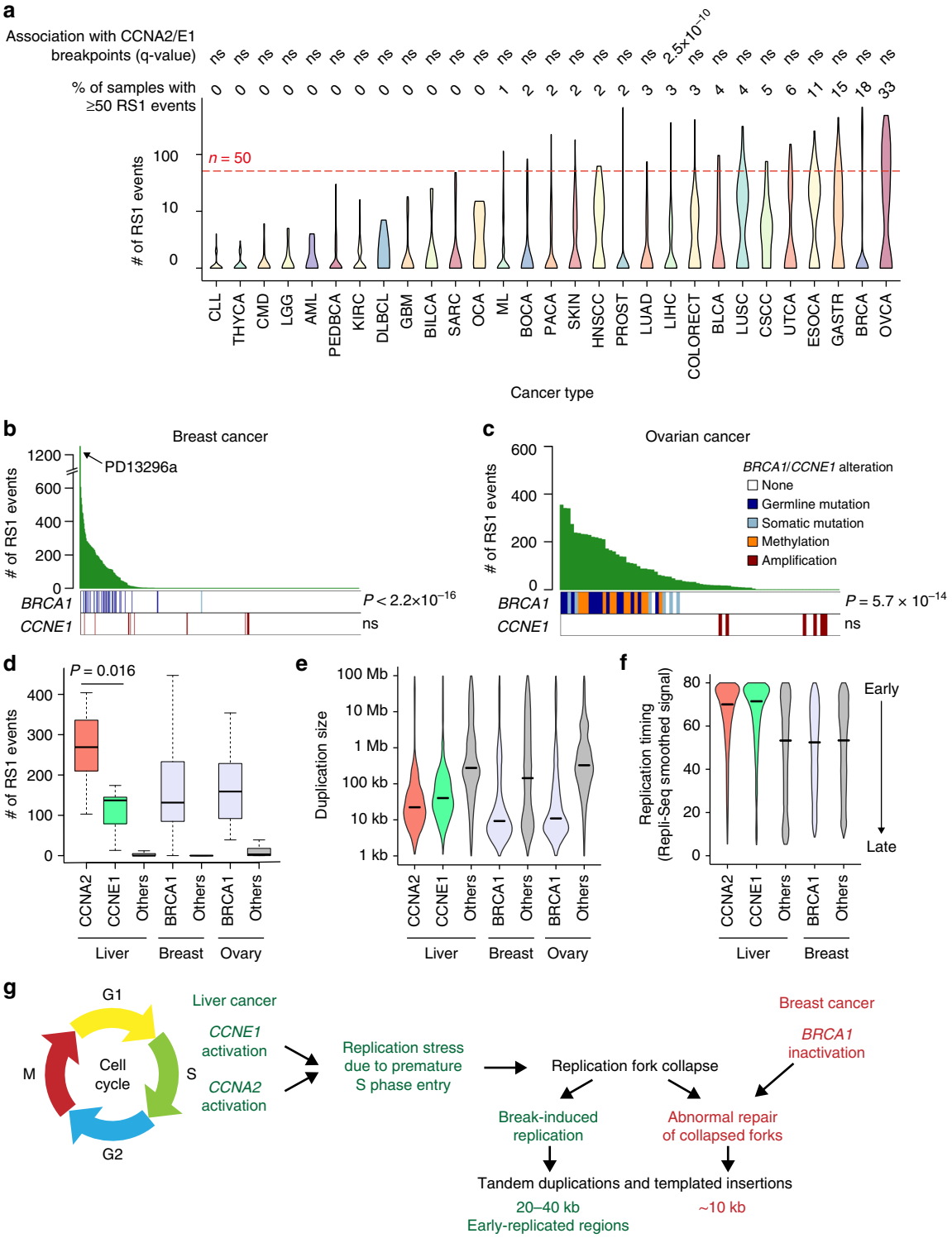


Fig. 5 Hotspot analysis of rearrangement signature 1 (RS1) breakpoints. **a** The density of RS1 breakpoints along chromosome 4 is displayed above replication timing, RNA-seq expression, H3K27Ac ChIP-seq profile and chromatin state. Replication timing was determined using Repli-Seq data from the liver cancer cell line HepG2. RNA-seq profile was generated from a normal liver sample. H3K27Ac and chromatin states in normal adult liver were obtained from the ROADMAP consortium. The legend for chromatin state color codes is displayed in **c**. Hotspots corresponding to highly expressed liver enzymes are annotated (UGDH, UDP-glucose 6-dehydrogenase; UGT, UDP glucuronosyltransferase family cluster; ALB, albumin; HSD17B, hydroxysteroid 17-Beta dehydrogenases 11 and 13, ADH, alcohol dehydrogenase cluster; ACSL1, acyl-CoA synthetase long chain family member 1). **b** RS1 breakpoint density in topologically associated domains (TADs). TADs were defined in human embryonic stem cells (H1) and classified based on gene expression in normal liver, replication timing and chromatin state. For each comparison, breakpoint density was normalized to be 1 in the group with the lowest density. Error bars indicate the 95% confidence interval. **c** Enrichment of rearrangement breakpoints in ROADMAP chromatin states for the 6 rearrangement signatures identified in HCC. For each signature, the fold-change between the observed and expected number of breakpoints falling within each chromatin state is represented, and chromatin states with a >2-fold enrichment are annotated. **d** Quantile-quantile plot of RS1 breakpoint enrichment p-values across 500 kb windows. **e** Proportion of *TERT* promoter alterations in CCN-HCC and other HCC analyzed by WGS. **f** Contribution of the 6 rearrangement signatures to *TERT* promoter rearrangements in CCN-HCC and other HCC



approved drugs designed to exploit synthetic lethality, have demonstrated benefit for patients carrying *BRCA1* mutations[38]. CCN-HCC do not harbor a DNA repair defect but share with *BRCA1*-altered tumors a signature of genomic instability that could conceivably confer these tumors sensitivity to PARP inhibitors. Finally, there are currently several compounds in phase I and II trials targeting the replication stress response pathway members *ATR*, *CHK1* and *WEE1*[39]. If brought to the clinic, such compounds would be promising for CCN-HCC treatment, given that the ATR pathway is strongly upregulated in CCN-HCC and overexpression of *CCNE1* has been shown to confer increased sensitivity to ATR inhibition[40].

We describe for the first time recurrent fusions involving *CCNA2* gene and recurrent rearrangements of *CCNE1* promoter region. *CCNA2* fusions are only the second recurrent fusion event identified in hepatocellular carcinoma, after the *PRKACA-DNAJB1* fusion characteristic of the rare fibrolamellar carcinoma subtype[41]. These fusions always involve the untranslated 5′ region of different partner genes upstream exons 3–8 of *CCNA2*, which constitutes an original mechanism leading to oncogene activation

Fig. 6 Pan-cancer analysis of the RS1 signature **a** Violin plots representing the number of rearrangements attributed to signature RS1 across patients within each cancer type in the ICGC PCAWG data set. For each cancer type, we assessed the association between tumors with ≥ 50 RS1 events and tumors with a rearrangement breakpoint < 80 kb from *CCNA2* or *CCNE1* gene using Fisher's exact tests. ns: not significant. The definition of cancer codes and number of samples per cancer type are available in Supplementary Data 9. **b** Number of RS1 events across 524 breast cancer genomes[30] and association with *BRCA1* alterations and *CCNE1* amplifications. PD13296a, the only tumor with both *BRCA1* mutation and *CCNE1* amplification, has the highest number of RS1 events in the series. **c** Number of RS1 events across 80 ovarian cancer genomes[75] and association with *BRCA1* alterations and *CCNE1* amplifications. *P*-values were obtained using one-sided Wilcoxon rank-sum tests. **d** Number of RS1 events in liver, breast and ovarian cancers with or without *CCNA2*, *CCNE1* and *BRCA1* alterations. The middle bar, median; box, interquartile range; bars extend to 1.5 times the interquartile range. **e**, Violin plots representing the distribution of tandem duplication sizes across liver, breast and ovarian cancers with or without *CCNA2*, *CCNE1* and *BRCA1* alterations. **f** Violin plots representing the replication timing of duplication and inter-chromosomal translocation breakpoint loci in liver and breast cancers with or without *CCNA2*, *CCNE1* and *BRCA1* alterations. Replication timing was determined using Repli-Seq data from the HepG2 cell line for liver cancer and from the MCF-7 cell line for breast cancer. **g** Proposed connexion beween rearrangement signatures in CCN-HCC and in *BRCA1*-inactivated breast and ovarian cancers

by truncating a regulatory N-terminal domain. Apart from liver cancers, none of the 2606 tumor genomes from the ICGC PCAWG dataset displayed a rearrangement breakpoint in *CCNA2* intron 2. Consistently, a recent RNA-seq analysis of 9,624 TCGA samples from 33 cancer types[42] did not reveal any *CCNA2* fusion in other cancer types. *CCNA2* fusions thus appear to be specific of liver cancers. Rearrangements affecting *CCNE1* promoter region result in the overexpression of cyclin E1 by bringing active enhancer regions upstream the transcription start site, mirroring the effect of viral enhancers. This mechanism was more frequent than *CCNE1* amplification in the liver cancer series we analyzed. Although HBV and AAV2 insertions were previously identified in *CCNA2* and *CCNE1*[14,16], the functional consequences of these insertions were unknown. By integrating WGS and RNA-seq data, we demonstrate here that viral insertions in *CCNA2*, like gene fusions, induce abnormal transcripts leading to truncated proteins lacking N-terminal regulatory domains. By contrast, viral insertions in *CCNE1* region lead to the overexpression of a full-length transcript and protein.

CCN-HCC display a characteristic transcriptional program, with a strong overexpression of E2F targets. Activation of the E2F pathway is expected in *RB1*-altered tumors and was already described in HCC[43]. However, E2F pathway is also activated in CCN-HCC without *RB1* inactivation event and may be partly explained by the ability of cyclin E/Cdk2 complexes to phosphorylate Rb. Interestingly, E2F-1 overexpression in the liver causes dysplasia and tumors in mice[43], and E2F1 was shown to inhibit c-Myc-driven apoptosis by activating PIK3CA/Akt/mTOR and c-Myb/COX-2 pathways[44].

A striking feature of CCN-HCC is the accumulation of hundreds of tandem duplications and templated insertion cycles. A recent study showed that *CCNE1* activation in U2OS cell lines leads to shortened G1 phase, early S phase entry and firing of normally silenced replication origins in highly expressed genes, prone to collapse and associated with DNA double-strand breaks formation[32]. Double-strand breaks formed following replication fork breakdown are primarily repaired by break-induced replication (BIR)[45]. In a cyclin E overexpression model of DNA replication stress, BIR was shown to be required for cell cycle progression and to induce duplications < 200 kb[46]. In addition, template switching may occur during BIR and generate complex chromosome rearrangements[24,25,47]. Thus, the nature of rearrangements identified in CCN-HCC and the enrichment of breakpoints in early-replicated, actively transcribed regions are consistent with a BIR mechanism induced by replication stress. However, future studies addressing the precise molecular mechanism generating templated insertions will be crucial to fully understand the relationship between replication stress and the RS1 rearrangement signature. The mechanism of tandem duplication formation in *BRCA1*-mutant cells was recently identified[35]. It involves abnormal repair of collapsed replication forks by a "replication restart bypass" mechanism with extension of the stalled leading strand by a migration bubble mechanism similar to BIR[48], terminated by end joining or by microhomology-mediated template switching. Thus, structural rearrangements induced by cyclin activation and *BRCA1* deficiency are initiated by replication fork collapse and processed by different repair mechanisms leading to a similar rearrangement signature with subtle differences regarding the size of rearrangements and breakpoint location. Interestingly, *BRCA1* inactivation and *CCNE1* amplification are mutually exclusive in ovarian cancers[49], and have been shown to be synthetically lethal[50]. The single breast tumor that we identified with both *BRCA1* mutation and *CCNE1* amplification (PD13296a) had the highest number of rearrangements related to the RS1 signature ($n = 1221$) across all the tumors we analyzed.

Contrary to *CCNA2* alterations that seem to be specific of liver cancers, *CCNE1* activation by high-level amplification is frequent across human cancers, in particular in gynecologic cancers[51]. Yet, *CCNE1* amplification in breast and ovarian cancers does not lead to the rearrangement phenotype that we observed in CCN-HCC. Several reasons may explain this discrepancy. First, adult hepatocytes are quiescent, rarely divide, and may thus be particularly sensitive to replication stress. Second, *CCNE1* is mostly activated by viral insertions and structural rearrangements of regulatory regions in HCC, rather than chromosome amplifications. These alterations may not have exactly the same functional consequence. Finally, we believe that viral insertions and structural rearrangements activating *CCNA2* or *CCNE1* are early events triggering hepatocarcinogenesis because they occur in patients without cirrhosis and in absence of other oncogenic event like *CTNNB1* mutations. *CCNE1* amplifications may occur later in breast and ovarian tumors, not leaving enough time for rearrangements to accumulate. Fujimoto et al. reported a positive correlation between the number of structural rearrangements and HBV insertion sites, suggesting that double-strand breaks generated by structural rearrangements may provide opportunities for HBV integration[11]. Here we describe the reciprocal relationship where viral insertions in cyclin genes lead to structural rearrangement formation due to replication stress.

The propensity of signature RS1 breakpoints to occur in enhancer-rich regions makes these rearrangements likely to activate oncogenes in trans. In this limited series of 22 CCN-HCC analyzed by WGS, we identified a single significantly recurrent hotspot at *TERT* promoter. However, the power to identify recurrent somatic rearrangement hotspots increases sharply with sample size[27], and future studies of larger CCN-HCC series may uncover additional sites under positive selection in CCN-HCC.

In conclusion, viral insertions and structural rearrangements activating *CCNA2* and *CCNE1* define a homogeneous subgroup of aggressive HCC developed in non-cirrhotic liver, sharing similar transcriptional profiles and frequent inactivation of *RB1*

and *PTEN*. These tumors display a specific rearrangement signature induced by replication stress that sustains tumor growth by activating *TERT* but may constitute a targetable vulnerability.

Methods

Description of the LICA-FR cohort. A series of 160 hepatocellular carcinoma (HCC) samples and their non-tumor counterparts were collected from patients surgically treated in four French hospitals located in Bordeaux and Paris region. The study was approved by institutional review board committees (CCPRB Paris Saint-Louis, 1997, 2004, and 2010, approval number 01–037; Bordeaux, 2010-A00498–31). Written informed consent was obtained in accordance with French legislation. All samples were immediately frozen in liquid nitrogen and stored at −80 °C. HCC were enriched in cases developed on a non-cirrhotic liver (107/160, 67%): 75 tumors developed in non-fibrotic (METAVIR F0-F1), 32 in chronic hepatitis (F2–F3) and 53 in cirrhotic liver (F4). Clinicopathological data were available for all cases. A diversity of risk factors were represented in our series, including alcohol ($n = 63$), metabolic syndrome ($n = 37$), HBV ($n = 30$), and HCV infection ($n = 30$). Twenty-nine patients had none of the above risk factors. These 160 samples were analyzed by RNA sequencing, 156 were analyzed by whole exome sequencing (including 96 were previously published[10]) and 45 by whole genome sequencing (35 were previously published[22]). Detailed clinical characteristics and sequencing details for each sample are provided in Supplementary Data 1.

Whole genome sequencing. Whole genome data from 45 tumors of the LICA-FR series were analyzed in this study, comprising 35 previously published[22] and 10 new cases. The whole genomes of 10 new tumor/normal pairs were sequenced for this project at the Center National de Recherche en Génomique Humaine (CNRGH, Evry, France) on an Illumina HiSeq X Five as paired-end 151 bp reads. Sequences were aligned to the hg19 version of the human genome using BWA[52] version 0.7.12. We used Picard tools version 1.108 (http://broadinstitute.github.io/picard/) to remove PCR duplicates and GATK[53] version v3.5 for local indel realignment and base quality recalibration, as recommended in GATK best practices[54]. We obtained an average depth of 119-fold for tumors (range 104–126) and 41-fold for matched non-tumor liver samples (range 38–43).

Whole exome sequencing. Whole exome data from 156 tumors of the LICA-FR series were analyzed in this study, comprising 96 previously published[10] and 60 new cases. Sequence capture, enrichment and elution of genomic DNA samples from the 60 new tumor/normal pairs was performed by IntegraGen (Evry, France). Agilent in-solution enrichment was used with the manufacturer's biotinylated oligonucleotide probe library SureSelect Human All-Exon kit v5 + UTRs ($n = 39$) or SureSelect Clinical Research Exome V2 ($n = 21$) according to the manufacturer's instructions. The eluted enriched DNA sample was sequenced on an Illumina HiSeq 2000 ($n = 39$) or HiSeq 4000 ($n = 21$) as paired-end 75 bp reads. Sequencing details for each sample are indicated in Supplementary Data 1.

Somatic mutation calling. We used MuTect2 to call somatic mutations from WES and WGS data by comparing each tumor sample with its matched non-tumor counterpart and a panel of normals (PON) file. We excluded mutations belonging to the ENCODE Data Analysis Consortium blacklisted regions (http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDacMapabilityConsensusExcludable.bed.gz) and regions covered by < 6 reads in the tumor or normal sample. We then selected only single nucleotide variants (SNVs) with a MuTect2 flag among "PASS", "clustered_events", "t_lod_fstar", "alt_allele_in_normal" or "homologous_mapping_event" and small insertions and deletions (indels) with a MuTect2 flag among "PASS", "clustered_events" or "str_contraction". To improve specificity in the calling of mutations with low variant allele frequency (VAF), we quantified the number of high quality variant reads in the tumor (mapping quality ≥ 20, base quality ≥ 20) and the number of variant reads in the non-tumor sample with no quality threshold using bamreadcount (https://github.com/genome/bam-readcount). Only variants matching the following criteria were finally retained: VAF ≥ 2% in the tumor with ≥ 3 variant reads, VAF ≤ 5% in the non-tumor samples with ≤ 2 variant reads, and a VAF ratio ≥ 5 between the tumor and non-tumor sample.

Copy-number and structural rearrangement analysis. We used MANTA[55] software to identify somatic structural rearrangements in WGS data. To keep only the most reliable events, we selected only rearrangements supported by ≥ 10 reads and with a variant allele fraction ≥ 5%. We used cgpBattenberg[56] algorithm to reconstruct copy-number profiles from WGS data. We used the circular binary segmentation algorithm implemented in the Bioconductor package DNAcopy[57] to reconstruct copy-number profiles from WES data.

RNA sequencing. RNA samples from the 160 tumors of the LICA-FR series were sequenced in several batches with slightly different protocols. RNA samples were enriched for polyadenylated RNA from 5 µg of total RNA, and the enriched samples were used to generate sequencing libraries with the Illumina TruSeq or Illumina TruSeq Stranded mRNA kit and associated protocol as provided by the manufacturer. Libraries were sequenced by IntegraGen (Evry, France) on an Illumina HiSeq 2000 or 4000 as paired-end 75 or 100 bp reads. Full Fastq files were aligned to the reference human genome hg19 using TopHat2[58]. Sequencing details for each sample and the parameters used for TopHat2 are indicated in Supplementary Data 1. We removed reads mapping to multiple locations, and we used HTSeq[59] to obtain the number of reads associated to each gene in the Gencode v19 database, restricting to protein-coding genes, pseudogenes, antisense and lincRNAs ($n = 42540$). We used the Bioconductor DESeq2 package[60] to import raw HTSeq counts for each sample into R statistical software and apply variance stabilizing transformation (VST) to the raw count matrix. FPKM scores (number of fragments per kilobase of exon model and millions of mapped reads) were calculated by normalizing the count matrix for the library size and the coding length of each gene. We used the area under the ROC curve (AUC) to identify and remove 2724 genes with a significant batch effect (AUC > 0.95 between one sequencing project and others).

Gene fusion detection. Fusions detected by TopHat2 (--fusion-search --fusion-min-dist 2000 --fusion-anchor-length 13 --fusion-ignore-chromosomes chrM) were filtered using the TopHatFusion-post algorithm. We kept only fusions validated by BLAST and with at least 10 split-reads or pairs of reads spanning the fusion event, and we removed fusions identified at least twice in a cohort of 36 normal liver samples.

Gene expression analysis. We used t-distributed stochastic neighbor embedding (t-SNE) to classify HCC based on their gene expression profiles. We selected the 1000 most variably expressed genes, and we used 1 minus the weighted Pearson correlation coefficient as the distance measure. Pairwise Pearson correlation was calculated using the wtd.cors function of the *weights* R package. We used standard deviation substracted by 0.2 as the weight, giving more variable genes greater influence. The resulting distance matrix was used to perform the t-SNE analysis using the R package *Rtsne*[61] with default parameters except the following: theta = 0, is_distance = T, pca = F, max_iter = 2000. We used the Bioconductor *limma* package[62] to test for differential expression between CCN-HCC and other HCC of all genes expressed in at least five samples (FPKM > 0). We applied a *q*-value threshold of ≤ 0.05 to define differentially expressed genes. We used an in-house adaptation of the GSEA method[63] to identify gene sets from the MSigDB v6 database overrepresented among upregulated and downregulated genes.

Viral insertion screening. AAV2 insertions had previously been screened by viral capture and whole exome sequencing in 83 tumors from the LICA-FR cohort[16]. We extended this screen to AAV2 and HBV insertions in all HCC from the LICA-FR cohort using RNA-seq and WES data. In the ICGC-JP cohort, AAV2 and HBV insertions had already been screened using WGS data and were provided by Fujimoto et al.[11] In the TCGA cohort, we screened AAV2 and HBV insertions using RNA-seq data from all tumors and WES data from 37 tumors showing viral reads or overexpression of *CCNA2* or *CCNE1* in RNA-seq data. For each tumor and matched normal sample, the sequence reads were mapped to the AAV2 (AF043303.1) and HBV (X02763, renumbered using the EcoR1 restriction site as the +1) reference genomes using BWA[52]. Read pairs with at least one read aligned on the virus were extracted using samtools[64], and aligned to a custom reference genome including human chromosomes and virus fasta sequences as pseudo-chromosomes. Tumors with ≥ 6 chimeric reads or read pairs aligned on both the human and viral genomes were further analyzed. All viral insertions were validated by visual inspection on IGV[65]. We used chimeric reads to identify insertion breakpoints at base resolution by mapping sequences on both sides of the junctions. Of the 12 LICA-FR tumors with viral insertions detected in *CCNA2* or *CCNE1*, 7 were previously analyzed by viral capture sequencing[16] and 3 were analyzed by whole genome sequencing. For these 10 tumors, we were able to extract reads covering the full length of the inserted viral genome and to reconstruct the complete human-virus-human chimeric sequence.

Consequences of cyclin A2 alterations on protein structure. All tumors from the LICA-FR series harboring AAV2 or HBV insertions in *CCNA2* were analyzed by WGS or viral capture[16] to determine the precise boundaries of viral insertion breakpoints. RNA-seq reads were then aligned on the reconstructed chimeric sequence with TopHat2[58], and we used Cufflinks v2.2.1[66] to identify and quantify the different transcripts. We used ElemeNT[67] to predict transcription initiation sites and Alamut Visual software (Interactive Biosoftware) to identify splicing signals on the chimeric DNA sequence. We used ATGpr[68] to identify translation initiation sites on abnormal transcripts resulting from viral insertion or gene fusions.

Western blot analysis of cyclin A2 and cyclin E1 proteins. Cell protein extracts were prepared using hot Laemmli buffer (50 mM Tris, pH = 6.8, 2% SDS, 5% glycerol, 2 mM DTT, 2.5 mM EDTA, 2.5 mM EGTA, Protease inhibitor cocktail complete MINI EDTA-free (Roche Applied Science), 1× HALT Phosphatase inhibitor (Perbio), 2 mM Na3VO4 and 10 mM NaF). Protein concentration was assessed using the BCA Protein Assay Kit (Pierce). Western blot analyses were

conducted using the following primary antibodies: CCNA2 N-ter (#211735, Abcam); CCNA2 C-ter (#32386, Abcam), CCNE1 (#33911, Abcam), and β-actin (#4967, Cell Signaling Technology) used as loading control. Proteins of interest were detected using an anti-rabbit IgG horseradish peroxidase–linked secondary antibody (#7074, Cell Signaling Technology) and the ECL Chemiluminescence Western Blotting Detection Kit (GE Healthcare), according to the provided protocol. Signal detection was performed using the ChemiDoc XRS system and the Image Lab software (Bio-Rad). All antibodies were used at 1:1000 dilution except secondary antibody, which was used at 1:2000.

Mutational and rearrangement signature analysis. We used the *Palimpsest* R package[69] to extract mutational and rearrangement signatures from WGS data. For point mutations, we quantified the contribution of the 10 mutational signatures referenced on the COSMIC website (https://cancer.sanger.ac.uk/cosmic/signatures) and described as operative in liver cancers (signatures 1, 4, 5, 6, 12, 16, 17, 22, 23, 24)[22] to each tumor genome. For structural rearrangements, we performed a de novo signature analysis across the 350 HCC genomes from the LICA-FR, TCGA and ICGC-JP datasets. We identified 6 rearrangement signatures that were very similar to the 6 signatures we previously obtained on a smaller dataset[22], except that the two initially described deletion signatures were now merged into signature RS5, and that a new signature emerged (RS6, dominated by inversions < 10 kb). We used *Palimpsest* to quantify the contribution of each signature to each tumor genome and to estimate the probability of each structural rearrangement being due to each process.

Identification of rearrangement hotspots. We identified 8466 breakpoints attributed to signature RS1 (probability > 0.5) across the 350 HCC genomes from the LICA-FR, TCGA and ICGC-JP datasets. To account for the uneven distribution of rearrangements in the genome, we then modeled the background distribution of breakpoints considering various genomic features as described by Glodzik et al.[27], with some modifications. In short, we divided the genome into 500 kb bins, and we characterized for each bin 17 genomic features likely to influence the density of rearrangements: replication timing in HepG2 cell line (ENCODE[70]), highly expressed (top 25%) and low-expressed (remaining 75%) genes in normal liver, average copy-number in the cohort, repetitive sequences (segmental duplications, ALU elements and other repeats), number of N bases in the reference genome, known fragile sites[71], chromatin staining, DNAse hyper-sensitive sites and 6 histone marks (H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27ac) in adult liver (ROADMAP[72]). All features were normalized to a mean of 0 and standard deviation of 1 across the bins. The total number of RS1 breakpoints were counted for each bin, and we used negative binomial regression to model the distribution of breakpoints according to the 17 normalized features. The model was trained across 4993 bins after removing bins containing validated cancer genes from the Cancer Gene Census[73] (https://cancer.sanger.ac.uk/census). For signature RS1, the most predictive features of a high breakpoint density were DNAse accessibility, H3K27 acetylation and early replication timing. We then used this model to estimate the expected number of breakpoints across 761 bins containing cancer genes, and we compared the number of observed breakpoints to the number of expected breakpoints using a one-sided binomial test. Finally, p-values were corrected for multiple testing using Benjamini-Hochberg procedure.

Chromatin state analysis. We used various genomic features to correlate with structural rearrangement density and to better understand the functional consequences of rearrangements. We used replication sequencing (Repli-seq) wavelet-smoothed signals downloaded generated by the ENCODE[70] consortium for the liver cancer cell line HepG2 to define early and late-replicating regions. We used ChIP-seq data for various histone modifications (H3K4me1, H3K4me3, H3K9me3, H3K27me3, H3K36me3, H3K27ac) and chromatin states derived from these modifications in normal adult liver by the ROADMAP consortium[72]. Topologically associated domain (TAD) boundaries in human embryonic stem cells (H1) were provided by Tsirigos et al.[74]

Pan-cancer analysis of structural rearrangement signatures. Somatic structural rearrangements called by a uniform pipeline over 2,606 tumor genomes were downloaded from the ICGC PanCancer Analysis of Whole Genomes (PCAWG) project[23,28,29]. Using *Palimpsest*[69], we identified 9 rearrangement signatures in this data set, including one (RS1-pancan) very similar to the RS1 signature identified in CCN-HCC, and we quantified the contribution of each signature to each tumor genome. In each cancer type, we tested if the presence of ≥ 50 rearrangements attributed to signature RS1-pancan was associated with the presence of rearrangement breakpoints < 80 kb from *CCNA2* or *CCNE1* gene using Fisher's exact test. We analyzed two additional series of breast ($n = 524$)[30] and ovarian ($n = 80$)[75] cancer genomes to correlate the amount of RS1-pancan events with *CCNE1* amplifications and *BRCA1* alterations.

Clinical associations. We tested the association of CCN-HCC in the LICA-FR cohort with gender, age, etiology, liver fibrosis, Edmonson grade, and vascular invasion using Wilcoxon rank sum test for continuous variables, Fisher's exact test for binary variables and Chi square test for trend for categorical variables. We used log-rank test and Kaplan–Meier method to compare overall survival between CCN-HCC and others, considering only HCC with curative resection (R0) and excluding patients who died within 3 months after surgery.

Computing codes. The functions used to perform the signatures analysis and associated figures are available as an open-source R package, Palimpsest, available on Github: https://github.com/FunGeST/Palimpsest.

URLs. ICGC data portal, https://dcc.icgc.org/; COSMIC database, https://cancer.sanger.ac.uk/cosmic; ENCODE project, https://www.encodeproject.org; GENCODE v19, http://www.gencodegenes.org/releases/19.html; ROADMAP project, http://www.roadmapepigenomics.org; NCI GDC data portal, https://portal.gdc.cancer.gov.

Data availability

The sequencing data reported in this paper have been deposited to the EGA (European Genome-phenome Archive) database (RNA-seq accession [EGAS00001002879]; WES accessions [EGAS00001000217], [EGAS00001001002] and [EGAS00001003063]; WGS accessions [EGAS00001002408], [EGAS00001000706] and [EGAS00001002888]) and the Inter- national Cancer Genome Consortium (ICGC) data portal (http://dcc.icgc.org/; release 27, April 2018).

Received: 23 July 2018 Accepted: 8 November 2018
Published online: 07 December 2018

References

1. European Association for the Study of the Liver & European Organisation for Research and Treatment of Cancer EASL-EORTC clinical practice guidelines: management of hepatocellular carcinoma. *J. Hepatol.* **56**, 908–943 (2012).
2. Llovet, J. M. et al. Sorafenib in advanced hepatocellular carcinoma. *New Engl. J. Med.* **359**, 378–390 (2008).
3. Bruix, J. et al. Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. *Lancet* **389**, 56–66 (2017).
4. Llovet, J. M. & Hernandez-Gea, V. Hepatocellular carcinoma: reasons for phase III failure and novel perspectives on trial design. *Clin. Cancer Res.* **20**, 2072–2079 (2014).
5. Llovet, J. M. et al. Hepatocellular carcinoma. *Nat. Rev. Dis. Prim.* **2**, 16018 (2016).
6. Nault, J. C. et al. Telomerase reverse transcriptase promoter mutation is an early somatic genetic alteration in the transformation of premalignant nodules in hepatocellular carcinoma on cirrhosis. *Hepatology* **60**, 1983–1992 (2014).
7. Guichard, C. et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. *Nat. Genet.* **44**, 694–698 (2012).
8. Fujimoto, A. et al. Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. *Nat. Genet.* **44**, 760–764 (2012).
9. Ahn, S.-M. et al. Genomic portrait of resectable hepatocellular carcinomas: Implications of RB1 and FGF19 aberrations for patient stratification. *Hepatology* **60**, 1972–1982 (2014).
10. Schulze, K. et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. *Nat. Genet.* **47**, 505–511 (2015).
11. Fujimoto, A. et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. *Nat. Genet.* **48**, 500–509 (2016).
12. Zucman-Rossi, J., Villanueva, A., Nault, J.-C. & Llovet, J. M. Genetic landscape and biomarkers of hepatocellular carcinoma. *Gastroenterology* **149**, 1226–1239.e4 (2015).
13. Wang, J., Chenivesse, X., Henglein, B. & Bréchot, C. Hepatitis B virus integration in a cyclin A gene in a hepatocellular carcinoma. *Nature* **343**, 555–557 (1990).
14. Sung, W.-K. et al. Genome-wide survey of recurrent HBV integration in hepatocellular carcinoma. *Nat. Genet.* **44**, 765–769 (2012).
15. Ding, D. et al. Recurrent targeted genes of hepatitis B virus in the liver cancer genomes identified by a next-generation sequencing-based approach. *PLoS Genet.* **8**, e1003065 (2012).
16. Nault, J.-C. et al. Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. *Nat. Genet.* **47**, 1187–1193 (2015).
17. Cancer Genome Atlas Research Network. Electronic address: wheeler@bcm.edu & Cancer Genome Atlas Research Network. Comprehensive and

integrative genomic characterization of hepatocellular carcinoma. *Cell* **169**, 1327–1341.e23 (2017).
18. Geley, S. et al. Anaphase-promoting complex/cyclosome-dependent proteolysis of human cyclin A starts at the beginning of mitosis and is not subject to the spindle assembly checkpoint. *J. Cell. Biol.* **153**, 137–148 (2001).
19. Fung, T. K., Yam, C. H. & Poon, R. Y. C. The N-terminal regulatory domain of cyclin A contains redundant ubiquitination targeting sequences and acceptor sites. *Cell Cycle* **4**, 1411–1420 (2005).
20. Sage, J., Miller, A. L., Pérez-Mancera, P. A., Wysocki, J. M. & Jacks, T. Acute mutation of retinoblastoma gene function is sufficient for cell cycle re-entry. *Nature* **424**, 223–228 (2003).
21. Garcia-Cao, I. et al. Systemic elevation of PTEN induces a tumor suppressive metabolic state. *Cell* **149**, 49–62 (2012).
22. Letouzé, E. et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. *Nat. Commun.* **8**, 1315 (2017).
23. Li, Y. et al. Patterns of structural variation in human cancer. Preprint at https://www.biorxiv.org/content/early/2017/08/27/181339, https://doi.org/10.1101/181339 (2017).
24. Lee, J. A., Carvalho, C. M. B. & Lupski, J. R. A. DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell* **131**, 1235–1247 (2007).
25. Hastings, P. J., Ira, G. & Lupski, J. R. A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* **5**, e1000327 (2009).
26. Carvalho, C. M. B. et al. Inverted genomic segments and complex triplication rearrangements are mediated by inverted repeats in the human genome. *Nat. Genet.* **43**, 1074–1081 (2011).
27. Glodzik, D. et al. A somatic-mutational process recurrently duplicates germline susceptibility loci and tissue-specific super-enhancers in breast cancers. *Nat. Genet.* **49**, 341–348 (2017).
28. Campbell, P. J., Getz, G., Stuart, J. M., Korbel, J. O. & Stein, L. D. Pan-cancer analysis of whole genomes. Preprint at https://www.biorxiv.org/content/early/2017/07/12/162784, https://doi.org/10.1101/162784 (2017).
29. Wala, J. A. et al. Selective and mechanistic sources of recurrent rearrangements across the cancer genome. Preprint at https://www.biorxiv.org/content/early/2017/09/14/187609, https://doi.org/10.1101/187609 (2017).
30. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature* **534**, 47–54 (2016).
31. Waszak, S. M. et al. Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. Preprint at https://www.biorxiv.org/content/early/2017/11/01/208330, https://doi.org/10.1101/208330 (2017).
32. Macheret, M. & Halazonetis, T. D. Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress. *Nature* **555**, 112–116 (2018).
33. Schlacher, K., Wu, H. & Jasin, M. A distinct replication fork protection pathway connects Fanconi anemia tumor suppressors to RAD51-BRCA1/2. *Cancer Cell* **22**, 106–116 (2012).
34. Pathania, S. et al. BRCA1 haploinsufficiency for replication stress suppression in primary cells. *Nat. Commun.* **5**, 5496 (2014).
35. Willis, N. A. et al. Mechanism of tandem duplication formation in BRCA1-mutant cells. *Nature* **551**, 590–595 (2017).
36. Forment, J. V. & O'Connor, M. J. Targeting the replication stress response in cancer. *Pharmacol. Ther.* https://doi.org/10.1016/j.pharmthera.2018.03.005 (2018).
37. Menghi, F. et al. The tandem duplicator phenotype as a distinct genomic configuration in cancer. *Proc. Natl Acad. Sci. USA* **113**, E2373–E2382 (2016).
38. Lord, C. J. & Ashworth, A. PARP inhibitors: synthetic lethality in the clinic. *Science* **355**, 1152–1158 (2017).
39. O'Connor, M. J. Targeting the DNA damage response in cancer. *Mol. Cell* **60**, 547–560 (2015).
40. Toledo, L. I. et al. A cell-based screen identifies ATR inhibitors with synthetic lethal properties for cancer-associated mutations. *Nat. Struct. Mol. Biol.* **18**, 721–727 (2011).
41. Honeyman, J. N. et al. Detection of a recurrent DNAJB1-PRKACA chimeric transcript in fibrolamellar hepatocellular carcinoma. *Science* **343**, 1010–1014 (2014).
42. Gao, Q. et al. Driver fusions and their implications in the development and treatment of human cancers. *Cell Rep.* **23**, 227–238.e3 (2018).
43. Conner, E. A. et al. Dual functions of E2F-1 in a transgenic mouse model of liver carcinogenesis. *Oncogene* **19**, 5054–5062 (2000).
44. Ladu, S. et al. E2F1 inhibits c-Myc-driven apoptosis via PIK3CA/Akt/mTOR and COX-2 in a mouse model of human liver cancer. *Gastroenterology* **135**, 1322–1332 (2008).
45. Kramara, J., Osia, B. & Malkova, A. Break-induced replication: the where, the why, and the how. *Trends Genet.* **34**, 518–531 (2018).
46. Costantino, L. et al. Break-induced replication repair of damaged forks induces genomic duplications in human cells. *Science* **343**, 88–91 (2014).
47. Smith, C. E., Llorente, B. & Symington, L. S. Template switching during break-induced replication. *Nature* **447**, 102–105 (2007).
48. Saini, N. et al. Migrating bubble during break-induced replication drives conservative DNA synthesis. *Nature* **502**, 389–392 (2013).
49. Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609–615 (2011).
50. Etemadmoghadam, D. et al. Synthetic lethality between CCNE1 amplification and loss of BRCA1. *Proc. Natl Acad. Sci. USA* **110**, 19489–19494 (2013).
51. Sanchez-Vega, F. et al. Oncogenic signaling pathways in The Cancer Genome Atlas. *Cell* **173**, 321–337.e10 (2018).
52. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
53. McKenna, A. et al. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
54. DePristo, M. A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
55. Chen, X. et al. Manta: rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics* **32**, 1220–1222 (2016).
56. Nik-Zainal, S. et al. The life history of 21 breast cancers. *Cell* **149**, 994–1007 (2012).
57. Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
58. Kim, D. et al. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* **14**, R36 (2013).
59. Anders, S., Pyl, P. T. & Huber, W. HTSeq—a Python framework to work with high-throughput sequencing data. *Bioinformatics* **31**, 166–169 (2015).
60. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550 (2014).
61. van der Maaten, L. & Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
62. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
63. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA* **102**, 15545–15550 (2005).
64. Li, H. et al. The sequence alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
65. Robinson, J. T. et al. Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
66. Trapnell, C. et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat. Protoc.* **7**, 562–578 (2012).
67. Sloutskin, A. et al. ElemeNT: a computational tool for detecting core promoter elements. *Transcription* **6**, 41–50 (2015).
68. Nishikawa, T., Ota, T. & Isogai, T. Prediction whether a human cDNA sequence contains initiation codon by combining statistical information and similarity with protein sequences. *Bioinformatics* **16**, 960–967 (2000).
69. Shinde, J. et al. Palimpsest: an R package for studying mutational and structural variant signatures along clonal evolution in cancer. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/bty388 (2018).
70. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
71. Bignell, G. R. et al. Signatures of mutation and selection in the cancer genome. *Nature* **463**, 893–898 (2010).
72. Roadmap Epigenomics Consortium. et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
73. Futreal, P. A. et al. A census of human cancer genes. *Nat. Rev. Cancer* **4**, 177–183 (2004).
74. Gong, Y. et al. Stratification of TAD boundaries reveals preferential insulation of super-enhancers by strong boundaries. *Nat. Commun.* **9**, 542 (2018).
75. Patch, A.-M. et al. Whole-genome characterization of chemoresistant ovarian cancer. *Nature* **521**, 489–494 (2015).

Acknowledgements

We thank Hidewaki Nakagawa for fruitful discussions and providing *TERT* promoter mutation data for the ICGC-JP series. We thank Rameen Beroukhim and Joachim Weischenfeldt for helping us access ICGC structural variant tables, and Aristotelis Tsirigos for providing topologically associated domain boundaries. We thank Tatiana Popova and Céline Vallot for critical discussion of the results. We thank the principal investigators of the liver cancer TCGA (Lewis Roberts, David Wheeler) and ICGC-JP (Tatsuhiro Shibata, Hidewaki Nakagawa) projects, and the ICGC consortium as a whole for providing the high quality data sets used in this study. We thank all the clinician surgeons and pathologists who have participated to this work: Jean Saric, Christophe Laurent, Laurence Chiche, Brigitte Le Bail, Claire Castain (CHU Bordeaux), Alexis Laurent, Daniel Cherqui, Daniel Azoulay (CHU Henri Mondor, Créteil,

APHP), Marianne Ziol, Nathalie Ganne-Carrié and Pierre Nahon (Jean Verdier Hospital, Bondy, APHP). We also thank the Réseau national CRB Foie (BB-0033-0085), the tumor banks of CHU Bordeaux (BB-0033-00036), Jean Verdier Hospital (APHP) and CHU Henri Mondor (APHP) for contributing to the tissue collection. This work was supported by INCa within the ICGC project, MUTHEC project (INCa translationnel PRTK2014), France Génomique, Cancéropole Ile de France (ExhauTrans project), ITMO Cancer AVIESAN (Alliance Nationale pour les Sciences de la Vie et de la Santé, National Alliance for Life Sciences & Health) within the framework of the Cancer Plan ("HTE program-HetColi network" and "Cancer et environnement program"), BPI France (ICE project), ANRS and the French Liver Biobanks network – INCa, BB-0033-00085, Hepatobio bank. The group is supported by the Ligue Nationale Contre le Cancer (Equipe Labellisée), Labex OncoImmunology (investissement d'avenir), Coup d'Elan de la Fondation Bettencourt-Shueller, the SIRIC CARPEM and Fondation Mérieux. QB and LM are supported by a fellowship from the HOB doctoral school and the ministry of Education and Research, TLB is supported by an "Attractivite IDEX" fellowship from IUH and CP is supported by a doctoral fellowship funded by ANRS.

Author contributions

J.Z-R. and E.L. conceived and directed the research. Q.B., L.M., C.P., S.I., J.Z-R., and E.L. designed the study and wrote the manuscript. C.P., I.M., G.C., and T.L-B. performed the experiments. Q.B., L.M., C.P., I.M., T.L-B., S.I., J.Z-R., and E.L. analyzed and interpreted the data. D.B., V.M., and J-F.D. generated whole-genome sequencing data. Q.B., L.M., C. P., V.R., J.S., E.T., D.B., V.M. S.I., and E.L. performed bioinformatics and statistical analysis. J-C.N., G.A., A.D-V, P.B-S., O.S., J-F.B., and J.C. provided essential biological resources and collected clinical data. All authors approved the final manuscript and contributed to critical revisions to its intellectual context.

Additional information

Supplementary Information accompanies this paper at https://doi.org/10.1038/s41467-018-07552-9.

Competing interests: The authors declare no competing interests.

Reprints and permission information is available online at http://npg.nature.com/reprintsandpermissions/

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

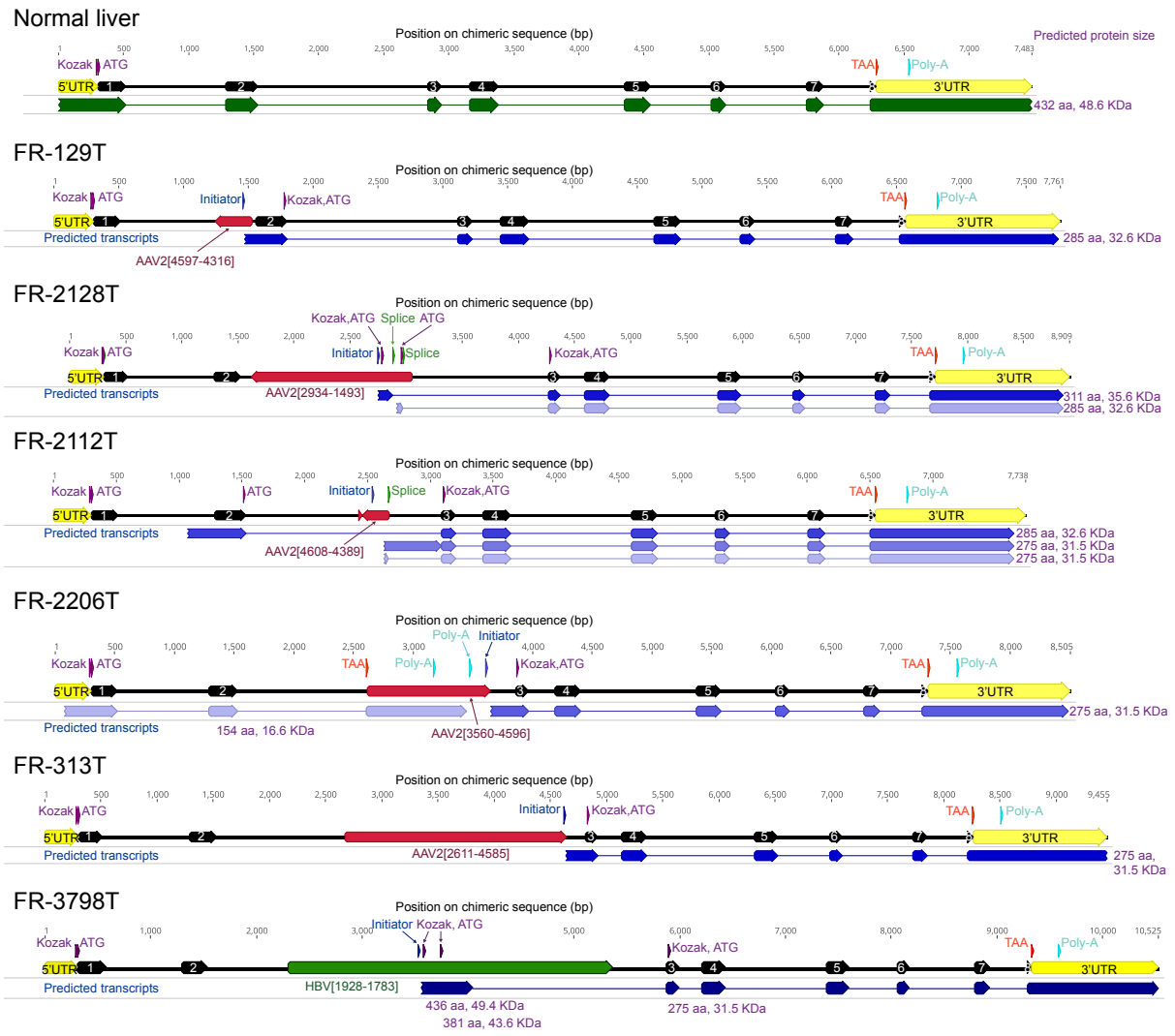
Supplementary Information

Cyclin A2 and E1 genomic alterations define a specific subclass of hepatocellular carcinomas

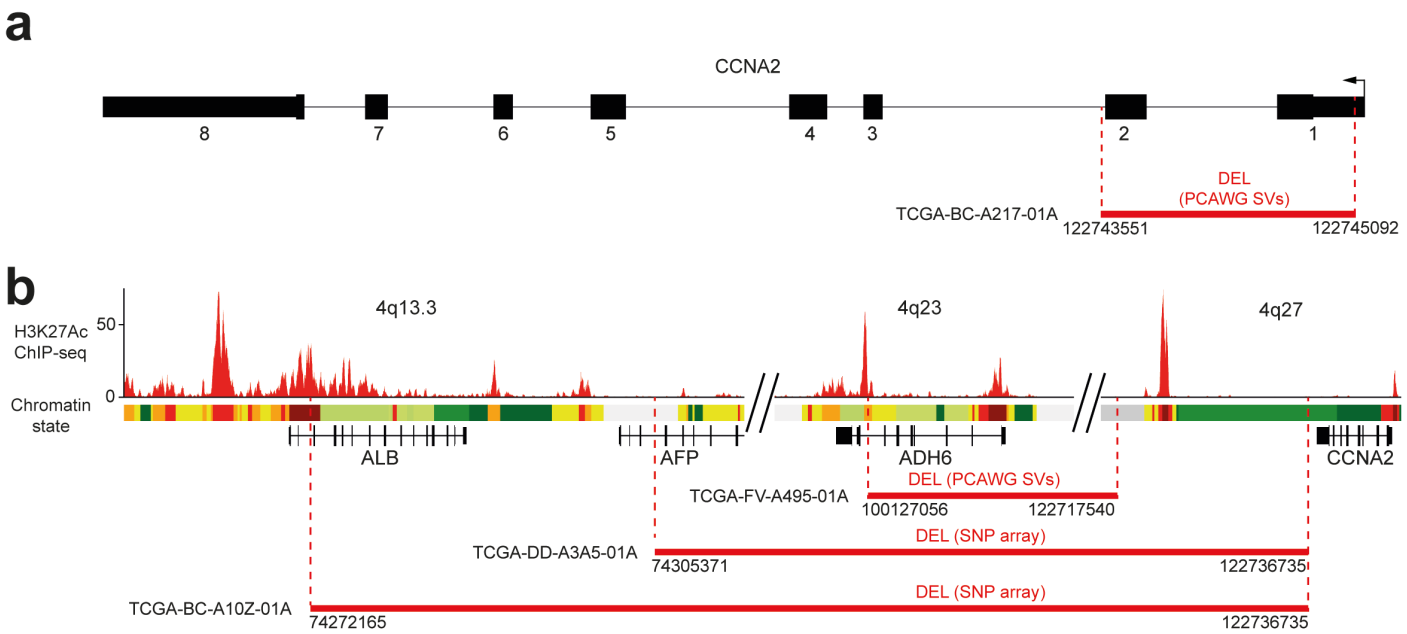
Supplementary Figures

Supplementary Tables

Supplementary Figure 1. Consequences of viral insertions in *CCNA2* gene
Five AAV2 and 1 HBV insertions were identified in *CCNA2* in the LICA-FR series. Precise insertion boundaries were identified by WGS or viral capture, and RNA-seq reads were aligned on the chimeric sequence. Here, the different transcripts predicted by Cufflinks are represented for each case, ordered by transcript abundance. Predicted functional elements (Transcription initiator, splice and poly-A sites, Kozak sequence, initiator and terminator codons) are annotated on the chimeric DNA sequence, and the predicted protein sizes resulting from the translation of each transcript are annotated on the right. Only the most abundant abnormal transcripts were represented in **Fig. 1c**.



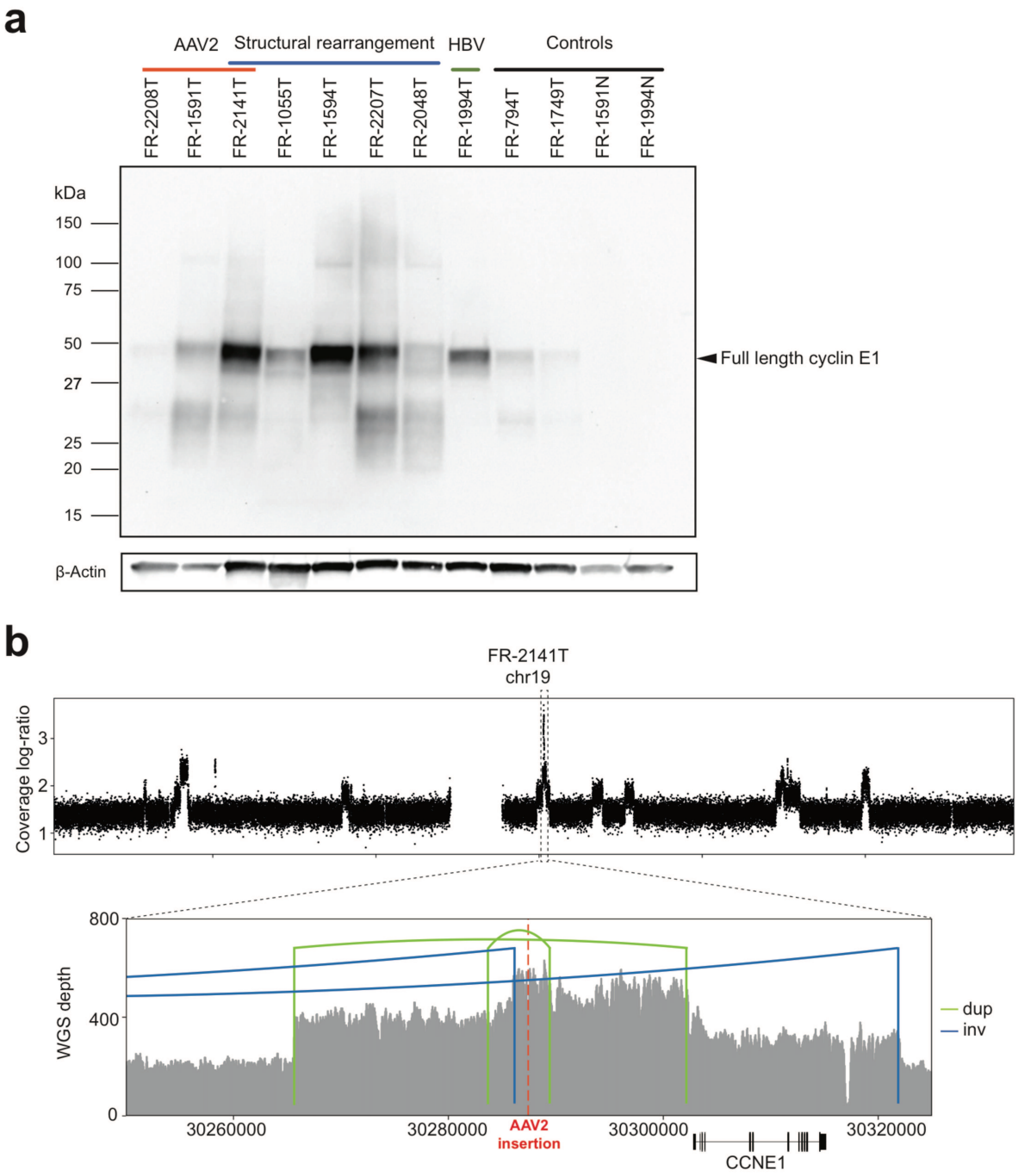
Supplementary Figure 2. Deletions associated with *CCNA2* deregulation
a Focal deletion of *CCNA2* exons 1 & 2 identified in TCGA tumor TCGA-BC-A217.
b Deletions identified in 3 TCGA tumors linking *CCNA2* downstream region with the highly expressed genes *ALB*, *AFP* and *ADH6*.



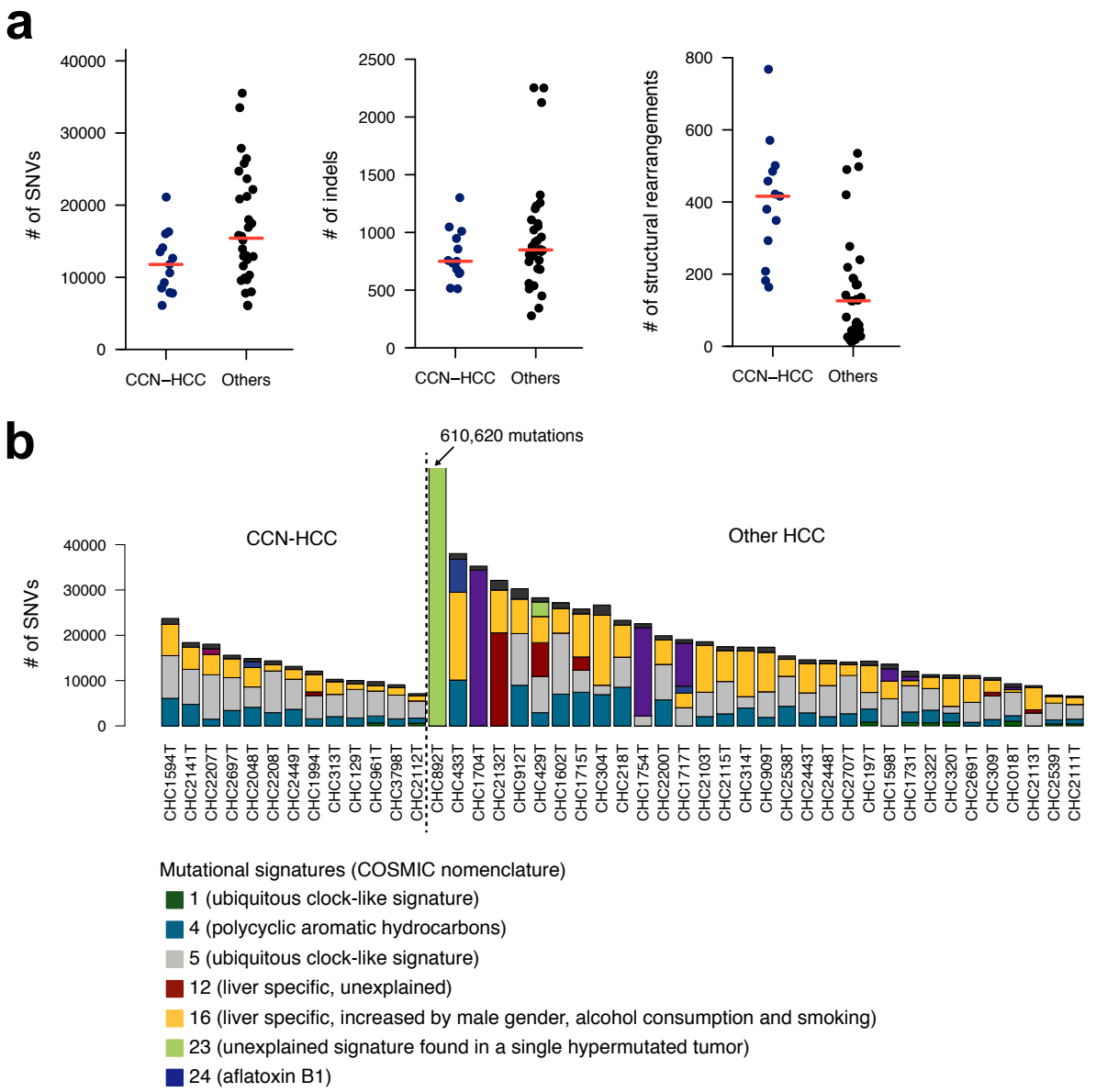
Supplementary Figure 3. Cyclin E1 overexpression induced by viral insertions and structural rearrangements

a Western blot analysis of cyclin E1. Tumors with viral insertions or structural rearrangements are compared with tumors without *CCNE1* alteration and non-tumoral liver controls.

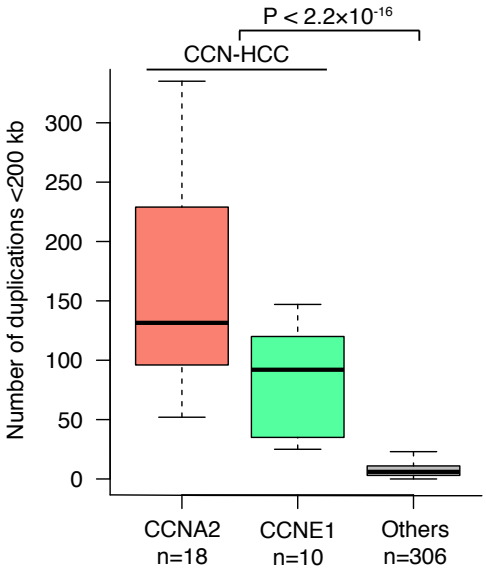
b Tumor FR-2141T displays both an AAV2 insertion in *CCNE1* regulatory region and a high-level amplification of the locus. The top panel displays the coverage log-ratio along chromosome 19 in this tumor. The bottom panel displays the coverage of WGS reads aligned to the chimeric sequence of *CCNE1* locus including AAV2 insertion together with structural rearrangement breakpoints. It shows that the most strongly amplified region includes *CCNE1* regulatory region, and in particular the locus of AAV2 insertion.



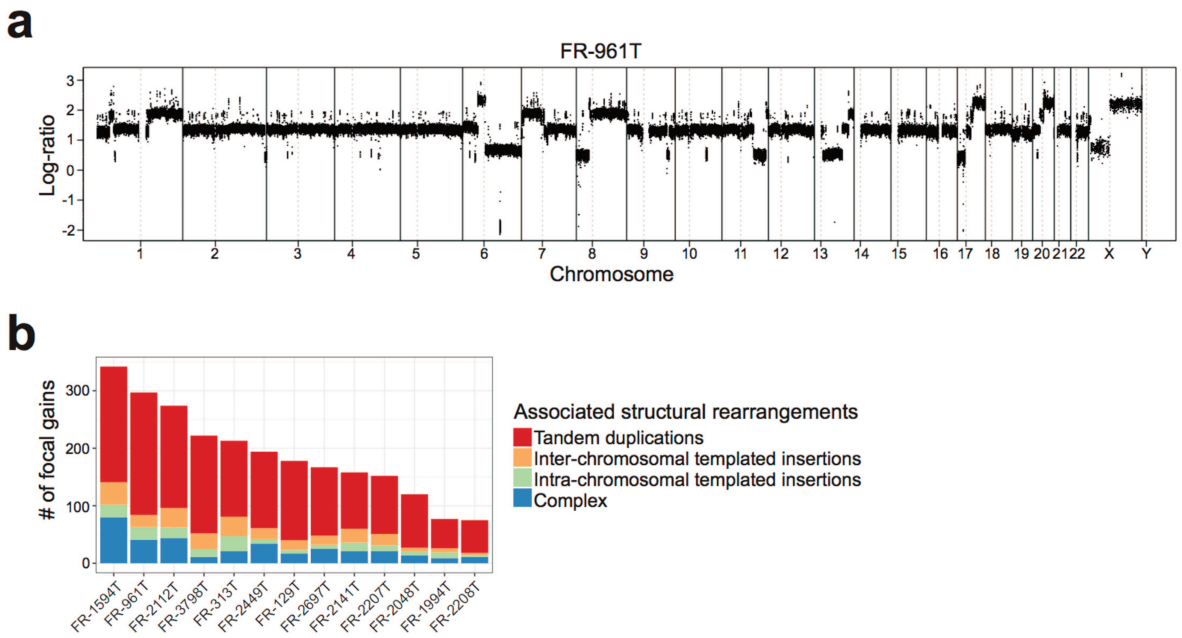
Supplementary Figure 4. Mutational signature analysis of CCN-HCC
a Comparison of the number of single-nucleotide variants (SNVs), indels and structural rearrangements in CCN-HCC vs others (LICA-FR data, 45 WGS).
b Contribution of the different mutational signatures known to be operative in liver cancers to the mutational burden of CCN-HCC and other HCC.



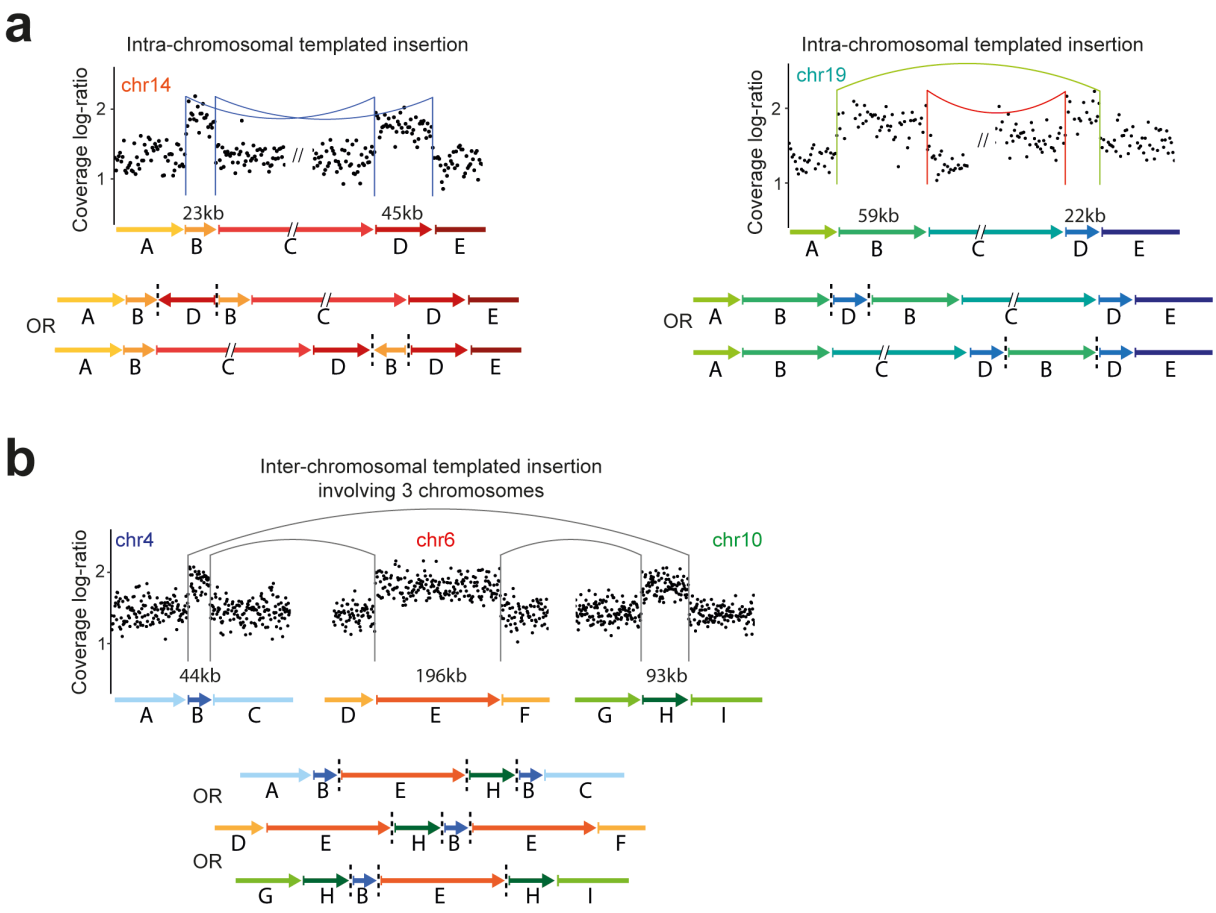
Supplementary Figure 5. SNP array analysis of focal duplications in CCN-HCC from the TCGA series.
Focal deletions (<200 kb) were quantified across 334 tumors from the TCGA series as a surrogate marker of the RS1 signature, and compared between *CCNA2*-activated, *CCNE1*-activated and other HCC.



Supplementary Figure 6. Characteristic copy-number profile of CCN-HCC
a Typical copy-number profile of a CCN-HCC, showing hundreds of focal gains scattered throughout the genome.
b Proportion of focal gains in each CCN-HCC attributed to each rearrangement mechanism.

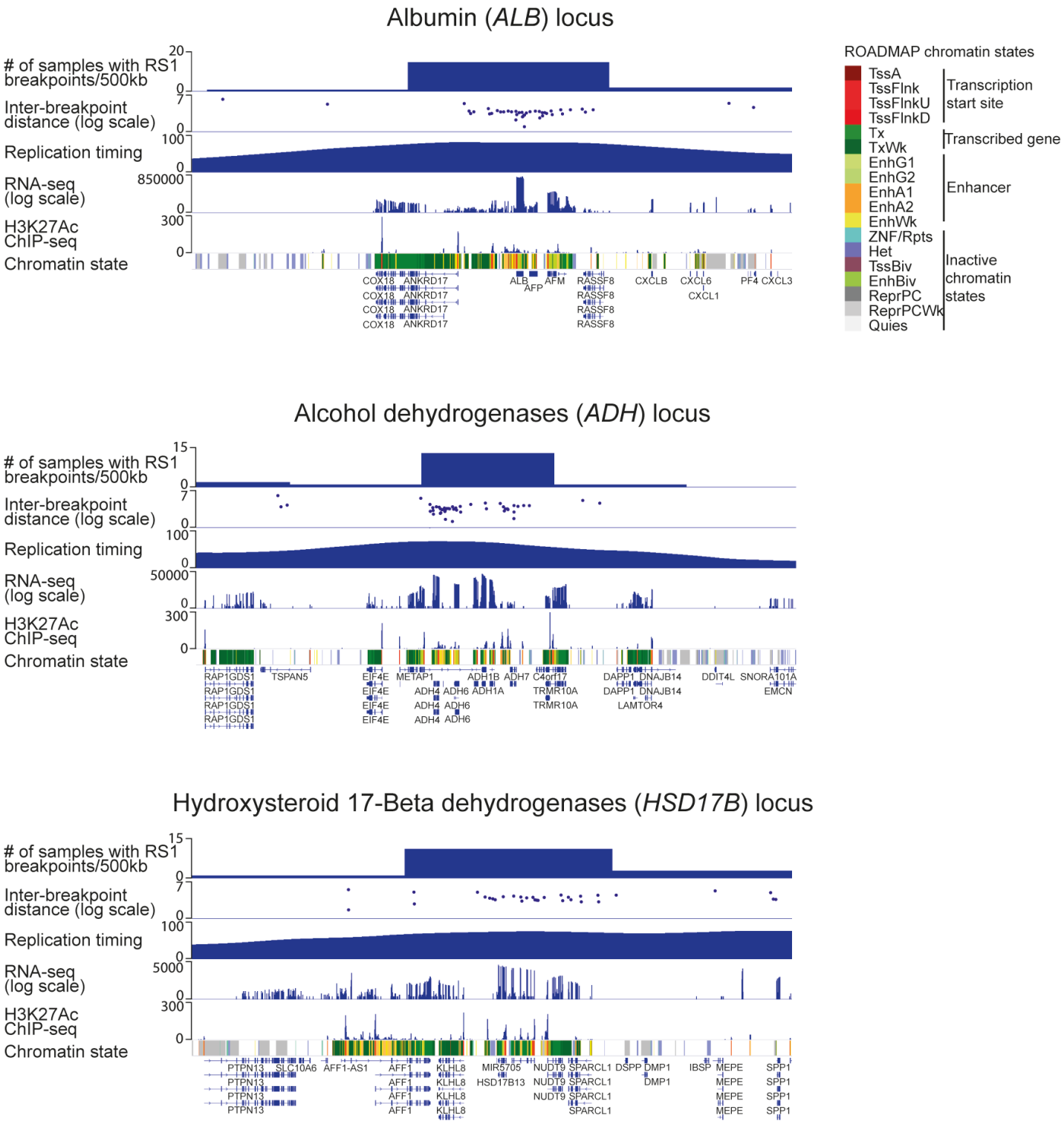


Supplementary Figure 7. Examples of intra-chromosomal templated insertions and inter-chromosomal templated insertion cycles involving several chromosomes
a Intra-chromosomal templated insertions appear as couples of inversions (left) or deletion and duplication (right) depending on the orientation of aberrant junctions.
b Example of inter-chromosomal templated insertion involving 3 different chromosomes.

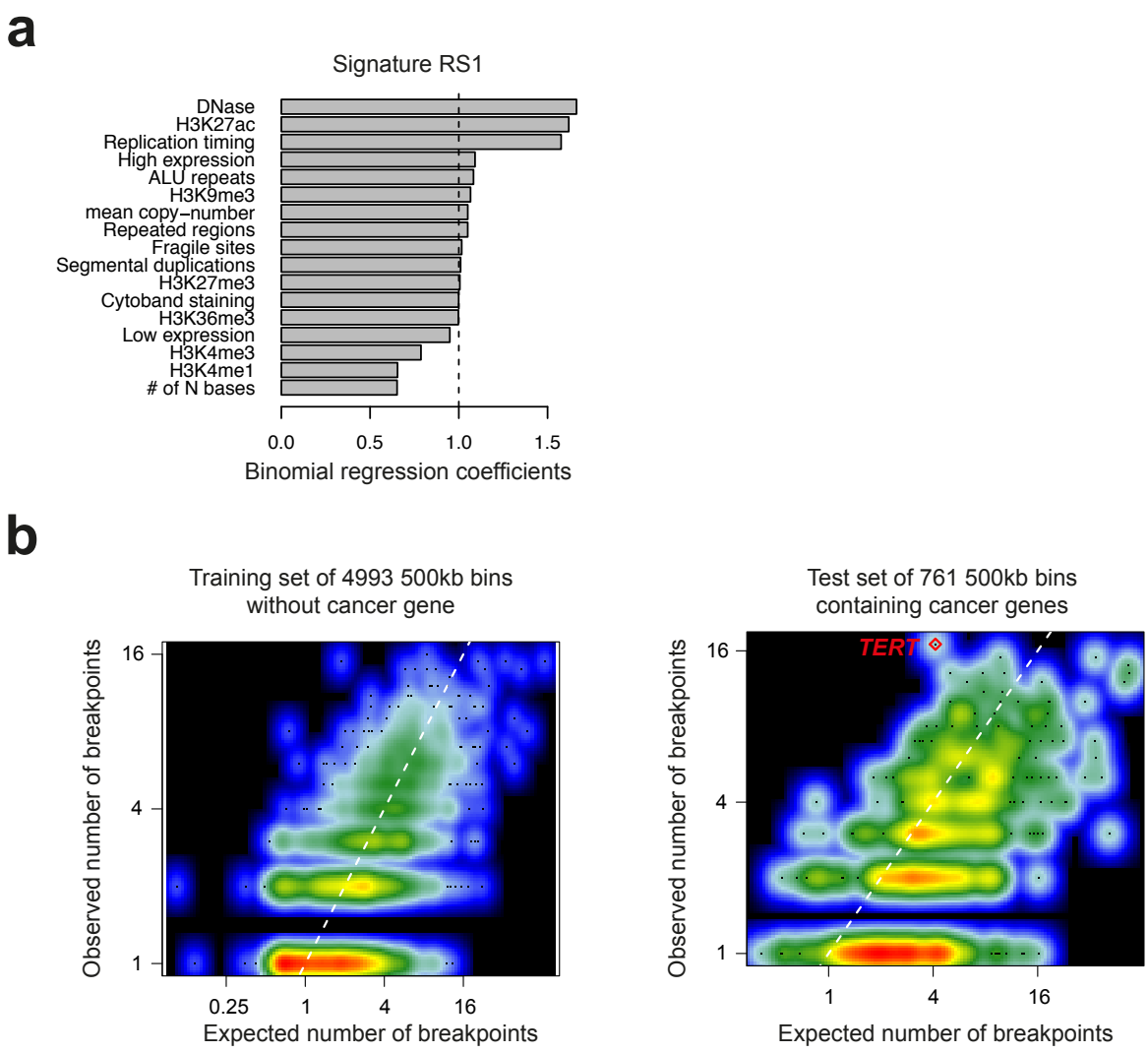


Supplementary Figure 8. RS1 breakpoint hotspots involving highly expressed liver enzymes

The density of RS1 breakpoints, replication timing, RNA-seq expression, H3K27Ac and ROADMAP chromatin states are displayed for 3 representative hotspots involving the very highly expressed liver enzymes albumin, alcohol dehydrogenases and hydroxysteroid 17-Beta dehydrogenases.



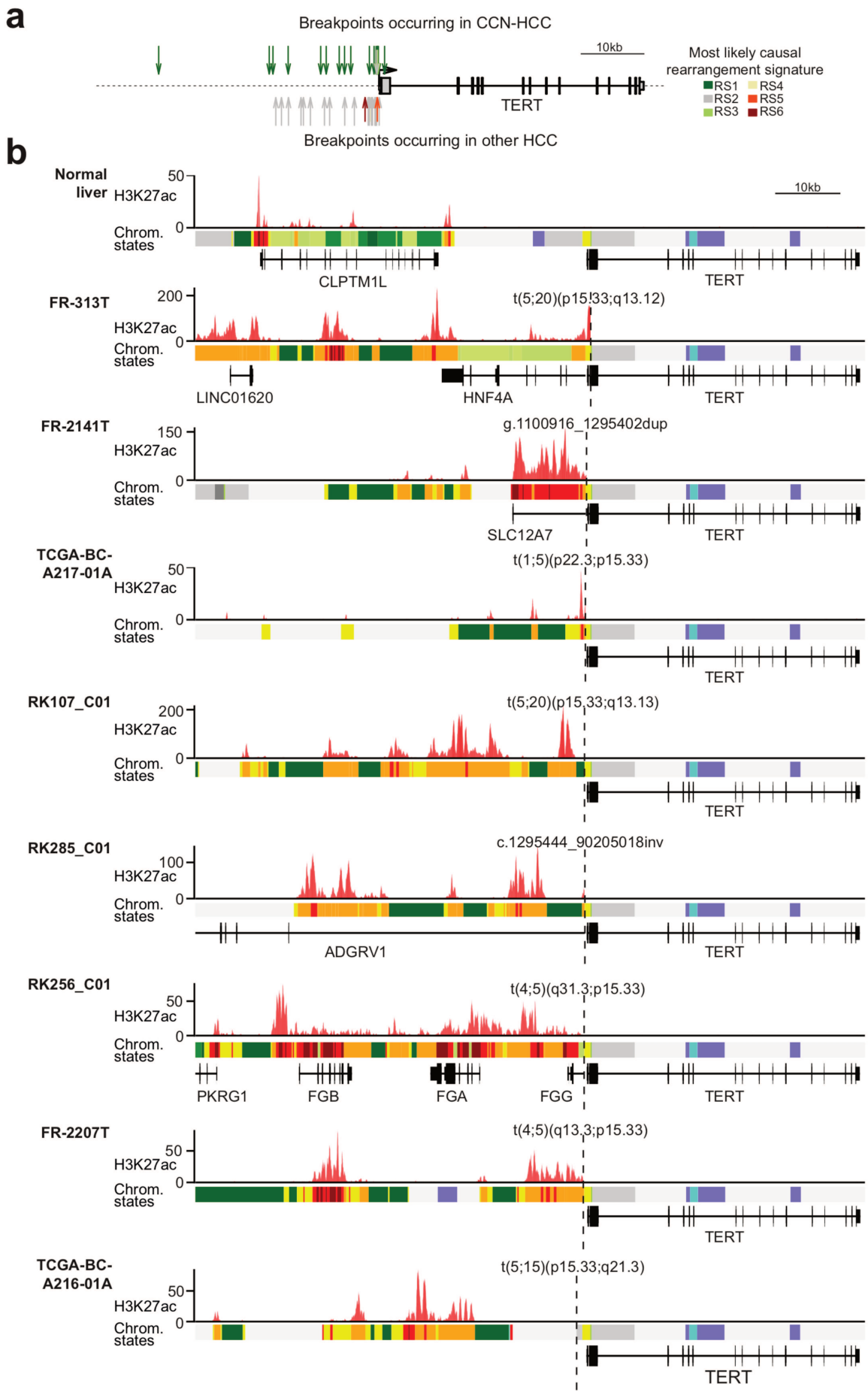
Supplementary Figure 9. Binomial regression modeling of rearrangement breakpoint density
a Regression coefficients of the 17 genomic features used to predict the density of signature RS1 breakpoints.
b Correlation between the number of observed RS1 breakpoints per 500 kb bin and the expected number predicted by the binomial regression model. Left: Within 4933 bins without any cancer gene used as training set. Right: Within 761 bins containing cancer genes (test set). The bin corresponding to *TERT* promoter region is highlighted in red.

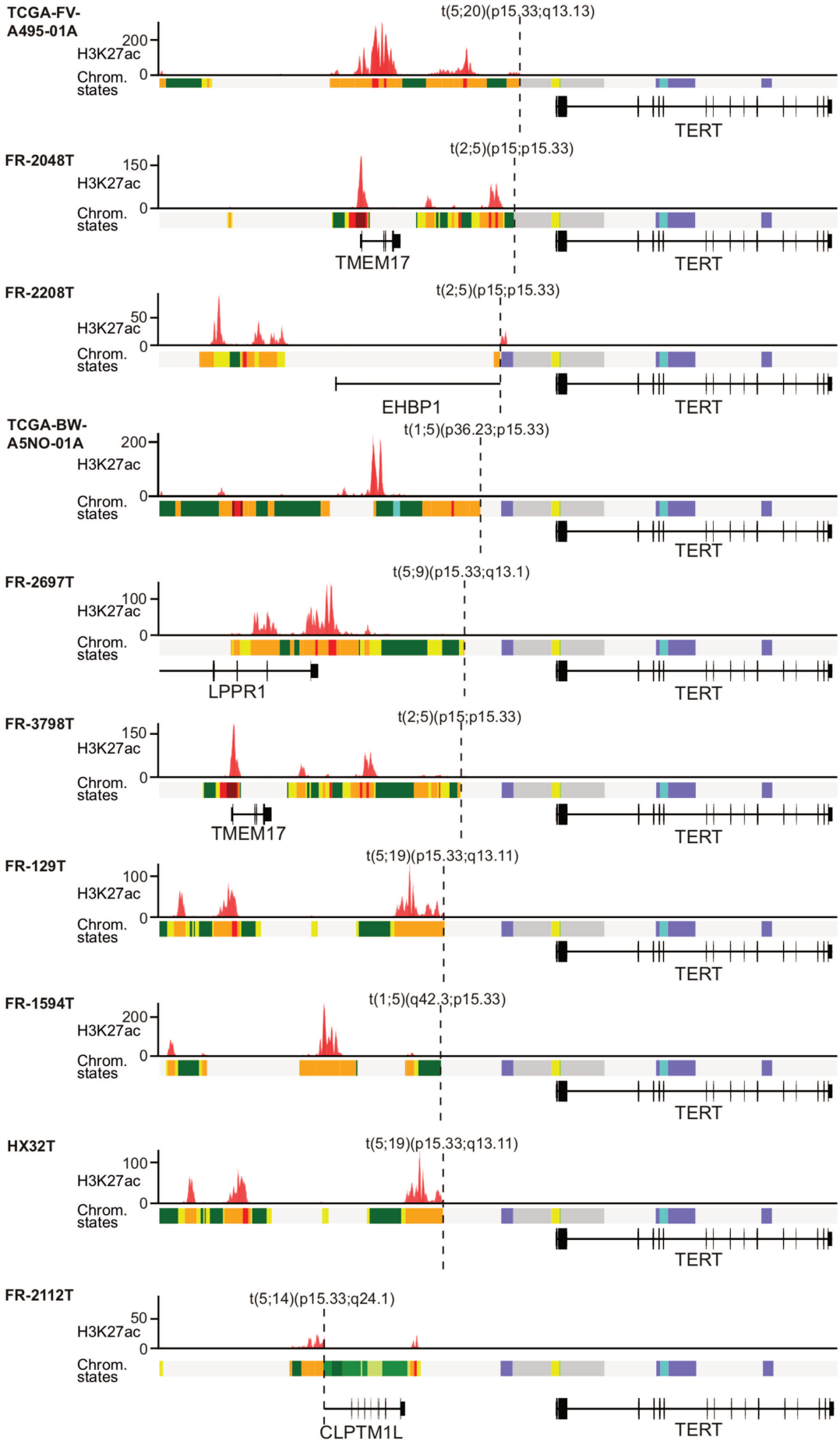


Supplementary Figure 10. Rearrangements affecting *TERT* promoter region in 350 HCC genomes

a Summary of rearrangement breakpoints in *TERT* regulatory region. Each arrow indicates a rearrangement breakpoint. The color indicates the rearrangement signature of the most likely causal process. Breakpoints occurring in CCN-HCC are represented above the scheme. Those occurring in other HCC are represented below.

b Functional consequences of structural rearrangements affecting *TERT* regulatory region in CCN-HCC. Chromatin states and H3K27 acetylation in normal adult liver (top) are compared with predicted chromatin states and H3K27 acetylation resulting from the 18 structural rearrangements of *TERT* regulatory regions identified in CCN-HCC. H3K27Ac chromatin immunoprecipitation sequencing (ChIP-seq) signal were obtained from the ROADMAP consortium. The color code of chromatin states is the same as in **Supplementary Fig. 8**.

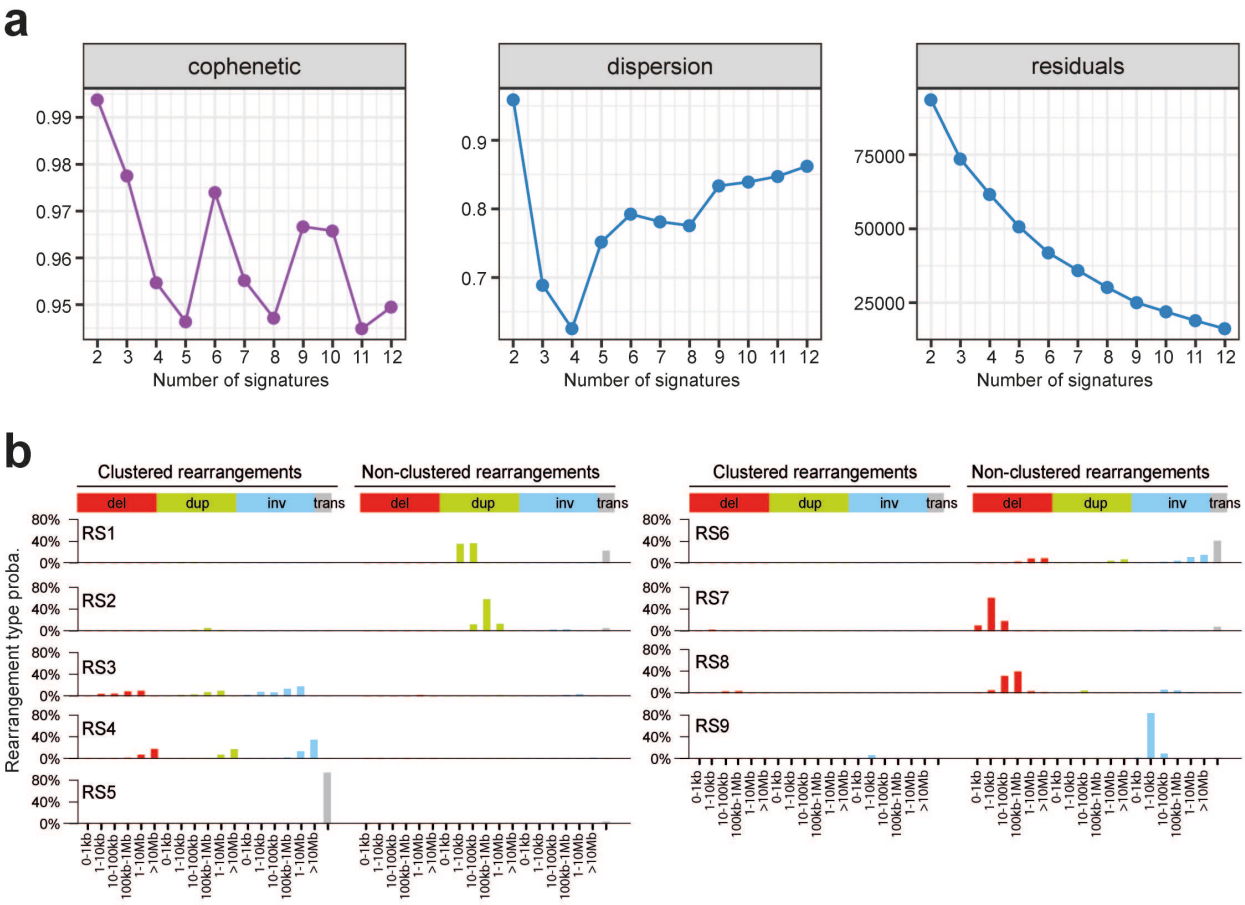

a
Breakpoints occurring in CCN-HCC
10kb
TERT
Most likely causal rearrangement signature
RS1
RS2
RS3
RS4
RS5
RS6
b
Breakpoints occurring in other HCC
Normal liver
H3K27ac
Chrom. states
CLPTM1L
TERT
FR-313T
t(5;20)(p15!33;q13.12)
LINC01620
HNF4A
FR-2141T
g.1100916_1295402dup
SLC12A7
TCGA-BC-A217-01A
t(1;5)(p22.3;p15.33)
RK107_C01
t(5;20)(p15.33;q13.13)
RK285_C01
c.1295444_90205018inv
ADGRV1
RK256_C01
t(4;5)(q31.3;p15.33)
PKRG1
FGB
FGA
FGG
FR-2207T
t(4;5)(q13.3;p15.33)
TCGA-BC-A216-01A
t(5;15)(p15.33;q21.3)


TCGA-FV-A495-01A
t(5;20)(p15.33;q13.13)
200
H3K27ac
0
Chrom. states
TERT
FR-2048T
t(2;5)(p15;p15.33)
150
H3K27ac
0
Chrom. states
TMEM17
TERT
FR-2208T
t(2;5)(p15;p15.33)
50
H3K27ac
0
Chrom. states
EHBP1
TERT
TCGA-BW-A5NO-01A
t(1;5)(p36.23;p15.33)
200
H3K27ac
0
Chrom. states
TERT
FR-2697T
t(5;9)(p15.33;q13.1)
100
H3K27ac
0
Chrom. states
LPPR1
TERT
FR-3798T
t(2;5)(p15;p15.33)
150
H3K27ac
0
Chrom. states
TMEM17
TERT
FR-129T
t(5;19)(p15.33;q13.11)
100
H3K27ac
0
Chrom. states
TERT
FR-1594T
t(1;5)(q42.3;p15.33)
200
H3K27ac
0
Chrom. states
TERT
HX32T
t(5;19)(p15.33;q13.11)
100
H3K27ac
0
Chrom. states
TERT
FR-2112T
t(5;14)(p15.33;q24.1)
50
H3K27ac
0
Chrom. states
CLPTM1L
TERT

Supplementary Figure 11. Rearrangement signatures identified in the pan-cancer ICGC series

a Non-negative matrix factorization (NMF) metrics used to determine the optimal number of signatures. With 9 signatures, we obtain good cophenetic coefficient and dispersion score.

b Frequency of the 38 structural rearrangement categories in the 9 signatures.



2. Signatures épigénétiques des carcinomes hépatocellulaires

Les mécanismes biologiques à l'origine des altérations de méthylation observées dans les tumeurs sont multiples et restent largement méconnus. Les résultats prometteurs obtenus en appliquant l'analyse en composantes indépendantes à l'étude du transcriptome des carcinomes hépatocellulaires m'ont confortée dans l'utilisation de cette approche appliquée au méthylome. Elle vise à mieux comprendre la diversité des facteurs qui influencent la méthylation des CHC. De par leur diversité clinique (âge, facteurs de risque, maladie chronique sous-jacente) et moléculaire (gènes *driver*, groupes moléculaires), les CHC sont un excellent modèle pour étudier la diversité des mécanismes épigénétiques à l'œuvre dans les tumeurs, et en particulier les effets des altérations d'un gène driver sur la méthylation des gènes. On sait par exemple que 32 % des CHC présentent une mutation dans un régulateur épigénétique et 28 % dans les complexes de remodelage de la chromatine (principalement *ARID1A* et *ARID2*) (Zucman-Rossi et al., 2015) mais les conséquences fonctionnelles de ces altérations restent mal comprises. L'utilisation de méthodes de déconvolution pourrait nous en apprendre davantage sur les conséquences de ces mutations.

2.1. Développement d'un outil pour extraire et interpréter les composantes de méthylation

Au cours de ma thèse, j'ai développé une série d'outils bioinformatiques pour appliquer l'analyse en composantes indépendantes aux données de méthylation, et interpréter la signification biologique des composantes identifiées. Bien que mis au point sur des données de puce méthylation (Illumina 450k), ces outils sont tout à fait applicables à des données de séquençage (whole genome bisulfite sequencing ou reduced representation bisulfite sequencing), et/ou à l'analyse de fenêtres génomiques plutôt que de CpG uniques.

L'analyse en composantes indépendantes des données de méthylation suit le même principe que pour les données transcriptomiques. La matrice de départ contient le taux de méthylation (beta-value, de 0 à 1) pour les 200,000 CpG les plus variants dans chacun des échantillons analysés. Elle est décomposée comme le produit de deux matrices, la première contenant

l'activité des composantes dans chaque échantillon, et la deuxième la contribution de la méthylation de chaque CpG dans chacune des composantes (cf. Figure 36).

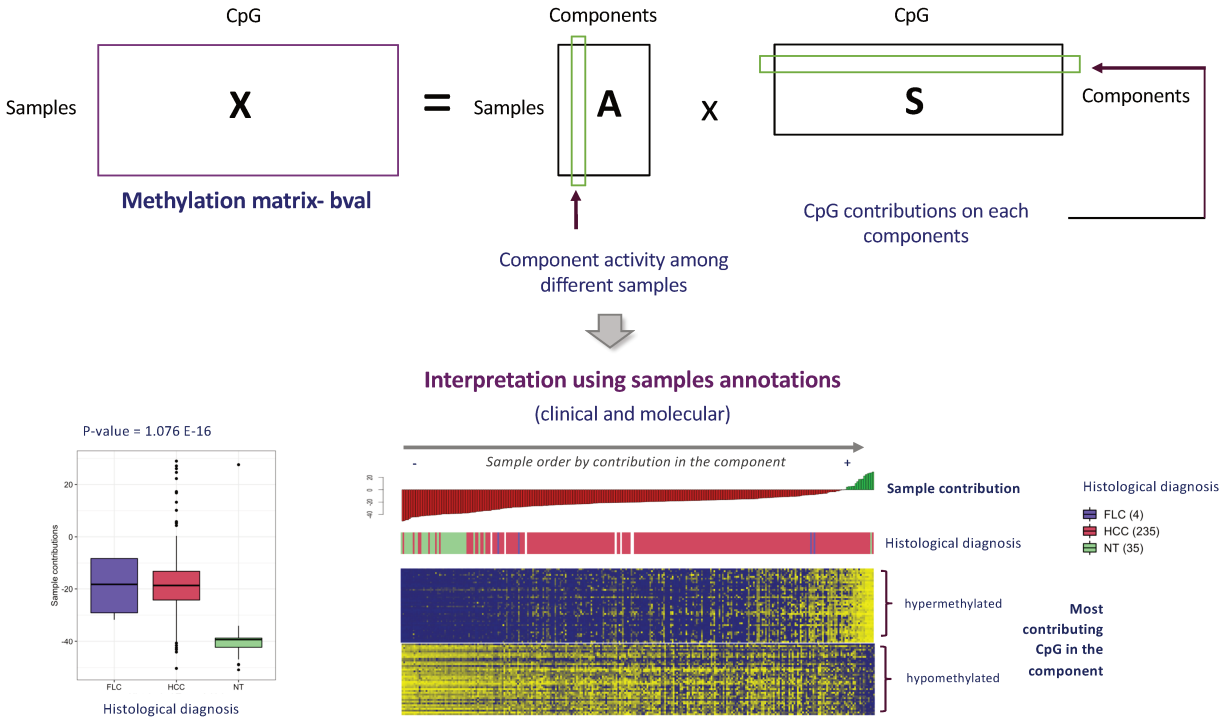


Figure 36 : ***Déconvolution de la matrice de méthylation par analyse en composantes indépendantes, et interprétation de l'activité des composantes dans les échantillons***. *Pour l'interprétation, les associations des composantes aux annotations cliniques et moléculaires sont calculées automatiquement via le test statistique approprié, et représentées sous forme de boxplot pour les variables catégorielles ou de graphe de corrélation pour les variables continues. Un heatmap est également généré représentant la méthylation des CpG les plus contributeurs dans les échantillons, ordonnés par intensité de la composante. Les variables cliniques et moléculaires significativement associées à la composante sont représentées sous forme d'annotations au-dessus du heatmap.*

L'interprétation de chaque composante implique alors l'analyse des échantillons et des CpG les plus contributeurs. L'analyse au niveau des échantillons est identique à celle retenue pour les données transcriptomiques et permet la caractérisation clinique et moléculaire de chaque composante obtenue. Les représentations graphiques retenues pour cette partie reprennent celles mobilisées pour l'ACI sur les données d'expression, utilisant certaines fonctions du package R MineICA (Biton et al.) complétées par celles que j'ai développées au laboratoire, avec notamment des boxplots et heatmaps représentant les échantillons ordonnés par contribution dans la composante avec les associations cliniques et moléculaires significatives (cf. Figure 36).

Pour interpréter la contribution des CpG à chaque composante, j'ai caractérisé en détail le contexte (épi)génomique des CpG les plus contributeurs. A partir des différentes sources disponibles, j'ai regroupé pour chaque CpG de la puce Illumina HumanMethylation 450 les informations suivantes :

a) la localisation par rapport aux gènes (GENCODE v.19 GRCh37/hg19 – Release : Dec. 2013) : CpG dans une région promotrice (TSS +/- 500 pb), dans le corps d'un gène ou intergénique ;
b) la localisation par rapport aux ilôts CpG (UCSC v.19 GRCh37/hg19 – Release : Apr. 2009) : CpG dans un ilôt, shore (2 kb de part et d'autre d'un ilôt), shelf (2kb pb de part et d'autre d'un shore) ou hors ilôt ;
c) l'état chromatinien : chromatine active/inactive et type de domaine chromatinien parmi les 18 définis par le consortium ROADMAP à partir de 6 marques d'histone analysées dans le foie normal (Roadmap Epigenomics Consortium et al., 2015) ;
d) le contexte de méthylation : grands domaines HMD ('Highly Methylated Domain'), PMD ('Partially Methylated Domain'), LMR ('Low Methylation Region') ou UMR ('UnMethylated Region') déterminés par whole genome bisulfite sequencing du foie normal (Salhab et al., 2018). J'ai également classé chaque CpG parmi les 48 catégories de séquences définies par Zhou *et al.* (Zhou et al., 2018) prenant en compte les bases situées directement avant et après le dinucléotide CpG et le nombre de CpG dans les 35 pb en amont et en aval ;
e) le timing de réplication, déterminé sur la lignée cellulaire de foie tumoral HEPG2 ;
f) le lien avec l'expression : CpG dont la méthylation est corrélée à l'expression d'un gène voisin, ce qui veut dire qu'il est situé dans une zone de régulation de la transcription. Les couples CpG-gène sont identifiés grâce au package ELMER évoqué dans l'introduction (Silva et al., 2019), en utilisant les échantillons analysés à la fois en RNAseq et sur puce de méthylation ;
g) la présence de motifs de liaison de facteur de transcription dans la région voisine du CpG.

A partir de ces informations et de la contribution des CpG dans chaque composante, j'ai implémenté différentes visualisations qui permettent de mieux comprendre les caractéristiques de chacune. Elles se basent successivement sur la sélection des CpG les plus

contributeurs à chaque composante, c'est-à-dire ceux avec une contribution inférieure à -0.005 ou supérieure à 0.005 (ce qui correspond dans notre cas à entre 0.6 % et 5.5 % des CpG), puis sur le calcul de l'enrichissement des différentes catégories (épi)génomiques listées ci-dessus associé à cette composante. Cet enrichissement est calculé sur la base d'un score qui correspond au rapport de la proportion de chaque catégorie dans les CpG les plus contributeurs à cette même proportion observée dans l'ensemble des 200,000 CpG analysé (cf. Figure 37).

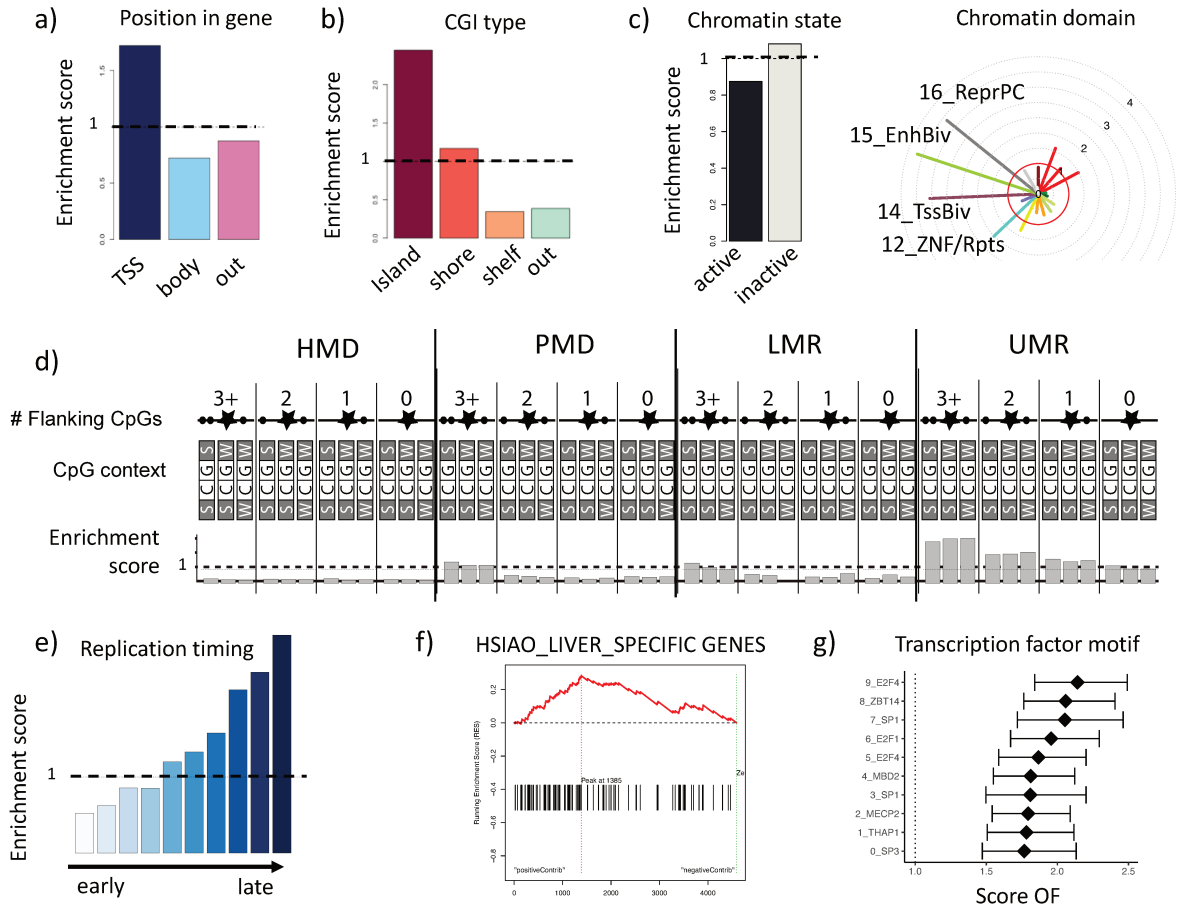


Figure 37 : ***Analyse des catégories (épi)génomiques enrichies parmi les CpG les plus contributeurs de la composante MC2 par rapport à l'ensemble des CpG analysés (200 000 plus variants)***. *a) Position par rapport aux gènes, b) Position par rapport aux ilôts CpG, c) Etats et domaines chromatiniens, d) Domaines de méthylation et contexte nucléotidique, e) Timing de réplication, f) Analyse d'enrichissement des gènes couplés aux CpG, et ordonnés par contribution des CpG dans la composante. Si un gène est couplé à plusieurs CpG, seul celui avec la plus forte contribution en valeur absolue est gardé pour le GSEA, g) Motif de liaison de facteur de transcription à proximité des CpG couplés à des gènes. Le score d'enrichissement de chaque catégorie correspond au ratio de la proportion dans les CpG les plus contributeurs (abs(value)>0.005) sur celle dans l'ensemble des 200,000 CpG analysés.*

Un autre point important pour la caractérisation des composantes est lié au type de changement de méthylation capturé par chacune des composantes. En effet, certaines

composantes sont principalement associées à des mécanismes d'hyper ou d'hypométhylation, d'autres à une combinaison des deux. Pour identifier les altérations de méthylation dominantes de chaque composante, j'ai restreint l'analyse aux Cpg les plus contributeurs et sélectionné les tumeurs présentant le plus fort delta de méthylation par rapport aux échantillons normaux. La visualisation graphique des résultats reprend les valeurs moyennes de méthylation de chaque CpG dans les tumeurs sélectionnées et dans les échantillons normaux (cf. Figure 38). Ces graphiques permettent d'identifier les composantes principalement associées à une hyperméthylation (≥75 % de CpG hyperméthylés dans les tumeurs, par ex. composante MC1), à une hypométhylation (≥75% de CpG hypométhylés dans les tumeurs, par ex. composante MC13) ou à une combinaison des deux (par ex. composante MC6).

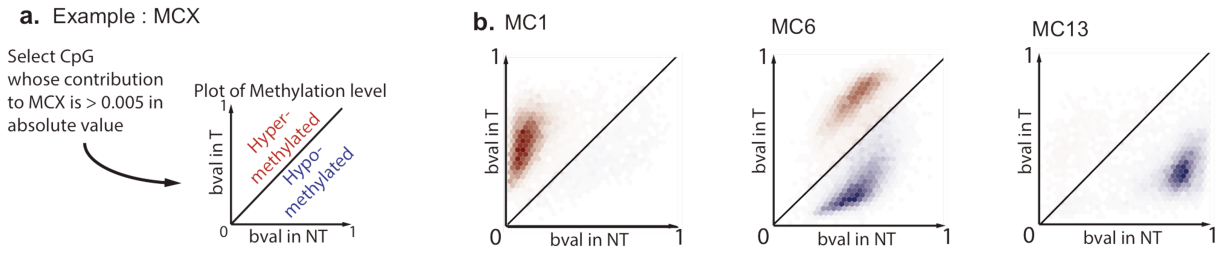


Figure 38 : ***Exemple de graphe permettant de caractériser le type de changement de méthylation.*** *a) Méthode de représentation, b) Exemples de composantes d'hyperméthylation (MC1), mixte (MC6) et d'hypométhylation (MC13).*

Toutes ces analyses d'enrichissement et représentations sont effectuées grâce à des scripts R que j'ai développés pour caractériser les dérégulations de la méthylation des CHC (partie 2.2). Pour mettre ces outils à disposition de la communauté scientifique, je suis actuellement en train de les implémenter dans un package R, *MethICA*, qui sera mis à disposition publiquement sur Github début 2020, au moment de la soumission de l'article.

2.2. Signatures des processus épigénétiques remodelant le méthylome des tumeurs hépatiques

J'ai ensuite utilisé l'analyse en composantes indépendantes pour explorer de manière approfondie les anomalies de méthylation des carcinomes hépatocellulaires. Pour cela, j'ai appliqué les outils bioinformatiques décrits ci-dessus à l'analyse de 3 jeux de données de méthylation des CHC (le jeu du laboratoire, LICA-FR, et les jeux de données publique TCGA-

LIHC et HEPTROMIC), rassemblant au total 738 CHC. Le package MethICA m'a permis de mettre en évidence 13 composantes de méthylation stables dans ces cohortes, correspondant à des mécanismes épigénétiques distinct. J'ai notamment identifié des signatures de processus connus associées au vieillissement et au cancer, mais aussi de nouvelles signatures d'hyper- et d'hypométhylation liées à des événements *driver* spécifiques et à des sous-groupes moléculaires. Ainsi, les altérations activatrices de *CTNNB1* ou des cyclines A2/E1 sont des modulateurs majeurs des profils de méthylation des CHC, associés notamment à des phénotypes hypométhylateurs plus ou moins étendus. Chaque signature cible préférentiellement certains types de CpG. J'ai notamment une signature d'hyperméthylation ciblant les domaines de la chromatine réprimés par Polycomb, ainsi que plusieurs signatures reflètant l'hyper/hypométhylation coordonnée d'*enhancers* suite à la dérégulation de voies de signalisation transcriptionnelles. Nous identifions également pour la première fois une signature de méthylation liée à l'inactivation du remodeleur de la chromatine ARID1A.

Cette approche innovante, basée sur l'analyse en composantes indépendantes, met en lumière la diversité des processus moléculaires qui remodèlent le méthylome du cancer du foie, et permet de définir les signatures de ces processus bien plus précisément que les approches classiques de classification et d'analyse de méthylation différentielle. Ce travail fait l'objet de l'article central de ma thèse, en cours de finalisation, qui sera soumis en février 2020.

Article 3 : "Signatures of epigenetic processes remodeling hepatocellular carcinoma methylomes" (under finalization)

Léa Meunier *et al.*

Signatures of epigenetic processes remodeling hepatocellular carcinoma methylomes

Meunier *et al.*

ABSTRACT

DNA methylation changes are widespread in human cancers, but the underlying molecular mechanisms remain incompletely understood. We developed an innovative statistical framework, MethICA, leveraging independent component analysis to identify sources of DNA methylation changes in tumors. Applied to a compendium of 738 hepatocellular carcinomas (HCC), MethICA unraveled 13 stable methylation components, characterized by their intensities across tumors and CpG sites. These included methylation signatures previously associated with ageing and cancer, but also new hyper- and hypomethylation signatures related to specific driver events and molecular subgroups. A hypermethylation component targeting chromatin domains repressed by Polycomb was specifically encountered in the G1 molecular group with frequent *AXIN1*, *RPS6KA3* and *BAP1* mutations. *CTNNB1* mutations activating ß-catenin were major modulators of methylation patterns in HCC, associated with a widespread hypomethylation of late-replicated partially methylated domains (PMD). By contrast, early-replicated highly methylated domains (HMD) were hypomethylated only in HCC with activation of cyclin A2/E1 and a strong replication stress. Several signatures reflected the coordinated hyper/hypomethylation of enhancers following the deregulation of transcriptional networks. *CTNNB1* mutations induced a hypomethylation of TCF7-bound enhancers in the vicinity of Wnt target genes. By contrast, *ARID1A* inactivation was associated with epigenetic silencing of differentiation-promoting transcriptional networks. This study sheds light on the diversity of molecular processes remodeling liver cancer methylomes, and provides a statistical framework to unravel the signatures of these processes.

INTRODUCTION

Aberrant DNA methylation is a hallmark of human cancers [1]. Typical alterations include a global hypomethylation and a local hypermethylation of CpG islands [2]. These changes play an active role in cancer notably by silencing tumor suppressor genes and favoring genome instability [3]. Genome-scale profiling of thousands of cancer methylomes by microarray or sequencing technologies revealed heterogenous patterns across and within cancer types. Most tumor types can be divided in a few DNA methylation-based subgroups showing more or less widespread hypo/hypermethylation changes [4]. In particular, CpG Island Methylator Phenotypes (CIMP), characterized by the coordinated hypermethylation of numerous CpG islands, have been described in several cancers [5–8]. Some DNA methylation signatures could be related to well-defined oncogenic events and epigenetic mechanisms. For example, *IDH1* mutations in glioma and SDH mutations in paraganglioma induce a hypermethylator phenotype due to the accumulation of oncometabolites that inhibit TET demethylases [9,10]. Other DNA methylation signatures with a clear molecular cause include the hypermethylator phenotype of *TET*-mutated acute myeloid

leukemias [11], and the CpG hypomethylator phenotype (CHOP) of *H3F3A* G34-mutated pediatric glioblastoma [12]. However, the molecular etiology of cancer-associated DNA methylation changes remains elusive in most cases. In addition, chromatin regulatory factors have emerged as frequent driver genes in various cancers [13], but have not been linked to precise methylation signatures so far.

A major hurdle to the identification of clear methylation signatures is that many different factors modulate the DNA methylation landscape of cancer cells and are intermingled in the methylome of the final tumor. The tissue and cell of origin determines the baseline methylation landscape [14,15]. Age-related processes operative throughout life induce methylation changes along cell divisions in healthy tissues [16–19]. Risk factors can have drastic effects on the methylome, exemplified by sun-exposed skin [20] and EBV-related gastric cancers [21]. DNA mutations or altered expression of epigenetic regulators can occur at different timings during tumorigenesis. Deregulated oncogenic pathways are often associated with differential methylation of regulatory elements [22]. Finally, solid tumors are a mixture of tumor and stromal cells with their specific methylation landscapes [23]. Thus, the DNA methylation profile of a tumor reflects the addition of many processes, operative at different strengths and during different lengths in tumor history. The classical approach to analyze cancer-associated methylation signatures has been to perform unsupervised classifications of tumor samples based on their DNA methylation profiles, and to characterize differentially methylated regions between tumor subgroups [4]. While this approach allows to identify the main sources of variation in a data set, it is not designed to capture more subtle signatures that may be operative in tumors dispersed across several subgroups. By contrast, blind source separation methods are dedicated to the deconvolution of independent signals and have shown promising applications to cancer biology [24]. Non-negative matrix factorization (NMF) is widely used to uncover signatures of mutational processes in cancer genomes [25]. Independent component analysis (ICA) has allowed the identification of biologically meaningful transcriptomic components in several cancers [26,27]. Yet, these methods have not been applied to methylation data so far.

Here, we present the MethICA statistical framework, leveraging ICA to disentangle independent sources of variation in methylation data, and we use it to explore the diversity of processes remodeling the methylomes of hepatocellular carcinomas (HCC). HCC, the 3rd most deadly cancer worldwide, is a heterogeneous disease that usually develops in a context of cirrhosis, related to diverse risk factors like hepatitis B (HBV) or hepatitis C virus (HCV) infection, alcohol intake or metabolic syndrome [28]. HCC are also diverse at the molecular level, with up to 6 distinct transcriptomic subgroups [29–31] and tens of driver genes belonging to 11 major pathways [32,33]. After *TERT* promoter (60% of HCC cases), *TP53* and *CTNNB1* (25-30%), chromatin remodeling is the most frequently altered pathway with recurrent mutations in *ARID1A* (13%) and *ARID2* (7%) genes. This clinical and molecular heterogeneity makes HCC a good model to study the diversity of DNA methylation components.

RESULTS

Independent component analysis of liver cancer methylomes

To unravel the diverse epigenetic processes remodeling liver cancer methylomes, we analyzed 3 independent data sets totaling 738 HCC and 104 non-tumor liver samples, all profiled with Illumina HumanMethylation450 beadchip arrays (**Fig. 1**). The in-house LICA-FR series comprised 239 HCC and 35 non-tumor liver samples with matched whole exome sequencing (WES, n=195) and RNA sequencing (RNA-seq, n=128) data, and extensive clinical annotations (**Supplementary Table 1**). We also analyzed the published TCGA LIHC (325 samples with matched WES and RNA-seq data) [31] and HEPTROMIC (243 samples) [34] data sets. We first performed independent component analysis (ICA) within each cohort to decompose the DNA methylation matrix as a mixture of 20 independent methylation components (**Fig. 1**). Each methylation component presumably captures the effect of one epigenetic process, and is characterized by an activation pattern across samples and across CpG sites. To evaluate the reproducibility of the results, we quantified the correlation of methylation components across the 3 data sets based on the contributions of CpG sites (**Supplementary Fig. 1**). Thirteen components were shared by at least two data sets (Pearson correlation ≥0.50), eleven of which were identified in the 3 data sets. In the following, we try to unravel the biological processes underlying this core set of 13 reproducible liver cancer methylation components.

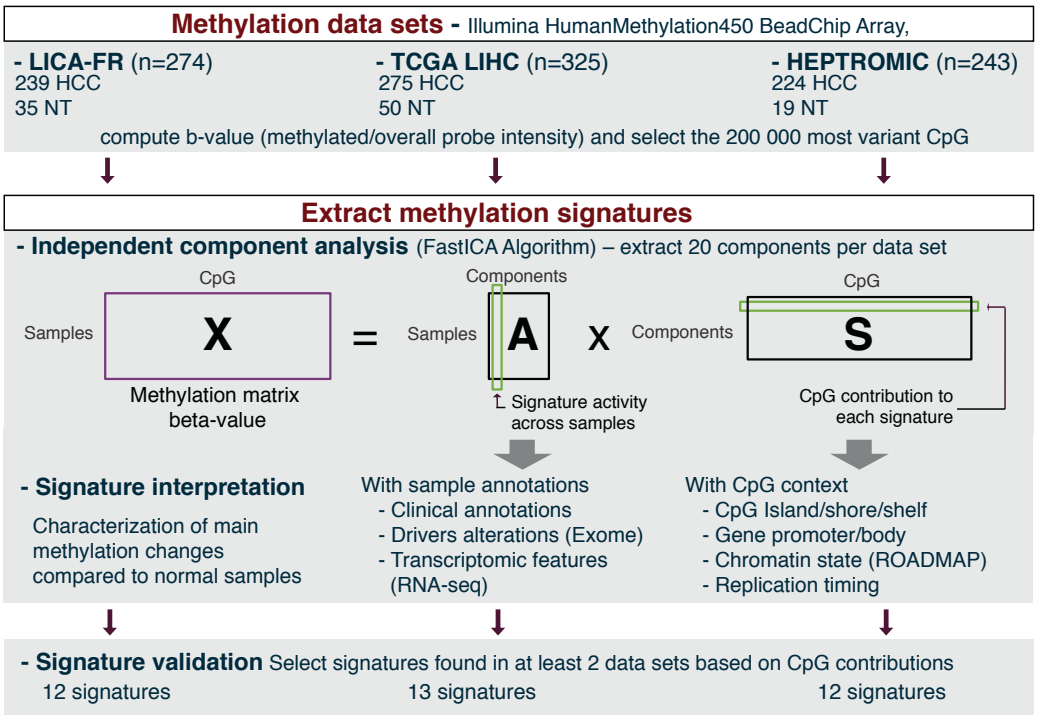


Fig. 1: MethICA analysis workflow. Three independent hepatocellular carcinoma data sets analyzed with the same methylation array are included in this study. Independent component analysis (ICA) is used to decompose the methylation beta value matrix X of dimension n × 200,000 (n samples and 200,000 most variant CpGs) as the product of a matrix A (size n × 20 methylation components, MC) giving the contributions of the samples to each MC (or activities of the MC in the n samples) and a matrix S (size 20 × 200,000) giving the

projections of the CpGs onto each MC. CpGs having the largest projection onto a component (providing the greatest contribution) are the most strongly influenced by the epigenetic process underlying the MC. To unravel the biological meaning of each component, we analyze the clinical and molecular annotations of the most contributing samples, and the (epi)genomic features of the most contributing CpGs.

DNA methylation changes associated with each component

We first characterized the main type of DNA methylation changes (hyper/hypo-methylation) associated with each methylation component (MC). We identified CpG sites with the strongest contribution to each MC, and we examined the DNA methylation changes across these CpG sites in the 5% of tumors with the strongest deviation from non-tumor liver tissues (**Fig. 2**). Components MC1-3 were dominated by hypermethylation, components MC10-13 were dominated by hypomethylation and components MC3-9 showed a combination of hyper- and hypomethylation. The range of methylation changes also varied strongly across components. For example, hypomethylation components MC10 and MC11 involved CpG sites that are highly methylated in normal liver (median beta-value > 0.87), whereas MC12 and MC13 involved CpG sites with intermediate methylation levels (median beta-value ~ 0.7). Hypermethylation components MC1 and MC2 both involved CpG sites that are unmethylated in normal liver (median beta-value = 0.14), but the median methylation gain was of 0.36 in MC1 vs 0.52 in MC2. Thus, each process displays its own dynamics of methylation changes, preferentially occurring at specific CpG sites.

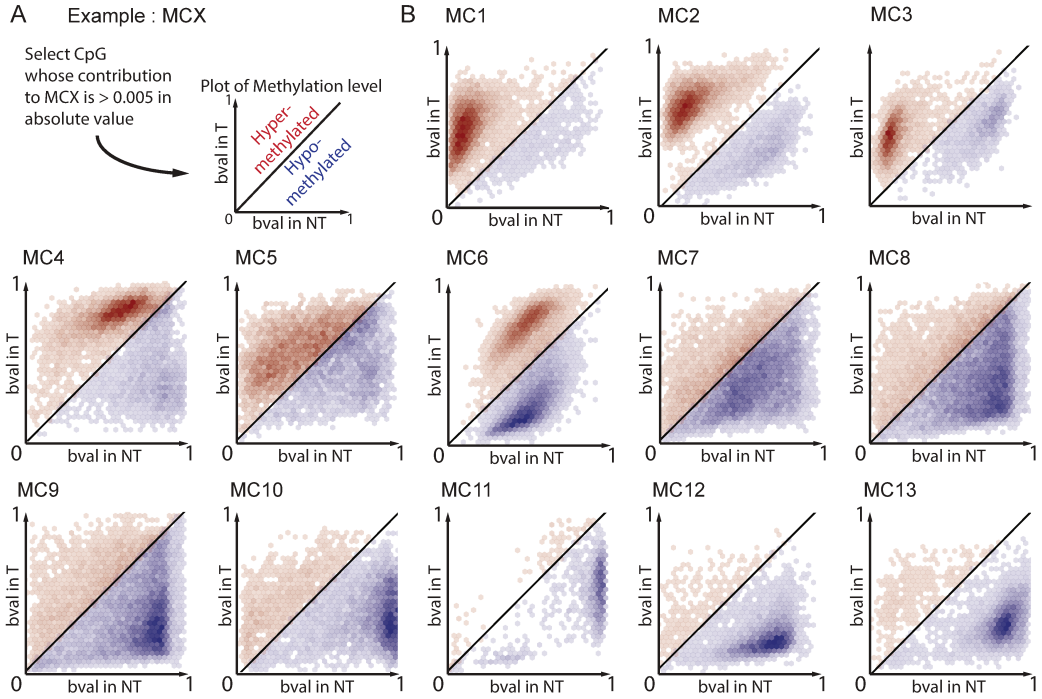


Fig. 2: Major DNA methylation changes associated with each component. (a) For each methylation component, the most contributive CpG sites were selected and their methylation compared between normal samples and the 5% of tumors with the strongest methylation changes. **(b)** Visualization of the main DNA methylation changes associated with each methylation component.

MCs are preferentially active in specific CpG contexts and chromatin states
We next examined whether the CpG sites with the strongest contribution to each component were preferentially located within specific CpG island (CGI)-based features (island, shore, shelf or outside CGI), gene-based features (TSS +/- 500 bases, gene body or intergenic) or chromatin states (**Fig. 3**). Chromatin states were defined by the ROADMAP consortium based on the ChIP-seq analysis of 7 different histone modifications in normal liver tissue [35]. Although histone marks are altered in cancer cells, we observed a good concordance between chromatin states defined in normal liver tissue and in the liver cancer cell line HepG2 (**Supplementary Fig. 2**). Thus, chromatin states defined in normal liver tissue likely reflect reasonably well the actual chromatin state at the time DNA methylation changes occur.

As expected, hypermethylation components were mostly active within CpG islands and at gene promoters, whereas hypomethylation components were associated with inactive chromatin domains. Components MC4-8, characterized by a more balanced combination of hyper- and hypo-methylation events, were enriched in enhancer regions. Beyond these general trends, each component displayed a specific pattern of enrichment within chromatin states, reproducible across the 3 cohorts (**Supplementary Fig. 3**), suggesting that they correspond to genuine biological processes preferentially active in certain epigenomic contexts.

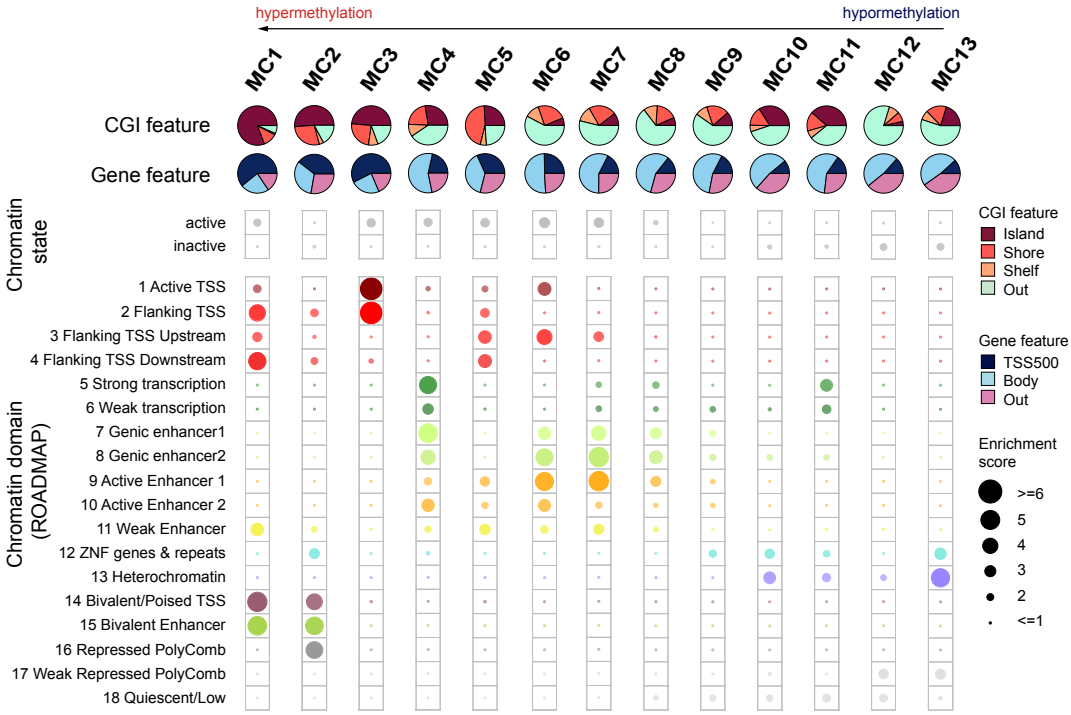


Fig. 3: Epigenomic features associated with each component. The most contributive CpG sites of each component were extracted. The first two lines indicate the proportion of these CpG sites falling within each CpG island and gene-based feature. Enrichment scores in active/inactive chromatin and across the 18 chromatin states defined by the ROADMAP consortium in normal liver are represented below.

DNA hypermethylation results from global and tumor-specific processes

To unravel the etiology of each process, we correlated the activity of components in tumor samples with diverse clinical and molecular features (**Fig. 4**). We first performed univariate linear regression analyses within the LICA-FR and TCGA-LIHC cohorts for which extensive clinical and molecular data was available (**Supplementary Table 2**). We then selected features significant in both cohorts for multivariate linear regression analysis to identify the most contributing features (**Supplementary Table 3**).

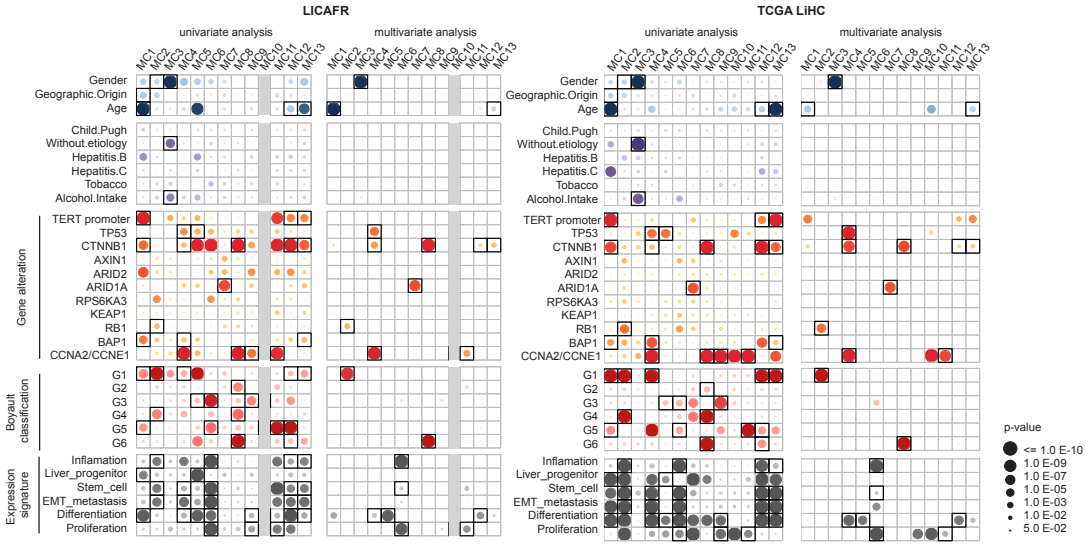


Fig. 4: Clinical and molecular features associated with each component. Clinical and molecular features significantly associated with each methylation component (MC) in the LICA-FR (left) and TCGA-LIHC (right) cohorts are represented. The size and color of each circle indicates the strength of the association in each cohort, in univariate (left) and multivariate (right) analyses.

Hypermethylation components MC1 and MC3 were strongly associated with patient characteristics (age and gender, respectively). MC3 splits perfectly males and females (P=2.7 × 10^{-58}, **Fig. 5a**). 95.5% of its most contributing CpG sites are located within the transcription start site (TSS) regions of X chromosome genes (**Fig. 5b**). These CpGs are unmethylated in males and hemi-methylated in females (**Fig. 5c**). Thus, MC3 is the signature of X chromosome inactivation in females and illustrates the ability of MethICA to identify signatures of well-defined epigenetic processes. MC1 increases linearly with age (P=8.9 × 10^{-6}, **Fig. 5d**) and is characterized by the hypermethylation of CpG islands preferentially located at bivalent TSS and enhancers (**Fig. 5e**). These regions display a co-existence of active (H3K4Me1 and/or H3K4Me3) and inactive (H3K27Me3) histone marks and have been shown to be prone to hypermethylation in cancer [36–39] and ageing [40,41]. Consistently, the most contributing CpG sites of MC1 were progressively hypermethylated with age both in HCC and normal liver, but the methylation increase was much faster in tumors (gain of 31.8% methylation per year vs 2.4% in normal tissue). Pan-cancer analysis revealed a similar association in other tissues (breast, lung, prostate and colorectal), with a systematically faster methylation increase in tumors as compared with normal tissue (**Supplementary**

Fig. 4). We conclude that MC1 reflects the progressive hypermethylation of bivalent chromatin domains that occurs naturally with age, but is sharply increased in tumors. Interestingly, this process was more active in well-differentiated HCC with activating *CTNNB1* and *TERT* promoter mutations (**Fig. 4**), although these associations were not significative independently from age in multivariate analysis.

Component MC2 was mostly active in CpG islands but also shores, and chromatin regions repressed by Polycomb proteins (marked by the repressive H3K27Me3 histone mark only) in addition to bivalent TSS and enhancers (**Fig. 5f**). Contrary to MC1, the hypermethylation component MC2 was not active in all HCC but essentially in the G1 molecular subgroup (P=6.7 × 10^{-6}, **Fig. 5g**,**h**). This subgroup, enriched in young patients of African origin, is characterized by an overexpression of genes expressed in fetal liver and controlled by parental imprinting [29]. G1 tumors display frequent alterations in *AXIN1*, *RPS6KA3* and *BAP1* genes, all of which were significantly associated with MC2, but not independently from G1 subgroup. By contrast, *RB1* mutations were associated with a stronger MC2 activity both in LICA-FR and TCGA cohorts. The specific hypermethylation signature of G1 tumors may reflect the epigenetic state of a different cell of origin, or the consequence of driver alterations enriched in this subgroup.

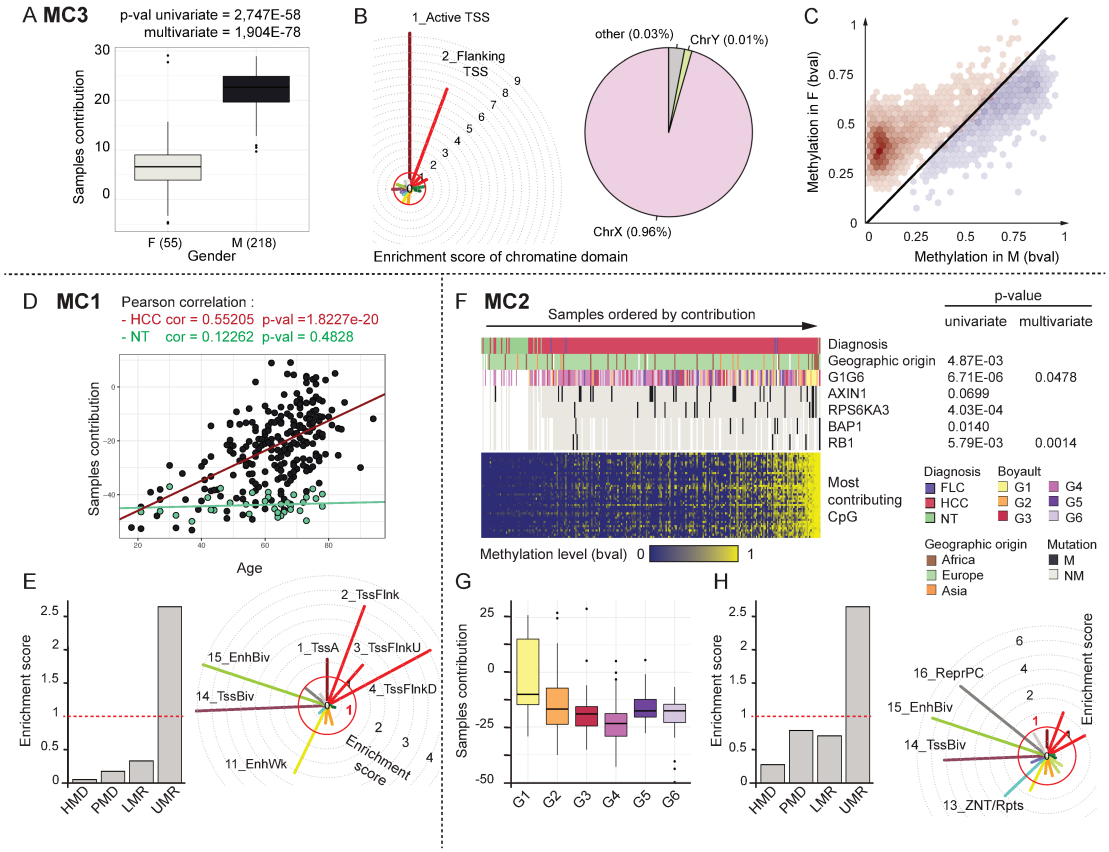


Fig. 5: Hypermethylation components operative in HCC. (a) Sample contribution to component MC3 allows to perfectly split males and females. **(b)** Enrichment of MC3 most contributive CpG sites chromatin states and chromosomes. **(c)** Average methylation of MC3 most contributive CpG sites in males (x-axis) and females (y-axis). **(d)** Hypermethylation component MC1 is strongly correlated with age in tumors. A modest increase is also observed in

males (x-axis) and females (y-axis). **(d)** Hypermethylation component MC1 is strongly correlated with age in tumors. A modest increase is also observed in normal samples, not significant in the LICA-FR but significant in TCGA-LIHC cohort (**Supplementary Fig. 5**). **(e)** The most contributive CpGs in MC1 are enriched in unmethylated regions within active transcription start sites (TSS) and bivalent TSS and enhancer chromatin domains. **(f,g)** Hypermethylation component MC2 is particularly active in the G1 expression subgroup, associated with *AXIN1*, *RPS6KA3* and *BAP1* mutations. It is also associated with *RB1* inactivation in multivariate analysis. **(h)** The most contributive CpG sites are enriched in unmethylated regions within bivalent TSS and enhancers, as well as chromatin domains repressed by Polycomb proteins.

Widespread hypomethylator phenotypes in *CTNNB1* and cyclin-activated tumors

Global loss of DNA methylation has been described in most cancer types (see [42] for review) including HCC [43], but the mechanism(s) by which this hypomethylation occurs and its potential role in cancer development remain incompletely understood.

Here, we identified 4 distinct hypomethylation components (MC10-13), preferentially affecting CpG sites within inactive chromatin states (Fig. 3). To further characterize these hypomethylation signatures, we analyzed the replication timing and the sequence contexts around their most contributive CpG sites. We first distinguished CpGs located in megabase-scale 'partially methylated domains' (PMDs [44]) and 'highly methylated domains' (HMDs) using WGBS data from normal liver [45]. We then classified the sequence context around each CpG dyad into 12 categories as described by Zhou *et al.* [19], taking into account the local CpG density (number of CpG sites within 35 bp on each side of the dyad) and the nucleotides directly flanking the CpG ('S' = C or G, 'W' = A or T).

Hypomethylation components MC12 and MC13 were both particularly active in late-replicated PMDs, but displayed different sequence context preferences (**Fig. 6a**). MC13 was enriched in CpG-dense regions, whereas MC12 was enriched in regions of low CpG density, in particular in 'WCGW' sequence context, consistent with the progressive loss of methylation occurring in healthy tissues along cell divisions [19]. Both MC12 and MC13 increased with age, but displayed an even stronger association with *CTNNB1* mutation in multivariate analysis (**Fig. 4**). As a result of these 2 processes, *CTNNB1*-mutated tumors displayed a massive hypomethylation of PMDs as compared with other HCC (**Fig. 6b,d**), with an average methylation in these regions of 45%, vs. 53% in other HCC and 71% in normal liver.

Strikingly, the 2 other hypomethylation components MC10 and MC11 were enriched in HMDs and dense CpG contexts (**Fig. 6a**), which have been described as hypomethylation-resistant in previous studies. These two components were particularly active in highly proliferative tumors, and specifically in the CCN-HCC subgroup driven by *CCNA2*/*E1* activation (**Fig. 6b,c**). In this subgroup, cyclin A2 or E1 activation leads to premature S phase entry and intense replication stress [46]. We hypothesize that, in these tumors, cancer cells are pushed to replicate so fast that even early-replicated HMDs become hypomethylated. Consistently, MC11 displays a stronger activity in early-replicated regions.

Altogether, our data indicate that several epigenetic processes are involved in the loss of DNA methylation in liver cancer cells. These processes are modulated by oncogenic alterations and lead to more or less extended hypomethylation patterns between molecular subgroups (**Fig. 6b**).

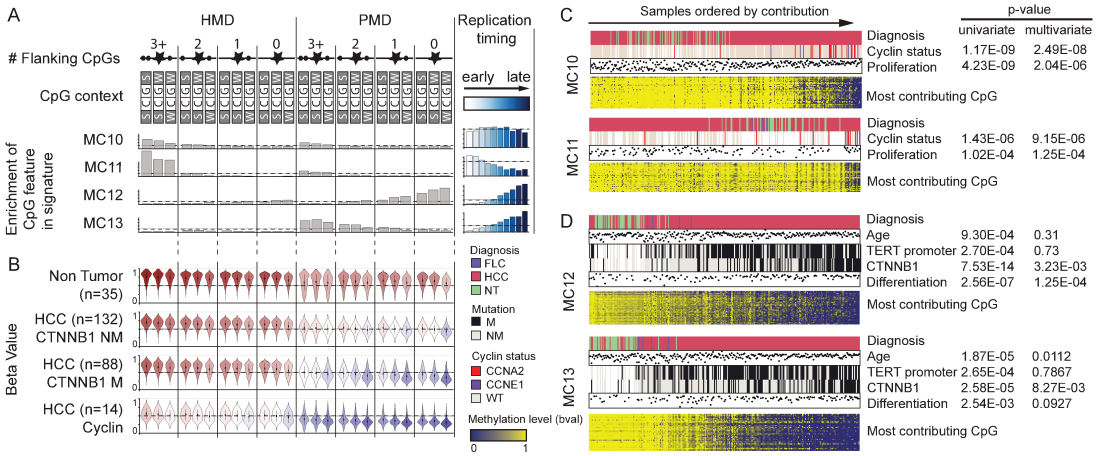


Fig. 6: Hypomethylation components operative in HCC. (a) Methylation domains and equence contexts enriched in the 4 hypomethylation components identified. **(b)** Distribution of methylation levels per CpG sequence context in normal liver and different HCC molecular subgroups. **(c,d)** Heatmaps showing the methylation of the most contributive CpG sites to components MC10 to MC13 across tumor and normal samples, ordered by component intensity, with associated clinico-molecular features.

Components MC4-8 reflect the coordinated regulation of enhancers by master regulators

We used ELMER tool [22,47] to identify CpG sites whose methylation level was linked with the expression of a nearby gene. We then quantified the transcriptional impact of each MC as the proportion of its 5% most contributing CpG sites significantly linked with the expression of a gene. Components MC4-8, enriched within enhancer regions marked by H3K27 acetylation, had the greatest transcriptional impact with on average 20% of their top CpGs linked with the expression of a gene, vs. 8.6% among hypermethylation components MC1-2 and 5.6% among hypomethylation components MC10-13.

To further explore the transcriptional changes associated with enhancer-related components MC4-8, we performed a gene-set enrichment analysis of genes linked with their most contributing CpG sites, coupled with a motif analysis to identify transcription factors responsible for their coordinated regulation.

Component MC8 was strongly associated with *CTNNB1* activating mutations (P=7.0 × 10^{-21}). In addition, different types of *CTNNB1* mutations activate ß-catenin with different strength [48], and the activity of component MC8 followed this gradient of activation (**Fig. 7a**). The most contributive CpG sites, preferentially located in active enhancers, were strongly correlated to the expression of adjacent genes (**Fig. 7b**) enriched in Wnt/ß-catenin targets overexpressed in the G6 subgroup (**Fig. 7c**). Motif analysis revealed an enrichment of TCF7 binding sites in the vicinity of MC8 top CpGs (**Fig. 7d**). TCF7 is a member of the TCF/LEF family of transcription factors, the main downstream

is a member of the TCF/LEF family of transcription factors, the main downstream effectors of Wnt signaling pathway. Thus, MC8 reveals a coordinated hypomethylation of enhancers bound by TCF7 in *CTNNB1*-mutated HCC, associated with the up-regulation of Wnt/ß-catenin pathway genes. A representative example is shown in **Fig. 7e,f** where the hypomethylation of a cluster of CpG sites, overlapping intragenic H3K27Ac and TCF7 ChIP-seq peaks, accompanies the overexpression of *AXIN2* in *CTNNB1*-mutated tumors. Whether this methylation signature precedes or follows the overexpression of Wnt target genes remains to be elucidated, but it certainly plays a role at least by stabilizing the transcriptional changes induced by *CTNNB1* mutations.

Component MC6 was linearly correlated with the level of immune infiltration in tumors (**Supplementary Fig. 5**) and was associated with two anti-correlated sets of CpGs. On one side, CpG sites located within hepato-specific enhancers, enriched in hepatocyte nuclear factors binding motifs, were hypomethylated in less infiltrated tumors. On the other side, CpG sites located near immune cell genes, enriched in JUN/FOS binding motifs, were hypomethylated in more infiltrated tumors. Thus, MC6 reflects the mixture of two cell types in HCC – liver and immune cells – each characterized by a specific enhancer methylation landscape.

Component MC7 was significantly associated with *ARID1A* mutations in both the LICA-FR (P=8.4×10^{-6}) and TCGA (P=1.3×10^{-5}) cohorts (**Fig. 7g**). *ARID1A*, member of the SWI/SNF chromatin remodeling complex, is recurrently mutated in HCC (13%, 4th most frequently altered gene [32]), but also in cirrhotic nodules. Indeed, *ARID1A* depletion has been shown to promote clonal expansion, notably in a context of chronic liver disease [49]. In mice, Arid1a interacts with several transcription factors that repress proliferation and maintain liver differentiation (C/ebpα, Hnf4a, Foxa2), and these pathways are down-regulated in *Arid1a* deficient cells [50]. Consistently, MC7 was characterized by a hypermethylation of enhancers, with an enrichment of several transcription factor binding motifs (**Fig. 7h**) including CEBPA, FOXA2, HNF4A but also NFIA implicated in the differentiation of several cell types [51–54]. Tumors with a strong activity of MC7 displayed a down-regulation of liver specific genes and an up-regulation of cell cycle genes (**Fig. 7i**). Thus, MC7 reveals the methylation changes related to *ARID1A* inactivation, with the hypermethylation of enhancers associated with a down-regulation of several differentiation pathways.

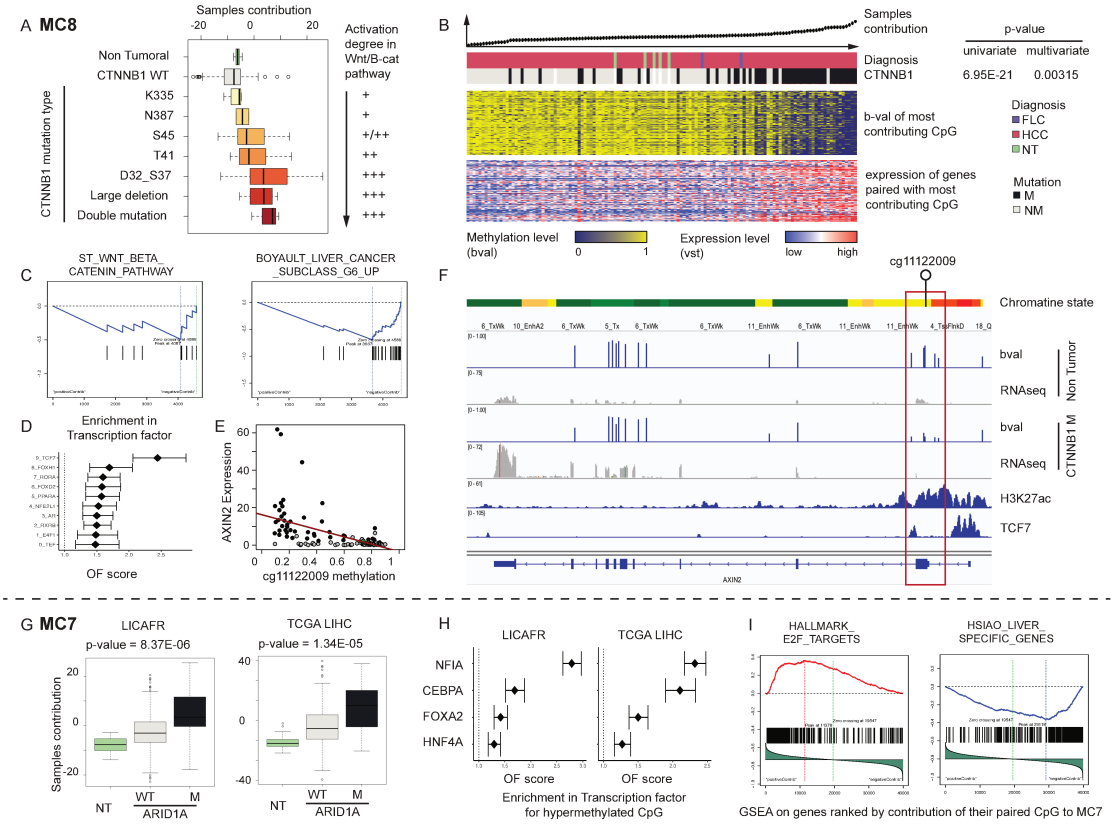


Fig. 7: Enhancer-related methylation components reflect the activity of transcriptional networks. Component MC8 is strongly correlated to somatic alterations that activate ß-catenin with different strength (**a**). Hypomethylation of its most contributive CpG sites is associated with the up-regulation of target genes (**b**), enriched in Wnt/ß-catenin targets (**c**). Motif analysis reveals an enrichment of TCF7 targets (**d**) exemplified by cg11122009 associated with *AXIN2* regulation (**e**,**f**). Component MC7, correlated with *ARID1A* inactivation (**g**) involves the hypermethylation of binding sites for several transcription factors related to differentiation (**h**). Hypermethylated genes are enriched in liver-specific genes while hypomethylated genes are enriched in cell cycle pathways like E2F targets (**i**).

DNA methylation-based classification of HCC reflects the combination of different components

We next performed unsupervised classifications of DNA methylation profiles to explore the contribution of each MC to the different tumor subgroups. Hierarchical clustering revealed 3 main clusters both in the LICA-FR (**Fig. 8a**) and TCGA-LIHC (**Fig. 8c**) cohorts: one cluster with few changes with respect to normal samples and two clusters with a hypermethylation of CpG islands (MC1), distinguished by the strength of hypomethylation in PMDs (MC12 and 13). Small subgroups with homogeneous methylation profiles are visible within these 3 main clusters, but it is difficult to define an optimal threshold to extract the most meaningful subgroups. By contrast, methylation components capture variations that are either widespread in the dataset (e.g. MC1) or restricted to small tumor groups (e.g. MC2 and MC10), as represented in the t-SNE plots (**Fig. 8b,d**). MCs thus reveal DNA methylation changes associated with relatively rare phenotypes like the G1

or CCN-HCC subgroups. They can also identify similarities across tumors belonging to different clusters. For example, *ARID1A*-mutated tumors are dispersed across the 3 clusters but MethICA was able to extract their common signature within MC7. Thus, independent DNA methylation component analysis is a powerful approach to refine the analysis of cancer methylomes beyond the main methylation clusters that only reflect a few dominant processes.

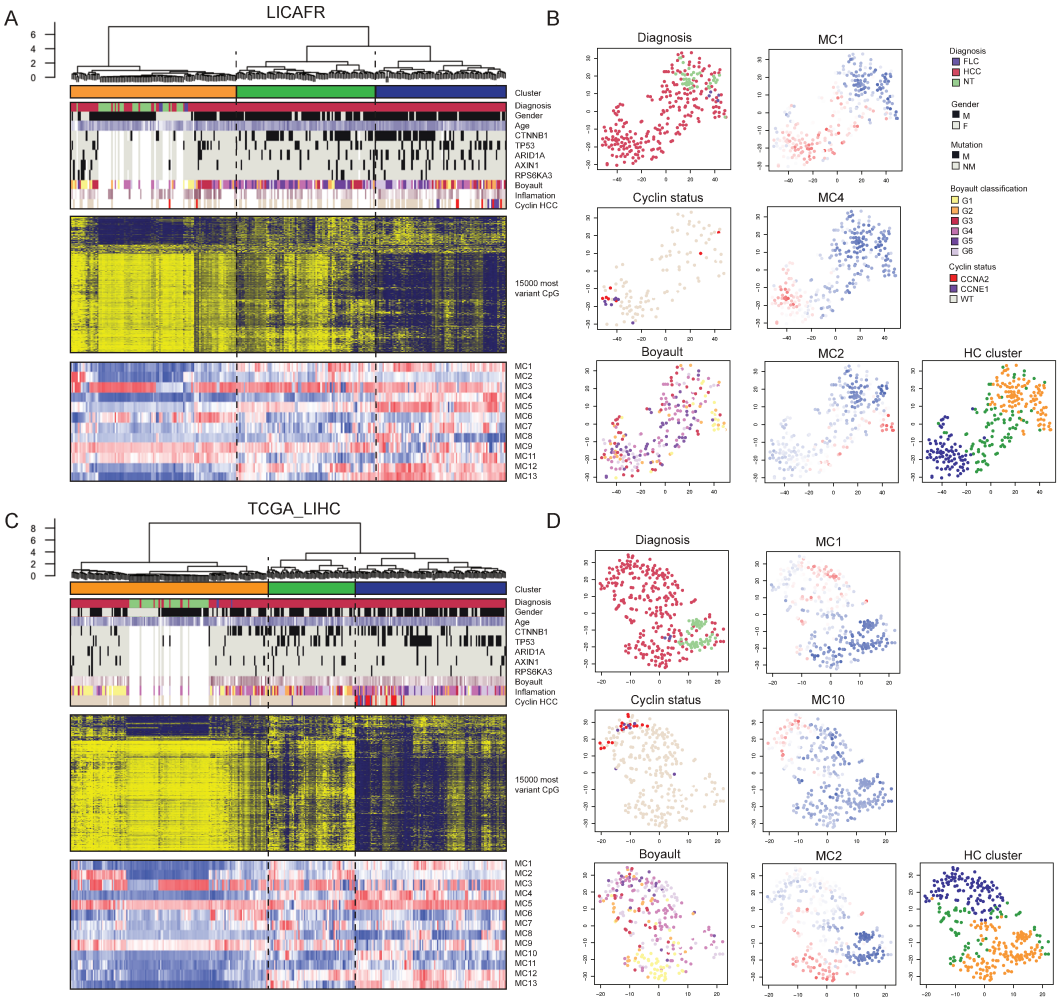


Fig. 8: DNA methylation-based classification of HCC. Tumors from the LICA-FR (**a**,**b**) and TCGA-LIHC (**c**,**d**) cohorts were classified according to the methylation levels of their 15,000 most variant CpGs using hierarchical clustering (**a**,**c**) or t-SNE (**b**,**d**). Associated clinico-molecular features and the intensity of each methylation component are represented by color codes.

DISCUSSION

Independent component analysis of the largest HCC series analyzed so far revealed 13 different methylation components operative with different strength across hepatocellular carcinomas and normal liver tissues. This represents a much greater diversity of methylation signatures than identified in previous studies. Early reports described global changes in HCC as compared with normal liver tissue, including hypermethylation of CpG islands enriched in Polycomb

repressive complex 2 (PRC2) target genes, and a widespread hypomethylation in open sea regions [6,55–62]. Recurrent target genes were also described including the tumor suppressor *CDKN2A*, inactivated by methylation in 53% of HCC [31]. Unsupervised classifications revealed between 3 and 7 HCC subgroups [31,63,64]. In particular, TCGA described 4 tumor subgroups based on hypermethylated probes and 3 subgroups, largely overlapping, based on hypomethylated probes. These subgroups display varying levels of hyper/hypomethylation with respect to normal samples and different activities of most of our methylation components. However, ICA allowed us to define more precise hypermethylation signatures, and to disentangle age-related processes from changes associated with specific tumor subgroups and driver alterations. The MC1 component captured the hypermethylation of CpG islands located in bivalent chromatin domains, known to occur naturally with ageing. This component increases with age in both normal liver and HCC, but the slope of this increase is much sharper in tumors. By contrast, MC2 is associated with the G1 transcriptomic subgroup and defines a strongly hypermethylated HCC entity. Further studies are required to determine whether this methylation signature results from a different cell of origin for this subgroup or is acquired during tumorigenesis.

Hypomethylation of open sea regions is a hallmark of cancer cells, but we show here that 4 independent processes are implicated in DNA hypomethylation in HCC. Two components (MC12 and MC13) are preferentially active in late-replicated partially methylated domains. PMD hypomethylation have been shown to occur in normal tissue along cell divisions [19,65]. Consistently, these components increase with age in our series but they are also correlated with *CTNNB1* activating mutations. This association is unexpected considering that *CTNNB1*-mutated HCC are better differentiated and less proliferative than other HCC subgroups. The two other hypomethylation components (MC10 and MC11) affect highly methylated domains and are particularly active in the CCN-HCC subgroup, driven by cyclin A2/E1 activation [46]. This finding demonstrates that replication stress not only favors the occurrence of genomic alterations but also impacts the methylome of cancer cells.

Finally, we identified 5 components (MC4-8) related to differential enhancer methylation. These components have been missed in previous studies, possibly because they involve fewer CpG sites and can be dispersed across several DNA methylation subgroups. Yet, they have the strongest transcriptional impact and constitute valuable markers of transcriptional network activity. MC8 reflects the precise level of activation of the Wnt/ß-catenin pathway induced by diverse *CTNNB1* mutations [48]. MC7 is to our knowledge the first methylation signature associated with *ARID1A* mutations. Motif analysis shed light on the transcription factors impacted by *ARID1A* deficiency, including several key regulators of differentiation.

Overall, ICA appears as a powerful tool for the analysis of DNA methylation signatures. All the utilities we developed for extracting and interpreting DNA methylation components are included in the *MethICA* package, available in Github.

MATERIAL AND METHODS

Cohorts used in this study

LICA-FR cohort

A series of 274 samples – 236 hepatocellular carcinoma (HCC) and 35 non-tumor liver tissues – were collected from patients surgically treated in four French hospitals located in Bordeaux and Paris region. The study was approved by institutional review board committees (CCPRB Paris Saint-Louis, 1997, 2004, and 2010, approval number 01–037; Bordeaux, 2010- A00498–31). Written informed consent was obtained in accordance with French legislation. Of the 274 HCC cases, 110 (40%) developed in non-fibrotic (METAVIR F0-F1), 65 (25%) in chronic hepatitis (F2–F3) and 94 (35%) in cirrhotic liver (F4). Clinicopathological data were available for all cases. The LICA-FR mostly comprises males (80%), with a median of 65 years, related to diverse risk factors including alcohol (43%), hepatitis B virus (HBV, 19%), and hepatitis C virus (HCV, 17%) infection. The 274 samples were analyzed using Illumina Infinium HumanMethylation450 arrays for this study (see below). Somatic mutations were available for 195 samples previously analyzed by whole genome (WGS) or whole exome sequencing (WES) [32,46,66–68]. RNA sequencing (RNA-seq) data was also available for 128 samples [46,68].

Detailed clinical characteristics and sequencing details for each sample are provided in **Supplementary table 1**.

TCGA-LIHC cohort

The TCGA-LIHC cohort comprises 275 HCC and 50 non-tumor liver tissues analyzed using Illumina Infinium HumanMethylation450 arrays, whole exome and RNA sequencing. Clinical annotations, DNA methylation (QC metrics and methylation beta values) and RNA-seq data (raw read counts per gene) were obtained from the TCGA data portal (https://tcga-data.nci.nih.gov). Cholangiocarcinomas and mixed forms of HCC were discarded to keep only pure HCC and non-tumor samples. Single somatic mutations and TERT promoter mutation were retrieved from the original article [31]. HCC cases in this cohort are predominantly of American, Canadian or Vietnamese origin, mostly males (64%), with a median of 63 years and related to diverse risk factors: alcohol (34%), HBV (14%), HCV (17%). Detailed clinical characteristics and sequencing details for each sample are available at the TCGA website.

HEPTROMIC cohort

The HEPTROMIC cohort comprises 221 surgically resected HCC and 19 non-tumor liver tissues analyzed using Illumina Infinium HumanMethylation450 arrays [34]. HCC cases in this cohort come from two institutions of the HCC Genomic Consortium: IRCCS Istituto Nazionale Tumori (Milan, Italy) and Hospital Clínic (Barcelona, Spain). HEPTROMIC cases are mostly males (78%), with a median of 66 years and a predominance of viral-related etiologies (HCV, 47%; HBV, 20%). No additional molecular data was available for this cohort

DNA methylation arrays

We analyzed the 274 samples from the LICA-FR cohort using Illumina Infinium HumanMethylation450 arrays. Microarray experiments were performed by

Integragen SA (Evry, France). In brief, genomic DNA was bisulfite-converted using the EZ-96 DNA Methylation Kit (Zymo Research, Irvine, CA, USA), whole-genome amplified, enzymatically fragmented, and hybridized to the BeadChip arrays in accordance with the manufacturer's instructions. The beta value (bval) DNA methylation scores for each locus were extracted together with detection p-values from Illumina GenomeStudio software. The bval gives an estimate of the methylation level of each CpG locus using the ratio of intensities between methylated and unmethylated probes. We removed CpGs with "NA" values or a detection p-value >0.05 in more than 20% of the samples, leaving 351,509 probes for analysis.
The 2 other cohorts (TCGA-LIHC and HEPTROMIC) were analyzed with the same methylation array. We retrieved the beta value and detection p-value matrices for these two datasets and selected reliable CpGs as for the LICA-FR cohort.

RNA-seq data processing
RNA-seq read counts per gene were obtained for the LICA-FR cohort as previously described [46], and directly from the TCGA website for the TCGA-LIHC cohort. We then applied the same pipeline to the raw counts of the two series to obtain normalized FPKM and vst matrices. We used DESeq2 [69] to import raw read counts into R statistical software and apply variance stabilizing transformation (VST) to the raw count matrix. FPKM scores (number of fragments per kilobase of exon model and millions of mapped reads) were calculated by normalizing the count matrix for the library size and the coding length of each gene.

Independent component analysis
We restricted each data set to the 200,000 most variant CpGs based on their standard deviation. We computed 20 independent methylation components (MC) in each cohort using the FastICA algorithm [70], as implemented in the *sklearn.decomposition* Python library (http://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html), with a first step of whitening of the matrix, the function of approximation to neg-entropy *logcosh*, and *parallel* algorithm. Since the FastICA algorithm involves random initialization, we performed 100 iterations and kept the results from the most stable iteration. A component was considered "stable" when a similar component (Pearson correlation of CpG contribution > 0.9) was identified in ≥50% of iterations. We selected the iteration giving the highest number of stable components and the highest average Pearson correlation score among stable components.
We next compared the results obtained for the 3 data sets. The similarity of two components from two different data sets was determined by calculating the absolute value of the Pearson correlation coefficient from the contribution of their common CpGs. We selected for further analysis the 13 most reliable components found in at least 2 of the 3 HCC data sets with a Pearson correlation score > 0.35.

Association between methylation components and (epi)genomic features
To better understand the preferential activity of each component towards specific regions, we analyzed the enrichment of their most contributive CpG sites across diverse types of (epi)genomic features. We selected the most contributive CpGs of each component (MC) by thresholding their absolute projections onto the MC: abs(projection) > 0.005. We next estimated the enrichment of these most

contributive CpGs across diverse (epi)genomic features. To do so, we calculated an enrichment score for each feature corresponding to the ratio between the proportion of the most contributive CpGs being located within the feature and the proportion of the 200,000 analyzed CpGs being located within the feature:

$$ES = \frac{N_{feature}^{contrib} / N^{contrib}}{N_{feature}^{all} / 200{,}000}$$

With $N_{feature}^{contrib}$ the number of most contributive CpGs located within the feature, $N^{contrib}$ the number of most contributive CpGs and $N_{feature}^{all}$ the number of CpGs located within the feature among the 200,000 analyzed CpGs.

The following (epi)genomic features were considered:

- CpG island-based features: CpG islands retrieved from UCSC database (release 19, GRCh37), shores (2 kb on each side of the islands) and shelves (2 kb outside shores)
- gene-based features: promoter (defined as transcription start site (TSS) +/- 500 bp) and gene body for each gene in GENCODE database (release 19 - GRCh37.p13)
- chromatin states in normal liver as defined by the ROADMAP consortium [35]. Eighteen chromatin states were defined by the consortium based on the genome-wide analysis of 6 histone marks (H3K4me3, H3K4me1, H3K36me3, H3K27me3, H3K9me3 and H3K27ac) using a multivariate Hidden Markov model (ChromHMM tool [71]). Each chromatin state corresponds to a particular combination of histone marks and is associated with a specific type of functional element (e.g. active TSS, genic enhancers, heterochromatin…). We downloaded the bed file of chromatin states in normal liver through the ROADMAP epigenomics website (http://www.roadmapepigenomics.org).
- DNA methylation domains derived from whole genome bisulfite sequencing of normal hepatocytes [45]. Genome-wide CpG methylation analyses have shown that the epigenome is organized in megabase-scale partially methylated domains (PMD, methylation between 50% and 80%) and highly methylated domains (HMD, methylation > 80%), as well as short (regulatory) lowly methylated (LMR, methylation between 10% and 50%) and unmethylated regions (UMR, methylation < 10%) [44,72]. We retrieved these domains in hepatocytes defined by Salhab *et al.* [45] using a Hidden Markov Model-based detection method called methylSeekR [72].
- replication timing in the liver cancer cell line HepG2. We used Repli-seq data generated by the ENCODE project [73] to characterize the replication timing of each CpG site. To do so, we downloaded the wavelet-smoothed Repli-seq signals for HepG2 cell line through the UCSC genome browser, and we segmented this signal into 10 deciles from the earliest (decile 1) to the latest (decile 10) replicated regions.

Association between methylation components and clinico-molecular annotations

We analyzed the association of each methylation component (MC) with more than 50 clinical and molecular features. For this part we chose to focus on the LICA-FR

and TCGA-LIHC cohorts, for which extensive clinical and molecular data were available. We first used linear models to identify features significantly correlated with sample contributions, using the *lm* function in R statistical software: *lm(sample contribution ~ annotation)*. Only positive associations were considered, i.e. features associated with an increased activity of a component. For example, mutation of a given driver gene was considered to be associated with a component only if mutated cases displayed a higher activity of the MC, to favor the identification of causal factors rather than indirect associations. This step was done separately in the LICA-FR and TCGA-LIHC cohorts. Clinico-molecular features that were significant (p-value < 0.01) in both cohorts were then included in multivariate analyses using also *lm* function: *lm(sample contribution ~ all selected annotations)*. We defined the most contributing features of each MC as those that remained significant (p-value < 0.05) in multivariate analysis in both cohorts.

Clinical features for the LICA-FR cohort are detailed in **Supplementary Table 1** and included patient information (gender, geographic origin, age), risk factors (alcohol intake, HBV or HCV infection, metabolic syndrome), underlying liver disease (METAVIR fibrosis stage; F0-F1: no fibrosis, F2-F3: moderate fibrosis, F4: cirrhotic liver) and various tumor characteristics like the number of nodules and size of the largest nodule, vascular invasion, Barcelona Clinic Liver Cancer stage (BCLC 0, A, B, C and D from the earliest to the terminal stage) and Edmonson grade (I-II = well differentiated, III-IV = poorly differentiated). In the TCGA cohort, some features were absent hence were only studied in the LICA-FR series: metabolic syndrome, number of nodules and largest nodule size, vascular invasion, BCLC stage and Edmonson grade. The Ishak fibrosis score was converted to METAVIR for comparison with the LICA-FR series as follows: "0,1,2 - No Fibrosis or Portal Fibrosis" = F0-F1; "3,4 - Fibrous Speta" = F2-F3; "5,6 - Nodular Formation, Incomplete Cirrhosis" and "Established Cirrhosis" = F4).

Molecular features analyzed in both cohorts included:

- driver alterations of 27 HCC driver genes defined by Schulze *et al.* [32] or characterized recently in the lab [46,68]: *TERT*, *CTNNB1*, *TP53*, *ARID1A*, *AXIN1*, *CDKN2A*, *ARID2*, *RPS6KA3*, *NFE2L2*, *KEAP1*, *PTEN*, *HNF1A*, *ALB*, *ACVR2A*, *RPL22*, *CDKN1A*, *RB1*, *TSC2*, *ATP10B*, *FGA*, *MEF2C*, *ZNRF3*, *EPHA4*, *TSC1*, *CCNA2*, *CCNE1*, *BAP1*). Mutational status for these 27 genes was derived from whole exome or whole genome sequencing, completed by *TERT* promoter screening by Sanger sequencing for both the TCGA-LIHC [31] and LICA-FR [74] cohorts.
- molecular subgroups of HCC, G1 to G6, defined by Boyault *et al.* from gene expression data [29]. We used the *MS.liverK* package [75] to predict the molecular group of each tumor based on RNA-seq expression data.
- selected transcriptional signatures related to hepatocellular carcinoma phenotypes were analyzed, including differentiation (*ALB*, *CDH1*, *APOF*, *CYP1A1*, *CYP2A6*, *UGT2B7*, *HNF1A*, *HNF4A*) and proliferation (*CDC20*, *GMNN*, *MKI67*, *RRM2*, *CCNA2*, *CCND1*, *CCNE1*, *AURKA*, *BUB1*, *PCNA*, *RAN*, *BIRC5*, *SPP1*) signatures defined by Nault *et al.* [76], as well as liver progenitor (*PROX1*, *AFP*, *EPCAM*, *IGF2*, *SALL4*, *PROM1*, *LGR5*, *GPC3*, *LIN28B*), stem cell (*CD47*, *CD44*, *KDR*, *IL6*, *NCAM2*, *THY1*, *KIT*) and epithelial-mesenchymal transition/metastasis (*SNAI2*, *ITGB3*, *TWIST1*, *ZEB2*, *PLAUR*,

VIM) signatures defined by Caruso *et al.* [77]. For each signature, a score was computed in each tumor as the mean expression of marker genes.
- immune infiltrate estimated from RNA-seq data using the MCPcounter tool [78]. The overall immune infiltrate was obtained by summing MCPcounter scores for all immune cell populations.

Linking CpG methylation with transcriptional networks

We used ELMER tool [22] to identify CpG-gene pairs, i.e. correlations between the methylation level of a CpG site and the expression of one or more nearby genes, leveraging samples with matched methylation array and RNA-seq data. The *get.pair* function of ELMER v2 package [47] was used in unsupervised mode to compare the expression of the 10 genes closest to each CpG site between the 40% samples with the highest/lowest methylation level for that CpG. We used a permutation size of 10,000 and selected CpG-gene pairs with an empirical p-value $Pe < 0.001$. We used an in-house adaptation of the GSEA (Gene Set Enrichment Analysis) method [79], modified to take as input a ranked gene list instead of an expression matrix, to identify gene sets associated with each methylation component (MC). For each MC, genes were ranked according to the contribution of their paired CpG. Genes paired with several CpGs were assigned to the CpG with the strongest contribution to the component (in absolute value). GSEA was then used to identify gene sets from the MSigDB v6 database overrepresented among genes paired with the most contributive CpGs. We used the *get.enriched.motif* of ELMER v2 package to identify transcription factor binding motifs enriched around the most contributive CpGs of each MC.

DNA methylation-based classification of hepatocellular carcinomas

We performed unsupervised classifications of the LICA-FR and TCGA-LIHC cohorts based on methylation profiles. Hierarchical clustering was performed on the 15 000 most variant probes (based on standard deviation) using R function *hclust* with Pearson's dissimilarity as distance metric and Ward.D2 linkage method. T-stochastic neighbor embedding (tSNE) was used to project the data set in two dimensions using the *Rtsne* package (https://github.com/jkrijthe/Rtsne). t-SNE was applied to a Pearson correlation matrix of CpGs with standard deviation > 0.25, with a theta value of zero over 2,000 iterations and perplexity of 19 for TCGA-LIHC and 12 for LICA-FR.

REFERENCES

1. Jones PA, Baylin SB. The epigenomics of cancer. Cell. 2007;128:683–92.

2. Esteller M. Cancer epigenomics: DNA methylomes and histone-modification maps. Nat Rev Genet. 2007;8:286–98.

3. Baylin SB, Jones PA. A decade of exploring the cancer epigenome - biological and translational implications. Nat Rev Cancer. 2011;11:726–34.

4. Witte T, Plass C, Gerhauser C. Pan-cancer patterns of DNA methylation. Genome Med. 2014;6:66.

5. Toyota M, Ahuja N, Ohe-Toyota M, Herman JG, Baylin SB, Issa JP. CpG island methylator phenotype in colorectal cancer. Proc Natl Acad Sci USA. 1999;96:8681–6.

6. Zhang C, Li Z, Cheng Y, Jia F, Li R, Wu M, et al. CpG island methylator phenotype association with elevated serum alpha-fetoprotein level in hepatocellular carcinoma. Clin Cancer Res. 2007;13:944–52.

7. Hinoue T, Weisenberger DJ, Lange CPE, Shen H, Byun H-M, Van Den Berg D, et al. Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res. 2012;22:271–82.

8. Hughes LAE, Melotte V, de Schrijver J, de Maat M, Smit VTHBM, Bovée JVMG, et al. The CpG island methylator phenotype: what's in a name? Cancer Res. 2013;73:5858–68.

9. Turcan S, Rohle D, Goenka A, Walsh LA, Fang F, Yilmaz E, et al. IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature. 2012;483:479–83.

10. Letouzé E, Martinelli C, Loriot C, Burnichon N, Abermil N, Ottolenghi C, et al. SDH mutations establish a hypermethylator phenotype in paraganglioma. Cancer Cell. 2013;23:739–52.

11. Figueroa ME, Abdel-Wahab O, Lu C, Ward PS, Patel J, Shih A, et al. Leukemic IDH1 and IDH2 mutations result in a hypermethylation phenotype, disrupt TET2 function, and impair hematopoietic differentiation. Cancer Cell. 2010;18:553–67.

12. Sturm D, Witt H, Hovestadt V, Khuong-Quang D-A, Jones DTW, Konermann C, et al. Hotspot mutations in H3F3A and IDH1 define distinct epigenetic and biological subgroups of glioblastoma. Cancer Cell. 2012;22:425–37.

13. Gonzalez-Perez A, Jene-Sanz A, Lopez-Bigas N. The mutational landscape of chromatin regulatory factors across 4,623 tumor samples. Genome Biol. 2013;14:r106.

14. Hoadley KA, Yau C, Hinoue T, Wolf DM, Lazar AJ, Drill E, et al. Cell-of-Origin Patterns Dominate the Molecular Classification of 10,000 Tumors from 33 Types of Cancer. Cell. 2018;173:291-304.e6.

15. Bormann F, Rodríguez-Paredes M, Lasitschka F, Edelmann D, Musch T, Benner A, et al. Cell-of-Origin DNA Methylation Signatures Are Maintained during Colorectal Carcinogenesis. Cell Rep. 2018;23:3407–18.

16. Rakyan VK, Down TA, Maslau S, Andrew T, Yang T-P, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20:434–9.

17. Horvath S. DNA methylation age of human tissues and cell types. Genome Biol.

2013;14:R115.

18. Day K, Waite LL, Thalacker-Mercer A, West A, Bamman MM, Brooks JD, et al. Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biol. 2013;14:R102.

19. Zhou W, Dinh HQ, Ramjan Z, Weisenberger DJ, Nicolet CM, Shen H, et al. DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat Genet. 2018;50:591–602.

20. Vandiver AR, Irizarry RA, Hansen KD, Garza LA, Runarsson A, Li X, et al. Age and sun exposure-related widespread genomic blocks of hypomethylation in nonmalignant skin. Genome Biol. 2015;16:80.

21. Kang GH, Lee S, Kim WH, Lee HW, Kim JC, Rhyu M-G, et al. Epstein-barr virus-positive gastric carcinoma demonstrates frequent aberrant methylation of multiple genes and constitutes CpG island methylator phenotype-positive gastric carcinoma. Am J Pathol. 2002;160:787–94.

22. Yao L, Shen H, Laird PW, Farnham PJ, Berman BP. Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol. 2015;16:105.

23. Chakravarthy A, Furness A, Joshi K, Ghorani E, Ford K, Ward MJ, et al. Pan-cancer deconvolution of tumour composition using DNA methylation. Nat Commun. 2018;9:3220.

24. Zinovyev A, Kairov U, Karpenyuk T, Ramanculov E. Blind source separation methods for deconvolution of complex signals in cancer biology. Biochem Biophys Res Commun. 2013;430:1182–7.

25. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500:415–21.

26. Teschendorff AE, Journée M, Absil PA, Sepulchre R, Caldas C. Elucidating the altered transcriptional programs in breast cancer using independent component analysis. PLoS Comput Biol. 2007;3:e161.

27. Biton A, Bernard-Pierrot I, Lou Y, Krucker C, Chapeaublanc E, Rubio-Pérez C, et al. Independent component analysis uncovers the landscape of the bladder tumor transcriptome and reveals insights into luminal and basal subtypes. Cell Rep. 2014;9:1235–45.

28. Llovet JM, Zucman-Rossi J, Pikarsky E, Sangro B, Schwartz M, Sherman M, et al. Hepatocellular carcinoma. Nat Rev Dis Primers. 2016;2:16018.

29. Boyault S, Rickman DS, de Reyniès A, Balabaud C, Rebouissou S, Jeannot E, et al. Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology. 2007;45:42–52.

30. Hoshida Y, Nijman SMB, Kobayashi M, Chan JA, Brunet J-P, Chiang DY, et al. Integrative Transcriptome Analysis Reveals Common Molecular Subclasses of Human Hepatocellular Carcinoma. Cancer Res. 2009;69:7385–92.

31. Cancer Genome Atlas Research Network. Electronic address: wheeler@bcm.edu, Cancer Genome Atlas Research Network. Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell. 2017;169:1327-1341.e23.

32. Schulze K, Imbeaud S, Letouzé E, Alexandrov LB, Calderaro J, Rebouissou S, et al. Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat Genet. 2015;47:505–11.

33. Fujimoto A, Furuta M, Totoki Y, Tsunoda T, Kato M, Shiraishi Y, et al. Whole-genome mutational landscape and characterization of noncoding and structural mutations in liver cancer. Nat Genet. 2016;48:500–9.

34. Villanueva A, Portela A, Sayols S, Battiston C, Hoshida Y, Méndez-González J, et al. DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma. Hepatology. 2015;61:1945–56.

35. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518:317–30.

36. Ohm JE, McGarvey KM, Yu X, Cheng L, Schuebel KE, Cope L, et al. A stem cell-like chromatin pattern may predispose tumor suppressor genes to DNA hypermethylation and heritable silencing. Nat Genet. 2007;39:237–42.

37. Schlesinger Y, Straussman R, Keshet I, Farkash S, Hecht M, Zimmerman J, et al. Polycomb-mediated methylation on Lys27 of histone H3 pre-marks genes for de novo methylation in cancer. Nat Genet. 2007;39:232–6.

38. Widschwendter M, Fiegl H, Egle D, Mueller-Holzner E, Spizzo G, Marth C, et al. Epigenetic stem cell signature in cancer. Nat Genet. 2007;39:157–8.

39. Easwaran H, Johnstone SE, Van Neste L, Ohm J, Mosbruger T, Wang Q, et al. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Res. 2012;22:837–49.

40. Rakyan VK, Down TA, Maslau S, Andrew T, Yang T-P, Beyan H, et al. Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. 2010;20:434–9.

41. Teschendorff AE, Menon U, Gentry-Maharaj A, Ramus SJ, Weisenberger DJ, Shen H, et al. Age-dependent DNA methylation of genes that are suppressed in stem cells is a hallmark of cancer. Genome Res. 2010;20:440–6.

42. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009;1:239–59.

43. Lin C-H, Hsieh S-Y, Sheen I-S, Lee W-C, Chen T-C, Shyu W-C, et al. Genome-wide Hypomethylation in Hepatocellular Carcinogenesis. Cancer Res. 2001;61:4238–43.

44. Lister R, Pelizzola M, Dowen RH, Hawkins RD, Hon G, Tonti-Filippini J, et al. Human DNA methylomes at base resolution show widespread epigenomic differences. Nature. 2009;462:315–22.

45. Salhab A, Nordström K, Gasparoni G, Kattler K, Ebert P, Ramirez F, et al. A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. Genome Biology. 2018;19:150.

46. Bayard Q, Meunier L, Peneau C, Renault V, Shinde J, Nault J-C, et al. Cyclin A2/E1 activation defines a hepatocellular carcinoma subclass with a rearrangement signature of replication stress. Nat Commun. 2018;9:5235.

47. Silva TC, Coetzee SG, Gull N, Yao L, Hazelett DJ, Noushmehr H, et al. ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA

methylation and transcriptome profiles. Bioinformatics. 2019;35:1974–7.

48. Rebouissou S, Franconi A, Calderaro J, Letouzé E, Imbeaud S, Pilati C, et al. Genotype-phenotype correlation of CTNNB1 mutations reveals different ß-catenin activity associated with liver tumor progression. Hepatology. 2016;64:2047–61.

49. Zhu M, Lu T, Jia Y, Luo X, Gopal P, Li L, et al. Somatic Mutations Increase Hepatic Clonal Fitness and Regeneration in Chronic Liver Disease. Cell. 2019;177:608-621.e12.

50. Sun X, Chuang J-C, Kanchwala M, Wu L, Celen C, Li L, et al. Suppression of the SWI/SNF Component Arid1a Promotes Mammalian Regeneration. Cell Stem Cell. 2016;18:456–66.

51. Piper M, Barry G, Hawkins J, Mason S, Lindwall C, Little E, et al. NFIA controls telencephalic progenitor cell differentiation through repression of the Notch effector Hes1. J Neurosci. 2010;30:9127–39.

52. Hiraike Y, Waki H, Yu J, Nakamura M, Miyake K, Nagano G, et al. NFIA co-localizes with PPARγ and transcriptionally controls the brown fat gene program. Nat Cell Biol. 2017;19:1081–92.

53. Singh PNP, Yadav US, Azad K, Goswami P, Kinare V, Bandyopadhyay A. NFIA and GATA3 are crucial regulators of embryonic articular cartilage differentiation. Development. 2018;145.

54. Chen K-S, Bridges CR, Lynton Z, Lim JWC, Stringer BW, Rajagopal R, et al. Transcription factors NFIA and NFIB induce cellular differentiation in high-grade astrocytoma. J Neurooncol. 2020;146:41–53.

55. Arai E, Ushijima S, Gotoh M, Ojima H, Kosuge T, Hosoda F, et al. Genome-wide DNA methylation profiles in liver tissue at the precancerous stage and in hepatocellular carcinoma. Int J Cancer. 2009;125:2854–62.

56. Hernandez-Vargas H, Lambert M-P, Le Calvez-Kelm F, Gouysse G, McKay-Chopin S, Tavtigian SV, et al. Hepatocellular carcinoma displays distinct DNA methylation signatures with potential as clinical predictors. PLoS ONE. 2010;5:e9749.

57. Stefanska B, Huang J, Bhattacharyya B, Suderman M, Hallett M, Han Z-G, et al. Definition of the landscape of promoter DNA hypomethylation in liver cancer. Cancer Res. 2011;71:5891–903.

58. Neumann O, Kesselmeier M, Geffers R, Pellegrino R, Radlwimmer B, Hoffmann K, et al. Methylome analysis and integrative profiling of human HCCs identify novel protumorigenic factors. Hepatology. 2012;56:1817–27.

59. Ammerpohl O, Pratschke J, Schafmayer C, Haake A, Faber W, von Kampen O, et al. Distinct DNA methylation patterns in cirrhotic liver and hepatocellular carcinoma. Int J Cancer. 2012;130:1319–28.

60. Shen J, Wang S, Zhang Y-J, Kappil M, Wu H-C, Kibriya MG, et al. Genome-wide DNA methylation profiles in hepatocellular carcinoma. Hepatology. 2012;55:1799–808.

61. Shen J, Wang S, Zhang Y-J, Wu H-C, Kibriya MG, Jasmine F, et al. Exploring genome-wide DNA methylation profiles altered in hepatocellular carcinoma using Infinium HumanMethylation 450 BeadChips. Epigenetics. 2013;8:34–43.

62. Song M-A, Tiirikainen M, Kwee S, Okimoto G, Yu H, Wong LL. Elucidating the

landscape of aberrant DNA methylation in hepatocellular carcinoma. PLoS ONE. 2013;8:e55761.

63. Mah W-C, Thurnherr T, Chow PKH, Chung AYF, Ooi LLPJ, Toh HC, et al. Methylation profiles reveal distinct subgroup of hepatocellular carcinoma patients with poor prognosis. PLoS ONE. 2014;9:e104158.

64. Cheng J, Wei D, Ji Y, Chen L, Yang L, Li G, et al. Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome Medicine. 2018;10:42.

65. Dmitrijeva M, Ossowski S, Serrano L, Schaefer MH. Tissue-specific DNA methylation loss during ageing and carcinogenesis is linked to chromosome structure, replication timing and cell division rates. Nucleic Acids Res. 2018;46:7022–39.

66. Guichard C, Amaddeo G, Imbeaud S, Ladeiro Y, Pelletier L, Maad IB, et al. Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat Genet. 2012;44:694–8.

67. Letouzé E, Shinde J, Renault V, Couchy G, Blanc J-F, Tubacher E, et al. Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nat Commun. 2017;8:1315.

68. Hirsch TZ, Negulescu A, Gupta B, Caruso S, Noblet B, Couchy G, et al. BAP1 mutations define a homogeneous subgroup of hepatocellular carcinoma with fibrolamellar-like features and activated PKA. J Hepatol. 2019;

69. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. 2014;15:550.

70. Hyvärinen A. Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans Neural Netw. 1999;10:626–34.

71. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9:215–6.

72. Burger L, Gaidatzis D, Schübeler D, Stadler MB. Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. 2013;41:e155.

73. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489:57–74.

74. Nault JC, Mallet M, Pilati C, Calderaro J, Bioulac-Sage P, Laurent C, et al. High frequency of telomerase reverse-transcriptase promoter somatic mutations in hepatocellular carcinoma and preneoplastic lesions. Nat Commun. 2013;4:2218.

75. Petitprez F, Meunier L, Letouzé E, Hoshida Y, Villanueva A, Llovet J, et al. MS.liverK: an R package for transcriptome-based computation of molecular subtypes and functional signatures in liver cancer. bioRxiv. 2019;540005.

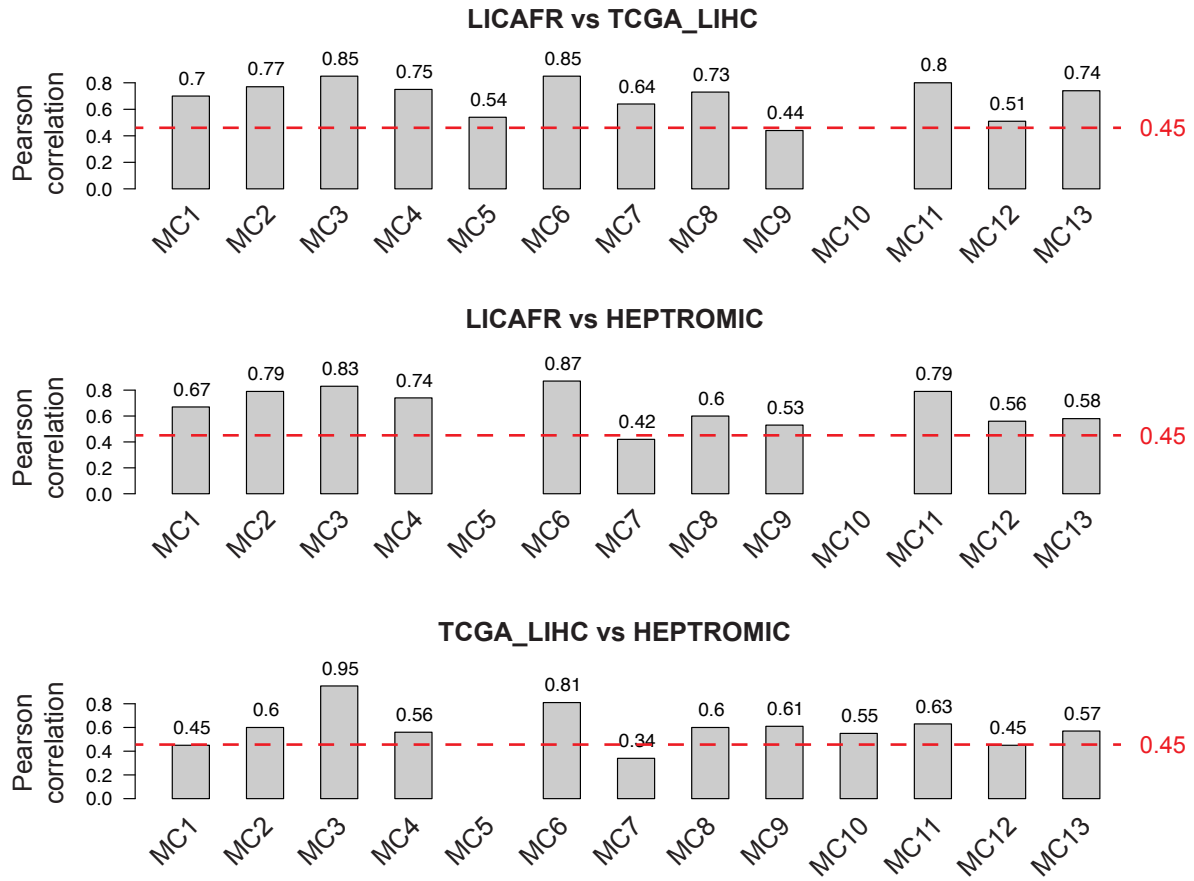
76. Nault J-C, Martin Y, Caruso S, Hirsch TZ, Bayard Q, Calderaro J, et al. Clinical Impact of Genomic Diversity From Early to Advanced Hepatocellular Carcinoma. Hepatology. 2019;

77. Caruso S, Calatayud A-L, Pilet J, La Bella T, Rekik S, Imbeaud S, et al. Analysis of Liver Cancer Cell Lines Identifies Agents With Likely Efficacy Against Hepatocellular Carcinoma and Markers of Response. Gastroenterology. 2019;157:760–76.

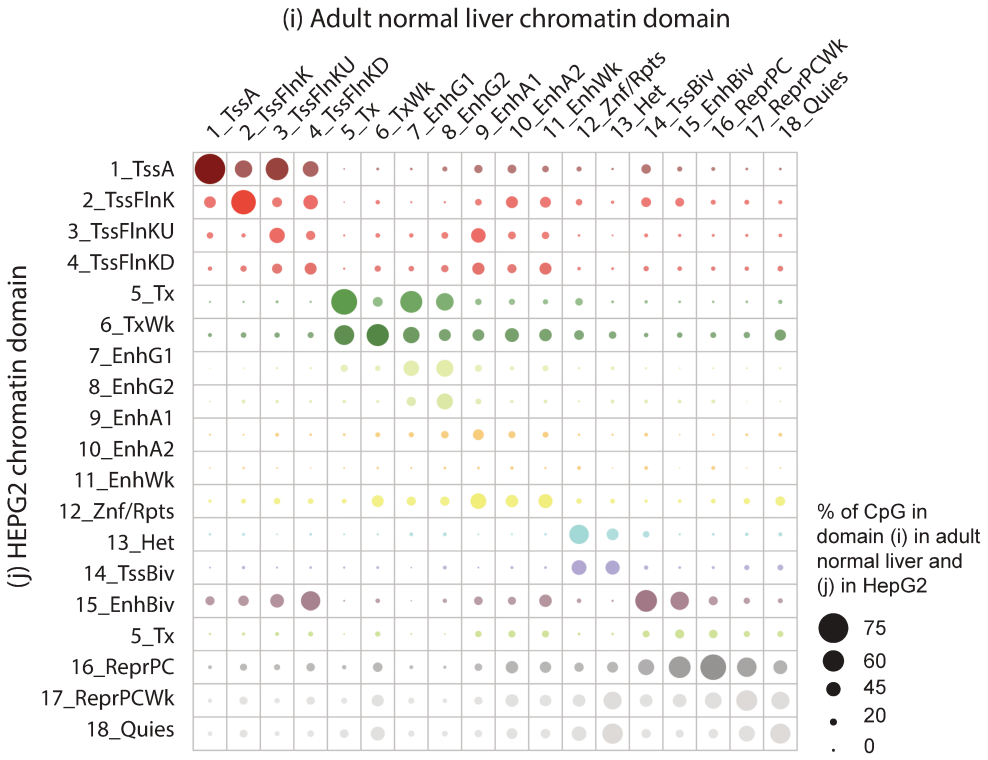
78. Becht E, Giraldo NA, Lacroix L, Buttard B, Elarouci N, Petitprez F, et al. Estimating the population abundance of tissue-infiltrating immune and stromal cell populations using gene expression. Genome Biology. 2016;17:218.

79. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA. 2005;102:15545–50.

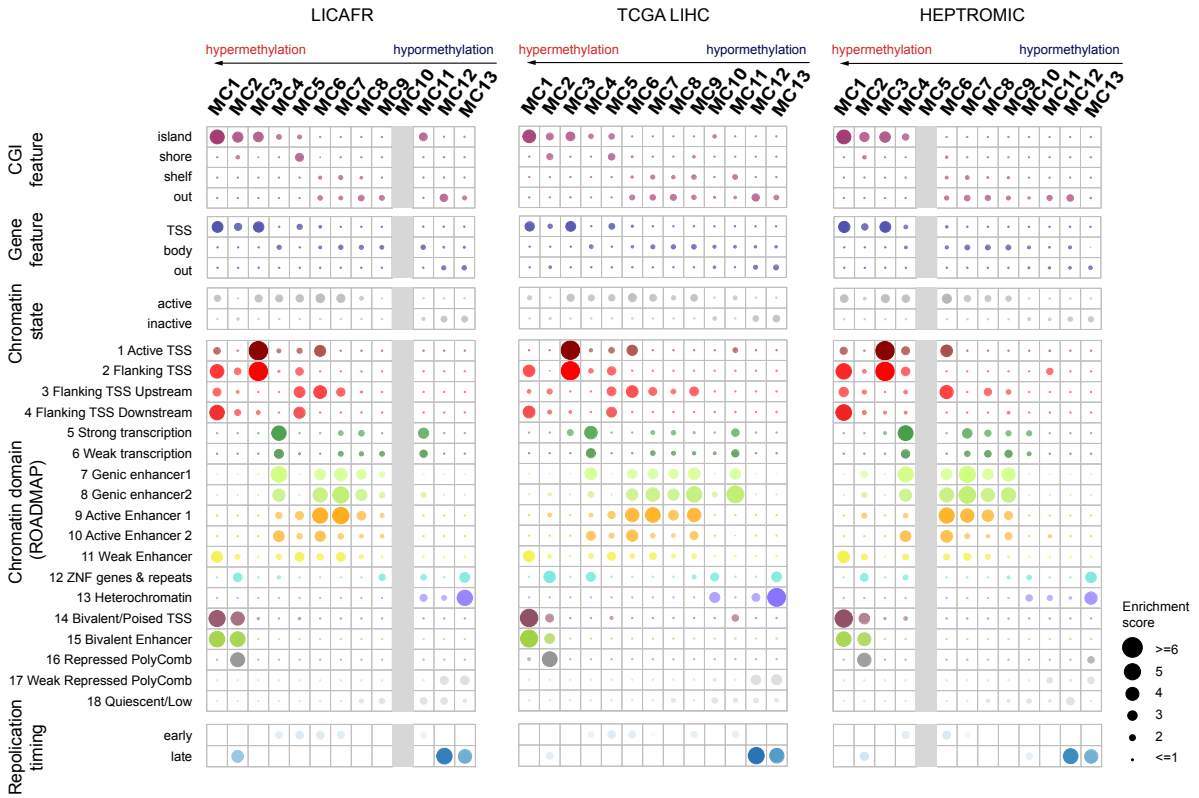
Supplementary Materials



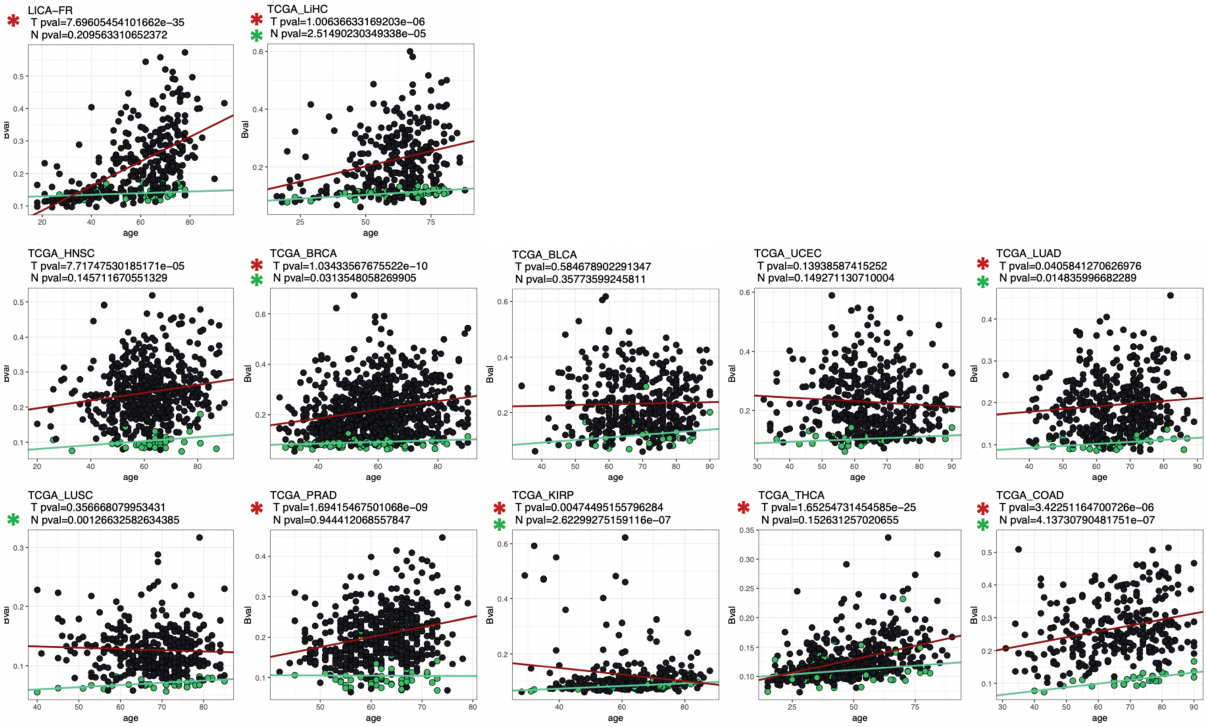
Supplementary Fig. 1: Reproducibility of methylation components across the 3 data sets. Pearson correlation coefficients were used to link each methylation component (MC) extracted in one cohort with its closest equivalent in the other two cohorts. The figure displays the Pearson correlation scores between cohorts for the 13 MCs identified in at least two cohorts.



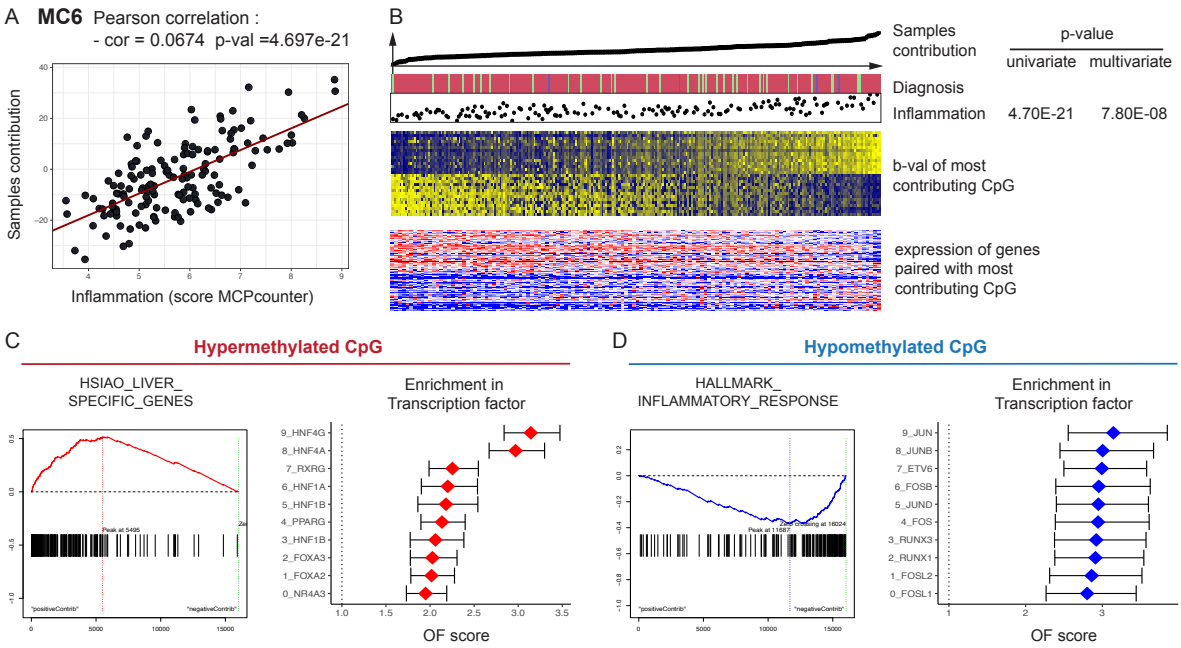
Supplementary Fig. 2: Correlation matrix of ROADMAP chromatin domains between normal liver and the cancerous HepG2 cell lines. Each column *i* indicates the proportion of CpG sites located within the i^{th} chromatin state in adult liver that are classified across of the 18 possible chromatin states in the HepG2 cell line.



Supplementary Fig. 3: Association of methylation components with CpG island- and gene-based features, chromatin states and replication timing in the LICA-FR, TCGA-LIHC and HEPTROMIC cohorts. The most contributive CpG sites of each component were extracted in each cohort, and the enrichment of these CpG sites across CpG-island features, gene-based features, chromatin states and replication deciles are shown below.



Supplementary Fig. 4: Correlation of MC1 most contributive CpG methylation with age across diverse normal and tumor tissues. The average methylation level of MC1 most contributive CpG sites was calculated in diverse tumor and normal tissues from TCGA. Linear regression was used to estimate the coefficient and significance of methylation changes separately in tumor and normal samples.



Supplementary Fig. 5: Methylation component 6 is a signature of tumor infiltration by immune cells. (a) Sample contribution to MC6 is strongly correlated to the level of infiltration by immune cells evaluated from RNA-seq data using MCPCounter tool. **(b)** Heatmap representing the methylation of the most contributive CpG sites and the expression of paired genes identified with ELMER. Samples are ordered according to the activity to the component and the level of inflammation estimated by MCPCounter is represented below. **(c)** Left: Gene set enrichment analysis of genes paired with CpG sites hypermethylated in samples with the highest activity of MC6 (strong immune infiltrate). Right: Transcription factor binding motif enrichment around CpG sites hypermethylated in samples with the highest activity of MC6. **(d)** Left: Gene set enrichment analysis of genes paired with CpG sites hypomethylated in samples with the highest activity of MC6 (strong immune infiltrate). Right: Transcription factor binding motif enrichment around CpG sites hypomethylated in samples with the highest activity of MC6.

A

CHCID	Diagnosis	Type	Gender	Age	Geographic Origin	Without. etiology	Alcohol. Intake	Hepatitis.B	Hepatitis.C	Hemochromatosis	Metabolic. syndrome	Tobacco	BMI	normal. Liver.Histology	Steatosis. non.tumoral.liver	Number. Nodules	nodule.size.mm	Largest.nodule.diameter	Edmonson.grade	Vascular.Invasion	Differenciation.WHO	Child.Pugh	BCLC	Inflammation
CHC018T	HCC	T	F	35	Africa	no	no	yes	no	no	no	no	NA	F2-F3	NA	UNIQUE	170	>50	III-IV	yes	medium	A	A	NA
CHC229T	HCC	T	F	65	Europe	no	no	no	yes	no	no	no	NA	F4	NA	UNIQUE	55	>50	III-IV	yes	medium	A	A	NA
CHC231T	HCC	T	M	66	Europe	no	yes	no	no	no	no	NA	NA	F4	NA	<=3	NA	<=50	I-II	no	good	A	A	NA
CHC013T	HCC	T	M	63	Europe	no	no	no	yes	no	no	no	NA	F4	NA	UNIQUE	30	<=50	I-II	yes	good	A	A	NA
CHC441T	HCC	T	M	77	Europe	no	no	no	no	yes	no	NA	NA	F0-F1	3	UNIQUE	40	<=50	III-IV	no	weak	A	A	no
CHC333T	HCC	T	M	73	Europe	no	yes	no	no	no	no	NA	NA	F4	NA	UNIQUE	42	<=50	I-II	no	good	NA	A	NA
CHC239T	HCC	T	F	21	Africa	no	no	yes	no	no	no	no	NA	F2-F3	NA	UNIQUE	100	>50	I-II	yes	good	A	A	NA
CHC399T	HCC	T	M	67	Europe	no	no	no	no	yes	no	NA	NA	F2-F3	NA	<=3	NA	<=50	I-II	no	medium	A	A	NA
CHC014T	HCC	T	M	30	Africa	no	no	yes	no	no	no	no	NA	F2-F3	NA	UNIQUE	130	>50	III-IV	yes	medium	A	A	NA
CHC043T	HCC	T	M	56	Asia	no	no	yes	no	no	no	NA	NA	F2-F3	NA	UNIQUE	50	<=50	III-IV	no	medium	A	A	NA
CHC037T	HCC	T	M	51	Africa	yes	no	no	no	no	no	NA	NA	F0-F1	NA	UNIQUE	120	>50	I-II	no	good	A	A	NA
CHC339T	HCC	T	F	26	Africa	no	no	yes	no	no	no	NA	NA	F2-F3	NA	UNIQUE	100	>50	I-II	yes	good	A	C	NA
CHC245T	HCC	R	M	64	Europe	no	no	yes	no	no	no	NA	NA	F4	NA	>=4	10	<=50	I-II	no	good	A	O	NA
CHC253T	HCC	T	M	67	Europe	no	no	no	no	no	yes	NA	[25;30]	F4	2	>=4	NA	>50	III-IV	yes	weak	A	C	NA
CHC158T	HCC	T	M	65	Europe	no	yes	yes	no	no	no	NA	NA	F4	NA	<=3	NA	<=50	I-II	no	good	A	A	NA
CHC445T	HCC	T	M	55	Europe	no	yes	no	yes	no	no	NA	NA	F4	NA	<=3	NA	<=50	I-II	no	NA	A	O	NA
CHC080T	HCC	T	M	43	Europe	no	yes	yes	yes	no	no	yes	NA	F4	NA	UNIQUE	35	<=50	III-IV	no	medium	A	A	NA
CHC335T	HCC	T	M	68	Europe	no	no	yes	no	no	no	NA	NA	F2-F3	NA	UNIQUE	160	>50	I-II	yes	good	A	A	NA
CHC230T	HCC	T	M	70	Europe	yes	no	no	no	no	no	no	NA	F0-F1	NA	UNIQUE	160	>50	I-II	no	good	A	A	NA
CHC228T	HCC	T	M	48	Europe	yes	no	no	no	no	no	NA	NA	F0-F1	NA	UNIQUE	145	>50	III-IV	yes	medium	A	C	NA
CHC010T	HCC	T	F	18	Europe	no	no	yes	no	no	no	NA	NA	F2-F3	NA	UNIQUE	100	>50	III-IV	no	medium	A	A	NA
CHC137T	HCC	T	M	71	Europe	no	no	yes	no	no	no	NA	NA	F2-F3	NA	UNIQUE	35	<=50	III-IV	no	medium	A	A	NA
CHC205T	HCC	R	M	46	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	100	>50	III-IV	no	weak	A	A	no
CHC218T	HCC	T	M	69	Europe	no	no	no	no	no	yes	no	[25;30]	F0-F1	1	NA	NA	>50	III-IV	yes	medium	NA	C	NA
CHC081T	HCC	T	F	76	Asia	no	no	yes	no	no	no	no	NA	F4	NA	UNIQUE	90	>50	I-II	no	good	A	A	NA
CHC031T	HCC	T	M	67	Europe	no	yes	no	no	no	no	yes	NA	F2-F3	NA	UNIQUE	16	<=50	I-II	no	good	A	O	NA
CHC242T	HCC	T	M	70	Europe	yes	no	no	no	no	no	no	NA	F0-F1	NA	UNIQUE	150	>50	I-II	no	good	A	A	NA
CHC059T	HCC	T	M	40	Europe	no	yes	no	no	no	no	NA	NA	F0-F1	NA	<=3	NA	>50	III-IV	yes	weak	A	C	NA
CHC220T	HCC	T	M	73	Europe	yes	no	no	no	no	NA	NA	NA	F0-F1	NA	UNIQUE	35	<=50	I-II	no	good	A	NA	NA
CHC206T	HCC	T	M	64	Europe	no	no	yes	no	no	no	NA	NA	F4	NA	<=3	17	<=50	III-IV	no	medium	A	B	no
CHC152T	HCC	T	M	64	Europe	no	no	yes	no	no	no	yes	NA	F4	NA	<=3	NA	<=50	NA	no	good	A	C	NA
CHC046T	HCC	T	M	61	Europe	no	no	yes	no	no	no	NA	NA	F4	NA	>=4	NA	>50	III-IV	yes	medium	A	C	NA
CHC211T	HCC	T	M	69	Europe	no	yes	no	no	no	no	NA	NA	F0-F1	NA	UNIQUE	80	>50	I-II	yes	medium	A	C	no
CHC437T	HCC	T	M	59	Europe	no	yes	no	no	no	yes	yes	NA	F4	0	UNIQUE	50	<=50	I-II	no	good	A	A	no
CHC725T	HCC	T	M	60	Europe	no	no	yes	no	no	no	no	NA	F4	NA	UNIQUE	27	<=50	III-IV	no	medium	A	A	NA
CHC317T	HCC	T	F	69	Europe	no	no	no	yes	no	no	NA	NA	F4	NA	>=4	NA	<=50	III-IV	no	medium	NA	A	NA
CHC789T	HCC	T	M	54	Europe	no	no	no	no	yes	no	NA	NA	F2-F3	0	<=3	NA	>50	I-II	yes	medium	A	B	no
CHC195T	HCC	T	M	71	Europe	no	yes	no	no	no	no	yes	[25;30]	F0-F1	1	UNIQUE	80	>50	I-II	no	medium	A	A	yes
CHC1196T	HCC	T	M	27	Africa	no	no	yes	no	no	no	yes	[<=25]	F2-F3	0	UNIQUE	90	>50	III-IV	yes	medium	A	A	yes
CHC398T	HCC	T	M	50	Africa	no	no	yes	no	no	no	NA	NA	F4	NA	<=3	NA	<=50	I-II	no	NA	NA	O	NA
CHC1010T	HCC	T	F	53	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	100	>50	III-IV	no	medium	A	A	yes
CHC1035T	HCC	T	M	68	Europe	no	yes	yes	no	no	no	NA	NA	F2-F3	1	UNIQUE	75	>50	I-II	no	medium	A	A	no
CHC1040T	HCC	T	M	73	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	0	UNIQUE	160	>50	III-IV	yes	medium	B	A	no
CHC1041T	HCC	T	M	69	Europe	yes	no	no	no	no	no	no	[<=25]	F0-F1	0	UNIQUE	100	>50	I-II	no	medium	A	O	no
CHC1044T	HCC	T	M	78	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	0	UNIQUE	16	<=50	III-IV	yes	weak	B	O	no
CHC1052T	HCC	T	M	75	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	1	UNIQUE	130	>50	III-IV	yes	medium	A	A	no
CHC1055T	HCC	T	M	68	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	2	<=3	200	>50	III-IV	yes	weak	A	B	yes
CHC1060T	HCC	T	M	66	Europe	no	no	no	no	no	yes	NA	NA	F4	2	UNIQUE	30	<=50	III-IV	no	medium	A	A	yes
CHC1061T	HCC	T	F	79	Europe	no	no	no	no	no	yes	NA	[>=30]	F0-F1	2	UNIQUE	150	>50	I-II	yes	medium	B	A	yes
CHC1062T	HCC	T	M	65	Europe	no	no	no	no	yes	no	no	[25;30]	F2-F3	1	UNIQUE	30	<=50	I-II	yes	medium	A	O	no
CHC1065T	HCC	T	M	77	Europe	yes	no	no	no	no	no	no	[<=25]	F0-F1	2	UNIQUE	35	<=50	I-II	yes	good	A	C	no
CHC1146T	HCC	T	M	60	Europe	no	yes	no	no	no	no	no	[<=25]	F2-F3	1	>=4	170	>50	III-IV	yes	medium	A	C	no
CHC1154T	HCC	T	M	43	Africa	no	no	yes	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	130	>50	I-II	yes	good	A	A	no
CHC1162T	HCC	T	M	60	Europe	no	yes	no	no	no	no	no	[25;30]	F4	0	<=3	NA	>50	I-II	no	good	A	A	NA
CHC1192T	HCC	T	M	40	Africa	no	no	yes	no	no	no	NA	[<=25]	F4	0	UNIQUE	70	>50	III-IV	yes	medium	A	A	no
CHC1199T	HCC	T	M	62	Europe	yes	no	no	no	no	no	no	[<=25]	F0-F1	1	UNIQUE	140	>50	I-II	yes	medium	A	C	no
CHC1201T	HCC	T	M	73	Europe	no	yes	no	no	no	yes	no	[>=30]	F4	3	UNIQUE	60	>50	I-II	no	medium	A	A	no
CHC255T	FLC	T	F	39	Europe	yes	no	no	no	no	no	yes	NA	F0-F1	2	NA	NA	>50	NA	no	good	A	NA	NA
CHC320T	HCC	T	M	65	Europe	no	yes	no	yes	no	no	NA	NA	F4	0	>=4	35	<=50	III-IV	no	medium	NA	B	no
CHC334T	FLC	T	F	24	Europe	yes	no	no	no	no	no	NA	NA	F0-F1	0	UNIQUE	90	>50	NA	no	NA	NA	NA	NA
CHC412T	FLC	T	F	51	NA	NA	NA	no	NA	NA	NA	NA	NA	NA	NA	>=4	NA	>50	NA	yes	NA	NA	NA	NA
CHC429T	HCC	T	F	64	Europe	yes	no	no	no	no	no	no	[25;30]	F0-F1	0	<=3	45	<=50	III-IV	yes	medium	A	A	no
CHC442T	FLC	T	F	27	Europe	yes	no	no	no	no	no	yes	NA	F0-F1	0	UNIQUE	100	>50	NA	yes	NA	NA	NA	NA
CHC451T	HCC	T	M	75	Europe	no	yes	no	yes	no	no	NA	NA	F2-F3	2	UNIQUE	30	<=50	I-II	no	good	A	A	NA
CHC613T	HCC	T	M	70	Europe	no	yes	no	no	no	yes	NA	[>=30]	F0-F1	1	UNIQUE	180	>50	I-II	yes	good	A	A	no
CHC614T	HCC	T	M	61	Europe	no	no	no	no	no	yes	no	[25;30]	F0-F1	1	UNIQUE	30	<=50	III-IV	yes	weak	A	A	no
CHC703T	HCC	T	M	55	Europe	no	yes	no	no	yes	no	NA	NA	F4	1	UNIQUE	26	<=50	III-IV	yes	weak	A	A	yes
CHC734T	HCC	T	M	65	Europe	no	yes	no	no	no	no	no	NA	F0-F1	1	<=3	80	>50	III-IV	no	medium	A	B	yes
CHC736T	HCC	T	M	77	Europe	no	no	yes	no	no	no	no	[<=25]	F0-F1	0	UNIQUE	160	>50	III-IV	yes	medium	B	A	no
CHC793T	HCC	T	M	61	Europe	no	no	no	no	yes	no	no	[25;30]	F0-F1	1	UNIQUE	80	>50	III-IV	yes	weak	A	A	no
CHC796T	HCC	T	M	76	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	1	UNIQUE	48	<=50	I-II	no	medium	A	A	no
CHC799T	HCC	R	F	67	Europe	yes	no	no	no	no	no	no	[<=25]	F0-F1	1	UNIQUE	45	<=50	I-II	yes	good	A	A	no
CHC805T	HCC	T	M	58	Europe	yes	no	no	no	no	no	NA	NA	F0-F1	1	UNIQUE	130	>50	I-II	no	good	A	A	yes
CHC884T	HCC	T	M	75	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	1	UNIQUE	130	>50	III-IV	yes	weak	B	C	no
CHC889T	HCC	T	M	71	Europe	no	no	no	no	yes	yes	NA	NA	F2-F3	0	<=3	NA	>50	I-II	yes	medium	B	C	yes
CHC891T	HCC	T	F	73	Europe	no	no	no	no	no	no	NA	NA	F4	1	UNIQUE	45	<=50	III-IV	yes	weak	A	A	yes
CHC892T	HCC	T	F	72	Europe	yes	no	no	no	no	no	no	[<=25]	F0-F1	1	UNIQUE	55	>50	I-II	no	good	A	A	yes
CHC896T	HCC	T	M	61	Europe	no	yes	no	no	no	yes	yes	[25;30]	F0-F1	NA	UNIQUE	110	>50	I-II	yes	good	A	C	NA
CHC898T	HCC	T	M	71	Europe	no	no	no	no	yes	no	no	NA	F2-F3	0	UNIQUE	80	>50	III-IV	yes	weak	A	A	no
CHC909T	HCC	T	M	70	Europe	yes	no	no	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	210	>50	III-IV	yes	medium	A	C	no
CHC912T	HCC	T	M	78	Europe	no	yes	no	yes	no	no	yes	[<=25]	F0-F1	0	UNIQUE	60	>50	III-IV	yes	weak	A	C	no
CHC983T	HCC	T	M	54	Europe	no	yes	no	no	no	no	NA	NA	F4	2	<=3	50	<=50	I-II	no	medium	A	B	yes
CHC051T	HCC	T	F	69	Europe	no	no	no	yes	no	no	NA	NA	F4	NA	UNIQUE	60	>50	III-IV	no	medium	A	A	NA
CHC121T	HCC	T	M	67	Europe	no	yes	no	no	no	yes	yes	[>=30]	F0-F1	0	UNIQUE	120	>50	I-II	no	medium	A	A	no
CHC155T	HCC	T	M	62	Europe	no	yes	no	no	no	yes	yes	[25;30]	F4	1	UNIQUE	35	<=50	I-II	no	good	A	A	NA
CHC258T	HCC	T	M	56	Europe	no	no	no	no	no	yes	NA	NA	F0-F1	0	UNIQUE	100	>50	I-II	no	good	A	A	NA
CHC301T	HCC	T	M	78	Europe	no	no	no	no	no	yes	NA	NA	F2-F3	0	<=3	NA	<=50	III-IV	no	medium	A	B	no
CHC302T	HCC	T	M	72	Europe	no	no	no	yes	no	no	NA	NA	F2-F3	1	UNIQUE	45	<=50	I-II	no	good	A	A	no
CHC303T	HCC	T	M	74	Europe	no	yes	no	no	no	no	NA	NA	F4	1	UNIQUE	90	>50	III-IV	yes	medium	NA	A	no
CHC306T	HCC	T	M	68	Europe	no	no	no	yes	no	no	NA	NA	F4	0	UNIQUE	20	<=50	I-II	no	good	NA	A	NA
CHC307T	HCC	T	M	54	Europe	no	yes	no	no	no	no	NA	NA	F4	1	<=3	30	<=50	III-IV	yes	medium	NA	A	NA
CHC326T	early.HCC	T	M	49	Europe	no	no	yes	no	no	no	NA	NA	F4	0	<=3	NA	<=50	I-II	no	good	NA	A	NA
CHC327T	HCC	T	M	63	Europe	no	no	no	yes	no	no	NA	NA	F4	1	>=4	25	<=50	I-II	no	good	NA	A	NA
CHC434T	HCC	T	F	71	Europe	yes	no	no	no	no	no	no	[<=25]	F0-F1	0	UNIQUE	80	>50	III-IV	no	weak	A	A	no
CHC609T	HCC	T	M	60	Europe	no	yes	yes	no	no	yes	yes	[>=30]	F2-F3	3	<=3	50	<=50	III-IV	yes	medium	A	B	no
CHC197T	HCC	T	M	73	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	NA	>=4	NA	>50	III-IV	yes	medium	A	B	no
CHC304T	HCC	T	M	77	Europe	no	yes	no	no	no	no	yes	NA	F0-F1	0	UNIQUE	180	>50	III-IV	yes	medium	A	C	no
CHC313T	HCC	T	F	43	Europe	no	no	no	yes	no	no	no	[<=25]	F0-F1	0	UNIQUE	130	>50	III-IV	yes	weak	A	A	no
CHC314T	HCC	T	M	71	Europe	no	yes	no	yes	no	no	NA	NA	F2-F3	0	UNIQUE	45	<=50	I-II	no	good	A	A	no
CHC322T	HCC	T	M	74	Europe	no	yes	no	no	no	no	NA	NA	F4	2	UNIQUE	40	<=50	III-IV	no	weak	A	A	no
CHC433T	HCC	T	M	70	Europe	no	yes	no	no	no	yes	NA	[>=30]	F0-F1	1	UNIQUE	180	>50	I-II	yes	good	A	A	no
CHC794T	HCC	T	M	73	Europe	no	no	no	no	yes	no	no	[<=25]	F0-F1	0	UNIQUE	160	>50	III-IV	yes	weak	A	C	yes
CHC798T	HCC	T	M	73	Europe	no	yes	no	no	no	yes	yes	[<=25]	F0-F1	0	UNIQUE	95	>50	I-II	no	good	A	A	no
CHC961T	HCC	T	M	57	Europe	no	yes	no	no	no	no	no	[<=25]	F0-F1	0	<=3	190	>50	III-IV	yes	medium	NA	B	no
CHC060T	HCC	T	M	68	Europe	no	yes	no	no	no	no	NA	NA	F4	0	UNIQUE	28	<=50	III-IV	yes	weak	A	C	no
CHC1137T	HCC	T	M	57	Europe	no	no	no	yes	no	no	no	[<=25]	F4	0	<=3	NA	>50	III-IV	yes	medium	A	B	no
CHC1568T	HCC	T	M	71	Asia	no	no	yes	no	no	no	yes	[25;30]	F4	1	UNIQUE	30	<=50	III-IV	yes	medium	A	A	yes

CHCID	Diagnosis	Type	Gender	Age	Geographic Origin	Without. etiology	Alcohol. Intake	Hepatitis.B	Hepatitis.C	Hemochromatosis	Metabolic. syndrome	Tobacco	BMI	normal. Liver.Histology	Steatosis. non.tumoral.liver	Number. Nodules	nodule.size.mm	Largest.nodule.diameter	Edmonson.grade	Vascular.Invasion	Differenciation.WHO	Child.Pugh	BCLC	Inflammation
CHC1604T	HCC	T	M	57	Europe	no	no	no	no	yes	no	yes	[<=25]	F2-F3	3	UNIQUE	18	<=50	III-IV	no	medium	A	O	yes
CHC1717T	HCC	T	M	50	Africa	no	no	yes	no	no	no	NA	[<=25]	F4	1	UNIQUE	55	>50	I-II	yes	medium	A	C	yes
CHC1763T	HCC	T	M	75	Europe	no	yes	no	no	no	yes	yes	[<=25]	F0-F1	0	UNIQUE	60	>50	III-IV	no	weak	A	B	no
CHC097T	HCC	T	M	56	Europe	no	no	no	yes	no	no	no	[<=25]	F0-F1	2	<=3	NA	>50	I-II	yes	medium	A	B	no
CHC1148T	HCC	T	M	69	Europe	no	yes	no	no	yes	no	yes	[<=25]	F0-F1	0	UNIQUE	90	>50	I-II	yes	medium	A	A	no
CHC1744T	HCC	T	M	50	Africa	no	no	yes	yes	no	no	no	[25;30]	F4	1	UNIQUE	70	>50	III-IV	yes	medium	A	B	no
CHC1152T	HCC	T	M	63	Europe	no	yes	no	no	no	no	yes	[25;30]	F4	1	UNIQUE	140	>50	III-IV	yes	medium	A	C	no
CHC1205T	HCC	T	M	72	Africa	no	yes	yes	no	no	yes	yes	[25;30]	F4	1	<=3	60	>50	III-IV	yes	medium	A	C	no
CHC1616T	HCC	T	F	77	Europe	no	no	no	yes	no	no	no	[25;30]	F4	0	UNIQUE	100	>50	III-IV	no	medium	A	A	no
CHC1720T	HCC	R	M	81	Europe	no	no	no	yes	no	no	yes	[25;30]	F4	2	<=3	100	>50	III-IV	no	medium	A	A	no
CHC1745T	HCC	T	F	69	Europe	no	no	no	yes	no	yes	yes	[>=30]	F4	2	UNIQUE	60	>50	III-IV	yes	medium	B	B	no
CHC432T	HCC	T	M	70	Europe	no	yes	no	no	no	no	yes	NA	F2-F3	1	UNIQUE	70	>50	I-II	yes	medium	A	A	no
CHC1207T	HCC	T	M	60	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	210	>50	I-II	yes	good	A	A	no
CHC1626T	HCC	T	M	75	Europe	no	yes	no	no	no	yes	no	[>=30]	F0-F1	1	UNIQUE	100	>50	III-IV	yes	weak	A	A	yes
CHC1725T	HCC	T	F	83	Africa	no	no	no	yes	no	no	yes	[<=25]	F2-F3	0	UNIQUE	60	>50	I-II	yes	medium	A	B	no
CHC1746T	HCC	T	M	75	Europe	no	no	no	no	yes	no	no	[<=25]	F2-F3	0	UNIQUE	40	<=50	III-IV	yes	medium	A	B	no
CHC1209T	HCC	T	M	66	Europe	no	no	no	yes	no	no	yes	[<=25]	F2-F3	0	UNIQUE	140	>50	I-II	yes	medium	A	A	no
CHC1594T	HCC	T	F	76	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	1	UNIQUE	100	>50	I-II	yes	good	A	A	no
CHC1629T	HCC	T	M	64	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	1	<=3	70	>50	I-II	yes	good	A	B	no
CHC1731T	HCC	T	F	55	Europe	yes	no	no	no	no	no	no	[25;30]	F0-F1	2	UNIQUE	110	>50	I-II	no	good	A	B	no
CHC1747T	HCC	T	M	54	Europe	no	yes	no	yes	no	no	yes	[<=25]	F4	2	UNIQUE	40	<=50	III-IV	yes	medium	A	C	no
CHC801T	HCC	T	M	78	Europe	no	no	no	no	no	yes	no	[<=25]	F0-F1	2	UNIQUE	50	<=50	I-II	no	medium	A	A	no
CHC1211T	HCC	T	F	32	Africa	no	no	yes	no	no	no	no	[25;30]	F0-F1	0	UNIQUE	130	>50	III-IV	no	medium	A	A	no
CHC1700T	HCC	T	M	62	Asia	no	no	yes	no	no	no	yes	[<=25]	F4	2	<=3	NA	<=50	I-II	yes	medium	A	A	yes
CHC1732T	HCC	T	M	49	Europe	no	yes	no	yes	no	no	yes	[<=25]	F4	0	UNIQUE	60	>50	III-IV	yes	weak	A	C	yes
CHC1749T	HCC	T	M	66	Asia	no	no	yes	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	150	>50	III-IV	yes	medium	A	B	no
CHC1183T	HCC	T	M	60	Europe	no	no	no	no	yes	no	no	[25;30]	F2-F3	0	UNIQUE	180	>50	I-II	no	good	A	A	no
CHC1597T	HCC	T	M	41	Europe	no	yes	yes	no	no	no	yes	[<=25]	F4	1	UNIQUE	105	>50	III-IV	yes	weak	A	C	yes
CHC1704T	HCC	T	M	43	Africa	no	no	yes	no	no	no	no	[<=25]	F2-F3	0	UNIQUE	140	>50	III-IV	yes	medium	A	B	no
CHC1734T	HCC	T	M	76	Europe	no	no	no	no	no	yes	yes	[<=25]	F0-F1	2	UNIQUE	40	<=50	III-IV	no	medium	A	B	yes
CHC1185T	HCC	T	M	53	Asia	no	yes	yes	no	no	no	yes	[<=25]	F4	2	UNIQUE	30	<=50	III-IV	no	medium	A	A	no
CHC1531T	HCC	T	M	78	Europe	no	yes	no	no	no	no	yes	NA	F0-F1	1	UNIQUE	60	>50	I-II	yes	medium	A	A	no
CHC1598T	HCC	T	F	76	Europe	no	no	yes	no	no	no	no	[<=25]	F0-F1	0	UNIQUE	90	>50	III-IV	yes	medium	A	A	no
CHC1705T	HCC	T	M	83	Europe	no	no	no	no	no	yes	yes	[>=30]	F0-F1	3	UNIQUE	90	>50	I-II	yes	good	A	B	no
CHC1736T	HCC	T	M	58	Europe	no	no	yes	no	yes	no	yes	[<=25]	F4	0	UNIQUE	45	<=50	III-IV	yes	weak	A	C	no
CHC1751T	HCC	T	M	52	Europe	no	yes	no	no	no	no	yes	[25;30]	F4	1	UNIQUE	80	>50	III-IV	yes	medium	A	B	no
CHC1186T	HCC	T	M	56	Africa	no	no	no	yes	no	no	no	[<=25]	F2-F3	1	<=3	85	>50	III-IV	yes	medium	A	A	no
CHC1539T	HCC	T	M	45	Europe	no	no	no	yes	no	no	NA	NA	F4	0	UNIQUE	32	<=50	III-IV	yes	medium	A	A	no
CHC1600T	HCC	T	M	69	Europe	no	yes	no	no	no	no	yes	[25;30]	F0-F1	0	UNIQUE	80	>50	I-II	yes	good	A	A	yes
CHC1708T	HCC	T	M	56	NA	NA	NA	NA	NA	NA	NA	NA	NA	F2-F3	2	NA	NA	>50	I-II	no	good	A	B	no
CHC1028T	HCC	T	M	62	Europe	no	yes	no	no	no	no	NA	NA	F4	0	UNIQUE	80	>50	I-II	yes	good	B	A	no
CHC1189T	HCC	T	M	62	Europe	no	yes	no	no	yes	no	no	[25;30]	F4	3	UNIQUE	60	>50	III-IV	yes	medium	A	A	yes
CHC1545T	HCC	T	M	77	Europe	no	yes	no	yes	no	no	yes	NA	F4	2	UNIQUE	40	<=50	I-II	no	good	A	A	NA
CHC1739T	HCC	T	M	55	Europe	no	yes	no	no	no	yes	yes	[25;30]	F4	2	UNIQUE	50	<=50	III-IV	no	medium	A	B	yes
CHC1754T	HCC	T	M	34	Africa	no	no	yes	no	no	no	no	[25;30]	F2-F3	0	UNIQUE	170	>50	III-IV	yes	medium	A	C	no
CHC1190T	HCC	T	F	68	Europe	no	yes	no	yes	no	no	yes	[<=25]	F2-F3	0	UNIQUE	22	<=50	I-II	yes	good	A	A	yes
CHC1602T	HCC	T	M	71	Europe	yes	no	no	no	no	no	yes	[25;30]	F0-F1	1	UNIQUE	75	>50	I-II	yes	good	A	A	no
CHC1741T	HCC	T	M	57	Europe	no	yes	no	no	no	no	yes	[<=25]	F4	0	UNIQUE	32	<=50	I-II	no	medium	A	B	no
CHC1756T	HCC	T	M	73	Europe	no	yes	yes	no	no	no	no	[25;30]	F4	1	<=3	45	<=50	III-IV	yes	weak	A	B	yes
CHC1079T	HCC	T	M	60	Europe	no	no	yes	no	no	no	NA	NA	F2-F3	1	UNIQUE	35	<=50	III-IV	yes	weak	A	A	yes
CHC1566T	HCC	T	M	68	Europe	no	yes	no	yes	no	no	yes	[25;30]	F4	1	UNIQUE	22	<=50	I-II	yes	medium	A	A	no
CHC1603T	HCC	T	M	78	Europe	no	yes	no	no	no	no	yes	[<=25]	F4	1	UNIQUE	50	<=50	III-IV	yes	weak	A	A	no
CHC1715T	HCC	T	M	72	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	60	>50	I-II	no	medium	A	B	no
CHC1742T	HCC	T	M	67	Europe	no	no	no	no	yes	no	yes	[<=25]	F0-F1	3	UNIQUE	33	<=50	III-IV	yes	weak	A	B	yes
CHC1757T	HCC	T	M	41	Europe	no	yes	no	no	no	yes	no	[>=30]	F4	2	UNIQUE	12	<=50	I-II	no	good	A	A	no
CHC2025T	HCC	T	F	58	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	2	UNIQUE	110	>50	III-IV	yes	weak	A	A	NA
CHC2112T	HCC	T	F	48	Europe	no	no	no	no	no	yes	no	[>=30]	F0-F1	1	UNIQUE	190	>50	III-IV	yes	medium	A	A	yes
CHC2029T	HCC	T	M	74	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	2	UNIQUE	60	>50	III-IV	yes	medium	A	A	no
CHC2113T	HCC	T	M	61	Europe	no	yes	no	no	no	no	no	[<=25]	F0-F1	0	UNIQUE	90	>50	III-IV	no	medium	A	A	yes
CHC2215T	HCC	T	M	65	Europe	no	no	no	no	yes	yes	yes	[25;30]	F0-F1	3	UNIQUE	20	<=50	III-IV	no	medium	A	A	no
CHC2034T	HCC	T	M	80	Europe	no	yes	no	no	no	yes	yes	[<=25]	F0-F1	0	UNIQUE	60	>50	III-IV	yes	medium	A	C	no
CHC2115T	HCC	T	M	75	Europe	no	yes	no	no	no	yes	yes	[25;30]	F0-F1	0	UNIQUE	100	>50	III-IV	yes	medium	A	C	no
CHC2127T	HCC	T	M	57	Europe	no	no	no	yes	no	no	yes	[25;30]	F0-F1	0	UNIQUE	100	>50	III-IV	yes	medium	A	B	no
CHC2043T	HCC	T	F	21	Europe	yes	no	no	no	no	no	no	[25;30]	F0-F1	0	UNIQUE	50	<=50	III-IV	no	medium	A	A	no
CHC2128T	HCC	T	F	53	Europe	no	no	no	no	no	yes	yes	[>=30]	F0-F1	1	UNIQUE	200	>50	I-II	yes	medium	A	A	no
CHC2048T	HCC	T	M	65	Europe	no	yes	no	no	no	no	no	[25;30]	F0-F1	0	UNIQUE	100	>50	III-IV	yes	medium	A	A	no
CHC2134T	HCC	T	F	57	Europe	no	no	no	yes	no	no	yes	[<=25]	F0-F1	0	UNIQUE	100	>50	III-IV	yes	medium	A	C	NA
CHC2052T	HCC	T	M	61	Europe	no	yes	no	no	no	yes	yes	[>=30]	F0-F1	1	UNIQUE	80	>50	III-IV	yes	medium	A	A	NA
CHC2141T	HCC	T	M	74	Europe	no	yes	no	no	no	no	yes	[25;30]	F2-F3	1	UNIQUE	65	>50	III-IV	no	medium	A	A	no
CHC2098T	HCC	T	M	85	Europe	no	yes	no	no	no	no	no	[25;30]	F0-F1	0	UNIQUE	70	>50	I-II	no	medium	A	A	NA
CHC2099T	HCC	T	M	73	Europe	no	no	no	no	no	yes	yes	[>=30]	F0-F1	0	UNIQUE	130	>50	III-IV	yes	medium	A	C	no
CHC2202T	HCC	T	F	48	Europe	yes	no	no	no	no	no	yes	[<=25]	F0-F1	0	<=3	15	>50	I-II	yes	good	A	B	no
CHC2103T	HCC	T	M	57	Europe	no	yes	no	yes	no	no	yes	[<=25]	F0-F1	1	UNIQUE	28	<=50	III-IV	yes	medium	A	A	NA
CHC2206T	HCC	T	M	90	Europe	yes	no	no	no	no	no	no	[<=25]	F0-F1	1	UNIQUE	40	<=50	III-IV	no	medium	A	A	no
CHC2111T	HCC	T	F	56	Europe	yes	no	no	no	no	no	yes	[<=25]	F0-F1	0	UNIQUE	60	>50	I-II	no	good	A	A	no
CHC2211T	HCC	T	F	37	Europe	no	no	no	no	no	no	no	[<=25]	F0-F1	2	>=4	105	>50	III-IV	yes	medium	A	B	no
CHC2216T	HCC	T	M	62	Europe	no	yes	no	no	no	yes	no	[>=30]	F0-F1	0	UNIQUE	150	>50	I-II	yes	good	A	B	no
CHC2208T	HCC	T	M	53	Europe	yes	no	no	no	no	no	yes	[<=25]	F0-F1	2	<=3	20	<=50	III-IV	no	medium	A	B	no
CHC1743T	HCC	T	M	64	Europe	no	yes	no	no	no	no	yes	[<=25]	F4	1	<=3	NA	<=50	I-II	no	medium	B	A	no
CHC1569T	HCC	T	M	84	Europe	no	yes	no	no	no	no	yes	[<=25]	F0-F1	0	<=3	NA	>50	III-IV	yes	weak	A	C	yes
CHC1611T	HCC	T	M	75	Europe	no	no	no	no	no	yes	NA	[<=25]	F0-F1	0	UNIQUE	150	>50	III-IV	no	medium	A	A	yes
CHC1719T	HCC	T	M	57	Europe	no	yes	no	no	no	no	no	[25;30]	F4	0	UNIQUE	12	>50	I-II	no	medium	A	A	yes
CHC1591T	HCC	T	M	60	Europe	no	no	no	no	no	yes	no	[>=30]	F0-F1	4	UNIQUE	120	>50	III-IV	yes	medium	A	A	no
CHC2200T	HCC	T	M	69	Europe	no	no	no	no	no	yes	yes	[>=30]	F0-F1	0	UNIQUE	110	>50	III-IV	yes	weak	A	A	no
CHC1053T	HCC	T	M	74	Europe	no	yes	no	no	no	no	no	[25;30]	F4	2	UNIQUE	35	<=50	III-IV	yes	medium	A	A	no
CHC1595T	HCC	T	M	74	Europe	no	no	no	no	no	yes	yes	[25;30]	F2-F3	2	UNIQUE	90	>50	III-IV	yes	medium	A	A	yes
CHC1601T	HCC	T	M	75	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	0	UNIQUE	40	<=50	I-II	no	medium	A	A	no
CHC1596T	HCC	T	M	66	Europe	no	yes	no	no	no	yes	NA	[25;30]	F4	1	UNIQUE	30	<=50	III-IV	no	weak	A	A	yes
CHC1624T	early.HCC	T	M	59	Europe	no	yes	no	no	no	no	NA	NA	F4	0	<=3	12	<=50	I-II	no	medium	B	A	yes
CHC2351T	early.HCC	T	M	69	NA	no	yes	no	no	no	yes	NA	[25;30]	F4	2	<=3	35	<=50	III-IV	yes	medium	NA	C	NA
CHC1177T	HCC	T	M	62	Europe	no	no	no	yes	no	yes	yes	[25;30]	F4	4	UNIQUE	20	<=50	III-IV	no	medium	A	O	no
CHC1712T	HCC	T	M	76	Europe	no	yes	no	no	no	no	no	[25;30]	F2-F3	3	UNIQUE	36	<=50	III-IV	no	medium	A	B	yes
CHC2358T	early.HCC	T	M	65	NA	no	no	no	yes	no	no	NA	[<=25]	F4	2	<=3	75	<=50	I-II	yes	good	NA	B	NA
CHC1180T	HCC	T	M	65	Europe	no	yes	no	no	no	no	no	[25;30]	F2-F3	1	UNIQUE	30	<=50	I-II	no	good	A	A	yes
CHC1714T	HCC	T	M	72	Europe	no	no	no	yes	no	no	yes	[<=25]	F0-F1	2	UNIQUE	26	<=50	III-IV	yes	weak	A	A	yes
CHC2362T	early.HCC	T	M	49	Europe	no	yes	no	no	no	yes	NA	NA	F4	0	<=3	NA	<=50	I-II	no	NA	NA	A	NA
CHC1737T	HCC	T	M	73	Europe	no	no	no	no	no	yes	yes	[>=30]	F2-F3	1	UNIQUE	32	<=50	III-IV	no	weak	A	B	yes
CHC1210T	HCC	T	F	44	Asia	no	no	yes	no	no	no	no	[25;30]	F2-F3	1	UNIQUE	70	>50	III-IV	no	weak	A	A	yes
CHC1738T	HCC	T	F	68	Europe	no	no	no	yes	no	no	no	[<=25]	F4	1	UNIQUE	20	<=50	I-II	no	medium	A	A	no
CHC1081T	HCC	T	M	53	Europe	no	yes	no	no	no	no	NA	NA	F4	0	<=3	NA	<=50	III-IV	yes	medium	A	A	NA
CHC1750T	HCC	T	M	54	Europe	no	yes	no	no	no	no	yes	[25;30]	F4	0	>=4	NA	<=50	III-IV	no	medium	A	A	no
CHC1530T	HCC	T	M	64	Europe	no	yes	yes	no	no	no	NA	NA	F2-F3	1	UNIQUE	75	>50	III-IV	yes	weak	A	C	no
CHC1753T	HCC	T	M	65	Africa	no	no	no	yes	no	no	yes	[<=25]	F2-F3	0	UNIQUE	25	<=50	III-IV	no	medium	A	A	yes
CHC902T	HCC	T	M	73	Europe	no	yes	no	no	yes	no	no	NA	F2-F3	1	UNIQUE	57	>50	I-II	no	medium	A	A	no
CHC1534T	HCC	T	M	67	Europe	no	yes	no	no	yes	no	NA	NA	F2-F3	0	UNIQUE	55	>50	III-IV	no	medium	A	A	NA
CHC1774T	early.HCC	T	M	65	Europe	no	yes	no	no	no	no	NA	NA	F4	NA	<=3	NA	<=50	I-II	no	NA	NA	A	NA
CHC923T	HCC	T	M	74	Europe	no	yes	no	no	no	yes	yes	NA	F2-F3	4	UNIQUE	75	>50	III-IV	no	weak	A	A	yes
CHC1085T	HCC	T	M	49	Europe	no	yes	yes	yes	no	no	NA	NA	F4	1	<=3	17	<=50	I-II	yes	good	B	A	NA
CHC1556T	HCC	T	F	54	Europe	yes	no	no	no	no	no	yes	[<=25]	F2-F3	1	UNIQUE	100	>50	III-IV	yes	medium	A	A	no
CHC1775T	early.HCC	T	M	65	Europe	no	yes	no	no	no	no	NA	NA	F4	NA	<=3	NA	<=50	I-II	no	NA	NA	A	NA

CHCID	Diagnosis	Type	Gender	Age	Geographic Origin	Without. etiology	Alcohol. Intake	Hepati tis.B	Hepat itis.C	Hemochr omatosis	Metabolic. syndrome	Toba cco	BMI	normal. Liver.His tology	Steatosis. non.tumo ral.liver	Number. Nodules	nodule.s ize.mm	Largest.n odule.dia meter	Edmonson .grade	Vascular .Invasion	Differenci ation.WH O	Child .Pug h	BCLC	Inflam mation
CHC1592T	HCC	T	F	69	Europe	no	no	no	yes	no	no	no	[<=25]	F4	0	UNIQUE	20	<=50	III-IV	yes	weak	A	O	no
CHC2449T	HCC	T	M	81	Europe	no	no	no	no	no	yes	yes	[>=30]	F0-F1	0	<=3	130	>50	III-IV	no	weak	A	C	NA
CHC2695T	HCC	T	M	94	Europe	yes	no	no	no	no	no	NA	NA	F0-F1	0	UNIQUE	83	>50	III-IV	no	medium	NA	NA	NA
CHC2415T	HCC	T	M	68	Europe	no	no	no	no	no	yes	no	[25;30]	F0-F1	0	UNIQUE	180	>50	III-IV	yes	weak	A	C	NA
CHC2448T	HCC	T	M	82	Europe	no	no	no	no	no	yes	no	[>=30]	F0-F1	1	UNIQUE	75	>50	I-II	yes	good	A	B	NA
CHC2687T	HCC	T	M	76	Europe	no	no	no	no	NA	yes	yes	[>=30]	F0-F1	0	UNIQUE	120	>50	III-IV	yes	medium	NA	NA	no
CHC2707T	HCC	T	M	79	Europe	no	no	no	no	yes	no	yes	[25;30]	F0-F1	0	UNIQUE	30	<=50	I-II	yes	good	A	A	NA
CHC1606T	HCC	R	F	77	Europe	yes	no	no	no	no	no	NA	NA	F0-F1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC2539T	HCC	T	F	41	Europe	yes	no	no	no	no	no	yes	[25;30]	F0-F1	0	>=4	190	>50	III-IV	yes	medium	A	B	NA
CHC2686T	HCC	T	F	52	NA	yes	no	no	no	no	no	NA	[<=25]	F0-F1	1	<=3	NA	>50	III-IV	yes	medium	NA	NA	NA
CHC2207T	HCC	T	M	49	Europe	no	no	no	no	no	yes	yes	[25;30]	F0-F1	1	UNIQUE	90	>50	III-IV	yes	medium	A	A	NA
CHC2560T	HCC	T	M	74	Europe	no	no	no	no	no	yes	yes	[25;30]	F0-F1	4	UNIQUE	70	>50	I-II	no	good	A	A	NA
CHC2132T	HCC	T	M	57	Europe	no	no	no	no	no	yes	yes	[25;30]	F0-F1	0	UNIQUE	180	>50	III-IV	no	medium	A	B	NA
CHC2491T	HCC	T	M	66	Europe	no	yes	no	no	no	no	NA	[<=25]	F4	0	>=4	30	<=50	III-IV	yes	medium	NA	NA	NA
CHC2558T	HCC	T	M	70	Europe	no	yes	no	no	no	yes	no	[>=30]	F0-F1	1	NA	NA	<=50	I-II	no	medium	A	A	NA
CHC2697T	HCC	T	M	64	Europe	no	yes	no	no	no	no	no	[<=25]	F0-F1	0	UNIQUE	110	>50	III-IV	yes	weak	A	C	NA
CHC2135T	HCC	T	F	57	Europe	no	yes	yes	no	no	yes	yes	[<=25]	F0-F1	4	UNIQUE	25	<=50	III-IV	no	medium	A	A	NA
CHC2706T	HCC	T	M	70	Europe	no	yes	no	no	no	no	yes	[>=30]	F4	1	UNIQUE	50	<=50	III-IV	yes	medium	A	A	NA
CHC2210T	HCC	T	M	66	Europe	no	no	no	no	no	yes	no	[25;30]	F0-F1	2	UNIQUE	50	<=50	III-IV	yes	weak	B	A	NA
CHC2538T	HCC	T	F	76	Europe	no	no	no	no	no	yes	no	[>=30]	F2-F3	2	UNIQUE	40	<=50	I-II	no	good	A	A	NA
CHC2443T	HCC	T	M	74	Europe	no	yes	no	no	no	yes	yes	[>=30]	F0-F1	1	UNIQUE	48	<=50	III-IV	yes	medium	A	A	NA
CHC2691T	HCC	T	M	68	Europe	no	yes	no	no	no	yes	NA	NA	F0-F1	3	UNIQUE	60	>50	III-IV	no	medium	NA	NA	no
CHC014N	NT	N	M	30	Africa	no	no	yes	no	no	no	no	NA	F2-F3	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC013N	NT	N	M	63	Europe	no	no	no	yes	no	no	no	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC898N	NT	N	M	71	Europe	no	no	no	no	yes	no	no	NA	F2-F3	0	NA	NA	NA	NA	NA	NA	A	NA	no
CHC235N	NT	N	F	66	Europe	no	no	no	yes	no	no	no	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC591N	NT	N	F	37	NA	no	no	no	no	no	no	no	[<=25]	F0-F1	1	NA	NA	NA	NA	NA	NA	NA	NA	no
CHC152N	NT	N	M	64	Europe	no	no	yes	no	no	no	yes	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC469N	NT	N	F	32	Europe	yes	no	no	no	no	no	NA	[<=25]	F0-F1	0	NA	NA	NA	NA	NA	NA	A	NA	no
CHC229N	NT	N	F	65	Europe	no	no	no	yes	no	no	no	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC028N	NT	N	M	64	Europe	no	no	no	yes	no	no	NA	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC566N	NT	N	F	55	NA	yes	no	no	no	no	NA	NA	NA	F0-F1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC245N	NT	N	M	64	Europe	no	no	yes	no	no	no	NA	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC333N	NT	N	M	73	Europe	no	yes	no	no	no	no	NA	NA	F4	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC239N	NT	N	F	21	Africa	no	no	yes	no	no	no	no	NA	F2-F3	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC226N	NT	N	M	42	Africa	no	no	yes	no	no	no	NA	NA	F2-F3	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC046N	NT	N	M	61	Europe	no	no	yes	no	no	no	NA	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC081N	NT	N	F	76	Asia	no	no	yes	no	no	no	no	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC168N	NT	N	M	67	Europe	no	yes	no	no	no	no	NA	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC173N	NT	N	M	61	Europe	no	no	no	no	yes	no	NA	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC043N	NT	N	M	56	Asia	no	no	yes	no	no	no	NA	NA	F2-F3	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC934N	NT	N	F	44	NA	yes	no	no	no	no	no	NA	NA	F0-F1	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC1196N	NT	N	M	27	Africa	no	no	yes	no	no	no	yes	[<=25]	F2-F3	0	NA	NA	NA	NA	NA	NA	A	NA	yes
CHC203N	NT	N	M	46	Europe	no	yes	no	no	no	no	NA	NA	F4	2	NA	NA	NA	NA	NA	NA	A	NA	no
CHC1040N	NT	N	M	73	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	0	NA	NA	NA	NA	NA	NA	B	NA	no
CHC1044N	NT	N	M	78	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	0	NA	NA	NA	NA	NA	NA	B	NA	no
CHC1052N	NT	N	M	75	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	1	NA	NA	NA	NA	NA	NA	A	NA	no
CHC1055N	NT	N	M	68	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	2	NA	NA	NA	NA	NA	NA	A	NA	yes
CHC1062N	NT	N	M	65	Europe	no	no	no	no	yes	no	no	[25;30]	F2-F3	1	NA	NA	NA	NA	NA	NA	A	NA	no
CHC1069N	NT	N	M	78	Europe	no	yes	no	no	no	no	NA	NA	F2-F3	1	NA	NA	NA	NA	NA	NA	A	NA	no
CHC1162N	NT	N	M	60	Europe	no	yes	no	no	no	no	no	[25;30]	F4	0	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC789N	NT	N	M	54	Europe	no	no	no	no	yes	no	NA	NA	F2-F3	0	NA	NA	NA	NA	NA	NA	A	NA	no
CHC326N	NT	N	M	49	Europe	no	no	yes	no	no	no	NA	NA	F4	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC051N	NT	N	F	69	Europe	no	no	no	yes	no	no	NA	NA	F4	NA	NA	NA	NA	NA	NA	NA	A	NA	NA
CHC302N	NT	N	M	72	Europe	no	no	no	yes	no	no	NA	NA	F2-F3	1	NA	NA	NA	NA	NA	NA	A	NA	no
CHC306N	NT	N	M	68	Europe	no	no	no	yes	no	no	NA	NA	F4	0	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC313N	NT	N	F	43	Europe	no	no	no	yes	no	no	no	[<=25]	F0-F1	0	NA	NA	NA	NA	NA	NA	A	NA	no

B

CHCID	TERT	CTNNB1	CTNNB1.type	TP53	ARID1A	AXIN1	CDKN2A	ARID2	RPS6KA3	NFE2L2	KEAP1	PTEN	HNF1A	ALB	ACVR2A	RPL22	CDKN1A	RB1	TSC2	ATP10B	FGA	MEF2C	ZNRF3	EPHA4	TSC1	BAP1	G1G6	MCPcounter_inflamation	Cyclin.Status	Liver_progenitor	Stem_cell	EMT_metastasis	Differentiation	Proliferation
CHC018T	NM	NM	NA	NM	NM	M	NM	NM	M	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G1	3,733	WT	11,042	6,037	6,258	10,791	9,598
CHC229T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	NA	NA	NA	NA	NA	NA	NA
CHC231T	M	NA	NM	NM	NA	NM	NA	NA	NA	NA	NA	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	NA	NA	NA	NA	NA	NA	NA
CHC013T	M	M	T41	M	M	NM	NM	NM	NM	NM	M	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	NA	NA	NA	NA	NA	NA	NA
CHC441T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	NA	NA	NA	NA	NA	NA	NA
CHC333T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	NA	NA	NA	NA	NA	NA	NA
CHC239T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G2	3,931	NA	7,888	6,251	6,230	9,976	9,958
CHC399T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G6	NA	NA	NA	NA	NA	NA	NA
CHC014T	NM	NM	NA	M	M	M	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	5,137	NA	NA	NA	NA	NA	NA
CHC043T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	6,871	NA	NA	NA	NA	NA	NA
CHC037T	M	M	S45	NM	NM	M	NM	NM	M	NM	NM	NA	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	6,119	NA	NA	NA	NA	NA	NA
CHC339T	NM	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G2	4,689	NA	9,490	6,774	6,692	9,548	9,323
CHC245T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	NA	NA	NA	NA	NA	NA	NA
CHC253T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	7,163	WT	6,115	10,225	10,819	5,502	10,985
CHC158T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	M	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC445T	M	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC080T	NA	NA	NM	M	NM	M	NA	NM	NM	NM	NA	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G2	NA	NA	NA	NA	NA	NA	NA
CHC335T	NA	M	D32_S37	NM	M	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G6	NA	NA	NA	NA	NA	NA	NA
CHC230T	M	M	K335	NM	M	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	NA	NA	NA	NA	NA	NA	NA
CHC228T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	M	G1	5,879	NA	NA	NA	NA	NA	NA
CHC010T	NM	NA	NM	NM	NM	NM	NA	NM	NM	NA	NA	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G1	NA	NA	NA	NA	NA	NA	NA
CHC137T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	5,071	NA	NA	NA	NA	NA	NA
CHC205T	M	NM	NA	NM	M	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	3,556	NA	8,785	6,938	6,974	11,502	10,597
CHC218T	M	NM	NA	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	6,348	WT	8,263	8,807	9,035	12,890	9,773
CHC081T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	M	NA	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC031T	NM	NM	NA	NM	NM	M	M	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	NA	NA	NA	NA	NA	NA	NA
CHC242T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NA	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G6	NA	NA	NA	NA	NA	NA	NA
CHC059T	NM	M	T41	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	NA	NA	NA	NA	NA	NA	NA
CHC220T	M	NA	NA	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC206T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	7,446	NA	NA	NA	NA	NA	NA
CHC152T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G6	NA	NA	NA	NA	NA	NA	NA
CHC046T	M	M	D32_S37	M	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	NA	NA	NA	NA	NA	NA	NA
CHC211T	M	M	Large_deletion	NM	M	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC437T	M	M	T41	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC725T	M	NM	NA	NM	NM	NM	NM	M	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC317T	M	M	T41	NM	NM	NM	NM	NM	NM	M	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	NA	NA	NA	NA	NA	NA	NA
CHC789T	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC195T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC1196T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G3	6,599	NA	NA	NA	NA	NA	NA
CHC398T	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G2	NA	NA	NA	NA	NA	NA	NA
CHC1010T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	G1	6,857	WT	9,365	8,942	7,910	12,591	9,405
CHC1035T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	6,272	NA	NA	NA	NA	NA	NA
CHC1040T	NM	M	S45	NM	M	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	4,499	WT	5,483	7,034	7,937	13,127	9,683
CHC1041T	M	M	D32_S37	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	G6	4,933	WT	6,142	7,090	6,554	12,498	9,642
CHC1044T	M	M	S45	NM	NM	NM	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	G1	NA	NA	NA	NA	NA	NA	NA
CHC1052T	M	M	Other	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	G4	4,213	WT	5,610	8,268	7,177	13,917	8,636
CHC1055T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	8,024	CCNE1	6,041	7,042	6,998	9,365	10,697
CHC1060T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1061T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	4,545	WT	5,433	7,312	8,037	11,355	8,682
CHC1062T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,429	NA	NA	NA	NA	NA	NA
CHC1065T	M	NM	NA	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1146T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	NA	NA	NA	NA	NA	NA	NA
CHC1154T	M	M	D32_S37	M	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	5,536	NA	NA	NA	NA	NA	NA
CHC1162T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1192T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	5,325	NA	NA	NA	NA	NA	NA
CHC1199T	M	NM	NA	NM	NM	M	NM	M	NM	NM	NM	NA	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G2	NA	NA	NA	NA	NA	NA	NA
CHC1201T	NM	NM	NA	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	G2	NA	NA	NA	NA	NA	NA	NA
CHC255T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	5,771	NA	8,338	8,031	8,696	9,046	8,771
CHC320T	M	M	D32_S37	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	6,520	WT	7,394	7,581	7,582	12,939	9,372
CHC334T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	7,202	NA	NA	NA	NA	NA	NA
CHC412T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	M	NA	6,305	NA	NA	NA	NA	NA	NA
CHC429T	M	M	D32_S37	NM	NM	NM	NM	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	4,348	WT	7,308	7,448	6,699	13,204	8,640
CHC442T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,815	NA	6,822	11,063	9,968	11,186	8,416
CHC451T	M	M	D32_S37	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	NA	NA	NA	NA	NA	NA	NA
CHC613T	M	M	D32_S37	NM	M	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	4,591	NA	6,658	7,976	7,542	11,206	9,402
CHC614T	NM	M	T41	NM	NM	NM	NM	M	NM	M	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	4,280	WT	7,572	6,410	6,895	13,564	9,585
CHC703T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	NA	NA	NA	NA	NA	NA	NA
CHC734T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC736T	NM	NM	NA	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	5,800	WT	5,779	7,417	7,096	11,290	9,889
CHC793T	NA	NM	NA	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	5,502	WT	6,199	6,897	7,444	11,625	10,525
CHC796T	M	M	D32_S37	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	4,308	WT	6,694	6,655	7,472	11,475	10,201
CHC799T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	NA	NA	NA	NA	NA	NA	NA
CHC805T	M	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC884T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	5,253	WT	5,338	7,007	6,786	12,472	9,763
CHC889T	NM	NM	NA	M	M	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	M	NM	G2	4,393	WT	6,223	7,685	6,691	12,870	8,501
CHC891T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	5,818	WT	9,892	8,270	7,567	10,056	10,466
CHC892T	M	NM	NA	NM	M	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,993	WT	7,741	8,143	7,832	13,704	8,606
CHC896T	NM	NM	NA	M	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	NA	NA	NA	NA	NA	NA	NA
CHC898T	M	M	D32_S37	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	4,433	WT	6,722	7,042	6,964	13,126	8,737
CHC909T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	G3	4,700	WT	7,341	6,914	7,121	13,441	10,509
CHC912T	NM	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	5,147	WT	7,024	7,372	6,901	10,696	9,193
CHC983T	M	M	Other	NM	M	NM	NM	NM	NM	NM	NM	NA	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G4	NA	NA	NA	NA	NA	NA	NA
CHC051T	NM	M	D32_S37	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	5,287	WT	6,996	7,578	8,077	11,615	8,540
CHC121T	M	M	S45	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	5,190	WT	5,019	6,762	6,010	12,094	9,737
CHC155T	M	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC258T	NM	NM	NA	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC301T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	G6	NA	NA	NA	NA	NA	NA	NA
CHC302T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	G5	6,308	WT	7,510	8,702	8,305	11,933	9,679
CHC303T	NM	M	D32_S37	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC306T	M	NM	NA	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,583	WT	6,666	9,168	8,306	13,260	8,455
CHC307T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC326T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	G4	7,098	NA	7,126	7,656	7,527	14,120	8,509
CHC327T	M	NM	NA	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	5,699	WT	9,004	8,065	7,440	12,321	10,556
CHC434T	NM	NM	NA	NM	M	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	5,700	NA	NA	NA	NA	NA	NA
CHC609T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	3,578	WT	6,178	6,214	6,533	13,753	9,298
CHC197T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	7,528	WT	6,360	7,364	7,269	11,764	9,788
CHC304T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	4,765	WT	9,353	6,250	7,488	11,192	10,724
CHC313T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,894	CCNA2	7,446	7,672	7,523	12,500	8,886
CHC314T	M	NM	NA	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,014	WT	5,376	7,618	7,438	12,689	8,226
CHC322T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	G4	5,912	WT	5,429	9,111	8,497	12,430	9,240
CHC433T	M	M	D32_S37	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	G5	4,478	WT	5,802	7,031	6,780	12,402	9,261
CHC794T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	M	NM	NM	NM	NM	NM	M	G3	5,125	WT	6,886	7,403	7,943	8,598	10,021
CHC798T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	M	NM	M	M	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	G6	NA	NA	NA	NA	NA	NA	NA
CHC961T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	4,898	CCNA2	6,524	7,476	7,614	13,388	10,107
CHC060T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	NA	NA	NA	NA	NA	NA	NA
CHC1137T	M	M	D32_S37	NM	NM	NM	NM	M	NM	M	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	5,133	WT	7,234	6,919	6,705	13,391	9,727
CHC1568T	NM	NM	NA	M	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,594	WT	6,452	7,860	7,846	13,925	9,591
CHC1604T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	G3	6,031	WT	5,482	7,811	8,182	9,435	11,116
CHC1717T	NM	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	6,061	WT	8,046	8,410	6,731	11,698	9,767
CHC1763T	M	M	Double_mutation	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	6,104	WT	7,213	8,307	7,435	13,391	10,162
CHC097T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1148T	M	M	Double_mutation	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	3,938	WT	6,113	6,688	7,306	12,890	9,831
CHC1744T	NA	M	Large_deletion	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	6,350	WT	11,070	7,990	7,420	10,371	9,982
CHC1152T	M	M	D32_S37	M	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	NA	NA	NA	NA	NA	NA	NA
CHC1205T	M	NM	NA	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	G3	NA	NA	NA	NA	NA	NA	NA

CHCID	TERT	CTNNB1	CTNNB1.type	TP53	ARID1A	AXIN1	CDKN2A	ARID2	RPS6KA3	NFE2L2	KEAP1	PTEN	HNF1A	ALB	ACVR2A	RPL22	CDKN1A	RB1	TSC2	ATP10B	FGA	MEF2C	ZNRF3	EPHA4	TSC1	BAP1	G1G6	MCPcounter_infiltration	Cyclin.Status	Liver_progenitor	Stem_cell	EMT_metastasis	Differentiation	Proliferation
CHC1616T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	G3	4,842	WT	6,559	7,174	6,916	11,116	10,100
CHC1720T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	G2	4,556	NA	6,276	7,045	6,679	11,626	10,337
CHC1745T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	6,486	WT	6,155	7,872	7,690	13,791	9,571
CHC432T	M	M	Double_mutation	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	6,157	WT	7,061	7,905	8,230	13,227	9,048
CHC1207T	M	M	T41	NM	M	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	4,090	WT	6,585	6,560	6,895	13,208	9,452
CHC1626T	M	M	N387	M	M	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	4,960	WT	4,039	6,611	7,003	11,823	10,756
CHC1725T	M	NM	NA	NM	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	G1	5,003	WT	11,198	8,007	7,172	10,976	9,027
CHC1746T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	4,475	WT	6,749	7,250	6,927	13,369	9,866
CHC1209T	M	M	S45	NM	NM	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	G6	NA	NA	NA	NA	NA	NA	NA
CHC1594T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	G2	6,054	CCNE1	7,166	9,068	7,690	13,462	9,626
CHC1629T	M	M	D32_S37	NM	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	G6	5,260	WT	7,231	8,337	7,344	11,998	9,298
CHC1731T	NM	M	D32_S37	NM	NM	NM	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	4,793	WT	7,674	6,945	7,090	13,470	8,199
CHC1747T	M	M	D32_S37	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	G5	5,879	WT	6,726	7,898	7,865	14,031	9,777
CHC801T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1211T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	4,496	WT	10,861	8,175	8,073	9,305	9,623
CHC1700T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	6,873	NA	NA	NA	NA	NA	NA
CHC1732T	M	NM	NA	M	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	M	NM	G3	5,958	WT	8,397	9,124	8,004	12,762	10,108
CHC1749T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	5,160	WT	5,472	7,663	8,101	12,183	10,797
CHC1183T	M	M	T41	NM	NM	NM	M	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	G5	4,802	WT	6,808	7,184	6,435	13,436	7,882
CHC1597T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	G3	5,364	NA	7,832	9,549	9,291	10,144	10,759
CHC1704T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	G3	5,941	WT	8,734	8,274	7,468	12,478	10,287
CHC1734T	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1185T	NM	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	5,344	WT	11,679	8,422	8,014	10,723	10,282
CHC1531T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1598T	NM	M	D32_S37	NM	NM	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	4,631	WT	6,499	7,185	7,159	13,434	8,603
CHC1705T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G5	NA	NA	NA	NA	NA	NA	NA
CHC1736T	NM	M	N387	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	5,896	WT	10,653	8,190	7,906	10,277	10,655
CHC1751T	M	M	T41	NM	NM	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1186T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1539T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,800	WT	8,070	9,808	8,759	13,938	9,388
CHC1600T	M	M	T41	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	8,003	WT	6,707	8,658	8,557	12,464	9,063
CHC1708T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	NA	NA	NA	NA	NA	NA	NA
CHC1028T	NM	M	S45	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1189T	M	M	T41	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G6	7,228	NA	7,342	8,856	9,037	11,284	10,103
CHC1545T	M	M	D32_S37	NM	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	5,044	WT	6,191	6,869	7,113	13,516	8,030
CHC1739T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	G5	4,806	WT	5,161	7,843	7,474	13,283	9,906
CHC1754T	M	NM	NA	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	4,644	WT	10,916	7,642	7,118	10,871	10,472
CHC1190T	NM	NM	NA	NM	NM	M	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1602T	M	NM	NA	M	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	6,053	WT	7,235	8,114	7,026	12,416	9,848
CHC1741T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	5,769	WT	7,286	7,142	7,224	13,661	8,401
CHC1756T	M	M	T41	M	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	5,049	WT	7,846	8,325	7,907	14,004	9,279
CHC1079T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,852	NA	NA	NA	NA	NA	NA
CHC1566T	M	M	D32_S37	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1603T	M	M	K335	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	G3	4,847	WT	7,079	8,164	6,831	12,565	9,426
CHC1715T	M	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	G5	5,297	WT	6,221	7,632	7,633	13,070	10,044
CHC1742T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	G3	4,954	WT	5,770	7,508	7,704	12,754	10,910
CHC1757T	M	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,517	WT	6,435	7,840	7,386	14,627	7,780
CHC2025T	M	M	S45	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G1	NA	NA	NA	NA	NA	NA	NA
CHC2112T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,300	CCNA2	5,803	6,790	6,799	12,901	10,153
CHC2029T	M	M	K335	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC2113T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	G5	6,556	WT	6,522	7,837	7,291	12,896	9,070
CHC2215T	M	NM	NA	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	6,496	NA	NA	NA	NA	NA	NA
CHC2034T	M	M	T41	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	NA	NA	NA	NA	NA	NA	NA
CHC2115T	M	NM	NA	NM	NM	NM	NM	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	6,641	WT	9,256	7,621	7,362	11,273	10,646
CHC2127T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC2043T	NM	M	Large_deletion	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	4,688	WT	9,574	6,860	7,311	11,905	8,203
CHC2128T	NM	NM	NA	NM	NM	M	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,965	CCNA2	6,292	9,071	8,665	13,390	10,005
CHC2048T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	G2	5,031	CCNE1	5,466	7,611	7,292	12,901	10,648
CHC2134T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA
CHC2052T	M	M	S45	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC2141T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	G2	5,966	CCNE1	7,306	8,570	8,137	12,985	10,254
CHC2098T	M	M	D32_S37	NM	NM	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC2099T	M	M	S45	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC2202T	NM	M	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC2103T	NM	M	T41	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	5,180	WT	6,969	7,056	6,681	14,163	9,794
CHC2206T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	8,261	CCNA2	6,905	9,956	8,416	12,520	9,911
CHC2111T	NM	M	Large_deletion	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	G6	4,823	WT	7,542	6,719	7,051	12,957	6,978
CHC2211T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	G2	5,178	NA	NA	NA	NA	NA	NA	NA
CHC2216T	NM	M	T41	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,419	NA	NA	NA	NA	NA	NA
CHC2208T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	M	NM	M	NM	NM	NM	NM	NM	NM	G3	6,884	CCNE1	6,776	9,873	9,072	12,213	10,698
CHC1743T	M	NM	NA	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1569T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1611T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1719T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1591T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	4,096	CCNE1	7,368	7,212	6,633	11,831	10,481
CHC2200T	NM	M	K335	M	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	G3	4,931	WT	9,038	8,124	7,416	12,403	11,161
CHC1053T	M	NM	NA	NM	M	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	M	NM	NM	NM	NM	G2	5,708	WT	7,079	7,202	6,099	13,186	9,377
CHC1595T	NA	M	D32_S37	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G6	NA	NA	NA	NA	NA	NA	NA
CHC1601T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	NA	NA	NA	NA	NA	NA	NA
CHC1596T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1624T	M	NM	NA	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC2351T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1177T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1712T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC2358T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1180T	M	NM	NA	M	M	NM	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	8,855	WT	3,745	10,724	11,910	6,233	11,087
CHC1714T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC2362T	M	NM	NA	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1737T	M	M	NA	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	8,868	WT	6,055	8,767	8,809	11,453	9,937
CHC1210T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	8,206	WT	5,959	8,852	8,774	9,775	10,096
CHC1738T	NM	NM	NA	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1081T	M	M	T41	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1750T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1530T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	6,848	WT	9,103	9,237	9,049	11,948	10,676
CHC1753T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,609	WT	6,785	10,097	9,437	12,074	9,617
CHC902T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1534T	M	NM	NA	M	M	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	7,016	WT	7,825	7,294	7,834	11,462	11,326
CHC1774T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC923T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	NA	NA	NA	NA	NA	NA	NA
CHC1085T	M	NM	NA	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1556T	NM	NM	NA	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	NA	NA	NA	NA	NA	NA	NA
CHC1775T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	NA	NA	NA	NA	NA	NA	NA
CHC1592T	M	NM	NA	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G1	NA	NA	NA	NA	NA	NA	NA
CHC2449T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	7,762	CCNA2	8,743	9,600	8,667	11,871	9,851
CHC2695T	NA	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,590	WT	7,184	8,962	8,843	12,293	9,244
CHC2415T	NM	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	G3	7,947	WT	4,694	9,670	9,782	7,573	10,552
CHC2448T	M	M	S45	NM	NM	NM	M	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	5,290	WT	7,154	7,546	7,738	13,362	8,424
CHC2687T	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NA	4,552	WT	6,546	6,972	7,351	12,269	9,723
CHC2707T	NA	NM	NA	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	5,873	WT	7,415	8,639	8,160	13,766	8,824
CHC1606T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	6,629	NA	7,240	7,686	7,737	14,068	8,437
CHC2539T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	5,228	WT	12,175	8,286	7,583	10,263	9,959
CHC2686T	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	M	NM	NM	NM	NM	NM	NM	NM	NA	5,139	WT	4,805	6,934	7,690	10,217	10,702
CHC2207T	NM	NM	NA	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	G4	4,595	CCNE1	5,430	7,365	6,672	13,332	10,352
CHC2560T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G4	6,028	WT	6,005	9,524	8,901	12,414	8,482
CHC2132T	M	NM	NA	M	NM	NM	NM	M	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	6,245	WT	9,045	7,382	7,455	10,775	10,365
CHC2491T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G2	7,924	WT	8,349	10,441	9,569	12,361	8,168
CHC2558T	M	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	4,557	WT	6,303	7,674	7,359	14,091	9,579
CHC2697T	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	5,679	CCNA2	6,036	8,564	7,792	12,680	9,385
CHC2135T	NM	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	G1	6,169	WT	10,141	9,558	7,802	10,228	10,329

CHCID	TERT	CTNNB1	CTNNB1.type	TP53	ARID1A	AXIN1	CDKN2A	ARID2	RPS6KA3	NFE2L2	KEAP1	PTEN	HNF1A	ALB	ACVR2A	RPL22	CDKN1A	RB1	TSC2	ATP10B	FGA	MEF2C	ZNRF3	EPHA4	TSC1	BAP1	G1G6	MCPcounter_inflamation	Cyclin.Status	Liver_progenitor	Stem_cell	EMT_metastasis	Differentiation	Proliferation
CHC2706T	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	G5	6,098	WT	6,834	9,006	8,818	12,663	10,033
CHC2210T	M	NM	NA	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	G3	5,562	WT	6,666	7,198	6,484	11,234	11,155
CHC2538T	M	M	D32_S37	NM	NM	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	G5	5,872	WT	5,899	6,886	7,202	13,203	9,763
CHC2443T	M	NA	NA	NM	NA	NM	NM	NM	NM	NM	NM	NM	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	G2	4,860	WT	7,430	7,849	7,608	12,539	9,541
CHC2691T	NA	M	T41	NM	M	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NM	NA	5,269	WT	6,350	7,903	7,793	12,692	9,036
CHC014N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC013N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC898N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC235N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC591N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC152N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC469N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6,425	NA	7,712	8,415	8,391	14,672	7,670
CHC229N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC028N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC566N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC245N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC333N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC239N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC226N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC046N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC081N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC168N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC173N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC043N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC934N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6,371	NA	7,906	8,747	8,173	14,757	7,312
CHC1196N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC203N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC1040N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC1044N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC1052N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC1055N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	6,752	NA	7,849	7,740	7,400	13,951	8,285
CHC1062N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC1069N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC1162N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC789N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC326N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	7,292	NA	8,668	9,160	8,883	14,032	8,955
CHC051N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC302N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
CHC306N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	7,141	NA	8,597	8,998	8,660	14,040	9,009
CHC313N	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA

Supplementary Table 1: LICAFR annotation data (A) Clinical (B) Molecular

A

annot	MC01	MC02	MC03	MC04	MC05	MC06	MC07	MC08	MC09	MC10	MC11	MC12	MC13
Gender	0,001993	0,007918	2,75E-58	0,000377	0,001683	0,007513	0,060813	0,693203	0,923175		0,488383	0,017449	0,000403
Geographic	0,001059	0,004874	0,897074	0,44218	0,088636	0,358144	0,06336	0,856897	0,427599		0,03552	0,082145	0,22455
Age	6,85E-19	0,187051	0,223452	0,043111	2,43E-06	0,696068	0,105217	0,833079	0,055567		0,103486	0,00093	1,88E-05
Alcohol Intake	0,288242	0,027682	0,000343	0,0568	0,00355	0,391634	0,079618	0,335743	0,284062		0,427231	0,15706	0,059527
Hepatitis B	0,000513	0,20042	0,939318	0,704865	0,001085	0,359614	0,173819	0,907513	0,057203		0,051239	0,020645	0,10721
Hepatitis C	0,19477	0,824767	0,166226	0,509159	0,074993	0,835554	0,680431	0,237401	0,545856		0,093845	0,111446	0,088312
Hemochromato	0,569986	0,455923	0,055948	0,835713	0,72915	0,585867	0,970387	0,945634	0,340789		0,744102	0,837181	0,445659
Metabolic	0,106944	0,330974	0,338488	0,258979	0,177245	0,779847	0,173664	0,325614	0,061227		0,589862	0,53976	0,121958
Without	0,736842	0,750585	7,2E-05	0,353913	0,637555	0,073133	0,600625	0,392991	0,820987		0,841124	0,573252	0,06669
Tobacco	0,360898	0,094308	0,031386	0,743655	0,823046	0,006688	0,035818	0,456544	0,069849		0,091711	0,584833	0,245781
BMI	0,043042	0,261518	0,72956	0,075044	0,13175	0,178094	0,647481	0,028095	0,00843		0,243583	0,082383	0,048589
BCLC	0,125484	0,289686	0,017826	0,166415	0,021145	0,001269	0,009835	0,427792	0,007188		0,045976	0,047079	0,80441
Normal Liver	0,866031	0,099118	0,005967	0,055163	0,314901	0,253052	0,014929	0,082326	0,315884		0,282868	0,418578	0,408182
Steatosis non	0,032315	0,435395	0,239382	0,007315	0,075192	0,188889	0,398144	0,212457	0,258126		0,030619	0,101899	0,153626
Child.Pugh	0,053299	0,436589	0,613494	0,630487	0,161895	0,240946	0,301001	0,803042	0,211159		0,262305	0,25473	0,136904
Edmonson	0,083473	0,476151	0,624615	0,876754	0,000712	0,012227	0,371604	0,658233	0,001364		0,022092	0,004754	0,291366
Differenciation	0,130449	0,605453	0,618225	0,65148	0,013597	0,00794	0,49249	0,527963	0,001319		0,097247	0,034931	0,900002
Number	0,290467	0,043516	0,102855	0,648474	0,114476	0,155076	0,010591	0,354573	0,167461		0,439832	0,298433	0,451714
Nodule size	0,534864	0,05645	0,162376	0,004579	0,67492	0,003526	0,627597	0,108077	0,561976		0,372779	0,475292	0,022199
Largest nodule	0,297278	0,287062	0,060689	0,169923	0,274565	0,232225	0,873323	0,038433	0,681801		0,329728	0,563379	0,160014
Vascular	0,721998	0,072534	0,449261	0,024821	0,76059	0,858301	0,48003	0,167488	0,163886		0,480509	0,850113	0,088743
TERT	6,18E-07	0,594317	0,00281	0,03119	0,004676	0,249865	0,594371	0,557264	0,867671		3,58E-06	0,00027	0,000265
TP53	0,918292	0,793342	0,767551	0,000542	0,002768	0,004566	0,216777	0,023812	0,025082		0,01526	0,08513	0,284872
CTNNB1	5,93E-05	0,411957	0,839737	0,006779	7,72E-07	5,18E-07	0,381371	6,95E-21	0,001294		4,99E-11	7,54E-14	2,58E-05
AXIN1	0,537461	0,069939	0,23382	0,590029	0,352328	0,124329	0,004085	0,073923	0,455364		0,45658	0,854639	0,187282
ALB	0,821859	0,585666	0,698794	0,78119	0,771125	0,581471	0,157187	0,922487	0,370989		0,910445	0,972788	0,840231
ARID2	1,55E-05	0,058332	0,148156	0,510545	0,424126	0,014326	0,005058	0,260369	0,000594		0,003736	0,000684	0,030162
ARID1A	0,23918	0,795949	0,571327	0,023143	0,024846	0,433723	8,37E-06	0,558755	0,022671		0,586605	0,068147	0,0088
ACVR2A	0,903204	0,640372	0,293828	0,350916	0,670675	0,593359	0,129648	0,430132	0,408108		0,869938	0,861838	0,948832
NFE2L2	0,055451	0,687905	0,247634	0,192565	0,547618	0,053873	0,013614	0,026511	0,350048		0,000691	0,056015	0,918216
RPS6KA3	0,620272	0,000403	0,215367	0,310978	0,761436	0,00071	0,144758	0,029305	0,897236		0,125634	0,139427	0,582618
KEAP1	0,317429	0,416405	0,436142	0,228253	0,26162	0,823318	0,401841	0,140601	0,183985		0,019825	0,496569	0,031609
RPL22	0,283815	0,876877	0,896715	0,265503	0,104176	0,417044	0,280471	0,004781	0,028868		0,30787	0,004754	0,080027
CDKN2A	0,474567	0,841309	0,400608	0,384611	0,604321	0,204528	0,301767	0,522777	0,385363		0,022873	0,199089	0,317157
CDKN1A	0,394773	0,230615	0,276928	0,062487	0,298484	0,86549	0,612357	0,034666	0,735339		0,083793	0,605086	0,501891
RB1	0,863443	0,00579	0,631554	0,989716	0,840037	0,558794	0,047109	0,053895	0,165816		0,421142	0,489223	0,230966
TSC2	0,047832	0,113558	0,442944	0,410323	0,178794	0,935362	0,360981	0,73102	0,137558		0,171921	0,009812	0,094773
ATP10B	0,281427	0,572282	0,74235	0,025567	0,073798	0,144663	0,074407	0,769806	0,229904		0,115924	0,251234	0,003995
FGA	0,667772	0,098333	0,273332	0,666556	0,373724	0,24716	0,493225	0,44908	0,060271		0,673973	0,673743	0,370666
MEF2C	0,698938	0,65363	0,957558	0,061221	0,149096	0,895346	0,622636	0,786492	0,159544		0,884453	0,393414	0,286286
HNF1A	0,382514	0,756325	0,131517	0,183184	0,629515	0,729634	0,147781	0,471653	0,050008		0,334573	0,959851	0,543713
ZNRF3	0,653177	0,281536	0,594848	0,361254	0,136772	0,64341	0,241007	0,905833	0,829388		0,38416	0,783473	0,379173
EPHA4	0,378667	0,414875	0,23104	0,776104	0,222922	0,513115	0,511256	0,995086	0,620035		0,955478	0,417878	0,666023
PTEN	0,540447	0,091301	0,56909	0,461659	0,350977	0,701421	0,151502	0,039639	0,048014		0,724308	0,555592	0,921817
TSC1	0,744637	0,289102	0,953831	0,501097	0,887747	0,033825	0,630064	0,273156	0,345929		0,085646	0,382074	0,822782
BAP1	0,000173	0,014047	0,157435	0,008299	0,02196	0,746092	0,786586	0,7976	0,225691		0,051901	0,017885	0,005536
CCNHCC	0,973801	0,084955	0,031013	2,09E-07	0,005552	0,263136	0,338197	6,86E-08	0,000101		1,43E-06	0,821344	0,40019
G1	2,45E-07	6,71E-06	0,066996	0,001309	0,000398	0,069062	0,367072	0,908964	0,064323		0,11113	0,00693	0,00051
G2	0,05175	0,016109	0,09895	0,005225	0,823529	0,411914	0,635246	0,00075	0,231605		0,512224	0,868852	0,144734
G3	0,196485	0,789395	0,013053	0,54542	0,01022	0,000121	0,336603	0,144366	0,065603		0,078281	0,001167	0,638118
G4	0,974082	0,038195	0,922765	0,06209	0,520351	0,07802	0,03387	0,000216	0,450482		0,228313	0,124177	0,012693
G5	0,046561	0,584762	0,60199	0,104273	0,061948	0,227144	0,197975	0,860281	0,340458		0,428798	0,011691	0,262586
G6	0,143642	0,031868	0,727461	0,159063	9,83E-05	0,010486	0,778112	1,13E-16	0,040919		0,000269	2,07E-05	0,000993
MCPcounter	0,040735	0,000118	0,046142	9,39E-05	0,032397	2,24E-21	0,818528	0,257916	0,821981		4,17E-05	0,00239	0,000181
Liver_progenitor	0,00015	0,025816	0,058139	0,053065	1,37E-08	0,275721	0,026147	0,670702	0,031917		0,091438	0,011572	0,023507
Stem_cell	0,022529	9,6E-05	0,24632	0,003885	0,01264	2,19E-16	0,326759	0,065317	0,367672		1,27E-07	0,000233	0,000142
EMT_metastasis	0,016691	7,51E-05	0,104288	2,97E-05	0,001861	5,33E-17	0,551264	0,211777	0,250285		5,11E-05	3,99E-05	2,28E-05
Differentiation	2,84E-06	0,05314	0,606277	0,004686	3,91E-06	1,66E-05	0,235569	0,552777	0,005688		0,004193	2,56E-07	0,002544
Proliferation	0,099403	0,166622	0,010002	0,093272	0,122359	1,78E-08	0,073694	0,02085	0,001381		0,000103	7,38E-05	0,898807

B

annot	MeC01	MeC02	MeC03	MeC04	MeC05	MeC06	MeC07	MeC08	MeC09	MeC10	MeC11	MeC12	MeC13
Gender	0,0159	0,0095	5E-88	0,0562	0,2211	0,0157	0,5598	0,1839	0,2231	0,4781	0,2257	0,1175	0,3968
Geographic.Origin	0,006	0,0123	0,0756	0,1528	0,2747	0,1892	0,0495	0,5744	0,2192	0,0342	0,685	0,3168	0,1091
tumor_grade	0,289	2E-07	0,3991	0,273	0,1793	0,4149	0,0615	0,3447	8E-05	0,1041	0,1692	0,0637	0,1282
Age	6E-11	0,1389	0,1185	0,0009	0,4859	0,3785	0,2625	0,0969	0,0149	0,0064	0,0242	0,0033	8E-10
tumor_stage	0,0748	0,0016	0,0708	0,2937	0,062	0,1367	0,0428	0,0623	0,0253	0,2877	0,0762	0,2811	0,1084
Child.Pugh	0,3327	0,4407	0,1284	0,5311	0,1074	0,1725	0,8058	0,2743	0,3809	0,4791	0,1767	0,6719	0,266
normal.Liver.Histology	0,0306	0,0057	0,0174	0,3878	0,0331	0,1915	0,1011	0,1075	0,2091	0,1081	0,5527	0,1588	0,1613
recod.BMI	0,0038	0,0036	0,4592	0,0377	0,1616	0,0502	0,2957	0,3717	0,2174	0,2649	0,4923	0,0993	0,0068
Without.etiology	0,0023	0,9829	2E-08	0,2035	0,6419	0,436	0,6605	0,9041	0,5579	0,9439	0,2439	0,0707	0,0347
Hepatitis.B	0,9058	0,007	0,0184	0,3115	0,2997	0,1217	0,7111	0,6635	0,7058	0,9917	0,7876	0,0754	0,8172
Hepatitis.C	2E-05	0,2237	0,0518	0,5431	0,0429	0,9201	0,9374	0,2831	0,3058	0,6151	0,7304	0,0022	0,0079
Tobacco	0,8837	0,9201	0,1423	0,1721	0,9206	0,8541	0,5668	0,9387	0,0405	0,0979	0,6776	0,1803	0,1499
Alcohol.Intake	0,6589	0,0758	2E-05	0,2762	0,414	0,0025	0,8032	0,3958	0,5591	0,8577	0,1403	0,6263	0,4191
CTNNB1.type	0,1468	0,0189	0,3461	0,1865	0,1237	0,1394	0,3847	0,0102	0,0972	0,3325	0,0577	0,1219	0,1134
TERT	7E-08	0,282	0,3235	0,053	0,9185	0,9299	0,0101	0,2604	0,5007	0,3586	0,8826	0,0003	3E-07
TP53	0,2486	0,4451	0,0154	9E-05	9E-05	0,0572	0,366	0,2787	0,337	0,0002	0,0541	0,3633	0,1818
CTNNB1	8E-06	0,0303	0,0295	0,0081	0,2013	0,0514	0,6439	7E-21	0,6241	0,0853	0,8624	3E-07	4E-05
AXIN1	0,9594	0,0058	0,2785	0,5207	0,5599	0,0061	0,0187	0,5532	0,1646	0,843	0,5644	0,8858	0,3599
ALB	0,0009	0,5871	0,2374	0,2379	0,643	0,2693	0,1926	0,5379	0,8016	0,1466	0,3358	0,0206	0,0138
ARID2	0,3728	0,234	0,4371	0,2613	0,8479	0,2081	0,1054	0,1294	0,1942	0,2337	0,3371	0,1016	0,3138
ARID1A	0,6485	0,0473	0,7582	0,5264	0,9489	0,7775	1E-05	0,1448	0,4396	0,6857	0,8527	0,9083	0,7989
ACVR2A	0,4471	0,4242	0,0222	0,2466	0,1471	0,4539	0,5199	0,0425	0,0357	0,7478	0,7381	0,172	0,0704
NFE2L2	0,0379	0,5757	0,7522	0,9411	0,3206	0,7211	0,882	0,2177	0,3431	0,5042	0,7123	0,1798	0,4123
RPS6KA3	0,1606	0,0193	0,1749	0,6774	0,0801	0,0582	0,0348	0,1516	0,4182	0,984	0,5999	0,7587	0,213
KEAP1	0,9489	0,4974	0,7179	0,2499	0,0808	0,0147	0,11	0,1581	0,5254	0,2509	0,1653	0,5438	0,9847
RPL22	0,3431	0,0268	0,1229	0,7517	0,1093	0,1287	0,0067	0,7973	0,2284	0,9134	0,3725	0,4796	0,2449
CDKN2A	0,5113	0,9575	0,7813	0,6153	0,3087	0,351	0,4939	0,0861	0,2258	0,6924	0,4475	0,9433	0,6363
CDKN1A	0,2954	0,4503	0,1306	0,2429	0,3333	0,3061	0,1261	0,6125	0,8826	0,2751	0,0479	0,3581	0,8811
RB1	0,5035	9E-05	0,7343	0,2188	0,1167	0,0011	0,0177	0,6411	0,0421	0,3761	0,9202	0,1902	0,1074
TSC2	0,3783	0,9457	0,1661	0,3749	0,6853	0,0017	0,7488	0,9287	0,2093	0,9401	0,6182	0,0947	0,7736
ATP10B	0,8785	0,7354	0,1836	0,4032	0,5705	0,4167	0,2105	0,1977	0,2388	0,2447	0,0831	0,1304	0,5922
FGA	0,2198	0,9412	0,72	0,1834	0,526	0,6329	0,7761	0,1266	0,8537	0,4619	0,4769	0,2831	0,1415
MEF2C	0,6796	0,699	0,9232	0,2902	0,8505	0,563	0,9354	0,7718	0,792	0,3485	0,12	0,7599	0,5483
HNF1A	0,7342	0,0804	0,0235	0,5147	0,9741	0,8933	0,7292	0,398	0,0034	0,2706	0,8723	0,1038	0,6653
ZNRF3	0,0797	0,7702	0,4528	0,7555	0,3891	0,5017	0,3279	0,7578	0,0558	0,6426	0,6332	0,6159	0,5127
EPHA4	0,8638	0,0315	0,2989	0,2696	0,213	0,6594	0,6064	0,7605	0,2078	0,7629	0,9819	0,1066	0,239
PTEN	0,3494	0,1852	0,8929	0,1043	0,9228	0,0401	0,0005	0,591	0,4629	0,6407	0,8777	0,0568	0,4678
TSC1	0,2245	0,9401	0,9638	0,9703	0,5904	0,6144	0,7251	0,0591	0,463	0,5958	0,9737	0,6308	0,7929
BAP1	0,0012	0,0035	0,04	2E-05	0,1199	0,2607	0,5589	0,1373	0,1313	0,1269	0,8086	3E-05	0,009
CCN.HCC	0,1301	0,0232	0,0037	3E-13	0,359	0,6141	0,4923	4E-10	7E-09	1E-09	7E-07	0,2979	2E-05
G1	2E-09	3E-12	0,0793	2E-10	0,9356	0,6613	0,1492	0,5681	0,0976	0,043	0,053	3E-10	3E-09
G2	0,1418	0,0359	0,1544	0,6933	0,5044	0,3822	0,0288	0,003	0,6788	0,0473	0,3394	0,0674	0,1173
G3	0,1632	0,1737	0,3203	0,2949	0,0058	0,0019	6E-05	0,4964	2E-06	0,0295	0,1578	0,4367	0,2624
G4	0,2852	2E-09	0,5743	0,5199	0,1226	0,0191	3E-05	2E-10	0,131	0,4686	0,6661	0,234	0,4871
G5	0,0001	0,9336	0,4642	3E-07	0,6895	0,0008	0,449	0,1992	0,0005	0,1369	3E-07	0,0004	0,0004
G6	0,0616	0,0669	0,0803	0,0513	0,4018	0,1158	0,6013	2E-27	0,2822	0,1938	0,2181	0,0002	0,0917
MCPcounter_inflamatio	0,0617	8E-11	0,3369	0,0122	0,4268	1E-49	0,0201	0,0727	0,7003	0,6134	0,7854	3E-06	0,0047
Liver_progenitor	2E-07	4E-11	0,344	1E-05	0,0007	0,6526	3E-11	0,0001	0,2451	0,2638	0,0118	3E-07	3E-06
Stem_cell	1E-05	6E-10	0,009	1E-06	0,7199	3E-22	0,0028	0,1985	0,1086	0,0819	0,0666	2E-13	7E-08
EMT_metastasis	9E-07	7E-08	0,1429	6E-08	0,3715	1E-29	0,1703	0,1531	0,0411	0,0516	0,0867	1E-12	5E-10
Differentiation	1E-06	2E-07	0,7423	9E-07	9E-05	6E-08	6E-06	1E-05	3E-05	0,9539	0,1821	5E-11	6E-09
Proliferation	0,488	2E-11	0,6317	0,0029	0,001	0,0002	1E-09	0,3936	3E-05	4E-09	0,0003	0,1203	0,332

Supplementary Table 2: Univariate analysis (A) LICAFR (B) TCGA

A

annot	MC01	MC02	MC03	MC04	MC05	MC06	MC07	MC08	MC09	MC10	MC11	MC12	MC13
Gender		0,659476	4,78E-55										
Geographic Origin	0,542679												
Age	1,07E-07											0,316275	0,011198
Alcohol Intake			0,521844										
Hepatitis B													
Hepatitis C													
Hemochromatosis													
Metabolic syndrome													
Without etiology													
Tobacco													
BMI													
BCLC													
Normal Liver Histology													
Steatosis non tumoral liver													
Child Pugh													
Edmonson grade													
Differenciation WHO													
Number Nodules													
nodule size (mm)													
Largest nodule diameter													
Vascular Invasion													
TERT	0,62751											0,729057	0,786753
TP53				8,41E-05									
CTNNB1	0,155361			0,000935				0,003151				0,00323	0,008274
AXIN1													
ALB													
ARID2													
ARID1A							8,37E-06						
ACVR2A													
NFE2L2													
RPS6KA3													
KEAP1													
RPL22													
CDKN2A													
CDKN1A													
RB1		0,001481											
TSC2													
ATP10B													
FGA													
MEF2C													
HNF1A													
ZNRF3													
EPHA4													
PTEN													
TSC1													
BAP1													
CCN.HCC				9,32E-11							9,81E-06		
G1		0,047899											
G2													
G3						0,012115							
G4													
G5													
G6								1,68E-06				0,372379	
MCPcounter inflamation						7,8E-08							
Liver_progenitor													
Stem_cell						0,007459							
EMT_metastasis						0,250152							
Differentiation	0,000763			0,012981	3,91E-06							0,000126	0,092721
Proliferation						1,85E-07			0,001381		0,000699		

B

annot	MeC01	MeC02	MeC03	MeC04	MeC05	MeC06	MeC07	MeC08	MeC09	MeC10	MeC11	MeC12	MeC13
Gender		0,3543384	1,905E-78										
Geographic Origin	0,6094441												
tumor grade													
Age	0,0008936									0,0001937		0,8566904	0,0030221
tumor stage													
Child Pugh													
Normal Liver Histology													
BMI													
Without etiology													
Hepatitis B													
Hepatitis C													
Tobacco													
Alcohol Intake			0,5242813										
TERT	0,0001569											0,0054522	0,0003596
TP53				1,123E-07						0,0298639			
CTNNB1	0,0353261			7,196E-06				9,121E-06				0,0444991	0,0219505
AXIN1													
ALB													
ARID2													
ARID1A							1,375E-05						
ACVR2A													
NFE2L2													
RPS6KA3													
KEAP1													
RPL22													
CDKN2A													
CDKN1A													
RB1		7,985E-05											
TSC2													
ATP10B													
FGA													
MEF2C													
HNF1A													
ZNRF3													
EPHA4													
PTEN													
TSC1													
BAP1													
CCNHCC				1,161E-14						2,488E-08	6,166E-06		
G1		3,95E-11											
G2													
G3						0,0045302							
G4													
G5													
G6								2,864E-13				0,4439664	
MCPcounter inflamation						1,019E-31							
Liver_progenitor													
Stem_cell						0,0396705							
EMT_metastasis						0,0003544							
Differentiation	0,5613904			1,156E-05	8,943E-05							7,341E-05	0,0081267
Proliferation						6,319E-07			2,671E-05	2,039E-06	0,0037492		

Supplementary Table 3: Multivariate analysis (A) LICAFR (B) TCGA

Discussion

Le carcinome hépatocellulaire est une tumeur complexe et hétérogène résultant de l'accumulation d'altérations driver et de la dérégulation de multiples mécanismes cellulaires. L'efficacité des traitements actuels est limitée, et l'un des espoirs thérapeutiques réside dans la médecine de précision ciblant les altérations présentes dans chaque tumeur. La régulation épigénétique, par son caractère modulable et réversible, est particulièrement attractive pour le développement de nouvelles thérapies. Cependant, les mécanismes épigénétiques altérés dans les cancers et leurs conséquences transcriptionnelles sont encore mal compris. Au cours de cette thèse, l'analyse détaillée de l'expression et de la méthylation des CHC en utilisant une méthode innovante de déconvolution m'a permis d'isoler des sources de variation à l'œuvre dans le transcriptome et le méthylome des CHC, et de mieux comprendre les connexions entre anomalies génétiques, épigénétiques et transcriptionnelles.

1. Analyse transcriptomique des carcinomes hépatocellulaires

Dans cette thèse, j'ai participé au développement d'un outil bioinformatique permettant de reproduire aisément les classifications et signatures transcriptomiques déjà décrites dans les carcinomes hépatocellulaires. La comparaison de ces classifications montre que malgré leurs points communs, chacune identifie des dérégulations différentes, et soulève la difficulté de former des groupes homogènes.

L'analyse en composantes indépendantes (ICA) appliquée aux données RNAseq de CHC, m'a permis de caractériser les différentes sources de variation qui modulent le transcriptome de ces tumeurs. Parmi les composantes identifiées, certaines sont associées aux caractéristiques cliniques des patients ou des tumeurs, aux altérations génomiques driver, à des voies de signalisations oncogéniques ou à des groupes moléculaires de CHC décrits au laboratoire (Boyault et al., 2007) ou par d'autres (Chiang et al., 2008). L'ICA permet en outre de quantifier précisément la combinaison et l'intensité de ces dérégulations dans chaque tumeur, s'affranchissant des limitations imposées par l'assignation un seul groupe à chaque tumeur.

Elle m'a également permis de mettre en évidence une nouvelle association entre gain du bras chromosomique 8q et activation de la voie du stress oxydatif. Cette association suggère un rôle des gènes *MYC* et *PVT1* dans l'activation de cette voie. Pour valider cette hypothèse, il serait intéressant d'introduire des copies supplémentaires de ces gènes dans des lignées cellulaires de CHC ne possédant que deux copies du chromosome 8, et d'évaluer leur impact sur l'activation de la voie du stress oxydatif.

Par ailleurs, les échantillons étudiés en RNA-seq ne représentent qu'une petite partie des échantillons disponibles au laboratoire. Nous disposons notamment d'une quantification de l'expression de 380 gènes par PCR quantitative (qPCR) dans plus de 700 tumeurs. Il sera intéressant d'étendre l'analyse des composantes pour affiner les corrélations à la lumière des annotations cliniques et moléculaires de l'ensemble de cette série. Dans cette optique, j'ai établi, à partir de 125 échantillons analysés à la fois en RNAseq et en Fluidigm Real-Time PCR, un prédicteur de la composante 13 reflétant l'activité de la voie Wnt/B-catenine. Pour cela j'ai séparé aléatoirement les 125 échantillons en deux groupes : un jeu d'apprentissage contenant 2/3 des échantillons, et un jeu de validation contenant le tiers restant. En utilisant la corrélation de l'expression des 380 gènes en qPCR et l'activité de la composante 13 établie en RNA-seq, j'ai identifié 11 gènes marqueurs de la composante, avec une corrélation > 0.6 (*AMACR*, *AQP9*, *AXIN2*, *HULC*, *LAMA3*, *LEF1*, *LGR5*, *MERTK*, *NKD1*, *RHBG* et *TBX3*). J'ai ensuite utilisé un modèle de régression linéaire multiple pour prédire l'activation de la composante 13 à partir de l'expression de ces gènes en qPCR. Le modèle obtenu à un r2 de 0,9195 dans la série d'apprentissage, c'est à dire qu'il explique 91,95 % de la variance observée, et l'erreur quadratique moyenne de la prédiction dans la série de validation est de 0,14236. Ces paramètres et la visualisation des prédictions en fonction des valeurs d'activation réelles (cf. Figure 39) montrent que cette composante peut être estimée de manière très fiable à partir des données qPCR. J'ai donc pu utiliser ces données pour confirmer l'association aux mutations de *CTNNB1* et affiner l'étude des différences fines entre les différents types de mutations activatrices. Il serait intéressant de développer des modèles de prédiction pour les autres composantes, particulièrement pour celles qui semblent associées à certains sous-groupes de tumeurs ou mutations faiblement représentés dans les jeux de données RNAseq.

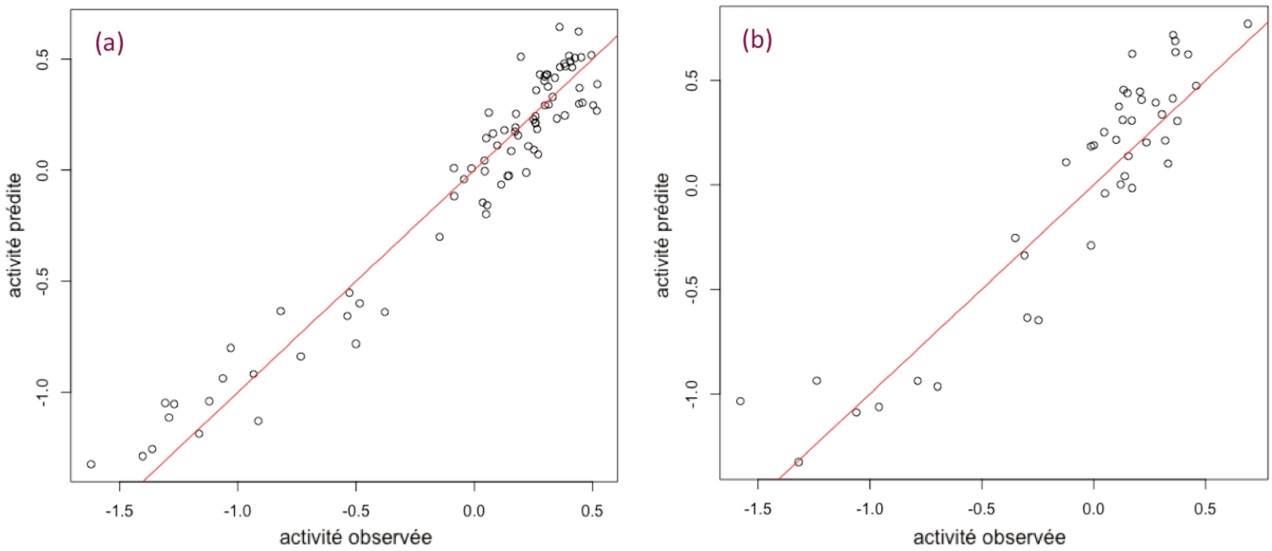


Figure 39 : ***Représentation de l'activité observée de la composante 13 dans les données RNA-seq en fonction de l'activité prédite par les données qPCR avec le modèle de régression linéaire multiple.*** *Le modèle a été mis en place sur le jeu de données d'apprentissage en utilisant les 11 gènes marqueurs sélectionnés. (a) Prédiction sur le jeu de données d'apprentissage et (b) sur le jeu de données de validation.*

L'analyse de sous-groupes de tumeurs présentant des signatures transcriptionnelles homogènes peut aussi permettre l'identification de nouvelles altérations *driver*. J'ai ainsi participé à la caractérisation d'un nouveau groupe de CHC initié par l'activation des cyclines A2 ou E1 (CCN-HCC). Ces tumeurs présentent une forte activité de la composante liée à la prolifération, ainsi qu'une suractivation de la voie ATR impliquée dans la réponse au stress réplicatif. En effet, l'activation des cyclines conduit ces tumeurs à entrer trop tôt en phase S, entraînant un stress réplicatif et une instabilité des fourches de réplication aboutissant à une signature de réarrangements structuraux particulière. Ce stress réplicatif intense pourrait offrir des possibilités thérapeutiques aux patients atteints de ce type de tumeur (Forment and O'Connor, 2018). On sait que les chimiothérapies conventionnelles affectent principalement les cellules en division active, et la signature de variants structuraux retrouvée dans les CCN-HCC, correspondant à des duplications en tandem, est marqueur de la réponse chimiothérapeutique dans les lignées cellulaires du cancer du sein (Menghi et al., 2016). Dans le cancer du sein, les tumeurs mutées BRCA1, impliquées dans la réparation de l'ADN, présentent la même signature de variants structuraux que les CHC altérés Cyclin. Il pourrait donc être intéressant de tester les médicaments approuvés pour ces tumeurs, comme les inhibiteurs de PARP, sur les CCN-HCC (Lord and Ashworth, 2017). De plus, il existe

actuellement plusieurs composés en essais de phase I et II qui ciblent les membres de la voie de réponse au stress de réplication ATR, CHK1 et WEE1. S'ils étaient amenés en clinique, ces composés seraient prometteurs pour le traitement des CCN-HCC (O'Connor, 2015).

Dans son article appliquant l'ACI à l'analyse du transcriptome des tumeurs de la vessie, Anne Biton a également comparé les composantes qu'elle a obtenues aux résultats de l'ACI appliquée à 10 jeux de données de différents types de cancer (Biton et al., 2014). Cette étude pan-cancer a permis d'identifier les composantes spécifiques de l'organe et du type de tumeurs analysé et les dérégulations communes entre les différents types de cancer. J'ai récupéré les composantes extraites dans cette étude et je les ai comparées à mes résultats d'ACI sur les carcinomes hépatocellulaires. De cette manière, j'ai pu constater que les processus ubiquitaires isolés par l'ACI dans l'étude des différents cancers le sont aussi dans celle des CHC. On retrouve notamment la composante liée au sexe du patient, la composante immunologique, liée à l'infiltration de la tumeur par les lymphocytes B et T, la composante caractéristique de la transition épithélio-mésenchymateuse et la composante associée au cycle cellulaire. En plus de celles-là, 4 autres composantes sont retrouvées dans tous les types de tumeurs dont le CHC, mais elles ne présentent pas d'associations claires. Il serait intéressant d'explorer plus avant ces composantes. Identifier des signatures moléculaires pan-cancer permet, dans le cadre de la médecine de précision, de proposer l'utilisation de médicaments ciblant des dérégulations communes à différents types de cancers, comme pour les tumeurs du foie CCN-HCC et les tumeurs du sein mutées *BRCA1*.

2. Analyse de la méthylation des carcinomes hépatocellulaires

L'utilisation de l'analyse en composantes indépendantes sur les données de puce de méthylation a permis de mieux comprendre les différents mécanismes influençant le méthylome des CHC. On retrouve des mécanismes connus, comme l'inactivation par la méthylation du second chromosome X chez la femme ou l'hyperméthylation de la chromatine bivalente liée à l'âge. D'autres composantes nouvelles sont associées à des groupes moléculaires et/ou altérations *driver* spécifiques. Ces composantes suggèrent de nouvelles

connexions entre altérations génétiques et épigénétiques, qu'il sera intéressant d'étudier au niveau fonctionnel. Par exemple, la signature MC2 définit un phénotype fortement hyperméthylateur dans les CHC. Elle est associée au sous-groupe transcriptomique G1 et à la présence de mutations de *AXIN1* et *RPS6KA3*. Dans l'introduction, on a pu voir que les phénotypes hyperméthylateurs dans les gliomes et les tumeurs du colon étaient liés à la dérégulation des déméthylases TET par mutation inactivatrice ou altération du cycle de Krebs, cependant on ne retrouve pas de diminution de l'expression des gènes codant pour les TET dans ce sous-groupe de tumeur. Pour élucider la cause moléculaire de cette composante, il sera intéressant d'étudier l'activité des différentes enzymes méthylases et déméthylases de l'ADN dans ce sous-groupe. Une autre signature qu'il serait intéressant de caractériser en profondeur est celle liée à la mutation *ARID1A*. Comment l'inactivation de ce gène perturbe-t-elle le remodelage de la chromatine et entraîne-t-elle les dérégulations observées dans la composante MC7 ? Pour mieux le comprendre, des analyses de ChIP-seq et/ou ATAC-seq de CHC mutés *ARID1A* seront extrêmement utiles.

La caractérisation globale des signatures pourra également être améliorée. Plusieurs paramètres génomiques pourront être ajoutés pour aider à l'interprétation, comme les informations relatives aux régions soumises à empreintes, aux régions répétées et les associations avec les signatures mutationnelles et signatures de variants structuraux caractérisées au laboratoire sur les données WGS (Bayard et al., 2018; Letouzé et al., 2017; Schulze et al., 2015). Il serait intéressant de voir si les mutations et variants structuraux entraînent directement des changements de méthylation sur les zones touchées ou si une perte de méthylation, par exemple dans les régions répétées où elle est censée protéger l'ADN des translocations, favorise l'apparition de certains réarrangements structuraux.

J'ai également commencé une analyse en composantes indépendantes pan-cancer sur les données de méthylation (puce illumina 450k) de 17 types de tumeurs provenant de différents tissus récupérés sur la plateforme TCGA (vessie - BLCA, sein - BRCA, cerveau - GBM/LGG, cellules de l'épiderme - HNSC, rein - KIRP, cellules du système hématopoïétique et réticulo-endothéliale - LAML, poumons - LUAD/LUSC, PAAD, prostate - PRAD, peau - SKCM, estomac et œsophage - STAD/ ESCA, thyroïde - THCA, utérus - UCEC et col de l'utérus - CESC). En comparant les composantes obtenues entre elles, j'ai pu identifier des signatures communes

à différents cancers (cf. Figure 40). Comme attendu les signatures MC1, reflétant l'hyperméthylation de la chromatine bivalente liée à l'âge, MC3, liée au sexe du patient et MC6 qui est caractéristique de l'infiltrat immunitaire sont retrouvées dans différents tissus. On remarque que la signature MC2, associée aux tumeurs du sous-groupe transcriptomique de CHC G1 (Boyault et al., 2007) est étonnamment retrouvée dans d'autres cancers. Les tumeurs G1 dans cette composante sont caractérisées par un phénotype hyperméthylateur. Il sera intéressant de voir si les composantes auxquelles la signature est corrélée sont aussi caractéristiques de phénotypes hyperméthylateurs dans les autres cancers, et est-ce que des altérations à l'origine de ce phénotype ont été trouvées. Les 3 signatures d'hypométhylation isolées dans le foie, MC11, MC12 et MC13 sont également présentes dans les tumeurs provenant d'autres types de tissus. La signature MC11 est associée aux CHC présentant une altération des cyclines A2 ou E1, la forte capacité à cycler de ces échantillons augmente leur âge mitotique, ce qui entraîne une perte de la méthylation des cellules (Zhou et al., 2018). Cette composante est probablement retrouvée dans d'autres cancers qui ont des sous-types tumoraux présentant également une forte prolifération et une activation du cycle cellulaire. Une autre signature liée aux CCN-CHC est la signature MC4. Contrairement à la signature MC11, elle n'est pas retrouvée dans d'autres cancers. Il sera intéressant de comprendre pourquoi les dérégulations isolées dans la signature MC4 sont spécifiques du foie contrairement à celle isolées dans la signature MC11. Je n'ai pu effectuer qu'une analyse préliminaire des résultats de l'ACI pan-cancer des données de méthylation. Approfondir ces analyses nous permettra d'en apprendre davantage sur les mécanismes ubiquitaires à l'œuvre dans les tumeurs et ceux spécifiques aux cancers du foie. Le package MethICA que je suis en train de développer, permettra de visualiser les résultats des ICA pan-cancer et d'aider à l'interprétation de chaque composante.

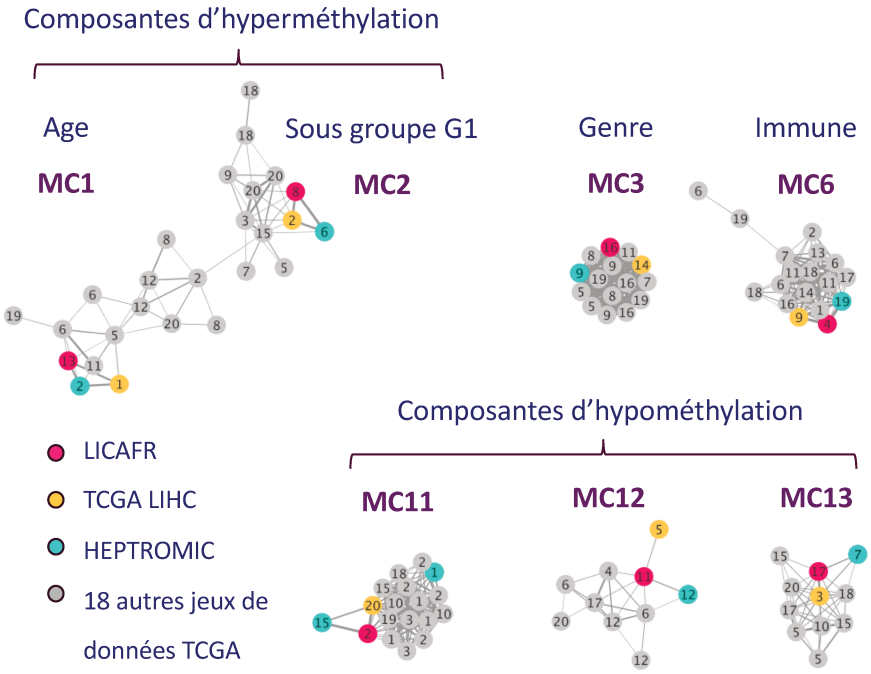


Figure 40 : ***Graphe représentant les corrélations entre les résultats Pan-cancer de l'ACI***. *Chaque cercle représente une composante, et deux composantes sont reliées entre elles si leur corrélation est > 0.35 (corrélation de Pearson calculée sur la contribution des CpG en commun).*

Dans ce travail j'ai exclusivement travaillé sur la méthylation, mais bien d'autres mécanismes épigénétiques interagissent pour réguler la transcription. Pour le moment, seules les données de contexte chromatinien déterminé sur un foie normal ou la lignée tumorale HEPG2 ont été utilisées. Générer ce type de données directement sur les tumeurs permettrait d'étudier plus précisément les modifications épigénétiques liées à l'accessibilité de la chromatine, le marquage des histones, la liaison des facteurs de transcription, et d'identifier les liens entre les différents niveaux de régulation épigénétique.

3. Intégration des données

Au cours de ma thèse, j'ai analysé de manière séparée les composantes transcriptionnelles et de méthylation. Le package ELMER m'a permis de mettre en lumière les liens entre méthylation des CpG et expression des gènes voisins et d'évaluer l'impact transcriptionnel des différentes composantes. Il serait intéressant d'étudier de manière plus globale l'interaction des composantes RNAseq et de méthylation. Une première analyse de la contribution des échantillons m'a permis de mettre en évidence, comme attendu, la forte corrélation entre les

composantes transcriptomiques et épigénétiques des composantes immunologiques (Pearson cor = 0.639, p-value = 5.17E-15), liées au sexe du patient (Pearson cor = 0.769, p-value = 1.83E-24), et à l'activation de la voie Wnt/B-catenin (Pearson cor = 0.857, p-value = 1.63E-35).

Pour aller plus loin, il serait intéressant de rechercher des composantes multi-omiques permettant de faire le lien entre les différentes signatures et altérations moléculaires identifiées. Pour cela, toutes les altérations détectées aux différents niveaux d'étude du génome pourront être combinées : les mutations somatiques, les variants structuraux (en termes d'altération driver mais aussi de signatures mutationnelles), les aberrations chromosomiques, les niveaux d'expression des microARNs et des protéines, ainsi que les signatures épigénétiques et transcriptomiques des ARNs codants et non codants. L'hétérogénéité des données et la complexité des variations inter-omiques sont deux défis majeurs pour la classification intégrative. Il existe diverses approches statistiques (Wang and Gu, 2016) permettant identifier des groupes d'altérations co-occurrentes impliquant plusieurs niveaux de dérégulation, afin d'élucider les mécanismes oncogéniques sous-jacents.

Parmi elles, une approche particulièrement intéressante est la méthode MOFA (Multi-Omics Factor Analysis) (Argelaguet et al., 2018). Cette méthode peut être considérée comme une généralisation de l'analyse en composantes principales (ACP) aux données multi-omiques. Elle permet d'identifier les axes d'hétérogénéité qui sont communs à plusieurs niveaux d'analyses omiques et ceux qui sont propres à un seul niveau de régulation. Cette méthode a été appliquée avec succès à une cohorte de 200 échantillons de patients atteints de leucémie lymphocytaire chronique, profilés pour les mutations somatiques, l'expression des gènes, la méthylation de l'ADN et les réponses aux médicaments *ex vivo*. Cela a permis d'identifier les principales dimensions de l'hétérogénéité de la maladie, y compris le statut des immunoglobulines à chaîne lourde dans les régions variables, la trisomie du chromosome 12 et les facteurs précédemment sous-estimés, tels que la réponse au stress oxydatif (Argelaguet et al., 2018). Il sera intéressant d'appliquer cette méthode aux données multi-omiques des carcinomes hépatocellulaires disponibles au laboratoire et de comparer les composantes multi-omiques aux résultats obtenus dans cette thèse. De plus, comme l'ACP, avec ses limites présentées en introduction, cherche non pas à isoler les sources de variation mais à expliquer

au maximum la variabilité observée des données, il serait donc prometteur d'adapter l'approche MOFA en remplaçant l'ACP par une analyse en composantes indépendantes.

4. Conclusion

L'utilisation de l'analyse en composantes indépendantes sur les données biologiques est un outil puissant pour isoler les dérégulations à l'œuvre dans les tumeurs. Dans cette thèse, j'ai appliqué l'ACI avec succès pour étudier le transcriptome et le méthylome des CHC. L'interprétation des composantes de méthylation nécessite la prise en compte du contexte chromatinien des CpG. Pour cela, j'ai développé des méthodes de visualisation qui facilitent leur caractérisation. Elles sont en cours d'implémentation dans un package R, qui sera disponible publiquement sur GitHub et pourra être réutilisé par le laboratoire et les autres équipes intéressés par l'utilisation de l'ICA sur les données de méthylation.

La poursuite de ces études permettra de mieux comprendre l'hétérogénéité tumorale des CHC, grâce à des méthodes d'analyse qui pourront être appliquées à d'autres types de tumeurs. L'intégration des données analysées avec des méthodes de déconvolution permet de mieux comprendre les dérégulations à l'œuvre dans les tumeurs et leurs interactions. Avec l'utilisation de plus en plus fréquente du RNA-seq pour caractériser le transcriptome des tumeurs en contexte clinique, il devient possible de quantifier précisément les différents niveaux de dérégulation dans chaque tumeur, et de corréler ceux-ci à la réponse au traitement. Cela ouvrira la voie au développement de nouvelles stratégies thérapeutiques prenant en compte non seulement les altérations moléculaires *driver*, mais aussi les caractéristiques moléculaires fines de chaque tumeur.

Bibliographie

Aapola, U., Shibuya, K., Scott, H.S., Ollila, J., Vihinen, M., Heino, M., Shintani, A., Kawasaki, K., Minoshima, S., Krohn, K., et al. (2000). Isolation and Initial Characterization of a Novel Zinc Finger Gene, DNMT3L, on 21q22.3, Related to the Cytosine-5-Methyltransferase 3 Gene Family. Genomics *65*, 293–298.

Abdi, H., and Williams, L.J. (2010). Principal Component Analysis.

Abou-Alfa, G.K., Meyer, T., Cheng, A.-L., El-Khoueiry, A.B., Rimassa, L., Ryoo, B.-Y., Cicin, I., Merle, P., Chen, Y., Park, J.-W., et al. (2018). Cabozantinib in Patients with Advanced and Progressing Hepatocellular Carcinoma. N. Engl. J. Med. *379*, 54–63.

Ahn, S.-M., Jang, S.J., Shim, J.H., Kim, D., Hong, S.-M., Sung, C.O., Baek, D., Haq, F., Ansari, A.A., Lee, S.Y., et al. (2014). Genomic portrait of resectable hepatocellular carcinomas: Implications of RB1 and FGF19 aberrations for patient stratification. Hepatology *60*, 1972–1982.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Anders, S., Pyl, P.T., and Huber, W. (2015). HTSeq—a Python framework to work with high-throughput sequencing data. Bioinformatics *31*, 166–169.

Aran, D., Sabato, S., and Hellman, A. (2013). DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. *14*, R21.

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., and Stegle, O. (2018). Multi-Omics Factor Analysis—a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. *14*, e8124.

Ascha, M.S., Hanouneh, I.A., Lopez, R., Tamimi, T.A.-R., Feldstein, A.F., and Zein, N.N. (2010). The incidence and risk factors of hepatocellular carcinoma in patients with nonalcoholic steatohepatitis. Hepatology *51*, 1972–1978.

Ashapkin, V.V., Kutueva, L.I., and Vanyushin, B.F. (2017). Aging as an Epigenetic Phenomenon. Curr. Genomics *18*, 385–407.

Babina, I.S., and Turner, N.C. (2017). Advances and challenges in targeting FGFR signalling in cancer. Nat. Rev. Cancer *17*, 318–332.

Bannister, A.J., and Kouzarides, T. (2011). Regulation of chromatin by histone modifications. Cell Res. *21*, 381–395.

Basham, K.J., Rodriguez, S., Turcu, A.F., Lerario, A.M., Logan, C.Y., Rysztak, M.R., Gomez-Sanchez, C.E., Breault, D.T., Koo, B.-K., Clevers, H., et al. (2019). A ZNRF3-dependent Wnt/β-catenin signaling gradient is required for adrenal homeostasis. Genes Dev. *33*, 209–220.

Bayard, Q., Meunier, L., Peneau, C., Renault, V., Shinde, J., Nault, J.-C., Mami, I., Couchy, G., Amaddeo, G., Tubacher, E., et al. (2018). Cyclin A2/E1 activation defines a

hepatocellular carcinoma subclass with a rearrangement signature of replication stress. Nat. Commun. *9*, 5235.

Baylin, S.B., and Jones, P.A. (2016). Epigenetic Determinants of Cancer. Cold Spring Harb. Perspect. Biol. *8*.

Bell, C.G., Lowe, R., Adams, P.D., Baccarelli, A.A., Beck, S., Bell, J.T., Christensen, B.C., Gladyshev, V.N., Heijmans, B.T., Horvath, S., et al. (2019). DNA methylation aging clocks: challenges and recommendations. Genome Biol. *20*, 249.

Bella, T.L., Imbeaud, S., Peneau, C., Mami, I., Datta, S., Bayard, Q., Caruso, S., Hirsch, T.Z., Calderaro, J., Morcrette, G., et al. (2019). Adeno-associated virus in the liver: natural history and consequences in tumour development. Gut.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002). A stability based method for discovering structure in clustered data. Pac. Symp. Biocomput. Pac. Symp. Biocomput. 6–17.

Berman, B.P., Weisenberger, D.J., Aman, J.F., Hinoue, T., Ramjan, Z., Liu, Y., Noushmehr, H., Lange, C.P.E., van Dijk, C.M., Tollenaar, R.A.E.M., et al. (2011). Regions of focal DNA hypermethylation and long-range hypomethylation in colorectal cancer coincide with nuclear lamina–associated domains. Nat. Genet. *44*, 40–46.

Bestor, T.H. (2000). The DNA methyltransferases of mammals. Hum. Mol. Genet. *9*, 2395–2402.

Bestor, T.H., and Ingram, V.M. (1983). Two DNA methyltransferases from murine erythroleukemia cells: purification, sequence specificity, and mode of interaction with DNA. Proc. Natl. Acad. Sci. U. S. A. *80*, 5559–5563.

Bhattacharjee, A., Richards, W.G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M., et al. (2001). Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses. Proc. Natl. Acad. Sci. U. S. A. *98*, 13790–13795.

Bhutani, N., Brady, J.J., Damian, M., Sacco, A., Corbel, S.Y., and Blau, H.M. (2010). Reprogramming towards pluripotency requires AID-dependent DNA demethylation. Nature *463*, 1042–1047.

Bhutani, N., Burns, D.M., and Blau, H.M. (2011). DNA Demethylation Dynamics. Cell *146*, 866–872.

Biden, K., Young, J., Buttenshaw, R., Searle, J., Cooksley, G., Xu, D.B., and Leggett, B. (1997). Frequency of mutation and deletion of the tumor suppressor gene CDKN2A (MTS1/p16) in hepatocellular carcinoma from an Australian population. Hepatology *25*, 593–597.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Biton, A., Bernard-Pierrot, I., Lou, Y., Krucker, C., Chapeaublanc, E., Rubio-Pérez, C., López-Bigas, N., Kamoun, A., Neuzillet, Y., Gestraud, P., et al. (2014). Independent Component Analysis Uncovers the Landscape of the Bladder Tumor Transcriptome and Reveals Insights into Luminal and Basal Subtypes. Cell Rep. *9*, 1235–1245.

Biton, A., Zinovyev, A., Barillot, E., and Radvanyi, F. MineICA: Independent component analysis of transcriptomic data. 29.

Bostick, M., Kim, J.K., Estève, P.-O., Clark, A., Pradhan, S., and Jacobsen, S.E. (2007). UHRF1 plays a role in maintaining DNA methylation in mammalian cells. Science *317*, 1760–1764.

Boyault, S., Rickman, D.S., Reyniès, A. de, Balabaud, C., Rebouissou, S., Jeannot, E., Hérault, A., Saric, J., Belghiti, J., Franco, D., et al. (2007). Transcriptome classification of HCC is related to gene alterations and to new therapeutic targets. Hepatology *45*, 42–52.

Brioude, F., Kalish, J.M., Mussa, A., Foster, A.C., Bliek, J., Ferrero, G.B., Boonen, S.E., Cole, T., Baker, R., Bertoletti, M., et al. (2018). Clinical and molecular diagnosis, screening and management of Beckwith–Wiedemann syndrome: an international consensus statement. Nat. Rev. Endocrinol. *14*, 229.

Bruix, J., and Llovet, J.M. (2002). Prognostic prediction and treatment strategy in hepatocellular carcinoma. Hepatology *35*, 519–524.

Bruix, J., Qin, S., Merle, P., Granito, A., Huang, Y.-H., Bodoky, G., Pracht, M., Yokosuka, O., Rosmorduc, O., Breder, V., et al. (2017). Regorafenib for patients with hepatocellular carcinoma who progressed on sorafenib treatment (RESORCE): a randomised, double-blind, placebo-controlled, phase 3 trial. Lancet Lond. Engl. *389*, 56–66.

Brunner, S.F., Roberts, N.D., Wylie, L.A., Moore, L., Aitken, S.J., Davies, S.E., Sanders, M.A., Ellis, P., Alder, C., Hooks, Y., et al. (2019). Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature *574*, 538–542.

Bryan, H.K., Olayanju, A., Goldring, C.E., and Park, B.K. (2013). The Nrf2 cell defence pathway: Keap1-dependent and -independent mechanisms of regulation. Biochem. Pharmacol. *85*, 705–717.

Bugli, C., and Lambert, P. (2007). Comparison between principal component analysis and independent component analysis in electroencephalograms modelling. Biom. J. Biom. Z. *49*, 312–327.

Burger, L., Gaidatzis, D., Schübeler, D., and Stadler, M.B. (2013). Identification of active regulatory regions from DNA methylation data. Nucleic Acids Res. *41*, e155–e155.

Calderaro, J., Letouzé, E., Bayard, Q., Boulai, A., Renault, V., Deleuze, J.-F., Bestard, O., Franco, D., Zafrani, E.-S., Nault, J.-C., et al. (2018). Systemic AA Amyloidosis Caused by Inflammatory Hepatocellular Adenoma. N. Engl. J. Med. *379*, 1178–1180.

Cardoso, J.-F. (1997). Infomax and maximum likelihood for blind source separation. IEEE Signal Process. Lett. *4*, 112–114.

Castillo, J., López-Rodas, G., and Franco, L. (2017). Histone Post-Translational Modifications and Nucleosome Organisation in Transcriptional Regulation: Some Open Questions. In Protein Reviews: Volume 18, M.Z. Atassi, ed. (Singapore: Springer), pp. 65–92.

Cedar, H., and Bergman, Y. (2009). Linking DNA methylation and histone modification: patterns and paradigms. Nat. Rev. Genet. *10*, 295–304.

Chang, D.Z., Kumar, V., Ma, Y., Li, K., and Kopetz, S. (2009). Individualized therapies in colorectal cancer: KRAS as a marker for response to EGFR-targeted therapy. J. Hematol. Oncol.J Hematol Oncol *2*, 18.

Chaudhary, P., Bhadana, U., Singh, R.A.K., and Ahuja, A. (2015). Primary hepatic angiosarcoma. Eur. J. Surg. Oncol. EJSO *41*, 1137–1143.

Chen, T., and Li, E. (2004). Structure and function of eukaryotic DNA methyltransferases. Curr. Top. Dev. Biol. *60*, 55–89.

Chen, S., Sanjana, N.E., Zheng, K., Shalem, O., Lee, K., Shi, X., Scott, D.A., Song, J., Pan, J.Q., Weissleder, R., et al. (2015). Genome-wide CRISPR screen in a mouse model of tumor growth and metastasis. Cell *160*, 1246–1260.

Chen, T., Hevi, S., Gay, F., Tsujimoto, N., He, T., Zhang, B., Ueda, Y., and Li, E. (2007). Complete inactivation of DNMT1 leads to mitotic catastrophe in human cancer cells. Nat. Genet. *39*, 391–396.

Cheng, X., and Blumenthal, R.M. (2011). Introduction—Epiphanies in Epigenetics. Prog. Mol. Biol. Transl. Sci. *101*, 1–21.

Cheng, J., Wei, D., Ji, Y., Chen, L., Yang, L., Li, G., Wu, L., Hou, T., Xie, L., Ding, G., et al. (2018). Integrative analysis of DNA methylation and gene expression reveals hepatocellular carcinoma-specific diagnostic biomarkers. Genome Med. *10*, 42.

Cheng, Y., Zhang, C., Zhao, J., Wang, C., Xu, Y., Han, Z., Jiang, G., Guo, X., Li, R., Bu, X., et al. (2010). Correlation of CpG island methylator phenotype with poor prognosis in hepatocellular carcinoma. Exp. Mol. Pathol. *88*, 112–117.

Chiang, D.Y., Villanueva, A., Hoshida, Y., Peix, J., Newell, P., Minguez, B., LeBlanc, A.C., Donovan, D.J., Thung, S.N., Sole, M., et al. (2008). Focal Gains of Vascular Endothelial Growth Factor A and Molecular Classification of Hepatocellular Carcinoma. Cancer Res. *68*, 6779–6788.

Coste, A. de L., Romagnolo, B., Billuart, P., Renard, C.-A., Buendia, M.-A., Soubrane, O., Fabre, M., Chelly, J., Beldjord, C., Kahn, A., et al. (1998). Somatic mutations of the β-catenin gene are frequent in mouse and human hepatocellular carcinomas. Proc. Natl. Acad. Sci. *95*, 8847–8851.

Coulouarn, C., Factor, V.M., and Thorgeirsson, S.S. (2008). Transforming growth factor-beta gene expression signature in mouse hepatocytes predicts clinical outcome in human cancer. Hepatol. Baltim. Md *47*, 2059–2067.

Dang, L., White, D.W., Gross, S., Bennett, B.D., Bittinger, M.A., Driggers, E.M., Fantin, V.R., Jang, H.G., Jin, S., Keenan, M.C., et al. (2010). Cancer-associated IDH1 mutations produce 2-hydroxyglutarate. Nature *465*, 966–966.

Das, M. (2018). Refametinib in RAS-mutated hepatocellular cancer. Lancet Oncol. *19*, e389.

Daughtry, B.L., and Chavez, S.L. (2016). Chromosomal instability in mammalian pre-implantation embryos: potential causes, detection methods, and clinical consequences. Cell Tissue Res. *363*, 201–225.

Davis, C.A., Hitz, B.C., Sloan, C.A., Chan, E.T., Davidson, J.M., Gabdank, I., Hilton, J.A., Jain, K., Baymuradov, U.K., Narayanan, A.K., et al. (2018). The Encyclopedia of DNA elements (ENCODE): data portal update. Nucleic Acids Res. *46*, D794–D801.

Day, K., Waite, L.L., Thalacker-Mercer, A., West, A., Bamman, M.M., Brooks, J.D., Myers, R.M., and Absher, D. (2013). Differential DNA methylation with age displays both common and dynamic features across human tissues that are influenced by CpG landscape. Genome Biol. *14*, R102.

Deltour, S., Chopin, V., and Leprince, D. (2005). Modifications épigénétiques et cancer. MS Médecine Sci. *21*, 405–411.

Dileep, V., Rivera-Mulia, J.C., Sima, J., and Gilbert, D.M. (2015). Large-Scale Chromatin Structure–Function Relationships during the Cell Cycle and Development: Insights from Replication Timing. Cold Spring Harb. Symp. Quant. Biol. *80*, 53–63.

Dobrovic, A., and Simpfendorfer, D. (1997). Methylation of the BRCA1 gene in sporadic breast cancer. Cancer Res. *57*, 3347–3350.

Duan, Y., Tian, L., Gao, Q., Liang, L., Zhang, W., Yang, Y., Zheng, Y., Pan, E., Li, S., and Tang, N. (2016). Chromatin remodeling gene ARID2 targets cyclin D1 and cyclin E1 to suppress hepatoma cell progression. Oncotarget *7*, 45863–45875.

Dudoit, S., and Fridlyand, J. (2002). A prediction-based resampling method for estimating the number of clusters in a dataset. Genome Biol. *3*, research0036.1.

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., et al. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Eden, A., Gaudet, F., Waghmare, A., and Jaenisch, R. (2003). Chromosomal Instability and Tumors Promoted by DNA Hypomethylation. Science *300*, 455–455.

Edmondson, H.A., and Steiner, P.E. (1954). Primary carcinoma of the liver: a study of 100 cases among 48,900 necropsies. Cancer *7*, 462–503.

Ehrlich, M. (2009). DNA hypomethylation in cancer cells. Epigenomics *1*, 239–259.

Eisen, M.B., Spellman, P.T., Brown, P.O., and Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. Proc. Natl. Acad. Sci. *95*, 14863–14868.

El–Serag, H.B., and Rudolph, K.L. (2007). Hepatocellular Carcinoma: Epidemiology and Molecular Carcinogenesis. Gastroenterology *132*, 2557–2576.

ENCODE Project Consortium (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Ernst, J., and Kellis, M. (2012). ChromHMM: automating chromatin state discovery and characterization. Nat. Methods *9*, 215–216.

Esteller, M., Hamilton, S.R., Burger, P.C., Baylin, S.B., and Herman, J.G. (1999). Inactivation of the DNA Repair Gene O6-Methylguanine-DNA Methyltransferase by Promoter Hypermethylation is a Common Event in Primary Human Neoplasia. Cancer Res. *59*, 793–797.

Esteller, M., Corn, P.G., Baylin, S.B., and Herman, J.G. (2001). A Gene Hypermethylation Profile of Human Cancer. Cancer Res. *61*, 3225–3229.

Evans, A.A., O'Connell, A.P., Pugh, J.C., Mason, W.S., Shen, F.M., Chen, G.C., Lin, W.Y., Dia, A., M'Boup, S., Dramé, B., et al. (1998). Geographic variation in viral load among hepatitis B carriers with differing risks of hepatocellular carcinoma. Cancer Epidemiol. Prev. Biomark. *7*, 559–565.

Farazi, P.A., Glickman, J., Jiang, S., Yu, A., Rudolph, K.L., and DePinho, R.A. (2003). Differential impact of telomere dysfunction on initiation and progression of hepatocellular carcinoma. Cancer Res. *63*, 5021–5027.

Feinberg, A.P., Ohlsson, R., and Henikoff, S. (2006). The epigenetic progenitor origin of human cancer. Nat. Rev. Genet. *7*, 21–33.

Ferlay, J., Colombet, M., Soerjomataram, I., Mathers, C., Parkin, D.M., Piñeros, M., Znaor, A., and Bray, F. (2019). Estimating the global cancer incidence and mortality in 2018: GLOBOCAN sources and methods. Int. J. Cancer *144*, 1941–1953.

Feuk, L., Carson, A.R., and Scherer, S.W. (2006). Structural variation in the human genome. Nat. Rev. Genet. *7*, 85–97.

Field, A.E., Robertson, N.A., Wang, T., Havas, A., Ideker, T., and Adams, P.D. (2018). DNA Methylation Clocks in Aging: Categories, Causes, and Consequences. Mol. Cell *71*, 882–895.

Forment, J.V., and O'Connor, M.J. (2018). Targeting the replication stress response in cancer. Pharmacol. Ther. *188*, 155–167.

Freytag, V., Carrillo-Roa, T., Milnik, A., Sämann, P.G., Vukojevic, V., Coynel, D., Demougin, P., Egli, T., Gschwind, L., Jessen, F., et al. (2017). A peripheral epigenetic signature of immune system genes is linked to neocortical thickness and memory. Nat. Commun. *8*.

Frezza, C., Pollard, P.J., and Gottlieb, E. (2011). Inborn and acquired metabolic defects in cancer. J. Mol. Med. *89*, 213–220.

Fujimoto, A., Totoki, Y., Abe, T., Boroevich, K.A., Hosoda, F., Nguyen, H.H., Aoki, M., Hosono, N., Kubo, M., Miya, F., et al. (2012). Whole-genome sequencing of liver cancers identifies etiological influences on mutation patterns and recurrent mutations in chromatin regulators. Nat. Genet. *44*, 760–764.

Fujimoto, A., Furuta, M., Totoki, Y., Tsunoda, T., Kato, M., Shiraishi, Y., Tanaka, H., Taniguchi, H., Kawakami, Y., Ueno, M., et al. (2016). Whole-genome mutational landscape

and characterization of noncoding and structural mutations in liver cancer. Nat. Genet. *48*, 500–509.

Galuppo, R., Ramaiah, D., Ponte, O.M., and Gedaly, R. (2014). Molecular Therapies in Hepatocellular Carcinoma: What can we target? Dig. Dis. Sci. *59*, 1688–1697.

Gama-Sosa, M.A., Slagel, V.A., Trewyn, R.W., Oxenhandler, R., Kuo, K.C., Gehrke, C.W., and Ehrlich, M. (1983). The 5-methylcytosine content of DNA from human tumors. Nucleic Acids Res. *11*, 6883–6894.

Gazin, C., Wajapeyee, N., Gobeil, S., Virbasius, C.-M., and Green, M.R. (2007). An elaborate pathway required for Ras-mediated epigenetic silencing. Nature *449*, 1073–1077.

Ginès, P., Graupera, I., Lammert, F., Angeli, P., Caballeria, L., Krag, A., Guha, I.N., Murad, S.D., and Castera, L. (2016). Screening for liver fibrosis in the general population: a call for action. Lancet Gastroenterol. Hepatol. *1*, 256–260.

Gräff, J., and Mansuy, I.M. (2008). Epigenetic codes in cognition and behaviour. Behav. Brain Res. *192*, 70–87.

Grazioli, L., Ambrosini, R., Frittoli, B., Grazioli, M., and Morone, M. (2017). Primary benign liver lesions. Eur. J. Radiol. *95*, 378–398.

Guan, B., Wang, T.-L., and Shih, I.-M. (2011). ARID1A, a factor that promotes formation of SWI/SNF-mediated chromatin remodeling, is a tumor suppressor in gynecologic cancers. Cancer Res. *71*, 6718–6727.

Guha, S., Rastogi, R., and Shim, K. (1998). CURE: An Efficient Clustering Algorithm for Large Databases. In Proceedings of the 1998 ACM SIGMOD International Conference on Management of Data, (New York, NY, USA: ACM), pp. 73–84.

Guichard, C., Amaddeo, G., Imbeaud, S., Ladeiro, Y., Pelletier, L., Maad, I.B., Calderaro, J., Bioulac-Sage, P., Letexier, M., Degos, F., et al. (2012). Integrated analysis of somatic mutations and focal copy-number changes identifies key genes and pathways in hepatocellular carcinoma. Nat. Genet. *44*, 694–698.

Haller, F., Bieg, M., Will, R., Körner, C., Weichenhan, D., Bott, A., Ishaque, N., Lutsik, P., Moskalev, E.A., Mueller, S.K., et al. (2019). Enhancer hijacking activates oncogenic transcription factor NR4A3 in acinic cell carcinomas of the salivary glands. Nat. Commun. *10*, 1–13.

Han, T.-S., Ban, H.S., Hur, K., and Cho, H.-S. (2018). The Epigenetic Regulation of HCC Metastasis. Int. J. Mol. Sci. *19*.

Hanahan, D., and Weinberg, R.A. (2000). The Hallmarks of Cancer. Cell *100*, 57–70.

Hanna, N., Parfait, B., Vidaud, D., and Vidaud, M. (2005). [Mutation mechanisms and their consequences]. Med. Sci. MS *21*, 969–980.

Hardy, T., Zeybel, M., Day, C.P., Dipper, C., Masson, S., McPherson, S., Henderson, E., Tiniakos, D., White, S., French, J., et al. (2017). Plasma DNA methylation: a potential

biomarker for stratification of liver fibrosis in non-alcoholic fatty liver disease. Gut *66*, 1321–1328.

Harrow, J., Frankish, A., Gonzalez, J.M., Tapanari, E., Diekhans, M., Kokocinski, F., Aken, B.L., Barrell, D., Zadissa, A., Searle, S., et al. (2012). GENCODE: The reference human genome annotation for The ENCODE Project. Genome Res. *22*, 1760–1774.

He, Y.-F., Li, B.-Z., Li, Z., Liu, P., Wang, Y., Tang, Q., Ding, J., Jia, Y., Chen, Z., Li, L., et al. (2011). Tet-Mediated Formation of 5-Carboxylcytosine and Its Excision by TDG in Mammalian DNA. Science *333*, 1303–1307.

Herman, J.G., Merlo, A., Mao, L., Lapidus, R.G., Issa, J.P., Davidson, N.E., Sidransky, D., and Baylin, S.B. (1995). Inactivation of the CDKN2/p16/MTS1 gene is frequently associated with aberrant DNA methylation in all common human cancers. Cancer Res. *55*, 4525–4530.

Herman, J.G., Umar, A., Polyak, K., Graff, J.R., Ahuja, N., Issa, J.-P.J., Markowitz, S., Willson, J.K.V., Hamilton, S.R., Kinzler, K.W., et al. (1998). Incidence and functional consequences of hMLH1 promoter hypermethylation in colorectal carcinoma. Proc. Natl. Acad. Sci. U. S. A. *95*, 6870–6875.

Hnisz, D., Weintraub, A.S., Day, D.S., Valton, A.-L., Bak, R.O., Li, C.H., Goldmann, J., Lajoie, B.R., Fan, Z.P., Sigova, A.A., et al. (2016). Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science *351*, 1454–1458.

Holliday, R., and Pugh, J.E. (1975). DNA modification mechanisms and gene activity during development. Science *187*, 226–232.

Honeyman, J.N., Simon, E.P., Robine, N., Chiaroni-Clarke, R., Darcy, D.G., Lim, I.I.P., Gleason, C.E., Murphy, J., Rosenberg, B.R., Teegan, L., et al. (2014). Detection of a Recurrent DNAJB1-PRKACA Chimeric Transcript in Fibrolamellar Hepatocellular Carcinoma. Science *343*, 1010–1014.

Horvath, S., and Raj, K. (2018). DNA methylation-based biomarkers and the epigenetic clock theory of ageing. Nat. Rev. Genet. *19*, 371–384.

Hoshida, Y., Nijman, S.M.B., Kobayashi, M., Chan, J.A., Brunet, J.-P., Chiang, D.Y., Villanueva, A., Newell, P., Ikeda, K., Hashimoto, M., et al. (2009). Integrative Transcriptome Analysis Reveals Common Molecular Subclasses of Human Hepatocellular Carcinoma. Cancer Res. *69*, 7385–7392.

Hsia, C.C., Kleiner, D.E., Axiotis, C.A., Bisceglie, A.D., Nomura, A.M.Y., Stemmermann, G.N., and Tabor, E. (1992). Mutations of p53 Gene in Hepatocellular Carcinoma: Roles of Hepatitis B Virus and Aflatoxin Contamination in the Diet. JNCI J. Natl. Cancer Inst. *84*, 1638–1641.

Hughes, L.A.E., Melotte, V., Schrijver, J. de, Maat, M. de, Smit, V.T.H.B.M., Bovée, J.V.M.G., French, P.J., Brandt, P.A. van den, Schouten, L.J., Meyer, T. de, et al. (2013). The CpG Island Methylator Phenotype: What's in a Name? Cancer Res. *73*, 5858–5868.

Hyun, K., Jeon, J., Park, K., and Kim, J. (2017). Writing, erasing and reading histone lysine methylations. Exp. Mol. Med. *49*, e324–e324.

Hyvärinen, A. (1999). Fast and robust fixed-point algorithms for independent component analysis. IEEE Trans. Neural Netw. *10*, 626–634.

Hyvärinen, A., and Oja, E. (2000). Independent component analysis: algorithms and applications. Neural Netw. *13*, 411–430.

Illingworth, R.S., and Bird, A.P. (2009). CpG islands – 'A rough guide.' FEBS Lett. *583*, 1713–1720.

Irizarry, R.A., Ladd-Acosta, C., Wen, B., Wu, Z., Montano, C., Onyango, P., Cui, H., Gabo, K., Rongione, M., Webster, M., et al. (2009). Genome-wide methylation analysis of human colon cancer reveals similar hypo- and hypermethylation at conserved tissue-specific CpG island shores. Nat. Genet. *41*, 178–186.

Iyer, P., Zekri, A.-R., Hung, C.-W., Schiefelbein, E., Ismail, K., Hablas, A., Seifeldin, I.A., and Soliman, A.S. (2010). Concordance of DNA methylation pattern in plasma and tumor DNA of Egyptian hepatocellular carcinoma patients. Exp. Mol. Pathol. *88*, 107–111.

Jain, A.K., Murty, M.N., and Flynn, P.J. (1999). Data Clustering: A Review. ACM Comput Surv *31*, 264–323.

Jutten, C., and Herault, J. (1991). Blind separation of sources, part I: An adaptive algorithm based on neuromimetic architecture. Signal Process. *24*, 1–10.

Jylhävä, J., Pedersen, N.L., and Hägg, S. (2017). Biological Age Predictors. EBioMedicine *21*, 29–36.

Kanayama, K., Chiba, T., Oshima, M., Kanzaki, H., Koide, S., Saraya, A., Miyagi, S., Mimura, N., Kusakabe, Y., Saito, T., et al. (2019). Genome-Wide Mapping of Bivalent Histone Modifications in Hepatic Stem/Progenitor Cells. Stem Cells Int. *2019*, 9789240.

Kaposi-Novak, P., Lee, J.-S., Gòmez-Quiroz, L., Coulouarn, C., Factor, V.M., and Thorgeirsson, S.S. (2006). Met-regulated expression signature defines a subset of human hepatocellular carcinomas with poor prognosis and aggressive phenotype. J. Clin. Invest. *116*, 1582–1595.

Kelly, A.D., and Issa, J.-P.J. (2017). The promise of epigenetic therapy: reprogramming the cancer epigenome. Curr. Opin. Genet. Dev. *42*, 68–77.

Khavari, D.A., Sen, G.L., and Rinn, J.L. (2010). DNA methylation and epigenetic control of cellular differentiation. Cell Cycle *9*, 3880–3883.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. *14*, R36.

Klein Hesselink, E.N., Zafon, C., Villalmanzo, N., Iglesias, C., van Hemel, B.M., Klein Hesselink, M.S., Montero-Conde, C., Buj, R., Mauricio, D., Peinado, M.A., et al. (2018). Increased Global DNA Hypomethylation in Distant Metastatic and Dedifferentiated Thyroid Cancer. J. Clin. Endocrinol. Metab. *103*, 397–406.

Klose, R.J., and Bird, A.P. (2006). Genomic DNA methylation: the mark and its mediators. Trends Biochem. Sci. *31*, 89–97.

Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., Ziller, M.J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317–330.

Ladd-Acosta, C., and Fallin, M.D. (2015). The role of epigenetics in genetic and environmental epidemiology. Epigenomics *8*, 271–283.

Larsen, F., Gundersen, G., Lopez, R., and Prydz, H. (1992). CpG islands as gene markers in the human genome. Genomics *13*, 1095–1107.

Lee, S.-I., and Batzoglou, S. (2003). Application of independent component analysis to microarrays. Genome Biol. *4*, R76.

Lee, T.-W., Girolami, M., Bell, A.J., and Sejnowski, T.J. (2000). A unifying information-theoretic framework for independent component analysis. Comput. Math. Appl. *39*, 1–21.

Lengauer, C., Kinzler, K.W., and Vogelstein, B. (1998). Genetic instabilities in human cancers. Nature *396*, 643–649.

Letouzé, E., Martinelli, C., Loriot, C., Burnichon, N., Abermil, N., Ottolenghi, C., Janin, M., Menara, M., Nguyen, A.T., Benit, P., et al. (2013). SDH mutations establish a hypermethylator phenotype in paraganglioma. Cancer Cell *23*, 739–752.

Letouzé, E., Shinde, J., Renault, V., Couchy, G., Blanc, J.-F., Tubacher, E., Bayard, Q., Bacq, D., Meyer, V., Semhoun, J., et al. (2017). Mutational signatures reveal the dynamic interplay of risk factors and cellular processes during liver tumorigenesis. Nat. Commun. *8*.

Levy, I., and Sherman, M. (2002). Staging of hepatocellular carcinoma: assessment of the CLIP, Okuda, and Child-Pugh staging systems in a cohort of 257 patients in Toronto. Gut *50*, 881–885.

Li, B., Carey, M., and Workman, J.L. (2007). The Role of Chromatin during Transcription. Cell *128*, 707–719.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P., and Mesirov, J.P. (2011). Molecular signatures database (MSigDB) 3.0. Bioinformatics *27*, 1739–1740.

Liberzon, A., Birger, C., Thorvaldsdóttir, H., Ghandi, M., Mesirov, J.P., and Tamayo, P. (2015). The Molecular Signatures Database Hallmark Gene Set Collection. Cell Syst. *1*, 417–425.

Liebermeister, W. (2002). Linear modes of gene expression determined by independent component analysis. Bioinformatics *18*, 51–60.

Lim, S.-O., Gu, J.-M., Kim, M.S., Kim, H.-S., Park, Y.N., Park, C.K., Cho, J.W., Park, Y.M., and Jung, G. (2008). Epigenetic changes induced by reactive oxygen species in hepatocellular carcinoma: methylation of the E-cadherin promoter. Gastroenterology *135*, 2128–2140, 2140.e1-8.

Lister, R., Pelizzola, M., Dowen, R.H., Hawkins, R.D., Hon, G., Tonti-Filippini, J., Nery, J.R., Lee, L., Ye, Z., Ngo, Q.-M., et al. (2009). Human DNA methylomes at base resolution show widespread epigenomic differences. Nature *462*, 315–322.

Liu, J., Jiang, J., Mo, J., Liu, D., Cao, D., Wang, H., He, Y., and Wang, H. (2019). Global DNA 5-Hydroxymethylcytosine and 5-Formylcytosine Contents Are Decreased in the Early Stage of Hepatocellular Carcinoma. Hepatology *69*, 196–208.

Llovet, J.M., and Hernandez-Gea, V. (2014). Hepatocellular carcinoma: reasons for phase III failure and novel perspectives on trial design. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *20*, 2072–2079.

Llovet, J.M., Ricci, S., Mazzaferro, V., Hilgard, P., Gane, E., Blanc, J.-F., de Oliveira, A.C., Santoro, A., Raoul, J.-L., Forner, A., et al. (2008). Sorafenib in advanced hepatocellular carcinoma. N. Engl. J. Med. *359*, 378–390.

Locarnini, S., and Zoulim, F. (2010). Molecular genetics of HBV infection. Antivir. Ther. *15 Suppl 3*, 3–14.

Lok, A.S., Seeff, L.B., Morgan, T.R., Di Bisceglie, A.M., Sterling, R.K., Curto, T.M., Everson, G.T., Lindsay, K.L., Lee, W.M., Bonkovsky, H.L., et al. (2009). INCIDENCE OF HEPATOCELLULAR CARCINOMA AND ASSOCIATED RISK FACTORS IN HEPATITIS C-RELATED ADVANCED LIVER DISEASE. Gastroenterology *136*, 138–148.

Lord, C.J., and Ashworth, A. (2017). PARP inhibitors: Synthetic lethality in the clinic. Science *355*, 1152–1158.

Love, M.I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. *15*, 550.

Luger, K., Mäder, A.W., Richmond, R.K., Sargent, D.F., and Richmond, T.J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. Nature *389*, 251–260.

Ma, S., and Dai, Y. (2011). Principal component analysis based methods in bioinformatics studies. Brief. Bioinform. *12*, 714–722.

Mah, W.-C., and Lee, C.G. (2014). DNA methylation: potential biomarker in Hepatocellular Carcinoma. Biomark. Res. *2*, 5.

Mah, W.-C., Thurnherr, T., Chow, P.K.H., Chung, A.Y.F., Ooi, L.L.P.J., Toh, H.C., Teh, B.T., Sauntharajah, Y., and Lee, C.G.L. (2014). Methylation Profiles Reveal Distinct Subgroup of Hepatocellular Carcinoma Patients with Poor Prognosis. PLoS ONE *9*.

van Malenstein, H., Gevaert, O., Libbrecht, L., Daemen, A., Allemeersch, J., Nevens, F., Van Cutsem, E., Cassiman, D., De Moor, B., Verslype, C., et al. (2010). A seven-gene set associated with chronic hypoxia of prognostic importance in hepatocellular carcinoma. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *16*, 4278–4288.

Masliah-Planchon, J., Bièche, I., Guinebretière, J.-M., Bourdeaut, F., and Delattre, O. (2015). SWI/SNF Chromatin Remodeling and Human Malignancies. Annu. Rev. Pathol. Mech. Dis. *10*, 145–171.

McGlynn, K.A., Tsao, L., Hsing, A.W., Devesa, S.S., and Fraumeni, J.F. (2001). International trends and patterns of primary liver cancer. Int. J. Cancer *94*, 290–296.

Meigs, T.E., Fedor-Chaiken, M., Kaplan, D.D., Brackenbury, R., and Casey, P.J. (2002). Galpha12 and Galpha13 negatively regulate the adhesive functions of cadherin. J. Biol. Chem. *277*, 24594–24600.

Meissner, A., Mikkelsen, T.S., Gu, H., Wernig, M., Hanna, J., Sivachenko, A., Zhang, X., Bernstein, B.E., Nusbaum, C., Jaffe, D.B., et al. (2008). Genome-scale DNA methylation maps of pluripotent and differentiated cells. Nature *454*, 766–770.

Menghi, F., Inaki, K., Woo, X., Kumar, P.A., Grzeda, K.R., Malhotra, A., Yadav, V., Kim, H., Marquez, E.J., Ucar, D., et al. (2016). The tandem duplicator phenotype as a distinct genomic configuration in cancer. Proc. Natl. Acad. Sci. U. S. A. *113*, E2373-2382.

Monti, S., Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003). Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data.

Mudbhary, R., Hoshida, Y., Chernyavskaya, Y., Jacob, V., Villanueva, A., Fiel, M.I., Chen, X., Kojima, K., Thung, S., Bronson, R.T., et al. (2014). UHRF1 overexpression drives DNA hypomethylation and hepatocellular carcinoma. Cancer Cell *25*, 196–209.

Mussa, A., Russo, S., Larizza, L., Riccio, A., and Ferrero, G.B. (2016). (Epi)genotype–phenotype correlations in Beckwith–Wiedemann syndrome: a paradigm for genomic medicine. Clin. Genet. *89*, 403–415.

Nakamura, M., Chiba, T., Kanayama, K., Kanzaki, H., Saito, T., Kusakabe, Y., and Kato, N. (2019). Epigenetic dysregulation in hepatocellular carcinoma: an up-to-date review. Hepatol. Res. *49*, 3–13.

Nan, X., Ng, H.H., Johnson, C.A., Laherty, C.D., Turner, B.M., Eisenman, R.N., and Bird, A. (1998). Transcriptional repression by the methyl-CpG-binding protein MeCP2 involves a histone deacetylase complex. Nature *393*, 386–389.

Nault, J.-C., and Zucman-Rossi, J. (2016). TERT promoter mutations in primary liver tumors. Clin. Res. Hepatol. Gastroenterol. *40*, 9–14.

Nault, J.-C., Reyniès, A.D., Villanueva, A., Calderaro, J., Rebouissou, S., Couchy, G., Decaens, T., Franco, D., Imbeaud, S., Rousseau, F., et al. (2013). A Hepatocellular Carcinoma 5-Gene Score Associated With Survival of Patients After Liver Resection. Gastroenterology *145*, 176–187.

Nault, J.-C., Datta, S., Imbeaud, S., Franconi, A., Mallet, M., Couchy, G., Letouzé, E., Pilati, C., Verret, B., Blanc, J.-F., et al. (2015). Recurrent AAV2-related insertional mutagenesis in human hepatocellular carcinomas. Nat. Genet. *47*, 1187–1193.

Nault, J.C., Couchy, G., Balabaud, C., Morcrette, G., Caruso, S., Blanc, J.-F., Bacq, Y., Caldéraro, J., Paradis, V., Ramos, J., et al. (2017). Recurrent INHBE-GLI1 fusions in hepatocellular adenomas with sonic hedgehog pathway activation. J. Hepatol. *66*, S14.

Neri, F., Rapelli, S., Krepelova, A., Incarnato, D., Parlato, C., Basile, G., Maldotti, M., Anselmi, F., and Oliviero, S. (2017). Intragenic DNA methylation prevents spurious transcription initiation. Nature *543*, 72–77.

Neuveut, C., Wei, Y., and Buendia, M.A. (2010). Mechanisms of HBV-related hepatocarcinogenesis. J. Hepatol. *52*, 594–604.

Ng, A.W.T., Poon, S.L., Huang, M.N., Lim, J.Q., Boot, A., Yu, W., Suzuki, Y., Thangaraju, S., Ng, C.C.Y., Tan, P., et al. (2017). Aristolochic acids and their derivatives are widely implicated in liver cancers in Taiwan and throughout Asia. Sci. Transl. Med. *9*.

Nicoglou, A., and Merlin, F. (2017). Epigenetics: A way to bridge the gap between biological fields. Stud. Hist. Philos. Sci. Part C Stud. Hist. Philos. Biol. Biomed. Sci. *66*, 73–82.

Oba, A., Shimada, S., Akiyama, Y., Nishikawaji, T., Mogushi, K., Ito, H., Matsumura, S., Aihara, A., Mitsunori, Y., Ban, D., et al. (2017). ARID2 modulates DNA damage response in human hepatocellular carcinoma cells. J. Hepatol. *66*, 942–951.

O'Connor, M.J. (2015). Targeting the DNA Damage Response in Cancer. Mol. Cell *60*, 547–560.

Ohm, J.E., and Baylin, S.B. (2007). Stem cell chromatin patterns -- an instructive mechanism for DNA hypermethylation? Cell Cycle Georget. Tex *6*, 1040–1043.

Oja, H., and Nordhausen, K. (2006). Independent Component Analysis. In Encyclopedia of Environmetrics, (John Wiley & Sons, Ltd), p.

Okano, M., Bell, D.W., Haber, D.A., and Li, E. (1999). DNA Methyltransferases Dnmt3a and Dnmt3b Are Essential for De Novo Methylation and Mammalian Development. Cell *99*, 247–257.

Paroush, Z., Keshet, I., Yisraeli, J., and Cedar, H. (1990). Dynamics of demethylation and activation of the α-actin gene in myoblasts. Cell *63*, 1229–1237.

Paterlini-Bréchot, P., Saigo, K., Murakami, Y., Chami, M., Gozuacik, D., Mugnier, C., Lagorce, D., and Bréchot, C. (2003). Hepatitis B virus-related insertional mutagenesis occurs frequently in human liver cancers and recurrently targets human telomerase gene. Oncogene *22*, 3911–3916.

Pellestor, F., Gatinois, V., Puechberty, J., Geneviève, D., and Lefort, G. (2014). Le chromothripsis - Une forme insoupçonnée de complexification extrême des remaniements chromosomiques. médecine/sciences *30*, 266–273.

Pérez, R.F., Tejedor, J.R., Bayón, G.F., Fernández, A.F., and Fraga, M.F. (2018). Distinct chromatin signatures of DNA hypomethylation in aging and cancer. Aging Cell *17*, e12744.

Peterson, C.L., and Laniel, M.-A. (2004). Histones and histone modifications. Curr. Biol. *14*, R546–R551.

Plentz, R.R., Park, Y.N., Lechel, A., Kim, H., Nellessen, F., Langkopf, B.H.E., Wilkens, L., Destro, A., Fiamengo, B., Manns, M.P., et al. (2007). Telomere shortening and inactivation of cell cycle checkpoints characterize human hepatocarcinogenesis. Hepatol. Baltim. Md *45*, 968–976.

Pollard, P.J., Spencer-Dene, B., Shukla, D., Howarth, K., Nye, E., El-Bahrawy, M., Deheragoda, M., Joannou, M., McDonald, S., Martin, A., et al. (2007). Targeted Inactivation of Fh1 Causes Proliferative Renal Cyst Development and Activation of the Hypoxia Pathway. Cancer Cell *11*, 311–319.

Pon, J.R., and Marra, M.A. (2015). Driver and passenger mutations in cancer. Annu. Rev. Pathol. *10*, 25–50.

Portela, A., and Esteller, M. (2010). Epigenetic modifications and human disease. Nat. Biotechnol. *28*, 1057–1068.

Poynard, T., Bedossa, P., and Opolon, P. (1997). Natural history of liver fibrosis progression in patients with chronic hepatitis C. The Lancet *349*, 825–832.

Prokhortchouk, E., and Defossez, P.-A. (2008). The cell biology of DNA methylation in mammals. Biochim. Biophys. Acta BBA - Mol. Cell Res. *1783*, 2167–2173.

Qu, Y.-L., Deng, C.-H., Luo, Q., Shang, X.-Y., Wu, J.-X., Shi, Y., Wang, L., and Han, Z.-G. (2019). Arid1a regulates insulin sensitivity and lipid metabolism. EBioMedicine *42*, 481–493.

Rakyan, V.K., Down, T.A., Maslau, S., Andrew, T., Yang, T.-P., Beyan, H., Whittaker, P., McCann, O.T., Finer, S., Valdes, A.M., et al. (2010). Human aging-associated DNA hypermethylation occurs preferentially at bivalent chromatin domains. Genome Res. *20*, 434–439.

Rasmussen, K.D., and Helin, K. (2016). Role of TET enzymes in DNA methylation, development, and cancer. Genes Dev. *30*, 733–750.

Rawla, P., Sunkara, T., Muralidharan, P., and Raj, J.P. (2018). Update in global trends and aetiology of hepatocellular carcinoma. Contemp. Oncol. *22*, 141–150.

Rebouissou Sandra, Franconi Andrea, Calderaro Julien, Letouzé Eric, Imbeaud Sandrine, Pilati Camilla, Nault Jean-Charles, Couchy Gabrielle, Laurent Alexis, Balabaud Charles, et al. (2016). Genotype-phenotype correlation of CTNNB1 mutations reveals different ß-catenin activity associated with liver tumor progression. Hepatology *64*, 2047–2061.

Renfree, M.B., Suzuki, S., and Kaneko-Ishino, T. (2013). The origin and evolution of genomic imprinting and viviparity in mammals. Philos. Trans. R. Soc. B Biol. Sci. *368*.

Roadmap Epigenomics Consortium, Kundaje, A., Meuleman, W., Ernst, J., Bilenky, M., Yen, A., Heravi-Moussavi, A., Kheradpour, P., Zhang, Z., Wang, J., et al. (2015). Integrative analysis of 111 reference human epigenomes. Nature *518*, 317.

Rosa, I., Denis, J., Renard, P., Lesgourgues, B., Dobrin, A.S., Becker, C., Faroux, R., Causse, X., Diaz, E., Le Dréau, G., et al. (2010). 585 A FRENCH MULTICENTRIC LONGITUDINAL DESCRIPTIVE STUDY OF HEPATOCELLULAR CARCINOMA

MANAGEMENT (THE CHANGH COHORT): PRELIMINARY RESULTS. J. Hepatol. *52*, S231–S232.

Salhab, A., Nordström, K., Gasparoni, G., Kattler, K., Ebert, P., Ramirez, F., Arrigoni, L., Müller, F., Polansky, J.K., Cadenas, C., et al. (2018). A comprehensive analysis of 195 DNA methylomes reveals shared and cell-specific features of partially methylated domains. Genome Biol. *19*.

Sasai, N., and Defossez, P.-A. (2009). Many paths to one goal? The proteins that recognize methylated DNA in eukaryotes. Int. J. Dev. Biol. *53*, 323–334.

Satoh, S., Daigo, Y., Furukawa, Y., Kato, T., Miwa, N., Nishiwaki, T., Kawasoe, T., Ishiguro, H., Fujita, M., Tokino, T., et al. (2000). AXIN1 mutations in hepatocellular carcinomas, and growth suppression in cancer cells by virus-mediated transfer of AXIN1. Nat. Genet. *24*, 245–250.

Saxonov, S., Berg, P., and Brutlag, D.L. (2006). A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters. Proc. Natl. Acad. Sci. *103*, 1412–1417.

Schaefer, M., Pollex, T., Hanna, K., Tuorto, F., Meusburger, M., Helm, M., and Lyko, F. (2010). RNA methylation by Dnmt2 protects transfer RNAs against stress-induced cleavage. Genes Dev. *24*, 1590–1595.

Schulze, K., Imbeaud, S., Letouzé, E., Alexandrov, L.B., Calderaro, J., Rebouissou, S., Couchy, G., Meiller, C., Shinde, J., Soysouvanh, F., et al. (2015). Exome sequencing of hepatocellular carcinomas identifies new mutational signatures and potential therapeutic targets. Nat. Genet. *47*, 505–511.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

Sharma, S., Kelly, T.K., and Jones, P.A. (2010). Epigenetics in cancer. Carcinogenesis *31*, 27–36.

Shen, J., Wang, S., Zhang, Y.-J., Kappil, M., Wu, H.-C., Kibriya, M.G., Wang, Q., Jasmine, F., Ahsan, H., Lee, P.-H., et al. (2012). Genome-wide DNA Methylation Profiles in Hepatocellular Carcinoma. Hepatol. Baltim. Md *55*, 1799–1808.

Shen, J., Song, R., Gong, Y., and Zhao, H. (2017). Global DNA hypomethylation in leukocytes associated with glioma risk. Oncotarget *8*, 63223–63231.

Shin, S.H., Kim, B., Jang, J.-J., Suh, K.S., and Kang, G.H. (2010). Identification of novel methylation markers in hepatocellular carcinoma using a methylation array. J. Korean Med. Sci. *25*, 1152–1159.

Sia, D., Villanueva, A., Friedman, S.L., and Llovet, J.M. (2017). Liver Cancer Cell of Origin, Molecular Class, and Effects on Patient Prognosis. Gastroenterology *152*, 745–761.

Silva, T.C., Coetzee, S.G., Gull, N., Yao, L., Hazelett, D.J., Noushmehr, H., Lin, D.-C., and Berman, B.P. (2019). ELMER v.2: an R/Bioconductor package to reconstruct gene regulatory networks from DNA methylation and transcriptome profiles. Bioinformatics *35*, 1974–1977.

Søes, S., Daugaard, I.L., Sørensen, B.S., Carus, A., Mattheisen, M., Alsner, J., Overgaard, J., Hager, H., Hansen, L.L., and Kristensen, L.S. (2014). Hypomethylation and increased expression of the putative oncogene ELMO3 are associated with lung cancer development and metastases formation. Oncoscience *1*, 367–374.

Song, M.-A., Tiirikainen, M., Kwee, S., Okimoto, G., Yu, H., and Wong, L.L. (2013). Elucidating the Landscape of Aberrant DNA Methylation in Hepatocellular Carcinoma. PLoS ONE *8*.

Sporn, M.B., and Liby, K.T. (2012). NRF2 and cancer: the good, the bad and the importance of context. Nat. Rev. Cancer *12*, 564–571.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., Nimwegen, E. van, Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Stein, R., Gruenbaum, Y., Pollack, Y., Razin, A., and Cedar, H. (1982). Clonal inheritance of the pattern of DNA methylation in mouse cells. Proc. Natl. Acad. Sci. U. S. A. *79*, 61–65.

Stratton, M.R., Campbell, P.J., and Futreal, P.A. (2009). The cancer genome. Nature *458*, 719–724.

Struyf, A., Hubert, M., and Rousseeuw, P. (1997). Clustering in an Object-Oriented Environment. J. Stat. Softw. *1*, 1–30.

Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al. (2005). Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. *102*, 15545–15550.

Sun, X., Wang, S.C., Wei, Y., Luo, X., Jia, Y., Li, L., Gopal, P., Zhu, M., Nassour, I., Chuang, J.-C., et al. (2017). Arid1a has context-dependent oncogenic and tumor suppressor functions in liver cancer. Cancer Cell *32*, 574-589.e6.

Szekely, G.J., and Rizzo, M.L. (2005). Hierarchical Clustering via Joint Between-Within Distances: Extending Ward's Minimum Variance Method. J. Classif. *22*, 151–183.

Teissandier, A., and Bourc'his, D. (2017). Gene body DNA methylation conspires with H3K36me3 to preclude aberrant transcription. EMBO J. *36*, 1471–1473.

Thomson, J.P., Ottaviano, R., Unterberger, E.B., Lempia inen, H., Muller, A., Terranova, R., Illingworth, R.S., Webb, S., Kerr, A.R.W., Lyall, M.J., et al. (2016). Loss of Tet1-Associated 5-Hydroxymethylcytosine Is Concomitant with Aberrant Promoter Hypermethylation in Liver Cancer. Cancer Res. *76*, 3097–3108.

Toh, T.B., Lim, J.J., and Chow, E.K.-H. (2019). Epigenetics of hepatocellular carcinoma. Clin. Transl. Med. *8*.

Tokoro, T., Kato, Y., Sugioka, A., and Mizoguchi, Y. (2014). Malignant transformation of hepatocellular adenoma over a decade. BMJ Case Rep. *2014*.

Tomasetti, C., and Vogelstein, B. (2015). Variation in cancer risk among tissues can be explained by the number of stem cell divisions. Science *347*, 78–81.

Totoki, Y., Tatsuno, K., Covington, K.R., Ueda, H., Creighton, C.J., Kato, M., Tsuji, S., Donehower, L.A., Slagle, B.L., Nakamura, H., et al. (2014). Trans-ancestry mutational landscape of hepatocellular carcinoma genomes. Nat. Genet. *46*, 1267–1273.

Toyota, M., Ahuja, N., Ohe-Toyota, M., Herman, J.G., Baylin, S.B., and Issa, J.-P.J. (1999). CpG island methylator phenotype in colorectal cancer. Proc. Natl. Acad. Sci. *96*, 8681–8686.

Trépo, C., Chan, H.L.Y., and Lok, A. (2014). Hepatitis B virus infection. Lancet Lond. Engl. *384*, 2053–2063.

Tseng, Y.-Y., Moriarity, B.S., Gong, W., Akiyama, R., Tiwari, A., Kawakami, H., Ronning, P., Reuland, B., Guenther, K., Beadnell, T.C., et al. (2014). PVT1 dependence in cancer with MYC copy-number increase. Nature *512*, 82–86.

Turcan, S., Rohle, D., Goenka, A., Walsh, L.A., Fang, F., Yilmaz, E., Campos, C., Fabius, A.W.M., Lu, C., Ward, P.S., et al. (2012). IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. Nature *483*, 479–483.

Unnikrishnan, A., Freeman, W.M., Jackson, J., Wren, J.D., Porter, H., and Richardson, A. (2019). The role of DNA methylation in epigenetics of aging. Pharmacol. Ther. *195*, 172–185.

Van der Auwera, I., Bovie, C., Svensson, C., Limame, R., Trinh, X.B., van Dam, P., Van Laere, S.J., Van Marck, E., Vermeulen, P.B., and Dirix, L.Y. (2009). Quantitative assessment of DNA hypermethylation in the inflammatory and non-inflammatory breast cancer phenotypes. Cancer Biol. Ther. *8*, 2252–2259.

Villanueva, A., Portela, A., Sayols, S., Battiston, C., Hoshida, Y., Méndez-González, J., Imbeaud, S., Letouzé, E., Hernandez-Gea, V., Cornella, H., et al. (2015). DNA methylation-based prognosis and epidrivers in hepatocellular carcinoma: HEPATOLOGY, Vol. XX, No. X, 2015. Hepatology *61*, 1945–1956.

Wagner, E.J., and Carpenter, P.B. (2012). Understanding the language of Lys36 methylation at histone H3. Nat. Rev. Mol. Cell Biol. *13*, 115–126.

Wang, D., and Gu, J. (2016). Integrative clustering methods of multi-omics data for molecule-based cancer classifications. Quant. Biol. *4*, 58–67.

Weber, M., Hellmann, I., Stadler, M.B., Ramos, L., Pääbo, S., Rebhan, M., and Schübeler, D. (2007). Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. Nat. Genet. *39*, 457–466.

Werner, R.J., Kelly, A.D., and Issa, J.-P.J. (2017). Epigenetics and Precision Oncology. Cancer J. Sudbury Mass *23*, 262–269.

Wheeler, D.A., and Roberts, L.R. (2017). Comprehensive and Integrative Genomic Characterization of Hepatocellular Carcinoma. Cell *169*, 1327-1341.e23.

Wigler, M., Levy, D., and Perucho, M. (1981). The somatic replication of DNA methylation. Cell *24*, 33–40.

Wilkinson, C.R.M., Bartlett, R., Nurse, P., and Bird, A.P. (1995). The fission yeast gene pmt1+ encodes a DNA methyltransferase homologue. Nucleic Acids Res. *23*, 203–210.

Wilson, V.L., Smith, R.A., Ma, S., and Cutler, R.G. (1987). Genomic 5-methyldeoxycytidine decreases with age. J. Biol. Chem. *262*, 9948–9951.

Witte, T., Plass, C., and Gerhauser, C. (2014). Pan-cancer patterns of DNA methylation. Genome Med. *6*.

Wolf, S.F., Jolly, D.J., Lunnen, K.D., Friedmann, T., and Migeon, B.R. (1984). Methylation of the hypoxanthine phosphoribosyltransferase locus on the human X chromosome: implications for X-chromosome inactivation. Proc. Natl. Acad. Sci. U. S. A. *81*, 2806–2810.

Wong, K.K., Lawrie, C.H., and Green, T.M. (2019). Oncogenic Roles and Inhibitors of DNMT1, DNMT3A, and DNMT3B in Acute Myeloid Leukaemia. Biomark. Insights *14*.

Yamada, N., Yasui, K., Dohi, O., Gen, Y., Tomie, A., Kitaichi, T., Iwai, N., Mitsuyoshi, H., Sumida, Y., Moriguchi, M., et al. (2016). Genome-wide DNA methylation analysis in hepatocellular carcinoma. Oncol. Rep. *35*, 2228–2236.

Yamashita, T., Forgues, M., Wang, W., Kim, J.W., Ye, Q., Jia, H., Budhu, A., Zanetti, K.A., Chen, Y., Qin, L.-X., et al. (2008). EpCAM and alpha-fetoprotein expression defines novel prognostic subtypes of hepatocellular carcinoma. Cancer Res. *68*, 1451–1461.

Yang, J.D., and Roberts, L.R. (2010). Hepatocellular carcinoma: a global view. Nat. Rev. Gastroenterol. Hepatol. *7*, 448–458.

Yang, M., and Pollard, P.J. (2013). Succinate: A New Epigenetic Hacker. Cancer Cell *23*, 709–711.

Yang, B., Guo, M., Herman, J.G., and Clark, D.P. (2003). Aberrant Promoter Methylation Profiles of Tumor Suppressor Genes in Hepatocellular Carcinoma. Am. J. Pathol. *163*, 1101–1107.

Yang, X., Han, H., De Carvalho, D.D., Lay, F.D., Jones, P.A., and Liang, G. (2014). Gene body methylation can alter gene expression and is a therapeutic target in cancer. Cancer Cell *26*, 577–590.

Yang, Z., Wong, A., Kuh, D., Paul, D.S., Rakyan, V.K., Leslie, R.D., Zheng, S.C., Widschwendter, M., Beck, S., and Teschendorff, A.E. (2016). Correlation of an epigenetic mitotic clock with cancer risk. Genome Biol. *17*, 205.

Yao, L., Shen, H., Laird, P.W., Farnham, P.J., and Berman, B.P. (2015). Inferring regulatory element landscapes and transcription factor networks from cancer methylomes. Genome Biol. *16*.

Zelic, R., Fiano, V., Grasso, C., Zugna, D., Pettersson, A., Gillio-Tos, A., Merletti, F., and Richiardi, L. (2015). Global DNA hypomethylation in prostate cancer development and progression: a systematic review. Prostate Cancer Prostatic Dis. *18*, 1–12.

Zhang, C., Li, Z., Cheng, Y., Jia, F., Li, R., Wu, M., Li, K., and Wei, L. (2007). CpG Island Methylator Phenotype Association with Elevated Serum α-Fetoprotein Level in Hepatocellular Carcinoma. Clin. Cancer Res. *13*, 944–952.

Zhang, C., Li, J., Huang, T., Duan, S., Dai, D., Jiang, D., Sui, X., Li, D., Chen, Y., Ding, F., et al. (2016). Meta-analysis of DNA methylation biomarkers in hepatocellular carcinoma. Oncotarget *7*, 81255–81267.

Zhang, K., Yan, J., Yi, B., Rui, Y., and Hu, H. (2019). High KCNQ1OT1 expression might independently predict shorter survival of colon adenocarcinoma. Future Oncol. *15*, 1085–1095.

Zhao, L.-H., Liu, X., Yan, H.-X., Li, W.-Y., Zeng, X., Yang, Y., Zhao, J., Liu, S.-P., Zhuang, X.-H., Lin, C., et al. (2016). Genomic and oncogenic preference of HBV integration in hepatocellular carcinoma. Nat. Commun. *7*, 12992.

Zheng, Y., Huang, Q., Ding, Z., Liu, T., Xue, C., Sang, X., and Gu, J. (2018). Genome-wide DNA methylation analysis identifies candidate epigenetic markers and drivers of hepatocellular carcinoma. Brief. Bioinform. *19*, 101–108.

Zhou, W., Dinh, H.Q., Ramjan, Z., Weisenberger, D.J., Nicolet, C.M., Shen, H., Laird, P.W., and Berman, B.P. (2018). DNA methylation loss in late-replicating domains is linked to mitotic cell division. Nat. Genet. *50*, 591–602.

Zhou, Y., Gan, F., Hou, L., Zhou, X., Adam Ibrahim, Y.A., and Huang, K. (2017). Modulations of DNMT1 and HDAC1 are involved in the OTA-induced cytotoxicity and apoptosis in vitro. Chem. Biol. Interact. *278*, 170–178.

Zhu, A.X., Kang, Y.-K., Yen, C.-J., Finn, R.S., Galle, P.R., Llovet, J.M., Assenat, E., Brandi, G., Lim, H.Y., Pracht, M., et al. (2018). REACH-2: A randomized, double-blind, placebo-controlled phase 3 study of ramucirumab versus placebo as second-line treatment in patients with advanced hepatocellular carcinoma (HCC) and elevated baseline alpha-fetoprotein (AFP) following first-line sorafenib. J. Clin. Oncol. *36*, 4003–4003.

Zilberman, D., Gehring, M., Tran, R.K., Ballinger, T., and Henikoff, S. (2007). Genome-wide analysis of Arabidopsis thaliana DNA methylation uncovers an interdependence between methylation and transcription. Nat. Genet. *39*, 61–69.

Zinovyev, A., Kairov, U., Karpenyuk, T., and Ramanculov, E. (2013). Blind source separation methods for deconvolution of complex signals in cancer biology. Biochem. Biophys. Res. Commun. *430*, 1182–1187.

Zucman-Rossi, J., Villanueva, A., Nault, J.-C., and Llovet, J.M. (2015). Genetic Landscape and Biomarkers of Hepatocellular Carcinoma. Gastroenterology *149*, 1226-1239.e4.

Liste des figures

Liste des tables