HAL
open science

Étude de l'association de variants génétiques rares aux formes précoces de cancer colorectal

Maud Blanluet

To cite this version:

Maud Blanluet. Étude de l'association de variants génétiques rares aux formes précoces de cancer colorectal. Sciences pharmaceutiques. 2018. dumas-01931231

HAL Id: dumas-01931231

https://dumas.ccsd.cnrs.fr/dumas-01931231v1

Submitted on 22 Nov 2018

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.

UNIVERSITE DE ROUEN NORMANDIE
UFR SANTE – Département PHARMACIE

Année 2018 N°

THESE

pour le **DIPLOME D'ETAT DE DOCTEUR EN PHARMACIE**
et le **DIPLOME DE DES DE BIOLOGIE MEDICALE**

Présentée et soutenue publiquement le 02 Octobre 2018

par Maud BLANLUET
Née le 27 Juillet 1988 à Paris, XXe

Etude de l'association de variants génétiques rares aux formes précoces de cancer colorectal

---

Président du jury et co-directeur de thèse :	Monsieur le Professeur Loïc Favennec
Directeur de thèse :	Monsieur le Professeur Thierry Frebourg
Membres du jury :	Madame le Docteur Stéphanie Baert-Desurmont
	Monsieur le Professeur Pierre Michel

Remerciements

Je tiens en premier lieu à exprimer ma profonde gratitude au Professeur Thierry Frebourg, directeur de cette thèse, pour la grande qualité de la formation dont j'ai eu la chance de bénéficier lors de mes années de spécialisation dans son service en Génétique Moléculaire. Je vous remercie grandement d'avoir suscité mon engouement pour l'oncogénétique et de m'avoir fait confiance en acceptant de m'encadrer pour mon Master 2 et pour ma thèse d'exercice.

Je souhaite remercier le Professeur Favennec de me faire l'honneur de présider le jury de cette thèse et de se porter garant de mon travail en acceptant d'en être le co-directeur. Je vous remercie de votre disponibilité et de votre soutien tout au long de mon internat à Rouen.

Je remercie également vivement le Docteur Stéphanie Baert-Desurmont d'avoir permis la constitution de cet échantillon de patients, qui repose sur un travail de collection de données et de prélèvements long d'une décennie, d'avoir accepté d'être rapporteur de mon travail de thèse, mais aussi de m'avoir formée au métier d'oncogénéticien moléculaire avec pédagogie, disponibilité, rigueur et bienveillance. Tu es une encadrante formidable et j'ai beaucoup appris à tes côtés.

Je remercie le Professeur Pierre Michel de me faire l'honneur de juger ce travail de thèse. Veuillez trouver ici l'expression de ma sincère et respectueuse reconnaissance.

Je remercie du fond du cœur Françoise Charbonnier, qui m'a épaulée tout au long de ce travail de thèse et qui a été lors de ces années à Rouen une rencontre exceptionnelle tant sur le plan scientifique que sur le plan humain. Merci de m'avoir transmis ton vaste savoir avec enthousiasme, merci de ton soutien inépuisable et de tes précieux conseils !

Je remercie également Camille Le Clézio et Sophie Coutant pour leur contribution à ce travail et leurs expertises respectives en statistique et en bio-informatique. Ce travail n'aurait pas pu voir le jour sans votre concours, merci de votre gentillesse et du temps précieux que vous m'avez consacré.

Je remercie tout le service de Génétique dans lequel j'ai eu la chance d'évoluer pendant mon internat à Rouen. Merci Pascale Saugier-Veber de m'avoir inculqué des bases solides, indispensables à une progression rigoureuse. Merci de votre soutien et de vos conseils avisés et amicaux. Merci Gaël Nicolas pour ton encadrement, ta sagacité et tes avis aiguisés (malgré l'accent). Merci Gaëlle Bougeard, Edwige Kasper (EK, ça y est je l'ai !!) et Pierre Macquère, le bâtisseur de Start Up, pour le plaisir que j'ai eu à travailler avec vous. Merci à toute l'équipe technique pour votre gentillesse et votre patience : Sandrine pour ta folie partagée, Steeve l'escapeur émérite (et Lucie), Gwen la bourreau des cœurs, Jackie le rayon de soleil, le duo de choc Tic et Tac, Nathalie le pilier de la neuro, Virginie et ta joie de vivre, Séverine maître es SMA, et Fanny Shiva réincarnée. Je remercie également les équipes de Cytogénétique et de Génétique Clinique, nos estimés et bienveillants collègues. J'ai été ravie de travailler avec vous tous pendant ces quelques années d'internat qui m'ont apporté enrichissement intellectuel et épanouissement.

Merci aussi aux équipes de recherche de l'U1245, qui ont pris soin de limiter les percussions impromptues quand je déboulais des escaliers à toute vitesse en claquant les portes. Merci de votre accueil pendant le M2 et après, tout spécialement à Pons / Marinette la reine de la décapitation par la congélation / aviatrice émérite / cavalière distinguée / grande défenseuse de baby devant l'éternel, ainsi qu'à Thomas pour nos discussions métaphysiques et nos récits rocambolesques (et sa Lucie), Laeti (et son Vincent) pour les fou rires et les soirées baby, Omar, Sabine, Alex, et les autres !

Merci bien sûr à la fine fleur rouennaise des internes de Génétique (la T T Tim \o/), j'ai nommé François Jamie MacroMan FPV-Raceur et babyfooteur pas trop mauvais, Gaby Gabz la bassiste métalleuse de-la-mort-qui-tue (et son Vincent le jeu-de-sociétaleux), Kéké le diplomate roi des kévinades et de la bonne chère, Aude la Bleue/violette/rose œnologue et bièrophile distinguée et les petits derniers Ferdi et Juliette, vous êtes nos seuls espoirs pour perpétuer l'ambiance génético-rouennaise ! Et merci Anne-Claire pour nos soirées bières et pas-potage !

Merci à mes potes d'enfance, de lycée, de fac, et à la grande famille des Messagers qui met régulièrement du SOLEIL dans mon cœur ! Merci à mes co-internes de bio et d'ailleurs, qui ont participé à rendre mon internat joyeux, rieur et plein de rebondissements ! Merci Chacha d'avoir illuminé mon année amiénoise de ton rire communicatif et de ton enthousiasme débordant. Merci Cédric et ta petite compresse (et ta Bruneaude), Beubeu l'organiste chanteur lyrique repousseur de cuir et tes oreilles absolues malmenées par notre groupe de ouf (et ta Camille), Thomas et tes leçons de baby dynamiques parfois piquantes mais toujours joviales, Valoche et nos ponts amiénois couloiresques, le Schnapps pour les TAB et ses jeux de mollets, Benoît le moine enfouisseur de cadavres enroulé dans du jambon, Julien le néopsy (mais qu'aux pâtes), et merci à tous ceux avec qui j'ai eu la joie de partager des gardes, des astreintes, des grosses merdes et/ou des petits bonheurs ! Merci aussi à Jérémie de m'avoir permis de sauver le monde, plusieurs fois, même morte.

Merci aux dynamiques équipes de Génétique de l'institut Curie, qui m'ont supportée pendant l'accouchement de cette thèse, et en particulier à Julien pour sa relecture affutée, son joyeux soutien et son amitié pour une petite M1 survoltée, qui a un peu grandi depuis.

Merci maintenant à tous ceux grâce à qui j'en suis là ! Merci maman et papa, pour votre amour et votre soutien indéfectible malgré les petits monstres que vous avez créé. Merci à Eric et Christine de prendre soin de vous et nous. Merci mes p'tits bro', Arthur et Fanou, pour les batailles à coup de doigts de pied, de phasmes et d'épées en latex. Merci aux générations de petits monstres avant nous et après nous, papi Claude, mamie Ninette et tous les cousins ! Merci et bizoux aux exilés du Grand Est, et courage Ariane !!! Merci à Arnaud, Elodie, Lauric et Enora de m'avoir acceptée avec mes handicaps, mon chat, et mes sacs de médocs pendant mon immobilisation forcée ! Merci au cousin Tybald d'avoir aussi mal évolué que moi malgré une décennie de perte de vue !

Merci Lucie (et ton Viktor), la meilleure des potes, avec qui on a vécu galères surmontables et joies colossales depuis le collège, avec qui on a gravit les montagnes des concours, écumé les mers de larmes, descendu les rapides de l'euphorie et plongé dans les abysses de l'inconnu.

Et enfin merci à la PO, en particulier à Max et Cécile, sans qui mon séjour à Rouen aurait définitivement et cruellement perdu en qualité.

L'Université de Rouen et l'UFR de Médecine et de Pharmacie de Rouen n'entendent donner aucune approbation ni improbation aux opinions émises dans cette thèse. Ces opinions sont propres à leurs auteurs.

ANNEE UNIVERSITAIRE 2017 - 2018

U.F.R. DE MEDECINE ET DE-PHARMACIE DE ROUEN

DOYEN :	**Professeur Pierre FREGER**
ASSESSEURS :	**Professeur Michel GUERBET**
	Professeur Benoit VEBER
	Professeur Guillaume SAVOYE

I - MEDECINE

PROFESSEURS DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mr Frédéric **ANSELME**	HCN	Cardiologie
Mme Isabelle **AUQUIT AUCKBUR**	HCN	Chirurgie plastique
Mme Gisèle **APTER**	Havre	Pédopsychiatrie
Mr Fabrice **BAUER**	HCN	Cardiologie
Mme Soumeya **BEKRI**	HCN	Biochimie et biologie moléculaire
Mr Ygal **BENHAMOU**	HCN	Médecine interne
Mr Jacques **BENICHOU**	HCN	Bio statistiques et informatique médicale
Mme Bouchra **LAMIA**	Havre	Pneumologie
Mr Olivier **BOYER**	UFR	Immunologie
Mme Sophie **CANDON**	HCN	Immunologie
Mr François **CARON**	HCN	Maladies infectieuses et tropicales
Mr Philippe **CHASSAGNE** *(détachement)*	HCN	Médecine interne (gériatrie)
Mr Vincent **COMPERE**	HCN	Anesthésiologie et réanimation chirurgicale
Mr Jean-Nicolas **CORNU**	HCN	Urologie
Mr Antoine **CUVELIER**	HB	Pneumologie
Mr Pierre **CZERNICHOW** *(surnombre)*	HCH	Epidémiologie, économie de la santé
Mr Jean-Nicolas **DACHER**	HCN	Radiologie et imagerie médicale
Mr Stéfan **DARMONI**	HCN	Informatique médicale
Mr Pierre **DECHELOTTE**	HCN	Nutrition
Mr Stéphane **DERREY**	HCN	Neurochirurgie
Mr Frédéric **DI FIORE**	CB	Cancérologie
Mr Fabien **DOGUET**	HCN	Chirurgie Cardio Vasculaire

Mr Jean **DOUCET**	SJ	Thérapeutique - Médecine interne et gériatrie
Mr Bernard **DUBRAY**	CB	Radiothérapie
Mr Philippe **DUCROTTE**	HCN	Hépato-gastro-entérologie
Mr Frank **DUJARDIN**	HCN	Chirurgie orthopédique - Traumatologique
Mr Fabrice **DUPARC**	HCN	Anatomie - Chirurgie orthopédique
Mr Eric **DURAND**	HCN	Cardiologie
Mr Bertrand **DUREUIL**	HCN	Anesthésiologie et réanimation chirurgicale
Mme Hélène **ELTCHANINOFF**	HCN	Cardiologie
Mr Manuel **ETIENNE**	HCN	Maladies infectieuses et tropicales
Mr Thierry **FREBOURG**	UFR	Génétique
Mr Pierre **FREGER**	HCN	Anatomie - Neurochirurgie
Mr Jean François **GEHANNO**	HCN	Médecine et santé au travail
Mr Emmanuel **GERARDIN**	HCN	Imagerie médicale
Mme Priscille **GERARDIN**	HCN	Pédopsychiatrie
Mr Michel **GODIN** *(surnombre)*	HB	Néphrologie
M. Guillaume **GOURCEROL**	HCN	Physiologie
Mr Dominique **GUERROT**	HCN	Néphrologie
Mr Olivier **GUILLIN**	HCN	Psychiatrie Adultes
Mr Didier **HANNEQUIN**	HCN	Neurologie
Mr Fabrice **JARDIN**	CB	Hématologie
Mr Luc-Marie **JOLY**	HCN	Médecine d'urgence
Mr Pascal **JOLY**	HCN	Dermato - Vénéréologie
Mme Annie **LAQUERRIERE**	HCN	Anatomie et cytologie pathologiques
Mr Vincent **LAUDENBACH**	HCN	Anesthésie et réanimation chirurgicale
Mr Joël **LECHEVALLIER**	HCN	Chirurgie infantile
Mr Hervé **LEFEBVRE**	HB	Endocrinologie et maladies métaboliques
Mr Thierry **LEQUERRE**	HB	Rhumatologie
Mme Anne-Marie **LEROI**	HCN	Physiologie
Mr Hervé **LEVESQUE**	HB	Médecine interne
Mme Agnès **LIARD-ZMUDA**	HCN	Chirurgie Infantile
Mr Pierre Yves **LITZLER**	HCN	Chirurgie cardiaque
Mr Bertrand **MACE**	HCN	Histologie, embryologie, cytogénétique

M. David **MALTETE**	HCN	Neurologie
Mr Christophe **MARGUET**	HCN	Pédiatrie
Mme Isabelle **MARIE**	HB	Médecine interne
Mr Jean-Paul **MARIE**	HCN	Oto-rhino-laryngologie
Mr Loïc **MARPEAU**	HCN	Gynécologie - Obstétrique
Mr Stéphane **MARRET**	HCN	Pédiatrie
Mme Véronique **MERLE**	HCN	Epidémiologie
Mr Pierre **MICHEL**	HCN	Hépato-gastro-entérologie
M. Benoit **MISSET**	HCN	Réanimation Médicale
Mr Jean-François **MUIR** *(surnombre)*	HB	Pneumologie
Mr Marc **MURAINE**	HCN	Ophtalmologie
Mr Philippe **MUSETTE**	HCN	Dermatologie - Vénéréologie
Mr Christophe **PEILLON**	HCN	Chirurgie générale
Mr Christian **PFISTER**	HCN	Urologie
Mr Jean-Christophe **PLANTIER**	HCN	Bactériologie - Virologie
Mr Didier **PLISSONNIER**	HCN	Chirurgie vasculaire
Mr Gaëtan **PREVOST**	HCN	Endocrinologie
Mr Bernard **PROUST**	HCN	Médecine légale
Mr Jean-Christophe **RICHARD**	HCN	Réanimation médicale - Médecine d'urgence
Mr Vincent **RICHARD**	UFR	Pharmacologie
Mme Nathalie **RIVES**	HCN	Biologie de la reproduction
Mr Horace **ROMAN**	HCN	Gynécologie - Obstétrique
Mr Jean-Christophe **SABOURIN**	HCN	Anatomie - Pathologie
Mr Guillaume **SAVOYE**	HCN	Hépato-gastrologie
Mme Céline **SAVOYE–COLLET**	HCN	Imagerie médicale
Mme Pascale **SCHNEIDER**	HCN	Pédiatrie
Mr Michel **SCOTTE**	HCN	Chirurgie digestive
Mme Fabienne **TAMION**	HCN	Thérapeutique
Mr Luc **THIBERVILLE**	HCN	Pneumologie
Mr Christian **THUILLEZ** *(surnombre)*	HB	Pharmacologie
Mr Hervé **TILLY**	CB	Hématologie et transfusion
M. Gilles **TOURNEL**	HCN	Médecine Légale
Mr Olivier **TROST**	HCN	Chirurgie Maxillo-Faciale

Mr Jean-Jacques **TUECH**	HCN	Chirurgie digestive
Mr Jean-Pierre **VANNIER** *(surnombre)*	HCN	Pédiatrie génétique
Mr Benoît **VEBER**	HCN	Anesthésiologie - Réanimation chirurgicale
Mr Pierre **VERA**	CB	Biophysique et traitement de l'image
Mr Eric **VERIN**	HB	Service Santé Réadaptation
Mr Eric **VERSPYCK**	HCN	Gynécologie obstétrique
Mr Olivier **VITTECOQ**	HB	Rhumatologie
Mme Marie-Laure **WELTER**	HCN	Physiologie

MAITRES DE CONFERENCES DES UNIVERSITES – PRATICIENS HOSPITALIERS

Mme Noëlle **BARBIER-FREBOURG**	HCN	Bactériologie – Virologie
Mme Carole **BRASSE LAGNEL**	HCN	Biochimie
Mme Valérie **BRIDOUX HUYBRECHTS**	HCN	Chirurgie Vasculaire
Mr Gérard **BUCHONNET**	HCN	Hématologie
Mme Mireille **CASTANET**	HCN	Pédiatrie
Mme Nathalie **CHASTAN**	HCN	Neurophysiologie
Mme Sophie **CLAEYSSENS**	HCN	Biochimie et biologie moléculaire
Mr Moïse **COEFFIER**	HCN	Nutrition
Mr Serge **JACQUOT**	UFR	Immunologie
Mr Joël **LADNER**	HCN	Epidémiologie, économie de la santé
Mr Jean-Baptiste **LATOUCHE**	UFR	Biologie cellulaire
Mr Thomas **MOUREZ**	HCN	Virologie
Mr Gaël **NICOLAS**	HCN	Génétique
Mme Muriel **QUILLARD**	HCN	Biochimie et biologie moléculaire
Mme Laëtitia **ROLLIN**	HCN	Médecine du Travail
Mr Mathieu **SALAUN**	HCN	Pneumologie
Mme Pascale **SAUGIER-VEBER**	HCN	Génétique
Mme Anne-Claire **TOBENAS-DUJARDIN**	HCN	Anatomie
Mr David **WALLON**	HCN	Neurologie

PROFESSEUR AGREGE OU CERTIFIE

Mme Mélanie **AUVRAY-HAMEL**	UFR	Anglais
Mr Thierry **WABLE**	UFR	Communication

I - PHARMACIE

PROFESSEURS

Mr Thierry **BESSON**	Chimie Thérapeutique
Mr Roland **CAPRON** (PU-PH)	Biophysique
Mr Jean **COSTENTIN** (Professeur émérite)	Pharmacologie
Mme Isabelle **DUBUS**	Biochimie
Mr François **ESTOUR**	Chimie Organique
Mr Loïc **FAVENNEC** (PU-PH)	Parasitologie
Mr Jean Pierre **GOULLE** (Professeur émérite)	Toxicologie
Mr Michel **GUERBET**	Toxicologie
Mme Isabelle **LEROUX - NICOLLET**	Physiologie
Mme Christelle **MONTEIL**	Toxicologie
Mme Martine **PESTEL-CARON** (PU-PH)	Microbiologie
Mme Elisabeth **SEGUIN**	Pharmacognosie
Mr Rémi **VARIN** (PU-PH)	Pharmacie clinique
Mr Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

MAITRES DE CONFERENCES

Mme Cécile **BARBOT**	Chimie Générale et Minérale
Mr Jérémy **BELLIEN** (MCU-PH)	Pharmacologie
Mr Frédéric **BOUNOURE**	Pharmacie Galénique
Mr Abdeslam **CHAGRAOUI**	Physiologie
Mme Camille **CHARBONNIER (LE CLEZIO)**	Statistiques
Mme Elizabeth **CHOSSON**	Botanique
Mme Marie Catherine **CONCE-CHEMTOB**	Législation pharmaceutique
Mme Cécile **CORBIERE**	Biochimie
Mr Eric **DITTMAR**	Biophysique
Mme Nathalie **DOURMAP**	Pharmacologie
Mme Isabelle **DUBUC**	Pharmacologie
Mme Dominique **DUTERTE- BOUCHER**	Pharmacologie
Mr Abdelhakim **ELOMRI**	Pharmacognosie

Mr François **ESTOUR**	Chimie Organique
Mr Gilles **GARGALA** (MCU-PH)	Parasitologie
Mme Nejla EL **GHARBI-HAMZA**	Chimie analytique
Mme Marie-Laure **GROULT**	Botanique
Mr Hervé **HUE**	Biophysique et mathématiques
Mme Laetitia **LE GOFF**	Parasitologie – Immunologie
Mme Hong **LU**	Biologie
Mme Marine **MALLETER**	Toxicologie
M. Jérémie **MARTINET** (MCU-PH)	Immunologie
Mme Sabine **MENAGER**	Chimie organique
Mme Tiphaine **ROGEZ-FLORENT**	Chimie analytique
Mr Mohamed **SKIBA**	Pharmacie galénique
Mme Malika **SKIBA**	Pharmacie galénique
Mme Christine **THARASSE**	Chimie thérapeutique
Mr Frédéric **ZIEGLER**	Biochimie

PROFESSEURS ASSOCIES

Mme Cécile **GUERARD-DETUNCQ**	Pharmacie officinale
Mr Jean-François **HOUIVET**	Pharmacie officinale

PROFESSEUR CERTIFIE

Mme Mathilde **GUERIN**	Anglais

ASSISTANT HOSPITALO-UNIVERSITAIRE

Mme Anaïs **SOAREZ**	Bactériologie

ATTACHES TEMPORAIRES D'ENSEIGNEMENT ET DE RECHERCHE

Mme Anne-Sophie **CHAMPY**	Pharmacognosie
M. Jonathan **HEDOUIN**	Chimie Organique
Mme Barbara **LAMY-PELLETER**	Pharmacie Galénique

LISTE DES RESPONSABLES DES DISCIPLINES PHARMACEUTIQUES

Mme Cécile **BARBOT**	Chimie Générale et minérale
Mr Thierry **BESSON**	Chimie thérapeutique
Mr Roland **CAPRON**	Biophysique
Mme Marie-Catherine **CONCE-CHEMTOB**	Législation et économie de la santé
Mme Elisabeth **CHOSSON**	Botanique
Mr Jean-Jacques **BONNET**	Pharmacodynamie
Mme Isabelle **DUBUS**	Biochimie
Mr Loïc **FAVENNEC**	Parasitologie
Mr Michel **GUERBET**	Toxicologie
Mr François **ESTOUR**	Chimie organique
Mme Isabelle **LEROUX-NICOLLET**	Physiologie
Mme Martine **PESTEL-CARON**	Microbiologie
Mme Elisabeth **SEGUIN**	Pharmacognosie
Mr Mohamed **SKIBA**	Pharmacie galénique
Mr Rémi **VARIN**	Pharmacie clinique
M. Jean-Marie **VAUGEOIS**	Pharmacologie
Mr Philippe **VERITE**	Chimie analytique

III – MEDECINE GENERALE

PROFESSEUR

Mr Jean-Loup **HERMIL**	UFR	Médecine générale

MAITRE DE CONFERENCE MEDECINE GENERALE

Mr Matthieu **SCHUERS** (MCU-MG)	UFR	Médecine générale

PROFESSEURS ASSOCIES A MI-TEMPS

Mr Emmanuel **LEFEBVRE**	UFR	Médecine Générale
Mme Elisabeth **MAUVIARD**	UFR	Médecine générale
Mr Philippe **NGUYEN THANH**	UFR	Médecine générale
Mme Marie Thérèse **THUEUX**	UFR	Médecine générale

MAITRE DE CONFERENCES ASSOCIE A MI-TEMPS

Mr Pascal **BOULET**	UFR	Médecine générale
Mr Emmanuel **HAZARD**	UFR	Médecine Générale
Mme Marianne **LAINE**	UFR	Médecine Générale
Mme Lucile **PELLERIN**	UFR	Médecine générale
Mme Yveline **SEVRIN**	UFR	Médecine générale

MONO-APPARTENANTS

PROFESSEURS

Mr Serguei **FETISSOV** (med)	Physiologie (ADEN)
Mr Paul **MULDER** (phar)	Sciences du Médicament
Mme Su **RUAN** (med)	Génie Informatique

MAITRES DE CONFERENCES

Mr Sahil **ADRIOUCH** (med)	Biochimie et biologie moléculaire (Unité Inserm905)
Mme Gaëlle **BOUGEARD-DENOYELLE** (med)	Biochimie et biologie moléculaire (UMR 1079**)**
Mme Carine **CLEREN** (med)	Neurosciences (Néovasc)
M. Sylvain **FRAINEAU** (phar)	Physiologie (Inserm U 1096)
Mme Pascaline **GAILDRAT** (med)	Génétique moléculaire humaine (UMR 1079)
Mr Nicolas **GUEROUT** (med)	Chirurgie Expérimentale
Mme Rachel **LETELLIER** (med)	Physiologie
Mme Christine **RONDANINO** (med)	Physiologie de la reproduction
Mr Antoine **OUVRARD-PASCAUD** (med)	Physiologie (Unité Inserm 1076)
Mr Frédéric **PASQUET**	Sciences du langage, orthophonie
Mme Isabelle **TOURNIER** (med)	Biochimie (UMR 1079)

CHEF DES SERVICES ADMINISTRATIFS : Mme Véronique **DELAFONTAINE**

HCN - Hôpital Charles Nicolle

HB - Hôpital de BOIS GUILLAUME

CB - Centre Henri Becquerel

CHS - Centre Hospitalier Spécialisé du Rouvray

CRMPR - Centre Régional de Médecine Physique et de Réadaptation

SJ – Saint Julien Rouen

Sommaire

1. INTRODUCTION 19
1.1 Le cancer : une maladie génétique 19
1.2 Les prédispositions génétiques au cancer 22
1.3 Les syndromes de prédisposition au cancer colorectal 25
1.3.1. Les polyposes adénomateuses 25
1.3.2. Les polyposes hamartomateuses 27
1.3.3. Les polyposes festonnées et les polyposes mixtes 27
1.3.4. Le syndrome de Lynch 28
1.4 L'hérédité manquante du cancer colorectal 29
1.5 Le séquençage de nouvelle génération 30
1.6 L'étude des variants rares 36
1.7 Les voies de la réparation de l'ADN 39
1.7.1. Le système MMR 39
1.7.2. Le système BER 40
1.7.3. Le système HR 42
1.7.4. Le système de réparation des ICLs 44
1.8 Stratégie de l'étude 47

2. MATERIEL ET METHODES 50
2.1 Sélection des patients 50
2.2 Sélection des témoins 51
2.3 Séquençage 51
2.3.1 Sélection des gènes analysés 51
2.3.2 Capture, séquençage et analyses bio-informatiques 57
2.3.3 Filtration et interprétation des variants détectés 59
2.4 Analyses statistiques 62
2.4.1 Contrôles qualité des individus 62
2.4.2 Contrôles qualité des variants 64
2.4.3 Tests d'association 65

3. RESULTATS 68
3.1 Résultats de l'analyse du panel de 201 gènes impliqués dans l'oncogenèse 69
3.1.1 Panel 201 gènes : détection de variants potentiellement délétères 69
3.1.2 Panel 201 gènes : analyse statistique par étude d'association cas-témoins 73
3.2 Résultats de l'analyse de WES 76
3.2.1 Analyse de WES par panel *in silico* de 14 gènes de rang diagnostique responsables de prédisposition héréditaire au CCR 76
3.2.2 Analyse de WES par panel *in silico* de 296 gènes impliqués dans la réparation de l'ADN 78

4. DISCUSSION 87

5. CONCLUSION ET PERSPECTIVES 102

6. BIBLIOGRAPHIE 104

7. SERMENT DE GALIEN 115

Liste des figures

Figure 1 : Le cancer, un processus génétique cumulatif (Chalela *et al*., 2017)..... 21
Figure 2 : Caractéristiques des prédispositions génétiques au cancer 23
Figure 3 : Répartition des formes de CCR..... 29
Figure 4 : *Next Generation Sequencing* avec étape de capture (Bamshad *et al*., 2011)..... 31
Figure 5 : Librairie (Illumina®) 32
Figure 6 : Formation du cluster (Illumina®)..... 34
Figure 7 : Répartition des variants génétiques en fonction de leur fréquence et de leur effet 37
Figure 8 : Le système MMR (Peltomäki, 2016) 40
Figure 9 : Le système BER (Weren *et al.*, 2018)..... 41
Figure 10 : Le système HR (Peterson et Côté, 2004)..... 43
Figure 11 : Résumé de la réparation des ICLs (Nalepa et Clapp, 2018) 44
Figure 12 : Le système ICL (Esteban-Jurado *et al*., 2016) 45
Figure 13 : Stratégie de l'étude 49
Figure 14 : Résultats de l'étude..... 68
Figure 15 : Arbre généalogique du patient DOCC1117 70
Figure 16 : Variations entre le gène *SBDS* et son pseudogène *SBDSP1* (Boocock *et al.,* 2002) 100

Liste des tableaux

Tableau 1 : Gènes responsables de FA (Nalepa et Clapp, 2018)..... 46
Tableau 2: Liste des 201 gènes du panel..... 53
Tableau 3 : Liste des 54 gènes impliqués dans la réparation de l'ADN parmi les 201 gènes du panel. 53
Tableau 4 : Liste des 296 gènes impliqués dans la réparation de l'ADN..... 55
Tableau 5 : Détection simultanée chez la patiente DOCC1544 d'un variant probablement pathogène et de 5 VSI..... 71
Tableau 6 : Détection simultanée chez la patiente DOCC2319 de 2 VSI..... 72
Tableau 7 : Détection simultanée chez la patiente DOCC239 de 3 VSI..... 73
Tableau 8 : Signaux d'association les plus importants détectés par le Burden test à l'échelle du gène 74
Tableau 9 : Variants faux sens du gène *WT1* prédits délétères par les 3 algorithmes 75
Tableau 10 : Burden test à l'échelle des gènes de la réparation de l'ADN du panel 201 gènes..... 75
Tableau 11 : Variations LOF de la voie ICL..... 79
Tableau 12 : CNV d'intérêt 80
Tableau 13 : Comparaison des proportions de porteurs 82
Tableau 14 : Burden test seqMeta comparant les charges en variants..... 83
Tableau 15 : SKAT test..... 84
Tableau 16 : Test d'enrichissement des voies de la réparation de l'ADN 85
Tableau 17 : Comparaison des proportions de porteurs de doubles occurrences dans les voies de la réparation de l'ADN..... 86

Glossaire

ACP	Analyse en Composante Principale
ACMG	*American College of Medical Genetics and Genomics*
BAM	*Binary Alignment Map* : Format de fichier après alignement des reads en NGS
BER	*Base Excision Repair* : Système de réparation de l'ADN par excision de base
CNRGH	Centre National de Recherche en Génomique Humaine
CCR	Cancer ColoRectal
ExAC	*Exome Aggregation Consortium*
FA	*Fanconi Anemia* : Anémie de Fanconi
FAP1	*Familial Adenomatous Polyposis* : Polypose Adénomateuse Familiale
FAP2	*MUTYH-associated Familial Adenomatous Polyposis* : Polypose Adénomateuse Familiale associée à *MUTYH*
FAP3	*NTHL1-associated Familial Adenomatous Polyposis :* Polypose Adénomateuse Familiale associée à *NTHL1*
FREX	*French Exome Project*
GWAS	*Genome Wide Association Study* : Etude d'association sur génome entier
HNPCC	*Hereditary Non-Polyposis Colorectal Cancer* : Cancer colorectal héréditaire non polyposique
HR	*Homologous Recombination* : Système de réparation de l'ADN par recombinaison homologue
HS	Hypersensilité aux agents génotoxiques
ICL	*Inter-strand CrossLink* : Système de réparation des liaisons inter-brins
INCa	Institut National du Cancer
lncRNA	Long ARN non codant
LOF	*Loss Of Function* : Perte de fonction
LOH	*Loss Of Heterozygosity* : Perte d'hétérozygotie
MAF	*Minor Allele Frequency* : Fréquence allélique de l'allèle alternatif
MMR	*Mismatch Repair* : Système de réparation des mésappariements
MSI	*MicroSatellite Instability*
MSS	*MicroSatellite Stability*
NER	*Nucleotide Excision Repair* : Système de réparation de l'ADN par excision de nucléotide
NGS	*Next-Generation Sequencing* : Séquençage haut débit
OR	*Odds Ratio*
SNP	*Single Nucleotide Polymorphism* : Variation nucléotidique polymorphique
SNV	*Single Nucléotide Variation* : Variation nucléotidique
VCF	*Variant Call Format* : fichier de sortie après l'appel des variants (*Variant calling*)
VSI	Variant de Signification Indéterminée
WES	*Whole Exome Sequencing* : Séquençage de l'exome entier

1. INTRODUCTION

1.1 Le cancer : une maladie génétique

Le cancer est une maladie génétique singulière. Il s'agit le plus souvent d'une maladie génétique somatique résultant de l'accumulation d'altérations génétiques et/ou épigénétiques au cours du temps. Contrairement aux maladies mendéliennes monogéniques, la transformation d'une cellule saine en cellule tumorale nécessite l'activation ou l'inactivation de nombreux gènes, afin de dépasser les multiples mécanismes de protection cellulaire. Le cancer est caractérisé par l'acquisition de plusieurs capacités biologiques : une prolifération dérégulée, une résistance à la mort cellulaire, un échappement à la sénescence, une instabilité génétique, l'induction d'une inflammation, une reprogrammation du métabolisme énergétique cellulaire, un échappement aux cellules immunitaires, une induction de l'angiogenèse, ainsi qu'une capacité invasive et métastatique (Hanahan et Weinberg, 2011). Ces capacités sont acquises par les cellules tumorales au cours d'une évolution multi-étapes. La contribution à l'oncogenèse des différentes variations observées dans les tumeurs n'est pas équivalente. Une minorité d'entre elles, qualifiées de mutations pilotes ou *driver mutations*, joue un rôle prépondérant dans la survenue du cancer. A ce jour, environ 140 gènes différents ont été décrits comme potentiels sièges de *driver mutations*. Leur nombre par tumeur est de l'ordre de la dizaine, et varie selon le type de cancer (Vogelstein *et al*., 2013). Ces altérations confèrent un avantage sélectif à la cellule dans laquelle elles surviennent. Les altérations génétiques contribuent à l'oncogenèse, sans qu'une seule soit cependant suffisante (Vogelstein et Kinzler, 2004; Vogelstein *et al.*, 2013). Les autres variations observées sont qualifiées de *passenger mutations* et sont le reflet de l'instabilité génétique de la cellule cancéreuse. Leur nombre peut varier d'une dizaine à des milliers, selon le type de tumeur et les gènes altérés. Elles ne subissent pas de pression de sélection positive et peuvent survenir aux stades pré- et post-néoplasiques (Stratton *et al*., 2009; Vogelstein *et al*., 2013).

Les gènes impliqués dans l'oncogenèse peuvent être classés schématiquement en 3 catégories : les proto-oncogènes, les gènes suppresseurs de tumeur et les gènes responsables de la stabilité du génome.

Les proto-oncogènes participent au développement du cancer lorsqu'une altération du gène rend la protéine constitutionnellement active, permet à celle-ci de s'activer dans des conditions anormales ou permet l'acquisition d'une nouvelle fonction oncogénique. Ils sont alors qualifiés d'oncogènes. Les principaux mécanismes d'activation sont les translocations chromosomiques, les amplifications géniques ou les mutations ponctuelles de type gain de fonction. L'altération d'un des deux allèles d'un proto-oncogène suffit généralement à conférer un avantage sélectif à la cellule.

A l'inverse, les gènes suppresseurs de tumeur participent à l'oncogenèse en subissant une pression de sélection positive lorsqu'ils perdent leur fonction. Le plus souvent une première altération, une insertion ou une délétion de grande ou de petite taille (indels), une mutation ponctuelle, ou une modification épigénétique, inactive un allèle, puis un second évènement inactive l'autre allèle, ce qui provoque une perte d'hétérozygotie (*Loss Of Heterozygosity*, LOH) (Berger *et al*., 2011; Knudson, 2002; Vogelstein et Kinzler, 2004). Plus rarement, le second évènement peut également être une altération ponctuelle. Une minorité des gènes suppresseurs de tumeur, dits haploinsuffisants, ne nécessite pas de second évènement pour que leur altération puisse concourir au développement tumoral (Santarosa et Ashworth, 2004). Les proto-oncogènes et les gènes suppresseurs de tumeur interviennent dans les mêmes processus cellulaires et confèrent tous deux un avantage sélectif à la cellule tumorale.

Le troisième type de gènes dont l'altération joue un rôle causal dans l'oncogenèse intervient de façon très différente puisqu'il s'agit des gènes responsables de la stabilité du génome, dits gènes *caretakers*. Certains de ces gènes participent directement aux processus de réparation de l'ADN, et sont indispensables à la correction des dommages de l'ADN, que ceux-ci soient endogènes, dus aux erreurs induites par la réplication, ou exogènes, dus aux mutagènes environnementaux. D'autres gènes sont indispensables au bon déroulement de la recombinaison mitotique ou de la ségrégation chromosomique, et sont nécessaires au maintien de l'intégrité du génome (Vogelstein et Kinzler, 2004; Vogelstein *et al*., 2013). L'inactivation de ces *caretakers* augmente considérablement le taux d'altérations géniques et chromosomiques en diminuant leur correction, et multiplie donc le risque de survenue de variation délétère activant des proto-oncogènes ou inactivant des gènes suppresseurs de tumeurs (Friedberg, 2003; Curtin, 2012).

Dans la très grande majorité des cas, le cancer est une maladie génétique purement somatique, c'est-à-dire confinée à un organe ou un tissu. Les altérations génétiques post-zygotiques sont acquises au cours de la vie et restreintes à un ou plusieurs clones cellulaires. Une altération somatique touchant un proto-oncogène ou un gène suppresseur de tumeur et soumise à une pression de sélection positive cause une expansion clonale initiant le processus néoplasique. La survenue d'autres variations est ensuite accélérée par la prolifération du clone cellulaire (Maley *et al.*, 2004; Nowell, 2002). Elles peuvent survenir spontanément, ou être induites par des facteurs environnementaux comme l'exposition aux UV, à des agents chimiques mutagènes ou à des virus, et ont échappé aux systèmes cellulaires de réparation de l'ADN. Les cancers, résultant le plus souvent d'altérations somatiques, ont donc une présentation épidémiologique sporadique, c'est-à-dire qu'ils surviennent de façon isolée au sein d'une famille. L'épidémiologie des cancers sporadiques montre une survenue généralement tardive avec une augmentation de leur incidence après 60 ans, reflétant ainsi le processus mutationnel cumulatif nécessaire à l'oncogenèse (Wang, 2016).

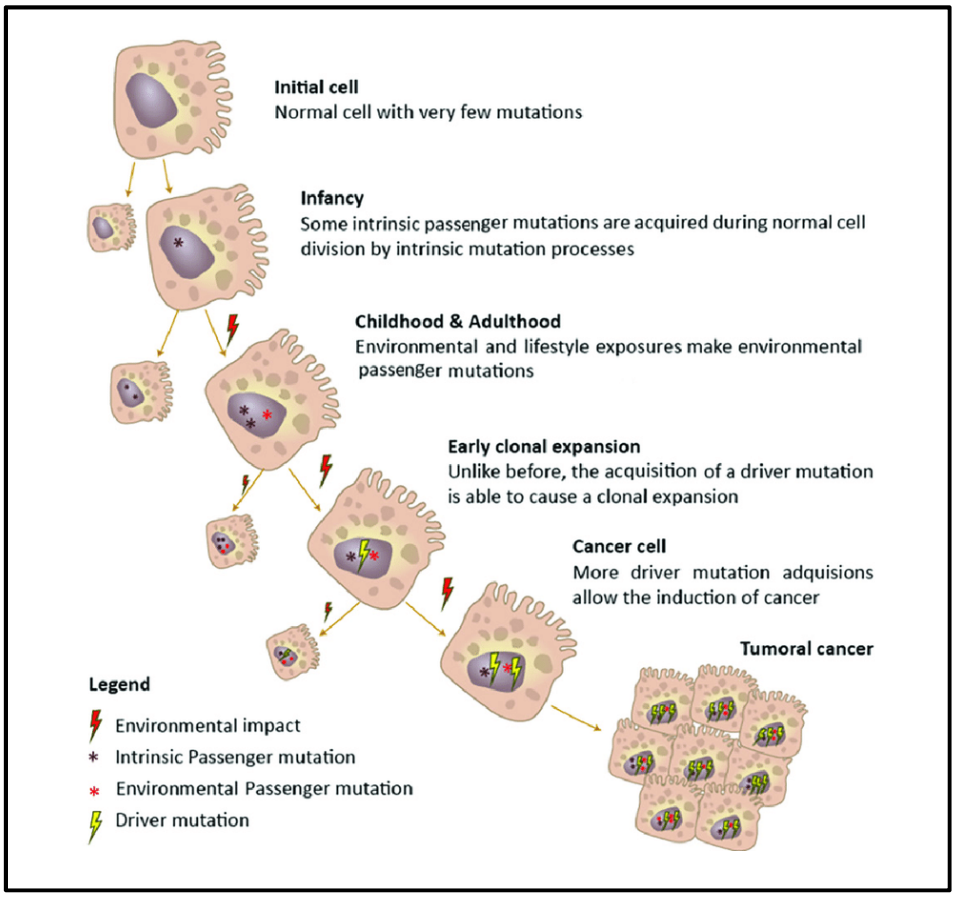


Figure 1 : Le cancer, un processus génétique cumulatif (Chalela et al., 2017)

1.2 Les prédispositions génétiques au cancer

Dans environ 5 % des cancers, d'après les données de l'Institut National du Cancer (INCa) publiées en 2017, l'existence d'une altération constitutionnelle d'un gène majeur impliqué dans l'oncogenèse entraîne une prédisposition héréditaire au cancer répondant à un mode de transmission mendélien. Ces formes mendéliennes de cancer ont trois caractéristiques cliniques : précocité, multiplicité des cancers primitifs, et agrégation familiale. Elles peuvent également être syndromiques, c'est-à-dire associées à un ensemble récurrent de signes cliniques et de symptômes liés ou non au cancer. La majorité des formes mendéliennes de cancers résulte de l'altération constitutionnelle hétérozygote d'un gène suppresseur de tumeur ou d'un gène intervenant dans la stabilité du génome. Ces cancers héréditaires ont le plus souvent une transmission dominante avec pénétrance incomplète (c'est-à-dire que le patient peut avoir le génotype à risque sans être atteint de la maladie) qui peut être expliqué par la nécessité d'un second événement somatique. C'est le cas par exemple du gène *RB1*, le premier gène suppresseur de tumeur identifié, dont l'altération hétérozygote est associée à des formes familiales de rétinoblastome (OMIM # 180200) (Knudson, 1971). En revanche, les cancers héréditaires dus à l'activation constitutionnelle d'un proto-oncogène ont généralement une pénétrance complète puisqu'un seul événement est suffisant. C'est le cas des formes familiales de cancer médullaire de la thyroïde (OMIM # 155240), associées à des variations hétérozygotes faux sens gain de fonction du proto-oncogène *RET*. Ces variations pathogènes sont à l'origine d'une activation constitutive des voies MAPK et PI3K responsable d'une croissance incontrôlée et d'une dé-différentiation cellulaire (Romei *et al*., 2016).

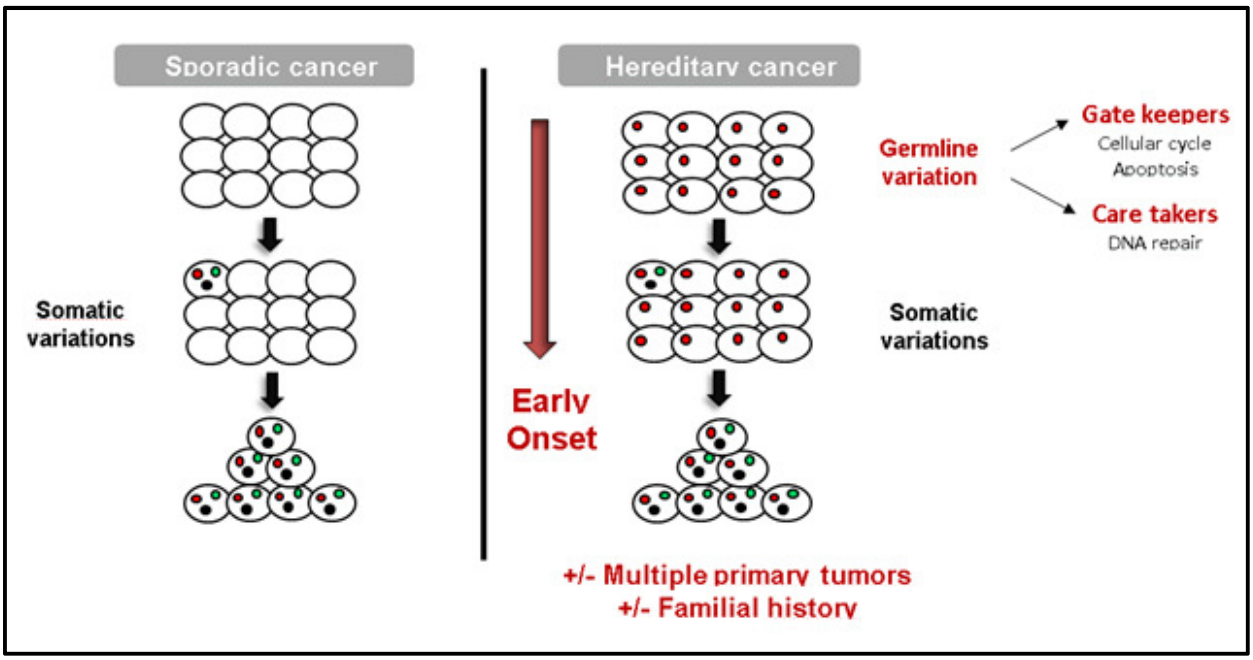


Figure 2 : Caractéristiques des prédispositions génétiques au cancer

Les premières descriptions de formes héréditaires de cancers sont tout d'abord purement phénotypiques et datent des années 1890, avec le dessin de l'arbre généalogique de la « famille G » atteinte de cancers colorectaux et gynécologiques sur au moins quatre générations par le Docteur Alfred Warthin (Warthin, 1985). Il fallut ensuite attendre une soixantaine d'années pour que le Docteur Henry Lynch mette également en évidence une agrégation familiale similaire de cancers digestifs. On lui doit par la suite la mise en œuvre d'une collaboration internationale destinée à réunir ces familles à risque élevé de cancer, ce qui lui permettra ainsi qu'à ses collègues d'identifier le syndrome de Lynch, le syndrome sein-ovaire, et d'autres syndromes génétiques de prédisposition au cancer moins fréquents (Garber et Offit, 2005; Wang, 2016). L'étude proprement dite des bases moléculaires des prédispositions génétiques au cancer s'accélère à la fin du XXe siècle, avec la publication de la théorie du *two-hit model* de Knudson (Knudson, 1971), l'avancée des techniques de génétique moléculaire (Watson, 1990), et l'essor des analyses de liaison génétique (Narod et Amos, 1990; Easton *et al.*, 1993). Pendant les années 1990, plus d'un nouveau gène de prédisposition au cancer est découvert tous les ans, dont 59 grâce à des analyses de liaison réalisées dans des familles à haut risque. Le reste d'entre eux a été identifié grâce à différentes stratégies ciblant des gènes candidats. Ceux-ci ont généralement été sélectionnés selon leur fonction, la voie biologique dans laquelle ils sont impliqués ou leur homologie de structure

avec un gène de prédisposition identifié (Sijbers *et al*., 1996; Meijers-Heijboer *et al.*, 2002; Armanios *et al.*, 2005; Seal *et al.*, 2006; Hao *et al.,* 2009).

Cependant la majorité des gènes candidats évoqués et explorés n'ont pas été associés à une augmentation significative du risque de cancer. La stratégie qui consiste à sélectionner des gènes candidats par les mutations somatiques détectées dans les tumeurs a également été moins performante qu'espérée puisqu'à ce jour elle n'a permis d'identifier qu'une douzaine de nouveaux gènes de prédisposition héréditaire au cancer (Malkin *et al*., 1990; Huff *et al*., 1991; Lohmann, 1999; Taylor *et al*., 2002; Rahman, 2014). C'est le cas par exemple du syndrome de Li Fraumeni, associé à des mutations constitutionnelles du gène *TP53*, et dont des mutations somatiques avaient été mises en évidence préalablement dans des cancers du spectre dont des sarcomes et des cancers du sein (Malkin *et al.*, 1990). A ce jour d'après les données publiées en 2017 par l'INCa, plus de 80 gènes de prédisposition génétique au cancer ont été identifiés, mais de nombreuses familles évocatrices de prédisposition héréditaire au cancer restent sans altération génétique constitutionnelle identifiée.

L'étude des bases moléculaires des prédispositions génétiques au cancer est actuellement l'une des applications médicales les plus importantes de la recherche sur le cancer. En effet, l'identification de la mutation causale est indispensable pour permettre un suivi médical et une prise en charge thérapeutique et chirurgicale adaptées, ainsi qu'un conseil génétique approprié. Les apparentés porteurs de la mutation pourront ainsi bénéficier de protocoles codifiés de détection précoce des tumeurs et se voir proposer des actes chirurgicaux prophylactiques tels que la mastectomie, l'ovariectomie, la colectomie et la gastrectomie (Jatoi et Benson, 2016; Seevaratnam *et al.*, 2012; Shaw et Clarke, 2016). Dans certains cas encore limités, une chémoprévention basée sur l'administration préventive de médicaments, vitamines ou minéraux peut réduire le risque de survenue d'un cancer donné. Des études encourageantes préconisent, par exemple, l'instauration d'un traitement chronique par acide acétylsalicylique chez les patients à risque de cancer colorectal (Burn *et al*., 2013; Din *et al*., 2012). Concernant le conseil génétique, les apparentés non porteurs de la mutation pourront être soulagés d'une lourde anxiété amplifiée par le caractère familial et précoce de la survenue de cancers héréditaires, ainsi que d'un suivi médical rapproché pesant, couteux et injustifié (Rahman, 2014).

1.3 Les syndromes de prédisposition au cancer colorectal

Avec 385 000 nouveaux cas de cancer estimés en France en 2015, dont 5% de formes héréditaires, les prédispositions au cancer représentent les maladies génétiques les plus fréquentes. Le Cancer ColoRectal (CCR) est l'un des plus communs puisqu'il se place au 2e rang des cancers incidents féminins et au 3e rang des cancers incidents masculins. Comme pour la plupart des cancers, le risque de survenue du CCR augmente avec l'âge. En France, d'après les données publiées par l'INCa, 94% des CCR se manifestent chez des patients de plus de 50 ans et la moyenne d'âge au moment du diagnostic est de 70 ans. Le CCR est l'un des cancers fréquents comportant une des proportions les plus importantes de formes héréditaires identifiées. Environ 5% des CCR sont associés à des variations génétiques pathogènes, à forte pénétrance, responsables de syndromes de prédisposition génétique à haut risque connus et caractérisés (Jasperson *et al.*, 2010). La majorité des CCR se développe à partir de polypes adénomateux préexistants (Stoffel et Boland, 2015). L'identification de ces variations est essentielle à la prise en charge des patients et de leur famille car elle permet d'assurer chez les porteurs une détection précoce des tumeurs colorectales grâce à la chromocoloscopie réalisée le plus souvent tous les 2 ans. Les syndromes responsables de CCR sont définis par des critères à la fois cliniques, histologiques et génétiques. Ils ont une pénétrance, un spectre tumoral et une étiologie génétique distincts, de ce fait, la prise en charge de chaque syndrome doit être adaptée (Jasperson *et al.*, 2010). Les prédispositions héréditaires au CCR se classent en plusieurs catégories.

1.3.1. Les polyposes adénomateuses

Les polyposes adénomateuses sont caractérisées par le développement d'au moins 10 polypes adénomateux, leur nombre pouvant s'élever à des milliers. Ces polyposes incluent la polypose adénomateuse familiale de transmission autosomique dominante par mutation du gène *APC* (*Familial Adenomatous Polyposis*, FAP1 ; OMIM # 175100), la polypose autosomique récessive de forme atténuée par mutation biallélique du gène *MUTYH (MUTYH-associated Familial Adenomatous Polyposis 2*, FAP2 ; OMIM # 608456) ou du gène *NTHL1 (NTHL1-associated Familial Adenomatous Polyposis 3*, FAP3 ; OMIM # 616415), et les polyposes autosomiques dominantes associées aux mutations des polymérases *POLE* (OMIM # 615083) et *POLD1* (OMIM # 612591), plus récemment identifiées.

Le gène suppresseur de tumeur *APC* (*WNT signaling pathway regulator*), impliqué dans la FAP1 est responsable de la dégradation de la β-caténine (Yan *et al*., 2016). Cette dernière est un facteur de transcription majeur responsable de la modulation de la prolifération cellulaire. La perte de fonction d'*APC* provoque une accumulation de β-caténine dans le noyau, ce qui induit une activation aberrante de la voie Wnt et autorise une prolifération épithéliale anarchique (Novellasdemunt *et al*., 2015; Stoffel et Boland, 2015). La forme classique de la FAP1 se caractérise par l'apparition vers 10-12 ans d'une centaine à des milliers de polypes dans le côlon et le rectum conduisant généralement à une colectomie prophylactique avant 15 ans (Vasen *et al*., 2008).

Le gène *MUTYH* (*MutY DNA glycosylase*) impliqué dans la FAP2 est un gène majeur du système de réparation de l'ADN par excision de base (*Base Excision Repair*, BER). La protéine codée est une glycosylase. Elle intervient dans la réparation des altérations de l'ADN dues aux espèces réactives de l'oxygène ou en réaction à la présence anormale de bases méthylées, désaminées ou hydroxylées (Sampson et Jones, 2009). Le phénotype de la FAP2 est caractérisé par une polypose modérée, avec le plus souvent moins de 100 polypes, généralement tardive, après 40 ans, et parfois associée à des polypes hyperplasiques (Vogt *et al*., 2009).

Le gène *NTHL1* (*Endonuclease III - like 1*), récemment décrit comme impliqué dans la FAP3, est également un gène majeur du système BER. La protéine codée est une glycosylase à activité endonucléase spécialisée dans la réparation des pyrimidines oxydées. Le phénotype de la FAP3 semble être très proche de celui de la FAP2. A ce jour 7 familles de patients FAP3, porteurs de 1 à plus de 50 adénomes, ont été décrits (Weren *et al.*, 2018). La survenue de CCR est tardive avec des cas de 40 à 65 ans. Des méningiomes, des cancers de la vessie et des carcinomes basocellulaires ont également été rapportés plusieurs fois associées aux altérations bialléliques de *NTHL1* (Weren *et al.,* 2018).

Les variations pathogènes dans les gènes *POLE* ou *POLD1* sont associées à un syndrome rare de polypose héréditaire atténuée découvert récemment (Palles *et al*., 2013). Ce syndrome, nommé *Polymerase Proofreading-Associated Polyposis* (PPAP), a un mode de transmission autosomique dominant et une forte pénétrance. Les gènes *POLE* et *POLD1* codent pour les unités catalytiques des ADN polymérases Polε et Polδ1, responsables de la synthèse respectivement du brin sens et du brin anti-sens lors de la réplication de l'ADN. Elles sont également impliquées dans les systèmes de réparation BER et de réparation des mésappariements de l'ADN (*MisMatch Repair*, MMR). Ces deux enzymes ont une activité de

correction d'épreuve qui permet de limiter les erreurs de réplication en les corrigeant immédiatement. Les variants pathogènes décrits dans le syndrome PPAP affectent les domaines de correction sur épreuve (*Proofreading*) de ces deux protéines, leur faisant ainsi très probablement perdre leur capacité de correction immédiate de leurs propres erreurs. L'analyse génétique somatique de ces tumeurs révèle un état « ultramuté » avec un taux de substitution environ 100 fois supérieur aux cancers non ultramutés. En revanche les microsatellites restent stables (Briggs et Tomlinson, 2013).

1.3.2. Les polyposes hamartomateuses

Elles incluent le syndrome de Peutz-Jeghers (*Peutz-Jeghers Syndrome* ; OMIM # 175200) dû à des altérations du gène suppresseur de tumeur *STK11*, la polypose juvénile (*Juvenile Polyposis Syndrome* ; OMIM # 174900), due à des altérations des gènes *SMAD4* ou *BMPR1A*, et le syndrome tumoral hamartomateux lié à *PTEN* (*Phosphatase and tensin homolog*), ou maladie de Cowden (OMIM # 158350). Ces entités sont caractérisées par le développement de polypes hamartomateux, qui correspondent à des lésions histologiquement identifiables initialement non néoplasiques et à risque de dégénérescence.

1.3.3. Les polyposes festonnées et les polyposes mixtes

La polypose festonnée (*Sessile Serrated Polyposis Cancer Syndrome* ; OMIM # 617108), précédemment appelée polypose hyperplasique (*Hyperplastic Polyposis*), est une entité distincte dont les bases génétiques ne sont actuellement pas caractérisées.

Il existe également des formes héréditaires de polyposes mixtes (*Polyposis syndrome, hereditary mixed*, 1 ; OMIM # 601228) de transmission autosomique dominante, caractérisées par le développement de multiples polypes de différentes histologies : polypes festonnés, polypes de Peutz-Jeghers, polypes juvéniles et polypes adénomateux, sans manifestation extra-colorectales décrites. La polypose mixte héréditaire a été associée à une duplication hétérozygote de 40 kb située en amont du gène *GREM1* responsable de la surexpression de la protéine GREM1. Cette duplication n'a, à ce jour, été identifiée que dans des familles d'origine Ashkenaze (Jaeger *et al.*, 2012). Un chevauchement phénotypique a également été décrit avec la polypose juvénile, des variations pathogènes du gène *BMPR1A* ayant été identifiées chez des patients atteints de polypose mixte (Cao *et al.*, 2006).

1.3.4. Le syndrome de Lynch

Le syndrome de Lynch, également appelé cancer colorectal héréditaire non polyposique (*Hereditary Non-Polyposis Colorectal Cancer*, HNPCC ; OMIM # 120435), initialement défini sur des critères cliniques (critères d'Amsterdam ou de Bethesda), représente la forme héréditaire de cancer colorectale la plus fréquente. Ce syndrome de transmission autosomique dominante est dû à l'altération constitutionnelle d'un des gènes du système MMR, que sont *MLH1* (*MutL homolog 1*), *MSH2* (*MutS homolog 2*), *MSH6* (*MutS homolog 6*) ou *PMS2* (*PMS1 homolog 2).* Ces gènes ont tous un rôle majeur dans le système MMR, responsable de la correction des bases mal appariées et des boucles d'insertion-délétions formées lors de la réplication de l'ADN. L'altération de ce système provoque une accumulation rapide d'altérations non corrigées touchant des séquences répétées dont certaines situées dans des gènes clefs de l'oncogenèse (Jasperson *et al.*, 2010), détectable par l'analyse de microsatellites. La signature tumorale du syndrome de Lynch est l'instabilité microsatellitaire (MSI), ou phénotype RER (*Replication ERror*), recherchée en routine par les laboratoires de génétique somatique des tumeurs. Des marquages Immuno-HistoChimiques (IHC) par des anticorps dirigés contre les protéines MLH1, MSH2, MSH6 et PMS2 permettent de cibler l'altération constitutionnelle grâce à la mise en évidence de l'extinction sélective d'une ou 2 de ces protéines dans la tumeur. Il est à noter que le phénotype RER/MSI, retrouvé dans 15% des CCR (Jasperson *et al.*, 2010), n'est pas spécifique du syndrome de Lynch puisqu'il peut résulter d'une altération somatique biallélique ou d'une altération épigénétique d'un des gènes MMR (Jasperson *et al.*, 2010; Mensenkamp *et al.*, 2014; Stoffel et Boland, 2015). D'autre part, environ 40 à 50% des patients remplissant les critères d'Amsterdam ou de Bethesda pour le syndrome de Lynch n'ont pas de déficience du système MMR dans la tumeur, ni de variations pathogènes détectables dans les gènes MMR (Stoffel and Boland, 2015). Ils sont alors considérés comme atteints de cancer colorectal familial de type X (Shiovitz *et al.*, 2014; Vasen e*t al.*, 2015). Pour l'instant, l'étiologie génétique de ces CCR familiaux sans instabilité des microsatellites (MSS) n'a pas été identifiée, à l'exception de variants pathogènes de *POLE* ou *POLD1* récemment décrits.

1.4 L'hérédité manquante du cancer colorectal

Un nombre conséquent de gènes a donc déjà été impliqué dans les prédispositions héréditaires au CCR et sont actuellement inclus dans les panels de routine diagnostique. Parmi tous les CCR, les études familiales et les études de jumeaux sont en faveur d'un déterminisme génétique dans environ 30% des cas (Grady, 2003; Lichtenstein *et al*., 2000). Des mutations constitutionnelles de gènes connus comme responsables de forme héréditaire de CCR ne sont cependant identifiées que dans environ 5% des cas de CCR tout venants (Figure 3). Il resterait donc à priori 25% des formes supposées héréditaires de CCR sans étiologie génétique identifiée (Jasperson *et al*., 2010). Or, la caractérisation des variations génétiques à l'origine des syndromes de prédisposition héréditaire au cancer est cruciale pour adapter la prise en charge, le suivi et le conseil génétique du patient et de sa famille.

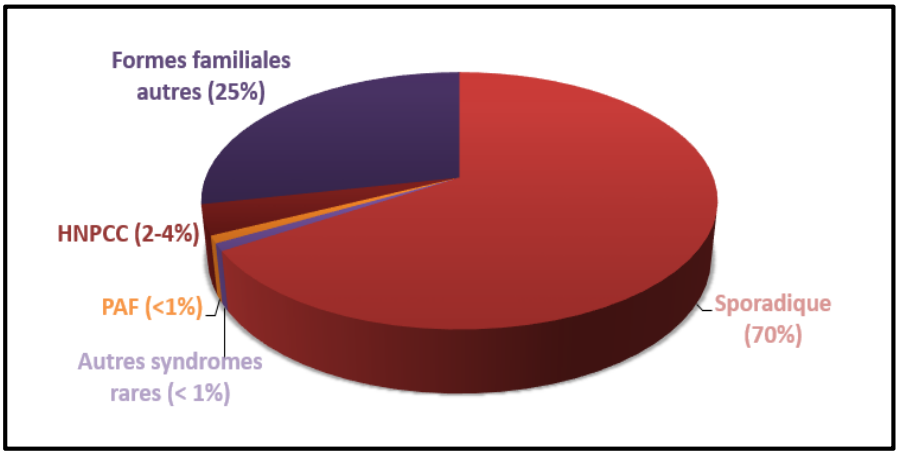


Figure 3 : Répartition des formes de CCR

Une des hypothèses concernant cette hérédité inexpliquée est l'implication d'altérations génétiques probablement plus fréquentes et potentiellement responsables d'une pénétrance plus faible que celles des syndromes de prédisposition déjà caractérisés. Une autre hypothèse est la co-occurrence de ces altérations avec un éventuel effet additif de celles-ci (Jasperson *et al.*, 2010). Ces hypothèses ont initialement été formulées pour expliquer l'héritabilité « manquante » des maladies complexes. Dans ce contexte, les études classiques par gènes candidats et les études de liaison intra-familiales se heurtent au problème des phénocopies (phénotype non héréditaire simulant un phénotype semblable à celui dû à une mutation génétique), très prégnantes dans le cas du CCR, et de la pénétrance incomplète. Pour

contourner ces difficultés, des études statistiques d'association cas-témoins ont été développées. Elles reposent sur la récurrence de variants en comparant la fréquence de variants détectés dans une population de patients avec la fréquence de ces mêmes variants dans une population de témoins. Grâce à l'évolution des techniques d'analyse du génome entier, cette hypothèse d'une implication de variants fréquents a tout d'abord été évaluée par des GWAS (*Genome Wide Association Study*) qui reposent sur l'étude simultanée de centaines de milliers de variations nucléotidiques polymorphiques (SNP), de fréquence allélique dans la population générale (*Minor Allele Frequency*, MAF) supérieure à 5%, chez des milliers d'individus. Plus de 31 SNP de faible pénétrance ont été statistiquement associés à une augmentation modérée du risque de CCR chez les patients par rapport aux témoins, correspondant à environ 5-10% des CCR sans étiologie génétique (COGENT Study *et al*., 2008; Houlston *et al.*, 2010; Dunlop *et al*., 2012, 2013; Peters *et al*., 2013; Baert-Desurmont *et al*., 2016). Cependant le risque de cancer associé à ces SNPs est dans la majorité faible, avec un Odds Ratio (OR) de l'ordre de 1,2-1,8. A titre de comparaison dans les études portant sur le cancer du sein, l'association de l'altération d'un gène à la survenue de cancer avec un un OR > 5 est considéré comme prédisposant à un haut risque (comme *TP53* : OR = 8,17 IC95% 3,74–18,26 p-value 2,23%), un OR compris entre 2 et 5 à un risque modéré (comme *ATM* : OR = 3,28 IC95% 2,06–5,21 p-value = 3,64%) et un OR < 2 sans intérêt clinique (Slavin *et al*., 2017). Or, plus le risque est faible, plus la puissance statistique donc la taille des échantillons de cas et de témoins doit être importante pour pouvoir montrer une association significative entre ce variant et la pathologie d'intérêt. D'autre part, pour des raisons technologiques, ces études de GWAS étaient limitées aux variant fréquents de MAF > 5%.

1.5 Le séquençage de nouvelle génération

Le bond technologique colossal qui s'est opéré ces dernières années avec le développement du séquençage de nouvelle génération (NGS ; *Next-Generation Sequencing*), aussi appelé séquençage de deuxième génération, séquençage haut débit ou séquençage massif en parallèle, a permis de faire évoluer conjointement le débit des analyses de génétique et les possibilités de stratégies d'analyse. Avant le développement de cette technique de biologie moléculaire, l'analyse des gènes de prédisposition au cancer, ou de toute autre maladie génétique reposait essentiellement sur le séquençage Sanger, dit de première

génération. Développée en 1975 et 1977, cette technique repose sur l'amplification d'une séquence d'intérêt par des amorces spécifiques suivie d'une étape de génération de fragments de taille différentes se terminant par une base marquée par un fluorochrome caractéristique de la base (Sanger et Coulson, 1975; Sanger *et al.,* 1977). Les fragments sont ensuite séparés par électrophorèse capillaire et la séquence peut être déterminée par lecture de la fluorescence en fonction de la taille du fragment. Cette technique efficace est encore actuellement très largement utilisée par les laboratoires de génétique moléculaire, mais elle nécessite la synthèse d'amorces spécifiques de chaque exon pour chaque gène à séquencer.

Le premier séquençage complet du génome humain aura nécessité 10 ans de 1993 à 2003 (Lander *et al*., 2001; Venter *et al.*, 2001). Aujourd'hui, de multiples techniques de NGS ont été développées par différentes industries pour rendre ce séquençage possible en quelques jours à peine. La plateforme de génomique du CHU de Rouen et le Centre National de Recherche en Génomique Humaine à Evry (CNRGH), où nous avons réalisé les analyses de cette étude, utilisent la technologie Illumina® avec capture. Cette technique repose sur une première étape de fragmentation de l'ADN. Une étape de PCR permet ensuite d'amplifier les fragments d'ADN et d'ajouter les séquences nécessaires à l'hybridation des amorces de séquençage universelles (Rd1 et Rd2 SP ; Figure 5). Ensuite vient l'étape de capture (Figure 4), caractérisée par l'hybridation des fragments de séquence d'intérêt sur des sondes d'ARN biotinylées spécifiques de ces régions cibles. Les locus d'ADN d'intérêt hybridés aux sondes biotinylées sont ensuite capturés par interaction entre la biotine et des billes magnétiques recouvertes de streptavidine.

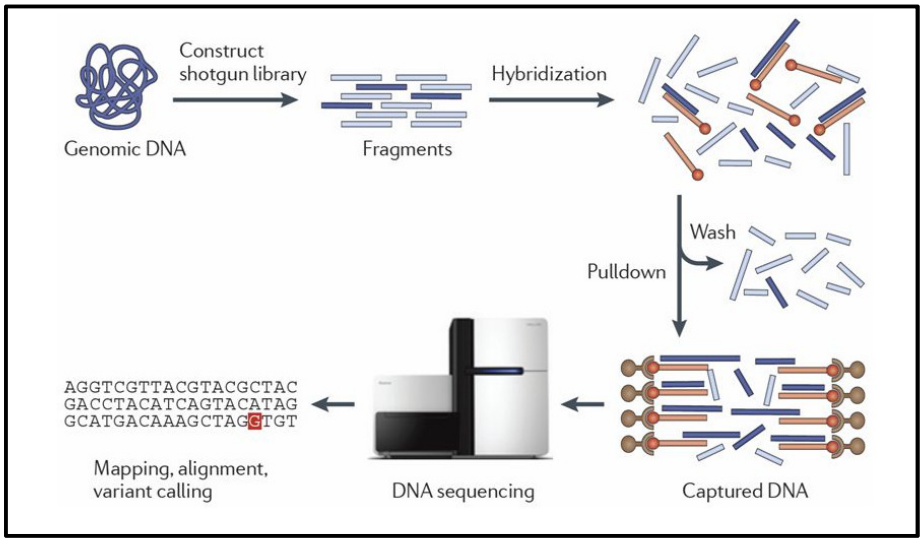
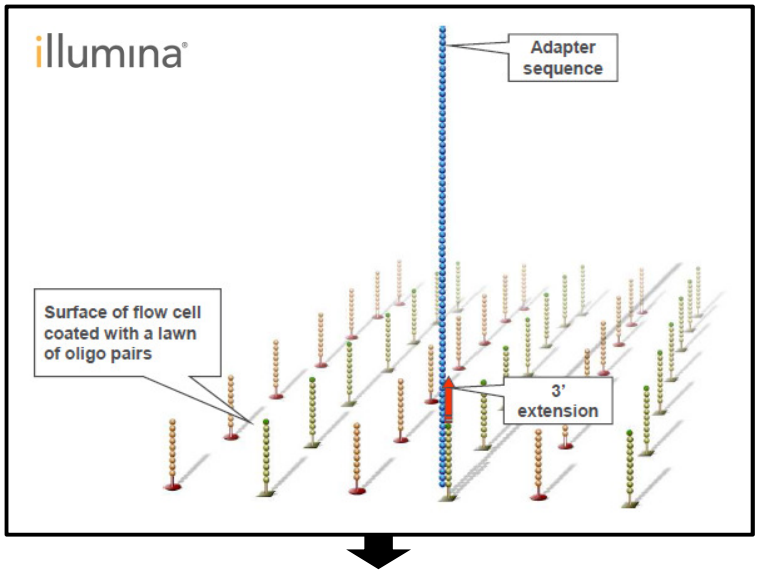


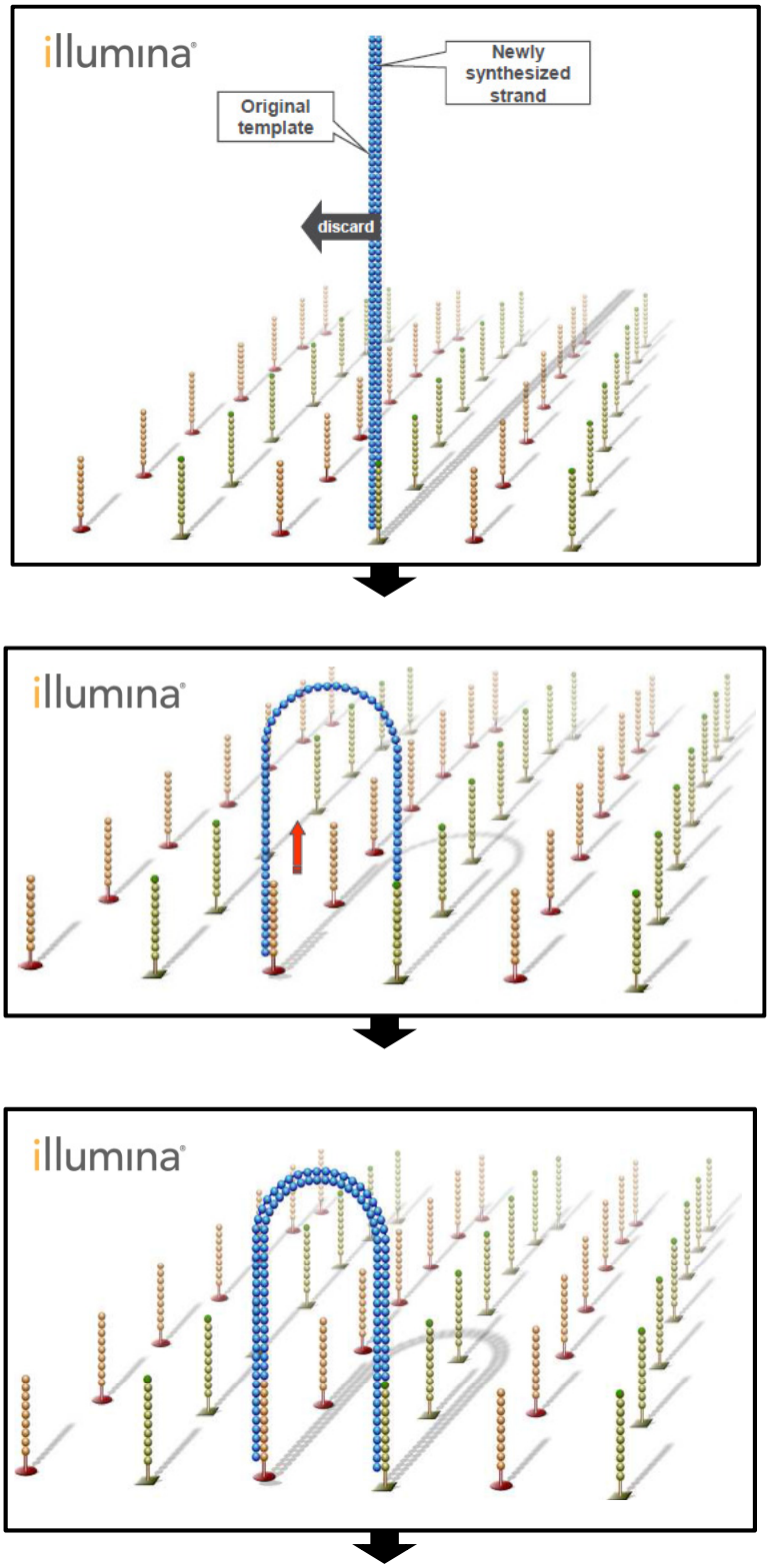
Figure 4 : Next Generation Sequencing avec étape de capture (Bamshad et al., 2011)

L'ADN enrichi en séquence cibles est ensuite élué et amplifié afin d'ajouter des index spécifiques de chaque patients (Index ; Figure 5), permettant le séquençage en parallèle de plusieurs individus, ainsi que les séquences d'attachement à la *flow-cell* (P5 et P7 ; Figure 5), permettant leur fixation au support solide du séquençage. Les index seront séquencés et permettront de ré-attribuer chaque fragment d'ADN à son propriétaire. L'ADN obtenu à l'issu de l'ensemble de ces étapes, alors appelé librairie (Figure 5), est prêt à être séquencé.

Figure 5 : Librairie (Illumina®)

Les librairies des différents patients sont ensuite mélangées de façon équimolaire. Lors du séquençage, les séquences d'ADN s'hybrident sur les amorces fixées à la *flow-cell* grâce aux séquences d'attachement. Le brin complémentaire du brin hybridé est ensuite synthétisé à partir de la séquence d'attachement. Le fragment d'intérêt est donc à présent fixé à la *flow-cell.* Il va ensuite former un pont avec l'amorce adjacente attachée à la *flow-cell* qui sera lui-même amplifié, et ainsi de suite jusqu'à ce qu'un cluster se forme (Figure 6).




illumina
Newly synthesized strand
Original template
discard
illumina
illumina

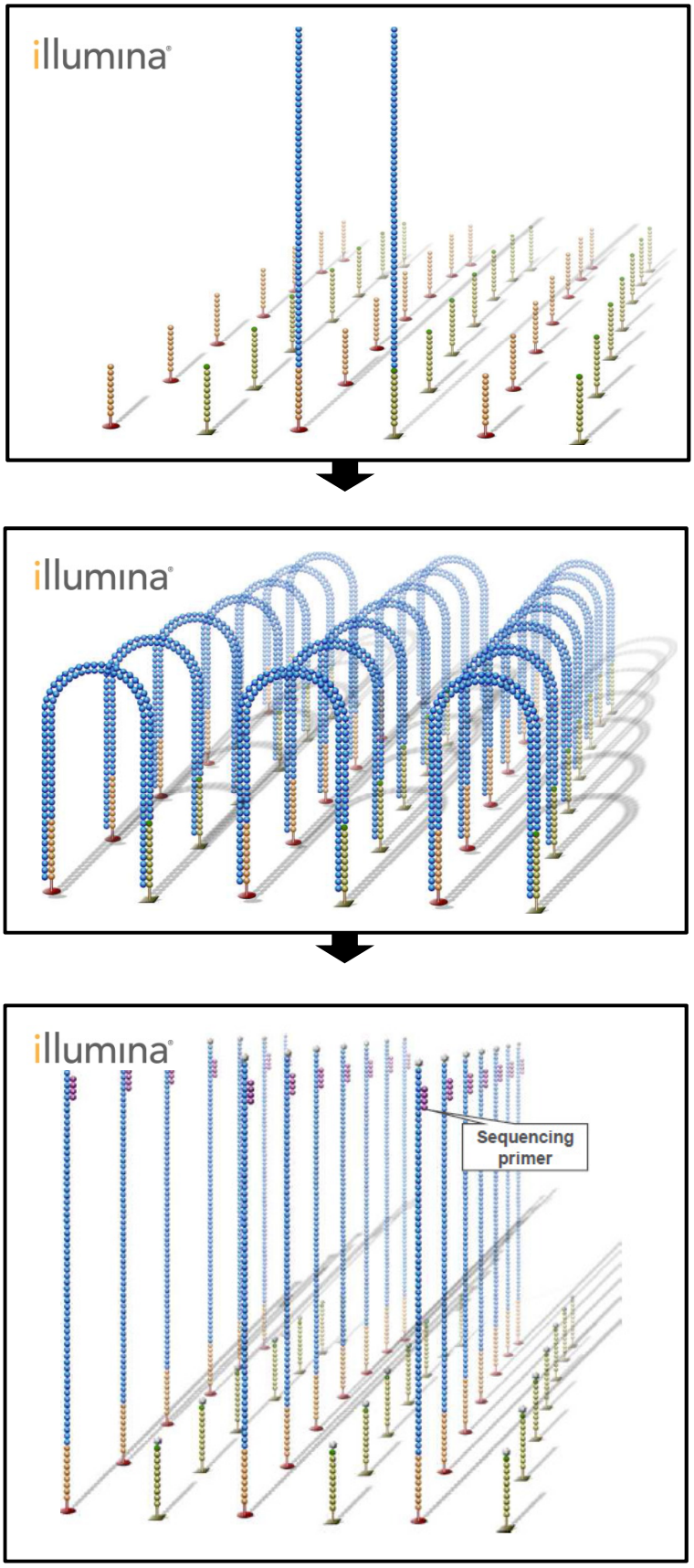


Figure 6 : Formation du cluster (Illumina®)

Ensuite, une succession de cycles de lavage et d'incorporation de nucléotides marqués de façon réversible permet le séquençage des molécules fixées à partir des amorces de séquençage. A chaque cycle, un seul nucléotide marqué est incorporé et une image de la fluorescence est prise. Le signal de fluorescence est identique pour un même fragment donc un même cluster et sera donc visible à la prise d'image par le séquenceur. Puis le fluorochrome du nucléotide marqué est clivé et un second est incorporé, jusqu'au séquençage de l'ensemble du fragment. Ces réactions ont lieu en parallèle pour des milliers de copies du même fragment, pour des milliers de fragments différents, et ce pour plusieurs patients.

En utilisant les sondes de capture de son choix, cette technologie permet de séquencer l'ensemble des exons de plusieurs gènes, pour plusieurs patients, en une seule analyse. On nomme panel de gènes les différents gènes que l'on souhaite ainsi séquencer en parallèle. Cette technologie permet même le séquençage de l'exome entier, ce qui correspond à l'analyse de l'ensemble des exons donc des séquences codantes des 23 000 gènes humains connus (*Whole Exome Sequencing* : WES), correspondant à 1,2 % du génome (3 milliards de base ou 3Gb). Malgré son coût qui s'est néanmoins considérablement réduit au cours des 2 dernières années, cette approche WES permet soit l'étude de l'ensemble des données de l'exome, complexe à analyser étant donné le nombre de variations, soit l'application de panels *in silico*. Cette dernière approche consiste à séquencer tout l'exome mais à ne sélectionner pour interprétation que les données de séquençage de gènes d'intérêt. L'avantage de cette méthode est de réduire le temps et les difficultés d'interprétation, tout en assurant la possibilité de revenir aux données initiales à distance, pour analyser d'autres gènes suite à l'évolution des connaissances par exemple.

Ces dernières années, l'avènement de cette technologie a permis de mettre en évidence l'étendue insoupçonnée de la variabilité du génome humain composé de 3Gb. En 2017, cette variabilité n'est bien caractérisée que pour l'exome, qui ne représente que 34 millions de base (34 Mb), soit 1,2% du génome. L'analyse massive de 60 706 exomes d'individus témoins indemnes de maladies mendéliennes lors de l'étude ExAC (*Exome Aggregation Consortium*) a identifié 7 404 909 variants différents. Parmi ces nombreux variants, 99% sont rares, c'est-à-dire avec une MAF inférieure à 1%, 72% étaient absents des études antérieures plus réduites (1000 Genomes et Exome Sequencing Project), et 54% étaient des singletons, soit des variations n'ayant été retrouvées qu'une seule fois dans cette large étude (Lek *et al.*, 2016). Chaque exome contient environ 23000 variations génétiques, dont 1600 variations rares avec

une MAF < 1% ou privées, c'est-à-dire non présentes dans les bases de données de populations. Parmi ces variations rares, environ 100-300 variations sont prédites délétères, de par leur nature ou suite à l'application d'algorithmes de prédiction de pathogénicité, basés sur la nature de la variation, la conservation de la base ou de l'acide aminé concernés au cours de l'évolution et sur l'écart physicochimique entre l'acide aminé initial et substitué. En moyenne 60 variants par exome sont responsables de la troncature de la protéine codée. La découverte de la grande variabilité de l'exome humain a provoqué un changement profond de paradigme dans l'exercice de la génétique moléculaire humaine. En effet, l'ère du NGS nous apprend que nous sommes tous porteurs d'un grand nombre de variations dont certaines sont rares, certaines ont un effet délétère sur la protéine, certaines sont neutres, voire même certaines sont protectrices. La difficulté aujourd'hui n'est plus d'identifier des variations mais de les interpréter correctement et, le cas échéant, de démontrer leur éventuelle causalité dans une pathologie à l'aide d'un faisceau d'arguments, ainsi que d'estimer le risque associé de développer une maladie (Kassahn *et al.*, 2014; Richards *et al.*, 2015).

1.6 L'étude des variants rares

Conceptuellement, la théorie de l'évolution voudrait que les variants pathogènes soient soumis à une pression de sélection négative et soient donc probablement rares (Lee *et al.*, 2014). Une des hypothèses de l'hérédité manquante est donc l'implication de variants rares conférant un risque intermédiaire dans les pathologies complexes (Manolio *et al.*, 2009) ; Figure 7). Des publications récentes soutiennent cette théorie, avec l'implication de variants rares de pénétrance modérée identifiés dans les gènes *RPS20* et *FAN1* dans des formes familiales de CCR (Nieminen *et al.*, 2014; Seguí *et al.*, 2015). Cependant l'inférence est compliquée par le fait que de tels variants sont présents chez très peu de sujets dans un échantillon de population. Pour atteindre une puissance statistique suffisante et mettre en évidence un éventuel effet significatif, il faut de fait des tailles d'échantillons de populations encore plus grandes que pour les GWAS. De plus, la pénétrance modérée de ces variants rend leur mode de transmission compliqué à caractériser. La présence ou l'absence de phénotype peut être lié à des interactions avec d'autres facteurs génétiques ou environnementaux (de la Chapelle, 2004).

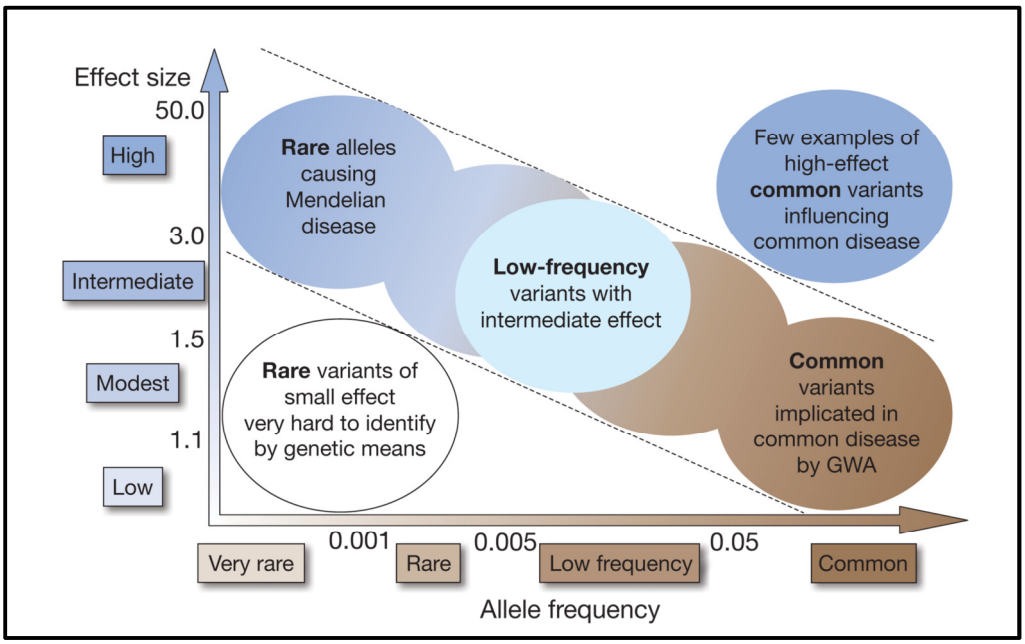


Figure 7 : Répartition des variants génétiques en fonction de leur fréquence et de leur effet (Manolio et al., 2009)

Il existe différentes stratégies que l'on peut combiner pour augmenter la puissance statistique des études d'association des variants rares sans avoir à constituer une cohorte monumentale. Une des approches possibles consiste à sélectionner rigoureusement les patients sur des critères d'inclusion stricts et drastiques afin d'obtenir une population homogène. Ensuite, la restriction d'inclusion aux patients avec un phénotype extrême permet d'augmenter considérablement la probabilité d'un déterminisme génétique important. En effet, il est légitime de penser que ces phénotypes extrêmes sous-tendent la présence de variants rares avec effet fort. De plus, l'hétérogénéité génétique serait moins importante dans une population présentant le même phénotype extrême (Barnett *et al*., 2013; Lee *et al*., 2014). D'après l'expérience du laboratoire de Génétique Moléculaire du CHU de Rouen basée sur l'analyse de 1644 cas index, la restriction de l'analyse des gènes *MLH1, MSH2, MSH6, PMS2, APC, MUTYH, STK11, SMAD4, BMPR1A* et *PTEN* aux patients évocateurs de formes héréditaires de CCR (tumeurs avec une déficience du système MMR avant 65 ans, agrégation familiale de CCR dont un cas avant 51 ans, CCR précoce avant 41 ans, polypose adénomateuse (plus de 10 adénomes), au moins un polype avancé (≥ 10mm, tubulovilleux, en dysplasie de haut grade), entre 2 et 10 adénomes avant 61 ans, polypose hamartomateuse documentée histologiquement, ou polypose festonnée avec plus de 5 lésions) permet d'obtenir

un taux de détection de variations causales d'environ 20%, bien supérieur aux 5% observés chez les patients atteints de CCR tout venants (Baert-Desurmont *et al*., 2018).

La constitution de l'échantillon de témoins revêt également une grande importance. Pour une étude pertinente, les populations de patients et de témoins doivent être appariées en fonction de leur origine ethnique. En effet l'étude de variants rares nécessite de filtrer l'ensemble des variants par fréquence dans l'échantillon total. Or les fréquences alléliques varient significativement selon l'origine ethnique des individus et l'absence d'appariement ethnique induit également un biais de stratification de population. Ce biais, dû à des structures de population différente, empêche de détecter une éventuelle différence de fréquence observée entre les patients et les témoins pour la pathologie étudiée (Barnett *et al*., 2013; Lee *et al*., 2014). Afin de ne pas biaiser la filtration de variants rares et l'interprétation d'une différence cas/témoins de fréquence observée, il est donc important de sélectionner rigoureusement une population témoins ethniquement appariée avec les patients.

Enfin, la mise en évidence de la récurrence d'une variation est problématique dans les études de variants rares. Des méthodes statistiques ont donc été développées afin de les agréger par gène, par voie biologique et/ou par réseaux de gènes. Ces tests sont appelés tests de charge, *Burden tests,* et permettent de cumuler les MAF des variants chez les cas d'une part et les témoins d'autre part. Ils ne prennent pas en compte la coexistence de variants délétères et protecteurs et sont appropriés lorsque les variants agrégés sont supposés tous avoir le même effet sur la protéine (Lee *et al*., 2014). Ils sont puissants lorsqu'une proportion importante des variants est causale et que les effets vont dans la même direction. Il existe également d'autres tests comme le SKAT test, qui prend en compte la variance des OR. Ces tests sont moins performants lorsque la plupart des variants sont causaux et ont le même effet, mais plus puissants en cas de faible fraction de variants causaux, ou de présence de variants à la fois délétères et protecteurs (Wu *et al*., 2011). Ces méthodes statistiques ont par exemple permis à notre l'Unité Inserm 1245 de la Faculté de Rouen de mettre en évidence l'association de variants rares du gène *SORL1* (S*ortilin Related Receptor 1*) avec des formes précoces familiales de la maladie d'Alzheimer, considérées comme un phénotype extrême (Nicolas *et al.*, 2016).

1.7 Les voies de la réparation de l'ADN

Les bases moléculaires connues dans les formes les plus fréquentes de prédisposition héréditaire au cancer colorectal impliquent l'altération de gènes des systèmes de réparation MMR et BER (Vogelstein et Kinzler, 2004). Malgré l'absence de consensus sur le sujet, de récentes publications ont suggéré l'implication dans le CCR de mutations hétérozygotes du gène *FAN1* (*FANCD2 and FANCI associated nuclease 1*), intervenant dans la réparation des liaisons inter-brins (*Inter-strandCrossLink* ; ICL), et du gène *BLM* (*Bloom syndrome RecQ like helicase*) (Seguí *et al.*, 2015; de Voer *et al.,* 2015). Ces deux gènes sont connus pour être impliqués dans les maladies génétiques récessives que sont la néphropathie interstitielle caryomégalique (MIM # 614817) et le syndrome de Bloom (MIM # 210900). La voie de la recombinaison homologue (*Homologous Recombination* ; HR) et la voie ICL, à laquelle appartiennent *FAN1* ainsi que de nombreux gènes responsables de FA, ont également été impliquées dans le CCR (Esteban-Jurado *et al.*, 2016; AlDubayan *et al*., 2018).

Chacun de ces systèmes de réparation de l'ADN est spécialisé dans la correction d'un ou plusieurs types de lésions particulières, mais il existe un chevauchement complexe des différents systèmes.

1.7.1. Le système MMR

Le système de réparation des mésappariements de l'ADN permet la correction des erreurs d'appariements monobases ainsi que des boucles d'insertions-délétions. L'hétérodimère hMutS, α constitué des protéines MSH2 et MSH6, ou β composé de MSH2 et MSH3, détecte ces anomalies. Le complexe hMutS subit ensuite un changement conformationnel et le complexe hMutL est recruté. L'hétérodimère hMutL principal est hMutLα, composé des protéines MLH1 et PMS2. Le complexe hMutS-hMutL est capable de recruter une machinerie d'excision du mésappariement. Une étape de synthèse, reposant sur l'activité de polymérases comme Polε et Polδ1, puis de ligation permettent finalement de restaurer l'ADN (Figure 8 ; Peltomäki, 2016).

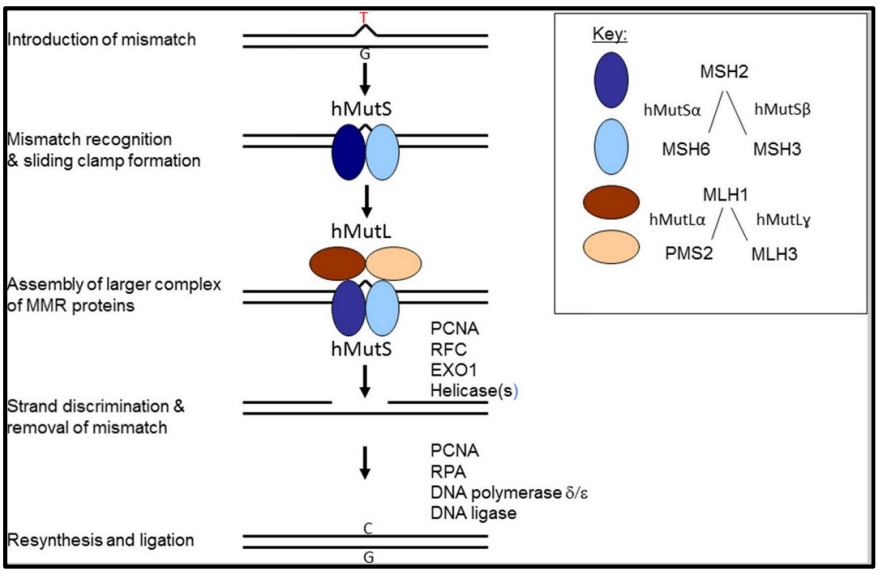


Figure 8 : Le système MMR (Peltomäki, 2016)

L'absence d'une des protéines clés de ce système, que sont MLH1, MSH2, MSH6 et PMS2, provoque une accumulation rapide d'altérations non corrigées touchant des séquences répétées du génome. L'inactivation mono allélique de l'un des gènes codants pour ces protéines est responsable du syndrome de Lynch, de transmission autosomique dominante (*Jasperson et al.*, 2010). L'abolition de la fonction de correction des erreurs de réplications de Polε et de Polδ1 est également responsable de prédisposition au CCR (syndrome PPAP).

1.7.2. Le système BER

Le système de réparation par excision de base est responsable de la réparation des bases de l'ADN altérées par oxydation, alkylation ou désamination, susceptibles de provoquer des mutations. La reconnaissance des lésions est réalisée par les *DNA glycosylases*, qui vont également retirer la base anormale en coupant la liaison N-glycosidique entre celle-ci et le désoxyribose, créant un site abasique. Ces glycosylases sont soit monofonctionnelles, comme MUTYH, c'est-à-dire seulement capables d'exciser la base endommagée, auquel cas une endonucléase est nécessaire pour couper la liaison phosphodiester adjacente, soit bifonctionnelles, c'est-à-dire capables en plus d'une activité endonucléase, comme NTHL1. La polymérase Polβ vient ensuite incorporer la base manquante (*short-patch BER pathway*), ou les polymérases Polε/Polδ remplacent la base manquante et une série de nucléotides

adjacents (*long-patch BER pathway*). Pour finir, une ligase remet en place la liaison phosphodiester (Figure 9 ; Weren *et al.*, 2018).

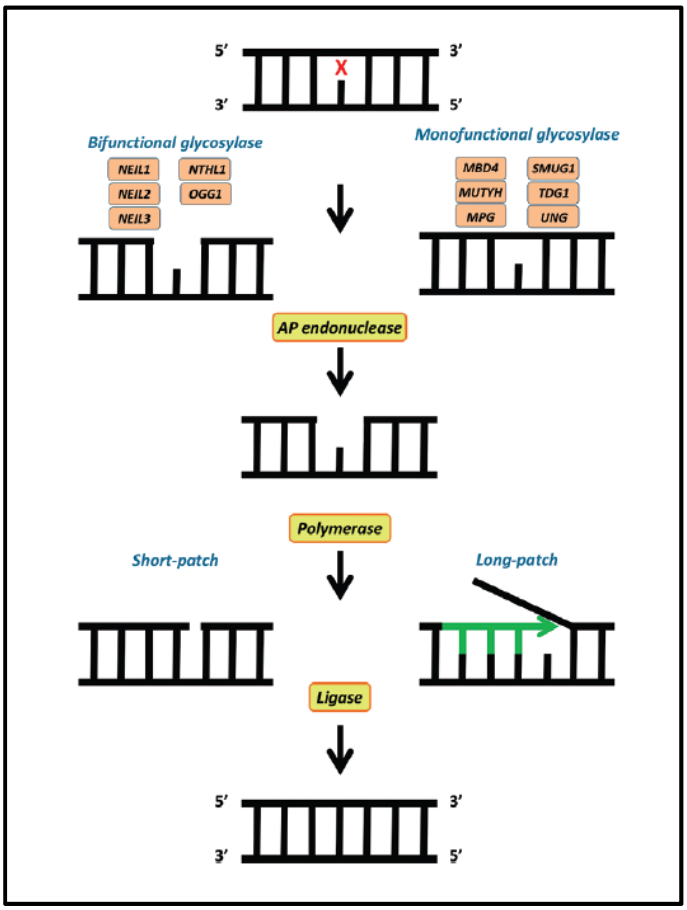


Figure 9 : Le système BER (Weren et al.*, 2018)*

Chacune des *DNA glycosylases* a une préférence pour un type de bases anormales spécifique. L'absence totale d'une de ces protéines, comme MUTYH ou NTHL1, peut être responsable de l'accumulation d'un type particulier de dommage non réparé (Svilar *et al.*, 2011). L'inactivation biallélique des gènes codants pour ces deux protéines sont responsables de formes familiales autosomiques récessives de polyposes (Weren *et al.*, 2018).

1.7.3. Le système HR

Le système de réparation par recombinaison homologue est l'un des systèmes permettant de réparer les cassures doubles brins de l'ADN. Après la reconnaissance de l'altération et la stabilisation des extrémités par le complexe MRN composé entre autre de MRE11, RAD50 et Nsb1, une endonucléase initie une résection 5'-3' à partir du point de cassure de chaque brin. La réparation de la cassure par le système de jonction des extrémités non-homologues (*Non Homologous End-Joining* ; NHEJ) ou par la HR est déterminée en fonction de l'étendue de cette résection. En cas de résection importante, la HR est favorisée. La protéine RPA recouvre l'ADN simple brin 3' formé pour le protéger. Des cofacteurs tels que BRCA2 et PALB2 permettent le recrutement de RAD51 qui se lie à l'ADN. Le complexe formé recherche ensuite une région homologue dans le noyau, dont la séquence va pouvoir être utilisée comme modèle (Moynahan et Jasin, 2010). A nouveau à l'aide des cofacteurs BRCA2 et PALB2, une activité recombinase induit l'invasion de l'ADN simple brin dans la région homologue et un appariement de l'ADN simple brin avec son brin complémentaire dans la région homologue (formation d'une *D-loop*). L'ADN manquant est ensuite synthétisé en utilisant ce dernier comme matrice (Figure 10 ; Peterson et Côté, 2004). Pour terminer, soit le brin néo-synthétisé s'hybride avec l'autre extrémité de la cassure qu'il utilise ensuite comme matrice pour reformer un chromosome complet, soit la seconde extrémité d'ADN pénètre dans la *D-loop* formant une double jonction de Holliday qui peut donner lieu ou non à un *crossing over* (Moynahan et Jasin, 2010).

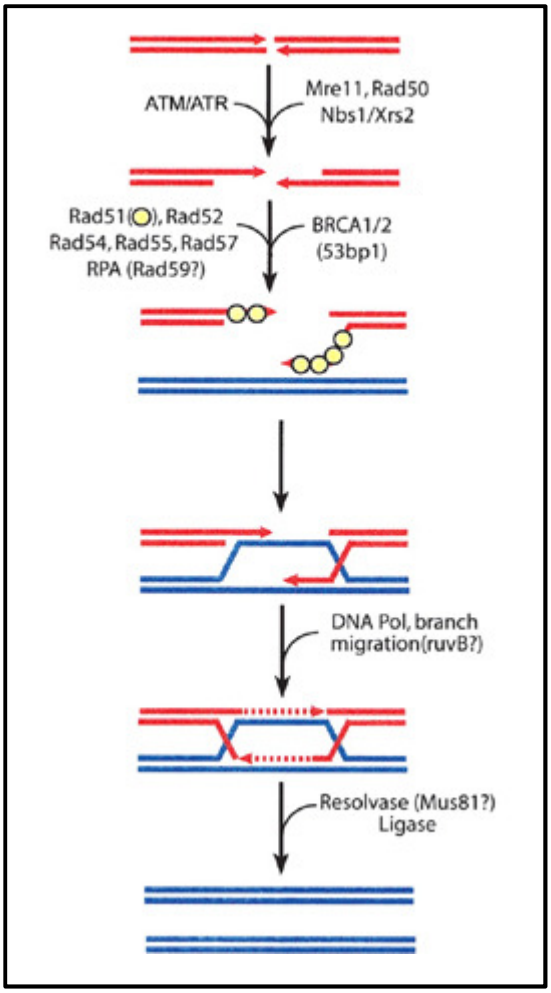


Figure 10 : Le système HR (Peterson et Côté, 2004)

Les protéines BRCA1, RAD51, mais aussi BRCA2 et PALB2 interviennent entre autre dans cette voie de réparation. L'altération d'un allèle du gène codant pour l'une de ces protéines est responsable du syndrome héréditaire de prédisposition au cancer du sein et de l'ovaire (Familial susceptibility to breast-ovarian cancer 1 : *BRCA1,* OMIM # 604370 ; 2 : *BRCA2,* OMIM # 612555 ; et Susceptibility to breast cancer : *PALB2,* OMIM # 114480), de transmission autosomique dominante. L'inactivation tumorale du deuxième allèle conduit à la tumorigénèse. Comme nous allons le voir, ces protéines sont également impliquées dans la voie ICL et leur altération biallélique constitutionnelle est responsable de FA (*Fanconi anemia, complementation group D1, BRCA2* : OMIM # 605724, et *Fanconi anemia, complementation group N, PALB2* : OMIM # 610832, *Fanconi anemia, complementation group S, BRCA1* : OMIM # 617883).

1.7.4. Le système de réparation des ICLs

Le système de réparation des liaisons inter-brins est responsable du clivage des ponts covalents se formant entre les deux brins de l'ADN double-brin, empêchant leur séparation. Ces liaisons covalentes anormales très fortement cytotoxiques sont causées par des agents alkylants bifonctionnels. Ces derniers, dotés de deux groupements alkyls, sont capables de réagir avec deux nucléotides et ainsi de former des ponts intra ou inter-caténaires. La détection des dommages est prise en charge par ATM, ATR et CHK1, qui phosphorylent différentes protéines dont FANCE, FANCM et FANCG pour faciliter la formation du complexe FA-core. Ce complexe permet la monoubiquitylation de l'hétérodimère FANCD2-FANCI par FANCL et UBE2T. C'est une étape clé qui permet le recrutement d'effecteurs nécessaire à la réparation des dommages ICLs (Figure 11 ; Nalepa et Clapp, 2018).

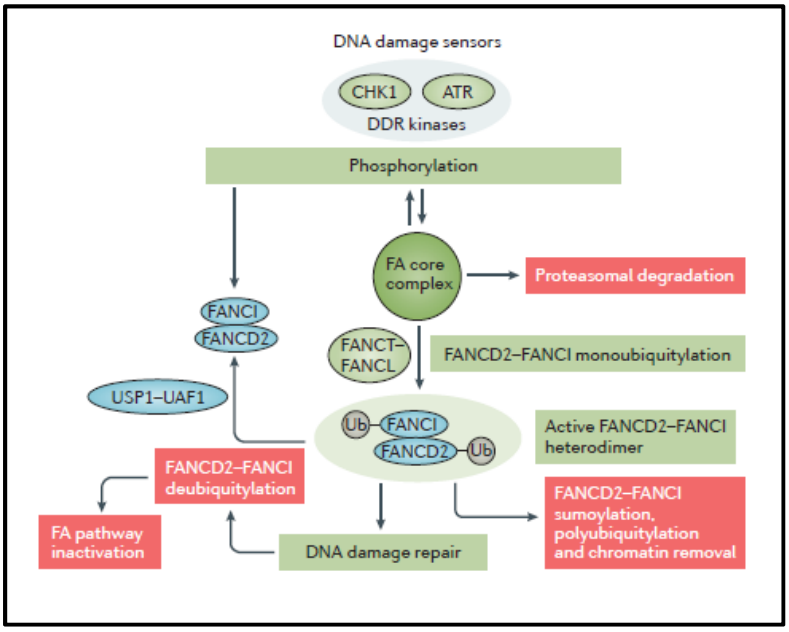


Figure 11 : Résumé de la réparation des ICLs (Nalepa et Clapp, 2018)

En effet, la réparation de ces dommages est particulièrement complexe car elle intéresse les deux brins de l'ADN. En plus de la voie ICL proprement dite, elle fait donc intervenir d'autres systèmes de réparation comme le système de réparation de l'ADN par excision de nucléotide (*Nucleotide Excision Repair* : NER), et la HR (Hashimoto *et al.*, 2016) (Figure 12).

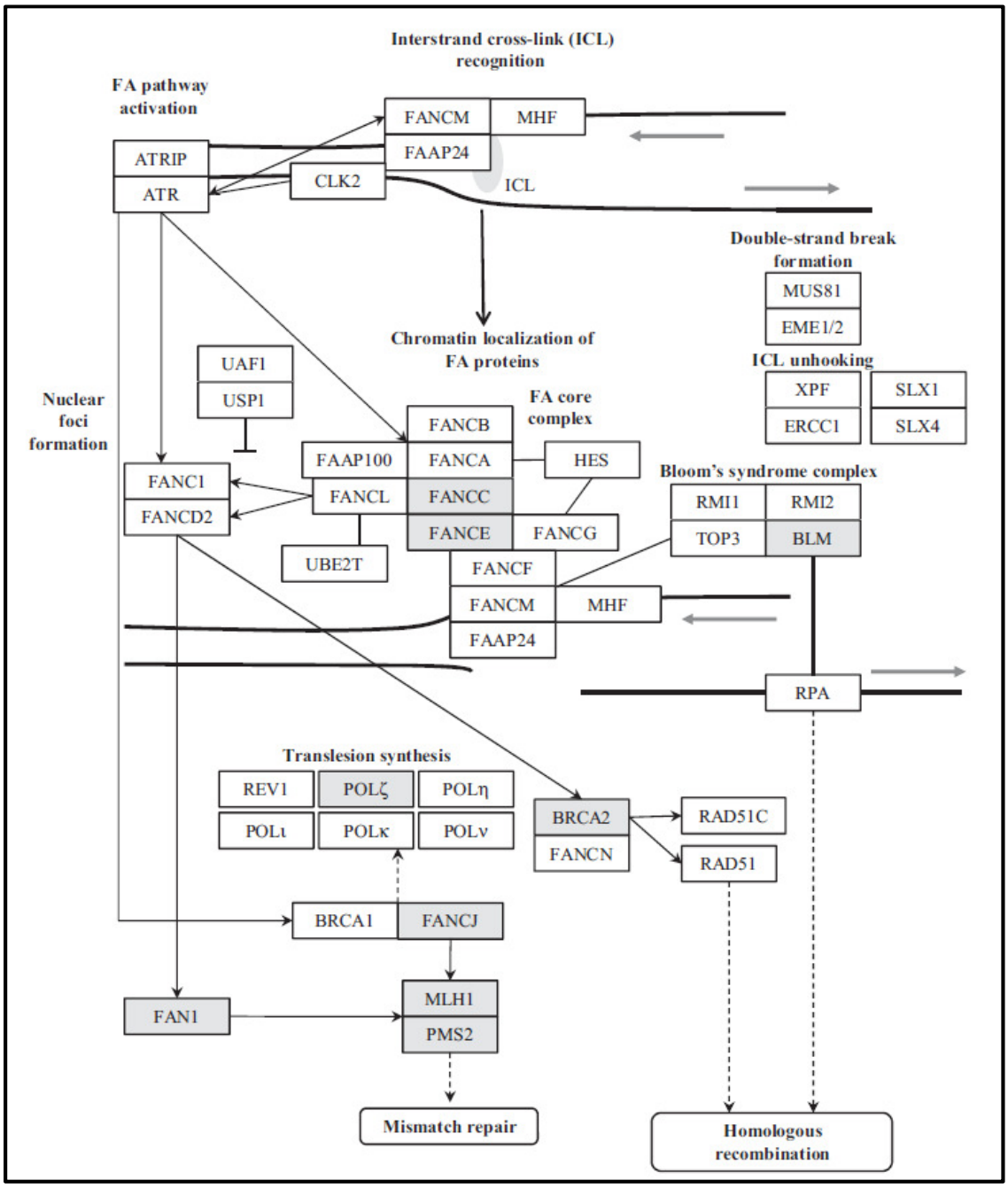


Figure 12 : Le système ICL (Esteban-Jurado et al.*, 2016)*

L'inactivation biallélique d'un nombre important des gènes impliqués dans la voie de réparation ICL provoque une hypersensibilité aux agents pontants inter-brins, responsable de l'Anémie de Fanconi (FA). Cette hypersensibilité est détectable par la mise en évidence de cassures chromosomiques multiples lors de l'exposition des lymphocytes de patients atteints à ces agents pontants. Le phénotype de la FA associe pancytopénie progressive, aplasie médullaire, malformations congénitales (petite taille, lésions cutanées, anomalies

urogénitales, anomalies des pouces, du cœur et du rein, dysmorphie…) et prédisposition aux hémopathies et aux tumeurs solides. A ce jour, 22 sous-types génétiques de FA ont été identifiés donc 21 décrits ci-après (Tableau 1). Le 22e gène de FA rapporté est le gène *RFWD3* ou *FANCW*, récemment impliqué dans la HR (Knies *et al.*, 2017). La FA est de transmission autosomique récessive, sauf pour les gènes *FANCB*, de transmission liée à l'X, et *RAD51,* dont les mutations rapportées sont hétérozygotes avec un effet dominant négatif (Nalepa et Clapp, 2018).

Tableau 1 : Gènes responsables de FA (Nalepa et Clapp, 2018)

FA gene alias	Alternative name	Established molecular function	Estimated frequency in FA	Cancer risk in heterozygotes
FANCA	–	FA core complex assembly	60%	Unknown
FANCB	–	FA core complex assembly	2% (X-linked)	Unknown
FANCC	–	FA core complex assembly	12%	Possible
FANCD1	*BRCA2*	DNA repair control and effector recruitment	2%	Yes
FANCD2	–	• Recruits DNA repair proteins • Is monoubiquitylated by the FA core complex • Forms a heterodimer with FANCI	2%	Unknown
FANCE	–	FA core complex assembly	2%	Unknown
FANCF	–	FA core complex assembly	2%	Unknown
FANCG	–	FA core complex assembly	10%	Unknown
FANCI	–	• Recruits DNA repair proteins • Is monoubiquitylated by the FA core complex • Forms a heterodimer with FANCD2	<2%	Unknown
FANCJ	*BRIP1*	DNA helicase essential for BRCA1-dependent DNA repair	<2%	Yes
FANCL	–	E3 ubiquitin ligase that monoubiquitylates FANCD2 and FANCI	Rare	Unknown
FANCM	–	• FA core complex assembly • DNA helicase involved in repair of Holliday junctions and replication forks • Recruits the BLM helicase during the DDR	Rare	Possible
FANCN	*PALB2*	Regulates BRCA1 localization and DNA repair activity	Rare	Yes
FANCO	*RAD51C*	DNA repair via homologous recombination	Rare	Yes
FANCP	*SLX4*	• Resolution of Holliday junctions • Interacts with several nucleases, including FANCQ	Rare	Unknown
FANCQ	*ERCC4*	DNA repair nuclease	Rare	Unknown
FANCR	*RAD51*	DNA repair via homologous recombination	Rare (dominant-negative)	Unknown
FANCS	*BRCA1*	DNA repair via homologous recombination	Rare	Yes
FANCT	*UBE2T*	E2 ubiquitin-conjugating enzyme for the FANCD2–FANCI monoubiquitylation by FANCL	Rare	Unknown
FANCU	*XRCC2*	DNA repair via homologous recombination	Rare	Unknown
FANCV	*REV7*	Translesion DNA synthesis	Rare	Unknown

1.8 Stratégie de l'étude

Même si les différents systèmes de réparation de l'ADN semblent majoritairement intervenir pour corriger des types distincts de lésions, de plus en plus de publications suggèrent leur chevauchement biologique et leur interdépendance (Haynes *et al.*, 2015; Hashimoto *et al.,* 2016; Shafirovich et Geacintov, 2017). L'altération de gènes codant pour des protéines impliquées dans différentes voies de la réparation pourrait donc avoir un impact cumulatif sur l'une ou plusieurs de ces voies. Nous avons souhaité évaluer l'implication des variants rares localisés dans les gènes de la réparation de l'ADN dans les CCR du sujet jeune, et nous avons donc choisi la stratégie suivante :

1. Nous avons constitué, grâce au recrutement national du laboratoire de génétique moléculaire du CHU de Rouen, un premier échantillon de population constitué de 84 patients avec un phénotype extrême caractérisé par la survenue d'un cancer colorectal jusqu'à l'âge 31 ans, et exempts de variation pathogène dans les gènes impliqués dans la prédisposition héréditaire au CCR. Nous avons comparé cet échantillon à un groupe témoin composé de 567 individus provenant de la population française appariés à nos patients et issus de la cohorte du French Exome Project (FREX).

2. Comme première étape de ce projet, nous avons choisi de restreindre notre étude à un panel de gènes impliqués dans le cancer séquencés par NGS après capture. Des études pangénomiques récentes ont en effet montré que le nombre de gènes contribuant au cancer est plus limité qu'attendu, et qu'ils sont impliqués dans des voies biologiques communes (Vogelstein *et al.*, 2013; Lawrence *et al.*, 2014). Près de la moitié des gènes de prédisposition au cancer décrits sont également affectés par des mutations somatiques dans la tumeur, ce qui confirme le chevauchement, au moins partiel, des voies biologiques impliquées dans les cancers héréditaires et sporadiques (Rahman, 2014). Nous avons déterminé, après revue bibliographique et analyse des bases de données, un panel extensif de 201 gènes du cancer. Les analyses NGS par panel de gènes permettent de plus d'obtenir une profondeur de séquençage plus importante que le WES, ce qui permet de détecter d'éventuelles variations en mosaïque (Hes *et al.*, 2008; Ruark *et al.*, 2013; Friedman *et al.*, 2015).

3. La qualité de toute analyse statistique est directement dépendante de la qualité des données exploitées. Nous avons donc appliqué de nombreux contrôles qualité pour éviter au maximum l'introduction de biais dans notre analyse. Pour renforcer notre puissance statistique après avoir agrégé les variants par gène, nous avons choisi de cumuler les variants des voies biologiques responsables de la réparation de l'ADN. Nous avons testé l'association entre la présence d'au moins un variant touchant au moins un gène impliqué dans la réparation de l'ADN et le développement de tumeurs précoces du côlon ou du rectum, en comparant la proportion de patients porteurs d'au moins un variant dans un gène de la réparation de l'ADN par rapport à la proportion observée chez les témoins.

4. Nous avons ensuite constitué un second échantillon de 93 patients, à nouveau avec un phénotype extrême caractérisé par la survenue d'un cancer colorectal jusqu'à l'âge 32 ans. Parmi ces patients 53 avaient déjà été inclus dans l'échantillon de patients analysés par le panel de 201 gènes, et 40 ont été nouvellement inclus. La même population de témoins a été utilisée.

5. A la lumière des résultats de la première étape du projet et afin de gagner en puissance statistique et d'élargir notre vision, nous avons choisi de séquencer l'exome de ces patients en nous focalisant sur 296 gènes impliqués dans les systèmes cellulaires de réparation de l'ADN (dont 54 communs avec le panel de 201 gènes) par l'application d'un panel *in silico*. L'objectif était de bénéficier des progrès technologiques récents de cette technique, de sa baisse de cout, et de la possibilité de ré-exploiter ultérieurement les données acquises pour s'adapter à l'évolution potentielle des connaissances.

6. A nouveau après l'application de nombreux contrôles qualité, nous avons réalisé une étude statistique pour évaluer l'association des variants rares dans les gènes de réparation de l'ADN à la survenue de CCR précoce. Nous avons réalisé un test de comparaison de la proportion de simples porteurs, un Burden test SeqMeta, un SKAT test, un test d'enrichissement, et un test de comparaison de la proportion de doubles porteurs. Etant donné les variants délétères détectés dans la population de patients, nous avons choisi d'appliquer ces tests d'association en agrégeant les variants rares de l'ensemble des gènes de la réparation de l'ADN d'une part et seulement ceux de la voie ICL d'autre part.

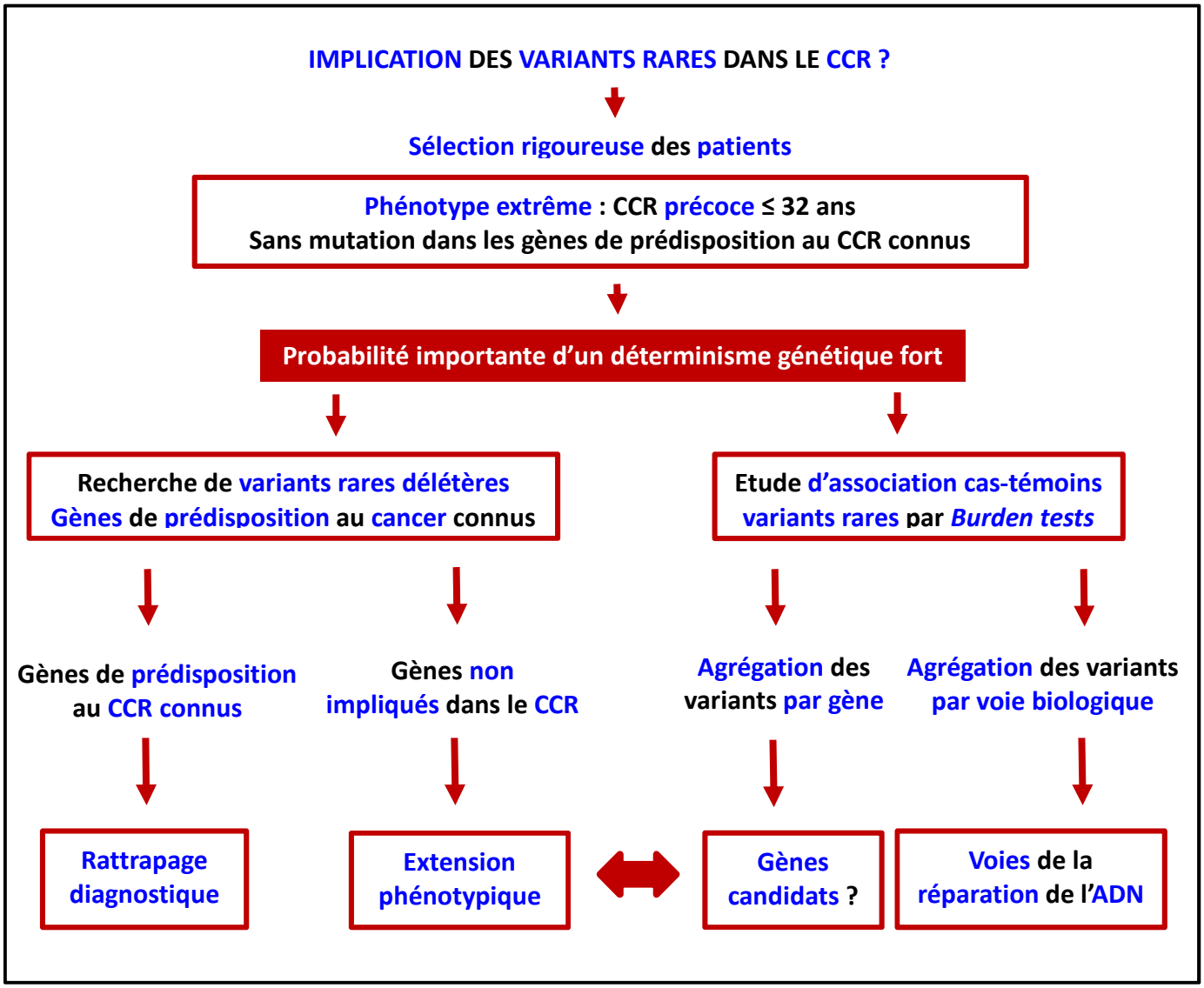


Figure 13 : Stratégie de l'étude

Gènes de prédisposition au CCR connus :

MLH1, MSH2, MSH6, PMS2, APC, MUTYH, STK11, SMAD4, BMPR1A, PTEN
NTHL1, POLE*, POLD1**

* Gènes récemment impliqués dans le CCR n'ayant pas été analysés chez tous les patients lors de l'inclusion dans l'étude

2. MATERIEL ET METHODES

2.1 Sélection des patients

Les critères d'inclusion des patients sont le diagnostic d'un cancer colorectal jusqu'à l'âge précoce de 31 ans. Un seuil initial de 31 ans a été fixé afin de concentrer l'étude sur les formes très précoces de CCR, pouvant être considérées comme phénotype extrême, évocateur d'un déterminisme génétique fort (Barnett *et al.*, 2013). Ces patients remplissent donc tous les critères de Bethesda révisés puisque le diagnostic de CCR est posé avant 50 ans. Les critères d'exclusion sont (i) un syndrome de Lynch, défini par la présence d'une mutation délétère d'un gène MMR (*MLH1, MSH2, MSH6,* ou *PSM2)*, en cas d'analyses non réalisées les patients avec une tumeur de phénotype MSI ont été exclus, (ii) une polypose adénomateuse, définie par un nombre d'adénomes histologiquement confirmés supérieur à 10, ou (iii) une polypose harmatomateuse, définie par la présence d'un hamartome prouvé histologiquement. L'inclusion des patients a été multicentrique, grâce au recrutement national du Centre Normand de Médecine Génomique et de Médecine Personnalisée, partenaire de différents Centres Français de lutte contre le cancer et Centres Hospitaliers Universitaires. Cette inclusion s'est faite de manière majoritairement rétrospective. Ce projet a été soumis au Comité de Protection des Personnes Nord-Ouest I, qui a donné son accord en Octobre 2015. Tous les patients inclus ont consenti de manière libre et éclairée à l'utilisation de leurs prélèvements dans le cadre de la recherche.

- <u>Sélection des patients analysés par le panel de 201 gènes</u>

Pour l'étude par séquençage haut débit d'un panel de 201 gènes, 84 cas index ont été inclus, dont 63% de femmes et 37% d'hommes. La moyenne d'âge au diagnostic de CCR de cette population de patients est de 27 ± 2,4 ans, avec un âge minimal de 14 ans et un âge maximal de 31 ans.

- Sélection des patients analysés par WES

Pour l'étude par NGS de l'exome entier, 93 cas index ont été inclus au total, dont 62% de femmes et 38% d'hommes. Parmi ceux-ci, 53 patients appartiennent également au 1er échantillon analysé par le panel 201 gènes, et 40 patients sont nouvellement inclus. Une exception concernant l'âge d'inclusion a été réalisée pour 3 patients analysés par exome. En effet l'un avait un CCR métastatique à 32 ans, suggérant l'apparition de la tumeur primitive avant 31 ans, et les deux autres n'avaient pas eu de chromocoloscopie les années précédentes, rendant probable une tumorigenèse initiée plus précocement. La moyenne d'âge au diagnostic de CCR de cette population de patients est de 27 ± 3,8 ans, avec un âge minimal de 14 ans et un âge maximal de 32 ans.

2.2 Sélection des témoins

Un total de 567 individus témoins issus de la population française a été inclus dans l'étude. Ces individus témoins ont été recrutés dans 6 villes françaises différentes dans le cadre du *French Exome Project* (FREX ; E Génin, JF Deleuze, R Redon, L Letenneur, D Campion, JC Lambert, France Génomique), auquel notre unité participe, et sont tous d'ascendance française. Leur moyenne d'âge est de 66,6 ±18,3 ans, avec un âge minimal de 19 ans et un âge maximal de 103 ans.

2.3 Séquençage

2.3.1 Sélection des gènes analysés

- Sélection des gènes du panel de 201 gènes

En 2015, nous avons recensé les gènes impliqués dans la transformation maligne, que leurs altérations rapportées soient constitutionnelles ou somatiques. Trois sources ont été utilisées à cet effet :

- Les articles publiés, avec une attention particulière portée sur les preuves génétiques et fonctionnelles permettant de conclure au caractère causal de l'altération du gène décrite dans le cancer.

- Les bases de données des tumeurs telles que la base de données COSMIC (*Catalogue of Somatic Mutations in Cancer*, http://cancer.sanger.ac.uk/cosmic), la base de données Broad Institute (http://www.tumorportal.org) et la base de données TCGA (*The Cancer Genome Atlas*, https://tcga-data.nci.nih.gov/tcga).
- Les différents panels de gènes du cancer développés par des équipes universitaires ou des sociétés commerciales pour détecter les mutations somatiques et/ou constitutionnelles.

Nous avons ainsi sélectionné 201 gènes de cancer (Tableau 2 et 3), en incluant la plupart des gènes déjà impliqués dans une prédisposition héréditaire au cancer. Il est à noter que ce panel étendu de gènes du cancer est destiné à être utilisé par plusieurs équipes de recherche au sein de notre unité Inserm, afin d'explorer la prédisposition au cancer de patients avec un phénotype extrême. Outre notre échantillon de patients avec cancer colorectal précoce, une cohorte de patients avec cancer du sein ou de l'ovaire précoce et une cohorte de patients avec cancers pédiatriques ou tumeurs primitives multiples, ont été également constituées afin d'explorer leur étiologie génétique non élucidée. La constitution de ce panel étendu pourra ainsi permettre l'identification de nouveaux gènes candidats mais aussi évaluer l'extension du spectre tumoral de syndromes de prédisposition au cancer déjà connus.

Tableau 2: Liste des 201 gènes du panel

AIP	*BLM*	*CTNNB1*	*FANCB*	*GNA11*	*KMT2B*	*NBN*	*POLH*	*RUNX1*	*STAG2*
AKT1	*BMPR1A*	*CYLD*	*FANCC*	*GNAQ*	*KMT2C*	*NF1*	*POT1*	*SBDS*	*STK11*
AKT2	*BRAF*	*DDB1*	*FANCD2*	*GNAS*	*KRAS*	*NF2*	*PPM1D*	*SDHA*	*SUFU*
AKT3	*BRCA1*	*DDB2*	*FANCE*	*H3F3A*	*MALT1*	*NKX2-1*	*PRKAR1A*	*SDHAF2*	*TCF7L2*
ALK	*BRCA2*	*DICER1*	*FANCF*	*HNF1A*	*MAP3K1*	*NRAS*	*PTCH1*	*SDHB*	*TERT*
AMER1	*BRIP1*	*DNMT3A*	*FANCG*	*HOXB13*	*MAX*	*NSD1*	*PTEN*	*SDHC*	*TGFBR2*
APC	*BUB1B*	*EGFR*	*FANCI*	*HRAS*	*MDM2*	*NTRK1*	*PTPN11*	*SDHD*	*TMEM127*
ARAF	*CARD11*	*EGLN1*	*FANCL*	*IDH1*	*MDM4*	*PALB2*	*PTPRD*	*SH2D1A*	*TNFAIP3*
ARID1A	*CASP8*	*EPCAM*	*FANCM*	*IDH2*	*MEN1*	*PAX7*	*RAD21*	*SLX4*	*TP53*
ARID5B	*CCND1*	*EPHB2*	*FAS*	*IGF2*	*MET*	*PDGFB*	*RAD50*	*SMAD3*	*TP63*
ATM	*CDH1*	*ERBB2*	*FAT1*	*IKZF1*	*MGMT*	*PDGFRA*	*RAD51*	*SMAD4*	*TRIM33*
ATR	*CDK4*	*ERCC1*	*FBXW7*	*INHA*	*MITF*	*PDGFRB*	*RAD51B*	*SMARCA4*	*TSC1*
ATRX	*CDKN1A*	*ERCC2*	*FGFR1*	*INHBA*	*MLH1*	*PHOX2B*	*RAD51C*	*SMARCB1*	*TSC2*
AURKA	*CDKN1B*	*ERCC3*	*FGFR2*	*INSR*	*MPL*	*PIK3CA*	*RAD51D*	*SMARCE1*	*VHL*
AXIN2	*CDKN2A*	*ERCC4*	*FGFR3*	*IRF4*	*MRE11A*	*PIK3CD*	*RAF1*	*SMC1A*	*WRN*
BAP1	*CDKN2B*	*ERCC5*	*FGFR4*	*IRS1*	*MSH2*	*PIK3R1*	*RB1*	*SMO*	*WT1*
BARD1	*CDKN2C*	*EXT1*	*FH*	*JAK1*	*MSH6*	*PIK3R2*	*RECQL4*	*SOS1*	*XPA*
BCL2	*CHEK2*	*EXT2*	*FHIT*	*JAK2*	*MUTYH*	*PML*	*RET*	*SPEN*	*XPC*
BCL2L11	*CREBBP*	*FAM157A*	*FLCN*	*JAK3*	*MYC*	*PMS2*	*RINT1*	*SPOP*	*XRCC2*
BCL9	*CTNNA1*	*FANCA*	*GATA2*	*KIT*	*MYCN*	*POLE*	*RNASEL*	*SPRED1*	*XRCC3*
BIRC5									

Tableau 3 : Liste des 54 gènes impliqués dans la réparation de l'ADN parmi les 201 gènes du panel

MMR	NER	ICL	HR	HS	Autres
MLH1	*DDB1*	*ATR*	*MRE11A*	*ATM*	*ATRX*
MSH2	*DDB2*	*BARD1*	*RAD21*	*BLM*	*CDKN1A*
MSH6	*ERCC1*	*BRCA1*	*RAD50*	*RECQL4*	*CDKN2A*
PMS2	*ERCC2*	*BRCA2*	*RAD51*	*WRN*	*CHEK2*
	ERCC3	*FANCA*	*RAD51B*		*MDM2*
	ERCC4	*FANCB*	*RAD51C*		*MDM4*
	ERCC5	*FANCC*	*RAD51D*		*MGMT*
	XPA	*FANCD2*	*SLX4*		*MUTYH*
	XPC	*FANCE*	*XRCC2*		*POLE*
		FANCF	*XRCC3*		*POLH*
		FANCG			*TERT*
		FANCI			*TP53*
		FANCL			
		FANCM			
		PALB2			

- Sélection des gènes des panels *in silico* appliqués aux données de WES

Panel in silico diagnostique :

Un panel *in silico* de 14 gènes responsables de syndromes de prédisposition héréditaire au CCR a été appliqué aux données de WES. Les gènes de ce panel diagnostique sont ceux actuellement analysés au laboratoire de Génétique Moléculaire du CHU de Rouen en cas de suspicion de forme héréditaire de CCR. Ces gènes sont : *APC, BMPR1A, MLH1, MSH2, MSH3, MSH6, MUTYH, NTHL1, PMS2, PTEN, POLD1, POLE, SMAD4* et *STK11*.

Panel in silico de 296 gènes impliqués dans la réparation de l'ADN :

A la lumière des résultats préliminaires du séquençage par panel de gènes, nous avons voulu élargir notre analyse en nous focalisant sur 296 gènes (Tableau 4) impliqués dans les systèmes cellulaires de réparation de l'ADN. Ces gènes ont été sélectionnés et classés selon la voie de réparation de l'ADN dans laquelle ils sont majoritairement impliqués. Deux sources principales ont été utilisées :

- Une revue extensive de la bibliographie
- Le panel de gènes de la réparation utilisé et détaillé par le consortium international ExAC (Das et Ghosh, 2017).

Tableau 4 : Liste des 296 gènes impliqués dans la réparation de l'ADN

BER	MMR	NER	HR	NHEJ	ICL	Pol	O	DDDR	DPCR
APEX1	*MLH1*	*ASF1A*	*BARD1*	*ABL1*	*BRCA2*	*MAD2L2*	*ATR*	*ALKBH1*	*TDP1*
APEX2	*MLH3*	*ATXN3*	*BCCIP*	*ATRX*	*BRIP1*	*PAPD7*	*ATRIP*	*ALKBH2*	*TDP2*
ERCC6L	*MSH2*	*CCNH*	*BRCA1*	*DCLRE1C*	*CENPS*	*PCNA*	*CDKN2A*	*ALKBH3*	
HMGB1	*MSH3*	*CDK7*	*CHD4*	*DNTT*	*FAAP100*	*POLA1*	*CETN3*	*MGMT*	
HMGB2	*MSH4*	*CDKN1A*	*DMC1*	*KAT5*	*FAAP20*	*POLA2*	*CHD1L*		
LIG3	*MSH5*	*CEBPG*	*EME1*	*LIG4*	*FAAP24*	*POLB*	*CHEK1*		
MBD4	*MSH6*	*CETN2*	*EME2*	*NHEJ1*	*FAN1*	*POLD1*	*CHEK2*		
MPG	*PMS1*	*CRY1*	*ESCO2*	*PRKDC*	*FANCA*	*POLD2*	*CIB1*		
MUTYH	*PMS2*	*DDB1*	*FIGNL1*	*PTTG1*	*FANCB*	*POLD3*	*CLK2*		
NEIL1	*RPA1*	*DDB2*	*GEN1*	*RBM14*	*FANCC*	*POLD4*	*CRY2*		
NEIL2	*RPA4*	*ERCC1*	*MATR3*	*SFPQ*	*FANCD2*	*POLE*	*CSNK1E*		
NEIL3		*ERCC2*	*MRE11*	*XRCC4*	*FANCE*	*POLE2*	*DCLRE1A*		
NTHL1		*ERCC3*	*MUS81*	*XRCC5*	*FANCF*	*POLE3*	*DCLRE1B*		
OGG1		*ERCC4*	*NBN*	*XRCC6*	*FANCG*	*POLE4*	*EMSY*		
PARP1		*ERCC5*	*NSMCE1*		*FANCI*	*POLG*	*HELQ*		
PARP2		*ERCC6*	*NSMCE2*		*FANCL*	*POLG2*	*HUS1*		
PARP3		*ERCC8*	*RAD21*		*FANCM*	*POLH*	*MDC1*		
PARP4		*GADD45A*	*RAD50*		*PALB2*	*POLI*	*MDM4*		
PNKP		*GADD45B*	*RAD51*		*SLX4*	*POLK*	*N4BP2*		
SMUG1		*GADD45G*	*RAD51AP1*		*UBE2T*	*POLL*	*NABP2*		
SUMO1		*GTF2H1*	*RAD51B*		*USP1*	*POLM*	*NIPBL*		
TDG		*GTF2H2*	*RAD51C*		*WDR48*	*POLN*	*PER1*		
UNG		*GTF2H3*	*RAD51D*			*POLQ*	*PRPF19*		
XRCC1		*GTF2H4*	*RAD52*			*REV1*	*RAD1*		
		GTF2H5	*RAD54B*			*REV3L*	*RAD17*		
		HTATIP2	*RAD54L*			*RFC1*	*RAD54L2*		
		LIG1	*RBBP4*			*RFC2*	*RAD9A*		
		MMS19	*RBBP8*			*RFC3*	*RDM1*		
		MNAT1	*RMI1*			*RFC4*	*RECQL*		
		RAD23A	*RMI2*			*RFC5*	*RECQL5*		
		RAD23B	*RUVBL2*			*TEP1*	*RIF1*		
		RPA2	*SEM1*			*TERF1*	*SMARCA1*		
		RPA3	*SLX1A*			*TERF2*	*SMARCAD1*		
		UVSSA	*SLX1B*			*TERT*	*SMARCAL1*		
		XAB2	*SMC1A*			*TINF2*	*TOPBP1*		
		XPA	*SMC1B*			*TOP1*	*TP53*		
		XPC	*SMC3*			*TOP2A*	*TP53BP1*		
			SMC5			*TOP2B*	*TP53I3*		
			SMC6			*TOP3A*			
			SYCP3			*TOP3B*			
			WDR33						
			WRNIP1						
			XRCC2						
			XRCC3						

SSBR	DSBR	MNP	Edit	CP	CS	Ub	HS
APTX	*APLF*	*DUT*	*ENDOV*	*CLSPN*	*CHAF1A*	*CUL4A*	*ATM*
	ATP23	*NUDT1*	*EXO1*	*CSNK1D*	*CHAF1B*	*CUL4B*	*BLM*
	MTOR	*NUDT3*	*FEN1*	*MAD1L1*	*H2AFX*	*HLTF*	*MPLKIP*
		RRM1	*SPO11*	*MAD2L1*	*HDAC8*	*MDM2*	*RECQL4*
		RRM2	*TREX1*	*RAD9B*	*MBD1*	*RAD18*	*WRN*
		RRM2B	*TREX2*	*RUVBL1*	*MBD2*	*RBX1*	
					MBD3	*RNF168*	
					MBD5	*RNF4*	
					MBD6	*RNF8*	
					SETMAR	*SHPRH*	
					SSRP1	*SPRTN*	
					SUPT16H	*UBA1*	
						UBE2A	
						UBE2B	
						UBE2D2	
						UBE2D3	
						UBE2I	
						UBE2L3	
						UBE2N	
						UBE2V1	
						UBE2V2	

Voies de réparation de l'ADN :

BER : *Base Excision Repair*

MMR : *MisMatch excision repair*

NER : *Nucleotide Excision Repair*

HR : *Homologous Recombination*

NHEJ : *Non-Homologous End-Joining*

ICL : *Interstand crosslink*

DDDR : *Direct DNA damage reversal*

DPCR : *DNA-protein crosslink repair*

SSBR : *Single Strand Break Repair*

DSBR : *Double Strand Break Repair*

MNP : *Modulation of Nucleotide Pools*

Pol : *Polymerases, telomerases, topoisomerases, accessory proteins*

Edit : *Editing and processing nucleases*

CP : *DNA damage Checkpoint*

CS : *Chromatine Structure*

Ub : *Ubiquitination (abnormal or short-lived proteins degradation)*

HS : *Hyper-sensitivity to DNA damaging agents implicated genes*

O : *Others > Suspected DNA repair function (structure homology…)*

2.3.2 Capture, séquençage et analyses bio-informatiques

- Capture, séquençage et analyses bio-informatiques de l'ADN des patients par le panel de 201 gènes

Nous avons mis au point fin 2015 un design à façon incluant tous les exons des 201 gènes sélectionnés, ainsi que les 25 paires de bases introniques flanquantes des exons. Nous avons choisi un chevauchement des sondes de 5x et un enrichissement spécifique des régions riches en GC. La taille de la région génomique capturée totale est d'environ 1 Mb.

Pour la plupart des patients, l'ADN a été extrait au laboratoire à partir d'un prélèvement périphérique de sang total sur EDTA. Le kit utilisé pour l'extraction est le kit *QuickGene DNA whole blood kit L* sur un appareil semi-automatique *QuickGene-610L* (*Autogen Inc.*).

La concentration en ADN double brin des échantillons avant et après préparation des librairies est évaluée par fluorimétrie en utilisant les kits *Quant-it™ dsDNA Assay Kit, Broad Range* et *Quanti-it™ dsDNA Assay Kit, High Sensitivity* (*Invitrogen*) et l'automate *Twinkle LB 970* (*Berthold Technologies*). Les étapes de préparation des librairies, de capture et de normalisation sont réalisées manuellement ou à l'aide d'une plateforme robotique dotée d'un robot *Sciclone® G3 NGS Workstation* (*PerkinElmer Inc.*) optimisée par l'utilisation du kit *SureSelect^QXT* (*Agilent Technologies*). La première étape est appelée « tagmentation » et correspond à la fragmentation enzymatique aléatoire de 50 ng d'ADN par échantillon combinée à l'ajout de séquences permettant l'hybridation des amorces des PCR d'enrichissement. Par la suite, une première PCR, d'une durée de 8 cycles, permet d'enrichir la préparation en matériel et d'ajouter les séquences nécessaires à l'hybridation des amorces de séquençage. Après purification par billes magnétiques grâce au kit *Agencourt AMPure XP* (*Beckman Coulter Life Sciences*), la taille et la concentration des fragments obtenus sont contrôlées par migration sur piste *D1000 ScreenTape* (*Agilent Technologies*) sur un appareil *TapeStation 2200* (*Agilent Technologies*). La quantité d'ADN de chaque échantillon peut ainsi être normalisée avant l'étape de capture. Après capture à l'aide de *Dynabeads®*

MyOne™ Streptavidin T1 (*Invitrogen*) et lavages successifs, une seconde PCR d'enrichissement est réalisée. Cette PCR de 14 cycles permet l'ajout d'index spécifiques de chaque patient. Après une seconde étape de purification par billes magnétiques, chaque librairie est à nouveau contrôlée par migration sur la Tape station et quantifiée par fluorimétrie, afin de réaliser un mélange équimolaire de l'ensemble des échantillons. Ce mélange équimolaire est finalement séquencé par synthèse en «paired-end» 2 x 150 pb sur un séquenceur *Illumina NextSeq500*. L'analyse primaire des données, comprenant l'évaluation de la qualité des données, l'analyse des images et l'appel des bases (*base calling)*, est réalisée en temps réel pendant le séquençage par l'outil RTA 2.0 (*Real Time Analysis*, *Illumina*). Les données issues du séquenceur sont ensuite prises en charge par le pipeline bio-informatique de la plateforme génomique de Rouen. Les outils BWA v0.7.6a (*Burrows-Wheeler Alignment tool*), PICARD v1.108 et GATK v3.4 (*Genome Analysis ToolKit*, *Broad Institute*) sont utilisés pour l'analyse secondaire afin de réaliser le démultiplexage, l'alignement et l'appel de variant (*variant calling)* (Van der Auwera *et al.,* 2013). Le génome de référence utilisé est celui d'*Homo Sapiens* version GRCh37. Nous avons choisi de retirer les duplicats de PCR afin de limiter l'enrichissement de nos résultats en artefacts. Cette méthode nous permet de conserver une profondeur de 500X, suffisante pour la détection d'éventuelles mosaïques.

- <u>Capture, séquençage et analyses bio-informatiques de l'ADN des patients et des témoins par WES</u>

Le séquençage des WES est effectué au Centre National de Recherche en Génomique Humaine (CNRGH) à Evry, dirigé par Monsieur Jean-François Deleuze, avec lequel notre laboratoire collabore étroitement depuis plusieurs années. Le séquençage des WES des patients et des témoins du projet FREX a été réalisé par ce centre, ce qui nous permet lors de l'étude statistique de comparer des données analysées dans des conditions proches et donc de réduire les éventuels biais techniques. Pour chaque individu sélectionné, 4 à 8 µg d'ADN sont envoyés par notre centre, accompagné du code anonyme de l'échantillon, du genre du patient, de la méthode d'extraction, ainsi que de la concentration et du volume de la prise d'essai correspondants. Après réception, le CNG s'assure de la bonne qualité des échantillons reçus : la concentration est vérifiée, l'intégrité de l'ADN est évaluée par migration sur un appareil *TapeStation 4200* (*Agilent Technologies*), une PCR est réalisée afin de détecter la présence d'éventuels inhibiteurs, et un génotypage permet de vérifier la concordance avec le genre renseigné. La capture des exomes est effectuée sur la plateforme robotique du CNRGH à

l'aide de kits *Agilent SureSelect Human All Exon V5+UTR* (*Agilent technologies*) pour les témoins, et de kits *Agilent SureSelect Human All Exon V6* (*Agilent technologies*) pour les patients, séquencés plus tardivement. Le séquençage a été réalisé par synthèse en «*paired-end*», en 75 ou 100 pb sur un séquenceur *Illumina HiSeq2000* ou *2500* pour les témoins, et en 2 x 150 pb sur un séquenceur *Illumina HiSeq4000* pour les patients. Les données brutes, après alignement sur le génome d'*Homo Sapiens* version GRCh37, démultiplexage et *variant calling* par BWA 0.7.5a, Picard 1.101 (*Broad Institute*) et GATK 3.3-0 (*Broad Institute*), sont transmises sous forme de FASTQ et de BAM à l'équipe bio-informatique de la plateforme génomique de Rouen pour être analysées.

Enfin l'analyse tertiaire, correspondant à l'annotation des variants, est effectuée pour l'ensemble des patients et les témoins par snpEff v4.2 (Cingolani *et al*., 2012). Ce pipeline est complété par le logiciel de détection des remaniements génomiques CANOES (Backenroth *et al*., 2014). Par ailleurs, un *variant calling* alternatif est réalisé par le logiciel *VarScan v2.3.9* (*Washington University*), dont les résultats sont annotés par *Alamut® Batch : v1.5.2 Standalone* (*Interactive BioSoftware*) afin d'optimiser la détection des mosaïques (Koboldt *et al*., 2009).

2.3.3 Filtration et interprétation des variants détectés

Afin de détecter des variants délétères rares voire privés susceptibles d'être à l'origine d'une prédisposition héréditaire au CCR à forte pénétrance donc à haut risque, nous avons réalisé une filtration bio-informatique séquentielle :

(i) Tout d'abord, seuls les variants avec une qualité suffisante sont conservés. Les seuils appliqués sont adaptés à la méthode de séquençage. Les variants détectés par le panel de 201 gènes conservés sont ceux avec un *Genotype Score* > 80 et une profondeur > 20. Les variants détectés par WES puis *panel in silico* de 296 gènes conservés sont ceux avec un *Genotype Score* > 20 et une profondeur > 10. Ces derniers scores sont plus permissifs car la profondeur de séquençage obtenue par cette technique est plus faible. (ii) Une filtration sur la MAF est ensuite réalisée afin de ne conserver que les variants rares, donc avec une MAF < 1% dans les bases de données de population ExAC (NFE : *Non Finish European*, à l'exclusion de la cohorte TCGA qui regroupe des patients atteints de cancers ; Lek *et al.,* 2016), ESP6500 (EA : *European American* ; *Exome Variant Server, NHLBI GO Exome Sequencing Project, Seattle, WA*) et 1000Genome (EUR : *Europeans ; Auton et al., 2015*).

(iii) Un autre filtre sélectionne les variants supposés strictement délétères, ce qui inclut tous les variants disruptifs impliquant une perte de fonction (*Loss Of Function* ; LOF) ainsi que les variants faux sens exoniques prédits délétères par les 3 logiciels de prédiction de pathogénicité que sont SIFT, Polyphen-2 HVAR et MutationTaster (versions utilisées par dbNSFP v3.3 ; Adzhubei *et al.*, 2010; Liu *et al.*, 2011, 2016; Schwarz *et al.*, 2014; Sim *et al.*, 2012). Les variants LOF conservés sont ceux qui provoquent un changement de cadre de lecture, la perte ou le gain d'un codon stop, et les variants introniques touchant les sites consensus d'épissage en -1, -2 ou +1, +2. Les variants affectant les derniers 5% de la protéine ont été exclu de la catégorie LOF, car la protéine est alors moins sensible à la troncature (MacArthur *et al.*, 2012). Le nettoyage des artefacts a été effectué manuellement par la lecture directe des fichiers BAM (*Binary Alignment Map*) de l'ensemble des variants identifiés par le panel de 201 gènes, et comparé au nettoyage automatique du VQSR, détaillé dans la partie statistique. Les variants identifiés par exome avec un mauvais score de VQSR ont été d'emblée éliminés, et les BAM des variants d'intérêt identifiés chez les patients ont été lus dans un second temps.

L'interprétation biologique et la classification de chaque variant sont finalement réalisées à l'aide du logiciel *Alamut® Visual* (*Interactive Biosoftware*), de l'interface web wIntervar (*Wang Genomic Lab*), et d'une revue de la bibliographie (Li & Wang, 2017). La classification en 5 classes issue des recommandations internationales de *l'American College of Medical Genetics and Genomics* (ACMG) et de *l'Association for Molecular Pathology* (AMP) est appliquée (Amendola *et al.,* 2016; Richards *et al.,* 2015). Cette classification est basée sur 28 critères incluant notamment la nature de la variation, la présence et la fréquence de la variation dans les bases de données de population (ExAC, ESP6500 et 1000Genomes) et de patients (InSIGHT : *International Society for Gastrointestinal Hereditary Tumours Database*, ClinVar, UMD : *Universal Mutation Database*), les résultats des algorithmes de prédiction de pathogénicité permettant de prédire *in silico* l'effet de la variation sur la protéine, les données fonctionnelles, et les données de ségrégation. Pour nous aider dans cette classification et dans l'interprétation, nous avons également appliqué des algorithmes de prédiction de l'impact de la variation sur l'épissage: SpliceSiteFinder-like, MaxEntScan, NNSPLICE 0.9, GeneSplicer et Human Splicing Finder 2.4.1 (Reese *et al.* 1997; Zhang 1998; Pertea *et al.* 2001; Yeo et Burge 2004; Desmet *et al.* 2009). Nous avons par ailleurs systématiquement consulté, pour chaque variant détecté, la valeur de la pLi (*probability of Loss of function intolerance*) fournie par la base de données ExAC du gène portant la

variation, ainsi que l'expression de celui-ci dans les tissus digestifs, grâce à la base de données GTEx (Lonsdale *et al*., 2013). La nomenclature des variants utilisée dans cette thèse est la nomenclature HGVS, recommandée et internationale. Les NM utilisés pour décrire ces variants correspondent à ceux de l'isoforme principale de la protéine codée par le gène, d'après la base de données APPRIS 2018_02.v27, construite par le *GENCODE consortium*, appartenant au *ENCODE project* (Rodriguez *et al.*, 2018).

2.3.4 Confirmation des variants d'intérêt

- Séquençage Sanger

Certains variants d'intérêt ont été vérifiés par séquençage Sanger. La vérification n'a pas été systématique puisque, d'après une étude de 2016, la confirmation par Sanger de variants dotés de bons score de qualité en NGS est plus à risque d'invalider un vrai positif du NGS plutôt que de mettre en évidence un faux positif (Beck *et al*., 2016). Des amorces de PCR ont été choisies pour chaque variant à partir du logiciel primer3 (http://bioinfo.ut.ee/primer3-0.4.0/). Une étape d'amplification de l'ADN par PCR est réalisée et purifiée, suivie d'une réaction de séquençage, également purifiée. Les fragments sont finalement migrés sur un séquenceur capillaire ABI3500 (*Applied Biosystems*), et les séquences sont analysées à l'aide du logiciel SeqA6 (*Applied Biosystems*).

- PCR digitale

Les CNV d'intérêt ont été vérifiés par une technique de PCR digitale en émulsion (*Digital Droplet PCR*, ddPCR) développée par le service de génétique et l'unité Inserm U1245 du CHU de Rouen. Cette technologie repose sur la quantification relative, dans la même réaction, d'un amplicon associé à une sonde d'hydrolyse émettant une fluorescence FAM localisé au locus du CNV d'intérêt, par rapport à un amplicon de référence, situé au locus du gène de ménage *HMBS*, associé à une sonde d'hydrolyse émettant une fluorescence VIC. Un protocole standard a été utilisé sur la plateforme ddPCR QX200 (Bio-Rad, Hercules, CA, USA). Ce protocole est composé d'une première étape de fragmentation de 40 ng d'ADN génomique par la chaleur à 98°C pendant 10 minutes. L'ADN fragmenté est mélangé au supermix ddPCR (BioRad) dédié, aux amorces de l'amplicon cible et de l'amplicon de référence, ainsi qu'aux sondes d'hydrolyse marquées FAM pour la séquence d'intérêt et VIC

pour la séquence référence. L'échantillon est ensuite divisé en 15 000 à 20 000 gouttelettes de 5 nL, qui servent chacune de puit réactionnel, par un système de générateur de gouttelettes QD200R. L'amplification par PCR est réalisée sur un thermocycleur BioRad S1000 : une étape initiale de dénaturation dure 95 min à 95°C, puis 40 cycles d'une étape de dénaturation à 94°C pendant 30 secondes, puis d'extension à 56°C pendant 1 minute sont réalisés, et enfin une dénaturation a lieu à 98°C pendant 10 minutes. La lecture et la quantification des gouttelettes ont été effectuées par le lecteur QX200. Le nombre de copies de la séquence d'intérêt est calculé par le rapport de la concentration de la cible marquée par FAM sur celle de la référence VIC. L'analyse des données a été réalisée à l'aide du module 2D du logiciel QuantaSoft (Bio-Rad).

2.4 Analyses statistiques

2.4.1 Contrôles qualité des individus

Ces contrôles s'effectuent en deux étapes. Tout d'abord, une première étape purement technique permet de s'assurer que des erreurs dans le processus de séquençage (artefact, perte d'un allèle, contamination) ne viendront pas fausser les résultats de certains de nos patients ou témoins. Différents critères de qualité sont systématiquement examinés pour chaque individu :

- Le ratio hétérozygotes/homozygotes : le ratio entre le nombre d'allèles alternatifs portés à l'état hétérozygote et homozygote parmi les variants connus et répertoriés dans dbSNP147 est calculé pour chaque individu. Si ce ratio observé s'écarte du ratio moyen de plus de 6 déviations standards, le ratio est considéré comme aberrant et l'individu est exclu de l'étude.
- Le ratio transition/transversion : un processus identique au ratio hétérozygotes/homozygotes est appliqué.
- Le nombre de singletons (variants mis en évidence une seule fois parmi l'ensemble des patients et témoins)
- La *Missingness* : Le taux de valeurs manquantes sur la zone de capture est déterminé pour chaque patient et témoin. S'il est supérieur à 15%, la capture est considérée comme de mauvaise qualité et l'individu est exclu de l'étude.

Une seconde étape méthodologique vient exclure les individus qui, en dépit d'une bonne qualité de séquençage, ne répondent pas aux critères d'éligibilité pour une analyse

d'association. En effet, pour mener à bien une analyse statistique d'association non biaisée il est nécessaire de s'assurer que les individus inclus dans l'étude (i) constituent un échantillon homogène permettant une estimation robuste des fréquences alléliques et (ii) sont indépendants les uns des autres.

A l'intersection entre les deux étapes du contrôle des individus, le nombre de singletons (variants mis en évidence une seule fois dans l'étude parmi l'ensemble des patients et témoins) est calculé pour chaque individu à la recherche d'un éventuel excès du nombre de singletons portés. En effet, un excès en singletons peut être le reflet de la survenue de nombreux artefacts de séquençage. En réalité, l'effet des artefacts de séquençage sur ce paramètre est minime à côté des effets de structure de population observés. Le nombre de singletons par individu met principalement en évidence les individus dont le fond génétique diffère fortement de la majorité de l'échantillon. Les singletons observés correspondent de fait à des allèles spécifiques d'autres populations, potentiellement fréquents voire fixés dans ces populations, mais absents de la population dont est issue la majorité de l'échantillon.

L'homogénéité génétique de la population est vérifiée par ailleurs à l'aide d'une analyse de stratification de population réalisée par Analyse en Composante Principale (ACP). Réalisée en parallèle d'une part sur un jeu de variants fréquents (MAF >5%) et un jeu de variants de faible fréquence (MAF<5%), cette analyse permet de résumer sur deux axes principaux l'ensemble de l'information génétique portée par les individus et de visualiser la distance génétique moyenne entre individus. Les individus isolés de la majorité de l'échantillon sur ces axes, dits *outliers*, sont exclus de l'analyse. Les analyses parallèles en variants rares et fréquents, jointes au dénombrement des singletons, permettent de conforter la décision d'exclusion.

La distance génétique entre les individus est également déterminée par le calcul de l'*Identity By State* (logiciel Plink). A l'opposé de l'analyse en composantes principales, qui met efficacement en évidence les individus divergeant de la majorité de l'échantillon, cette analyse permet une estimation du degré d'apparentement entre individus afin d'exclure d'éventuels doublons ou apparentés au premier ou au deuxième degré.

2.4.2 Contrôles qualité des variants

Après avoir exclu les individus à risque de biaiser l'étude statistique, seuls sont conservés les variants de MAF < 1% sur l'ensemble de l'échantillon. Six paramètres de qualité sont ensuite évalués pour caractériser la qualité de chaque variant et décider de sa conservation ou non dans l'étude statistique. Encore une fois, une première partie de ces critères a pour objectif de déceler les variants de mauvaise qualité intrinsèque. La seconde partie vient vérifier des critères d'éligibilité pour l'analyse statistique.

Les trois critères de qualité intrinsèque suivant sont appliqués :

- **Variant Quality Score Log-Odds (VQSLOD)** : le VQSLOD est un score de probabilité de qualité attribué à un variant à l'échelle d'un individu et calibré par un algorithme d'apprentissage automatique, le VQSR (*Variant Quality Score Recalibration*), résumant l'information de qualité apportée par l'ensemble des individus sur ce variant. Ce dernier s'entraîne à distinguer les vrais variants et les faux positifs à partir de variables reflétant la qualité des étapes d'alignement ou de détection des variants sur des bases de données de variants connus (HapMap et 1000G par exemple) puis applique le modèle estimé à nos données pour fournir pour chaque variant un score lié à la probabilité que le variant soit vrai plutôt que faux. Conformément aux recommandations de bonnes pratiques de GATK, les variables utilisées pour calculer le VQSLOD des SNVs et des indels sont :
    - *QualByDepth* : qualité du *variant calling* pondéré par la profondeur de lecture à cette position
    - *StrandOddsRatio & FisherStrand* : biais de *strand* estimé par un OR et un test de Fisher, comparant la proportion d'allèle alternatif portés dans chaque sens de lecture.
    - *MappingQualityRankSumTest* : comparaison de la qualité d'alignement des *reads* de l'allèle de référence par rapport à ceux de l'allèle alternatif
    - *ReadPosRankSum* : recherche d'un biais de position des allèles de références et alternatifs sur leurs *reads* respectifs dans l'objectif de détecter des allèles alternatifs issus d'erreurs de séquençage spécifiques des fins de *reads*.
    - *Inbreeding coeff* : évaluation de la déviation du modèle de Hardy-Weinberg permettant de mettre en évidence un excès d'hétérozygotie ou d'homozygotie

Conformément aux recommandations de GATK, seuls les SNPs jusqu'à un seuil de VQSLOD correspondant à une sensibilité de 99,5 % et les INDELs jusqu'à un seuil de sensibilité de 99 % sont conservés.

- **Genotype Quality** fournis par *HaplotypeCaller* : ce score est appliqué à l'échelle du génotype, individu par individu, et non à l'échelle globale du variant. Les génotypes de qualité inférieure à 20 sont considérés comme trop incertains et donc manquants.
- **Balance allélique** : exclusion des variants avec une balance allélique moyenne inférieure à 90% pour les variants portés à l'état homozygote, et en moyenne supérieure à 75% ou inférieure à 25% pour les variants portés à l'état hétérozygote.

Pour ne pas biaiser les analyses d'association, les variants doivent par ailleurs vérifier les critères suivants :

- *Missingness* : exclusion des variants dont le génotype est manquant pour plus de 5% des individus de l'échantillon.
- *Missingness différentielle* : exclusion des variants dont la valeur p (*p-value*) obtenue avec le test de Fisher, comparant le nombre de génotypes manquants entre les patients et les témoins, est inférieure à 10^{-7}.
- *Test d'équilibre de Hardy-Weinberg chez les témoins* : exclusion des variants en déséquilibre dont la *p-value* au test de $\chi 2$ réalisé entre effectifs observés et théoriques est inférieure à 10^{-7}.

2.4.3 Tests d'association

Nous avons effectué une étude d'association cas-témoins sur une population rendue homogène après l'application de la stratification de population. Afin d'augmenter la puissance statistique de l'étude, nous avons choisi d'agréger les variants à l'échelle du gène et à l'échelle de listes de gènes. En effet, à l'image des tests classiques d'association entre un variant génétique et une maladie, la puissance des tests d'association du type *Burden test*, dépend à la fois de l'Odds Ratio (OR) moyen attendu et de la fréquence allélique cumulée des variants agrégés dans ce *Burden test*. En agrégeant les variants à l'échelle d'un gène ou d'une liste de gènes, on augmente la fréquence allélique cumulée des variants. Si par ailleurs l'ensemble des variants agrégés partagent le même OR et qu'aucun effet dilution ne joue, la puissance s'en trouve mécaniquement augmentée.

Différents tests statistiques ont été effectués, répondant à des questions différentes :

- ***Collapsing test*** : comparaison de la proportion de patients porteurs d'au moins un variant dans les gènes considérés, par rapport à la proportion de témoins porteurs.
- ***Burden test* de type seqMeta** : comparaison du nombre total de variants portés par les patients au nombre total de variants portés par les témoins dans les gènes considérés

Pour ces deux tests, il est important que l'effet des variants agrégés soit homogène, sinon l'odds ratio calculé est une moyenne des odds ratio de chacun des variants agrégés. L'inclusion de variants neutres, protecteurs ou de faible risque viendrait diluer l'effet de variants à haut risque.

- **SKAT test** : test analysant la variance des OR de l'ensemble des variants agrégés. Ce test présente l'avantage de permettre la détection de gènes ou listes de gènes avec présence simultanée de variants protecteurs et à risque, mais n'est pas très sensible aux variants très rares (Wu *et al*., 2011).
- **Test d'enrichissement** : comparaison du nombre de variants détectés dans les voies de réparation de l'ADN par rapport au reste de l'exome, chez les patients et les témoins. Contrairement aux trois tests précédents, qui s'intéressent à l'association intrinsèque entre les variants d'un gène ou d'une liste de gènes et une maladie, ce test cherche à démontrer la spécificité d'un signal d'association à un gène ou liste de gène, en concurrence avec le reste de l'exome.
- ***Collapsing* test des co-occurrences** : comparaison de la proportion de patients porteurs de doubles occurrences dans les voies de réparation de l'ADN, par rapport à la proportion de témoins porteurs.

A l'exclusion du *Burden test* et du SKAT test pour lesquels la librairie seqMeta du logiciel R a été utilisée, nous avons utilisé un test exact de Fisher pour calculer un OR, un intervalle de confiance (IC95%) et une *p-value*.

Le seuil de significativité des tests statistiques approprié aux données d'un panel de gènes est le même que pour un exome, puisque l'on considère qu'il est possible de tester tous les gènes ou toutes les combinaisons de gènes jusqu'à atteindre une valeur significative. Il est

nécessaire d'appliquer la correction de Bonferroni au seuil de *p-value* communément admis de $p = 0,5$. Cette correction dépend du nombre de tests réalisés donc des environs 20 000 gènes du génome humain. Le seuil de significativité qui doit être retenu est donc de $p = 0,05 / 20\,000 = 2,5.10^{-6}$ (Lee *et al.*, 2014; Wang *et al.*, 2015).

3. RESULTATS

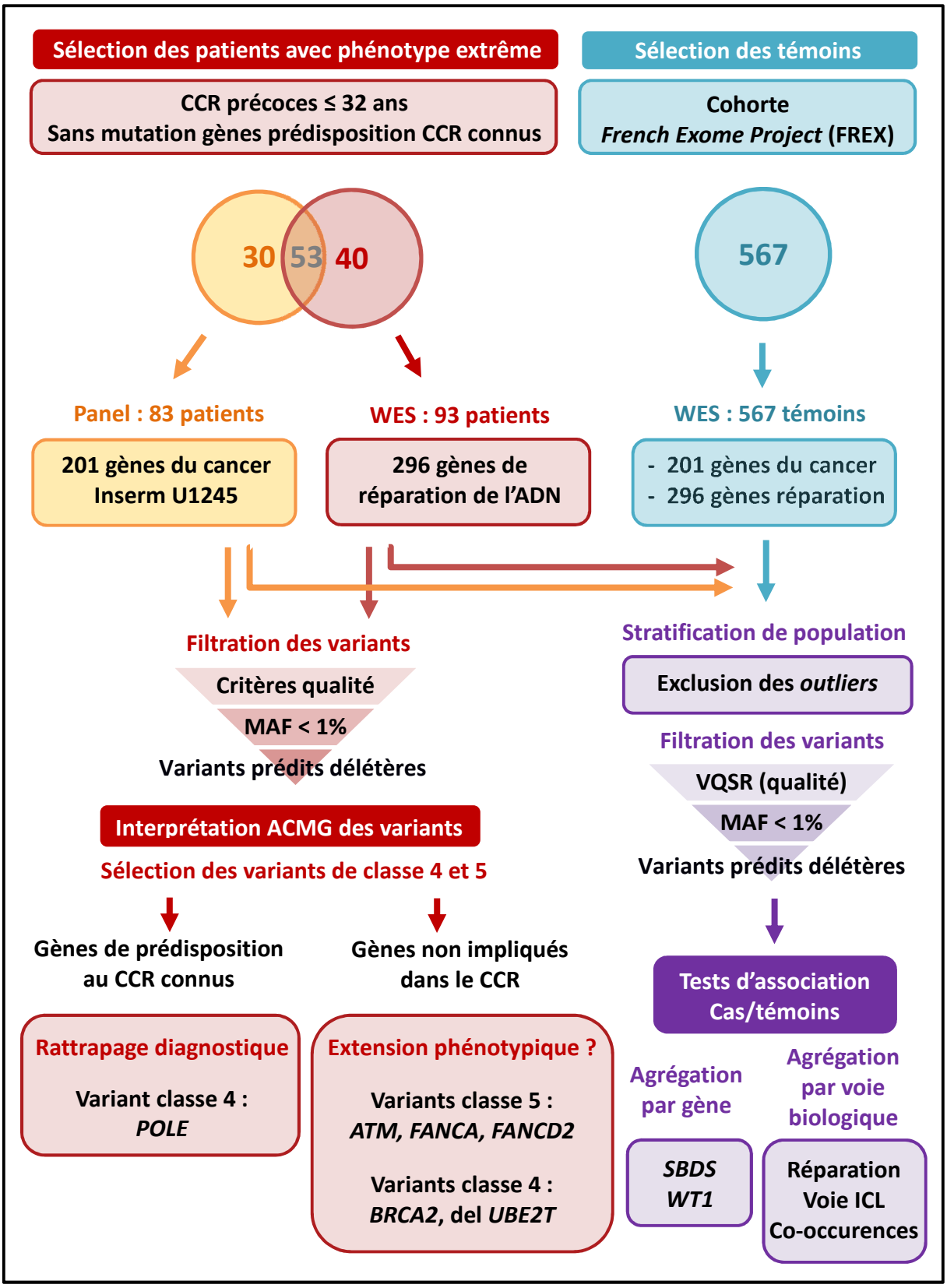


Figure 14 : Résultats de l'étude

3.1 Résultats de l'analyse du panel de 201 gènes impliqués dans l'oncogenèse

Le premier échantillon de 83 patients présentant un CCR précoce a été analysé par un panel de 201 gènes sélectionnés pour leur implication connue dans l'oncogenèse. La couverture moyenne de séquençage obtenue sur ce panel est de 47,14% à la profondeur de 500X, et de 99,98% à la profondeur de 100X. Le *on-target* moyen, c'est-à-dire la proportion de séquences localisées dans la cible donc les gènes du panel 201, est de 69,70%.

3.1.1 Panel 201 gènes : détection de variants potentiellement délétères

Dans un premier temps, nous avons recherché les variations potentiellement pathogènes (classe 4) ou pathogènes (classe 5) présentes parmi ces 201 gènes. L'objectif était de détecter des altérations qui auraient été méconnues par les analyses moléculaires réalisées au préalable au titre du diagnostic chez ces patients mais également d'identifier des altérations non connues à ce jour comme responsables de formes héréditaires de CCR.

Chez ces 83 patients, nous avons détecté un total de 172 variations rares LOF ou faux sens prédits délétères selon les 3 algorithmes de prédiction de pathogénicité utilisés. Après lecture des données qualités et des fichiers BAM, 58/172 variants (dont 16 différents) se sont révélés être des artefacts. Les 114 autres variants hétérozygotes identifiés ont été détectés chez 66 patients et 48 de ces variants sont situés dans les 54 gènes impliqués dans la réparation de l'ADN inclus dans le panel de 201 gènes. Selon les recommandations de l'ACMG, parmi les 114 variations retenues pour analyse, 4 variations détectées chez trois patients différents ont été classées en probablement pathogènes (classe 4) ou pathogènes (classe 5), alors que 94 variations sont considérées de signification indéterminée (VSI : classe 3), et 16 variants sont bénins ou probablement bénins (respectivement classe 1 ou 2).

- Le premier variant identifié, de classe 4, est situé sur le gène *POLE* (NM_006231) : c.857C>T, p.(Pro286Leu). Cette variation faux sens située dans l'exon 9, situé dans le domaine exonucléasique de la protéine, est prédite délétère par les 3 algorithmes utilisés et est absente des bases de données de population et de patients. Le patient DOCC1117 présentant ce variant, est atteint d'un adénocarcinome du cæcum et du côlon droit diagnostiqué à 28 ans.

Il s'agit d'un des rares patients inclus avec une histoire familiale de cancer du côlon puisque sa mère avait 9 adénomes coliques tubuleux et a développé un cancer du rectum à 48 ans. Son grand-père maternel a développé un cancer du côlon à 58 ans et deux des frères de celui-ci ont eu respectivement un cancer du côlon à 60 ans et une colectomie sans plus de précision (Figure 15). L'analyse de la ségrégation familiale est en cours.

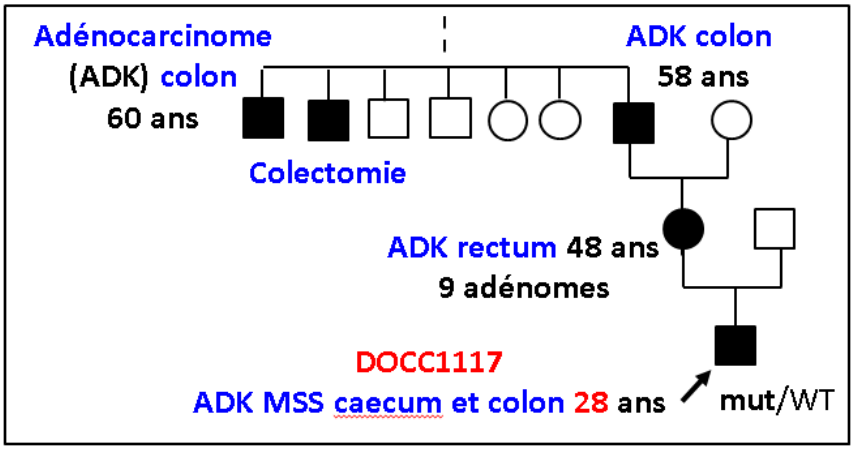


Figure 15 : Arbre généalogique du patient DOCC1117

- Le deuxième variant, de classe 5, est la variation c.7928-2A>G, p. ?, située dans le gène *ATM* (NM_000051.3). C'est une substitution située sur le site accepteur d'épissage de l'intron 53. Les algorithmes utilisés prédisent un saut de l'exon 54, qui n'est pas en phase, ce qui serait alors responsable de l'apparition d'un codon stop prématuré. Cette variation est absente des bases de données de population et est rapportée probablement pathogène dans la base de données de patients ClinVar. La patiente DOCC1070 porteuse a eu un adénocarcinome du cæcum et colon droit diagnostiqué à 26 ans.

- Le troisième variant, de classe 4, est une délétion hétérozygote de 5 nucléotides dans l'exon 18 situé dans le domaine OB1 (IPR015187) du gène *BRCA2* (NM_000059) : c.8218_8222del, p.(Leu2740Glnfs*22). Ce domaine est l'un des 5 domaines de la protéine BRCA2 capable de lier l'ADN simple brin *in vitro*. Cette délétion, absente des bases de données de population et de patients, entraine un décalage du cadre de lecture avec prédiction de l'utilisation d'un codon stop prématuré 22 acides aminés plus loin. Elle est donc prédite responsable d'une perte de fonction de la protéine. Le variant c.8218_8223delinsC, p.(Leu2740Glufs*22) est prédit être responsable du même impact sur la protéine codée que ce

variant et il est rapporté pathogène (classe 5) dans la base de donnée BRCA Share (autrefois dite UMD-BRCA2 pour *Universal Mutation Database BRCA2*). La patiente DOCC1104, porteuse de la variation, a développé un adénocarcinome du côlon gauche et sigmoïde à 24 ans, sans adénomes associés.

- Le dernier variant de classe 4 détecté, c.2590C>T, p.(Gln864*) (Tableau 5), est une variation non-sens hétérozygote responsable de l'apparition d'un codon stop prématuré dans l'exon 11 de *BRCA2,* ce qui aboutit très probablement à une perte de fonction de la protéine. Ce variant est également situé dans le domaine OB1 et absent des bases de données de population et de patients. La patiente DOCC1544 porteuse de ce variant a présenté un adénocarcinome du côlon gauche et sigmoïde à 26 ans. D'origine asiatique, elle a été ultérieurement écartée de l'étude statistique par l'analyse de stratification de population. De façon remarquable, cette patiente est également porteuse de 5 variants rares de classe 3 (Tableau 3), dont deux affectent d'autres gènes impliqués dans la réparation de l'ADN : *FANCL* (*Fanconi anemia complementation group L* ; OMIM # 614083) du système ICL et *MLH1* du système MMR. Le variant *FANCL* est une duplication frameshift de 4 pb dans le dernier exon de la protéine, à l'origine de l'ajout de 3 acides aminés terminaux par rapport à la séquence sauvage. Le variant faux sens *MLH1* est prédit délétère par les 3 algorithmes de prédiction utilisés et classé VSI par le groupe expert international InSIGHT (*International Society for Gastrointestinal Hereditary Tumours Database*). A noter que cette patiente est également porteuse d'un variant situé sur un site consensus d'épissage dans l'intron 10 de *MUTYH*, rare chez les caucasiens mais de MAF supérieure à 1% dans la population asiatique (ExAC EAS : population d'Asie de l'Est). Ce variant fréquent affecte le site consensus accepteur d'épissage de l'intron 10 du gène *MUTYH* et est prédit responsable d'une altération de l'épissage.

Tableau 5 : Détection simultanée chez la patiente DOCC1544 d'un variant probablement pathogène et de 5 VSI

ID	Gène	Ref. Seq.	ADN	Protéine	ExAC (EAS non TCGA)	ESP6500 (siv2; ALL)	1000G (2014, ALL)
DOCC1544	*BRCA2*	NM_000059	c.2590C>T	p.(Gln864*)	.	.	.
DOCC1544	*FANCL*	NM_018062	c.1096_1099dup	p.(Thr367Asnfs*13)	0.01%	0.25%	.
DOCC1544	*MLH1*	NM_000249	c.649C>T	p.(Arg217Cys)	0.38%	.	0.0599%
DOCC1544	*SDHA*	NM_004168	c.991G>A	p.(Ala331Thr)	0.18%	.	0.0798%
DOCC1544	*TSC2*	NM_000548	c.3803G>A	p.(Arg1268His)	0.18%	.	0.0599%
DOCC1544	*MUTYH*	NM_012222.2	c.925-2A>G	p.?	1,45%	.	0.300%

En marge de ces variants de classe 4 et 5, parmi les 95 VSI détectés chez les 83 patients, trois autres variations ont retenu notre attention :

- Le premier VSI d'intérêt est une variation indel en phase localisée dans l'exon 33 de *FANCD2* (*Fanconi anemia complementation group D2* ; OMIM # 227646 ; NM_033084) : c.3260indelGTAA, p.? (Tableau 6). Elle est absente des bases de données de population et de patients et est prédite responsable de la création d'un site donneur d'épissage fort par les différents algorithmes de prédiction d'épissage utilisés. La patiente présentant cette variation a eu un cancer de la jonction rectosigmoïdienne à 24 ans sans adénome associé. Cette patiente est également porteuse d'un variant faux sens *MSH2* de classe 3 prédit délétère par les 3 algorithmes, situé dans l'exon 5, inclus dans le domaine de connexion (IPR007860) et classée VSI par le groupe expert international InSIGHT (*International Society for Gastrointestinal Hereditary Tumours Database*).

Tableau 6 : Détection simultanée chez la patiente DOCC2319 de 2 VSI

ID	Gène	Ref. Seq.	ADN	Protéine	ExAC (NFE non TCGA)	ESP6500 (siv2; EA)	1000G (2014, EUR)
DOCC2319	*FANCD2*	NM_033084	c.3260indelGTAA	p.?	.	.	.
DOCC2319	*MSH2*	NM_000251	c.815C>T	p.(Ala272Val)	0.04%	0.06%	.

- Le deuxième VSI d'intérêt, identifié chez la patiente DOCC2391, est un variant qui touche le site consensus accepteur d'épissage de l'intron 5 du gène *MUTYH* (c.454-1G>A, p.?). Il est prédit altérer l'épissage en provoquant un possible saut de l'exon 5, altérant le domaine *DNA glycosylase* (IPR011257) de *MUTYH*. Cette patiente, atteinte d'un adénocarcinome du côlon gauche et sigmoïde diagnostiqué à 30 ans, porte également le 3e VSI d'intérêt, qui correspond à une variation non-sens dans le dernier exon du gène *XRCC2* (*X-ray repair cross complementing group 2*), ainsi qu'un variant faux sens du gène *IRS1* prédit délétère par les 3 algorithmes (Tableau 7). Il a été impliqué dans la FA mais est classé VSI dans la base de données ClinVar.

Tableau 7 : Détection simultanée chez la patiente DOCC239 de 3 VSI

ID	Gène	Ref. Seq.	ADN	Protéine	ExAC (NFE non TCGA)	ESP6500 (siv2; EA)	1000G (2014, EUR)
DOCC2391	*MUTYH*	NM_012222	c.454-1G>A	p.?	.	.	.
DOCC2391	*XRCC2*	NM_005431	c.643C>T	p.(Arg215*)	0,007%	0.01%	.
DOCC2391	*IRS1*	NM_005544	c.1117C>T	p.(Arg373Cys)	0.04%	0.02%	0.1%

3.1.2 Panel 201 gènes : analyse statistique par étude d'association cas-témoins

En vue de l'étude statistique, les individus *outliers* ou identifiés comme à risque de biaiser l'étude d'association par les contrôles qualité ont été exclus. Les contrôles de qualité appliqués ont permis d'identifier 7 patients trop éloignés génétiquement du reste de la population. La consultation de leurs dossiers et de la MAF ExAC de leurs variants a permis de confirmer leur divergence génétique (deux patients d'origine asiatique et cinq patients d'origine africaine). Pour permettre une adéquation entre population de patients et de témoins, ces patients ont été exclus. Par ailleurs, le calcul de l'*Identity By State* a permis d'identifier un patient en doublon, qui avait été inclus dans l'étude sous deux dénominations différentes, ainsi que 15 témoins FREX avec une contamination inter-échantillons, présentant et partageant donc plus de variants. Ces 16 individus ont donc été également exclus de l'étude. A l'issue des contrôles qualité à l'échelle individuelle, l'analyse statistique a donc porté sur 76 patients et 552 témoins. Des contrôles de la qualité des variants ont ensuite été réalisés afin d'éliminer le plus d'artefacts possibles avant les tests d'association. Nous avons pu comparer les contrôles automatiques réalisés (détaillés dans le matériel et méthode), avec notre nettoyage manuel des artefacts par lecture de BAM, portants sur les variants LOF et les faux sens prédits délétères par les trois algorithmes utilisés. Les 16 artefacts détectés manuellement ont bien été identifiés et éliminés de l'étude. Un 17e variant, manuellement classé artefact, posa question, mais le séquençage Sanger de confirmation réalisé était bien en faveur de sa présence. Cette comparaison nous a permis de valider cette méthode automatisée de nettoyage des résultats, qui pourra ainsi être utilisée à plus grande échelle lors de l'analyse d'exomes.

Chez les 76 patients inclus dans l'étude statistique, 1 345 variants ont été identifiés avant filtration. Afin de limiter la dilution d'un éventuel effet d'association par des variants neutres, nous avons choisi de nous concentrer sur les variants rares LOF ou faux sens prédits délétères par les 3 algorithmes (3/3). Un test de *Collapsing* permettant l'utilisation du test

exact de Fisher a ensuite été utilisé. Ce test repose sur l'agrégation de ces variants à l'échelle du gène et la comparaison de la proportion de patients porteurs d'au moins un variant à celle des témoins.

Après avoir agrégé les variants détectés par gène, le signal d'association avec l'OR le plus élevé et la *p-value* la plus basse concernant les variants rares LOF (Type de variant : LOF) a été détecté sur le gène *SBDS* (*ribosome maturation factor,* tableau 8). Celui concernant les variants rares LOF plus les variants faux sens prédits délétères par les 3 algorithmes (Type de variant : LOF + 3/3) a été détecté sur le gène *WT1* (*Wilms Tumor 1,* tableau 8). Cependant, les *p-values* obtenues sont respectivement de 0,026 et 0,040. Celles-ci sont donc supérieures au seuil fixé de $2{,}5.10^{-6}$, et les intervalles de confiance des OR comprennent la valeur 1. Par conséquent aucune des associations mises en évidence après avoir agrégé les variants par gène n'est plus que nominalement significative.

Tableau 8 : Signaux d'association les plus importants détectés par le Burden test à l'échelle du gène

Type de variant	Gène	n cas porteurs	n cas totaux	n témoins porteurs	n témoins totaux	OR	Intervalle de confiance	p value
LOF	*SBDS*	3	76	3	552	7,477	0,983 - 56,884	0,026
LOF + 3/3	*WT1*	2	76	1	552	14,769	0,760 - 874,540	0,040

Le variant LOF *SBDS* détecté chez les 3 patients (DOCC1317, DOCC1412 et 28611.001) est identique à celui détecté chez les 3 témoins. Il s'agit d'une transition nucléotidique située sur un site consensus donneur d'épissage en position +2 dans l'intron 3 du gène *SBDS* (NM_016038) : c.258+2T>C, p.?. Les algorithmes de prédiction d'épissage appliqués prédisent une abolition du site donneur d'épissage original. Ce variant répertorié dans la base dbSNP (rs113993993) est présent dans les bases de données de population ExAC (*Non Finish European* non TCGA), ESP6500 (European American) et 1000G (Europeans) avec une fréquence allélique respective de 0,40%, 0,37 et 0,20%, en cohérence avec la MAF observée chez les témoins dans notre échantillon (0,27%).

Les deux variants *WT1*, prédits délétères par les 3 logiciels de prédiction de pathogénicité, sont des transitions faux sens dans les exons 6 et 7, inclus dans le domaine N-

terminal (IPR000976) de *WT1* (Tableau 9). L'une a été détectée à la fois chez un patient et un témoin.

Tableau 9 : Variants faux sens du gène WT1 prédits délétères par les 3 algorithmes

ID	Gène	Ref. Seq.	ADN	Protéine	ExAC (NFE non TCGA)	ESP6500 (siv2; EA)	1000G (EUR)
DOCC1083 B00GQCK	*WT1*	NM_024426	c.1048T>C	p.(Cys350Arg)	0.06%	0.08%	.
DOCC1645	*WT1*	NM_024426	c.1126C>T	p.(Pro376Ser)	.	.	.

Afin de gagner en puissance statistique, nous avons ensuite effectué un test de charge avec agrégation des variants à l'échelle d'une liste de gènes en analysant les résultats de notre panel de 201 gènes selon deux classes : les gènes impliqués dans la réparation de l'ADN et les autres gènes. Les tests d'association ont été réalisés en agrégeant respectivement les variants issus de ces deux catégories et en comparant la proportion de patients porteurs (cas) d'au moins un variant rare dans les gènes de la réparation de l'ADN, par rapport à la proportion de témoins.

Le signal d'association avec l'OR le plus élevé et la *p-value* la plus basse concerne les variants LOF détectés dans les gènes impliqués dans la réparation de l'ADN (Tableau 10). Cependant, l'association n'est pas statistiquement significative : la *p-value* atteint 21% et l'intervalle de confiance de l'OR comprend la valeur 1. L'OR observé après avoir additionné les variants LOF et faux sens prédits délétères par les 3 algorithmes, est encore plus faible.

Tableau 10 : Burden test à l'échelle des gènes de la réparation de l'ADN du panel 201 gènes

Type de variant	Voie biologique	n cas porteurs	n cas totaux	n témoins porteurs	n témoins totaux	OR	Intervalle de confiance	p value
LOF	*DNA repair*	5	76	20	552	1,870	0,530 - 5,350	0,211
LOF + 3/3	*DNA repair*	33	76	217	552	1,180	0,710 - 1,980	0,533
LOF	Autres	5	76	29	552	1,270	0,370 - 3,470	0,590
LOF + 3/3	Autres	49	76	308	552	1,440	0,885 - 2,470	0,175

3.2 Résultats de l'analyse de WES

Le second échantillon constitué de 93 patients présentant un CCR précoce a été analysé par WES. Parmi ceux-ci, 53 patients appartiennent également au 1er échantillon analysé par le panel 201 gènes. La profondeur moyenne de séquençage sur le WES chez les patients analysés est de 168X, et 99,8% des bases sont couvertes avec une profondeur de plus de 30X. Le *on-target* moyen est de 74,1%. Chez ces 93 patients, nous avons détecté à l'échelle de l'exome entier un total de 42 758 variations LOF, dont 10 982 artefacts, et 36 171 faux sens prédits délétères selon les 3 algorithmes de prédiction de pathogénicité utilisés, dont 9 604 artefacts. Les artefacts ont été classés grâce à l'algorithme du VQSR validé par les analyses sur le panel de 201 gènes. Après filtration des variants rares par la fréquence allélique (MAF ExAC NFE < 1%), nous avons détecté 1637 LOF, dont 216 artefacts, et 10 315 faux sens rares prédits délétères selon les 3 algorithmes de prédiction de pathogénicité utilisés, dont 148 artefacts.

3.2.1 Analyse de WES par panel *in silico* de 14 gènes de rang diagnostique responsables de prédisposition héréditaire au CCR

Dans un premier temps, nous avons souhaité vérifier qu'aucune variation probablement pathogène (classe 4) ou pathogène (classe 5) située dans un gène connu de prédisposition au CCR, susceptible d'expliquer la survenue de cancer colorectal précoce chez ces patients, n'avait été omise en amont de l'inclusion dans cette étude. L'utilisation du WES permet l'analyse simultanée du nombre de gènes souhaité. Un panel *in silico* de 14 gènes responsables de syndromes de prédisposition héréditaire au CCR connus et bien caractérisés a donc été appliqué aux données de WES. Les gènes de ce panel diagnostique sont : *APC, BMPR1A, MLH1, MSH2, MSH3, MSH6, MUTYH, NTHL1, PMS2, PTEN, POLD1, POLE, SMAD4* et *STK11*. Pour ces 93 patients, au total 2 111 variations ont été identifiées, dont 10 artefacts mis en évidence par l'algorithme VQSR, préalablement testé sur le panel de 201 gènes. Après filtration sélectionnant les variants de fréquence allélique inférieure à 5%, nous avons obtenus 138 variants dont 6 variants pathogènes (classe 5), 80 VSI (classe 3), 47 variants probablement bénins (classe 2) ou bénins (classe 1), et 5 artefacts. Il n'a pas été mis en évidence de variation de classes 4 ou 5 dans les gènes responsables de prédisposition au CCR de transmission autosomique dominante.

Six variations hétérozygotes de classe 5 ont en revanche été identifiées dans les gènes *MUTYH* et *NTHL1,* impliqués dans le système de réparation de l'ADN par excision de base (BER : *Base Excision Repair*). Ces deux gènes sont responsables de prédisposition au CCR de transmission autosomique récessive. Ces variations hétérozygotes, identifiées chez des patients différents, ne peuvent donc pas expliquer à elles-seules la survenue de CCR précoce. Un second évènement non détecté, de type mosaïcisme ou variant de structure, a été activement recherché, par lecture des BAM et *variant calling* alternatif par l'outil VarScan, sans succès. De façon intéressante, les trois variants pathogènes détectés dans le gène *NTHL1* (NM_002528.5) sont identiques chez trois patients sans polypose connue. Ca variant n'a en revanche pas été détecté dans l'échantillon de population de témoins analysé. Le variant identifié est le non-sens c.268C>T, p.(Gln290*), situé à proximité du domaine DNA glycosylase (IPR011257) de la protéine. Représenté à 0,25% dans la cohorte ExAC NFE non TGCA, il est rapporté pathogène dans la base de donnée de patients ClinVar. Les patients porteurs sont le patient 17-01760 et la patiente 33578.001, qui ont déclaré un adénocarcinome d'emblée métastatique respectivement à 18 et 22 ans, et le patient DOCC1590, atteint d'adénocarcinome du rectum à 28 ans. Il faut cependant noter qu'un variant *NTHL1* différent mais également rapporté pathogène, le c.709+1G>A, p. ? a été identifié chez un des témoins. Ce variant, de MAF ExAC NFE non TGCA 0,0019%, est prédit pour abolir le site donneur d'épissage de l'intron 4, aboutissant probablement à un saut de l'exon 4.

Malgré l'absence d'autres altérations de rang diagnostique, un variant d'intérêt a été identifié dans le gène *POLE* (NM_006231). Il s'agit du variant faux sens c.902A>G, p.(Asp301Gly) porté par le patient 32666.001. Celui-ci a déclaré un adénocarcinome du rectum à 23 ans et n'a pas d'antécédents familiaux connus. Ce variant est situé dans le domaine critique exonucléasique de la protéine et prédit pathogène par les 3 algorithmes de prédictions de pathogénicité utilisés. Il est absent des bases de données de population et n'a jamais été rapporté en pathologie. En l'absence de données fonctionnelles ou somatiques en faveur ou défaveur de son implication, celui-ci est à ce jour classé VSI. En attendant les résultats de l'investigation diagnostique en cours, le patient porteur a été exclu de l'analyse statistique ultérieure.

3.2.2 Analyse de WES par panel *in silico* de 296 gènes impliqués dans la réparation de l'ADN

Afin d'explorer la piste de l'oligogénisme suggérée par la cooccurrence de variants dans des gènes impliqués dans les voies de la réparation de l'ADN observée lors de l'analyse du panel de 201 gènes, nous avons sélectionné 296 gènes d'intérêts appartenant à ces voies biologiques. L'objectif est ainsi d'évaluer de façon plus exhaustive que précédemment l'implication des voies de réparation de l'ADN dans la survenue de CCR précoce.

3.1.2.1 WES : Détection de variants potentiellement délétères

Nous avons, d'une part, recherché les variations potentiellement pathogènes (classe 4) ou pathogènes (classe 5) présentes parmi ces 296 gènes afin de détecter des altérations non connues à ce jour comme responsables de formes héréditaires de CCR. Après filtration des variants rares présents dans ces 296 gènes par la fréquence allélique (MAF ExAC NFE non TGCA < 1%), nous avons détecté 61 variations LOF, dont 34 artefacts (17 différents), et 205 faux sens rares prédits délétères selon les trois algorithmes de prédiction de pathogénicité utilisés, dont 11 artefacts (5 différents).

Selon les recommandations de l'ACMG, parmi les 221 variations retenues pour interprétation, 5 variations détectées chez 5 patients différents ont été classées en classe 4 et classe 5. Le variant *ATM* de classe 5 et les deux variants *BRCA2* de classe 4, préalablement détectés par le panel 201 gènes, ont à nouveau été mis en évidence.

- Le quatrième variant, de classe 5 (Tableau 11), a été identifié dans l'exon 13 du gène *FANCA* (*Fanconi anemia, complementation group A* ; OMIM # 227650 ; NM_000135.2). La variation tronquante c.1115_1118del, p.(Val372Alafs*42), est responsable d'un décalage du cadre de lecture et de l'apparition d'un codon stop prématuré. Elle est présente à 0,0092% dans la cohorte ExAC NFE non TGCA, et est rapportée pathogène dans la FA, de transmission récessive, dans la base de données ClinVar. De plus le patient 33764.001, qui porte cette variation, est le père d'un enfant qui est atteint d'Anémie de Fanconi. Le patient 33764.001 a développé un adénocarcinome du colon à 27 ans.

- Le cinquième variant de classe 5 (Tableau 11) a été identifié dans l'intron 28 du gène *FANCD2* (NM_033084.3). La variation d'épissage c.2715+1G>A, p. ?, est prédite abolir le

site donneur, résultant probablement en un saut de l'exon 28 qui n'est pas en phase et aboutirait donc à un décalage du cadre de lecture. Bien que présente à 0,03% dans la cohorte ExAC NFE non TGCA, elle a déjà été décrite dans la FA. Le patient 29469.001 a développé un adénocarcinome du colon détecté à 32 ans. Son oncle a fait un cancer cutané sans précision et sa tante un cancer de l'utérus.

Les variations pathogènes mises en évidence par le panel 201 gènes et le panel *in silico* des 296 gènes de la réparation, ainsi que la bibliographie nous ont incités à nous intéresser particulièrement aux gènes de la voie ICL (22 gènes/296). Les gènes de la réparation de l'ADN impliqués en pathologie le sont généralement par un mécanisme de perte de fonction. Nous avons donc décidé de nous intéresser plus étroitement aux variations LOF détectés dans les gènes de cette voie. Sept variants LOF ont été identifiés chez 7 patients différents (Tableau 11).

Tableau 11 : Variations LOF de la voie ICL

ID	Gène	Ref. Seq.	ADN	Protéine	ExAC (NFE non TCGA)	ESP6500 (siv2; EA)	Classe
DOCC1104	***BRCA2***	**NM_000059.3**	**c.8218_8222del**	**p.Leu2740fs**	**absent**	**absent**	**4**
DOCC1544	***BRCA2***	**NM_000059.3**	**c.2590C>T**	**p.Gln864***	**absent**	**absent**	**4**
33764-001	***FANCA***	**NM_000135.2**	**c.1115_1118del**	**p.Val372fs**	**0,0092%**	**absent**	**5**
29469-001	***FANCD2***	**NM_033084.3**	**c.2715+1G>A**	**p. ?**	**0,03%**	**0,023%**	**5**
DOCC1713	*FAAP20*	NM_001282671.1	c.94C>T	p.Arg32*	absent	absent	3
32047-001	*FAN1*	NM_014967.4	c.1121dup	p.Tyr374fs	0,0037%	absent	3
6832-003	*FANCI*	NM_001113378.1	c.2457-2A>G	p. ?	absent	absent	3

Grâce à l'algorithme CANOES, nous avons également recherché la présence de variations de nombre de copie (CNV) situées dans ces 296 gènes. Nous avons détecté 14 remaniements de type délétion ou duplication, dont 8 remaniements récurrents, situés dans des régions répétées et associés à de faibles scores de qualité, qui sont probablement des polymorphismes ou des artefacts. Sur les 6 restants, nous avons identifié deux remaniements d'intérêt ayant de bons scores de qualité (Tableau 12). Ils sont tous deux situés dans des gènes impliqués dans la voie ICL et ne sont pas polymorphiques. Leur existence a été confirmée par PCR digitale. Le patient 25166.001 porte une délétion partielle de 2,86 kb du gène *UBE2T (Ubiquitin-conjugating enzyme E2T)*, qui s'étend de l'exon 2, correspondant au premier exon

codant du gène, à l'exon 6. Ce patient a déclaré un adénocarcinome du colon à 21 ans. Son père était porteur de 2 polypes et a fait un cancer de la prostate à 53 ans. Le second patient, DOCC1643, porte une duplication de 226,67 kb, incluant entièrement les gènes *FAN1* et *HERC2P10*, et partiellement les gènes *MTMR10* et *LOC100288637*. Ce patient a déclaré un adénocarcinome du colon transverse à 24 ans et porte un adénome colique tubuleux inférieur à 10mm.

Tableau 12 : CNV d'intérêt

ID	Gène	Remaniement	Taille (kb)	Coordonnées
25166.001	*UBE2T*	Délétion	2,86	Chr1 : 202302171-202305031
DOCC1643	*HERC2P10, FAN1, MTMR10, LOC100288637*	Duplication	226,67	Chr15 : 31042462-31269133

Par ailleurs, aucun variant supplémentaire n'a été identifié parmi les 40 nouveaux patients analysés par WES dans le gène *SBDS,* sélectionné comme gène candidat après l'étude statistique sur le panel de 201 gènes. En revanche, le variant faux sens *WT1* c.1048T>C, p.(Cys350Arg) identifié chez un patient et un témoin par le panel 201 gènes a également été mis en évidence chez un des 40 nouveaux patients analysés. Le patient porteur 2827.001 a été atteint d'un adénocarcinome du colon à 26 ans et son frère d'un cancer de l'estomac à un âge non spécifié.

3.1.2.2 WES : Analyse statistique par étude d'association cas-témoins

D'autre part, afin d'objectiver une potentielle association entre l'occurrence de variants rares dans les gènes de la réparation de l'ADN et la survenue d'un CCR précoce, des tests d'associations cas-témoins ont été réalisés.

Une étude de stratification de population, nécessaire à toute étude génétique statistique robuste, a préalablement été réalisée sur les données WES. Les contrôles de qualité des individus ont permis d'identifier 14 patients *outliers*, c'est-à-dire génétiquement éloignés du reste de la population. Pour permettre une adéquation entre population de patients et de témoins et ainsi limiter les biais, ces patients ont été exclus de l'étude. Le patient 32666.001, porteur d'un variant potentiellement de rang diagnostique, a également été exclu de l'étude statistique. Le calcul de l'*Identity By State* n'a pas identifié d'individus apparentés dans

l'échantillon de patients. Les 15 témoins FREX inter-contaminés préalablement identifiés ont été éliminés de l'étude. A l'issue des contrôles qualité à l'échelle individuelle, l'analyse statistique a donc porté sur 78/93 patients et 552/567 témoins. Des contrôles qualité des variants ont ensuite été réalisés afin d'éliminer le plus d'artefacts possibles avant les tests d'association.

Chez les 78 patients et les 552 témoins inclus dans l'étude statistique, 4 999 variants rares (de MAF inférieure à 1%) différents ont été identifiés et 3 316 variants ont été conservés après étude des contrôles qualités de ces variants. Plusieurs tests statistiques ont été réalisés. Ils reposent tous sur l'agrégation des variants rares par voie biologique. Afin de limiter la dilution d'un éventuel effet d'association par des variants neutres, nous avons choisi de nous concentrer sur les variants rares LOF et sur les variants rares faux sens prédits délétères par les 3 algorithmes utilisés (3/3). Les tests ont également été réalisés sur les variants synonymes (SYN), supposés neutres, afin de valider l'absence d'un éventuel biais technique ou d'échantillonnage entre les cas et les témoins. Ces tests ont été effectués d'une part en agrégeant les variants situés dans des gènes sélectionnés pour leur implication dans l'ensemble des voies de réparation de l'ADN (*DNA repair*) et d'autre part en agrégeant ceux des gènes de la voie ICL uniquement. Afin d'évaluer le potentiel impact de la rareté du variant, nous avons choisi d'analyser en premier lieu tous les variants rares, soit de MAF inférieure ou égale à 1%, puis de restreindre l'analyse aux variants les plus rares, avec un seuil fixé à une MAF inférieure ou égale à 1/1000, autrement dit aux singletons observés une seule fois sur l'ensemble de notre échantillon.

Le premier test réalisé est la comparaison de la proportion de patients porteurs (cas) d'au moins un variant rare dans les gènes de la réparation de l'ADN ou dans la voie ICL, par rapport à la proportion de témoins (Tableau 13). Il s'agit du même test que celui réalisé sur les données obtenues par le panel 201 gènes. Il n'y a pas de résultats significatifs à l'échelle de l'exome. Concernant les variants ≤ 1% des gènes de la réparation de l'ADN, les synonymes sont ceux avec l'OR et la p-value les plus significatifs. Les variants LOF et LOF + 3/3 ont un effet du même ordre de grandeur, relativement faible (OR inférieur à 2). Lorsque l'on se concentre sur les variants de MAF ≤ 1‰, leur OR augmente un peu et la p-value diminue, surtout pour les variants LOF +3/3. Pour les variants synonymes, l'OR et la p-value diminuent. Concernant les variants situés dans les gènes impliqués dans l'ICL, les variants LOF sont ceux avec l'OR et la p-value les plus significatifs. Ces 6 variants LOF correspondent aux 7 variants détaillés dans le sous chapitre précédent, avec exclusion du variant *BRCA2* porté par la patiente *outlier* DOCC1544. La délétion partielle du gène *UBE2T* ne fait pas non plus partie de ces LOF, puisque le patient porteur 25166.001 est également un *outlier*. La restriction des variants rares ICL analysés à ceux de MAF ≤ 1‰ s'accompagne également d'une hausse de l'OR et d'une réduction de la p-value, sauf pour les synonymes. Globalement l'effet le plus important est celui des LOF de MAF ≤ 1‰, pour lesquels l'OR est de 3,7 et la p-value de 1,6%, et l'effet des variants situés dans les gènes impliqués dans l'ICL est plus important que celui des variants impliqués dans l'ensemble des voies de réparation de l'ADN.

Tableau 13 : Comparaison des proportions de porteurs

Type de variant	MAF	Voie biologique	n cas porteurs	n cas totaux	n témoins porteurs	n témoins totaux	OR	Intervalle de confiance	p value
LOF	≤ 1%	*DNA repair*	16	78	88	552	1,360	0,699 – 2,522	0,328
LOF + 3/3	≤ 1%	*DNA repair*	70	78	468	552	1,569	0,719 – 3,917	0,305
SYN	≤ 1%	*DNA repair*	76	78	509	552	3,206	0,806 – 27,873	0,103
LOF	**≤ 1‰**	*DNA repair*	10	78	50	552	1,475	0,637 – 3,122	0,302
LOF + 3/3	**≤ 1‰**	*DNA repair*	53	78	307	552	1,690	0,999 – 2,926	0,050
SYN	**≤ 1‰**	*DNA repair*	58	78	334	552	1,891	1,084 – 3,417	0,018
LOF	≤ 1%	ICL	6	78	15	552	2,976	0,916 – 8,443	0,035
LOF + 3/3	≤ 1%	ICL	20	78	117	552	1,282	0,701 – 2,266	0,38
SYN	≤ 1%	ICL	26	78	127	552	1,672	0,961 – 2,855	0,05
LOF	**≤ 1‰**	**ICL**	**6**	**78**	**12**	**552**	**3,738**	**1,116 – 11,167**	**0,016**
LOF + 3/3	**≤ 1‰**	ICL	11	78	48	552	1,722	0,768 – 3,573	0,143
SYN	**≤ 1‰**	ICL	11	78	57	552	1,424	0,641 – 2,921	0,329

Le deuxième test réalisé est la comparaison de la charge en variants rares par *Burden test* de type seqMeta entre les patients porteurs (cas) et les témoins porteurs (Tableau 14). Contrairement au précédent, ce test prend en compte le nombre de variants portés par chaque individu. Construit sur une régression logistique, ce test présente l'avantage de permettre un ajustement sur les axes d'ACP pour prendre en compte la stratification de population. Mais contrairement au test de Fisher précédent, il repose sur une approximation Gaussienne. Les résultats sont proches de ceux de la comparaison des proportions de porteurs, et ne sont pas significatifs à l'échelle de l'exome. Cependant, on peut noter que l'ajustement sur les axes d'ACP permet de réduire l'effet apparent des variants synonymes, qui se dissout alors dans la structure de population, alors que l'effet des variant LOF se trouve lui exacerbé par cet ajustement. Au total, l'effet le plus important est à nouveau celui des LOF de MAF ≤ 1‰, pour lesquels l'OR est de 5,88 (avec un IC95% de 1,474-23,455) et une p-value de 1,2%. Les OR et les p-values des variants situés dans les gènes impliqués dans l'ICL sont à nouveau plus importants que ceux des variants impliqués dans l'ensemble des voies de réparation de l'ADN.

Tableau 14 : Burden test seqMeta comparant les charges en variants

Type de variant	MAF	Voie biologique	n allèles cas	n allèles tot cas	n allèles témoins	n allèles tot témoins	OR	Intervalle de confiance	p value
LOF	≤ 1%	*DNA repair*	19	2 x 78	96	2 x 552	1,412	0,813 – 2,453	0,221
LOF + 3/3	≤ 1%	*DNA repair*	162	156	1027	1104	1,112	0,934 – 1,323	0,234
SYN	≤ 1%	*DNA repair*	268	156	1564	1104	1,176	1,034 – 1,337	0,014
LOF	**≤ 1‰**	*DNA repair*	12	156	52	1104	1,681	0,807 – 3,501	0,165
LOF + 3/3	**≤ 1‰**	*DNA repair*	85	156	464	1104	1,270	0,997 – 1,618	0,053
SYN	**≤ 1‰**	*DNA repair*	91	156	512	1104	1,256	0,979 – 1,610	0,073
LOF	≤ 1%	ICL	6	156	15	1104	4,200	1,142 – 15,444	0,030
LOF + 3/3	≤ 1%	ICL	23	156	139	1104	1,119	0,718 – 1,744	0,620
SYN	≤ 1%	ICL	31	156	156	1104	1,342	0,902 – 1,998	0,147
LOF	**≤ 1‰**	**ICL**	**6**	**156**	**12**	**1104**	**5,88**	**1,474 – 23,455**	**0,012**
LOF + 3/3	**≤ 1‰**	ICL	11	156	50	1104	1,554	0,728 – 3,314	0,254
SYN	**≤ 1‰**	ICL	13	156	61	1104	1,518	0,780 – 2,955	0,219

Le troisième test réalisé est un SKAT test analysant la variance des OR de chacun des variants agrégés. Contrairement au test précédent, ce test prend donc en compte la cooccurrence éventuelle de variants délétères et protecteurs dans les mêmes gènes ou les mêmes voies étudiées, qui auraient alors un effet de sens opposés sur l'OR (Tableau 15). Ce test est également ajusté selon l'axe d'ACP pour prendre en compte la stratification de population. Les résultats ne sont pas significatifs à l'échelle de l'exome, mais les p-values obtenues sont globalement plus significatives que celles des autres tests. Concernant les variants ≤ 1% des gènes de la réparation de l'ADN, l'effet le plus important est celui des LOF, ce qui n'est plus le cas lors de la restriction des variants rares analysés à ceux de MAF ≤ 1‰. Concernant les variants situés dans les gènes impliqués dans l'ICL, les variants LOF sont de nouveau ceux avec la *p-value* la plus significative, et leur effet est renforcé lors de la restriction de l'analyse aux variants de MAF ≤ 1‰. Les *p-values* associées aux variants situés dans les gènes impliqués dans l'ICL sont à nouveau plus significatives que celles associées aux variants impliqués dans l'ensemble des voies de réparation de l'ADN.

Tableau 15 : SKAT test

Type de variant	MAF	Voie biologique	p value
LOF	≤ 1%	*DNA repair*	0,021
LOF + 3/3	≤ 1%	*DNA repair*	0,245
SYN	≤ 1%	*DNA repair*	0,181
LOF	**≤ 1‰**	*DNA repair*	0,025
LOF + 3/3	**≤ 1‰**	*DNA repair*	0,020
SYN	**≤ 1‰**	*DNA repair*	0,024
LOF	≤ 1%	ICL	0,005
LOF + 3/3	≤ 1%	ICL	0,156
SYN	≤ 1%	ICL	0,082
LOF	**≤ 1‰**	**ICL**	**0,002**
LOF + 3/3	**≤ 1‰**	ICL	0,056
SYN	**≤ 1‰**	ICL	0,056

Le quatrième test réalisé est un test de Fisher d'enrichissement en variants dans les voies de réparation de l'ADN (IN) par rapport au reste de l'exome (OUT ; Tableau 16). Il permet de comparer la proportion de variants dans ces voies à la proportion de variants totaux dans l'exome, chez les cas et les témoins. Les résultats ne sont pas significatifs à l'échelle de l'exome. Concernant les variants ≤ 1% des gènes de la réparation de l'ADN, l'OR le plus important est celui des variants LOF, cependant avec un OR pourtant très faible, les synonymes sont ceux avec la *p-value* la plus significative, résultat du grand nombre de variants synonymes observés. Lorsque l'on se concentre sur les variants de MAF ≤ 1‰, l'OR et la p-value des variants LOF et LOF +3/3 s'améliorent. C'est l'inverse pour les variants synonymes. Concernant les variants situés dans les gènes impliqués dans l'ICL, les variants LOF sont de nouveau ceux avec l'OR et la *p-value* les plus significatifs. La restriction des variants rares ICL analysés à ceux de MAF ≤ 1‰ s'accompagne d'une hausse globale de l'OR et d'une réduction de la *p-value*, sauf pour les synonymes, ce qui suggère à nouveau un problème de stratification de population au niveau des variants synonymes. Au total, l'effet le plus important est à nouveau celui des LOF de MAF ≤ 1‰, pour lesquels l'OR est de 2,88 (avec un IC95% de 0,882 – 8,332) et une *p-value* de 4%. Les OR et les p-values des variants situés dans les gènes impliqués dans l'ICL sont à nouveau plus importants que ceux des variants impliqués dans l'ensemble des voies de réparation de l'ADN.

Tableau 16 : Test d'enrichissement des voies de la réparation de l'ADN

Type de variant	MAF	Voie biologique	n allèles cas IN	n allèles cas OUT	n allèles témoins - IN	n allèles témoins - OUT	OR	Intervalle de confiance	p value
LOF	≤ 1%	*DNA repair*	19	988	96	6003	1,202	0,691 – 1,992	0,499
LOF + 3/3	≤ 1%	*DNA repair*	162	7827	1027	51716	1,042	0,876 – 1,233	0,634
SYN	≤ 1%	*DNA repair*	268	14750	1564	96885	1,126	0,983 – 1,283	0,076
LOF	**≤ 1‰**	*DNA repair*	12	501	52	2880	1,326	0,640 – 2,540	0,375
LOF + 3/3	**≤ 1‰**	*DNA repair*	85	3776	464	23310	1,131	0,884 – 1,432	0,291
SYN	**≤ 1‰**	*DNA repair*	91	5633	512	32980	1,041	0,822 – 1,305	0,727
LOF	≤ 1%	ICL	6	1001	15	6084	2,431	0,771 – 6,649	0,106
LOF + 3/3	≤ 1%	ICL	23	7963	139	52600	1,093	0,670 – 1,709	0,643
SYN	≤ 1%	ICL	31	14987	156	98277	1,303	0,856 – 1,926	0,194
LOF	**≤ 1‰**	**ICL**	**6**	**507**	**12**	**2920**	**2,88**	**0,882 – 8,332**	**0,040**
LOF + 3/3	**≤ 1‰**	ICL	11	3850	50	23724	1,356	0,636 – 2,641	0,354
SYN	**≤ 1‰**	ICL	13	5711	61	33425	1,247	0,628 – 2,297	0,508

Le dernier test réalisé est la comparaison de la proportion de patients (cas) porteurs d'une double occurrence dans les gènes de la réparation de l'ADN, c'est-à-dire d'au moins deux variants rares, par rapport à la proportion de témoins porteurs d'une double occurrence (Tableau 17). Etant donné le peu de doubles occurrences dans la voie de l'ICL, les résultats ne sont pas présentés. Il n'y a pas de résultats significatifs à l'échelle de l'exome. Concernant les variants des gènes de la réparation de l'ADN, l'OR le plus important est celui des variants LOF, et la p-value la plus significative est celle des synonymes. En réduisant les variants analysés à ceux de MAF ≤ 1‰, l'OR augmente, sauf pour les synonymes, et la p-value s'améliore pour tous. Concernant les variants de la voie ICL, bien que l'OR soit beaucoup plus important que pour les autres tests, le petit nombre de cooccurrences rend les résultats des variants LOF peu stables, ce qui est illustré par un intervalle de confiance de l'OR très étendu. Etant donné le peu de doubles occurrences de MAF inférieure à 1/1000 dans la voie de l'ICL, les résultats ne sont pas présentés.

Tableau 17 : Comparaison des proportions de porteurs de doubles occurrences dans les voies de la réparation de l'ADN

Type de variant	MAF	Voie biologique	n cas double porteurs	n cas totaux	n témoins double porteurs	n témoins totaux	OR	Intervalle de confiance	p value
LOF	≤ 1%	*DNA repair*	3	78	8	552	2,714	0,454 – 11,623	0,145
LOF + 3/3	≤ 1%	*DNA repair*	49	78	299	552	1,429	0,856 – 2,421	0,181
SYN	≤ 1%	*DNA repair*	66	78	418	552	1,761	0,909 – 3,693	0,087
LOF	**≤ 1‰**	*DNA repair*	2	78	2	552	7,196	0,514 – 100,641	0,077
LOF + 3/3	**≤ 1‰**	*DNA repair*	23	78	113	552	1,623	0,911 – 2,821	0,078
SYN	**≤ 1‰**	*DNA repair*	26	78	131	552	1,606	0,923 – 2,739	0,070
LOF	≤ 1%	ICL	0	78	0	552	/	/	/
LOF + 3/3	≤ 1%	ICL	3	78	20	552	1,064	0,197 – 3,716	1
SYN	≤ 1%	ICL	5	78	23	552	1,574	0,453 – 4,412	0,375

4. DISCUSSION

L'objectif de ce travail était de caractériser les bases génétiques de la prédisposition au cancer colorectal en étudiant par NGS un échantillonnage de patients avec un phénotype extrême caractérisé par la survenue précoce de CCR, inférieure ou égale à 32 ans, sans mutation détectable avec les techniques usuelles utilisées en routine diagnostique. Une première phase a consisté par l'analyse d'un panel restreint de 201 gènes impliqués dans le cancer chez 83 patients. Puis ce travail a été étendu par l'application d'un panel *in silico* de 296 gènes impliqués dans la réparation de l'ADN aux données d'exome de 93 patients. Nos résultats, certes préliminaires compte-tenu de la taille de l'échantillon analysé, illustrent néanmoins le sauvetage de diagnostic moléculaire, posent la question du déterminisme oligogénique du CCR ainsi que de l'implication de la voie ICL dans la genèse du cancer colorectal. Cette étude nous a également permis de structurer une approche statistique rigoureuse pour rechercher un enrichissement de variants rares chez les patients atteints de cancers précoces.

L'analyse des gènes impliqués dans les syndromes connus de prédisposition génétique au CCR nous a permis de faire un unique rattrapage diagnostique. Cela illustre la qualité du recrutement multicentrique de cet échantillonnage de patients. La variation de rang diagnostique détectée est le faux sens de classe 4 c.857C>T, p.(Pro286His) situé dans l'exon 9 du gène *POLE*. Lors de l'inclusion du patient dans l'étude, ce gène de découverte récente n'était pas analysé au titre du diagnostic moléculaire en routine. Le patient porteur a présenté un cancer du côlon droit à 28 ans et a des antécédents familiaux compatibles avec une transmission autosomique dominante. L'exon 9 dans lequel se situe cette variation fait partie du domaine exonucléasique de correction d'épreuve de la polymérase ε, où ont été détectés les variants des gènes *POLE* et *POLD1* impliqués en pathologie (Palles *et al.*, 2013). Le variant que nous avons détecté est localisé dans la partie du domaine exonucléasique spécifiquement responsable de la liaison de la polymérase à l'ADN, de même que le variant pathogène constitutionnel analogue *POLD1* p.(Pro327Leu), connu comme étant responsable de polypose (Palles *et al.,* 2013; Spier *et al.*, 2015). Nous sommes confiants pour conclure que cette mutation est à l'origine de la prédisposition héréditaire au cancer colorectal dans la

famille de ce patient et pour l'utiliser afin de codifier la prise en charge des apparentés dans cette famille. Les analyses de ségrégation sont en cours.

Un autre variant intéressant, c.902A>G, p.(Asp301Gly), a été identifié dans le même domaine exonucléasique de ce gène chez un patient avec un adénocarcinome du rectum à 23 ans sans antécédents familiaux connus. La responsabilité dans le CCR des gènes *POLE* et *POLD1* ayant été récemment mise en évidence, nous avons encore peu de recul sur la causalité d'un nombre important de leurs variations. Contrairement à la majorité des gènes intervenant dans la réparation de l'ADN impliqués en pathologie, les variations pathogènes décrites ne sont pas des pertes de fonction totales de la protéine, de type LOF, mais des altérations d'un domaine particulier, le domaine exonucléasique, qui corrige les erreurs de réplication. L'abolition de la fonction de correction augmente le taux de mutations non réparées lors de la réplication, favorisant la survenue du cancer (Briggs et Tomlinson, 2013; Palles *et al.*, 2013). Des tests fonctionnels ou la recherche du caractère ultramuté de la tumeur par l'analyse de sa charge mutationnelle, encore en développement, devront être réalisés chez ce patient, afin de permettre d'étayer ou non la pathogénicité du variant de signification inconnue détecté.

Lors de cette analyse, nous avons identifié le variant *NTHL1* connu pathogène c.268C>T, p.(Gln290*) à l'état hétérozygote chez trois patients différents et, très étonnamment, chez aucun témoin. Ce variant pathogène, qui à l'état homozygote ou hétérozygote composite prédispose à la polypose colique et au CCR, a été identifié dans la bibliographie au sein de 7 familles différentes. Tous les patients avaient entre 8 et 51 polypes adénomateux, et un âge de survenue du CCR entre 40 et 67 ans (Rivera *et al*., 2015; Weren *et al*., 2015; Chubb *et a*l., 2016; Belhadj *et al.*, 2017). Un autre variant pathogène, identifié chez un témoin, a été impliqué à l'état hétérozygote composite en association avec le variant récurrent c.268C>T, p.(Gln290*) chez une patiente avec des tumeurs multiples en rafale à type entre autre de CCR à 41 ans, de *nevi* à 42 et 55 ans, de cancer de la vessie à 47 ans, de méningiomes multiples à 47 ans et de cancer de la tête et du cou à 55 ans (Rivera *et al.*, 2015). L'hypothèse d'une prédisposition au CCR due à la variation récurrente du gène *NTHL1* à l'état hétérozygote est peu probable étant donné sa fréquence allélique dans la population générale. Une étude a mis en évidence six occurrences à l'état hétérozygote dans une population de 2 140 patients atteints de CCR et, dans les mêmes proportions, 6 occurrences parmi les 2 176 témoins analysés (Smith *et al.,* 2013). Un second évènement en *trans* a donc été recherché par lecture directe des fichiers BAM, filtration et lecture des *reads* chimériques,

et par recherche de variation en mosaïque dans le sang par *variant calling* alternatif, sans succès. Nous ne pouvons pas écarter des évènements introniques ou situés dans des séquences régulatrices, et une étude de l'ARN pourrait être menée pour avancer sur cette hypothèse. Il serait également intéressant de rechercher une perte d'hétérozygotie ou un deuxième évènement dans la tumeur. Il est possible que l'éventuel second événement soit survenu en mosaïque et il serait informatif d'analyser le tissu sain adjacent à la tumeur.

Avant de réaliser l'analyse statistique cas-témoins, nous avons eu la surprise d'identifier, après filtration, des variants rares hétérozygotes de classe 4 dans le domaine OB1 du gène *BRCA2* chez deux patientes (DOCC1104 et DOCC1544, tableau 3) ayant présenté des CCR sans antécédents personnels et familiaux connus évocateurs d'une prédisposition héréditaire au cancer du sein et de l'ovaire. La perte de fonction d'un allèle du gène *BRCA2* est responsable du syndrome sein-ovaire associé à une prédisposition au cancer du sein et de l'ovaire avec une forte pénétrance chez la femme (OMIM # 612555 ; Petrucelli *et al*. 2010) alors que l'inactivation biallélique de *BRCA2* est, elle, impliquée dans l'Anémie de Fanconi (*Fanconi anemia complementation group D1 ;* OMIM # 605724), maladie de transmission autosomique récessive associant pancytopénie progressive, aplasie médullaire, malformations congénitales et prédisposition aux hémopathies et aux tumeurs solides. Le gène *BRCA2* est impliqué dans deux voies chevauchantes de réparation de l'ADN : l'ICL et la recombinaison homologue (HR). Ces deux voies de réparation sont particulièrement interdépendantes puisque le système HR est nécessaire à la correction des liaisons inter brins initiée par les gènes de la voie du FA (Deans et West, 2011).

La prévalence des mutations *BRCA1/2* dans la population générale est estimée entre 1/400 et 1/800 (Petrucelli *et al.* 2010). D'ailleurs, deux variants LOF *BRCA2*, c.7558C>T et c.9382C>T, qui sont des variants non-sens dans les exons 15 et 25, ont également été mis en évidence chez nos 552 témoins français, correspondant à une fréquence de 1/276. Avec deux patientes sur 83, la fréquence dans notre échantillon de patients est d'environ 1/40. Les études concernant l'implication de variants hétérozygotes de *BRCA2* dans la prédisposition au CCR sont contradictoires. En effet, il n'existe pas à ce jour d'argument formel démontrant de façon consensuelle une augmentation du risque de CCR chez les porteurs et porteuses de mutation du gène *BRCA2* (Sopik *et al.,* 2015). En particulier, une étude prospective, incluant 7 015 patientes dont 5 481 altérations pathogènes *BRCA1* et 1 474 *BRCA2,* a d'ailleurs conclut à un risque de cancer du côlon 5 fois plus important chez les patientes *BRCA1* de moins de 50 ans

(2 638 patientes), et aucun sur-risque pour les patientes *BRCA2* (Phelan *et al.,* 2014), bien que dans cette étude une puissance insuffisante ne puisse être exclue, compte-tenu du nombre limité de patientes *BRCA2* de moins de 50 ans (554 patientes). Une étude focalisée sur les CCR considérés comme précoces, diagnostiqués jusqu'à 55 ans, conclue, suite à l'analyse par WES d'une cohorte de 847 patients et 1 609 témoins, à l'absence d'enrichissement de variants pathogènes dans les gènes *BRCA2* et *TP53* (Dobbins *et al*., 2016). Ainsi, les variations de classe 4 de *BRCA2* détectées dans notre série de patients atteints de CCR précoces pourraient être considérées comme des découvertes secondaires sans rapport avec le phénotype. Cependant, l'association entre variants LOF *BRCA2* et CCR précoce a déjà été suggérée par plusieurs études. Deux variations LOF *BRCA2* associées à un CCR précoce ont été rapportées dans une cohorte de 274 CCR familiaux (Hansen *et al.*, 2017). Quinze variants LOF *BRCA1/2* ont également été décrits dans une cohorte de 1 260 patients évocateurs de syndrome de Lynch, dont 60% des porteurs avec antécédent personnel et 47% avec une histoire familiale de CCR (Yurgelun *et al.*, 2015). Les auteurs posent l'hypothèse de l'existence d'un sous-ensemble de patients *BRCA1/2* qui pourraient avoir un phénotype atypique mimant le syndrome HNPCC. Une autre étude implique un variant LOF *BRCA2* dans une famille de CCR MSS, sur 48 testées, sur les arguments de la ségrégation familiale et de la LOH tumorale, et décrit trois variants faux sens comme potentiels facteurs de risque de CCR (Garre *et al.*, 2015). Enfin, une publication a rapporté chez deux sœurs avec des CCR MSS précoces, diagnostiqués respectivement à 31 ans et 35 ans pour l'une, et 27 ans pour la seconde, des mutations LOF hétérozygotes composites de *BRCA2,* mais sans phénotype clinique évident de FA (Degrolard-Courcet *et al*. 2014). Les analyses cytogénétiques effectuées à postériori ont cependant montré une hypersensibilité aux agents pontant de l'ADN, confirmant le diagnostic biologique de FA. A la lumière de cette publication, nous avons recherché un 2e évènement *BRCA2* chez nos deux patientes. Un 2e variant situé dans l'exon 18 de *BRCA2* et absent des bases de données de population et de patients a été identifié chez la patiente DOCC1104. Cependant celui-ci est faux sens, et prédit possiblement délétère seulement par l'algorithme Polyphen HVAR, il est donc difficile de l'impliquer sans études fonctionnelles. Cette patiente n'avait pas de phénotype de FA, cependant la recherche d'une éventuelle hypersensibilité aux agents pontant de l'ADN n'a pas été réalisée.

De façon très intéressante, l'identification chez la seconde patiente présentant une mutation classe 4 de *BRCA2* (DOCC 1544) d'autres mutations dans les gènes de la réparation de l'ADN (Tableau 5) pose la question de la contribution de l'oligogénisme dans le

déterminisme du CCR précoce. Cette patiente cumule un variant de classe 4 dans le gène *BRCA2*, un variant de classe 3 dans le gène *FANCL* impliqué dans le système ICL, un variant de classe 3 dans le gène *MLH1*, impliqué dans le système MMR, et un variant de classe 3 dans le gène *MUTYH*, impliqué dans le système BER. Le variant *FANCL* est potentiellement pathogène et possiblement de classe 4 puisqu'une variation voisine, responsable de la même variation protéique p.Thr367Asnfs*13, a été rapportée pathogène dans la FA, avec des études fonctionnelles en faveur d'un variant hypomorphe (Ali *et al.*, 2009). Le variant *MLH1* est lui classé VSI par le groupe expert international InSIGHT (*International Society for Gastrointestinal Hereditary Tumours Database)* et rapporté neutre dans la base de données de patients UMD-MLH1 (*Universal Mutation Database*). Des études ARN ont montré que le variant d'épissage *MUTYH* était responsable de la troncature de la protéine avec altération de son transport au noyau (Tao *et al.*, 2004). Les auteurs ont également identifié ce dernier variant à l'état homozygote chez un patient japonais et ont évoqué sa possible implication dans le cancer gastrique de ce patient. Cependant ce variant n'est pas rare dans la population asiatique et l'épissage de *MUTYH* est soumis à différents épissages alternatifs, rendant son épissage particulièrement complexe à analyser (Plotz *et al.*, 2012).

Chez 5 autres patients des variants LOF dans d'autres gènes responsables de FA impliqués dans le système ICL/HR ont été également détectés. Le variant non-sens *XRCC2* c.643C>T, le variant frameshift *FANCA* c.1115_1118delTTGG, et le variant d'épissage *FANCD2* c.2715+1G>A ont déjà été rapportés comme responsables de FA. Le variant *XRCC2* détecté a été rapporté pathogène à l'état homozygote chez le premier patient décrit atteint d'une forme atypique de FA (*Fanconi anemia complementation group U* ; OMIM # 617247) due à des mutations bialléliques du gène *XRCC2* (Shamseldin *et al.* 2012 ; Park *et al*. 2016). La responsabilité de ce gène dans cette pathologie a ensuite été confirmée grâce à des modèles murins et cellulaires. Par ailleurs, certains de ses polymorphismes ont été associés à une augmentation du risque de CCR (Krupa *et al*., 2011 ; Li *et al.,* 2014). Le variant *FANCA* a été rapporté de façon récurrente à l'état hétérozygote composite chez plusieurs patients atteints de FA et l'individu porteur est le père d'un enfant atteint de FA (Castella *et al*., 2011). Le variant *FANCD2* a lui été décrit responsable de l'abolition du site donneur d'épissage conduisant à l'utilisation d'un site cryptique donneur responsable de la rétention de 27 pb de l'intron 28 dans l'ARN messager. Il s'agit donc d'une variation hypomorphe qui a cependant été impliquée plusieurs fois à l'état hétérozygote composite dans la FA (Kalb *et al*., 2007). Le variant d'épissage *FANCI* c.2457-2A>G et la délétion des exons 2 à 6 du gène *UBE2T* n'ont

encore jamais été décrits dans la FA cependant ce sont des LOF et ces deux gènes sont responsables respectivement de FA du groupe I (*Fanconi anemia, complementation group I* ; OMIM # 609053) et du groupe T (*Fanconi anemia, complementation group T* ; OMIM # 616435) par perte de fonction. Un enrichissement en variants potentiellement pathogènes dans les gènes de la FA (*BRCA2, BRIP1, FANCC, FANCE* et *REV3L*) a déjà été décrit dans des formes héréditaires de CCR, avec 6 familles porteuses sur 40 analysées par WES. Le test d'enrichissement réalisé utilisant les témoins de la cohorte ExAC est un test de χ^2 avec une p-value de 0,25% pour les LOF + faux sens prédits délétères, et une p-value de 0,5% pour les LOF seuls (Esteban-Jurado *et al.*, 2016).

En plus de ces 5 LOF dans des gènes de FA, nous avons identifié une variation indel en phase dans le gène *FANCD2,* c.3260delinsGTAA, prédite pour altérer l'épissage, et deux autres LOF ont été identifiés dans *FAN1* et *FAAP20*. Les gènes *FAN1* et *FAAP20* ne sont pas responsables de FA mais les protéines codées appartiennent à la même voie de réparation de l'ADN que ceux impliqués dans la FA, le système ICL. Par ailleurs, bien que cette observation n'ai pas été répliquée, deux variations pathogènes tronquantes et trois variations faux sens du gène *FAN1* ont été impliquées dans la survenue de formes familiales de CCR dans cinq familles différentes *(Seguí et al.*, 2015). Aucune LOH en regard de *FAN1* n'a cependant été identifiée dans les CCR de ces patients. Le gène *FAN1* est responsable d'une maladie rénale rare de transmission autosomique récessive (*karyomegalic intersitial nephritis* ; OMIM # 614817) (Zhou *et al.,* 2012). A ce jour, il n'y a pas de prédisposition au cancer formellement associée aux mutations du gène *FAN1*.

Ces variations dans la voie de l'ICL sont toutes associées à d'autres variations prédites pathogènes de type LOF non impliquées en pathologie, ou à des faux sens non connus prédits délétères par les trois algorithmes utilisés, ce qui est compréhensible étant donné le nombre important de gènes de la réparation analysés. Aucun gène ni aucune autre voie biologique de la réparation particuliers ne semble être récurrent ou se détacher parmi ces co-occurrences. Cependant, il est légitime de se poser la question d'un éventuel oligogénisme.

La seule exception à la récurrence est l'identification de deux variants *MUTYH* à l'état mono-allélique, détectés chez les patients DOCC1544 et DOCC2391, en association à d'autres variations dans les gènes de la voie ICL. Chez la patiente DOCC1544, la variation *MUTYH* est associée à une variation tronquante *BRCA2* et une variation *FANCL*. Chez le patient DOCC2391, la variation *MUTYH* est associée au variant non-sens *XRCC2*. Ces cooccurrences sont particulièrement intéressantes pour discuter d'un modèle oligogénique. En

effet, l'augmentation du risque de CCR chez les patients porteurs d'un variant LOF *MUTYH* mono-allélique reste controversée. Une méta-analyse regroupant 16 études différentes, afin de maximiser la puissance statistique, ne permet pas de conclure à un OR significatif (Win, *et al*. 2011a). Alors qu'une étude de 114 cas index de CCR et 2 179 apparentés a rapporté un sur-risque de CCR avec un OR à 2,04 et une *p-value* inférieure à 0,001 (Win *et al*., 2011b). Il est possible que cette discordance soit attribuable à un effet additif du risque de survenue du cancer induit par d'autres variants concomitants. Un cas de déterminisme digénique du CCR impliquant *MUTYH* a d'ailleurs déjà été rapporté. Il s'agit de l'association d'un variant faux sens *MUTYH* et d'un variant d'épissage *OGG1*, un autre gène impliqué dans le système BER (Morak *et al*., 2011). Les voies de réparation de l'ADN sont intrinsèquement liées et enchevêtrées. Il est donc crédible que plusieurs impacts mineurs dans différentes voies puissent altérer l'efficacité globale des mécanismes de réparation de l'ADN. Pour étayer cette hypothèse d'oligogénisme, des tests statistiques spécifiques supplémentaires devront être réalisés. La contribution de l'oligogénisme dans le déterminisme des cancers avec effet cumulatif de variant rares a été récemment évoquée chez les patients atteints de sarcome (Ballinger *et al*., 2016), bien que l'étude, publiée dans une revue de haut niveau, présente des limites méthodologiques sur le plan statistique. Cette étude d'association a en effet conclut que les patients atteints de sarcome avaient significativement plus de variants pathogènes dans les 72 gènes de cancer testés que les témoins (Ballinger *et al*., 2016). Cependant leur méthode statistique est critiquable puisqu'ils nomment OR la comparaison de la proportion de variants rares délétères par rapport aux variants rares synonymes chez les cas et les témoins au lieu de comparer la proportion de porteurs parmi les cas et les témoins. Cette méthode permet d'augmenter artificiellement considérablement leurs *p-values*.

Nous avons donc effectué des tests statistiques rigoureux afin d'étayer notre hypothèse d'oligogénisme impliquant les gènes de la réparation de l'ADN et contribuant aux formes précoces de CCR.

L'analyse statistique réalisée à partir des données de 76 patients, atteints de CCR précoces rigoureusement sélectionnés et séquencés par le panel de 201 gènes, ne permet pas de mettre en évidence d'association significative entre les variants rares LOF et/ou faux sens prédits délétères de ces gènes et la survenue de CCR précoce, ni même de déceler une tendance. Il est cependant intéressant de noter la baisse de l'OR observée lors de l'ajout des variants faux sens prédits délétères aux variants LOF touchant les gènes de réparation de

l'ADN (Tableau 10). Ce résultat est cohérent avec les données très récemment publiées par le consortium ExAC rapportant que les gènes de réparation de l'ADN présentent de nombreuses mutations dans la population générale, dont 30,4% de variants faux sens, avec plus de la moitié d'entre eux prédits délétères par SIFT et au moins possiblement délétères par Polyphen-2 HVAR (Das et Ghosh 2017). Cela pose la question de la pertinence des algorithmes de prédiction de pathogénicité collégialement utilisés et qui peuvent classer des mutations faux sens en mutations délétères par excès.

La seconde analyse statistique a été réalisée à partir des données de 78 patients, également atteints de CCR précoces rigoureusement sélectionnés, mais séquencés par WES avec application d'un panel *in silico* de 296 gènes impliqués dans la réparation de l'ADN. Les analyses réalisées à partir de l'ensemble des gènes de la réparation de l'ADN montrent globalement un effet plus fort avec un OR supérieur et une *p-value* plus significative pour les variations synonymes, par rapport aux LOF et faux sens prédits délétères (3/3). En revanche lorsque le test est restreint aux singletons (MAF inférieure à 1/1000), ou lors de l'application du test seqMeta, qui permet de prendre en compte la stratification de population, cet effet disparaît. Cela est rassurant et peut être expliqué par le fait que la population de patients est essentiellement rouennaise, alors que les témoins viennent de différents endroits de France. Les structures de populations semblent donc être plus fines que la seule échelle origine française. Cet effet n'est pas observé lorsque les analyses sont restreintes à la voie ICL dont les synonymes ont un OR faible et de *p-value* non significative. L'effet de l'association des variants de la voie ICL au CCR précoce est, quel que soit le test employé, plus important que pour les gènes de la réparation globaux. L'effet est fort pour les LOF, ce qui est cohérent avec les variants pathogènes d'intérêt identifiés dans cette voie. Il est également intéressant de noter que la force de l'association est systématiquement plus importante lorsque l'on restreint l'analyse aux variants ICL de MAF inférieure à 1/1000. En effet, chez les patients la restriction ne change pas le nombre de variants analysés, alors que cela en supprime plusieurs chez les témoins. Cela sous-tend une implication supérieure des variants très rares voir privés au CCR précoce. La rareté de ces variants est en faveur d'une pression de sélection à leur encontre. Bien que non significatives à l'échelle de l'exome, les valeurs d'association obtenues sont bien plus importantes que lors de l'analyse par le panel de 201 gènes et montrent une tendance. Pour la comparaison des proportions cas témoins de porteurs de variants LOF de MAF inférieure à 1/1000 dans la voie ICL, l'OR obtenu est de 8,7 avec un intervalle de confiance 95% ne contenant pas la valeur 1 et une p-value de 1,6%. Les résultats

du *Burden test* SeqMeta ont un OR plus faible et une *p-value* un peu plus significative avec un OR à 5,9 et une *p-value* à 1,2%. Cela est probablement dû à l'ajustement que permet ce test sur l'analyse de stratification de population. Le SKAT test permet d'obtenir une *p-value* inférieure aux autres tests : 2,1% pour les LOF de l'ensemble des gènes de la réparation, et 0,2% pour les LOF de l'ICL de MAF inférieure à 1/1000. C'est étonnant puisque le SKAT test sert à prendre en compte la variance de l'OR, ce qui permet de limiter l'effet de cooccurrence de variations pathogènes et protectrices, tirant l'OR vers un milieu neutre dans les autres tests alors que ce n'est pas le cas. Cette *p-value* plus significative serait donc en faveur de l'existence de variants protecteurs dans les gènes impliqués dans la réparation de l'ADN, ce qui n'a encore jamais été décrit. Un test d'enrichissement des variants dans les gènes de la réparation de l'ADN ou de l'ICL par rapport aux variants détectés dans l'ensemble de l'exome a également été appliqué. Cependant l'OR le plus important, obtenu pour les LOF ICL de MAF inférieure à 1/1000 est plus faible que pour les autres tests et l'intervalle de confiance 95% comprend la valeur de 1. Le dernier test réalisé pour évaluer l'oligogénisme est un test de cooccurrence des variants. Cependant le nombre d'événements est globalement trop faible pour obtenir une *p-value* inférieure à 5% pour les gènes de l'ensemble de la réparation de l'ADN. L'OR très élevé observé pour la cooccurence de variants LOF de la voie ICL est intéressant mais le petit nombre de cas porteurs parmi les patients ne permet pas de conclure formellement à un oligogénisme. Les cooccurrences détaillées dans l'analyse manuelle des résultats sont survenues le plus souvent chez des patients *outliers*, qui ont été exclus de l'étude statistique. Il se peut que ce soit l'effet du hasard, mais les patients *outliers* portent globalement plus de variants que les autres et il y a donc plus de probabilité que des variants soient identifiés dans des gènes d'intérêt. Quoiqu'il en soit si les cooccurrences observées concourent au CCR par déterminisme oligogénique, la taille de l'échantillon, *outliers* exclus, est insuffisante pour le démontrer.

Globalement malgré une tendance en faveur de l'implication des variants rares LOF de la voie ICL, les résultats des analyses statistiques sur les données obtenues par panel de gènes et par exome ne sont pas significatifs à l'échelle de l'exome. C'est également le cas des rares études portant sur l'impact de ces variants rares touchant des gènes impliqués dans la réparation de l'ADN. En effet, la *p-value* du test d'enrichissement réalisé pour comparer les variants potentiellement délétères de la voie ICL (*BRCA2, BRIP1, FANCC*, *FANCE* et *REV3L*) des 6 familles porteuses sur 40 analysées par WES aux variants des témoins de la cohorte ExAC était de seulement 0,25% pour les LOF et faux sens prédits délétères, et de

0,5% pour les LOF seuls (Esteban-Jurado *et al.*, 2016). Une autre étude portant sur l'analyse de 40 gènes impliqués dans la réparation de l'ADN à partir de données de WES a été réalisée pour comparer une cohorte de 680 patients atteints de CCR avec 27 728 témoins appariés ExAC NFE non TGCA. Cette étude a été répliquée avec une cohorte de 1 661 patients atteints de CCR et 1456 patients atteints de forme dite précoce de CCR, c'est-à-dire de survenue avant 56 ans. Les auteurs concluent à un enrichissement en variants pathogènes dans les gènes *ATM* (OR = 2,81 p = 3,5% ; OR = 2,97 p = 0,13%) et *PALB2* (OR = 4,91 p = 2,4% ; OR = 3,42 p = 3,4%) chez les patients, mais sans effet sur la précocité du cancer. Une LOH tumorale a été systématiquement observée en cas de variation pathogène constitutionnelle d'*ATM* et jamais de *PALB2*. D'autres gènes leur ont paru intéressants, dont certains gènes impliqués dans l'ICL comme *BRCA2, FANCE, FANCL, SLX4*, mais l'OR obtenu comprenait la valeur 1 et les p-values étaient situées entre 5 et 30% (AlDubayan *et al.*, 2018). De façon intéressante, nous avons également identifié dans notre cohorte un variant LOF de classe 5 dans le gène *ATM*. La mise en évidence de LOH à ce locus dans la tumeur permettrait d'étayer les résultats de cette publication. Les auteurs ont également conclut à l'implication des altérations de l'ensemble des gènes de la voie HR dans le CCR (OR = 1,78 IC95% = 1,07-2,86 et p-value 2%), avec une association encore plus forte pour la population de CCR précoce (OR = 1,78 IC95% = 1,19-2,36 et p-value 0,2%). Les gènes sélectionnés par les auteurs comme appartenant à la voie HR étaient : *ATM, BARD1, BLM, BRCA1, BRCA2, BRIP1, MRE11, NBN, PALB2, RAD51, RAD51C, RAD51D, RAD54L* et *XRCC3.* Dans notre étude, les gènes *BRCA2, BRIP1, PALB2* ont été classés dans la voie ICL, et les gènes *ATM* et *BLM* dans la catégorie des gènes responsables d'une hypersensibilité aux agents génotoxiques, conformément à la classification utilisée dans la publication du consortium ExAC (Das et Ghosh, 2017). Ces différences de classement illustrent la difficulté de la classification de certains gènes à cause de la multiplicité des fonctions de certaines protéines et du chevauchement important des voies de réparation de l'ADN.

Une des premières limites de notre travail est l'effectif réduit de patients, qui limite la puissance de l'analyse statistique. En effet les OR obtenus sont relativement faibles et une forte puissance statistique serait nécessaire pour obtenir des résultats significatifs à l'échelle du WES. La population de témoins que nous avons utilisés pour l'étude est également critiquable. Bien que d'origine française donc appariée à nos cas, il s'agit d'un échantillon d'individus de la population générale. Nous ne pouvons pas exclure la survenue de cancer chez certains d'entre eux car nous n'avons pas d'information sur leur phénotype. Les OR de

survenue de CCR précoce identifiés chez les cas sont relativement faibles, ce qui est étonnant alors que la population de patients est rigoureusement sélectionnée et présente un CCR très précoce. Cela peut être expliqué par une pénétrance incomplète des variants qui peuvent donc être identifiés chez les témoins. Le modèle oligogénique, de transmission non mendélienne, qui suppose un effet lors de l'accumulation de variants, implique que les témoins puissent être porteurs de certains de ces variants concourant à une éventuelle prédisposition, ce qui augmente encore la puissance statistique nécessaire pour le démontrer. La population de témoins est âgée en moyenne de 66 ans contre 27 ans pour les patients. Il est également légitime de se poser la question d'un éventuel biais induit par cet écart. Bien qu'essentiellement *passenger,* des mutations somatiques s'accumulent en effet avec l'âge, y compris dans les cellules hématopoïétiques (Holstege *et al*., 2014). Cependant ces mutations étant sous clonales, la balance allélique de ces mutations devrait être plus faible que celle des variations constitutionnelles. Comme nous appliquons un filtre sélectionnant les variants hétérozygotes de balance allélique comprise entre 25 et 75%, cela ne devrait pas impacter le nombre de variants détectés chez les témoins.

Les variations pathogènes identifiées dans la voie ICL chez les patients non *outliers* n'étant qu'au nombre de 6, les résultats de cette étude sont quoiqu'il en soit en faveur d'une hétérogénéité importante des éventuelles étiologies, génétiques ou non, responsables de la survenue précoce de CCR. En effet, il n'a pas été identifié de variations délétères, de gènes ou de voies biologiques altérés de façon significativement récurrente chez la majorité des patients de cet échantillon. Différentes situations peuvent mener à cette observation. Il est possible que le déterminisme génétique soit important dans ces formes précoces de CCR, mais qu'il y ait une véritable hétérogénéité génétique des variants, des gènes et des voies biologiques impliquées. Nous n'analysons ici les données qu'avec l'angle de vue de la réparation de l'ADN, or les voies de l'apoptose, de la ségrégation des chromosomes et du cycle cellulaire peuvent aussi concourir à la survenue précoce de CCR. Auquel cas plus l'hétérogénéité est importante, plus la taille des échantillons doit être grande pour obtenir une puissance statistique suffisante. Le critère de survenue précoce du CCR a été décrit par d'autres équipes comme générateur d'une population hétérogène (Hahn *et al*., 2016). Bien que l'âge de 32 ans que nous avons retenu soit bien plus précoce que le seuil de 51 ans souvent utilisé dans la bibliographie, il est peut-être nécessaire d'être encore plus stringent dans les critères de sélection, en se restreignant à une survenue pédiatrique ou en sélectionnant des patients sur des critères cliniques ou biologiques plus stricts. L'approche par sélection de cas

de CCR familiaux est également intéressante, mais elle est probablement inadaptée à la recherche de variants ou de gènes avec une pénétrance modérée. Cette dernière complique en effet considérablement la mise en évidence d'une héritabilité (Hahn *et al.*, 2016). A l'inverse, il est également probable que la force de l'effet de la voie ICL ou même des gènes de la réparation de l'ADN soit diluée par notre méthode de comparaison des variations. En effet, afin de ne pas biaiser l'étude statistique, nous sommes contraints de donner le même poids biologique à tous les variants identifiés au sein d'une même catégorie. Lors de l'agrégation des variants par voie biologique, nous considérons ainsi de manière égale les variants dans les différents gènes inclus dans ces voies. Or pour l'ICL par exemple, la majorité des variants LOF portés par les témoins se situent dans le gène *BRIP1* (3) et dans le gène *FANCM* (5), alors qu'aucune variation pathogène dans ces gènes n'a été identifiée chez les patients. De même, au sein d'un gène, tous les variants du même type sont considérés de la même façon. Concernant les LOF, on sait que des mécanismes de sauvetage de la protéine par saut d'exons ou autres ont été décrits dans certains gènes. Il est aussi à noter, en particulier pour les faux sens, que certains des variants qui sont localisés dans un domaine crucial n'ont pas nécessairement la même pathogénicité qu'un variant localisé à un autre endroit. C'est le cas par exemple pour les variations *POLE* et *POLD1* qui ne sont responsables de phénotype que lorsqu'elles surviennent dans le domaine exonucléasique (Palles *et al.,* 2013). Il est possible que les faibles OR observés soient le reflet de la dilution de l'effet des variants LOF réellement pathogènes par des variants LOF bénins. L'approche statistique systématique risque donc de noyer un potentiel effet de certains variants, dans la masse d'autres non pertinents. Il faut également garder à l'esprit que, bien que nos méthodes d'analyse par NGS soient très performantes, nous sommes encore loin de détecter toutes les variations du génome. Les variations hors des séquences codantes, introniques ou dans les séquences régulatrices de gènes d'intérêts, ainsi que les variations hors des gènes ne sont pas détectées par le panel de 201 gènes ni le WES. Cependant l'analyse de ces régions ne sera vraiment pertinente que lorsque nous aurons progressé en parallèle sur l'interprétation des variants susceptibles d'y être identifiés. De même, certaines variations de structure ne sont à ce jour pas détectées par les techniques d'analyses utilisées (Hahn *et al.*, 2016).

Il est également possible que le déterminisme génétique soit finalement peu important dans les formes précoces de CCR. L'absence de résultat statistique significatif pourrait être expliquée par le fait qu'une fraction importante des patients présentant un CCR précoce inclus dans cette étude, n'ait pas de mutations constitutionnelles d'un gène majeur intervenant dans

le déterminisme génétique du CCR mais une altération biallélique exclusivement somatique. La contribution des altérations somatiques dans la genèse des cancers du sujet jeune a été récemment rappelée par l'étude du groupe d'Albert de La Chapelle, montrant que la majorité des CCR du sujet jeune de présentation sporadique et de phénotype RER positif s'explique par une mutation biallélique d'un gène MMR (Haraldsdottir *et al.*, 2014). Il s'agit d'une information majeure et rassurante pour les familles puisque notre étude suggère fortement que la majorité des CCR même de survenue très précoce ne sont probablement pas liés à un déterminisme héréditaire. Le rôle de l'environnement dans l'oncogenèse ne doit également pas être négligé, et les facteurs environnementaux délétères auxquels les patients inclus dans notre étude ont pu être exposés n'ont pas été évalués.

Malgré l'absence de résultats statistiques significatifs, l'étude statistique effectuée à l'échelle du gène nous a permis de nous intéresser aux gènes *SBDS* et *WT1*, bien que déjà impliqués dans des maladies mendéliennes sans rapport connu avec le CCR. Le variant récurrent *SBDS* c.258+2T>C, p. ? détecté chez 3 patients et 3 témoins est décrit pathogène dans le syndrome Shwachman-Diamond (SDS ; OMIM # 260400 ; Taneichi *et al.,* 2006; Boocock *et al.,* 2002). Ce syndrome, de révélation à l'âge pédiatrique, de transmission autosomique récessive et associé à une instabilité chromosomique est caractérisé par une insuffisance médullaire, une insuffisance pancréatique, des anomalies squelettiques et un risque augmenté d'hémopathies malignes. Il a d'ailleurs été montré que la perte de SBDS perturbait la réponse cellulaire aux dommages de l'ADN et aux stress chimiques, suggérant un potentiel rôle suppresseur de tumeur (Ball *et al.*, 2009). Le variant détecté est également décrit à l'état hétérozygote comme facteur de risque d'anémie aplasique (Calado *et al.*, 2007). Il résulte de la conversion génique du gène *SBDS* avec son pseudogène *SBDSP1* (*ribosome maturation factor pseudogene 1*), qui produit un long ARN non codant (*lncRNA*, figure 16). Il est associé à l'abolition du site donneur d'épissage original en 258+1 au profit de l'utilisation du site d'épissage de *SBDSP1* en 250+1, ce qui aboutit à une protéine SBDS tronquée (Boocock *et al.*, 2002).

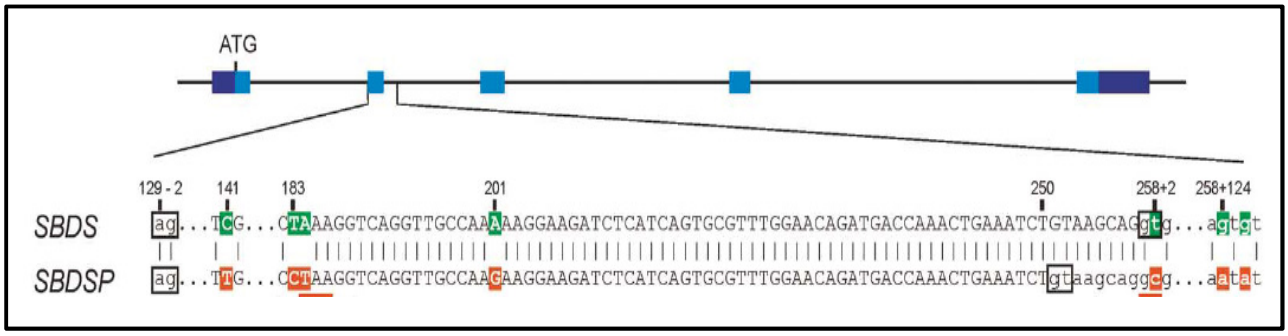


Figure 16 : Variations entre le gène SBDS et son pseudogène SBDSP1 (Boocock et al., 2002)

La protéine SBDS est impliquée dans la maturation des ribosomes et pléiotrope, c'est-à-dire avec de nombreuses et diverses fonctions, qui restent cependant encore mal connues (In *et al.*, 2016). Le *lncRNA* SBDP1 issu de la transcription du pseudogène est augmenté dans les tissus et cellules de CCR et a été impliqué dans la croissance tumorale, la prolifération cellulaire, l'invasion et la tumorigenèse du CCR. Une des hypothèses concernant l'implication de *SBDSP1* dans le CCR est l'altération par ce *lncRNA* de la potentielle fonction suppresseur de tumeur de *SBDS* (Shi *et al.*, 2017). De façon intéressante, la mutation d'épissage mise en évidence dans notre échantillon mime la séquence de ce pseudogène impliqué au niveau somatique.

De même, nous ne pouvons exclure formellement l'implication de variants *WT1* dans le déterminisme génétique du CCR. Deux variants faux sens situés dans le domaine N-terminal de *WT1*, impliqué dans la régulation de la transcription et l'homodimérisation de WT1 ont été identifiés chez trois patients et un témoin. Il est à noter que nous avons détecté une troisième variation faux sens prédite délétère par les 3 algorithmes, c.695C>T, p.(Pro232Leu) dans l'échantillon de patients, située dans le même domaine de WT1, mais ce dernier variant, présent à 0,013% dans la population africaine (ExAC AFR), avait été exclu de l'analyse statistique par l'analyse de stratification de population car le patient porteur est d'origine africaine. La protéine WT1 est un facteur de transcription impliquée dans la cascade de signalisation de la voie Wnt, essentielle au développement du système uro-génital, et très peu exprimée dans le côlon d'après les données GTEx. Les variants pathogènes constitutionnels *WT1* prédisposent à la tumeur de Wilms ou néphroblastome (OMIM # 194070), une forme pédiatrique de cancer du rein le plus souvent sporadique et de

présentation clinique isolée ou syndromique. Cependant de façon intéressante, l'expression de WT1 lors du checkpoint mitotique prolonge l'inhibition du complexe de destruction formé par APC et AXIN1 empêchant l'adressage de la β-caténine au protéasome, et retardant ainsi la sortie de mitose (Shandilya *et al.* 2014; Shandilya et Roberts 2015). La voie Wnt est impliquée dans le CCR puisque la FAP1 est due à des variations pathogènes dans le gène *APC*. Par ailleurs un polymorphisme de *WT1* a déjà été impliqué comme facteur de risque, et facteur pronostique dans le CCR par une étude cas-témoins (Sangkhathat *et al.*, 2015). Les patients porteurs n'étant pas atteints de tumeur de Wilms, outre la pénétrance incomplète, il est possible que les variations faux sens détectées n'aient qu'un effet hypomorphe, insuffisant pour prédisposer au néphroblastome.

SBDS et *WT1* sont donc des gènes candidats intéressants, cependant des études de ségrégation et des études de réplication rigoureuses seront nécessaires pour étayer leur implication potentielle dans le CCR.

5. CONCLUSION ET PERSPECTIVES

En conclusion, les résultats de cette étude sont en faveur d'une hétérogénéité de l'étiologie de la survenue de CCR précoces, ce qui rend l'implication formelle de variants, gènes, ou voies biologiques difficile. Différentes pistes peuvent être exploitées pour poursuivre ce travail. Les OR mis en évidence jusqu'ici sont relativement faibles, et nécessitent donc une population de patients beaucoup plus importante pour atteindre des valeurs de *p-value* significatives à l'échelle du WES. Le phénotype de CCR précoce étant rare, une collaboration nationale ou internationale pourrait permettre d'augmenter considérablement l'échantillon de patients analysés et donc d'accroître la puissance statistique. Par ailleurs, le rôle dans l'oncogenèse des altérations somatiques de gènes impliqués dans les voies de la réparation de l'ADN n'est pas restreint aux CCR. Une étude d'association réalisée sur une population de 98 patients chinois atteints de cancer de la vessie séquencés par un panel de 222 gènes, dont 54 gènes impliqués dans la réparation de l'ADN, comparée aux données des bases de données de population asiatiques ExAC et GnomAD, a récemment conclu à l'association d'altérations pathogènes dans ces 54 gènes avec un risque de survenue de cancer plus précoce 3 fois plus élevé, malgré une *p-value* non significative de 17%, et associé à un pronostic défavorable avec une p-value de 1,3% (Na *et al*., 2018). De telles conclusions, basées sur une étude avec des cas-contrôles non appariés et sans *p-value* significative à l'échelle du WES, ne peuvent, comme dans notre cas, être interprétés comme plus d'une simple tendance. Une autre étude, portant sur l'analyse des données de WES de 412 patientes atteintes de cancer de l'ovaire de haut grade dans la cohorte TGCA, conclue à un enrichissement en variants dans les gènes impliqués dans la réparation de l'ADN (*p-value* = $1{,}8 \times 10^{-3}$), bien que non significatif à l'échelle de l'exome. Ils décrivent également une association entre variants pathogènes de *FANCM* et cancer de l'ovaire (OR = 2,5 IC95% 1,3 - 5 et p = 6% ; Dicks *et al*., 2017). Ces études posent la question de l'intérêt d'agréger différents types de cancers précoces ensemble, afin d'augmenter le nombre de cas de cancers précoces donc rares analysés. Il pourrait ainsi être intéressant de constituer une cohorte de patients atteints de cancers précoces de différents types pour s'affranchir en partie de la rareté du phénotype. Le risque est d'augmenter encore l'hétérogénéité mise en évidence par cette étude. Etant donnée la tendance mise en évidence, des études supplémentaires concernant l'association des altérations très rares voire privées des gènes de la voie ICL aux formes

précoces de CCR devront être menées. L'analyse de la tumeur des patients porteurs de variants pathogènes dans les gènes de la voie ICL serait également intéressante pour rechercher une éventuelle LOH, mais en cas de déterminisme oligogénique impliquant les gènes de la réparation, il n'est pas certain que le *two-hits model* de Knudson s'applique toujours. Il est en effet possible que l'altération mineure de protéines impliquées dans la réparation fournisse un terrain génétique propice à la tumorigenèse, sans qu'un second événement soit nécessaire. L'estimation du déterminisme oligogénique impliquant les variants rares n'en est qu'à ses débuts et le développement d'outils bio-informatiques et statistiques adaptés suivra probablement le bond technologique qui a déjà permis un progrès fulgurant en génétique moléculaire. Quoiqu'il en soit, la poursuite de cette étude bénéficiera du socle méthodologique et bio-informatique adapté aux études statistiques d'association développé et inauguré par ce travail, indispensable aujourd'hui à toute implication de gènes en pathologie (Broderick *et al*., 2017).

6. BIBLIOGRAPHIE

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat. Methods *7*, 248–249.

AlDubayan, S.H., Giannakis, M., Moore, N.D., Han, G.C., Reardon, B., Hamada, T., Mu, X.J., Nishihara, R., Qian, Z., Liu, L., et al. (2018). Inherited DNA-Repair Defects in Colorectal Cancer. Am. J. Hum. Genet. *102*, 401–414.

Ali, A.M., Kirby, M., Jansen, M., Lach, F.P., Schulte, J., Singh, T.R., Batish, S.D., Auerbach, A.D., Williams, D.A., and Meetei, A.R. (2009). Identification and characterization of mutations in FANCL gene: a second case of Fanconi anemia belonging to FA-L complementation group. Hum. Mutat. *30*, E761-770.

Armanios, M., Chen, J.-L., Chang, Y.-P.C., Brodsky, R.A., Hawkins, A., Griffin, C.A., Eshleman, J.R., Cohen, A.R., Chakravarti, A., Hamosh, A., et al. (2005). Haploinsufficiency of telomerase reverse transcriptase leads to anticipation in autosomal dominant dyskeratosis congenita. Proc. Natl. Acad. Sci. U.S.A. *102*, 15960–15964.

Auton, A., Abecasis, G.R., Altshuler, D.M., Durbin, R.M., Abecasis, G.R., Bentley, D.R., Chakravarti, A., Clark, A.G., Donnelly, P., Eichler, E.E., et al. (2015). A global reference for human genetic variation. Nature *526*, 68–74.

Baert-Desurmont, S., Charbonnier, F., Houivet, E., Ippolito, L., Mauillon, J., Bougeard, M., Abadie, C., Malka, D., Duffour, J., Desseigne, F., et al. (2016). Clinical relevance of 8q23, 15q13 and 18q21 SNP genotyping to evaluate colorectal cancer risk. European Journal of Human Genetics *24*, 99–105.

Baert-Desurmont, S., Coutant, S., Charbonnier, F., Macquere, P., Lecoquierre, F., Schwartz, M., Blanluet, M., Vezain, M., Lanos, R., Quenez, O., et al. (2018). Optimization of the diagnosis of inherited colorectal cancer using NGS and capture of exonic and intronic sequences of panel genes. European Journal of Human Genetics 1.

Ball, H.L., Zhang, B., Riches, J.J., Gandhi, R., Li, J., Rommens, J.M., and Myers, J.S. (2009). Shwachman-Bodian Diamond syndrome is a multi-functional protein implicated in cellular stress responses. Human Molecular Genetics *18*, 3684–3695.

Ballinger, M.L., Goode, D.L., Ray-Coquard, I., James, P.A., Mitchell, G., Niedermayr, E., Puri, A., Schiffman, J.D., Dite, G.S., Cipponi, A., et al. (2016). Monogenic and polygenic determinants of sarcoma risk: an international genetic study. The Lancet Oncology *17*, 1261–1271.

Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.

Barnett, I.J., Lee, S., and Lin, X. (2013). Detecting Rare Variant Effects Using Extreme Phenotype Sampling in Sequencing Association Studies. Genetic Epidemiology *37*, 142–151.

Beck, T.F., Mullikin, J.C., NISC Comparative Sequencing Program, and Biesecker, L.G. (2016). Systematic Evaluation of Sanger Validation of Next-Generation Sequencing Variants. Clin. Chem. *62*, 647–654.

Belhadj, S., Mur, P., Navarro, M., González, S., Moreno, V., Capellá, G., and Valle, L. (2017). Delineating the Phenotypic Spectrum of the NTHL1-Associated Polyposis. Clin. Gastroenterol. Hepatol. *15*, 461–462.

Berger, A.H., Knudson, A.G., and Pandolfi, P.P. (2011). A continuum model for tumour suppression. Nature *476*, 163–169.

Boocock, G.R.B., Morrison, J.A., Popovic, M., Richards, N., Ellis, L., Durie, P.R., and Rommens, J.M. (2002). Mutations in SBDS are associated with Shwachman–Diamond syndrome. Nature Genetics *33*, 97–101.

Briggs, S., and Tomlinson, I. (2013). Germline and somatic polymerase ε and δ mutations define a new class of hypermutated colorectal and endometrial cancers. J. Pathol. *230*, 148–153.

Broderick, P., Dobbins, S.E., Chubb, D., Kinnersley, B., Dunlop, M.G., Tomlinson, I., and Houlston, R.S. (2017). Validation of Recently Proposed Colorectal Cancer Susceptibility Gene Variants in an Analysis of Families and Patients—a Systematic Review. Gastroenterology *152*, 75-77.e4.

Burn, J., Mathers, J.C., and Bishop, D.T. (2013). Chemoprevention in Lynch syndrome. Fam. Cancer *12*, 707–718.

Calado, R.T., Graf, S.A., Wilkerson, K.L., Kajigaya, S., Ancliff, P.J., Dror, Y., Chanock, S.J., Lansdorp, P.M., and Young, N.S. (2007). Mutations in the SBDS gene in acquired aplastic anemia. Blood *110*, 1141–1146.

Cao, X., Eu, K.W., Kumarasinghe, M.P., Li, H.H., Loi, C., and Cheah, P.Y. (2006). Mapping of hereditary mixed polyposis syndrome (HMPS) to chromosome 10q23 by genomewide high-density single nucleotide polymorphism (SNP) scan and identification of BMPR1A loss of function. Journal of Medical Genetics *43*, e13–e13.

Castella, M., Pujol, R., Callén, E., Trujillo, J.P., Casado, J.A., Gille, H., Lach, F.P., Auerbach, A.D., Schindler, D., Benítez, J., et al. (2011). Origin, functional role, and clinical impact of Fanconi anemia FANCA mutations. Blood *117*, 3759–3769.

Chalela, R., Curull, V., Enríquez, C., Pijuan, L., Bellosillo, B., and Gea, J. (2017). Lung adenocarcinoma: from molecular basis to genome-guided therapy and immunotherapy. Journal of Thoracic Disease *9*, 2142–2158.

de la Chapelle, A. (2004). Genetic predisposition to colorectal cancer. Nature Reviews Cancer *4*, 769–780.

Chubb, D., Broderick, P., Dobbins, S.E., Frampton, M., Kinnersley, B., Penegar, S., Price, A., Ma, Y.P., Sherborne, A.L., Palles, C., et al. (2016). Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. Nature Communications *7*, 11883.

Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92.

COGENT Study, Houlston, R.S., Webb, E., Broderick, P., Pittman, A.M., Di Bernardo, M.C., Lubbe, S., Chandler, I., Vijayakrishnan, J., Sullivan, K., et al. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. Nat. Genet. *40*, 1426–1435.

Curtin, N.J. (2012). DNA repair dysregulation from cancer driver to therapeutic target. Nat. Rev. Cancer *12*, 801–817.

Das, R., and Ghosh, S.K. (2017). Genetic variants of the DNA repair genes from Exome Aggregation Consortium (EXAC) database: significance in cancer. DNA Repair *52*, 92–102.

Deans, A.J., and West, S.C. (2011). DNA interstrand crosslink repair and cancer. Nat Rev Cancer *11*, 467–480.

Degrolard-Courcet, E., Sokolowska, J., Padeano, M.-M., Guiu, S., Bronner, M., Chery, C., Coron, F., Lepage, C., Chapusot, C., Loustalot, C., et al. (2014). Development of primary early-onset colorectal cancers due to biallelic mutations of the FANCD1/BRCA2 gene. Eur. J. Hum. Genet. *22*, 979–987.

Desmet, F.-O., Hamroun, D., Lalande, M., Collod-Béroud, G., Claustres, M., and Béroud, C. (2009). Human Splicing Finder: an online bioinformatics tool to predict splicing signals. Nucleic Acids Research *37*, e67–e67.

Dicks, E., Song, H., Ramus, S.J., Oudenhove, E.V., Tyrer, J.P., Intermaggio, M.P., Kar, S., Harrington, P., Bowtell, D.D., Group, A.S., et al. (2017). Germline whole exome sequencing and large-scale replication identifies FANCM as a likely high grade serous ovarian cancer susceptibility gene. Oncotarget *8*, 50930–50940.

Din, F.V.N., Valanciute, A., Houde, V.P., Zibrova, D., Green, K.A., Sakamoto, K., Alessi, D.R., and Dunlop, M.G. (2012). Aspirin inhibits mTOR signaling, activates AMP-activated protein kinase, and induces autophagy in colorectal cancer cells. Gastroenterology *142*, 1504-1515.e3.

Dobbins, S.E., Broderick, P., Chubb, D., Kinnersley, B., Sherborne, A.L., and Houlston, R.S. (2016). Undefined familial colorectal cancer and the role of pleiotropism in cancer susceptibility genes. Fam. Cancer *15*, 593–599.

Dunlop, M.G., Dobbins, S.E., Farrington, S.M., Jones, A.M., Palles, C., Whiffin, N., Tenesa, A., Spain, S., Broderick, P., Ooi, L.-Y., et al. (2012). Common variation near CDKN1A, POLD3 and SHROOM2 influences colorectal cancer risk. Nat. Genet. *44*, 770–776.

Dunlop, M.G., Tenesa, A., Farrington, S.M., Ballereau, S., Brewster, D.H., Koessler, T., Pharoah, P., Schafmayer, C., Hampe, J., Völzke, H., et al. (2013). Cumulative impact of common genetic variants and other risk factors on colorectal cancer risk in 42 103 individuals. Gut *62*, 871–881.

Easton, D.F., Bishop, D.T., Ford, D., and Crockford, G.P. (1993). Genetic linkage analysis in familial breast and ovarian cancer: results from 214 families. The Breast Cancer Linkage Consortium. Am. J. Hum. Genet. *52*, 678–701.

Esteban-Jurado, C., Franch-Expósito, S., Muñoz, J., Ocaña, T., Carballal, S., López-Cerón, M., Cuatrecasas, M., Vila-Casadesús, M., Lozano, J.J., Serra, E., et al. (2016). The Fanconi anemia DNA damage repair pathway in the spotlight for germline predisposition to colorectal cancer. European Journal of Human Genetics *24*, 1501–1505.

Friedberg, E.C. (2003). DNA damage and repair. Nature *421*, 436–440.

Friedman, E., Efrat, N., Soussan-Gutman, L., Dvir, A., Kaplan, Y., Ekstein, T., Nykamp, K., Powers, M., Rabideau, M., Sorenson, J., et al. (2015). Low-level constitutional mosaicism of a de novoBRCA1 gene mutation. Br. J. Cancer *112*, 765–768.

Garber, J.E., and Offit, K. (2005). Hereditary cancer predisposition syndromes. J. Clin. Oncol. *23*, 276–292.

Garre, P., Martín, L., Sanz, J., Romero, A., Tosar, A., Bando, I., Llovet, P., Diaque, P., García-Paredes, B., Díaz-Rubio, E., et al. (2015). BRCA2 gene: a candidate for clinical testing in familial colorectal cancer type X. Clin. Genet. *87*, 582–587.

Grady, W.M. (2003). Genetic testing for high-risk colon cancer patients. Gastroenterology *124*, 1574–1594.

Hahn, M.M., de Voer, R.M., Hoogerbrugge, N., Ligtenberg, M.J.L., Kuiper, R.P., and van Kessel, A.G. (2016). The genetic heterogeneity of colorectal cancer predisposition - guidelines for gene discovery. Cell Oncol (Dordr) *39*, 491–510.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hansen, M.F., Johansen, J., Sylvander, A.E., Bjørnevoll, I., Talseth-Palmer, B.A., Lavik, L.A.S., Xavier, A., Engebretsen, L.F., Scott, R.J., Drabløs, F., et al. (2017). Use of multigene-panel identifies pathogenic variants in several CRC-predisposing genes in patients previously tested for Lynch Syndrome: HANSEN et al. Clinical Genetics.

Hao, H.-X., Khalimonchuk, O., Schraders, M., Dephoure, N., Bayley, J.-P., Kunst, H., Devilee, P., Cremers, C.W.R.J., Schiffman, J.D., Bentz, B.G., et al. (2009). SDH5, a gene required for flavination of succinate dehydrogenase, is mutated in paraganglioma. Science *325*, 1139–1142.

Haraldsdottir, S., Hampel, H., Tomsic, J., Frankel, W.L., Pearlman, R., de la Chapelle, A., and Pritchard, C.C. (2014). Colon and Endometrial Cancers With Mismatch Repair Deficiency Can Arise From Somatic, Rather Than Germline, Mutations. Gastroenterology *147*, 1308-1316.e1.

Hashimoto, S., Anai, H., and Hanada, K. (2016). Mechanisms of interstrand DNA crosslink repair and human disorders. Genes Environ *38*.

Haynes, B., Saadat, N., Myung, B., and Shekhar, M.P.V. (2015). Crosstalk between translesion synthesis, Fanconi anemia network, and homologous recombination repair pathways in interstrand DNA crosslink repair and development of chemoresistance. Mutat Res Rev Mutat Res *763*, 258–266.

Hes, F.J., Nielsen, M., Bik, E.C., Konvalinka, D., Wijnen, J.T., Bakker, E., Vasen, H.F.A., Breuning, M.H., and Tops, C.M.J. (2008). Somatic APC mosaicism: an underestimated cause of polyposis coli. Gut *57*, 71–76.

Holstege, H., Pfeiffer, W., Sie, D., Hulsman, M., Nicholas, T.J., Lee, C.C., Ross, T., Lin, J., Miller, M.A., Ylstra, B., et al. (2014). Somatic mutations found in the healthy blood compartment of a 115-yr-old woman demonstrate oligoclonal hematopoiesis. Genome Res. *24*, 733–742.

Houlston, R.S., Cheadle, J., Dobbins, S.E., Tenesa, A., Jones, A.M., Howarth, K., Spain, S.L., Broderick, P., Domingo, E., Farrington, S., et al. (2010). Meta-analysis of three genome-wide association studies identifies susceptibility loci for colorectal cancer at 1q41, 3q26.2, 12q13.13 and 20q13.33. Nat. Genet. *42*, 973–977.

Huff, V., Miwa, H., Haber, D.A., Call, K.M., Housman, D., Strong, L.C., and Saunders, G.F. (1991). Evidence for WT1 as a Wilms tumor (WT) gene: intragenic germinal deletion in bilateral WT. Am. J. Hum. Genet. *48*, 997–1003.

In, K., Zaini, M.A., Müller, C., Warren, A.J., von Lindern, M., and Calkhoven, C.F. (2016). Shwachman–Bodian–Diamond syndrome (SBDS) protein deficiency impairs translation re-initiation from *C/EBPα* and *C/EBPβ* mRNAs. Nucleic Acids Research *44*, 4134–4146.

Jaeger, E., Leedham, S., Lewis, A., Segditsas, S., Becker, M., Cuadrado, P.R., Davis, H., Kaur, K., Heinimann, K., Howarth, K., et al. (2012). Hereditary mixed polyposis syndrome is caused by a 40-kb upstream duplication that leads to increased and ectopic expression of the BMP antagonist GREM1. Nat. Genet. *44*, 699–703.

Jasperson, K.W., Tuohy, T.M., Neklason, D.W., and Burt, R.W. (2010). Hereditary and Familial Colon Cancer. Gastroenterology *138*, 2044–2058.

Jatoi, I., and Benson, J.R. (2016). Management of women with a hereditary predisposition for breast cancer. Future Oncol *12*, 2277–2288.

Kalb, R., Neveling, K., Hoehn, H., Schneider, H., Linka, Y., Batish, S.D., Hunt, C., Berwick, M., Callen, E., Surralles, J., et al. (2007). Hypomorphic mutations in the gene encoding a key Fanconi anemia protein, FANCD2, sustain a significant group of FA-D2 patients with severe phenotype. Am. J. Hum. Genet. *80*, 895–910.

Kassahn, K.S., Scott, H.S., and Caramins, M.C. (2014). Integrating massively parallel sequencing into diagnostic workflows and managing the annotation and clinical interpretation challenge. Hum. Mutat. *35*, 413–423.

Knies, K., Inano, S., Ramírez, M.J., Ishiai, M., Surrallés, J., Takata, M., and Schindler, D. (2017). Biallelic mutations in the ubiquitin ligase RFWD3 cause Fanconi anemia. J. Clin. Invest. *127*, 3013–3027.

Knudson, A.G. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc. Natl. Acad. Sci. U.S.A. *68*, 820–823.

Knudson, A.G. (2002). Cancer genetics. Am. J. Med. Genet. *111*, 96–102.

Krupa, R., Sliwinski, T., Wisniewska-Jarosinska, M., Chojnacki, J., Wasylecka, M., Dziki, L., Morawiec, J., and Blasiak, J. (2011). Polymorphisms in RAD51, XRCC2 and XRCC3 genes of the homologous recombination repair in colorectal cancer--a case control study. Mol. Biol. Rep. *38*, 2849–2854.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Lawrence, M.S., Stojanov, P., Mermel, C.H., Robinson, J.T., Garraway, L.A., Golub, T.R., Meyerson, M., Gabriel, S.B., Lander, E.S., and Getz, G. (2014). Discovery and saturation analysis of cancer genes across 21 tumour types. Nature *505*, 495–501.

Lee, S., Abecasis, G.R., Boehnke, M., and Lin, X. (2014). Rare-Variant Association Analysis: Study Designs and Statistical Tests. The American Journal of Human Genetics *95*, 5–23.

Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.

Li, X.-B., Luo, H., Huang, J., Zhang, J.-D., Yang, Z.-X., and Sun, X.-W. (2014). XRCC2 gene polymorphisms and its protein are associated with colorectal cancer susceptibility in Chinese Han population. Med. Oncol. *31*, 245.

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. N. Engl. J. Med. *343*, 78–85.

Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: A lightweight database of human nonsynonymous SNPs and their functional predictions. Human Mutation *32*, 894–899.

Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Human Mutation *37*, 235–241.

Lohmann, D.R. (1999). RB1 gene mutations in retinoblastoma. Hum. Mutat. *14*, 283–288.

Lonsdale, J., Thomas, J., Salvatore, M., Phillips, R., Lo, E., Shad, S., Hasz, R., Walters, G., Garcia, F., Young, N., et al. (2013). The Genotype-Tissue Expression (GTEx) project. Nature Genetics *45*, 580–585.

MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al. (2012). A Systematic Survey of Loss-of-Function Variants in Human Protein-Coding Genes. Science *335*, 823–828.

Maley, C.C., Galipeau, P.C., Li, X., Sanchez, C.A., Paulson, T.G., and Reid, B.J. (2004). Selectively advantageous mutations and hitchhikers in neoplasms: p16 lesions are selected in Barrett's esophagus. Cancer Res. *64*, 3414–3427.

Malkin, D., Li, F.P., Strong, L.C., Fraumeni, J.F., Nelson, C.E., Kim, D.H., Kassel, J., Gryka, M.A., Bischoff, F.Z., and Tainsky, M.A. (1990). Germ line p53 mutations in a familial syndrome of breast cancer, sarcomas, and other neoplasms. Science *250*, 1233–1238.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., et al. (2009). Finding the missing heritability of complex diseases. Nature *461*, 747–753.

Meijers-Heijboer, H., van den Ouweland, A., Klijn, J., Wasielewski, M., de Snoo, A., Oldenburg, R., Hollestelle, A., Houben, M., Crepin, E., van Veghel-Plandsoen, M., et al. (2002). Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of BRCA1 or BRCA2 mutations. Nat. Genet. *31*, 55–59.

Mensenkamp, A.R., Vogelaar, I.P., van Zelst-Stams, W.A.G., Goossens, M., Ouchene, H., Hendriks-Cornelissen, S.J.B., Kwint, M.P., Hoogerbrugge, N., Nagtegaal, I.D., and Ligtenberg, M.J.L. (2014). Somatic mutations in MLH1 and MSH2 are a frequent cause of mismatch-repair deficiency in Lynch syndrome-like tumors. Gastroenterology *146*, 643-646.e8.

Morak, M., Massdorf, T., Sykora, H., Kerscher, M., and Holinski-Feder, E. (2011). First evidence for digenic inheritance in hereditary colorectal cancer by mutations in the base excision repair genes. Eur. J. Cancer *47*, 1046–1055.

Moynahan, M.E., and Jasin, M. (2010). Mitotic homologous recombination maintains genomic stability and suppresses tumorigenesis. Nat Rev Mol Cell Biol *11*, 196–207.

Na, R., Wu, Y., Jiang, G., Yu, H., Lin, X., Wang, M., Conran, C.A., Fantus, R.J., Zhang, N., Liu, S., et al. (2018). Germline mutations in DNA repair genes are associated with bladder cancer risk and unfavourable prognosis. BJU Int.

Nalepa, G., and Clapp, D.W. (2018). Fanconi anaemia and cancer: an intricate relationship. Nat. Rev. Cancer *18*, 168–185.

Narod, S.A., and Amos, C. (1990). Estimating the power of linkage analysis in hereditary breast cancer. Am. J. Hum. Genet. *46*, 266–272.

Nicolas, G., Charbonnier, C., Wallon, D., Quenez, O., Bellenguez, C., Grenier-Boley, B., Rousseau, S., Richard, A.-C., Rovelet-Lecrux, A., Le Guennec, K., et al. (2016). SORL1 rare variants: a major risk factor for familial early-onset Alzheimer's disease. Molecular Psychiatry *21*, 831–836.

Nieminen, T.T., O'Donohue, M.-F., Wu, Y., Lohi, H., Scherer, S.W., Paterson, A.D., Ellonen, P., Abdel-Rahman, W.M., Valo, S., Mecklin, J.-P., et al. (2014). Germline mutation of RPS20, encoding a ribosomal protein, causes predisposition to hereditary nonpolyposis colorectal carcinoma without DNA mismatch repair deficiency. Gastroenterology *147*, 595-598.e5.

Novellasdemunt, L., Antas, P., and Li, V.S.W. (2015). Targeting Wnt signaling in colorectal cancer. A Review in the Theme: Cell Signaling: Proteins, Pathways and Mechanisms. Am. J. Physiol., Cell Physiol. *309*, C511-521.

Nowell, P.C. (2002). Tumor progression: a brief historical perspective. Semin. Cancer Biol. *12*, 261–266.

Palles, C., Cazier, J.-B., Howarth, K.M., Domingo, E., Jones, A.M., Broderick, P., Kemp, Z., Spain, S.L., Guarino, E., Guarino Almeida, E., et al. (2013). Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. Nat. Genet. *45*, 136–144.

Park, J.-Y., Virts, E.L., Jankowska, A., Wiek, C., Othman, M., Chakraborty, S.C., Vance, G.H., Alkuraya, F.S., Hanenberg, H., and Andreassen, P.R. (2016). Complementation of hypersensitivity to DNA interstrand crosslinking agents demonstrates that XRCC2 is a Fanconi anaemia gene. J. Med. Genet. *53*, 672–680.

Peltomäki, P. (2016). Update on Lynch syndrome genomics. Fam Cancer *15*, 385–393.

Pertea, M., Lin, X., and Salzberg, S.L. (2001). GeneSplicer: a new computational method for splice site prediction. Nucleic Acids Res. *29*, 1185–1190.

Peters, U., Jiao, S., Schumacher, F.R., Hutter, C.M., Aragaki, A.K., Baron, J.A., Berndt, S.I., Bézieau, S., Brenner, H., Butterbach, K., et al. (2013). Identification of Genetic Susceptibility Loci for Colorectal Tumors in a Genome-Wide Meta-analysis. Gastroenterology *144*, 799-807.e24.

Peterson, C.L., and Côté, J. (2004). Cellular machineries for chromosomal DNA repair. Genes Dev. *18*, 602–616.

Petrucelli, N., Daly, M.B., and Feldman, G.L. (2010). Hereditary breast and ovarian cancer due to mutations in BRCA1 and BRCA2. Genet. Med. *12*, 245–259.

Phelan, C.M., Iqbal, J., Lynch, H.T., Lubinski, J., Gronwald, J., Moller, P., Ghadirian, P., Foulkes, W.D., Armel, S., Eisen, A., et al. (2014). Incidence of colorectal cancer in BRCA1 and BRCA2 mutation carriers: results from a follow-up study. Br. J. Cancer *110*, 530–534.

Plotz, G., Casper, M., Raedle, J., Hinrichsen, I., Heckel, V., Brieger, A., Trojan, J., and Zeuzem, S. (2012). MUTYH gene expression and alternative splicing in controls and polyposis patients. Human Mutation *33*, 1067–1074.

Rahman, N. (2014). Realizing the promise of cancer predisposition genes. Nature *505*, 302–308.

Reese, M.G., Eeckman, F.H., Kulp, D., and Haussler, D. (1997). Improved Splice Site Detection in Genie. Journal of Computational Biology *4*, 311–323.

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet. Med. *17*, 405–424.

Rivera, B., Castellsagué, E., Bah, I., van Kempen, L.C., and Foulkes, W.D. (2015). Biallelic NTHL1 Mutations in a Woman with Multiple Primary Tumors.

Rodriguez, J.M., Rodriguez-Rivas, J., Di Domenico, T., Vázquez, J., Valencia, A., and Tress, M.L. (2018). APPRIS 2017: principal isoforms for multiple gene sets. Nucleic Acids Research *46*, D213–D217.

Romei, C., Ciampi, R., and Elisei, R. (2016). A comprehensive overview of the role of the *RET* proto-oncogene in thyroid carcinoma. Nature Reviews Endocrinology *12*, 192–202.

Ruark, E., Snape, K., Humburg, P., Loveday, C., Bajrami, I., Brough, R., Rodrigues, D.N., Renwick, A., Seal, S., Ramsay, E., et al. (2013). Mosaic PPM1D mutations are associated with predisposition to breast and ovarian cancer. Nature *493*, 406–410.

Sampson, J.R., and Jones, N. (2009). MUTYH-associated polyposis. Best Pract Res Clin Gastroenterol *23*, 209–218.

Sanger, F., and Coulson, A.R. (1975). A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. Journal of Molecular Biology *94*, 441–448.

Sanger, F., Nicklen, S., and Coulson, A.R. (1977). DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences *74*, 5463–5467.

Sangkhathat, S., Maneechay, W., Chaiyapan, W., Kanngern, S., and Boonpipattanapong, T. (2015). Association of Wilms' tumor 1 gene single-nucleotide polymorphism rs16754 with colorectal cancer. Mol Clin Oncol *3*, 1401–1405.

Santarosa, M., and Ashworth, A. (2004). Haploinsufficiency for tumour suppressor genes: when you don't need to go all the way. Biochim. Biophys. Acta *1654*, 105–122.

Schwarz, J.M., Cooper, D.N., Schuelke, M., and Seelow, D. (2014). MutationTaster2: mutation prediction for the deep-sequencing age. Nature Methods *11*, 361–362.

Seal, S., Thompson, D., Renwick, A., Elliott, A., Kelly, P., Barfoot, R., Chagtai, T., Jayatilake, H., Ahmed, M., Spanova, K., et al. (2006). Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. Nat. Genet. *38*, 1239–1241.

Seevaratnam, R., Coburn, N., Cardoso, R., Dixon, M., Bocicariu, A., and Helyer, L. (2012). A systematic review of the indications for genetic testing and prophylactic gastrectomy among patients with hereditary diffuse gastric cancer. Gastric Cancer *15 Suppl 1*, S153-163.

Seguí, N., Mina, L.B., Lázaro, C., Sanz-Pamplona, R., Pons, T., Navarro, M., Bellido, F., López-Doriga, A., Valdés-Mas, R., Pineda, M., et al. (2015). Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. Gastroenterology *149*, 563–566.

Shafirovich, V., and Geacintov, N.E. (2017). Removal of oxidatively generated DNA damage by overlapping repair pathways. Free Radical Biology and Medicine *107*, 53–61.

Shamseldin, H.E., Elfaki, M., and Alkuraya, F.S. (2012). Exome sequencing reveals a novel Fanconi group defined by XRCC2 mutation. J. Med. Genet. *49*, 184–186.

Shandilya, J., and Roberts, S.G.E. (2015). A role of WT1 in cell division and genomic stability. Cell Cycle *14*, 1358–1364.

Shandilya, J., Toska, E., Richard, D.J., Medler, K.F., and Roberts, S.G. (2014). WT1 Interacts with MAD2 and Regulates Mitotic Checkpoint Function. Nat Commun *5*, 4903.

Shaw, P.A., and Clarke, B.A. (2016). Prophylactic Gynecologic Specimens from Hereditary Cancer Carriers. Surg Pathol Clin *9*, 307–328.

Shi, D., Liang, L., Zheng, H., Cai, G., Li, X., Xu, Y., and Cai, S. (2017). Silencing of long non-coding RNA SBDSP1 suppresses tumor growth and invasion in colorectal cancer. Biomedicine & Pharmacotherapy *85*, 355–361.

Shiovitz, S., Copeland, W.K., Passarelli, M.N., Burnett-Hartman, A.N., Grady, W.M., Potter, J.D., Gallinger, S., Buchanan, D.D., Rosty, C., Win, A.K., et al. (2014). Characterisation of Familial Colorectal Cancer Type X, Lynch syndrome, and non-familial colorectal cancer. Br J Cancer *111*, 598–602.

Sijbers, A.M., de Laat, W.L., Ariza, R.R., Biggerstaff, M., Wei, Y.F., Moggs, J.G., Carter, K.C., Shell, B.K., Evans, E., de Jong, M.C., et al. (1996). Xeroderma pigmentosum group F caused by a defect in a structure-specific DNA repair endonuclease. Cell *86*, 811–822.

Sim, N.-L., Kumar, P., Hu, J., Henikoff, S., Schneider, G., and Ng, P.C. (2012). SIFT web server: predicting effects of amino acid substitutions on proteins. Nucleic Acids Research *40*, W452–W457.

Slavin, T.P., Maxwell, K.N., Lilyquist, J., Vijai, J., Neuhausen, S.L., Hart, S.N., Ravichandran, V., Thomas, T., Maria, A., Villano, D., et al. (2017). The contribution of pathogenic variants in breast cancer susceptibility genes to familial breast cancer risk. NPJ Breast Cancer *3*, 22.

Smith, C.G., West, H., Harris, R., Idziaszczyk, S., Maughan, T.S., Kaplan, R., Richman, S., Quirke, P., Seymour, M., Moskvina, V., et al. (2013). Role of the oxidative DNA damage repair gene OGG1 in colorectal tumorigenesis. J. Natl. Cancer Inst. *105*, 1249–1253.

Sopik, V., Phelan, C., Cybulski, C., and Narod, S.A. (2015). BRCA1 and BRCA2 mutations and the risk for colorectal cancer. Clin. Genet. *87*, 411–418.

Spier, I., Holzapfel, S., Altmüller, J., Zhao, B., Horpaopan, S., Vogt, S., Chen, S., Morak, M., Raeder, S., Kayser, K., et al. (2015). Frequency and phenotypic spectrum of germline mutations in POLE and seven other polymerase genes in 266 patients with colorectal adenomas and carcinomas. Int. J. Cancer *137*, 320–331.

Stoffel, E.M., and Boland, C.R. (2015). Genetics and Genetic Testing in Hereditary Colorectal Cancer. Gastroenterology *149*, 1191-1203.e2.

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

Svilar, D., Goellner, E.M., Almeida, K.H., and Sobol, R.W. (2011). Base Excision Repair and Lesion-Dependent Subpathways for Repair of Oxidative DNA Damage. Antioxid Redox Signal *14*, 2491–2507.

Taneichi, H., Kanegane, H., Futatani, T., Otsubo, K., Nomura, K., Sato, Y., Hama, A., Kojima, S., Kohdera, U., Nakano, T., et al. (2006). Clinical and genetic analyses of presumed Shwachman-Diamond syndrome in Japan. Int. J. Hematol. *84*, 60–62.

Tao, H., Shinmura, K., Hanaoka, T., Natsukawa, S., Shaura, K., Koizumi, Y., Kasuga, Y., Ozawa, T., Tsujinaka, T., Li, Z., et al. (2004). A novel splice-site variant of the base excision repair gene MYH is associated with production of an aberrant mRNA transcript encoding a truncated MYH protein not localized in the nucleus. Carcinogenesis *25*, 1859–1866.

Taylor, M.D., Liu, L., Raffel, C., Hui, C., Mainprize, T.G., Zhang, X., Agatep, R., Chiappa, S., Gao, L., Lowrance, A., et al. (2002). Mutations in SUFU predispose to medulloblastoma. Nat. Genet. *31*, 306–310.

Vasen, H.F.A., Möslein, G., Alonso, A., Aretz, S., Bernstein, I., Bertario, L., Blanco, I., Bülow, S., Burn, J., Capella, G., et al. (2008). Guidelines for the clinical management of familial adenomatous polyposis (FAP). Gut *57*, 704–713.

Vasen, H.F.A., Tomlinson, I., and Castells, A. (2015). Clinical management of hereditary colorectal cancer syndromes. Nature Reviews Gastroenterology & Hepatology *12*, 88–97.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

de Voer, R.M., Hahn, M.-M., Mensenkamp, A.R., Hoischen, A., Gilissen, C., Henkes, A., Spruijt, L., van Zelst-Stams, W.A., Marleen Kets, C., Verwiel, E.T., et al. (2015). Deleterious Germline BLM Mutations and the Risk for Early-onset Colorectal Cancer. Scientific Reports *5*.

Vogelstein, B., and Kinzler, K.W. (2004). Cancer genes and the pathways they control. Nature Medicine *10*, 789–799.

Vogelstein, B., Papadopoulos, N., Velculescu, V.E., Zhou, S., Diaz, L.A., and Kinzler, K.W. (2013). Cancer genome landscapes. Science *339*, 1546–1558.

Vogt, S., Jones, N., Christian, D., Engel, C., Nielsen, M., Kaufmann, A., Steinke, V., Vasen, H.F., Propping, P., Sampson, J.R., et al. (2009). Expanded extracolonic tumor spectrum in MUTYH-associated polyposis. Gastroenterology *137*, 1976-1985.e1-10.

Wang, Q. (2016). Cancer predisposition genes: molecular mechanisms and clinical impact on personalized cancer care: examples of Lynch and HBOC syndromes. Acta Pharmacol. Sin. *37*, 143–149.

Wang, Q., Lu, Q., and Zhao, H. (2015). A review of study designs and statistical methods for genomic epidemiology studies using next generation sequencing. Frontiers in Genetics *6*.

Warthin, A.S. (1985). Classics in oncology. Heredity with reference to carcinoma as shown by the study of the cases examined in the pathological laboratory of the University of Michigan, 1895-1913. CA Cancer J Clin *35*, 348–359.

Watson, J.D. (1990). The human genome project: past, present, and future. Science *248*, 44–50.

Weren, R.D., Ligtenberg, M.J., Geurts van Kessel, A., De Voer, R.M., Hoogerbrugge, N., and Kuiper, R.P. (2018). NTHL1 and MUTYH polyposis syndromes: two sides of the same coin? J. Pathol. *244*, 135–142.

Weren, R.D.A., Ligtenberg, M.J.L., Kets, C.M., de Voer, R.M., Verwiel, E.T.P., Spruijt, L., van Zelst-Stams, W.A.G., Jongmans, M.C., Gilissen, C., Hehir-Kwa, J.Y., et al. (2015). A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. Nat. Genet. *47*, 668–671.

Win, A.K., Hopper, J.L., and Jenkins, M.A. (2011a). Association between monoallelic MUTYH mutation and colorectal cancer risk: a meta-regression analysis. Fam Cancer *10*, 1–9.

Win, A.K., Cleary, S.P., Dowty, J.G., Baron, J.A., Young, J.P., Buchanan, D.D., Southey, M.C., Burnett, T., Parfrey, P.S., Green, R.C., et al. (2011b). Cancer risks for monoallelic MUTYH mutation carriers with a family history of colorectal cancer. International Journal of Cancer *129*, 2256–2262.

Wu, M.C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test. The American Journal of Human Genetics *89*, 82–93.

Yan, H.H.N., Lai, J.C.W., Ho, S.L., Leung, W.K., Law, W.L., Lee, J.F.Y., Chan, A.K.W., Tsui, W.Y., Chan, A.S.Y., Lee, B.C.H., et al. (2016). RNF43 germline and somatic mutation in serrated neoplasia pathway and its association with BRAF mutation. Gut.

Yeo, G., and Burge, C.B. (2004). Maximum Entropy Modeling of Short Sequence Motifs with Applications to RNA Splicing Signals. Journal of Computational Biology *11*, 377–394.

Yurgelun, M.B., Allen, B., Kaldate, R.R., Bowles, K.R., Judkins, T., Kaushik, P., Roa, B.B., Wenstrup, R.J., Hartman, A.-R., and Syngal, S. (2015). Identification of a Variety of Mutations in Cancer Predisposition Genes in Patients With Suspected Lynch Syndrome. Gastroenterology *149*, 604-613.e20.

Zhang, M.Q. (1998). Statistical features of human exons and their flanking regions. Hum. Mol. Genet. *7*, 919–932.

Zhou, W., Otto, E.A., Cluckey, A., Airik, R., Hurd, T.W., Chaki, M., Diaz, K., Lach, F.P., Bennett, G.R., Gee, H.Y., et al. (2012). FAN1 mutations cause karyomegalic interstitial nephritis, linking chronic kidney failure to defective DNA damage repair. Nat. Genet. *44*, 910–915.

7. SERMENT DE GALIEN

SERMENT DE GALIEN

Je jure d'honorer ceux qui m'ont instruit dans les préceptes de mon art et de leur témoigner ma reconnaissance en restant fidèle à leur enseignement.

D'exercer dans l'intérêt de la Santé publique ma profession avec conscience et de respecter non seulement la législation en vigueur mais aussi les règles de l'Honneur, de la Probité et du Désintéressement.

De ne jamais oublier ma responsabilité et mes devoirs envers le malade et sa dignité humaine.

De ne dévoiler à personne les secrets qui m'auraient été confiés ou dont j'aurais eu connaissance dans l'exercice de ma Profession.

En aucun cas, je ne consentirai à utiliser mes connaissances et mon état pour corrompre les mœurs et favoriser les actes criminels.

Que les hommes m'accordent leur estime si je suis fidèle à mes promesses.

Que je sois méprisé de mes Confrères si je manque à mes engagements.

BLANLUET Maud

Etude de l'association de variants génétiques rares aux formes précoces de cancer colorectal

Th. D. Pharm., Rouen, 2018, 116 p.

RESUME

L'analyse des gènes responsables de formes mendéliennes de CCR reste parfois négative chez des patients présentant un Cancer ColoRectal (CCR) de survenue très précoce, situation clinique pourtant suggestive d'un déterminisme génétique fort. Plusieurs études ont suggéré la contribution de variants rares affectant des gènes de réparation de l'ADN au déterminisme du CCR mais le manque de significativité statistique à l'échelle de l'exome limite l'interprétation de ces résultats. Afin de déterminer la contribution des variations génétiques rares affectant la réparation de l'ADN au CCR, nous avons séquencé par *Next Generation Sequencing* (NGS) un panel de 201 gènes impliqués dans le cancer chez 84 patients atteints de CCR très précoce d'origine française rigoureusement sélectionnés (CCR diagnostiqué jusqu'à l'âge de 32 ans, exclusion du syndrome de Lynch et des polyposes adénomateuses et hamartomateuses). Nous avons également séquencé l'exome de 93 patients, sélectionnés selon les mêmes critères, et de 567 témoins FREX (*French Exome Project*). Nous avons filtré les variants rares prédits pathogènes, affectant les 201 gènes du cancer ou les 296 gènes de réparation de l'ADN selon l'analyse réalisée, et nous avons comparé leurs fréquences relatives entre les cas et les témoins. A l'aide d'un test statistique de charge en variants (*Burden test SeqMeta),* nous avons obtenu un *Odds Ratio* (OR) de 1,412 (p = 0,165) pour les variations *Loss of Function* (LOF) rares (*Minor Allele Frequency*, MAF < 1%) affectant les gènes de la réparation de l'ADN, et un OR de 1,681 (p = 0,017) pour les variations LOF très rares (MAF < 1‰). L'effet observé est plus fort lorsque le test est restreint aux variants très rares affectant le système de réparation des liaisons inter-brins (*Inter-strand CrossLink*, ICL), avec un OR de 5,88 (p = 0,012). Fait remarquable, la comparaison de la cooccurrence d'au moins deux variations LOF très rares affectant la réparation de l'ADN entre patients et témoins donne un OR de 7,196 (p = 0,077). Ces résultats posent la question de la contribution de la voie ICL au CCR et de la place de l'oligogénisme, basé sur l'effet cumulatif de plusieurs variations affectant en particulier la réparation de l'ADN, dans la prédisposition au CCR.

MOTS CLES

Oncogénétique – Cancer colorectal – Réparation de l'ADN – Etude d'association – Oligogénisme

---

JURY

Président du jury, co-directeur de thèse :	Monsieur le Professeur Loïc Favennec
Directeur de thèse :	Monsieur le Professeur Thierry Frebourg
Membres du jury :	Madame le Docteur Stéphanie Baert-Desurmont
	Monsieur le Professeur Pierre Michel

DATE DE SOUTENANCE : 02 Octobre 2018