Classification et caractérisation des cancers colorectaux par approches omiques

Laetitia Marisa

▶ **To cite this version:**

Laetitia Marisa. Classification et caractérisation des cancers colorectaux par approches omiques. Cancer. Université Pierre et Marie Curie - Paris VI, 2015. Français. NNT : 2015PA066235 . tel-01230962

HAL Id: tel-01230962
https://theses.hal.science/tel-01230962

Submitted on 19 Nov 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

École Doctorale de Physiologie et Physiopathologie ED 394

Thèse de doctorat

Discipline : Génomique, Cancérologie, Médecine, Santé

présentée par

Laetitia Marisa

Classification et Caractérisation des Cancers Colorectaux par Approches Omiques

---

dirigée par Alex Duval et co-encadrée par Aurélien de Reyniès

Soutenue le 13 octobre 2015 devant le jury composé de :

M. Florent Soubrier	Hôpital Pitié-Salpêtrière	Président
M^{me} Béatrice Romagnolo	Institut Cochin	Rapporteur
M. François Radvanyi	Institut Curie	Rapporteur
M. Pierre Laurent-Puig	Université Paris Descartes	Examinateur
M. Charles Ferté	Institut Gustave Roussy	Examinateur
M. Alex Duval	Centre de Recherche Saint-Antoine	Directeur
M. Aurélien de Reyniès	Ligue Nationale Contre le Cancer	Co-Encadrant

Ligue Nationale Contre le Cancer Programme Carte d'Identité des Tumeurs
14 rue Corvisart
75 013 Paris

INSERM - Centre de Recherche Hôpital Saint-Antoine
Équipe Instabilité des Microsatellites et Cancer UMRS 938
34 rue Crozatier
75012 Paris

Université Pierre et Marie Curie
École Doctorale Physiologie Physiopathologie et Thérapeutique ED 394
Centre de Recherche des Cordeliers
15 rue de l'École de Médecine
75006 Paris

A mes deux petites têtes, Lula et Yarol, mes 2 plus belles expériences de biologie.

"John Constable said that we see nothing until we truly understand, but we can also say that we understand nothing until we truly see.", Pr Jeremy Jass, pathologiste gastro-intestinal (1951-2008)

Remerciements

Je tiens en tout premier lieu à remercier infiniment le professeur Jacqueline Godet, directrice du programme CIT et présidente de la Ligue Nationale Contre le Cancer, d'avoir accepté que je réalise cette thèse dans le cadre de mon travail. Je vous suis très reconnaissante de votre confiance et de votre soutien tout au long de ces années.

Je remercie tout particulièrement Alex Duval pour avoir accepté d'encadrer cette thèse. Bien que le manuscrit ne ressemble pas à ce que l'on aurait souhaité tous les deux, la présentation d'une sous-classification multiomique des tumeurs colorectales avec instabilité microsatellitaire, j'ai beaucoup appris lors de nos travaux ensemble, j'ai appliqué tout ce que j'ai appris avec toi pour réaliser ce travail. Je te remercie de toujours prendre en compte mon avis et de m'impliquer dans tes projets toujours très motivants.

Je remercie aussi tout particulièrement Aurélien de Reyniès pour co-encadrer ce travail, pour être de bon conseil dans mes analyses, et pour arriver à simplifier les choses quand le stress et les boucles infinies qui se créent dans mon cerveau me submergent.

Je remercie également très sincèrement Pierre Laurent-Puig et Valérie Boige sans qui le travail présenté dans ce manuscrit n'aurait pas abouti. J'ai beaucoup apprécié de travailler avec vous deux sur ce projet. En dehors de votre grande sympathie, de travailler avec vous m'a permis de donner une dimension beaucoup plus clinique, plus proche du patient, à mon travail.

Je remercie également tous les porteurs des projets CIT côlon avec qui j'ai collaboré pour ce travail et aux co-signataires de l'article pour tout le travail qui a été réalisé pour aboutir à ces résultats.

Je tiens également à exprimer mes sentiments les plus sincères aux autres personnes de l'équipe CIT actuelle qui mettent une ambiance tip top dans notre bocal : Mira Ayadi, qui organise tout ça à merveille, Nabila Elarouci, qui nous facilite la vie grandement en nous préparant les annotations aux petits oignions, Fabien Petel qui œuvre pour que l'infrastructure informatique réponde à nos exigences, Sylvie Job pour avoir aboli les déjeuners devant l'écran d'ordinateur, Aurélie Kamoun pour nous abreuver de sucreries et de petits canards, Noémie Robil pour sa constante bonne humeur, et les 2 petits nouveaux Rémy Nicolle et Yuna Blum qui ont déjà apporté beaucoup de sang neuf dans l'équipe.

Et j'ai aussi une pensée particulière pour tous les anciens membres de CIT : Eric Letouzé, qui avec quelques phrases motivantes m'a permis d'avancer dans ce travail, Renaud Schiappa le Marseillais qui s'est arraché les cheveux pour intégrer les annotations de ce projet, à Julien Laffaire qui me manque dans l'équipe, à Anne-Sophie Valin sans qui je n'aurais même pas penser à postuler pour rentrer dans l'équipe CIT, à Mickaël Guedj

avec qui j'ai débuté à CIT et qui a toujours été très inspirant, Laure Vescovo avec qui j'ai débuté ce long travail de classification, Emilie Thomas qui avait toujours des supers astuces, et Jacqueline Métral, la pétillante, qui a orchestré le programme avec une poigne de fer jusqu'à l'année dernière.

Des remerciements aussi pour toute l'équipe d'Alex, et plus spécialement à Ada Collura, merci pour ta gentillesse et ton aide dans mes projets, à la super anatomopathologiste Magali Svrcek qui a eu la gentillesse de m'accueillir dans son laboratoire pour me faire découvrir l'envers du décor de ce que j'étudie sous forme de fichier tabulé, à Nizar Elmurr pour m'avoir fait plongé dans le monde des miRNA et aux autres thésardes de ma promotion, Anais et Sahra, à Olivier Buhart et à Kristelle Wanherdrick pour leur aide pour les échantillons et les annotations, et à Agathe Guilloux pour son aide en statistiques sur les études de HSP110.

Et enfin de tendres pensées pour mes proches :

à mes parents, ma soeur et ma belle-famille pour leur soutien et leur aide précieuse avec les enfants.

à Caro pour m'avoir motivé à combattre mes démons et à m'inscrire en thèse pour enfin arrêter d'en parler au futur antérieur.

à mon Tof, ma force tranquille, même si tu avais raison, mais t'as (presque) toujours raison en même temps..., merci de m'avoir aidée, motivée et soutenue pendant la phase de rédaction malgré tout. En plus d'être un papa Parfait, tu es un compagnon Parfait (enfin presque).

à mes petits cocos, mes boosters, vous avez été super mignons pendant toute cette période, même si votre maman était moins disponible.

Table des matières

I Introduction **9**

II Revue Bibliographique **13**

1 Le cancer colorectal en clinique : caractéristiques épidémiologiques, cliniques et anatomo-pathologiques **15**
1.1 Une problème de santé majeur 15
1.1.1 Incidence et mortalité 15
1.1.2 Les facteurs de risque, la part de la génétique et de l'environnement 16
1.2 Physiopathologie et anatomopathologie cancéreuse du côlon 18
1.2.1 Progression tumorale : la séquence adénome-carcinome 18
1.2.2 Les cellules d'origine 20
1.2.3 Les lésions précurseurs 24
1.2.4 Les différents types histologiques de carcinomes 24
1.3 La classification clinique 26
1.3.1 La classification TNM 26
1.3.2 Valeur pronostique 26
1.3.3 Implication sur le choix de traitement 27

2 Le cancer colorectal à l'échelle moléculaire : caractérisation et classification moléculaire **31**
2.1 Les mutations clés de la tumorigenèse colique 31
2.1.1 Le modèle d'oncogenèse de Fearon et Vogelstein 31
2.1.2 Les autres mutations clés pour la tumorigenèse 33
2.2 Les trois formes d'instabilité du génome tumoral observées 35
2.2.1 L'instabilité chromosomique, phénotype et mécanismes 35
2.2.2 L'instabilité des microsatellites 37
2.2.3 L'hyperméthylation de l'ADN 41
2.2.4 Évolution du modèle de tumorigenèse à la lumière des instabilités . 43
2.3 Les Classifications moléculaires proposées 44

3 Le cancer colorectal à l'ère de la génomique : l'apport des technologies à haut débit **47**
3.1 Étude des tumeurs par les technologies à haut débit 47
3.1.1 Émergence et évolution de la génomique 47
3.1.2 Principe des omiques 48
3.1.3 Les limites des omiques 53
3.1.4 L'analyse des données omiques appliquées au cancer 55
3.2 La recherche de marqueurs pronostiques dans les cancers colorectaux . . . 60

3.2.1 Aperçu de l'état de l'art 61
3.2.2 Focus sur Oncotype DX 63
3.2.3 Focus sur Coloprint 63
3.3 Caractérisation et classification des tumeurs colorectales par l'utilisation d'omiques 64
3.3.1 Caractérisations par signatures supervisées 64
3.3.2 Les classifications non supervisées de tumeurs colorectales basées sur les omiques 65

III Résultats **69**
Article PLoS Medicine : Mise en évidence de l'existence de sous-types moléculaires des cancers du côlon par l'établissement d'une classification robuste des données d'expression 71
Établissement d'une classification consensus des cancers colorectaux à partir de six systèmes de classification différents 105
L'approche développée 105
La classification consensus obtenue 106
Comparaison de la classification CIT et de la classification consensus 108

IV Discusssion et Conclusion Générale **109**

Appendix **115**

A Méthode de mesure des instabilités du génome **117**

B Étude des cellules souches dans les lignées cellulaires colorectales **119**

C Etude du pronostic inter et intra sous-type MSI **149**

Table des figures **211**

Bibliographie **213**

Résumé **229**

Première partie

Introduction

Le cancer colorectal (CCR) est le 3ème cancer le plus fréquent en France. Malgré les avancées en terme de dépistage, de diagnostic et de traitements, il constitue la 2ème cause de mortalité par cancer après celui du poumon. En pratique clinique, le pronostic et le traitement des patients sont basés uniquement, encore aujourd'hui, sur la classification histo-pathologique tumor-node-metastasis (TNM).Cette classification définit 4 stades en fonction du degré d'invasion de la paroi colique, avec un risque de décès de plus en plus accru avec l'augmentation du stade. Toutefois, cette classification ne permet pas de prédire précisément la rechute d'une grande partie des patients.

Sur le plan biologique, un cancer colorectal est une tumeur d'origine épithéliale provenant dans la très grande majorité des cas d'adénomes évoluant en carcinome. En raison du peu de variations observées au niveau anatomopathologique, le CCR a été considéré pendant longtemps comme une entité homogène d'adénocarcinomes. Mais il est maintenant bien établi que le CCR est une maladie hétérogène depuis la découverte dans les dernières décennies des différentes formes d'instabilité du génome que sont (i) l'instabilité chromosomique (CIN, Chromosomal instability, 70% des CCR), (ii) l'instabilité des microsatellites (MSI, Microsatellite instability, 15% des CCR) et (iii) le phénotype d'hyperméthylation d'îlots CpG (CIMP, CpG island methylator phenotype, 15% des CCR). Ces formes seraient associées à différentes séquences d'altérations lors de la carcinogenèse. Confortant l'idée d'hétérogénéité de la maladie, il est également établi que le pronostic en fonction de ces formes n'est pas le même et que la réponse de ces cancers aux traitements actuellement en vigueur est différente.

L'hétérogénéité moléculaire et clinique reste néanmoins encore peu et mal décrite dans les CCR, bien que l'identification de sous-types de CCR ayant des caractéristiques pronostiques et de réponse au traitement distinctes aurait un intérêt clinique majeur. Les avancées technologiques à haut-débit des dix dernières années ouvrent de nouvelles perspectives pour caractériser les mécanismes d'oncogenèse et identifier des marqueurs associés au pronostic des cancers. Il est maintenant possible d'analyser les cellules cancéreuses, ainsi que les cellules de leur micro-environnement, tant en ce qui concerne l'expression de gènes (ARN messagers) que les aberrations génomiques (mutations ou variations du nombre de copie d 'ADN) ou épigénomiques (anomalies de méthylation ou d'expression de microARNs). De telles données peuvent donc permettre d'identifier des sous-types de tumeurs ayant des profils moléculaires distincts qui pourraient expliquer des différences de pronostic et de réponse au traitement. De plus, la délimitation de sous-types moléculaires homogènes constitue un premier pas vers la médecine personnalisée, préalable à la recherche de marqueurs théranostiques spécifiques pouvant être les cibles de drogues et/ou à la stratification thérapeutique.

Mon projet de thèse s'est intéressé à décrire finement, via de telles approches, l'hétérogénéité de la maladie sur le plan moléculaire et clinique. Ce travail a été réalisé sur la large cohorte de données générées grâce à la collaboration entre plusieurs acteurs majeurs de la recherche sur le cancer colorectal en France et le programme Cartes d'Identité des Tumeurs (CIT), programme de génomique des cancers initié et financé par la Ligue Nationale contre le Cancer. Dans ce rapport, je commencerai par évoquer l'état de l'art, en m'efforçant de décrire la prise en charge des CCR en clinique et les besoins pour son amélioration, ce qui est connu de l'hétérogénéité cellulaires et moléculaires des CCR et l'apport des analyses des données omiques. Puis je présenterai les résultats obtenus : l'établissement d'une classification sur les données transcriptomiques pour décrire l'hétérogénéité, puis l'apport de cette classification dans un effort d'établir une classification consensus internationale suite aux publications simultanées de travaux similaires. Enfin, je conclurai en discutant l'intérêt et les limites de ces résultats.

Deuxième partie

Revue Bibliographique

Chapitre 1

Le cancer colorectal en clinique : caractéristiques épidémiologiques, cliniques et anatomo-pathologiques

1.1 Une problème de santé majeur

1.1.1 Incidence et mortalité

Le cancer colorectal est parmi les cancers les plus fréquents et les plus mortels dans le monde avec les plus fortes incidences observées dans les pays occidentaux [1] (Figure 1.1).

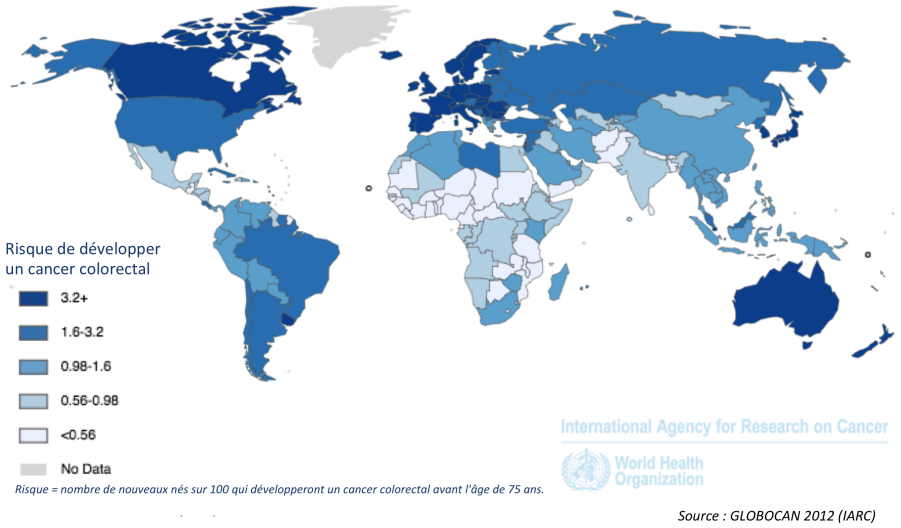


FIGURE 1.1 – Incidence du cancer colorectal dans le monde selon Globocan1.

1. Selon Globocan, un projet visant à fournir des estimations les plus à jour possible de l'incidence, la mortalité et la prévalence de la majorité des cancers au niveau international `http://globocan.iarc.fr/Pages/Map.aspx`

En France, selon les données de l'INCA de 2012 [2], le cancer colorectal est le 3ème cancer le plus fréquent, après les cancers de la prostate et du sein, avec 42150 nouveaux cas par an. En comparaison, parmi les cancers gastro-intestinaux, le 2ème cancer le plus fréquent est celui du pancréas qui ne compte que 9500 nouveaux cas par an. En terme de mortalité, il se place au 2ème rang des causes de décès par cancer après celui du poumon, avec 17 722 décès par an, la survie relative à 5 ans étant de seulement 56%. Malgré ce bilan alarmant, une stabilisation de l'incidence est observée et la mortalité diminue, notamment grâce à un diagnostic plus précoce, une amélioration de la prise en charge thérapeutique et à une diminution de la mortalité opératoire. Le bénéfice attendu de la mise en place du dépistage organisé n'a pas encore d'impact sur les chiffres.

1.1.2 Les facteurs de risque, la part de la génétique et de l'environnement

La part des facteurs héréditaires contribue relativement peu à la plupart des cancers, l'environnement ayant un rôle majeur. Toutefois, pour les cancers colorectaux, un risque significativement accru a été observé pour les patients ayant des antécédents familiaux, avec près d'un tiers des patients ayant une composante héréditaire (Lichtenstein et al., 2000). En effet, la majorité des cancers colorectaux (~70%) sont très probablement dus à des facteurs environnementaux et sont dits sporadiques. De 3 à 5% sont dus à une maladie génétique héréditaire, bien décrite, fortement pénétrante, de type Mendélienne et sont dits héréditaires. Les ~25% restants sont associés à des antécédents familiaux et sont dits familiaux. Ils sont probablement dus à une combinaison de facteurs génétiques moins pénétrants et de facteurs environnementaux (Burt, 2007; Lichtenstein et al., 2000) (Figure 1.2) .

Les formes héréditaires à forte pénétrance

Parmi les cas héréditaires fortement pénétrants, les syndromes héréditaires les plus communs sont la polypose adénomateuse familiale ou FAP (pour Familial adenomatous polyposis), responsable d'1% des CCR, et le syndrome de Lynch, appelé aussi Hereditary non-polyposis colon cancer (HNPCC), responsable de 3 à 5% des CCR (Burt, 2007).

La FAP se manifeste par la formation de plusieurs centaines de polypes dans le côlon, dès l'adolescence, les polypes finissant pas devenir cancéreux avec le temps. C'est une maladie autosomale dominante due à la mutation du gène *adenomatous polyposis coli* APC. Dans le cas du syndrome de Lynch, qui a été l'un des premiers syndromes héréditaires décrits, les patients sont généralement jeunes et présentent pas ou peu de polypes. Les tumeurs ont des caractéristiques anatomo-pathologiques spécifiques, comme notamment l'infiltration lymphocytaire (Fearon, 2011). Les localisations possibles de développement du cancer ne sont pas restreintes au côlon mais peuvent toucher les ovaires, l'estomac, l'intestin grêle, le pancréas, les reins, le cerveau, l'urètre et le canal biliaire, les femmes ayant plus de risque de développer un cancer de l'endomètre. C'est aussi une maladie autosomale dominante due à une mutations dans un des gènes du système de réparation de l'ADN qui confère aux tumeurs une instabilité génétique au niveau des microsatellites (MSI) (voir chapitre suivant2). (Fearon, 2011)

D'autres syndromes beaucoup plus rares (<1%), regroupés sous le nom de syndromes de polypose hamartomateuse, augmentent le risque de CCR comme le syndrome de Turcot,

2. Données de l'Institut National du Cancer (INCA) établies en 2012 : http ://www.e-cancer.fr/cancerinfo/les-cancers/cancers-du-colon/quelques-chiffres-sur-les-cancers-colorectaux

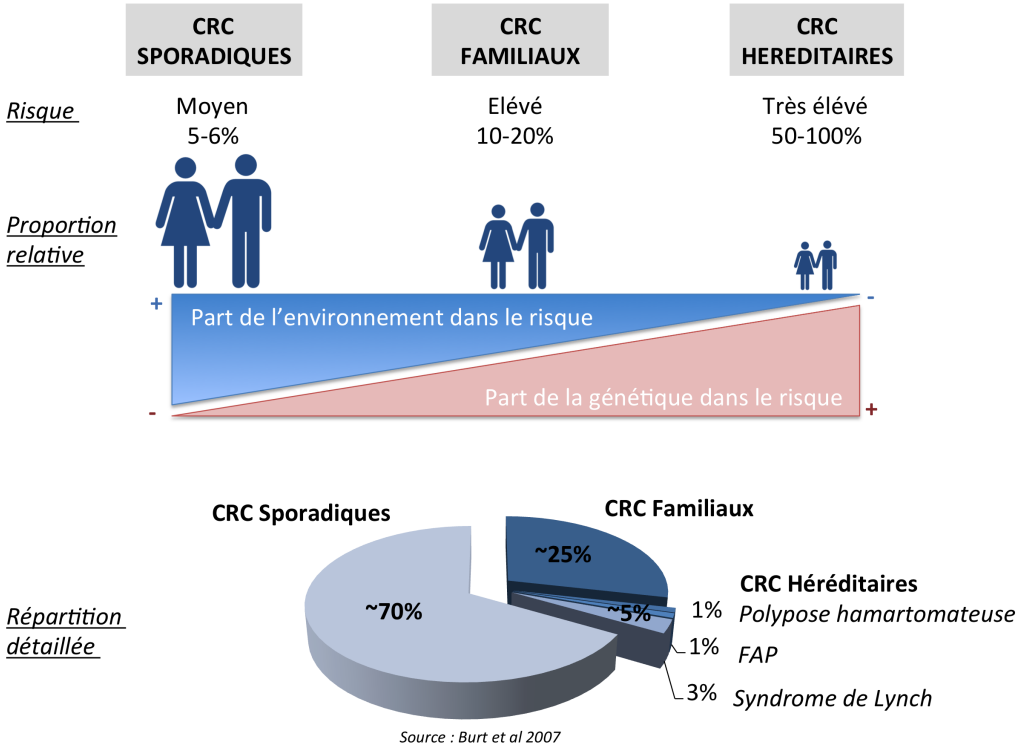


Figure 1.2 – Formes héréditaires, familiales et sporadiques des cancers colorectaux.

le syndrome de Peutz-Jeghers dû à une mutation du gène STK1 et la polypose associée à MUTYH (ou MAP, MUTYH-associated polyposis) dû à une mutation du gène MUTYH.

Pour les formes héréditaires, le risque est extrêmement élevé. Les patients atteints de FAP ont un risque de développer la maladie au cours de leur vie de 100% et ceux atteints du syndrome de Lynch de 50 à 80% (Burt, 2007) (Figure 1.2).

Les formes familiales

Les CCR familiaux sont hétérogènes, regroupant des tumeurs liées à des syndromes héréditaires encore non identifiés, des tumeurs sporadiques liées à des comportements socio-culturels laissant penser à de l'hérédité (comme par exemple la surconsommation de viandes rouges) et des tumeurs liées à une combinaison de facteurs génétiques et environnementaux. Les gènes impliqués dans ce dernier type ne sont pas encore clairement définis (Armelao and de Pretis, 2014). Le risque de développer au cours de sa vie un cancer colorectal dans la population générale est de 5-6%, ce risque est augmenté de 2 à 4 fois si un contexte familiale est retrouvé (Burt, 2007; Armelao and de Pretis, 2014).

Les autres facteurs de risque

En plus des antécédents familiaux, les autres facteurs de risque sont l'âge (la moyenne d'âge au diagnostic est de 71 ans), la présence de polypes, les maladies inflammatoires du côlon, les antécédents personnels de cancers et certaines habitudes de vie (certains types de régimes comme ceux enrichis en viandes rouges ou en acides gras insaturés, l'inactivité physique, l'obésité, la consommation de cigarettes, la consommation importante

d'alcool) [3].

1.2 Physiopathologie et anatomopathologie cancéreuse du côlon

1.2.1 Progression tumorale : la séquence adénome-carcinome

Le côlon normal

Le côlon constitue avec le rectum et l'anus la dernière portion du système gastro-intestinal. Il fait suite à l'intestin grêle. Il est responsable des dernières étapes du processus de digestion. Les principales fonctions du côlon sont l'absorption de l'eau restante et des nutriments, et la compaction et l'évacuation des résidus alimentaires non digérés dans l'intestin grêle.

Il se divise en 4 segments : le caecum/côlon ascendant, le côlon transverse (avant l'angle splénique), le côlon descendant et le côlon sigmoïde (Figure 1.3). Les 2 premiers segments constituent le côlon proximal, et les 2 derniers segments constituent le côlon distal (droit et gauche sont parfois employés pour proximal et distal). Cette séparation en côlon proximal et en côlon distal reflète l'appartenance à des structures embryonnaires distinctes, l'intestin primitif moyen et l'intestin primitif postérieur respectivement. Selon Weinberg (2013), l'origine embryonnaire du tissu est un déterminant clé de la biologie des tumeurs, tout comme la cellule d'origine et le phénotype de la cellule ayant été transformée.

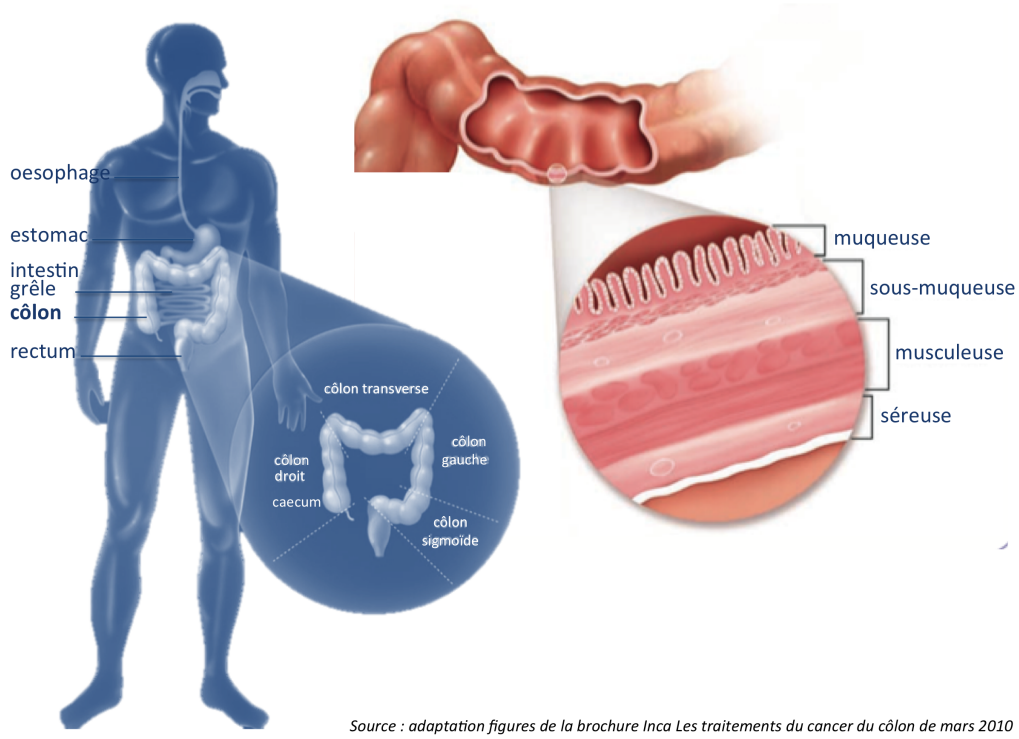


FIGURE 1.3 – Anatomie et structure du côlon.

3. http ://www.e-cancer.fr/cancerinfo/les-cancers/cancers-du-colon/les-facteurs-de-risque

Le côlon forme un tube dont la paroi est constituée de quatre couches concentriques différentes (Barker et al., 2008) :

la muqueuse à la lumière du côlon, consistant en un épithélium simple spécialisé, organisé en cryptes (ou glandes), qui est responsable de l'absorption de l'eau et des nutriments ;

la sous-muqueuse aussi appelée stroma ou chorion, contenant de nombreux vaisseaux sanguins et lymphatiques, des fibres nerveuses et diverses cellules immunitaires ;

la musculeuse constituée de plusieurs couches de muscles, qui va avec le système nerveux entérique permettre l'évacuation du contenu du côlon vers le rectum ;

la séreuse qui constitue une partie du péritoine (la membrane qui entoure les viscères) ;

La transformation cancéreuse

L'initiation de la tumorigenèse colique se fait à partir de l'épithélium de la muqueuse. Dans la majorité des cas, le cancer colorectal provient d'une tumeur bénigne, un adénome, qui se développe à partir des cellules de l'épithélium puis évolue ensuite lentement et finit par devenir cancéreuse (Figure 1.4). Dans les cas très évolués, il finit par former des métastases à distance (dans le foie, les poumons et le péritoine principalement).

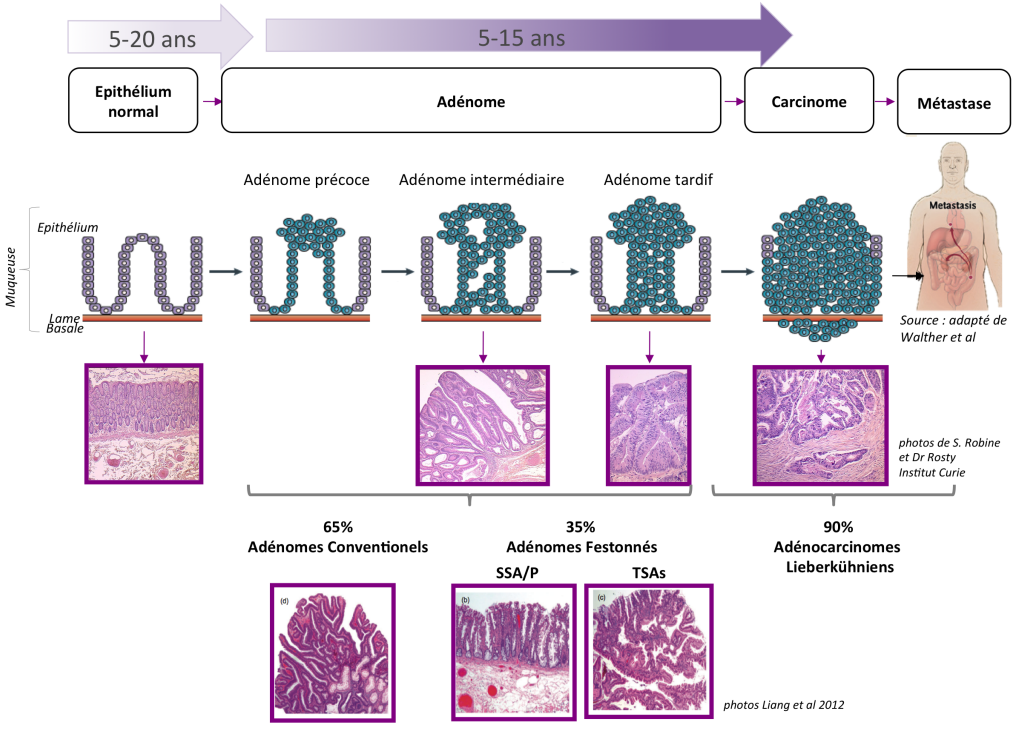


Figure 1.4 – Séquence de progression tumorale adénome-carcinome.

Plus spécifiquement, une partie de l'épithélium commence par devenir hyperplasique, c'est-à-dire que le nombre de cellules augmente mais l'architecture globale de la crypte ou la morphologie des cellules ne sont pas altérées. Puis, généralement, l'épithélium devient dysplasique, c'est-à-dire que les cellules deviennent anormales cytologiquement, ce qui se

traduit notamment par des tailles et formes de noyaux anormales et une activité mitotique accrue. L'épithélium continue alors à évoluer par la multiplication des cellules pour former une masse tumorale bénigne, aussi appelée adénome ou polype ou polype adénomateux [4]. Ces formes contiennent tous les types cellulaires du tissu normal mais leur assemblage est anormal. La tumeur est bénigne car elle reste confinée dans la limite de la lame basale qui sépare l'épithélium du stroma sous-jacent. Si la tumeur traverse la membrane basale et envahit les tissus avoisinants, elle est alors considérée comme cancéreuse. Dans ce cas, les adénomes se transforment en adénocarcinomes caractérisés par leur agressivité et leur capacité à faire des métastases. Seule une fraction des adénomes progressent en cancer, et la progression prendrait plusieurs années, voire des décennies. (Weinberg, 2013).

1.2.2 Les cellules d'origine

Les différents types cellulaires de l'épithélium normal

L'épithélium intestinal est en perpétuel renouvellement (tous les 4-5 jours), ce tissu étant celui qui se renouvelle le plus dans le corps (van der Flier and Clevers, 2009; Barker et al., 2008). Il est formé de cryptes organisées en 3 compartiments spatialement distincts : le compartiment de prolifération à la base des cryptes, le compartiment d'amplification et de lignage cellulaire, et le compartiment de différentiation cellulaire (Figure 1.5).

L'épithélium est constitué de plusieurs types de cellules se répartissant différemment en fonction des compartiments (Figure 1.5). On distingue d'une part les cellules non différenciées, les cellules souches (CS) et les cellules progénitrices (ou "transit-amplifying cells", TA) et, d'autre part, les cellules différenciées. (van der Flier and Clevers, 2009)

Les cellules non différenciées Les **cellules souches** se trouvent à la base des cryptes. Elles sont responsables, par leur renouvellement perpétuel, de la diversité cellulaire des cryptes. Les études convergent vers 4 à 6 cellules souches par crypte. Une CS se définit par sa longévité (souvent la vie entière) et par sa capacité de multipotence, c'est-à-dire à générer les cellules différenciées de son tissu. Elle est donc capable de se renouveler de manière illimitée et, a priori, de manière asymétrique pour donner d'une part une cellule fille très cyclante et une autre en remplacement de la cellule mère.

Les CS sont souvent considérées comme quiescentes, car elles se diviseraient très peu fréquemment. Un certain nombre de cellules comme les cheveux suivent ce modèle. Malgré tout, selon Barker et al. (2008), il n'y a pas de raison a priori qu'une CS soit quiescente et ne procède qu'à des divisions asymétriques. Notamment les cellules souches les plus étudiées chez la drosophile sont très activement cyclantes. De même les cellules souches embryonnaires cyclent très rapidement mais ne font a priori jamais de division asymétrique.

La progénie des cellules souches, les **progénitrices TA**, se développent par plusieurs cycles de mitoses (près de 300 cellules par crypte par jour, contrairement aux CS, les TA ne font qu'un nombre limité de divisions) tout en migrant vers le haut des cryptes. En arrivant près de la lumière, les TA arrêtent leur cycle et finalisent leur processus de différenciation (Merlos-Suárez et al., 2011).

Les cellules différenciées Quatre types principaux de cellules différenciées de morphologie différentes sont retrouvés, 1 type absorbant et les 3 autres sécrétant :

4. Le terme adénome est réservé aux polypes néoplasiques, alors que polype est le terme générique qu'il soit néoplasique ou pas.

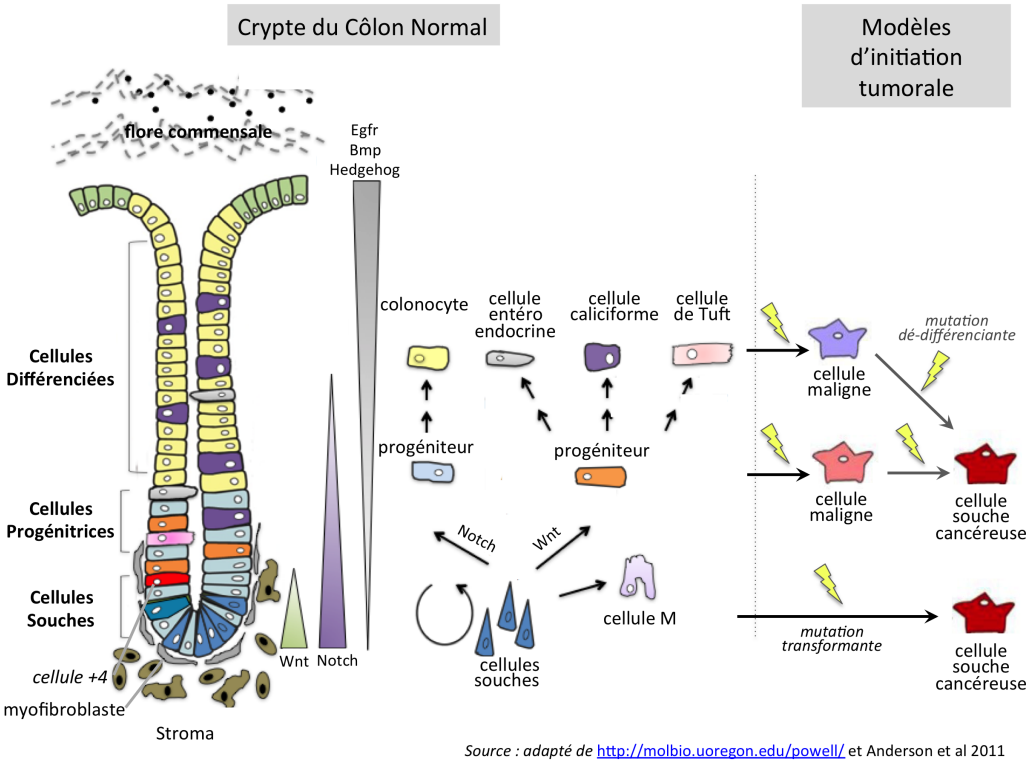


Figure 1.5 – Schéma d'une crypte colique avec le lignage des différents types cellulaires et les modèles possibles de transformation cancéreuse. A gauche, la crypte avec l'ensemble des différents types cellulaires. La cellule rouge correspond à la cellule +4. Les triangles représentent la localisation et l'intensité de l'expression des gènes de la voie de signalisation indiquée. Au milieu, les différents lignages permettant la différenciation dans les différents types cellulaires connus. A droite, les différents modèles d'initiation de la tumorigenèse en fonction du type de cellule d'origine (cellules souches, progénitrices ou différenciées). L'éclair jaune représente un événement mutationnel.

les colonocytes ou entérocytes : ce sont des cellules absorbantes, très abondantes dans l'intestin grêle, moins dans le côlon que l'on retrouve dans le haut des cryptes sur la partie directement en contact de la lumière.

les cellules caliciformes ou "Goblet cells" : elles sécrètent du mucus. Le nombre de ces cellules augmente du côlon au rectum de manière conjointe avec le degré de compaction des selles.

les cellules entéroendocrines : elles représentent moins de 1% des cellules épithéliales. Elles contrôlent la physiologie des intestins par la sécrétion de variété d'hormones comme la sérotonine, la substance P et la sécrétine.

les cellules de Paneth : très abondantes dans l'intestin grêle, elles ne sont retrouvées que dans le caecum, très rarement dans le reste du côlon. Elles ont un rôle dans la protection contre les agents pathogènes par la sécrétion d'agents antimicrobiens. Elles s'intercalent avec les cellules souches à la base des cryptes (Anderson et al., 2011).

D'autres types cellulaires existent mais sont moins bien caractérisés comme les cellules Tuft et les cellules M. Les cellules différenciées migrent rapidement vers le haut de la crypte puis meurent ou sont rejetées dans la lumière du côlon après 4 à 8 jours (Anderson et al.,

2011).

L'identification des cellules souches

Bien que l'existence des CS soit reconnue depuis longtemps dans les intestins, leur identité a été difficilement établie et fait encore débat. Un groupe de cellules souches fait toutefois consensus, les CBC (crypt base columnar cells) qui ont été très bien caractérisées (Romagnolo, 2012) et correspondent au modèle "stem cell zone" proposé par Cheng et Leblond au début des années 70. Elles sont petites, indifférenciées et très cyclantes, intercalées entre les cellules de Paneth, si présentes. Un autre modèle, le modèle "+4 position", positionne les CS comme les cellules en position relative +4 par rapport au bas de la crypte, les 1ères positions étant occupées par les cellules de Paneth. Ce modèle permettrait d'expliquer la position des cellules de Paneth qui migreraient vers le bas et les autres vers le haut (Medema and Vermeulen, 2011; Barker et al., 2008). Plusieurs marqueurs ont été identifiés mais très peu semblent être cohérents d'une étude à l'autre, ou ne sont pas exclusifs des CS (*CD34*, *DCAMKL1*, *EphB* receptors, *Msi-1*, *LGR5*) (Anderson et al., 2011; Di Franco Simone et al., 2011). *LGR5* serait un des marqueurs le plus consistant mais il ne marquerait que les CS se divisant activement (Anderson et al., 2011).

Le concept de niche

L'environnement des cellules souches, ou *niche*, semble avoir un impact important sur le maintien, la régulation et l'auto-renouvellement de ces cellules mais la réalité de ce concept n'est pas encore clairement définie. En effet, en plus des cellules épithéliales adjacentes aux CS, de nombreuses cellules sont retrouvées dans le fond des cryptes qui pourraient interagir avec les CS : les myofibroblastes, les neurones entériques, les cellules endothéliales, les lymphocytes intraépithéliaux de même que d'autres composants comme la membrane basale (Romagnolo, 2012). Notamment, les myofibroblastes auraient un rôle important dans le maintien des CS pour leur rôle dans l'organogenèse et la réparation des tissus. Les voies Wnt, BMP, Notch et Sonic Hedgehog auraient un rôle majeur dans le maintien de la niche, les effecteurs de ces voies étant distribués différemment le long de l'axe de la crypte (Di Franco Simone et al., 2011).

La théorie des cellules souches cancéreuses

Dans ce contexte, le modèle classique de développement d'une tumeur, le modèle clonal ou stochastique, considère que toute cellule est capable de produire une tumeur (Figure 1.6). Les tumeurs sont toutefois hétérogènes d'un point de vue cellulaire avec divers degrés de différenciation des cellules et différents aspects morphologiques. La cellule initiatrice devrait donc être capable de donner naissance à une diversité de cellules (Anderson et al., 2011). Ces dernières années a émergé l'idée que les cancers seraient une maladie des CS, la théorie des cellules souches cancéreuses (CSC pour Cancer Stem Cells), en anglais *Cancer Stem Cell Therory*. La première connexion entre cancer et CS remonte au début du 19ème siècle quand des similarités entre tumeurs et tissus embryonnaires ont été montrées. Cela a donné naissance à la théorie "*embryonal rest*", postulant que les cancers sont causés par des cellules ayant des propriétés comparables à celles de l'embryon (Nguyen et al., 2012).

Le théorie des CSC postule que les tumeurs primaires et les tumeurs métastatiques se développent à partir d'une petite population de cellules cancéreuses ayant la capacité d'auto-renouvellement et de multipotence qui sont capables d'initier et de maintenir

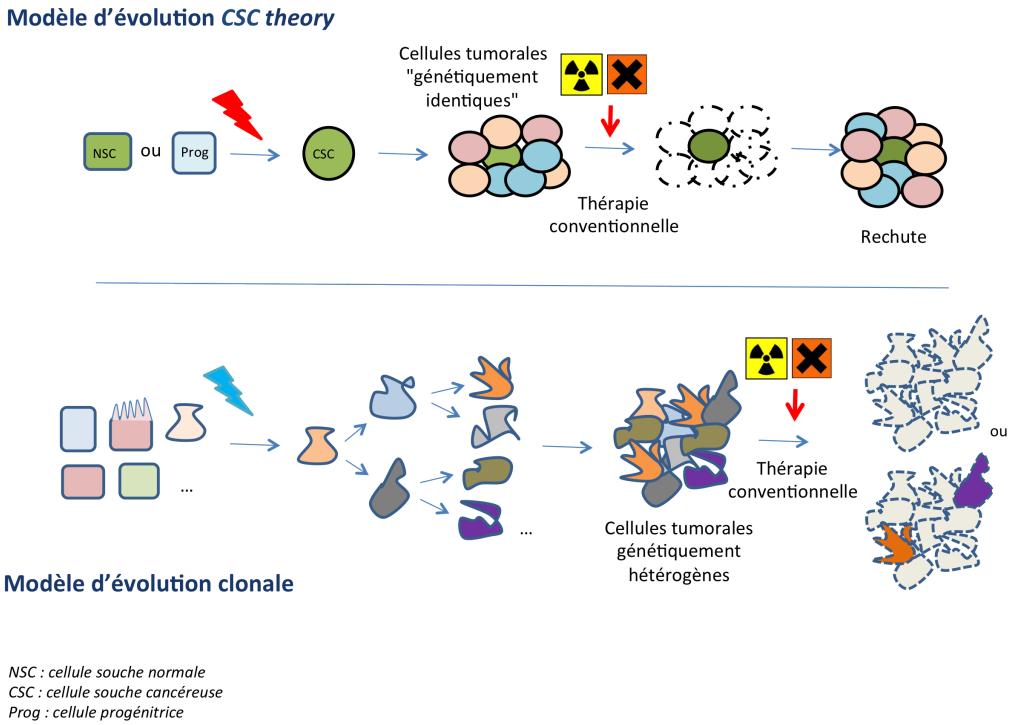


Figure 1.6 – Modèles de développement tumoral

une tumeur (Anderson et al., 2011). Le concept de CSC est souvent confondu avec le concept plus restreint qu'une cellule souche normale devienne cancéreuse, on parle alors de "*cancerous stem cell*" (Nguyen et al., 2012). En effet, les CSC peuvent provenir de CS normale, seul une mutation (génétique ou épigénétique) serait alors nécessaire pour la transformation tumorale. Mais elles peuvent également être issues de cellules progénitrices ou de cellules différenciées. Dans ces cas, 2 mutations sont nécessaires pour la transformation, une mutation transformante et une mutation dédifférenciant en CS car l'auto-renouvellement et la capacité à donner divers types cellulaires doivent être maintenus (Figure 1.5) (Anderson et al., 2011). Dans ce modèle, seules les CSC ont le pouvoir de reproduire la tumeur, les autres cellules tumorales étant différenciées. L'hétérogénéité tumorale proviendrait alors de mutations dans la CSC et la différenciation de sa progénie. L'identification des CSC n'est pourtant pas facile et plusieurs marqueurs différents des cellules normales ont été mis en évidence (*CD24*, *CD29*, *CD44*, *CD133*, *CD166*, *ESA*, *LGR5*, *ALDH1*, *EPCAM*)(Di Franco Simone et al., 2011; Anderson et al., 2011).

L'autre aspect important de cette théorie est son impact clinique quant aux traitements des patients. En effet, cette théorie pourrait expliquer les cas de chimiorésistance et de récurrence, les chimiothérapies classiques ne s'attaquant qu'aux cellules en division. Si les CSCs sont quiescentes alors elles échapperaient au traitement et recoloniseraient le tissu. Il est donc important pour améliorer les thérapies actuelles de viser spécifiquement ces cellules (Di Franco Simone et al., 2011).

1.2.3 Les lésions précurseurs

Les différents type d'adénomes

Toute excroissance au-dessus de la muqueuse adjacente est communément appelée polype. Tout polype n'est pas précancéreux, seuls les adénomes sont néoplasiques. Initialement deux grands groupes étaient classiquement définis : les polypes adénomateux, dysplasique et potentiellement précancéreux, et les polypes hyperplasiques, non néoplasiques. Mais dans les années 2000, suite à la mise en évidence que des polypes hyperplasiques pouvaient être précurseurs de tumeurs, et à plusieurs démonstrations de l'existence de lésions moléculaires précancéreuses dans ces polypes (notamment de l'instabilité microsatellitaire, des aberrations de méthylation, des mutations de *BRAF* et *KRAS*), une nouvelle voie de carcinogenèse à partir de polype hyperplasique a été proposée, la voie dite festonnée ou *serrated* en anglais (Snover, 2011).

Depuis 2003, la nouvelle classification des polypes distingue 2 grands groupes : les adénomes conventionnels (ou adénomateux) et les adénomes festonnés incluant les polypes hyperplasiques et une diversité de polypes néoplasiques. De 65 à 90% des CCR proviendraient de polypes conventionnels et de 10 à 35% des polypes festonnés suivant les publications. Les tumeurs bénignes finalement forment donc un groupe varié. (Snover, 2011; Liang et al., 2013)

Les adénomes conventionnels

Au sein des polypes conventionnels, 3 grands types sont distingués : les tubuleux (75%), les tubuleux villeux(20%) et les villeux (5-10%) (Fleming et al., 2012). L'adénome tubuleux est le plus petit et le plus commun des adénomes mais c'est le moins susceptible de se transformer en cancer. L'adénome villeux a le plus de risque de devenir cancéreux. Il est habituellement plat avec une base large et peut devenir assez gros, ce qui le rend difficile à enlever. L'adénome tubulo-villeux présente à la fois des caractéristiques de l'adénome villeux et de l'adénome tubuleux. Le risque qu'il se transforme en cancer semble être à mi-chemin entre celui de l'adénome villeux et celui de l'adénome tubuleux (Fleming et al., 2012).

Les adénomes festonnés

Les polypes festonnés ont une morphologie en dent de scie ou *sawtooth-like*. Trois grands types sont également distingués : les hyperplasiques non néoplasiques (HP), les "traditional serrated adenomas" présentant une dysplasie cytologique (TSA) et les "sessile serrated adenomas" (SSA) qui ont une histologie festonnée mais ne présentent pas de dysplasie cytologique. Les tumeurs issues de ces polypes ont des caractéristiques moléculaires et cliniques différentes. Notamment elles ont un moins bon pronostic et sont plus associées aux mutations de *BRAF* et *KRAS* et à l'instabilité microsatellitaire. Les SSA sont plus associés à l'instabilité microsatellitaire et à la mutation de *BRAF*, et les TSA à la mutation de *KRAS* (Liang et al., 2013).

1.2.4 Les différents types histologiques de carcinomes

Une faible hétérogénéité morphologique

Le cancer colorectal est plutôt homogène histologiquement. Plus de 90% des cancers colorectaux sont des adénocarcinomes (du grec adéno qui signifie glande et carcino cancer). Les autres formes regroupent des tumeurs carcinoïdes provenant de cellules nerveuses

digestives, sarcomes des tissus mous, lymphomes et métastases d'autres organes comme l'ovaire, la prostate, l'estomac ou le sein. Il existe tout de même plusieurs sous-types histologiques d'adénocarcinomes mais les plus fréquents, les adénocarcinomes Lieberkühniens, représentent 95% des adénocarcinomes. Les autres formes plus rares sont les adénocarcinomes mucineux (ou colloïdes) et les adénocarcinomes dits à cellules en bague à chaton, ou "*signet ring cell*", (adénocarcinome ayant plus de 50% des cellules tumorales présentant du mucus intra-cytoplasmique). (Fleming et al., 2012).

Par ailleurs, le type d'adénome à l'origine du carcinome se retrouve dans sa morphologie. On peut donc distinguer des tumeurs de type festonné ou conventionnel. Les carcinomes festonnés étant identifiés par la présence de structures cribriformes (en forme de crible) et trabéculaires (en forme de sillons), la présence de mucus intra- et extracellulaire, des noyaux ronds ou ovoïdes avec une large membrane nucléaire, un faible ratio noyau vs cytoplasme, la préservation de la polarité nucléaire et un cytoplasme clair ou éosinophile (qui a une affinité pour l'éosine, un colorant orange-rosé) (Jass, 2007; Laiho et al., 2007).

Un faible hétérogénéité selon le grade

Une mesure couramment utilisée par les anatomopathologistes est le grade qui mesure le niveau de différenciation des cellules tumorales. Une faible différenciation indique une perte de ressemblance avec le tissu parent, ce qui correspond pour les adénocarcinomes à une perte du développement glandulaire (Jass, 2007). Les adénocarcinomes sont divisés en bien (>95% de structures glandulaires), modérément (50-95%) ou peu différenciés (5-50%). Les termes de bas-grade (bien et modérément différencié) et haut-grade (peu différencié) sont couramment utilisés. Le grade est défini sur le composant tumoral le moins différencié en dehors du front d'invasion. Près de 70% des adénocarcinomes sont modérément différenciés, 20% peu différenciés et 10% bien différenciés (Fleming et al., 2012). Par convention, les adénocarcinomes mucineux et les adénocarcinomes dits à cellules en bague à chaton sont peu différenciés (Hamilton et al., 2006).

Les autres aspects histologiques participant de l'hétérogénéité

La morphologie à la marge d'invasion de la tumeur peut prendre des formes diverses. Notamment un aspect d'invasion diffus rendant difficile la délimitation de la tumeur du tissu sain avec la présence de cellules tumorales isolées ou groupés en petits amas, se détachant à la marge (on parle de bourgeonnement tumoral ou "tumor budding"). Le bourgeonnement tumoral a été associé à un plus mauvais pronostic et est considéré comme la traduction morphologique du phénomène de transition épithélio-mésenchymateuse (EMT en anglais) (Svrcek and Fléjou, 2012).

L'infiltration lymphocytaire est aussi un aspect permettant de différencier les tumeurs. Une infiltration lymphocytaire péri-tumorale très marquée a initialement été utilisée pour caractériser les tumeurs associées à un syndrome de Lynch. Une infiltration intra-tumorale avec la présence de lymphocytes intraépithéliaux, appelés TILS pour *Tumor Infiltrating Lymphocytes*, notamment la présence de lymphocytes T cytotoxiques, constitue aussi un marqueur des tumeurs avec instabilité microsatellitaire d'origine héréditaire mais également d'origine sporadique. Par ailleurs, la présence de lymphocytes T cytotoxiques est très variable suivant les tumeurs. La relative rareté des lymphocytes cytotoxiques dans certaines tumeurs, comparativement au tissu normal qui en contient également, a amené certains auteurs à suggérer qu'ils étaient détruits par les cellules tumorales (O'Connell et al., 1996; Jass, 2007).

1.3 La classification clinique

En clinique, classer les tumeurs est primordial pour savoir comment traiter et prendre en charge le patient.

1.3.1 La classification TNM

Avant les années 1930, l'établissement du pronostic était basé sur le grade histologique (basé sur la différentiation cellulaire), mais sa valeur pronostique était limitée. En 1932, Cuthbert Dukes publia la classification des cancers du rectum qui constituera la base des systèmes de classifications actuels. Le succès de cette classification résida dans sa simplicité et, de part sa valeur pronostique et prédictive de la mortalité après opération, dans son impact pour le choix thérapeutique. Cette classification prend en compte la profondeur de pénétration dans la paroi et la présence ou non de métastases dans les ganglions lymphatiques régionaux. La classification de Dukes fut longtemps utilisée pour les CCR mais elle a été supplantée depuis quelques années par la classification plus détaillée TNM (Tumor-Node-Metastasis) et n'est plus recommandée en clinique (Hamilton et al., 2006; Haq et al., 2009).

La classification TNM est la classification proposée par l'American Joint Committee on Cancer (AJCC) qui, depuis 1959, travaille sur la définition de standards de classification des cancers. Elle mesure l'étendue locale et à distance des cellules cancéreuses à 3 niveaux : 1) l'envahissement de la tumeur primaire (T), 2) l'envahissement ganglionnaire périphérique et le nombre de ganglions touchés (N, pour node) et 3) la dissémination métastatique (M) (Figure 1.7). Pour chaque niveau, un chiffre est donné puis le tout est combiné en 4 stades allant de I à IV. Schématiquement, les stades I et II sont caractérisées par un envahissement de la paroi intestinale allant de la sous-muqueuse à la totalité de la paroi ; les stades III par un envahissement de la paroi avec un envahissement ganglionnaire locorégional ; et les stades IV par la présence de métastase à distance, généralement dans le foie, les poumons ou le péritoine. Le stade 0 correspond aux adénomes, c'est-à-dire que les cellules tumorales ne dépassent pas la lame basale.

Selon les données de l'INCA de 2010, la répartition en fonction des 4 stades est quasiment identique avec 26% de stades I, 21% de stades II, 21% de stades III et 26% de stades IV (6% non déterminé) [5]

1.3.2 Valeur pronostique

L'intérêt majeur de la classification TNM est sa forte association au pronostic (Figure 1.7). En effet, la survie spécifique à 5 ans est de 93% pour les stades I, de 83% pour les stades II, de 60% pour les stades III et seulement de 8% pour les stades IV qui sont quasiment incurables. Des raffinements de cette classification, sous-divisant les stades existants, lors de la 6ème édition en 2002, permettent de mieux définir le pronostic (O'Connell et al., 2004). Elle constitue l'un des facteurs pronostiques majeurs des cancers coliques, et reste le principal outil du clinicien pour la décision thérapeutique après chirurgie.

Toutefois, d'autres facteurs apportent de l'information complémentaire au pronostic : le grade (le degré de différenciation des cellules), la qualité de la résection/exérèse chirurgical, l'invasion lymphovasculaire et nerveuse, la présentation sous forme d'occlusion ou d'une

5. Survie attendue des patients atteints de cancer : état des lieux 2010 (http ://www.e-cancer.fr/publications/69-epidemiologie/578-survie-attendue-des-patients-atteints-de-cancer-etat-des-lieux-2010), labeltnmrepartition

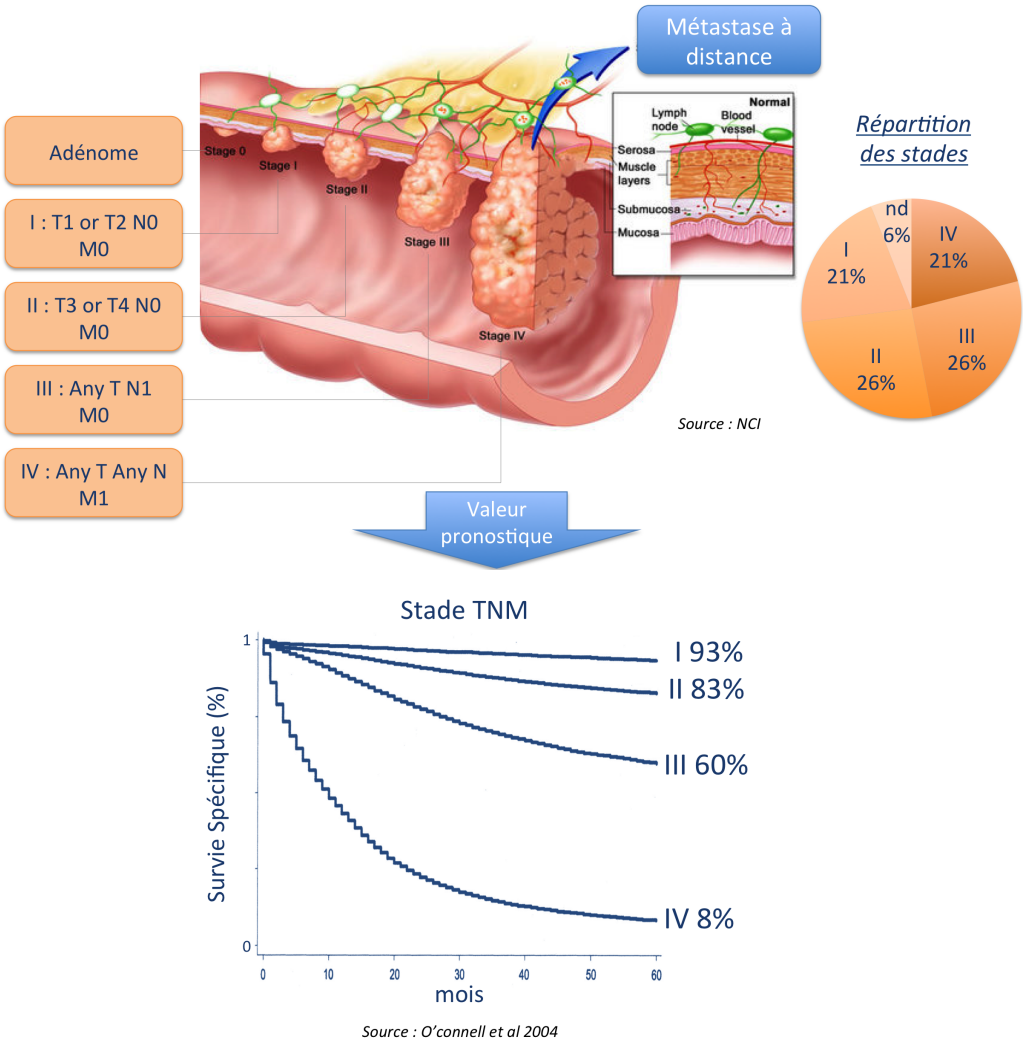


FIGURE 1.7 – Classification clinique selon le stade TNM et sa valeur pronostique.

perforation tumorale (stage T4) et le nombre de ganglions envahis (O'Connell et al., 2004; Sablin et al., 2009).

1.3.3 Implication sur le choix de traitement

La classification TNM constitue la base du choix thérapeutique pour les oncologues. Globalement, les stades I et II, de meilleur pronostic, ne subissent qu'une simple chirurgie, les stades III une chirurgie suivie d'une chimiothérapie adjuvante. Pour les stades IV métastatiques, le traitement est adapté en fonction du patient, avec en général une chimiothérapie couplée à une thérapie ciblée si les caractéristiques moléculaires de la tumeurs s'y prêtent.

Les traitements standards La chirurgie est le traitement principal du cancer du côlon. Elle consiste à enlever la portion du côlon atteinte par la tumeur ainsi que le réseau de ganglions qui en dépend. Le côlon n'étant pas un organe vital, il est possible de vivre normalement même si on en enlève une grande partie, voire la totalité. Concernant le traitement

médicamenteux, les stades III et IV et quelques stades II à haut risque sont traités par chimiothérapie au 5-fluoro-uracile, appelé 5-FU, qui a été, depuis 1957, et jusqu'en 1996, le seul médicament disponible. Il agit principalement sur la synthèse d'ADN en bloquant l'activité de la thymidylate synthase, et donc la synthèse de la pyrimidine, conduisant à un arrêt du cycle cellulaire et à l'apoptose. Il est généralement associé à l'acide folique, qui accroît son efficacité (5FU/LV, LV pour leucovorine, un acide folique réduit) (Kelly and Goldberg, 2005). Plusieurs traitements sont apparus après 1996 et aujourd'hui le traitement de référence, depuis la publication des résultats sur l'essai MOSAIC (André et al., 2004), est l'oxaliplatine (un analogue du platine induisant des ponts intra-brins d'ADN et résultant à l'apoptose) combiné au 5FU/LV, ou FOLFOX. Toutefois, son utilisation est difficile en pratique du fait de nombreux effets secondaires rapportés. L'irinotécan, un inhibiteur de la topoisomérase I, a également été montré associé à un bénéfice pour le patient lorsque combiné avec le 5FU/LV, FOLFIRI, mais a lui aussi de nombreux effets secondaires et est généralement utilisé en deuxième ligne de traitement (Saltz et al., 2000; Kelly and Goldberg, 2005).

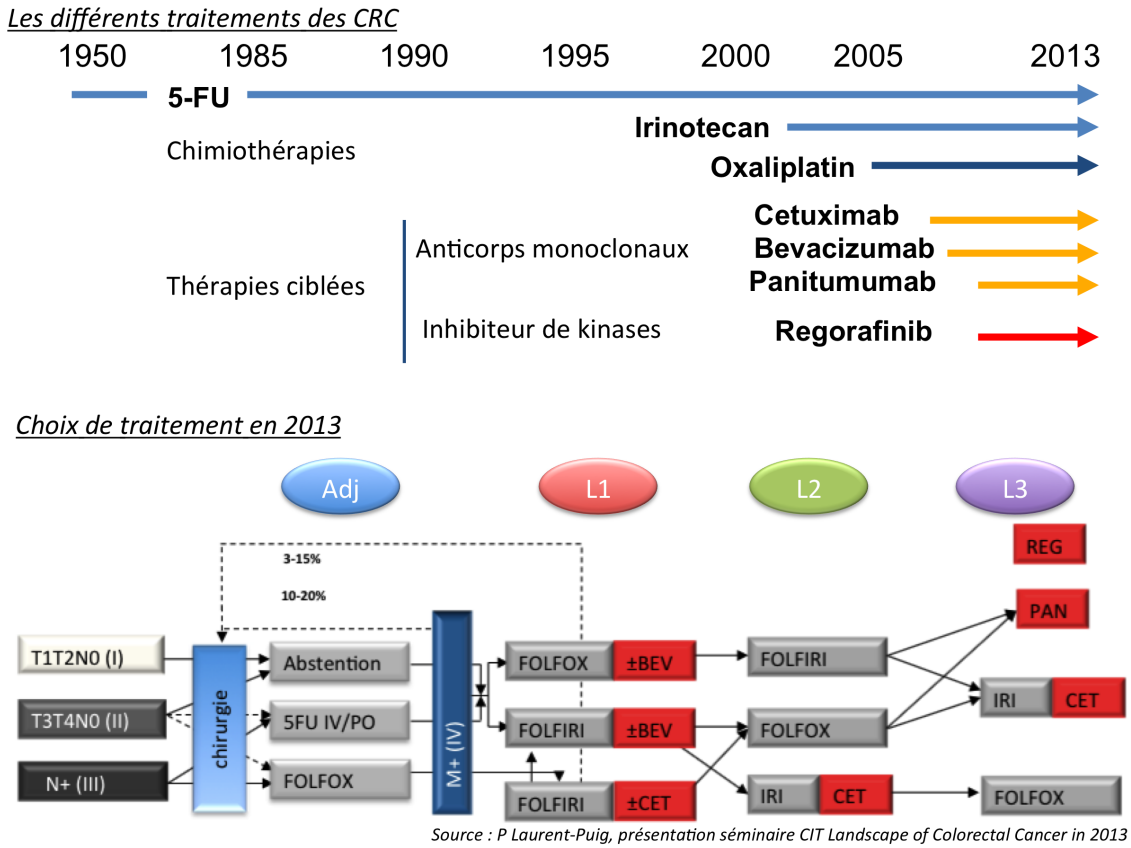


FIGURE 1.8 – Traitements des cancers colorectaux. adj, traitement adjuvant ; L, ligne de traitement ; N+, tumeur avec un envahissement ganglionnaire ; M+, tumeur métastatique

Émergence de la médecine personnalisée ou de précision En complément des chimiothérapies, les progrès de la recherche ont permis de développer de nouveaux médicaments ciblant spécifiquement les cellules cancéreuses, appelés thérapies ciblées ou traitements ciblés. Il y a 2 grandes familles, les anticorps monoclonaux et les inhibiteurs de kinases. Trois médicaments de la famille des anticorps monoclonaux et 1 des inhibiteurs

de kinases ont montré leur efficacité : le bevacizumab (commercialisé sous le nom Avastin), visant le régulateur clé de l'angiogenèse VEGFR, le cetuximab (commercialisé sous le nom Erbitux) et le panitumumab (commercialisé sous le nom Vectibix) visant le facteur de croissance EGFR, et le Regorafenib (commercialisé sous le nom de Stivarga) inhibant plusieurs kinases dont une ciblant VEGFR. Ces traitements sont indiqués en seconde ligne de traitement ou pour les CCR métastatiques.

En clinique, un des critères pour mieux traiter une maladie est de bien la classer. La classification des tumeurs selon le système TNM s'est imposée car elle constitue le meilleur facteur pronostique utilisé en clinique aujourd'hui. D'autres facteurs cliniques sont pronostiques mais ne sont pas aussi efficaces que le stade de la tumeur.
Malgré cette classification, près de 10% à 20% des patients de stade II et de 30% à 40% des stades III développent des récurrences. Le paradigme "one treatment fits all" trouve ici ses limites. Cela soulève la question de l'hétérogénéité moléculaire des tumeurs et suggère l'existence de voies de carcinogenèse distinctes. Cela pose aussi la question de l'hétérogénéité intra-tumorale comme source de résistance aux traitements.
En effet, le stade TNM ne prend pas en compte la diversité histologique et cytologique, uniquement le degré de progression de la tumeur. Or, comme on l'a vu dans ce chapitre, d'un point de vue histologique, bien que les CCR soient assez homogènes, il existe une diversité des cellules, des lésions d'origine et de la part de la génétique dans ces cancers. Diverses voies de progression tumorale sont donc envisageables. Il peut donc être nécessaire de raffiner la classification en s'aidant des approches moléculaires ce qui sera abordé dans les chapitres suivants.

Chapitre 2

Le cancer colorectal à l'échelle moléculaire : caractérisation et classification moléculaire

L'oncologie moléculaire est une discipline plutôt récente. C'est avec la découverte en 1975 du premier proto-oncogène [1] que l'analyse moléculaire est vraiment entrée dans l'étude des cancers (Weinberg, 2013) (Figure 2.1). Depuis, les cancers ont été décrits comme des maladies dues à l'accumulation progressives d'altérations génétiques ou épigénétiques notamment dans des oncogènes et dans des gènes suppresseurs de tumeurs. Les oncogènes sont généralement impliqués dans la régulation de la prolifération et de la survie cellulaire, leur activation induisant la transformation tumorale. Les gènes suppresseurs de tumeurs, à l'inverse, inhibent en condition normale la croissance et la prolifération, leur inactivation levant donc l'inhibition et favorisant le développement tumoral. Tout un pan de la recherche en cancérologie s'est consacré à identifier ces gènes.

Pour le cancer colorectal, en 1990, dans un célèbre article publié dans la revue Cell, Fearon and Vogelstein (1990) ont établi, en terme moléculaire, la première séquence de transformation d'un épithélium colique sain en adénocarcinome invasif. Ils montrèrent que le processus requiert l'accumulation progressive et par étapes de plusieurs mutations touchant des gènes clés. Depuis la publication de ce modèle, d'autres mutations et surtout différentes formes d'instabilité ont été découvertes et décrites dans le cancer colorectal, qui remirent en question l'unicité du modèle.

2.1 Les mutations clés de la tumorigenèse colique

2.1.1 Le modèle d'oncogenèse de Fearon et Vogelstein

Dans le modèle de Fearon and Vogelstein (1990) (Figure 2.2), des gènes/altérations ont été proposés à chaque étape de la tumorigenèse, de l'épithélium normal au carcinome, en passant par les différents stades de progression de l'adénome.

Le principal gène identifié impliqué dans l'initiation des cancers colorectaux est le gène suppresseur de tumeur *APC* (Adenomatous Polyposis Coli). La perte d'*APC* conduit à une activation constitutionnelle de la voie de signalisation Wnt/β-caténine et à une prolifération anarchique et continue des cellules coliques. La forme sauvage d'*APC* forme un complexe protéique qui fixe la β-caténine et l'empêche donc de se transloquer vers le noyau pour

1. Gène ayant la capacité après son altération d'induire un cancer à partir de cellules normales

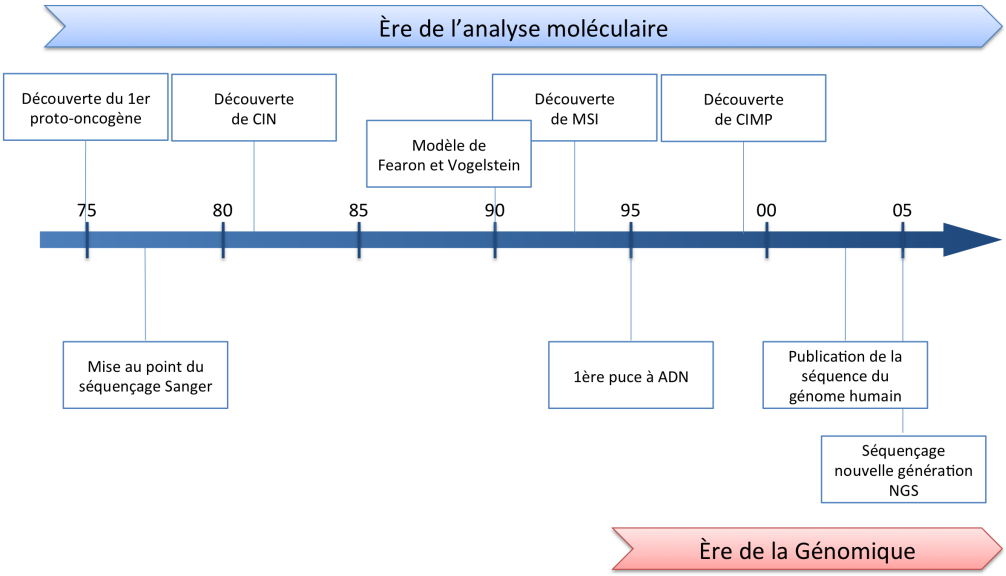


FIGURE 2.1 – Chronologie des principales découvertes moléculaires liées à l'oncologie du cancer colorectal

activer ses cibles, incluant certains gènes de prolifération cellulaire. Le polype qui résulte de cette mutation est plus susceptible d'acquérir d'autres mutations ou épimutations, notamment des pertes de méthylation de l'ADN, lesquelles sont souvent observées dès les formes très précoces d'adénome. Cette hypométhylation pourrait être à l'origine de l'instabilité du génome, conduisant à des pertes et gains de chromosomes, ce qui rendrait le polype plus susceptible d'acquérir d'autres mutations le permettant d'évoluer vers une forme néoplasique.

Ensuite, une mutation activatrice dans le gène *KRAS*, conduisant à une augmentation de la prolifération cellulaire, est responsable de la progression du polype hyperplasique en polype dysplasique. *KRAS* est une enzyme impliquée dans les voies de signalisation en réponse à divers facteurs de croissance dont l'*EGF*, epidermal growth factor. Il est intéressant de noter que l'article mentionne que, bien qu'ayant un rôle plus tardif dans l'oncogenèse dans la plupart des tumeurs, *KRAS* est un oncogène capable de conférer des propriétés néoplasiques et pourrait être l'événement initiateur d'un sous-ensemble de tumeurs.

D'autres altérations survenant plus tardivement s'ensuivent. La perte du chromosome 18q permet le passage à des formes très avancées d'adénomes avec l'implication des gènes suppresseurs de tumeurs localisés sur ce chromosome comme *DCC* (deleted in colorectal cancer) et *SMAD4*/*SMAD2*, participant à la voie de signalisation TGF-β. Le gène suppresseur de tumeur TP53 localisé sur le chromosome 17, également fréquemment perdu, conduit ensuite à la formation d'un adénocarcinome. Les cellules cancéreuses présentent alors une forte instabilité caractérisée par des gains et des pertes de fragments chromosomiques qui se rajoutent à ces différentes mutations et confèrent aux cellules la capacité d'envahir les tissus adjacents.

Malgré la présentation séquentielle de l'acquisition des mutations, qui semble être la voie préférentielle, l'article souligne que c'est plus l'accumulation que l'ordre qui semble avoir une importance, les altérations observées pouvant l'être à différents stades de la progression tumorale dans certaines tumeurs. Ceci a été confirmé par la suite par plusieurs études qui ont montré que seulement 10% des CRC présentaient des mutations conjointes des 3 gènes *APC*, *KRAS*, *TP53* (Jass, 2007). Ceci indique que ce modèle est trop simplifié et ne reflète la biologie que d'une partie des CRC, et que d'autres voies seraient impliquées dans la carcinogenèse colique.

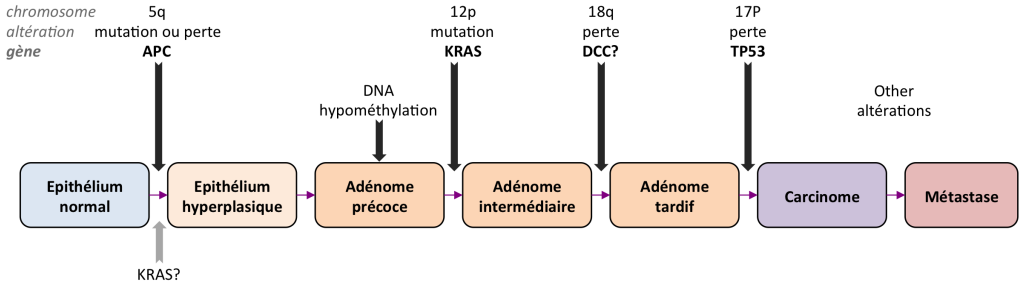


Figure 2.2 – Le Modèle de tumorigenèse Colorectal de (Fearon and Vogelstein, 1990).

2.1.2 Les autres mutations clés pour la tumorigenèse

Les "Drivers" et "Passengers"

Greenman et al. (2007) ont introduit en 2007 la notion d'événements "drivers" et "passengers" afin de distinguer les mutations participant de la carcinogenèse, donc sélectionnées, des événements qui apparaissent par hasard, lors de la division incessante des cellules ou indirectement lors d'un remaniement chromosomique ou de méthylation. Les "drivers" correspondent à des mutations dans des gènes clés associés à l'initiation tumorale, la progression et/ou la maintenance du cancer. Le challenge dans la recherche de mutations d'intérêt pour la compréhension de l'oncogenèse des CRC étant de distinguer les mutations "drivers" des mutations "passengers".

Les mutations somatiques

Wood et al. (2007) ont été les premiers à décrire le "paysage" mutationnel des cancers colorectaux en réalisant le séquençage des séquences codantes des gènes de la base de données RefSeq. Ces travaux ont montré que chaque tumeur présente près de 80 mutations en moyenne dans les exons. Les mutations les plus fréquentes touchaient les gènes du modèle de Fearon et Vogelstein (*APC*, *TP53*, *KRAS*) ainsi que *PIK3CA* et *FBXW7*. Les autres mutations ne sont trouvées que dans peu de tumeurs mais un certain nombre d'entre elles pourraient toutefois être des *drivers*. (Wood et al., 2007). D'autres mutations ont été répertoriées notamment dans les oncogènes *BRAF* (5-10%), *CMYC* (5-10%) et dans des gènes suppresseurs de tumeurs *PTEN* (10%) et *SMAD4*(10%) (Fearon, 2011).

En 2012, le TCGA (The Cancer Genome Atlas), un programme américain financé par le NIH visant à caractériser à l'échelle moléculaire par approches à haut-débit l'ensemble des types de cancers et à rendre les données accessibles à tous, a ensuite, par le séquençage des exons de 276 échantillons, trouvés des mutations à fréquence significative dans les gènes *POLE*, *ARID1A*, *SOX9*, *FAM123B* et *TCF7L2*. Ils ont reporté l'ensemble des mutations

en tenant compte d'une partie de l'hétérogénéité de la maladie, l'hypermutabilité des tumeurs, chaque groupe ayant des mutations associées distinctes (Cancer Genome Atlas Network, 2012).

Les mutations germinales

Comme on l'a vu dans le premier chapitre, près de 5% des patients atteints d'un cancer colorectal ont une mutation germinale, héritée, à l'origine du cancer. Les gènes impliqués sont connus : *APC* étant le gène du syndrome FAP, les gènes du système MMR *MLH1/MSH2/MSH6/PMS2* pour les formes du syndrome de Lynch, *MUTYH* pour les formes MAP, *STK1* pour les formes Peutz-Jeghers syndrome. La table 2.1 liste l'ensemble des gènes connus à l'origine de syndromes héréditaires. Ces gènes constituent donc des gènes clés pour la carcinogenèse qui peuvent être intéressants à regarder pour l'étude des formes sporadiques.

Genes	**Syndromes Associés**
APC	Polypose Adénomateuse Familiale (FAP)
	Syndrome de Gardner
	Syndome de Turcot's
	FAP atténuée
MLH1	Syndrome de Lynch
	Syndome de Turcot
PMS2	Syndrome de Lynch
	Syndome de Turcot
MSH2	Syndrome de Lynch
MSH6	Syndrome de Lynch
STK11	Syndrome de Peutz-Jeghers
PTEN	Syndome de polypose juvénile
	Maladie de Cowden
MYH	Polypose associée à MYH
SMAD4	Syndome de polypose juvénile

TABLE 2.1 – Liste des gènes impliqués dans des formes héréditaires de CRC d'après Fearon (2011); Burt (2007).

Pronostic des mutations somatiques

Le pronostic associé aux mutations diverge beaucoup entre les études. L'une des raisons de ces divergences vient probablement de l'hétérogénéité des CRC incluant notamment les différents types d'instabilité génomique développés ci-après. Parmi ces mutations, *BRAF* a été associé à un mauvais pronostic, et *KRAS* est prédictif de réponse au traitement anti-EGFR (Roth et al., 2010). Selon certaines études, la mutation de *PIK3CA* est associée à un mauvais pronostic (Liao et al., 2012b). D'autres études l'associe à un bon pronostic sous traitement à l'aspirine (Liao et al., 2012a).

2.2 Les trois formes d'instabilité du génome tumoral observées

Un des paradoxes qu'intéresse l'oncologie moléculaire est de comprendre comment il est possible de développer un cancer alors que l'ADN se révèle être l'élément le plus stable des cellules, protégé par de nombreux mécanismes de défense contre l'instabilité. En effet, en considérant la rareté des mutations dans les cellules normales et la fréquence élevée de mutations dans les cancers humains, le taux de mutations spontanées n'est pas suffisant pour expliquer ce nombre observé dans les tumeurs. Loeb (2001) proposa que les cellules cancéreuses arborent un phénotype mutateur, ou "mutator phenotype", qui est le résultat de mutations dans des gènes impliqués dans la stabilité du génome.

Dans les cancers colorectaux, trois formes différentes d'instabilité du génome sont principalement observées :

— l'instabilité chromosomique (CIN, Chromosomal INstability),
— l'instabilité microsatellitaire (MSI, Microsatellite Instability) et
— le phénotype hyperméthylateur (CIMP, CpG Island Methylator Phenotype).

Il est à noter que l'instabilité systématique dans les microsatellites est associée à un phénotype hypermutateur en dehors des microsatellites. Une 4ème forme correspondant à un phénotype hypermutateur sans contexte MSI a été récemment décrite (Cancer Genome Atlas Network, 2012).

Ces 3 formes sont retrouvées dans les cancers colorectaux dans des proportions distinctes (Figure 2.3). L'instabilité CIN est la forme majoritaire 75-85%, l'instabilité MSI compte pour 15-20%, ces 2 formes étant quasiment exclusives. La forme CIMP, retrouvée dans 20-30% des cas, se retrouve majoritairement associée à MSI ($\tilde{7}5\%$ des tumeurs MSI sont CIMP). Un certain nombre de tumeurs ne présenteraient aucune des 3 instabilités et pourraient donc présenter d'autres formes d'instabilité non encore identifiées (Geigl et al., 2008).

2.2.1 L'instabilité chromosomique, phénotype et mécanismes

Une grande majorité des tumeurs colorectales, comme d'autres cancers, ont un contenu chromosomique anormal, elles sont aneuploïdes. La spécificité du cancer colorectal est l'observation de pertes de matériel chromosomique plus que de gains. Des pertes induisent une perte d'hétérozygotie (on parle alors de LOH pour Loss Of Heterozygosity) suggérant un rôle plus important des gènes suppresseurs que des oncogènes dans la pathologie ou l'implication d'oncogènes soumis à empreinte parentale. La 1ère découverte des pertes chromosomiques dans le cancer colorectal remonte au début des années 1980 (Reichmann et al., 1981; Muleris et al., 1985) (Figure 2.1).

Le Phénotype CIN

L'instabilité chromosomique est une expression mal définie. Elle est très souvent utilisée pour désigner un phénotype défini par la présence dans les cellules tumorales de multiples réarrangements des chromosomes et/ou du nombre de chromosomes. En pratique, ceci correspond à observer de l'aneuploïdie ou de la polyploïdie (Walther et al., 2008).

Selon Geigl et al. (2008) et Rajagopalan et al. (2003) la définition exacte du CIN est le taux avec lequel les chromosomes, ou de larges portions de chromosome, sont gagnés ou perdus. L'état aneuploïde observé dans une image statique du contenu chromosomique d'une tumeur peut ne pas être dû à de l'instabilité mais à des proportions de clones aneuploïdes différents. La notion d'instabilité intracellulaire est donc importante dans la

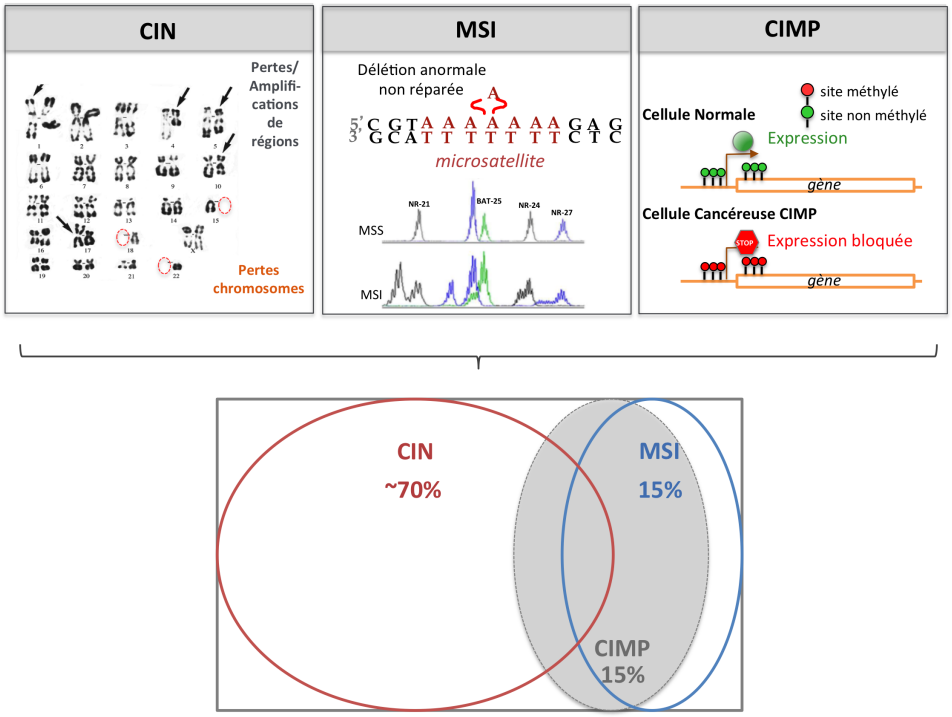


FIGURE 2.3 – Schéma des différentes instabilités et leur répartition dans la population des CRC.

définition. Le problème est que c'est difficilement mesurable. C'est pourquoi la proportion d'altérations du génome est généralement utilisée pour définir une tumeur comme CIN positive ou négative.

Dans le cas du CRC, cette instabilité se traduit par des pertes d'hétérozygotie par la perte de chromosome entier ou de bras entier très fréquemment observée dans les chromosome 1,5,8,17 et 18 comme mentionné dans le modèle de (Fearon and Vogelstein, 1990).

L'instabilité CIN est retrouvée dans les adénomes suggérant son importance pour l'initiation et/ou la maintenance du cancer (Geigl et al., 2008; Pino and Chung, 2010).

Les approches pour évaluer le CIN sont diverses incluant la cytométrie, le caryotype, l'analyse de perte d'hétérozygotie, l'hybridation fluorescente *in situ* et, plus récemment, l'hybridation génomique comparative, CGH (pour Comparative Genomic Hybridization) (Pino and Chung, 2010), les SNP arrays et le séquençage de l'ADN.

Mécanismes induisant du CIN

Trois mécanismes sont connus pour induire ce type d'instabilité (Pino and Chung, 2010) :

- un défaut de la ségrégation chromosomique lors de la division cellulaire avec une distribution des chromosomes déséquilibrée dans les cellules filles
- un défaut de stabilité des télomères
- une déficience du système de réponse aux dommages de l'ADN

Mais les causes de ces défauts eux-mêmes ne sont pas encore élucidées.

La mutation d'*APC* a été proposée comme un potentiel initiateur du CIN de par son rôle dans la régulation du cytosquelette pouvant mener à un défaut de ségrégation. Les gènes *AURKA*, *BUB1*, *BUBR1*, *CENPA*, *PLK1* et *PTTG* ont été impliqués comme régulant la ségrégation chromosomique. Les gènes *TP53*, *MRE11*, *BRCA*, *ATM/ATR* participent à la réponse aux altérations de l'ADN et *TERC* à la régulation des télomères.

Les tumeurs considérées CIN positives sont plus fréquemment mutées dans des gènes suppresseurs de tumeur comme *APC*, *TP53*, *SMAD4*, *SMAD2* et *DCC* ou dans des oncogènes comme *KRAS*, *CTNNB1* et *PIK3CA* (Pino and Chung, 2010).

Bénéfices possibles de CIN

CIN peut conférer des avantages sélectifs aux les cellules tumorales. Notamment, il peut augmenter la probabilité d'inactiver des gènes suppresseurs de tumeurs, et ainsi donner des avantages pour la croissance, la prolifération ou l'évitement de la mort cellulaire en changeant massivement l'expression des gènes (toutefois des mécanismes de compensation comme le dosage génique et la réduplication du chromosome perdu peuvent contrer cet effet) et enfin conférer aux cellules un mécanisme adaptatif aux changements d'environnements efficace, telle que l'adaptation aux traitements (Rajagopalan et al., 2003).

Impact clinique du CIN

L'étude des altérations chromosomiques n'a pas fait ressortir d'impact pronostique robuste spécifique. Par contre, le phénotype CIN pris dans sa globalité est associé à un pronostic défavorable (Pino and Chung, 2010). Cet impact pronostique a été présumé être dû à la contre-sélection du phénotype MSI, connu pour avoir un pronostic plus favorable. Cependant une méta-analyse de 63 études ayant évalué le CIN a confirmé qu'il est associé à un plus mauvais pronostic en tenant compte du statut MSI et qu'il devrait être évalué comme un marqueur pronostique indépendant du statut MSI (Walther et al., 2008).

2.2.2 L'instabilité des microsatellites

Près de 15% des tumeurs CRC ne présentent pas d'instabilité chromosomique, elles ont un contenu chromosomique diploïde ou presque.

Découverte du phénotype

En 1993, suite à la publication de Fearon and Vogelstein (1990), en recherchant de nouveaux gènes suppresseurs de tumeurs, trois équipes, par des approches très différentes, ont montré qu'une partie des CRC présentaient un raccourcissement au niveau de microsatellites, séquences d'ADN constituées d'une répétition de motifs composés de 1 à 6 nucléotides et dispersées dans tout le génome. Ionov et al. (1993) cherchaient des régions de délétion par PCR avec des amorces aléatoires en comparant la tumeur à son tissu adjacent. Au lieu d'observer sur gel l'absence de bandes, ils observèrent des bandes migrant plus loin. L'analyse approfondie de ces bandes a montré que les séquences contenaient des séquences répétées simples. Thibodeau et al. (1993), en recherchant des gènes suppresseurs de tumeurs par analyses des microsatellites dinucléotides, utiles pour la détection de LOH, sur les régions 5q, 15q, 17p et 18q, observèrent aussi des délétions au niveau des microsatellites. Ils utilisèrent le terme de MIN pour Microsatellite INstability, que l'on rencontre dans certains articles. Ils montrèrent également que ces tumeurs étaient majoritairement dans le côlon proximal et que les patients avaient un meilleur pronostic. La découverte des deux équipes représentait une voie inédite pour le développement de

tumeurs qui n'impliquait pas la perte d'hétérozygotie. Et enfin, Aaltonen et al. (1993) et Peltomäki et al. (1993) ont découvert une association des tumeurs héréditaires liées au syndrome de Lynch avec un marqueur du chromosome 2p. Ils utilisèrent ce marqueur pour vérifier que le deuxième "hit" [2] seraient une LOH à ce locus mais trouvèrent finalement eux aussi de l'instabilité microsatellitaire à ce locus, ce phénomène étant observé dans la grande majorité des tumeurs héréditaires. Au final, il a été établi que 15% des tumeurs colorectales présentaient ce mécanisme inédit de pathogénicité et qu'il correspondrait à la majorité des formes héréditaires répertoriées. (Boland and Goel, 2010).

Origine

Très rapidement après ces découvertes, ce phénotype a été associé à une perte fonctionnelle du système de réparation des mésappariements de l'ADN (MMR, pour Mismatch Repair) (Fishel et al., 1993; Leach et al., 1993). Le système MMR a pour fonction de scanner l'ADN néosynthétisé et de réparer les erreurs d'appariement survenues lors de la réplication, incluant les mésappariements de bases (incorporation du mauvais nucléotide) et les courtes insertions/délétions (1-16 nucléotides). Son action est d'autant plus importante dans les régions répétées où la probabilité d'erreur en raison de "glissements" (*stuttering* ou *strand slippage*) des ADN polymérases est augmentée par la nature même de la séquence. Les protéines impliquées dans le système MMR sont : MSH2, MSH3, MSH6, MLH1, MLH3, PMS1 et PMS2. MSH2 et MSH6 forment le complexe MutSα capable de reconnaître les simples mésappariements et les petites boucles liées à l'insertion ou la délétion de quelques nucléotides (généralement 1 ou 2 nucléotides) (Figure 2.4). MSH2 et MSH3 forment le complexe MutSβ capable de reconnaître les grandes boucles d'insertions ou de délétions (de 2 à 16 nucléotides). Après l'étape de reconnaissance, MutSα ou MutSβ recrutent le complexe MutLα , l'hétérodimère formé par MLH1 et PMS2. MutLα forme alors un anneau coulissant qui scanne l'ADN à la recherche d'entailles ou de sites hypométhylés caractérisant le brin néosynthétisé. La partie du brin néo-synthétisé est alors excisée, l'ADN polymérase de nouveau recrutée, et un nouveau brin est alors synthétisé. (Boland and Goel, 2010; Hamelin et al., 2008; Weinberg, 2013).

L'acquisition de la déficience du système MMR dans ces cancers reposent sur 2 voies principales : l'inactivation par mutation et l'inactivation par méthylation des gènes du MMR. Les formes héréditaires passent par une mutation dans un allèle des gènes des protéines du MMR et un second "hit" au cours de la vie dans le même gène va alors inactiver le système MMR. Le syndrome de Lynch est associé à des mutations germinales des gènes *MLH1* ou *MSH2*, moins fréquemment dans *MSH6* et très rarement dans *PMS2*. Les formes sporadiques passent par l'inactivation du gène *MLH1* par méthylation biallélique majoritairement. Cette inactivation par méthylation surviendrait comme une conséquence du phénotype CIMP (voir section2.2.3).

Une partie des tumeurs MSI ne présentent toutefois pas d'anomalie génétique ou épigénétique dans les gènes MMR ce qui pourrait impliquer d'autres types d'événements épigénétiques comme les microARN ou la méthylation d'histones. Récemment, la mutation du gène *SETD2* codant pour une triméthyltransferase des histones, a été décrit comme engendrant un phénotype MSI (Yamamoto and Imai, 2015).

2. Selon le modèle de Knudson, un deuxième événement est nécessaire pour déclencher le processus de transformation cellulaire appelé deuxième "hit".

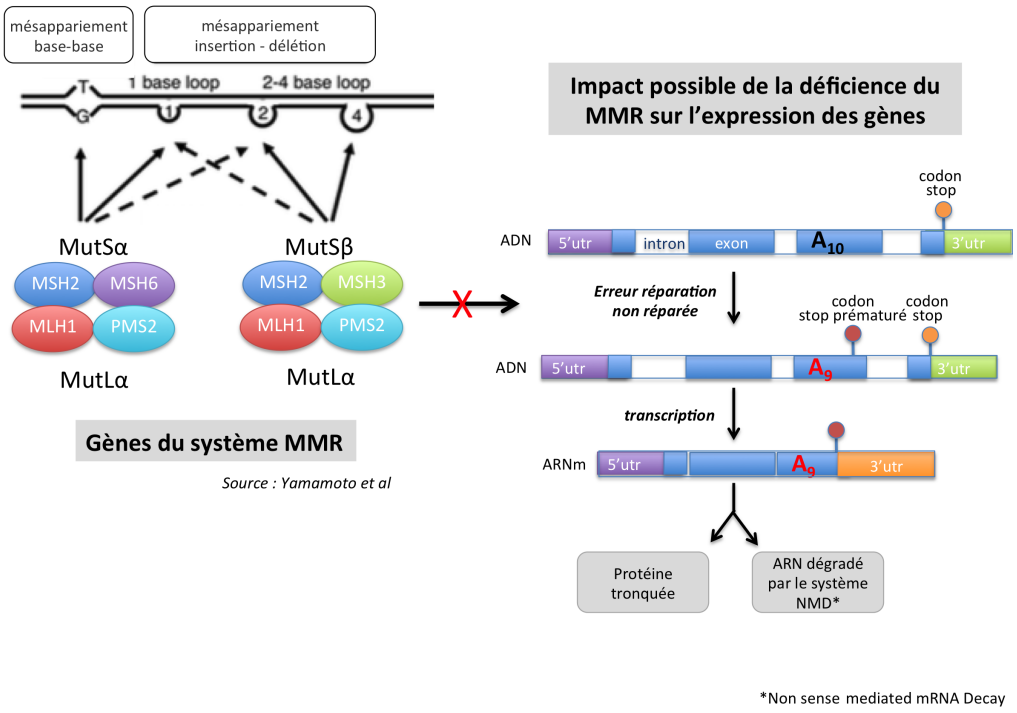


Figure 2.4 – Système de réparation des mésappariements de l'ADN et l'impact de sa défaillance dans les tumeurs de type MSI.

Le phénotype MSI : l'instabilité microsatellitaire et nucléotidique

Le phénotype MSI correspond donc à une variabilité, le plus souvent à un raccourcissement, de la taille d'un certain nombre de microsatellites. Les microsatellites sont abondants le long du génome, près de 3% du génome, et sont retrouvés aussi dans les régions géniques majoritairement non codantes mais aussi codantes (près de 4000 gènes contiennent un microsatellite mononucléotidique, de taille supérieur à 7, dans leur séquence codante). Ils sont très polymorphiques entre individus mais, au sein d'un même individu, ils sont invariants (Duval and Hamelin, 2002). Une description des méthodes de mesure du statut MSI des tumeurs est fournie en annexe A.

En plus de cette instabilité facilement identifiable, un nombre important de mutations en dehors des régions répétées du génome sont également retrouvées. Ces mutations sont en majorité peu récurrentes au sein des cohortes tumorales. Quelques mutations sont toutefois assez fréquentes comme *APC* (dans une moindre mesure que pour les tumeurs non MSI), *BRAF*, *KRAS*, *CTNNB1* et PIK3CA(Cancer Genome Atlas Network, 2012). Ces mutations pourraient avoir un rôle important dans l'initiation de la tumorigenèse. La mutation activatrice de *BRAF* (un gène qui code pour une serine/thréonine kinase dans la voie de signalisation MAP/ERK) se retrouve très majoritairement dans les cancers MSI et est mutuellement exclusive des mutations KRAS (qu'on retrouve majoritairement dans les cancers MSS (non MSI, MicroSatellite Stable) et très peu dans les tumeurs MSI). La mutation de *BRAF* se retrouve uniquement chez les patients atteints de cancers MSI non héréditaires. Pour le gène *APC*, beaucoup moins fréquemment muté dans les cancers MSI, la grande majorité des cancers MSI possèdent des taux d'expression normaux de la protéine APC. Dans ce cas, une mutation de la *CTNNB1*, la β-caténine la rend incapable d'être re-

connue par APC et a donc les mêmes conséquences qu'une perte de fonction d'*APC* (Vilar and Gruber, 2010; Boland and Goel, 2010; Hamelin et al., 2008). En outre, l'équipe d'Alex Duval a rapporté des mutations fréquentes du facteur de croissance *TCF7L2*, l'effecteur principal de la voie Wnt (Cuilliere-Dartigues et al., 2006). Très récemment a été identifié une 4ème forme d'instabilité. Elle correspond à une hypermutabilité mais uniquement en-dehors des microsatellites. Elle est associée à des mutations dans les gènes *POLE* et *POLD1*, gènes impliqués dans la réparation des erreurs lors de la réplication "proofreading reparation" (Cancer Genome Atlas Network, 2012; Kim et al., 2013; Palles et al., 2013).

Impact du MSI, dérégulation des gènes cibles

Le processus d'instabilité microsatellitaire n'est qu'indirectement oncogénique car c'est l'accumulation de nombreuses mutations au niveau de gènes cibles d'instabilité qui semble être responsable du processus tumoral. En effet comme mentionné plus haut, des microsatellites sont localisés dans les exons de certains gènes. Des insertions ou délétions de nucléotides aux niveaux de ces séquences entraînent un décalage dans le cadre de lecture et donc l'apparition d'une protéine tronquée, le plus souvent non fonctionnelle (Figure 2.4). Si ces protéines participent à des fonctions telles que le contrôle de la prolifération cellulaire, de la réponse aux facteurs de croissance ou de l'apoptose, leur inactivation risque alors de jouer un rôle dans le développement ou la progression tumorale MSI. Le premier gène identifié comme cible de l'instabilité microsatellitaire est le récepteur du TGFβ (*TGFBR2*). Markowitz et al. (1995) ont montré que le gène TGFBR2 ne s'exprimait pas dans les lignées cellulaires et les xénogreffes MSI et que cela était dû à la présence dans sa partie codante d'une séquence microsatellite très fréquemment mutée dans un contexte MSI. TGFβ et son récepteur TGFBR2 agissent dans ce cas comme des suppresseurs de tumeurs, et leur perte de fonction dans un contexte MSI favorise la prolifération des cellules épithéliales. D'autres gènes ont également été montrés comme étant fréquemment la cible de MSI, notamment *ACVR2*, *AXIN2*, *BAX*, *IGF2R*, *TCF7L2* (souvent nommé *TCF4*), MSH6, MLH3, ainsi que d'autres gènes impliqués dans différentes voies de signalisation comme la transduction du signal, l'apoptose et l'inflammation, la régulation de la transcription, la signalisation des dommages de l'ADN et leur réparation (Duval and Hamelin, 2002; Hamelin et al., 2008). Par ailleurs, des séquences répétées dans les introns proches des séquences d'épissage peuvent également être sources de mutations ayant un impact sur l'expression de la protéine, comme cela a été décrit par le laboratoire d'Alex Duval pour le gène codant la protéine chaperonne *HSPH1* (souvent mentionnée HSP110) (Dorard et al., 2011), ou encore les gènes *MRE11* et *ATM*. Toutefois, les cancers MSI générant pléthore de mutations, le défi n'est plus de découvrir de nouveaux gènes cibles mutés dans ces cancers, mais de déterminer ceux qui jouent un rôle significatif au cours de l'initiation et de la progression tumorale.

Impact clinique du MSI

Le phénotype MSI a commencé à être pris en compte en pratique clinique en raison d'associations au pronostic et à la réponse au traitement et en raison de son intérêt pour identifier les syndromes de Lynch. En effet, l'association à la survie observée par Thibodeau et al. (1993) a été confirmée quelques années plus tard sur une large cohorte de 607 patients atteints de CCR MSI et non MSI (Gryfe et al., 2000). Les patients ayant des tumeurs MSI présentaient un meilleur pronostic et étaient moins sujets à former des métastases dans les ganglions lymphatiques ou à distance. Une méta-analyse réalisée par Popat et al. (2005), regroupant les données de 32 études différentes, soit au total 7642 patients, dont 1277

CCR MSI, a confirmé de manière significative l'avantage pronostique de MSI. Quant à l'association à la réponse au traitement, le phénotype MSI a été associé, mais de manière plus controversée, à la réponse au traitement au 5 FU : les MSI de stade II ou III ne répondraient pas au 5-FU, le 5-FU pourrait même être délétère (Ribic et al., 2003; Sargent et al., 2010). Le bénéfice du 5-FU serait en fait dépendant de la forme héréditaire ou sporadique des tumeurs, les syndromes de Lynch présenteraient une meilleure survie avec le 5-FU que sans le 5-FU (Sinicrope and Sargent, 2012). Inversement, les tumeurs de patients atteints de cancer colorectal métastatique sont plus susceptibles de répondre favorablement à un traitement à l'irinotécan (cf. 1.3.3) si leur tumeur est de phénotype MSI (Fallik et al., 2003; Hamelin et al., 2008). Il existe encore très peu de données pour comprendre les raisons qui sont à l'origine du meilleur pronostic des cancers MSI et de leur réponse différente à la chimiothérapie. Dans ce champ de recherche, les travaux de l'équipe d'Alex Duval, auxquels j'ai participé, ont rapporté des premières pistes intéressantes, relativement à la découverte de mutations dues à MSI qui affectent la chaperonne HSP110 dans ces tumeurs de manière dominante négative conférant un bon pronostic aux patients traités par chimiothérapie ou de stade 3 lorsque que la délétion est assez large(Dorard et al., 2011).

Sur le plan anatomopathologique, des caractéristiques morphologiques existent permettant de différentier les tumeurs MSI. Les principales caractéristiques, communes aux formes héréditaires et sporadiques, sont l'infiltration lymphocytaire (en pratique clinique le marqueur le plus sensible du phénotype MSI), la sécrétion de mucus et la faible différentiation. Au sein des MSI, des différences entre les formes héréditaires et sporadiques sont également retrouvées, en adéquation avec des évidences convergentes vers deux voies de carcinogenèses distinctes pour ces deux formes. L'infiltration lymphocytaire, la dédifférentiation, et la coexistence d'adénomes sont plus marquées dans les syndromes de Lynch, tandis que la sécrétion de mucus, la faible différentiation, l'hétérogénéité tumorale, la morphologie festonnée et la coexistence de polypes festonnés sont plus évidents dans les formes sporadiques (Jass, 2004).

2.2.3 L'hyperméthylation de l'ADN

Alors que la majorité de l'attention s'était portée sur les événements génétiques, ces 2 dernières décennies se sont focalisées sur les événements n'impliquant pas de changements de la séquence d'ADN, le domaine de l'épigénétique. L'épigénétique recouvre la méthylation de cytosines de l'ADN qui va nous intéresser ci-après, mais aussi des changements structuraux des histones ou de la chromatine et des altérations dans l'expression des ARN non codants, comme les microARN ou les ARN longs non codants.

Découverte

En 1999, en recherchant des gènes suppresseurs de tumeurs inactivés par méthylation, Toyota et al. (1999) a mis en évidence une nouvelle forme d'instabilité passant par la méthylation de plusieurs gènes spécifiquement dans les tumeurs, qu'ils nommèrent "CpG island methylator phenotype" (CIMP) dans un sous-groupe de CRC. En effet, ils ont montré qu'une majorité des événements de méthylation seraient dus à l'âge, et pourraient être associés à l'initiation de la carcinogenèse, et une plus faible partie seraient spécifiques aux tumeurs et retrouvés seulement dans un sous-groupe de tumeurs. Ils ont proposé une nouvelle voie de tumorigenèse passant par l'inactivation simultanée par méthylation de plusieurs gènes suppresseurs de tumeurs et apparaissant être responsable de la majorité des tumeurs sporadiques MSI par la survenue de la méthylation de MLH1. Mais cette voie

ne se réduit pas aux MSI car près de la moitié des tumeurs CIMP ne sont pas MSI et ne présentent pas de méthylation de MLH1. Le CIMP étant retrouvé dans les adénomes, il a été suggéré que ce serait un événement précoce dans la carcinogenèse.

Le Phénotype CIMP

En raison de controverses et de résistances, il fallut attendre en 2005 la publication de Samowitz et al. (2005) sur la validation du phénotype CIMP sur une large cohorte et l'article conjoint "CIMP at last" (Issa et al., 2005) pour que l'existence du phénotype CIMP soit définitivement reconnu. Aujourd'hui, le phénotype CIMP est défini comme une vaste hyperméthylation des sites CpG au niveau des promoteurs résultant dans l'inactivation d'un certain nombre de gènes (Curtin et al., 2011). Mais comme le phénotype CIN, sa définition reste assez floue (Hughes et al., 2013). Pour sa caractérisation, il n'y a donc pas encore de consensus. Plusieurs panels de marqueurs ont été proposés (cf. Annexe A (page 118)), le plus utilisé en pratique étant le panel proposé par Weisenberger et al. (2006). Ces panels donnent des résultats légèrement différents. Toutefois, l'association à la mutation de BRAF est observée quelque soit le panel. Les différences vont reposer sur la définition d'un sous-groupe CIMP-low, ayant des événements de méthylation, mais dans une plus faible mesure que les CIMP, qui serait lui associé à la mutation de KRAS (Curtin et al., 2011).

Le phénotype CIMP n'est pas spécifique des CRC, il a été décrit dans divers types de cancers, les mieux étudiés étant les glioblastomes et les leucémies. Il semble toutefois présenter des spécificités en fonction du type de cancer notamment dans les gènes ciblés. (Hughes et al., 2013).

Mécanismes induisant le phénotype CIMP

L'une des raisons à la difficulté de définir le CIMP est la difficulté à déterminer les mécanismes responsables de sa survenue. Toyota et al. (1999) proposait 2 mécanismes possibles : soit une défaillance au niveau de la méthylation *de novo* avec une mutation dans une DNA-méthyltransférase, soit une perte de protection contre la méthylation *de novo* par la perte de facteurs trans-activateurs.

Pour les gliomes et les leucémies, il a été démontré que *IDH1* et les gènes affectant la même voie métabolique, comme *IDH2* et *TET2* étaient impliqués de manière causale.

Pour les cancers colorectaux, l'événement causal n'a pas encore été identifié mais ne semble pas a priori faire intervenir cette même voie. *BRAF* semble jouer un rôle primordial mais il n'a pas pu être mis en évidence son implication causale (Curtin et al., 2011). Très récemment, 2 articles ont permis de mieux comprendre la séquence pouvant aboutir à ce phénotype avec d'un côté, pour les tumeurs mutées *BRAF*, l'intervention du facteur de transcription MAFG et de la DNA-méthyltransférase DNTM3B nécessaires à l'extinction de *MLH1* (Fang et al., 2014), et d'autre part, pour les tumeurs mutées *KRAS*, l'intervention de ZNF304 et de la DNA-méthyltransférase DNMT1 dans l'extinction de INK4-ARF (Serra et al., 2014).

Impact clinique du CIMP

Les tumeurs ayant un phénotype CIMP présentent des caractéristiques cliniques particulières, comme une localisation plus proximale des tumeurs. Ce phénotype est associé à la mutation BRAF. Il est également observé dans les polypes festonnés de type "sessile serrated polyps". Ces associations sont indépendantes du statut MSI bien que la plupart des

MSI sporadiques soient CIMP (le phénotype CIMP est très peu retrouvé dans les formes héréditaires) (Curtin et al., 2011). Le phénotype CIMP a été associé à un bon ou mauvais pronostic en fonction du type de cancer et du type de tumeurs. Il a été plutôt associé à un pronostic favorable dans les CRC et les gliomes. Toutefois, des études cas-témoins ont reporté un mauvais pronostic de CIMP associé à un phénotype MSS, mais cela pourrait être un effet confondant de la mutation de BRAF. (Hughes et al., 2013)

2.2.4 Évolution du modèle de tumorigenèse à la lumière des instabilités

La découverte de ces différentes formes d'instabilité du génome amène à réévaluer le modèle de Fearon et Vogelstein, lequel apparaît plus spécifique de l'initiation et de la progression des tumeurs présentant de l'instabilité chromosomique. Plusieurs études ont permis de mettre à jour le modèle de Fearon avec les connaissances des instabilités (Pino and Chung, 2010; Vilar and Gruber, 2010; Walther et al., 2009). Il n'existe encore aucun modèle de tumorigenèse représentant le CIMP non MSI. La Figure 2.5 présente une synthèse des modèles de tumorigenèse colique en fonction des instabilités. Ces formes d'instabilité sont bien retrouvées assez tôt dans la carcinogenèse, ce qui serait en faveur de leur participation au mécanisme de l'oncogenèse.

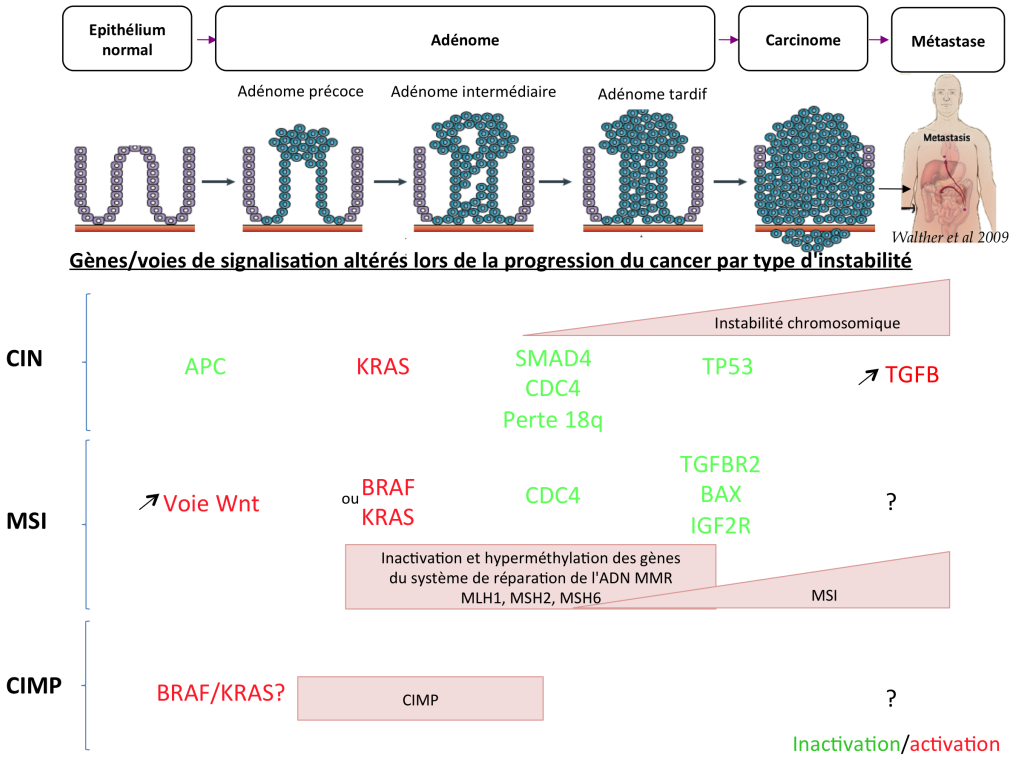


Figure 2.5 – Modèles de tumorigenèse colique en fonction des instabilités. Gènes et voies de signalisation impliqués dans la progression tumorale définis à partir de Walther et al. (2009); Markowitz and Bertagnolli (2009); Vilar and Gruber (2010); Yamamoto and Imai (2015).

2.3 Les Classifications moléculaires proposées

A la lumière de toutes ces découvertes, il était important d'intégrer cette information afin d'en évaluer l'intérêt clinique et de mieux caractériser la diversité des CRC. En effet, les classes moléculaires sont importantes car elles sont censées refléter des scénarios oncogéniques distincts.

Jass (2007) proposa une classification très fine et très détaillée mêlant les aspects cliniques, morphologiques et moléculaires, en se basant sur le phénotype MSI et le phénotype CIMP, CIN étant utilisé comme une annotation, à partir des données de la littérature et de ses connaissances d'anatomo-pathologiste. Il définit 5 sous-types (Figure 2.6), 2 MSI , (CIMP-/MSI) et (CIMP+/MSI/BRAF muté) pour les héréditaires Lynch, et 3 MSS (CIMP+/MSI-low ou MSS/BRAF muté), (CIMP-low/MSS ou MSI-low/KRAS muté) et (CIMP-/MSS). Notamment, il évalua le lien entre ces sous-types et la lésion précancéreuse. Les MSI non héréditaires et les (CIMP+/MSI-low ou MSS/BRAF muté) proviendraient de polypes festonnés alors que les (CIMP-/MSS) et les MSI héréditaires proviendraient de polypes conventionnels adénomateux. Les (CIMP-low/MSS ou MSI-low/KRAS muté) peuvent provenir des 2 types de lésions.

Shen et al. (2007) identifièrent 3 sous-types distincts à partir d'une classification hiérarchique intégrée de facteurs génétiques (MSI et les mutations de *KRAS*, *BRAF*, *TP53*) et épigénétiques (méthylation de 27 régions promotrices) : CIMP1, CIMP2, and CIMP-negative. CIMP1 était caractérisé par l'instabilité MSI, des mutations de *BRAF* et de rares mutations de KRAS. CIMP2 était associé à des mutations de KRAS et rarement à MSI et aux mutations de *BRAF* et *TP53*. Et CIMP-negative avait une forte fréquence de mutations de *TP53* et peu de MSI, de mutations de *BRAF* ou de mutations de *KRAS*.

L'année suivante Ogino and Goel (2008) proposèrent une classification moléculaire basée sur les corrélations aux phénotypes d'instabilité. Ils proposèrent 6 sous-types à partir de CIMP et MSI (le CIN n'ayant pas été utilisé en raison du manque de mesure standard). Neuf sous-types étant possibles, ils regroupèrent ceux ayant des caractéristiques communes : (MSI/CIMP+), (MSI/CIMP-low ou CIMP-), (MSI-low/CIMP-low), (MSS/CIMP-low), (MSS ou MSI-low/CIMP+) et (MSS ou MSI-low/CIMP-).

Les CRC sur le plan moléculaire sont hétérogènes en terme de mutations et en terme de mécanismes d'instabilité génomique observés. Les classifications proposées à partir de ces marqueurs moléculaires sont intéressantes mais ne reposent que sur l'analyse de quelques marqueurs. Celle de Jass (2007), se basant en plus sur les caractéristiques anatomopathologiques, est toutefois très intéressante car proche de la biologie des tumeurs sous-jacentes et révèle qu'il y a une vraie hétérogénéité de la maladie contrairement à ce qu'a pu laisser croire le modèle de Fearon et Vogelstein. Avec le séquençage du génome humain, ces 15 dernières années ont vu émerger les technologies à haut-débit permettant d'interroger la totalité des gènes en terme d'expression et d'altérations. Cette révolution a ouvert de nouvelles perspectives quant à la caractérisation et à la classification des tumeurs.

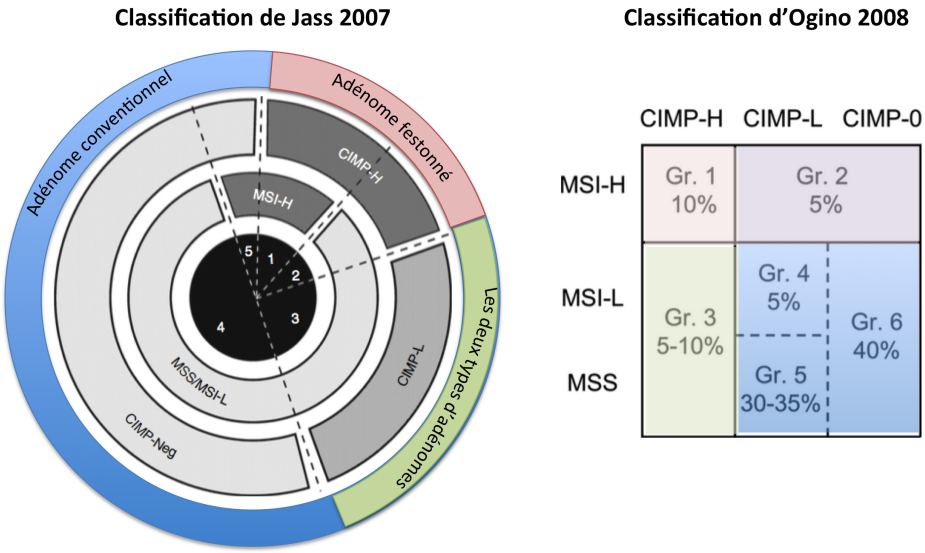


Figure 2.6 – Les classifications moléculaires proposées intégrant les instabilités génomiques

Chapitre 3

Le cancer colorectal à l'ère de la génomique : l'apport des technologies à haut débit

La recherche en biologie est dans une période de révolution. Ces 15 dernières années, le séquençage du génome humain, conjointement au développement des nouvelles technologies de puces à ADN, ou *microarrays*, ont permis l'étude de la cellule à l'échelle du génome entier. Initialement développées pour l'étude de l'expression de l'ensemble des ARNm (transcriptome), ces technologies se sont vite déclinées pour étudier l'ensemble des omiques possibles (génome, méthylome, mirnome, ...). Par ailleurs, depuis quelques années, on assiste de nouveau à une avancée majeure dans ces technologies avec l'émergence du séquençage à haut débit de nouvelle génération. De très nombreuses publications utilisant ces technologies sont déjà parues, en particulier en cancérologie. Ceci a permis de générer énormément d'informations, mais aussi beaucoup d'inconsistances. Dans le cas du cancer colorectal, beaucoup d'efforts ont été fournis pour trouver des marqueurs pronostiques et prédictifs de la réponse aux traitements et pour caractériser la séquence adénome-carcinome-métastase ainsi que les différences inter-carcinomes.

3.1 Étude des tumeurs par les technologies à haut débit

3.1.1 Émergence et évolution de la génomique

Selon le *Oxford English Dictionary*, le terme de "génome" a été pour la première fois employé en 1920 par un botaniste, Hans Winkler, afin de définir l'ensemble des chromosomes pour distinguer l'origine parentale des chromosomes de formes hybrides (Lederberg, 2001). Le terme de génomique a été lui utilisé pour la 1ère fois en 1944 par Victor McKusick et Frank Ruddle comme nom de leur journal. Le terme génomique est aujourd'hui souvent employé dans un sens large pour désigner l'étude moléculaire à l'échelle du génome entier par les technologies à haut-débit. Le suffixe -ome est utilisé pour tout ce qui a trait à la génomique pour sa signification l'"ensemble de".

La publication de la séquence du génome humain en avril 2003 a constitué un tournant important dans le développement de la génomique. C'est grâce à la technique de séquençage développée par Sanger en 1977, et aux améliorations apportées par la suite, que le *Human Genome Project* a pu commencer à être envisagé puis lancé en 1990 (Collins et al., 2003) (Figures 3.1 et 2.1).

Quelques années auparavant étaient développées les 1ères puces à ADN. Cette avancée

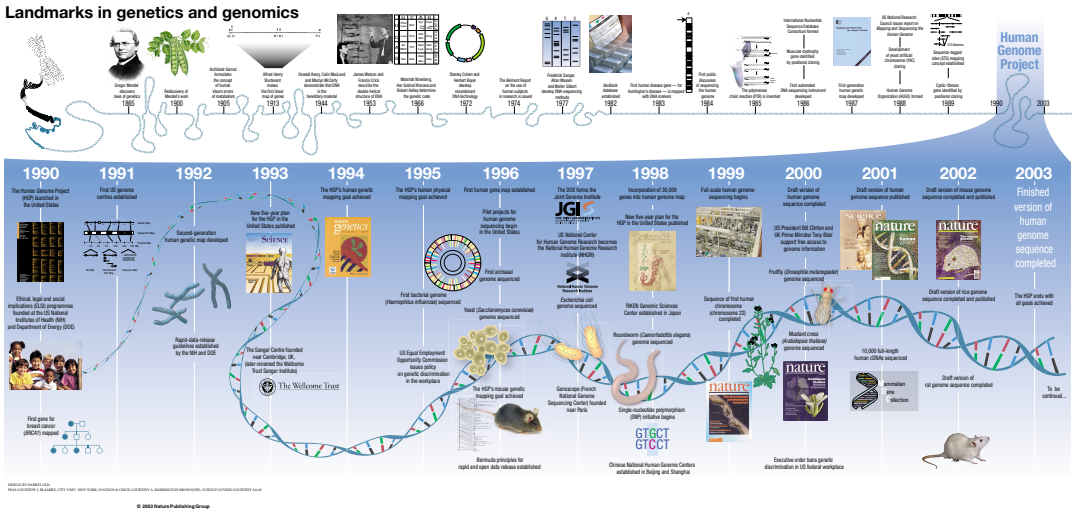


Figure 3.1 – Les principales avancées de la génétique et de la génomique selon Collins et al. (2003) *(zoomer dans le fichier pdf pour visualiser).*

a pu voir le jour grâce au développement de la technique de Southern blot et au développement de banques de séquences sur des plaques de titration (Lander, 1999). La 1ère puce à ADN fut proposée par Pat Brown et son équipe. Elle était composée de 45 sondes d'Arabidopsis thaliana, déposées par un robot à l'aide d'aiguilles sur une lame de verre (Schena et al., 1995). Ils annonçaient, qu'avec la densité de dépôt des robots, près de 20,000 sondes pouvaient être déposées, ou *spottées* (en raison de la forme ronde du dépôt), sur une lame de verre (cf. Figure 3.2 pour le principe de fixation des sondes). En parallèle, Stephen Fodor et ses collègues adaptèrent la technique de photolithographie pour synthétiser in situ, à très haute densité, des sondes oligonucléotidiques, ce qui donnera naissance à la première puce Affymetrix (Lipshutz et al., 1999). D'autres compagnies, comme Agilent, ont également développé une approche par synthèse in situ en déposant par goutte un à un les nucléotides (Figure 3.2). Avec le séquençage du génome humain, les sondes pouvaient alors cibler l'ensemble des gènes.

En 2005, une autre révolution s'amorça, le développement d'une nouvelle génération de séquençage, ou *next-generation sequencing* (NGS), donnant la possibilité de séquencer des génomes individuels en peu de temps et à un moindre coût (Meyerson et al., 2010). On peut maintenant l'utiliser pour séquencer tous les omiques ADN et ARN, et ainsi avoir l'information de séquences ou d'expression la plus exhaustive possible avec l'accès aux mutations et variants.

Jusqu'à très récemment, les microarrays étaient considérées comme le *gold standard*, en particulier pour l'étude du transcriptome. Toutefois on commence à assister à un revirement, notamment en raison de l'amélioration des approches d'analyse de données de séquençage, la balance semble pencher aujourd'hui vers le séquençage (Wang et al., 2009; Robinson et al., 2015).

3.1.2 Principe des omiques

Les technologies omiques permettent d'étudier l'ensemble du génome à différents niveaux biologiques :

— au niveau des altérations chromosomiques (variations du nombre de copies de ré-

gions ou de chromosomes entiers).
— au niveau de la séquence ADN : le séquençage du génome entier ("whole genome") apporte à la fois une information de séquence et de structure. L'exome étudie principalement la séquence des parties codantes (exons et bornes introniques).
— au niveau des transcrits des gènes : le transcriptome permet d'obtenir en une seule fois les niveaux d'expression de l'ensemble des gènes exprimés dans un tissu.
— au niveau des protéines : le protéome est l'ensemble des protéines exprimées dans un tissu

L'étude du métabolome, c'est-à-dire l'ensemble des métabolites [1] contenus dans la cellules, commence également à émerger.

Par extension, l'épigénomique correspond à l'analyse de marques épigénétiques sur l'ensemble du génome. L'épigénétique regroupe toutes les modifications pouvant être transmises, indépendamment de modifications de la séquence primaire de l'ADN : la méthylation de l'ADN, les modifications des histones et l'expression des miRNA/lncRNA :
— le méthylome étudie le niveau de méthylation de l'ADN au niveau de sites CpG répartis le long du génome.
— les modifications des histones peuvent être étudiées par immunoprécipitation de la chromatine couplée avec une puce ADN (ChIP-on-chip).
— le mirnome étudie le niveau d'expression des microARN, qui sont des petits ARN non codants capables de réguler l'expression d'un grand nombre de gènes en se fixant sur les ARNm.

Principe général

Les microarrays Toutes les technologies de type microarrays reposent sur le même principe. L'hybridation de 2 séquences complémentaires, la cible et la sonde, pour former un ADN double brin (Figure 3.2). La cible est la molécule que l'on cherche à mesurer, et la sonde, la séquence représentative de la séquence que l'on cherche à mesurer. Dans ce cas, ce sont les cibles qui sont marquées, contrairement aux expériences de Southern ou Northern blot où le marquage se fait sur la sonde.

Les sondes sont des séquences nucléotidiques d'ADN, généralement fixées ou synthétisées in situ sur un support solide (lame de verre, plastique ou silicone en fonction de la technologie) [2]. En fonction du support et du type de puces, les séquences vont être plus ou moins longues, de 25 à 70 nucléotides pour les puces de type oligoarray et jusqu'à des longueurs variables pour les puces "*spottées*".

Les cibles peuvent être des séquences d'ADN ou des ARN (transformés en ADNc). Elles sont marquées puis déposées sur le support en condition optimale pour qu'elles s'accrochent spécifiquement à leurs sondes (bain avec agitation, mise à température TM, ...). Les quantités relatives des cibles sont alors mesurées par un scanner mesurant l'intensité du marqueur. Les quantités sont relatives car les cibles n'accrochent pas la même quantité de marqueurs, c'est dépendant du contenu de la base marquée dans la séquence cible. Toutefois, les intensités de signal mesurées sont proportionnelles à la quantité de transcrits présents dans l'échantillon. Certaines approches utilisent la compétition de l'hybridation entre 2 types de cibles différentes (une condition d'intérêt vs une référence, ou directement 2 conditions différentes) ce qui requiert la mesure de 2 signaux pour une même sonde. Les principales plateformes que sont Affymetrix, Illumina, Agilent, Nimblgene diffèrent par le

1. Petites molécules organiques intermédiaires ou issues du métabolisme.
2. La technologie proposée par Illumina fixe les sondes sur des billes qui sont alors dispersées et donc réparties aléatoirement sur la surface. Cela a l'intérêt de minimiser l'effet de biais spaciaux sur les puces qui peuvent se produire préférentiellement sur certaines localisations du support

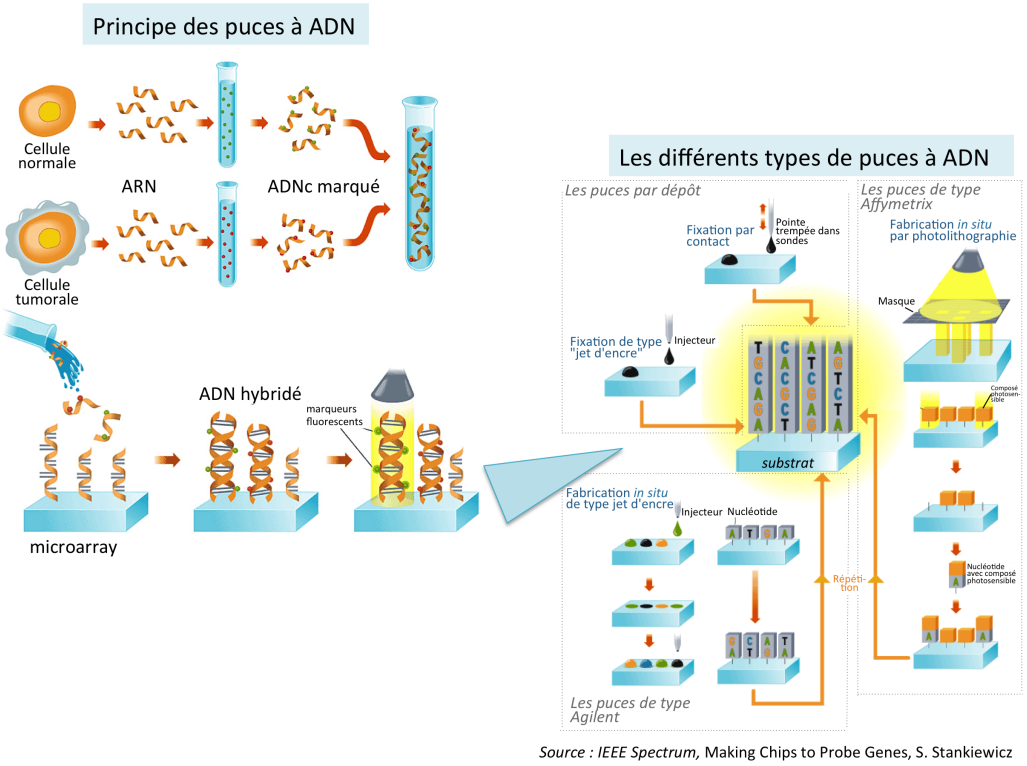


Figure 3.2 – Le principe des puces à ADN et les différentes types de puces.

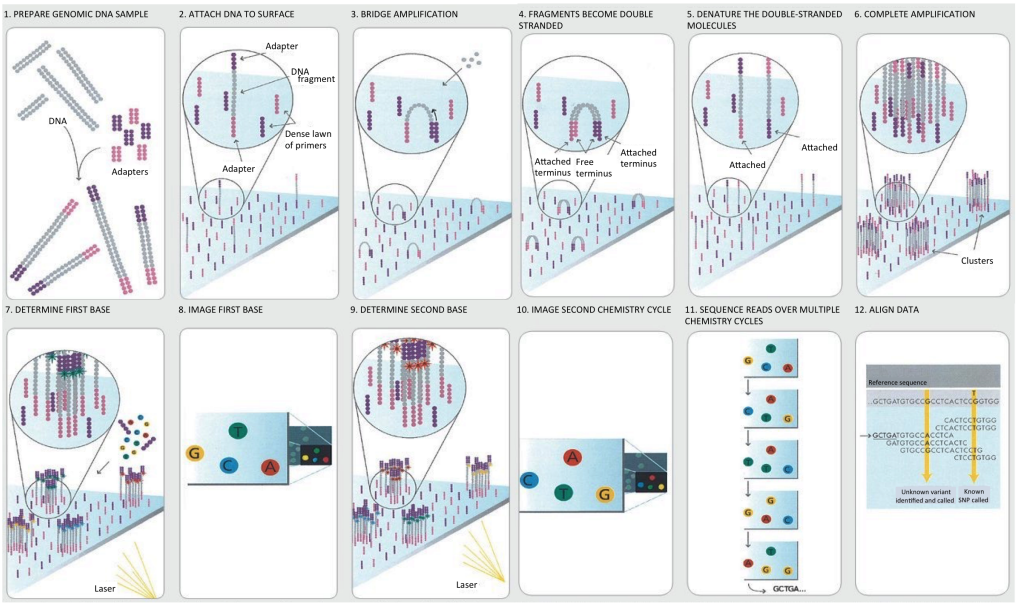
type de sondes, le type de synthèse des sondes, le support et le type de marquage des cibles.

Le séquençage Pour le séquençage, le principe est de fragmenter les cibles, de déterminer l'enchaînement des bases de la séquence sur une longueur fixée (de 75 à 100 bases en général) puis de faire un assemblage des séquences (par l'alignement sur la séquence connue du génome humain) (Figure 3.3). Les valeurs obtenues sont des comptages des séquences alignées (*reads*) à une localisation donnée. Le séquençage peut être réalisé sur l'ADN, soit à l'échelle du génome entier, soit à l'échelle des exons uniquement, ou sur les ARN. Il a l'avantage de pouvoir identifier différentes altérations génomiques par une seule technique (substitutions nucléotidiques, réarrangements, anomalies du nombre de copies) et l'ensemble des variants issus de mutations, de polymorphismes ou d'événements d'édition pour les ARN.

Une fois les puces scannées ou les cibles séquencées et alignées, commence alors l'analyse des données. Une première étape cruciale de l'analyse est le prétraitement pour filtrer et normaliser les données afin qu'elles soient de qualité suffisante et comparables entre elles. Beaucoup d'articles ont été écrits au début des années 2000 pour définir des méthodes de prétraitement appropriées pour faire face aux biais techniques spécifiques de chaque technologie. Aujourd'hui, pour les puces, on est arrivé à peu près à certains consensus. Pour le séquençage, par contre, il y a encore beaucoup de littérature sur le sujet.

Étude du transcriptome

Plus spécifiquement, pour l'étude du transcriptome, une des technologies qui est très largement utilisée est celle proposée par Affymetrix, des puces à oligonucléotides synthé-



Source : Illumina

Figure 3.3 – Principe du séquencage de nouvelle génération. Exemple du principe de séquençage d'Illumina.

tisés in situ par photolithographie (Figure 3.4). Une des puces ayant été très utilisée est la puce Affymetrix GeneChip Human Genome U133 Plus 2.0 Array, dont les sondes ont été élaborées pour mesurer plus spécifiquement les parties 3' des ARNm, moins sujettes à la dégradation que la partie 5' des ARNm. Elle permet la mesure de 54 635 sondes, soit près de 47 000 transcrits et 20 000 gènes. Pour chaque cible, entre 11 et 16 paires de 25-mers (on parle de *probe sets*) sont mesurées et combinées en une seule valeur. Une paire correspond à une séquence "Perfect Match" (PM) représentant la séquence exacte et une séquence "Mismatch" (MM) avec un changeant de la base centrale. Les MM avaient été prévus initialement pour avoir une mesure du bruit de fond et donc corriger les valeurs des sondes PM. Toutefois, en pratique, les MM accrocheraient d'autres séquences que le PM et donc introduiraient plus de bruit qu'elles n'en filtreraient. La méthode classique de normalisation des données RMA s'est affranchie de leur utilisation.

Étude du génome

Pour l'étude des altérations du génome, les toutes premières puces utilisées ont été les puces CGH (Comparative genomic hybridization), ou *"CGH array"* en anglais. Le principe de la puce CGH est le même que celui de la CGH classique, mais l'hybridation se fait sur une puce à ADN et non plus sur les chromosomes directement. Des fragments d'ADN, de taille, de séquence et de position génomique connues, couvrant le génome, vont constituer les sondes fixées sur la puce. La couverture du génome et la résolution dépendent du nombre de sondes présentes sur la puce et de leur taille. Deux types principaux de puces CGH existent :

— les puces BACs (bacterial artificial chromosome) / PACs (P1 artificial chromosome) (la taille des sondes est assez longue, de 100 à 200 kb et la résolution est de 1 Mb).
— les puces à oligonucléotides (la taille des sondes est de 20 à 70 nucléotides et le niveau de résolution est de 20 à 30 kb sur la totalité du génome, avec une densification

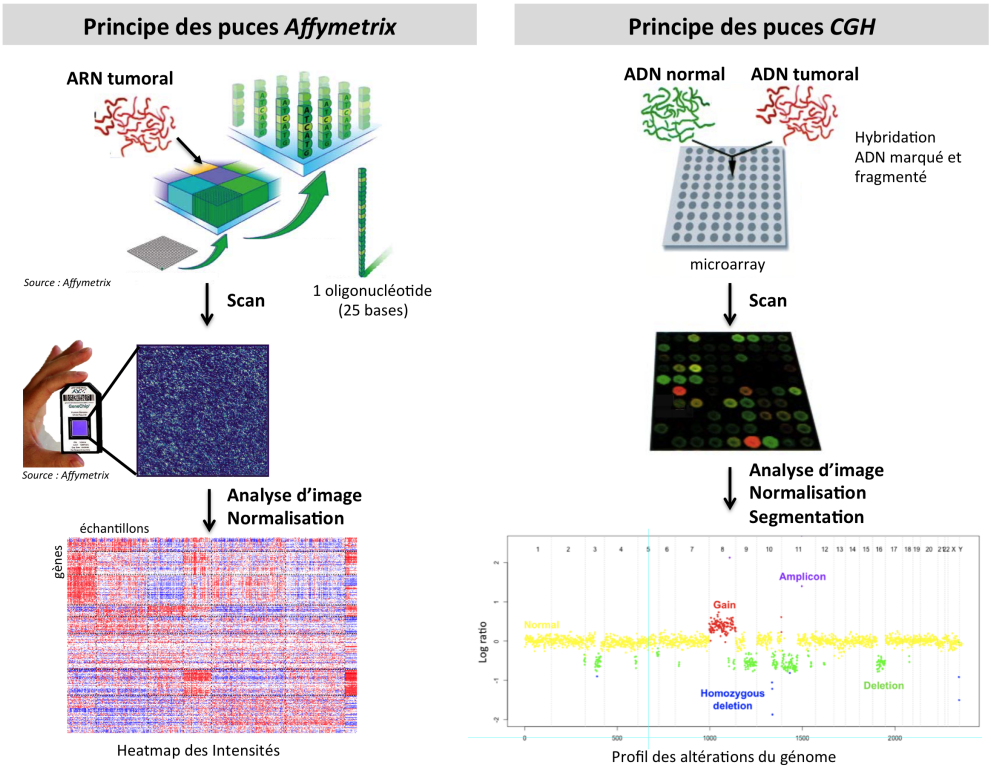


Figure 3.4 – Principe des puces Affymetrix et CGH.

accrue du nombre de sondes dans les régions codantes du génome).

Dans ce cas, on hybride compétitivement l'ADN extrait de l'échantillon tumoral avec l'ADN d'une référence, soit l'échantillon apparié non tumoral, soit plus classiquement une référence commerciale constituée d'un pool d'ADN de plusieurs individus (Figure 3.4). Un gain va alors être détecté par une surreprésentation du marqueur de la tumeur par rapport au marqueur de l'échantillon normal sur les clones couvrant la région remaniée. A l'inverse, une perte sera détectée par une surreprésentation du marqueur de l'échantillon normal.

Les puces CGH ne sont quasiment plus utilisées aujourd'hui. Elles ont fait place aux puces SNP qui permettent de mesurer de manière indirecte les altérations du génome. Les SNP (Single Nucleotide Polymorphism) sont des variations constitutives inter-individus d'une base à un locus donné de l'ADN, présentes avec une fréquence supérieure à 1-2% dans la population. Les sondes des puces SNP mesurent le génotype à ce locus, en général seulement 2 bases sont possibles. L'avantage de ces puces est qu'elles apportent à la fois une information sur le nombre de copies et sur le génotypage de la tumeur. Elles permettent donc de détecter les pertes d'hétérozygotie par isodisomie (en anglais "*copy neutral LOH*"), c'est-à-dire une LOH due à la perte d'un chromosome puis à la duplication à l'identique du chromosome restant. Et enfin, elles permettent d'estimer de façon fiable la ploïdie, ce que ne permettent pas les puces CGH. Ni les puces CGH, ni les puces SNP, ne permettent de détecter les translocations réciproques, c'est-à-dire les échanges de matériel entre chromosomes sans gain ou perte de matériel chromosomique.

Étude de l'épigénome

Pour l'étude du mirnome, le principe va être le même que pour le transcriptome. Les différences résident dans la manière d'extraire les miARN. Les puces vont contenir des sondes représentatives des miRNA. Le séquençage des miARN va être réalisé sur la totalité de la séquence du miRNA mature sur une longueur plus courte que pour l'exome.

Pour l'étude de la méthylation, le principe est le même que les puces SNP. L'ADN va être traité au bisulfite qui transforme les cytosines en uraciles mais uniquement si elles ne sont pas protégées par un événement de méthylation. Par conséquent, les sondes vont soit mesurer spécifiquement la forme avec un U et celle avec un C indépendamment, soit la sonde va mesurer la partie juste en amont du site CpG, et le marquage va être réalisé en ajoutant un C ou A marqué avec des fluorophores différents.

3.1.3 Les limites des omiques

Les données omiques apportent un nombre considérable d'information mais certaines limites sont à considérer pour exploiter et appréhender au mieux ce type de données.

Les biais techniques

Une difficulté est l'existence d'effets liés à chaque lot ("*batch effect*") ce qui pose problème quand plusieurs lots distincts doivent être analysés ensemble. Ces effets doivent être identifiés et corrigés lors de la normalisation des données. Toutefois si l'effet "*batch*" est confondu avec la variable d'intérêt, il sera alors impossible d'identifier les effets biologiques recherchés.

Qualité des échantillons biologiques

La qualité des analyses dépend de la qualité des échantillons biologiques, notamment pour l'ARN qui est moins stable que l'ADN. Les tissus sont collectés suite à une chirurgie ou une biopsie et de nombreux facteurs peuvent affecter la qualité des échantillons : le type et la durée de l'anesthésie, la chirurgie en elle-même comme l'ischémie (la diminution de l'apport sanguin artériel à un organe), la durée et le mode de transport, et la durée et le mode de conservation (Ma et al., 2012). Trois principales périodes peuvent avoir un impact et affecter différemment la qualité des échantillons : la période in vivo avant l'exérèse de la pièce chirurgicale, la période ex vivo entre l'exérèse chirurgicale et la congélation, et la période entre la décongélation et le traitement de l'échantillon (comme l'extraction des ARN). L'impact se mesure à deux niveaux : l'impact sur l'intégrité des ARN et l'impact sur l'expression des gènes qui lors de la période in vivo et ex vivo peut être modifiée. L'impact de la période in vivo est difficilement contrôlable (liée au temps opératoire). L'impact de la période ex vivo semble avoir un effet limité sur la dégradation des ARN mais par contre peut modifier l'expression de certains gènes, dont les gènes de l'inflammation et du cycle cellulaire. L'effet sur l'expression des gènes semble dépendre du tissu. Cet effet semble important dans le côlon (20% des gènes à 30 min) mais limité dans l'ovaire ou le sein (<1% des gènes) (Ma et al., 2012). L'intégrité des ARN est presque uniquement impactée après décongélation où une dégradation des ARN assez rapide et forte est constatée en l'absence d'un traitement spécifique. Les ARN les plus longs sont les plus sensibles à la dégradation. Un recueil rapide et une bonne conservation des échantillons sont donc indispensables pour limiter l'impact sur l'expression des gènes et la dégradation des ARN.

L'hétérogénéité intra-tumorale

Un autre écueil de l'utilisation de ces technologies dans le cas de l'étude de tumeurs tient à l'hétérogénéité du contenu cellulaire des échantillons tumoraux. Cela complique beaucoup l'interprétation des données obtenues par ces technologies.

La contamination par des cellules non tumorales, comme le stroma tumoral, le tissu adjacent normal ou les cellules immunitaires, participe de cette hétérogénéité. Cette contamination peut affecter notamment les résultats de transcriptome (Cleator et al., 2006) et de variations du nombre de copies ADN (Peiffer et al., 2006). Il est courant aujourd'hui de coupler un examen histologique du matériel passé en omiques, pour garantir histologiquement le contenu en cellules tumorales.

Par ailleurs, il existe au sein même de la tumeur une hétérogénéité régionale. L'une des premières études mettant en évidence par approches omiques l'hétérogénéité est l'étude de Navin et al. (2010). Ils montrèrent, dans le cancer du sein par approche CGH, qu'il existait des tumeurs "*monogénomiques*" avec un clone majoritaire quelle que soit la région de la tumeur et des tumeurs "*polygénomiques*" contenant plusieurs sous-populations clonales pouvant occuper la même région ou des régions différentes. Puis par une approche omique "single cell", permettant de faire de la génomique à l'échelle d'une seule cellule isolée de la tumeur, la même équipe montra que, dans les tumeurs "*monogénomiques*", une importante diversité des altérations génomiques était observée entre les cellules étudiées, même dans des régions identiques (Navin et al., 2011). Une autre étude qui fait référence dans le domaine est celle de Gerlinger et al. (2012) dans laquelle ils ont analysé moléculairement plusieurs régions et métastases de tumeurs du rein. Ils ont montré la coexistence de plusieurs clones présentant des mutations somatiques et des altérations chromosomiques différentes en fonction de la région interrogée. Seul un tiers des mutations somatiques étaient communes aux différents clones. L'hétérogénéité se retrouvait également au niveau de l'expression des gènes avec un profil d'expression en faveur soit d'un bon, soit d'un mauvais pronostic dans les différentes régions d'une même tumeur. Ceci suggère que l'analyse d'une partie d'une tumeur ne rend pas forcément compte de la complexité de la tumeur dans sa globalité.

En ce qui concerne le cancer colorectal, Dalerba et al. (2011) utilisa l'approche par "*single cell*" pour caractériser l'hétérogénéité tumorale. En se basant sur l'étude de la diversité cellulaire de l'épithélium normal, ils établirent des marqueurs spécifiques de chaque type cellulaire par l'étude du transcriptome de cellules isolées, puis caractérisèrent la composition des tumeurs en ces cellules. Ils montrèrent que les tumeurs étaient composées des diverses formes cellulaires de l'épithélium normal. Par ailleurs, ils montrèrent par xénogreffe, qu'une cellule isolée, présentant des marqueurs des cellules des compartiments des cellules souches et progénitrices, était capable de redonner naissance aux diverses formes de cellules de l'épithélium, soulignant que l'hétérogénéité clonale n'était donc pas la seule à prendre en compte. Très récemment, Sottoriva et al. (2015) proposèrent et validèrent un nouveau modèle d'évolution tumorale, le "*big bang model*", en étudiant la diversité intra-tumorale de 11 tumeurs et 4 adénomes colorectaux. Dans le modèle clonal d'évolution classique, par accumulation d'altérations, les clones sont sélectionnés pour l'avantage qu'ils ont acquis et finissent par devenir dominants. L'hétérogénéité intra-tumorale n'est alors qu'un état transitoire entre la sélection et l'expansion. Ce modèle requiert l'acquisition de diverses mutations "drivers". Dans le modèle "big bang", l'idée est que l'hétérogénéité tumorale n'est que le reflet de développement de clones en fonction du temps de progression de la tumeur, sans pression de sélection. De ce fait, les événements d'altérations les plus fréquents sont les plus précoces, les événements les moins représentés sont ceux survenus le plus tardivement. Les mutations initiatrices se retrouvent donc parmi les événements

les plus fréquents.

Cette hétérogénéité intra-tumorale explique certainement les raisons pour lesquelles l'identification de marqueurs pronostiques et prédictifs d'un traitement est difficile. En effet, le traitement adéquat devrait cibler les anomalies communes aux différents clones ou combiner plusieurs molécules ciblant différents clones sous peine de sélectionner des sous-clones résistants aux traitements.

Enfin, les tumeurs analysées sont fixées à un instant donné. Le transcriptome d'une cellule est dynamique et change rapidement en réponse à son environnement ou à des perturbations. Les données de microarrays ne rendent donc compte que d'une photographie de la réalité.

3.1.4 L'analyse des données omiques appliquées au cancer

Dans l'analyse des données de tumeurs, trois grands types d'analyses sont souvent réalisés :

— les approches pour l'identification de sous-types moléculaires de tumeurs, qui correspondent à de l'analyse dite non supervisée,
— les approches visant à la recherche de biomarqueurs, ou de signatures moléculaires, associés au diagnostic, à la survie ou à la réponse aux traitements, qui correspondent à l'analyse dite supervisée de sélection de variables, et enfin
— les approches de prédiction de classe, c'est-à-dire les approches visant à attribuer un échantillon à un groupe défini, appartenant également aux analyses de type supervisées.

Les approches pour l'identification de sous-groupes de tumeurs

De nombreuses méthodes existent pour appréhender la diversité des cancers. Les approches non-supervisées visent à regrouper les échantillons en fonction de la similitude de leur profil moléculaire. L'idée sous-jacente est que les tumeurs ayant des profils d'expression similaires auraient donc a priori des comportements biologiques semblables et proviendraient de voies de carcinogenèse communes, et partageraient des spécificités stromales, immunitaires ou tumorales identiques.

Il existe 2 grandes catégories de méthodes, les méthodes hiérarchiques et les méthodes de partitionnement. Les méthodes hiérarchiques produisent une hiérarchie des échantillons, c'est-à-dire un arbre de classification allant du plus petit groupe formé par un échantillon à un grand groupe regroupant l'ensemble des échantillons. Le nombre de groupes n'est pas préspécifié et est choisi a posteriori en fonction de l'arbre obtenu. Ces méthodes sont dites divisives, ou descendantes, lorsque tous les échantillons forment initialement un groupe qui est récursivement divisé en plus petits groupes. A l'opposé, elles sont dites agglomératives, ou ascendantes, lorsque chaque échantillon forme un groupe qui au fur et à mesure est fusionné à un autre groupe. Les méthodes de partitionnement, comme les k-means ou les cartes de Kohonen (les "Self Organized Map", SOM), recherchent elles à séparer les données en fonction d'un nombre de groupes spécifié en recherchant à optimiser un critère donné, par approche itérative. En pratique, les méthodes hiérarchiques agglomératives sont les plus utilisées en raison de leur simplicité et de la facilité d'interprétation des résultats. Elles ont l'avantage pas rapport aux méthodes de partitionnement de ne pas avoir à spécifier le nombre de groupes. A contrario elles sont plus rigides, une mauvaise décision dans l'arbre se répercute sur l'ensemble de l'ascendance ou descendance et n'est pas corrigée a posteriori (Dudoit and Gentleman, 2002).

a) Classification par approche hiérarchique ascendante Deux mesures sont nécessaires pour définir la hiérarchie : celle permettant de mesurer la similitude entre deux échantillons et celle permettant de mesurer la similitude entre deux groupes d'échantillons.

La similitude entre les profils moléculaires de 2 échantillons est représentée mathématiquement par une mesure de distance ou de dissimilarité entre les échantillons. De nombreuses distances existent, le choix de la distance a un impact considérable sur les résultats de la classification et il n'est pas toujours facile de comprendre les spécificités de chacune (Brazma and Vilo, 2000). En conséquence, le choix est généralement fait sur des critères a priori en fonction de la connaissance de la pathologie, ou, en recherchant à optimiser un critère donné comme par exemple le regroupement d'un groupe connu d'échantillons ou de réplicats techniques. Les distances les plus classiquement utilisées en génomique sont la distance Euclidienne et la distance de Pearson. La distance Euclidienne donne le même poids à tous les gènes et correspond à une extension à N dimension de la distance physique :

$$dis(S1, S2) = \sqrt{\sum_{i=1}^{g} |x1_i - x2_i|^2}, \tag{3.1}$$

avec S1 et S2, 2 échantillons, i le gène considéré et $x1$, $x2$ les mesures du gène des 2 échantillons

La distance de Pearson tient compte du niveau de variation entre les gènes et va elle plutôt prendre en compte les tendances globales de la série de gènes mesurés (Figure 3.5). Elle correspond à (1- le coefficient de Corrélation de Pearson (r)). r se distribue entre 1 et -1 en fonction de la forte corrélation ou anti-corrélation des échantillons, une valeur proche de 0 indiquant aucune corrélation entre les échantillons :

$$d(S1, S2) = 1 - r\,(S1, S2) = \frac{\sum_{i=1}^{p}(x1_i - \overline{x1})(x2_i - \overline{x2})}{\sqrt{\sum_{i=1}^{p}(x1_i - \overline{x1})^2 \sum_{i=1}^{p}(x2_i - \overline{x2})^2}} \quad , \tag{3.2}$$

où $\overline{x} = \frac{1}{p}\left(\sum_{i=1}^{p} x_i\right)$ est la moyenne des valeurs des gènes d'un échantillons x.

La distance se distribue donc entre 0 et 2, deux échantillons parfaitement corrélés auront une distance de 0 alors que des échantillons anti-corrélés seront les plus distants (1-(-1)=2). Elle est beaucoup plus intuitive en biologie et permet d'avoir des résultats cohérents. Si l'on souhaite regrouper des échantillons anti-corrélés avec des échantillons corrélés positivement, ce qui peut être plus intéressant à l'échelle des gènes, la distance peut être redéfinie ainsi :

$$d(S1, S2) = 1 - |r\,(S1, S2)| \tag{3.3}$$

La figure 3.5 représente visuellement les différences entre les 2 distances et l'impact qu'elles ont sur la classification finale.

Ensuite pour regrouper les échantillons entre eux, il existe plusieurs algorithmes possibles. La différence va résider dans la manière de calculer la distance entre les groupes. L'algorithme *single* va prendre la distance entre les paires les plus proches, *complete* entre les paires les plus éloignées, *average* la distance moyenne entre l'ensemble des paires et *ward* la distance entre les barycentres des groupes ce qui va permettre de minimiser l'inertie intra-classe [3] (Figure 3.5).

Le dendrogramme est la représentation graphique d'une classification hiérarchique. Il se présente comme un arbre dont les feuilles sont les échantillons et les branches correspondent aux distances entre échantillons ou groupes d'échantillons. La *heatmap* est

3. L'inertie d'un cluster mesure la concentration des points du groupe autour du centre de gravité.

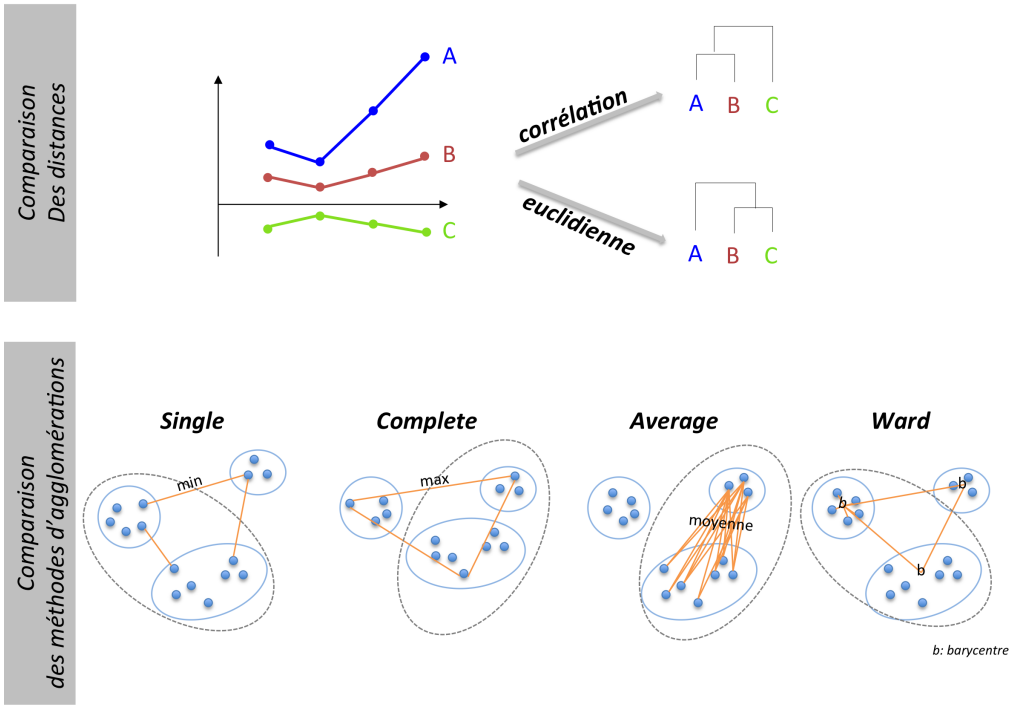


Figure 3.5 – Différence entre la distance Euclidienne et la distance de Pearson et comparaison des méthodes d'agglomération.

également une représentation classiquement utilisée en génomique pour illustrer les classifications. Elle permet de représenter l'expression des gènes en niveaux de couleurs et ainsi de visualiser les gènes aux niveaux d'expression communs et différents entre les groupes d'échantillons (Figure 3.6). L'étape critique ensuite est de définir le nombre de groupes, c'est-à-dire le niveau de coupe dans l'arbre. Des méthodes ont été développées pour aider à la décision (Shannon et al., 2003).

Les difficultés de la classification non supervisée hiérarchique résident dans le fait que le nombre de groupes est inconnu, dans la sélection des gènes et de la distance (Dudoit and Gentleman, 2002). De plus, ces méthodes ne sont pas basées sur des approches probabilistes qui permettraient de définir des tests pour estimer le niveau de coupe. Et enfin, elles définiront toujours des groupes même s'il n'y en a pas (Shannon et al., 2003). Par ailleurs, les biais techniques peuvent influencer la classification considérablement. Il est donc toujours important de vérifier la cohérence des groupes, la disponibilité d'annotations des échantillons étant cruciale pour s'assurer que ce ne sont pas les biais techniques qui expliquent la classification.

b) Robustesse d'une classification Un élément important dans ce type d'approche est de s'assurer de la stabilité des partitions obtenues. En effet les résultats peuvent varier, parfois de manière très importante, en fonction des paramètres choisis (la distance entre échantillons, la méthode d'agglomération entre groupes d'échantillons, le nombre de gènes et le nombre d'échantillons). Il est donc important d'évaluer la stabilité de la partition obtenue sous différentes conditions expérimentales. Le concept de classification consensus a été proposé il y a plusieurs décennies pour évaluer la robustesse d'une classification, et Monti et al. 2003 ont été parmi les premiers à l'utiliser avec des données omiques. L'idée

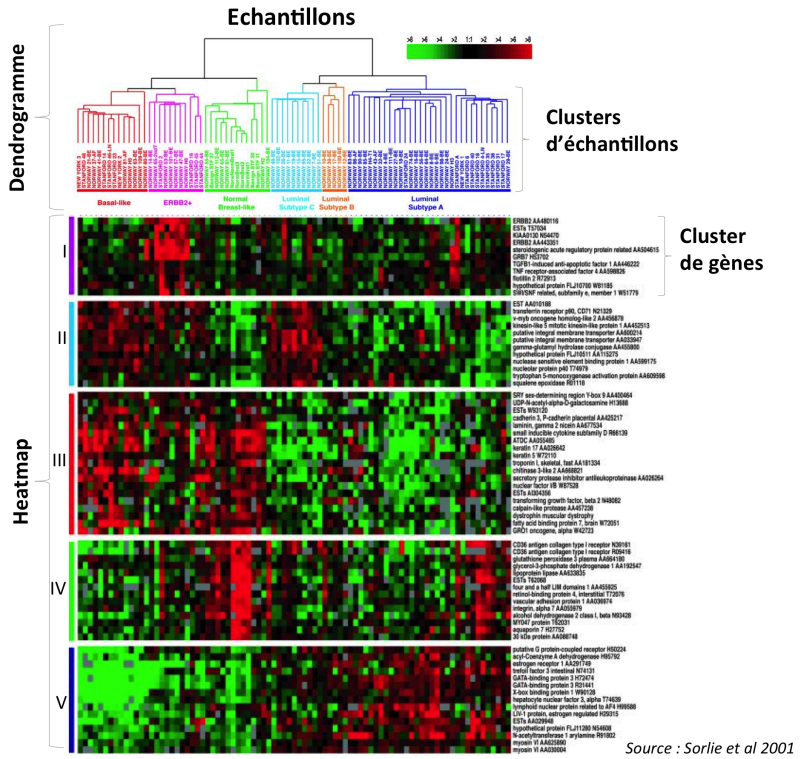


Figure 3.6 – Exemple de dendrogramme et de *heatmap* obtenus par classification hiérarchique agglomérative. Cet exemple illustre la visualisation de classification de cancers du sein proposée par Sørlie et al. (2001). En haut, l'arbre représente le dendrogramme des échantillons, colorés en fonction des sous-types définis. En dessous est représentée la *heatmap*, en niveau de vert et de rouge en fonction de la sous- et surexpression des gènes respectivement, de 5 clusters (ou groupes) de gènes d'intérêt. Cette représentation permet notamment de voir que les sous-types luminaux partagent un certain nombre de gènes dérégulés de la même manière entre ces différents sous-types comparativement aux sous-type non luminaux (cf. clusters III et V).

est de reproduire la procédure de classification un certain nombre de fois en changeant les paramètres (le nombre de gènes et/ou d'échantillons et/ou la méthode de distance ou d'agrégation) puis d'agréger les résultats obtenus par des valeurs de proportion de co-classification de chaque échantillon entre eux et d'obtenir une partition basée sur la matrice des co-classifications. Les échantillons qui sont les plus fréquemment co-classés ensemble vont donc définir des groupes stables. Cette approche est aujourd'hui très couramment utilisée dans l'analyse de données de microarrays (Wilkerson and Hayes, 2010). Toutefois un autre moyen de s'assurer de la robustesse d'une classification est de la valider sur un jeu de données indépendant, en retrouvant les groupes avec les mêmes profils moléculaires et les mêmes associations cliniques et moléculaires. Ceci n'est possible que si des jeux de données sont disponibles, ce qui est le cas pour les cancers colorectaux, et que si des annotations d'intérêt sont fournies, ce qui est par contre malheureusement plus rare.

Caractérisation des sous-types tumoraux

Une fois les sous-groupes définis, il est important pour confirmer leur intérêt biologique de les caractériser en définissant l'ensemble des altérations moléculaires (à l'échelle des

gènes et des voies de signalisation) et particularités cliniques qui leur sont spécifiquement associées.

Pour les associations aux données cliniques ou à des mesures de types catégorielles, comme une altération du nombre de copies ADN (gain/perte) ou une mutation, le test classiquement utilisé est le test exact de Fisher, ou lorsqu'il ne peut pas être utilisé en raison du trop grand nombre de modalités de l'annotation considérée, le test de Chi-2 qui est sa version non exacte. Ces tests statistiques permettent d'analyser des tables de contingences, c'est-à-dire la table dénombrant le croisement des modalités de la classification et les modalités de l'annotation considérée. Le test évalue l'indépendance des 2 variables et pose l'hypothèse nulle que les proportions relatives d'une variable sont les mêmes quelque soit les valeurs de la deuxième variable.

Pour définir les signatures moléculaires, c'est-à-dire l'ensemble des gènes significativement sous- ou sur-exprimés, des tests de comparaison de moyenne sont appropriés. Le t-test avec inégalité des variances, aussi appelé le test de Welch, lorsqu'il n'y a que 2 groupes, ou l'ANOVA, dans les cas de plus de deux groupes, sont classiquement utilisés. Des améliorations du t-test pour s'adapter aux données de microarrays ont été proposées, notamment en cherchant à mieux estimer la variance par gène, comme le t-test dit modéré Limma (Smyth, 2004). Ce dernier est apparu comme apportant des améliorations significatives au t-test classique et comparativement à d'autres tests, en terme de puissance et de taux de faux positifs et ce quelle que soit la taille du jeu de données (Jeanmougin et al., 2010). Par ailleurs, vu le nombre de tests réalisés dans le cas des microarrays, on se trouve confronté au problème de tests multiples. En effet, un gène va être considéré différentiellement exprimé avec un seuil α donné de se tromper. En général, on prend le seuil α de 0,05 ou 0,01. Mais ce seuil est valable pour un seul test. Si on réalise 100 tests, on multiplie par autant la chance de se tromper, et donc 5 faux-positifs seront attendus pour 100 tests au seuil α de 5%, c'est-dire que 5 tests sont attendus avoir une p-value inférieure à α alors que le gène n'est pas différentiellement exprimé. Dans les cas de puces de type Affymetrix, sur 50000 sondes dans une comparaison sans différence entre 2 groupes, on s'attendra à obtenir 2 500 tests significatifs au seuil de α 5% et correspondront donc à des faux positifs. C'est pourquoi il est nécessaire d'appliquer une procédure de correction des p-values ou d'évaluation du nombre de faux positifs lorsque l'on réalise un grand nombre de tests. Il existe deux grands types de méthodes : celles qui contrôlent le FWER (Family Wise Error Rate), la probabilité d'avoir au moins un faux-positif, qui sont très conservatives, c'est-à-dire qu'elles ne laissent pas passer de faux positifs, ce qui va engendrer en contrepartie beaucoup de faux négatifs, et celles qui contrôlent le FDR (False Discovery Rate), la proportion de faux-positifs parmi les gènes ressortis régulés, comme les méthodes de Benjamini and Hochberg (1995) ou de Benjamini et Yekutieli (2001).

Les analyses de *pathways* permettent de tester s'il y a, parmi les gènes dérégulés dans le sous-type considéré, un enrichissement en gènes de n'importe quel "*pathway*" (liste de gènes), comme par exemple l'ensemble des gènes d'une voie de signalisation. La stratégie classique est d'analyser tous les *pathways* répertoriés et de s'intéresser à tous ceux significativement enrichis. D'abord, pour chaque gène, on obtient une statistique en comparant le sous-type d'intérêt à son complément. Il y a alors deux types d'approches classiquement utilisées : (i) celles se basant sur une sélection de gènes, obtenue par application d'un seuil sur la statistique précitée, et testant la surreprésentation des gènes du *pathway* dans cette liste présélectionnée (le test hypergéométrique ou méthodes dérivées comme GoStat (Beissbarth and Speed, 2004)) et (ii) celles ne demandant pas de fixer un seuil mais se basant sur une agrégation, pour les gènes du *pathway*, de la statistique précitée aboutissant à un score d'enrichissement (comme GSEA (Subramanian et al., 2005)) (Hol-

mans, 2010). Différents types d'algorithmes existent et proposent des modes d'agrégation différents, liés à des hypothèses statistiques différentes. Un troisième type d'approche est apparu plus récemment qui permet d'utiliser des informations de relation entre les gènes du *pathway*, sa topologie (Khatri et al., 2012). Il faut tenir compte dans l'interprétation des résultats de la qualité (niveau de preuve expérimentale, curation, ...) des *pathways* utilisés.

Une autre stratégie plus dirigée peut être également utilisée si on s'intéresse à des *pathways* spécifiques. Des méthodes d'enrichissement dites *self-contained* (les précédentes sont dites compétitives) sont alors requises pour tester l'enrichissement. Elles vont tester si l'enrichissement est différent par rapport à l'enrichissement obtenu sur des données simulant une non-association (Ramanan et al., 2012).

Prédiction de sous-types tumoraux : les méthodes supervisées

Pour pouvoir assigner de nouveaux échantillons à la classification obtenue, il est classique de construire un prédicteur. C'est un problème assez complexe avec deux difficultés majeures : la sélection des gènes (les variables) et le sur-apprentissage.

Une approche qui est très couramment utilisée en cancérologie est l'approche par centroïde (Cover and Hart, 1967). Pour chaque groupe, la moyenne d'expression des gènes est calculée, ce qui définit le centroïde de la classe, un vecteur de données de la longueur du nombre de gènes retenus. Puis on affecte un nouvel échantillon en mesurant la distance du profil d'expression de l'échantillon aux centroïdes de chacun des groupes et on attribue l'échantillon au groupe du centroïde le plus proche. Tibshirani et al. (2002) proposa une méthode améliorée très utilisée qui permet de réduire le bruit des centroïdes et d'ajouter l'étape cruciale de sélection de gènes dans le processus de prédiction.

Toutefois, il est classique pour éviter le surapprentissage d'utiliser des approches de *cross-validation* ou de *bootstrap* ou de Monte Carlo ou encore de *leave-one-out* (voir Slawski et al. 2008). Le prédicteur va alors être construit sur une partie des données puis testé sur le reste et ce plusieurs fois. Pour la sélection de variables, on sélectionne en général les gènes les plus associés, soit sur la p-value du test d'analyse différentiel soit sur l'AUC (Aera under the curve), l'aire sous la courbe de la spécificité en fonction de la sensibilité, calculé pour chacun des gènes, qui semble être efficace (Lauss et al., 2010). Il est préférable de retirer la redondance dans les gènes sélectionnés avant de construire le prédicteur (Ding and Peng, 2005).

Les mesures de la qualité de la prédiction sont la sensibilité, la spécificité, la valeur prédictive positive (VPP) et la valeur prédictive négative (VPN). En fonction de la question posée, on pourra favoriser l'une ou l'autre des mesures 3.7.

3.2 La recherche de marqueurs pronostiques dans les cancers colorectaux

Comme décrit dans le 1er chapitre, la classification TNM ne permet pas de bien identifier 10 à 20% des patients de stade II, pour lesquels une chimiothérapie adjuvante pourrait être bénéfique, et 30-40% des stades III pour lesquels la chimiothérapie adjuvante n'est pas suffisante ou, inversement, pour lesquels une simple chirurgie pourrait être suffisante. L'émergence des microarrays a donc apporté beaucoup d'espoir en permettant d'étudier l'association au pronostic d'un très large panel de marqueurs. Une multitude d'études pronostiques sur des données transcriptomiques ont été publiées utilisant diverses approches (Manceau and Laurent-Puig, 2012; Cardoso et al., 2007; Sanz-Pamplona et al., 2012).

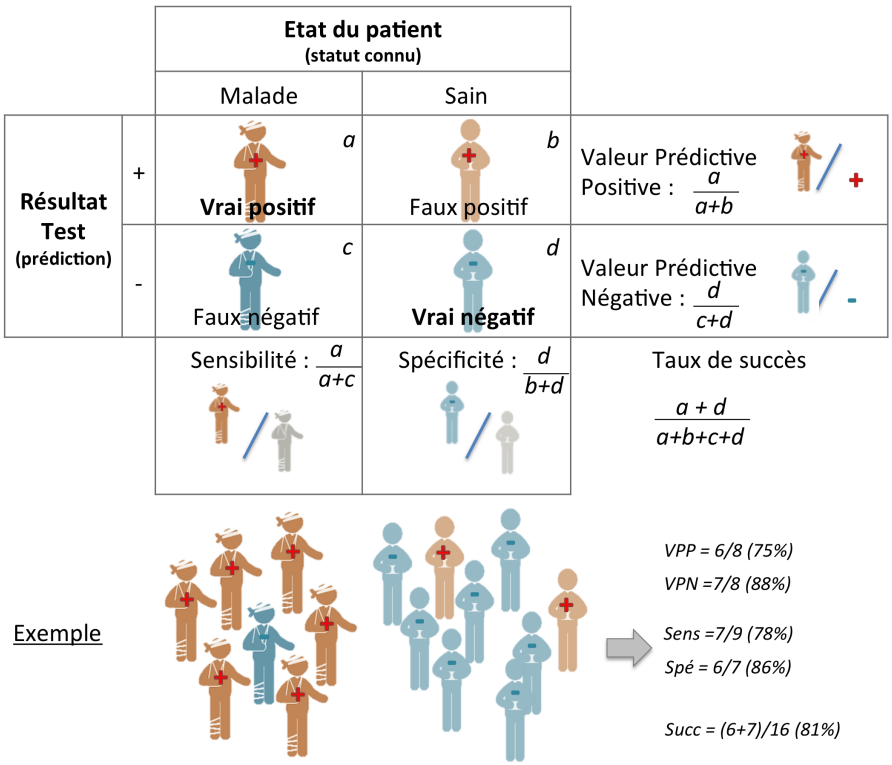


Figure 3.7 – Ensemble des variables décrivant la qualité d'une prédiction.

3.2.1 Aperçu de l'état de l'art

La majorité des études ont utilisé des approches supervisées, comparant soit les événements de rechute (la RFS, la DFS, l'OS, la CSS ou la SAR [4]), soit les tumeurs de stades I vs IV, les tumeurs envahissant les ganglions lymphatiques N+ vs les tumeurs N- ou les métastases vs les tumeurs primaires, soit l'environnement tumoral entre des tumeurs métastatiques et des tumeurs non métastatiques (Manceau and Laurent-Puig, 2012). Quelques études ont proposé des approches non supervisées en recherchant des sous-types "naturellement" pronostiques (Oh et al., 2012) ou à l'inverse pour obtenir une signature pronostique indépendante de la classification (Salazar et al., 2010).

Cardoso et al. (2007) et Roth et al. (2012) ont montré la faible concordance des résultats au niveau des gènes et au niveau de la prédiction des patients de toutes ces études. Selon la revue de Cardoso et al. (2007), seulement 13 gènes ont été retrouvés associés au pronostic dans au moins 2 études, dont 6 étaient exprimés dans des sens de régulation contraires. La méta-analyse du pronostic de 31 signatures proposée par Sanz-Pamplona et al. (2012) a révélé que la majorité de ces signatures sont bien associées au pronostic, en moyenne entre les différents jeux de données, mais qu'elles avaient des valeurs faibles de discrimination (faible spécificité et sensibilité). Ce qui en soit n'empêche pas d'avoir une utilité clinique car ce sont les valeurs prédictives négative et positive (VPN et VPP) qui sont importantes. Toutefois, les différentes signatures présentaient également de faibles VPP. La faible concordance des résultats a été également pointée du doigt dans l'article Michiels et al. (2005) qui par l'étude de la stabilité de différentes signatures et de leurs prédictions conseillait de prendre avec précaution les résultats des études menées dans le

4. RFS : Relapse Free Survival, Survie sans rechute ; DFS, Disease Free Survival ou Survie sans maladie ; OS : Overall Survival, Survie globale ; SAR, *Survival After Relapse*, Survie Après Rechute

cancer. En effet, seulement 2 sur les 7 étudiées classifiaient mieux que par chance. Les signatures pronostiques sont extrêmement dépendantes de la cohorte utilisée pour l'apprentissage, et la manière de valider les prédicteurs est cruciale. Notamment lorsqu'il n'y a pas de jeu de validation indépendant disponible, il est recommandé de faire de la validation par rééchantillonage aléatoire quand la taille de la cohorte le permet. Toutefois, aujourd'hui, la mise en accès publiques de très nombreux jeux de données sur le site Gene Expression Omnibus (GEO) du NCBI (ou SRA pour les données de séquençage) ou encore ArrayExpress de l'EBI rendent possible la validation sur un jeu de données indépendant. Les problèmes d'analyse du pronostic spécifiquement générés par les données de microarrays ont amené à proposer en 2005 un guide appelé REMARK pour "REporting recommendations for tumor MARKer prognostic studies" (McShane et al., 2005; Altman et al., 2012) afin qu'il y ait une transparence sur la manière d'établir les prédicteurs. Dupuy and Simon (2007) proposèrent une liste de "do" et "don't" pour l'analyse du pronostic afin d'établir des bonnes pratiques d'analyse. Par ailleurs, pour qu'un test soit utilisable en clinique, il doit répondre aux 3 critères établis par le ACCE Model Project [5] pour évaluer les tests génétiques : la validité analytique, la validité clinique et l'utilité clinique (Figure 3.8).

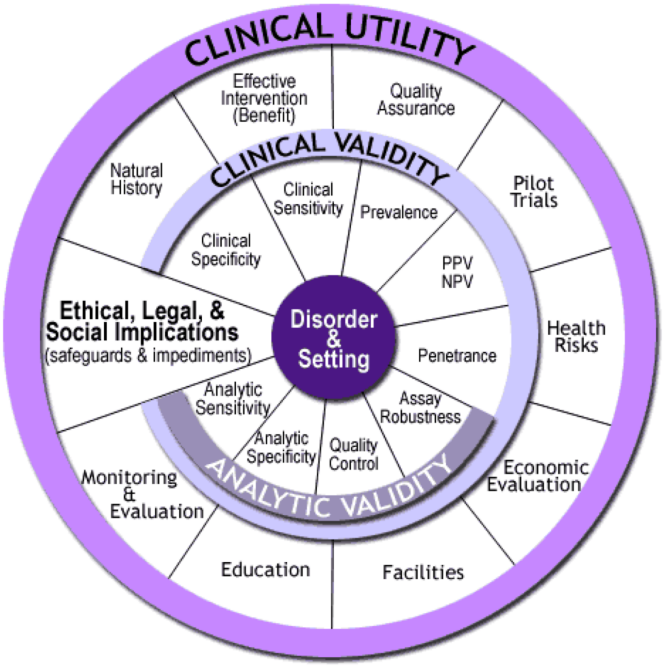


Figure 3.8 – Principaux critères pour évaluer un test génétique selon le projet *AACE model* visant à définir le processus analytique pour évaluer les données scientifiques sur les nouveaux tests génétiques.

Parmi toutes les signatures pronostiques des CCR publiées, quasiment aucune ne s'est avérée valide et applicable en pratique clinique. Les facteurs en cause pouvant être la taille des cohortes, l'hétérogénéité inter-tumorale, la non validation sur des jeux de données indépendants, la qualité des annotations fournies, la qualité des échantillons ou encore

5. http://www.cdc.gov/genomics/gtesting/ACCE/acce_proj.htm

la représentativité de la portion de tumeur extraite (soit l'hétérogénéité intra-tumorale et la contamination par le stroma et les cellules immunitaires). Seuls deux prédicteurs, Oncotype DX et Coloprint se démarquent pour leur utilité clinique validée concernant le pronostic des stades II/III.

3.2.2 Focus sur Oncotype DX

De manière analogue au prédicteur développé pour le cancer du sein, O'Connell et al. (2010) adoptèrent une approche supervisée sur une très large cohorte de 1851 patients de stade II et III provenant de 4 études indépendantes, en utilisant des échantillons fixés au formol. Ils firent une première sélection de 761 gènes candidats associés au pronostic (dans la littérature, les bases de données et dans des jeux de données de microarrays publiques) ou appartenant à des voies de signalisation considérées comme fonctionnellement importantes pour le cancer du côlon. Sur ces 761 gènes, 375 ont été retenus en fonction de leur association à la récurrence et au bénéfice de la chimiothérapie par modèle de Cox. Afin de retirer l'effet de confusion possible avec le stade, les gènes significativement différents entre les stades II et III ont été exclus. Sur l'étude de ces gènes en RT-PCR dans les 4 jeux de données indépendants, 48 gènes ont été retrouvés associés significativement à un risque de récidive et 66 associés à un bénéfice de la chimiothérapie. De ces gènes, les meilleurs panels pour la récurrence et le bénéfice du traitement ont été sélectionnés, en se basant notamment sur la puissance de l'association et la consistance d'une étude à l'autre, puis utilisés pour définir les scores pour prédire la rechute (RS, *Recurrence Score*) et la réponse au traitement (TS, *Treatment Score*). Le score RS contient 12 gènes, 7 gènes associés à la récurrence et 5 de référence, et le TS est composé de 11 gènes, 6 gènes associés au bénéfice du traitement et 5 référence. La valeur prédictive de chaque score a été mesurée par rééchantillonage, plutôt que sur des séries indépendantes, ce qui laisse penser que les résultats sont peut être plus optimistes que la réalité. Les scores ont ensuite été validés dans une seconde publication sur l'essai thérapeutique QUASAR sur des patients de stade II (Gray et al., 2011). Le score RS a été validé comme pronostic dans les patients avec des tumeurs de stade II après chirurgie et complémentant la valeur pronostique du stade T et du MSI. Par contre, le score TS n'a pas été prédictif d'un bénéfice de la chimiothérapie. Les mêmes conclusions ont été obtenue sur une étude clinique prospective sur des patients de stades II et III (Yothers et al., 2013) où le score RS était prédictif de la récurrence dans les 2 stades de manière indépendante des facteurs cliniques et pathologiques conventionnels, et également dans 2 autres études (Venook et al., 2013; Srivastava et al., 2014). Le fait de se mettre dans des conditions non idéale mais proche de la réalité du terrain en utilisant beaucoup d'échantillons doit en partie expliquer la reproductibilité de l'approche, car cela limite le surapprentissage sur un jeu particulier.

3.2.3 Focus sur Coloprint

Salazar et al. (2010) adoptèrent une approche mixte non supervisée puis supervisée pour définir une signature pronostique pour les patients de Stade II/III. A partir de données de microarrays de tumeurs congelées de 188 patients de tout stade, inclus prospectivement, la classification par classification hiérarchique a mis en évidence 3 groupes, un groupe enrichi en MSI et en mutation BRAF de meilleur pronostic, un groupe non MSI enrichi en mutation BRAF de plus mauvais pronostic et un groupe non MSI et non BRAF de pronostic intermédiaire. L'originalité de leur approche est d'avoir établi leur signature pronostique uniquement sur le 3ème groupe afin d'obtenir une signature non biaisée par les 2 premiers groupes qui risquaient d'expliquer à eux seuls le pronostic. Un ensemble

de 18 gènes très fortement associés à la survie sans rechute métastatique ont été sélectionnés de manière optimisée par analyse différentielle entre les patients développant une métastase après traitement et les patients ne rechutant pas. Ces 18 gènes ont ensuite été utilisés pour construire un prédicteur basé sur l'approche des centroïdes les plus proches. Les échantillons étant plus proches du centroïde établi sur les patients non rechutant sont classés comme ayant un risque faible de rechute, alors que ceux plus proches de l'autre centroïde sont classés comme ayant un risque élevé. Ils validèrent son pouvoir prédictif sur le même jeu de données par *cross-validation* puis sur un jeu de données indépendant provenant d'un autre centre clinique de 206 échantillons majoritairement de stade II. Les groupes définis étaient très associés au pronostic pour les stades II et III confondus, ou indépendamment, et la valeur pronostique du prédicteur restait significative avec l'ajout du stade dans le modèle. Dans les stades II, il était meilleur que le stade et, parmi les patients ayant été traités par chimiothérapie, 68% étaient classés à faible risque. Ils conclurent que ces patients pourraient ne pas avoir besoin d'une chimiothérapie. De plus, le prédicteur classait comme attendu la grande majorité des MSI dans le groupe de faible risque, bien que ce groupe ait été exclu lors de la sélection des gènes, ce qui confirme la puissance de l'approche développée.

Suite à cette publication, un test diagnostique a été développé. Le travail de validation clinique s'est poursuivi en appliquant le test sur 3 jeux de données supplémentaires pour lesquelles la valeur pronostique du prédicteur a été confirmée pour les stades II et III, la valeur pronostique étant supérieure à celles des critères cliniques classiques (Maak et al., 2013; Kopetz et al., 2015). Un essai thérapeutique de phase III est en cours de recrutement pour évaluer Coloprint sur 785 patients avec une tumeur de stade II (PARCS study, NCT00903565)(Chee and Meropol, 2014). Le test est commercialement disponible [6].

3.3 Caractérisation et classification des tumeurs colorectales par l'utilisation d'omiques

Les approches pronostiques non supervisées auguraient de l'intérêt de définir des sous-types mais la grande majorité des publications se sont arrêtées à l'étude du pronostic. Contrairement au cancer du sein pour lequel dès le début des microarrays une classification a été proposée (Perou et al., 2000), les classifications du cancer colorectal se sont arrêtées à l'intégration des marqueurs moléculaires (cf. section 2.3) ou à la caractérisation des entités déjà identifiées. Ce n'est que très récemment que des classifications non supervisées basées sur les données d'expression ont commencé à voir le jour dans la littérature.

3.3.1 Caractérisations par signatures supervisées

L'essentiel des travaux menés au début de l'exploitation des puces dans les années 2000 se sont portés sur la caractérisation de marqueurs différenciant chaque stade de la progression tumorale. Les études ont comparé les profils d'expression entre les tumeurs et les tissus sains, les adénomes et les adénocarinomes (Manceau and Laurent-Puig, 2012; Cardoso et al., 2007). La méta-analyse de Chan et al. (2008) et la revue de Cardoso et al. (2007) ont permis d'y voir un peu plus clair dans ce dédale d'études. Cardoso et al. (2007) établirent une liste très exhaustives des publications par type d'étude et de données génomiques et en dégagèrent les gènes différentiellement exprimés entre carcinome et tissu normal et adénome et entre MSI et MSS, ainsi que et les altérations du génome à différents

6. http ://www.agendia.com/healthcare-professionals/colon-cancer/

stades de la carcinogénèse, concordants entre plusieurs études. Notamment, la β-caténine (*CTNNB1*) et *SMAD4* étaient retrouvés surexprimés dans les MSI par rapport aux MSS et *TP53* sous-exprimé dans 3 études. Chan et al. (2008) rassembla la liste des gènes différentiellement régulés entre carcinome et normal, adénome et normal et carcinome et adénome en utilisant les données de 26 études et conclut que 573 gènes sur les 6537 gènes reportés différents entre carcinome et normal étaient confirmé par plusieurs études, 39 sur les 1101 entre adénomes et normal et 5 sur les 538 entre carcinome et adénome. La taille des cohortes et le manque d'ajustement des p-values pour les tests multiples peuvent expliqué ses différences. Parmi les gènes les plus fréquents retrouvés entre cancer et normal, *TGFβI* et *MYC* a été trouvé surexprimé dans les cancers dans 9 et 7 études respectivement et *CA2* sous exprimés dans 11 études.

3.3.2 Les classifications non supervisées de tumeurs colorectales basées sur les omiques

Finalement très peu de classifications, dans le sens de découverte de nouvelles entités, ont été réalisées dans le cas du cancer colorectal. Sur les données d'expression, les classifications se sont limitées à la séparation MSI/MSS et au sein des MSI entre les MSI sporadique et les MSI héréditaires. L'une des limites majeures à la découverte de classes est la taille trop limitée de nombreux jeux de données, ce qui peut expliquer ce constat.

Basées sur le transcriptome

Contrairement à la majorité des cancers fréquents comme le cancer du sein (Perou et al., 2000; Sørlie et al., 2001), du poumon (Wilkerson and Hayes, 2010) ou du cerveau (Verhaak et al., 2010) pour lesquels des classifications moléculaires basées sur le transcriptome ont été proposées depuis longtemps, quasiment aucune classification n'avait été proposée jusqu'en 2011. Les seules classifications avant 2011 reportaient seulement la différence entre les MSI et les MSS comme Kruhøffer et al. (2005) qui par classification hiérarchique, complètement non supervisée (sans sélection de gènes), de 101 échantillons tumoraux et 17 échantillons normaux, définit un groupe isolant les normaux et 3 groupes de tumeurs isolant les MSI, sporadiques et héréditaires rassemblés, et séparant les MSS en 2 groupes sans association avec la localisation proximale ou distale dans le côlon. En dehors des 2 classifications proposées dans le cadre de la recherche de marqueurs pronostiques (Salazar et al., 2010; Oh et al., 2012), où aucune analyse approfondie des sous-types identifiés n'a été menée (Salazar et al. 2010 montre seulement une association d'un sous-type de mauvais pronostic à la mutation de BRAF), Loboda et al. (2011) sont les premiers à avoir proposé une classification différente de MSI/MSS. En se basant sur les gènes associés à la 1ère composante de l'analyse en composante principale (ACP), réalisée sur les données d'expression de 326 cancers du côlon, ils identifièrent 2 sous-groupes très fortement associés à la signature de la transition Epithélio-mésenchymateuse (EMT) qu'ils nommèrent *mesenchymal* et *epithelial*. Confirmant la part importante de l'EMT dans les cancers colorectaux, ils trouvèrent une anti-corrélation du microARN mir-200 connu pour réguler l'EMT. Toutefois, cette classification manquait d'association à des annotations moléculaires connues et ne permettait pas de replacer la classification dans le contexte moléculaire avec les instabilités MSI, CIN et CIMP et les autres mutations connues. L'année d'après, une autre publication (Perez-Villamil et al., 2012) proposa une classification en 4 sous-types, basées sur 88 tumeurs de stade I à IV : un sous-type de tumeurs ayant un pourcentage de stroma plus faible et plus de β-caténine nucléaire, de bon pronostic, et un sous-type caractérisé par MSI, la mutation de BRAF et l'histologie mucineuse, les 2

autres sous-types ne présentant pas d'associations spécifiques aux annotations disponibles. Ils montrèrent que la classification était indépendante du stade TNM, en mentionnant que cela impliquait que les sous-types sont établis dès l'initiation de la tumeur. Enfin, le TCGA proposa une classification sur les données d'expression en 3 groupes : MSI/CIMP, invasive et CIN. Mais l'objectif principal de l'article du TCGA était de décrire les mutations et les événements au niveau de l'ADN. Ils identifièrent un sous-groupe présentant un nombre très important de mutations, regroupant l'ensemble des tumeurs MSI et une petite fraction de tumeurs MSS. La classification des données d'expression n'a pas été plus approfondie, elle n'était présentée que pour montrer l'association des tumeurs hypermutées au groupe MSI/CIMP en figure supplémentaire. (Cancer Genome Atlas Network, 2012).

Par contre fin 2012 et en 2013, une multitude d'études proposant des classifications plus raffinées et détaillées sont parues quasiment en même temps, dont celle qui fait l'objet de mon travail de thèse(Schlicker et al., 2012; Marisa et al., 2013; Sadanandam et al., 2013; Melo et al., 2013; Budinska et al., 2013; Roepman et al., 2013). Elles seront développées dans les résultats.

Basées sur le génome

En dehors du transcriptome, l'émergence des microarrays avec les CGH arrays, puis les puces SNP, ont permis d'évaluer finement les profils de remaniements chromosomiques dans les cancers colorectaux. Il n'y a pas eu de classification proposée, les classifications sur les événements ADN étant plus difficiles à établir, seulement un catalogue des altérations fréquemment observées. Une différence de profil très nette est observée entre les tumeurs de type MSS et MSI, les MSS présentant de nombreuses pertes de bras chromosomiques alors que les MSI ne sont quasiment pas remaniés et, lorsqu'il y a des remaniements, il s'agit de gains (8q, 7q, 17q) (Camps et al., 2006).

Basées sur le méthylome

Le travail de Hinoue et al. (2012) est l'une des classifications omiques les plus abouties pour le cancer colorectal. Ils confirmèrent les classes trouvées par Toyota et al. (1999) mais à la résolution du génome entier. Ils définirent 4 sous-types de tumeurs en se basant sur les profils de méthylation des sites CpG de l'ADN : CIMP-high, CIMP-low et 3/4 qui ont des profils intermédiaires entre les tissus normaux adjacents et les cancers CIMP-low.

Basées sur le mirnome

Comme pour les données d'expression, la plupart des études du mirnome ont été supervisées entre les normaux et les tumeurs et entre les MSI et les MSS. Les seules classifications proposées séparent les MSI des MSS et les MSI sporadiques des héréditaires. (Balaguer et al., 2011).

Basées sur les données de séquençage

Avec l'arrivée des NGS, l'ensemble des mutations ont été caractérisées. Il n'y a pas de classification particulière, elles sont juste présentées ordonnées selon leur fréquence. La majorité des mutations étaient déjà connues, mais ces analyses ont permis de caractériser plus finement le contingent de mutations des tumeurs de type MSI et hypermutées. Seules les mutations de *POLE* et *POLD1* associées à l'hypermutabilité ont apporté de nouvelles informations (Cancer Genome Atlas Network, 2012; Kim et al., 2013; Donehower et al., 2013).

Avec l'émergence des technologies à haut-débit, un nombre important de travaux ont comparé les profils d'expressions dans les CCR et ont cherché à trouver des marqueurs pronostiques pour améliorer la prise en charge thérapeutique des patients, reposant quasiment uniquement sur la prédiction du stade TNM, non satisfaisante pour un certain nombre de patients. On aurait pu s'attendre à ce que la quantité d'informations permettent de trouver des marqueurs utiles pour la pratique clinique mais pour l'instant le bilan est plutôt mitigé. Aucun biomarqueur ou signature fiable est encore utile et applicable en clinique, à l'exception peut-être de Oncoprint DX et Coloprint dont la validité clinique a été démontrée. En dehors de considérations techniques, comme la taille des cohortes ou la qualité des échantillons et données, l'une des raisons possibles est que les CCR sont hétérogènes et que cette hétérogénéité empêche de trouver des marqueurs applicables à l'ensemble des tumeurs. Les classifications basées sur les données transcriptomiques ont considérablement enrichies la compréhension de beaucoup de cancers dont le cancer du sein, du poumon et du cerveau. Étonnamment, aucune classification moléculaire robuste basée sur l'expression des gènes n'avait encore été établie pour les cancers du côlon lors de l'initiation de ce travail. Une telle classification permettrait d'élucider l'hétérogénéité de la maladie de manière plus fine que les trois formes d'instabilité MSI, CIN et CIMP.

Troisième partie

Résultats

Étude de l'hétérogénéité moléculaire des CC par analyse non supervisée des données transcriptomiques

En clinique, comme développé dans la partie de revue bibliographique, la classification TNM est le seul outil pronostic utilisé, mais elle n'est pas suffisante pour bien traiter près de la moitié des patients de stade II/III. Malgré l'observation de plusieurs formes d'instabilité, de nombreuses mutations de gènes clés et de diverses cellules et formes d'adénomes, l'hétérogénéité moléculaire des CRC n'a pas été décrite de manière approfondie. Les études basées sur les technologies à haut-débit auraient eu les moyens de décrire finement cette hétérogénéité mais la grande majorité des études publiées se sont quasiment toutes soit intéressées au pronostic, soit limitées à des approches comparatives entre les entités de cancers colorectaux déjà connues. Toutefois, les données omiques constituent, grâce à la quantité d'information qu'elles apportent, un outil incomparable pour élucider la question de l'hétérogénéité de la maladie. Le but de mon travail de thèse a donc été d'étudier de manière approfondie cette hétérogénéité inter-tumorale via l'utilisation de données omiques. Il a consisté principalement en l'établissement d'une classification moléculaire puis, suite à la publication de ces résultats, à la participation à un travail collaboratif pour définir une classification consensus entre plusieurs classifications publiées en même temps par plusieurs équipes dans le monde, lequel vient d'être accepté pour publication.

Article PLoS Medicine - Mise en évidence de l'existence de sous-types moléculaires des cancers du côlon par l'établissement d'une classification robuste des données d'expression

Le travail que j'ai mené pour cette étude s'inscrit dans le cadre du consortium sur les cancers du côlon (CC) du programme Carte d'Identité des Tumeurs (CIT) de la Ligue Nationale Contre le Cancer. Grâce à ce consortium, une très large base de données omiques de tumeurs de 750 patients atteints de CC, provenant de divers centres partout en France, a été générée. Les transcriptomes, analysés sur puces Affymetrix (U133plus 2) et les altérations génomiques, analysées sur puces CGH, étaient disponibles pour 566 et 464 tumeurs respectivement. Un important effort d'intégration des annotations cliniques et de caractérisation moléculaire des tumeurs pour les mutations et instabilités du génome classiquement observées a été réalisé conjointement par les collaborateurs des projets et par les curateurs de la base de données CIT. Afin d'évaluer l'hétérogénéité des tumeurs, l'objectif du travail a été d'établir une classification des cancers du côlon, robuste, à partir des données

transcriptomiques de ces tumeurs, les ARNm étant en grande partie le reflet de ce qui se passe dans les cellules et étant généralement impactés de manière directe ou indirecte par les altérations génétiques ou épigénétiques.

L'approche développée Pour établir une classification robuste, en dehors de la taille de la cohorte, qui doit être assez grande et représentative de la population générale des cancers colorectaux pour avoir une représentation de sous-groupes de tumeurs moins fréquents, deux critères sont apparus importants à prendre en compte dans les analyses : (i) la classification doit être établie par des méthodes permettant de s'assurer qu'elle est stable indépendamment des paramètres de classification et (ii) elle doit être validée sur des données indépendantes.

Pour établir la classification, j'ai recherché l'approche qui permettait à la fois de minimiser les biais techniques, d'obtenir des résultats facilement interprétables et robustes. La qualité des données étant primordiale, je me suis tout d'abord assurer de corriger les biais techniques potentiels dû à des différences d'extraction et de centre. Ensuite j'ai réalisé une approche par classification hiérarchique sur 443 échantillons (une partie des données a été réservée pour la validation) à partir des gènes les plus variants en terme d'expression. La classification hiérarchique a l'avantage d'organiser les groupes les uns par rapport aux autres et, à l'aide de "heatmap", permet de mieux appréhender la structure des profils d'expression. Et enfin, pour rentre la partition robuste, j'ai adopté une approche de classification consensus en réitérant 1000 fois la procédure de classification. Les itérations ont été faites en faisant varier à la fois le nombre d'échantillons et le nombre de gènes, tout en restant dans des proportions qui évitent de perdre l'impact de sous-groupes moins fréquents. A partir des partitions obtenues, une partition consensus par classification hiérarchique des fréquences de co-classements des tumeurs les unes par rapport aux autre a été établie, les tumeurs se co-classant le plus fréquemment ensemble se retrouvant affectées dans le même groupe. La sélection du nombre de sous-types a été définie en fonction d'un critère quantitatif (basé sur la distribution attendue de la distribution des fréquences de co-classements), de la visualisation de la matrice de co-classement et du nombre de tumeurs dans les sous-types.

Une fois la classification obtenue, j'ai cherché à caractériser finement les sous-types en identifiant les associations aux annotations cliniques, aux instabilités génomiques MSI/CIMP et CIN (défini à partir des données de CGH), aux mutations des gènes classiquement mutés (TP53, KRAS, BRAF) et à des voies de signalisation. J'ai également comparé les profils d'altérations du génome des sous-types entre eux. Enfin, il m'est apparu intéressant d'essayer de trouver des pistes quand à la composition cellulaire de ces sous-types, notamment en terme de cellules souches (suite à un travail réalisé avec le laboratoire d'Alex Duval cf. Annexe B (page 119)) et quant à l'origine des adénomes en raison de la voie festonnée qui commençait à émerger dans la littérature et de l'association aux sous-types de la classification proposée par Jass (2007). Pour cela, j'ai recherché des signatures d'intérêt dans la littérature et j'ai annoté chaque tumeur en fonction de ces signatures. Me trouvant limitée par le peu de signatures moléculaires disponibles, les signatures que j'ai retenues pour l'origine de la lésion précancéreuse étaient : une signature d'expression établie par comparaison de tumeurs classées histologiquement festonnées ou conventionnelles (Laiho et al., 2007) et une signature de la mutation *BRAF* (Popovici et al., 2012) qui isole, en plus des tumeurs mutées pour *BRAF*, des tumeurs non mutées *BRAF* mais mutées *KRAS*, ce qui me paraissait pouvoir refléter indirectement la diversité des adénomes festonnés. Pour la caractérisation relative aux cellules souches, j'ai retenue une signature de cellules souches intestinales normales (Merlos-Suárez et al., 2011) et une signature établie entre les cellules

du haut et du bas de cryptes coliques (Kosinski et al., 2007). Des profils de sous-types *normal-like* ayant été décrits dans le cancer du sein (Sørlie et al., 2001), j'ai également regardé les profils d'expression des tumeurs comparativement à ceux d'échantillons non tumoraux adjacents aux tumeurs et évalué leur distance.

Les sous-types définis Au final, 6 sous-types ont été identifiés, 4 sous-types bien délimités et 2 sous-types avec des tumeurs présentant de fortes co-classifications entre eux, suggérant que ces tumeurs partagent des signatures de gènes communes (voir Figure 9) :

un sous-type "C2-MSI" (19%) enrichi en tumeurs de type MSI (68%) et, comme déjà décrit, mutées pour BRAF (40%), présentant un phénotype CIMP (59%) et localisées dans le côlon proximal (72%). Ce sous-type présente de plus une surexpression des voies de l'immunité et du cycle cellulaire, consistant avec le fait que les cancers de types MSI ont la particularité d'être très cyclant et immunogène ;

un sous-type "C4-Cancer Stem Cell" (10%) enrichi également, à un niveau moindre, en tumeurs mutées pour BRAF (22%), présentant un phénotype CIMP (34%) et localisées dans le côlon proximal (57%) mais ne présentant pas d'enrichissement spécifique en tumeurs de type MSI. Ce sous-type présente un enrichissement en tumeurs surexprimant des signatures de cellules souches, des voies de signalisation liées à l'invasion, de la transition épithélio-mésenchymateuse (EMT) et du cycle cellulaire, et un enrichissement en tumeurs de stade IV.

un sous-type "C3-KRAS muté" (13%) très enrichi en tumeurs mutées pour KRAS (87%), légèrement pour CIMP (6%), localisées dans le côlon proximal (59%) et présentant une sous-expression des voies de l'EMT, de l'angiogenèse et des voies de l'immunité et une surexpression des voies du métabolisme.

un sous-type "C6-CIN norm-like" (10%) dont la majorité des tumeurs ont un profil d'expression très proche des profils des tissus normaux.

un sous-type "C5-CIN Wnt up" (27%) présentant une surexpression de la voie Wnt et des voies de l'angiogenèse.

un sous-type "C1-CIN Immune down" (21%) présentant globalement une sous expression des voies de signalisation liées au cancer par rapport aux autres tumeurs et en particulier la sous expression des voies de l'immunité.

Les 3 derniers sous-types présentent la caractéristique commune d'être très enrichis en tumeurs ayant une instabilité chromosomique et mutées sur le gène *TP53*. Ces 3 sous-types ne se distinguent pas sur le plan anatomo-clinico-moléculaire mais présentent toutefois des spécificités de dérégulation d'expression substantielles. L'évaluation de la dérégulation de la signature moléculaire de tumeurs provenant d'adénomes conventionnels et d'adénomes festonnés, associés à la signature KRAS muté-like, pouvant être attendu dans des tumeurs provenant d'adénomes festonnés, laisse penser qu'une origine différente des sous-types est possible, en définissant deux grands groupes : les sous-types non CIN (C4, C2, C3) et le sous-type CIN normal-like (C6) proviendraient préférentiellement de polypes de types festonnés, tandis que les deux derniers sous-types CIN (C1 et C5) proviendraient des adénomes dit conventionnels. L'étude des données génomiques n'a pas révélé d'association à des altérations spécifiques par sous-types. C'est le taux d'altérations, matérialisé par le statut CIN, qui distinguent le plus les différents sous-types. A l'exception des tumeurs C2 qui présentent peu d'instabilité du génome et plutôt des gains, les autres sous-types présentent des profils de fréquences d'altérations similaires. Les sous-types C4 et C3 présentent des fréquences d'altération un peu plus faibles, avec C3 ne présentant pas de gains

des 13q et 20q et de pertes du 18q. Enfin aucune association à la classification TNM n'a été retrouvée, seul un léger enrichissement en tumeurs métastatiques est retrouvé dans le groupe C4.

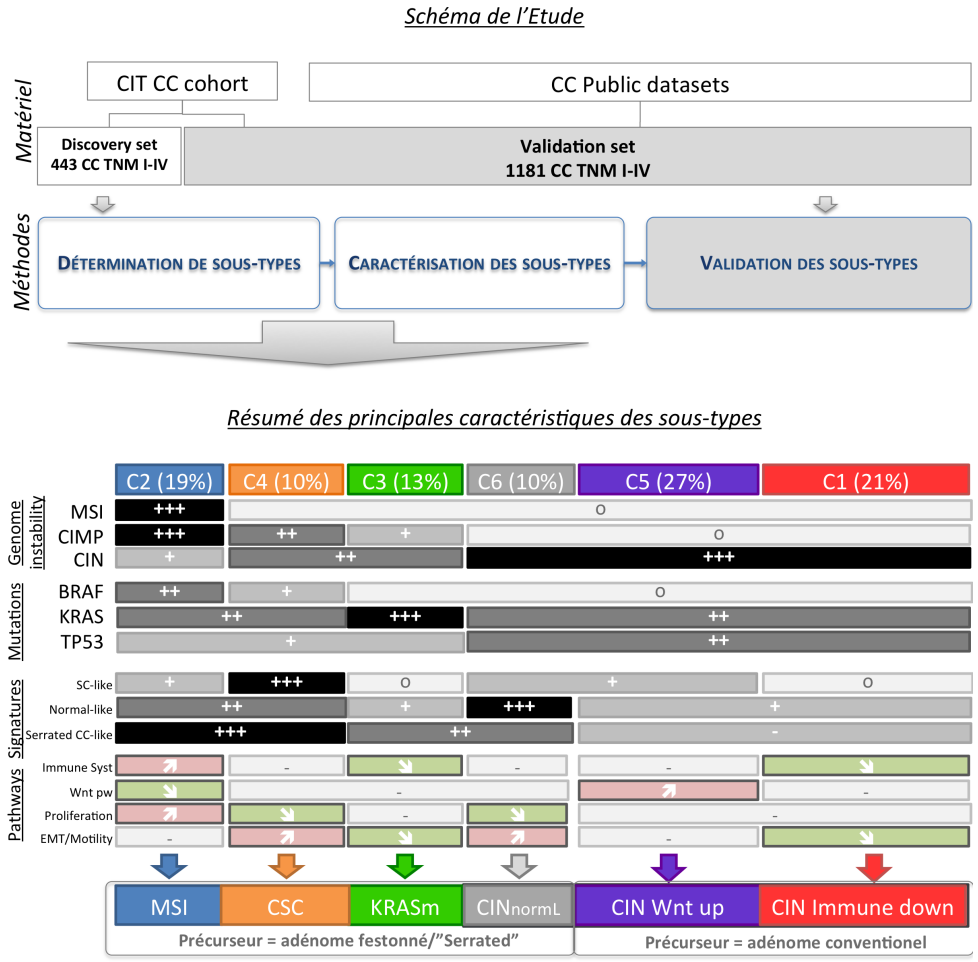


FIGURE 9 – Schéma de l'approche de classification et des caractéristiques des différents sous-types

Validation de la classification sur des données indépendantes Afin de valider l'existence de ces sous-types, j'ai ensuite collecté un grand nombre de données publiques (1058 tumeurs et 123 de notre cohorte gardées pour pouvoir valider les associations à des annotations non disponibles dans les bases de données publiques tels que les mutations et le statut CIMP), pour la majorité sur la même technologie afin de s'affranchir des biais de technologies, annotées pour la survie ou le statut MSI ou autres annotations d'intérêt. Ici, la difficulté a été de définir un prédicteur des sous-types afin d'assigner les tumeurs collectées aux six sous-types avec la contrainte de minimiser le nombre de gènes pour que cela soit plus facilement transférable en clinique. L'approche que j'ai défini a été de lister les gènes les plus spécifiques des sous-types ordonnés selon leur AUC et de construire l'ensemble des prédicteurs en prenant de 1 à 10 paires de gènes up et down régulés, puis d'évaluer l'ensemble des prédicteurs par une approche basée sur les centroïdes les plus proches et par cross-validation et de sélectionner le meilleur selon le taux de bon

classement (brièvement, le jeu de données est divisé en 10 sous-ensembles, le prédicteur est construit sur 9 sous-ensembles et testé sur le dernier, la procédure est répétée sur tous les sous-ensembles). Chaque sous-type présentait bien les mêmes profils d'expression et les mêmes caractéristiques anatomo-cliniques et moléculaires principales sur l'ensemble des données de validation.

Étude de la valeur pronostique de la classification Enfin, il est apparu intéressant de faire une analyse approfondie du pronostic en raison, d'une part, de l'intérêt en pratique clinique, et, d'autre part, pour évaluer les signatures pronostiques en intra-sous types. J'ai donc investigué le pronostic de ces sous types en comparaison des autres co-variables pronostiques et des prédicteurs moléculaires pronostiques qui ressortaient efficaces dans la littérature, OncotypeDX et ColoprintTM. L'étude a tout d'abord montré que les sous-types "C4-Cancer Stem Cell" et "C6-CIN normal-like" présentaient un plus mauvais pronostic. Le pronostic de la classification était de manière intéressante indépendante du stade TNM, seule variable anatomo-clinique pronostique. L'étude de la valeur pronostique du prédicteur OncotypeDX a montré que la classification apportait de l'information complémentaire au prédicteur et au stade TNM. Et enfin Oncotype DX prédisait la quasi-totalité des tumeurs "C4-Cancer Stem Cell" à haut risque de rechute mais ne permettait pas de séparer les patients de mauvais pronostic dans les sous-type C2 et C1.

OPEN ACCESS Freely available online

PLOS | MEDICINE

Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value

Laetitia Marisa[1], Aurélien de Reyniès[1], Alex Duval[2,3], Janick Selves[4], Marie Pierre Gaub[5,6], Laure Vescovo[1], Marie-Christine Etienne-Grimaldi[7], Renaud Schiappa[1], Dominique Guenot[5], Mira Ayadi[1], Sylvain Kirzin[4], Maurice Chazal[8], Jean-François Fléjou[2,3,9], Daniel Benchimol[10], Anne Berger[11], Arnaud Lagarde[12], Erwan Pencreach[5,6,13], Françoise Piard[14], Dominique Elias[15], Yann Parc[3,16], Sylviane Olschwang[12,17,18,19], Gérard Milano[7], Pierre Laurent-Puig[20]*¶, Valérie Boige[15,20]¶

1 "Cartes d'Identité des Tumeurs" Program, Ligue Nationale Contre le Cancer, Paris, France, **2** Unité Mixte de Recherche S938, Centre de Recherche Hôpital Saint-Antoine, INSERM, Paris, France, **3** Université Pierre et Marie Curie–Paris 6, Paris, France, **4** Unité Mixte de Recherche 1037, Centre de Recherche en Cancérologie de Toulouse, Université de Toulouse III, INSERM, Toulouse, France, **5** Unité de Recherche Physiopathologie et Médecine Translationnelle EA 4438, Université de Strasbourg, Strasbourg, France, **6** Laboratoire de Biochimie et Biologie Moléculaire, Hôpital de Hautepierre, Hôpitaux Universitaires de Strasbourg, Strasbourg, France, **7** Laboratoire d'Oncopharmacologie EA 3836, Centre Antoine Lacassagne, Nice, France, **8** Clinique Saint-George, Nice, France, **9** Service d'Anatomie et Cytologie Pathologiques, Hôpital Saint-Antoine, Assistance Publique-Hôpitaux de Paris, Paris, France, **10** Centre Hospitalier Universitaire de Nice, Nice, France, **11** Hôpital Européen Georges Pompidou, Paris, France, **12** Unité Mixte de Recherche S910, Faculté de Médecine La Timone, INSERM, Marseille, France, **13** Centre de Ressources Biologiques, Hôpitaux Universitaires de Strasbourg, Strasbourg, France, **14** Service de Pathologie, Centre Hospitalier Universitaire, Dijon, France, **15** Institut Gustave Roussy, Villejuif, France, **16** Service de Chirurgie Générale et Digestive, Hôpital Saint-Antoine, Assistance Publique-Hôpitaux de Paris, Paris, France, **17** Pôle DACCORD, Hôpital La Timone, Marseille, France, **18** Département d'Oncologie, Hôpital Clairval, Marseille, France, **19** Département de Gastroentérologie, Hôpital Ambroise Paré, Marseille, France, **20** Unité Mixte de Recherche S775, Paris Sorbonne Cité, Université Paris Descartes, INSERM, Paris, France

Abstract

Background: Colon cancer (CC) pathological staging fails to accurately predict recurrence, and to date, no gene expression signature has proven reliable for prognosis stratification in clinical practice, perhaps because CC is a heterogeneous disease. The aim of this study was to establish a comprehensive molecular classification of CC based on mRNA expression profile analyses.

Methods and Findings: Fresh-frozen primary tumor samples from a large multicenter cohort of 750 patients with stage I to IV CC who underwent surgery between 1987 and 2007 in seven centers were characterized for common DNA alterations, including *BRAF*, *KRAS*, and *TP53* mutations, CpG island methylator phenotype, mismatch repair status, and chromosomal instability status, and were screened with whole genome and transcriptome arrays. 566 samples fulfilled RNA quality requirements. Unsupervised consensus hierarchical clustering applied to gene expression data from a discovery subset of 443 CC samples identified six molecular subtypes. These subtypes were associated with distinct clinicopathological characteristics, molecular alterations, specific enrichments of supervised gene expression signatures (stem cell phenotype–like, normal-like, serrated CC phenotype–like), and deregulated signaling pathways. Based on their main biological characteristics, we distinguished a deficient mismatch repair subtype, a *KRAS* mutant subtype, a cancer stem cell subtype, and three chromosomal instability subtypes, including one associated with down-regulated immune pathways, one with up-regulation of the Wnt pathway, and one displaying a normal-like gene expression profile. The classification was validated in the remaining 123 samples plus an independent set of 1,058 CC samples, including eight public datasets. Furthermore, prognosis was analyzed in the subset of stage II–III CC samples. The subtypes C4 and C6, but not the subtypes C1, C2, C3, and C5, were independently associated with shorter relapse-free survival, even after adjusting for age, sex, stage, and the emerging prognostic classifier Oncotype DX Colon Cancer Assay recurrence score (hazard ratio 1.5, 95% CI 1.1–2.1, $p = 0.0097$). However, a limitation of this study is that information on tumor grade and number of nodes examined was not available.

Conclusions: We describe the first, to our knowledge, robust transcriptome-based classification of CC that improves the current disease stratification based on clinicopathological variables and common DNA markers. The biological relevance of these subtypes is illustrated by significant differences in prognosis. This analysis provides possibilities for improving prognostic models and therapeutic strategies. In conclusion, we report a new classification of CC into six molecular subtypes that arise through distinct biological pathways.

Please see later in the article for the Editors' Summary.

Citation: Marisa L, de Reyniès A, Duval A, Selves J, Gaub MP, et al. (2013) Gene Expression Classification of Colon Cancer into Molecular Subtypes: Characterization, Validation, and Prognostic Value. PLoS Med 10(5): e1001453. doi:10.1371/journal.pmed.1001453

Academic Editor: Christopher Kemp, Fred Hutchinson Cancer Research Center, United States of America

Received November 6, 2012; **Accepted** April 10, 2013; **Published** May 21, 2013

Copyright: © 2013 Marisa et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The Ligue Nationale Contre le Cancer, a non-governmental charity organization, through the Cartes d'Identité Tumeurs program, funded experiments from sample extraction to transcriptome and genome arrays, mutational characterization, and statistical analyses. The funder had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: YP gave expert advice to Coloplast (France) for less than 1,000 euros last year. The other authors declare that no competing interests exist.

Abbreviations: CC, colon cancer; CGH, comparative genomic hybridization; CIMP, CpG island methylator phenotype; CIN, chromosomal instability; CIT, Cartes d'Identité des Tumeurs; CRC, colorectal cancer; dMMR, deficient mismatch repair; GEP, gene expression profile; HR, hazard ratio; MMR, mismatch repair; MSI, microsatellite instability; NT, normal tissue; pMMR, proficient mismatch repair; RFS, relapse-free survival; TCGA, The Cancer Genome Atlas; TNM, tumor node metastasis.

* E-mail: Pierre.laurent-puig@parisdescartes.fr

¶ Pierre Laurent-Puig and Valérie Boige are joint senior authors on this work.

Introduction

Despite advances in screening, diagnosis, and treatment, colorectal cancer (CRC) is the third most common cancer and the fourth-leading cause of cancer death worldwide [1]. Pathological staging is the only prognostic classification used in clinical practice to select patients for adjuvant chemotherapy [2]. However, pathological staging fails to predict recurrence accurately in many patients undergoing curative surgery for localized CRC. In fact, 10%–20% of patients with stage II CRC, and 30%–40% of those with stage III CRC, develop recurrence. Among the molecular markers that have been extensively investigated for colon cancer (CC) characterization and prognosis, microsatellite instability (MSI), caused by defective function of the DNA mismatch repair (MMR) system, is the only marker that was reproducibly found to be a significant prognostic factor in early CRC in both a meta-analysis and a prospective trial [3,4]. Many studies have exploited microarray technology to investigate gene expression profiles (GEPs) in CRC in recent years, but no established signature has been found that is useful for clinical practice, especially for predicting prognosis [5–8]. GEP studies on CRC have been only poorly reproducible, possibly because CRC is composed of distinct molecular entities that may develop through multiple pathways on the basis of different molecular features [9–11]. As a consequence, there may be several prognostic signatures for CRC, each corresponding to a different entity. Indeed, GEP studies that include unsupervised hierarchical clustering, and integrated genetic/epigenetic analysis—including the more recent classification based on high-throughput methylome data [12]—have identified at least three distinct molecular subtypes of CC [7,9–13]. Therefore, CC should no longer be considered as a homogeneous entity. However, the molecular classification of CC currently used, which is based on a few common DNA markers (MSI, CpG island methylator phenotype [CIMP], chromosomal instability [CIN], and *BRAF* and *KRAS* mutations) [9–11], needs to be refined, and a standard and reproducible molecular classification is still not available.

In this study, we exploited a large, multicenter, and extensively characterized series of CC samples to establish a robust molecular classification based on genome-wide mRNA expression analysis. Then we assessed the associations between molecular subtypes and clinicopathological factors, common DNA alterations, and prognosis. To confirm the robustness of the subtypes obtained, we further validated our molecular classification in a large independent set.

Methods

Ethics Committee Approval

The use of the tumor collection was approved by the following ethics committees and institutional boards: lle de France II (2008-135; AFSSAP 2008-A01058-47), Marseille (PHRC2005, COS-IPC of 27 September 2007), Strasbourg (Comité Consultatif de Protection des Personnes dans la Recherche Biomedicale d'Alsace, 2004-63 and CPP-EST4 [DC-2009-1016 and AC-2008-438]), the Human Research Ethics Committee of Saint-Antoine Hospital (INCa; TUM0203—project 2010-1-RT-02), the Toulouse Hospital board (CRB–Cancer Toulouse, DC-2008-463, AC-2008-820, CPP2), and Nice (PHRC1997, CHUNice-948). The informed consent of the patients was recorded as required by a French law in force until 2007. Since the last inclusion in this study was 2007, the standard hospital blanket consent was considered sufficient.

Patients

The French national Cartes d'Identité des Tumeurs (CIT) program involves a multicenter cohort of 750 patients with stage I to IV CC who underwent surgery between 1987 and 2007 in seven centers. Fresh-frozen primary tumor tissue samples were retrospectively collected at the Institut Gustave Roussy (Villejuif), the Hôpital Saint Antoine (Paris), the Hôpital Européen Georges Pompidou (Paris), the Hôpital de Hautepierre (Strasbourg), the Hôpital Purpan (Toulouse), and the Institut Paoli-Calmettes (Marseille), and prospectively collected at the Centre Antoine Lacassagne (Nice). Patients who received preoperative chemotherapy and/or radiation therapy and those with primary rectal cancer were excluded from this study. Clinical and pathologic data were extracted from the medical records and centrally reviewed for the purpose of this study. Patients were staged according to the American Joint Committee on Cancer tumor node metastasis (TNM) staging system [2] and monitored for relapse (distant and/or locoregional recurrence; median follow-up of 51.5 mo). Patient and tumor characteristics are summarized in Table 1 and detailed in Table S1.

Of the 750 tumor samples of the CIT cohort, 566 fulfilled RNA quality requirements for GEP analysis (Figure S1). The 566 samples were split into a discovery set ($n = 443$) and a validation set ($n = 123$), well balanced for the main anatomoclinical characteristics (Table 1). The validation set also included 906 CC samples available from seven public datasets (GSE13067, GSE13294, GSE14333, GSE17536/17537, GSE18088, GSE26682, and GSE33113). These datasets corresponded to all available public datasets fulfilling the following criteria: available GEP data obtained using a similar chip platform (Affymetrix U133 Plus 2.0 chips) with raw data CEL files, and tumor location and either common DNA alteration ($n = 457$) and/or patient outcome ($n = 449$) data available. Within the discovery ($n = 443$) and the validation ($n = 1,029$) sets, 359 and 416 patients with stage II–III CC and documented relapse-free survival (RFS) were available for survival analysis, respectively (Figure S1). The dataset from The Cancer Genome Atlas (TCGA) [13], although obtained using a non-Affymetrix platform and therefore analyzed separately, was added to the validation set because of the extensive DNA alteration annotations provided for 152 CC samples.

Gene Mutations, MMR Status, and CIMP Analysis

The seven most frequent mutations in codons 12 and 13 of *KRAS* were assessed as previously described [14]. The *BRAF* c.1799T>A (p.V600E) mutation was assessed by allelic discrimination using TaqMan probes and the same protocol as that for

Table 1. Patient and tumor characteristics of the different sets.

Characteristics	CIT Cohort Patients ($n=750$)	CIT Discovery Dataset ($n=443$)	Validation Datasets		CIT Cohort *p*-Value	All Cohorts *p*-Value
			CIT ($n=123$)	Public ($n=906$)		
Mean age (sd, range), years	67 (14, 19–97)	67 (14, 22–97)	68 (12, 42–90)	68 (13, 23–95)	0.21	0.25
Sex (male/female) (percent)	429/321 (57/43)	237/206 (53/47)	73/50 (59/41)	347/330 (51/49)	0.24	0.24
TNM stage (percent)						
I	52 (7)	27 (6)	10 (8)	48 (11)	0.058	<0.001
II	351 (47)	198 (45)	66 (54)	205 (46)		
III	265 (35)	164 (37)	41 (33)	113 (25)		
IV	82 (11)	54 (12)	6 (5)	83 (18)		
NA	0	0	0	457		
Location (percent)						
Proximal	305 (41)	176 (40)	48 (39)	125 (51)	0.97	0.014
Distal	445 (59)	267 (60)	75 (61)	122 (49)		
NA	0	0	0	659		
Adjuvant chemotherapy[a] (percent)						
Yes	257 (42)[b]	161 (45)[b]	42 (40)[b]	91 (51)	0.42	0.31
No	357 (58)	200 (55)	64 (60)	87 (49)		
NA	2	1	6	140		
Median follow-up (sd, range), months	51.5 (37, 0–201)	50 (39, 0–201)	58 (37, 0–146)	48 (26, 0–143)	0.33	<0.001
Relapse[a] (percent)						
Yes	179 (29)	109 (30)	30 (29)	75 (24)	0.81	0.08
Distant/locoregional/both	149/23/7	83/22/4	29/0/1	—		
No	428 (71)	250 (70)	72 (71)	239 (76)		
NA	9	3	5	4		
dMMR (percent)	118/701 (17)	61/409 (15)	14/110 (13)	126/418 (30)	0.67	<0.001
CIMP+ (percent)	102/555 (18)	74/380 (19)	17/116 (15)	—	0.3	—
***KRAS*-mutant (percent)**	261/680 (38)	172/392 (41)	45/121 (37)	—	0.57	—
***BRAF*-mutant (percent)**	70/634 (11)	44/424 (11)	7/120 (6)	—	0.12	—
***TP53*-mutant (percent)**	226/451 (50)	135/245 (55)	55/106 (52)	—	0.66	—

p-Values are Chi-squared test *p*-values comparing the discovery and validation sets in the CIT cohort only and in all cohorts (excluding samples for which data were not available).
[a]Among patients with stage II–III CC.
[b]Only fluorouracil and folinic acid.
NA, not available; sd, standard deviation.
doi:10.1371/journal.pmed.1001453.t001

KRAS mutations. *TP53* mutations (exons 4–9) were assessed as previously described [15]. MSI was analyzed using a panel of five different microsatellite loci from the Bethesda reference panel [16]. MSI-high tumors were further classified as deficient MMR (dMMR), and both MSI-low and MSS tumors as proficient MMR (pMMR). CIMP status was determined using a panel of five markers (CACNA1G, IGF2, NEUROG1, RUNX3, and SOCS1) as previously described [17]. Experimental procedures are detailed in Text S1. Common DNA alterations are summarized in Table 1 and detailed in Table S1.

Gene Expression Analysis

The GEP of 566 primary CC samples were determined on Affymetrix U133 Plus 2.0 chips. For 19 patients, adjacent non-tumor tissue (normal tissue [NT]) was also available and was tested. The methods used for RNA purification, quality control, fluorescent probe production, hybridization, and raw data processing were as previously described [18]. Each dataset was normalized independently in batches using the robust multi-array average method implemented in the R package affy [19]. For the CIT dataset, residual technical batch effects were corrected using the ComBat method implemented in the SVA R package [20]. Data are available via the NCBI Gene Expression Omnibus (http://www.ncbi.nlm.nih.gov/geo/; accession number GSE39582).

Array-Based Comparative Genomic Hybridization Analysis

A total of 464 of the 750 primary CC samples from the CIT cohort could be analyzed for array-based comparative genomic hybridization (CGH) on a BAC array containing 4,434 bacterial artificial chromosome clones with a median resolution of 0.6 Mb.

DNA labeling, hybridization, and data processing were as previously described [21]. CIN was defined from CGH profiles: samples with at least 20% gain or loss of whole chromosomes or fractions of chromosomes were scored as CIN+ (see Text S1 for details).

Unsupervised Subtype Discovery Based on Gene Expression Analysis

Unsupervised classification of the discovery set was performed using hierarchical clustering (Ward linkage and 1 − Pearson correlation coefficient distance used) on the most variant class of probe sets ($n = 1,459$). To obtain a robust classification, we used a consensus unsupervised approach [22] implemented in the R package ConsensusClusterPlus. The consensus clusters were obtained from 1,000 resampling iterations of the hierarchical clustering, by randomly selecting a fraction of the samples and of the most variant probe sets (90%). The optimal number of clusters was selected according to the approach criteria detailed in Text S1.

Validation Set Subtype Assignment

Validation datasets were independently assigned to GEP subtypes according to a standard distance-to-centroid approach [23]. A centroid-based predictor was built by a 10-fold cross-validation approach, resulting in the selection of the five top up-regulated and five top down-regulated genes specific to each subtype, yielding 57 genes (three genes were shared by two subtypes). The approach was implemented in the R package citccmst, and is detailed in Text S1.

Molecular Subtype Characterization

The Chi-squared test and logistic regression were used to study associations between anatomoclinical features, common DNA alterations, and subtypes. Each molecular subtype was further characterized according to (i) GEP of NT counterparts from our dataset; (ii) previously published supervised signatures based on intestinal stem cell phenotype [24,25], *BRAF* mutation [26], and serrated CRC phenotype [27], as described in Text S1; (iii) cancer-relevant signaling pathways retrieved from the Kyoto Encyclopedia of Genes and Genomes (see Text S1); and (iv) CGH alteration frequencies.

Recurrence Risk Group Assignment according to Other Molecular Predictors

The ColoPrint and Oncotype DX prognostic classifiers [7,8] were adapted and applied to our overall datasets as described in Text S1.

Survival Analysis

Survival analysis was intentionally restricted to the subgroup of patients with stage II–III tumors because reliable prognostic biomarkers are most needed for these patients. Indeed, most stage I patients will not derive benefit from adjuvant chemotherapy because of their excellent prognosis after curative surgery, and most stage IV patients, already metastatic, will die from their disease and therefore should be analyzed independently for progression-free survival. RFS was defined as the time from surgery to the first recurrence and was censored at 5 y. Survival was analyzed according to the Kaplan-Meier method, and differences between survival distributions were assessed with the log-rank test. Univariate and multivariate models were computed using Cox proportional-hazards regression (R package survival) (see Text S1 for details).

Results

Unsupervised Analysis of Gene Expression Profiles Revealed Six Subtypes of Colon Cancer

Consensus unsupervised analysis of the GEP data from the 443 samples of the discovery set revealed six clusters of samples based on the most variant probe sets ($n = 1,459$): C1 ($n = 95$, 21%), C2 ($n = 83$, 19%), C3 ($n = 56$, 13%), C4 ($n = 46$, 10%), C5 ($n = 118$, 27%), and C6 ($n = 45$, 10%) (Figure 1; Table S2). The consensus matrix showed that C2, C3, C4, and C6 appeared as well-individualized clusters, whereas there was more classification overlap between C1 and C5 (Figure 1A). Based on cluster expression centroid classification and the gene expression heatmap (Figure 1B and 1C), cluster C4 appeared to be the most distinct. The other clusters subdivided into C2 and C3 on one side of the cluster expression centroid classification (Figure 1B), and C6, C5, and C1 on the other. The GEPs of C1 and C5 showed overlap but displayed slightly distinct gene deregulations. This was confirmed in the supervised selection of the cluster-discriminant probe sets shown in the gene expression heatmap in Figure S2 and detailed in Table S3.

Clinical and Molecular Relevance of Colon Cancer Subtypes

Associations with anatomoclinical and DNA alterations data are shown in Figures 1C and S3A and in Table S4. Tumors classified as C1, C5, and C6 were more frequently CIN+, CIMP−, *TP53*-mutant, and distal ($p<0.001$), without any other molecular or clinicopathological features able to discriminate these three clusters clearly. Tumors classified as C2, C4, and C3 were more frequently CIMP+ (59%, 34%, and 18%, respectively, versus <5% in other clusters) and proximal. C2 was enriched for dMMR (68%) and *BRAF*-mutant tumors (40%). C3 was enriched for *KRAS*-mutant tumors (87%). No association between clusters and TNM stage was found, except enrichment for metastatic (31%) tumors in C4.

The analyses of CGH arrays revealed that CIN+ samples shared a typical DNA copy alteration pattern including +7, −8p, +8q, +13q, −17p, −18, +20q. Differences between subtypes mainly reflected their relative content of CIN+ samples. However, some specific alterations were observed for the two CIN subtypes, C5 (+2, +11, +17q) and C1 (−10q, −14q, −15q) (Figure S4).

Signaling Pathways Associated with Colon Cancer Subtypes

We analyzed cancer-related signaling pathways from the Kyoto Encyclopedia of Genes and Genomes database for specific deregulation in each subtype signature (Figure 2). As expected, up-regulated immune system and cell growth pathways were found in C2, the subtype enriched for dMMR tumors. C4 and C6 both showed down-regulation of cell growth and death pathways and up-regulation of the epithelial–mesenchymal transition/motility pathways. Most signaling pathways were down-regulated in C1 and C3. In C5, cell communication, Wnt, and metabolism pathways were up-regulated. In C1, cell communication and immune pathways were down-regulated.

Exploratory Analysis of Cell and Precursor Origins of the Subtypes

These six molecular subtypes were further investigated using GEP data from NT and previously published supervised signatures based on DNA alterations and cellular phenotypes to explore the subtype origins. Based on the growing amount of data suggesting

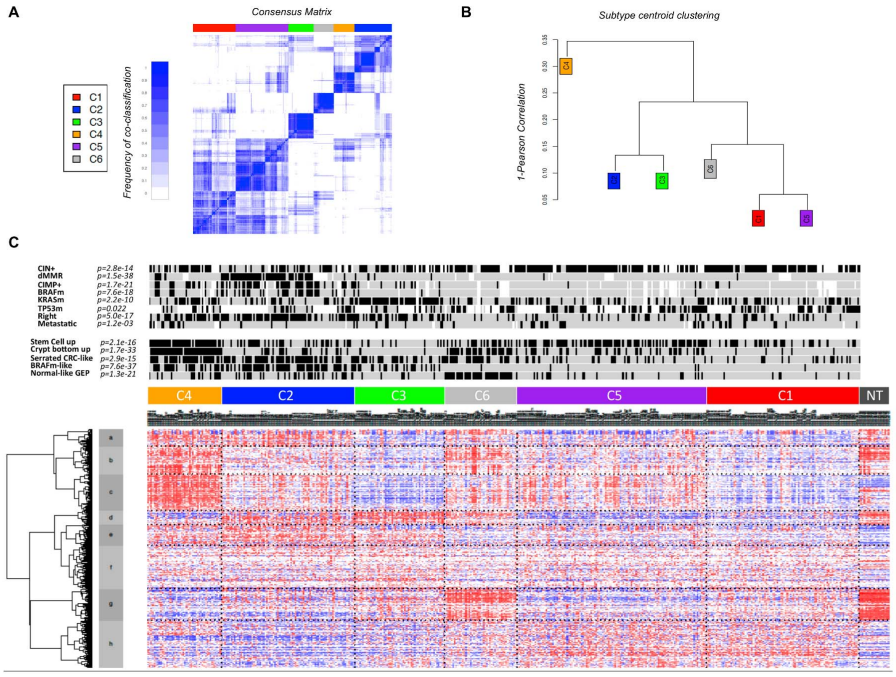


Figure 1. Unsupervised gene expression analysis of the discovery set of 443 colon cancers. (A) Consensus matrix heatmap defining six clusters of samples for which consensus values range from 0 (in white, samples never clustered together) to 1 (dark blue, samples always clustered together). (B) Distance between clusters according to the hierarchical clustering of the 1,459 probe sets based on the centroids of each cluster. (C) GEP heatmap of the 1,459 probe sets ordered by subtype, with annotations associated with each subtype.
doi:10.1371/journal.pmed.1001453.g001

that cancer is closely linked to stem cells, a mouse-derived intestinal stem cell signature [24] and a human colon top and bottom crypt signature were selected and applied to our GEP data [25]. C4 appeared highly enriched for tumors displaying "stem cell phenotype–like" GEPs (91%) and up-regulating of the bottom crypt signature (96%). (Figure S3A). This finding was consistent with the pathways specifically deregulated in C4 (cell cycle pathway down-regulated and cell communication pathway up-regulated).

As previously described for breast cancer [23], we also investigated the existence of a "normal-like" subtype using the GEP centroid from NT samples. C6 was enriched for normal-like GEP tumors, although 86% of them were CIN+.

Serrated CC, in contrast to conventional CC, may arise through a recently introduced serrated neoplasia pathway [27]. We therefore applied the supervised signature, described by Laiho et al. [27], comparing gene expression of serrated to conventional CC to our GEP data. Most of the tumors classified as C2, C3, C4, and C6 displayed a "serrated CC phenotype–like" GEP, whereas those in C1 and C5 displayed a "conventional CC phenotype–like" GEP. A strong association between *BRAF* mutations and the serrated adenoma pathway has been reported [28], and a *BRAF*-mutant-like supervised signature has been described by Popovici et al. [26] that identifies a *BRAF* wild-type subgroup, 30% of which were *KRAS* mutants and 13% of which were double wild-type CC. This signature was also applied to our GEP data: subtypes C2, C3, and C4 were enriched in *BRAF*-mutant-like GEP tumors.

A schematic summary of the subtype characteristics is shown in Figure 3. The six subtypes were named according their main respective biological characteristic as follows: C1, "$CIN_{Immune\text{-}Down}$"; C2, "dMMR"; C3, "KRASm" (for "*KRAS*-mutant"); C4 "CSC" (for "cancer stem cell"); C5, "CIN_{WntUp}"; and C6, "CIN_{normL}".

Validation of the Subtypes across Nine Colon Cancer Datasets

To validate our six-subtype classification, a 57-gene centroid classifier was built from the discovery set by a 10-fold cross-validation approach (<5% misclassification; Figure S5; Table S5). We applied this signature to the Affymetrix validation set of 1,029 samples (Table 1). All subtypes were found in the same proportions as in the discovery set, and the main associations between the different clusters and anatomoclinical/DNA/GEP characteristics described above were confirmed (Figures S2B and S3B), except for the enrichment of C4 with *BRAF*-mutant and stage IV tumors. When applied to the Agilent TCGA dataset ($n = 152$) [13], the molecular and clinical characteristics of the subtypes were all confirmed (Figure S3C). To further validate the six-subtype

Category		Pathways	C2	C4	C3	C6	C5	C1
Cell communication		KEGG Focal adhesion	3.9E-02	4.40E-16	3.7E_14		3.1E-04	3.9E-11
		KEGG ECM-receptor interaction		4.90E-14	2.6E-14	2.7E-02	3.7E-05	1.2E-12
		KEGG Tight junction		2.8E-03	2.4E-03			2.6E-02
		KEGG Gap junction	3.5E-03	1.3E-03	1.4E-02			4.4E-02
Cell growth and death		KEGG Cell cycle	1.6E-04	2.7E-06		4.8E-14		
		KEGG p53 signaling pathway	3.4E-03	2.6E-02		2.5E-07		
		KEGG Apoptosis	1.9E-02					
Immune system		KEGG Antigen processing and presentation	7.20E-20		6.2E-05			2.8E-11
		KEGG Toll-like receptor signaling pathway	5.1E-09	1.8E-02	2.5E-03	4.5E-02		5.8E-07
		KEGG Hematopoietic cell lineage	4.8E-06	7.5E-03	1.2E-05	1.7E-02		1.1E-07
		KEGG Natural killer cell mediated cytotoxicity	1.3E-09					1.2E-02
		KEGG T cell receptor signaling pathway	3.0E-03					
Motility		KEGG Axon guidance		1.8E-02	7.1E-03			9.8E-03
		KEGG Regulation of actin cytoskeleton	2.5E-02	5.3E-06	4.0E-04			7.1E-04
		GO epithelial to mesenchymal transition	4.2E-03	1.3E-02	1.6E-03	2.4E-02		1.7E-02
		GO regulation of epithelial to mesenchymal transition	5.4E-04	2.4E-04	1.1E-04	9.8E-03	3.7E-02	9.2E-04
		GO mesenchyme development	2.2E-02	3.5E-03	5.3E-06	1.3E-03		8.3E-03
Replication and Repair		KEGG DNA replication		3.4E-02		7.9E-06		
		KEGG Mismatch repair				3.8E-02		
Signal transduction		KEGG Wnt signaling pathway	4.9E-05				5.9E-03	
		GO Wnt-protein binding		4.5E-05	2.0E-04	2.8E-02	8.6E-05	2.9E-03
		GO Wnt receptor signaling pathway	2.0E-02	9.8E-07	1.2E-09		3.5E-03	6.6E-05
		KEGG TGF-beta signaling pathway		5.3E-04	7.6E-03			8.7E-03
		KEGG MAPK signaling pathway	2.5E-02	3.5E-03	1.4E-02		3.3E-02	7.3E-03
		KEGG Hedgehog signaling pathway	1.6E-02			2.6E-02		
		KEGG VEGF signaling pathway	2.2E-02					
Angiogenesis		KEGG Renin-angiotensin system			4.4E-03		1.4E-02	6.5E-03
		GO sprouting angiogenesis	3.5E-05	9.1.E-16	5.5E-18		2.8E-03	1.9E-11
		GO angiogenesis			1.2E-02		7.3E-03	4.2E-02
Metabolism	Carbohydrate	KEGG Pentose and glucuronate interconversions		3.1E-14	3.2E-12	6.7E-15	2.0E-15	1.1E-02
		KEGG Fructose and mannose metabolism	4.3E-02	1.1E-03	2.9E-02		1.4E-03	
		KEGG Starch and sucrose metabolism		5.3E-12	3.1E-08	8.6E-13	3.3E-12	5.5E-03
		KEGG Butanoate metabolism		8.3E-03	6.8E-05			2.9E-02
	Energy	KEGG Nitrogen metabolism	1.2E-02	1.3E-04	4.8E-06	6.8E-04	2.7E-08	
		KEGG Androgen and estrogen metabolism		5.5E-13	4.9E-10	2.5E-11	2.1E-14	3.9E-03
	Lipid	KEGG Arachidonic acid metabolism	3.1E-03	4.8E-04	3.7E-03	5.8E-03	9.2E-04	6.8E-03
		KEGG Linoleic acid metabolism	4.4E-03	3.9E-04	2.1E-04	1.1E-02	2.5E-06	
	Xenobiotics	KEGG Metabolism of xenobiotics by cytochrome P450		2.7E-14	1.2E-12	2.8E-15	2.5E-13	1.1E-02

Legends

Enrichment p-value in	up regulated genes	down regulated genes
[0.01; 0.05[		
[0.01: 1E-04[		
[1E-04;1E-10[		
<1.0E-10		

Figure 2. Signaling pathways associated with each molecular subtype. The enrichment of Kyoto Encyclopedia of Genes and Genomes (KEGG) and GeneOntology (GO) pathways and gene sets related to cancer hallmarks was tested in each subtype signature (1,000 top differentially up- and down-expressed genes, separately). The hypergeometric test *p*-values for enrichment in up- and down-regulated signatures are indicated in red and green, respectively. ECM, extracellular matrix.
doi:10.1371/journal.pmed.1001453.g002

classification in the validation dataset, we performed the same consensus clustering approach with the whole validation set; the subtypes generated were highly concordant with the six assigned subtypes (Chi-squared test, $p<10^{-16}$).

Prognostic Value of the Six-Subtype Classification

Further investigation of the clinical relevance of our classification included a prognostic analysis based on RFS restricted to stage II and III tumors. The prognosis of each of our six subtypes in the discovery set ($n=359$) differed, but not significantly so, with patients whose tumors were classified as C4 and C6 having a relatively poorer outcome (5-y RFS rates of 52% and 61%, respectively, compared to 70%, 77%, 65%, and 70% for C1, C2, C3, and C5, respectively; $p=0.18$) (Figure 4A). The prognostic value of the six-subtype classification was significant in the validation set ($n=416$) ($p=0.0009$), with a worse prognosis confirmed for patients with C4 and C6 tumors (Figure 4B); The six-subtype classification was also significant for the discovery and the validation sets combined ($p=0.0003$) (Figure 4C). To compare the prognostic value of our classification to other prognostic covariates, we recoded our classification by combining C4 and C6 into a single high-risk group, versus all other subtypes as the low-risk group. This binary classification led to an even stronger association of the high-risk group versus the low-risk group with RFS (hazard ratio [HR] 1.7, 95% CI 1.1–2.6, $p=0.014$, in the discovery set; HR 2.3, 95% CI 1.5–3.5, $p=0.00012$, in the validation set; HR 2, 95% CI 1.5–2.7, $p=7.1\times10^{-6}$, in the overall dataset) (Figures 4D and S6) and remained an independent prognostic factor, together with TNM stage, in the multivariate analysis (discovery and validation sets analyzed separately and merged) (Tables 2 and S6). The binary classification also remained an independent prognostic factor ($p<0.01$) when common DNA alterations (MMR status, CIMP, and *BRAF* and *KRAS* mutations) were added to the model (Table S7).

Prognostic Classifiers within Subtypes

The Oncotype DX recurrence score [8] is an emerging prognostic classifier, and we attempted to assess its prognostic value with our data. This score had prognostic value in both the discovery and validation sets, and in the overall dataset ($p=3.4\times10^{-6}$; Figure S6). In particular, 97% of the C4 samples were classified as high risk by the Oncotype DX score. However, this score was not prognostic for all of the subtypes (Figure S7). In a multivariate stepwise analysis, both our recoded classification and the Oncotype DX score remained independently prognostic, together with TNM stage (Table 2).

We also attempted an exploratory analysis of the signature described by Salazar et al. [7] by investigating 17 of the 18 probe sets available on the Affymetrix U133 Plus 2.0 chips. We found no significant prognostic value of this 17-gene expression signature in our series (Figure S6).

Discussion

Using a large comprehensively characterized multicenter cohort of CC patients, we identified six robust molecular subtypes of CC

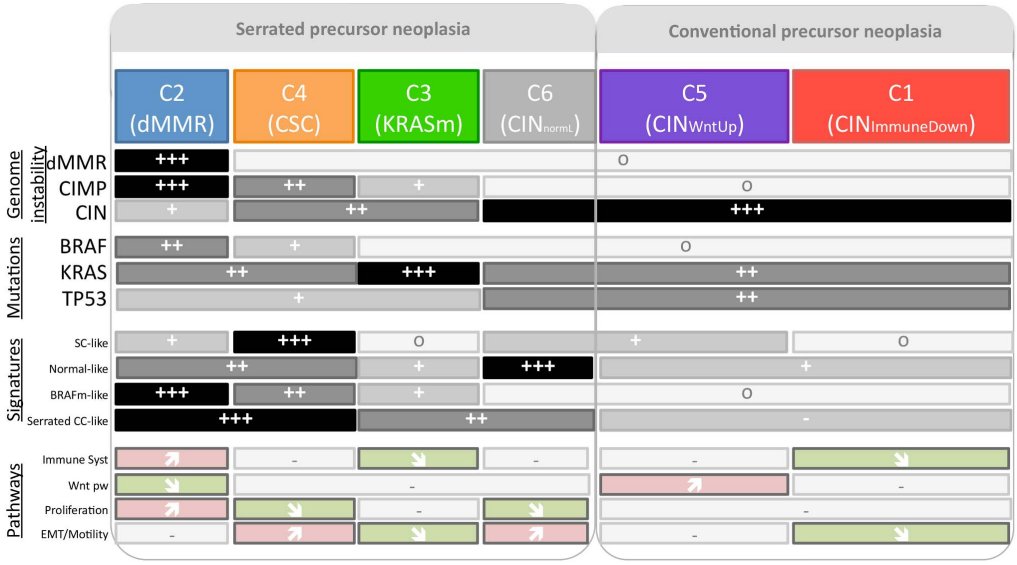


Figure 3. Summary of the main characteristics of the six subtypes. Symbols correspond to the relative frequency within the subtype (o: very low frequencies [~0%]; +++: very high frequencies; +/++: intermediate frequencies), and arrows indicate significant enrichment of subtype up- and down-regulated genes in most of the pathways of the given category. EMT, epithelial–mesenchymal transition; SC, stem cell; Wnt pw, Wnt pathway.
doi:10.1371/journal.pmed.1001453.g003

individualized by distinct clinicobiological characteristics. Importantly, this six-subtype classification was validated in nine independent datasets. Furthermore, classification into high- and low-risk subtypes was of prognostic value.

Although retrospective, our cohort was very representative of the clinicopathological characteristics and common DNA alteration frequencies observed in the population of patients with CC.

Our findings clearly demonstrate that anatomoclinical factors and common DNA alterations alone are helpful for highlighting subtype characteristics, but they are not sufficient to define boundaries between subtypes and to describe the molecular heterogeneity of CC. Our classification successfully identified the dMMR tumor subtype, and also individualized five other distinct subtypes among pMMR tumors, including three CIN+ CIMP− subtypes representing slightly more than half of the tumors. As expected, mutation of *BRAF* was associated with the dMMR subtype, but was also frequent in the C4 CIMP+ poor prognosis subtype. *TP53*- and *KRAS*-mutant tumors were found in all the subtypes; nevertheless, the C3 subtype, highly enriched in *KRAS*-mutant CC, was individualized and validated, suggesting a specific role of this mutation in this particular subgroup of CC. There was no significant association between our classification and pathological stage, suggesting that tumor subtype is established at the initial stages.

Exploratory analysis of each subtype GEP with previously published supervised signatures and relevant deregulated signaling pathways improved the biological relevance of the classification. Indeed, this analysis suggested that different types of CC may arise from distinct cell origins, and distinguished between the two main pathways, defined as the serrated and the conventional precursor neoplasia pathways. Interestingly, we not only individualized the dMMR subtype among the serrated precursor neoplasia subtypes, but also within the C4 CSC and the C3 KRASm subtypes. This finding is consistent with the serrated polyp classification showing two main groups: the sessile serrated adenomas, commonly associated with dMMR tumors, and the traditional serrated adenomas, commonly associated with *KRAS*-mutant tumors [29]. However, the proportion of serrated precursor neoplasia tumors that we found was higher than expected, indicating that further pathological investigations are required.

Another interesting finding is the reproducible association between the stem cell signature and the poor prognosis C4 subtype. Almost half of the top genes deregulated in C4—including *secreted frizzled-related protein 2 (SFRP2)*, described as a key factor in stem cell regulation [30] and belonging to the Frizzled gene family, and *growth arrest-specific 1 (GAS1)*—were included in the poor prognosis cluster signature reported by Oh et al. [31]; these genes may therefore be markers of the aggressiveness of CC cells and may constitute potential therapeutic targets.

The C6 CIN_{normL} subtype was more difficult to characterize; it belongs to the CIN+ subgroup but has a GEP and RFS that are distinct from those of the other two CIN subtypes. Several genes up-regulated in C6, in particular *carbonic anhydrase II (CA2)* and *solute carrier family 4, sodium bicarbonate cotransporter, member 4 (SLC4A4)*, were also included in the prognostic classifier described by Lin et al [32].

The two other CIN subtypes, C1 and C5, were more difficult to distinguish from each other. They show common clinical and DNA alteration characteristics. They share some gene expression patterns, leading to lower co-classification rates than for the other subtypes. Moreover, these two subtypes are combined if the number of clusters is set to five instead of six. As a result, the

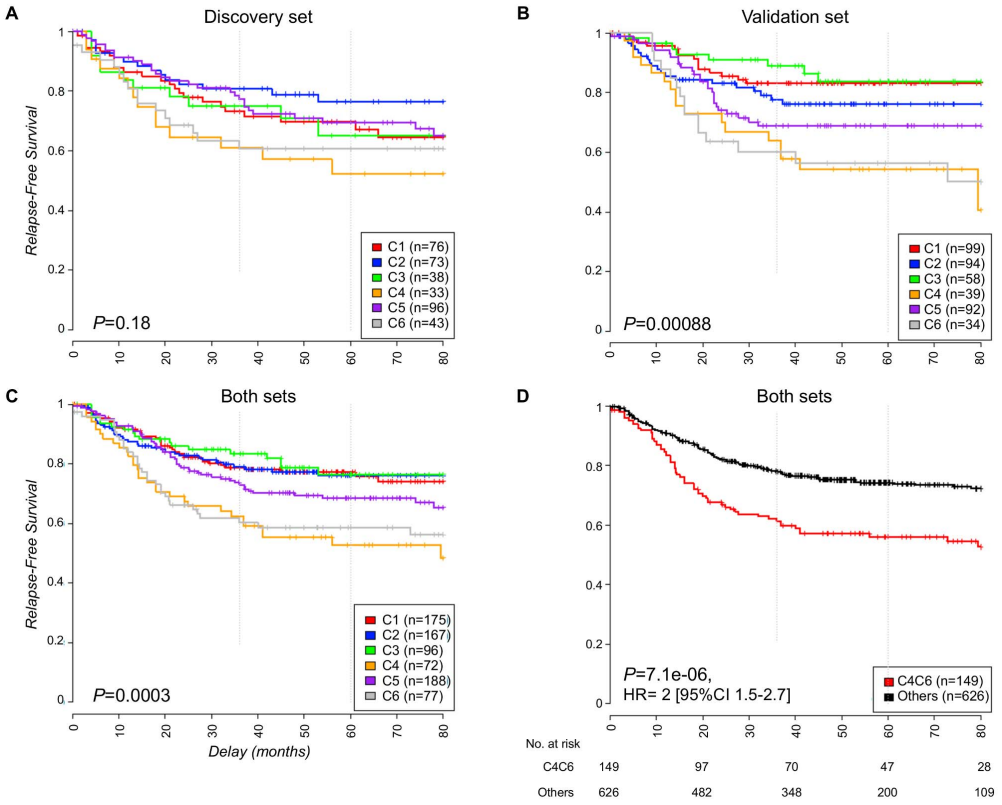


Figure 4. Kaplan-Meier relapse-free survival. This figure shows RFS in (A) the discovery dataset, (B) the validation dataset, (C) the overall dataset, and (D) the overall dataset for C4 and C6 subtypes combined versus the other subtypes; the numbers at risk on the time axis are given.
doi:10.1371/journal.pmed.1001453.g004

division of C1 and C5 into two distinct subtypes can be questioned (Figure S8). However, the C1 and C5 subtypes are also clearly associated with distinct gene expression signatures (Table S3; Figure S2) and display specific pathway deregulation (immunity and epithelial–mesenchymal transition pathways; Figure 2). In addition, only four out of 507 samples in the validation set classified as subtype C1 or C5 had a mixed assignment C1/C5, as a result of being close to both the C1 and C5 centroids (see Text S1). Altogether, these observations supported these two clusters being representative of two distinct molecular entities.

The biological relevance of our subtypes was highlighted by significant differences in prognosis. In our unsupervised hierarchical clustering, patients whose tumors were classified as C4 or C6 had poorer RFS than the other patients. Thus, our study, like others [7,31], supports the idea that the unsupervised analysis of transcripts in primary tumors yields information of prognostic value. The prognostic value of our signature was statistically significant in the validation and the overall datasets, independently of TNM stage, with a worse prognosis confirmed for C4 and C6 subtypes. Subtype C4 was enriched in CIMP+ *BRAF*-mutant tumors and may correspond to the poor prognostic cluster reported by Salazar et al. containing the same proportion of *BRAF*-mutant tumors [7].

Prognostic analyses based solely on common DNA alterations can distinguish between risk groups, but are still inadequate, as most CCs are pMMR CIMP− *BRAF*wt (75% in our series; data not shown). Indeed, the markers *BRAF*-mutant, CIMP+, and dMMR may be useful for classifying a small proportion of cases, but are uninformative for a large number of patients. This was illustrated in the study by Salazar et al. in which *BRAF* mutation was found in both good and poor outcome clusters, but was rare in the intermediate prognosis cluster used to build the ColoPrint prognostic classifier [7].

The ColoPrint and Oncotype DX prognostic classifiers were developed recently to improve risk prediction in early-stage CRC [7,8]. ColoPrint was validated in three independent datasets of stage II–IIIA CC, and the robustness of the signature is currently being evaluated prospectively [33,34]. The corresponding 17 probe sets available on Affymetrix chips did not identify risk groups in our series (data not shown). Oncotype DX has been validated as a prognostic score in the QUASAR and CALGB9581 trials, and more recently in an independent cohort of patients with

Table 2. Univariate and multivariate analyses of relapse-free survival according to clinical annotations, the six-subtype classification, and the Oncotype DX prognostic classifier in the overall dataset.

Variables	Value[a]	Univariate Analysis						Multivariate Model 1					Multivariate Model 2				
		n	*n* Event	HR	95% CI	Modality *p*-Value (Wald)	Model *p*-Value (Log-Rank)	*n*	HR	95% CI	Modality *p*-Value (Wald)	Model *p*-Value (Log-Rank)	*n*	HR	95% CI	Modality *p*-Value (Wald)	Model *p*-Value (Log-Rank)
TNM stage[b,c] (ref = II)	III	775	202	2	1.5–2.6	0.0000011	0.00000064	775	1.9	1.4–2.5	0.0000077	1.2×10^{-09}	775	1.8	1.4–2.4	0.000022	6.0×10^{-11}
CIT classification recoded[b,c] (ref = others)	C4C6	775	202	2	1.5–2.7	0.000011	0.0000071		1.8	1.3–2.5	0.00011			1.5	1.1–2.1	0.0097	
CIT classification[b] (ref = C1)	C2	775	202	1.1	0.66–1.7	0.83	0.0003										
	C3	775	202	0.94	0.54–1.6	0.83											
	C4	775	202	2.3	1.4–3.8	0.00063											
	C5	775	202	1.4	0.89–2.1	0.15											
	C6	775	202	2.1	1.3–3.4	0.0031											
Tumor location (ref = distal)	Proximal	623	173	0.85	0.62–1.1	0.29	0.29										
Sex[b] (ref = female)	Male	775	202	1.2	0.88–1.5	0.3	0.3										
Age[b]	—	774	202	1	0.99–1	0.84	0.84										
Oncotype DX recurrence score[c] (ref = low risk)	High risk	775	202	1.9	1.4–2.5	0.0000050	0.0000034							1.6	1.2–2.1	0.0027	

Analyses of RFS were performed using Cox regression. The multivariate models reported correspond to the best multivariate models obtained using a backward–forward selection procedure (R function step). Multivariate model 1 included all given clinical annotations (except tumor location, which was less well filled in) and the classification. Multivariate model 2 included only variables of prognostic interest, i.e., TNM stage, the CIT recoded classification, and the Oncotype DX classifier. Only samples for which all the variables were available were included in multivariate models.

[a]Value indicates the modality of the annotation associated with the HR.

[b]Variables used in multivariate model 1.

[c]Variables used in multivariate model 2.

ref, reference.

doi:10.1371/journal.pmed.1001453.t002

stage III CC [35–37]. Although not identified by a genome-wide gene expression approach, the Oncotype DX score's prognostic value was confirmed in our overall stage II–III CC dataset but not in every subtype: it had prognostic value for the C3, C4, and C6 subtypes, and marginally in C5; it did not have prognostic value in C1 and C2, which represent 44% of our overall dataset. Our classification added prognostic information that remained significant in the multivariate analysis adjusted for TNM and Oncotype DX score. This suggests that the "one size fits all" prognostic signature approach can be difficult to apply because of the heterogeneity of CC. This may explain, in part, the poor concordance of GEP prognostic signatures in CC [38].

Our multivariate analysis has some limitations. In particular, some established predictors of CC prognosis, notably tumor grade and number of nodes examined, were not included because this information was not available for a substantial proportion of cases [39]. Thus, the significance and robustness of the signature as a prognostic classification requires further confirmation, ideally with large prospective patient cohorts included in adjuvant trials.

In conclusion, we report a new classification of CC into six robust molecular subtypes that arise through distinct biological pathways and represent novel prognostic subgroups. Our study clearly demonstrates that these gene signatures reflect the molecular heterogeneity of CC. This classification therefore provides a basis for the rational design of robust prognostic signatures for stage II–III CC and for identifying specific, potentially targetable markers for the different subtypes.

Supporting Information

Figure S1 Discovery and validation sets used in the study. The data used in this study were collected from the CIT program cohort (a French multicenter cohort) and from publicly available datasets. There were 750 CC samples from the CIT program suitable for common DNA alteration characterization, and 566 of these provided tumor RNA samples satisfying stringent quality control criteria. These RNA were hybridized on an Affymetrix chip (asterisk) and used for molecular subtype determinations. The discovery set was composed of 443 tumors from the CIT cohort. The validation set was composed of the remaining CIT cohort CC samples, CC samples from seven Affymetrix publicly available datasets (indicated with their NCBI GEO accession number), and CC samples from the non-Affymetrix TCGA program (performed on an Agilent platform). For survival analyses, only stage II and III patients were considered, stage I and IV patients not being informative as almost all survive or die, respectively; there were thus 359 cases in the CIT discovery set and 416 in the CIT validation set and three public datasets included in this analysis. pbs, probe sets.
(PDF)

Figure S2 Heatmaps of subtype-discriminant probe set expression profiles in the discovery set and in the Affymetrix validation set. (A) Heatmap of the discovery set samples ordered according to gene hierarchical clustering (1 − Pearson metric, Ward linkage) and by subtypes. (B) Heatmap of Affymetrix validation set samples ordered as in (A). For each subtype, discriminant probe sets were selected from the discovery set using a moderated *t*-test, comparing the given subtype to the other subtypes, with an adjusted $p<10^{-5}$ and a |log fold change|>0.5, yielding 1,108 discriminant probe sets.
(PDF)

Figure S3 Associations between molecular subtypes and anatomoclinical characteristics, DNA alterations, and supervised signature annotations in the discovery and validation sets. Associations were assessed in (A) the Affymetrix discovery set, (B) the Affymetrix validation set, and (C) the TCGA, non-Affymetrix, validation set. For each subtype and variable, the proportion of each modality is represented (dark grey: "positive/true/yes" proportion; white: "negative/false/no" proportion; grey: "data not available" proportion), and the percent of the main feature (dark grey) within each subtype is indicated. The Chi-squared test *p*-values are indicated in red.
(PDF)

Figure S4 Subtype genomic alteration profiles along the genome. CC molecular subtypes present different copy-number-change profiles. The profiles were established using genome-wide array-based CGH available for 356 samples. (A) Frequencies of gains (frequency>0) and losses (frequency<0) observed at a given location on the genome are shown for all samples (first row; darker bars are loci with an alteration frequency higher than 20%) and by subtype (darker bars are significantly differentially altered regions, displayed in [B]). (B) Subtype-specific genomic regions of copy-number change. Bars represent significant *p*-values (adjusted *p*-value<0.01), after a logarithmic transformation, for the differences in the proportions of samples with each chromosomal abnormality between the different subtypes. For all samples, regions having an alteration frequency higher than 20% are displayed.
(PDF)

Figure S5 Determination of subtype prediction centroids. (A) Percentage of misclassification of the discovery set as a function of the number of top up- and down-regulated gene pairs used in the centroids. Misclassification is computed for the validation set by a 10-fold cross-validation procedure and is plotted by subtype (top) and averaged (bottom). (B) Heatmap of the 57-gene centroids used to assign a new dataset.
(PDF)

Figure S6 Prognostic value of the recoded CIT classification and of the Oncotype DX–like and Coloprint-like prognostic classifiers in the discovery and validation sets for patients with TNM stage II or III CC. RFS according to the recoded molecular subtype classification (C4/C6 subtypes versus other subtypes) in each TNM stage category (left, TNM II; middle, TNM III; right, TNM II–III), RFS of high- and low-risk patients as predicted by the Oncotype DX–like classifier, and RFS of patients belonging to cluster 1 and cluster 2 of the ColoPrint 17-gene expression signature in the discovery set (top), the validation set (middle), and the both datasets combined (bottom).
(PDF)

Figure S7 Prognostic Oncotype DX–like classifier within each CIT molecular subtype in the combined discovery and validation sets. RFS curves of high- and low-risk patients as predicted by the Oncotype DX–like classifier within each of the six CIT molecular subtypes.
(PDF)

Figure S8 Selection of the number of clusters. (A) Cumulative distribution function plot for each tested number of clusters; (B) cumulative distribution function delta area plot; (C) consensus matrix for different numbers of clusters ($k=5$ to 8).
(PDF)

Table S1 Patient and tumor characteristics. CIMP+/−, 3–5 methylated markers/0–2 methylated markers; CIN+/−, CIN>20%/CIN≤20%; F/M, female/male; WT/M, wild-type/mutant.
(XLS)

Table S2 List of the 1,459 most variant probe sets used to perform unsupervised analysis. The GeneCluster column corresponds to the gene cluster letters in Figure 1; logFC_CjvsCx corresponds to the gene expression log2 fold changes of subtype Cj versus the other subtypes; adjpv.CjvsCx corresponds to the adjusted *p*-values of the moderated *t*-test comparing Cj versus the other subtypes.
(XLS)

Table S3 List of the 1,108 subtype-discriminant probe sets. For each subtype, discriminant probe sets were selected from the discovery set using a moderated *t*-test, comparing a given subtype to the other subtypes (adjusted $p<10^{-5}$ and a |log fold change|>0.5), yielding 1,108 discriminant probe sets. The GeneCluster column corresponds to the gene cluster letters in Figure S2; logFC_CjvsCx corresponds to the gene expression log2 fold changes of subtype Cj versus the other subtypes; adjpv.CjvsCx corresponds to the adjusted *p*-values of the moderated *t*-test comparing Cj versus the other subtypes.
(XLS)

Table S4 Associations of anatomoclinical characteristics, DNA alterations, and supervised signature annotations with the six subtypes based on logistic regression. Associations were assessed by logistic regression using a multinomial logit model (function mlogit, R package mlogit).
(XLS)

Table S5 List of the 57 genes used to assign subtypes. The 57 genes selected to build the subtype predictor, given with each subtype's centroid values.
(XLS)

Table S6 Univariate and multivariate Cox analyses including the classification and clinical annotations. Associations of the classification and clinical annotations with RFS were assessed by Cox proportional-hazards regression analyses on (A) the discovery set, (B) the validation set, and (C) both sets. Univariate Cox analyses were performed on each variable independently. The best multivariate model was determined by using a backward–forward selection approach to restrict the multivariate model to the most informative variables for the subset of samples for which all the variables were available. Value indicates the modality of the annotation associated with the HR. H.R., Cox HR.
(XLS)

Table S7 Univariate and multivariate Cox analyses including the classification and other molecular annotations. Associations with RFS of the six-subtype classification—including *BRAF*, *KRAS*, and *TP53* mutations, MMR status, and CIMP status—were assessed by Cox proportional-hazards regression analyses on the discovery set. Univariate Cox analyses were performed on each variable independently. The best multivariate model was determined by using a backward–forward approach to restrict the multivariate model to the most informative variables for the subset of samples for which all the variables were available. The *TP53* mutation variable was excluded from the multivariate analysis, as only 201 samples were characterized and as it was not significantly associated to outcome. Value indicates the modality of the annotation associated with the HR. H.R., Cox HR.
(XLS)

Text S1 Supplementary methods.
(PDF)

Acknowledgments

We thank Prof. Jacqueline Godet and Dr. Jacqueline Métral for their constant support during the course of this work. The expertise of the personnel serving the RNA extraction and qualification (Saint-Louis Hospital, Paris), Affymetrix expression array (Institut de Génétique et de Biologie Moléculaire et Cellulaire, Strasbourg), and array-based CGH CIT platforms and of the CIT bioinformatics team is gratefully acknowledged. We also thank all the pathologists, biologists, and clinicians of each center who have participated to the collection of samples and clinical data. Some of the results reported herein were presented in part at the annual meeting of the American Association for Cancer Research, Chicago, Illinois, March 31–April 4, 2012 (abstract 5065).

Author Contributions

Conceived and designed the experiments: AD MPG JS SO GM PLP VB. Performed the experiments: MA AL EP. Analyzed the data: LM ADR LV RS VB PLP. Contributed reagents/materials/analysis tools: MC JFF DB AB FP DE YP EP AL LM ADR LV RS. Wrote the first draft of the manuscript: LM ADR VB PLP. Contributed to the writing of the manuscript: LM ADR VB PLP AD JS MPG GM DG MCEG SK SO. ICMJE criteria for authorship read and met: LM ADR AD JS MPG LV MCEG RS DG MA SK MC JFF DB AB AL EP FP DE YP SO GM PLP VB. Agree with manuscript results and conclusions: LM ADR AD JS MPG LV MCEG RS DG MA SK MC JFF DB AB AL EP FP DE YP SO GM PLP VB. Enrolled patients: AD MPG JS SO GM PLP VB.

References

1. Greenlee RT, Murray T, Bolden S, Wingo PA (2000) Cancer statistics, 2000. CA Cancer J Clin 50: 7–33.
2. American Joint Committee on Cancer (1997) AJCC cancer staging manual, 5th edition. Philadelphia: Lippincott-Raven.
3. Popat S, Hubner R, Houlston RS (2005) Systematic review of microsatellite instability and colorectal cancer prognosis. J Clin Oncol 23: 609–618.
4. Hutchins G, Southward K, Handley K, Magill L, Beaumont C, et al. (2011) Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. J Clin Oncol 29: 1261–1270.
5. Wang Y, Jatkoe T, Zhang Y, Mutch MG, Talantov D, et al. (2004) Gene expression profiles and molecular markers to predict recurrence of Dukes' B colon cancer. J Clin Oncol 22: 1564–1571.
6. Eschrich S, Yang I, Bloom G, Kwong KY, Boulware D, et al. (2005) Molecular staging for survival prediction of colorectal cancer patients. J Clin Oncol 23: 3526–3535.
7. Salazar R, Roepman P, Capella G, Moreno V, Simon I, et al. (2011) Gene expression signature to improve prognosis prediction of stage II and III colorectal cancer. J Clin Oncol 29: 17–24.
8. O'Connell MJ, Lavery I, Yothers G, Paik S, Clark-Langone KM, et al. (2010) Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. J Clin Oncol 28: 3937–3944.
9. Jass JR (2007) Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. Histopathology 50: 113–130.
10. Shen L, Toyota M, Kondo Y, Lin E, Zhang L, et al. (2007) Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. Proc Natl Acad Sci U S A 104: 18654–18659.
11. Kang GH (2011) Four molecular subtypes of colorectal cancer and their precursor lesions. Arch Pathol Lab Med 135: 698–703.
12. Hinoue T, Weisenberger DJ, Lange CP, Shen H, Byun HM, et al. (2012) Genome-scale analysis of aberrant DNA methylation in colorectal cancer. Genome Res 22: 271–282.
13. Cancer Genome Atlas Network (2012) Comprehensive molecular characterization of human colon and rectal cancer. Nature 487: 330–337.
14. Lièvre A, Bachet JB, Boige V, Cayre A, Le Corre D, et al. (2008) *KRAS* mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. J Clin Oncol 26: 374–379.
15. Cabelguenne A, Blons H, de Waziers I, Carnot F, Houllier AM, et al. (2000) *p53* alterations predict tumor response to neoadjuvant chemotherapy in head and neck squamous cell carcinoma: a prospective series. J Clin Oncol 18: 1465–1473.
16. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, et al. (1998) A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 58: 5248–5257.

17. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, et al. (2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with *BRAF* mutation in colorectal cancer. Nat Genet 38: 787–793.
18. de Reyniès A, Assié G, Rickman DS, Tissier F, Groussin L, et al. (2009) Gene expression profiling reveals a new classification of adrenocortical tumors and identifies molecular predictors of malignancy and survival. J Clin Oncol 27: 1108–1115.
19. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, et al. (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31: e15.
20. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127.
21. Guedj M, Marisa L, de Reynies A, Orsetti B, Schiappa R, et al. (2012) A refined molecular taxonomy of breast cancer. Oncogene 31: 1196–1206.
22. Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data. Mach Learn 52: 91–118.
23. Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, et al. (2003) Repeated observation of breast tumor subtypes in independent gene expression data sets. Proc Natl Acad Sci U S A 100: 8418–8423.
24. Merlos-Suárez A, Barriga FM, Jung P, Iglesias M, Céspedes MV, et al. (2011) The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. Cell Stem Cell 8: 511–524.
25. Kosinski C, Li VS, Chan AS, Zhang J, Ho C, et al. (2007) Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. Proc Natl Acad Sci U S A 104: 15418–15423.
26. Popovici V, Budinska E, Tejpar S, Weinrich S, Estrella H, et al. (2012) Identification of a poor-prognosis *BRAF*-mutant-like population of patients with colon cancer. J Clin Oncol 30: 1288–1295.
27. Laiho P, Kokko A, Vanharanta S, Salovaara R, Sammalkorpi H, et al. (2007) Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. Oncogene 26: 312–320.
28. Snover DC (2011) Update on the serrated pathway to colorectal carcinoma. Hum Pathol 42: 1–10.
29. Liang JJ, Bissett I, Kalady M, Bennet A, Church JM (2012) Importance of serrated polyps in colorectal carcinogenesis. ANZ J Surg. E-pub ahead of print. doi:10.1111/j.1445-2197.2012.06269.x
30. Alfaro MP, Pagni M, Vincent A, Atkinson J, Hill MF, et al. (2008) The Wnt modulator sFRP2 enhances mesenchymal stem cell engraftment, granulation tissue formation and myocardial repair. Proc Natl Acad Sci U S A 105: 18366–18371.
31. Oh SC, Park YY, Park ES, Lim JY, Kim SM, et al. (2011) Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. Gut 61: 1291–1298.
32. Lin YH, Friederichs J, Black MA, Mages J, Rosenberg R, et al. (2007) Multiple gene expression classifiers from different array platforms predict poor prognosis of colorectal cancer. Clin Cancer Res 13: 498–507.
33. Salazar R, Tabernero J, Moreno V, Nitsche U, Bachleitner-Hofmann T, et al. (2012) Validation of a genomic classifier (ColoPrint) for predicting outcome in the T3-MSS subgroup of stage II colon cancer patients [abstract 3510]. American Society of Clinical Oncology Annual Meeting; 1–5 Jun 2012; Chicago, Illinois, US.
34. (2011) A prospective study for the assessment of recurrence risk in stage II colon cancer patients using ColoPrint (PARSC). ClinicalTrials.gov: NCT00903565. Available: http://clinicaltrials.gov/ct2/show/NCT00903565. Accessed 16 April 2013.
35. Gray RG, Quirke P, Handley K, Lopatin M, Magill L, et al. (2011) Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. J Clin Oncol 29: 4611–4619.
36. Venook AP, Niedzwiecki D, Lopatin M, Lee M, Friedman PN, et al. (2011) Validation of a 12-gene colon cancer recurrence score (RS) in patients (pts) with stage II colon cancer (CC) from CALGB 9581 [abstract]. 2011 ASCO Annual Meeting Proceedings (Post-Meeting Edition). J Clin Oncol 29 (May 20 Suppl): 3518.
37. O'Connell M, Lee M, Lopatin M, Yothers G, Clark-Langone K, et al. (2012) Validation of the 12-gene colon cancer recurrence score (RS) in NSABP C07 as a predictor of recurrence in stage II and III colon cancer patients treated with 5FU/LV (FU) and 5FU/LV+oxaliplatin (FU+Ox) [abstract]. 2012 ASCO Annual Meeting Proceedings (Post-Meeting Edition). J Clin Oncol 30 (May 20 Suppl): 3512.
38. Roth A, Di Narzo AF, Tejpar S, Bosman F, Popovici V, et al. (2012) Validation of two gene-expression risk scores in a large colon cancer cohort and contribution to an improved prognostic method [abstract 3509]. American Society of Clinical Oncology Annual Meeting; 1–5 Jun 2012; Chicago, Illinois, US.
39. Weiser MR, Gönen M, Chou JF, Kattan MW, Schrag D (2011) Predicting survival after curative colectomy for cancer: individualizing colon cancer staging. J Clin Oncol 29: 4796–4802.

Editors' Summary

Background. Cancer of the large bowel (colorectal cancer) is the third most common cancer in men and the second most common cancer in women worldwide. Despite recent advances in the screening, diagnosis, and treatment of colorectal cancer, an estimated 608,000 people die every year from this form of cancer—8% of all cancer deaths. The prognosis and treatment options for colorectal cancer depend on five pathological stages (0–IV), each of which has a different treatment option and five year survival rate, so it is important that the stage is correctly identified. Unfortunately, pathological staging fails to accurately predict recurrence (relapse) in patients undergoing surgery for localized colorectal cancer, which is a concern, as 10%–20% of patients with stage II and 30%–40% of those with stage III colorectal cancer develop recurrence.

Why Was This Study Done? Previous studies have investigated whether there are any possible gene expression profiles (identified through microarray techniques) that can help predict prognosis of colorectal cancer, but so far, there have been no firm conclusions that can aid clinical practice. In this study, the researchers used genetic information from a French multicenter study to identify a standard, reproducible molecular classification based on gene expression analysis of colorectal cancer. The authors also assessed whether there were any associations between the identified molecular subtypes and clinical and pathological factors, common DNA alterations, and prognosis.

What Did the Researchers Do and Find? The researchers used genetic information from a cohort of 750 patients with stage I to IV colorectal cancer who underwent surgery between 1987 and 2007 in seven centers in France. The researchers identified relevant clinical and pathological staging information for each patient from the medical records and calculated recurrence-free survival (the time from surgery to the first recurrence) for patients with stage II or III disease. In the genetic analysis, 566 tumor samples were suitable—443 were used in a discovery set, to create the classification, and the remainder were used in a validation set, to test the classification. The researchers also used information from eight public datasets to validate their findings.

Using these methods, the researchers classified the colon cancer samples into six molecular subtypes (based on gene expression data) and, on further analysis and validation, were able to distinguish the main biological characteristics and deregulated pathways associated with each subtype. Importantly, the researchers found that that these six subtypes were associated with distinct clinical and pathological characteristics, molecular alterations, specific gene expression signatures, and deregulated signaling pathways. In the prognostic analysis based on recurrence-free survival, the researchers found that patients whose tumors were classified in one of two clusters (C4 and C6) had poorer recurrence-free survival than the other patients.

What Do These Findings Mean? These findings suggest that it is possible to classify colorectal cancer into six robust molecular subtypes that might help identify new prognostic subgroups and could provide a basis for developing robust prognostic genetic signatures for stage II and III colorectal cancer and for identifying specific markers for the different subtypes that might be targets for future drug development. However, as this study was retrospective and did not include some known predictors of colorectal cancer prognosis, such as tumor grade and number of nodes examined, the significance and robustness of the prognostic classification requires further confirmation with large prospective patient cohorts.

Additional Information. Please access these websites via the online version of this summary at http://dx.doi.org/10.1371/journal.pmed.1001453.

- The American Cancer Society provides information about colorectal cancer and also about how colorectal cancer is staged
- The US National Cancer Institute also provides information on colon and rectal cancer and colon cancer stages

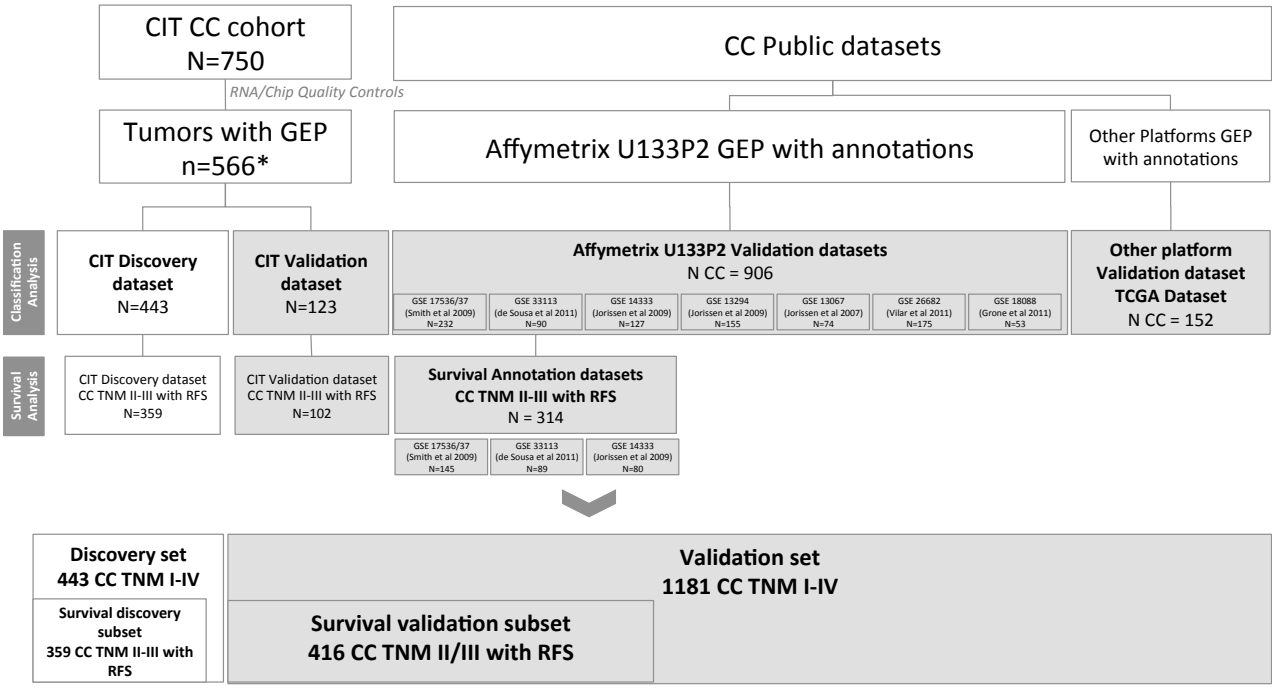

CIT CC cohort
N=750
CC Public datasets
RNA/Chip Quality Controls
Tumors with GEP
n=566*
Affymetrix U133P2 GEP with annotations
Other Platforms GEP
with annotations
Classification Analysis
CIT Discovery
dataset
N=443
CIT Validation
dataset
N=123
Affymetrix U133P2 Validation datasets
N CC = 906
GSE 17536/37
(Smith et al 2009)
N=232
GSE 33113
(de Sousa et al 2011)
N=90
GSE 14333
(Jorissen et al 2009)
N=127
GSE 13294
(Jorissen et al 2009)
N=155
GSE 13067
(Jorissen et al 2007)
N=74
GSE 26682
(Vilar et al 2011)
N=175
GSE 18088
(Grone et al 2011)
N=53
Other platform
Validation dataset
TCGA Dataset
N CC = 152
Survival Analysis
CIT Discovery dataset
CC TNM II-III with RFS
N=359
CIT Validation dataset
CC TNM II-III with RFS
N=102
Survival Annotation datasets
CC TNM II-III with RFS
N = 314
GSE 17536/37
(Smith et al 2009)
N=145
GSE 33113
(de Sousa et al 2011)
N=89
GSE 14333
(Jorissen et al 2009)
N=80
Discovery set
443 CC TNM I-IV
Survival discovery
subset
359 CC TNM II-III with
RFS
Validation set
1181 CC TNM I-IV
Survival validation subset
416 CC TNM II/III with RFS

Figure S1

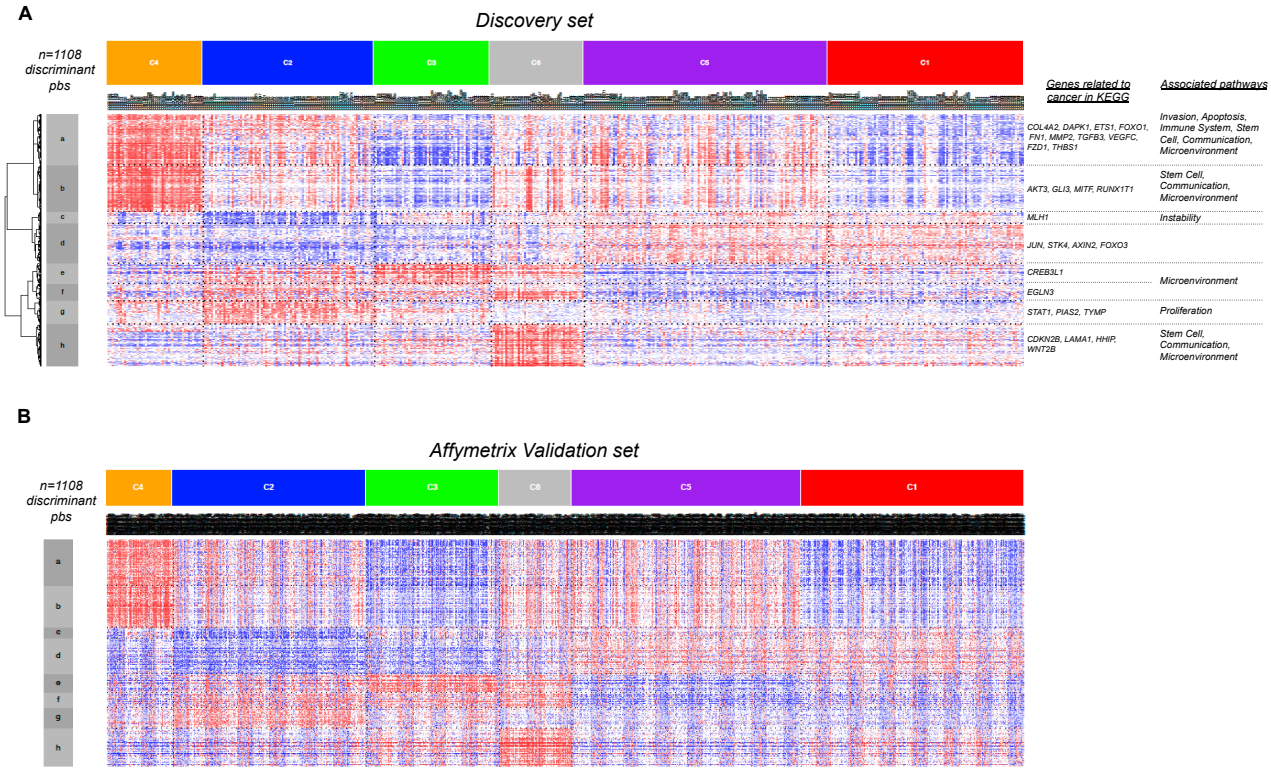
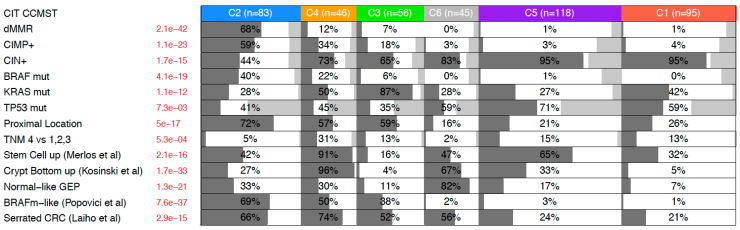
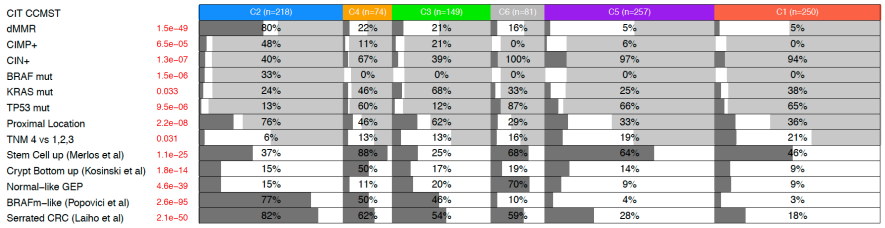

A
Discovery set
n=1108 discriminant pbs
C4
C2
C3
C6
C5
C1
Genes related to cancer in KEGG
Associated pathways
COL4A2, DAPK1, ETS1, FOXO1, FN1, MMP2, TGFB3, VEGFC, FZD1, THBS1
Invasion, Apoptosis, Immune System, Stem Cell, Communication, Microenvironment
AKT3, GLI3, MITF, RUNX1T1
Stem Cell, Communication, Microenvironment
MLH1
Instability
JUN, STK4, AXIN2, FOXO3
CREB3L1
Microenvironment
EGLN3
STAT1, PIAS2, TYMP
Proliferation
CDKN2B, LAMA1, HHIP, WNT2B
Stem Cell, Communication, Microenvironment
B
Affymetrix Validation set
n=1108 discriminant pbs
C4
C2
C3
C6
C5
C1

Figure S2

A Affymetrix Discovery set



B Affymetrix Validation set



C TCGA Validation dataset

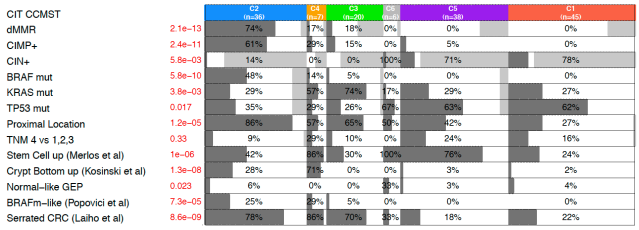


Figure S3

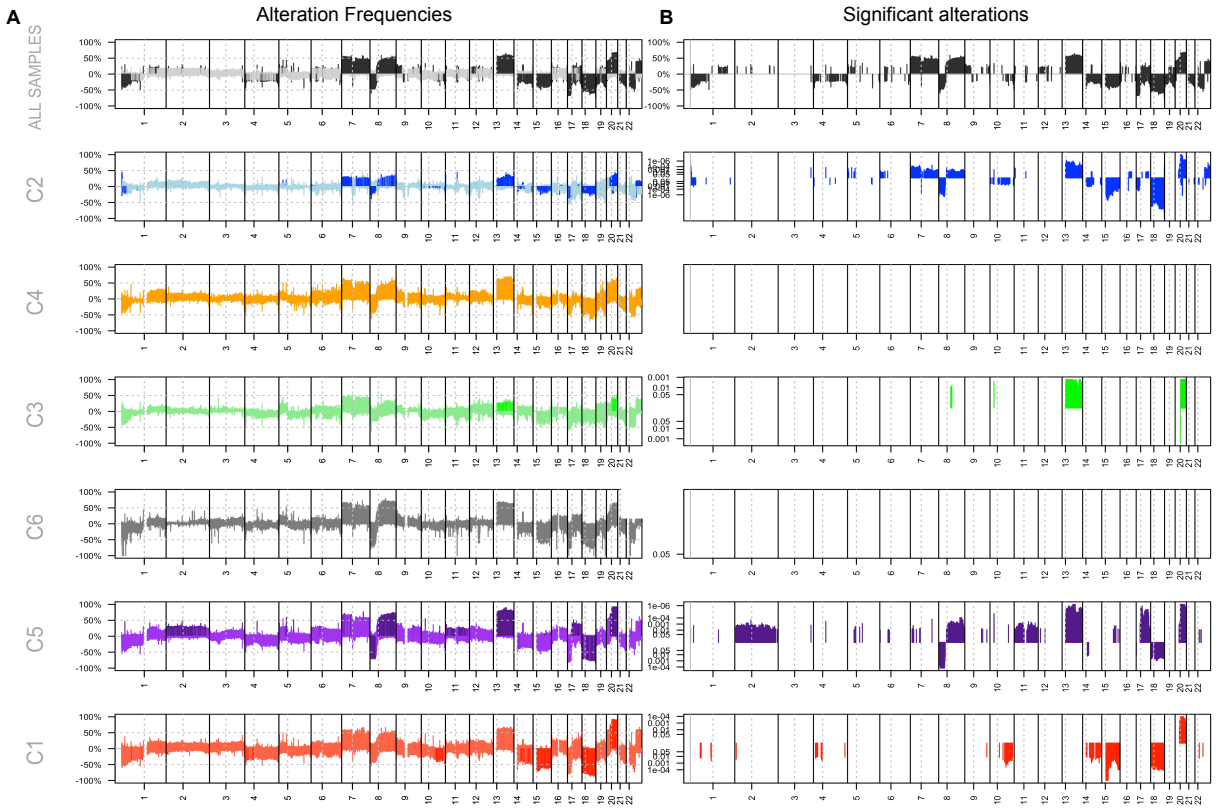

A
Alteration Frequencies
B
Significant alterations
ALL SAMPLES
C2
C4
C3
C6
C5
C1

Figure S4

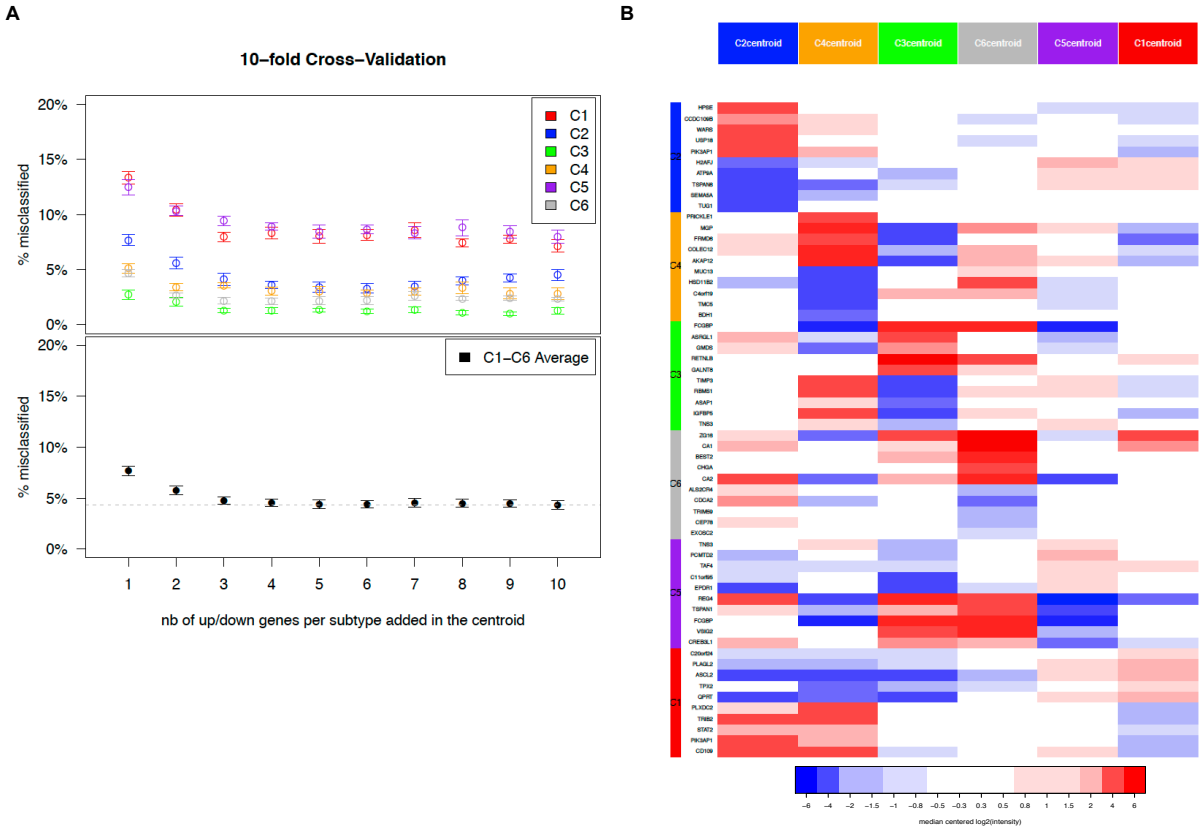

A
10−fold Cross−Validation
% misclassified
20%
15%
10%
5%
0%
C1
C2
C3
C4
C5
C6
C1−C6 Average
1
2
3
4
5
6
7
8
9
10
nb of up/down genes per subtype added in the centroid
B
C2centroid
C4centroid
C3centroid
C6centroid
C5centroid
C1centroid
median centered log2(intensity)

Figure S5

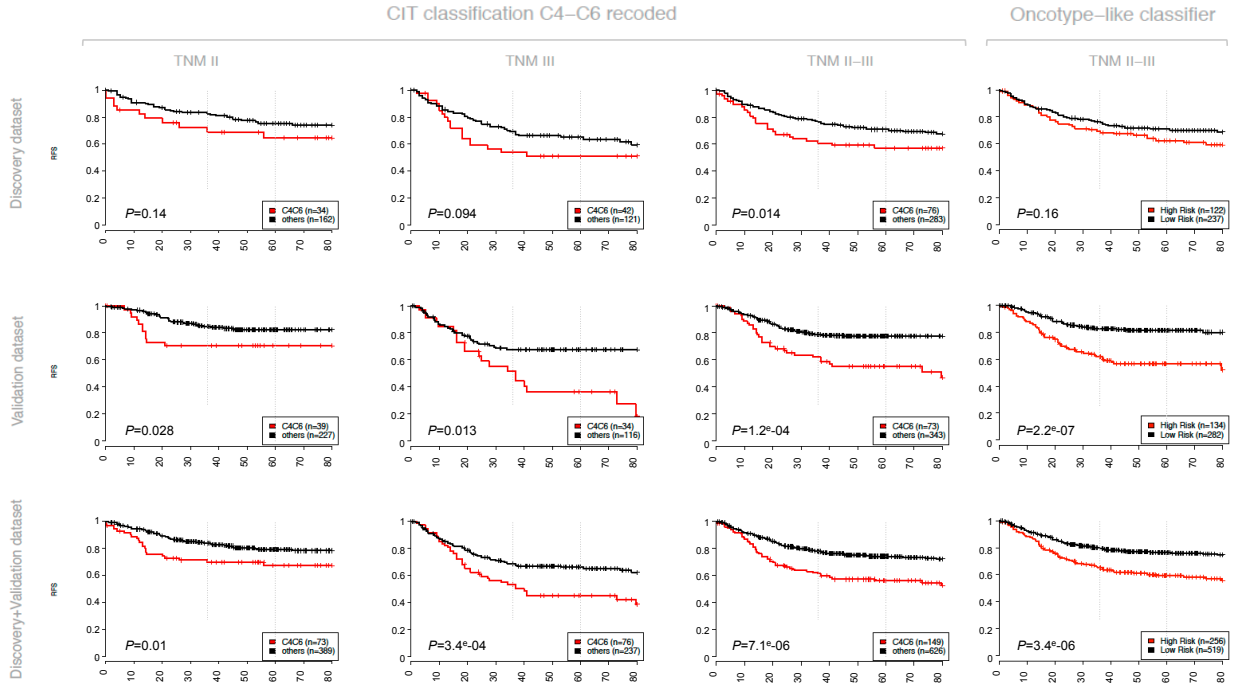

CIT classification C4–C6 recoded
Oncotype–like classifier
TNM II
TNM III
TNM II–III
TNM II–III
Discovery dataset
Validation dataset
Discovery+Validation dataset
RFS
P=0.14
C4C6 (n=34)
others (n=162)
P=0.094
C4C6 (n=42)
others (n=121)
P=0.014
C4C6 (n=76)
others (n=283)
P=0.16
High Risk (n=122)
Low Risk (n=237)
P=0.028
C4C6 (n=39)
others (n=227)
P=0.013
C4C6 (n=34)
others (n=116)
P=1.2e-04
C4C6 (n=73)
others (n=343)
P=2.2e-07
High Risk (n=134)
Low Risk (n=282)
P=0.01
C4C6 (n=73)
others (n=389)
P=3.4e-04
C4C6 (n=76)
others (n=237)
P=7.1e-06
C4C6 (n=149)
others (n=626)
P=3.4e-06
High Risk (n=256)
Low Risk (n=519)

Figure S6

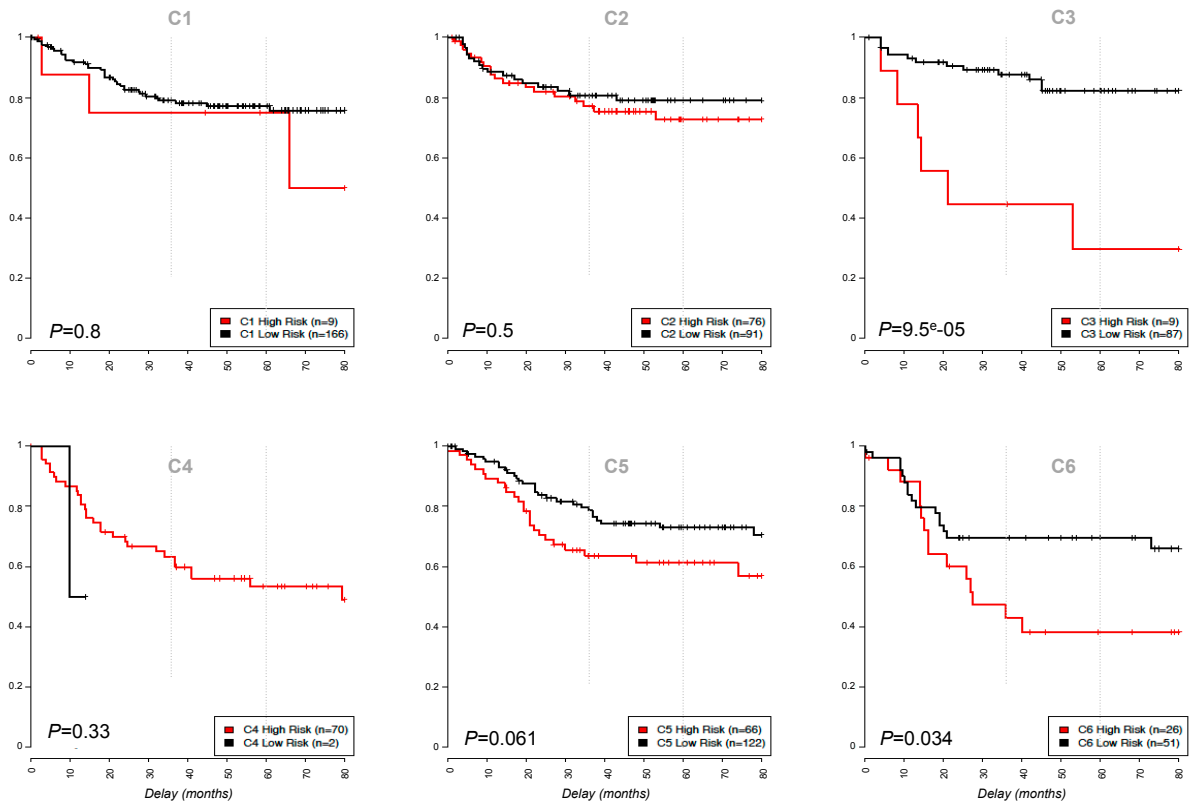

C1
P=0.8
C1 High Risk (n=9)
C1 Low Risk (n=166)
C2
P=0.5
C2 High Risk (n=76)
C2 Low Risk (n=91)
C3
P=9.5e-05
C3 High Risk (n=9)
C3 Low Risk (n=87)
C4
P=0.33
C4 High Risk (n=70)
C4 Low Risk (n=2)
C5
P=0.061
C5 High Risk (n=66)
C5 Low Risk (n=122)
C6
P=0.034
C6 High Risk (n=26)
C6 Low Risk (n=51)
Delay (months)

Figure S7

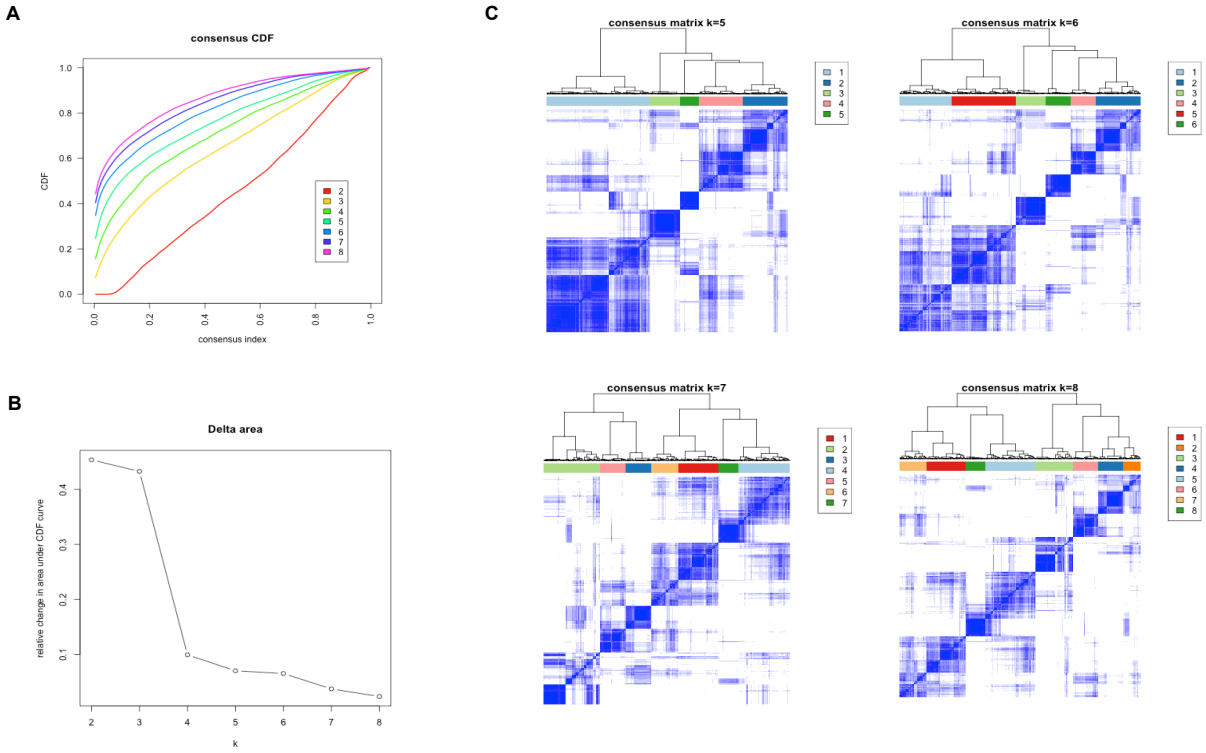

A
consensus CDF
CDF
consensus index
B
Delta area
relative change in area under CDF curve
k
C
consensus matrix k=5
consensus matrix k=6
consensus matrix k=7
consensus matrix k=8

Figure S8

TEXT S1. SUPPLEMENTARY METHODS

Molecular Characterizations

MMR Status determination

Microsatellite instability was analyzed using a panel of five different microsatellite loci from the Bethesda reference panel [1]. Tumors were characterized on the basis of high-frequency MSI (MSI-H) if two or more of the five markers showed instability, low-frequency MSI (MSI-L) if only one of the five markers showed instability, and MSS if none of the five markers showed instability. MSI-H tumors were further classified as deficient MMR (dMMR), and both MSI-L and MSS as proficient MMR (pMMR).

CIMP Status determination

The CIMP status was determined using the panel of five markers described in Weisenberg *et al* [2]: *CACNA1G*, *IGF2*, *NEUROG1*, *RUNX3* and *SOCS1*. After DNA bisulfite treatment, two multiplex methylation-specific PCR were performed. Fragment analysis was carried out by capillary electrophoresis on automatic sequencer (Beckman Coulter®, Danvers, MA, USA). Methylator phenotype positive cases (CIMP +) had ≥ 3 methylated promoters while CIMP - ones had less than 2 methylated promoters, according to established criteria (22).

Chromosomal Instability (CIN) Status definition

The CIN status was assigned according to CGH alteration profile. A CIN rate was designed as the mean of the per chromosome rate of gained or lost clones (mean(number of clones with a Gain or Loss/total number of clones of the chromosome)). A tumor having an alteration rate superior to 20% was considered CIN+, otherwise CIN-. This cut-off of 20% was chosen based on unsupervised hierarchical clustering of GNL profiles, which delimited a group of tumors with no/very low instability, which displayed a CIN rate inferior to 20%.

Gene expression data normalization

The CIT cohort CEL files were first normalized using the Robust Multi-array Average (RMA) [3] method implemented in the R package affy. Then to remove potential multicenter batch effects, data were corrected using ComBat method [4] implemented in the R package sva, with Centre and RNA extraction method as batch effects and with tumoral and MMR status as features of interest.

Each Affymetrix public datasets used for validation were independently normalized by the RMA method as well.

Molecular subtype determination

Unsupervised Probe set selection

The 1459 probe sets used for subtype determination fulfill the three following criteria:

(1) to be expressed in at least 5% of the samples (i.e. 5th decile of normalized intensities across samples > log2(15))

(2) to have a variance significantly different from the median variance of all probe sets (i.e. variance test p-value<0·01)

Variance test: For each probe set (P) we tested whether its variance across samples was different from the median of the variances of selected probe sets in (1). The statistic used was $((n-1)\times Var(P) / Var_{med})$, where n refers to the number of samples. This statistic was compared to a percentile of the Chi-squared distribution with (n-1) degrees

of freedom (this criteria is used in the BRB ArrayTools filtering tool, described in the User's Manual [5]) and yielded a *p*-value for each probe set.

(3) to have a high robust coefficient of variation (rCV > 0·186).

rCV : rCV for each probe set was calculated by dividing the standard deviation by the mean, eliminating the highest and lowest expression value across the samples for each probe set.

rCV threshold determination : the cut-off point was defined using Gaussian mixture model clustering approach (R package mclust [6]) which defined 4 groups of rCV, the most variant one containing 1459 probe sets, the minima rCV being 0·186.

Consensus Unsupervised class discovery approach

The subtypes were determined using the consensus clustering approach described in Monti et al [7] and implemented in the R package ConsensusClusterPlus [8]. In brief, a clustering analysis is performed n times on subsets of the probe sets and of the samples selected randomly. Then all derived partitions for a given number of clusters k are summarized by clustering the (samples x samples) co-classification matrix*. The whole data were first gene median centered and the parameters used were set as follows:

- Clustering algorithm: hierarchical clustering
- Clustering metrics: (1-Pearson correlation) distance and Ward linkage
- n resamplings: 1000
- Proportion of samples and probe sets used in each resampling: 90%,
- k tested: from 2 to 8.

As described in Monti et al, [7] the choice of the number of clusters can be based on the delta area plot and should correspond to the number of clusters k where the Cumulative distribution (CDF) levels off and the corresponding relative increase in the CDF area gets closes to zero. Following this procedure, in our case, several values of k could reasonably have been selected (Figure S7), and, at inspecting the consensus matrices progression as suggested, the more balanced partition appeared to be for k=6.

* giving for each pair of samples the proportion of partitions in which these two samples were co-clustered.

Molecular subtype prediction

To assign a subtype to each sample from the validation series, we developed a centroid-based predictor using the most discriminating probe sets (over and under expressed) of each subtype.

The selection of the probe sets used in the centroids was performed among the probe sets selected in the 2 first steps of the subtype determination approach and having an Affymetrix grade A annotation (NetAffx [9] Annotations version na31 were used) and then as follows for each subtype:

- Probe sets significantly differentially expressed in samples of the given subtype compared to samples of other subtypes according to the Limma moderated t-test [10] or the Welch t-test (adjusted *p*-value<1e-5 and |log2 fold change|>0·5) were retained
- Then the selected probe sets were ordered according to their AUC score (computed using the R package PresenceAbsence [11]) and only those with a score superior to 0.7 were kept.
- To avoid the selection of highly correlated probe sets (redundancy) we clustered probe sets using hierarchical clustering (distance=1-Pearson, linkage method=Ward), cut the

tree to isolate uncorrelated clusters (tree cut-off (1-correlation) = 0.9) and kept one probe set per cluster, the one having the best AUC and a gene symbol annotation.

- To select the probe sets to use in the centroid, we proceeded by a 10-fold cross-validation approach. The discovery dataset was split into 10 subsets. The top up/down regulated pairs of probe sets were used to build centroids on 9 of the 10 subsets and the assignment (see below) was then computed on the remaining subset. This procedure was repeated for each subset and for each number of probe set pairs tested (from 1 to 10). The lowest global misclassification was obtained for 5 top up/down pairs (Figure S4A).

This procedure yields 57 probe sets (corresponding to 57 unique genes), 3 probe sets being specific to several subtypes but with inverted regulation (Table S2, Figure S4B).

Then using those probe sets, 6 centroids were computed on the gene-median centered discovery dataset and for each validation dataset (RMA normalized and gene-median centered), the distance to the 6 centroids of each sample was computed and samples were assigned to the closest centroid subtype. The decision rule was based on the diagonal quadratic discriminant analysis method (DQDA) and is defined as follows:

$$DQDA(X) = \underset{j\in\{C1,\ldots,C6\}}{Arg\ min}\left(\sum_{i=1}^{N}\frac{\left(x_i-\mu_{j,i}\right)^2}{v_{j,i}}\right)+\sum_{i=1}^{N}\log(v_{j,i})$$

where N is the number of genes (here N=57), x the expression normalized values, $\mu_{j,i}$ and $\nu_{j,i}$ the mean and the variance of the gene i across samples of the subtype j from the discovery data set (i.e. the centroid).
The confidence of the prediction was evaluated by identifying outliers (too distant samples) and mixed assignment samples (when a sample is close to several centroids). More specifically, a sample is said to be an outlier if its distance to the closest centroid is superior to n times the median absolute deviation (mad) of the distances of the samples used to compute the centroid; n is defined as the maximum (distances to centroid-$median_{distances\ to\ centroid}$)/$mad_{distances\ to\ centroid}$). A sample has a mixed assignment if the difference of its distance to centroid is inferior to the 1st decile of the difference between centroids on data used to compute centroids.
Among the 1029 samples of the validation data set, only 13 samples had an uncertain assignment and no outliers were found.
The subtype prediction procedure is implemented in the R package citccmst that will be available at the R CRAN repository (http://cran.r-project.org/).

N.B.: This prediction procedure has been designed from Affymetrix U133P2 data set and applied to Affymetrix U133P2 data sets so the prediction of other platform datasets should require caution and adjustment (as gene symbol mapping, re-computing the centroid using those selected genes and using another distance metrics).

Molecular subtype characterization
i) Non-tumoral Colonic Mucosa GEP tumors:
To evaluate the similarity of GEP tumors to colon normal tissue, the distance of each tumor samples to the centroid of the 1459 probe sets of the normal mucosa samples was computed.

A tumor was assigned Normal-like GEP if its distance was amongst the 25% closest to the NC centroid (metrics 1-Pearson correlation, Ward linkage, median gene centered data).

ii) Annotation with published supervised signatures
Tumors were assigned to molecular and cellular phenotypes as follows:
For all signatures used, genes were matched to our probe sets by the Gene Symbol annotation and only the most variant probe set (maximal rCV) was selected.

Stem cell signature up regulated tumors:
The signature used is the Merlos-Suarez et al [12] Intestinal Stem Cells (ISC) signature (in their table S1). As describe in their article, an ISC score was computed by gene centering the data (median) and computing the mean expression of all genes of the signature. A tumor was assigned Stem Cell signature up regulated when this score was superior to the mean of all scores.

Cell from crypt signature up regulated tumors:
The signature used is composed of a selection of the genes highly up regulated in bottom crypt given in Kosinski et al [13] (in their table 3, *p*-value paired t-test < 1e-5 and |logFC|>2). As only some of those genes were highly up regulated in our tumors, a hierarchical clustering approach was preferred over mean expression score and allows us to divide our samples into 2 groups, those with a subset of those bottom gene highly up regulated were assigned Crypt Cell Signature up regulated.

Popovici BRAF mutated like tumors:
As described in their article [14], the genes given in the Table 2 were used and if the mean of G1 genes was smaller than the mean of G2 the tumor was assigned *BRAF*m-like.

Laiho et al Serrated CRC tumors:
A centroid of the probe sets of their signature [15] (Table S3) was computed on the original data set (GSE4045) and our tumors were assigned Serrated or Conventional adenoma depending on the distance to the closest centroid (metrics 1-Pearson correlation, median gene centered).

iii) Cancer pathway analysis
KEGG pathways and some gene sets from Gene Ontology selected to be related to cancer hallmarks (Cell communication, growth/death, Immune system, Motility, Replication and repair, Angiogenesis, Metabolism and main cancer signal transduction pathways) were tested for enrichment of the top 1000 up and top 1000 down regulated genes of every subtypes (genes were selected based on Limma t-test *p*-values and |FC|>1·5) by computing a hypergeometric test (*p*-value <0·05).

iv) CGH alteration frequency profiles
CGH array chip and experiment have already been described here [16]. Raw log2-ratio values were filtered (i) using a signal-to-noise threshold of 2·0 for the reference channel and (ii) when the individual single intensities for the sample or reference was less than 1·0 or at saturation (i.e. 65,000). The remaining values were normalized using the lowess within-print tip group method [17] and the values of clone replicates were averaged if their standard deviation was less than 0.25 otherwise filtered. Then to define region of loss and gain, for each sample the normalized values were smoothed to obtain segments using tilingArray method [18] and the DNA copy number was determined as follows*:* the level (L_N)

corresponding to a normal (i.e. diploid) copy number is determined as the first mode of the distribution of the smoothed log2-ratio values across all autosomes; the standard deviation (SD) of the difference between normalized and smoothed values is calculated; then for all clones in a segment, the 'GNL' copy number status (G: gain - N: normal - L: loss) is determined based on the segment smoothed log-ratio value (X): if $X > L_N + SD$ then status=gain (G), if $X < L_N - SD$ then status=loss (L), else status=normal; in a given segment, outlier clones that yielded normalized log2-ratio values (Y) such that $Y > L_N + 3 \times SD$ (respectively $Y < L_N - 3 \times SD$) are classified as gains (respectively losses).
Alteration frequencies profiles in Figure S3 were obtained using the 356 CGH arrays available for samples of the discovery dataset and by computing the proportion of samples harbouring a gain or a loss of copy, at each clone of the array for all samples and by subtypes. Frequently altered genomic regions (Figure S3 B) in the whole dataset were determined by identifying regions for which the proportion of alteration (in gain or loss) exceed 20%. Subtypes specific regions were determined by applying at each clone a test of proportion comparing the proportion of alteration (gain and loss) in the samples of a given subgroup versus in samples of the others corrected for multiple-testing by FDR (Benjamini and Hochberg) [19], a subtype specific genomic regions being defined as a set of consecutive clones significantly more altered in the subtype of interest (*p*-values < 0·01).

Molecular Subtype Robustness
Internal robustness:

- The subtypes were obtained using a consensus clustering procedure using both gene and sample resampling (1000 random subselections of 90% of the samples and 90% of the genes), such that these results are stable under conditions of gene and sample resampling.
- The subtypes were obtained from a large set (n=443) of samples processed with the same experimental procedure, as part of the Cartes d'Identité des Tumeurs program.
- Moreover, we have tested that our classification results were also repeatedly obtained using different metrics (Euclidian/Pearson).

External robustness:
The subtypes were validated on a large dataset collected under different conditions, from numerous centers: clinical and biological characteristics of the subtypes were found to be conserved in this validation set.

Survival Analyses
Survival analyses were restricted to the subgroup of patients with stage II-III tumors. Additional prognostic biomarkers are most needed for these patients. This is because the vast majority of stage I CRC patients will never relapse after curative surgery and will not derive benefit from adjuvant chemotherapy because the prognosis is excellent. Also, most stage IV CRC patients are already metastatic and will die from their disease.

Relapse-Free Survival was used and defined as the time from surgery to the first recurrence.

Survival curves were obtained according to the method of Kaplan and Meier (function Surv, R package survival) and differences between survival distributions were assessed by Log-rank test using an endpoint of five years/60 months (function survdiff, R package survival). The proportional-hazards assumption was tested to examine the model's appropriateness (function cox.zph, R package survival).

For the analysis of associations with patient outcome, univariate and multivariate models were computed using Cox proportional-hazards regression (function coxph, R package survival). Univariate analyses were performed to assess the marginal value of each variable independently from the others. For multivariate analyses, first a multivariate analysis using all variables (excluding those with insufficient data as not to reduce the power of the analysis) was performed. Next, to select the best multivariate model, a backward-forward step procedure was computed to restrict the multivariate model to the most informative variables as described in Venables & Ripley, 2002[20] (function *step(), R package stats*). Only samples for which all the variables were available were included in multivariate models.

Recurrence Risk group assignment according to O'Connell and Salazar predictors

O'Connell et al [21] Oncotype classifier:

The O'Connell Recurrence Risk (RS) score is composed of 12 genes among which 5 reference genes and 7 genes associated to recurrence. For the reference genes, when several probe set were possible, the less variant one was selected. For the other genes, data were median gene centered and aggregated by mean if several probe sets were available. Then the recurrence genes intensities for each sample were subtracted by the mean of the reference gene per sample and the formula given in O'Connell et al (Figure 3 and supplemental method) was applied for each sample RS_u = 0·15* mean(BGN,FAP,INHBA)- 0·3*mean(MKI67,MYC,MYBL2)+ 0·15* GADD45B)

This score was then rescaled RS=44*(RSu+0.82). RS ranged from 8 to 82 so ranging its distribution between 0 and 100 was not necessary. A tumor was predicted with high risk if the score was superior or equal to 41 as mentioned in the article.

Salazar et al [22] predictor:

Among the 18 genes from their classifier, only 17 are found in Affymetrix annotations. As no centroid was available and down/up regulations were not mentioned, we computed a hierarchical clustering of the probe sets average matching those 17 genes to obtain 2 clusters.

References

1. Boland CR, Thibodeau SN, Hamilton SR, Sidransky D, Eshleman JR, et al. (1998) A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res 58:5248-57.

2. Weisenberger DJ, Siegmund KD, Campan M, Young J, Long TI, et al. (2006) CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with *BRAF* mutation in colorectal cancer. Nat Genet 38:787-93.

3. Irizarry RA, Hobbs B, Collin F, et al. (2003) Exploration, normalization, and summaries of high density oligonucleotide array probe level data. *Biostatistics* 4:249-264.

4. Johnson WE, Li C, Rabinovic A (2007) Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics 8: 118–127.

5. Simon R, and Peng Lam A. (2003) BRB-ArrayTools software v3.1 User's Manual linus.nci.nih.gov/BRB-ArrayTools.html.

6. Chris Fraley and Adrian E. Raftery (2006) MCLUST Version 3 for R: Normal Mixture Modeling and Model-based Clustering.Technical Report No. 504, Department of Statistics, University of Washington (revised 2009)

7. Monti S, Tamayo P, Mesirov J, Golub T (2003) Consensus clustering: A resampling-based method for class discovery and visualization of gene expression microarray data. *Machine learning* 52:91-118.

8. Matt Wilkerson (2011). ConsensusClusterPlus: ConsensusClusterPlus. R package version 1.6.0.

9. Liu G, Loraine AE, Shigeta R, Cline M, Cheng J, Valmeekam V, Sun S, Kulp D, Siani-Rose MA (2003) NetAffx: Affymetrix probesets and annotations. *Nucleic Acids Res*;31(1):82-6.

10. Smyth, G.K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol* 3, Article3

11. Freeman, Elizabeth (2007) PresenceAbsence: An R Package for Presence-Absence Model Evaluation. USDA Forest Service, Rocky Mountain Research Station, 507 25th street, Ogden, UT, USA.

12. Merlos-Suárez A, Barriga FM, Jung P, et al. (2011) The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell* 8:511-24.

13. Kosinski C, Li VS, Chan AS, Zhang J, et al. (2007) Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proc Natl Acad Sci U S A* 104:15418-23.

14. Popovici V, Budinska E, Tejpar S, et al. (2012) Identification of a poor-prognosis *BRAF*-mutant-like population of patients with colon cancer. *J Clin Oncol* 30:1288-95.

15. Laiho P, Kokko A, Vanharanta S, et al. (2007) Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene* 26:312-20.

16. Guedj M, Marisa L, de Reynies A, et al. (2012) A refined molecular taxonomy of breast cancer. *Oncogene* 1;31(9):1196-206.

17. Yang YH, Dudoit S, Luu P, et al. (2002) Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* 30:e15.

18. Wolfgang Huber and Joern Toedling and Lars M. Steinmetz (2006) Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics* 22, 1963-1970.

19. Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*. 1995; 57 289-300.

20. Venables, W. N. and Ripley, B. D. (2002) _Modern Applied Statistics with S. Fourth edition. Springer.

21. O'Connell M, Lee M, Lopatin M, et al. (2012) Validation of the 12-gene colon cancer recurrence score (RS) in NSABP C07 as a predictor of recurrence in stage II and III colon cancer patients treated with 5FU/LV (FU) and 5FU/LV+oxaliplatin (FU+Ox). *J Clin Oncol* 30 (suppl; abstr 3512).

22. Salazar R, Josep Tabernero, Victor Moreno, et al. (2012) Validation of a genomic classifier (ColoPrint) for predicting outcome in the T3-MSS subgroup of stage II colon cancer patients. *J Clin Oncol* 30 (suppl; abstr 3510).

Établissement d'une classification consensus des cancers colorectaux à partir de six systèmes de classification différents

En même temps que la publication de ce travail, 5 autres publications sur la classification transcriptomique paraissaient (Schlicker et al., 2012; Melo et al., 2013; Sadanandam et al., 2013; Budinska et al., 2013; Roepman et al., 2013). Chacune des classifications proposait un nombre de sous-types différents et des associations cliniques et moléculaires différentes (Figure 10). Il n'était pas évident de mettre en lumière les identités et les inconsistances entre chaque partition proposée.

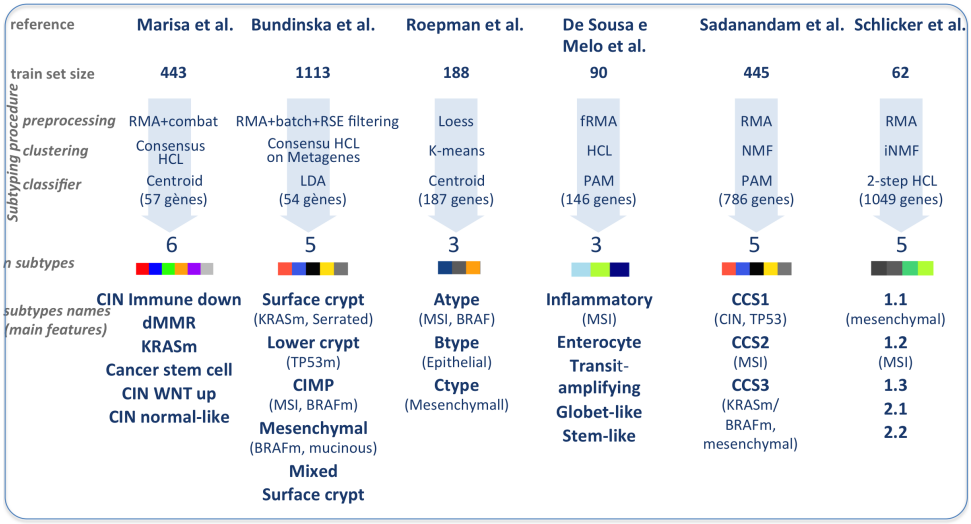


FIGURE 10 – Les six classifications utilisées pour établir la classification consensus

Sous l'impulsion de Sabine Tejpar, clinicienne et chercheuse à l'université de Leuven, et de Stephen Friend, à la tête du groupe Sage Bionetwork, une organisation promouvant la science ouverte ("open science") et la collaboration entre équipes, un consortium de travail s'est organisé, regroupant les acteurs des 6 classifications publiées et orchestré par des chercheurs de Sage, auquel j'ai participé. Le but était de proposer une classification consensus des 6 classifications. Le travail réalisé, que je cosigne, est accepté pour publication dans Nature Medicine (Guinney et al., 2015). Toutefois n'ayant pas réalisé les analyses principales, je le présente ici dans l'esprit de discuter de mon 1er travail de classification et pour l'intérêt scientifique.

L'approche développée

L'approche qui a été retenue par le consortium a été de rassembler un très large jeu de données et que chaque groupe applique indépendamment leur système de classification (Figure 11). Une approche pour combiner les 6 différentes partitions a alors été élaborée. Brièvement, un réseau d'associations entre les différents sous-types de chaque système a été construit. Les nœuds sont représentés par la prévalence de chacun des sous-types et les arcs l'association entre les nœuds, définie par le coefficient de similarité de Jaccard (taille de l'intersection entre les 2 groupes d'échantillons sur la taille de leur union). Pour définir des sous-types consensus, une classification par partitionnement adapté au réseau (Markov Cluster Algorithm) a été appliqué et la procédure a été reproduite 1000 fois en

prenant aléatoirement une partie des échantillons. La partition consensus finale est celle qui donne la meilleur performance de partitionnement selon la mesure des moyennes des "Silhouettes" de chaque nœud (c'est-à-dire pour chaque nœud, on prend la distance b(n) minimale entre le nœud et les autres clusters à laquelle on soustrait la distance a(n) entre ce nœud et les autres nœuds du cluster que l'on divise par la valeur maximale entre a(n) et b(n) ; la moyenne de l'ensemble des valeurs de chacun des nœuds est utilisée comme critère à maximiser pour sélectionner le nombre de groupe).

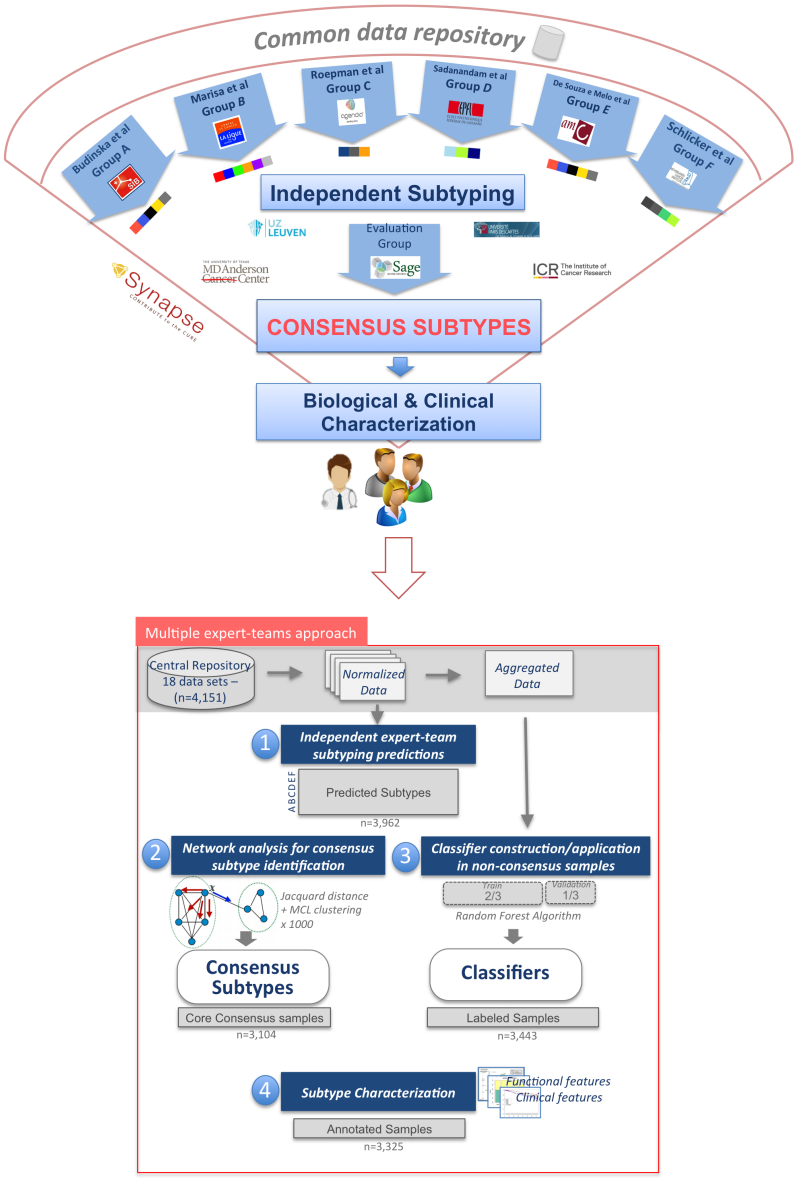


FIGURE 11 – Description globale et détaillée de l'approche de classification consensus

La classification consensus obtenue

La classification consensus établie sur 3962 tumeurs se compose de 4 sous-types consensus : 14% ont été classées CMS1 (MSI immune), 37% CMS2 (Canonical) , 13% CMS3 (Metabolic), 23% CMS4 (Mesenchymal). 13% n'ont pas été assignés car ils étaient trop

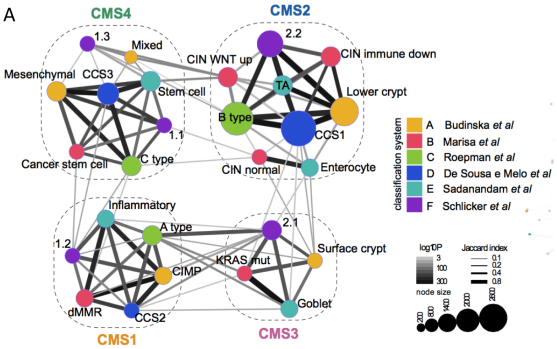
discordants dans les différents systèmes de classification. Une étude plus approfondie a confirmé que ces tumeurs ne formaient pas un 5ème sous-type et a montré que 7% avaient des caractéristiques mixtes entre plusieurs sous-types. Une caractérisation très approfondie des sous-types a été réalisée en utilisant l'ensemble des annotations disponibles et l'ensemble des données du TCGA (exome, methylome, mirnome et protéome) dont les principales caractéristiques sont résumées dans la Figure 12 :

CMS1 (MSI immune) enrichi en tumeurs MSI (76%), hypermutées (94%) et ayant une forte activation des gènes de l'immunité

CMS2 (Canonical) enrichi en tumeurs épithéliales, très instables génomiquement et présentant une activation des voies Wnt et MYC ; de fréquents gains de copies d'ADN d'oncogènes et pertes de copies d'ADN de gènes suppresseur de tumeurs y sont retrouvés

CMS3 (Metabolic) enrichi en tumeurs épithéliales très fréquemment muté KRAS (68%) avec une forte dérégulation des voies métaboliques ; il rassemble des tumeurs aux profils génomiques et épigénomiques particulier avec peu d'altérations le long du génome et un enrichissement du phénotype CIMP (16%) et CIMP-Low (40%, déterminé par classification non supervisée des données méthylomes du TCGA) et du phénotype MSS hypermuté (28%) ;

CMS4 (Mesenchymal) enrichi en tumeurs ayant une activation de la voie TGFβ, de l'angiogenèse et de l'EMT, ayant plus d'invasion stromale et diagnostiquées plus tardivement (TNM III/IV)



CMS1 MSI Immune	CMS2 Canonical	CMS3 Metabolic	CMS4 Mesenchymal
14%	37%	13%	23%
MSI, CIMP high Hypermutation	SCNA high	Mixed MSI status SCNA low, CIMP low	SCNA high
BRAF mutations		*KRAS* mutations	
Immune infiltration and activation	WNT and MYC activation	Metabolic deregulation	Stromal infiltration TGF beta activation Angiogenesis
Worse survival after relapse	Better survival after relapse		Worse relapse-free and overall survival

Source : Guinney et al

FIGURE 12 – Résumé des résultats des quatre sous-types consensus. A) Réseau de l'ensemble des sous-types de chaque système de classifications. B) Principales caractéristiques

Les associations aux mutations BRAF, KRAS et aux instabilités MSI, CIN et CIMP ont été retrouvées. Mais aucune autre association à des mutations ou des réarrangements n'a été identifiée par l'étude des données TCGA, suggérant une non corrélation entre le génotype et le phénotype. L'étude de la survie a montré un plus mauvais pronostic du

sous-type CMS4 en OS, RFS et SAR [7] et un meilleur pronostic de CMS2 en SAR.

Comparaison de la classification CIT et de la classification consensus

Dans cette classification consensus, la classification CIT est très concordante, avec 87% d'identité obtenue en regroupant les sous-types C1,C5 et C6. Le recoupement entre les différentes classifications et avec la classification consensus est montré dans la Figure 13 : CMS1 correspond au sous-type C2-MSI (89% des C2), CMS2 correspond aux C1-ImmuneDown (100%), C5-WntUp (72%) et C6-normL (75%), CMS3 au C3-KRASm (89%) et enfin le CMS4 à C4 (93%). La classification CIT se distingue sur 2 points : d'une part, elle subdivise le grand groupe CMS1 (37% des tumeurs) en 3 sous-groupes, et d'autre part, les groupes CMS4 et C4 ne sont pas complètement concordants, notre groupe C4 isolant une sous-population du sous-type CMS4 et CMS4 scindant notre sous-type C5-Wnt-Up en 2 parties (Figure 13). Les associations cliniques et moléculaires entre les 2 classifications convergent en grande partie, la classification consensus apportant plus d'information notamment par l'analyse des autres omiques du jeu de données TCGA. Notamment, pour le groupe CMS3 une association intéressante a en plus été observée avec le phénotype CIMP-low et l'hypermutabilité en dehors du contexte MSI. Pour le groupe CMS4/C4, quelques différences sont observées : les légers enrichissements en mutation BRAF, en CIMP et en localisation proximale de la classification CIT ne sont plus retrouvés.

On peut noter également que notre groupe C6 était commun à un seul autre système de classification (E, Sadanandam et al) et notre groupe C4 était quasiment délimité comme celui du groupe F (Schlieker et al) (cf. Figure 13).

Les voies associées aux adénomes festonnés n'ont pas été investiguées dans la classification consensus. Pour les voies des cellules souches, la signature de Merlos-Suárez et al. (2011) est légèrement associée en analyse de *pathways* (p=0.048).

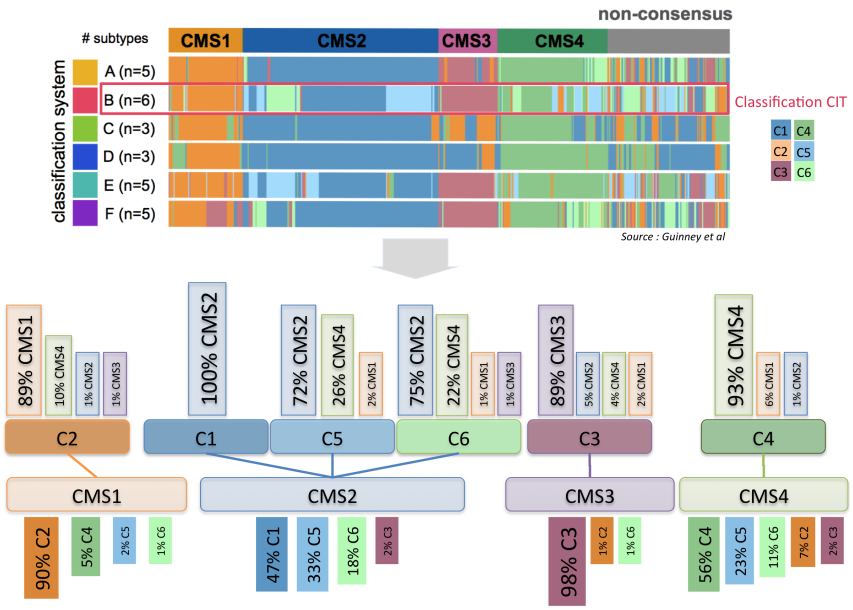


FIGURE 13 – Comparaison des classifications indépendantes et de la classification consensus

7. Suvie Après Rechute

Quatrième partie

Discusssion et Conclusion Générale

Mon travail de classification sur la cohorte CIT a permis de délimiter 6 sous-types de manière robuste car obtenus à partir d'une très large cohorte représentative de la population, définis par des approches permettant de s'assurer de la stabilité des résultats et enfin validés sur une très large série de données indépendantes publiques. En définissant 6 sous-types, la classification a permis, tout en isolant le groupe attendu de tumeurs avec de l'instabilité microsatellitaire MSI, de scinder le grand groupe CIN (des tumeurs ayant une instabilité chromosomique) en 5 sous-types distincts. Ce travail a permis de donner une vue intégrée de la part des différentes formes d'instabilité génomique et des mutations connues dans la diversité des cancers colorectaux. En effet, bien que ces caractéristiques moléculaires soient associées à la classification, elles ne permettent pas à elles seules de bien délimiter des sous-types. De plus, il est important de noter que la classification TNM n'est pas retrouvée associée à ces 6 sous-types, signifiant que le stade est un prédicteur transversal et que la biologie moléculaire *per se* de la tumeur a un poids plus important que le phénotype d'invasion de la paroi colique dans la classification obtenue.

Suite à la publication de ce travail, un consortium international auquel j'ai participé s'est organisé dans le but de déterminer une classification consensus à partir des 6 classifications différentes qui ont été publiées quasiment en même temps dans différents journaux. Les résultats de la classification consensus ont permis de définir 4 sous-types consensus. Ces groupes sont très concordants avec notre classification à l'exception du groupe CMS4 Mesenchymal/C4 Stem Cell, notre groupe C4 correspond à un sous-groupe de CMS4. La classification consensus à l'avantage de mieux caractériser l'ensemble des groupes par l'énorme travail de caractérisation en intégrant notamment tous les omiques mis à disposition par le TCGA.

Trois sous-types sont très clairement identifiés en terme de profils d'expression, avec peu de tumeurs se co-classant dans des groupes différents, et en terme de caractéristiques cliniques et moléculaires : C2-MSI, C3-KRASm et C4-Cancer_Stem_Cell. Ces 3 sous-types ont également été retrouvés par la classification consensus.

Le sous-type C2-MSI était déjà connu mais il n'est pas délimité aussi clairement qu'attendu avec près de 30% de tumeurs diagnostiquées comme non-MSI. Ces proportions sont également retrouvées dans le groupe MSI (CMS1) de la classification consensus. Par contre l'association à l'hypermutabilité est quasiment parfaite ce qui suggèrent que soit une partie des MSS dans le groupe ont été mal diagnostiqués et serait des MSI, soit que des tumeurs MSS avec instabilité nucléotidique ont des profils d'expression comparables aux tumeurs MSI. Par ailleurs, un certain nombre de tumeurs MSI ne sont pas dans ce groupe. Une analyse de l'association au syndrome de Lynch sur nos données montre que ces tumeurs seraient un sous-ensemble des MSI héréditaires. Ceci est intéressant car les tumeurs issues de patients atteints du syndrome de Lynch présentent des différences sur le plan anatomopathologique avec les formes sporadiques (Jass, 2007, 2004) et donc devraient avoir des profils transcriptomiques distincts. De plus, ceci suggère qu'une sous-classification intra-MSI est à envisager pour identifier la diversité des MSI.

Le sous-type C3-KRASm est très intéressant car il isole une sous-population de tumeurs presque toutes mutées pour le gène KRAS. Le gène KRAS est retrouvé muté dans l'ensemble des sous-types dans des proportions d'environ 30%. Ce sous-type correspond donc à une sous-population des mutants KRAS. Ceci laisse penser à un rôle précurseur de la mutation dans la carcinogenèse de ce sous-type, comme APC pour les tumeurs CIN. Ceci est consistent avec l'observation de Fearon and Vogelstein (1990)] que pour un sous-groupe de tumeurs, la mutation de KRAS pourrait être la mutation initiatrice de la tumorigenèse, et avec l'association de cette mutation aux adénomes de type festonnés (les TSA, "traditional serrated adenoma"). La classification consensus apporte de manière très inté-

ressante que ce sous-type est enrichi en tumeurs CIMP-Low, consistent avec l'observation d'une association de ce phénotype à la mutation KRAS (Curtin et al., 2011) et en tumeurs MSS hypermutées, le nouveau type d'instabilité décrite dans les cancers colorectaux. De plus, l'analyse consensus confirme une surexpression des gènes métaboliques, suggérant un "swicth" métabolique dans la veine de l'effet Warburg (Levine and Puzio-Kuter, 2010; Nakajima and Van Houten, 2013). Ce phénotype moléculaire a été également retrouvé dans un sous-type du cancer de l'estomac (Lei et al., 2013).

Le sous-type C4-Cancer Stem Cell est lui le sous-type le plus distinct en terme de profil d'expression avec des caractéristiques moléculaires propres aux tumeurs plus agressives. Ces caractéristiques ont également été retrouvées par la classification consensus. La composition en cellules souches n'a pas été investiguée par la classification consensus mais une association significative à la signature de Merlos-Suárez et al. (2011) est toutefois observée. Cette analyse reste très hypothétique mais il était intéressant que divers critères de "souchitude" soient convergents dans ce sous-type comme la quiescence, suggérée par la sous-expression des gènes du cycle cellulaire, la sur-expression de gènes d'une signature de cellule souche, la sur-expression de gène d'une signature spécifique du haut et bas de la crypte (Kosinski et al., 2007) et l'association à la signature EMT (en référence aux cellules souches mésenchymateuses (Uccelli et al., 2008)). De plus, le mauvais pronostic de ce sous-type peut être un signe de résistance plus développée de ces tumeurs à la chimiothérapie qui constitue une des caractéristiques des cellules souches cancéreuses. Ces éléments sont limités et ne constituent que des pistes de réflexion. Des expériences plus directes permettraient d'apporter des preuves plus tangibles comme la visualisation sur lame des tumeurs annotées C4 par des marqueurs de cellules souches cancéreuses comme LGR5, mais vu la difficulté d'identifier les marqueurs des cellules souches, ça pourrait ne pas être concluant. L'association à l'EMT peut également suggérer une certaine agressivité en raison de son lien avec le mécanisme de dissémination métastatique. La signature EMT de ce sous-type peut ne pas refléter l'agressivité ou la "souchitude", mais la composition en stroma. C'est l'hypothèse qu'ont validée Isella et al. (2015). Ils ont montré par l'utilisation du transcriptome de xénogreffes (ce qui permet de séparer le signal tumoral humain et du signal stromal murin) qu'une partie de la signature correspondait à la composante stromale des échantillons (déplétée dans le signal humain issu des xénogreffes).

Par ailleurs, ce sous-type n'est pas délimité de la même manière par la classification consensus. Le groupe CMS4 inclus en plus de notre sous-type C4, une partie du sous-type C5 CIN Wt up. Ceci peut se comprendre car les 2 sous-types au regard des profils d'expression ont des similitudes, C4 surexprime ou sousexprime plus fortement une grande partie des gènes communs à C5. Un continuum entre ces 2 sous-types pourrait expliquer ces résultats. Ce qui par conséquent rendrait moins probable l'origine festonnée de C4. De plus, la signature cellule souche était également retrouvée dans le sous-type C5 mais la signature du bas de la crypte ne convergeait pas. Notre sous-type est donc peut-être un sous-type de CMS4 ou alors CMS4 rend compte d'une signature différente, comme la signature stromale, et ne reflète pas forcément un sous-type. Ceci peut expliquer la différence d'associations à BRAF et à CIMP entre notre classification et la classification consensus.

Les 3 autres groupes (C1, C5, C6) se regroupent par leur caractéristique d'instabilité chromosomique et sont regroupés en un seul sous-type CMS2-Canonical dans la classification consensus, ce qui n'exclut donc pas leur existence. On voit ici la limite de la classification consensus qui peut-être contrainte par le nombre de classes des partitions d'origine.

Le groupe C6-Norm-like est toutefois particulier. Cette similitude aux profils d'expres-

sion d'échantillons de tissus normaux adjacents laisse perplexe mais ce phénomène a déjà été décrit dans d'autres cancers comme le sein et le foie (Sørlie et al., 2001). De plus, le sous-type est associé à un mauvais pronostic validé sur le jeu de donnée indépendant et les profils génomiques de ces tumeurs sont très instable. Enfin, il a été également isolé dans la classification de Sadanandam et al. (2013).

Les deux derniers groupes "C1-CIN Immune down" et "C5-CIN Wnt up" ont été beaucoup plus difficiles à caractériser cliniquement et moléculairement. Ils partagent des profils de gènes dérégulés similaires ce qui peut expliquer le taux de co-classement intermédiaire pour les tumeurs de ces 2 groupes. La séparation en 2 groupes peut donc être discutée. Toutefois l'analyse de "pathways" et les gènes spécifiques de ces 2 entités sont très distincts.

La caractérisation fine du génome par les altérations ou des mutations dans notre cohorte ou dans la cohorte consensus n'a pas apporté plus d'information. Il existe beaucoup d'altérations et de mutations, mais aucune n'est apparue spécifique d'un sous-type contrairement à ce qui peut s'observer dans certains cancers comme les gliomes avec les mutations IDH et la co-délétion 1p/19q (Suzuki et al., 2015). Cela suggère que ces altérations constituent des événement plutôt "passengers" et confirme la part importante de l'étude des données d'expression dans les cancers colorectaux.

L'un des intérêts qu'une classification peut avoir en pratique clinique est son pouvoir pronostic. Le mauvais pronostic du groupe C4-Stem_Cell a été confirmé dans le groupe CMS4. Notre étude du pronostic a montré que la combinaison des 2 sous-types de mauvais pronostic apporte de l'information complémentaire au stade, qui n'est toutefois pas détrôné, ce qui reste vrai en ajoutant au modèle le prédicteur pronostique Oncotype.

Par ailleurs, les classifications permettent de stratifier la recherche de marqueurs pronostiques, approche dont j'ai pu montré l'intérêt dans 2 autres travaux de ma thèse (Collura et al., 2014; Manceau et al., 2015) (mis en Annexe C (page 149) car ils ne portent pas sur des données omiques mais sur des mutations).

L'autre intérêt pour la pratique clinique est que la classification permet la recherche de cibles de traitement intra-sous-type. Un fameux exemple est la réponse au Trastuzumab dans le cancer du sein pour les tumeurs ERBB2+. De manière intéressante, il a été montré que le bénéfice du traitement était observé dans les tumeurs ERBB2+ négative pour le récepteur aux œstrogènes (ER-) (Prat et al., 2014). La classification initiale de Sørlie et al. (2001) ne distinguait pas bien les 2 formes de ERBB2+. Or une classification, que l'on a proposée (Guedj et al., 2012), divisait bien les ERBB2+ en 2 sous-types en fonction du statut de ER et nous avons validé que la réponse au Trastuzumab était bien spécifique du sous-types ERBB2+/ER- (travaux non publiés). Un autre exemple récent dans le cancer colorectal est l'efficacité d'anti-PD1 spécifiquement dans des tumeurs métastatiques de type MSI (Le et al., 2015).

Au final, les 2 classifications proposées apportent des informations convergentes sur la biologie des cancers colorectaux. La classification consensus a le grand avantage par la taille de la cohorte d'avoir une validité plus importante quant aux associations observées. Les 2 classifications ont quelques spécificités sur la manière de délimiter le sous-type de mauvais pronostic C4/CMS4 et notre classification subdivise le grand groupe "canonique". Les spécificités de notre classification pour C4 peuvent refléter une réalité biologique que les autres classifications n'ont pas bien délimité. En effet, un consensus n'est pas forcément la panacée, il peut donner une vision lissée de la réalité. L'intérêt des spécificités de l'une ou l'autre des classifications pourra être tranché par l'intérêt que chacune va avoir notamment dans la stratification d'études de la réponse au traitement. Ceci fait parti des perspectives de notre projet. Nous prévoyons d'étudier le pronostic et le pouvoir prédictif de notre classification dans un premier temps sur un essaie thérapeutique, l'essai PETACC08 qui

visait à estimer le bénéfice du Cextucimab. Pour ce la, nous sommes en train de développer un prédicteur applicable à des tumeurs fixées, qui sont les données les plus accessibles en clinique.

Conclusion

Grâce à l'étude des données génomiques, j'ai pu mettre en évidence que le cancer du côlon est bien une maladie hétérogène présentant différents sous-types avec des caractéristiques distinctes laissant fortement penser qu'ils proviennent de différentes voies de carcinogenèse. Cette classification est confortée par un travail conséquent d'un consortium international de classification consensus qui a permis de mieux caractériser chacun des groupes. Toutefois des spécificités de notre classification pouvant être d'intérêt pour la clinique sont perdues dans la classification consensus. Ces résultats sont importants pour la compréhension de la maladie et pour envisager des traitements plus personnalisés et pour la recherche de marqueurs pronostiques. De ce travail, un prédicteur sur données fixées est en cours de réalisation pour son utilisation dans des essais thérapeutiques afin i) de valider l'existence de ces sous-types et leur proportion dans la population générale, ii) d'évaluer correctement le pronostic des sous-types et iii) d'évaluer la réponse au traitement en fonction des sous-types.

Comme je l'ai compris de la phrase de Jeremy Jass en épigraphe, maintenant que l'on voit la réalité de l'existence de l'hétérogénéité des cancers colorectaux, on peut commencer à comprendre la biologie qu'il y a derrière.

Appendix

Annexe A

Méthode de mesure des instabilités du génome

Les méthodes d'évaluation du statut MSI

La mesure du statut MSI peut se faire soit directement par allélotypage au niveau de microsatellites, soit indirectement par la recherche de défaillance de gènes du MMR.

Pour définir le statut MSI d'une tumeur par allélotypage, une grande variété de méthodes et de critères quant à la sélection des microsatellites marqueurs, à leur nombre et au seuil, ont été utilisés. En 1997, lors d'un workshop du National Cancer Institute à Bethesda visant à faire le point et unifier le domaine, un panel de référence de 5 marqueurs microsatellitaires, 3 dinucléotides (D5S346, D2S123, D17S250), 2 mononuclotidique (BAT25, BAT26) a été préconisé pour la définition du MSI. Selon ces directives, MSI (ou MSI-High) était défini comme l'instabilité de 2 ou plus des 5 microsatellites, MSS (pour MicroSatellite Stable) comme aucune instabilité sur aucun des 5 marqueurs et MSI-L (low) représentant le phénotype intermédiaire (Boland et al., 1998). Toutefois, l'utilisation du tissu normal et de marqueurs dinucléotides ont mené à accepter, lors de la révision en 2002 des critères Bethesda [1], un autre panel plus simple utilisant peu de matériel de par l'utilisation de PCR. Ce nouveau panel est constitué de cinq marqueurs mononucléotidiques quasi-monomorphes dans la population mondiale (BAT-25, BAT-26, NR-21, NR-24 et NR-27), mis au point dans le laboratoire d'accueil (Suraweera et al., 2002; Umar et al., 2004). La pentaplex est hautement sensible et spécifique, et ne nécessite pas une comparaison de l'ADN tumoral et germinal. Par ailleurs, ce test requiert très peu de matériel biologique et il est hautement reproductible. Ce panel nommé "PCR pentaplex" est, à l'heure actuelle, la méthode de référence utilisée pour déterminer le statut MSI des tumeurs.

La détermination du statut MSI des tumeurs peut être également réalisée de manière indirecte en étudiant l'expression des protéines du système MMR par immunohistochimie (IHC). L'extinction/perte d'expression d'au moins une des 4 principales protéines MMR (MLH1, MSH2, MSH6 et PMS2) est recherchée par comparaison entre le tissu tumoral et la muqueuse saine adjacente. En général, une perte d'expression de MLH1/MSH2 est associée à une perte d'expression de PMS2/MSH6 mais la réciproque n'est pas vraie. L'IHC est une méthode présentant les avantages d'être facile à mettre en place, rapide, peu onéreuse, et peut donc être utilisée en routine mais elle aurait plus un intérêt pour évaluer les formes héréditaires (les syndromes de Lynch) associées à une mutation des gènes du MMR.

1. HNPCC workshop de 2002 organisé par le NCI à Bethesda.

Les méthodes d'analyse du CIMP

L'une des méthodes classiquement utilisées pour définir le phénotype CIMP d'une tumeur est la méthode proposée par Weisenberger et al. (2006) : 5 marqueurs sont utilisés CACNA1G, IGF2, NEUROG1, RUNX3, and SOCS1. Après le traitement au bisulfite de l'ADN, le taux de méthylation est mesuré par la technologie Methylight. Un échantillon est assigné CIMP+ si plus de 3 promoteurs sont méthylés, CIMP- si moins de 2 promoteurs sont méthylés.

Les autres panels proposés sont indiqués dans la table ci-dessous extraites de Curtin et al. (2011) :

Étude	Panel de marqueurs CIMP	Notes
Toyota et al.	CDKN2A (p16), MINT1, MINT2, MINT12, MINT17	Pioneering work to identify markers that distinguish CIMP from age-related methylation
Park et al.	MINT25, MINT27, MINT31, MLH1, THBS1 CDKN2A, MINT1, MINT2, MINT31, MLH1	So-called "classic" or traditional panel
Weisenberger et al.	CACNA1G, IGF2, NEUROG1, RUNX3, SOCS1	"New" panel based on stepwise screen of 195 markers
Ogino et al.	CACNA1G, CDKN2A, CRABP1, MLH1, NEUROG1	Selected markers to distinguish high-level from low-level methylation
Shen et al.	CIMP1 : MINT1, MLH1, RIZ1, TIMP3, BRAF mutation ; CIMP2 : MINT2, MINT27, MINT31, Megalin, KRAS mutation	Examined 27 CpG sites, proposed optimal epigenetic and genetic markers to identify CIMP1, CIMP2, or CIMP-
Tanaka et al.	CACNA1G, CDKN2A, CHFR, CRABP1, HIC1, IGF2, IGFBP3, MGMT, MINT1, MINT31, NEUROG1, p14, RUNX3, SOCS1, WRN	Correlation structures of markers and CIMP differ by KRAS and BRAF status
Ang et al.	Total of 202 CpG sites differentially methylated between tumor and normal	Comprehensive DNA methylation profiling in 807 cancer genes
Kaneda and Yagi	Group 1 : IGF2, LOX, MINT1, MINT2, MINT31, MLH1, RUNX3, SOCS1 ; Group 2 : ADAMTS1, DUSP26, EDIL3, ELMO1, FBN2, HAND1, IGFBP3, NEUROG1, RASSF2, STOX2, THBD, UCHL1	Comprehensive DNA epigenotyping of genomewide regions indentified two groups (high and intermediate to low methylation)

Annexe B

Étude des cellules souches dans les lignées cellulaires colorectales

Manuscript
Click here to download Manuscript: Collura et al revised 03-09-2012.doc
Click here to view linked References

EXTENSIVE CHARACTERIZATION OF SPHERE MODELS ESTABLISHED FROM COLORECTAL CANCER CELL LINES

Ada Collura [1,2], Laetitia Marisa [3], Diletta Trojan [1,2], Olivier Buhard [1,2], Anaïs Lagrange [1,2], Arnaud Saget[1,2], Marianne Bombled [4], Patricia Méchighel [1,2], Mira Ayadi [3], Martine Muleris [1,2], Aurélien de Reynies [3], Magali Svrcek [1,2,5,6], Jean-François Fléjou [1,2,5,6], Jean-Claude Florent [4], Florence Mahuteau-Betzer [4], Anne-Marie Faussat [2], Alex Duval [1,2 $]

(1) Inserm, UMRS_938 - Centre de Recherche Saint-Antoine, Equipe “Instabilité des Microsatellites et Cancers”, F-75012, Paris, France ;
(2) Université Pierre et Marie Curie-Paris6, Paris, France ;
(3) Programme “Cartes d'Identité des Tumeurs”, Ligue Nationale Contre le Cancer, Paris, France;
(4) UMR 176 CNRS/Institut Curie, Univ. Paris-Sud, Orsay, France.
(5) AP-HP, Hôpital Saint-Antoine, Service d’Anatomie et Cytologie Pathologiques, Paris, France;
(6) AP-HP, Hôpital Saint-Antoine, Tumorothèque CancerEst, F-75012, Paris, France ;

[$] Address for correspondence:

Alex Duval
INSERM UMRS 938
Team "Microsatellite Instability and Cancer"
Hôpital Saint-Antoine
184, rue du Faubourg Saint-Antoine, F75571
Paris cedex 12
Tel: 33 (0)1 49 28 66 80
Fax: 33 (0)1 49 28 66 81
Email: alex.duval@inserm.fr

Ada Collura
INSERM UMRS 938
Team "Microsatellite Instability and Cancer"
Hôpital Saint-Antoine
184, rue du Faubourg Saint-Antoine, F75571
Paris cedex 12
Tel: 33 (0)1 49 28 66 72
Fax: 33 (0)1 49 28 66 81
Email: ada.collura@inserm.fr

Keywords: Colorectal cancer, colon cancer cell lines, spheres, microarray.

ABSTRACT (250 Words)

Links between cancer and stem cells have been proposed for many years. As the cancer stem cell (CSC) theory became widely studied, new methods were developed to culture and expand cancer cells with conserved determinants of “stemness”. These cells show increased ability to grow in suspension as spheres in serum-free medium supplemented with growth factors and chemicals. The physiological relevance of this phenomenon in established cancer cell lines remains unclear. Cell lines have traditionally been used to explore tumor biology and serve as preclinical models for the screening of potential therapeutic agents. Here, we grew cell forming spheres (CFS) from 25 established colorectal cancer cell lines. The molecular and cellular characteristics of CFS were compared to the bulk of tumor cells. CFS could be isolated from 72% of the cell lines. Both CFS and their parental CRC cell lines were highly tumorigenic. Compared to their parental cells they showed similar expression of putative cancer stem cell markers. The ability of CRC cells to grow as CFS was greatly enhanced by prior treatment with 5-fluorouracil. At the molecular level, CFS and parental CRC cells showed identical gene mutations and very similar genomic profiles, although microarray analysis revealed changes in CFS gene expression that were independent of DNA copy-number. We identified a CFS gene expression signature common to CFS from all CRC cell lines and that was predictive of disease relapse in CRC patients. In conclusion, CFS models derived from CRC cell lines possess interesting phenotypic features that may have clinical relevance for drug resistance and disease relapse.

INTRODUCTION

The cancer stem cell (CSC) theory has generated much interest in both the research and clinical communities, notably for colorectal cancer (CRC) [1-3]. According to this hierarchical model, CSCs are defined by their ability to self-renew indefinitely, while also being able to differentiate and generate both tumorigenic and non-tumorigenic daughter cells that constitute the bulk of the tumor [4]. CSCs are thought to have a low rate of division and proliferation that helps them resist various chemotherapies and radiation. Both these forms of treatment preferentially affect highly proliferative cells, thus potentially making CSCs a major reason for the failure of anticancer treatment [5]. In tumors including CRC, the presence and survival of CSCs has been suggested as a key mechanism underlying chemoresistance and disease relapse [6-7]. This original interpretation of the CSC theory has recently been challenged, however. Some authors have highlighted the plasticity of the CSC phenotype and suggested that it could be induced through dedifferentiation processes influenced by the tumor cell environment [8].

Cancer cell lines have been widely used to explore tumor biology and as preclinical models for the screening of potential therapeutic agents. They are a valuable resource that can be used repeatedly and have also been well characterized with respect to mutational and gene expression profiles [9]. Similar frequencies of gene mutation have been reported in primary tumors and in cancer cell lines derived from the same primary site. Cancer cell lines are not contaminated with stromal tissue, which can sometimes affect the interpretation of data obtained from primary tumors [10]. Furthermore, cancer cell lines often faithfully represent the tumor from which they were isolated [11-12]. It remains to be determined whether cancer cell lines are relevant biological tools to study the role of CSCs in tumorigenesis. A number of authors have hypothesized the existence of cancer stem-like cells in these cellular models [13-14], however their phenotype is still poorly characterized. Moreover, it is not known whether cancer stem-like cells from cell lines have any clinical relevance [15-16].

In order to study cells from cancer cell lines that could display a stem-like phenotype, the first requirement is to have a system in which they can be propagated. The ability to grow in suspension as spheres in serum-free medium supplemented with specific growth factors and chemicals has been described for the expansion of neuronal stem cells [17]. Sphere culture has also been proposed as a valuable method for isolating cancer cells with conserved stemness determinants that are able to propagate in defined medium [18-21]. In the present study we have used this method to grow cell forming spheres (CFS) from a panel of 25 CRC cell lines. These cell lines were selected to reflect the heterogeneity of CRC in terms of showing microsatellite stability (MSS) or instability (MSI). CRC is a complex tumor entity that includes distinctive molecular phenotypes associated with different clinical features, including response to chemotherapy [22-24]. We investigated the cellular and molecular phenotypes of CFS derived from this panel of CRC cell lines, with particular reference to treatment resistance and CSC features.

RESULTS

The ability of CRC cell lines to grow as CFS in serum-free medium enriched with growth factors and chemical supplements is highly variable

We tested the ability of established MSI and MSS CRC cell lines (n=25) to grow as CFS in sphere-medium. CFS were obtained from 10 MSI and 8 MSS cell lines (total 18/25, 72%) (**Fig. 1A**). In the remaining 7 CRC cell lines (4 MSI, 3 MSS), a small number of cells remained afloat but died after 1-4 passages (1-6 weeks) in sphere-medium (data not shown). Morphologically, the CFS were quite heterogeneous and ranged from densely packed spheres with almost indiscernible individual cells to more loosely packed spheres or individual floating cells (**Fig. 1B**).

CFS display identical clonal genomic alterations to their parental CRC cells

Extensive analysis using SNP microarrays revealed that genomic alterations due to chromosomal instability (CIN) were similar between parental CRC cell lines and their related CFS. This was observed both with MSI and MSS CRC cell lines displaying low and high levels of CIN, respectively. SNP analysis was performed on 13 CFS/adherent cell lines (LIM2405, HCT8, HCT116, HCT15, TC-7 CO115, RKO, LS411, V9P, HT29, SW620, Colo320) (**Fig. 2A** and data not shown). The status of DNA microsatellites that constitute accurate markers of the 'history' of each tumor cell in CRC cell lines displaying MSI was also analyzed. Mutation analysis of DNA microsatellite sequences in 10 $CFS^{Positive}$ MSI cell lines (HCT116, CO115, HCT15, HCT8, ISHI, LIM2405, LS411, RKO, TC-7, TC71) revealed identical patterns of alteration in non-coding (BAT26, BAT25, NR21, NR25, NR27) and coding (*TGFBR2, RAD50, MSH6, MSH3, MBD4, BAX, ATR, BLM*) repeats between parental cells and their corresponding CFS (**Fig. 2B**). Thus, CRC cell lines and their corresponding CFS progeny are clonally identical, indicating the CFS phenotype arises from the selection of pre-existing clones in the parental cell lines that are able to grow under specific conditions of serum deprivation.

Acquisition of the CFS phenotype by CRC cells is associated with specific changes in gene expression

The gene expression profiles of 13 $CFS^{Positive}$ CRC cell lines (8 MSI: LIM2405, HCT8, HCT116, HCT15, TC-7 CO115, RKO, LS411; 5 MSS: V9P, HT29, SW620, Colo320, FET) and their corresponding CFS populations were compared using microarrays. Principal component analysis of the profiles revealed the CFS grouped together with their parental CRC cell lines (**Supplementary Fig. S1A**). A total of 264 genes displayed significant down- or up-regulation in CFS (*P*-value < .001, paired moderated *t*-test) and are listed in **Supplementary Table S1**. Specific signaling and metabolic pathways were associated with the $CFS^{Positive}$ gene expression signature (**Table 1**). As expected, a number of genes from this signature reflected the different culture conditions used for adherent and CFS cells and were linked mainly to cell metabolism and growth factor pathways (**Table 1**). The CFS signature also included genes related to stemness and to cellular mechanisms associated with treatment resistance, such as transmembrane transporters, apoptosis and DNA damage (**Table 1**). Specific chromosomal regions were enriched with genes from the CFS signature (**Supplementary Fig. S1B**). These regions displayed a similar genomic status (DNA copy-number) in CFS and the parental CRC cell line, suggesting that gene deregulation occurred *via* epigenetic processes that were independent of DNA copy number. Regions that were frequently down or up-regulated in CFS (> 30%) are shown in the lower panel of Figure S1B.

Both CFS and their parental CRC cell lines are highly tumorigenic and they show similar expression of CSC markers

Serial engraftments of 8 CRC cell lines (5 $CFS^{Positive}$: HCT15, HCT116, LoVo, V9P, RKO; 3 $CFS^{Negative}$: LS174T, LIM1215, KM12) were performed in nude mice. All 8 cell lines tested were found to be highly tumorigenic, even when only 200 CRC cells were injected (**Fig. 3A**). CFS cells derived from HCT116 and LoVo cell lines were also highly tumorigenic. Low numbers of injected CFS cells (200 or 500 cells) derived from both these MSI CRC cell lines were less tumorigenic compared to the parental cell lines (**Fig. 3B** and data not shown). However, no difference was apparent when 1000 cells were injected. The expression of putative colorectal CSC markers (CD44, CD133, CD166, CD24, CD29, EPCAM, ALDH, OLFM4, LGR5) evaluated using arrays was not significantly different between 13 pairs of CFS cells and their corresponding parental cell lines (8 MSI: LIM2405, HCT8, HCT116, HCT15, TC-7 CO115, RKO, LS411; 5 MSS: V9P, HT29, SW620, Colo320, FET) (**Fig. 4A**). The expression of 6 putative colorectal CSC markers (CD44, CD166, CD133, CD24, EpCAM, CD29) was also compared using flow cytometry in the 25 parental CRC cell lines **(Supplementary Table S2)**. Overall, the expression of these markers was highly variable and none was expressed exclusively in $CFS^{positive}$ cell lines. Moreover, their expression was highly variable over time, as shown for CD44 and CD166 expression in cell sub-populations sorted by FACS from LS174T and HCT116 parental cells (**Fig. 4B**). Finally, we quantified the expression of putative colorectal CSC markers (CD44, CD133, CD166, CD24, CD29, EPCAM) in 15 primary CRCs and 12 CRC tumor xenografts established from these primary CRCs and grown in nude mice. Most markers were expressed at significantly higher levels in CRC cell lines compared to primary CRCs and/or tumor xenografts (**Fig. 4C**). Overall, these results demonstrate that the CFS phenotype derived from CRC cell lines is only weakly related to the putative CSC phenotype from primary tumors. In line with this, a CFS assay showed similar results using sorted populations of $CD166^{+}CD44^{+}EpCAM^{high}$ or $CD166^{-}CD44^{-}EpCAM^{low}$ HCT116 cells (**Fig. 4D**).

The ability of CRC cells to grow as CFS is strongly increased by prior treatment with 5-Fluorouracil

The ability of LoVo and HCT116 cell lines to grow as CFS in sphere-medium increased following treatment with the chemotherapeutic agent 5-Fluorouracil (5-FU) for 5 days at $IC_{10\%}$ (5 µM for HCT116 and 7,5 µM for LoVo cells) (**Fig. 5A** and **5B** and data not shown). In contrast, 5-FU-resistant clones from the LIM1215 and LS174T $CFS^{Negative}$ cell lines remained unable to grow in sphere-medium after 5-FU treatment at IC_{10} (**Fig. 5A**). CRC cell lines were also compared to their CFS counterparts for resistance to 5-FU. In 6 CRC cell lines tested (FET, HCT116, LS411, V9P, TC71, LIM2405), the CFS displayed greater resistance to 5-FU than their parental cells (**Fig. 5C**). Both $CFS^{Positive}$ and $CFS^{Negative}$ CRC cell lines displayed marked differences in resistance to this drug (data not shown). LIM2405 showed a strong predilection to grow as CFS in sphere-medium (data not shown), yet the adherent cells and CFS showed similar resistance to 5-FU (**Fig. 5C,** bottom and right panel).

CFS established from CRC cell lines share a gene signature that predicts disease relapse in CRC patients

Fifty-five genes that were differentially expressed between CFS and their corresponding CRC cell lines displayed a high level of up- or down-regulation (Log2-fold change for CFS/parental cell line > 0.8; **Supplementary Table S1 and Figure 6).** This 55-gene CFS expression signature predicted disease relapse in a previously

described, retrospective series of stage II and III CRC patients [25]. Tumors were classified into two groups according to the expression level of the 55 genes (T-CFS^{High} or T-CFS^{Low}; for further details, see the Materials and Methods section and **Figure 6A**), where T-CFS^{High} and T-CFS^{Low} correspond to tumors displaying high or low levels of expression of the CFS signature, respectively. Patients with T-CFS^{High} tumors showed significantly shorter disease-free survival (DFS) compared to T-CFS^{Low} patients (**Supplementary Figure S3A**, bottom and left panel). To confirm these clinical findings, the same analysis was performed on another, independent cohort of stage II and III CRC patients [26]. A trend for similar clinical impact of the 55-gene CFS expression signature was observed (**Supplementary Figure S2A**, bottom and right panel). Following normalization of the data, the two patient series were combined to achieve greater statistical power. In the overall series and in univariate analysis, the survival of stage II and III CRC patients was associated with the expression level of the 55-gene CFS signature (**Fig. 7**, left panel and **Supplementary Figure S2A**, top and left panel).

Since the 55-gene signature showed a trend for association with patient outcome in the second dataset, we sought to identify a novel prognostic signature by selecting genes that were individually associated with outcome in the first dataset. This allowed us to define an 8-gene CFS prognostic signature (**Fig. 6B** and **Supplementary Table S1**) that showed stronger associations with disease relapse in both the individual and combined patient cohorts (**Fig. 7,** right panel and **Supplementary Figure S2B**, top and right panel). To ensure this finding was related to the CFS 55-gene selection and not to the methodology, the same approach was repeated 1,000 times using signatures from 55 genes selected at random. This analysis did not increase the false positive rate in the second dataset (data not shown). In the overall series and in multivariate Cox analysis that included TNM stage, the survival of patients with stage II or III CRC was confirmed to be associated with the expression level of 8 genes from the CFS signature **(Table 2**).

CFS can be used as models to identify new drugs that are more efficient at killing CFS than 5-fluorouracil

Because of the current clinical interest in CFS, the Institut Curie-CNRS chemical library (8560 compounds) was screened using the HCT116 (MSI) and FET (MSS) CRC cell lines (**Supplementary Fig. S3A**). Fifteen new compounds were identified with the ability to kill both parental CRC cells and CFS from HCT116 and/or FET at low concentration ($IC_{50\%}$ < 1µM; **Supplementary Fig. S3B**). Of note, 8/15 (53%) of these new compounds belong to the Nitrofurans' family previously reported to display anticancer and antioxidant properties (entries 1-8, **Supplementary Table S3**) [29-30]. Since the library contains 240 of this class of compound (2.8% of the total), Nitrofurans were significantly over-represented amongst the 15 drugs with high effectiveness against CFS ($P = 2.4x10^{-9}$, Fisher's exact test). In contrast to results obtained with 5-FU (see **Fig. 5B**), the ability of HCT116 and FET cell lines to grow in serum-free medium did not increase following treatment with 3 of the new compounds for 5 days at $IC_{10\%}$ (**Supplementary Fig. S3C**).

DISCUSSION

Not all CRC cell lines contain CFS. This result may reflect the true heterogeneity of CRC but may also be due to events that occurred *in vitro* during or after the establishment of CRC cell lines. All CRC cell lines tested here were highly tumorigenic in mouse xenograft assays. We did not observe increased tumorigenicity of CFS-positive compared to CFS-negative CRC cell lines, nor of CFS compared to their parental cell line counterpart. Surprisingly, at lower numbers of injected cells, CFS were less tumorigenic compared to their parental cell line. We have no obvious explanation for this result. In any case, CRC cell lines are not ideal models for evaluating tumorigenicity because of their high level of heterogeneity. Compared to primary colorectal tumors, aberrant expression of putative CSC markers was observed in the CRC cell lines investigated here, but was not different to that of CFS, as already reported for HCT116 [16]. Moreover, the current transcriptome analyses revealed only weak correlations between CFS and the expression of stem cell gene markers. These results highlight the fact that CSC markers are not specific for CFS. The CFS phenotype is therefore quite different to that of putative CSCs from primary tumors. It would be interesting to perform similar experiments relating to the expression of CSC markers, tumorigenicity and gene signatures using primary tumors and their sphere counterparts.

Although clearly different to putative CSCs from primary tumors, an interesting question raised by this study is whether the CFS models established from cancer cell lines have clinical relevance. The differential expression of genes in CFS relative to their parental cells is partly a result of the presence of growth factors in one medium and not in the other. This could involve growth factor pathways as well as CSC-like genes. In support of this, it was recently shown that CFS cultures contain only a fraction of CSCs and therefore cannot be regarded as pure CSC models. With this in mind, our data highlight that CFS have retained interesting phenotypical characteristics, including increased resistance to 5-FU in standard culture conditions. Moreover, the 55-gene CFS signature identified here was common to all CRC cell lines and was predictive for disease relapse in CRC patients. Although validated in two independent CRC series, these findings require confirmation in additional studies using larger cohorts of patients. In summary, CFS models derived from CRC cell lines have interesting phenotypical features and may have clinical relevance for drug resistance and disease relapse, but are unlikely to serve as models for putative CSCs in primary CRC.

The genomic and mutational analyses of CRC cell lines and CFS performed here, including the study of microsatellite DNA repeats in MSI cell lines, help to explain the origin of CFS. The status of DNA microsatellites constitutes an accurate marker of the 'history' of each tumor cell. These genetic markers, including the non-coding microsatellite repeats (BAT26, BAT25, NR21, NR25, and NR27), were identical between CFS and their parental CRC cells. The CFS phenotype therefore corresponds to a cellular state achieved only by some clones under specific culture conditions or exposure to drugs. Considering also the transcriptome analyses, our results indicate that only a fraction of cells from CRC cell lines have the capacity to rapidly adjust the expression of specific genes and hence to persist as CFS under challenging growth conditions. These data agree with other recent studies demonstrating widespread plasticity and dedifferentiation processes that affect some tumor cells under specific environmental conditions, particularly CRC cells [31-32,4,33-35].

The expression signature of normal intestinal stem cells, also shared by cells with a stem-like cell phenotype within primary CRC, was recently found to be predictive of disease relapse in CRC patients [36]. However, another group reported the CSC gene signature in CRC may reflect the differentiation status of malignant tissue [37], rather than reflecting the number of CSCs as suggested in the former study. The results presented here relate only to the properties of CFS cultures derived from CRC cell lines, not from primary tumors. We speculate that the gene expression pattern observed in our CFS model corresponds to that of poorly differentiated 'progenitor cells', or dedifferentiated CRC cells, due to the altered cell culture conditions. This pattern could be expressed in an important fraction of cancer cells within the primary tumor. The risk of developing a recurrence of CRC might be associated with the expression of specific genes in these tumor cells and which are required for tumor regeneration following cancer therapy. Interestingly, 5 of the genes included in the limited 8-gene CFS signature are predicted to have indirect associations with signaling pathways. *LDLR* and *HMGCS1* participate in SREBP control of the lipid pathway by stimulating lipid synthesis, while *FGFR4* is a growth factor reported to be associated with poor prognosis and aggressive disease in many different cancer types including colon cancer [27]. Although still poorly described, *AHNAK2* and *FAM46A* are both upregulated in cisplatin-resistant gastric cancer cell lines [28].

The clinical and molecular heterogeneity of CRC is a major limitation of this type of study. Principal component analyses of the transcriptomic profiles showed that considerable heterogeneity remains between CFS displaying MSI or MSS. The response to chemotherapy is thought to be different between patients with MSI and MSS colon tumors, even when considering tumors with the same stage of disease [37]**.** The CFS gene signature identified here was shared by both MSI and MSS CRC. Nevertheless, we could not evaluate the clinical relevance of MSI-specific or MSS-specific CFS signatures for patients with these tumor subgroups because MSI status is not contained in the publicly available Affymetrix U133P2 datasets with Recurrence Free Survival annotations used for this study. This is a subject for future investigation.

In conclusion, the current findings support the existence of CFS subpopulations within CRC cell lines and provide a framework to explain the origin of these tumor cells. They also suggest that CRC cells acquire the CFS phenotype through mechanisms that are only weakly related to CSC, but which could nevertheless be important for the development of chemoresistance by CRC cells *in vivo* in primary tumors. The screening of 8,560 potential anticancer agents using an assay involving CFS populations derived from CRC cell lines may therefore be a useful approach to identify novel drugs for clinical application. The 15 new drugs identified in this study, including several new compounds belonging to the Nitrofurans' family, are currently being tested for toxicity and efficacy in pre-clinical studies using animal models.

MATERIALS AND METHODS

Tissue collection and preparation of xenografts

Human colon tissue fragments were obtained in accordance with the ethical standards of the institutional committee on human experimentation from 15 patients undergoing a colon resection for CRC at the Saint-Antoine hospital in Paris. A

biobank collection of 30 tumors stored at -80°C was used to obtain fresh tumor tissue after engraftment in 5-week old female nude mice (nu/nu). Tumor implantation procedures were performed as previously described [38]. Twelve tumor xenografts were grown for between 1 and 4 months after engraftment. Cancer tissues were intensively washed four times in PBS solution containing antibiotics and then incubated overnight in DMEM/F12 (PAA) containing penicillin (500U/ml), streptomycin (500 µg/ml), amphotericin B (0,25 µg/ml), ceftazidime (50µg/ml). Enzymatic digestion was performed using collagenase (1,5 mg/ml, SIGMA) and hyaluronidase (20 µg/ml, SIGMA) in PBS for 1 hour. These digests were used for FACS analysis.

Cell culture

Colon cancer cell lines were cultured in DMEM media supplemented with 10% FCS (20% for Caco-2 cell line), 100 units/ml penicillin G and 100 µg/ml streptomycin. For the culture of CFS, cell lines were grown in serum-free DMEM/F12 media supplemented with 100 units/ml penicillin G, 100 µg/ml streptomycin, 6 g/l glucose, 1 mg/ml $NaHCO_3$, 5 mM HEPES, 2mM L-glutamine, 4µg/ml Heparin, 4 mg/ml BSA, 60 mmol/l putrescine, 20 nmol/l progesterone, 30 nmol/l sodium selenite, 25 µg/ml, insulin, 100 µg/ml apo-transferrin and human recombinant EGF and FGF-2 (Sigma), both at a final concentration of 20 µg/ml (sphere-medium).

CFS formation assay

The CFS capacity of colon cancer cell lines tested in this study was derived from monolayer culture or floating culture (for Colo320 colon cancer cell line). It was assessed by plating $2x10^5$ cells in a T25 flask (8000 cells/cm^2). In $CFS^{Positive}$ cell lines, CFS were observed 3-7 days after plating. To obtain pure CSC-like cells the culture period in sphere-medium was extended to 10 passages. To evaluate the CFS capacity of sorted $CD166^+CD44^+EpCAM^{high}$ or $CD166^-CD44^-EpCAM^{low}$ HCT116 cell subpopulations (**Fig. 4C**), 1000 cells/well were plated in 96-well culture dishes in 200µl of sphere-medium. The number of CFS in each well was evaluated after 5 days.

Proliferation and chemosensitivity assay

Rates of proliferation and sensitivity to 5-FU were assessed using the cell proliferation reagent WST-1 (Roche). Briefly, 10^4 cells of each cell line or from CFS cultures were plated per well in 24-well plates in 2 ml of media with or without 5-FU. After 5 days, WST-1 reagent was added at a 1:10 final dilution and incubated for 4h at 37°C. The relative survival fraction of cells was compared between treated and untreated cells.

CFS assay following 5-FU treatment

Inoculation of $2x10^6$ cells into a T75 flask was made with different concentrations of 5-FU to obtain 10% cell survival after 5 days of incubation. Cells were then washed detached and 10^5 cells were inoculated into a T25 flask with sphere-medium (4000 cells/cm^2). After 3 days observation, photographs were taken to determine the proportion of CFS amongst the 5-FU resistant cells. For experimental controls, untreated cells were plated at the same concentration in sphere-medium.

Chemical screening

The multi-step strategy used to screen the Institut Curie/CNRS chemical library is shown in Fig. 3A. This bank contains 8,560 compounds in a 96-well format at 10 mg/ml in DMSO (*i.e.* mean concentration 10 mM). Screening was performed at 10 and 1 µM final concentrations in 96-well plates and in a final volume of 200 µl of medium. CRC cell lines were incubated with the chemical bank at 700 cells/well for 5 days in standard culture conditions. The Wst-1 assay was used to indirectly estimate cell survival according to the indicated procedure (ROCHE). Confirmation of the 15 compounds (validation step) was obtained by starting with drug powders in order to reach the correct initial concentration and then performing a cell survival test in 24-well plates with 2ml of medium. CRC cell lines and CFS were incubated with the chemical bank at 700 and 1500 cells/well respectively for 5 days in standard culture conditions.

Subcutaneous transplantation of colon cancer cell lines

Colon cancer cell lines and CFS were suspended in 200 µl PBS-Matrigel (1:1) mixture. They were injected subcutaneously in the flank of 5 week old nude mice (nu/nu; 1 injection/flank). Experiments were performed in triplicates (3 mice/each sample). Tumor formation was evaluated using a caliper starting on the third week after injection and then weekly for 4 weeks. Animals were sacrificed when the tumor size was between 15 and 20 mm in diameter, or 7 weeks after the injection.

Microsatellite analysis

Non-coding microsatellite repeat markers were used to detect instability in five microsatellites (NR27, NR21, NR24, Bat25, Bat26 comprising the pentaplex PCR system) in CRC cell lines and their CFS counterparts, as previously described [39]. The *TGFBR2*, *BAX*, *MSH3* and *MSH6* genes containing coding repeats were amplified as described before [40]. Four other genes containing coding mononucleotide repeat were also amplified with specific primers (sequences available on request). Amplified PCR products were run on an Applied Biosystems PRISM 3100 Genetic Analyzer automated capillary electrophoresis DNA sequencer. Allelic sizes were estimated using gene mapper software (Applied Biosystems).

RNA and DNA extraction

Total RNA from CFS and CRC cell lines was extracted using Trizol (Invitrogen) and DNA was extracted using a standard phenol-chloroform procedure. Both RNA and DNA were assessed for integrity and quantity following stringent quality control criteria (CIT program protocols http://cit.ligue-cancer.net).

Flow cytometry

Flow cytometry was performed on adherent cell lines or CFS cultures after dissociation with accutase (PAA). Cells were washed once in PBS supplemented with 1% BSA (Sigma) and resuspended in PBS/ 1% BSA at a concentration of 10^6 cells/100 µl. Cells were stained with IgG-PE/PECy5/FITC/APC (BD Biosciences) to detect non-specific binding of antibodies and autofluorescence. Primary antibodies used were: CD44-FITC 1:75 (clone G44-26 BD Biosciences), CD133-PE 1:100 (clone AC133 Miltenyi Biotec), EpCAM-APC 1:200 (clone HEA-125 Miltenyi Biotec), CD24-FITC 1:100 (clone ML5 BD Biosciences) and CD29-PECy5 1:100 (clone MAR4 BD Biosciences). Cells were incubated for 30 min at 4°C in the dark and then

washed in buffer (PBS, 1% BSA, 1mM EDTA). Expression of cell surface markers was detected with a FACScan flow cytometer (BD Biosciences). Cell line suspensions were sorted according to their CD166 and/or CD44 and/or EpCAM expression with a FACS Coulter (BD Biosciences). Separated subpopulations were reanalyzed for purity.

Genomic and Gene Expression Arrays and analysis

Data preparation

Gene expression analysis using arrays was carried out on the IGBMC microarray platform (Strasbourg, France). Total RNA was amplified, labeled and hybridized to Affymetrix Human Genome U133 plus2 GeneChips following the manufacturer's protocol (Affymetrix, Santa Clara, CA). The chips were scanned with the Affymetrix GeneChip Scanner 3000 and raw intensities were quantified from subsequent images using GCOS 1.4 software (Affymetrix). Data were normalized using the Robust Multi-array Average method and implemented in the R package affy [41].

Genomic arrays were performed on the Integragen Platform (Evry, France). DNAs from 13 CFS/adherent cell lines (*i.e.* LIM2405, HCT8, HCT116, HCT15, TC-7, CO115, RKO, LS411, V9P, HT29, SW620, Colo320, FET) were hybridized on IlluminaSNP HumanCNV610 chips according to instructions provided by the array manufacturer (Illumina, San Diego, CA). Data were normalized and processed as described in supplemental methods [42]**.** Data are available in ArrayExpress database (www.ebi.ac.uk/arrayexpress).
All analyses were performed using R software (http://www.R-project.org).

Unsupervised analysis of gene expression data

To evaluate the distance between CRC cell lines and CFS, PCA and consensus hierarchical clustering analysis of the 13 pairs (26 samples) were performed on probe-sets present (log2 intensity > log2(3.5)) in at least 5% of the samples and having a robust coefficient of variation significantly different from the median variance of all probe-sets (rCV>.05 and *P*-value variance test <.01).

CFS signature

Genes differentially expressed between CRC cell lines and CFS were assessed using Limma paired moderated *t*-test [43]. Genes having a *P* value <.001 define the CFS signature (*n* =359 probe sets).

CFS functional analysis

All KEGG pathways and gene sets functionally related to stemness from KEGG, Biocarta, GeneOntology, Molecular Signatures database and Stanford Microarray database were tested for enrichment of up- and down-regulated genes in CFS. Enrichment of the top 100 to 500 up/down deregulated probe sets was evaluated by computing a hypergeometric test. The median *P* value across up/down top probe set lists was used to select pathways and gene sets of interest (*P* value <.05).

Analysis of deregulated regions

To define up/down regulation regions separately for each CFS/CRC cell line pair, the genome was segmented into overlapping windows of 5 Mb. In each window, the enrichment of up- or down-regulated genes (log2(FC)>0.5) for the given pair was assessed by a Fisher test between up/down genes in the windows and up/down regulated genes in the rest of the genome. To compute the frequency across pairs, each region was assigned -1(down-regulated)/0 (not modified)/1 (up-regulated) depending on the significance (*P*-value < .05) of the enrichment.

Survival analysis

Two publicly available Affymetrix U133P2 datasets with Recurrence Free Survival annotations were used: [25] dataset GSE17536 and GSE17537 comprising 148 samples of Stage II/III CRC), and [26] GSE14333 comprising 99 samples of Stage B and C CRC not contained in the previous dataset. These were normalized by RMA and by clinical center. To evaluate the survival impact of the CFS signature in those datasets, a subset of genes from the original signature (see **Supplementary Table S1**) was selected based upon their high fold-change (|log2(CFS/cell line)|>0.8; *n* = 55). To define a prognostic CFS signature, the 55-gene signature was reduced to probe sets significantly associated with prognosis in the first dataset using univariate Cox models (log rank *P*-value <.05; n =8). For both CFS signatures an average expression score per sample was then defined. Normalized intensity values of selected genes were each centered to zero by subtracting the median expression of each gene. Genes that were down regulated in CFS were multiplied by (-1), thus up and down regulated genes can both be used in the score. All genes from the given signature were then averaged. A higher score corresponded to a higher deregulation of CSF genes. The score was then divided into high and low score groups by taking tails of the score distribution in the considered dataset, *i.e.* 30% of the highest and 30% of the lowest scores.

Survival curves were calculated according to the Kaplan-Meier method with an end-point at 5 years. Differences between curves were assessed using the log-rank test. Univariate and multivariate associations for outcome were performed using the Cox regression model.

GRANT SUPPORT

This work was supported by the 'Carte d'Identité des Tumeurs' (CIT) program (http://cit.ligue-cancer.net) from the Ligue Nationale Contre le Cancer and by grants from the 'Institut National du Cancer' (INCa) (To AD). AD group has the label de « La Ligue Contre le Cancer ». AC is a recipient of an INCa fellowship (Institut National du Cancer). AL is a recipient of a MESR fellowship (Ministère de l'Enseignement Supérieur et de le Recherche.

CONFLICTS OF INTEREST

No potential conflicts of interests were disclosed.

REFERENCES

1. Todaro M, Alea MP, Di Stefano AB, Cammareri P, Vermeulen L, Iovino F, Tripodo C, Russo A, Gulotta G, Medema JP, Stassi G (2007) Colon cancer stem cells dictate tumor growth and resist cell death by production of interleukin-4. Cell Stem Cell 1 (4):389-402. doi:S1934-5909(07)00118-X [pii]
10.1016/j.stem.2007.08.001
2. O'Brien CA, Pollett A, Gallinger S, Dick JE (2007) A human colon cancer cell capable of initiating tumour growth in immunodeficient mice. Nature 445 (7123):106-110. doi:nature05372 [pii]
10.1038/nature05372
3. Ricci-Vitiani L, Lombardi DG, Pilozzi E, Biffoni M, Todaro M, Peschle C, De Maria R (2007) Identification and expansion of human colon-cancer-initiating cells. Nature 445 (7123):111-115. doi:nature05384 [pii]
10.1038/nature05384
4. Reya T, Morrison SJ, Clarke MF, Weissman IL (2001) Stem cells, cancer, and cancer stem cells. Nature 414 (6859):105-111. doi:10.1038/35102167
35102167 [pii]
5. Dean M, Fojo T, Bates S (2005) Tumour stem cells and drug resistance. Nat Rev Cancer 5 (4):275-284. doi:nrc1590 [pii]
10.1038/nrc1590
6. Eppert K, Takenaka K, Lechman ER, Waldron L, Nilsson B, van Galen P, Metzeler KH, Poeppl A, Ling V, Beyene J, Canty AJ, Danska JS, Bohlander SK, Buske C, Minden MD, Golub TR, Jurisica I, Ebert BL, Dick JE (2011) Stem cell gene expression programs influence clinical outcome in human leukemia. Nat Med 17 (9):1086-1093. doi:nm.2415 [pii]
10.1038/nm.2415
7. Al-Hajj M, Becker MW, Wicha M, Weissman I, Clarke MF (2004) Therapeutic implications of cancer stem cells. Curr Opin Genet Dev 14 (1):43-47. doi:10.1016/j.gde.2003.11.007
S0959437X03001734 [pii]
8. Medema JP, Vermeulen L (2011) Microenvironmental regulation of stem cells in intestinal homeostasis and cancer. Nature 474 (7351):318-326. doi:nature10212 [pii]
10.1038/nature10212
9. Liu Y, Bodmer WF (2006) Analysis of P53 mutations and their expression in 56 colorectal cancer cell lines. Proc Natl Acad Sci U S A 103 (4):976-981. doi:0510146103 [pii]
10.1073/pnas.0510146103
10. Volchenboum SL, Li C, Li S, Attiyeh EF, Reynolds CP, Maris JM, Look AT, George RE (2009) Comparison of primary neuroblastoma tumors and derivative early-passage cell lines using genome-wide single nucleotide polymorphism array analysis. Cancer Res 69 (10):4143-4149. doi:0008-5472.CAN-08-3112 [pii]
10.1158/0008-5472.CAN-08-3112
11. Douglas EJ, Fiegler H, Rowan A, Halford S, Bicknell DC, Bodmer W, Tomlinson IP, Carter NP (2004) Array comparative genomic hybridization analysis of colorectal cancer cell lines and primary carcinomas. Cancer Res 64 (14):4817-4825. doi:10.1158/0008-5472.CAN-04-0328
64/14/4817 [pii]
12. Willson JK, Bittner GN, Oberley TD, Meisner LF, Weese JL (1987) Cell culture of human colon adenomas and carcinomas. Cancer Res 47 (10):2704-2713
13. Kondo T (2007) Stem cell-like cancer cells in cancer cell lines. Cancer Biomark 3 (4-5):245-250
14. Yeung TM, Gandhi SC, Wilding JL, Muschel R, Bodmer WF (2010) Cancer stem cells from colorectal cancer-derived cell lines. Proc Natl Acad Sci U S A 107 (8):3722-3727. doi:0915135107 [pii]
10.1073/pnas.0915135107

15. Dittfeld C, Dietrich A, Peickert S, Hering S, Baumann M, Grade M, Ried T, Kunz-Schughart LA (2009) CD133 expression is not selective for tumor-initiating or radioresistant cell populations in the CRC cell lines HCT-116. Radiother Oncol 92 (3):353-361. doi:S0167-8140(09)00338-7 [pii]
10.1016/j.radonc.2009.06.034
16. Kai K, Nagano O, Sugihara E, Arima Y, Sampetrean O, Ishimoto T, Nakanishi M, Ueno NT, Iwase H, Saya H (2009) Maintenance of HCT116 colon cancer cell line conforms to a stochastic model but not a cancer stem cell model. Cancer Sci 100 (12):2275-2282. doi:CAS1318 [pii]
10.1111/j.1349-7006.2009.01318.x
17. Reynolds BA, Weiss S (1992) Generation of neurons and astrocytes from isolated cells of the adult mammalian central nervous system. Science 255 (5052):1707-1710
18. Lobo NA, Shimono Y, Qian D, Clarke MF (2007) The biology of cancer stem cells. Annu Rev Cell Dev Biol 23:675-699. doi:10.1146/annurev.cellbio.22.010305.104154
19. Fan X, Ouyang N, Teng H, Yao H (2011) Isolation and characterization of spheroid cells from the HT29 colon cancer cell line. Int J Colorectal Dis 26 (10):1279-1285. doi:10.1007/s00384-011-1248-y
20. Vermeulen L, Todaro M, de Sousa Mello F, Sprick MR, Kemper K, Perez Alea M, Richel DJ, Stassi G, Medema JP (2008) Single-cell cloning of colon cancer stem cells reveals a multi-lineage differentiation capacity. Proc Natl Acad Sci U S A 105 (36):13427-13432. doi:0805706105 [pii]
10.1073/pnas.0805706105
21. Jung P, Sato T, Merlos-Suarez A, Barriga FM, Iglesias M, Rossell D, Auer H, Gallardo M, Blasco MA, Sancho E, Clevers H, Batlle E (2011) Isolation and in vitro expansion of human colonic stem cells. Nat Med 17 (10):1225-1227. doi:nm.2470 [pii]
10.1038/nm.2470
22. Lengauer C, Kinzler KW, Vogelstein B (1997) Genetic instability in colorectal cancers. Nature 386 (6625):623-627. doi:10.1038/386623a0
23. Duval A, Hamelin R (2002) Mutations at coding repeat sequences in mismatch repair-deficient human cancers: toward a new concept of target genes for instability. Cancer Res 62 (9):2447-2454
24. Pino MS, Chung DC (2010) The chromosomal instability pathway in colon cancer. Gastroenterology 138 (6):2059-2072. doi:S0016-5085(10)00170-8 [pii]
10.1053/j.gastro.2009.12.065
25. Smith JJ, Deane NG, Wu F, Merchant NB, Zhang B, Jiang A, Lu P, Johnson JC, Schmidt C, Bailey CE, Eschrich S, Kis C, Levy S, Washington MK, Heslin MJ, Coffey RJ, Yeatman TJ, Shyr Y, Beauchamp RD (2010) Experimentally derived metastasis gene expression profile predicts recurrence and death in patients with colon cancer. Gastroenterology 138 (3):958-968. doi:S0016-5085(09)01964-7 [pii]
10.1053/j.gastro.2009.11.005
26. Jorissen RN, Gibbs P, Christie M, Prakash S, Lipton L, Desai J, Kerr D, Aaltonen LA, Arango D, Kruhoffer M, Orntoft TF, Andersen CL, Gruidl M, Kamath VP, Eschrich S, Yeatman TJ, Sieber OM (2009) Metastasis-Associated Gene Expression Changes Predict Poor Outcomes in Patients with Dukes Stage B and C Colorectal Cancer. Clin Cancer Res 15 (24):7642-7651. doi:1078-0432.CCR-09-1431 [pii]
10.1158/1078-0432.CCR-09-1431
27. Spinola M, Leoni V, Pignatiello C, Conti B, Ravagnani F, Pastorino U, Dragani TA (2005) Functional FGFR4 Gly388Arg polymorphism predicts prognosis in lung adenocarcinoma patients. J Clin Oncol 23 (29):7307-7311. doi:JCO.2005.17.350 [pii]
10.1200/JCO.2005.17.350
28. Kang HC, Kim IJ, Park JH, Shin Y, Ku JL, Jung MS, Yoo BC, Kim HK, Park JG (2004) Identification of genes with differential expression in acquired drug-resistant gastric cancer cells using high-density oligonucleotide microarrays. Clin Cancer Res 10 (1 Pt 1):272-284
29. Hayakawa I, Shioya R, Agatsuma T, Furukawa H, Sugano Y (2004) Thienopyridine and benzofuran derivatives as potent anti-tumor agents possessing different structure-activity relationships. Bioorg Med Chem Lett 14 (13):3411-3414. doi:10.1016/j.bmcl.2004.04.079
S0960894X04005943 [pii]

30. Baraldi PG, Romagnoli R, Giovanna Pavani M, del Carmen Nunez M, Bingham JP, Hartley JA (2002) Benzoyl and cinnamoyl nitrogen mustard derivatives of benzoheterocyclic analogues of the tallimustine: synthesis and antitumour activity. Bioorg Med Chem 10 (5):1611-1618. doi:S0968089601004254 [pii]
31. Gottesman MM (2002) Mechanisms of cancer drug resistance. Annu Rev Med 53:615-627. doi:10.1146/annurev.med.53.082901.103929
53/1/615 [pii]
32. Zhou S, Schuetz JD, Bunting KD, Colapietro AM, Sampath J, Morris JJ, Lagutina I, Grosveld GC, Osawa M, Nakauchi H, Sorrentino BP (2001) The ABC transporter Bcrp1/ABCG2 is expressed in a wide variety of stem cells and is a molecular determinant of the side-population phenotype. Nat Med 7 (9):1028-1034. doi:10.1038/nm0901-1028
nm0901-1028 [pii]
33. Ito K, Hirao A, Arai F, Matsuoka S, Takubo K, Hamaguchi I, Nomiyama K, Hosokawa K, Sakurada K, Nakagata N, Ikeda Y, Mak TW, Suda T (2004) Regulation of oxidative stress by ATM is required for self-renewal of haematopoietic stem cells. Nature 431 (7011):997-1002. doi:nature02989 [pii]
10.1038/nature02989
34. Bao S, Wu Q, McLendon RE, Hao Y, Shi Q, Hjelmeland AB, Dewhirst MW, Bigner DD, Rich JN (2006) Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. Nature 444 (7120):756-760. doi:nature05236 [pii]
10.1038/nature05236
35. Diehn M, Cho RW, Lobo NA, Kalisky T, Dorie MJ, Kulp AN, Qian D, Lam JS, Ailles LE, Wong M, Joshua B, Kaplan MJ, Wapnir I, Dirbas FM, Somlo G, Garberoglio C, Paz B, Shen J, Lau SK, Quake SR, Brown JM, Weissman IL, Clarke MF (2009) Association of reactive oxygen species levels and radioresistance in cancer stem cells. Nature 458 (7239):780-783. doi:nature07733 [pii]
10.1038/nature07733
36. Merlos-Suarez A, Barriga FM, Jung P, Iglesias M, Cespedes MV, Rossell D, Sevillano M, Hernando-Momblona X, da Silva-Diz V, Munoz P, Clevers H, Sancho E, Mangues R, Batlle E (2011) The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. Cell Stem Cell 8 (5):511-524. doi:S1934-5909(11)00110-X [pii]
10.1016/j.stem.2011.02.020
37. Sinicrope FA, Foster NR, Thibodeau SN, Marsoni S, Monges G, Labianca R, Kim GP, Yothers G, Allegra C, Moore MJ, Gallinger S, Sargent DJ (2011) DNA mismatch repair status and colon cancer recurrence and survival in clinical trials of 5-fluorouracil-based adjuvant therapy. J Natl Cancer Inst 103 (11):863-875. doi:djr153 [pii]
10.1093/jnci/djr153
38. Poupon MF, Arvelo F, Goguel AF, Bourgeois Y, Jacrot M, Hanania N, Arriagada R, Le Chevalier T (1993) Response of small-cell lung cancer xenografts to chemotherapy: multidrug resistance and direct clinical correlates. J Natl Cancer Inst 85 (24):2023-2029
39. Buhard O, Suraweera N, Lectard A, Duval A, Hamelin R (2004) Quasimonomorphic mononucleotide repeats for high-level microsatellite instability analysis. Dis Markers 20 (4-5):251-257
40. Gayet J, Zhou XP, Duval A, Rolland S, Hoang JM, Cottu P, Hamelin R (2001) Extensive characterization of genetic alterations in a series of human colorectal cancer cell lines. Oncogene 20 (36):5025-5032. doi:10.1038/sj.onc.1204611
41. Irizarry RA, Bolstad BM, Collin F, Cope LM, Hobbs B, Speed TP (2003) Summaries of Affymetrix GeneChip probe level data. Nucleic Acids Res 31 (4):e15
42. Letouze E, Sow A, Petel F, Rosati R, Figueiredo BC, Burnichon N, Gimenez-Roqueplo AP, Lalli E, de Reynies A (2012) Identity by descent mapping of founder mutations in cancer using high-resolution tumor SNP data. PLoS One 7 (5):e35897. doi:10.1371/journal.pone.0035897
PONE-D-12-05227 [pii]
43. Smyth GK (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol 3:Article3. doi:10.2202/1544-6115.1027

FIGURE LEGENDS

Figure 1. (A) Results of the CFS assay in 25 CRC cell lines (MSS=11; MSI=14). (B) Morphological features of CFS from 9 CRC cell lines (MSI in black, MSS in red) grown in sphere-medium. The morphology of CFS derived from T71 is peculiar and these grow in suspension as single cells.

Figure 2. (A) Extensive microarray analysis revealed that chromosomal aberrations due to CIN were similar between 13 parental CRC cell lines (LIM2405, HCT8, HCT116, HCT15, TC-7, CO115, RKO, LS411, V9P, HT29, SW620, Colo320, FET) and their corresponding CFS. Examples of chromosomal instability in HCT116 (MSI cell line), HT29 (MSS cell line) and their corresponding CFS. (B) Microsatellite instability in HCT116 parental cells and their corresponding CFS. Mutation analysis of DNA microsatellites showed identical patterns of alteration in both non-coding (BAT26, BAT25, NR21, NR25, NR27) and coding (*TGFBR2, RAD50, MSH6, MSH3, MBD4, BAX, ATR, BLM*) repeats in the parental CRC cell lines and their corresponding CFS. This was observed in HCT116 and in 10 $CFS^{Positive}$ MSI CRC cell lines (HCT116, CO115, HCT15, HCT8, ISHI, LIM2405, LS411, RKO, TC-7, TC71 not shown).

Figure 3. (A) Analysis of tumor growth following injection of a high number (*n* = $3x10^5$) of cells from CRC cell lines (5 $CFS^{positive}$ and 3 $CFS^{negative}$, left top panel) or low number (n = 200) of cells (2 $CFS^{positive}$ and 2 $CFS^{negative}$, right top panel). (B) Comparison of tumor growth following injection of CRC cell line HCT116 and the corresponding CFS at 200 cells (* p=0,034 at 7 days and p=0,012 at 21 days), 500 cells (** p=0,006 at 21 days) and 1000 cells.

Figure 4. (A) Expression of putative colorectal CSC markers (CD44, CD133, CD166, CD24, CD29, EPCAM, ALDH, OLFM4, LGR5) were investigated using gene expression arrays in 13 CFS/parental cell lines (8 MSI: LIM2405, HCT8, HCT116, HCT15, TC-7 CO115, RKO, LS411; 5 MSS: V9P, HT29, SW620, Colo320, FET). (B) Expression of CD166 and CD44 markers was investigated by flow cytometry in different subpopulations of sorted cells from the HCT116 and LS174T parental cells. CD166-/CD44-, CD166+/CD44-, CD166+/CD44+ and CD166-/CD44+ (HCT116 only) cells were sorted by FACS and re-inoculated in medium. After a few days of growth, cells from sorted subpopulations were analyzed again for marker expression. The CD expression pattern was compared to the cell sorted profile. (C) Expression of 6 markers in CRC cell lines (*n =25*), primary CRCs (*n* =15) and/or tumor xenografts established directly from the primary CRC (*n* =12). (D) CFS ability of HCT116 after sorting of two subpopulations of $CD166^{+}CD44^{+}EpCAM^{high}$ and $CD166^{-}CD44^{-}EpCAM^{low}$ cells.

Figure 5. (A) Percent of CFS in sphere-medium from LoVo, HCT116, LIM1215 and LS174T cells before and after treatment with 5-Fluorouracil for 5 days at $IC_{10\%}$ (7,5 µM, 7,5 µM, 5 µM and 20 µM respectively). (B) Representative image of spheres (arrows) from HCT116 cells treated with 5-fluorouracil for 5 days at $IC_{10\%}$ (top) and untreated (bottom) after 3 days in sphere-medium (size bar = 100 µm). (C) Cell survival curves for 6 CRC cell lines and their corresponding CFS following 5 days of exposure to increasing concentrations of 5-Fluorouracil. Dotted lines indicate the IC_{50} value in each case.

Figure 6. Heatmaps of the 55-gene (A) and 8-gene signatures (B). Visualization of the expression level for each gene in each sample relative to the median gene intensity (C) across all samples. Green corresponds to down-regulation and red to up-regulation. CL, parental CRC cell line.

Figure 7. Kaplan-Meier analysis of disease-free survival for CRC patients classified according to expression levels for the 55-gene (left panel) or 8-gene (right panel) CFS signature.

Supplementary Figure S1. (A) Clustering of MSI and MSS CRC samples (CFS and their corresponding cell lines) according to gene expression was evaluated by Principal Component Analysis (PCA). With both subtypes, CFS grouped closely together with their parental cell lines. (B) Gene deregulation across chromosomes in CFS compared to parental CRC cell lines. Regions down-regulated in CFS are in green, while regions up-regulated in CFS are in red (upper panel). Specific chromosomal regions were enriched with genes from the CFS signature. These regions displayed similar genomic status (*i.e.* the same DNA copy-number) in CFS and the parental CRC cell line, suggesting that gene deregulation occurred mainly *via* processes that were independent of DNA copy number. Regions frequently down or up-regulated in CFS (> 30%) are listed (lower panel).

Supplementary Figure S2. Survival impact of the 55-gene (A) or 8-gene (B) CFS signature in CRC patients. A and B top-panel shows analysis performed in the combined patient series (Smith+Jorissen). A and B bottom panel shows analysis performed in a single patient series (left: Smith; right: Jorissen). A score per sample of up/down regulation of the CFS signature was computed as the mean of absolute normalized intensity values of selected genes centered to zero per gene. The score was then classified into high and low score groups, corresponding to the highest 30% and lowest 30% scores in the considered dataset.

Supplementary Figure S3. (A) Flow chart for the multi-step strategy applied to screen the Institut Curie/CNRS chemical library and identify the resulting 15 compounds. (B) Example of survival of HCT116 and FET CRC cell lines and their corresponding CFS following treatment with 3 different selected compounds from the library. (C) CFS abilities of HCT116 and FET before (untreated) and after treatment with 5-Fluorouracil or selected compounds from the library for 5 days at $IC_{10\%}$.

Table 1. Specific signaling and metabolic pathways significantly associated with the CFS gene signature. Abbreviations: Bioc = Biocarta, GO = GeneOntology, MSigDB = Molecular Signatures Database, SMD = Stanford Microarray Database, KEGG = Kyoto Encyclopedia of Genes and Genomes. C2 = MSigDB curated gene sets, C5 = MSigDB GO gene sets. *P*-value up (down) = hypergeometric test *P*-value using top up (down) regulated genes in CFS *vs* Cell line.

Table 2. Association of the 8-gene CFS High/Low score and TNM stage and prognosis in the combined dataset (Smith + Jorissen). H.R.: Cox Hazard Ratio, 95% C.I.: 95 Percent Confidence Interval of HR. Value: modality of the annotation associated to H.R.

Supplementary Table S1. List of 264 annotated genes displaying differential expression between CFSs and CRC cell lines (P-value < 0.001). P value = Welch *t* test. Genes included in the 55 (1 + 2) or 8 (2) CFS gene signatures are indicated.

Supplementary Table S2. List of colon cancer cell lines, xenograft tumors and primary tumors used in this work and the results of expression of the 6 putative CSC markers (CD44, CD133, CD166, CD24, CD29 and EpCAM) by flow cytometry in these models. All analyses were performed at least twice. ND= not determined; 0% = not detected (CD marker not expressed in tumor cells).

Supplementary Table S3. The 15 compounds identified from the Institut Curie /CNRS chemical library as putative drugs targeting CFS are listed. The indicated IC50 is for CFS in HCT116 (MSI; in red) and/or FET (MSS; in blue).

ABBREVIATIONS

CRC Colorectal cancer

MSI Microsatellite instability

MSS Microsatellite stability

CFS Cell-forming-spheres

CSC Cancer stem cells

MMR Mismatch-repair

CIN Chromosomal Instability

DFS Disease-free survival

5-FU 5-Fluorouracil

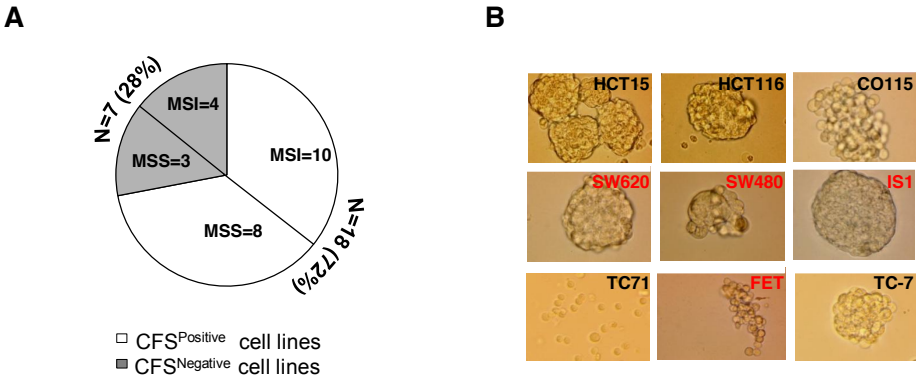

A
B
MSI=4
MSI=10
MSS=3
MSS=8
N=7 (28%)
N=18 (72%)
CFSPositive cell lines
CFSNegative cell lines
HCT15
HCT116
CO115
SW620
SW480
IS1
TC71
FET
TC-7

Figure 1

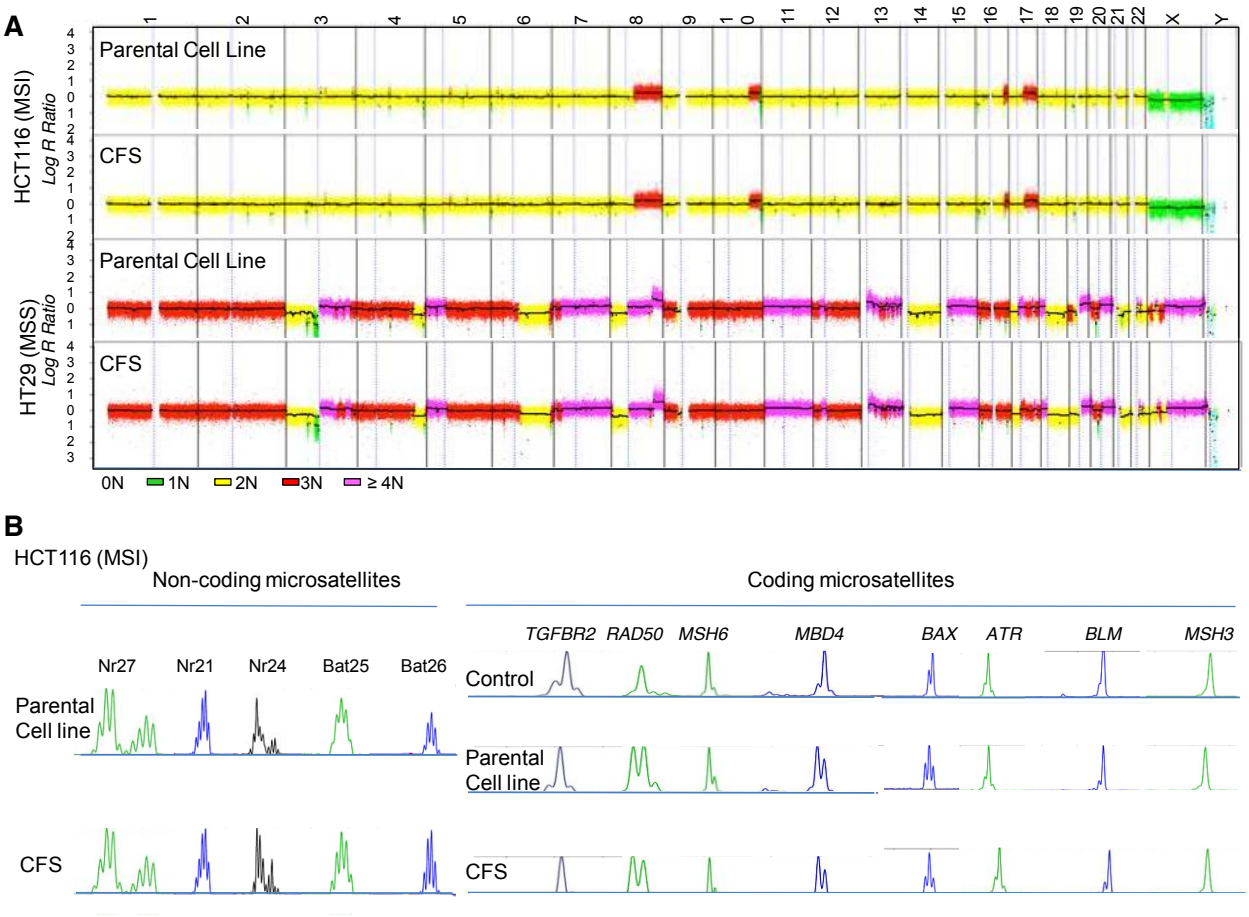

A
Parental Cell Line
CFS
HCT116 (MSI)
Log R Ratio
Parental Cell Line
CFS
HT29 (MSS)
Log R Ratio
0N
1N
2N
3N
≥ 4N
B
HCT116 (MSI)
Non-coding microsatellites
Coding microsatellites
Nr27
Nr21
Nr24
Bat25
Bat26
TGFBR2
RAD50
MSH6
MBD4
BAX
ATR
BLM
MSH3
Parental Cell line
Control
Parental Cell line
CFS
CFS

Figure 2

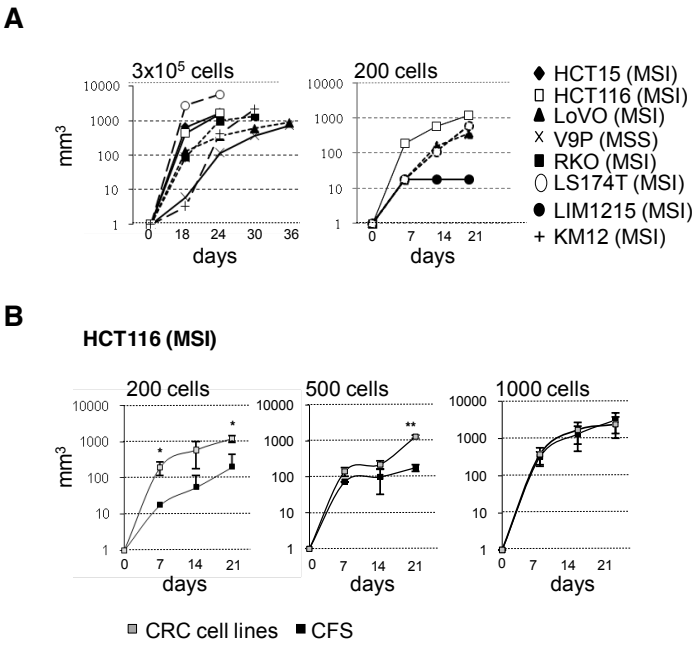

A
3x10^5 cells
200 cells
mm3
days
HCT15 (MSI)
HCT116 (MSI)
LoVO (MSI)
V9P (MSS)
RKO (MSI)
LS174T (MSI)
LIM1215 (MSI)
KM12 (MSI)
B
HCT116 (MSI)
200 cells
500 cells
1000 cells
CRC cell lines
CFS

Figure 3

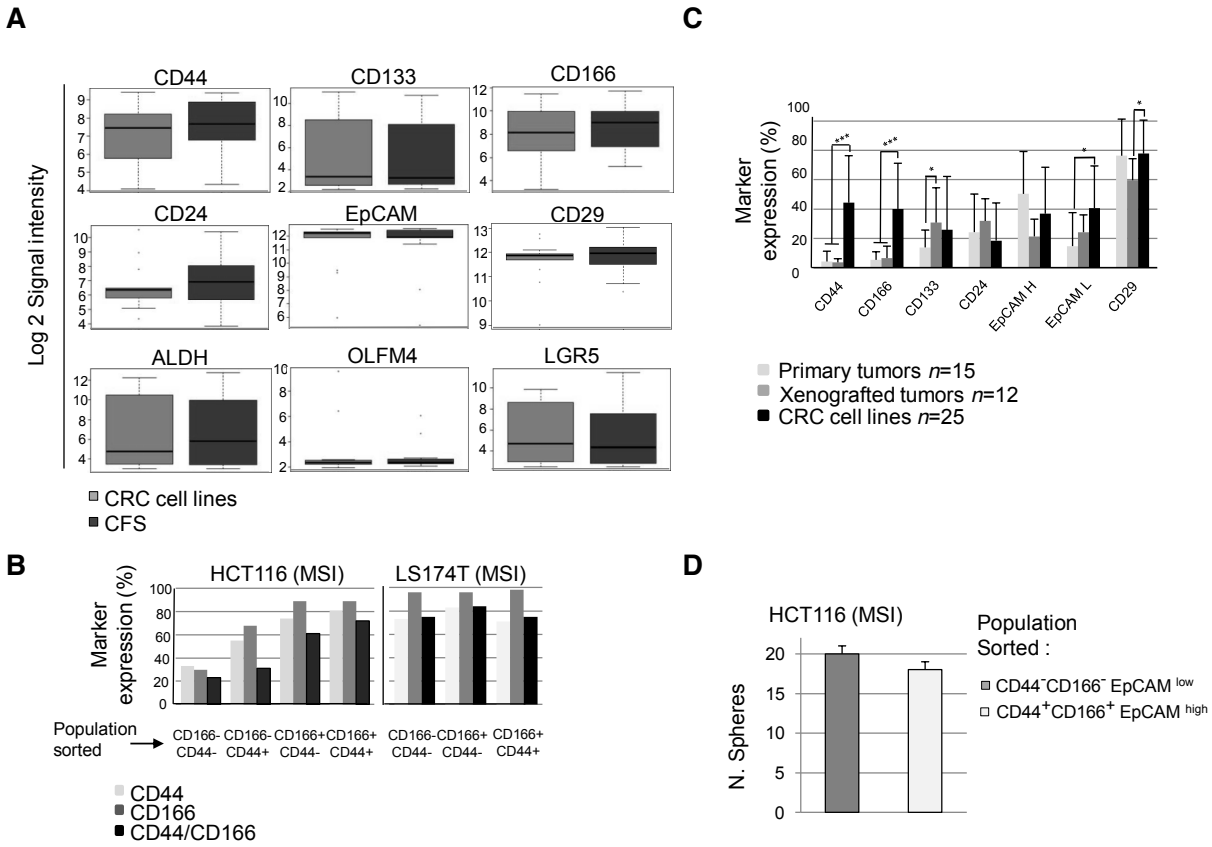

A
CD44
CD133
CD166
CD24
EpCAM
CD29
ALDH
OLFM4
LGR5
Log 2 Signal intensity
CRC cell lines
CFS
B
HCT116 (MSI)
LS174T (MSI)
Marker expression (%)
Population sorted
CD166- CD44-
CD166- CD44+
CD166+ CD44-
CD166+ CD44+
CD166- CD44-
CD166+ CD44-
CD166+ CD44+
CD44
CD166
CD44/CD166
C
Marker expression (%)
CD44
CD166
CD133
CD24
EpCAM H
EpCAM L
CD29
Primary tumors n=15
Xenografted tumors n=12
CRC cell lines n=25
D
HCT116 (MSI)
N. Spheres
Population Sorted :
CD44-CD166- EpCAM low
CD44+CD166+ EpCAM high

Figure 4

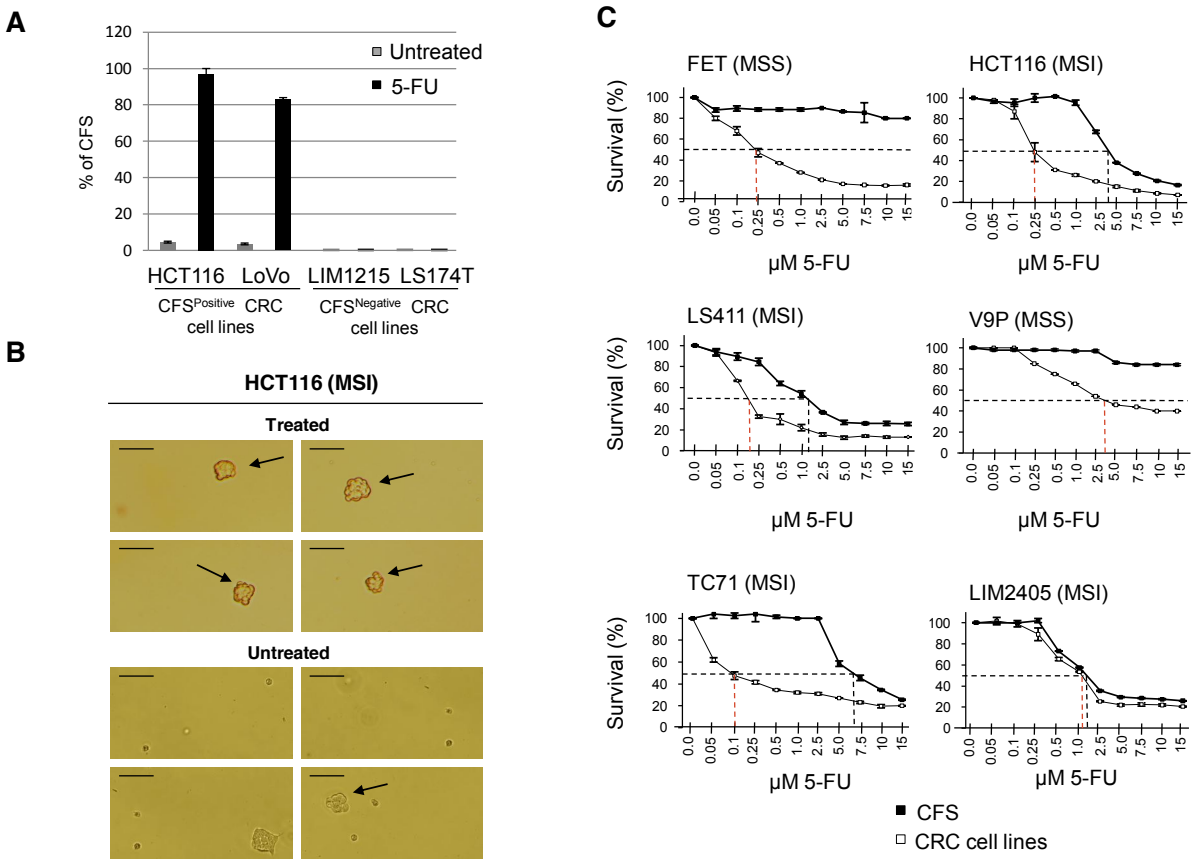

A
% of CFS
Untreated
5-FU
HCT116
LoVo
LIM1215
LS174T
CFSPositive CRC cell lines
CFSNegative CRC cell lines
B
HCT116 (MSI)
Treated
Untreated
C
FET (MSS)
HCT116 (MSI)
LS411 (MSI)
V9P (MSS)
TC71 (MSI)
LIM2405 (MSI)
Survival (%)
µM 5-FU
CFS
CRC cell lines

Figure 5

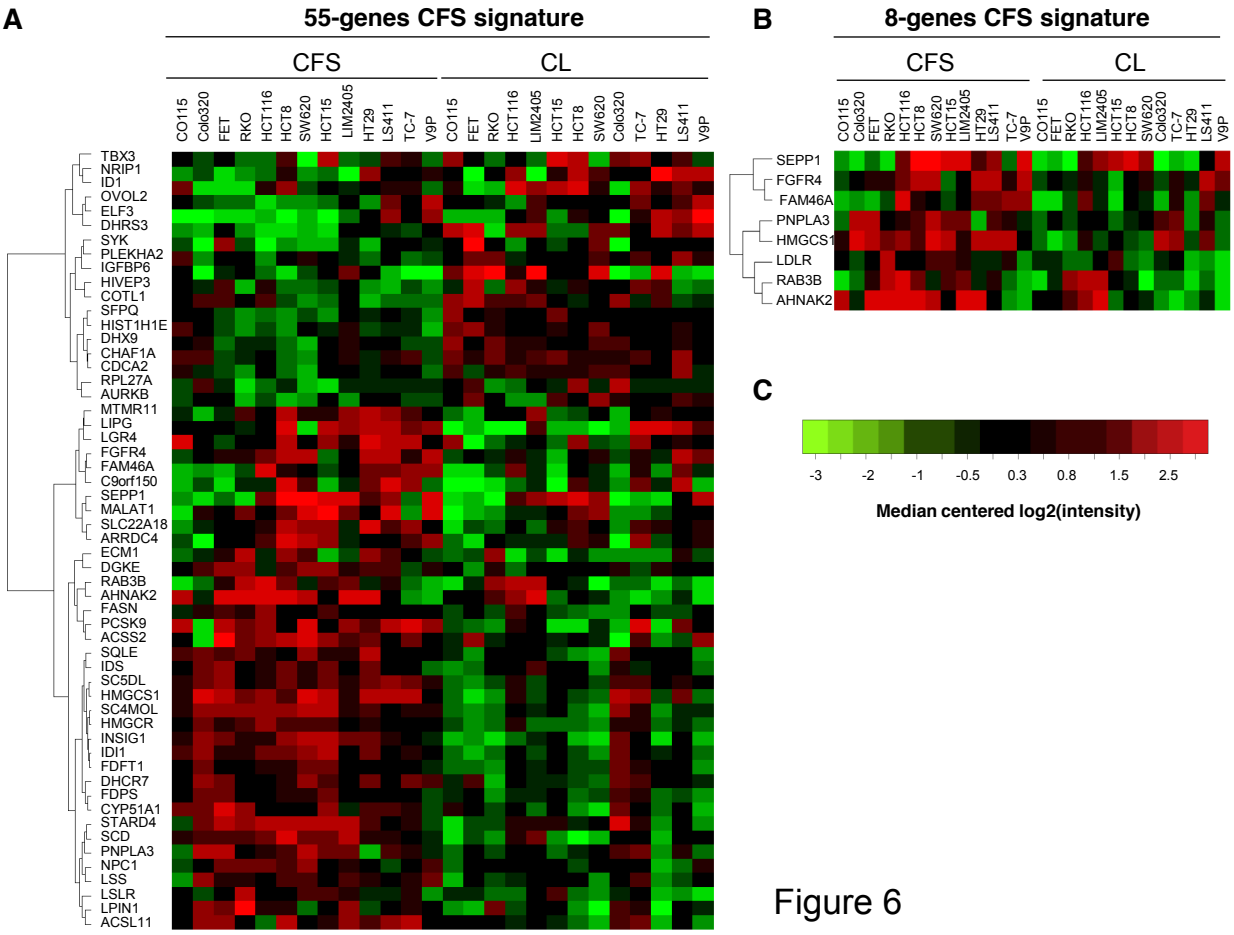

A
55-genes CFS signature
CFS
CL
CO115
Colo320
FET
RKO
HCT116
HCT8
SW620
HCT15
LIM2405
HT29
LS411
TC-7
V9P
TBX3
NRIP1
ID1
OVOL2
ELF3
DHRS3
SYK
PLEKHA2
IGFBP6
HIVEP3
COTL1
SFPQ
HIST1H1E
DHX9
CHAF1A
CDCA2
RPL27A
AURKB
MTMR11
LIPG
LGR4
FGFR4
FAM46A
C9orf150
SEPP1
MALAT1
SLC22A18
ARRDC4
ECM1
DGKE
RAB3B
AHNAK2
FASN
PCSK9
ACSS2
SQLE
IDS
SC5DL
HMGCS1
SC4MOL
HMGCR
INSIG1
IDI1
FDFT1
DHCR7
FDPS
CYP51A1
STARD4
SCD
PNPLA3
NPC1
LSS
LSLR
LPIN1
ACSL11
B
8-genes CFS signature
SEPP1
FGFR4
FAM46A
PNPLA3
HMGCS1
LDLR
RAB3B
AHNAK2
C
-3
-2
-1
-0.5
0.3
0.8
1.5
2.5
Median centered log2(intensity)

Figure 6

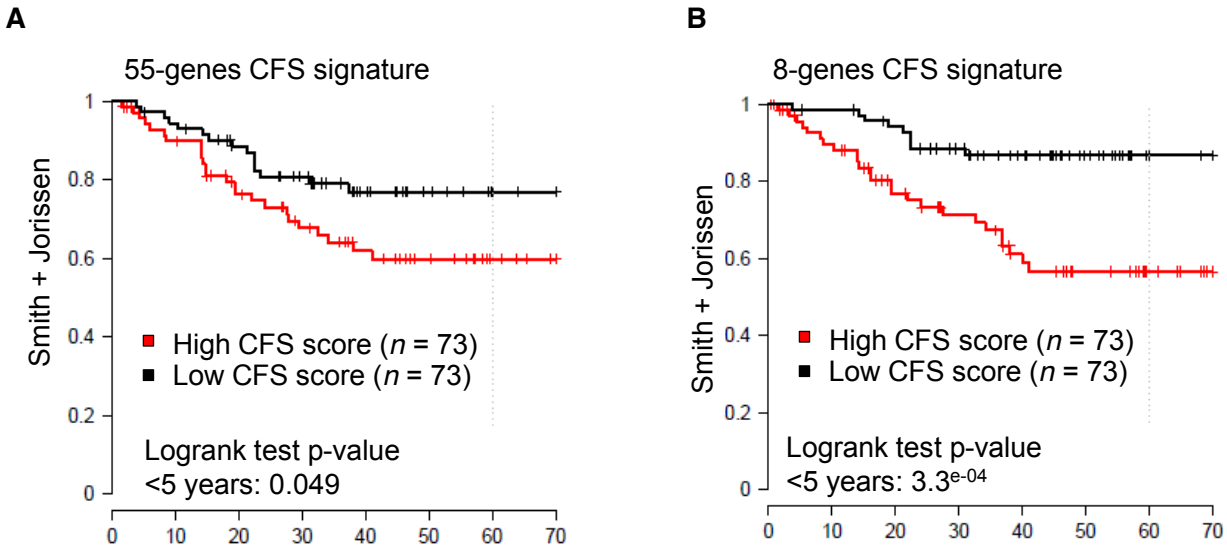

A
55-genes CFS signature
Smith + Jorissen
High CFS score (n = 73)
Low CFS score (n = 73)
Logrank test p-value
<5 years: 0.049
B
8-genes CFS signature
Smith + Jorissen
High CFS score (n = 73)
Low CFS score (n = 73)
Logrank test p-value
<5 years: 3.3e-04

Figure 7

Table 1

Table 1

	Category	Subcategory	Src	Gene Sets	*P* value up	*P* value down
Significant KEGG pathways	Metabolism	Overview	KEGG	Biosynthesis of steroids	2,6E-27	-
		Metabolism of Terpenoids and Polyketides	KEGG	Terpenoid biosynthesis	1,3E-07	-
		Energy Metabolism	KEGG	Reductive carboxylate cycle (CO2 fixation)	2,4E-04	-
		Lipid Metabolism	KEGG	Fatty acid biosynthesis	1,4E-03	-
			KEGG	Synthesis and degradation of ketone bodies	3,3E-03	-
			KEGG	Glycerolipid metabolism	3,4E-03	-
			KEGG	Biosynthesis of unsaturated fatty acids	4,3E-03	-
			KEGG	Fatty acid metabolism	9,6E-03	-
			KEGG	Glycerophospholipid metabolism	2,7E-02	-
		Carbohydrate Metabolism	KEGG	Propanoate metabolism	5,6E-03	-
			KEGG	Pyruvate metabolism	8,6E-03	-
			KEGG	Citrate cycle (TCA cycle)	3,4E-02	-
		Amino Acid Metabolism	KEGG	Valine, leucine and isoleucine degradation	1,0E-02	-
		Glycan Biosynthesis and Metabolism	KEGG	Glycosaminoglycan degradation	1,3E-02	-
			KEGG	Other glycan degradation	1,0E-02	-
		Metabolism of Other Amino Acids	KEGG	Glutathione metabolism	3,9E-02	-
	Organismal Systems	Endocrine System	KEGG	Adipocytokine signaling pathway	4,0E-03	-
			KEGG	PPAR signaling pathway	2,9E-02	-
	Environmental Information Processing	Signal Transduction	KEGG	MAPK signaling pathway	1,6E-02	-
	Genetic Information Processing	Folding, Sorting and Degradation	KEGG	Proteasome	-	1,8E-06
			KEGG	Ubiquitin mediated proteolysis	-	9,2E-04
		Replication and Repair	KEGG	DNA replication	-	1,0E-06
			KEGG	Mismatch repair	-	1,4E-04
			KEGG	Nucleotide excision repair	-	1,7E-03
	Metabolism	Carbohydrate Metabolism	KEGG	Glycolysis / Gluconeogenesis	-	7,4E-04
		Nucleotide Metabolism	KEGG	Pyrimidine metabolism	-	3,6E-02
	Cellular Processes	Cell Growth and Death	KEGG	Cell cycle	-	1,1E-02
	Environmental Information Processing	Signal Transduction	KEGG	Hedgehog signaling pathway	-	2,7E-02
			KEGG	TGF-beta signaling pathway	-	4,0E-02
Stem Cell Targeted Gene Sets	Stem Cell		MSigDB	BHATTACHARYA_EMBRYONIC_STEM_CELL (C2)	3,5E-02	-
			MSigDB	JAATINEN_HEMATOPOIETIC_STEM_CELL_UP (C2)	4,4E-02	-
			MSigDB	WONG_EMBRYONIC_STEM_CELL_CORE (C2)	-	2,2E-06
			MSigDB	OSWALD_HEMATOPOIETIC_STEM_CELL_IN_COLLAGEN_GEL_DN (C2)	-	7,4E-05
			MSigDB	BYSTRYKH_HEMATOPOIESIS_STEM_CELL_QTL_TRANS (C2)	-	2,9E-03
			MSigDB	LIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_QTL (C2)	-	1,3E-02
			MSigDB	GAL_LEUKEMIC_STEM_CELL_DN (C2)	-	4,2E-02
			MSigDB	LIANG_HEMATOPOIESIS_STEM_CELL_NUMBER_LARGE_VS_TINY_DN (C2)	-	4,4E-02
	Treatment Resistance	Drug Resistance	MSigDB	CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_3 (C2)	7,3E-09	-
			MSigDB	MASSARWEH_TAMOXIFEN_RESISTANCE_UP (C2)	2,9E-07	-
			MSigDB	KANG_CISPLATIN_RESISTANCE_UP (C2)	1,5E-03	-
			MSigDB	KANG_CISPLATIN_RESISTANCE_UP (C2)	1,5E-03	-
			GO	drug transporter activity	9,2E-03	-
			MSigDB	WHITESIDE_CISPLATIN_RESISTANCE_UP (C2)	1,3E-02	-
			MSigDB	CREIGHTON_ENDOCRINE_THERAPY_RESISTANCE_5 (C2)	1,3E-02	-
			MSigDB	WHITESIDE_CISPLATIN_RESISTANCE_UP (C2)	1,3E-02	-
			MSigDB	RIGGINS_TAMOXIFEN_RESISTANCE_UP (C2)	2,6E-02	-
			MSigDB	GO RESPONSE_TO_DRUG (C5)	4,4E-02	-
		Cell cycle	GO	cell cycle	-	1,7E-11
			GO	mitotic cell cycle	-	2,9E-08
			GO	cell cycle process	-	3,6E-08
			GO	positive regulation of ubiquitin-protein ligase activity during mitotic cell cycle	-	6,9E-07
			GO	negative regulation of ubiquitin-protein ligase activity during mitotic cell cycle	-	5,7E-06
			SMD	cancerModules Cell cycle (expression clusters)	-	1,4E-05
			Bioc	Cyclins and Cell Cycle Regulation	-	5,0E-03
			KEGG	Cell cycle	-	1,1E-02
			Bioc	Cell Cycle: G1/S Check Point	-	1,5E-02
			Bioc	Regulation of p27 Phosphorylation during Cell Cycle Progression	-	1,6E-02
			GO	M phase of mitotic cell cycle	-	2,1E-02
			GO	S phase of mitotic cell cycle	-	2,1E-02
			GO	interphase of mitotic cell cycle	-	2,3E-02
			Bioc	Ubiquitylation in the Control of Cell Cycle	-	4,8E-02
			GO	regulation of cell cycle	-	5,0E-02
		Transporter	MSigDB	GO TRANSITION_METAL_ION_TRANSMEMBRANE_TRANSPORTER_ACTIVITY (	7,0E-03	-
			GO	drug transporter activity	9,2E-03	-
			MSigDB	GO LIPID_TRANSPORTER_ACTIVITY (C5)	2,8E-02	-
			MSigDB	GO DI___TRI_VALENT_INORGANIC_CATION_TRANSMEMBRANE_TRANSPORT	3,3E-02	-
			MSigDB	REACTOME_METAL_ION_SLC_TRANSPORTERS (C2)	3,5E-02	-
			MSigDB	REACTOME_GLUCOSE_AND_OTHER_SUGAR_SLC_TRANSPORTERS (C2)	4,4E-02	-
			GO	phosphatidylinositol transporter activity	4,8E-02	-
			GO	secondary active monocarboxylate transmembrane transporter activity	4,8E-02	-
		Apoptosis	MSigDB	CONCANNON_APOPTOSIS_BY_EPOXOMICIN_UP (C2)	1,8E-03	-
			MSigDB	BROCKE_APOPTOSIS_REVERSED_BY_IL6 (C2)	2,4E-03	-
			MSigDB	GO ANTI_APOPTOSIS (C5)	2,6E-02	-
			MSigDB	GO REGULATION_OF_APOPTOSIS (C5)	4,3E-02	-
			MSigDB	REACTOME_APOPTOSIS (C2)	-	3,0E-06
			MSigDB	REACTOME_APOPTOSIS_INDUCED_DNA_FRAGMENTATION (C2)	-	1,2E-02
			MSigDB	CONCANNON_APOPTOSIS_BY_EPOXOMICIN_DN (C2)	-	1,3E-02
			MSigDB	DEBIASI_APOPTOSIS_BY_REOVIRUS_INFECTION_DN (C2)	-	1,7E-02
		DNA Damage	MSigDB	REACTOME_P53_INDEPENDENT_DNA_DAMAGE_RESPONSE (C2)	-	2,5E-07
			GO	response to DNA damage stimulus	-	1,1E-03
			SMD	cancerModules DNA damage response	-	1,2E-02
			MSigDB	KYNG_DNA_DAMAGE_BY_4NQO (C2)	1,3E-02	3,0E-02

Table 2

Table 2

Association between 8-gene CFS$^{High/Low}$ score and TNM stage and to prognosis in combined dataset Smith + Jorissen											
		Chi2 test *P* value	CFS^{High} (*n*=73)	CFS^{Low} (*n*=73)							
TNM Stage	2 (n=82)	0,097	35 (49%)	46 (63%)							
	3 (n=65)		38 (51%)	27 (37%)							
				Cox Univariate Analysis				Cox Multivariate Analysis			
	Annotation	Value	n samples	n events	H.R.	95%C.I.	p-value	H.R.	95%C.I.	p-value	model p-value
Jorissen + Smith 8-genes signature	TNM Stage	3	145	34	2,1	1-4.2	0,037	1,8	0.88-3.5	0,11	4,10E-04
	SC score	CFS^{Low}	145	34	0,27	0.13-0.58	8,0E-04	0,29	0.14-0.64	1,8E-03	
H.R.: Cox Hazard Ratio, 95% C.I.: 95 Percent Confidence Interval of HR. Value : modality of the annotation associated to H.R.											

Annexe C

Etude du pronostic inter et intra sous-type MSI

Pronostic de la mutation de la protéine chaperonne dans les tumeurs de type MSI

ACCEPTED MANUSCRIPT

PATIENTS WITH COLORECTAL TUMORS WITH MICROSATELLITE INSTABILITY AND LARGE DELETIONS IN HSP110 T_{17} HAVE IMPROVED RESPONSE TO 5-FLUOROURACIL-BASED CHEMOTHERAPY

RUNNING HEAD: HSP110 PREDICTS SURVIVAL OF MSI CRC PATIENTS

ADA COLLURA [1,2,£], ANAÏS LAGRANGE [1,2], MAGALI SVRCEK [1,2,3], LAETITIA MARISA [4], OLIVIER BUHARD [1,2], AGATHE GUILLOUX [1,2], KRISTELL WANHERDRICK [1,2], CORALIE DORARD [1,2], ANNA TAIEB [1,2], ARNAUD SAGET [1,2], MARIE LOH [5], RICHIE SOONG [5], NIKOLAJS ZEPS [6,7], CAMERON PLATELL [6,7], ANDREW MEWS [6,7], BARRY IACOPETTA [7], AURÉLIE DE THONEL [8,9], RENAUD SEIGNEURIC [8,9], GUILLAUME MARCION [8,9], CAROLINE CHAPUSOT [10], COME LEPAGE [11], ANNE-MARIE BOUVIER [11], MARIE-PIERRE GAUB [12], GÉRARD MILANO [13], JANICK SELVES [14], PATRICK SENET [15], PATRICE DELARUE [15], HAYAT ARZOUK [16,17], CLAIRE LACOSTE [16,17], ARNAUD COQUELLE [16,17], LEILA BENGRINE-LEFÈVRE [18], CHRISTOPHE TOURNIGAND [18], JÉRÉMIE H LEFÈVRE [2,19], YANN PARC [2,19], DENIS S. BIARD [20], JEAN-FRANÇOIS FLEJOU [1,2,3], CARMEN GARRIDO [8,9,21], ALEX DUVAL [1,2,£]

(1) INSERM, UMRS 938 - Centre de Recherche Saint-Antoine, Equipe "Instabilité des Microsatellites et Cancers", Equipe labellisée par la Ligue Nationale contre le Cancer, F-75012, Paris, France;
(2) Université Pierre et Marie Curie-Paris 6, Paris, France;
(3) AP-HP, Hôpital Saint-Antoine, Service d'Anatomie et Cytologie Pathologiques, Paris, France. Plateforme de microdissection de l'Institut Fédératif de Recherche 65 et Tumorothèque des Hôpitaux Universitaires Paris-Est, Paris, France.
(4) Programme "Cartes d'Identité des Tumeurs", Ligue Nationale Contre le Cancer, Paris, France;
(5) Cancer Science Institute of Singapore, National University of Singapore, Singapore;
(6) Bendat Family Comprehensive Cancer Centre, St John of God HealthCare, Subiaco, 6008 Australia;
(7) School of Surgery M507, University of Western Australia, 35 Stirling Hwy, Nedlands, 6009, Australia;
(8) INSERM, UMRS 866, 21033 Dijon, France;
(9) University of Burgundy, Esplanade Erasme, 21078 Dijon, France;
(10) Service de Pathologie, CHU Dijon, France;
(11) Burgundy Cancer Registry, INSERM U866, Burgundy University, Dijon University Hospital, BP 87900 21079 Dijon, France;
(12) INSERM, U682, Développement et Physiopathologie de l'Intestin et du Pancréas, 67200 Strasbourg, France;
(13) Laboratoire d'Oncopharmacologie, EA 3836, Centre Antoine Lacassagne, Nice, France ;
(14) INSERM, Unité 563, Centre de Recherche sur le Cancer de Toulouse, France;
(15) UMR 6303 CNRS-Université de Bourgogne, 21078 Dijon Cedex ;
(16) IRCM, Institut de Recherche en Cancérologie de Montpellier, Montpellier, F-34298, France ; INSERM, U896, Montpellier, F-34298, France ;
(17) Université Montpellier1, Montpellier, F-34298, France ; Institut régional du Cancer Montpellier, Montpellier, F-34298, France.
(18) Service d 'oncologie médicale, Hôpital Henri Mondor, Université Paris Est Créteil ;
(19) AP-HP, Service de Chirurgie Générale et Digestive, Hôpital Saint-Antoine, Paris, France ;
(20) CEA, DSV, iMETI, SEPIA, Fontenay-aux-Roses,France;
(21) Anticancer center Georges François Leclerc, Dijon.

[£] Address for correspondence: Alex Duval and Ada Collura. Emails: **alex.duval@inserm.fr**, **ada.collura@inserm.fr**. INSERM UMRS 938, Paris, France.

Conflicts of interest : The authors disclose no conflicts.

Author Contributions-List (Initials) : study concept and design (AC[£], AD); acquisition of data (AC[£], AL, OB, KW, CD, AT, AS, AdT, CL, AC); analysis and interpretation of data (AC[£], AL, LM, OB, AG, AdT, AC, CG); drafting of the manuscript (AC[£], AD); critical revision of the manuscript for important intellectual content (AC[£], AD, BI, LM, AG, CG); statistical analysis (LM, AG, AD); obtained funding (AD, CG); administrative, technical, or

ACCEPTED MANUSCRIPT

material support (AC[£] AL, ML, RS, NZ, CP, AM, BI, RS, GM, CC, CL, AMB, MPG, GM, JS, PS, PD, HA, CL, AC, LBL, CT, JHL, YP, DSB, JFF); study supervision (AD).

Financial Disclosure

No funding bodies had any role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. This work was supported by the 'Carte d'Identité des Tumeurs' (CIT) program (http://cit.ligue-cancer.net) from the Ligue Nationale Contre le Cancer and by grants from the 'Institut National du Cancer' (INCa) (To AD). AD group has the label de « La Ligue Contre le Cancer ». AC is a recipient of an INCa fellowship (Institut National du Cancer). AL is a recipient of a MESR fellowship (Ministère de l'Enseignement Supérieur et de le Recherche).

Acknowledgments

AC is a recipient of an INCa fellowship (Institut National du Cancer). We thank Dr. Nizar El-Murr for critical reading of the manuscript. We thank All authors declare no conflict of interest. A posthumous thank to Pr. Françoise Piard, for her help and contribution to this work.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Abstract

BACKGROUND & AIMS: Patients with colorectal tumors with microsatellite instability (MSI) have better prognoses than patients with tumors without MSI, but have a poor response to 5-fluorouracil-based chemotherapy. A dominant-negative form of HSP110 (HSP110DE9) expressed by cancer cells with MSI, via exon skipping caused by somatic deletions in the T_{17} intron repeat, sensitizes the cells to 5-fluorouracil and oxaliplatin. We investigated whether *HSP110* T_{17} could be used to identify patients with colorectal cancer who would benefit from adjuvant chemotherapy with 5-fluorouracil and oxaliplatin.

METHODS: We characterized the interaction between HSP110 and HSP110DE9 using surface plasmon resonance. Using PCR and fragment analysis, we examined how the size of somatic allelic deletions in *HSP110* T_{17} affected the HSP110 protein expressed by tumor cells. We screened a 329 consecutive patients with stage II–III colorectal tumors with MSI who underwent surgical resection, at tertiary medical centers, for *HSP110* T_{17}.

RESULTS: HSP110 and HSP110DE9 interacted in a 1:1 ratio. Tumor cells with large deletions in T_{17} had increased ratios of HSP110DE9:HSP110, due to the loss of expression of full-length HSP110. Deletions in *HSP110* T_{17} were mostly bi-allelic in primary tumor samples with MSI. Patients with stage II–III cancer who received chemotherapy and had large *HSP110* T_{17} deletions

(≥5 bp; 18/77 patients, 23.4%) had longer times of relapse-free survival than patients with small or no deletions (≤4 bp; 59/77 patients, 76.6%) in multivariate analysis (hazard ratio, 0.16; 95% confidence interval, 0.012–0.8; *P*=.03). We found a significant interaction between chemotherapy and T_{17} deletion (*P*=.009).

CONCLUSIONS: About 25% of patients with stage II–III colorectal tumors with MSI have an excellent response to chemotherapy, due to large, bi-allelic deletions in the T_{17} intron repeat of *HSP110* in tumor DNA.

KEYWORDS: therapeutic response, prognostic factor, outcome, treatment

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Introduction

Colorectal cancer (CRC) is the second most common cause of cancer-related death worldwide. CRC is a molecularly heterogeneous disease, with the majority of cases (80-85%) displaying chromosomal instability (CIN) in conjunction with microsatellite stability (MSS). A significant fraction (15-20%) shows widespread instability at DNA repeats (MSI for Microsatellite Instability), due to a defective mismatch repair (MMR) system [1-3]. Clinically, MSI tumors have been reported to show improved prognosis but a bad response to 5-fluorouracil-based chemotherapy [4-6].

In a recent study, we reported specific mutation of the molecular chaperone HSP110 in MSI CRC [7]. In colon cancers, chaperone proteins including HSP110 promote the survival of malignant cells [8-10]. We provided evidence that an HSP110 mutant, referred to as HSP110DE9, was specifically expressed in MSI CRCs. This protein was generated from an aberrantly spliced mRNA lacking exon 9, thus encoding a truncated HSP110 protein. HSP110DE9 was found to abrogate the chaperone activity of HSP110 in a dominant-negative manner. Its forced overexpression caused MSI CRC cells to become sensitized to 5-fluorouracil and oxaliplatin. A fraction of stage II-III MSI CRC patients showing high tumor expression of HSP110DE9 had significantly longer relapse-free survival (RFS) compared to those with low expression [7]. However, these last results were preliminary since based on the analysis of a small series of MSI CRC patients treated at a single center.

One of the most tantalizing questions raised by our previous study is whether *HSP110* mutation provides a mechanistic basis for the possibility that a fraction of MSI CRCs displaying high expression of mutant HSP110DE9 relatively to HSP110wt (wild-type) might be more responsive to chemotherapy. We have shown in earlier work that aberrant expression of HSP110DE9 in MSI cancer cells was associated with somatic deletion of the T_{17} repeat within intron 8 of *HSP110* [7]. Decreasing length of the *HSP110* T_{17} repeat was likely to correlate with increased synthesis of HSP110DE9 due to exon 9 skipping. Nevertheless, such a correlation was not precisely examined in primary tumor samples. In the present study, we investigated whether the mutation status of *HSP110* T_{17} might be useful to identify a fraction of CRC patients who would benefit from adjuvant chemotherapy. Results obtained indicated that accurate measurement of the *HSP110* T_{17} deletion in tumor DNA was likely to constitute an unbiased way to evaluate the HSP110DE9/HSP110 ratio in tumor cells and the clinical impact of *HSP110* status in MSI colon cancer patients.

Materials and Methods

Ethics Statement. This study was approved by the institutional review board/ethics committee of the participating centers. Informed consent was recorded in each case. Patients who received preoperative chemotherapy and/or radiation therapy were excluded from this study.

Surface plasmon resonance. Kinetics and affinity constant measurements of HSP110DE9/HSP110wt interaction were performed in a BIAcore BX100 (GE Healthcare). All experiments were performed with immobilized His-tagged HSP110 (10µg/mL) using an NTA sensor chip at 25°C. For affinity constant determination, curves were analyzed with a global fit 1:1 binding model with drifting baseline.

ACCEPTED MANUSCRIPT

Modeling human HSP110 and the protein-protein interface HSP110/HSP110DE9. The structures of hHSP110 and HSP110DE9 were built by homology using the experimental structure of yeast Hsp110 Sse1 (Protein Data Bank code: 3C7N) [11] as a template for the software MODELLER [12]. The models were fitted with 3C7N and relaxed by molecular dynamics simulations. We used the 3C7N monomer A to produce a set of 3D models for HSP110 and HSP110DE9 sequences. All-atom molecular dynamics simulations in explicit water were carried out with the GROMACS software package using the GROMOS96 ffG43a1 force field and the Simple Point Charge (SPC) water model (Hochschulverlag AG, Zurich, Switzerland). The time step (0.001 ps) and the list of neighbors were updated every 0.005 ps with the 'grid' method and a cut-off radius of 1 nm. The coordinates of all the atoms in the simulation box were saved every 2 ps. The initial velocities were chosen randomly. We used the NPT ensemble with a cubic box of initial side equal to 15.754 nm with 125,055 SPC water molecules keeping a minimum distance of 0.9 nm between the solute and each face of the box. The charge of hHSP110-HSP110DE9 was neutralized by adding 19 Na^+ counter-ions. The temperature and pressure were kept using the Berendsen method and isotropic coupling for the pressure (T=310K, τ_T=0.1 ps; P_0=1 bar, coupling time τ_P=1 ps). The electrostatic term was computed using the Particle Mesh Ewald (PME) algorithm (radius: 1 nm) with the Fast Fourier Transform optimization. The "cut off" algorithm was applied for the non-coulomb potentials with a radius of 1 nm. The production period was 70 ns.

Extensive analysis of a panel of HCT116 sub-clones for both HSP110DE9/HSP110wt mRNA expression ratio and T_{17} deletion status. The HCT116 cell line was cultured in DMEM media as described [7]. Single cell sub-cloning was performed using MoFlo Astrios (Beckman Coulter), spotting 1 cell/well in 96 well plates containing 200 µl of DMEM media. The HSP110DE9/HSP110wt mRNA expression ratio was evaluated from 10^6 cells of each single HCT116 cell sub-cloned line. Total RNA was purified with an RNeasy Mini Kit (Qiagen). Complementary DNAs were synthesized using the High Capacity cDNA reverse transcription kit (Applied Biosystems). Primers, internal probes and thermal cycling conditions were used as described [7]. DNA from each single HCT116 cell sub-cloned line was extracted using QIamp DNA Mini Kit (Qiagen).

Western blotting. Total protein extractions were obtained from adherent cells lysed with 1X Laemli sample buffer (BioRad). Protein extractions were treated with DNAse set (Qiagen). Proteins were separated in SDS-polyacrylamide gels and transferred to nitrocellulose membrane (Hybond ECL GE healthcare). Membranes were first probed using primary antibodies: HSP110 (EPR4576 Abcam); HSC70 (13D3 Abcam); 14-3-3 (Santa-Cruz). Next, membranes were incubated with secondary HRP-coupled antibodies (Jackson ImmunoResearch Laboratories) before revelation.

EBV-based vectors construction. We introduced HSP110wt or HSP110DE9 mutant open reading frames (ORF) into puromycin-resistant pEBV plasmids [13] downstream of a CAG promoter (pEBVCAG-puro or pBD2347, unpublished) to obtain pEBVCAG-HSP110wt-puro (pBD2640) and pEBVCAG-HSP110DE9-puro (pBD2570) plasmids, respectively. For PCR amplification, we used the same forward primer for both constructs (5'-ATGTCGGTGGTGGGGTTGGACGTGGGC-3'). As reverse primers, we employed 5'-CTAGTCCAAGTCCATATTAACAGAATT-3' for HSP110wt and 5'-TCATGAACACTGTAATGCACATCC-3' for HSP110DE9. Cells were transfected with JetPrime (Ozyme) according to the manufacturer's recommendations. 24h later, cells medium was

ACCEPTED MANUSCRIPT

supplemented with puromycine (0.5 µg/ml for HCT116 or 2.0 µg/ml for SW480 and FET cell lines).

Chemosensitivity assay. Rates of sensitivity to 5-fluorouracil and oxaliplatin were assessed using WST-1 (Roche). Briefly, $2x10^4$ cells of each cell line were plated per well in 24-well plates in 2 ml of media with or without drugs. After 72h, WST-1 reagent was added and incubated for 4h at 37°C. The absorbance was measured at 450 nm. The reference wavelength was at 750 nm.

Patients and specimens. We identified 329 patients who underwent surgical resection for histologically proven stage II or stage III MSI CRC from 1998-2007 in one of the 6 clinical centers involved in this study (Hôpital Saint-Antoine, Paris; CHU de Dijon, Dijon; CHU de Toulouse Purpan, Toulouse; Centre Antoine Lacassagne, Nice; St John of God Pathology, Subiaco, Australia; National University Hospital, Singapore) (Supplementary **Table S1**). In all French clinical centers, MSI was prospectively identified at diagnosis using the pentaplex PCR method [14-18]. The detection of MSI in tumors from Singapore and Australia was prospectively identified using immunohistochemistry and for the same time period (1998-2007), but this was confirmed using the same PCR pentaplex method [14, 15] (Supplementary **Table S1**). In a subgroup of 166 patients, the methylation status of the *MLH1* promoter was examined (Supplementary **Table S1**). We also analyzed a retrospective cohort of 258 MSS CRC patients collected from the same clinical centers and matched for tumor stage, year of diagnosis, age, gender and primary tumor location (**Table 1** and Supplementary **Table S1**).

Extensive clinical follow up and treatment details were available for all MSI CRC patients included in this study. The study was conducted according to the recommendations of the institutional authorities. Patients who received adjuvant therapy (n=77) were mainly treated with 5-fluorouracil-based chemotherapy, *i.e.* 5-fluorouracil plus leucovorin, either alone (LV5FU2, n=42; FUFOL, n=2) or in combination with other drugs (oxaliplatin, n=32) (Supplementary **Table S1**; in 3 cases, the exact data concerning the used 5-FU-based chemotherapy regimen were unknown, *i.e.* 5-FU alone or in combination with oxaliplatin). Adjuvant chemotherapy was systematically proposed for stage III CRC patients and administrated in the absence of contraindication. In stage II CRC patients, it was proposed in "at risk" individuals according to the following criteria, *i.e.* perforated cancer, pT4N0 with vascular emboli, and/or obstructive colorectal tumor, as reported [19]. Recurrence was uniformly assessed, *i.e.* physical examination with biological tests and measurement of carcino-embryonic antigen level, Chest X-ray and abdominal ultrasonography or computed tomography every 3 months during the first 3 years after surgery, then every 6 months for 2 years, and then annually. Patient follow-up was defined as the time elapsed between surgery and the last hospital contact or disease recurrence.

DNA extraction from primary tumor samples. We processed frozen tissues (30 mm^3) using the Qiamp protocol (Qiagen). Formalin-fixed and paraffin-embedded tissues were processed as described [20].

Laser capture microdissection. Six serial 5 mm paraffin-embedded sections from 27 cases were cut and then mounted onto membrane slides (PALM Membranes Slides, Bernreid, Germany), as described [17].

Mutation analysis of the *HSP110* T_{17} and other microsatellite sequences contained in target genes in cohorts of patients with MSI or MSS CRC. The polymorphic status of the *HSP110* T_{17} repeat was evaluated in a series of 50 MMR-proficient lymphoblastoid cell lines from

ACCEPTED MANUSCRIPT

healthy controls (CEPH institute, Paris, France) using PCR and fragment analysis (data not shown). The mutation status of *HSP110* T_{17} was initially evaluated in the set of 98 MSI CRC samples from the Saint-Antoine hospital that had previously been investigated for HSP110DE9 expression (MSI-Set1) [7]. The mutation status of *HSP110* T_{17} was subsequently examined in a second set of 231 MSI CRC samples from the five other clinical centers involved in this study (MSI-Set2). The multi-centered cohort of 258 MSS CRCs was used to evaluate whether *HSP110* T_{17} mutations also occurred in MMR-proficient tumors. **Table 1** summarizes the overall clinical features of patients and their tumor characteristics. A total of 15 other genes containing mononucleotide repeat sequences were also analyzed [21,22].

Statistical analysis. All statistical analyses were stratified according to clinical centers in order to take into account the potential heterogeneity of different centers. RFS was used and this was defined as the time from surgery to the date of first recurrence (relapse, or death from CRC) or last contact. Patients who were alive without relapse at the last follow-up were considered as censored cases, *i.e.* they were included in the "at risk" set in the survival probability estimations until they were lost to follow-up. Survival curves were obtained according to the method of Kaplan and Meier and differences between survival distributions were assessed by log-rank test using an endpoint of five years. Univariate and multivariate models were computed using Cox proportional-hazards regression. For multivariate analyses, only those variables with information available for all sample groups were included in models. Interaction between *HSP110* T_{17} deletions and adjuvant chemotherapy was assessed using the likelihood ratio test. Graphical and statistical methods were used to examine whether proportional hazards assumptions were satisfied [7]. Survival analyses were performed using the R package survival.

The cut-off value that resulted in maximal survival difference between patient groups with large and small deletions in the *HSP110* T_{17} was determined by the minimal *p*-value approach. To overcome a possible false detection of a cut-off, which is known to constitute a risk for this approach [23], we first investigated the robustness of our 5 bp cut-off by bootstrap. In addition, we applied the *p*-value correction of Lausen and Schumacher (1996) to test for the existence of a threshold in the effect of *HSP110* T_{17} deletions on RFS.

Differences between *HSP110* Del^L (Large T_{17} deletions) and Del^S (Small T_{17} deletions) groups and other clinical annotations were tested for statistical significance using the Cochran-Mantel-Haenszel chi-squared test for categorical variables, or an unpaired Student's t-test for continuous variables.

For all analyses, *p*-values of less than .05 were considered to indicate statistical significance.

Results

HSP110DE9/HSP110wt interaction *in vitro.* To better understand the molecular basis for any potential prognostic value of HSP110DE9, we studied the HSP110DE9/HSP110wt interaction by computer modeling and surface plasmon resonance (BIACORE). *In silico* determination showed that one molecule of HSP110DE9 interacted with one molecule of HSP110wt (**Fig. 1A**). Our MD calculations indicated the HSP110 amino acids ASP^{633}, GLN^{707} and GLU^{708} were essential for this interaction. Interestingly, these amino acids are located within the peptide-binding domain of HSP110wt (also called chaperone domain), which

ACCEPTED MANUSCRIPT

could explain its inactivation through HSP110DE9 interaction. Likewise, our BIACORE studies also indicated a 1:1 association between HSP110wt and HSP110DE9 that was favored in the presence of ADP (HSP110 has an ATP binding domain) (**Fig. 1B**). Thus, HSP110wt and HSP110DE9 are able to physically interact with each other and they interact with each other in 1:1 ratio.

HSP110DE9/HSP110wt expression ratio and T_{17} deletion status in MSI CRC cell lines. We analyzed a panel of 33 HCT116 sub-clones for their T_{17} deletion status (DT) (**Fig. 1C**). A pronounced increase in the expression of HSP110DE9/HSP110wt mRNA ratio was observed in sub-clones with 4 bp or larger T_{17} deletions (**Fig. 1C**). In line with the mRNA results, a correlation was also observed at the protein level in 4 HCT116 sub-clones displaying small or large *HSP110* T_{17} deletions (**Fig. 1C**). HCT116 and LS174T displaying small or large T_{17} deletions were also analyzed and similar results were obtained (**Fig. 1D**). Most likely, these results simply mean that the predominant product of the gene switches to HSP110DE9 as the deletion gets longer (≥4 bp). In both quantitative RT-PCR and western-blotting experiments, the difference in HSP110DE9/HSP110wt expression ratio between cell lines with small or large T_{17} deletions was mainly due to a decrease in HSP110wt expression whereas the expression of HSP110DE9 stayed relatively similar; firstly, small deletions in T_{17} resulting from MSI allow the aberrant expression of HSP110DE9 through exon skipping. This is followed by the loss of expression of HSP110wt once the size of the deletion increases beyond a certain point.

***HSP110* T_{17} status of MSS and MSI CRCs.** None of the 258 MSS CRCs investigated here was found to contain a mutation in the *HSP110* T_{17} repeat. In all cases, allelic lengths were either T_{16} or T_{17} (**Fig. 2A**) and thus within the polymorphic zone observed in lymphoblastoid cell lines (data not shown). In contrast, the large majority of MSI tumors (n=319/329, 97%) showed deletions of up to 7 base pairs that were outside the polymorphic zone (*P*<.001 compared to MSS; **Fig. 2B** and **2C**). For the majority of analyzed primary tumor samples, only one mutated allele type was detected and this corresponded to the main clonal population present in the tumors. For the cases that displayed multi-allelic profiles, the peak associated with the larger T_{17} deletion and that did not appear to result from a stuttering of Taq polymerase was used for classification purposes. The mutation frequencies of the 15 other genes representing reported targets for MSI-driven instability were highly variable in MSI colon tumors (**Fig. 2D**).

***HSP110* T_{17} deletions are mostly bi-allelic in MSI CRCs.** The analysis of MSI CRC cell lines showed that the majority of these models (n=11/13, 85%) displayed bi-allelic alterations with no remaining detectable wild-type *HSP110* T_{17} allele (Supplementary **Fig. S1A**). Primary tumor samples were not micro- or even macro-dissected. Therefore, the peaks corresponding to T_{17} alleles simply reflected what is usually observed when performing routine analysis of this DNA repeat in MSI primary CRCs. The presence of bi-allelic mutations for *HSP110* T_{17} was demonstrated by first performing microdissection of 3 primary tumor samples that were highly contaminated with normal cells. We compared *HSP110* T_{17} and *TGFBR2* A_{10} mutation profiles which display mainly bi-allelic mutations in MSI cancer cells [24] before and after microdissection of the tumor tissue. In all 3 cases, microdissection led to disappearance of the *HSP110* and *TGFBR2* wild-type alleles (**Fig. 2E**). We also compared the mutation profile of *HSP110* T_{17} with that of *TGFBR2* A_{10} and the pentaplex panel in 7 other primary tumors that displayed varying peak intensities for the wild type T_{17} allele. The results showed a high overall level of correlation for the mutation status of all these DNA repeats

ACCEPTED MANUSCRIPT

(Supplementary **Fig. S1B**). We confirmed these results by performing microdissection experiments on 24 additional primary CRCs (Supplementary **Fig. S1C**). Sixteen primary tumor samples that contained low levels of contamination with normal cells were also found to display bi-allelic deletions in *HSP110* T_{17}, even without microdissection (Supplementary **Fig. S1D**). Consequently, these results demonstrate that *HSP110* T_{17} deletions are usually bi-allelic in MSI primary colon tumors.

Varied chemosensitivity of both MSI and MSS CRC cell lines based on the expression of HSP110DE9. We performed chemosensitivity assays using stably transfected MSI and MSS CRC cell lines that overexpressed HSP110wt or HSP110DE9 protein. Importantly, these cells were transfected using an EBV-based vector allowing stable expression of genes [13]. This allows the possibility to work with polyclonal models, thus avoiding the biases inherent with single cell approaches. Using this method, we demonstrated that both MSI (HCT116) and MSS (SW480 and FET, in which *HSP110* T_{17} is not mutated) CRC cells became more sensitive to chemotherapy (5-FU, oxaliplatin) when they overexpressed HSP110DE9 as compared to HSP110wt (**Fig. 3A**).

***HSP110* T_{17} mutation status and the survival of stage II-III MSI CRC patients.** The first (MSI-Set1) and second set (MSI-Set2) of patients showed distinct clinical characteristics in terms of gender, age and tumor location (**Table 1**). This was due to the fact that Saint-Antoine hospital in Paris is a reference center for the treatment of Lynch patients, whereas recruitment at the other clinical centers was enriched for elderly patients known to contain more women and sporadic MSI tumors. For each set, the threshold value for the size of *HSP110* T_{17} deletions that resulted in maximal survival difference between patient groups with long and short deletions was independently determined. In both sets, the same 5 bp deletion cut-off identified that stage II-III MSI patients under chemotherapy with large deletions had excellent survival (Supplementary **Fig. S2A**). The 329 MSI CRC patients were therefore classified into two groups displaying large deletions (DT ≥ 5 bp; *HSP110* Del^L; n=76/329, 23%) or small deletions (0 ≤ DT ≤ 4; *HSP110* Del^S; n=253/329, 77%).

In the overall cohort, *HSP110* T_{17} mutation status showed a trend for association with RFS in multivariate analysis (HR, 0.57 [95% CI, 0.29-1.1], *P*=.096; Supplementary **Table S2**). Stage II-III patients who received chemotherapy and had large *HSP110* T_{17} deletions (n=18/77, 23.4%) showed excellent RFS compared to patients with small deletions (5-year RFS of 94% *vs* 64%, respectively; Log-rank *P*=.04; **Fig. 3B**) in multivariate analysis (HR, 0.16 [95% CI, 0.012-0.8], *P*=.03; **Table 2**). In contrast and as expected due to the chemosensitizing effect of HSP110DE9, no significant influence of T_{17} status was observed in stage II-III patients who did not receive chemotherapy (Supplementary **Fig. S2B**). A significant interaction between chemotherapy and T_{17} mutation status was observed in both univariate (*P*=.03 for interaction) and multivariate models (*P*=.009 for interaction) (**Table S3**). The association between survival and T_{17} deletion status remained significant in the subgroup of 42 patients treated with 5-fluorouracil alone (**Fig. 3C**). The interaction between chemotherapy with 5-fluorouracil alone and T_{17} mutation status also remained significant (*P*=.014 and *P*=.007 for interaction using univariate and multivariate models, respectively; data not shown).

No significant associations were found between the size of *HSP110* T_{17} deletions and tumor stage (*P*=.89) or the methylation status of the *MLH1* promoter (*P*=.52) (**Table 3**). The only positive association observed for the other clinical parameters was between *HSP110* Del^L and proximal tumor location (*P*=.026; **Table 3**).

ACCEPTED MANUSCRIPT

The Lausen and Schumacher corrected *p*-value for Stage II-III patients who received chemotherapy was .057 (data not shown). Although not reaching significance, it indicates an homogeneity with regards to RFS within the two groups displaying large or small deletions. When considering the other target genes for MSI (**Fig. 2D**), we observed that, in each case, the difference in survival between MSI mutated and wild type groups remained not significant for stage II-III patients who received chemotherapy (**Table 4**).

Discussion

In our previous study [7], we described the novel finding that a chaperone protein, *i.e.* HSP110, was mutated in human colon cancer. We showed this mutation was specific and occurred frequently in MSI tumors, leading to abrogation of HSP110 chaperone activity and of its anti-apoptotic function. We presented evidence showing that forced overexpression of the HSP110DE9 mutant protein in CRC cell lines led them becoming sensitized to chemotherapy. Finally, we used quantitative RT-PCR to calculate the HSP110DE9/HSP110wt ratio in a small series of primary tumor samples. This RNA ratio suggested that HSP110DE9/HSP110wt expression might be discriminant for the response of MSI CRC patients to chemotherapy.

In the present work, we further demonstrated the dominant negative effect of this dominant negative mutant. We have characterized the 1:1 molecular interaction between HSP110wt and HSP110DE9 that suggests the HSP110DE9/HSP110wt ratio in tumor cells must be greater than 1 to neutralize all HSP110wt and thereby to obtain a significant chemosensitization. In both cell lines and primary tumors, we have shown this situation occurs in a fraction of MSI samples due to large deletions of the *HSP110* T_{17} repeat that allow both the aberrant expression of HSP110DE9 mutant and the complete silencing of HSP110wt in tumor cells. Our data clearly show that *HSP110* T_{17} mutations are usually bi-allelic in primary MSI CRCs. Careful examination of T_{17} status in a consecutive, multi-centered series of patients whose positive MSI status was prospectively identified at the time of diagnosis confirmed that, in line with our molecular data, only patients with large T_{17} deletions (5 bp or more) and representing a minority but nevertheless important fraction of MSI CRC patients (*i.e.* about 25%) appeared to benefit from 5-FU-based adjuvant chemotherapy. Of particular interest, the association between survival and T_{17} deletion status remained significant in the subgroup of patients treated with 5-fluorouracil alone.

It has yet to be established that adjuvant chemotherapy confers a clear survival advantage to stage II CRC patients. However, the publication by André et *al*. [19] proposed that a group of patients with "at risk" stage II colorectal cancer may benefit from adjuvant chemotherapy (*e.g.* those with perforation, pT4N0 with vascular emboli, and/or obstructive colorectal cancer). The very good survival following chemotherapy observed for MSI CRC patients with large HSP110 T17 deletions highlights the potential clinical importance of this predictive biomarker with respect to the use of adjuvant therapy for stage II CRC patients displaying large T17 deletions. It is worth noting that none of the 4 chemotherapy-treated stage II MSI patients with large *HSP110* T17 deletions suffered a relapse. However, it is not possible to reach any firm conclusions given the very small number of stage II patients analyzed who received chemotherapy. An interesting clinical follow-up would therefore be to independently confirm if stage II MSI patients with large *HSP110* deletions derive any survival benefit from adjuvant chemotherapy.

ACCEPTED MANUSCRIPT

There are several limitations to this study. Given the number of subjects with the large *HSP110* deletion (*i.e.* about 25% of MSI CRC patients), the clinical relevance of our findings has to be further examined. Importantly, it is however strongly supported by mechanistic data. Besides, we show that 15 other mutated genes including some that might play a role in drug response (*e.g. ATR, ATM*) have no prognostic significance, highlighting the particular role of *HSP110* in predicting response to chemotherapy. We did not assess whether the prognostic impact of *HSP110* T_{17} deletion was independent of other molecular features such as *BRAF* mutation and the CIMP phenotype that have been putatively associated with CRC prognosis and are significantly associated with the MSI phenotype [25, 26]. Finally, the screening for *HSP110* T_{17} deletions was performed retrospectively and individuals were not randomized to receive chemotherapy, thus potentially introducing some bias. Nevertheless, our MSI tumor cohort is one of the largest ever investigated for survival and was consecutive and prospectively collected. This should minimize the biases often inherent to studies that investigate non-consecutive cohorts compiled in a retrospective manner. We observed no associations between *HSP110* mutation status and clinical variables such as disease stage or Lynch syndrome that were recently suggested to influence the survival of MSI CRC patients [27]. Moreover, there is no evidence to suggest that the level of instability in mononucleotide repeats such as T_{17} in *HSP110* differs between Lynch and sporadic MSI tumors [28]. Interestingly, a recent immunohistochemistry-based study of a small tumor cohort has largely confirmed our findings [29], highlighting that expression of HSP110wt is likely to constitute a prognostic factor in MSI CRC. Despite observing an inverse correlation between HSP110wt expression and the size of *HSP110* T_{17} deletions in their tumor series, these workers failed to observe a significant impact of T_{17} mutation status on patient survival. This could be due to several reasons, including the manner in which somatic T_{17} deletions were assessed in tumors. As we have shown here, the analysis must be standardized so that it takes into account the polymorphisms within this DNA marker and the heterogeneity of tumor cell content. Other reasons could be the relatively small size of the study cohort and the design of survival analyses.

Although defective MMR is the well-established mechanism by which MSI tumors develop, the precise downstream events that functionally explain differences in the clinical behavior of MSI cancers remain unclear. Confirmation of the present findings using our genomic approach or alternative standardized methods should lead to reconsideration of the clinical behavior of MSI CRCs in terms of response to 5-fluorouracil-based adjuvant chemotherapy towards a more individualized medicine.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

References

1. Aaltonen LA, Peltomaki P, Leach FS, et al. Clues to the pathogenesis of familial colorectal cancer. Science 1993;260:812-6.
2. Ionov Y, Peinado MA, Malkhosyan S, et al. Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. Nature 1993;363:558-61.
3. Fishel R, Lescoe MK, Rao MR, et al. The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. Cell 1993;75:1027-38.
4. Ribic CM, Sargent DJ, Moore MJ, et al. Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. N Engl J Med 2003;349:247-57.
5. Carethers JM, Smith EJ, Behling CA, et al. Use of 5-fluorouracil and survival in patients with microsatellite-unstable colorectal cancer. Gastroenterology 2004;126:394-401.
6. Sargent DJ, Marsoni S, Monges G, et al. Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. J Clin Oncol 2010;28:3219-26.
7. Dorard C, de Thonel A, Collura A, et al. Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. Nat Med 2011;17:1283-89.
8. Kai M, Nakatsura T, Egami H, et al. Heat shock protein 105 is overexpressed in a variety of human tumors. Oncol Rep 2003;10:1777-82.
9. Slaby O, Sobkova K, Svoboda M, et al. Significant overexpression of Hsp110 gene during colorectal cancer progression. Oncol Rep 2009;21:1235-41.
10. Rerole AL, Gobbo J, De Thonel A, et al. Peptides and Aptamers Targeting HSP70: A Novel Approach for Anticancer Chemotherapy. Cancer Res 2011;71:484-95.
11. Schuermann JP, Jiang J, Cuellar J, et al. Structure of the Hsp110:Hsc70 nucleotide exchange machine. Mol Cell 2008;31:232-43.
12. Sali A, Blundell TL. Comparative protein modelling by satisfaction of spatial restraints. J Mol Biol 1993;234:779-815.
13. Biard DS, Despras E, Sarasin A, et al. Development of new EBV-based vectors for stable expression of small interfering RNA to mimick human syndromes: application to NER gene silencing. Mol Cancer Res 2005;3:519-29.
14. Suraweera N, Duval A, Reperant M, et al. Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. Gastroenterology 2002;123:1804-11.
15. Buhard O, Cattaneo F, Wong YF, et al. Multipopulation analysis of polymorphisms in five mononucleotide repeats used to determine the microsatellite instability status of human tumors. J Clin Oncol 2006;24:241-51.
16. Hampel H, Frankel WL, Martin E, et al. Screening for the Lynch syndrome (hereditary nonpolyposis colorectal cancer). N Engl J Med 2005;352:1851-60.
17. Svrcek M, Buhard O, Colas C, et al. Methylation tolerance due to an O6-methylguanine DNA methyltransferase (MGMT) field defect in the colonic mucosa: an initiating step in the development of mismatch repair-deficient colorectal cancers. Gut 2010;59:1516-26.
18. Svrcek M, El-Bchiri J, Chalastanis A, et al. Specific clinical and biological features characterize inflammatory bowel disease associated colorectal cancers showing microsatellite instability. J Clin Oncol 2007;25:4231-8.
19. Andre T, Boni C, Mounedji-Boudiaf L, et al. Oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment for colon cancer. N Engl J Med 2004;350:2343-51.
20. Weiss MM, Hermsen MA, Meijer GA, et al. Comparative genomic hybridisation. Mol Pathol 1999;52:243-51.
21. El-Bchiri J, Guilloux A, Dartigues P, et al. Nonsense-mediated mRNA decay impacts MSI-driven carcinogenesis and anti-tumor immunity in colorectal cancers. PLoS ONE 2008;3:e2583.
22. Ionov Y, Nowak N, Perucho M, et al. Manipulation of nonsense mediated decay identifies gene mutations in colon cancer Cells with microsatellite instability. Oncogene 2004;23:639-45.
23. Altman DG, Lausen B, Sauerbrei W, et al. Dangers of using "optimal" cutpoints in the evaluation of prognostic factors. J Natl Cancer Inst 1994;86:829-35.
24. Markowitz S, Wang J, Myeroff L, et al. Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. Science 1995;268:1336-8.
25. Colas C, Coulet F, Svrcek M, et al. Lynch or not Lynch? Is that always a question? Adv Cancer Res 2012;113:121-66.

ACCEPTED MANUSCRIPT

26. Gavin PG, Colangelo LH, Fumagalli D, et al. Mutation profiling and microsatellite instability in stage II and III colon cancer: an assessment of their prognostic and oxaliplatin predictive value. Clin Cancer Res 2012;18:6531-41.
27. Sinicrope FA, Foster NR, Thibodeau SN, et al. DNA mismatch repair status and colon cancer recurrence and survival in clinical trials of 5-fluorouracil-based adjuvant therapy. J Natl Cancer Inst 2011;103:863-75.
28. You JF, Buhard O, Ligtenberg MJ, et al. Tumours with loss of MSH6 expression are MSI-H when screened with a pentaplex of five mononucleotide repeats. Br J Cancer 2010;103:1840-5.
29. Kim JH, Kim KJ, Rhee YY, et al. Expression status of wild-type HSP110 correlates with HSP110 T deletion size and patient prognosis in microsatellite-unstable colorectal cancer. Mod Pathol 2013.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Figure Legends

Figure 1. (A) Modeling of HSP110 and the protein-protein interface of HSP110DE9/HSP110wt. The HSP110 structure and mutant HSP110DE9 were built by homology using the experimental structure of yeast Hsp110 Sse1 (Protein Data Bank code: 3C7N) as a template in the software MODELLER. Right panel, details of the interaction surface with the strongest interaction sites highlighted. **(B)** Characterization of the HSP110/HSP110DE9 protein-protein interaction by surface plasmon resonance. His-tagged HSP110 was immobilized with an NTA sensor chip and HSP110DE9 was injected in running buffer alone (red curve), with ADP (2mM, green curve), ATP (2mM, blue curve) or both ATP and ADP (black curve). **(C)** Distribution of *HSP110* T_{17} genotypes CRC cell line and HSP110DE9/HSP110wt mRNA ratio in 33 HCT116 sub-clones (top panel); western blotting analysis of 4 HCT116 sub-clones (bottom left panel), ratio of HSP110wt/HSP110DE9 expression was calculated in each case (bottom right panel). **(D)** Amplification plots corresponding to HSP110wt and HSP110DE9 RT-PCR products in HCT116 (small T_{17} status) and LS174T (large T_{17} status). Results are expressed ($E= 2^{-dCT}$) as n-fold difference in HSP110DE9 relative to HSP110wt expression (dCT), where dCT was determined by substrating the average CT value of the HSP110DE9 mRNA from the average CT value of the HSP110wt mRNA (top panel); western blotting analysis of HSP110wt and HSP110DE9 mutant proteins in 2 cell lines (HCT116 and LS147T; bottom left panel); ratio of HSP110wt/HSP110DE9 expression was calculated in each case (bottom right panel).

Figure 2. Fragment analysis of the intronic T_{17} in MSS and MSI primary colon tumors. (A) Distribution of *HSP110* T_{17} genotypes in 258 stage II and stage III MSS colorectal tumors (polymorphic zone). (**B)** Distribution of mutated *HSP110* T_{17} alleles in 329 stage II and stage III MSI CRCs. **(C)** Examples of 7 *HSP110* T_{17} mutated alleles. DT: size of T_{17} somatic deletions in tumor DNA (in bp). The polymorphic zone relates to the allelic variations observed for the *HSP110* T_{17} repeat in MSS patients. **(D)** Mutation of the intronic *HSP110* T_{17} and 15 other coding microsatellites contained in target genes for MSI in colon tumors. Although T_{17} deletions are observed in the great majority of MSI CRCs, 23% of them displayed large T_{17} deletions with a clinical impact in MSI CRC patients (*i.e.* response to chemotherapy). **(E)** Fragment analysis of *HSP110* T_{17} and *TGFBR2* A_{10} microsatellites profiles (left and right panel) of 3 primary CRCs before and after microdissection.

Figure 3. (A) Chemosensitivity assay using stably transfected MSI (HCT116) and MSS (SW480, FET) CRC cell lines that overexpressed HSP110wt or HSP110DE9 protein (left panel). The absorbance values (OD) obtained when plating the same number of untreated cells (SW480, FET, HCT116) that overexpressed either HSP110wt or HSP110DE9 are shown as controls (right panel) **(B)** Survival analysis of stage II and stage III MSI CRC patients treated with (left panel) or without (right panel) chemotherapy. Patients were classified into two groups according to the size of deletion in the T_{17} intronic repeat (DT ≥ 5 bp, T_{11}, T_{10}, T_9 for MSI *HSP110* Del^L patients; 0≤ DT < 5, T_{17} to T_{12} for MSI *HSP110* Del^S patients). **(C)** Survival analysis of stage II-III MSI colorectal cancer patients according to the size of deletions in the T_{17} DNA repeat and who were treated with 5-fluorouracil alone. Forty-two MSI stage II and III CRC patients received adjuvant chemotherapy with 5-FU alone (all data are shown in Supplementary Table S1. Other patients received 5-FU-based chemotherapy in which 5-FU was combined with oxaliplatin).

ACCEPTED MANUSCRIPT

Table 1. Description of MSI and MSS cohorts used in this work.

		MSI				MSS	
		Set 1	Set 2		Total	Total	*P*
		No. (%)	No. (%)	*P*	No. (%)	No. (%)	*Value*
		(N=98)	(N=231)	Value	(N=329)	(N=258)	MSS vs MSI
Sex							
	Female	48 (49)	149 (65)	.0099	197 (60)	137 (53)	.11
	Male	50 (51)	82 (35)		132 (40)	121 (47)	
Age at diagnostic	-	71	76	.023	75	73	.38
Tumor Stage							
	II	71 (73)	157 (68)	.36	229 (70)	187 (72)	.47
	III	26 (27)	74 (32)		100 (30)	71 (28)	
Tumor Location					(n=324)		
	Distal colon	25 (26)	45 (20)	.049	70 (22)	57 (22)	.91
	Proximal colon	67 (69)	179 (79)		246 (76)	196 (76)	
	Rectum	5 (5)	3 (1)		8 (2)	5 (2)	
Lynch syndrome					(n=166)		
	yes	41 (43)	3 (4)	.003	44 (27)	0 (0)	<.001
	no	54 (57)	68 (96)		122 (73)	258 (100)	
Chemotherapy performed							
	yes	23 (23)	54 (23)	.99	77 (23)	ND	NA
	no	75 (77)	177 (77)		252 (77)	ND	
Chemotherapy type					(n=318)		

ACCEPTED MANUSCRIPT

	FOLFOX	13 (13)	17 (8)		30 (9)	ND	
	FUFOL	0 (0)	2 (1)		2 (1)	ND	
	LV5FU2	10 (10)	32 (15)	.31	42 (13)	ND	NA
	None	75 (77)	169 (77)		244 (77)	ND	
	Others	0 (0)	0 (0)		0 (0)	ND	
Relapse							
	yes	11 (11)	60 (26)	.003	71 (22)	ND	NA
	no	87 (89)	171 (74)		258 (78)	ND	
Median Follow-up Time [IQR] months	-	42 [21-51]	60 [29-60]	.009	50 [24-60]	ND	NA
Clinical Center							
	Australia	0 (0)	44 (19)		44 (13)	0 (0)	
	Dijon	0 (0)	133 (49)		133 (34)	97 (38)	
	Nice	0 (0)	14 (6)	<.001	14 (4)	20 (8)	<.001
	Paris-SA	98 (100)	0 (0)		98 (30)	104 (40)	
	Singapore	0 (0)	50 (22)		50 (15)	35 (14)	
	Toulouse	0 (0)	10 (4)		10 (3)	2 (1)	

Abbreviations: MSI, Microsatellite instability; MSS, Microsatellite stable; FOLFOX, FOLinic acid, Fluorouracil and OXaliplatin regimen; LV5FU2, Fluorouracil and Leucovorin regimen; FUFOL, Fluorouracil and Folinic acid regimen.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 2. Association of clinical and molecular annotations to outcome (relapse-free survival) for patients under chemotherapy.

		COX UNIVARIATE ANALYSIS				COX MULTIVARIATE ANALYSIS[a]			
Variable	**Available data**			**modality**	**model**			**modality**	**model**
	n (n relapse)	**H.R.**	**(95%C.I.)**	**P value (Wald)**	**P value (Log-rank)**	**H.R.**	**(95%C.I.)**	**P value (Wald)**	**P value (Log-rank)**
HSP110Del (Large vs Small)	77 (20)	0.16	(0.02-1.2)	.073	.040	0.1	(0.012-0.8)	.03	
TNM Stage (III)	77 (20)	2.1	(0.66-6.5)	.21	.20	2.1	(0.54-8.1)	.29	
Chemotherapy type (LV5FU2)	72 (17)	1.8	(0.56-5.6)	.33	.32	1.1	(0.32-3.9)	.87	
Tumor location (Proximal colon)	77 (20)	2.3	(0.66-8.1)	.19	.40	3.5	(0.89-14)	.074	.15
Tumor location (Rectum)		1.6	(0.16-15)	.7		1.4	(0.14-14)	.78	
Gender (Male)	77 (20)	0.74	(0.28-1.9)	.53	.53	1.2	(0.33-4)	.82	
Age recoded (≥75y)	77 (20)	0.73	(0.16-3.4)	.69	.69	0.73	(0.15-3.6)	.7	
Lynch syndrome (Lynch)	27 (6)	0.61	(0.1-3.7)	.59	.59				

[a] Multivariate models included variables available for most samples. Therefore, the model was estimated on 72 patients (n relapse=17).

Abbreviations: H.R., Cox Hazard Ratio; 95% C.I., 95 Percent Confidence Interval of HR; LV5FU2, Fluorouracil and Leucovorin regimen.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 3. Associations of clinical annotations to *HSP110* deletion status in MSI colorectal cancer patients.

		***HSP110* Del^{L}**	***HSP110* Del^{S}**	***P* value**
		No. (%) (n=76)	**No. (%) (n=253)**	
Sex				
	Female	48 (63)	149 (59)	.43
	Male	28 (37)	104 (41)	
Age at diagnosis (median)	-	77.5	74	.12
Tumor stage				
	II	54 (71)	175 (69)	.89
	III	22 (29)	78 (31)	
Tumor location				
	Distal colon	10 (13)	60 (24)	
	Proximal colon	65 (87)	181 (73)	.026
	Rectum	0 (0)	8 (3)	
Relapse				
	yes	12 (16)	59 (23)	.37
	no	64 (84)	194 (77)	
Chemotherapy performed				
	yes	18 (24)	59 (23)	.97
	no	58 (76)	194 (77)	
Chemotherapy type				
	FOLFOX	7 (9)	23 (9)	
	FUFOL	0 (0)	2 (1)	.65

ACCEPTED MANUSCRIPT

	LV5FU2	11 (15)	31 (13)	
	None	56 (76)	188 (77)	
Lynch syndrome				
	yes	12 (27)	32 (26)	.52
	no	33 (73)	89 (74)	

Abbreviations: *HSP110* $Del^{L/S}$, Large/Small *HSP110* deletion (DT≥5bp/DT<5bp); FOLFOX, FOLinic acid, Fluorouracil and OXaliplatin regimen; LV5FU2, Fluorouracil and Leucovorin regimen; FUFOL, Fluorouracil and Folinic acid regimen.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Table 4. Target gene mutations detected in MSI colorectal cancer and their impact on patients' survival (RFS).

		Stage 2&3		Stage 2		Stage 3		Under chemotherapy	
Gene name	**Mutation frequencies (%)**	**n./n.mut**	**p-value (Log-Rank)**	**n./n.mut**	**p-value (Log-Rank)**	**n./n.mut**	**p-value (Log-Rank)**	**n./n.mut**	**p-value (Log-Rank)**
TGFBR2	**88%**	185/162	.22	130/111	.87	55/51	.25	46/42	.31
SLC35F5	**54%**	37/20	.98	24/14	.2	13-juin	.44	10-juin	.64
MSH3	**54%**	154/83	.7	109/57	.4	45/26	.14	36/21	.32
BAX	**44%**	154/67	.42	108/45	.27	46/22	.68	34/14	.36
GRK4	**41%**	163/67	.51	113/43	.74	50/24	.24	41/17	.24
RAD50	**39%**	185/72	.34	132/52	.61	53/20	.38	43/13	.78
ATR	**38%**	163/62	.25	113/42	.52	50/20	.34	38/14	.71
MBD4	**38%**	157/59	.17	112/45	.4	45/14	.076	35/13	.08
GRB14	**37%**	165/61	.53	116/43	.45	49/18	.75	40/15	.66
HSP110	**23%** [a]	329/76 [a]	.26	229/54 [a]	.59	100/22 [a]	.02	77/18 [a]	.04
MSH6	***23%***	174/40	.47	120/31	.87	54/9	.78	40/9	.078
BLM	***21%***	182/39	.58	127/26	.36	55/13	.99	43/9	.46
CDX2	**18%**	123/27	.29	110/20	.35	40/7	.6	34/4	.18
RECQL	**13%**	164/21	.64	119/16	.41	45/5	.16	37/7	.21
RIZ	**4%**	161/6	.014	117/2	.00092	44/4	.79	37/3	.89
TFDP2	**2%**	161/3	.57	116/1	.75	45/2	.55	34/1	.55

ACCEPTED MANUSCRIPT

[a] mutated samples are those displaying large deletions in the *HSP110* T17 (≥5 bp).
Abbreviations: n, total number of tumor samples analyzed in each case; n.mut, number of mutated samples.

ACCEPTED MANUSCRIPT

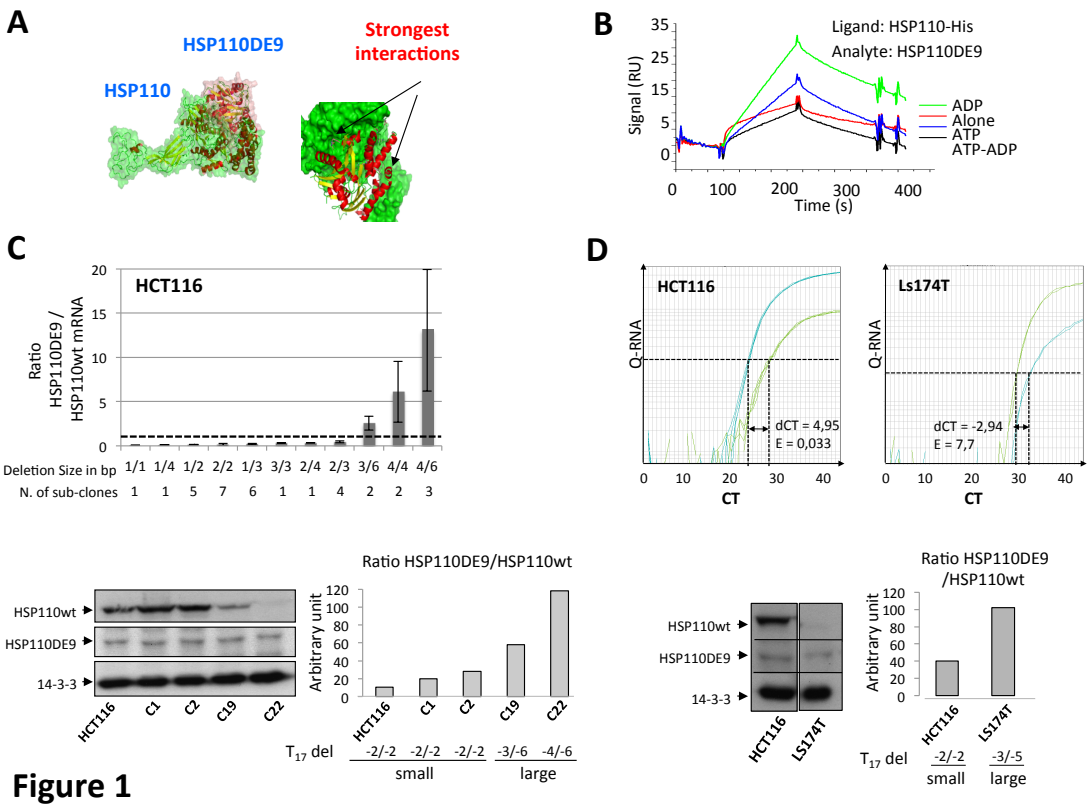

A
HSP110DE9
HSP110
Strongest interactions
B
Ligand: HSP110-His
Analyte: HSP110DE9
Signal (RU)
Time (s)
ADP
Alone
ATP
ATP-ADP
C
HCT116
Ratio HSP110DE9 / HSP110wt mRNA
Deletion Size in bp 1/1 1/4 1/2 2/2 1/3 3/3 2/4 2/3 3/6 4/4 4/6
N. of sub-clones 1 1 5 7 6 1 1 4 2 2 3
D
HCT116
Ls174T
Q-RNA
dCT = 4,95
E = 0,033
dCT = -2,94
E = 7,7
CT
HSP110wt
HSP110DE9
14-3-3
HCT116
C1
C2
C19
C22
Ratio HSP110DE9/HSP110wt
Arbitrary unit
T17 del
-2/-2 -2/-2 -2/-2 -3/-6 -4/-6
small
large
Ratio HSP110DE9 /HSP110wt
LS174T
-2/-2
-3/-5

Figure 1

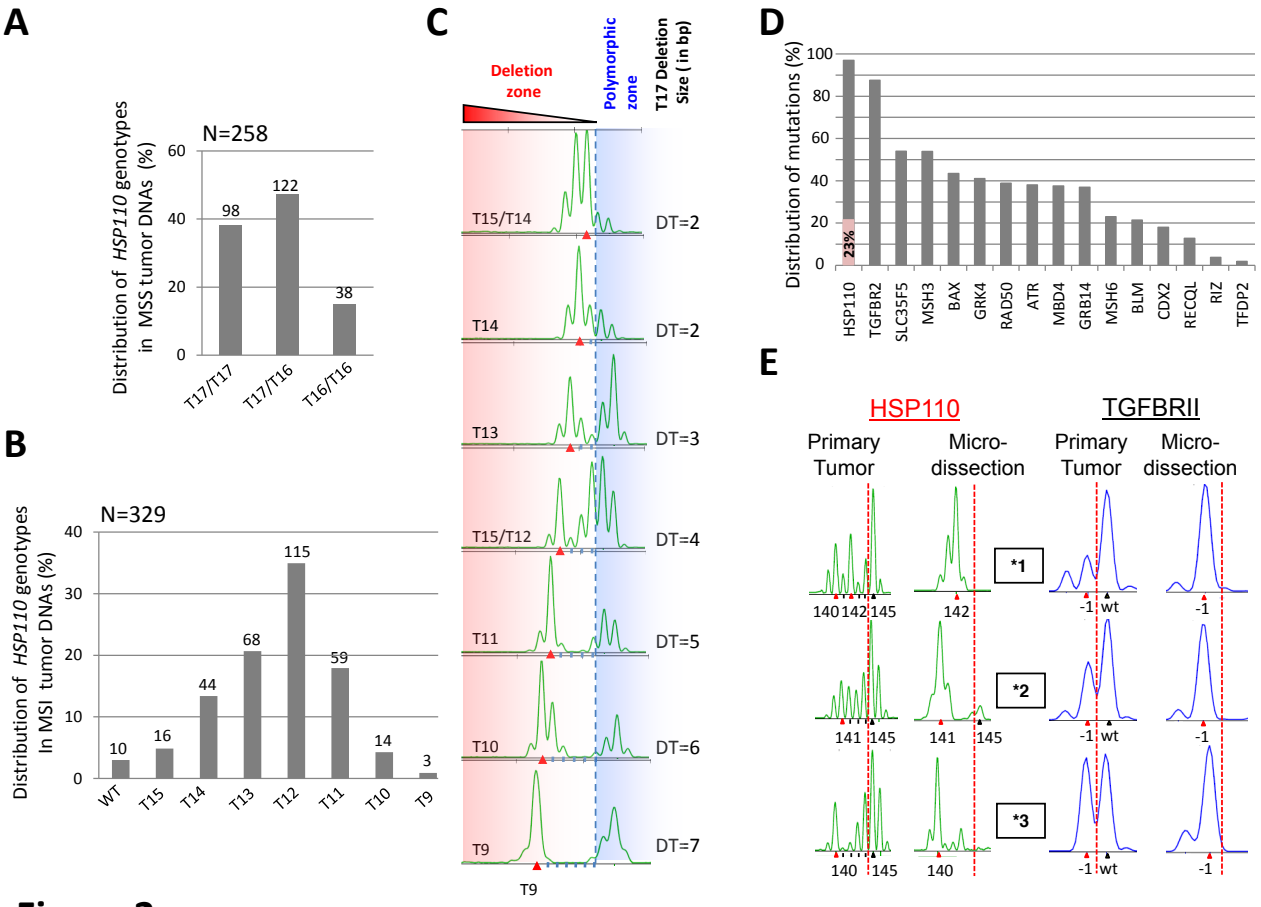

A
N=258
Distribution of HSP110 genotypes in MSS tumor DNAs (%)
98
122
38
T17/T17
T17/T16
T16/T16
B
N=329
Distribution of HSP110 genotypes In MSI tumor DNAs (%)
10
16
44
68
115
59
14
3
WT
T15
T14
T13
T12
T11
T10
T9
C
Deletion zone
Polymorphic zone
T17 Deletion Size (in bp)
T15/T14
DT=2
T14
DT=2
T13
DT=3
T15/T12
DT=4
T11
DT=5
T10
DT=6
T9
DT=7
T9
D
Distribution of mutations (%)
23%
HSP110
TGFBR2
SLC35F5
MSH3
BAX
GRK4
RAD50
ATR
MBD4
GRB14
MSH6
BLM
CDX2
RECQL
RIZ
TFDP2
E
HSP110
TGFBRII
Primary Tumor
Micro-dissection
Primary Tumor
Micro-dissection
*1
*2
*3
140 142 145
142
-1 wt
-1
141 145
141 145
-1 wt
-1
140 145
140
-1 wt
-1

Figure 2

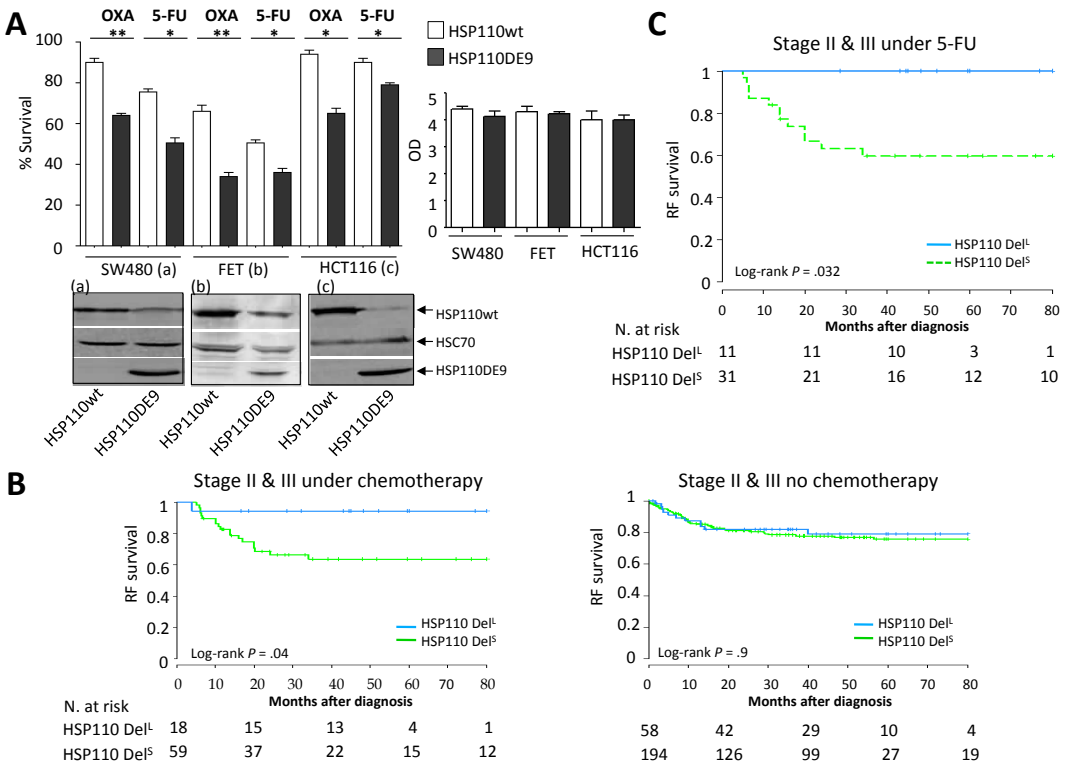

A
OXA
5-FU
OXA
5-FU
OXA
5-FU
HSP110wt
HSP110DE9
% Survival
OD
SW480 (a)
FET (b)
HCT116 (c)
SW480
FET
HCT116
(a)
(b)
(c)
HSP110wt
HSC70
HSP110DE9
C
Stage II & III under 5-FU
RF survival
Log-rank P = .032
HSP110 Del^L
HSP110 Del^S
Months after diagnosis
N. at risk
HSP110 Del^L 11 11 10 3 1
HSP110 Del^S 31 21 16 12 10
B
Stage II & III under chemotherapy
RF survival
Log-rank P = .04
HSP110 Del^L
HSP110 Del^S
Months after diagnosis
N. at risk
HSP110 Del^L 18 15 13 4 1
HSP110 Del^S 59 37 22 15 12
Stage II & III no chemotherapy
RF survival
Log-rank P = .9
HSP110 Del^L
HSP110 Del^S
Months after diagnosis
58 42 29 10 4
194 126 99 27 19

Figure 3

ACCEPTED MANUSCRIPT

Supplementary Figure Legends

Figure S1. (A) Fragment analysis of the intronic *HSP110* microsatellite in colon cancer cell lines. As expected, the T_{17} deletion status of HCT116 (DT=2, *e.g.* homozygous 2 bp deletion) was similar to the most frequent T_{17} genotype observed in this population (**Fig. 1C**). **(B)** Fragment analysis of 7 primary MSI tumor samples (*HSP110* T_{17}, *TGFBR2* A_{10} and 5 non-coding microsatellite DNA repeats from the pentaplex). * indicates the tumor samples that were further analyzed for fragment analysis using microdissection (see **Fig. 2E**). Assuming the existence of bi-allelic mutations in *TGFBR2* A_{10} and the pentaplex panel, samples 1-3 were found to be weakly contaminated with stromal cells and to display bi-allelic mutations in *HSP110* T_{17}. Samples 4-7 were highly contaminated with non-tumor cells and displayed wild type T_{17} alleles as a result of this contamination. **(C)** Additional molecular profiles (n=24) are presented. In each case, results before and after microdissection are shown, highlighting that *HSP110* T17 deletions are usually bi-allelic in MSI CRCs. Expectedly, some differences were observed in few samples after microdissection: this is due to the fact that we amplified a population of tumor cells that was probably poorly represented in the primary tumor sample. **(D)** Sixteen primary tumor samples that contained low levels of contamination with normal cells were also found to display bi-allelic deletions in *HSP110* T17, even without microdissection (3 of these 16 cases were microdissected as positive controls. They are indicated with an asterisk *).

Figure S2. (A) Deletion cut-off analysis for stage II-III MSI patients treated with chemotherapy. **(B)** Deletion cut-off analysis for stage II-III MSI patients who did not receive chemotherapy. Using a bootstrap approach, 100 subsamples containing 85% of patients were randomly selected and the best cut-off for each of these subsamples was 5 bp in 99% of the subsamples of patients tested. The same threshold value was found when considering all patients (data not shown). The 329 MSI CRC patients were therefore classified into two groups displaying large deletions (DT ≥ 5 bp; *HSP110* Del^L; n=76/329, 23%) or small deletions (0 ≤ DT < 5; *HSP110* Del^S; n=253/329, 77%).

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Supplementary Table S1. Details of the patient characteristics and their tumors.

Sample Name	Clinical Center	MSI Status	Deletion HSP110	Deletion HSP110 Class	AJCC Staging	Tumor Location	Gender	Age at diagnosis	Lynch Syndrome *	RFS event	RFS delay (in months)	Chemotherapy Adj Performed	5-Fluorouracil-based Chemotherapy
aus001	Australie	MSI	3	Small	2	RC	F	82.59	NA	0	51.29	N	None
aus002	Australie	MSI	3	Small	3	RC	M	71.53	NA	0	50.04	N	None
aus003	Australie	MSI	4	Small	3	NA	F	81.48	NA	0	48.53	N	None
aus004	Australie	MSI	5	Large	2	RC	F	60.6	NA	0	48.2	Y	LV5FU2
aus005	Australie	MSI	5	Large	2	RC	F	67.39	NA	0	44.58	Y	LV5FU2
aus006	Australie	MSI	4	Small	2	RC	F	66.67	NA	0	10.32	N	None
aus007	Australie	MSI	4	Small	3	LC	F	82.93	NA	1	16.34	N	None
aus008	Australie	MSI	4	Small	2	RC	F	50.36	NA	0	41.85	Y	LV5FU2
aus009	Australie	MSI	4	Small	2	RC	F	72.62	NA	1	1.25	N	None
aus010	Australie	MSI	2	Small	3	NA	F	81.82	NA	0	38.86	N	None
aus011	Australie	MSI	1	Small	2	RC	M	58.56	NA	0	37.91	N	None
aus012	Australie	MSI	6	Large	2	RC	F	82.47	NA	0	35.57	N	None
aus013	Australie	MSI	5	Large	3	RC	F	90.86	NA	0	34.75	N	None
aus014	Australie	MSI	3	Small	3	NA	F	89.47	NA	1	2.24	N	None
aus015	Australie	MSI	2	Small	2	RC	F	79.03	NA	0	34.42	N	None
aus016	Australie	MSI	4	Small	2	RC	F	57.02	NA	0	33.04	Y	FOLFOX
aus017	Australie	MSI	3	Small	3	LC	F	89.01	NA	0	32.88	N	None
aus018	Australie	MSI	4	Small	2	RC	M	41.61	NA	0	32.15	N	None
aus019	Australie	MSI	6	Large	2	RC	M	81.3	NA	0	29.56	N	None
aus020	Australie	MSI	5	Large	3	RC	F	80.39	NA	0	28.93	N	None
aus021	Australie	MSI	3	Small	3	RC	M	42.78	NA	0	28.9	Y	FOLFOX
aus022	Australie	MSI	5	Large	2	RC	F	72.23	NA	0	28.5	Y	LV5FU2
aus023	Australie	MSI	4	Small	2	RC	F	74.99	NA	0	27.58	N	None
aus024	Australie	MSI	6	Large	3	RC	F	90.72	NA	0	26.66	N	None
aus025	Australie	MSI	3	Small	2	RC	M	74.8	NA	0	25.68	N	None
aus026	Australie	MSI	3	Small	3	RC	F	81.25	NA	0	24.07	N	None
aus027	Australie	MSI	3	Small	3	LC	M	50.47	NA	0	23.93	Y	FOLFOX
aus028	Australie	MSI	3	Small	2	RC	M	81.43	NA	0	21.99	N	None
aus029	Australie	MSI	4	Small	2	NA	M	82.55	NA	0	21.76	N	None
aus030	Australie	MSI	3	Small	2	RC	M	61.35	NA	0	19	N	None
aus031	Australie	MSI	3	Small	2	RC	M	63.28	NA	0	17.95	N	None
aus032	Australie	MSI	5	Large	3	RC	F	64.65	NA	0	16.57	Y	FOLFOX
aus033	Australie	MSI	4	Small	2	RC	M	80.36	NA	0	15.78	N	None
aus034	Australie	MSI	2	Small	2	RC	F	80.92	NA	0	15.55	N	None
aus035	Australie	MSI	4	Small	2	RC	F	84.06	NA	0	7.5	N	None
aus036	Australie	MSI	2	Small	3	RC	F	74.47	NA	1	5.79	Y	NA
aus037	Australie	MSI	3	Small	2	RC	F	87.33	NA	0	13.41	N	None
aus038	Australie	MSI	4	Small	2	LC	F	66.81	NA	0	14.17	Y	LV5FU2

ACCEPTED MANUSCRIPT

aus039	Australie	MSI	4	Small	2	RC	M	63.15	NA	0	12.16	Y	LV5FU2
aus040	Australie	MSI	2	Small	3	RC	M	63.78	NA	0	11.9	Y	FOLFOX
aus041	Australie	MSI	2	Small	2	RC	F	59.2	NA	0	9.21	N	None
aus042	Australie	MSI	1	Small	2	RC	M	84.44	NA	0	8.88	N	None
aus043	Australie	MSI	3	Small	2	LC	F	71.87	NA	0	7.5	N	None
aus044	Australie	MSI	4	Small	2	RC	F	62	NA	0	6.81	N	None
dij001	Dijon	MSI	2	Small	2	RC	F	73	NA	0	59.51	N	None
dij002	Dijon	MSI	4	Small	2	RC	F	66	none	1	14.85	N	None
dij003	Dijon	MSI	5	Large	2	RC	F	85	none	0	59.51	N	None
dij004	Dijon	MSI	4	Small	2	RC	M	77	none	1	8.18	N	None
dij005	Dijon	MSI	4	Small	2	RC	F	93	NA	0	59.51	N	None
dij006	Dijon	MSI	4	Small	2	RC	M	80	none	1	46.49	N	None
dij007	Dijon	MSI	3	Small	2	LC	F	66	NA	1	20.11	Y	FOLFOX
dij008	Dijon	MSI	5	Large	3	RC	F	68	NA	0	59.51	Y	LV5FU2
dij009	Dijon	MSI	4	Small	2	RC	F	90	none	0	59.51	N	None
dij010	Dijon	MSI	5	Large	2	RC	F	83	none	0	59.51	N	None
dij011	Dijon	MSI	3	Small	2	LC	M	53	none	0	59.51	Y	FOLFOX
dij012	Dijon	MSI	1	Small	2	RC	F	82	NA	0	53.98	N	None
dij013	Dijon	MSI	2	Small	2	RC	F	86	none	0	59.51	N	None
dij014	Dijon	MSI	3	Small	3	RC	F	90	none	1	2.56	N	None
dij015	Dijon	MSI	4	Small	2	RC	M	57	NA	1	13.86	Y	LV5FU2
dij016	Dijon	MSI	4	Small	2	RC	M	76	none	1	16.07	N	None
dij017	Dijon	MSI	4	Small	2	RC	F	88	NA	1	12.32	N	None
dij018	Dijon	MSI	2	Small	2	RC	F	83	none	0	59.51	N	None
dij019	Dijon	MSI	2	Small	2	LC	F	53	none	0	59.51	N	None
dij020	Dijon	MSI	2	Small	2	RC	M	85	none	0	59.51	N	None
dij021	Dijon	MSI	3	Small	2	RC	M	79	none	1	9.72	N	None
dij022	Dijon	MSI	4	Small	2	RC	M	72	none	0	59.51	N	None
dij023	Dijon	MSI	3	Small	3	RC	F	80	none	1	10.12	N	None
dij024	Dijon	MSI	3	Small	3	RC	F	89	none	1	9.3	N	None
dij025	Dijon	MSI	4	Small	3	RC	F	75	NA	1	2.17	N	None
dij026	Dijon	MSI	5	Large	2	RC	F	86	NA	0	59.51	N	None
dij027	Dijon	MSI	0	Small	2	RC	M	60	NA	0	59.51	N	None
dij028	Dijon	MSI	6	Large	2	RC	F	93	NA	1	1.71	N	None
dij029	Dijon	MSI	0	Small	3	RC	F	81	NA	0	59.51	N	None
dij030	Dijon	MSI	3	Small	3	RC	F	67	NA	0	6.83	Y	FOLFOX
dij031	Dijon	MSI	4	Small	2	RC	F	80	NA	0	59.51	N	None
dij032	Dijon	MSI	4	Small	2	RC	M	77	NA	0	59.51	N	None
dij033	Dijon	MSI	5	Large	2	RC	F	82	NA	0	59.51	N	None
dij034	Dijon	MSI	3	Small	3	RC	F	65	NA	1	10.84	Y	FOLFOX
dij035	Dijon	MSI	4	Small	3	RC	F	77	NA	0	59.51	N	None
dij036	Dijon	MSI	0	Small	2	RC	M	57	NA	1	56.51	N	None

ACCEPTED MANUSCRIPT

dij037	Dijon	MSI	5	Large	2	RC	F	68	NA	0	59.51	Y	FOLFOX
dij038	Dijon	MSI	3	Small	2	RC	F	72	NA	0	59.51	N	None
dij039	Dijon	MSI	3	Small	2	RC	F	79	NA	0	2.14	N	None
dij040	Dijon	MSI	5	Large	3	RC	F	61	NA	0	59.51	Y	FOLFOX
dij041	Dijon	MSI	4	Small	2	LC	M	77	NA	1	25.3	N	None
dij042	Dijon	MSI	3	Small	2	LC	F	86	NA	0	0.39	N	None
dij043	Dijon	MSI	1	Small	2	RC	F	80	NA	0	59.51	N	None
dij044	Dijon	MSI	5	Large	2	RC	F	65	NA	0	59.51	N	None
dij045	Dijon	MSI	4	Small	2	LC	F	44	NA	0	59.51	Y	FOLFOX
dij046	Dijon	MSI	5	Large	2	RC	M	85	NA	0	59.51	N	None
dij047	Dijon	MSI	2	Small	2	RC	F	78	NA	0	59.51	N	None
dij048	Dijon	MSI	4	Small	2	RC	F	95	NA	1	6.21	N	None
dij049	Dijon	MSI	4	Small	2	RC	F	77	NA	1	36.9	N	None
dij050	Dijon	MSI	5	Large	2	RC	F	82	NA	0	59.51	N	None
dij051	Dijon	MSI	2	Small	2	RC	M	88	NA	1	0.46	N	None
dij052	Dijon	MSI	5	Large	2	RC	F	73	NA	1	3.19	N	None
dij053	Dijon	MSI	5	Large	3	RC	F	82	NA	1	3.91	Y	FOLFOX
dij054	Dijon	MSI	1	Small	2	RC	M	84	NA	0	54.6	N	None
dij055	Dijon	MSI	4	Small	2	RC	F	71	NA	0	11.73	Y	FOLFOX
dij056	Dijon	MSI	4	Small	2	RC	F	92	NA	0	0	N	None
dij057	Dijon	MSI	2	Small	3	LC	M	64	NA	0	11.76	Y	FOLFOX
dij058	Dijon	MSI	4	Small	3	RC	F	86	NA	0	1.77	N	None
dij059	Dijon	MSI	4	Small	2	RECTUM	M	49	NA	0	51.52	Y	FOLFOX
dij060	Dijon	MSI	5	Large	3	RC	M	83	NA	0	18.53	Y	FOLFOX
dij061	Dijon	MSI	3	Small	2	RC	F	67	NA	0	23.75	N	None
dij062	Dijon	MSI	4	Small	2	RC	M	83	NA	0	14.19	N	None
dij063	Dijon	MSI	5	Large	2	RC	F	78	NA	0	14.39	N	None
dij064	Dijon	MSI	4	Small	2	RC	F	52	NA	1	11.27	Y	LV5FU2
dij065	Dijon	MSI	3	Small	3	RC	F	82	none	1	4.7	N	None
dij066	Dijon	MSI	3	Small	3	LC	F	60	none	0	59.51	Y	LV5FU2
dij067	Dijon	MSI	4	Small	2	RC	M	88	none	0	59.51	N	None
dij068	Dijon	MSI	4	Small	2	RC	F	91	none	1	19.29	N	None
dij069	Dijon	MSI	3	Small	2	RC	F	85	none	0	59.51	N	None
dij070	Dijon	MSI	6	Large	3	RC	F	87	none	1	9.2	N	None
dij071	Dijon	MSI	2	Small	2	RC	F	85	none	1	8.54	N	None
dij072	Dijon	MSI	2	Small	3	RC	F	79	none	1	10.35	N	None
dij073	Dijon	MSI	5	Large	2	RC	M	80	none	1	3.29	N	None
dij074	Dijon	MSI	2	Small	2	RC	F	77	none	0	59.51	N	None
dij075	Dijon	MSI	5	Large	2	RC	F	75	none	0	59.51	N	None
dij076	Dijon	MSI	2	Small	3	RC	F	88	none	0	59.51	N	None
dij077	Dijon	MSI	4	Small	2	RC	F	88	none	0	59.51	N	None
dij078	Dijon	MSI	0	Small	2	RC	F	81	NA	0	59.51	N	None

ACCEPTED MANUSCRIPT

dij079	Dijon	MSI	3	Small	2	RC	F	69	none	0	59.51	N	None
dij080	Dijon	MSI	4	Small	2	RC	F	83	none	0	59.51	N	None
dij081	Dijon	MSI	2	Small	2	RC	F	79	none	0	59.51	N	None
dij082	Dijon	MSI	2	Small	3	RC	F	89	none	1	6.87	N	None
dij083	Dijon	MSI	3	Small	2	RC	M	82	none	0	59.51	N	None
dij084	Dijon	MSI	3	Small	2	RC	F	70	none	1	5.91	N	None
dij085	Dijon	MSI	4	Small	3	RC	M	82	none	0	59.51	N	None
dij086	Dijon	MSI	3	Small	2	RC	M	64	none	0	59.51	N	None
dij087	Dijon	MSI	3	Small	2	LC	M	63	NA	0	59.51	N	None
dij088	Dijon	MSI	4	Small	2	LC	M	46	LYNCH	0	59.51	N	None
dij089	Dijon	MSI	2	Small	3	RC	F	80	none	1	7.92	N	None
dij090	Dijon	MSI	5	Large	2	RC	F	83	none	1	6.8	N	None
dij091	Dijon	MSI	5	Large	2	RC	F	77	none	0	59.51	N	None
dij092	Dijon	MSI	4	Small	3	RC	F	85	none	0	59.51	N	None
dij093	Dijon	MSI	2	Small	3	RC	F	64	NA	0	59.51	Y	LV5FU2
dij094	Dijon	MSI	4	Small	2	RC	F	90	NA	0	59.51	N	None
dij095	Dijon	MSI	3	Small	2	RC	F	65	NA	0	59.51	N	None
dij096	Dijon	MSI	3	Small	3	LC	F	13	NA	0	59.51	N	None
dij097	Dijon	MSI	5	Large	3	RC	M	73	NA	0	59.51	N	None
dij098	Dijon	MSI	4	Small	3	RC	F	74	NA	1	6.47	Y	LV5FU2
dij099	Dijon	MSI	4	Small	3	RC	F	80	NA	0	2.53	N	None
dij100	Dijon	MSI	4	Small	3	LC	M	75	NA	0	59.51	N	None
dij101	Dijon	MSI	4	Small	2	RC	F	94	NA	1	0.26	N	None
dij102	Dijon	MSI	4	Small	3	LC	M	76	NA	1	8.31	N	None
dij103	Dijon	MSI	6	Large	3	RC	F	82	NA	0	59.51	Y	LV5FU2
dij104	Dijon	MSI	4	Small	2	RC	F	76	NA	0	55.49	N	None
dij105	Dijon	MSI	3	Small	3	RC	F	83	NA	1	13.86	Y	LV5FU2
dij106	Dijon	MSI	4	Small	2	RC	F	81	NA	1	0.26	N	None
dij107	Dijon	MSI	4	Small	2	RC	F	86	NA	0	1.35	N	None
dij108	Dijon	MSI	4	Small	2	LC	M	74	NA	0	37.85	N	None
dij109	Dijon	MSI	4	Small	2	RC	M	81	none	0	59.51	N	None
dij110	Dijon	MSI	4	Small	2	RC	M	52	none	0	59.51	N	None
dij111	Dijon	MSI	6	Large	2	RC	M	75	none	0	59.51	N	None
dij112	Dijon	MSI	4	Small	2	RC	M	67	none	1	28.91	N	None
dij113	Dijon	MSI	5	Large	2	RC	M	88	none	1	3.61	N	None
dij114	Dijon	MSS	0	Small	3	LC	M	60	NA	NA	NA	NA	NA
dij115	Dijon	MSS	0	Small	2	RC	M	72	NA	NA	NA	NA	NA
dij116	Dijon	MSS	0	Small	2	RC	F	72	NA	NA	NA	NA	NA
dij117	Dijon	MSS	0	Small	3	RC	F	86	NA	NA	NA	NA	NA
dij118	Dijon	MSS	0	Small	3	RC	F	70	NA	NA	NA	NA	NA
dij119	Dijon	MSS	0	Small	3	RC	F	82	NA	NA	NA	NA	NA
dij120	Dijon	MSS	0	Small	2	RC	F	71	NA	NA	NA	NA	NA

ACCEPTED MANUSCRIPT

dij121	Dijon	MSS	0	Small	3	LC	M	71	NA	NA	NA	NA	NA
dij122	Dijon	MSS	0	Small	2	RC	F	73	NA	NA	NA	NA	NA
dij123	Dijon	MSS	0	Small	3	RC	F	57	NA	NA	NA	NA	NA
dij124	Dijon	MSS	0	Small	2	RC	F	76	NA	NA	NA	NA	NA
dij125	Dijon	MSS	0	Small	2	RC	F	72	NA	NA	NA	NA	NA
dij126	Dijon	MSS	0	Small	2	RC	F	93	NA	NA	NA	NA	NA
dij127	Dijon	MSS	0	Small	2	LC	F	50	NA	NA	NA	NA	NA
dij128	Dijon	MSS	0	Small	2	RC	M	79	NA	NA	NA	NA	NA
dij129	Dijon	MSS	0	Small	3	RC	F	90	NA	NA	NA	NA	NA
dij130	Dijon	MSS	0	Small	2	RECTUM	M	72	NA	NA	NA	NA	NA
dij131	Dijon	MSS	0	Small	2	LC	F	63	NA	NA	NA	NA	NA
dij132	Dijon	MSS	0	Small	3	RC	F	75	NA	NA	NA	NA	NA
dij133	Dijon	MSS	0	Small	3	RC	F	64	NA	NA	NA	NA	NA
dij134	Dijon	MSS	0	Small	2	RC	F	80	NA	NA	NA	NA	NA
dij135	Dijon	MSS	0	Small	3	RC	F	60	NA	NA	NA	NA	NA
dij136	Dijon	MSS	0	Small	2	RC	M	72	NA	NA	NA	NA	NA
dij137	Dijon	MSS	0	Small	3	RC	F	75	NA	NA	NA	NA	NA
dij138	Dijon	MSS	0	Small	3	LC	F	55	NA	NA	NA	NA	NA
dij139	Dijon	MSS	0	Small	2	RC	M	62	NA	NA	NA	NA	NA
dij140	Dijon	MSS	0	Small	3	RC	M	77	NA	NA	NA	NA	NA
dij141	Dijon	MSS	0	Small	2	RC	M	71	NA	NA	NA	NA	NA
dij142	Dijon	MSS	0	Small	2	RC	F	63	NA	NA	NA	NA	NA
dij143	Dijon	MSS	0	Small	3	RC	M	78	NA	NA	NA	NA	NA
dij144	Dijon	MSS	0	Small	3	RC	F	79	NA	NA	NA	NA	NA
dij145	Dijon	MSS	0	Small	3	RC	F	75	NA	NA	NA	NA	NA
dij146	Dijon	MSS	0	Small	2	RC	F	95	NA	NA	NA	NA	NA
dij147	Dijon	MSS	0	Small	2	RC	M	76	NA	NA	NA	NA	NA
dij148	Dijon	MSS	0	Small	2	RC	M	77	NA	NA	NA	NA	NA
dij149	Dijon	MSS	0	Small	3	RC	F	80	NA	NA	NA	NA	NA
dij150	Dijon	MSS	0	Small	3	RC	F	76	NA	NA	NA	NA	NA
dij151	Dijon	MSS	0	Small	2	RC	F	73	NA	NA	NA	NA	NA
dij152	Dijon	MSS	0	Small	3	RC	F	79	NA	NA	NA	NA	NA
dij153	Dijon	MSS	0	Small	2	RC	F	70	NA	NA	NA	NA	NA
dij154	Dijon	MSS	0	Small	2	LC	M	46	NA	NA	NA	NA	NA
dij155	Dijon	MSS	0	Small	3	RC	F	84	NA	NA	NA	NA	NA
dij156	Dijon	MSS	0	Small	2	RC	F	94	NA	NA	NA	NA	NA
dij157	Dijon	MSS	0	Small	2	RC	F	91	NA	NA	NA	NA	NA
dij158	Dijon	MSS	0	Small	3	RC	F	73	NA	NA	NA	NA	NA
dij159	Dijon	MSS	0	Small	2	LC	M	58	NA	NA	NA	NA	NA
dij160	Dijon	MSS	0	Small	2	RC	M	60	NA	NA	NA	NA	NA
dij161	Dijon	MSS	0	Small	3	RC	F	74	NA	NA	NA	NA	NA
dij162	Dijon	MSS	0	Small	2	RC	F	76	NA	NA	NA	NA	NA

ACCEPTED MANUSCRIPT

dij163	Dijon	MSS	0	Small	3	RC	F	85	NA	NA	NA	NA	NA
dij164	Dijon	MSS	0	Small	3	RC	F	79	NA	NA	NA	NA	NA
dij165	Dijon	MSS	0	Small	2	LC	F	44	NA	NA	NA	NA	NA
dij166	Dijon	MSS	0	Small	2	RC	F	92	NA	NA	NA	NA	NA
dij167	Dijon	MSS	0	Small	2	RC	F	70	NA	NA	NA	NA	NA
dij168	Dijon	MSS	0	Small	2	RC	M	77	NA	NA	NA	NA	NA
dij169	Dijon	MSS	0	Small	2	RC	F	77	NA	NA	NA	NA	NA
dij170	Dijon	MSS	0	Small	2	RC	F	90	NA	NA	NA	NA	NA
dij171	Dijon	MSS	0	Small	2	LC	F	81	NA	NA	NA	NA	NA
dij172	Dijon	MSS	0	Small	2	RC	M	81	NA	NA	NA	NA	NA
dij173	Dijon	MSS	0	Small	3	RC	M	68	NA	NA	NA	NA	NA
dij174	Dijon	MSS	0	Small	2	RC	M	83	NA	NA	NA	NA	NA
dij175	Dijon	MSS	0	Small	2	RC	F	64	NA	NA	NA	NA	NA
dij176	Dijon	MSS	0	Small	2	RC	M	71	NA	NA	NA	NA	NA
dij177	Dijon	MSS	0	Small	2	RC	M	70	NA	NA	NA	NA	NA
dij178	Dijon	MSS	0	Small	2	RC	M	80	NA	NA	NA	NA	NA
dij179	Dijon	MSS	0	Small	2	RC	F	87	NA	NA	NA	NA	NA
dij180	Dijon	MSS	0	Small	2	RC	F	76	NA	NA	NA	NA	NA
dij181	Dijon	MSS	0	Small	2	RC	M	84	NA	NA	NA	NA	NA
dij182	Dijon	MSS	0	Small	2	RC	F	80	NA	NA	NA	NA	NA
dij183	Dijon	MSS	0	Small	2	RC	F	86	NA	NA	NA	NA	NA
dij184	Dijon	MSS	0	Small	2	RC	F	77	NA	NA	NA	NA	NA
dij185	Dijon	MSS	0	Small	3	LC	M	70	NA	NA	NA	NA	NA
dij186	Dijon	MSS	0	Small	2	LC	M	48	NA	NA	NA	NA	NA
dij187	Dijon	MSS	0	Small	3	RC	F	74	NA	NA	NA	NA	NA
dij188	Dijon	MSS	0	Small	3	RC	F	88	NA	NA	NA	NA	NA
dij189	Dijon	MSS	0	Small	2	RC	M	75	NA	NA	NA	NA	NA
dij190	Dijon	MSS	0	Small	2	RC	F	78	NA	NA	NA	NA	NA
dij191	Dijon	MSS	0	Small	2	RC	F	82	NA	NA	NA	NA	NA
dij192	Dijon	MSS	0	Small	2	RC	F	61	NA	NA	NA	NA	NA
dij193	Dijon	MSS	0	Small	2	RC	M	52	NA	NA	NA	NA	NA
dij194	Dijon	MSS	0	Small	3	RC	F	71	NA	NA	NA	NA	NA
dij195	Dijon	MSS	0	Small	2	RC	F	74	NA	NA	NA	NA	NA
dij196	Dijon	MSS	0	Small	2	RC	F	64	NA	NA	NA	NA	NA
dij197	Dijon	MSS	0	Small	2	RC	F	52	NA	NA	NA	NA	NA
dij198	Dijon	MSS	0	Small	3	RC	F	65	NA	NA	NA	NA	NA
dij199	Dijon	MSS	0	Small	2	RC	F	76	NA	NA	NA	NA	NA
dij200	Dijon	MSS	0	Small	2	RC	M	56	NA	NA	NA	NA	NA
dij201	Dijon	MSS	0	Small	2	RC	M	48	NA	NA	NA	NA	NA
dij202	Dijon	MSS	0	Small	2	RC	F	73	NA	NA	NA	NA	NA
dij203	Dijon	MSS	0	Small	2	RC	F	81	NA	NA	NA	NA	NA
dij204	Dijon	MSS	0	Small	2	RC	F	87	NA	NA	NA	NA	NA

ACCEPTED MANUSCRIPT

dij205	Dijon	MSS	0	Small	2	RC	M	57	NA	NA	NA	NA	NA
dij206	Dijon	MSS	0	Small	2	RC	M	75	NA	NA	NA	NA	NA
dij207	Dijon	MSS	0	Small	2	RC	F	79	NA	NA	NA	NA	NA
dij208	Dijon	MSS	0	Small	2	RC	M	68	NA	NA	NA	NA	NA
dij209	Dijon	MSS	0	Small	3	RC	F	77	NA	NA	NA	NA	NA
dij210	Dijon	MSS	0	Small	2	RECTUM	F	78	NA	NA	NA	NA	NA
nic001	Nice	MSI	4	Small	3	RC	F	83	none	0	42	N	None
nic002	Nice	MSI	2	Small	2	RC	F	46	none	0	133	Y	FUFOL
nic003	Nice	MSI	2	Small	3	RC	M	82	none	0	72	Y	FUFOL
nic004	Nice	MSI	4	Small	2	RC	M	65	none	0	17	N	None
nic005	Nice	MSI	3	Small	3	RC	F	82	none	0	31	N	None
nic006	Nice	MSI	4	Small	3	RC	F	86	none	0	46	N	None
nic007	Nice	MSI	3	Small	2	RC	F	77	none	0	83	N	None
nic008	Nice	MSI	5	Large	2	RC	M	85	none	1	5	N	None
nic009	Nice	MSI	3	Small	2	RC	M	50	none	0	85	Y	LV5FU2
nic010	Nice	MSI	3	Small	2	LC	F	72	none	0	87	N	None
nic011	Nice	MSI	5	Large	3	RC	F	79	none	0	77	Y	LV5FU2
nic012	Nice	MSI	4	Small	3	RC	F	80	none	0	41	N	None
nic013	Nice	MSI	5	Large	3	LC	F	82	none	0	72	N	None
nic014	Nice	MSI	2	Small	2	LC	F	43	none	0	57	N	None
nic015	Nice	MSS	0	Small	2	LC	M	68	NA	NA	NA	NA	NA
nic016	Nice	MSS	0	Small	3	LC	M	74	NA	NA	NA	NA	NA
nic017	Nice	MSS	0	Small	2	LC	M	74	NA	NA	NA	NA	NA
nic018	Nice	MSS	0	Small	3	LC	M	75	NA	NA	NA	NA	NA
nic019	Nice	MSS	0	Small	3	LC	F	74	NA	NA	NA	NA	NA
nic020	Nice	MSS	0	Small	2	LC	M	73	NA	NA	NA	NA	NA
nic021	Nice	MSS	0	Small	2	LC	M	71	NA	NA	NA	NA	NA
nic022	Nice	MSS	0	Small	2	LC	M	73	NA	NA	NA	NA	NA
nic023	Nice	MSS	0	Small	2	LC	M	69	NA	NA	NA	NA	NA
nic024	Nice	MSS	0	Small	3	LC	M	73	NA	NA	NA	NA	NA
nic025	Nice	MSS	0	Small	2	LC	M	89	NA	NA	NA	NA	NA
nic026	Nice	MSS	0	Small	2	LC	M	73	NA	NA	NA	NA	NA
nic027	Nice	MSS	0	Small	2	LC	F	72	NA	NA	NA	NA	NA
nic028	Nice	MSS	0	Small	2	LC	F	68	NA	NA	NA	NA	NA
nic029	Nice	MSS	0	Small	2	LC	F	70	NA	NA	NA	NA	NA
nic030	Nice	MSS	0	Small	2	LC	M	68	NA	NA	NA	NA	NA
nic031	Nice	MSS	0	Small	2	LC	F	73	NA	NA	NA	NA	NA
nic032	Nice	MSS	0	Small	2	LC	F	88	NA	NA	NA	NA	NA
nic033	Nice	MSS	0	Small	3	LC	M	74	NA	NA	NA	NA	NA
nic034	Nice	MSS	0	Small	2	LC	M	74	NA	NA	NA	NA	NA
psa001	Paris-SA	MSI	4	Small	3	RC	M	81	none	0	1	N	None
psa002	Paris-SA	MSI	3	Small	2	RC	F	75	none	0	65	N	None

ACCEPTED MANUSCRIPT

psa003	Paris-SA	MSI	4	Small	2	RC	M	88	none	0	2	N	None
psa004	Paris-SA	MSI	4	Small	2	RC	F	69	none	0	57	N	None
psa005	Paris-SA	MSI	5	Large	2	RC	M	78	none	0	33	N	None
psa006	Paris-SA	MSI	4	Small	3	RC	M	74	none	1	20	Y	LV5FU2
psa007	Paris-SA	MSI	3	Small	2	RC	M	57	LYNCH	0	20	N	None
psa008	Paris-SA	MSI	3	Small	2	LC	F	81	LYNCH	0	0	N	None
psa009	Paris-SA	MSI	4	Small	2	LC	M	49	LYNCH	0	48	N	None
psa010	Paris-SA	MSI	2	Small	2	LC	M	61	LYNCH	0	17	N	None
psa011	Paris-SA	MSI	4	Small	3	RC	F	79	none	0	7	N	None
psa012	Paris-SA	MSI	5	Large	2	LC	F	88	none	0	5	N	None
psa013	Paris-SA	MSI	5	Large	3	RC	F	58	LYNCH	0	32	Y	FOLFOX
psa014	Paris-SA	MSI	4	Small	3	RC	M	73	LYNCH	1	5	Y	LV5FU2
psa015	Paris-SA	MSI	4	Small	3	LC	M	32	LYNCH	0	76	Y	LV5FU2
psa016	Paris-SA	MSI	5	Large	2	RC	M	46	none	0	73	N	None
psa017	Paris-SA	MSI	7	Large	3	RC	F	78	none	0	60	Y	LV5FU2
psa018	Paris-SA	MSI	4	Small	3	LC	M	56	LYNCH	0	20	Y	FOLFOX
psa019	Paris-SA	MSI	5	Large	2	RC	F	75	none	0	65	N	None
psa020	Paris-SA	MSI	5	Large	2	RC	M	48	LYNCH	0	35	N	None
psa021	Paris-SA	MSI	4	Small	3	RC	F	91	none	0	57	N	None
psa022	Paris-SA	MSI	3	Small	2	RC	M	76	none	1	3	N	None
psa023	Paris-SA	MSI	5	Large	3	RC	M	59	none	0	43	Y	LV5FU2
psa024	Paris-SA	MSI	6	Large	2	RC	F	88	none	0	51	N	None
psa025	Paris-SA	MSI	6	Large	2	RC	F	72	none	0	45	N	None
psa026	Paris-SA	MSI	5	Large	3	RC	F	67	none	0	52	Y	LV5FU2
psa027	Paris-SA	MSI	3	Small	2	RC	F	79	none	0	46	N	None
psa028	Paris-SA	MSI	3	Small	3	RC	F	59	none	1	24	Y	LV5FU2
psa029	Paris-SA	MSI	5	Large	2	LC	F	79	LYNCH	0	51	N	None
psa030	Paris-SA	MSI	5	Large	2	RC	M	65	LYNCH	0	51	N	None
psa031	Paris-SA	MSI	2	Small	3	RECTUM	M	61	LYNCH	0	35	Y	LV5FU2
psa032	Paris-SA	MSI	4	Small	2	LC	M	71	none	0	2	N	None
psa033	Paris-SA	MSI	7	Large	2	RC	M	24	LYNCH	0	37	N	None
psa034	Paris-SA	MSI	4	Small	2	RC	M	77	none	0	43	N	None
psa035	Paris-SA	MSI	4	Small	2	RC	M	67	LYNCH	0	35	N	None
psa036	Paris-SA	MSI	5	Large	2	RC	M	71	none	0	31	N	None
psa037	Paris-SA	MSI	2	Small	2	LC	F	70	LYNCH	0	39	N	None
psa038	Paris-SA	MSI	4	Small	3	LC	M	46	LYNCH	1	10	Y	FOLFOX
psa039	Paris-SA	MSI	3	Small	3	LC	F	26	NA	0	16	Y	FOLFOX
psa040	Paris-SA	MSI	4	Small	2	RC	F	83	none	0	9	N	None
psa041	Paris-SA	MSI	4	Small	2	LC	M	31	NA	0	28	Y	FOLFOX
psa042	Paris-SA	MSI	5	Large	2	RC	F	67	LYNCH	0	28	N	None
psa043	Paris-SA	MSI	4	Small	2	RC	F	66	none	0	20	Y	FOLFOX
psa044	Paris-SA	MSI	5	Large	2	RC	M	74	none	0	36	N	None

ACCEPTED MANUSCRIPT

psa045	Paris-SA	MSI	5	Large	2	LC	F	87	none	0	48	N	None
psa046	Paris-SA	MSI	3	Small	3	RC	M	44	LYNCH	0	91	Y	FOLFOX
psa047	Paris-SA	MSI	4	Small	2	RC	F	85	none	0	33	N	None
psa048	Paris-SA	MSI	2	Small	2	RC	F	89	none	0	24	N	None
psa049	Paris-SA	MSI	5	Large	3	RC	F	51	NA	0	45	Y	LV5FU2
psa050	Paris-SA	MSI	2	Small	2	LC	M	71	LYNCH	0	6	Y	FOLFOX
psa051	Paris-SA	MSI	1	Small	2	RECTUM	F	28	LYNCH	0	42	N	None
psa052	Paris-SA	MSI	5	Large	2	RC	M	65	LYNCH	1	13	N	None
psa053	Paris-SA	MSI	2	Small	2	LC	F	73	LYNCH	0	49	N	None
psa054	Paris-SA	MSI	3	Small	2	RC	M	87	none	0	46	N	None
psa055	Paris-SA	MSI	4	Small	2	RC	M	60	none	0	72	N	None
psa056	Paris-SA	MSI	1	Small	2	RC	F	80	LYNCH	0	42	N	None
psa057	Paris-SA	MSI	5	Large	3	RC	F	97	none	0	2	N	None
psa058	Paris-SA	MSI	4	Small	2	LC	F	85	none	0	48	N	None
psa059	Paris-SA	MSI	5	Large	3	LC	M	52	LYNCH	0	60	Y	FOLFOX
psa060	Paris-SA	MSI	3	Small	2	RC	M	67	LYNCH	0	60	N	None
psa061	Paris-SA	MSI	2	Small	3	RC	F	74	none	1	6	Y	FOLFOX
psa062	Paris-SA	MSI	2	Small	2	RC	F	85	none	0	48	N	None
psa063	Paris-SA	MSI	5	Large	2	NA	M	83	none	1	40	N	None
psa064	Paris-SA	MSI	4	Small	2	LC	F	60	LYNCH	0	42	N	None
psa065	Paris-SA	MSI	0	Small	2	LC	M	26	none	0	60	N	None
psa066	Paris-SA	MSI	4	Small	2	RECTUM	M	69	LYNCH	0	21	N	None
psa067	Paris-SA	MSI	5	Large	2	RC	M	86	none	0	17	N	None
psa068	Paris-SA	MSI	4	Small	2	RC	M	50	LYNCH	0	42	N	None
psa069	Paris-SA	MSI	5	Large	2	RC	F	90	none	0	1	N	None
psa070	Paris-SA	MSI	4	Small	2	RC	F	51	LYNCH	0	60	N	None
psa071	Paris-SA	MSI	5	Large	2	RC	F	80	none	0	60	N	None
psa072	Paris-SA	MSI	3	Small	2	RC	M	82	none	0	1	N	None
psa073	Paris-SA	MSI	5	Large	2	LC	M	48	LYNCH	0	72	N	None
psa074	Paris-SA	MSI	4	Small	2	RECTUM	M	59	LYNCH	0	24	N	None
psa075	Paris-SA	MSI	4	Small	2	RC	F	73	none	0	12	N	None
psa076	Paris-SA	MSI	5	Large	2	RC	M	57	LYNCH	0	24	N	None
psa077	Paris-SA	MSI	3	Small	2	LC	M	37	LYNCH	0	36	N	None
psa078	Paris-SA	MSI	3	Small	2	LC	F	53	LYNCH	0	44	N	None
psa079	Paris-SA	MSI	2	Small	2	RC	M	66	LYNCH	0	54	N	None
psa080	Paris-SA	MSI	3	Small	3	LC	M	83	none	1	19	N	None
psa081	Paris-SA	MSI	2	Small	2	RC	F	77	none	0	59	N	None
psa082	Paris-SA	MSI	3	Small	2	RC	F	66	none	0	54	N	None
psa083	Paris-SA	MSI	2	Small	3	RC	M	81	none	0	48	Y	LV5FU2
psa084	Paris-SA	MSI	4	Small	2	RC	F	86	none	0	69	N	None
psa085	Paris-SA	MSI	5	Large	2	RC	F	82	none	1	13	N	None
psa086	Paris-SA	MSI	3	Small	2	RC	M	57	LYNCH	0	45	N	None

ACCEPTED MANUSCRIPT

psa087	Paris-SA	MSI	1	Small	2	RC	M	83	none	0	1	N	None
psa088	Paris-SA	MSI	5	Large	3	RC	M	67	none	0	24	N	None
psa089	Paris-SA	MSI	3	Small	2	RC	F	77	LYNCH	0	19	N	None
psa090	Paris-SA	MSI	6	Large	2	LC	M	49	LYNCH	0	42	N	None
psa091	Paris-SA	MSI	3	Small	2	RC	F	103	none	1	30	N	None
psa092	Paris-SA	MSI	0	Small	2	RECTUM	F	55	LYNCH	0	60	N	None
psa093	Paris-SA	MSI	2	Small	3	RC	M	59	LYNCH	0	22	Y	FOLFOX
psa094	Paris-SA	MSI	4	Small	3	RC	F	84	none	0	39	Y	FOLFOX
psa095	Paris-SA	MSI	4	Small	2	RC	F	85	none	0	3	N	None
psa096	Paris-SA	MSI	6	Large	2	RC	F	75	none	0	17	N	None
psa097	Paris-SA	MSI	2	Small	2	RC	F	89	none	0	48	N	None
psa098	Paris-SA	MSI	4	Small	3	LC	M	56	LYNCH	0	26	Y	FOLFOX
psa099	Paris-SA	MSS	0	Small	2	RC	F	70	NA	NA	NA	NA	NA
psa100	Paris-SA	MSS	0	Small	2	LC	M	73	NA	NA	NA	NA	NA
psa101	Paris-SA	MSS	0	Small	3	RC	F	90	NA	NA	NA	NA	NA
psa102	Paris-SA	MSS	0	Small	2	RC	M	65	NA	NA	NA	NA	NA
psa103	Paris-SA	MSS	0	Small	2	LC	F	54	NA	NA	NA	NA	NA
psa104	Paris-SA	MSS	0	Small	2	RECTUM	F	83	NA	NA	NA	NA	NA
psa105	Paris-SA	MSS	0	Small	2	RC	M	44	NA	NA	NA	NA	NA
psa106	Paris-SA	MSS	0	Small	2	RC	F	84	NA	NA	NA	NA	NA
psa107	Paris-SA	MSS	0	Small	2	RC	F	69	NA	NA	NA	NA	NA
psa108	Paris-SA	MSS	0	Small	2	RC	F	75	NA	NA	NA	NA	NA
psa109	Paris-SA	MSS	0	Small	2	RC	F	74	NA	NA	NA	NA	NA
psa110	Paris-SA	MSS	0	Small	2	RC	F	62	NA	NA	NA	NA	NA
psa111	Paris-SA	MSS	0	Small	3	RC	F	81	NA	NA	NA	NA	NA
psa112	Paris-SA	MSS	0	Small	2	RC	F	80	NA	NA	NA	NA	NA
psa113	Paris-SA	MSS	0	Small	2	RC	M	62	NA	NA	NA	NA	NA
psa114	Paris-SA	MSS	0	Small	2	RC	M	72	NA	NA	NA	NA	NA
psa115	Paris-SA	MSS	0	Small	3	RC	F	80	NA	NA	NA	NA	NA
psa116	Paris-SA	MSS	0	Small	2	RC	F	90	NA	NA	NA	NA	NA
psa117	Paris-SA	MSS	0	Small	3	RC	F	71	NA	NA	NA	NA	NA
psa118	Paris-SA	MSS	0	Small	2	RECTUM	M	63	NA	NA	NA	NA	NA
psa119	Paris-SA	MSS	0	Small	2	RC	M	59	NA	NA	NA	NA	NA
psa120	Paris-SA	MSS	0	Small	3	RC	M	81	NA	NA	NA	NA	NA
psa121	Paris-SA	MSS	0	Small	2	LC	M	73	NA	NA	NA	NA	NA
psa122	Paris-SA	MSS	0	Small	2	LC	M	64	NA	NA	NA	NA	NA
psa123	Paris-SA	MSS	0	Small	2	RC	F	76	NA	NA	NA	NA	NA
psa124	Paris-SA	MSS	0	Small	2	RC	F	67	NA	NA	NA	NA	NA
psa125	Paris-SA	MSS	0	Small	2	RC	F	38	NA	NA	NA	NA	NA
psa126	Paris-SA	MSS	0	Small	2	LC	M	47	NA	NA	NA	NA	NA
psa127	Paris-SA	MSS	0	Small	2	RC	M	80	NA	NA	NA	NA	NA
psa128	Paris-SA	MSS	0	Small	2	RC	F	74	NA	NA	NA	NA	NA

ACCEPTED MANUSCRIPT

psa129	Paris-SA	MSS	0	Small	2	RC	M	48	NA	NA	NA	NA	NA
psa130	Paris-SA	MSS	0	Small	3	RC	F	74	NA	NA	NA	NA	NA
psa131	Paris-SA	MSS	0	Small	2	RC	M	43	NA	NA	NA	NA	NA
psa132	Paris-SA	MSS	0	Small	2	RC	F	64	NA	NA	NA	NA	NA
psa133	Paris-SA	MSS	0	Small	2	RC	M	56	NA	NA	NA	NA	NA
psa134	Paris-SA	MSS	0	Small	2	RC	F	85	NA	NA	NA	NA	NA
psa135	Paris-SA	MSS	0	Small	2	RC	M	42	NA	NA	NA	NA	NA
psa136	Paris-SA	MSS	0	Small	2	RC	F	76	NA	NA	NA	NA	NA
psa137	Paris-SA	MSS	0	Small	2	RC	M	34	NA	NA	NA	NA	NA
psa138	Paris-SA	MSS	0	Small	2	RC	M	84	NA	NA	NA	NA	NA
psa139	Paris-SA	MSS	0	Small	2	RC	M	76	NA	NA	NA	NA	NA
psa140	Paris-SA	MSS	0	Small	3	RC	M	73	NA	NA	NA	NA	NA
psa141	Paris-SA	MSS	0	Small	3	RC	M	40	NA	NA	NA	NA	NA
psa142	Paris-SA	MSS	0	Small	3	RC	M	71	NA	NA	NA	NA	NA
psa143	Paris-SA	MSS	0	Small	2	RC	F	67	NA	NA	NA	NA	NA
psa144	Paris-SA	MSS	0	Small	2	RC	M	73	NA	NA	NA	NA	NA
psa145	Paris-SA	MSS	0	Small	2	RC	F	80	NA	NA	NA	NA	NA
psa146	Paris-SA	MSS	0	Small	2	RC	F	68	NA	NA	NA	NA	NA
psa147	Paris-SA	MSS	0	Small	2	RC	M	84	NA	NA	NA	NA	NA
psa148	Paris-SA	MSS	0	Small	2	LC	F	58	NA	NA	NA	NA	NA
psa149	Paris-SA	MSS	0	Small	2	LC	M	56	NA	NA	NA	NA	NA
psa150	Paris-SA	MSS	0	Small	2	RC	M	83	NA	NA	NA	NA	NA
psa151	Paris-SA	MSS	0	Small	3	RC	F	52	NA	NA	NA	NA	NA
psa152	Paris-SA	MSS	0	Small	2	RC	M	53	NA	NA	NA	NA	NA
psa153	Paris-SA	MSS	0	Small	2	RC	F	79	NA	NA	NA	NA	NA
psa154	Paris-SA	MSS	0	Small	2	RC	M	76	NA	NA	NA	NA	NA
psa155	Paris-SA	MSS	0	Small	3	RC	M	60	NA	NA	NA	NA	NA
psa156	Paris-SA	MSS	0	Small	2	RC	F	71	NA	NA	NA	NA	NA
psa157	Paris-SA	MSS	0	Small	2	RC	F	55	NA	NA	NA	NA	NA
psa158	Paris-SA	MSS	0	Small	2	RC	F	88	NA	NA	NA	NA	NA
psa159	Paris-SA	MSS	0	Small	2	RC	M	85	NA	NA	NA	NA	NA
psa160	Paris-SA	MSS	0	Small	2	RC	M	83	NA	NA	NA	NA	NA
psa161	Paris-SA	MSS	0	Small	3	RC	M	75	NA	NA	NA	NA	NA
psa162	Paris-SA	MSS	0	Small	3	RC	M	69	NA	NA	NA	NA	NA
psa163	Paris-SA	MSS	0	Small	3	RC	M	60	NA	NA	NA	NA	NA
psa164	Paris-SA	MSS	0	Small	2	RC	M	70	NA	NA	NA	NA	NA
psa165	Paris-SA	MSS	0	Small	2	RECTUM	F	77	NA	NA	NA	NA	NA
psa166	Paris-SA	MSS	0	Small	2	RC	M	74	NA	NA	NA	NA	NA
psa167	Paris-SA	MSS	0	Small	3	RC	M	81	NA	NA	NA	NA	NA
psa168	Paris-SA	MSS	0	Small	2	RC	F	88	NA	NA	NA	NA	NA
psa169	Paris-SA	MSS	0	Small	2	RC	M	80	NA	NA	NA	NA	NA
psa170	Paris-SA	MSS	0	Small	2	RC	M	61	NA	NA	NA	NA	NA

ACCEPTED MANUSCRIPT

psa171	Paris-SA	MSS	0	Small	2	RC	F	80	NA	NA	NA	NA	NA
psa172	Paris-SA	MSS	0	Small	2	RC	M	80	NA	NA	NA	NA	NA
psa173	Paris-SA	MSS	0	Small	2	RC	M	81	NA	NA	NA	NA	NA
psa174	Paris-SA	MSS	0	Small	2	RC	M	83	NA	NA	NA	NA	NA
psa175	Paris-SA	MSS	0	Small	2	RC	F	79	NA	NA	NA	NA	NA
psa176	Paris-SA	MSS	0	Small	2	RC	F	56	NA	NA	NA	NA	NA
psa177	Paris-SA	MSS	0	Small	3	RC	F	71	NA	NA	NA	NA	NA
psa178	Paris-SA	MSS	0	Small	2	RC	M	82	NA	NA	NA	NA	NA
psa179	Paris-SA	MSS	0	Small	2	RC	F	72	NA	NA	NA	NA	NA
psa180	Paris-SA	MSS	0	Small	2	RC	F	79	NA	NA	NA	NA	NA
psa181	Paris-SA	MSS	0	Small	2	RC	M	70	NA	NA	NA	NA	NA
psa182	Paris-SA	MSS	0	Small	3	RC	M	84	NA	NA	NA	NA	NA
psa183	Paris-SA	MSS	0	Small	3	RC	F	79	NA	NA	NA	NA	NA
psa184	Paris-SA	MSS	0	Small	2	RC	F	78	NA	NA	NA	NA	NA
psa185	Paris-SA	MSS	0	Small	2	RC	F	88	NA	NA	NA	NA	NA
psa186	Paris-SA	MSS	0	Small	2	RC	F	77	NA	NA	NA	NA	NA
psa187	Paris-SA	MSS	0	Small	3	RC	M	46	NA	NA	NA	NA	NA
psa188	Paris-SA	MSS	0	Small	2	RC	M	75	NA	NA	NA	NA	NA
psa189	Paris-SA	MSS	0	Small	2	RC	M	63	NA	NA	NA	NA	NA
psa190	Paris-SA	MSS	0	Small	2	LC	M	28	NA	NA	NA	NA	NA
psa191	Paris-SA	MSS	0	Small	2	RC	M	45	NA	NA	NA	NA	NA
psa192	Paris-SA	MSS	0	Small	2	LC	M	61	NA	NA	NA	NA	NA
psa193	Paris-SA	MSS	0	Small	2	RC	F	74	NA	NA	NA	NA	NA
psa194	Paris-SA	MSS	0	Small	2	RC	F	53	NA	NA	NA	NA	NA
psa195	Paris-SA	MSS	0	Small	3	RC	F	81	NA	NA	NA	NA	NA
psa196	Paris-SA	MSS	0	Small	2	RC	M	58	NA	NA	NA	NA	NA
psa197	Paris-SA	MSS	0	Small	2	RC	F	56	NA	NA	NA	NA	NA
psa198	Paris-SA	MSS	0	Small	2	RC	F	81	NA	NA	NA	NA	NA
psa199	Paris-SA	MSS	0	Small	3	LC	M	41	NA	NA	NA	NA	NA
psa200	Paris-SA	MSS	0	Small	2	RC	M	49	NA	NA	NA	NA	NA
psa201	Paris-SA	MSS	0	Small	3	RC	F	100	NA	NA	NA	NA	NA
psa202	Paris-SA	MSS	0	Small	3	RC	M	52	NA	NA	NA	NA	NA
sin001	Singapore	MSI	0	Small	2	LC	F	83	NA	0	21.7	N	None
sin002	Singapore	MSI	6	Large	2	LC	M	42	NA	0	121	N	None
sin003	Singapore	MSI	1	Small	2	LC	M	72	NA	1	92	N	None
sin004	Singapore	MSI	3	Small	2	RC	F	51	NA	0	108	N	None
sin005	Singapore	MSI	3	Small	2	LC	M	48	NA	0	109	N	None
sin006	Singapore	MSI	1	Small	2	RC	F	41	NA	0	106	N	None
sin007	Singapore	MSI	3	Small	2	LC	F	69	NA	0	88	N	None
sin008	Singapore	MSI	2	Small	2	RC	F	70	NA	0	138.17	N	None
sin009	Singapore	MSI	2	Small	2	LC	F	64	NA	0	91.95	N	None
sin010	Singapore	MSI	1	Small	3	RC	F	65	NA	1	34	Y	LV5FU2

ACCEPTED MANUSCRIPT

sin011	Singapore	MSI	0	Small	3	LC	M	60	NA	0	207.26	Y	LV5FU2
sin012	Singapore	MSI	1	Small	3	RC	F	64	NA	1	20	Y	LV5FU2
sin013	Singapore	MSI	4	Small	2	LC	M	39	NA	0	0.2	N	None
sin014	Singapore	MSI	3	Small	2	RC	M	70	NA	1	16	Y	LV5FU2
sin015	Singapore	MSI	4	Small	3	RC	F	60	NA	0	217.74	Y	LV5FU2
sin016	Singapore	MSI	3	Small	3	LC	M	62	NA	0	105.06	Y	LV5FU2
sin017	Singapore	MSI	4	Small	2	LC	M	72	NA	0	213.24	N	None
sin018	Singapore	MSI	4	Small	2	LC	M	79	NA	1	15.54	N	None
sin019	Singapore	MSI	6	Large	3	RC	F	44	NA	0	184.82	Y	LV5FU2
sin020	Singapore	MSI	0	Small	2	LC	M	52	NA	0	0.3	N	None
sin021	Singapore	MSI	3	Small	2	RC	M	57	NA	0	189.06	Y	LV5FU2
sin022	Singapore	MSI	4	Small	2	LC	F	43	NA	0	153.81	N	None
sin023	Singapore	MSI	3	Small	2	RC	M	71	NA	0	63.14	Y	LV5FU2
sin024	Singapore	MSI	4	Small	3	RECTUM	M	40	NA	0	170.37	Y	LV5FU2
sin025	Singapore	MSI	4	Small	3	LC	M	76	NA	0	0.03	N	None
sin026	Singapore	MSI	4	Small	2	LC	F	61	NA	0	147.8	N	None
sin027	Singapore	MSI	4	Small	3	RC	F	82	NA	0	0.33	N	None
sin028	Singapore	MSI	4	Small	2	RC	F	75	NA	1	29.11	N	None
sin029	Singapore	MSI	4	Small	3	LC	M	32	NA	1	6.34	Y	LV5FU2
sin030	Singapore	MSI	4	Small	3	RC	M	79	NA	0	1.15	N	None
sin031	Singapore	MSI	4	Small	3	RECTUM	M	58	NA	1	6.01	Y	LV5FU2
sin032	Singapore	MSI	0	Small	2	LC	F	70	NA	0	156.11	Y	LV5FU2
sin033	Singapore	MSI	5	Large	2	LC	M	60	NA	0	81.41	N	None
sin034	Singapore	MSI	4	Small	3	LC	F	42	NA	0	155.35	Y	LV5FU2
sin035	Singapore	MSI	2	Small	3	LC	F	84	NA	1	8.61	N	None
sin036	Singapore	MSI	1	Small	3	RC	M	63	NA	1	5.26	N	None
sin037	Singapore	MSI	1	Small	2	RC	F	92	NA	1	1.68	N	None
sin038	Singapore	MSI	1	Small	2	RC	F	88	NA	0	48.16	N	None
sin039	Singapore	MSI	4	Small	2	RC	M	80	NA	0	51.68	N	None
sin040	Singapore	MSI	2	Small	3	RC	F	74	NA	1	9.86	N	None
sin041	Singapore	MSI	7	Large	2	LC	M	71	NA	0	144.19	N	None
sin042	Singapore	MSI	5	Large	2	RC	F	80	NA	0	144.84	N	None
sin043	Singapore	MSI	4	Small	2	LC	F	85	NA	0	142.21	N	None
sin044	Singapore	MSI	2	Small	2	RC	M	63	NA	0	122.96	Y	LV5FU2
sin045	Singapore	MSI	4	Small	2	RC	M	42	NA	0	120.57	N	None
sin046	Singapore	MSI	4	Small	2	RC	M	31	NA	0	117.97	N	None
sin047	Singapore	MSI	3	Small	3	RC	F	74	NA	0	113.93	Y	LV5FU2
sin048	Singapore	MSI	3	Small	2	RC	F	51	NA	0	108.05	N	None
sin049	Singapore	MSI	2	Small	2	RC	F	52	NA	0	92.71	N	None
sin050	Singapore	MSI	5	Large	2	RC	F	87	NA	0	62.02	N	None
sin051	Singapore	MSS	0	Small	2	LC	M	42	NA	NA	NA	NA	NA
sin052	Singapore	MSS	0	Small	2	LC	M	71	NA	NA	NA	NA	NA

ACCEPTED MANUSCRIPT

sin053	Singapore	MSS	0	Small	2	RC	F	53	NA	NA	NA	NA	NA
sin054	Singapore	MSS	0	Small	2	LC	M	49	NA	NA	NA	NA	NA
sin055	Singapore	MSS	0	Small	2	RC	F	39	NA	NA	NA	NA	NA
sin056	Singapore	MSS	0	Small	2	RC	F	70	NA	NA	NA	NA	NA
sin057	Singapore	MSS	0	Small	3	RC	F	69	NA	NA	NA	NA	NA
sin058	Singapore	MSS	0	Small	3	LC	M	60	NA	NA	NA	NA	NA
sin059	Singapore	MSS	0	Small	3	RC	F	67	NA	NA	NA	NA	NA
sin060	Singapore	MSS	0	Small	2	LC	M	39	NA	NA	NA	NA	NA
sin061	Singapore	MSS	0	Small	2	RC	M	72	NA	NA	NA	NA	NA
sin062	Singapore	MSS	0	Small	3	RC	F	52	NA	NA	NA	NA	NA
sin063	Singapore	MSS	0	Small	3	LC	M	62	NA	NA	NA	NA	NA
sin064	Singapore	MSS	0	Small	2	LC	M	73	NA	NA	NA	NA	NA
sin065	Singapore	MSS	0	Small	2	LC	M	79	NA	NA	NA	NA	NA
sin066	Singapore	MSS	0	Small	3	RC	F	45	NA	NA	NA	NA	NA
sin067	Singapore	MSS	0	Small	2	LC	M	52	NA	NA	NA	NA	NA
sin068	Singapore	MSS	0	Small	2	LC	F	42	NA	NA	NA	NA	NA
sin069	Singapore	MSS	0	Small	2	RC	M	71	NA	NA	NA	NA	NA
sin070	Singapore	MSS	0	Small	3	RC	F	83	NA	NA	NA	NA	NA
sin071	Singapore	MSS	0	Small	2	RC	F	74	NA	NA	NA	NA	NA
sin072	Singapore	MSS	0	Small	3	LC	M	30	NA	NA	NA	NA	NA
sin073	Singapore	MSS	0	Small	2	RC	M	73	NA	NA	NA	NA	NA
sin074	Singapore	MSS	0	Small	2	LC	M	60	NA	NA	NA	NA	NA
sin075	Singapore	MSS	0	Small	3	LC	F	42	NA	NA	NA	NA	NA
sin076	Singapore	MSS	0	Small	3	RC	M	62	NA	NA	NA	NA	NA
sin077	Singapore	MSS	0	Small	2	RC	F	87	NA	NA	NA	NA	NA
sin078	Singapore	MSS	0	Small	3	RC	F	73	NA	NA	NA	NA	NA
sin079	Singapore	MSS	0	Small	2	LC	M	71	NA	NA	NA	NA	NA
sin080	Singapore	MSS	0	Small	2	RC	F	80	NA	NA	NA	NA	NA
sin081	Singapore	MSS	0	Small	2	RC	M	62	NA	NA	NA	NA	NA
sin082	Singapore	MSS	0	Small	2	RC	M	75	NA	NA	NA	NA	NA
sin083	Singapore	MSS	0	Small	3	RC	F	65	NA	NA	NA	NA	NA
sin084	Singapore	MSS	0	Small	2	RC	F	65	NA	NA	NA	NA	NA
sin085	Singapore	MSS	0	Small	2	RC	F	87	NA	NA	NA	NA	NA
tou001	Toulouse	MSI	4	Small	2	RC	M	78.2	LYNCH	1	17	Y	NA
tou002	Toulouse	MSI	4	Small	3	RC	F	35.8	NA	1	10	Y	NA
tou003	Toulouse	MSI	4	Small	2	RC	F	66.4	none	0	53	N	N
tou004	Toulouse	MSI	1	Small	2	RC	F	88.3	none	0	67	N	N
tou005	Toulouse	MSI	5	Large	2	RC	M	81.8	none	0	48	N	N
tou006	Toulouse	MSI	3	Small	3	RC	M	79.3	none	0	42	N	N
tou007	Toulouse	MSI	3	Small	3	RC	M	75.3	none	0	87	N	N
tou008	Toulouse	MSI	4	Small	2	RC	F	68	none	1	5	N	N
tou009	Toulouse	MSI	4	Small	2	LC	F	92.4	none	0	91	N	N

ACCEPTED MANUSCRIPT

tou010	Toulouse	MSI	6	Large	2	RC	F	74.5	LYNCH	1	14	N	N
tou011	Toulouse	MSS	0	Small	2	LC	M	71.3	NA	NA	NA	NA	NA
tou012	Toulouse	MSS	0	Small	2	LC	M	67.5	NA	NA	NA	NA	NA

Abbreviations : MSI, Microsatellite Instable tumor; MSS, Microsatellite stable tumor; RC, right colon/proximal colon; LC, left colon/distal colon; F, Female; M, Male; Large, *HSP110* Deletion size ≥5bp; Small *HSP110* Deletion size <5bp ; NA : Non Applicable or Non Available ; N : No ; Y : Yes. * Patients that were likely to have Lynch syndrome were those that displayed no methylation of the *MLH1* promoter in their tumor.

ACCEPTED MANUSCRIPT

ACCEPTED MANUSCRIPT

Supplementary Table S2. Associations of clinical and molecular annotations to outcome (relapse-free survival) for Stage II & III MSI-CRC Patients.

Variable	Available data	COX UNIVARIATE ANALYSIS				COX MULTIVARIATE ANALYSIS[a]			
				modality	model			modality	model
				P value	P value			P value	P value
	n (n relapse)	H.R.	(95%C.I.)	(Wald)	(Log-rank)	H.R.	(95%C.I.)	(Wald)	(Log-rank)
TNM Stage (III)	329 (70)	2.3	(1.4-3.7)	<.001	<.001	2.2	(1.3-3.8)	<.001	
Age recoded (≥75y)	329 (70)	1.5	(0.89-2.5)	.13	.13	1.5	(0.8-2.7)	.21	
Chemotherapy (yes)	329 (70)	1.4	(0.83-2.4)	.21	.21	1.3	(0.64-2.6)	.48	
HSP110Del (Large vs Small)	329 (70)	0.7	(0.37-1.3)	.26	.26	0.57	(0.29-1.1)	.096	
Tumor location (Proximal colon)	324 (68)	1.8	(0.85-3.6)	.13	.28	1.8	(0.87-3.9)	.11	<.001
Tumor location (Rectum)		1.1	(0.14-8.6)	.94		0.92	(0.11-7.4)	.94	
Gender (Male)	329 (70)	0.93	(0.57-1.5)	.76	.76	1.1	(0.64-1.8)	.8	
5-FU based Chemotherapy type *	72 (17)	1.8	(0.56-5.6)	.33	.32				
Lynch syndrome (Lynch)	166 (35)	0.69	(0.24-2)	.5	.50				

[a] Multivariate models included variables available for most samples. Therefore, the model was estimated on 324 patients (n relapse=68).

Abbreviations: H.R., Cox Hazard Ratio; 95% C.I., 95 Percent Confidence Interval of HR; LV5FU2, Fluorouracil and Leucovorin regimen. *LV5FU2 alone or FOLFOX (LV5FU2 + Oxaliplatin)

ACCEPTED MANUSCRIPT

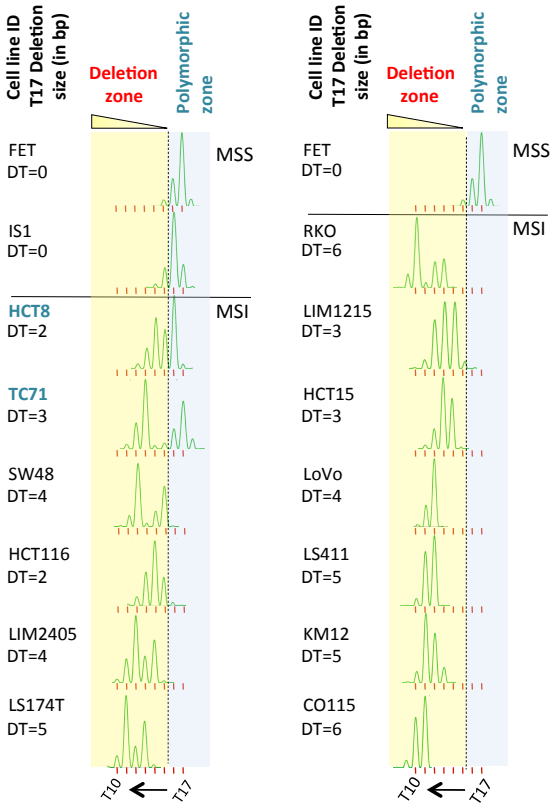
ACCEPTED MANUSCRIPT

Supplementary Table 3. Interaction between *HSP110* deletion size and chemotherapy performed.

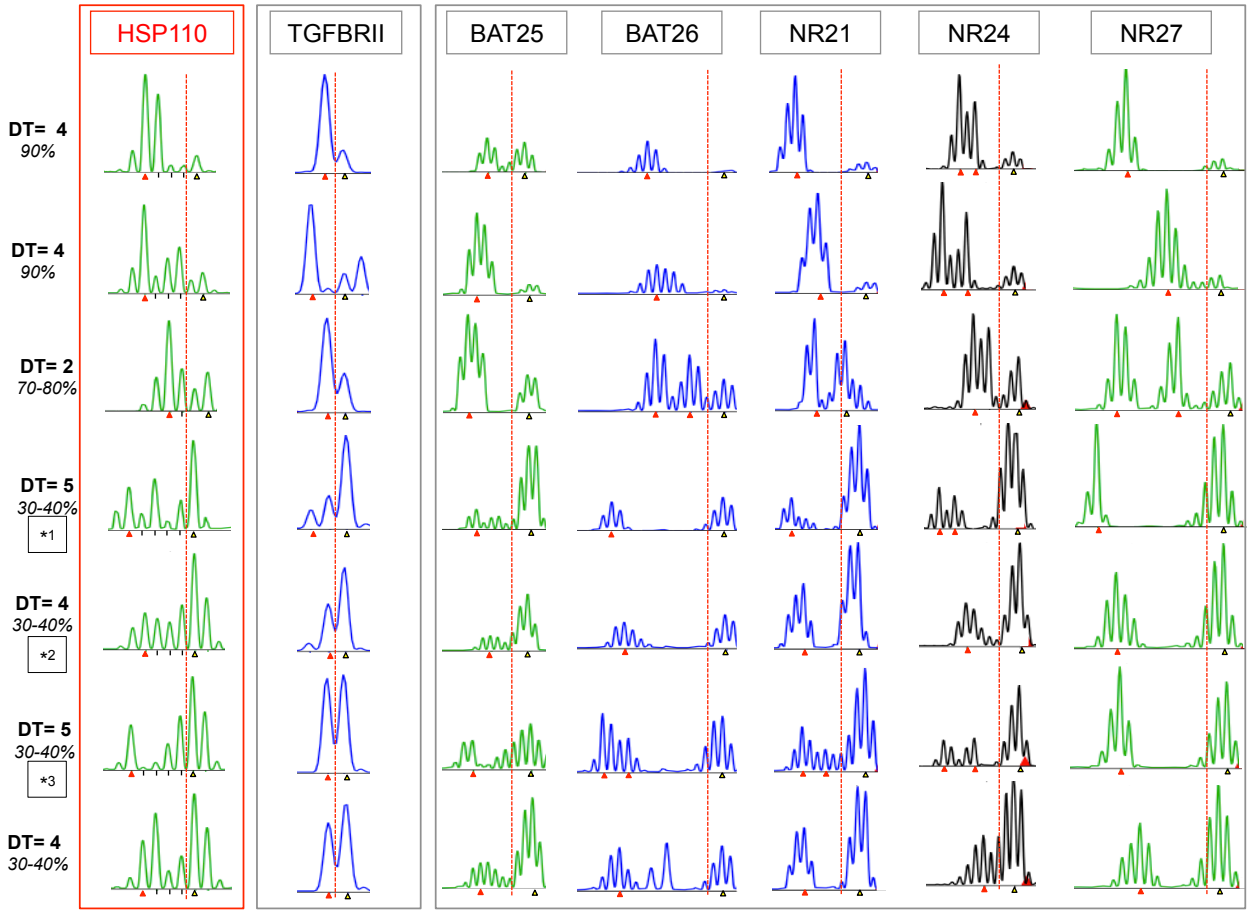
		Under chemotherapy		No chemotherapy		*P* value for interaction
		***HSP110* Del^{L}**	***HSP110* Del^{S}**	***HSP110* Del^{L}**	***HSP110* Del^{S}**	
No. of relapse/No. At Risk		janv-18	19/59	nov-58	40/194	
5-Year survival		0.94	0.64	0.79	0.76	
HR (95% CI)						
	univariate	0.16 (0.02-1.2)	1 [Reference]	1 (0.53-2.1)	1 [Reference]	.030
	multivariate[a]	0.089 (0.011-0.71)	1 [Reference]	0.99 (0.48-2)	1 [Reference]	.0090

[a] Multivariate models included HSP110 deletion, Stage, Tumor Location, Age (recoded 75y) and Gender.

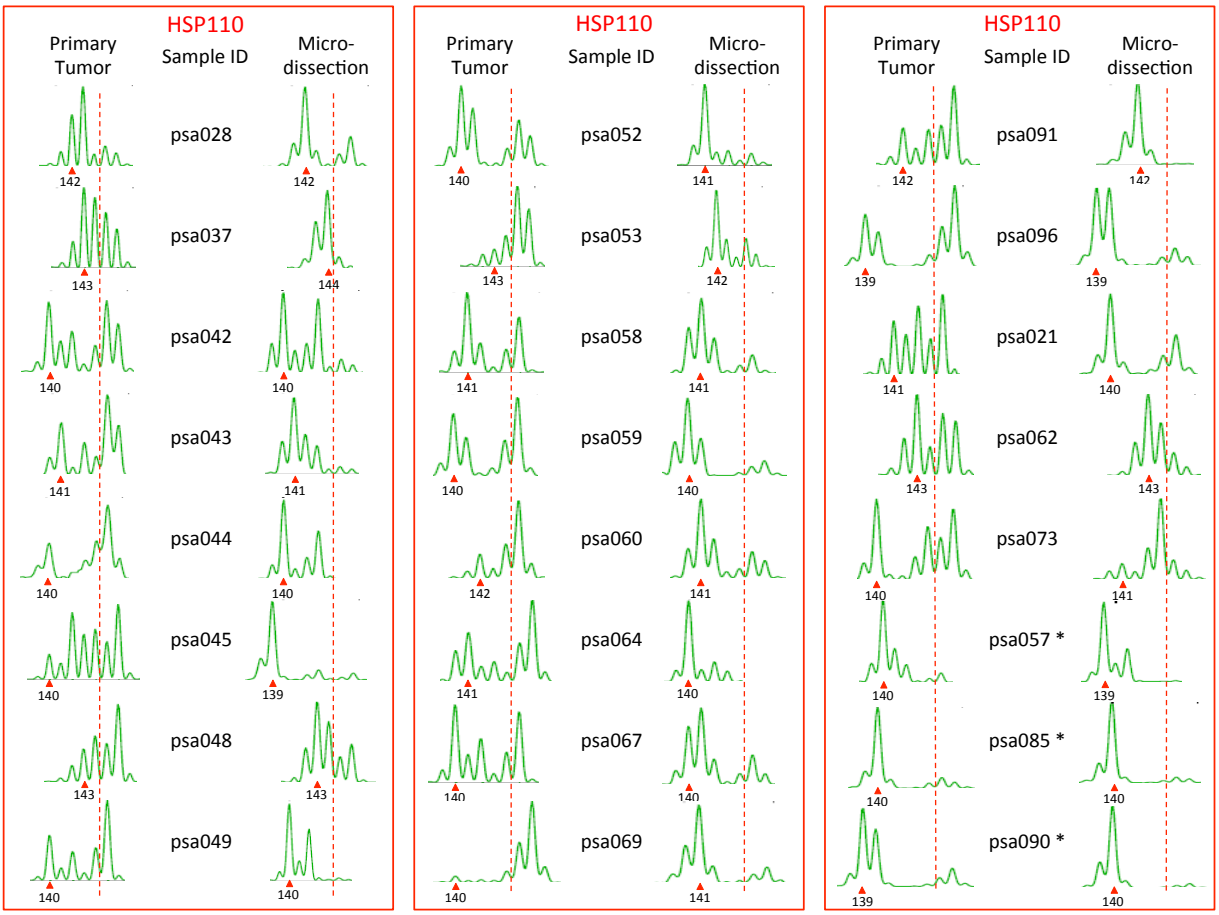
ACCEPTED MANUSCRIPT



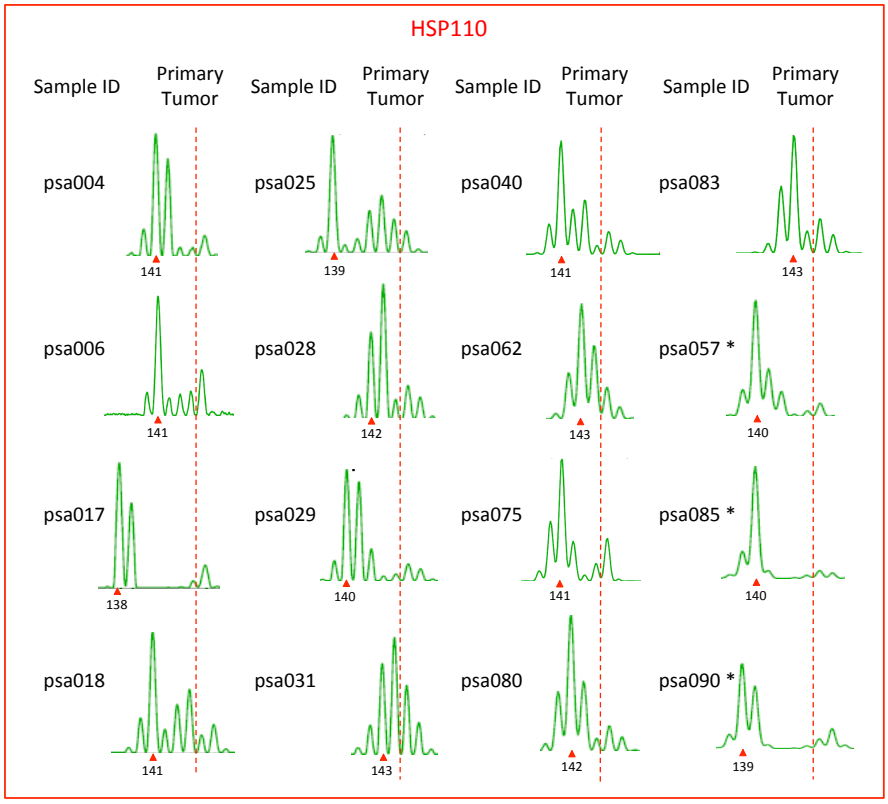
Supplementary Figure S1A



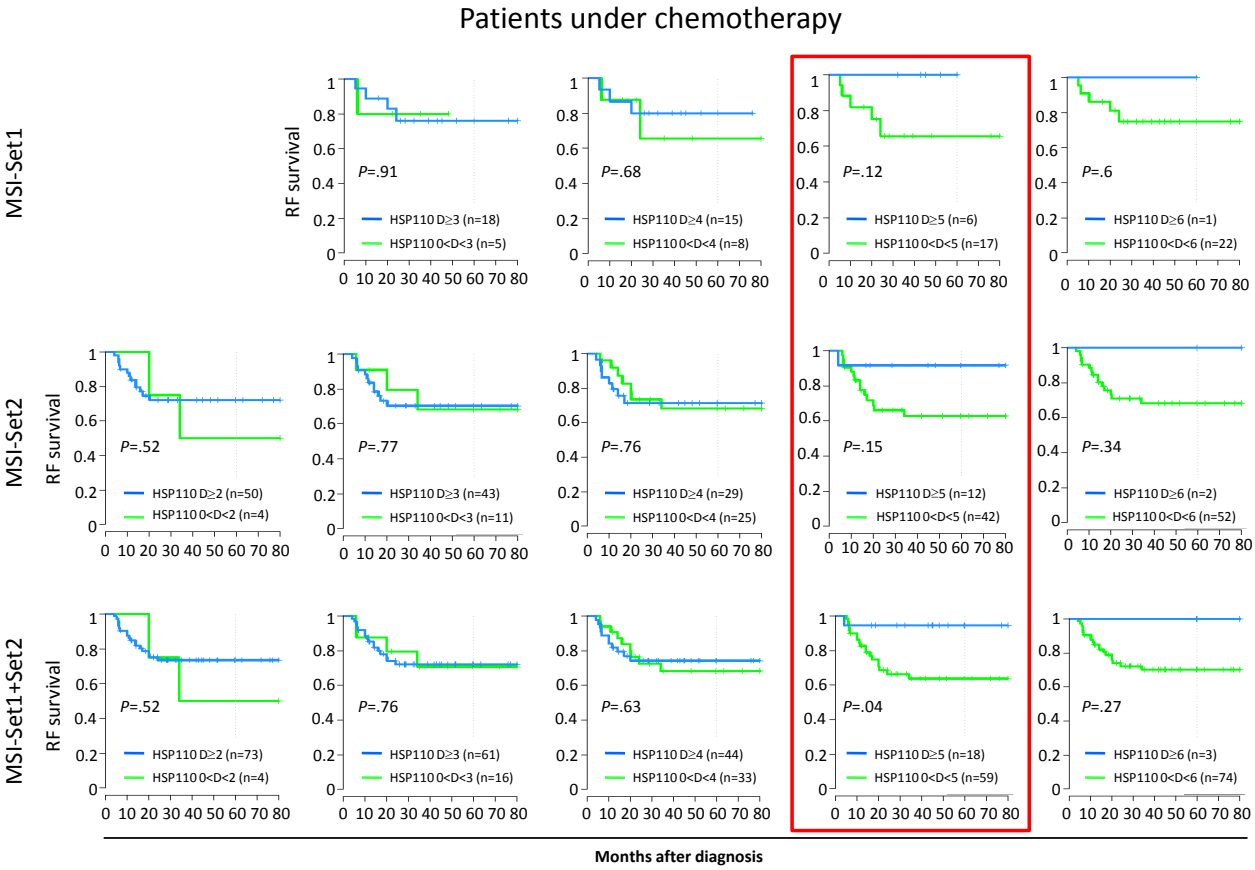
Supplementary Figure S1B



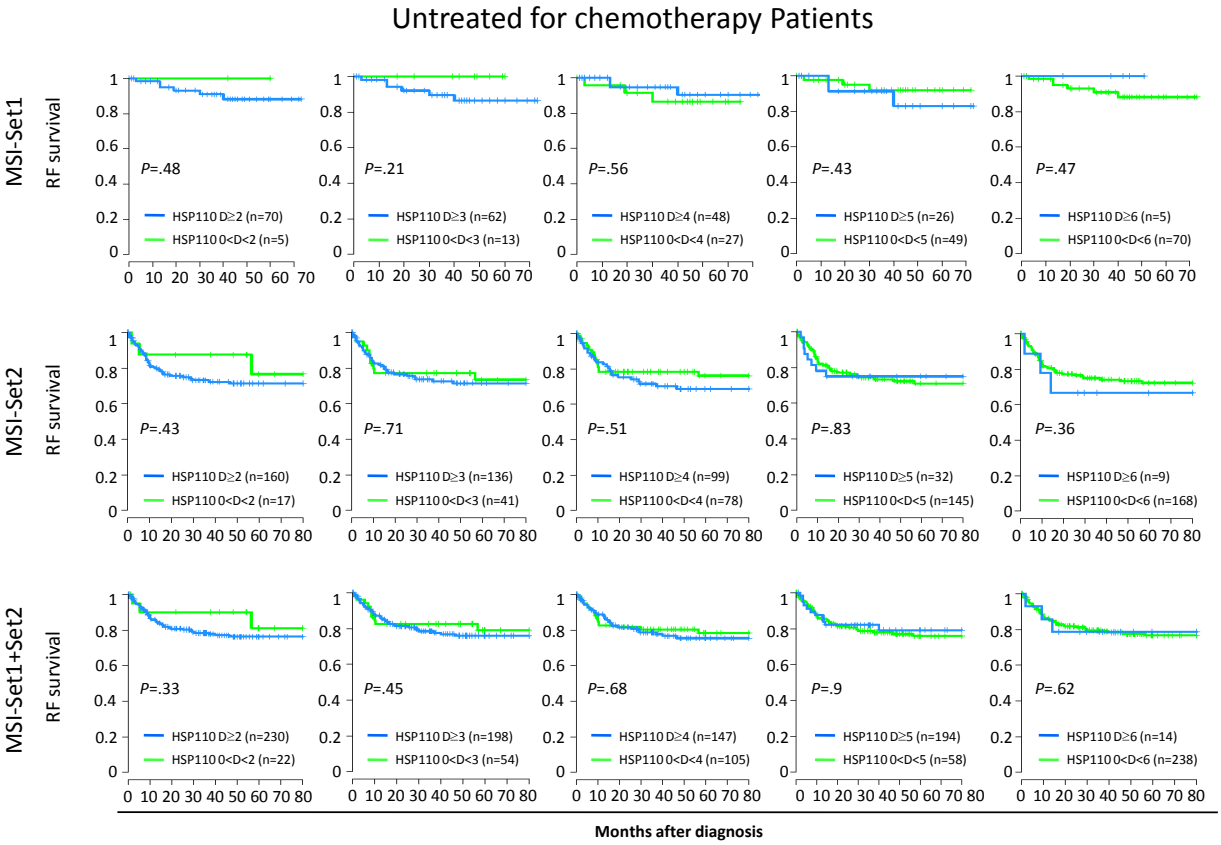
Supplementary Figure S1C



Supplementary Figure S1D



Supplementary Figure S2A



Supplementary Figure S2B

Pronostic de la mutation de PIK3CA intra-sous type MSS vs MSI

Cancer Medicine

Open Access

ORIGINAL RESEARCH

PIK3CA mutations predict recurrence in localized microsatellite stable colon cancer

Gilles Manceau[1,2,3]*, Laetitia Marisa[4]*, Valérie Boige[1,5], Alex Duval[3,6], Marie-Pierre Gaub[7,8], Gérard Milano[9], Janick Selves[10], Sylviane Olschwang[11,12,13,14], Valérie Jooste[15], Michè le Legrain[7], Delphine Lecorre[1], Dominique Guenot[8], Marie-Christine Etienne-Grimaldi[9], Sylvain Kirzin[10], Laurent Martin[16], Come Lepage[15], Anne-Marie Bouvier[15] & Pierre Laurent-Puig[1]

[1]Unité Mixte de Recherche S1147, Paris Sorbonne Cité, Université Paris Descartes, INSERM, Paris, France
[2]Assistance Publique-Hôpitaux de Paris, Service de Chirurgie Digestive et Hépato-Bilio-Pancréatique, Hôpital Pitié-Salpêtrière, Paris, France
[3]Institut Universitaire de Cancérologie, Université Pierre et Marie Curie-Paris 6, Paris, France
[4]"Cartes d'Identité des Tumeurs" Program, Ligue Nationale Contre le Cancer, Paris, France
[5]Institut Gustave Roussy, Villejuif, France
[6]Unité Mixte de Recherche S938, Centre de Recherche Hôpital Saint-Antoine, INSERM, Paris, France
[7]Laboratoire de Biochimie et Biologie Moléculaire, Hôpitaux Universitaires de Strasbourg, Hôpital de Hautepierre, Strasbourg, France
[8]EA 3430 Progression tumorale et microenvironnement. Approches translationnelles et Epidémiologie. Fédération de Médecine Translationnelle de Strasbourg, Université de Strasbourg, Strasbourg, France
[9]Laboratoire d'Oncopharmacologie EA 3836, Centre Antoine Lacassagne, Nice, France
[10]Unité Mixte de Recherche 1037, Centre de Recherche en Cancérologie de Toulouse, Université de Toulouse III, INSERM, Toulouse, France
[11]Unité Mixte de Recherche S910, Faculté de Médecine La Timone, INSERM, Marseille, France
[12]Pôle DACCORD, Hôpital La Timone, Marseille, France
[13]Département d'Oncologie, Hôpital Clairval, Marseille, France
[14]Département de Gastroentérologie, Hôpital Ambroise Paré, Marseille, France
[15]Registre Bourguignon des Cancers Digestifs, INSERM U866, CHU Dijon, France
[16]Service d'anatomie et de cytologie pathologiques, CHU Dijon, France

Keywords
Biomarker, colon cancer, microsatellite instability, mismatch repair, mutations, PIK3CA, prognosis

Correspondence
Pierre Laurent-Puig, UMR-S775 Molecular basis of response to xenobiotics, 45 Rue des Saints-Pères, 75006 Paris, France.
Tel: 33142862081; Fax: 33142862072;
E-mail: pierre.laurent-puig@parisdescartes.fr

Funding Information
This work was supported by a grant from the Ligue Nationale Contre le Cancer.

Received: 8 April 2014; Revised: 28 September 2014; Accepted: 29 September 2014

doi: 10.1002/cam4.370

*Equally contributed as first author.

Abstract

PIK3CA, which encodes the p110α catalytic subunit of PI3Kα, is one of the most frequently altered oncogenes in colon cancer (CC), but its prognostic value is still a matter of debate. Few reports have addressed the association between *PIK3CA* mutations and survival and their results are controversial. In the present study, we aimed to clarify the prognostic impact of *PIK3CA* mutations in stage I–III CC according to mismatch repair status. Fresh frozen tissue samples from two independent cohorts with a total of 826 patients who underwent curative surgical resection of CC were analyzed for microsatellite instability and screened for activating point mutations in exon 9 and 20 of *PIK3CA* by direct sequencing. Overall, 693 tumors (84%) exhibited microsatellite stability (MSS) and 113 samples (14%) harbored *PIK3CA* mutation. In the retrospective training cohort ($n = 433$), patients with *PIK3CA*-mutated MSS tumors ($n = 47$) experienced a significant increased 5-year relapse-free interval compared with *PIK3CA* wild-type MSS tumors ($n = 319$) in univariate analysis (94% vs. 68%, Log-rank $P = 0.0003$) and in multivariate analysis (HR = 0.12; 95% confidence interval, 0.029–0.48; $P = 0.0027$). In the prospective validation cohort ($n = 393$), the favorable prognostic impact of *PIK3CA* mutations in MSS tumors ($n = 327$) was confirmed (83% vs. 67%, Log-rank $P = 0.04$). Our study showed that *PIK3CA* mutations are associated with a good prognosis in patients with MSS stage I–III CC.

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

Introduction

The phosphoinositide 3-kinase (PI3K)/AKT/mTOR signaling pathway is critical for cell growth, survival, and malignant transformation. It is inappropriately activated in many different cancer types [1]. Activation is often mediated by mutations occurring in *PIK3CA*, which encodes the p110α catalytic subunit of a heterodimeric class IA PI3K called PI3Kα. This gene is one of the most frequently mutated genes (16%) in colorectal cancer (CRC), after notably *TP53* (51%), *APC* (37%), and *KRAS* (36%) [2].

In the signal transduction, after ligand binding to a tyrosine kinase receptor, activated PI3Kα phosphorylates phosphatidylinositol 4,5-bisphosphate (PIP2) at the 3′-position of the inositol ring, allowing the generation of phosphatidylinositol 3,4,5-triphosphate (PIP3). This second messenger binds and recruits the 3-phosphoinositide-dependent protein kinase-1 (PDK1), which thereafter activates AKT/PKB, a serine/threonine kinase involved in regulating many biological processes such as cell survival, growth, and metabolism [3]. There are three major mutational hotspots (at codons 542, 545, and 1047) in exons 9 and 20 of *PIK3CA*, partially encoding the helical domain and the C-terminal kinase domain of the protein, respectively [4]. Activating point mutations in these amino-acid residues elevate the enzymatic activity of PI3K and contribute to tumorigenesis through cell proliferation, decreased apoptosis and autophagy, loss of contact inhibition, induction of angiogenesis, and increased tumor invasion [5–7].

As one of the most commonly deregulated pathways in solid human cancers, targeting the PI3K/AKT/mTOR pathway could be of important therapeutic interest [8]. Indeed, a number of PI3K or dual PI3K-mTOR inhibitors have been, or will soon be, introduced into clinical trials as antitumor agents for the treatment of CRC and other malignancies [9–11]. Preclinical studies demonstrated that *PIK3CA* mutations predict response to these agents, presumably due to oncogene addiction [12–15]. However, prognosis of CRC patients harboring *PIK3CA* mutation remains unclear and results from previous studies dealing with this issue seem conflicted [16–22]. Thus, the prognostic role of *PIK3CA* as an independent predictor of recurrence and/or survival in patients with CRC remains to be determined.

The discrepancy of published results could be in part explained by the well-known molecular heterogeneity of CRC. Among the different individualized molecular subgroups, tumors with microsatellite instability (MSI) accounts for ~15% of all CRCs and are characterized by defective DNA mismatch repair and genomic instability. The remaining 85% of microsatellite stable (MSS) CRC display chromosomal instability resulting in hyperploidy associated with allelic losses. Both groups show specific particularities in terms of natural history, tumor location, pathological features, mechanisms of carcinogenesis, and genetic mutation patterns [23].

The aim of this multicenter study was to investigate the prognostic value of *PIK3CA* mutations in nonmetastatic colon cancer (CC) according to MSI status.

Material and Methods

Study population

The French national "Cartes d'Identité des Tumeurs" (CIT) program involved a multicenter cohort of 782 patients with stage I–IV CRC who underwent surgery between 1987 and 2007 in seven centers. Fresh frozen primary tumor tissue samples were retrospectively collected. Clinicopathological data were extracted from the medical charts and centrally reviewed for all patients. This retrospective cohort was used as a training cohort.

For validation purpose, the prospective cohort from the population-based registry of digestive cancer in the Côte-d'Or area (Burgundy, France) previously describe elsewhere was used [24, 25]. Tumor samples from this validation patient cohort included all CC resected between 1998 and 2002 for which frozen tissue material was available and suitable for molecular analysis.

Patients with rectal cancer (located within 15 cm from the anal verge), distant metastasis or who received neoadjuvant therapy (either chemotherapy and/or radiotherapy) were excluded from analysis. Thus, only stage I–III CC according to the 7th edition of the AJCC/UICC tumor-node-metastasis (TNM) classification operated in a curative intent were further considered for evaluation [26]. *PIK3CA* status had also to be determined for each sample. This study was approved by the Institutional Review Boards of all participating centers.

Mutation analysis

Genomic DNA was extracted from fresh frozen tissues. Exons 9 and 20 of the *PIK3CA* gene were selected for screening by direct Sanger sequencing on both strands because of the high frequency of somatic mutations known to be clustered in these regions [4]. All samples found to be mutated were PCR-amplified and sequenced in a second, independent experiment. PCR conditions for amplification and primer sequences are available upon request.

The seven most common somatic mutations of *KRAS* located within codon 12 (G12D, G12V, G12C, G12A,

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

G12S, and G12R) and codon 13 (G13D), as well as *BRAF* V600E mutation, were assessed by allelic discrimination using TaqMan-specific probes as previously described [27].

Microsatellite status and CpG island methylator phenotype analysis

In the retrospective CIT cohort, MSI status was assessed according to the panel of five microsatellites approved by the consensus conference (D2S123, D5S346, D17S250, BAT25, and BAT26) [28]. In the validation cohort, MSI status was determined as previously described [29].

We used the MSP (gel-based methylation-specific PCR) method with the panel of the five markers CACNA1G, IGF2, NEUROG1, RUNX3, and SOCS1 to determine CpG island methylator phenotype (CIMP) status [30]. After DNA bisulfite treatment, two multiplex methylation-specific PCR were performed. Capillary electrophoresis on automatic sequencer (ABI 3130 Genetic analyzer; Applied Biosystems, Foster City, CA, USA) was used for fragment analysis. The methylation status of each gene was determined as detailed by Weisenberg and colleagues [29]. Methylator phenotype-positive cases (CIMP+) were defined as those with ≥3 methylated promoters and CIMP—cases as those with <2 methylated promoters.

Statistic analysis

The distribution of patient and tumor characteristics was compared across cohorts by the Chi-squared test or the Fisher's exact test for categorical data, as appropriate, and by the Welch's *t*-test for continuous data.

Relapse-free interval (RFI) was defined as the time from CC resection to locoregional and/or distant recurrence, whichever came first. Patients alive with no evidence of disease at last follow-up and patients who died without any recurrence were censored. Overall survival (OS) was defined as the period of time between CC surgery and death. Survival curves were plotted according to the method of Kaplan and Meier and differences between survival distributions were assessed by the log-rank test. Univariate and multivariate models for survival analysis were computed using Cox proportional hazards regression. All the variables that were significant in univariate analysis were included in the multivariate model. The proportional hazards assumptions were tested to examine the appropriateness of the models.

All *P*-values were two-sided and statistical significance was assumed for $P \leq 0.05$. All statistical analyses were performed using the *R* statistical environment (http://www.R-project.org). Survival analyses were performed using the *R* package survival.

Results

Patient characteristics

A total of 826 samples from stage I–III CC with successful mutation analysis for *PIK3CA* were obtained from both cohorts ($n = 433$ in the training cohort and $n = 393$ in the validation cohort). Patients' characteristics are depicted in Table 1. Overall, 133 tumors exhibited MSI phenotype (67 [15%] in the training cohort and 66 [17%] in the validation cohort, $P = 0.67$). Among MSS tumors, patients included in the validation cohort were significantly older (69 vs. 73 years, $P = 0.00021$), had less advanced stage (45% vs. 35% for stage III, $P = 0.015$) and were less likely to receive adjuvant chemotherapy (45% vs. 32%, $P = 0.00071$), as compared with the training cohort. Among MSI tumors, patients in the validation cohort were significantly older (74 vs. 80 years, $P = 0.0016$). Significant more patients had a Lynch syndrome in the training cohort (4% vs. 1%, $P = 0.0087$). Only one patient, included in the validation cohort, had a familial adenomatous polyposis.

PIK3CA mutations analysis and associations with other molecular and clinicopathological features

Among the 826 tumors suitable for *PIK3CA*, 113 (14%) tumors displayed a mutation in exon 9 and/or 20 (59 in the training cohort and 54 in the validation cohort). Most *PIK3CA* mutations were located in exon 9 with 38 tumors in the training cohort (64% of the mutated samples) and 32 tumors in the validation cohort (59% of the mutated samples). *PIK3CA* mutations in exon 20 were detected in 6% of the tumors in both cohorts. Only four tumors (three MSS tumors and one MSI tumor, all in the training cohort) harbored *PIK3CA* mutation in both exons 9 and 20, which accounted for 0.5% of all studied samples and 4% of all mutated samples (Table 2).

The determination of *KRAS* and *BRAF* mutational status could be ascertained for 817 (99%) and 780 (94%) tumors, respectively (Table 3). In the whole population, we identified *KRAS* mutation at codon 12 or 13 and V600E *BRAF* mutation in 37% and 11% of cases, respectively, with no statistical difference between the two cohorts, either globally or according to MSI status. As expected, *KRAS* and *BRAF* mutations were mutually exclusive. Concomitant *PIK3CA* and *KRAS* mutations were found in 55 tumors (7%), whereas only 11 tumors (1%) were mutated for both *PIK3CA* and *BRAF*.

We assessed the relationship between *PIK3CA* mutation and clinicopathological and molecular features in each cohort according to MSI status (Table 4). The frequency

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

Table 1. Clinical and pathological characteristics of patients according to microsatellite status in the two cohorts.

Clinical or pathological features	All cases	*n* (%)	Total CIT (%)	Total Dijon (%)	*P* value	MSS CIT (%)	MSS Dijon (%)	*P* value	MSI CIT (%)	MSI Dijon (%)	*P* value
Gender											
Female	826	361 (44)	189 (44)	172 (44)	0.97	156 (43)	130 (40)	0.49	33 (49)	42 (64)	0.13
Male		465 (56)	244 (56)	221 (56)		210 (57)	197 (60)		34 (51)	24 (36)	
Age (years)[1]	826	826	69 [24–96]	73 [33–95]	$1.6 \times 10-6$	69 [25–96]	73 [36–95]	0.00021	74 [24–92]	80 [33–91]	0.0016
Tumor location[2]											
Distal	826	472 (57)	249 (58)	223 (57)	0.88	234 (64)	214 (65)	0.74	15 (22)	9 (14)	0.28
Proximal		354 (43)	184 (42)	170 (43)		132 (36)	113 (35)		52 (78)	57 (86)	
TNM stage											
I	826	76 (9)	33 (8)	43 (11)	0.0052	25 (7)	35 (11)	0.015	8 (12)	8 (12)	0.23
II		431 (52)	211 (49)	220 (56)		175 (48)	176 (54)		36 (54)	44 (67)	
III		319 (39)	189 (44)	130 (33)		166 (45)	116 (35)		23 (34)	14 (21)	
Adjuvant CT											
No	823	540 (66)	257 (59)	283 (72)	0.00014	203 (55)	222 (68)	0.00071	54 (82)	61 (92)	0.12
Yes		283 (34)	175 (41)	108 (28)		163 (45)	103 (32)		12 (18)	5 (8)	
Associated syndrome											
FAP	806	1 (0)	0 (0)	1 (0)	0.0087	0 (0)	1 (0)	0.96	0 (0)	0 (0)	0.0002
Lynch syndrome		22 (3)	18 (4)	4 (1)		0 (0)	0 (0)		18 (35)	4 (6)	
None		783 (97)	395 (96)	388 (99)		362 (100)	326 (100)		33 (65)	62 (94)	

MSS, microsatellite stable; MSI, microsatellite instable; CT, chemotherapy; FAP, familial adenomatous polyposis.

[1]Median [range].

[2]Proximal colon included cecum, ascending colon, hepatic flexure, and transverse colon, and distal colon included splenic flexure, descending, and sigmoid colon.

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

Table 2. Location of *PIK3CA* mutations according to microsatellite status in the two cohorts.

	CIT cohort		Dijon cohort		CIT cohort vs. Dijon cohort	
PIK3CA mutant types	MSS tumors (%)	MSI tumors (%)	MSS tumors (%)	MSI tumors (%)	*P* value MSS tumors	*P* value MSI tumors
Exon 9 mutant	30 (64)	4 (33)	30 (70)	2 (18)	0.24	0.39
Exon 20 mutant	14 (30)	7 (58)	13 (30)	9 (82)		
Exon 9 and 20 mutant	3 (6)	1 (8)	0 (0)	0 (0)		

MSS, microsatellite stable; MSI, microsatellite instable.

Table 3. Molecular characteristics of colon cancers according to microsatellite status in the two cohorts.

Molecular features	All cases	*n* (%)	Total CIT (%)	Total Dijon (%)	*P* value	MSS CIT (%)	MSS Dijon (%)	*P* value	MSI CIT (%)	MSI Dijon (%)	*P* value
PIK3CA status											
Mutant	826	113 (14)	59 (14)	54 (14)	0.96	47 (13)	43 (13)	0.99	12 (18)	11 (17)	0.97
Wild-type		713 (86)	374 (86)	339 (86)		319 (87)	284 (87)		55 (82)	55 (83)	
KRAS status											
Mutant	817	301 (37)	164 (38)	137 (35)	0.37	147 (40)	129 (40)	0.92	17 (27)	8 (12)	0.062
Wild-type		516 (63)	263 (62)	253 (65)		216 (60)	195 (60)		47 (73)	58 (88)	
BRAF status											
Mutant	780	85 (11)	35 (9)	50 (13)	0.12	7 (2)	12 (4)	0.37	28 (43)	38 (58)	0.11
Wild-type		695 (89)	353 (91)	342 (87)		316 (98)	315 (96)		37 (57)	27 (42)	
CIMP status											
CIMP−	763	606 (79)	301 (81)	305 (78)	0.37	278 (90)	288 (88)	0.68	23 (37)	17 (26)	0.26
CIMP+		157 (21)	71 (19)	86 (22)		32 (10)	38 (12)		39 (63)	48 (74)	

MSS, microsatellite stable; MSI, microsatellite instable; CIMP, CpG island methylator phenotype.

of *PIK3CA* mutations was not different between MSS and MSI tumors (13% vs. 18%, $P = 0.36$ in the training cohort and 13% vs. 17%, $P = 0.57$ in the validation cohort). Among MSS tumors, *PI3KCA* mutations were significantly more frequently located in the proximal colon and associated with CIMP-positive tumors in the training cohort (51% vs. 34%, $P = 0.033$ and 23% vs. 8%, $P = 0.0063$ respectively), and significantly more frequent in female in the validation cohort (56% vs. 37%, $P = 0.032$). A significant association between *PIK3CA* and *KRAS* mutations (58% vs. 37%, $P = 0.014$) was found in the validation cohort. Among MSI tumors, *PIK3CA* mutations were significantly more frequent in patients with early-stage disease in the training cohort (33% vs. 7% for stage I, $P = 0.034$).

PI3KCA mutation and patient survival according to microsatellite status

We further examined the prognostic impact of *PIK3CA* mutations in nonmetastatic CC after curative resection. The median follow-up was 51 months (range, 1–192 months) in the training cohort and 61 months (range, 1–143 months) in the validation cohort. The 5-year RFI was similar in the training cohort and in the validation cohort, either globally (63% vs. 68%, $P = 0.24$), and also within MSS and MSI tumors (63% vs. 66%, $P = 0.58$ and 67% vs. 81%, $P = 0.23$, respectively).

Within the MSS CC subgroup of the training cohort, patients with *PIK3CA*-mutated CC experienced a significantly higher RFI than those without *PIK3CA* mutation (5-year RFI 94% vs. 68%, Log-rank $P = 0.0003$; Hazard Ratio [HR] = 0.12; 95% confidence interval [CI], 0.029–0.48) on univariate analysis (Fig. 1A). This finding was confirmed in the validation cohort (5-year RFI 83% vs. 67%, Log-rank $P = 0.04$; HR = 0.45; 95% CI, 0.21–0.97; $P = 0.04$) (Fig. 1B). Similarly, OS was significantly higher in patients with MSS *PIK3CA*-mutated CC than those without *PIK3CA* mutation in the training cohort (5-year OS 88% vs. 75%, Log-rank $P = 0.04$; HR = 0.43; 95% CI, 0.19–0.98) (Fig. 1C), and a strong tendency was also observed in the validation cohort (5-year OS 77% vs. 62%, Log-rank $P = 0.052$; HR = 0.61, 95% CI, 0.37–1) (Fig. 1D). A subgroup analysis according to TNM stage (stage I–II and stage III) was also performed for RFI and OS in both cohorts (Figs. S1 and S2).

We found that the type of *PIK3CA* mutation had no differential effect on RFI (Fig. S3). Five-year RFI of

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

Table 4. Clinical, pathological, and molecular characteristics of tumors according to *PIK3CA* and microsatellite status in the two cohorts.

Features	MSS tumors						MSI tumors					
	CIT cohort			Dijon cohort			CIT cohort			Dijon cohort		
	PIK3CAm	PIK3CAwt	*P* value	PIK3CAm	PIK3CAwt	*P* value	PIK3CAm	PIK3CAwt	*P* value	PIK3CAm	PIK3CAwt	*P* value
Gender												
Female	18 (38)	138 (43)	0.63	24 (56)	106 (37)	0.032	6 (50)	27 (49)	0.79	7 (64)	35 (64)	0.73
Male	29 (62)	181 (57)		19 (44)	178 (63)		6 (50)	28 (51)		4 (36)	20 (36)	
Age (years)[1]	69 [25–96]	69 [37–96]	0.62	73 [49–94]	73 [36–95]	0.75	70 [32–88]	75 [24–92]	0.093	77 [67–88]	80 [33–91]	0.78
TNM.stage												
I	3 (7)	22 (7)	0.54	5 (12)	30 (11)	0.19	4 (33)	4 (7)	0.034	1 (9)	7 (13)	0.89
II	26 (55)	149 (47)		28 (65)	148 (52)		4 (33)	32 (58)		8 (73)	36 (65)	
III	18 (38)	148 (46)		10 (23)	106 (37)		4 (33)	19 (35)		2 (18)	12 (22)	
Tumor.location												
Distal	23 (49)	211 (66)	0.033	23 (53)	191 (67)	0.11	2 (17)	13 (24)	0.89	0 (0)	9 (16)	0.34
Proximal	24 (51)	108 (34)		20 (47)	93 (33)		10 (83)	42 (76)		11 (100)	46 (84)	
Adjuvant CT												
No	28 (60)	175 (55)	0.65	32 (74)	190 (67)	0.45	10 (83)	44 (81)	0.79	10 (91)	51 (93)	0.68
Yes	19 (40)	144 (45)		11 (26)	92 (33)		2 (17)	10 (19)		1 (9)	4 (7)	
KRAS status												
Mutant	22 (48)	125 (39)	0.36	25 (58)	104 (37)	0.014	6 (55)	11 (21)	0.053	2 (18)	6 (11)	0.87
Wild-type	24 (52)	192 (61)		18 (42)	177 (63)		4 (45)	42 (79)		9 (82)	49 (89)	
BRAF status												
Mutant	1 (2)	6 (2)	0.6	1 (2)	11 (4)	0.95	3 (27)	25 (46)	0.41	6 (55)	32 (59)	0.96
Wild-type	46 (98)	270 (98)		42 (98)	273 (96)		8 (73)	29 (54)		5 (45)	22 (41)	
CIMP status												
CIMP–	33 (77)	245 (92)	0.0063	37 (86)	251 (89)	0.8	5 (50)	18 (35)	0.57	3 (27)	14 (26)	1
CIMP+	10 (23)	22 (8)		6 (14)	32 (11)		5 (50)	34 (65)		8 (73)	40 (74)	

CT, chemotherapy; m, mutated; wt, wild-type. Figures in brackets represent the percentages.
[1]Median [range].

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

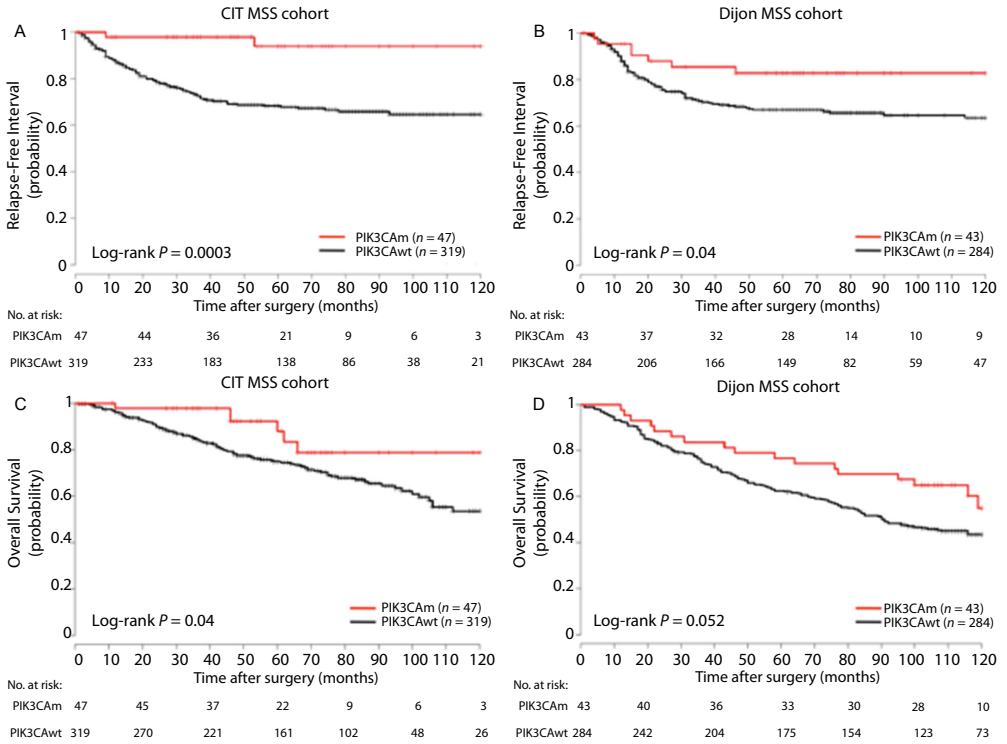


Figure 1. Kaplan–Meier curves for relapse-free interval in the CIT cohort (A) and in the Dijon cohort (B), for overall survival in the CIT cohort (C) and in the Dijon cohort (D) according to *PIK3CA* status in microsatellite stable tumors.

patients with *PIK3CA* mutation in exon 9 or exon 20 were 94% and 93%, respectively, in the training cohort ($P = 0.50$), and 79% and 92% ($P = 0.37$), respectively, in the validation cohort. Notably, no recurrence occurred for the three patients in the CIT cohort whose tumor harbored concomitant *PIK3CA* exon 9 and exon 20 mutations.

We performed a multivariate analysis, adjusting for all significant prognostic variables ($P \leq 0.05$) including TNM stage and *PIK3CA* mutations (Table 5). Risk of recurrence remained significantly lower in the training cohort for *PIK3CA*-mutated MSS CC patients as compared with wild-type *PIK3CA* MSS CC patients (HR = 0.12; 95% CI, 0.029–0.48; $P = 0.0027$). Although not significant, a trend toward decreased recurrence risk was observed for *PIK3CA*-mutated MSS CC patients in the validation cohort (HR = 0.49; 95% CI, 0.23–1.1; $P = 0.074$), and *PIK3CA* mutation variable remained with TNM stage in a backward–forward step selection to reduce the multivariate model to the only informative variables.

Within the MSI tumors, patients with *PIK3CA* mutation in the training cohort experienced a significantly decreased RFI than those without *PIK3CA* mutation on univariate analysis (5-year RFI 67% vs. 87%, Log-rank $P = 0.043$, HR = 3.4; 95% CI, 0.96–12; $P = 0.058$) (Fig. S4A). This finding was not confirmed in the validation cohort (5-year RFI 100% vs. 80%, Log-rank $P = 0.44$) (Fig. S4B).

In the two cohorts, we found that the impact of *PIK3CA* mutation was not different between patients with MSS CC who received adjuvant chemotherapy and those who had no adjuvant treatment (Fig. S5).

Discussion

Over the past several decades, significant progress has been achieved in the treatment of CC, mostly due to improvements in surgical techniques and chemotherapeutic regimens [31, 32]. These advances have contributed to increase cancer-specific survival (CSS), but

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

Table 5. Cox proportional hazards model for RFI among microsatellite stable tumors in the two cohorts.

	Univariate analysis					Multivariate analysis[1]			
Variables	*n*	*n* events	HR	95% CI	*P* value	*n*	HR	95% CI	*P* value
CIT cohort									
TNM stage									
II	366	99	5.1	0.7–38	0.11	366	5.4	0.74–39	0.096
III	366	99	11	1.5–77	0.019		11	1.5–80	0.017
PIK3CA									
Mutated	366	99	0.12	0.029–0.48	0.0029		0.12	0.029–0.48	0.0027
Gender									
Male	366	99	1.1	0.73–1.6	0.67				
Age									
–	365	99	1	0.98–1	0.84				
Tumor location									
Proximal colon	366	99	1	0.66–1.5	0.99				
KRAS									
Mutated	363	99	1.2	0.81–1.8	0.37				
BRAF									
Mutated	323	78	1.5	0.36–6	0.6				
CIMP									
CIMP+	310	74	0.98	0.45–2.1	0.95				
Dijon cohort									
TNM stage									
II	327	97	2.5	0.89–6.8	0.084	327	2.5	0.89–6.8	0.084
III	327	97	4.2	1.5–12	0.0058		4.2	1.5–12	0.0058
PIK3CA									
Mutated	327	97	0.45	0.21–0.97	0.042		0.49	0.23–1.1	0.074
Gender									
Male	327	97	1.1	0.71–1.6	0.77				
Age									
-	327	97	0.99	0.98–1	0.46				
Tumor location									
Proximal colon	327	97	0.7	0.44–1.1	0.11				
KRAS									
Mutated	324	95	1.3	0.86–1.9	0.22				
BRAF									
Mutated	327	97	0.31	0.043–2.2	0.24				
CIMP									
CIMP+	326	97	0.89	0.45–1.8	0.75				

CIMP, CpG island methylator phenotype; HR, hazard ratio; CI, confidence interval; *P* value, Wald test *P* value.
[1]Multivariate models include significant variables ($P < 0.05$).

patient outcome is still difficult to predict. Thus, prognostic biomarkers are required to guide physicians for patient management and follow-up after CC curative resection. Currently, the most recognized prognostic factor in CRC is the AJCC/UICC TNM staging system, defined by the depth of bowel wall invasion and by the presence of metastases in lymph nodes or more distant sites. However, there remains considerable heterogeneity in outcome within the different stages of this classification [33]. Many studies assessed the potential prognostic impact of several somatic mutations including *KRAS*, *BRAF,* and *TP53* mutations after curative surgery [34–36]. So far, none of these has been identified as a reproducible prognostic biomarker, except V600E *BRAF* mutation, which seems to be an independent biomarker of poor prognosis in MSS stage III CCs [37]. Regarding *PI3KCA* mutations, reports are scarce and results are equivocal [16–20]. None used an independent validation group to support their conclusions. Some of them included patients with stage IV disease [16, 18, 19], while others evaluated also patients with rectal cancer [16, 18, 19, 21]. In a series of 418 CRCs, Abubaker and colleagues reported that *PIK3CA* mutations were not associated with OS [16]. Day and colleagues found similar results in a series of 585 stage II–III CRC regarding disease-free survival (DFS) [21]. Nevertheless, this finding is in contradiction with other

 © 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

publications showing that any *PIK3CA* mutation induced a significant decrease in survival [17, 18]. Furthermore, studies that evaluated disease outcome according to the type of *PIK3CA* mutation have led to different conclusions. Fariña Sarasqueta and colleagues found that mutations located in exon 20 conferred poorer survival in stage III patients [20]. But recently, Liao and colleagues emphasized that the prognostic impact of *PI3KCA* mutations was only restricted to the small proportion of CRCs harboring concomitant mutations in both exon 9 and 20 [19]. Finally, in a large series of 627 stage III CC, Ogino and colleagues found that *PIK3CA* mutation was neither a prognostic biomarker nor a predictive biomarker of response to adjuvant chemotherapy [22]. It should be noted that these patients have been enrolled in a randomized controlled trial and received either the Roswell Park regimen of 5-fluoro-uracil (FU)/leucovorin (LV) or the regimen of irinotecan/FU/LV. These adjuvant treatments are, however, not those recommended in case of stage III disease (i.e., a combination of FU and oxaliplatin) [38].

We found in the present study that *PI3KCA* mutations had a favorable prognostic impact in MSS stage I–III CC. This finding is based on two large homogenous groups of patients, excluding those with rectal cancer or distant metastatic disease. Indeed, prognosis and management of patients with rectal cancer differ from that of patients with CC, as neoadjuvant chemoradiotherapy and quality of surgery have a significant impact on local recurrence [39, 40]. Moreover, prognostic biomarkers and chemotherapy regimens differ greatly between localized and advanced CC. Especially, targeted therapies can be used for patients with stage IV CRC, and discussions are still ongoing as to whether *PIK3CA* mutations including those present at the exon 20 are predictive biomarkers of response to anti-EGFR monoclonal antibodies [41].

Although retrospective, our training cohort showed quite similar molecular features than those of our prospective validation cohort, either for all patients or for patients with MSS CC. The mutation rates of *PIK3CA*, *KRAS,* and *BRAF* were consistent with those reported in the literature and were obtained from frozen tumoral tissues [2]. In the literature, *PIK3CA* mutations have been reported to be significantly more frequent in women [24], in elderly patients [21], in proximal tumors [21, 24], in *KRAS*-mutated tumors, [17, 21, 24] and in MSI tumors [16]. In contrast with these reports, we did not find a consistent association between *PIK3CA* mutations and one particular clinical or molecular characteristic in our two cohorts of patients. The fact that the prognostic value of *PIK3CA* mutations was not confirmed on multivariate analysis could be explained by the clinical characteristics of the patients included in the validation cohort. The MSS tumors included in this second cohort were associated with a spontaneous better prognosis with significantly less stage III CC and less indications for adjuvant chemotherapy. One might assume that this group was underpowered to detect a statistical difference.

Very recently, Liao and colleagues highlighted that the use of aspirin after diagnosis among patients with *PIK3CA*-mutated CRC was associated with a significant longer CCS and OS compared with patients with *PIK3CA*-wild-type CRC, with an 82% reduction in CRC deaths and a 45% reduction in deaths from all causes [42]. In this study based on two large prospective combined cohorts, the authors concluded that this gene could be used as a predictive biomarker for the prescription of aspirin therapy in adjuvant setting. Similarly, Domingo and colleagues also found in a large randomized trial comparing rofecoxib with placebo after primary CRC resection that regular use of low-dose aspirin after CRC diagnosis was associated with a reduced rate of recurrence in patients with *PIK3CA*-mutated tumors compared with *PIK3CA*-wild-type tumors (HR = 0.11; 95% CI; 0.001–0.832; $P = 0.027$) [43]. But, in a series of 1487 CRC patients including 185 patients with *PIK3CA*-mutated tumors, Kothari and colleagues did not confirm the relationship between aspirin use and improved survival in patients with stage II-III disease [44]. We were not able to assess patients' survival according to *PIK3CA* status and aspirin treatment since information regarding aspirin therapy was not available in our database. However, definitive conclusion about the predictive value of *PIK3CA* mutations for aspirin treatment in nonmetastatic CRC can only be given with the results of a randomized trial.

The negative prognostic impact of *PIK3CA* mutations in MSI CC was not confirmed in the validation cohort. MSI CRCs have a significantly better prognosis with higher survival rates compared to MSS CRCs [23]. In adjuvant setting, this phenotype is associated with a halving of the risk of recurrence [45]. Therefore, for clinical practice, effective and accurate biomarkers of disease relapse after curative surgery are particularly required for patients with MSS tumors. MSI CRCs also constitute a heterogeneous group of CC, including both tumors with germline mutation of mismatch repair genes and tumors with CIMP and hypermethylation of the *MLH1* gene promoter. The two cohorts were too small to analyze the influence of *PIK3CA* mutations according to the different subgroups of MSI tumors.

As for the study of Ogino and colleagues, we found that *PIK3CA* mutations were not a predictive biomarker for response to adjuvant chemotherapy [22]. In our study, *PIK3CA* mutation was a good prognostic biomarker, either in patients with MSS CC treated with adjuvant chemotherapy or those without adjuvant treatment. The

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

type of adjuvant chemotherapy regimen was not recorded in our study, but it seems likely that a majority of patients received an oxaliplatin-based regimen, as recommended.

Our results seem to be in complete contrast with previous reports in CRC. Indeed, in experimental models, *PIK3CA* gain-of-function mutations have been shown to cause increased phosphorylation of AKT, aberrant activation of the PI3K/AKT/mTOR signaling pathway, and to promote oncogenic transformation. One would expect that proto-oncogene activation (or tumor suppressor gene inactivation) would clinically be associated with aggressive tumor behavior and unfavorable prognosis. However, in understanding of cancer biology, such reasoning seems too simplistic and contradicted by the well-known example of MSI phenotype in CRC [23]. Finally, the good prognostic value of *PIK3CA* mutations has been emphasized in other cancer types, such as breast cancer, endometrial cancer, ovarian clear cell carcinoma, and esophageal squamous cell carcinoma [46–49]. Notably, Kalinsky and colleagues showed in a series of 509 primary breast tumors with a median follow-up of more than 12 years that patients with *PIK3CA*-mutated tumors had a less aggressive phenotype with a significant improvement in OS and CSS [46]. Similarly, Shigaki and colleagues found in a series of 219 patients who had undergone curative resection of stage I–III esophageal squamous cell carcinoma that patients with *PIK3CA* mutations experienced significantly longer DFS, CSS, and OS than those with wild-type *PIK3CA* [49]. One possible explanation is that *PIK3CA* mutations could result in oncogene-induced senescence [46], but the biological mechanisms underlying this effect are still unclear. Finally, in CRC, Baba and colleagues reported in a series of 717 samples that phosphorylated AKT expression was significantly associated with *PIK3CA* mutations, and that patients with AKT-activated tumors had a significantly improved CSS in multivariate analysis [50]. Surprisingly in this study, the authors used the data from the Nurses' Health Study and the Health Professionals Follow-up Study, which are the same cohorts as those used in the studies of Ogino and colleagues and Liao and colleagues that led to diametrically opposite conclusions regarding the prognostic impact of *PIK3CA* mutations [17, 19].

In summary, our study suggests that *PIK3CA* mutations are associated with better outcome in patients with resected MSS stage I-III CC. Our results may have clinical implications and provide useful information for the postoperative management of patients. Those with high-risk stage II or stage III *PIK3CA*-mutated MSS CC may not require adjuvant chemotherapy. Nevertheless, the mechanisms explaining this favorable prognostic impact of *PIK3CA* mutations remain to be elucidated. We cannot exclude that this could be the reflect of the predictive value of aspirin therapy. These results warrant confirmation in further translational studies.

Acknowledgments

None.

Conflict of Interest

None declared.

References

1. Vivanco, I., and C. L. Sawyers. 2002. The phosphatidylinositol 3-Kinase AKT pathway in human cancer. Nat. Rev. Cancer 2:489–501.
2. Forbes, SA., N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, et al. 2011. COSMIC: mining complete cancer genomes in the catalogue of somatic mutations in cancer. Nucleic Acids Res. 39:D945–D950.
3. Manning, B. D., and L. C. Cantley. 2007. AKT/PKB signaling: navigating downstream. Cell 129:1261–1274.
4. Samuels, Y., Z. Wang, A. Bardelli, N. Silliman, J. Ptak, S. Szabo, et al. 2004. High frequency of mutations of the PIK3CA gene in human cancers. Science 304:554.
5. Samuels, Y., L.A., Diaz Jr., O. Schmidt-Kittler, J. M. Cummins, L. Delong, I. Cheong, et al. 2005. Mutant PIK3CA promotes cell growth and invasion of human cancer cells. Cancer Cell 7:561–573.
6. Ikenoue, T., F. Kanai, Y. Hikiba, T. Obata, Y. Tanaka, J. Imamura, et al. 2005. Functional analysis of PIK3CA gene mutations in human colorectal cancer. Cancer Res. 65:4562–4567.
7. Guo, X. N., A. Rajput, R. Rose, J. Hauser, A. Beko, K. Kuropatwinski, et al. 2007. Mutant PIK3CA-bearing colon cancer cells display increased metastasis in an orthotopic model. Cancer Res. 67:5851–5858.
8. Engelman, J. A. 2009. Targeting PI3K signalling in cancer: opportunities, challenges and limitations. Nat. Rev. Cancer 9:550–562.
9. Ihle, N., T. R. Williams, S. Chow, W. Chew, M. I. Berggren, G. Paine-Murrieta, et al. 2004. Molecular pharmacology and antitumor activity of PX-866, a novel inhibitor of phosphoinositide-3-kinase signaling. Mol. Cancer Ther. 3:763–772.
10. Howes, A. L., G. G. Chiang, E. S. Lang, C. B. Ho, G. Powis, K. Vuori, and R. T. Abraham. 2007. The phosphatidylinositol 3-kinase inhibitor, PX-866, is a potent inhibitor of cancer cell motility and growth in three-dimensional cultures. Mol. Cancer Ther. 6:2505–2514.
11. Ihle, N. T., and G. Powis. 2010. Inhibitors of phosphatidylinositol-3-kinase in cancer therapy. Mol. Aspects Med. 31:135–144.

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

12. Ihle, N., T. R. Lemos Jr , P. Wipf, A. Yacoub, C. Mitchell, D. Siwak, et al. 2009. Mutations in the phosphatidylinositol-3-kinase pathway predict for antitumor activity of the inhibitor PX-866 whereas oncogenic Ras is a dominant predictor for resistance. Cancer Res. 69:143–150.
13. Tanaka, H., M. Yoshida, H. Tanimura, T. Fujii, K. Sakata, Y. Tachibana, et al. 2011. The selective class I PI3K inhibitor CH5132799 targets human cancers harboring oncogenic PIK3CA mutations. Clin. Cancer Res. 17:3272–3281.
14. Zhang, H., G. Liu, M. Dziubinski, Z. Yang, S. P. Ethier, and G. Wu. 2008. Comprehensive analysis of oncogenic effects of PIK3CA mutations in human mammary epithelial cells. Breast Cancer Res. Treat. 112:217–227.
15. Serra, V., B. Markman, M. Scaltriti, P. J. Eichhorn, V. Valero, M. Guzman, et al. 2008. NVP-BEZ235, a dual PI3K/mTOR inhibitor, prevents PI3K signaling and inhibits the growth of cancer cells with activating PI3K mutations. Cancer Res. 68: 8022–8030.
16. Abubaker, J., P. Bavi, S. Al-Harbi, M. Ibrahim, A. K. Siraj, N. Al-Sanea, et al. 2008. Clinicopathological analysis of colorectal cancers with PIK3CA mutations in Middle Eastern population. Oncogene 27:3539–3545.
17. Ogino, S., K. Nosho, G. J. Kirkner, K. Shima, N. Irahara, S. Kure, et al. 2009. PIK3CA mutation is associated with poor prognosis among patients with curatively resected colon cancer. J. Clin. Oncol. 27:1477–1484.
18. Kato, S., S. Iida, T. Higuchi, T. Ishikawa, Y. Takagi, M. Yasuno, et al. 2007. PIK3CA mutation is predictive of poor survival in patients with colorectal cancer. Int. J. Cancer 121:1771–1778.
19. Liao, X., T. Morikawa, P. Lochhead, Y. Imamura, A. Kuchiba, M. Yamauchi, et al. 2012. Prognostic role of PIK3CA mutation in colorectal cancer: cohort study and literature review. Clin. Cancer Res. 18:2257–2268.
20. Farina Sarasqueta, A. E., C. Zeestraten, T. vanWezel , G. Van Lijnschoten , R. Van Eijk , J. W. Dekker, et al. 2011. PIK3CA kinase domain mutation identifies a subgroup of stage III colon cancer patients with poor prognosis. Cell. Oncol. (Dordr.) 34:523–531.
21. Day, F. L., R. N. Jorissen, L. Lipton, D. Mouradov, A. Sakthianandeswaren, M. Christie, et al. 2013. PIK3CA and PTEN gene and exon mutation-specific clinicopathologic and molecular associations in colorectal cancer. Clin. Cancer Res. 19:3285–3296.
22. Ogino, S., X. Liao, Y. Imamura, M. Yamauchi, N. J. McCleary, K. Ng, et al. 2013. Predictive and prognostic analysis of PIK3CA mutation in stage III colon cancer intergroup trial. J. Natl. Cancer Inst. 105:1789–1798.
23. Popat, S., R. Hubner, and R. S. Houlston. 2005. Systematic review of microsatellite instability and colorectal cancer prognosis. J. Clin. Oncol. 23:609–618.
24. Barault, L., N. Veyrie, V. Jooste, D. Lecorre, C. Chapusot, J. M. Ferraz, et al. 2008. Mutations in the RAS-MAPK, PI(3)K (phosphatidylinositol-3-OH kinase) signaling network correlate with poor survival in a population-based series of colon cancers. Int. J. Cancer 122:2255–2259.
25. Chauvenet, M., V. Cottet, C. Lepage, V. Jooste, J. Faivre, and A. M. Bouvier. 2011. Trends in colorectal cancer incidence: a period and birth-cohort analysis in a well-defined French population. BMC Cancer 11:282.
26. Edge, S. B., D. R. Byrd, C. C. Compton, A. G. Fritz, F. L. Green, and A. Trotti. 2009. Colon and rectum. Pp. 143–164 *in* S. Edge, D. Byrd, and C. Compton, et al., eds. AJCC cancer staging manual. 7th ed. Springer, New York, NY.
27. Lievre, A., J. B. Bachet, V. Boige, A. Cayre, D. Le Corre, E. Buc, et al. 2008. KRAS mutations as an independent prognostic factor in patients with advanced colorectal cancer treated with cetuximab. J. Clin. Oncol. 26:374–379.
28. Boland, C. R., S. N. Thibodeau, S. R. Hamilton, D. Sidransky, J. R. Eshleman, R. W. Burt, et al. 1998. A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition: development of international criteria for the determination of microsatellite instability in colorectal cancer. Cancer Res. 58:5248–5257.
29. Chapusot, C., L. Martin, P. Laurent-Puig, T. Ponnelle, N. Cheynel, A. M. Bouvier, et al. 2004. What is the best way to assess microsatellite instability status in colorectal cancer? Study on a population base of 462 colorectal cancers. Am. J. Surg. Pathol. 28:1553–1559.
30. Weisenberger, D., J. K. D. Siegmund, M. Campan, J. Young, T. I. Long, M. A. Faasse, et al. 2006. CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. Nat. Genet. 38:787–793.
31. West, N. P., E. J. Morris, O. Rotimi, A. Cairns, P. J. Finan, and P. Quirke. 2008. Pathology grading of colon cancer surgical resection and its association with survival: a retrospective observational study. Lancet Oncol. 9:857–865.
32. Andre, T., C. Boni, M. Navarro, J. Tabernero, T. Hickish, C. Topham, et al. 2009. Improved overall survival with oxaliplatin, fluorouracil, and leucovorin as adjuvant treatment in stage II or III colon cancer in the MOSAIC trial. J. Clin. Oncol. 27:3109–3116.
33. Gunderson, L., L. J. M. Jessup, D. J. Sargent, F. L. Greene, and A. K. Stewart. 2010. Revised TN categorization for colon cancer based on national survival outcomes data. J. Clin. Oncol. 28:264–271.
34. Andreyev, H., J. A. R. Norman, D. Cunningham, J. Oates, B. R. Dix, B. J. Iacopetta, et al. 2001. Kirsten ras mutations in patients with colorectal cancer: the 'RASCAL II' study. Br. J. Cancer 85:692–696.
35. Ogino, S. J., A. Meyerhardt, N. Irahara, D. Niedzwiecki, D. Hollis, L. B. Saltz, et al. 2009. KRAS mutation in stage III

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

colon cancer and clinical outcome following intergroup trial CALGB 89803. Clin. Cancer Res. 15:7322–7329.
36. Russo, A., V. Bazan, B. Iacopetta, D. Kerr, T. Soussi, and N. Gebbia. 2005. The TP53 colorectal cancer international collaborative study on the prognostic and predictive significance of p53 mutation: influence of tumor site, type of mutation, and adjuvant treatment. J. Clin. Oncol. 23:7518–7528.
37. Ogino, S., K. Shima, J. A. Meyerhardt, N. J. McCleary, K. Ng, D. Hollis, et al. 2012. Predictive and prognostic roles of BRAF mutation in stage III colon cancer: results from intergroup trial CALGB 89803. Clin. Cancer Res. 18:890–900.
38. Schmoll, H., J. E. Van Cutsem, A. Stein, V. Valentini, B. Glimelius, K. Haustermans, et al. 2012. ESMO Consensus Guidelines for management of patients with colon and rectal cancer. a personalized approach to clinical decision making. Ann. Oncol. 23:2479–2516.
39. Quirke, P., R. Steele, J. Monson, R. Grieve, S. Khanna, J. Couture, et al. 2009. Effect of the plane of surgery achieved on local recurrence in patients with operable rectal cancer: a prospective study using data from the MRC CR07 and NCIC-CTG CO16 randomised clinical trial. Lancet 373:821–828.
40. Gerard, J., P. T. Conroy, F. Bonnetain, O. Bouche, O. Chapet, M. T. Closon-Dejardin, et al. 2006. Preoperative radiotherapy with or without concurrent fluorouracil and leucovorin in T3-4 rectal cancers: results of FFCD 9203. J. Clin. Oncol. 24:4620–4625.
41. De Roock, W., B. Claes, D. Bernasconi, J. De Schutter, B. Biesmans, G. Fountzilas, et al. 2010. Effects of KRAS, BRAF, NRAS, and PIK3CA mutations on the efficacy of cetuximab plus chemotherapy in chemotherapy-refractory metastatic colorectal cancer: a retrospective consortium analysis. Lancet Oncol. 11:753–762.
42. Liao, X., P. Lochhead, R. Nishihara, T. Morikawa, A. Kuchiba, M. Yamauchi, et al. 2012. Aspirin use, tumor PIK3CA mutation, and colorectal-cancer survival. N. Engl. J. Med. 367:1596–1606.
43. Domingo, E., D. N. Church, O. Sieber, R. Ramamoorthy, Y. Yanagisawa, E. Johnstone, et al. 2013. Evaluation of PIK3CA mutation as a predictor of benefit from nonsteroidal anti-inflammatory drug therapy in colorectal cancer. J. Clin. Oncol. 31:4297–4305.
44. Kothari, N, R. D. Kim, P. Gibbs, T. J. Yeatman, M. J. Schell, J. Desai, et al. 2014. Regular aspirin (ASA) use and survival in patients with PIK3CA-mutated metastatic colorectal cancer (CRC). J. Clin. Oncol. 32(Suppl. 3).
45. Hutchins, G., K. Southward, K. Handley, L. Magill, C. Beaumont, J. Stahlschmidt, et al. 2011. Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer. J. Clin. Oncol. 29:1261–1270.
46. Kalinsky, K., L. M. Jacks, A. Heguy, S. Patil, M. Drobnjak, U. K. Bhanot, et al. 2009. PIK3CA mutation associates with improved outcome in breast cancer. Clin. Cancer Res. 15:5049–5059.
47. Dong, Y., X. Yang, O. Wong, X. Zhang, Y. Liang, Y. Zhang, et al. 2012. PIK3CA mutations in endometrial carcinomas in Chinese women: phosphatidylinositol 3′-kinase pathway alterations might be associated with favorable prognosis. Hum. Pathol. 43:1197–1205.
48. Rahman, M., K. Nakayama, M. T. Rahman, N. Nakayama, M. Ishikawa, A. Katagiri, et al. 2012. Clinicopathologic and biological analysis of PIK3CA mutation in ovarian clear cell carcinoma. Hum. Pathol. 43:2197–2206.
49. Shigaki, H., Y. Baba, M. Watanabe, A. Murata, T. Ishimoto, M. Iwatsuki, et al. 2013. PIK3CA mutation is associated with a favorable prognosis among patients with curatively resected esophageal squamous cell carcinoma. Clin. Cancer Res. 19:2451–2459.
50. Baba, Y., K. Nosho, K. Shima, M. Hayashi, J. A. Meyerhardt, A. T. Chan, et al. 2011. Phosphorylated AKT expression is associated with PIK3CA mutation, low stage, and favorable outcome in 717 colorectal cancers. Cancer 117:1399–1408.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Figure S1. Kaplan-Meier curves for relapse-free interval in the CIT cohort (A) and in the Dijon cohort (B), for overall survival in the CIT cohort (C) and in the Dijon cohort (D) according to *PIK3CA* status in stage I–II microsatellite stable tumors.
Figure S2. Kaplan-Meier curves for relapse-free interval in the CIT cohort (A) and in the Dijon cohort (B), for overall survival in the CIT cohort (C) and in the Dijon cohort (D) according to *PIK3CA* status in stage III microsatellite stable tumors.
Figure S3. Kaplan-Meier curves for relapse-free interval in the CIT cohort (A) and in the Dijon cohort (B), according to *PIK3CA* exon-specific mutation in microsatellite stable tumors.
Figure S4. Kaplan-Meier curves for relapse-free interval in the CIT cohort (A) and in the Dijon cohort (B), according to *PIK3CA* status in microsatellite instable tumors.
Figure S5. Kaplan-Meier curves for relapse-free interval according to *PIK3CA* status in microsatellite stable tumors for patients operated on without adjuvant chemothrapy in the CIT cohort (A) and in the Dijon cohort (D), and for those treated with adjuvant chemotherapy in the CIT cohort (B) and in the Dijon cohort (D).

© 2015 The Authors. *Cancer Medicine* published by John Wiley & Sons Ltd.

Table des figures

1.1 Incidence du cancer colorectal dans le monde 15
1.2 Formes héréditaires, familiales et sporadiques des cancers colorectaux. . . . 17
1.3 Anatomie et structure du côlon. 18
1.4 Séquence de progression tumorale adénome-carcinome. 19
1.5 Schéma d'une crypte colique 21
1.6 Modèles de développement tumoral 23
1.7 Classification clinique selon le stade TNM et sa valeur pronostique. 27
1.8 Traitements des cancers colorectaux 28

2.1 Chronologie des principales découvertes moléculaires liées à l'oncologie du cancer colorectal 32
2.2 Le Modèle de tumorigenèse Colorectal de (Fearon and Vogelstein, 1990). . . 33
2.3 Schéma des différentes instabilités et leur répartition dans la population des CRC. 36
2.4 Système de réparation des mésappariements de l'ADN et l'impact de sa défaillance dans les tumeurs de type MSI. 39
2.5 Modèles de tumorigenèse colique en fonction des instabilités 43
2.6 Les classifications moléculaires proposées intégrant les instabilités génomiques 45

3.1 Les principales avancées de la génétique et génomique 48
3.2 Le principe des puces à ADN et les différentes types de puces 50
3.3 Principe du séquencage de nouvelle génération 51
3.4 Principe des puces Affymetrix et CGH. 52
3.5 Différence entre la distance Euclidienne et la distance de Pearson et comparaison des méthodes d'agglomération. 57
3.6 Exemple de dendrogramme et de *heatmap* obtenus par classification hiérarchique agglomérative 58
3.7 Ensemble des variables décrivant la qualité d'une prédiction. 61
3.8 Principaux critères pour évaluer un test génétique selon le projet *AACE model* 62
9 Schéma de l'approche de classification et des caractéristiques des différents sous-types 74
10 Les six classifications utilisées pour établir la classification consensus 105
11 Description globale et détaillée de l'approche de classification consensus . . 106
12 Résumé des résultats des quatre sous-types consensus 107
13 Comparaison des classifications indépendantes et de la classification consensus 108

Bibliographie

Aaltonen, L. A., Peltomäki, P., Leach, F. S., Sistonen, P., Pylkkänen, L., Mecklin, J. P., Järvinen, H., Powell, S. M., Jen, J., and Hamilton, S. R. (1993). Clues to the pathogenesis of familial colorectal cancer. *Science (New York, N.Y.)*, 260(5109) :812–816.

Altman, D. G., McShane, L. M., Sauerbrei, W., and Taube, S. E. (2012). Reporting Recommendations for Tumor Marker Prognostic Studies (REMARK) : explanation and elaboration. *PLoS medicine*, 9(5) :e1001216.

Anderson, E. C., Hessman, C., Levin, T. G., Monroe, M. M., and Wong, M. H. (2011). The role of colorectal cancer stem cells in metastatic disease and therapeutic response. *Cancers*, 3(1) :319–339.

André, T., Boni, C., Mounedji-Boudiaf, L., Navarro, M., Tabernero, J., Hickish, T., Topham, C., Zaninelli, M., Clingan, P., Bridgewater, J., Tabah-Fisch, I., and de Gramont, A. (2004). Oxaliplatin, Fluorouracil, and Leucovorin as Adjuvant Treatment for Colon Cancer. *New England Journal of Medicine*, 350(23) :2343–2351.

Armelao, F. and de Pretis, G. (2014). Familial colorectal cancer : A review. *World Journal of Gastroenterology : WJG*, 20(28) :9292–9298.

Balaguer, F., Moreira, L., Lozano, J. J., Link, A., Ramirez, G., Shen, Y., Cuatrecasas, M., Arnold, M., Meltzer, S. J., Syngal, S., Stoffel, E., Jover, R., Llor, X., Castells, A., Boland, C. R., Gironella, M., and Goel, A. (2011). Colorectal cancers with microsatellite instability display unique miRNA profiles. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 17(19) :6239–6249.

Barker, N., Wetering, M. v. d., and Clevers, H. (2008). The intestinal stem cell. *Genes & Development*, 22(14) :1856–1864.

Beissbarth, T. and Speed, T. P. (2004). GOstat : find statistically overrepresented Gene Ontologies within a group of genes. *Bioinformatics (Oxford, England)*, 20(9) :1464–1465.

Boland, C. R. and Goel, A. (2010). Microsatellite Instability in Colorectal Cancer. *Gastroenterology*, 138(6) :2073–2087.e3.

Boland, C. R., Thibodeau, S. N., Hamilton, S. R., Sidransky, D., Eshleman, J. R., Burt, R. W., Meltzer, S. J., Rodriguez-Bigas, M. A., Fodde, R., Ranzani, G. N., and Srivastava, S. (1998). A National Cancer Institute Workshop on Microsatellite Instability for cancer detection and familial predisposition : development of international criteria for the determination of microsatellite instability in colorectal cancer. *Cancer Research*, 58(22) :5248–5257.

Brazma, A. and Vilo, J. (2000). Gene expression data analysis. *FEBS letters*, 480(1) :17–24.

Budinska, E., Popovici, V., Tejpar, S., D'Ario, G., Lapique, N., Sikora, K. O., Di Narzo, A. F., Yan, P., Hodgson, J. G., Weinrich, S., Bosman, F., Roth, A., and Delorenzi, M. (2013). Gene expression patterns unveil a new level of molecular heterogeneity in colorectal cancer. *The Journal of Pathology*, 231(1) :63–76.

Burt, R. (2007). Inheritance of Colorectal Cancer. *Drug discovery today. Disease mechanisms*, 4(4) :293–300.

Camps, J., Armengol, G., Rey, J. d., Lozano, J. J., Vauhkonen, H., Prat, E., Egozcue, J., Sumoy, L., Knuutila, S., and Miró, R. (2006). Genome-wide differences between microsatellite stable and unstable colorectal tumors. *Carcinogenesis*, 27(3) :419–428.

Cancer Genome Atlas Network (2012). Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, 487(7407) :330–337.

Cardoso, J., Boer, J., Morreau, H., and Fodde, R. (2007). Expression and genomic profiling of colorectal cancer. *Biochimica Et Biophysica Acta*, 1775(1) :103–137.

Chan, S. K., Griffith, O. L., Tai, I. T., and Jones, S. J. M. (2008). Meta-analysis of Colorectal Cancer Gene Expression Profiling Studies Identifies Consistently Reported Candidate Biomarkers. *Cancer Epidemiology Biomarkers & Prevention*, 17(3) :543–552.

Chee, C. E. and Meropol, N. J. (2014). Current status of gene expression profiling to assist decision making in stage II colon cancer. *The Oncologist*, 19(7) :704–711.

Cleator, S. J., Powles, T. J., Dexter, T., Fulford, L., Mackay, A., Smith, I. E., Valgeirsson, H., Ashworth, A., and Dowsett, M. (2006). The effect of the stromal component of breast tumours on prediction of clinical outcome using gene expression microarray analysis. *Breast Cancer Research*, 8(3) :R32.

Collins, F. S., Green, E. D., Guttmacher, A. E., Guyer, M. S., and US National Human Genome Research Institute (2003). A vision for the future of genomics research. *Nature*, 422(6934) :835–847.

Collura, A., Lagrange, A., Svrcek, M., Marisa, L., Buhard, O., Guilloux, A., Wanherdrick, K., Dorard, C., Taieb, A., Saget, A., Loh, M., Soong, R., Zeps, N., Platell, C., Mews, A., Iacopetta, B., De Thonel, A., Seigneuric, R., Marcion, G., Chapusot, C., Lepage, C., Bouvier, A., Gaub, M., Milano, G., Selves, J., Senet, P., Delarue, P., Arzouk, H., Lacoste, C., Coquelle, A., Bengrine ?Lefèvre, L., Tournigand, C., Lefèvre, J. H., Parc, Y., Biard, D. S., Fléjou, J., Garrido, C., and Duval, A. (2014). Patients With Colorectal Tumors With Microsatellite Instability and Large Deletions in HSP110 T17 Have Improved Response to 5-Fluorouracil ?Based Chemotherapy. *Gastroenterology*, 146(2) :401–411.e1.

Cover, T. and Hart, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1) :21 –27.

Cuilliere-Dartigues, P., El-Bchiri, J., Krimi, A., Buhard, O., Fontanges, P., Fléjou, J.-F., Hamelin, R., and Duval, A. (2006). TCF-4 isoforms absent in TCF-4 mutated MSI-H colorectal cancer cells colocalize with nuclear CtBP and repress TCF-4-mediated transcription. *Oncogene*, 25(32) :4441–4448.

Curtin, K., Slattery, M. L., and Samowitz, W. S. (2011). CpG Island Methylation in Colorectal Cancer : Past, Present and Future. *Pathology Research International*, 2011 :e902674.

Dalerba, P., Kalisky, T., Sahoo, D., Rajendran, P. S., Rothenberg, M. E., Leyrat, A. A., Sim, S., Okamoto, J., Johnston, D. M., Qian, D., Zabala, M., Bueno, J., Neff, N. F., Wang, J., Shelton, A. A., Visser, B., Hisamori, S., Shimono, Y., van de Wetering, M., Clevers, H., Clarke, M. F., and Quake, S. R. (2011). Single-cell dissection of transcriptional heterogeneity in human colon tumors. *Nature Biotechnology*, 29(12) :1120–1127.

Di Franco Simone, S., Mancuso, P., Benfante, A., Spina, M., Iovino, F., Dieli, F., Stassi, G., and Todaro, M. (2011). Colon Cancer Stem Cells : Bench-to-Bedside New Therapeutical Approaches in Clinical Oncology for Disease Breakdown. *Cancers*, 3(2) :1957–1974.

Ding, C. and Peng, H. (2005). Minimum redundancy feature selection from microarray gene expression data. *Journal of Bioinformatics and Computational Biology*, 3(2) :185–205.

Donehower, L. A., Creighton, C. J., Schultz, N., Shinbrot, E., Chang, K., Gunaratne, P. H., Muzny, D., Sander, C., Hamilton, S. R., Gibbs, R. A., and Wheeler, D. (2013). MLH1-silenced and non-silenced subgroups of hypermutated colorectal carcinomas have distinct mutational landscapes. *The Journal of pathology*, 229(1) :99–110.

Dorard, C., de Thonel, A., Collura, A., Marisa, L., Svrcek, M., Lagrange, A., Jego, G., Wanherdrick, K., Joly, A. L., Buhard, O., Gobbo, J., Penard-Lacronique, V., Zouali, H., Tubacher, E., Kirzin, S., Selves, J., Milano, G., Etienne-Grimaldi, M.-C., Bengrine-Lefèvre, L., Louvet, C., Tournigand, C., Lefèvre, J. H., Parc, Y., Tiret, E., Fléjou, J.-F., Gaub, M.-P., Garrido, C., and Duval, A. (2011). Expression of a mutant HSP110 sensitizes colorectal cancer cells to chemotherapy and improves disease prognosis. *Nature medicine*, 17(10) :1283–1289.

Dudoit, S. and Gentleman, R. (2002). Cluster analysis in DNA microarray experiments. *Bioconductor Short Course Winter*.

Dupuy, A. and Simon, R. M. (2007). Critical Review of Published Microarray Studies for Cancer Outcome and Guidelines on Statistical Analysis and Reporting. *JNCI Journal of the National Cancer Institute*, 99(2) :147–157.

Duval, A. and Hamelin, R. (2002). Mutations at coding repeat sequences in mismatch repair-deficient human cancers : toward a new concept of target genes for instability. *Cancer Research*, 62(9) :2447–2454.

Fallik, D., Borrini, F., Boige, V., Viguier, J., Jacob, S., Miquel, C., Sabourin, J.-C., Ducreux, M., and Praz, F. (2003). Microsatellite instability is a predictive factor of the tumor response to irinotecan in patients with advanced colorectal cancer. *Cancer Research*, 63(18) :5738–5744.

Fang, M., Ou, J., Hutchinson, L., and Green, M. R. (2014). The BRAF Oncoprotein Functions through the Transcriptional Repressor MAFG to Mediate the CpG Island Methylator Phenotype. *Molecular Cell*, 55(6) :904–915.

Fearon, E. R. (2011). Molecular Genetics of Colorectal Cancer. *Annual Review of Pathology : Mechanisms of Disease*, 6(1) :479–507.

Fearon, E. R. and Vogelstein, B. (1990). A genetic model for colorectal tumorigenesis. *Cell*, 61(5) :759–767.

Fishel, R., Lescoe, M. K., Rao, M. R., Copeland, N. G., Jenkins, N. A., Garber, J., Kane, M., and Kolodner, R. (1993). The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell*, 75(5) :1027–1038.

Fleming, M., Ravula, S., Tatishchev, S. F., and Wang, H. L. (2012). Colorectal carcinoma : Pathologic aspects. *Journal of Gastrointestinal Oncology*, 3(3) :153–173.

Geigl, J. B., Obenauf, A. C., Schwarzbraun, T., and Speicher, M. R. (2008). Defining 'chromosomal instability'. *Trends in genetics : TIG*, 24(2) :64–69.

Gerlinger, M., Rowan, A. J., Horswell, S., Larkin, J., Endesfelder, D., Gronroos, E., Martinez, P., Matthews, N., Stewart, A., Tarpey, P., Varela, I., Phillimore, B., Begum, S., McDonald, N. Q., Butler, A., Jones, D., Raine, K., Latimer, C., Santos, C. R., Nohadani, M., Eklund, A. C., Spencer-Dene, B., Clark, G., Pickering, L., Stamp, G., Gore, M., Szallasi, Z., Downward, J., Futreal, P. A., and Swanton, C. (2012). Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England Journal of Medicine*, 366(10) :883–892.

Gray, R. G., Quirke, P., Handley, K., Lopatin, M., Magill, L., Baehner, F. L., Beaumont, C., Clark-Langone, K. M., Yoshizawa, C. N., Lee, M., Watson, D., Shak, S., and Kerr, D. J. (2011). Validation study of a quantitative multigene reverse transcriptase-polymerase chain reaction assay for assessment of recurrence risk in patients with stage II colon cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 29(35) :4611–4619.

Greenman, C., Stephens, P., Smith, R., Dalgliesh, G. L., Hunter, C., Bignell, G., Davies, H., Teague, J., Butler, A., Stevens, C., Edkins, S., O'Meara, S., Vastrik, I., Schmidt, E. E., Avis, T., Barthorpe, S., Bhamra, G., Buck, G., Choudhury, B., Clements, J., Cole, J., Dicks, E., Forbes, S., Gray, K., Halliday, K., Harrison, R., Hills, K., Hinton, J., Jenkinson, A., Jones, D., Menzies, A., Mironenko, T., Perry, J., Raine, K., Richardson, D., Shepherd, R., Small, A., Tofts, C., Varian, J., Webb, T., West, S., Widaa, S., Yates, A., Cahill, D. P., Louis, D. N., Goldstraw, P., Nicholson, A. G., Brasseur, F., Looijenga, L., Weber, B. L., Chiew, Y.-E., DeFazio, A., Greaves, M. F., Green, A. R., Campbell, P., Birney, E., Easton, D. F., Chenevix-Trench, G., Tan, M.-H., Khoo, S. K., Teh, B. T., Yuen, S. T., Leung, S. Y., Wooster, R., Futreal, P. A., and Stratton, M. R. (2007). Patterns of somatic mutation in human cancer genomes. *Nature*, 446(7132) :153–158.

Gryfe, R., Kim, H., Hsieh, E. T., Aronson, M. D., Holowaty, E. J., Bull, S. B., Redston, M., and Gallinger, S. (2000). Tumor microsatellite instability and clinical outcome in young patients with colorectal cancer. *The New England journal of medicine*, 342(2) :69–77.

Guedj, M., Marisa, L., de Reynies, A., Orsetti, B., Schiappa, R., Bibeau, F., MacGrogan, G., Lerebours, F., Finetti, P., Longy, M., Bertheau, P., Bertrand, F., Bonnet, F., Martin, A. L., Feugeas, J. P., Bièche, I., Lehmann-Che, J., Lidereau, R., Birnbaum, D., Bertucci, F., de Thé, H., and Theillet, C. (2012). A refined molecular taxonomy of breast cancer. *Oncogene*, 31(9) :1196–1206.

Guinney, J., Dienstmann, R., Wang, X., de Reynies, A., Schlicker, A., Soneson, C., Marisa, L., Roepman, P., Nyamundanda, G., Angelino, P., Bot, B., Morris, J., Simon, I., Gerster, S., Fessler, E., de Sousa e Melo, Fand Missiaglia, E., Ramay, H., Barras, D., Homicsko, K., Maru, D., Manyam, G., Broom, B., Boige, V., Perez-Villamil, B., Laderas, T., Salazar, R., Gray, J., Hanahan, D., Tabernero, J., Bernards, R., Friend, S.,

Laurent-Puig, P., Medema, J., Sadanandam, A., Wessels, L., Delorenzi, M., Kopetz, S., Vermeulen, L., and Tejpar, S. (2015). The consensus molecular subtypes of colorectal cancer submitted. *Nature Medicine*.

Hamelin, R., Chalastanis, A., Colas, C., El Bchiri, J., Mercier, D., Schreurs, A.-S., Simon, V., Svrcek, M., Zaanan, A., Borie, C., Buhard, O., Capel, E., Zouali, H., Praz, F., Muleris, M., Fléjou, J.-F., and Duval, A. (2008). Clinical and molecular consequences of microsatellite instability in human cancers. *Bulletin Du Cancer*, 95(1) :121–132.

Hamilton, S. R., Weltgesundheitsorganisation, and International Agency for Research on Cancer, editors (2006). *Pathology and genetics of tumours of the digestive system : [... reflects the views of a working group that convened for an Editorial and Consensus Conference in Lyon, France, November 6 - 9, 1999]*. Number 2 in World Health Organization classification of tumours. IARC Press, Lyon, reprinted edition.

Haq, A. I., Schneeweiss, J., Kalsi, V., and Arya, M. (2009). The Dukes staging system : a cornerstone in the clinical management of colorectal cancer. *The Lancet Oncology*, 10(11) :1128.

Hinoue, T., Weisenberger, D. J., Lange, C. P. E., Shen, H., Byun, H.-M., Van Den Berg, D., Malik, S., Pan, F., Noushmehr, H., van Dijk, C. M., Tollenaar, R. A. E. M., and Laird, P. W. (2012). Genome-scale analysis of aberrant DNA methylation in colorectal cancer. *Genome Research*, 22(2) :271–282.

Holmans, P. (2010). Statistical methods for pathway analysis of genome-wide data for association with complex genetic traits. *Advances in Genetics*, 72 :141–179.

Hughes, L. A. E., Melotte, V., de Schrijver, J., de Maat, M., Smit, V. T. H. B. M., Bovée, J. V. M. G., French, P. J., van den Brandt, P. A., Schouten, L. J., de Meyer, T., van Criekinge, W., Ahuja, N., Herman, J. G., Weijenberg, M. P., and van Engeland, M. (2013). The CpG island methylator phenotype : what's in a name ? *Cancer Research*, 73(19) :5858–5868.

Ionov, Y., Peinado, M. A., Malkhosyan, S., Shibata, D., and Perucho, M. (1993). Ubiquitous somatic mutations in simple repeated sequences reveal a new mechanism for colonic carcinogenesis. *Nature*, 363(6429) :558–561.

Isella, C., Terrasi, A., Bellomo, S. E., Petti, C., Galatola, G., Muratore, A., Mellano, A., Senetta, R., Cassenti, A., Sonetto, C., Inghirami, G., Trusolino, L., Fekete, Z., De Ridder, M., Cassoni, P., Storme, G., Bertotti, A., and Medico, E. (2015). Stromal contribution to the colorectal cancer transcriptome. *Nature Genetics*, 47(4) :312–319.

Issa, J.-P. J., Shen, L., and Toyota, M. (2005). CIMP, at last. *Gastroenterology*, 129(3) :1121–1124.

Jass, J. R. (2004). HNPCC and sporadic MSI-H colorectal cancer : a review of the morphological similarities and differences. *Familial cancer*, 3(2) :93–100.

Jass, J. R. (2007). Classification of colorectal cancer based on correlation of clinical, morphological and molecular features. *Histopathology*, 50(1) :113–130.

Jeanmougin, M., de Reynies, A., Marisa, L., Paccard, C., Nuel, G., and Guedj, M. (2010). Should We Abandon the t-Test in the Analysis of Gene Expression Microarray Data : A Comparison of Variance Modeling Strategies. *PLoS ONE*, 5(9) :e12336.

Kelly, H. and Goldberg, R. M. (2005). Systemic therapy for metastatic colorectal cancer : current options, current evidence. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 23(20) :4553–4560.

Khatri, P., Sirota, M., and Butte, A. J. (2012). Ten years of pathway analysis : current approaches and outstanding challenges. *PLoS computational biology*, 8(2) :e1002375.

Kim, T.-M., Laird, P., and Park, P. (2013). The Landscape of Microsatellite Instability in Colorectal and Endometrial Cancer Genomes. *Cell*, 155(4) :858–868.

Kopetz, S., Tabernero, J., Rosenberg, R., Jiang, Z.-Q., Moreno, V., Bachleitner-Hofmann, T., Lanza, G., Stork-Sloots, L., Maru, D., Simon, I., Capellà, G., and Salazar, R. (2015). Genomic classifier ColoPrint predicts recurrence in stage II colorectal cancer patients more accurately than clinical factors. *The Oncologist*, 20(2) :127–133.

Kosinski, C., Li, V. S. W., Chan, A. S. Y., Zhang, J., Ho, C., Tsui, W. Y., Chan, T. L., Mifflin, R. C., Powell, D. W., Yuen, S. T., Leung, S. Y., and Chen, X. (2007). Gene expression patterns of human colon tops and basal crypts and BMP antagonists as intestinal stem cell niche factors. *Proceedings of the National Academy of Sciences of the United States of America*, 104(39) :15418–15423.

Kruhøffer, M., Jensen, J. L., Laiho, P., Dyrskjøt, L., Salovaara, R., Arango, D., Birkenkamp-Demtroder, K., Sørensen, F. B., Christensen, L. L., Buhl, L., Mecklin, J.-P., Järvinen, H., Thykjaer, T., Wikman, F. P., Bech-Knudsen, F., Juhola, M., Nupponen, N. N., Laurberg, S., Andersen, C. L., Aaltonen, L. A., and Ørntoft, T. F. (2005). Gene expression signatures for colorectal cancer microsatellite status and HNPCC. *British Journal of Cancer*, 92(12) :2240–2248.

Laiho, P., Kokko, A., Vanharanta, S., Salovaara, R., Sammalkorpi, H., Järvinen, H., Mecklin, J.-P., Karttunen, T. J., Tuppurainen, K., Davalos, V., Schwartz, S., Arango, D., Mäkinen, M. J., and Aaltonen, L. A. (2007). Serrated carcinomas form a subclass of colorectal cancer with distinct molecular basis. *Oncogene*, 26(2) :312–320.

Lander, E. S. (1999). Array of hope. *Nature Genetics*, 21 :3–4.

Lauss, M., Frigyesi, A., Ryden, T., and Höglund, M. (2010). Robust assignment of cancer subtypes from expression data using a uni-variate gene expression average as classifier. *BMC Cancer*, 10(1) :532.

Le, D. T., Uram, J. N., Wang, H., Bartlett, B. R., Kemberling, H., Eyring, A. D., Skora, A. D., Luber, B. S., Azad, N. S., Laheru, D., Biedrzycki, B., Donehower, R. C., Zaheer, A., Fisher, G. A., Crocenzi, T. S., Lee, J. J., Duffy, S. M., Goldberg, R. M., de la Chapelle, A., Koshiji, M., Bhaijee, F., Huebner, T., Hruban, R. H., Wood, L. D., Cuka, N., Pardoll, D. M., Papadopoulos, N., Kinzler, K. W., Zhou, S., Cornish, T. C., Taube, J. M., Anders, R. A., Eshleman, J. R., Vogelstein, B., and Diaz, L. A. (2015). PD-1 Blockade in Tumors with Mismatch-Repair Deficiency. *New England Journal of Medicine*, 372(26) :2509–2520.

Leach, F. S., Nicolaides, N. C., Papadopoulos, N., Liu, B., Jen, J., Parsons, R., Peltomäki, P., Sistonen, P., Aaltonen, L. A., and Nyström-Lahti, M. (1993). Mutations of a mutS homolog in hereditary nonpolyposis colorectal cancer. *Cell*, 75(6) :1215–1225.

Lederberg, J. (2001). 'Ome Sweet 'Omics– A Genealogical Treasury of Words | The Scientist Magazine®.

Lei, Z., Tan, I. B., Das, K., Deng, N., Zouridis, H., Pattison, S., Chua, C., Feng, Z., Guan, Y. K., Ooi, C. H., Ivanova, T., Zhang, S., Lee, M., Wu, J., Ngo, A., Manesh, S., Tan, E., Teh, B. T., So, J. B. Y., Goh, L. K., Boussioutas, A., Lim, T. K. H., Flotow, H., Tan, P., and Rozen, S. G. (2013). Identification of molecular subtypes of gastric cancer with different responses to PI3-kinase inhibitors and 5-fluorouracil. *Gastroenterology*, 145(3) :554–565.

Levine, A. J. and Puzio-Kuter, A. M. (2010). The Control of the Metabolic Switch in Cancers by Oncogenes and Tumor Suppressor Genes. *Science*, 330(6009) :1340–1344.

Liang, J. J., Bissett, I., Kalady, M., Bennet, A., and Church, J. M. (2013). Importance of serrated polyps in colorectal carcinogenesis : Clinical and molecular aspects of serrated polyps. *ANZ Journal of Surgery*, 83(5) :325–330.

Liao, X., Lochhead, P., Nishihara, R., Morikawa, T., Kuchiba, A., Yamauchi, M., Imamura, Y., Qian, Z. R., Baba, Y., Shima, K., Sun, R., Nosho, K., Meyerhardt, J. A., Giovannucci, E., Fuchs, C. S., Chan, A. T., and Ogino, S. (2012a). Aspirin Use, Tumor PIK3ca Mutation, and Colorectal-Cancer Survival. *New England Journal of Medicine*, 367(17) :1596–1606.

Liao, X., Morikawa, T., Lochhead, P., Imamura, Y., Kuchiba, A., Yamauchi, M., Nosho, K., Qian, Z. R., Nishihara, R., Meyerhardt, J. A., Fuchs, C. S., and Ogino, S. (2012b). Prognostic role of PIK3ca mutation in colorectal cancer : cohort study and literature review. *Clinical Cancer Research : An Official Journal of the American Association for Cancer Research*, 18(8) :2257–2268.

Lichtenstein, P., Holm, N. V., Verkasalo, P. K., Iliadou, A., Kaprio, J., Koskenvuo, M., Pukkala, E., Skytthe, A., and Hemminki, K. (2000). Environmental and Heritable Factors in the Causation of Cancer ? Analyses of Cohorts of Twins from Sweden, Denmark, and Finland. *New England Journal of Medicine*, 343(2) :78–85.

Lipshutz, R. J., Fodor, S. P., Gingeras, T. R., and Lockhart, D. J. (1999). High density synthetic oligonucleotide arrays. *Nature genetics*, 21 :20–24.

Loboda, A., Nebozhyn, M. V., Watters, J. W., Buser, C. A., Shaw, P. M., Huang, P. S., Van't Veer, L., Tollenaar, R. A. E. M., Jackson, D. B., Agrawal, D., Dai, H., and Yeatman, T. J. (2011). EMT is the dominant program in human colon cancer. *BMC medical genomics*, 4 :9.

Loeb, L. A. (2001). A Mutator Phenotype in Cancer. *Cancer Research*, 61(8) :3230–3239.

Ma, Y., Dai, H., and Kong, X. (2012). Impact of warm ischemia on gene expression analysis in surgically removed biosamples. *Analytical Biochemistry*, 423(2) :229–235.

Maak, M., Simon, I., Nitsche, U., Roepman, P., Snel, M., Glas, A. M., Schuster, T., Keller, G., Zeestraten, E., Goossens, I., Janssen, K.-P., Friess, H., and Rosenberg, R. (2013). Independent validation of a prognostic genomic signature (ColoPrint) for patients with stage II colon cancer. *Annals of Surgery*, 257(6) :1053–1058.

Manceau, G. and Laurent-Puig, P. (2012). Signatures moléculaires des cancers colorectaux. In Cremoux, P. d., editor, *Signatures moléculaires des cancers*. John Libbey Eurotext.

Manceau, G., Marisa, L., Boige, V., Duval, A., Gaub, M.-P., Milano, G., Selves, J., Olschwang, S., Jooste, V., le Legrain, M., Lecorre, D., Guenot, D., Etienne-Grimaldi, M.-C., Kirzin, S., Martin, L., Lepage, C., Bouvier, A.-M., and Laurent-Puig, P. (2015). PIK3ca mutations predict recurrence in localized microsatellite stable colon cancer. *Cancer Medicine*, 4(3) :371–382.

Marisa, L., de Reyniès, A., Duval, A., Selves, J., Gaub, M. P., Vescovo, L., Etienne-Grimaldi, M.-C., Schiappa, R., Guenot, D., Ayadi, M., Kirzin, S., Chazal, M., Fléjou, J.-F., Benchimol, D., Berger, A., Lagarde, A., Pencreach, E., Piard, F., Elias, D., Parc, Y., Olschwang, S., Milano, G., Laurent-Puig, P., and Boige, V. (2013). Gene expression classification of colon cancer into molecular subtypes : characterization, validation, and prognostic value. *PLoS medicine*, 10(5) :e1001453.

Markowitz, S., Wang, J., Myeroff, L., Parsons, R., Sun, L., Lutterbaugh, J., Fan, R. S., Zborowska, E., Kinzler, K. W., and Vogelstein, B. (1995). Inactivation of the type II TGF-beta receptor in colon cancer cells with microsatellite instability. *Science (New York, N.Y.)*, 268(5215) :1336–1338.

Markowitz, S. D. and Bertagnolli, M. M. (2009). Molecular origins of cancer : Molecular basis of colorectal cancer. *The New England Journal of Medicine*, 361(25) :2449–2460.

McShane, L. M., Altman, D. G., Sauerbrei, W., Taube, S. E., Gion, M., and Clark, G. M. (2005). REporting recommendations for tumor MARKer prognostic studies (REMARK). *Nature Clinical Practice Oncology*, 2(8) :416–422.

Medema, J. P. and Vermeulen, L. (2011). Microenvironmental regulation of stem cells in intestinal homeostasis and cancer. *Nature*, 474(7351) :318–326.

Melo, F. D. S. E., Wang, X., Jansen, M., Fessler, E., Trinh, A., de Rooij, L. P. M. H., de Jong, J. H., de Boer, O. J., van Leersum, R., Bijlsma, M. F., Rodermond, H., van der Heijden, M., van Noesel, C. J. M., Tuynman, J. B., Dekker, E., Markowetz, F., Medema, J. P., and Vermeulen, L. (2013). Poor-prognosis colon cancer is defined by a molecularly distinct subtype and develops from serrated precursor lesions. *Nature Medicine*, advance online publication.

Merlos-Suárez, A., Barriga, F. M., Jung, P., Iglesias, M., Céspedes, M. V., Rossell, D., Sevillano, M., Hernando-Momblona, X., da Silva-Diz, V., Muñoz, P., Clevers, H., Sancho, E., Mangues, R., and Batlle, E. (2011). The intestinal stem cell signature identifies colorectal cancer stem cells and predicts disease relapse. *Cell Stem Cell*, 8(5) :511–524.

Meyerson, M., Gabriel, S., and Getz, G. (2010). Advances in understanding cancer genomes through second-generation sequencing. *Nature Reviews. Genetics*, 11(10) :685–696.

Michiels, S., Koscielny, S., and Hill, C. (2005). Prediction of cancer outcome with microarrays : a multiple random validation strategy. *Lancet (London, England)*, 365(9458) :488–492.

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. *Machine Learning*, 52(1-2) :91–118.

Muleris, M., Salmon, R. J., Zafrani, B., Girodet, J., and Dutrillaux, B. (1985). Consistent deficiencies of chromosome 18 and of the short arm of chromosome 17 in eleven cases

of human large bowel cancer : a possible recessive determinism. *Annales De Génétique*, 28(4) :206–213.

Nakajima, E. C. and Van Houten, B. (2013). Metabolic symbiosis in cancer : refocusing the Warburg lens. *Molecular Carcinogenesis*, 52(5) :329–337.

Navin, N., Kendall, J., Troge, J., Andrews, P., Rodgers, L., McIndoo, J., Cook, K., Stepansky, A., Levy, D., Esposito, D., Muthuswamy, L., Krasnitz, A., McCombie, W. R., Hicks, J., and Wigler, M. (2011). Tumour evolution inferred by single-cell sequencing. *Nature*, 472(7341) :90–94.

Navin, N., Krasnitz, A., Rodgers, L., Cook, K., Meth, J., Kendall, J., Riggs, M., Eberling, Y., Troge, J., Grubor, V., Levy, D., Lundin, P., Månér, S., Zetterberg, A., Hicks, J., and Wigler, M. (2010). Inferring tumor progression from genomic heterogeneity. *Genome Research*, 20(1) :68–80.

Nguyen, L. V., Vanner, R., Dirks, P., and Eaves, C. J. (2012). Cancer stem cells : an evolving concept. *Nature Reviews Cancer*, 12(2) :133–143.

O'Connell, J., O'Sullivan, G. C., Collins, J. K., and Shanahan, F. (1996). The Fas counterattack : Fas-mediated T cell killing by colon cancer cells expressing Fas ligand. *The Journal of Experimental Medicine*, 184(3) :1075–1082.

O'Connell, J. B., Maggard, M. A., and Ko, C. Y. (2004). Colon cancer survival rates with the new American Joint Committee on Cancer sixth edition staging. *Journal of the National Cancer Institute*, 96(19) :1420–1425.

O'Connell, M. J., Lavery, I., Yothers, G., Paik, S., Clark-Langone, K. M., Lopatin, M., Watson, D., Baehner, F. L., Shak, S., Baker, J., Cowens, J. W., and Wolmark, N. (2010). Relationship between tumor gene expression and recurrence in four independent studies of patients with stage II/III colon cancer treated with surgery alone or surgery plus adjuvant fluorouracil plus leucovorin. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(25) :3937–3944.

Ogino, S. and Goel, A. (2008). Molecular Classification and Correlates in Colorectal Cancer. *The Journal of Molecular Diagnostics : JMD*, 10(1) :13–27.

Oh, S. C., Park, Y.-Y., Park, E. S., Lim, J. Y., Kim, S. M., Kim, S.-B., Kim, J., Kim, S. C., Chu, I.-S., Smith, J. J., Beauchamp, R. D., Yeatman, T. J., Kopetz, S., and Lee, J.-S. (2012). Prognostic gene expression signature associated with two molecularly distinct subtypes of colorectal cancer. *Gut*, 61(9) :1291–1298.

Palles, C., Cazier, J.-B., Howarth, K. M., Domingo, E., Jones, A. M., Broderick, P., Kemp, Z., Spain, S. L., Guarino, E., Salguero, I., Sherborne, A., Chubb, D., Carvajal-Carmona, L. G., Ma, Y., Kaur, K., Dobbins, S., Barclay, E., Gorman, M., Martin, L., Kovac, M. B., Humphray, S., Consortium, T. C., Consortium, T. W., Lucassen, A., Holmes, C. C., Bentley, D., Donnelly, P., Taylor, J., Petridis, C., Roylance, R., Sawyer, E. J., Kerr, D. J., Clark, S., Grimes, J., Kearsey, S. E., Thomas, H. J. W., McVean, G., Houlston, R. S., and Tomlinson, I. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature Genetics*, 45(2) :136–144.

Peiffer, D. A., Le, J. M., Steemers, F. J., Chang, W., Jenniges, T., Garcia, F., Haden, K., Li, J., Shaw, C. A., Belmont, J., Cheung, S. W., Shen, R. M., Barker, D. L., and Gunderson, K. L. (2006). High-resolution genomic profiling of chromosomal aberrations using Infinium whole-genome genotyping. *Genome Research*, 16(9) :1136–1148.

Peltomäki, P., Lothe, R. A., Aaltonen, L. A., Pylkkänen, L., Nyström-Lahti, M., Seruca, R., David, L., Holm, R., Ryberg, D., and Haugen, A. (1993). Microsatellite instability is associated with tumors that characterize the hereditary non-polyposis colorectal carcinoma syndrome. *Cancer Research*, 53(24) :5853–5855.

Perez-Villamil, B., Romera-Lopez, A., Hernandez-Prieto, S., Lopez-Campos, G., Calles, A., Lopez-Asenjo, J.-A., Sanz-Ortega, J., Fernandez-Perez, C., Sastre, J., Alfonso, R., Caldes, T., Martin-Sanchez, F., and Diaz-Rubio, E. (2012). Colon cancer molecular subtypes identified by expression profiling and associated to stroma, mucinous type and different clinical behavior. *BMC Cancer*, 12(1) :260.

Perou, C. M., Sørlie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Rees, C. A., Pollack, J. R., Ross, D. T., Johnsen, H., Akslen, L. A., Fluge, O., Pergamenschikov, A., Williams, C., Zhu, S. X., Lønning, P. E., Børresen-Dale, A. L., Brown, P. O., and Botstein, D. (2000). Molecular portraits of human breast tumours. *Nature*, 406(6797) :747–752.

Pino, M. S. and Chung, D. C. (2010). The chromosomal instability pathway in colon cancer. *Gastroenterology*, 138(6) :2059–2072.

Popat, S., Hubner, R., and Houlston, R. S. (2005). Systematic Review of Microsatellite Instability and Colorectal Cancer Prognosis. *Journal of Clinical Oncology*, 23(3) :609–618.

Popovici, V., Budinska, E., Tejpar, S., Weinrich, S., Estrella, H., Hodgson, G., Van Cutsem, E., Xie, T., Bosman, F. T., Roth, A. D., and Delorenzi, M. (2012). Identification of a poor-prognosis BRAF-mutant-like population of patients with colon cancer. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 30(12) :1288–1295.

Prat, A., Bianchini, G., Thomas, M., Belousov, A., Cheang, M. C. U., Koehler, A., Gómez, P., Semiglazov, V., Eiermann, W., Tjulandin, S., Byakhow, M., Bermejo, B., Zambetti, M., Vazquez, F., Gianni, L., and Baselga, J. (2014). Research-Based PAM50 Subtype Predictor Identifies Higher Responses and Improved Survival Outcomes in HER2-Positive Breast Cancer in the NOAH Study. *Clinical Cancer Research*, 20(2) :511–521.

Rajagopalan, H., Nowak, M. A., Vogelstein, B., and Lengauer, C. (2003). The significance of unstable chromosomes in colorectal cancer. *Nature Reviews. Cancer*, 3(9) :695–701.

Ramanan, V. K., Shen, L., Moore, J. H., and Saykin, A. J. (2012). Pathway analysis of genomic data : concepts, methods, and prospects for future development. *Trends in Genetics*, 28(7) :323–332.

Reichmann, A., Martin, P., and Levin, B. (1981). Chromosomal banding patterns in human large bowel cancer. *International Journal of Cancer. Journal International Du Cancer*, 28(4) :431–440.

Ribic, C. M., Sargent, D. J., Moore, M. J., Thibodeau, S. N., French, A. J., Goldberg, R. M., Hamilton, S. R., Laurent-Puig, P., Gryfe, R., Shepherd, L. E., Tu, D., Redston,

M., and Gallinger, S. (2003). Tumor microsatellite-instability status as a predictor of benefit from fluorouracil-based adjuvant chemotherapy for colon cancer. *The New England journal of medicine*, 349(3) :247–257.

Robinson, D. G., Wang, J. Y., and Storey, J. D. (2015). A nested parallel experiment demonstrates differences in intensity-dependence between RNA-seq and microarrays. *Nucleic Acids Research*, page gkv636.

Roepman, P., Schlicker, A., Tabernero, J., Majewski, I., Tian, S., Moreno, V., Snel, M. H., Chresta, C. M., Rosenberg, R., Nitsche, U., Macarulla, T., Capella, G., Salazar, R., Orphanides, G., Wessels, L. F., Bernards, R., and Simon, I. M. (2013). Colorectal cancer intrinsic subtypes predict chemotherapy benefit, deficient mismatch repair and epithelial-to-mesenchymal transition : Molecular subtypes in colorectal cancer. *International Journal of Cancer*, 134(3) :552–562.

Romagnolo, B. (2012). Paneth cell niche is no longer a driving force. *Médecine Sciences : M/S*, 28(12) :1058–1060.

Roth, A., Di Narzo, A. F., Tejpar, S., Bosman, F., Popovici, V. C., Wirapati, P., Xie, T., Estrella, H., Pavlicek, A., Mao, M., et al. (2012). Validation of two gene-expression risk scores in a large colon cancer cohort and contribution to an improved prognostic method. *ASCO Annual Meeting Proceedings*, 30(15_suppl) :3509.

Roth, A. D., Tejpar, S., Delorenzi, M., Yan, P., Fiocca, R., Klingbiel, D., Dietrich, D., Biesmans, B., Bodoky, G., Barone, C., Aranda, E., Nordlinger, B., Cisar, L., Labianca, R., Cunningham, D., Van Cutsem, E., and Bosman, F. (2010). Prognostic role of KRAS and BRAF in stage II and III resected colon cancer : results of the translational study on the PETACC-3, EORTC 40993, SAKK 60-00 trial. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 28(3) :466–474.

Sablin, M.-P., Italiano, A., and Spano, J.-P. (2009). Colorectal cancers : prognostic and predictive factors of response to treatment. *Bulletin Du Cancer*, 96(4) :417–423.

Sadanandam, A., Lyssiotis, C. A., Homicsko, K., Collisson, E. A., Gibb, W. J., Wullschleger, S., Ostos, L. C. G., Lannon, W. A., Grotzinger, C., Del Rio, M., Lhermitte, B., Olshen, A. B., Wiedenmann, B., Cantley, L. C., Gray, J. W., and Hanahan, D. (2013). A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nature Medicine*.

Salazar, R., Roepman, P., Capella, G., Moreno, V., Simon, I., Dreezen, C., Lopez-Doriga, A., Santos, C., Marijnen, C., Westerga, J., Bruin, S., Kerr, D., Kuppen, P., Velde, C. v. d., Morreau, H., Velthuysen, L. V., Glas, A. M., Veer, L. J. V., and Tollenaar, R. (2010). Gene Expression Signature to Improve Prognosis Prediction of Stage II and III Colorectal Cancer. *Journal of Clinical Oncology*.

Saltz, L. B., Cox, J. V., Blanke, C., Rosen, L. S., Fehrenbacher, L., Moore, M. J., Maroun, J. A., Ackland, S. P., Locker, P. K., Pirotta, N., Elfring, G. L., and Miller, L. L. (2000). Irinotecan plus Fluorouracil and Leucovorin for Metastatic Colorectal Cancer. *New England Journal of Medicine*, 343(13) :905–914.

Samowitz, W. S., Albertsen, H., Herrick, J., Levin, T. R., Sweeney, C., Murtaugh, M. A., Wolff, R. K., and Slattery, M. L. (2005). Evaluation of a large, population-based sample supports a CpG island methylator phenotype in colon cancer. *Gastroenterology*, 129(3) :837–845.

Sanz-Pamplona, R., Berenguer, A., Cordero, D., Riccadonna, S., Solé, X., Crous-Bou, M., Guinó, E., Sanjuan, X., Biondo, S., Soriano, A., Jurman, G., Capella, G., Furlanello, C., and Moreno, V. (2012). Clinical Value of Prognosis Gene Expression Signatures in Colorectal Cancer : A Systematic Review. *PLoS ONE*, 7(11) :e48877.

Sargent, D. J., Marsoni, S., Monges, G., Thibodeau, S. N., Labianca, R., Hamilton, S. R., French, A. J., Kabat, B., Foster, N. R., Torri, V., Ribic, C., Grothey, A., Moore, M., Zaniboni, A., Seitz, J.-F., Sinicrope, F., and Gallinger, S. (2010). Defective mismatch repair as a predictive marker for lack of efficacy of fluorouracil-based adjuvant therapy in colon cancer. *Journal of clinical oncology : official journal of the American Society of Clinical Oncology*, 28(20) :3219–3226.

Schena, M., Shalon, D., Davis, R. W., and Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science (New York, N.Y.)*, 270(5235) :467–470.

Schlicker, A., Beran, G., Chresta, C. M., McWalter, G., Pritchard, A., Weston, S., Runswick, S., Davenport, S., Heathcote, K., Alferez Castro, D., Orphanides, G., French, T., and Wessels, L. F. (2012). Subtypes of primary colorectal tumors correlate with response to targeted treatment in colorectal cell lines. *BMC medical genomics*, 5(1) :66.

Serra, R. W., Fang, M., Park, S. M., Hutchinson, L., and Green, M. R. (2014). A KRAS-directed transcriptional silencing pathway that mediates the CpG island methylator phenotype. *eLife*, 3 :e02313.

Shannon, W., Culverhouse, R., and Duncan, J. (2003). Analyzing microarray data using cluster analysis. *Pharmacogenomics*, 4(1) :41–52.

Shen, L., Toyota, M., Kondo, Y., Lin, E., Zhang, L., Guo, Y., Hernandez, N. S., Chen, X., Ahmed, S., Konishi, K., Hamilton, S. R., and Issa, J.-P. J. (2007). Integrated genetic and epigenetic analysis identifies three different subclasses of colon cancer. *Proceedings of the National Academy of Sciences of the United States of America*, 104(47) :18654–18659.

Sinicrope, F. A. and Sargent, D. J. (2012). Molecular Pathways : Microsatellite Instability in Colorectal Cancer : Prognostic, Predictive, and Therapeutic Implications. *Clinical Cancer Research*, 18(6) :1506–1512.

Slawski, M., Daumer, M., and Boulesteix, A.-L. (2008). CMA : a comprehensive Bioconductor package for supervised classification with high dimensional data. *BMC bioinformatics*, 9 :439.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3 :Article3.

Snover, D. C. (2011). Update on the serrated pathway to colorectal carcinoma. *Human pathology*, 42(1) :1–10.

Sottoriva, A., Kang, H., Ma, Z., Graham, T. A., Salomon, M. P., Zhao, J., Marjoram, P., Siegmund, K., Press, M. F., Shibata, D., and Curtis, C. (2015). A Big Bang model of human colorectal tumor growth. *Nature Genetics*, 47(3) :209–216.

Srivastava, G., Renfro, L. A., Behrens, R. J., Lopatin, M., Chao, C., Soori, G. S., Dakhil, S. R., Mowat, R. B., Kuebler, J. P., Kim, G., Mazurczak, M., Lee, M., and Alberts, S. R. (2014). Prospective Multicenter Study of the Impact of Oncotype DX Colon Cancer Assay Results on Treatment Recommendations in Stage II Colon Cancer Patients. *The Oncologist*, 19(5) :492–497.

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., and Mesirov, J. P. (2005). Gene set enrichment analysis : a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43) :15545–15550.

Suraweera, N., Duval, A., Reperant, M., Vaury, C., Furlan, D., Leroy, K., Seruca, R., Iacopetta, B., and Hamelin, R. (2002). Evaluation of tumor microsatellite instability using five quasimonomorphic mononucleotide repeats and pentaplex PCR. *Gastroenterology*, 123(6) :1804–1811.

Suzuki, H., Aoki, K., Chiba, K., Sato, Y., Shiozawa, Y., Shiraishi, Y., Shimamura, T., Niida, A., Motomura, K., Ohka, F., Yamamoto, T., Tanahashi, K., Ranjit, M., Wakabayashi, T., Yoshizato, T., Kataoka, K., Yoshida, K., Nagata, Y., Sato-Otsubo, A., Tanaka, H., Sanada, M., Kondo, Y., Nakamura, H., Mizoguchi, M., Abe, T., Muragaki, Y., Watanabe, R., Ito, I., Miyano, S., Natsume, A., and Ogawa, S. (2015). Mutational landscape and clonal architecture in grade II and III gliomas. *Nature Genetics*, 47(5) :458–468.

Svrcek, M. and Fléjou, J.-F. (2012). Le « tumor budding » ou bourgeonnement tumoral dans les cancers colorectaux. *Cancéro digest.*

Sørlie, T., Perou, C. M., Tibshirani, R., Aas, T., Geisler, S., Johnsen, H., Hastie, T., Eisen, M. B., van de Rijn, M., Jeffrey, S. S., Thorsen, T., Quist, H., Matese, J. C., Brown, P. O., Botstein, D., Lønning, P. E., and Børresen-Dale, A. L. (2001). Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. *Proceedings of the National Academy of Sciences of the United States of America*, 98(19) :10869–10874.

Thibodeau, S. N., Bren, G., and Schaid, D. (1993). Microsatellite instability in cancer of the proximal colon. *Science (New York, N.Y.)*, 260(5109) :816–819.

Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10) :6567–6572.

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. *Proceedings of the National Academy of Sciences of the United States of America*, 96(15) :8681–8686.

Uccelli, A., Moretta, L., and Pistoia, V. (2008). Mesenchymal stem cells in health and disease. *Nature Reviews Immunology*, 8(9) :726–736.

Umar, A., Boland, C. R., Terdiman, J. P., Syngal, S., de la Chapelle, A., Rüschoff, J., Fishel, R., Lindor, N. M., Burgart, L. J., Hamelin, R., Hamilton, S. R., Hiatt, R. A., Jass, J., Lindblom, A., Lynch, H. T., Peltomaki, P., Ramsey, S. D., Rodriguez-Bigas, M. A., Vasen, H. F. A., Hawk, E. T., Barrett, J. C., Freedman, A. N., and Srivastava, S. (2004).

Revised Bethesda Guidelines for Hereditary Nonpolyposis Colorectal Cancer (Lynch Syndrome) and Microsatellite Instability. *Journal of the National Cancer Institute*, 96(4) :261–268.

van der Flier, L. G. and Clevers, H. (2009). Stem Cells, Self-Renewal, and Differentiation in the Intestinal Epithelium. *Annual Review of Physiology*, 71(1) :241–260.

Venook, A. P., Niedzwiecki, D., Lopatin, M., Ye, X., Lee, M., Friedman, P. N., Frankel, W., Clark-Langone, K., Millward, C., Shak, S., Goldberg, R. M., Mahmoud, N. N., Warren, R. S., Schilsky, R. L., and Bertagnolli, M. M. (2013). Biologic determinants of tumor recurrence in stage II colon cancer : validation study of the 12-gene recurrence score in cancer and leukemia group B (CALGB) 9581. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 31(14) :1775–1781.

Verhaak, R. G. W., Hoadley, K. A., Purdom, E., Wang, V., Qi, Y., Wilkerson, M. D., Miller, C. R., Ding, L., Golub, T., Mesirov, J. P., Alexe, G., Lawrence, M., O'Kelly, M., Tamayo, P., Weir, B. A., Gabriel, S., Winckler, W., Gupta, S., Jakkula, L., Feiler, H. S., Hodgson, J. G., James, C. D., Sarkaria, J. N., Brennan, C., Kahn, A., Spellman, P. T., Wilson, R. K., Speed, T. P., Gray, J. W., Meyerson, M., Getz, G., Perou, C. M., Hayes, D. N., and Cancer Genome Atlas Research Network (2010). Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer cell*, 17(1) :98–110.

Vilar, E. and Gruber, S. B. (2010). Microsatellite instability in colorectal cancer ?the stable evidence. *Nature Reviews Clinical Oncology*, 7(3) :153–162.

Walther, A., Houlston, R., and Tomlinson, I. (2008). Association between chromosomal instability and prognosis in colorectal cancer : a meta-analysis. *Gut*, 57(7) :941–950.

Walther, A., Johnstone, E., Swanton, C., Midgley, R., Tomlinson, I., and Kerr, D. (2009). Genetic prognostic and predictive markers in colorectal cancer. *Nature Reviews Cancer*, 9(7) :489–499.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq : a revolutionary tool for transcriptomics. *Nature Reviews. Genetics*, 10(1) :57–63.

Weinberg, R. (2013). *The Biology of Cancer, Second Edition.* Garland Science.

Weisenberger, D. J., Siegmund, K. D., Campan, M., Young, J., Long, T. I., Faasse, M. A., Kang, G. H., Widschwendter, M., Weener, D., Buchanan, D., Koh, H., Simms, L., Barker, M., Leggett, B., Levine, J., Kim, M., French, A. J., Thibodeau, S. N., Jass, J., Haile, R., and Laird, P. W. (2006). CpG island methylator phenotype underlies sporadic microsatellite instability and is tightly associated with BRAF mutation in colorectal cancer. *Nature Genetics*, 38(7) :787–793.

Wilkerson, M. D. and Hayes, D. N. (2010). ConsensusClusterPlus : a class discovery tool with confidence assessments and item tracking. *Bioinformatics*, 26(12) :1572–1573.

Wood, L. D., Parsons, D. W., Jones, S., Lin, J., Sjöblom, T., Leary, R. J., Shen, D., Boca, S. M., Barber, T., Ptak, J., Silliman, N., Szabo, S., Dezso, Z., Ustyanksky, V., Nikolskaya, T., Nikolsky, Y., Karchin, R., Wilson, P. A., Kaminker, J. S., Zhang, Z., Croshaw, R., Willis, J., Dawson, D., Shipitsin, M., Willson, J. K. V., Sukumar, S., Polyak, K., Park, B. H., Pethiyagoda, C. L., Pant, P. V. K., Ballinger, D. G., Sparks,

A. B., Hartigan, J., Smith, D. R., Suh, E., Papadopoulos, N., Buckhaults, P., Markowitz, S. D., Parmigiani, G., Kinzler, K. W., Velculescu, V. E., and Vogelstein, B. (2007). The genomic landscapes of human breast and colorectal cancers. *Science (New York, N.Y.)*, 318(5853) :1108–1113.

Yamamoto, H. and Imai, K. (2015). Microsatellite instability : an update. *Archives of Toxicology*, 89(6) :899–921.

Yothers, G., O'Connell, M. J., Lee, M., Lopatin, M., Clark-Langone, K. M., Millward, C., Paik, S., Sharif, S., Shak, S., and Wolmark, N. (2013). Validation of the 12-gene colon cancer recurrence score in NSABP C-07 as a predictor of recurrence in patients with stage II and III colon cancer treated with fluorouracil and leucovorin (FU/LV) and FU/LV plus oxaliplatin. *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology*, 31(36) :4512–4519.

Résumé

Résumé

Le cancer du côlon (CC) est l'un des cancers les plus fréquents et les plus mortels en France et dans le monde. Près de la moitié des patients décèdent dans les 5 ans suivant le diagnostic. La classification clinique en stade histologique et la classification moléculaire selon les formes d'instabilité du génome (l'instabilité des microsatellites (MSI), l'instabilité chromosomique (CIN) et l'hyperméthylation des promoteurs (CIMP)) ne suffisent pas à définir des entités homogènes du point de vue moléculaire et à prédire de manière efficace la récidive. Pour améliorer la prise en charge des patients, il apparaît indispensable de mieux appréhender la diversité de la maladie afin de trouver des marqueurs pronostiques et prédictifs efficaces.

Mon travail de thèse a donc été d'étudier la diversité des CC à l'échelle moléculaire par l'utilisation d'approches omiques sur une large cohorte de patients. Il a abouti à l'établissement d'une classification transcriptomique robuste de ce cancer dans son ensemble, validée sur des données indépendantes, et à la caractérisation fine de chacun des sous-types. Six sous-types ont ainsi été définis présentant des caractéristiques clinico-pathologiques, des altérations moléculaires de l'ADN, des enrichissements de signatures liées aux lésions et cellules d'origines, des voies de signalisation dérégulées et des survies bien distinctes. Les résultats de ce travail ont été confortés par un travail de classification consensus mis en place avec un consortium de travail international auquel j'ai participé.

Ces résultats ont permis de confirmer que le cancer colorectal n'est pas une maladie homogène. Ils ouvrent de nouvelles perspectives pour l'établissement de signatures pronostiques et la recherche de cibles pour de nouveaux traitements ainsi que pour l'évaluation de la réponse au traitement au sein d'essais cliniques.

Mots-clefs

cancer, colon, classification, omiques, MSI, CIMP, CIN, médecine personnalisée

Classification of Colorectal Cancer by Omics Approaches

Abstract

Colon cancer (CC) is one of the most frequent and most deadly cancer in France and worldwide. Nearly half of patients die within 5 years after diagnosis. Clinical stage based on histological features and molecular classification based genomic instabilities (microsatellite instability (MSI), chromosomal instability (CIN) and hypermethylation of the promoters (ICPM)) are not sufficient to define homogeneous molecular entities and to predict recurrence effectively. To improve patient care, it is essential to better understand the diversity of the disease so that effective prognostic and predictive markers could be found.

My PhD work has been focused on studying the diversity of CC at the molecular level through the use of omics approaches on a large cohort of tumor samples. It led to the establishment of a robust transcriptomic classification of these cancers, validated on independent data sets, and to a detailed characterization of each of the subtypes. Six subtypes have been defined and were associated with distinct clinicopathological characteristics and molecular alterations, specific enrichments of supervised gene expression signatures related to cell and lesions of origin, specific deregulated signaling pathways and distinct survival. The results of this work have been strengthened by a consensus classification defined by an international consortium working group in which I've been involved.

These results confirm that colorectal cancer is an heterogeneous disease. They provide a renewed framework to develop prognostic signatures, discover new treatment targets, identify new therapeutic strategies and assess response to treatment in clinical trials.

Keywords

cancer, colon, classification, omics, MSI, CIMP, CIN, personalized medicine