



**HAL**  
open science

# In-silico analyses of cancer genomic instabilities and DNA repair deficiencies for diagnostics and treatment choice

Alexandre Eeckhoutte

► **To cite this version:**

Alexandre Eeckhoutte. In-silico analyses of cancer genomic instabilities and DNA repair deficiencies for diagnostics and treatment choice. Cancer. Université Paris Cité, 2021. English. NNT : 2021UNIP7361 . tel-04167003

**HAL Id: tel-04167003**

**<https://theses.hal.science/tel-04167003>**

Submitted on 20 Jul 2023

**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.



Université de Paris

**Ecole doctorale Hématologie Oncogénèse et Biothérapies n°561**

***Equipe Mélanome Uvéal et Réparation de l'ADN  
Institut Curie, INSERM U830, Paris***

# **In-silico analyses of cancer genomic instabilities and DNA repair deficiencies for diagnostics and treatment choice**

Par **Alexandre Eeckhoutte**

Thèse de doctorat de Bioinformatique

Dirigée par Tatiana Popova

Et par Marc-Henri Stern

Présentée et soutenue publiquement le 16/12/2021

Devant un jury composé de :

**Frédérique Penault-Llorca – Rapporteur** : Professeur des Universités-Praticien Hospitalier, INSERM 1240, Université Clermont-Auvergne

**Daniel Gautheret – Rapporteur** : Professeur des Universités, Université Paris-Saclay

**Eric Letouzé – Examineur** : Directeur de Recherche, Institut de Recherche en Cancérologie Nantes-Atlantique

**Ludmila Danilova – Examineur** : PhD, Research associate, Johns Hopkins School of Medicine

## Abstract

### **In-silico analyses of cancer genomic instabilities and DNA repair deficiencies for diagnostics and treatment choice**

This work is devoted to the analysis of cancer genomic instability and DNA repair deficiency as a part of clinical investigation and the implementation of a diagnostic tool.

The first part of the thesis describes Homologous Recombination Deficiency (HRD) and the genomic aberration patterns in tumors which emerge upon this deficiency. Homologous Recombination (HR) is a DNA repair pathway that can fix double strand breaks by using a sister chromatid and sister chromosome as a DNA template. Tumors inactivated for some major genes involved in HR such as *BRCA1*, *BRCA2*, *RAD51C* and *PALB2* were shown to present high genomic instability. Detecting inactivation of HR pathway in tumors nowadays became extremely important because of new therapeutic options targeting HRD (PARP inhibitors) and established clinical practice of familial genetic supervision. The comprehensive testing for HR gene inactivation is complicated because of the diversity of inactivation mechanisms including germline and somatic mutations, promoter methylation and structural variants. Genomic signatures based on the specific aberration profiles of HRD tumors have been developed. These signatures capturing tumor BRCAness or genomic HRD covering all possible causes are described. A new tool called shallowHRD to evaluate HRD in tumors is presented. ShallowHRD is based on low-coverage Whole Genome Sequencing (<1X). This DNA sequencing technique was chosen because of a low-cost and efficiency that allows constructing the copy-number alteration profile of a tumor even for FFPE samples. ShallowHRD exploits large-scale intra-chromosomal breaks in copy number profiles characteristic of HRD. ShallowHRD shows ~90% of sensitivity and specificity, which is comparable to most state-of-the-art methods. Due to the intrinsic advantages of shallowHRD it is already implemented for ovarian cancer in Institut Curie. The new incoming cases and the large retrospective cohorts analyzed using shallowHRD allowed to refine HRD and quality diagnostics for clinical precision.

The second part describes the background and the results of testing *CDK12* (cyclin-dependent kinase 12) as a cancer predisposition gene in epithelial ovarian cancer (EOC). *CDK12* is a tumor suppressor gene, which inactivation was consistently associated with specific genomic instability in EOC and prostate cancers. *CDK12* is an RNA processing protein not directly implicated in DNA repair and exact functional link between genomic instability and *CDK12* inactivation is not yet deciphered. The investigation of a cohort of unrelated 416 patients with EOC for *CDK12* revealed no germline deleterious mutation and a proportion of missense

mutations similar to the representative non-Finnish European population. Overall, no evidence was found to indicate that *CDK12* is a cancer predisposition gene.

The third part describes a background and the results of testing *MBD4* (methyl binding domain 4) as a cancer predisposition gene for Uveal Melanoma (UM) by massive-parallel sequencing. *MBD4* is a DNA-glycosylase that helps to maintain genomic stability principally by repressing mCpG > TpG mutations and its inactivation in tumors was consistently associated with high rate of CpG > TpG mutations and marked the response to immunotherapy. The results of screening of the cohort of 1093 germline DNA of unrelated UM patient and 193 UM tumors for *MBD4* mutation are presented. *MBD4* germline inactivating mutations were found in 0.7% of UM cases and *MBD4* is proved to be an UM predisposing gene with a 10-fold relative risk. The study concluded that *MBD4* must be included into cancer gene panel for clinical testing and further explored in UM and other diseases.

## Résumé

### **Analyses in-silico d'instabilités génomiques et de déficiences de réparation de l'ADN dans les cancers pour le diagnostic et le choix de traitement**

Ce travail est consacré à l'analyse d'instabilités génomiques et de déficiences de réparation de l'ADN dans les cancers dans un contexte clinique et de la mise en place d'un outil de diagnostic.

La première partie de la thèse décrit le déficit en recombinaison homologue (HRD) et les motifs d'aberration génomique dans les tumeurs qui émergent lors de ce déficit. Les cassures doubles brins de l'ADN sont des lésions très toxiques qui nécessitent d'être réparées par les cellules. La recombinaison homologue (HR) est l'une des principales voies de réparation de l'ADN permettant la réparation fidèle de ces cassures double brins en utilisant une chromatide sœur comme matrice durant la phase S à G2 du cycle cellulaire ou un chromosome homologue lors de la méiose. Les tumeurs inactivées pour certains gènes majeurs impliqués dans HR tels que *BRCA1*, *BRCA2*, *RAD51C* et *PALB2* se sont avérées présenter une instabilité génomique élevée. La détection de l'inactivation de la voie HR dans les tumeurs est aujourd'hui devenue extrêmement importante en raison de nouvelles options thérapeutiques ciblant l'HRD (inhibiteurs de PARP et les chimiothérapies par sels de platine) et des pratiques cliniques pour la surveillance génétique familiale. Les tests complets d'inactivation des gènes HR sont cependant compliqués en raison de la diversité des mécanismes d'inactivation, y compris les mutations germinales et somatiques, la méthylation du promoteur pour *BRCA1* et *RAD51C* ainsi que les variants structurelles.

Des signatures génomiques basées sur les profils d'aberrations spécifiques des tumeurs HRD ont été développées. Trois signatures génomiques structurelles de 1<sup>ère</sup> génération ont été identifiées en 2012, se basant respectivement sur des réarrangements de grandes tailles (LST), sur un déséquilibre allélique au niveau des régions télomériques (TAI) et sur un excès des grandes régions avec perte d'hétérozygotie (LOH) ne parcourant pas tout le chromosome. Le test Myriad myChoice® CDx combine ces trois signatures et est actuellement le seul test disponible commercialement permettant la prescription d'inhibiteurs de PARP dans des tumeurs sauvages pour les gènes *BRCA1* et *BRCA2*. Sa mise en place est cependant onéreuse, environ 3500 euros. D'autres signatures de seconde génération ont également été développée grâce à l'émergence des nouvelles technologies de séquençage. Celles-ci comprennent des signatures de remaniements structuraux (RS3 et RS5), mais également des signatures mutationnelles avec la signature 3 représentant une distribution quasi-uniforme des 96 types de substitutions possibles ainsi que des signatures de petites délétions avec

microhomologie à la jonction (ID6 et ID8). Deux classificateurs, HRDetect et CHORD prennent en compte l'ensemble de ces signatures et présentent les meilleures performances actuellement pour prédire l'HRD. Le coût de séquençage et de stockage des données ainsi que la complexité de ces analyses sont les principaux facteurs limitants pour leur implémentation en routine pour la clinique.

Dans cette thèse un nouvel outil pour évaluer l'HRD dans les tumeurs, *shallowHRD*, est présenté. *ShallowHRD* est basée sur le séquençage du génome entier à faible couverture (sWGS ; <1X). Cette technique de séquençage d'ADN a été choisie en raison de son faible coût et de son efficacité qui permet de construire le profil d'altération du nombre de copies d'une tumeur, même pour les échantillons FFPE. *ShallowHRD* détecte automatiquement un niveau minimum entre deux segments génomiques afin d'optimiser le profil de nombre de copies de la tumeur et exploiter des ruptures intrachromosomiques de grandes échelles relatives dans les profils caractéristiques de l'HRD. *ShallowHRD* montre ~ 90% de sensibilité et de spécificité sur les échantillons de tumeur du sein du projet The Cancer Genome Atlas (TCGA), ce qui est comparable à la plupart des méthodes de pointes. L'accumulation d'un grand nombre de cas a permis d'améliorer *shallowHRD* afin de s'approcher encore davantage de la précision nécessaire en clinique, en rendant plus robuste la détection d'un niveau minimum de différence entre deux niveaux copies et en assurant une meilleure optimisation du profil génomique. Des critères de qualités ont également été développés pour aider au diagnostic. La comparaison des résultats entre *shallowHRD* et le test approuvé cliniquement Myriad myChoice® CDx montre une importante correspondance entre les deux méthodes de 92,5%. En outre, dans 54 xénogreffes issues de tumeurs (PDX) du sein triple-négatives, le statut de la recombinaison homologe prédit significativement la réponse au cisplatine - 70,96% des PDX HRD ont une maladie stable ou une réponse aux sels de platine pour 27,27% pour les PDX compétents en HR. En raison des avantages intrinsèques de *shallowHRD*, notre méthode est déjà mise en œuvre pour le cancer de l'ovaire à l'Institut Curie, généralement dans le même séquençage qu'un panel de gène d'intérêt développé à l'Institut Curie, et peut permettre d'aider à améliorer le diagnostic pour une meilleure prise en charge de patient en routine.

La deuxième partie décrit le contexte et les résultats de l'investigation de *CDK12* (cyclin-dependent kinase 12) en tant que gène de prédisposition au cancer épithélial de l'ovaire (EOC). L'EOC est associé à des facteurs de risques connus notamment l'âge et la génétique d'un individu. Les gènes *BRCA1* et *BRCA2* expliquent une large majorité des cancers des EOC héréditaires, avec des mutations dans d'autres gènes de l'HR. De plus des gènes impliqués dans la réparation de l'ADN, comme ceux de la famille MMR (*MLH1*, *MSH2*, *MSH6*, *PMS2*), expliquent une plus petite proportion des cas héréditaires. Néanmoins certains de ces

cas ne sont toujours pas expliqués. *CDK12* est un gène suppresseur de tumeur maintenant la stabilité génomique, régulant la transcription de plusieurs gènes de réparation de l'ADN et faisant partie des dix gènes les plus mutés dans les cancers sévères de haut-grade de l'ovaire. Il correspond donc à un gène de prédisposition possible pour l'EOC. Son inactivation est systématiquement associée à une instabilité génomique spécifique caractérisée par de nombreuses duplications en tandem entre 0.3 et 3Mb dans les EOC et cancer de la prostate.

Pour répondre si *CDK12* est un gène de prédisposition à l'EOC, une étude d'une cohorte de 416 ADN germinale de patients non apparentés atteints d'EOC a été mise en place. Toutes les régions codantes de *CDK12* ont été étudiées par séquençage massif parallèle dans des mélanges équimolaires de 8 ADN germinales. Cette approche permet de diminuer les coûts de séquençage en regardant les mutations dans *CDK12* de plusieurs échantillons dans un seul séquençage. L'expérience a été construite pour que la couverture de séquençage soit au minimum de 320X par base, soit 20X par allèle. L'appel de mutations a été développé pour être sensible et ces dernières ont été conservées si elles étaient appelées au moins par un des trois logiciels d'appel de variants sélectionnés ou script personnel. Ces mutations ont ensuite été validées ou invalidées via un séquençage Sanger. Cette étude n'a montré aucune mutation délétère germinale de *CDK12* au sein de notre cohorte. De plus une proportion de mutations faux-sens similaire à la population représentative européenne non finlandaise a été retrouvée dans notre cohorte (Fisher's exact test :  $p$  value = 0.1453). Les variants faux-sens ont été investigués pour les tumeurs disponibles au sein de l'Institut Curie pour la perte de l'allèle sauvage. Des quatre tumeurs disponibles, une seule présentait une perte de l'allèle sauvage. De plus, aucune de ces tumeurs ne présentait le profil génomique caractéristique de l'inactivation de *CDK12*. En outre, l'exploration du TCGA dans 511 cas de cancer de l'ovaire n'a pas mis en évidence la présence de mutations délétères germinales. Dans une cohorte de cancers de la prostate résistant à la castration, *CDK12* a été trouvé inactivé dans 7% des cas, sans qu'aucune mutation germinale délétère ne soit trouvée. Dans l'ensemble, aucune preuve n'indique que *CDK12* est un gène de prédisposition au cancer de l'ovaire. Le nombre de cas dans notre cohorte est cependant insuffisant pour conclure complètement dans ce sens.

La troisième partie décrit le contexte et l'investigation de *MBD4* (Methyl-CpG Binding Domain 4) en tant que gène de prédisposition au cancer pour le mélanome uveal (UM). *MBD4* est une ADN glycosylase qui aide à maintenir la stabilité génomique principalement en réprimant les mutations mCpG>TpG. En 2018 notre laboratoire a étudié une patiente UM métastatique montrant une réponse exceptionnelle à l'immunothérapie avec des anticorps anti-PD1 (*Program cell death 1*). Ce patient présentait un taux de mutation tumorale très important avec un phénotype CpG > TpG. Conformément aux fonctions de *MBD4* et son phénotype mutationnel observé, une mutation germinale délétère dans le gène *MBD4* avec une

inactivation somatique du gène a été trouvée dans ce patient. L'investigation du TCGA pour le phénotype mutationnel CpG > TpG a permis de trouver deux tumeurs supplémentaires, une d'UM et une de glioblastome, les deux portant une mutation germinale dans *MBD4* et une inactivation somatique du gène.

Une large cohorte de 1093 ADN germinaux de patients UM non apparentés et de 193 tumeurs UM a été constituée afin d'étudier si *MBD4* est un gène de prédisposition à l'UM. L'approche développée est similaire à celle employée pour l'investigation de *CDK12* en tant que gène de prédisposition. Des mélanges équimolaires de 8 ADN germinaux et de 3 ou 4 ADN tumorales ont été constitués. Les variants trouvés par au moins un des quatre logiciels d'appel de variant ont été ensuite validés par séquençage en Sanger. Des mutations germinales délétères *MBD4* ont été trouvées dans 0,7% des cas d'UM et *MBD4* a été prouvé prédisposant à l'UM avec un risque relatif de 9,15. Nous avons en outre confirmé que les tumeurs inactivées pour *MBD4* possèdent toutes le même phénotype hypermutateur observé précédemment. Pas de différences marquantes sur la survie globale, la survie sans le développement de métastases ou un âge minimum plus bas que la population générale pour le développement d'UM n'a été observé entre patients compétents et déficients pour *MBD4*. De plus aucun cas familial d'UM n'arbore une mutation germinale délétère de *MBD4*. Cependant en raison du faible nombre de tumeur inactivées *MBD4* dans notre cohorte nous ne pouvons tirer des conclusions sur ces caractéristiques. L'inactivation de *MBD4* est toujours associée avec une monosomie du chromosome 3 ou une isodisomie 3. De plus, l'inactivation de *MBD4* est toujours associée à l'inactivation de gènes impliqués déjà dans l'oncogenèse classique de l'UM, avec une inactivation dans des gènes de la voie Gαq et des mutations mutuellement exclusives des gènes *BAP1*, *SF3B1* et *EIF1AX*. L'étude a conclu que *MBD4* doit être inclus dans les panels de gènes pour les tests cliniques. En outre nous avons regardé toutes les mutations germinales délétères présentes dans la littérature et avons trouvé plusieurs dans l'UM mais aussi dans d'autres maladies comme le glioblastome ou la leucémie myéloïde aiguë. Cela indique que *MBD4* doit être exploré encore davantage dans l'UM mais aussi dans d'autres maladies. En coopération avec une équipe de l'hôpital La-Pitié-Salpêtrière nous allons évaluer dans les gliomes si *MBD4* est un gène de prédisposition. Au-delà du caractère prédisposant de *MBD4*, les tumeurs inactivées pour ce gène sont distinctes avec un phénotype génomique particulier et des traitements pourraient potentiellement cibler cette spécificité, parmi lesquels l'immunothérapie. La patiente ayant répondu de manière exceptionnelle aux anticorps anti-PD1 a développé une résistance secondaire à l'immunothérapie. De nouveaux traitements dans un contexte *MBD4*<sup>-/-</sup> devraient être explorés et nous avons déjà mis en évidence certains médicaments qui montrent des effets prometteurs dans un contexte *MBD4* déficient.



## Remerciements

Je tiens tout d'abord à remercier mes directeurs de thèse, Tatiana Popova et Marc-Henri Stern. Je remercie Tatiana pour le temps passé à me suivre au quotidien, de m'avoir aidé, corrigé et m'orienté dans mes projets, lors de ces trois années de thèse. Je remercie Marc-Henri pour sa disponibilité à chacun de mes questionnements, ses conseils et l'opportunité qu'il m'a donnée de m'accueillant au sein de son équipe et me permettre de travailler sur différents projets de recherche. Sans leurs confiances et leurs encadrements rien de tout cela n'aurait été possible et je n'aurais pas pu profiter de cette incroyable expérience humaine et professionnelle.

Je remercie l'ensemble des chercheurs m'ayant fait l'honneur d'avoir accepté de faire partie des membres de mon jury. Je remercie les professeurs Frédérique Penault-Llorca et Daniel Gautheret d'avoir accepté d'être rapporteurs de ma thèse et de corriger cette dernière. Je remercie également les docteurs Eric Letouzé et Ludmila Danilova d'en être les examinateurs.

Je remercie l'ensemble des membres de la DRUM team de m'avoir accueilli. Je me sens privilégié d'avoir pu travailler au sein de cette équipe à bien des égards. Professionnellement chaque personne que j'ai pu croiser a toujours fait preuve de patience pour m'expliquer, m'apprendre, échanger et me soutenir, toujours avec bienveillance et sollicitude. Très vite vous avez également fait du laboratoire pour moi un lieu de travail agréable et stimulant duquel je retire à la fois des connaissances mais également aujourd'hui des amitiés. À Anne-Céline Derrien, je n'aurai pu rêver avoir une meilleure co-thésarde. À Alexandre Houy pour sa gentillesse et sa patience. À Manuel Rodrigues pour toutes les conversations et son sens du partage. À Thomas Chabot et Olivier Ganier pour leurs soutiens et leurs conseils. À André Bortolini pour sa disponibilité et sa passion. À Stéphane Dayot pour son accueil et son aide. À toutes les personnes qui sont passés dans cette équipe au cours de ces trois années : Agathe Garcia, Anaïs Le Ven, Thibault Verrier, Elodie Manié, Lenha Mobuchon, Manon Reverdy, Sophie Gadrat, Dorine Bellanger, Aude Tible-Sicard, Antoine Chouteau et Noémie Lefrancq.

Je remercie l'ensemble des membres de l'U830, à ceux avec qui j'ai pu longuement discuter mais également ceux que je n'ai malheureusement pas assez croisé au cours de ses années. Je remercie spécialement Fatima Mechta-Grigoriou, Olivier Delattre et Géraldine Gentric pour leur implication et participation au 4<sup>ème</sup> congrès international de l'ADELIS.

Je remercie l'ensemble des équipes de l'Institut Curie qui contribue à faire de ce centre d'oncologie un centre d'excellence, des techniciens aux médecins du département d'Oncologie Médicale et à la plateforme de séquençage. Un grand merci au département de génétique, Julien Masliah-Planchon, Romane Beaurepere, Adrien Briaux, Celine Callens, Ivan Bièche et Dominique Stoppa-Lyonnet. Nos aller-retours permanents ont grandement contribué à l'amélioration de mes résultats et pour je l'espère une meilleure prise en charge des patients. Merci à François-Clément Bidard qui m'a suivi tout au long de la thèse. Un grand merci également aux patients pour leur confiance dans l'Institut Curie sans laquelle il ne serait possible d'avancer.

Je remercie La Ligue contre le cancer d'avoir financé ma thèse pendant trois ans.

À ma famille. À ma mère et mon père qui ont toujours été là pour moi pour me supporter et m'encourager. Sachez que sans toujours l'exprimer j'ai conscience de la chance que j'ai de vous avoir. À mes grand-mères, pour leurs gentillesse et leurs amours. À ma tante et mon cousin, chaque occasion de se voir est un véritable plaisir. À ceux qui ne sont plus.

À ma seconde famille. À mes amis d'enfance, Guillaume, Alexandre, Kevin, Pierre et Jessim, qui partagent ma vie depuis bientôt 18 ans. À mes amis de classe préparatoire, Nathan, Quentin, Jean-Baptiste, Pierryves, Arnaud et Raphael. À mes amis de l'ENSAT et mes numéros dix parisien, ils se reconnaîtront en lisant ces lignes.

# Table of Contents

Table of Contents .....	1
Table of figures .....	3
Abbreviations.....	4
<b>INTRODUCTION.....</b>	<b>8</b>
<b>DNA DOUBLE-STRAND BREAKS AND HOMOLOGOUS RECOMBINATION .....</b>	<b>9</b>
Endogenous and exogenous causes of DNA double-strand breaks .....	9
Double-strand breaks detection and signaling.....	10
Different double-strand break repair pathways.....	11
Homologous recombination pathway.....	11
Replication fork protection for genomic stability.....	14
The Fanconi anemia pathway.....	15
Conclusion .....	16
<b>HOMOLOGOUS RECOMBINATION DEFICIENCY IN CANCERS.....</b>	<b>17</b>
Cancer predisposition syndromes associated with homologous recombination genes .....	17
Prevalence of HR genes inactivation in cancer.....	17
Genomic instability in BRCA1/2-/- and HRD phenotype .....	18
Treatment for tumors with homologous recombination deficiency.....	20
Large-scale rearrangements based genomic signatures of homologous recombination deficiency .....	22
Large-scale alterations signatures in Myriad myChoice® CDx test.....	25
Small-scale somatic alteration signatures of homologous recombination deficiency .....	25
Rearrangement signature of homologous recombination deficiency.....	28
Combining signatures to extensively describe HRD .....	29
RAD51-foci, a functional signature of HRD.....	31
Possible etiology of HRD signatures .....	32
Conclusion .....	33
<b>ARTICLES .....</b>	<b>35</b>
<b>Introduction: “shallowHRD: detection of homologous recombination deficiency from shallow whole genome sequencing” .....</b>	<b>35</b>
Workflow and pipeline .....	35
<b>Conclusion: “shallowHRD: detection of homologous recombination deficiency from shallow whole genome sequencing” .....</b>	<b>39</b>
Additional results of <i>shallowHRD</i> in different cohorts.....	39
<b>Introduction: “Lack of evidence for CDK12 as an ovarian cancer predisposing gene” .....</b>	<b>44</b>

The cyclin-dependent kinase 12 .....	44
CDK12 functions .....	44
CDK12 in cancers .....	46
CDK12 as a potential cancer predisposing gene for epithelial ovarian carcinoma .....	47
<b>Conclusion: “Lack of evidence for CDK12 as an ovarian cancer predisposing gene” .....</b>	<b>49</b>
<b>Introduction: “Germline MBD4 Mutations and Predisposition to Uveal Melanoma” .....</b>	<b>50</b>
<b>Conclusion: “Germline MBD4 Mutations and Predisposition to Uveal Melanoma” .....</b>	<b>51</b>
<b>CONCLUSIONS AND PERSPECTIVES.....</b>	<b>52</b>
<b>References .....</b>	<b>56</b>

## Table of figures

FIGURE 1: End resection of the double-strand break determine the repair pathway.....	12
FIGURE 2: All principal DSB repair pathway and possible occurrence of alternative repair pathways	14
FIGURE 3: Simplified model for fork protection of reversed fork upon replication fork stalling.....	15
FIGURE 4: PARPi activity and potential synthetic lethality in HRD.....	21
FIGURE 5: Number of Large-scale State Transitions (LST) in a large in-house breast cancer cohort...	23
FIGURE 6: LOH signature in Epithelial Ovarian Cancer.....	24
FIGURE 7: Cisplatin response in serous ovarian cancer according to the number of allelic imbalances extending to the telomere end .....	25
FIGURE 8: Small-scale signature associated with HRD .....	27
FIGURE 9: Rearrangement signatures in 560 Whole Genome Sequencing of breast cancers.....	29
FIGURE 10: Parameters and performance of HRDetect classifier on 560 breast cancers .....	31
FIGURE 11: Alternative DSB repair pathway leads to small indels than flanked by (micro-)homology	33
FIGURE 12: Read Depth analysis to detect Copy Number changes with Next-Generation Sequencing	36
FIGURE 13: Genomic profile from sWGS of the chromosome 7 of a breast tumor generated with a Read-Depth approach with ControlFREEC and for different size of windows .....	37
FIGURE 14: Genomic profile from tumor sWGS generated with controlFREEC.....	38
FIGURE 15: Comparison between the diagnosis of the Myriad myChoice® CDx and <i>shallowHRD</i> .....	40
FIGURE 16: Response to cisplatin in 54 PDX of TNBC .....	41
FIGURE 17: Number of LGAs for 32 sWGS of breast cancer with <i>BRCA1</i> VUS and two cases with <i>BRCA1</i> inactivation.....	42
FIGURE 18: Plan of treatment for the RadioParp study .....	43
FIGURE 19: Results of <i>shallowHRD</i> on 22 sWGS of TNBC from the RadioParp study .....	43
FIGURE 20: The inhibition of CDK12 leads to elongation defect by premature cleavage and polyadenylation in a gene-length dependent manner .....	45
FIGURE 21: CDK12 inactivated tumors display a distinct genomic profile in metastatic castration-resistant prostate cancers.....	47

# Abbreviations

53BP1: p53 binding protein

AI: allelic imbalance

Alt-EJ: Alternative End-joining

AML: acute myeloblastic leukemia

ADP: adenosine diphosphate

ATP: adenosine triphosphate

AP: apurinic/aprimidinic

ATM: Ataxia-Telangiectasia Mutated

ATR: Ataxia Telangiectasia and Rad3-related

AS: assembly

BARD1: BRCA1-associated ring domain

BLC: Basal-like breast Carcinomas

BER: Base excision repair

BIR: Break Induced Replication

BRCA1: Breast Cancer gene 1

BRCA2: Breast Cancer gene 2

BRIP1: BRCA1 Interacting Helicase 1

CCNK: cyclin K

CDK: Cyclin-dependent kinase

CDK12: Cyclin-dependent kinase 12

CDK12-TDs: CDK12 associated tandem duplications phenotype

CDK13: Cyclin-dependent kinase 13

CFS: common fragile site

CNA: copy number alteration

C-NHEJ: canonical non-homologous end-joining

CHK1: checkpoint protein 1

CHK2: Checkpoint protein 2

CI: confidence interval

CTD: carboxy-terminal-domain

CtIP: C-terminal Interaction Protein  
DDR: DNA damage response  
DMNT1: DNA Methyltransferase 1  
DNA: deoxyribonucleic acid  
DNA-PK: DNA-dependent protein kinase  
DBS: doublet-base substitution signature  
DSB: double-strand break  
DSBR: Double-Strand Break Repair  
EOC: epithelial ovarian cancer  
ER: Estrogen Receptor  
FDA: US Food and Drug Administration  
FEN1: flap endonuclease 1  
FFPE: formalin-fixed, paraffin-embedded  
GIS: genomic instability score  
HER2: human epidermal growth factor receptor 2  
HR: homologous recombination  
HRD: homologous recombination deficient  
HRP: homologous recombination proficient  
H2A: histone 2 A  
H2AX: histone 2 A X  
ICL: interstrand crosslink  
ICGC: International Cancer Genome Consortium  
ID: Insertion-Deletion signature  
indel: small insertion/deletion  
IR: Ionizing Radiation  
LGA: large genomic aberrations  
LOH: loss-of-heterozygosity  
LST: large-scale state transition  
MLH1: MutL Homolog 1  
MLH6: MutL Homolog 6  
MRN: MRE11-RAD50-NSB1  
MMR: Mismatched repair

MSI: microsatellite instability  
NER: Nucleotide Excision Repair  
NMF: non-negative matrix factorization  
OS: overall survival  
PALB2: Partner And Localizer of BRCA2  
PARP: Poly (ADP-ribose) polymerase  
PARP1: Poly (ADP-ribose) polymerase 1  
PARP2: Poly (ADP-ribose) polymerase 2  
PARPi: Poly (ADP-ribose) polymerase inhibitors  
PCAWG: Pan-Cancer Analysis Whole-Genome  
PCNA: proliferating nuclear antigen  
PDX: patient derived xenograft  
PR: progesterone receptor  
RAD50: recombinase 50  
RAD51: recombinase 51  
RAD52: recombinase 52  
RAD54: recombinase 54  
RNA: Ribonucleic acid  
ROC: Receiver operating characteristic  
ROS: reactive oxygen species  
RP: read-pair approach  
RPA: replication protein A  
RR: relative risk  
RS: rearrangement signature  
SBS: single-base substitution signature  
SCE: sister chromatid exchange  
SDSA: Synthesis-Dependent Strand Annealing  
SNP: single nucleotide polymorphism  
SNV: Single Nucleotide Variant  
SR: split-read approach  
ssDNA: single-strand DNA  
SSA: Single-Strand Annealing



SSB: single-strand break

SSL: Single-strand DNA lesion

SV: Structural Variations

TCGA: The Cancer Genome Atlas

TDG: Thymine DNA Glycosylase

TMB: tumor mutation burden

TNBC: triple-negative breast cancer

UM: Uveal Melanoma

VUS: Variant of Unknown Significance

WGD: whole genome duplication

WGS: Whole Genome Sequencing

XRCC2: X-Ray Repair Cross Complementing 2

XRCC3: X-Ray Repair Cross Complementing 3

# INTRODUCTION

Cancer is one the most important cause of human death worldwide with 10 million cancer-related death identified in 2020<sup>1</sup>. Many types of cancers exist and originate at different locations in the body, ranging from frequent cancer such as breast cancer to rare disease such as uveal melanoma. Understanding how tumors form and develop but also what characterize them is paramount for patient care as it allows to create and improve clinical diagnosis and appropriated treatment.

Cancers are by nature genetics diseases that emerge upon deoxyribonucleic acid (DNA) changes. Altered DNA represent a hallmark of cancers<sup>2</sup>. DNA is damaged several times a day inside human cell. If not handle correctly, those damages can lead to cell inactivity, cell death or tumorigenesis. In the latter case, cells undergo genetic and epigenetics changes that result in their transformation and uncontrolled multiplication. These alterations arise from genotoxic exposure, consequences of cell functioning in an abnormal situation or a defect in DNA repair pathways. Different DNA repair pathways exist that are DNA damage specific and occur at different moment of the cell cycle. DNA repair deficiency is a major cause of cancers<sup>3</sup>, which can be used as a biomarker for treatments that tackle this specificity.

DNA modifications in cancer are systematized as follows: Single Nucleotide Variant (SNV), small insertion/deletion (indel) and Structural Variation (SV) - with large indel, duplication, inversion and translocations. Whole genome duplication (WGD) with the multiplication of the entire set of chromosomes is a frequent event in cancers. Across and even within cancers types the burden of somatic mutations is highly variable, ranging from 0.001 to 400 per megabase (Mb)<sup>4</sup>. SVs correspond to the juxtaposition of non-contiguous chromosomal segments through genomic rearrangement or more than 50bp. High variations in the burden of SVs can as well be observed, ranging from no SV to one thousand in some breast tumors genomes<sup>5</sup>.

The first chapter of the introduction describes the reparation of DNA Double-Strand Break, a highly cytotoxic lesion, with a focus on Homologous Recombination, a crucial DNA repair pathway for genomic stability. The second chapter presents the concept of Homologous Recombination Deficiency, related to the inactivation of this major reparation pathway and the different genomic signatures induced by this deficiency.

During this thesis I also had the opportunity to bioinformatically investigate the prevalence of mutations in two genes, *CDK12* and *MBD4*, that play an active part in genomic stability processes and that are of specific interest in two cancer diseases, Epithelial Ovarian Cancer (EOC) and Uveal Melanoma (UM), respectively. Those two genes are directly introduced before the two publications resulting from their investigations.

## **DNA DOUBLE-STRAND BREAKS AND HOMOLOGOUS RECOMBINATION**

A DNA Double Strand-Break (DSB) represent the most cytotoxic type of lesions. DSB appears when both backbones of complementary DNA strands are severed in one location of the genome<sup>6</sup>. DSB can arise from both exogenous and endogenous origins. The main pathways for DSB repair are Homologous Recombination (HR) and Canonical-Non-Homologous End Joining (C-NHEJ). HR is a replication associated reparation pathway that use a sister chromatid as a template, principally during mid S-phase to G2-phase but can also use a homologous chromosome during meiosis. C-NHEJ does not use a template DNA and is effective through the entire cell cycle. This chapter will focus on the Homologous Recombination pathway, including how DSBs can appear, what cellular processes arise upon this type of lesion and what favors a DSB repair by HR.

### **Endogenous and exogenous causes of DNA double-strand breaks**

The most important factor for endogenous DBSs is related to DNA replication. DNA replication starts from many individual replication origins that form bidirectional replication forks<sup>7,8</sup>. When encountering a single-strand DNA lesion (SSL), the replication fork can slow down or stall. The slowing or stalling of the replication forks is referred as “replication stress”. DSB might appear from SSL either during the interaction of the replication fork with a nicked DNA strand or by the processing of a stalled replication fork intermediate<sup>9-11</sup>.

One source of endogenous SSL is Reactive Oxygen Species (ROS), by-products of cellular metabolism that induce base modifications, single-strand breaks (SSB), protein-DNA adducts, and intra/interstrand DNA crosslinks<sup>12,13</sup>. ROS can also create two SSBs close to one another on different strands that can evolve into a DSB<sup>14,15</sup>. Other endogenous causes of replication forks ROS are due to complex DNA structures such as G4-structures<sup>16</sup>, or collisions between DNA replication and RNA at gene transcription, forming hybrids known as R-loops<sup>17</sup>. These regions may induce DSBs, notably at large DNA regions susceptible to replication stress known as Common Fragile Sites (CFS)<sup>18</sup>.

Finally, DSBs can be directly programmed by the cell. V(D)J recombination, a key mechanism for T-cell receptor and immunoglobulins diversity<sup>19</sup> and meiosis<sup>20</sup> are good example of programmed DSBs.

The two major exogenous origins of DSBs are chemicals and Ionizing Radiation (IR). IR affects directly the chromosome and produce DSBs or induce ROS with subsequent endogenous damage to DNA<sup>21,22</sup>. Major DNA damaging chemicals that introduce DSBs include DNA alkylating agent preventing a correct linkage of the DNA helix and leading to DNA breakage<sup>23</sup>,

cross-linking agent like cisplatin blocking DNA strands separation<sup>24</sup> and topoisomerase inhibitors such as camptothecin<sup>25</sup>, preventing DNA winding and condensation.

### Double-strand breaks detection and signaling

The first step for DSB repair is the detection of the damage. The MRE11-RAD50-NSB1 (MRN) complex sits at the top of the cell response to DSB in human. It serves as the primary sensor for DSB that first arrives at the break<sup>26</sup>. Another identified sensor complex is the Ku70/80 that accumulate at DSBs<sup>27,28</sup>. Both orchestrate and participate to DNA Damage Response (DDR), notably by activating downstream signaling pathways that involves several kinases<sup>29,30</sup>. The Ataxia-Telangiectasia Mutated (ATM) kinase, one of the principals signaling proteins of DSB repair is recruited and activated by the MRN complex<sup>31</sup>. The DNA-dependent protein kinase (DNA-PK) is recruited by Ku70/80 and both serves for C-NHEJ repair<sup>32</sup>. In replicative stress another kinase, the Ataxia Telangiectasia and Rad3-related (ATR), is recruited and its activation partly depends on the MRN complex<sup>33,34</sup>.

Upon DSB, the chromatin is reorganized, allowing the DNA to open and change the histones<sup>35-37</sup>. This helps with the recruitment of DDR proteins, the signaling and subsequent DSB repair. The histone H2AX, a variant of the histone family H2A, is transformed into  $\gamma$ H2AX at both sides of the break<sup>38,39</sup>. A cascade of events will lead to the bidirectional spread of  $\gamma$ H2AX around the DSB, with more MRN complexes recruited and creating a DDR competent domain<sup>40,41</sup>.  $\gamma$ H2AX can be found at the break site within few minutes after the damage<sup>39</sup> and is a widely used marker for DNA damage<sup>42</sup>. The activation of  $\gamma$ H2AX can be done by ATM<sup>42</sup>, ATR<sup>43</sup> or DNA-PKs<sup>44</sup>.

In response to DNA damage and DSB, cells activate checkpoints blocking or slowing the cell-cycle. This provides time for the DNA repair machinery to fix the damage, preventing DNA damage to remain in the cell during replication and chromosome segregation.

Upon directly created DSB, notably by ionizing radiation, the Checkpoint protein 2 (CHK2)/ATM kinase signaling is activated. CHK2 serves as a signal distributor to downstream targets that will block G1/S<sup>45</sup> or G2/M transition<sup>46</sup>. Upon replicative stress leading to DSB, it is the Checkpoint protein (CHK1)/ATR kinase signaling pathway that is activated. CHK1 also promotes the blockage of G1/S or G2/M transitions<sup>45,47</sup>. A crosstalk exists between ATR and ATM mediated pathways, distinct but with overlapping function.

Another important actor in cell-cycle blockage is p53, nicknamed “the guardian of the genome”. P53 most prominent roles are apoptosis<sup>48,49</sup> and cell cycle arrest in damaged cell. Upon DSB, ATM directly cause the rapid accumulation of p53<sup>50</sup>. P53 can be activated by both ATM/ATR and CHK1/CHK2 upon DSB<sup>51-53</sup>. After activation, it promotes G1 and G2/M cell cycle arrest<sup>54-56</sup>. It is therefore a crucial tumor suppressor present in many cancers.

## Different double-strand break repair pathways

Different pathways exist to repair DSBs that highly depend on the cell-cycle phase. Homologous Recombination (HR) relies on the physical proximity of a sister chromatid for a faithful reparation and therefore can only act from S to G2 phase of the cell-cycle. HR can also occur during meiosis and use the homologous chromosome<sup>57</sup>. Canonical-Non-Homologous End-Joining (C-NHEJ) relies on the direct ligation of the DSB ends. Because C-NHEJ does not need a template DNA, it is active throughout the cell-cycle. HR and C-NHEJ are considered the main reparation pathways for DSB in a normal cell<sup>21</sup>. Additional alternative DSB reparation pathways exist, including Single-Strand Annealing (SSA) and Alternative End-Joining (Alt-EJ). Alt-EJ is used to describe slow and error prone repair pathways with mechanistic differences but largely overlapping<sup>58,59</sup>. Alt-EJ can pair small regions of microhomology of less than 20bp to anneal two DNA ends with its key DNA polymerase Pol- $\theta$ <sup>60,61</sup>. Alt-EJ without the use of Pol- $\theta$  also exists and introduces large deletions but is less well characterized<sup>62,63</sup>. SSA requires larger homology regions than Alt-EJ of more than 20bp<sup>64</sup>. SSA repair pathway involves the annealing of the two homology regions by RAD52<sup>65</sup> and then the removal of the non-homologous 3' DNA tails, with potential DNA gaps filling by polymerase<sup>66</sup>. Whether SSA and Alt-EJ are backup pathways in healthy cells or are privilege at certain moment or lesion is still debatable and under investigation<sup>67</sup>.

## Homologous recombination pathway

### **DNA end resection in HR and the choice of the double-strand break repair pathway**

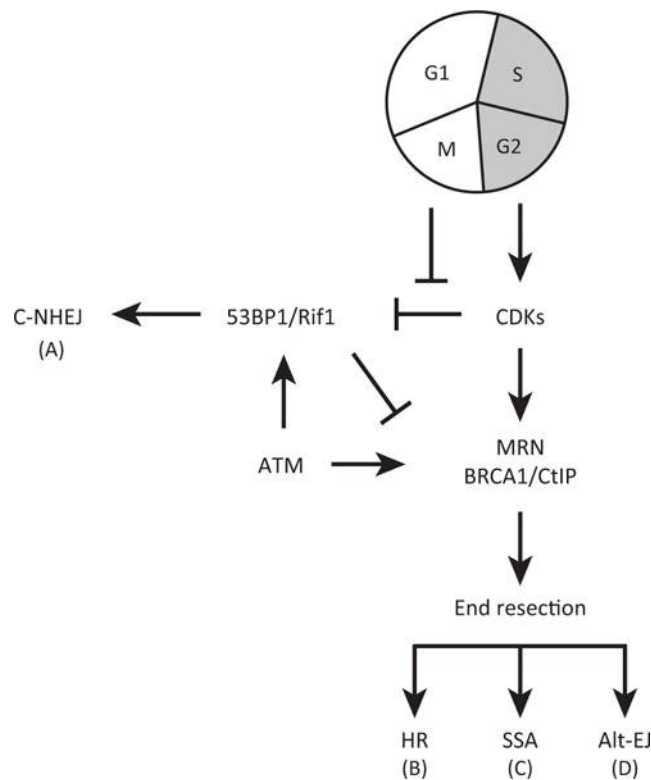
The choice of the pathway for DSB repair depends on the end-resection of the DNA, which is directly mediated by the cell-cycle phase and the molecular actors present at the break (Figure 1). HR relies on the physical proximity of a sister chromatid or homologous chromosome and therefore can only act from S to G2 phase of the cell-cycle or during programmed chromosome recombination.

The initiation of the end-resection of the DSB and is mainly regulated by the p53 binding protein 1 (53BP1).  $\gamma$ H2AX histone directs the accumulation of 53BP1 at the DSB site on the chromatin<sup>68</sup>. Independently and in parallel, ATM promotes the formation of the Shieldin complex with 53BP1 and other proteins, which binds to the ssDNA of the DSB<sup>69-71</sup>. This complex sterically prevents the action of nucleases, protecting DNA ends from resection and promoting C-NHEJ<sup>70</sup>. Additionally, the heterodimer Ku70/80 at the DSB attenuates any possible resection of the DNA ends<sup>72</sup>.

During S-phase, the high level of CDKs favors the phosphorylation of the C-terminal Interaction Protein (CtIP) and its interaction with Breast Cancer gene 1 (BRCA1)<sup>73,74</sup>. This promotes the

formation of a complex with BRCA1 and BRCA1 Associated RING Domain 1 (BARD1)<sup>75</sup> that consequently prompt the displacement of Ku70/80 and the withdrawal of the Shieldin complex from the DSB site, providing access to DNA ends for nucleases<sup>69,76,77</sup>.

Unprotected DNA ends are then resected in two phases. In the first phase CtIP and the MRE11 from the MRN complex create a short (~20 bp in mammalian cells) 3' end resected ssDNA<sup>73</sup>. At this step the DNA ends are available for Alt-EJ, which needs small DNA end resection<sup>78</sup>. The second phase is a long-range end resection ensured by nucleases and helicases from two distinct yet similar downstream pathways<sup>79-82</sup>. This new resected DSBs can no longer be processed by C-NHEJ, thereby promoting HR. Another DSB repair pathway SSA may occur at this step when the longer resected DNA ends are available<sup>83</sup>.



**FIGURE 1: End resection of the double-strand break determine the repair pathway**

The end resection of the double-strand break is directly dependent on the cell-cycle and regulate the choice between C-NHEJ and the other end-resection dependent repair pathways. C-NHEJ: Canonical-Non-Homologous End-Joining; HR: Homologous Recombination; SSA: Single-Strand Annealing; Alt-EJ: Alternative nonhomologous End-Joining. Extracted from Ceccaldi et al<sup>84</sup>.

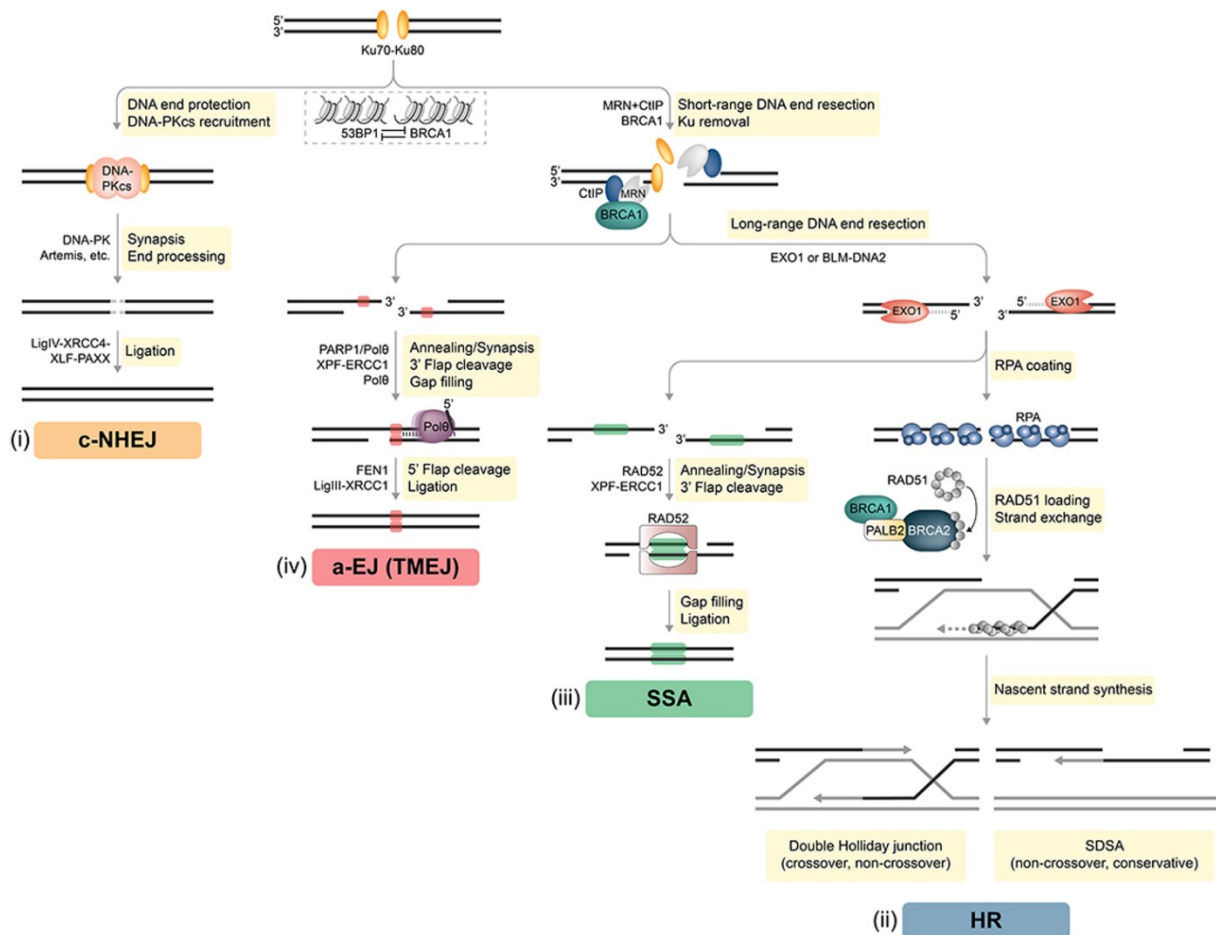
### DNA strand invasion by RAD51 and reparation homologous recombination

ssDNA intermediates created by the different end resections are rapidly passively coated and protected with the Replication Protein A (RPA). ssDNA paired with RPA cannot be associated with another ssDNA<sup>85,86</sup>. RPA blocks the RAD51 nucleoprotein filament from loading on the ssDNA, preventing a reparation by HR. To promote HR, Breast Cancer gene 2 (BRCA2) is

recruited to the DSB by a complex comprising of BRCA1-BRIP1 (BRCA1 Interacting Helicase 1) and the Partner And Localizer of BRCA2 (PALB2). BRCA2 and PALB2 promotes the active recruitment of RAD51 to replace RPA<sup>87-90</sup>. Interestingly, BRCA2 mediator role seems to be close to that of the RAD52 protein in yeast<sup>91</sup> but no clear role of RAD52 in human HR has been established for now.

Further on, RAD51 and its paralogs (RAD51B, RAD51C, RAD51D, XRCC2, and XRCC3) form a dynamic structure called nucleoprotein filament around the ssDNA<sup>92,93</sup>. The filament will invade the intact dsDNA of the sister chromatin and search for a homology sequence. This part of the mechanism remains to be completely described but RAD54 is thought to stimulate locally the displacement of the RAD51, while perturbing locally the sister chromatid dsDNA and stimulating the homology recognition of RAD51<sup>94</sup>. Once the homology sequence is found, the DNA is synthesized using the intact template by polymerases and ligases forming D-loop also called a Holliday junction. This can be done through two different mechanisms: by Synthesis-Dependent Strand Annealing (SDSA) or with the formation of a double Holliday junction by Double-Strand Break Repair (DSBR)<sup>95</sup>.

In SDSA only one end of DSBs invades the sister chromatid with RAD51. The other end of the DSB is passive and is annealed to the nascent displaced strand, facilitating HR termination. This will result in only non-crossover events<sup>96</sup>. In opposition, the DSBR, historically called the “canonical” DSB HR repair, occurs when the second end is captured and annealed to the same displaced template strand. The DNA is then synthesized in both directions and is forming a double Holliday junction. Depending on the resolution of the double Holiday junction, these might result in a cross-over events between sister chromatids. This resolution can occur according to two axes in the planar form created by the double junctions. The horizontal solution does not lead to genomic crossing and is assured by the BLM-TOP3 $\alpha$ -RM1 complex that will “dissolve” the junctions<sup>97</sup>. Most Holliday junctions are processed this way. However, the cleavage of the junction can be done vertically, with the action of resolvases SLX1-SLX4, MUS81-EME1 or GEN1, potentially resulting in crossover events between sister chromatin and potential loss-of-heterozygosity (LOH)<sup>98,99</sup>. HR is considered mostly as an error-free pathway. The final step of the Homologous Recombination is the dissociation of the RAD51 nucleoprotein of the DNA<sup>100</sup>.



**FIGURE 2: All principal DSB repair pathway and possible occurrence of alternative repair pathways**

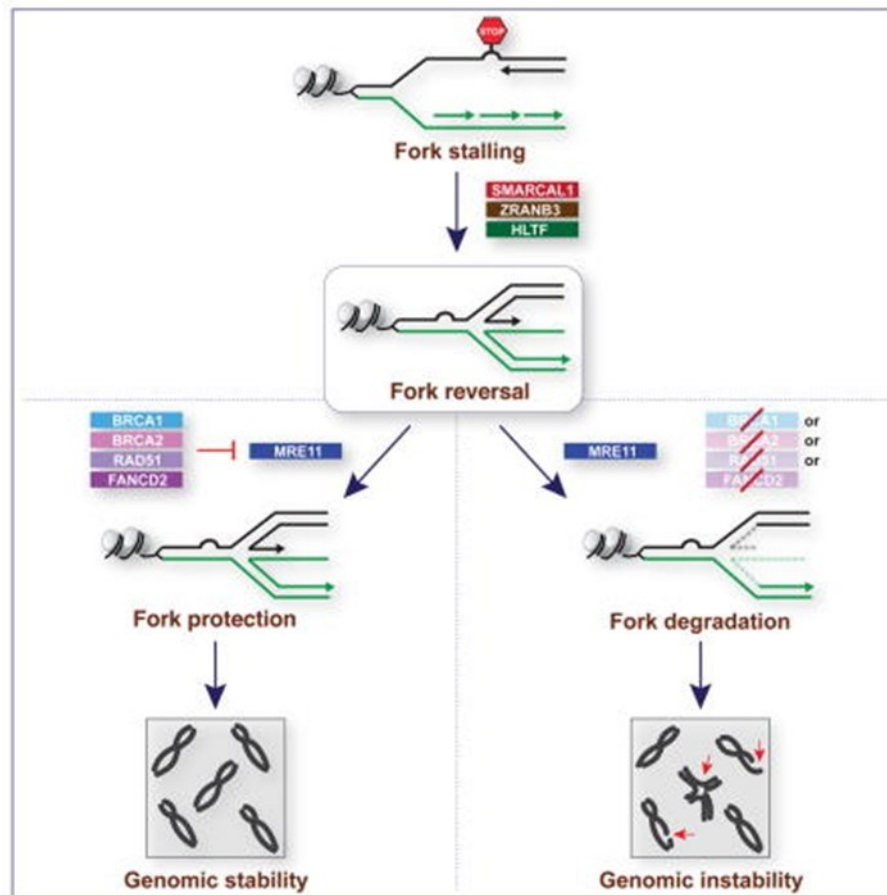
C-NHEJ: Canonical-Non-Homologous End-Joining; HR: Homologous Recombination; SSA: Single-Strand Annealing; a-EJ: Alternative nonhomologous End-Joining; TMEJ: Theta-Mediated End-Joining, another name for Alt-EJ; SDSA: Synthesis-Dependent Strand Annealing. Extracted from Trenner et al<sup>101</sup>.

### Replication fork protection for genomic stability

BRCA1, BRCA2 and RAD51 fulfil another major role in genome integrity: replication fork protection. Replication fork reversal by different translocases is a key protective mechanism of the cell to stabilize and restart the stalled replication fork upon DNA lesions without DNA breakage<sup>102-105</sup>. Especially important in replicative stress<sup>106</sup>, *BRCA1* and *BRCA2* genes protect the reversed replication forks by stabilizing RAD51 on the regressed arm, preventing nucleolytic degradation (Figure 3)<sup>107-110</sup>. Replication fork can also restart in an HR-error-free way without fork collapsing<sup>111,112</sup>. This can however lead to Sister Chromatid Exchange (SCE)<sup>113</sup>. The degradation of this mechanism leads to genomic instability and potential cell-death. The role of *BRCA1*, *BRCA2* and *RAD51* in genomic stability therefore extends beyond DSB repair. In the case of fork collapsing, the DSB can be repaired with a form of HR called Break-Induced Replication (BIR), with a strand invasion as an initial step<sup>114</sup>. FANCD2-FANCI,



actors implicated a complex DNA repair pathway called Fanconi anemia, are also thought to play a role in replication fork protection<sup>115,116</sup>.



**FIGURE 3: Simplified model for fork protection of reversed fork upon replication fork stalling**

SMARCAL1, ZRANB3, HLTF are fork remodelers that promote replication fork reversal. BRCA1, BRCA2, RAD51 and FANCD2 protect reversed replication fork. Extracted from Tagliatela et al<sup>110</sup>.

### The Fanconi anemia pathway

HR is closely associated with the major pathway to repair Interstrand crosslinks (ICL) in human, the Fanconi anemia pathway, with many proteins present in both<sup>117</sup>. ICLs are cytotoxic DNA lesions representing covalent linkage between nucleotide residues from opposite strand of DNA, which block DNA strand separation, stalling cellular processes such as DNA replication or transcription. ICLs need either to be bypassed or repaired, especially during replication stress. In human, the Fanconi anemia pathway orchestrates ICL reparation by Nucleotide Excision Repair (NER), translesion DNA synthesis, HR and alternative repair pathway<sup>117</sup>. The nucleotic incision of the ICL is done by the NER pathway, which creates a DSB as intermediate, repaired mainly by HR<sup>118,119</sup>. It insures supplementary roles for the stabilization of stalled replication fork and fragile sites protection. Nineteen proteins have been identified as Fanconi anemia proteins (from FANCA to FANCT), including HR proteins BRCA2 (FANCD1),

BRIP1 (FANCI), PALB2 (FANCD1), RAD51C (FANCD2), RAD51 (FANCD3) and BRCA1 (FANCD4). Other interplay between HR and Fanconi anemia pathway include FANCD2-FANCI, which plays a central role in Fanconi pathway by initiating the repair of ICL and participate in DNA end resection that disables C-NHEJ in favor of HR<sup>120,121</sup>. However, cells inactivated for FANCD2-FANCI have a mild effect on the reparation of DSB by HR<sup>122,123</sup>.

### Conclusion

Homologous Recombination is a complex reparation pathway involving many actors that depends on other complex molecular machineries, including DSB signaling. Cancer inactivating mutations were observed in many DNA repair genes. One longstanding question was to decipher which gene inactivation led to similar defects in the DNA repair. This is important to know, first, to refine gene function and, second, for patient care and drug development.

# HOMOLOGOUS RECOMBINATION DEFICIENCY IN CANCERS

## Cancer predisposition syndromes associated with homologous recombination genes

Cancer predisposition syndromes (or hereditary cancer predispositions) corresponds to genetic alterations that increases the likelihood of developing a cancer relative to the general population.

*BRCA1* and *BRCA2* are the principal genes predisposing to breast and ovary cancers. Approximately 13% and 1.2% of women of the general population will develop a breast cancer or an epithelial ovarian carcinoma (EOC) during their life, respectively<sup>124</sup>. Meanwhile, for women with inherited deleterious variants for *BRCA1* and *BRCA2* the likelihood to develop a breast cancer by the age of 80 (cumulative risk) is estimated to be 57-72% (39-44% for EOC) and 49-69% (11-17% for EOC), respectively<sup>125,126</sup>. Additionally, *BRCA2* mutations and in a lesser extent *BRCA1* mutations, predispose men to breast cancer with a Relative Risk (RR,  $RR = \frac{\text{likelihood under condition}}{\text{likelihood in general population}}$ ) of 22 and 11 at the age of 80, respectively<sup>127</sup>. The same goes for prostate cancers, with a reported RR = 2.5-4.65 for *BRCA2* mutation carriers<sup>127,128</sup>, and for pancreatic cancer, with reported RR = 2.26 and RR = 3.51 for *BRCA1* and *BRCA2* mutation carriers, respectively<sup>129</sup>.

Deleterious germline mutations of other genes directly implicated in HR repair were also shown to increase the risks of EOC, breast and pancreatic cancers. The RR for breast cancer in *PALB2* deleterious mutation carriers is estimated to range between 5 to 9 depending on the age<sup>130</sup>. For EOC and pancreatic cancer, the associated RR are estimated to be 2.91 and 2.37, respectively<sup>131</sup>. *BARD1* is reported in several studies as a cancer predisposing gene<sup>132,133</sup>. *BRIP1* represents a 10 % cumulative risk for EOC<sup>134,135</sup> and was also reported in few familial cases of colon cancer<sup>136</sup>. Finally, RAD51 paralogs *RAD51C* and *RAD51D* predispose to EOC with a less clear involvement in breast cancer predisposition<sup>137,138</sup>. *RAD51B* and *XRCC2* are also reported in familial cases of breast and ovarian cancers<sup>139,140</sup>. The predisposing nature of *XRCC3* is more mitigated<sup>140,141</sup>.

## Prevalence of HR genes inactivation in cancer

All the cancer predisposing genes listed above behave like “tumor suppressors” and the somatic inactivation of the second allele according to the Knudson's two-hit hypothesis leads to tumorigenesis. In the context of hereditary deleterious mutation, the wild-type allele is mostly deleted via a large-scale chromosomal loss leading to the Loss-Of-Heterozygosity (LOH). Somatic bi-allelic inactivation of HR genes was also observed in breast and ovarian cancers, with somatic deleterious mutation accompanied by LOH and in some cases two deleterious

somatic mutations<sup>142</sup>. Certain HR genes can be inactivated by hypermethylation of the promoter, including frequently silenced *BRCA1* and some rather rare silencing of *RAD51C*, with a high prevalence of concurrent LOH<sup>143-145</sup>.

*BRCA1* and *BRCA2* bi-allelic inactivation are present in several cancers including EOC, breast, prostate and pancreatic cancers. In High-Grade Serous Ovarian Carcinoma (HGSOC), a frequent and aggressive subtype of EOC, *BRCA1* and *BRCA2* inactivation represents ~29% of the cases, a third of those being inactivated by *BRCA1* promoter methylation<sup>146,147</sup>. Germline mutations in *BRCA1/BRCA2* explain around 22% of all HGSOC cases<sup>146,147</sup> while accounting for a lesser proportion of all EOC, approximately 14-15%<sup>147,148</sup>. In breast cancer, the inactivation of BRCA-genes explains around 16% of the cases with two third originating from *BRCA1/2* germline mutations<sup>149,150</sup>. *BRCA1* inactivation is mainly associated with Triple-Negative Breast Cancer (TNBC), an aggressive subtype of breast cancer characterized by the lack of expression of Estrogen Receptor (ER), Progesterone Receptor (PR) and lack of over-expression of human epidermal growth factor receptor 2 (HER2)<sup>149</sup>, while *BRCA2* inactivation is more represented in the luminal breast cancer subtype, characterized by the expression of hormone receptors (ER and/or PR)<sup>151,152</sup>. Breast tumors that overexpress HER2 (HER2+) are rarer in the context of *BRCA1/2* mutations<sup>142,153</sup>. *BRCA* genes inactivation is also present in other cancer types. In prostate cancer *BRCA2* deletion is more present, reported in 5.3 to 13% of cases, compared to *BRCA1* that represents less than 1% of the cases<sup>154,155</sup>. The same kind of repartition can be observed for *BRCA2* and *BRCA1* inactivation in pancreatic tumors, encompassing 3.5% and 0.9% of the cases, respectively<sup>156</sup>.

The prevalence of other HR with bi-allelic inactivation is rather low, with *RAD51C* and *PALB2* being the most represented among those. *RAD51C* is inactivated in 2% of the cases from The Cancer Genome Atlas (TCGA) HGSOC cohort<sup>146</sup> and is described in breast carcinomas<sup>157</sup>. Likewise, *PALB2* bi-allelic inactivation is displayed in breast tumors in a few percent of the cases<sup>157,158</sup>. Interestingly, no tumor with *RAD51* inactivation was found, suggesting that it is an essential gene for cells<sup>159</sup>.

### Genomic instability in *BRCA1/2*-/- and HRD phenotype

*BRCA1* or *BRCA2* bi-allelic inactivation leads to impaired DNA-reparation by HR and an accumulation of unrepaired DSBs resulting in chromosomal abnormalities<sup>160-162</sup>. Indeed, evidence of spontaneous chromosomal breaks with important level of aneuploidy and genetic exchange between non-homologous chromosomes were observed in cell lines, cancer cells and in mouse models<sup>163-166</sup>. Engineered mouse with *BRCA2* mutation showed an increase number of DNA deletions at the DSB that seemed characteristic of this inactivation<sup>162</sup>. The genomic phenotype observed in *BRCA1/2*-/- cells is often referred to as genomic instability

phenotype, because of numerous chromosomal changes observed upon *BRCA1/2* inactivation and compared to other tumors/cells/models.

The Homologous Recombination Deficiency (HRD) phenotype is linked to the impairment of DSB repair by HR. The concept appeared when the inactivation of *BRCA1/2* was described in a fraction of breast and ovarian carcinomas rising the question of the scope of *BRCA1/2*-like phenotype in tumors. The term BRCAness was proposed to designate the similar phenotype that sporadic tumors share with cancers developing in *BRCA1/2* germline mutation carriers<sup>167</sup>. Now HRD and BRCAness account for the same genomic instability phenotype associated with the impairment of DSB repair by HR.

Modern DNA sequencing techniques provided the background for extensive analysis of genetic events across tumors and genomes in association with DNA repair gene inactivation<sup>54,168</sup>. Recently, the HRD phenotype was exhaustively described, including the spectrum of inactivated genes associated with the HRD phenotype in breast cancers<sup>150</sup> and at the pan-cancer level<sup>169</sup>. Indeed, *BRCA1/2*-/- breast and ovarian tumors were characterized by the increased level of structural chromosomal rearrangements of small, intermediate, and large-scale (described in detail in sections 5,6,7,8 of the current chapter). Similar phenotypes were consistently found in tumors with *RAD51C* and *PALB2* bi-allelic inactivation. Thus, *BRCA1*, *BRCA2*, *RAD51C* and *PALB2* are the only genes, which bi-allelic inactivation is proven to be unambiguously associated to the HRD phenotype. Some other genes from HR pathway have to the moment weaker evidence of being associated with HRD. These are *RAD51B*, *RAD51D*, *XRCC2*, *XRCC3* (all *RAD51* paralogs), *BRIP1* and *BARD1*, which lack a strong statistical validation due to the rarity of their bi-allelic inactivation in tumors.

Some genes from the HR pathway were however proven to be not associated to HRD if inactivated in cancers. These are the genes implicated in the DSB signaling such as *ATM*, *ATR*, *MRE11*, *RAD50* and *NSB1*, which are also critical for genome integrity. These genes can lead to different disorders such as cancer-prone syndromes Ataxia-Telangiectasia for *ATM*<sup>170</sup> or Nijmegen breakage syndrome for *NSB1*<sup>171</sup>. The inactivation of these genes does not lead to a HRD phenotype similar to *BRCA1/2* inactivation in tumors and their sensitivity to drug targeting HRD remains unclear<sup>172</sup>. The same goes for the genes implicated in the Fanconi anemia pathway but not directly related to HR pathway. Fanconi anemia hereditary disorder leads to bone marrow failure and mostly acute myeloblastic leukemia (AML), which displays a different genomic phenotype than BRCAness.

The scope of HRD phenotype across tumor types, besides breast, ovarian, prostate and pancreatic carcinomas, need to be further clarified. However, few cases with this phenotype could be observed in various cancer type including colon, skin, lung, kidney, liver, esophagus,

lymphoid, head and neck cancers<sup>169</sup>. Investigating these cases is important for future personal medicine approaches.

## Treatment for tumors with homologous recombination deficiency

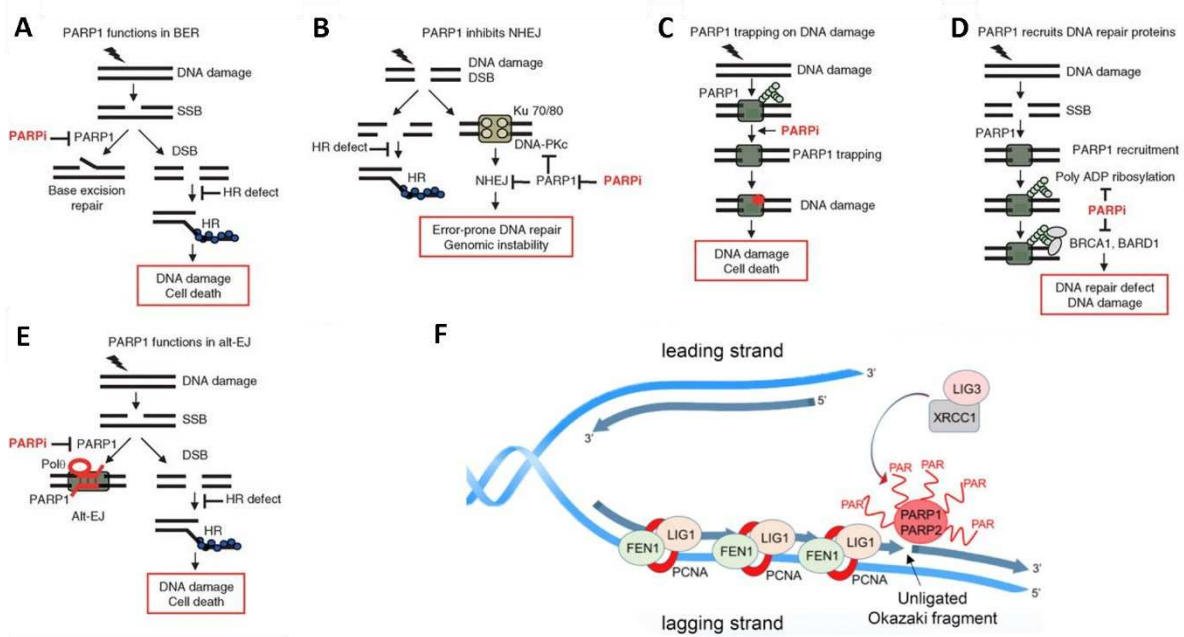
### Sensitivity to DNA damaging agents

*BRCA1* and *BRCA2* deficient tumors were long known to be sensitive to DNA damaging agents. Indeed, those tumors showed good response to platinum salts chemotherapy such as cisplatin and carboplatin<sup>173,174</sup>. Platinum salts are cross-linking agents that can create interstrand crosslinks (ICL). The processing of ICL by the Fanconi pathway creates DSBs as intermediate of the pathway that cannot be repaired in an HRD context, resulting in a specific strong cytotoxic effect in HR-deficient cells.

BRCA-/- tumors were also shown to respond well to poly-ADP-ribose-polymerase inhibitors (PARPi) that display great performance in clinics to treat breast, ovarian, prostate and pancreatic cancers<sup>175-177</sup>. In EOC, *BRCA1/2* germline mutation carriers harbor a longer progression-free survival (PFS) with no tumor progression or death compared to BRCA wild type patients upon PARPi treatment, with a reported hazard ratio of 0.27<sup>178</sup>.

PARPi is based on the synthetic lethality interaction between PARP inhibition and HRD. PARP1 is participating in Single-Strand Break (SSB) reparation with Base Excision Repair (BER)<sup>10</sup>. When PARP1 is inhibited, SSB are not corrected and subsequently transformed into DSB upon replication fork passage during S-phase. Those DBSs can normally be repaired by HR. In a HRD context other error-prone DSB repair mechanisms will be utilized, leading to an accumulation of mutations, genomic instability and potential apoptosis of the cell (Figure 4A).

Moreover, PARPi activity is broader than the catalytic inhibition of PARP1 in BER. PARPi can also block PARP1 and PARP2 onto a DNA damaged site. The DNA-PARP complex resulting from this stabilization, referred to as trapping, prevents DNA processing and has a cytotoxic effect for the cell<sup>179,180</sup> (Figure 4C). Additionally, PARPi also influence other DNA repair pathways. PARP1 was shown to limit C-NHEJ<sup>181</sup> and PARPi therefore promote C-NHEJ (Figure 4B). PARPi also impairs the recruitment of the BRCA1-BARD1 complex<sup>182</sup>, which might have some implication on DNA repair, even in an HRD context (Figure 4D). Finally, the inhibition of PARPs also disrupts Alt-EJ<sup>183,184</sup>. Alt-EJ was shown to be an important alternative pathway in HRD<sup>185</sup>, therefore also potentially explaining PARPi synthetic lethality (Figure 4E). More recent publications also uncovered that PARP1 helps to fix unligated DNA replication intermediates (Figure 4F). Hence PARPi could also alter DNA replication and may create a DSB, providing an additional rationale for PARPi toxicity in HR-deficient cancer cells<sup>186,187</sup>.



**FIGURE 4: PARPi activity and potential synthetic lethality in HRD**

**A|** PARPi impairs BER promoting toxicity in HRD **B|** PARPi promotes C-NHEJ which may promote additional toxicity in HR **C|** PARPi can trap PARP1 on DNA lesions **D|** BRCA1-BARD1 complex rely on PARP mediated recruitment and is impaired by PARPi, with potential subsequent DNA repair defect **E|** PARPi impairs Alt-EJ, a major alternative pathway of HR **F|** PARP1 facilitates the repair of unligated replication intermediate known as Okazaki fragment which may explain additional toxicity of PARPi in HRD. Extracted from Konstantinopoulos et al<sup>184</sup> (A to E) and Hanzlikova et al<sup>186</sup> (F).

The response to platinum-based chemotherapy and PARPi depending on the HR inactivation mechanism was assessed in different cohorts. Germline versus somatic inactivation of BRCA1/2 seems to display similar response rate<sup>188</sup>. A differential response between BRCA1-/- and BRCA2-/- tumors is dubious depending on the study<sup>188-190</sup>. In a TNBC cohort, mutated and hypermethylated BRCA1-/- cases responded similarly to chemotherapy<sup>144</sup>. BRCA-null tumors are mostly considered without distinctions in clinical settings. The location of the inactivating mutation seems however to influence tumor sensitivity<sup>191-193</sup>.

The association of PARPi with other type of treatments is currently investigated in a clinical context. This includes, but is not limited to, inhibition treatments for DNA reparation actors like ATR<sup>194</sup>, Pol- $\theta$ <sup>195,196</sup> or targeting treatment for known oncogenes<sup>197</sup>.

### Resistance to treatments

Although PARPi through synthetic lethality has been associated with good response in BRCA-/- tumors, resistance to this treatment is common in clinic<sup>198,199</sup>. One mechanism that was reported for PARPi resistance was mutation reversion. Secondary mutations can restore the wild-type version of the genes, therefore conferring late resistance to PARPi<sup>200-203</sup>. Another potential resistance mechanism is through the alteration of DNA resection, the key step for DSB repair pathway choice. 53BP1 and the Shieldin complex indirectly protect the DNA from

end resection and their inactivation has been associated with partial restoration of HR and PARPi resistance in BRCA-/- tumor<sup>70,204,205</sup>. Mechanisms of resistance to PARPi and chemotherapy are numerous, not listed exhaustively above and many probably remain to be unraveled. Alternative reparation pathway providing PARPi resistance may also explain unresponsive BRCA-/- tumors. Those resistance mechanisms are important and should be considered for optimal medical care throughout the history of tumor treatment.

Because of the sensitivity of HR-deficient tumors, detecting it is instrumental to predict and better advice for clinical care.

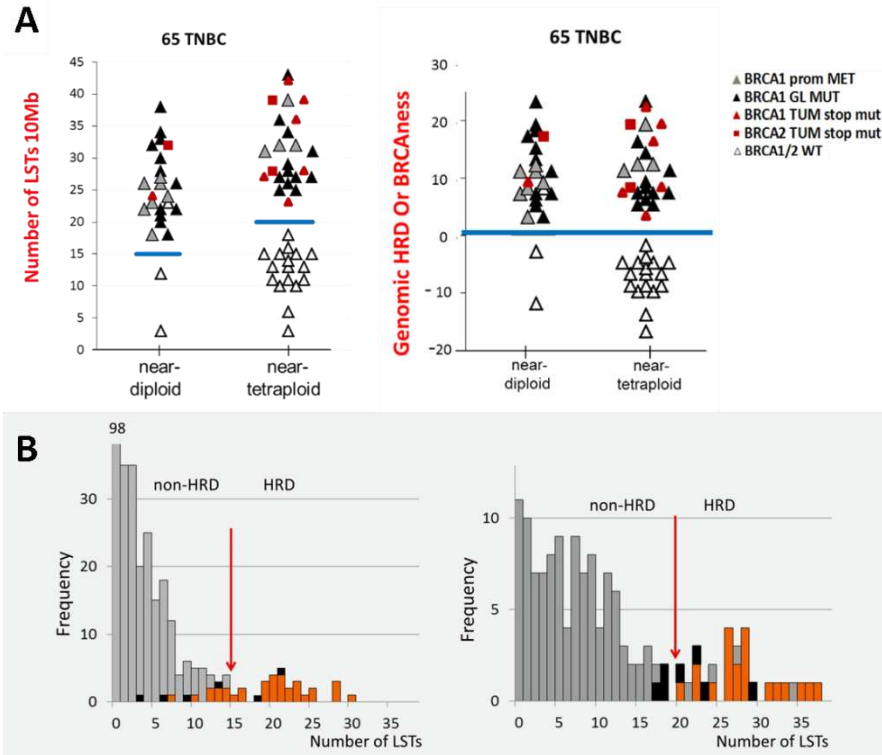
### Large-scale rearrangements based genomic signatures of homologous recombination deficiency

Genomic instability resulting from HRD is translated in highly rearranged tumor copy number profiles, which can be captured using various profile characterization methods. In 2012, three large-scale genomic signatures were described and were shown to be efficient biomarkers of HRD. The commercially available Myriad myChoice® CDx test combines these three signatures in one HRD test that is now the only FDA approved test allowing the prescription of PARPi treatment in BRCA-wildtype tumors.

#### Large-scale State Transitions (LST) signature

LST genomic signature of HRD was developed based on the series of 65 TNBCs and SNP-array technology<sup>206</sup>. Genomic profiles of tumors were mined using SNP-arrays and absolute copy number profiles were obtained. Large-scale State Transition (LST) was defined as a copy number break within chromosome arm between two contiguous genomic regions of at least 10Mb; small segments less than 3 Mb were filtered and/or smoothed; to call LST, a distance between large segments should not exceed 3Mb. The number of LSTs were calculated for each tumor genome. Additionally, each tumor was characterized by the genomic content corresponding to the estimated number of chromosomes and the tumors were classified into 2 groups: near-diploid (<50 chromosomes) and near-tetraploid (>= 50 chromosomes) - near-tetraploid genomes are those who underwent whole genome duplication<sup>207</sup>). The number of LST and tumor ploidy were shown to consistently separate TNBC tumors with and without HRD (Figure 5A, left panel). Using a defined ploidy-dependent cut-off of 15 and 20 LSTs for near-diploid and near-tetraploid tumors, respectively, genomic BRCAness score could be obtained (Figure 5A, right panel). The LST signature was furtherly validated in a larger cohort of 456 breast tumors, including all breast cancer subtypes (Figure 5B)<sup>142</sup>. *BRCA1*, *BRCA2* and *RAD51C* inactivation was shown to share large-scale instability phenotype in all subtypes of breast and ovarian cancers. Cisplatin treatment responders displayed mostly an elevated number of LSTs in TNBC<sup>142</sup>.



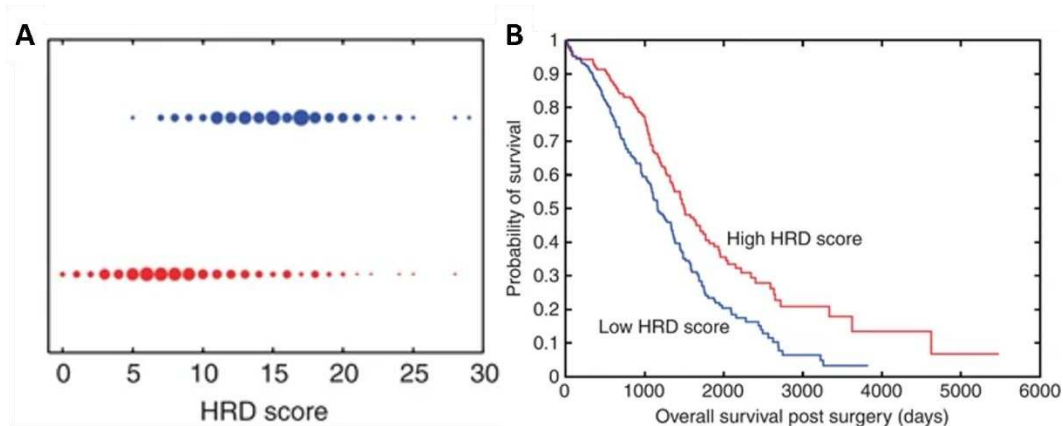


**FIGURE 5: Number of Large-scale State Transitions (LST) in a large in-house breast cancer cohort**

**A]** Number of LST for 65 TNBC. Left panel: absolute number of LST depending on the ploidy detected. Right panel: Corrected number of LST depending on the defined ploidy dependent cut-off. **B]** Number of LST for 456 breast carcinomas (399 luminal and 56 HER2+). Left panel: 317 near-diploid tumors. Right panel: 139 near-tetraploid tumors. Black bars: No HRD cause identified; Orange bars: HRD identified with altered *BRCA1/BRCA2* or *RAD51C* methylation; Grey bars: cases that were not tested completely or partially; Red arrow: cut-off for HRD as defined in Popova et al<sup>206</sup>. Extracted from personal data and Manié et al<sup>142,206</sup>.

### Loss of Heterozygosity (LOH) signature

Abkevich et al. investigated HRD in SNP-arrays through the prism of Loss-Of-Heterozygosity (LOH) profile<sup>208</sup>. Comparing the length of LOH adjusted for chromosome length, they extracted different features comprised of (1) small LOH (<15Mb), (2) large LOH (>15 Mb) but not covering the entire chromosome arm and (3) LOH spanning the entirety of chromosome. 15Mb was chosen arbitrary but the exact cut-off does not significantly impact the signature. The number of LOH covering the entire chromosome correlated with functional *BRCA1/2* while the number of large LOH longer than 15Mb but less than the whole chromosome arm correlated with *BRCA1/2* inactivated cases. Similar observation was made with *RAD51C* promotor methylation. Therefore, Abkevich et al. proposed the latter feature as a signature of HRD and could correlate their score to Overall Survival (OS) post-surgery. The results of this method are presented in Figure 6.

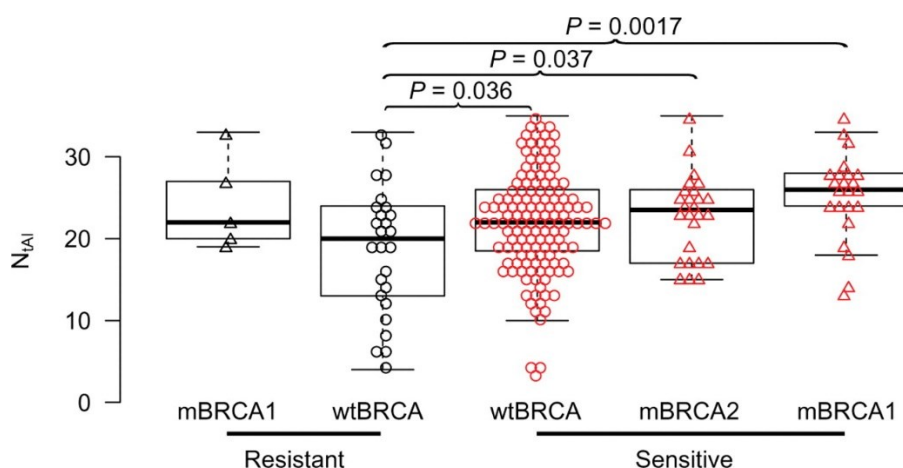


**FIGURE 6: LOH signature in Epithelial Ovarian Cancer**

A| x axis: Number of LOH regions defined by Abkevich et al. Blue circles: Samples inactivated for BRCA1 or BRCA2. Red circles: Samples with intact BRCA1 and BRCA2. The size of the circle is proportional to the number of people having the same number of LOH regions. 434 samples of Epithelial Ovarian Cancer with 146 BRCA1/2 inactivated. B| Kaplan–Meier plot of OS post-surgery for HRD score split at it's median. Generated from 507 samples of Epithelial Ovarian Cancer from the TCGA with available copy number data and survival information. Extracted from Abkevich et al<sup>208</sup>.

### Telomeric Allelic Imbalance (TAI) signature

Birkbak et al. built their signature based on the SNP-array profile and using response to cisplatin as a surrogate marker of HRD, as *BRCA1/2* inactivated cases are known to be more sensitive to platinum salts<sup>209</sup> SNP-arrays applied to tumor biopsy before treatment were mined for genomic aberrations extracting Allelic Imbalance (AI) profiles from tumor biopsy before treatment. AI corresponds to the uneven contribution of the alleles and thus evidence copy number alteration. Allelic Imbalance at the telomere regions (TAI) emerged as the genomic feature that was significantly associated with tumor response to cisplatin. The correlation was higher in the TNBC subtype. The association with cisplatin-sensitivity remained significant even in wild-type *BRCA1/2* tumors, evidencing cisplatin sensitivity in HRD besides *BRCA1/2* mutations. The results of this method are represented in Figure 7.



**FIGURE 7: Cisplatin response in serous ovarian cancer according to the number of allelic imbalances extending to the telomere end**

wtBRCA: no mutation BRCA1 or BRCA2. mBRCA1: mutated for BRCA1. mBRCA2: mutated for BRCA2. Extracted from Birkbak et al<sup>209</sup>.

### Large-scale alterations signatures in Myriad myChoice® CDx test

The three signatures described above, LST, LOH and TAI, were further validated for their association with HRD and the combination of those different scores was proposed to bring additional robustness<sup>210</sup>. The commercially available Myriad myChoice® CDx test combines these three signatures in one HRD test in addition to mutations of *BRCA1* and *BRCA2* detected by gene sequencing. It is the only FDA approved HRD test and is now required for the prescription of PARPi treatment for BRCA-wildtype tumors<sup>178,211,212</sup>. This test provides an evidence for additional cases to prescribe PARPi, as HRD is not limited to *BRCA1/BRCA2* mutation. The combination of these three signatures was recently explored based on next generation sequencing technic and showed a good performance when estimated from Whole Exome Sequencing copy number profiles<sup>213</sup>.

### Small-scale somatic alteration signatures of homologous recombination deficiency

High-throughput sequencing contributed enormously to characterization of fine-scale genomic alterations in cancers. Small-scale somatic alterations include bases substitutions and small insertions and deletions (indels). Single base substitutions can be of 6 types: C>A, C>G, C>T, T>A, T>C, T>G. Taken together with the 3' and 5' nucleotide context, these substitutions represent 96 possible combinations. Indels are defined as DNA fragments of less than 50bp that are either included or lost in a genomic location (the "small" size of indel is defined by the typical size of the read in the sequencing technique used). Indels are characterized by the size and the presence of homology/repetitive/unspecific DNA sequences at junction. Analysis of the frequency profiles of those alterations in a large set of tumors allowed to decipher recurrent

patterns (mutational signatures) using the non-negative matrix factorization (NMF) approach<sup>214,215</sup>.

Single-base substitutions signatures were initially obtained for a large cohort of tumor Whole-Exome Sequencing data<sup>215</sup>. The most up-to-date study has unraveled 49 single-base substitution (SBS) signatures, 11 Doublet-Base Substitution (DBS) signatures and 17 small Insertion-Deletion (ID) signatures<sup>4</sup>. They were formalized on 2,780 whole genomes sequencing from the Pan-Cancer Analysis Whole-Genome project (PCAWG)<sup>54</sup> and were verified on 1,865 additional whole genomes and 19,184 exomes from the International Cancer Genome Consortium (ICGC) and The Cancer Genome Atlas (TCGA) project<sup>168</sup>.

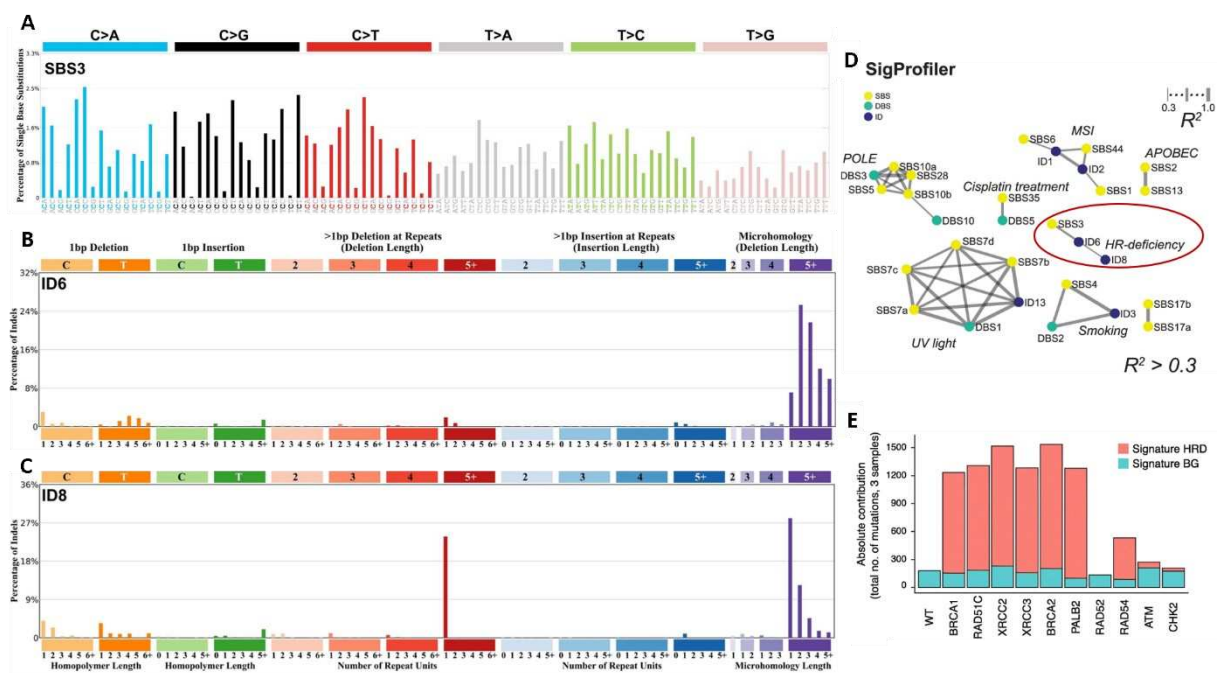
### Single-Base Substitution signatures of homologous recombination deficiency

The most consistent SBS signature associated with HRD is the so-called signature 3 (SBS3). SBS3 is characterized by an almost uniform distribution of all 96 base substitution types (Figure 8A). SBS3 was first associated with *BRCA1/2* depletion in breast, ovarian and pancreatic cancers<sup>150,215</sup>. Several tumors inactivated for *RAD51C* and *PALB2* also showed a strong prevalence of signature 3<sup>216-218</sup>. However, a large part of the samples harboring a high SBS3 prevalence did not have mutations in *BRCA1/2* nor epigenetic silencing of *BRCA1*. Using whole genome sequencing of isogenic cell lines inactivated for HR genes and DNA damage checkpoint gene, Poti et al. confirmed the prevalence of SBS3 signature for tumors inactivated for *BRCA1*, *BRCA2*, *RAD51C* and *PALB2*, but also highlighted comparable level of SBS3 for other RAD51 paralogs *XRCC2* and *XRCC3* (Figure 8E)<sup>219</sup>. No association between *ATM/CHK2* inactivation and signature 3 was found, comforting the difference in genomic profiles between HR genes and DSB signaling associated genes (Figure 8E)<sup>219</sup>. *RAD54* inactivation in cell line showed a moderate signature 3, while *RAD52*<sup>-/-</sup> cell-line didn't (Figure 8E)<sup>219</sup>.

SBS signatures only report the frequency of the mutations, not the actual mutation rate. A major caveat of SBS3 as HRD biomarker is the lack of specificity: due to quasi-uniform frequency distribution SBS3 can be confounded with the background mutation burden. A tool SigMA was developed for HRD detection based on the prevalence of SBS3 for exome and gene panels sequencing data<sup>220</sup>. The software showed a sensitivity of 74% and a specificity of 90% for HRD with the MSK-IMPACT gene panel (FDA approved) and found that the prevalence of SBS3 signature in tumors was associated with better response to PARPi. Another mutational signature, SBS8, have also been associated to HRD<sup>150</sup>, but the specificity of this signature is contested in other studies<sup>4,221,222</sup>.

## Insertion-Deletion signatures of homologous recombination deficiency

An increase number of indels was also found associated with *BRCA1*, *BRCA2* and *PALB2* inactivation in tumors, and with the inactivation of *RAD51* paralogs *RAD51C*, *XRCC2* and *XRCC3* in cell-lines (Figure 8E)<sup>4,150,219</sup>. The type of indels corresponded to the signatures ID6 and ID8, which are both related to deletions of more than 5bp<sup>4</sup>. ID6 has (micro)homology at breakpoint of more than 2bp, while ID8 has shorter to no microhomology at deletion (0-3bp) (Figure 8B and 8C). ID6 was strongly associated with SBS3 while ID8 displayed weaker association (Figure 8D)<sup>4</sup>. In cell-lines, ID6 signature with higher proportion of deletions with microhomology was strongly associated with *BRCA2* and *PALB2* inactivation as compared to *BRCA1*, *RAD1C*, *XRCC2* and *XRCC3* that showed a more equal repartition between microhomology, repeat at the break and no homology at the deletions (Figure 8E)<sup>219</sup>. Confirming the previous findings, inactivation of *ATM* and *CHK2* in cell lines were not associated to those two signatures, nor *RAD52* and only *RAD54* to a lesser extent (Figure 8E)<sup>219</sup>.



**FIGURE 8: Small-scale signature associated with HRD**

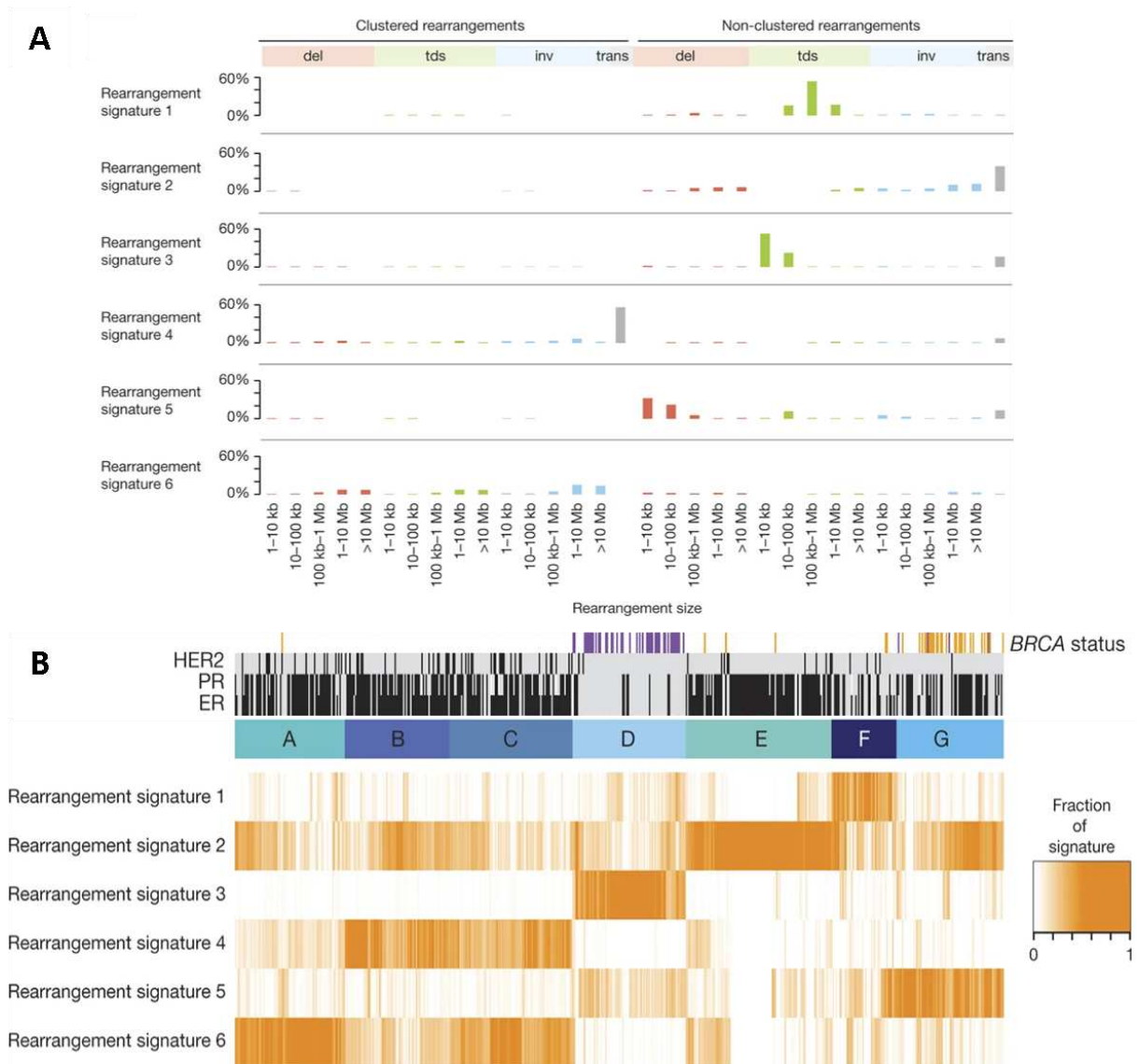
**A**] Single-Base Substitutions Signature 3 as defined by Alexandrov et al. **B**] Insertions and Deletions Signature 6 represented by Alexandrov et al. **C**] Insertions and Deletions Signature 8 represented by Alexandrov et al. **D**] Association of the different signatures by Alexandrov et al<sup>4</sup>. **E**] Representation of two de-novo signatures in cell-lines. The signature HRD is highly similar to SBS3 and signature BG represent the average substitution profile of a wild-type cell<sup>219</sup>.

## Rearrangement signature of homologous recombination deficiency

HRD is also characterized by an increased level of structural rearrangement of several types. In 2016, Nik-Zainal et al. pathed the way for this by looking at the Whole-Genome Sequencing (WGS) of 560 breast cancers<sup>150</sup>. An increased number of Structural Variations (SVs) in BRCA-/- tumors was observed with translocations, ~10kb size deletions and in some cases ~10kb size Tandem Duplications (TDs). To investigate SVs and to classify tumors based on SVs, 32 classes of large rearrangement (>1kb) were constructed as combination of SV type (insertion, deletions, translocations, inversions) and size. Based on the deconvolution of SV counts in 560 breast cancer genomes, 6 Rearrangement Signatures (RS) were extracted (Figure 9A) and subsequent clustering of tumors revealed 7 groups with similar SV frequency profiles (Figure 9B).

Strikingly, *BRCA1* and *BRCA2* inactivated samples were classified in different groups. BRCA1 and BRCA2 groups were both characterized by the RS5 (large deletions of <100kb with microhomology at junction) but BRCA1 to a lesser extent. The most predominant SVs associated with *BRCA1* inactivation are Tandem Duplications (TDs) <100kb (RS3), which were not found in *BRCA2*-/- cases. The TDs has a peak of 2bp microhomology at junction. Another study confirmed similar rearrangement signatures in BRCA-/- tumors<sup>169</sup>. In cell lines, *BRCA1* inactivation was also strongly associated with <10kb tandem duplications<sup>219,223</sup> while *PALB2*, *RAD51C*, *XRCC2* and *XRCC3* inactivation were more associated to large deletions with a signature resembling the RS5<sup>219</sup>.

From these studies we can conclude, that (1) the genomic instability in BRCA1-/- tumors consists in increased number of deletions <100kb, TDs<100kb, inter- and intra-chromosomal translocations; (2) genomic instability in BRCA2-/- tumors consists in increased number of deletions <100kb and inter- and intra-chromosomal translocations; (3) both gene inactivation lead to the accumulation of small indels with high prevalence of more than 5bp deletions with microhomology; (4) single base substitutions spectrum of HRD is characterized by almost uniform distribution.



**FIGURE 9: Rearrangement signatures in 560 Whole Genome Sequencing of breast cancers**

**A]** Rearrangement signatures extracted from the analysis of 560 Whole Genome Sequencing of breast cancers. Y axis: probability of a rearrangement X axis, rearrangement according to the type and the size. Del: deletions; tds: tandem duplication; inv: inversion; trans: translocation. **B]** Cluster groups based on unsupervised hierarchical clustering according to their proportion of rearrangement signatures for each WGS. Each column represents one Whole Genome Sequencing. BRCA1<sup>-/-</sup> cases are indicated in purple and are mostly in group D. BRCA2<sup>-/-</sup> cases are indicated in yellow and are mostly in group G. HER2: human epidermal growth factor receptor 2; PR: progesterone receptor; ER: estrogen receptor. Black bars: negative for the expression of ER or PR or lack of overexpression for HER2. Extracted from Nik-Zainal<sup>150</sup>.

### Combining signatures to extensively describe HRD

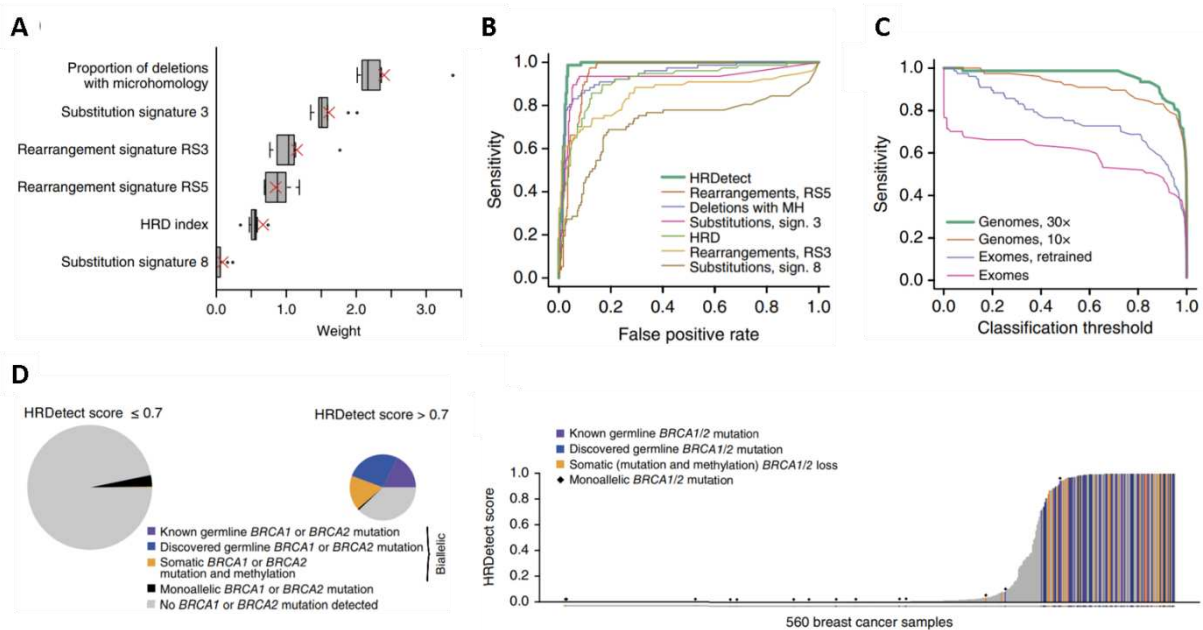
WGS studies revealed specific genomic phenotypes of HRD providing nearly an exhaustive picture of its alteration landscape. Based on this, several teams proposed to integrate all the mutational signatures to create a high precision classifier. HRDetect was developed subsequently to the work of Nik-Zainal et al<sup>224</sup>. This classifier was trained on 77 WGS with bi-

allelic inactivation of *BRCA1/2* and 234 WGS of sporadic breast cancer. All rearrangement signatures previously described alongside eleven base-substitution (SBS1, SBS2, SBS3, SBS5, SBS6, SBS8, SBS13, SBS17, SBS18, SBS20, SBS26) and four small indels signatures were extracted (insertions, deletions with microhomology, deletions at repeats, other type of deletions). The HRD index of the Myriad myChoice® CDx test was also included. A tenfold cross-validation was used to build the classifier HRDetect and six signatures were detected that provide the best separation of *BRCA1/2*-/- and sporadic cancers. In the decreasing order of contribution: deletions flanked by microhomology, SBS3, RS3, RS5, HRD myChoice® CDx index and SBS8 (Figure 10A).

HRDetect supplants all other individual mutational signatures described until now, with a sensitivity and specificity of almost 100% (Figure 10B). *BRCA1* and *BRCA2* phenotypes are reported differently even if not being discriminated formally in HRDetect. The application of HRDetect on *in-silico* down-sampled WGS in the breast cancer cohort still showed 86% sensitivity for *BRCA1/2* inactivated tumors (Figure 10C). No other bi-allelic gene inactivation was associated with a high HRDetect score because of their absence in the cohort (Figure 10D). HRDetect had similar results in ovarian and pancreatic tumors. The association between HRDetect score and actual sensitivity of tumors to treatments showed high correlations in at least two different studies<sup>225,226</sup>. The cost of sequencing and storage along with complexity of the analysis are the principal limiting factor for clinical application of this comprehensive approach.

A more recent pan-cancer classifier for HRD and differentiation of *BRCA1*-/- and *BRCA2*-/- phenotypes named CHORD was built. Similar features contributed to the classifier compared to HRDetect, with different weights however that may be due to its pan-cancer approach<sup>169</sup>. The most important feature was the deletions with microhomology at the breakpoint >1bp and the duplication of <100kb. Interestingly, this classifier looked only at the actual frequency of all possible small-scale alterations and SVs, without relying on the mutational signatures by Alexandrov et al. and Davies et al<sup>4,224</sup>. CHORD showed excellent performance comparable to that of HRDetect on the same samples (99% correspondence), with minimal bias regarding tumor type. CHORD encompasses broader datasets compared to the publication by Davies and colleagues in 2017. Using this power, they could compare bi-allelic inactivation of genes and their scores called by CHORD. On top of *BRCA1* and *BRCA2*, *PALB2* and *RAD51C* genes were both significantly enriched in tumors classified as HRD. They both exhibited *BRCA2*-like phenotype. In a smaller proportion to be significant, bi-allelic mutation of *RAD51B* and *XRCC2* exhibited *BRCA2*-like phenotype for two patients each. Bi-allelic inactivation of genes implicated in *BRCA1*-binding domain namely *BARD1*, *BRIP1* and surprisingly *FANCA* and *FAM175A* were found each in one patient, all presenting *BRCA1*-like phenotype<sup>169</sup>.





**FIGURE 10: Parameters and performance of HRDetect classifier on 560 breast cancers**

**A**| Weight of genomic features used for the construction of HRDetect. The red cross indicates the final weight used in HRDetect after the training of the classifier. **B**| ROC curves for the performance of HRDetect compared to other methods on 371 breast cancer samples. **C**| Comparison of the performance of HRDetect for different technology. HRDetect were retrained on WES. **D**| HRDetect scores according to the mutation status of 560 breast cancer samples. Extracted from Davies et al<sup>224</sup>.

## RAD51-foci, a functional signature of HRD

HRD tumors acquire a specific “scar” of genomic instability during their oncogenesis. The timing of genomic alteration events and their regularity is unknown. Inactivated HR can be restored at some timepoint of tumor evolution reversing HRD. This happens quite frequently upon treatment by DNA damaging agents. However, the genomic scar remains unchanged and all the genomic HRD signatures described above are of little help to distinguish historic and actual HR status.

It has long been attempted to develop a simple functional test of ongoing HRD. RAD51, one of the key proteins of HR, forms foci at the break site that can be visualized in immunofluorescent microscopy<sup>227</sup>. The absence of RAD51 foci was proposed as a marker of ongoing HRD to predict the response to HRD targeting treatments like platinum salts or PARPi. This functional approach assesses actual HR status at the time evaluation. In 2018 Cruz et al. showed the presence of RAD51 foci in PDX and patient samples with initial *BRCA1/2*-/- tumors displaying PARPi resistance<sup>228</sup>. Recent study in TNBC showed high correlation between RAD51 foci test, HRDetect prediction and PARPi sensitivity<sup>229</sup>. This highlights the potential and feasibility of functional assays in clinics. If developed to the clinical application this test could be a cost-efficient method for routine HRD testing.

## Possible etiology of HRD signatures

In this final section of the chapter, I will try to connect mutational and structural variant signatures observed in HRD to known or suspected etiology shaping the phenotype. A recapitulation of those signatures is presented in Table 1.

**Table 1: Mutational signatures and possible etiology associated with HRD**

Signature	Description	Depleted genes	Possible aetiology
SBS3	Uniform distribution of mutations across all 96 bases substitution	<i>BRCA1, BRCA2, PALB2, RAD51C</i> ( <i>XRCC2, XRCC3, RAD54</i> )	Pol-θ mediated
ID6	Deletions majorly ≥5 bp and ≥2bp microhomology at junction	<i>BRCA1, BRCA2, PALB2</i> ( <i>RAD51C, XRCC2, XRCC3, RAD54</i> )	Pol-θ mediated
ID8	Deletions majorly ≥5 bp and 0-3bp microhomology at junction	<i>BRCA1, BRCA2, PALB2</i>	DSB repair by C-NHEJ and/or Pol-θ mediated
RS3	1–100 kb tandem duplications and microhomology peak at 2bp at breakpoint junctions	<i>BRCA1</i> ( <i>BARD1, BRIP1</i> )	Pol-θ mediated
RS5	<100 kb deletions and microhomology peak at 2bp at breakpoint junctions and with larger microhomology (>10 bp)	<i>BRCA2, PALB2, BRCA1</i> ( <i>PALB2, RAD51B, RAD51C, XRCC2, XRCC3</i> )	SSA mediated

Brackets: mutational signatures associated with genes inactivation found in cell-lines or by the classifier CHORD. Adapted from Stok et al<sup>230</sup>.

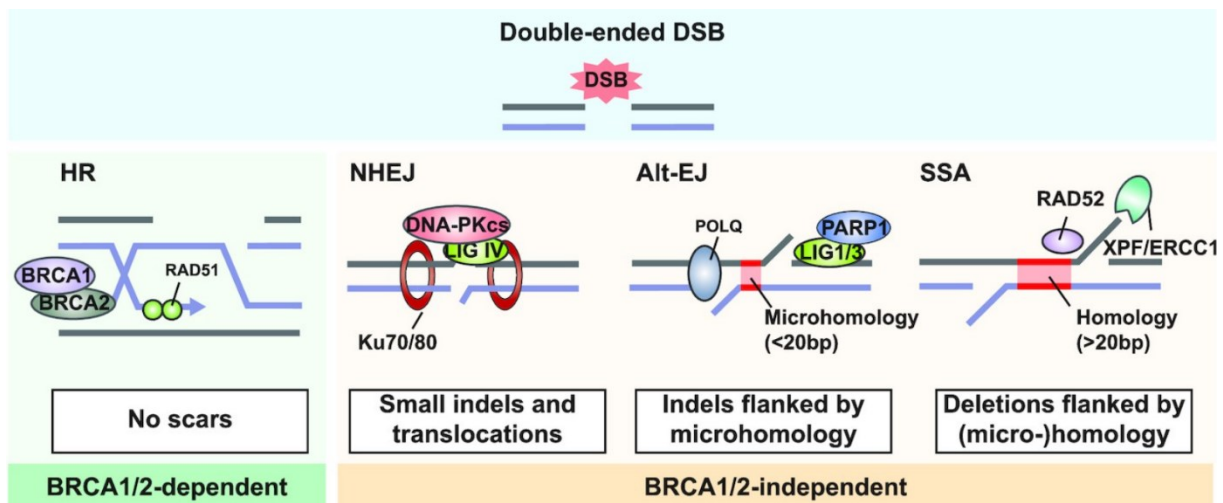
The accumulation of certain genomic alterations in HRD is mainly attributed to DSB reparation by alternative pathways. DSB in a HR deficient context can be handled by C-NHEJ, Alt-EJ and SSA. Those pathways are reportedly more error-prone than HR especially Alt-EJ and SSA (Figure 11).

C-NHEJ is active through all the cell-cycle. C-NHEJ fast establishment plays an important protection role in genome integrity, preventing chromosomal translocations<sup>231</sup> or promoting rearrangements under replicative stress<sup>232</sup>. C-NHEJ may results in small deletions/insertions (1-5bp) with no to little random microhomology flanking the break sites<sup>15,233</sup>. Thus, ID8 signature with 0 to 3 bp microhomology could be enriched via C-NHEJ DSB repair. Interestingly, this signature contributes mildly to BRCA deficient tumors landscape, emphasizing the potential lower contribution of C-NHEJ in HRD cells compared to other DSB repair pathway<sup>224</sup>.

Another possible reparation pathway is Alt-EJ. HR-deficient tumors have been reported to strongly rely on this pathway<sup>185</sup> and the inactivation of Pol-θ is reported synthetic lethal in HR deficient context<sup>195</sup>. Alt-EJ leads to frequent deletions of 20 to 200 bp between homology

regions<sup>234,235</sup> which footprint could be caught by ID6 signature. Moreover, Pol-θ is a low fidelity polymerase favoring base substitution<sup>236</sup> and may contribute at least partly to the SBS3 signature. ID6 and SBS3 are strongly correlated<sup>4</sup> and highly contribute to the HRD classifiers<sup>4,150,169</sup>. Alt-EJ is an important alternative repair pathway in HR-deficient tumors and may be at the origin of these signatures<sup>185,237</sup>. Additionally, several proposed models points Alt-EJ as the cause of TDs observed in BRCA1-/- tumors<sup>223,238</sup>.

Another annealing repair pathway relying on larger DNA-end resection is SSA. SSA can use the same substrate as HR and compete with this pathway. BRCA1-PALB2 complex was shown to directly promote HR and suppress SSA<sup>239</sup>. SSA leads to deletions between the homology regions, reportedly larger than deletions resulted from Alt-EJ<sup>240</sup>. ID6 and RS5 signatures both including large deletions flanked by homology of >2bp and >10bp (RS5) are likely to be attributed to SSA-mediated DSB repair.



**FIGURE 11: Alternative DSB repair pathway leads to small indels than flanked by (micro-)homology**

Extracted from Stok et al<sup>230</sup>.

## Conclusion

The concept of HRD being initially associated with cancer predisposition syndrome in *BRCA1/2* mutation carrier, has now expanded to general phenotype found in many cancers from nearly 50% of cases to occasional occurrence. The comprehensive characterization of the genomic phenotype permitted by modern DNA sequencing lead to quasi-complete description of genomic alterations in HRD tumors. Two classifiers, HRDetect and CHORD, build based on WGS showed great performance in HRD detection and allowed the attribution of genomic phenotype to inactivated genes from HR pathway and detection of HRD tumors without any apparent mutation in known genes. The list of genes which inactivation could potentially lead

to cancer HRD is not yet exhaustive due to the rarity of their mutations. HRD got attention in the recent years because of the invention of the targeted treatment, PARPi, to which HRD tumors are particularly sensitive. Because HRD is frequent in breast cancers (which is one of the common cancers) and PARPi became widely available for patients, the organization of routine HRD testing in clinics is of importance. Simple and efficient methods for HRD detection, which can be implemented in clinics, are needed.

# ARTICLES

## Introduction: “shallowHRD: detection of homologous recombination deficiency from shallow whole genome sequencing”

Genomic HRD phenotype is nowadays well described and shown to be detectable by many approaches. The exhaustive testing of HRD through genomic signatures can be obtained from high covered WGS (>30X). This is however technically complex, the data generation is costly, its processing is long, and its storage cumbersome. This complicates the implementation of an exhaustive genomic HRD testing in a clinical setting. Our goal was to develop a robust method, which could be routinely applicable in clinics for retrospective studies and for actual patient diagnostics. Sequencing facilities are now widely in place in cancer centers and the reduction of the sequencing cost make WGS at a low coverage (<3X) available at affordable prices. Denoted here as shallow WGS (sWGS), whole genome sequencing at a low coverage (down to <1X) can potentially be used to detect HRD from the cancer genomic profile. However, its low coverage makes it difficult to robustly detect Allelic Imbalance, LOH, mutations, structural variations or rearrangement signatures of HRD (SBS3, ID6, ID8, RS5, RS3). The performance of those approaches directly relies on high coverage and multiple reads at a given location. On the other hand, sWGS is reportedly suitable for reconstructing Copy Number tumor genomic profile without a matched normal sample, even for formalin-fixed, paraffin-embedded (FFPE) samples largely used in clinic<sup>241,242</sup>.

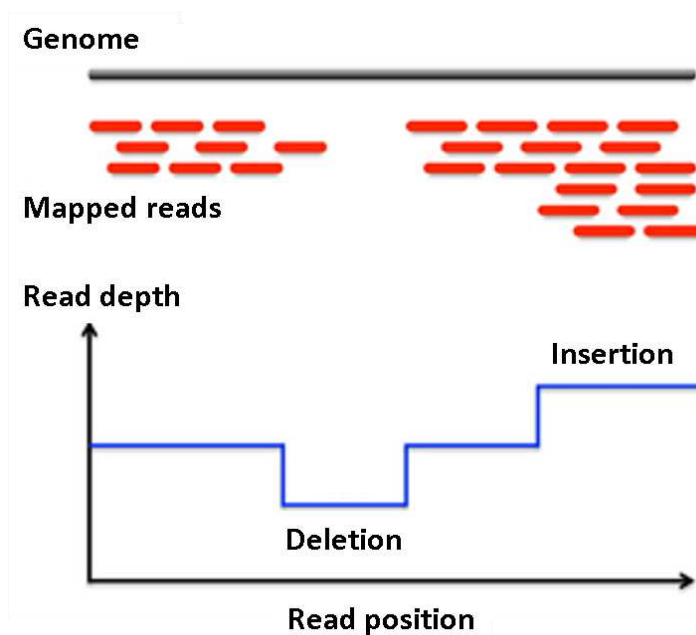
We developed *shallowHRD* a new method of HRD testing based on sWGS and Copy Number Aberrations (CNAs). A CNA corresponds to somatic gains or losses of a chromosome region. SNParrays showed a high level of correspondence between HRD and the number of large CNAs through the LST signature, justifying therefore to tackle this in sWGS. The *shallowHRD* method and an application was published in *Bioinformatics* as an application note.

### Workflow and pipeline

#### Alignment and read depth approach

To process sWGS, we first aligned the sequencing files to a human reference genome. This was done using *bwa-mem*, a high-efficient aligner and one of the most commonly used for small-reads DNA sequencing<sup>243-245</sup>. The construction of the tumor copy number genomic profile and the subsequent investigation of CNA was then done through a Read-Depth (RD) analysis.

RD analysis is based on the underlying assumption that the coverage of a genomic region is positively correlated with its number of chromosomal copies<sup>246</sup>. This principle is represented in Figure 12. For this method, the number of reads in fixed windows of a given size along the genome is counted. Those windows can be overlapping, providing a better resolution at the cost of computational time. RD analysis is adapted for sWGS using large windows to count reads, which compensated the lack of coverage and allowed to build a high-quality genomic profile<sup>247,248</sup>. The number of reads inside each window is influenced by several biases that needs to be corrected<sup>249,250</sup>.



**FIGURE 12: Read Depth analysis to detect Copy Number changes with Next-Generation Sequencing**

The number of reads is positively correlated to the number of chromosomal copies in a genomic region. It allows to construct a genomic profile and infers Copy Number Aberrations. Adapted from Valsesia et al<sup>251</sup>.

### Profile normalization

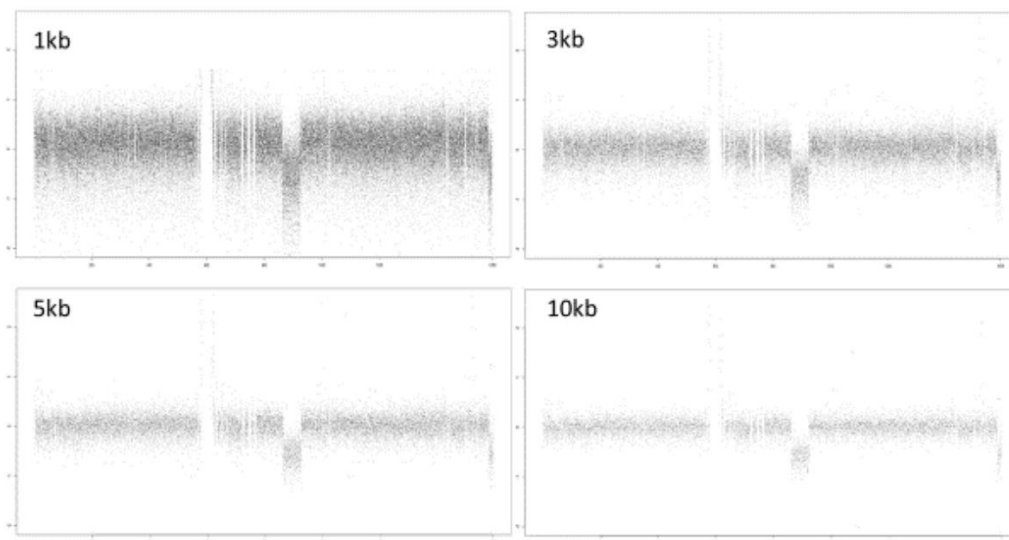
To build a copy number genomic profile from the aligned reads we tested several tools including ControlFREEC developed in Institut Curie<sup>252</sup>, QDNAseq<sup>241</sup> and ichorCNA<sup>253</sup>. These three tools do not need a matched normal sample to correct for read count biases. They rely on different models to correct the two main biases observed in a Read-Depth approach: GC-content and mappability. GC-content is the proportion of base that are either G or C in a given genomic region. Both AT-rich and GC-rich regions present a lower coverage<sup>250</sup>. "Mappability" of a chromosomal region corresponds to the uniqueness of a DNA genomic region. This bias comes from repetitive regions that introduce mistakes or ambiguous assignment of reads at

multiple locations during alignment<sup>249</sup>. GC-content and mappability biases are corrected either independently and sequentially by ControlFREENC and ichorCNA or jointly by QDNAseq.

Additionally, QDNAseq filters out problematic genome regions based on the ENCODE blacklist regions<sup>254</sup> and their own list they developed on the healthy cases<sup>241</sup>. Despite those differences, the genomic profiles for the three RD approaches exhibited minimal differences for varying size of windows. QDNAseq presented a smoother copy number profile for small genomic regions, but the effect of this improvement on large CNA is dubious.

### Selecting of a bin size

Within profile normalization pipeline it was necessary to select bin (window) size adapted for sWGS. For a read size of 100bp and a coverage of 1X, the expected number of reads for windows of 5kb is 50. This expected number can variate depending on the genomic region and alignment errors. Decreasing the size of the window for read count below 5kb introduce variations, affecting the stability of CNA profile. Increasing the window size is however well tolerated and produce similar CNA profiles up to 200kb. The profile of chromosome 7 for a breast tumor for different window size is presented in Figure 13. Variations are important between close dots for windows of 1kb and 3kb, which encouraged us to take larger windows. In-house, our current preferred bin size is 50 kb to anticipate a coverage that would drop way below 1X.

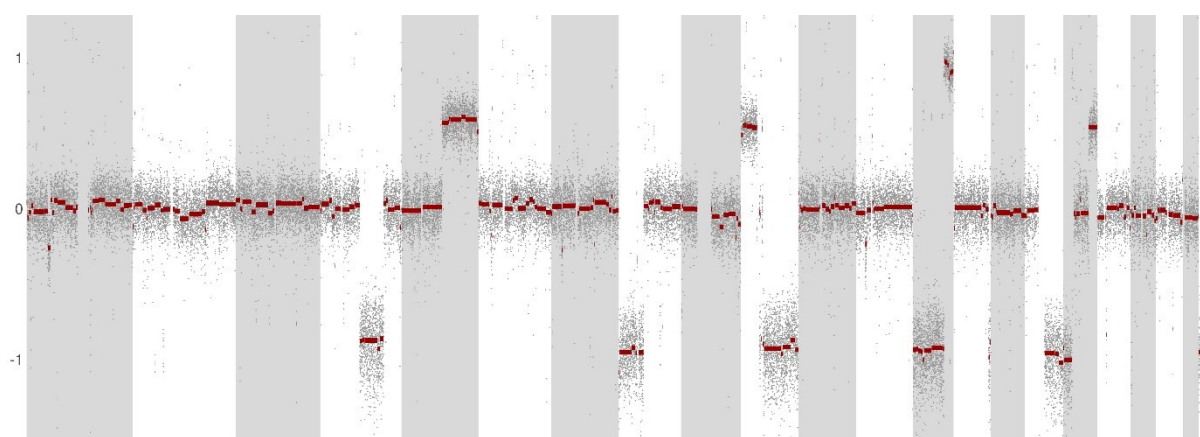


**FIGURE 13: Genomic profile from sWGS of the chromosome 7 of a breast tumor generated with a Read-Depth approach with ControlFREENC and for different size of windows**

Personal data.

## CNA profile segmentation and optimization

Copy number profile segmentation is essential for robust estimation of large-scale copy number breaks. This segmentation can be done by several approaches. ControlFREEEC uses a LASSO-based algorithm<sup>255</sup>, QDNAseq relies on the Circular Binary Segmentation algorithm<sup>256,257</sup> while ichorCNA use Hidden Markov Models<sup>253</sup>. The segmentation resulting from ichorCNA was too oversimplified compared to the two other tools, which could result in missing altered genomic regions. We have chosen QDNAseq and controlFREEEC which had slightly oversensitive segmentation that is further optimized using *shallowHRD*. A segmented copy number profile generated by a Read Depth approach is represented below in Figure 14. The segmented profile is then optimized to eliminate over-segmentation.



**FIGURE 14: Genomic profile from tumor sWGS generated with controlFREEEC**

y-axis: Value of read count corrected for GC-content and normalized. Grey and white stripes correspond to chromosome (1 to 22). Red line: segment that associated contiguous windows to the same copy number level. Grey point: Read count of a fixed window normalized for GC-content and mappability. Personal data.


In sWGS, the absolute copy number of a segment is difficult to infer because of the lack of allelic balance information that can be extracted with the low coverage. Hence, we based our method on relative copy number changes and a minimal difference between segments, designated as  $M$  in the article.  $M$  is based on the overall pairwise comparison of the large segments in the initial segmentation. It is used to optimize the genomic profile by uniting the segments belonging to the same copy number and extract large CNA to predict HRD.

The article describing *shallowHRD* was published in *Bioinformatics* as an application note. The detailed computation for profile optimization is provided in this article. ControlFREEEC was used because it was available for cloud computing. The supplementary data of the article are included as they are an important part for the overall validation of the results present in the article. A few relevant cohorts that we analyzed were presented in conclusion of the article.



## Genome analysis

# ShallowHRD: detection of homologous recombination deficiency from shallow whole genome sequencing

Alexandre Eeckhoutte <sup>1,2,\*</sup>, Alexandre Houy<sup>1,2</sup>, Elodie Manié<sup>1,2</sup>, Manon Reverdy<sup>1,2</sup>, Ivan Bièche<sup>3</sup>, Elisabetta Marangoni<sup>2,4</sup>, Oumou Goundiam<sup>2,4</sup>, Anne Vincent-Salomon<sup>5</sup>, Dominique Stoppa-Lyonnet<sup>1,6</sup>, François-Clément Bidard<sup>7,8</sup>, Marc-Henri Stern<sup>1,2,3</sup> and Tatiana Popova<sup>1,2</sup>

<sup>1</sup>DNA Repair and Uveal Melanoma (D.R.U.M.), Inserm U830, Institut Curie, Paris 75248, France, <sup>2</sup>Institut Curie, PSL Research University, Paris 75005, France, <sup>3</sup>Department of Genetics, Institut Curie, Paris 75248, France, <sup>4</sup>Department of Translational Research, Institut Curie PSL Research University, Paris 75248, France, <sup>5</sup>Department of Biopathology, Institut Curie PSL Research University, Paris 75005, France, <sup>6</sup>Faculty of Medicine, University of Paris, Paris, France, <sup>7</sup>Department of Medical Oncology, Institut Curie PSL Research University, Paris 75248, France and <sup>8</sup>Versailles Saint Quentin en Yvelines University, Paris Saclay University, Versailles 78035, France

\*To whom correspondence should be addressed.

Associate Editor: Inanc Birol

Received on January 3, 2020; revised on March 25, 2020; editorial decision on April 13, 2020; accepted on April 14, 2020

## Abstract

**Summary:** We introduce *shallowHRD*, a software tool to evaluate tumor homologous recombination deficiency (HRD) based on whole genome sequencing (WGS) at low coverage (shallow WGS or sWGS; ~1X coverage). The tool, based on mining copy number alterations profile, implements a fast and straightforward procedure that shows 87.5% sensitivity and 90.5% specificity for HRD detection. *shallowHRD* could be instrumental in predicting response to poly(ADP-ribose) polymerase inhibitors, to which HRD tumors are selectively sensitive. *shallowHRD* displays efficiency comparable to most state-of-art approaches, is cost-effective, generates low-storable outputs and is also suitable for fixed-formalin paraffin embedded tissues.

**Availability and implementation:** *shallowHRD* R script and documentation are available at <https://github.com/aeckhou/hou/shallowHRD>.

**Contact:** alexandre.eeckhoutte@curie.fr

**Supplementary information:** [Supplementary data](#) are available at *Bioinformatics* online.

## 1 Introduction

Aggressive subtypes of breast and ovarian cancers are frequently associated with homologous recombination deficiency (HRD) making these tumors sensitive to poly(ADP-ribose) polymerase inhibitors (Coleman *et al.*, 2019). HRD arises upon inactivation of *BRCA1/2*, *RAD51C* or *PALB2* and is characterized by specific tumor genome instability (Nik-Zainal *et al.*, 2016; Staaf *et al.*, 2019). Even though HRD genes are mostly known, exhaustive testing of their inactivation is difficult. This motivates developing surrogate genomic markers of HRD. Recent developments based on high throughput sequencing, HRDetect, Signature 3, SigMA, scarHRD, achieved excellent capacity to evaluate HRD (Davies *et al.*, 2017; Gulhan *et al.*, 2019; Polak *et al.*, 2017; Sztupinski *et al.*, 2018). However, these methods are technically complex, time- and data-storage consuming, often need a matched normal sample and can be costly.

We introduce *shallowHRD*, a software for HRD testing based on the number of large-scale genomic alterations (LGA) obtained from

whole genome sequencing (WGS) at low coverage (shallow WGS or sWGS; ~1X). sWGS robustly detect copy number alterations (CNAs), even in fixed-formalin paraffin embedded (FFPE) samples and liquid biopsies (Van Roy *et al.*, 2017) at low cost and with easy-storable outputs. The concept of LGAs follows single-nucleotide polymorphism (SNP) array approaches, exploiting an increased number of large-scale intra-chromosomal CNAs characteristic of HRD (Abkevich *et al.*, 2012; Birkbak *et al.*, 2012; Popova *et al.*, 2012).

## 2 Materials and methods

### 2.1 Data

In-house sWGS of breast and ovarian cancers (26 primary tumors, 39 patient-derived xenografts from frozen blocks and 4 primary tumors FFPE) and down-sampled to ~1X WGS (108 normal tissues, 79 primary tumors from the TCGA breast cancer) were processed

by Control-FREEC (v11.5) (Boeva *et al.*, 2012) (Supplementary Material).

## 2.2 shallowHRD

The tool takes as input ‘sample\_name.bam\_ratio.txt’, which includes CNA profile  $\{x, g\}_{1, N}$  where  $x$  is normalized read counts in a sliding window,  $g$  is genomic coordinate and the profile segmentation with  $S_i, Z_i$  segment median and size (in megabases, Mb).

### 2.2.1 Workflow

- CNA cut-off is detected and the profile segmentation is optimized as follows: Segments are defined as ‘large’ if  $Z_i \geq (Q_3 - Q_1)/2$ , where  $Q_1, Q_3$  are quartiles of  $Z_i$  ( $Z_i > 3$  Mb) distribution.  $M$  is detected as the first local minimum of  $(S_i - S_j)$  density, where  $i, j$  are large segments (Supplementary Fig. S1). CNA cut-off =  $\min(\max(0.025, M), 0.45)$ . Adjacent segments are merged if  $(S_i - S_{i+1}) < \text{CNA cut-off}$ ; starting from the largest segment.
- LGAs, defined as intra-chromosome arm CNA breaks with adjacent segments  $Z_i, Z_{i+1} \geq 10$  Mb, are counted after removing segments  $< 3$  Mb.
- The sample is annotated as ‘non-HRD’ (LGA  $< 15$ ), ‘borderline’ ( $15 \leq \text{LGA} \leq 19$ ) or ‘HRD’ (LGA  $> 19$ ).
- Sample quality is defined by  $M$  and  $cMAD$ ,  $cMAD = \text{median}((x - S_x))$ , where  $S_x$  corresponds to the segment enclosing  $x$ , before optimization: ‘bad’ ( $cMAD > 0.5$  |  $cMAD > 0.14$  and  $M > 0.45$ ), ‘average’ ( $cMAD > 0.14$  and  $M < 0.45$  |  $cMAD < 0.14$  and  $M > 0.45$ ) or ‘normal or highly contaminated’ ( $M < 0.025$ ) (Supplementary Material and Fig. S2).
- CCNE1 amplification is called if  $S_c \geq 4 \cdot \text{CNA cut-off}$ , where  $c$  is the segment enclosing the gene (4 was set arbitrarily).

shallowHRD output contains: (A) Tumor genome profile. (B) Density plot for CNA cut-off. (C) CNA segmentation summary. (D) Sample quality and HRD diagnostics (Supplementary Fig. S3).

## 3 Results

In-house sWGS and down-sampled WGS of normal samples (TCGA) were employed to develop the sWGS methodology similar to the large-scale state transitions (LST) in SNP-arrays (Popova *et al.*, 2012) (Section 2). LGAs inferred from sWGS corresponded well to the LSTs with identical HRD calls for 8 primary tumors tested (76–97% match in segments  $\geq 10$  Mb) (Supplementary Fig. S4). sWGS coverage  $> 0.3X$  provide adequate quality, also for FFPE (Supplementary Figs. S2 and 5).

Validation by down-sampled WGS (TCGA) showed LGA to be coherent to SNP-arrays LST ( $r = 0.92$ ; slope = 0.88;  $P < 2.2e-16$ , Pearson) with increased discrepancy in average quality samples



Fig. 1. shallowHRD validation in down-sampled WGS of the TCGA (A) and performance (B). Proven/No HRD: cases with/without inactivation of BRCA1/2, RAD51C or PALB2 (Supplementary Material); HRD (red) and non-HRD (blue) cases in SNP-arrays; LGAs: large-scale genomic alterations; WES: whole exome sequencing. \*Low specificity could be due to non-complete annotation of HRD

( $n = 13$ ), and HRD diagnostics discordant in three and borderline in four cases (Fig. 1A; Supplementary Material, Supplementary Figs. S6 and 7, Supplementary Table S1). CCNE1 amplification was found in four non-HRD cases, in-line with previous observations of almost mutual exclusivity with HRD (Goundiam *et al.*, 2015). Thus, sWGS LGAs is suitable to take over the SNP-array LSTs, which is a clinically validated method for HRD detection.

Tumor content for sWGS limits to  $> 0.3$  as estimated from the TCGA and *in silico* dilution series (Supplementary Material, Supplementary Figs. S8 and 9).

Fifteen and 20 LGAs represent soft and stringent cut-offs with sensitivity of 87.5% and 81.25% (16 cases HRD) and specificity of 90.5% and 95.2% (63 non-HRD cases), respectively, which is compatible with other state-of-the-art approaches (Fig. 1B).

To conclude, shallowHRD implements a fast and straightforward evaluation of tumor HRD in breast, ovarian and other cancers such as pancreatic or prostatic, performing similar to most state-of-the-art approaches, the technique is cheap and suitable for all type of samples.

## Acknowledgements

The results here are in part-based upon data generated by the TCGA Research Network: <http://cancergenome.nih.gov/>. The authors thank The Seven Bridges Cancer Genomics Cloud for computational facilities.

## Funding

This work was supported by the Ligue Nationale Contre le Cancer (to A.E.).

### Conflict of Interest:

E. Manić, T. Popova and M.-H. Stern are co-inventors of the LST method (US20170260588, US20150140122 and exclusive Licence to Myriad Genetics).

## References

- Abkevich, V. *et al.* (2012) Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. *Br. J. Cancer*, 107, 1776–1782.
- Birkbak, N.J. *et al.* (2012) Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. *Cancer Discov.*, 2, 366–375.
- Boeva, V. *et al.* (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics*, 28, 423–425.
- Coleman, R.L. *et al.* (2019) Veliparib with first-line chemotherapy and maintenance therapy in ovarian cancer. *N. Engl. J. Med.*, 381, 2403–2415.
- Davies, H. *et al.* (2017) HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. *Nat. Med.*, 23, 517–525.
- Goundiam, O. *et al.* (2015) Histo-genomic stratification reveals the frequent amplification/overexpression of CCNE1 and BRD4 genes in non-BRCAness high grade ovarian carcinoma. *Int. J. Cancer*, 137, 1890–1900.
- Gulhan, D.C. *et al.* (2019) Detecting the mutational signature of homologous recombination deficiency in clinical samples. *Nat. Genet.*, 51, 912–919.
- Nik-Zainal, S. *et al.* (2016) Landscape of somatic mutations in 560 breast cancer whole-genome sequences. *Nature*, 534, 47–54.
- Polak, P. *et al.* (2017) A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. *Nat. Genet.*, 49, 1476–1486.
- Popova, T. *et al.* (2012) Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. *Cancer Res.*, 72, 5454–5462.
- Staa, J. *et al.* (2019) Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. *Nat. Med.*, 25, 1526–1533.
- Sztupinski, Z. *et al.* (2018) Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. *NPJ Breast Cancer*, 4, 16.
- Van Roy, N. *et al.* (2017) Shallow whole genome sequencing on circulating cell-free DNA allows reliable noninvasive copy-number profiling in neuroblastoma patients. *Clin. Cancer Res.*, 23, 6305–6314.

## Supplementary Data

ShallowHRD: Detection of Homologous Recombination Deficiency from shallow Whole Genome Sequencing (A. Eeckhoutte et al.)

Supplementary Methods .....	2
In-house Whole Genome Sequencing (sWGS).....	2
WGS from the TCGA .....	2
Configuration file Control-FREEC .....	2
Quality control .....	3
HRD annotation .....	3
Soft and stringent HRD cut-offs and borderline HRD.....	3
Estimation of tumor content in WGS.....	4
<i>In silico</i> dilution series based on sWGS.....	4
Supplementary References.....	5
Supplementary Figures .....	6
Figure S1. Distribution of pairwise differences between segment medians in CNA profile .....	6
Figure S2. Two parameters characterizing quality of CNA profile .....	7
Figure S3. Reports of shallowHRD .....	9
Figure S4. Consistency in the large CNA segments in sWGS and in SNP-arrays .....	10
Figure S5. FFPE profiles from sWGS analyzed by shallowHRD .....	11
Figure S6. Large-scale CNA correspondence between sWGS and SNP-arrays .....	12
Figure S7. Tumor with high number of copy-neutral LOH .....	13
Figure S8. Example of <i>in silico</i> dilution series .....	14
Figure S9. Tumor content and performance of shallowHRD .....	16
Supplementary Table S1. Validation cohort of down-sampled WGS from the TCGA.....	17

## Supplementary Methods

### In-house Whole Genome Sequencing (sWGS)

DNA was extracted from frozen blocs (26 primary tumors, 39 Patient-Derived Xenografts, PDX) and Fixed-Formalin Paraffin Embedded (FFPE) tissues (4 primary tumors) and was sequenced on HiSeq2500 or NovaSeq (Illumina; 100bp paired-end library; coverage 0.06-1.65X; 4-6X for FFPE) and aligned on hg19 and hg38 by BWA-MEM (v0.7.15) (Li and Durbin, 2009); PDX were purified from mouse reads using Xenofilter (Kluin, et al., 2018). Optical/PCR duplicates were filtered by PicardTools (v1.140) (<http://broadinstitute.github.io/picard/>) and supplementary alignments were removed by Samtools (v1.9) (Li, et al., 2009).

### WGS from the TCGA

WGS from the breast cancer TCGA-BRCA cohort (Weinstein, et al., 2013) (108 normal tissues, 79 primary tumors) were down-sampled to 1X by Sambamba (v0.5.9) (Tarasov, et al., 2015) on the Cancer Genomics Cloud of SevenBridges (Lau, et al., 2017).

### Configuration file Control-FREEC

[general]

ploidy = 2,4  
window = 40000  
step = 20000

breakPointThreshold = 0.65  
breakPointType = 2  
forceGCcontentNormalisation = 1

uniqueMatch = FALSE  
contaminationAdjustment = TRUE

samtools = /path/to/samtools

chrFiles = /path/to/chromFa/  
chrLenFile = /path/to/hg19.len  
gemMappabilityFile = /path/to/out100m2\_hg19.gem

outputDir = /path/to/outputDir

BedGraphOutput = FALSE

[sample]

mateFile = /path/to/file.bam  
inputFormat = BAM  
matesOrientation = FR

## Quality control

GC-corrected and normalized read counts profiles of sWGS and their sensitive segmentations (number of segments 300-1600), were annotated manually as “good” (n=55), “average” (n=6) or “bad” (n=8). Based on this annotation quality thresholds were defined.

Bad quality cases represented mainly sequencing failure independent of coverage, with frequent (n=5) poorly detectable local minimum in  $M$ , separating fluctuations of segments with equal copy numbers from one copy difference. Average quality was mainly due to a low coverage (0.06-0.3X) displaying high fluctuations in the number of reads per window characterized by  $cMAD$ .  $cMAD > 0.14$  and  $M > 0.45$  indicate low quality samples (Fig. S2). Coverage 0.3X is a low limit for sWGS to ensure prominent CNA profile.

After evaluating 108 down-sampled WGS of normal samples, a lower boundary for *CNA cut-off* was set to 0.025, to avoid CNA detection in normal and over-segmentation in low tumor content samples.

## HRD annotation

***In-house cases:*** In-house tumor cases were partially tested on the Institute Curie platform.

***TCGA cohort:*** HRD annotation of the TCGA cohort was previously described (Manie, et al., 2016). Briefly, mutations in *BRCA1/2*, *RAD51C* and *PALB2* genes were searched in whole exome sequencing (WES) data; gene inactivation was considered proven when deleterious mutation and LOH (Loss Of Heterozygosity) were observed at the gene locus or two deleterious mutations found in the gene; missense mutations annotated as pathogenic in COSMIC database were considered deleterious. *BRCA1* and *RAD51C* promoter methylation was checked using the gene expression; cases with outlier low expression were annotated as HRD due to promoter methylation.

### ***Specificity of HRD calls in SNP-array LST and scarHRD:***

LST was validated on the TCGA cohort, which at the time of publication (Manie, et al., 2016) was not completely available for direct search and verification of the reported mutations. This explains relatively low specificity of LST method shown in Fig.1B. In the current validation set of the TCGA down-sampled WGS, specificity of LST method was very close to LGA in sWGS (predictions of SNP-array based method are indicated by colors in Fig.1A).

For scarHRD (Sztupinski, et al., 2018), the methylation of *RAD51C* promoter was not assessed, which might led to missing HRD cases.

## Soft and stringent HRD cut-offs and borderline HRD

Two cut-offs, soft and stringent, were introduced on the LGA number to call HRD or nonHRD. The reason for this is the appearance of HRD in breast and ovarian tumors: while the majority of cases with BRCAness (HRD) have LGA number far higher than 20, small proportion of mainly *BRCA2* mutated tumors display near-diploid genome with ~15 large-scale chromosomal breaks. From the other hand, nonHRD tumors with near-tetraploid genomes can display 15-20 large-scale chromosomal breaks. When the tumor ploidy is known, there is no problem to distinguish these two situations.

For sWGS, ploidy estimation is problematic and could introduce additional uncertainty. To overcome this issue and to bring additional attention to the low confidence of the call we introduced borderline HRD.

The Supplementary Table S1 recapitulate all the TCGA down-sampled WGS cases processed including their ID, HRD diagnostic with shallowHRD and SNP-arrays, automatic quality detection and correspondence of large segment between sWGS and SNP-array. Cases with contradictory calls are commented.

### **Estimation of tumor content in WGS**

We used estimation of tumor content inferred from the SNP-arrays by GAP method (Popova, et al., 2009) and ichorCNA (Adalsteinsson, et al., 2017) to directly estimate tumor content in sWGS and in the dilution series using window of 50kb on all autosomal chromosomes.

### ***In silico* dilution series based on sWGS**

To obtain tumor content limitation for *shallowHRD* we performed *in silico* dilution of 7 in-house sWGS by 1 sWGS with quasi-normal genome (Supplementary Figure S8A). These 8 cases were sequenced in the same batch. The dilution series was done using picardTools MergeSam (<http://broadinstitute.github.io/picard/>), recursively merging seven times the BAM file of the tumor with “quasi-normal” profile with the BAM files of other cases. The effect of the dilution is shown in Supplementary Figs. S8 B, C and D.

Three chromosomes which carried some CNA in the “quasi-normal” profile (chromosomes 3, 5 and 17) after controlFREEC processing were masked for CNA cut-off determination and LGA counting. The number of LGAs according to those dilutions is represented in Supplementary Figure S9B. *shallowHRD* presents relatively stable results with mild variation in LGA counts even for high number of sequential dilutions in good quality cases.

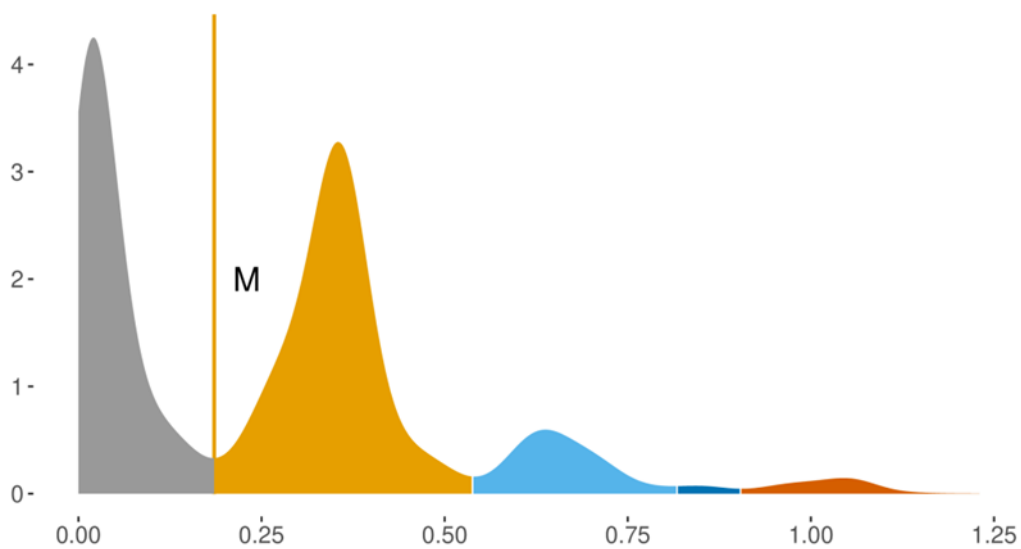
The tumor content estimation was based on the initial tumor content from SNP-array and calculated as proportion of mapped reads in the undiluted sWGS and the diluter sWGS. Estimations of tumor content inferred with ichorCNA (designed for cfDNA) were taken for comparison.

Even though sWGS show stable results around very low tumor content (~0.1), 0.3 could be considered as a good limit for the method application. Tumor cellularity is not directly assessed in *shallowHRD*, but rather taken into account to some extent in the automatic quality control procedure.

## Supplementary References

- Adalsteinsson, V.A., *et al.* (2017) Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors, *Nat Commun*, **8**, 1324.
- Chin, S.F., *et al.* (2018) Shallow whole genome sequencing for robust copy number profiling of formalin-fixed paraffin-embedded breast cancers, *Exp Mol Pathol*, **104**, 161-169.
- Kluin, R.J.C., *et al.* (2018) XenofilteR: computational deconvolution of mouse and human reads in tumor xenograft sequence data, *BMC Bioinformatics*, **19**, 366.
- Lau, J.W., *et al.* (2017) The Cancer Genomics Cloud: Collaborative, Reproducible, and Democratized-A New Paradigm in Large-Scale Computational Research, *Cancer Res.*, **77**, e3-e6.
- Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754-1760.
- Li, H., *et al.* (2009) The Sequence Alignment/Map format and SAMtools, *Bioinformatics*, **25**, 2078-2079.
- Manie, E., *et al.* (2016) Genomic hallmarks of homologous recombination deficiency in invasive breast carcinomas, *Int. J. Cancer*, **138**, 891-900.
- Popova, T., *et al.* (2009) Genome Alteration Print (GAP): a tool to visualize and mine complex cancer genomic profiles obtained by SNP arrays, *Genome Biol*, **10**, R128.
- Robbe, P., *et al.* (2018) Clinical whole-genome sequencing from routine formalin-fixed, paraffin-embedded specimens: pilot study for the 100,000 Genomes Project, *Genet Med*, **20**, 1196-1205.
- Scheinin, I., *et al.* (2014) DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly, *Genome Res.*, **24**, 2022-2032.
- Sztupinski, Z., *et al.* (2018) Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer, *NPJ Breast Cancer*, **4**, 16.
- Tarasov, A., *et al.* (2015) Sambamba: fast processing of NGS alignment formats, *Bioinformatics*, **31**, 2032-2034.
- Van Roy, N., *et al.* (2017) Shallow Whole Genome Sequencing on Circulating Cell-Free DNA Allows Reliable Noninvasive Copy-Number Profiling in Neuroblastoma Patients, *Clin Cancer Res*, **23**, 6305-6314.
- Weinstein, J.N., *et al.* (2013) The Cancer Genome Atlas Pan-Cancer analysis project, *Nat Genet*, **45**, 1113-1120.

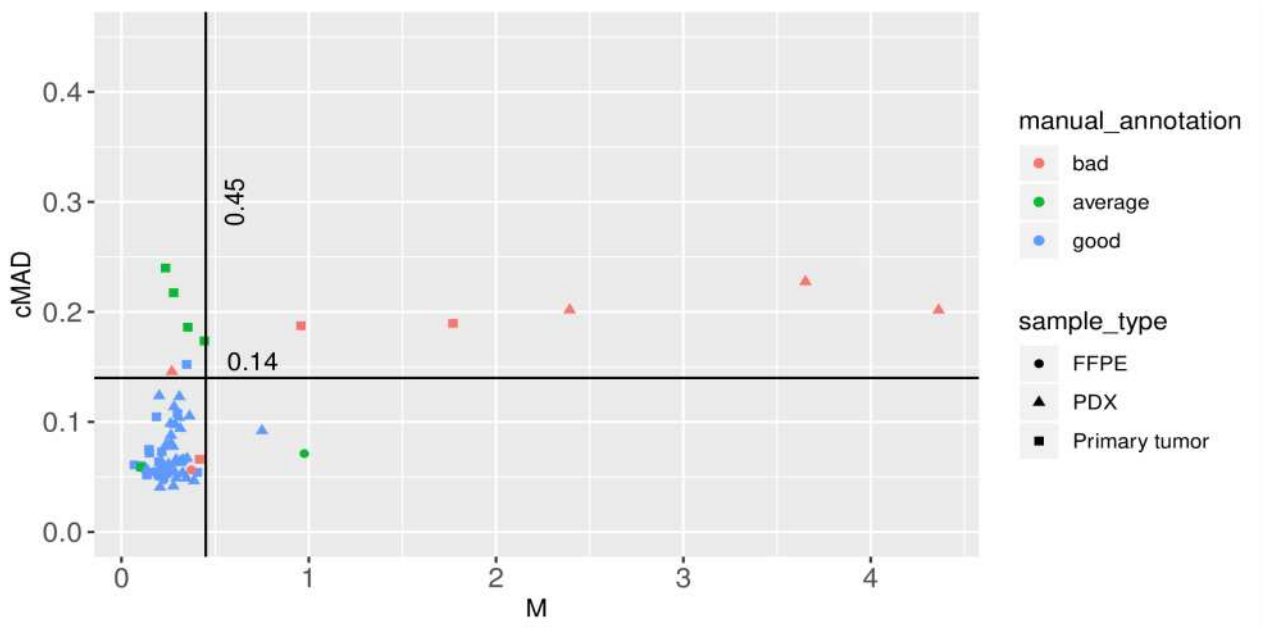
## Supplementary Figures



**Figure S1. Distribution of pairwise differences between segment medians in CNA profile**

An example of density plot of pairwise differences between segment medians is shown (only segments  $> 3\text{Mb}$  were considered). The first (grey) pick corresponds to fluctuations in segment medians related to the same copy number; the second (yellow) pick corresponds to fluctuations around the one copy difference; the third (light-blue) pick corresponds to the difference in two copies, etc. The first minimum  $M$  is detected (yellow vertical line). Here  $M$  corresponds to *CNA cut-off* used to optimize copy number segmentation and define genomic alterations. A prominent  $M$  evidences high signal to noise ratio in CNA profile and pure copy number states (without sub-clones).





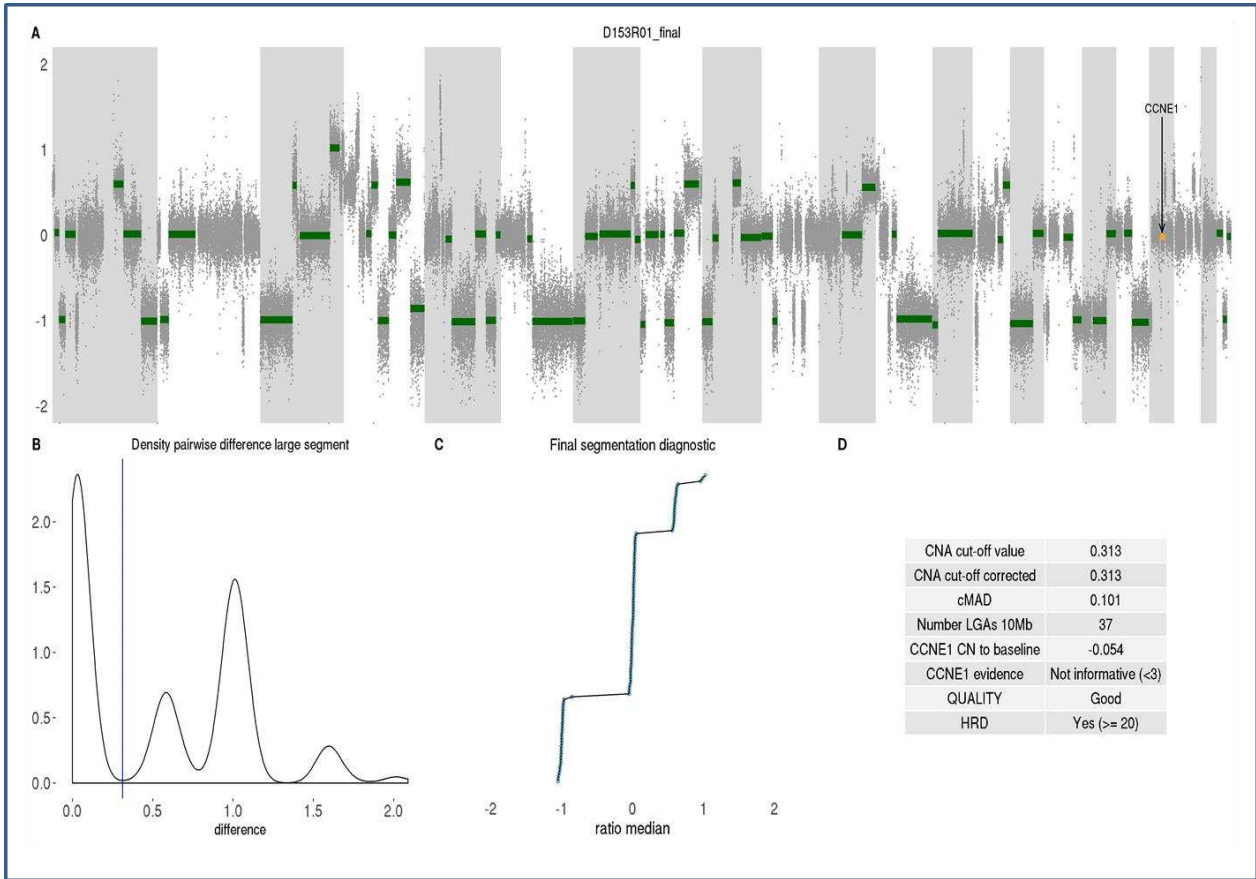
**Figure S2. Two parameters characterizing quality of CNA profile**

In-house sWGS CNA profiles (69 cases) manually annotated as of “bad”, “average” or “good” quality were characterized by 2 parameters:  $M$  defining *CNA cut-off*, and  $cMAD$  characterizing intra-segmental variation. These two parameters could be considered as sWGS quality markers. Two thresholds were defined:  $cMAD=0.14$  and  $M=0.45$  for automatic attribution of sample quality.

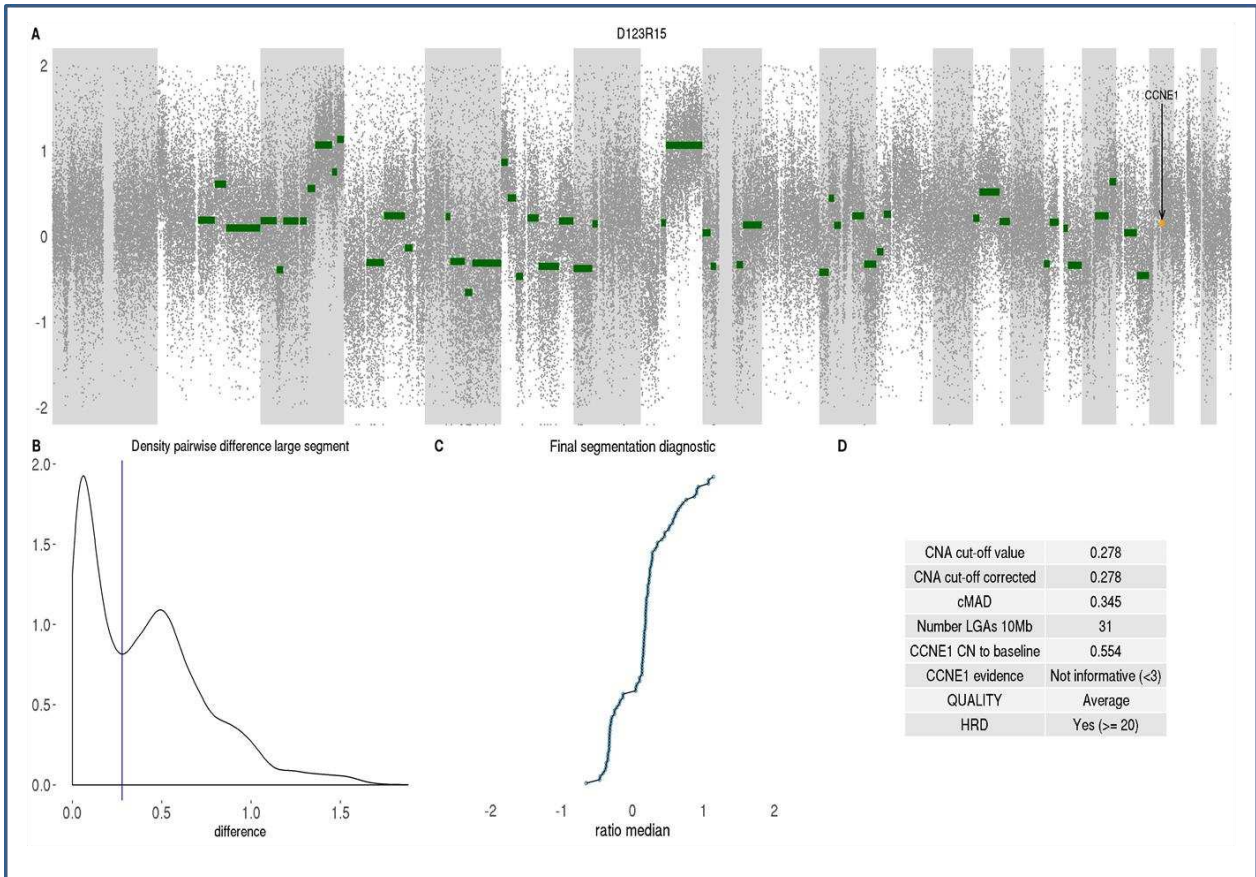
Bad quality cases represented mainly sequencing failure independent of coverage, with frequent (n=5) poorly detectable local minimum in  $M$ , separating fluctuations of segments with equal copy numbers from one copy difference. Average quality was mainly due to a low coverage (0.06-0.3X) displaying high fluctuations in the number of reads per window characterized by  $cMAD$ .

FFPE samples were among “good” (n=3) and “average” (n=1) quality regarding the thresholds, while two cases were actually annotated manually as “average” and “bad” (the latter due to low tumor content) (see Fig.S5 for details).

## I.



## II.

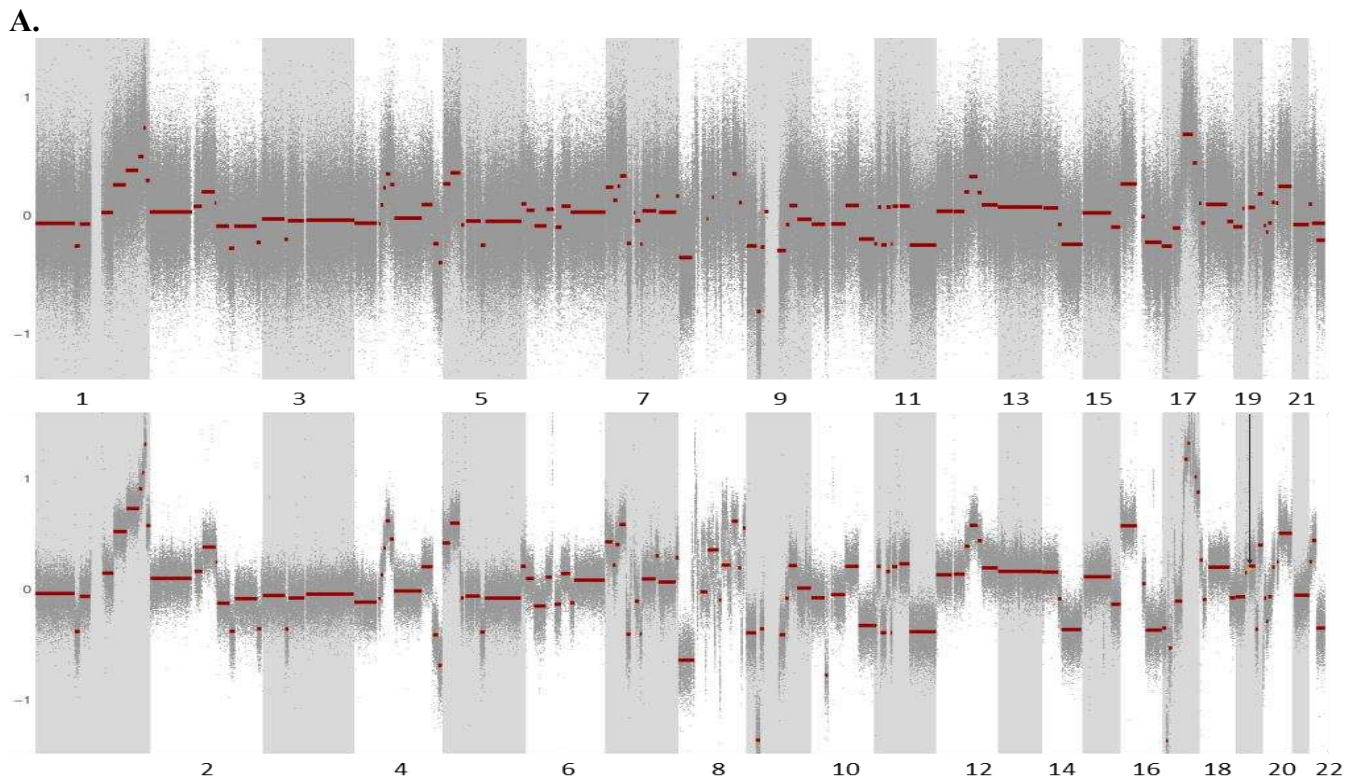


### **Figure S3. Reports of shallowHRD**

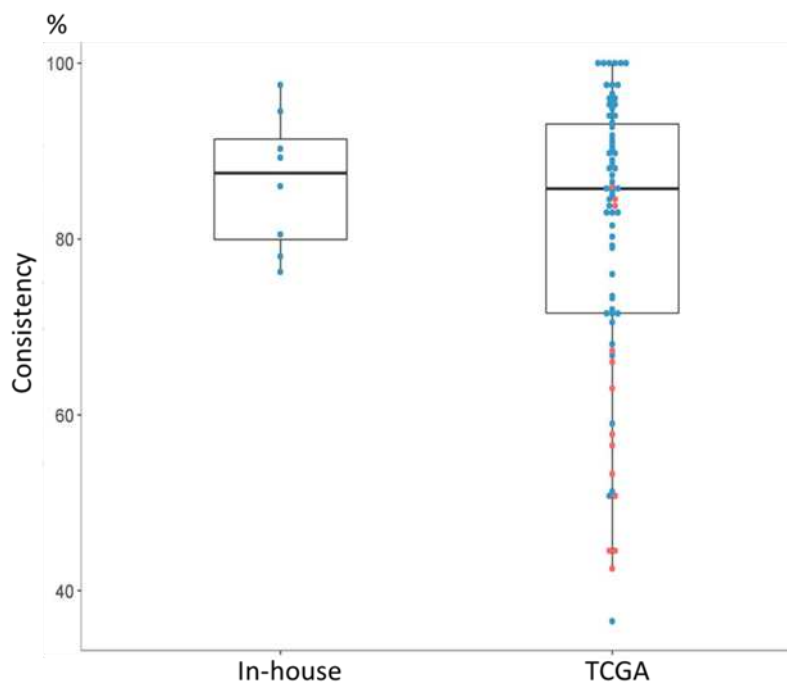
#### **I. PDX of breast cancer with *BRCA1* germline mutation; II. Primary ovarian tumor with unknown status of *BRCA1/2***

*shallowHDR* report contains the following items:

- A.** Tumor genomic profile with LGAs indicated in green.
- B.** Density plot of pairwise differences between large segments used to define the *CNA cut-off*.
- C.** Visual representation of the final segmentation, where segment medians were ordered and represented by the dots. Clear stepwise profile evidences high signal to noise ratio and proper segmentation (good quality, panel I); fuzzy profile with blurred steps evidences high unspecific variation in CNA medians with ambiguous copy number levels (average or poor quality, panel II).
- D.** Quality and Homologous Recombination Deficiency diagnostics including *M*, *cMAD* and LGA number with HRD status.



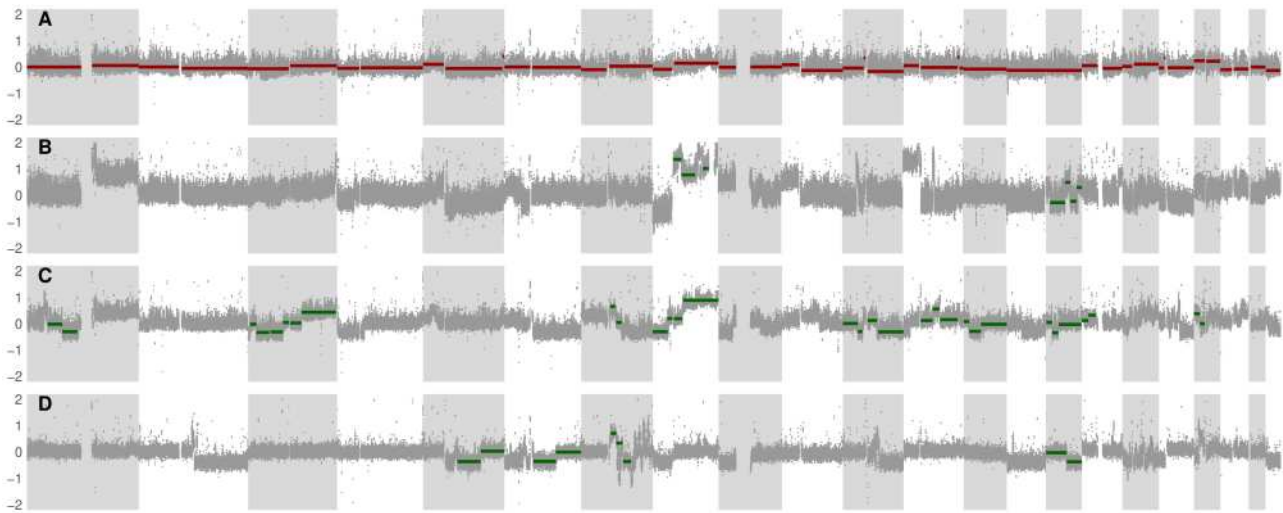
**B.**



**Figure S4. Consistency in the large CNA segments in sWGS and in SNP-arrays**

**A.** Segmented CNA profiles on SNP-array (upper panel) and sWGS (lower panel) of the in-house tumor sample. Segmentation for SNP-array was optimized to absolute copy numbers using GAP method (Popova, et al., 2009) and sWGS profile was optimized by *shallowHRD* using *CNA cut-off*. Segments were considered consistent if they were both  $\geq 10$ Mb in size and their boundaries were within 3Mb. sWGS CNA profile reproduced 86% of the large segments detected by SNP-arrays.

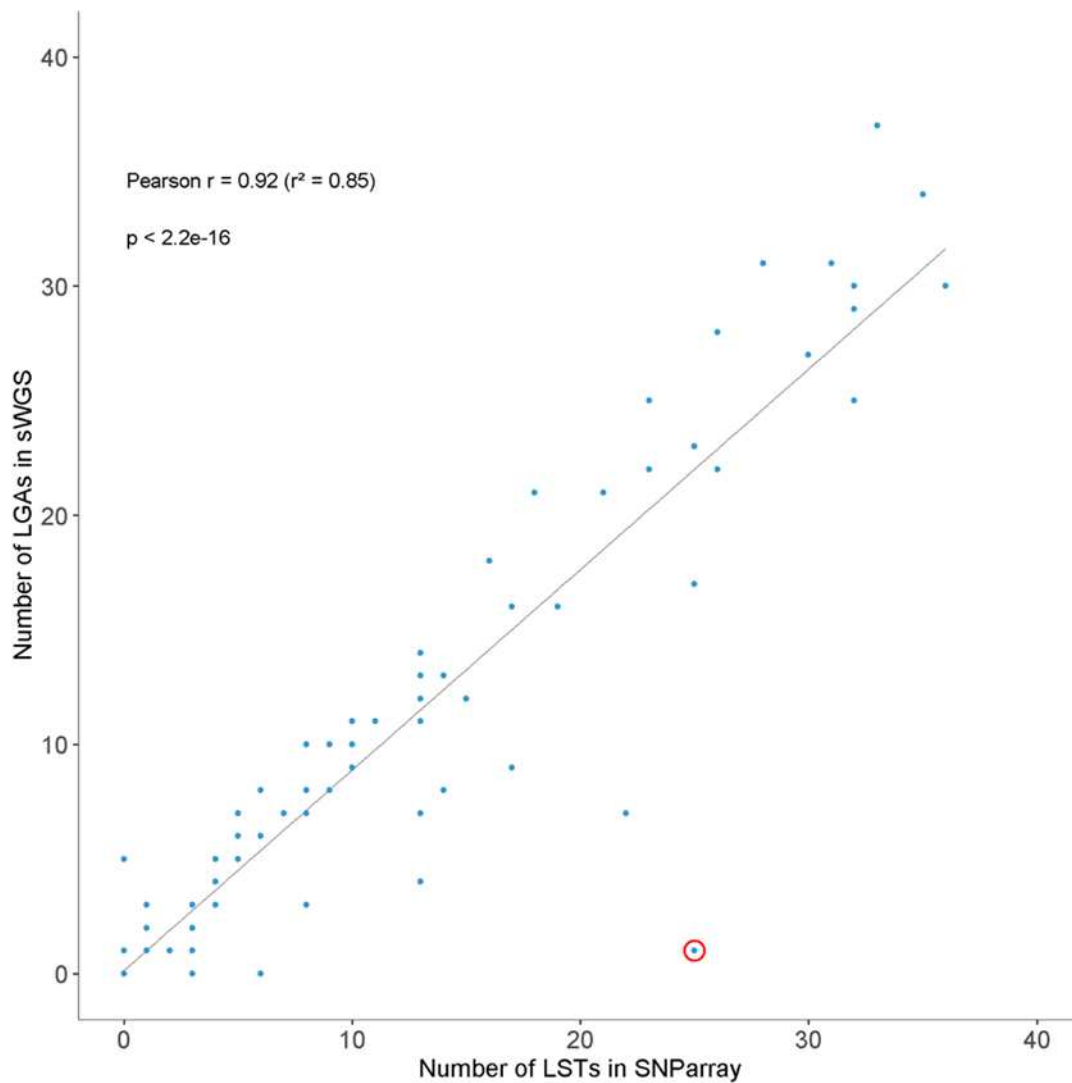
**B.** Overall large segments consistency (estimated as described in Figure S4A) in 8 in-house cases and 79 TCGA down-sampled WGS processed by *shallowHRD* and SNP-arrays. Red dots are cases automatically detected of average quality by *shallowHRD*.



**Figure S5. FFPE profiles from sWGS analyzed by shallowHRD**

sWGS profiles of four FFPE cases analyzed by *shallowHRD* are shown. Segments in green correspond to LGA. The entire segmentation is indicated in red for the profile A because no LGA was detected. Profiles A, C and D are detected as “good” quality while the profile B is detected as “average” quality. Manual annotation classified sWGS of profile A as “bad” because of a low tumor content and profile B as “average. Samples B and C were correctly predicted as nonHRD and HRD (BRCA2-/-), respectively. Sample D with unknown status was predicted as nonHRD.

Overall, the limited number of cases does not allow us drive definitive conclusion but support that sWGS and therefore *shallowHRD* is applicable for FFPE cases. Moreover, several studies, including a pilot study for the 100,000 Genomes Project, investigated the use of FFPE samples for sWGS and presented good results for FFPE with WGS and CNAs interpretation (Chin, et al., 2018 ; Robbe, et al., 2018; Scheinin, et al., 2014 ).

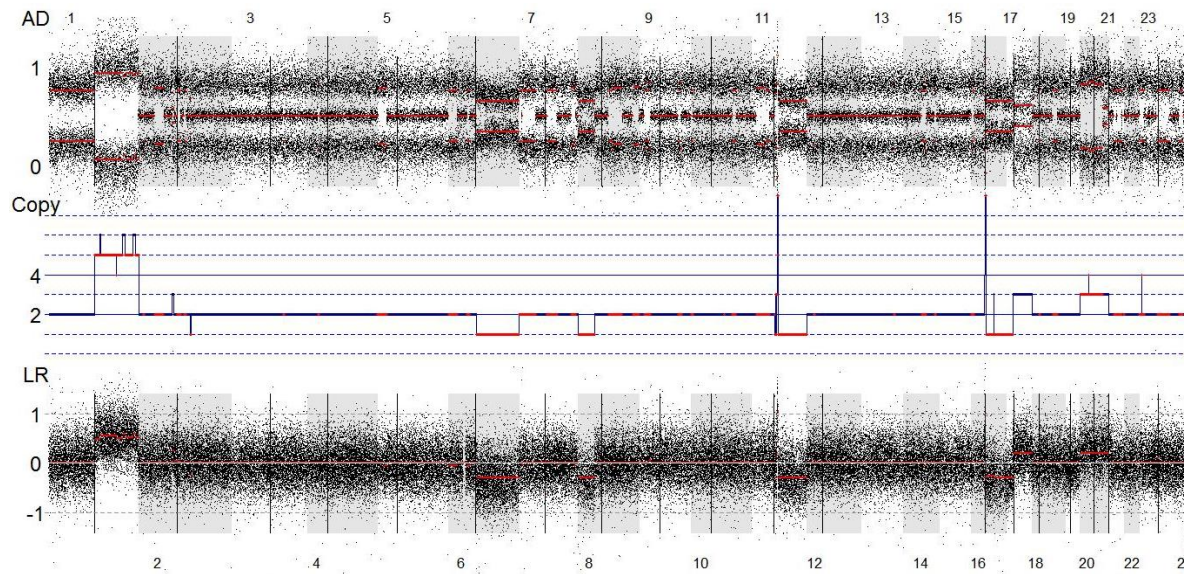


**Figure S6. Large-scale CNA correspondence between sWGS and SNP-arrays**

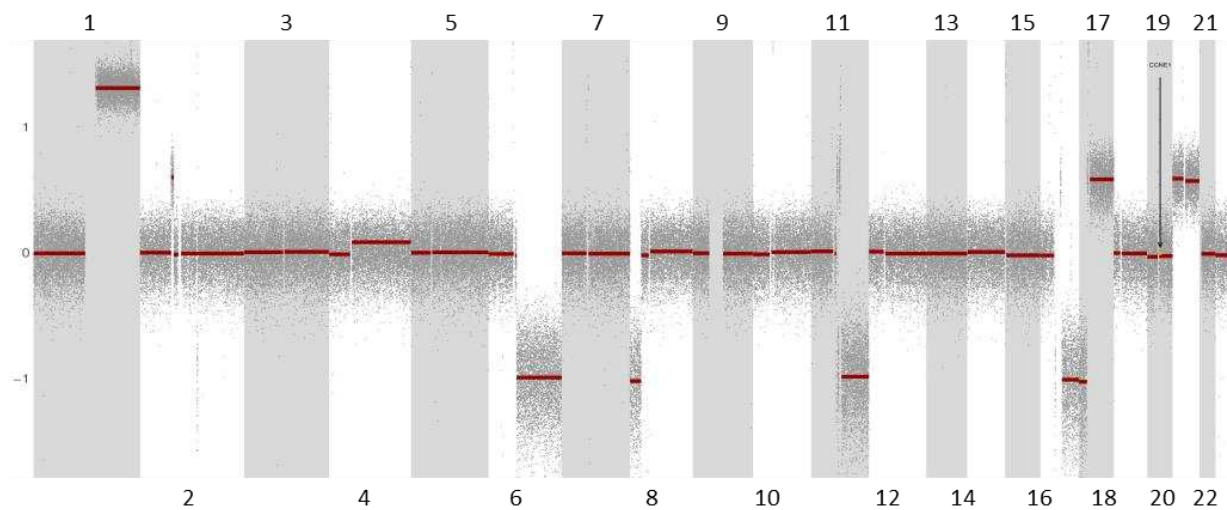
Number of LGAs in down-sampled WGS versus the number of LSTs in SNP-arrays is shown for 79 TCGA cases. The most discordant case, circled in red, is characterized by a high number of copy-neutral Loss Of Heterozygosity (see Fig. S9). Detailed information is summarized in Supplementary table S1.



**A.**

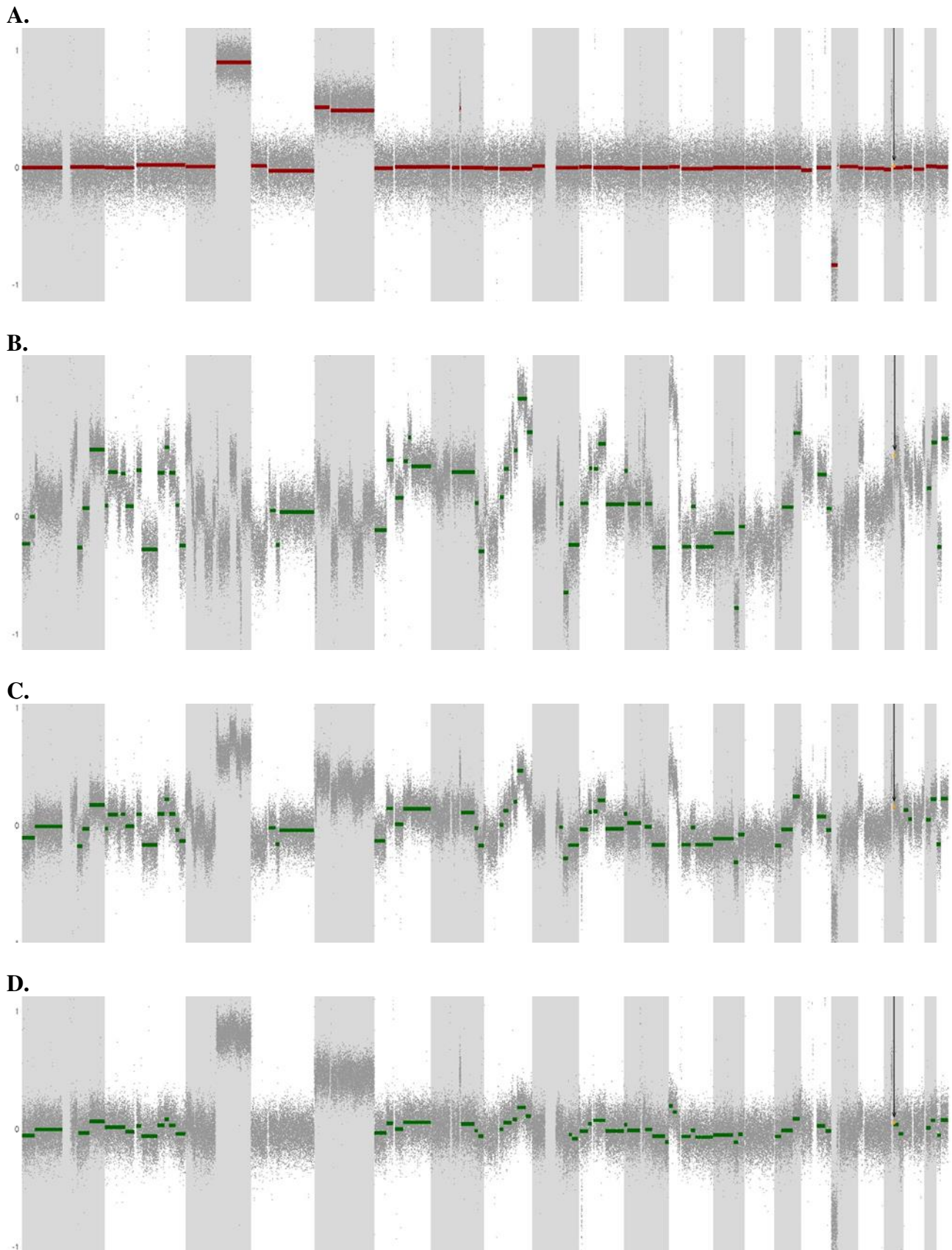


**B.**



**Figure S7. Tumor with high number of copy-neutral LOH**

**A.** SNP-array copy number profile mined by GAP (Popova, et al., 2009) with numerous large-scale breakpoints detected due to copy-neutral Loss Of Heterozygosity (LOH) (ID: TCGA-EW-A1J5-01A). Top panel represents B-Allele Frequency; bottom panel represents Log ratio related to copy number alteration profile and absolute copy numbers detected by GAP software in the middle (red segments correspond to LOH). **B.** Down-sampled WGS profile of the same tumor analyzed by *shallowHRD* with a few copy number breakpoints recognized, which led to nonHRD prediction.



**Figure S8. Example of *in silico* dilution series**

Tumor with almost flat CNA profile (A), tumor with HRD (B) and *in silico* dilutions (C, D) of the tumor with HRD by the tumor with flat profile used as a "quasi-normal" counterpart.



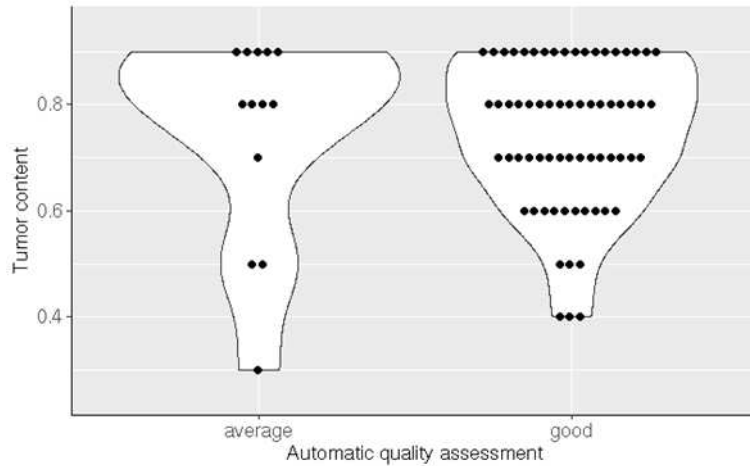
**A.** In-house tumor used to make serial dilutions considered as a “quasi-normal” case. The chromosomes 3, 5 and 17 bear CNAs in this tumor and were masked from the analysis by *shallowHRD*, as detailed in Supplementary Notes. No LGA was found in this case and all segments of the final segmentation after *shallowHRD* processing are represented with red lines.

**B.** The tumor with HRD (shown in Supplementary Figure S6B) with SNA-array estimated tumor content 0.7. LGAs are indicated with green lines.

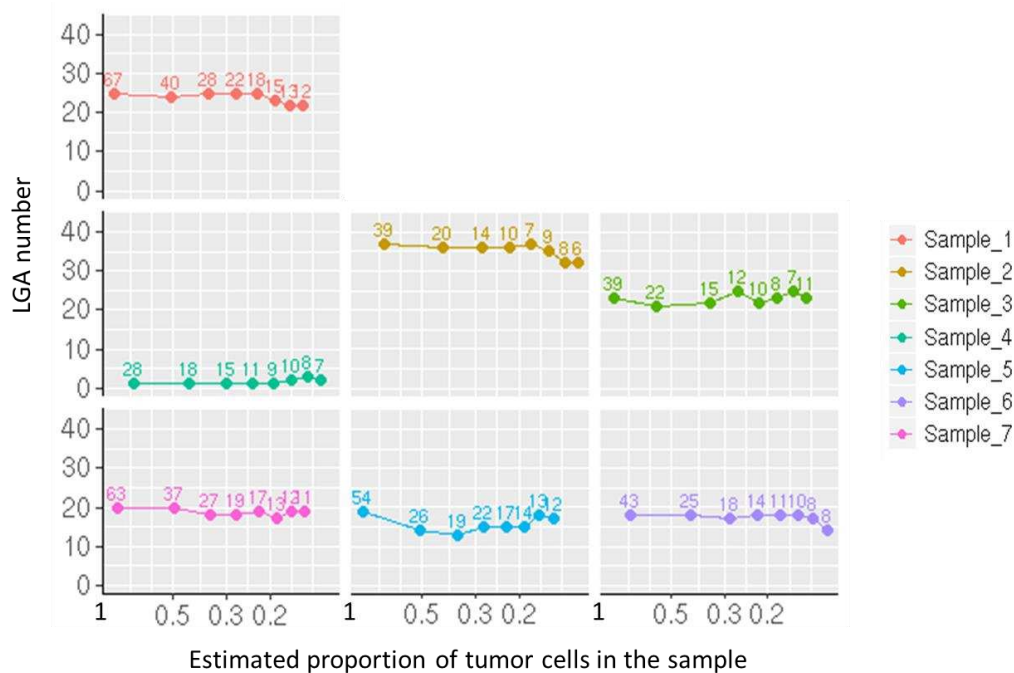
**C.** The tumor with HRD (B) diluted twice with “quasi-normal” case (A) and estimated tumor content 0.28. Tumor content was estimated regarding the initial tumor content and the proportion of reads of the tumor and “quasi-normal” sWGS.

**D.** The tumor with HRD (B) diluted seven times with “quasi-normal” case (A). Tumor content in this case is estimated to be 0.11.

**A.**



**B.**



**Figure S9. Tumor content and performance of shallowHRD**

**A.** Tumor content and sample quality were shown for down-sampled WGS TCGA cases (n=79). Tumor content was taken from the corresponding SNP-arrays as estimated by GAP method (Popova, et al., 2009) and the quality assessment was automatically produced by *shallowHRD*. One case with tumor content of ~0.3 was of average quality with a low concordance between sWGS and SNP-array. Four cases of good quality had a tumor content of 0.4 and worked nicely.

**B.** *In silico* dilution series of in-house sWGS analyzed by *shallowHRD*. Each panel corresponds to the dilution series of one sample. Quasi-normal sample used for dilution had CNAs in chr 3, 5, 17, which were excluded from further analysis (Fig.S5A). Estimated proportion of tumor cells was shown in two ways: (1) x-axis, which is related to the tumor content estimated from the dilution (proportion of mapped reads in the undiluted sWGS and the “quasi-normal” diluter sWGS) and (2) by labels upon each point showing the percent of tumor cells evaluated directly from the diluted WGS using ichorCNA (Supplementary Methods).

**Supplementary Table S1. Validation cohort of down-sampled WGS from the TCGA**

TCGA ID	Tumor type	Tumor content	Proven HRD	LST diagnostic	N LST	shallowHRD diagnostic	N LGA	% large segments conserved	Detected quality	Prediction sWGS Comments
TCGA-AR-A0TU	TNBC	0.7	HRD	HRD	35	HRD	34	91	good	TP
TCGA-AO-A0J2	TNBC	0.8	HRD	HRD	32	HRD	30	73	good	TP
TCGA-A2-A04T	TNBC	0.8	HRD	HRD	31	HRD	31	79	good	TP
TCGA-A2-A3Y0	TNBC	0.7	HRD	HRD	30	HRD	27	74	good	TP
TCGA-AN-A0AT	TNBC	0.7	HRD	HRD	28	HRD	31	88	good	TP
TCGA-AN-A04D	TNBC	0.9	HRD	HRD	26	HRD	28	100	good	TP
TCGA-A7-A0CE	TNBC	0.9	HRD	HRD	25	HRD	23	91	good	TP
TCGA-AR-A256	TNBC	0.8	HRD	HRD	25	Borderline	17	51	good	Borderline Low quality sWGS <sup>1</sup>
TCGA-AO-A124	TNBC	0.9	HRD	HRD	23	HRD	25	81	good	TP
TCGA-BH-A0WA	TNBC	0.8	HRD	HRD	23	HRD	22	67	good	TP
TCGA-EW-A1PB	TNBC	0.6	HRD	HRD	21	HRD	21	95	good	TP
TCGA-B6-A0RG	luminal	0.9	HRD	nonHRD	13	nonHRD	11	72	good	FN LGA/LST consistent <sup>2</sup>
TCGA-AO-A0J6	TNBC	0.8	HRD	HRD	33	HRD	37	57	average	TP
TCGA-A2-A04P	TNBC	0.8	HRD	HRD	32	HRD	25	58	average	TP
TCGA-C8-A12L	TNBC	0.8	HRD	HRD	32	HRD	29	63	average	TP
TCGA-AO-A0J4	TNBC	0.7	HRD	nonHRD	17	nonHRD	9	42	average	FN Low quality sWGS
TCGA-A2-A0D0	TNBC	0.8	-	HRD	36	HRD	30	83	good	FP LGA/LST consistent <sup>3</sup>
TCGA-E2-A14P	HER2+	0.6	-	HRD	26	HRD	22	76	good	FP LGA/LST consistent <sup>3</sup>
TCGA-EW-A1J5	luminal	0.8	-	HRD	25	nonHRD	1	93	good	TN LGA/LST inconsistent (Fig S9)
TCGA-A2-A0EY	HER2+	0.7	-	nonHRD	19	Borderline	16	86	good	Borderline <sup>4</sup>
TCGA-E2-A1LL	TNBC	0.7	-	nonHRD	18	HRD	21	71	good	TP LGA/LST inconsistent
TCGA-C8-A12Q	HER2+	0.6	-	HRD	17	Borderline	16	89	good	Borderline
TCGA-C8-A130	luminal	0.8	-	nonHRD	16	Borderline	18	36	good	Borderline Low quality sWGS <sup>1</sup>
TCGA-B6-A0RU	TNBC	0.6	-	HRD	15	nonHRD	12	80	good	TN LGA/LST inconsistent
TCGA-B6-A0RE	TNBC	0.8	-	nonHRD	14	nonHRD	13	51	good	TN
TCGA-B6-A0RE	TNBC	0.8	-	nonHRD	14	nonHRD	8	59	good	TN
TCGA-AC-A2BK	TNBC	0.9	-	nonHRD	13	nonHRD	13	86	good	TN
TCGA-AO-A0JL	TNBC	0.8	-	nonHRD	13	nonHRD	14	71	good	TN
TCGA-AO-A0JM	HER2+	0.7	-	nonHRD	13	nonHRD	12	88	good	TN
TCGA-A7-A13D	TNBC	0.7	-	nonHRD	11	nonHRD	11	85	good	TN
TCGA-BH-A1FC	TNBC	0.9	-	nonHRD	10	nonHRD	11	71	good	TN
TCGA-BH-A0H7	luminal	0.8	-	nonHRD	10	nonHRD	10	72	good	TN
TCGA-B6-A0I2	TNBC	0.6	-	nonHRD	10	nonHRD	9	96	good	TN
TCGA-BH-A18R	HER2+	0.6	-	nonHRD	10	nonHRD	11	97	good	TN
TCGA-A2-A0YG	HER2+	0.6	-	nonHRD	9	nonHRD	10	83	good	TN
TCGA-BH-A0DK	luminal	0.6	-	nonHRD	9	nonHRD	8	100	good	TN
TCGA-A8-A09X	luminal	0.4	-	nonHRD	9	nonHRD	8	94	good	TN
TCGA-E2-A15E	luminal	0.9	-	nonHRD	8	nonHRD	7	79	good	TN
TCGA-BH-A0E0	TNBC	0.7	-	nonHRD	8	nonHRD	8	94	good	TN
TCGA-BH-A0HX	luminal	0.7	-	nonHRD	8	nonHRD	8	89	good	TN

TCGA-BH-A0GY	luminal	0.6	-	nonHRD	8	nonHRD	10	90	good	TN
TCGA-A2-A04X	HER2+	0.7	-	nonHRD	7	nonHRD	7	92	good	TN
TCGA-EW-A1P8	TNBC	0.7	-	nonHRD	7	nonHRD	7	96	good	TN
TCGA-A2-A0D1	HER2+	0.9	-	nonHRD	6	nonHRD	6	87	good	TN
TCGA-BH-A0HB	luminal	0.8	-	nonHRD	6	nonHRD	8	84	good	TN
TCGA-A8-A07I	HER2+	0.9	-	nonHRD	5	nonHRD	7	85	good	TN
TCGA-E9-A1NH	luminal	0.8	-	nonHRD	5	nonHRD	5	91	good	TN
TCGA-B6-A0WX	TNBC	0.5	-	nonHRD	5	nonHRD	6	90	good	TN
TCGA-E2-A156	luminal	0.9	-	nonHRD	4	nonHRD	3	85	good	TN
TCGA-B6-A0RI	luminal	0.8	-	nonHRD	4	nonHRD	4	96	good	TN
TCGA-E2-A152	HER2+	0.8	-	nonHRD	4	nonHRD	4	83	good	TN
TCGA-A2-A3XX	TNBC	0.7	-	nonHRD	4	nonHRD	3	68	good	TN
TCGA-A7-A0D9	luminal	0.9	-	nonHRD	3	nonHRD	2	91	good	TN
TCGA-E2-A15K	luminal	0.9	-	nonHRD	3	nonHRD	2	84	good	TN
TCGA-AO-A0JJ	luminal	0.5	-	nonHRD	3	nonHRD	1	88	good	TN
TCGA-AR-A0TX	HER2+	0.4	-	nonHRD	3	nonHRD	3	93	good	TN
TCGA-BH-A0H0	luminal	0.8	-	nonHRD	2	nonHRD	1	98	good	TN
TCGA-A7-A26J	luminal	0.9	-	nonHRD	1	nonHRD	1	98	good	TN
TCGA-A7-A26J	luminal	0.9	-	nonHRD	1	nonHRD	1	86	good	TN
TCGA-B6-A0X4	luminal	0.9	-	nonHRD	1	nonHRD	1	100	good	TN
TCGA-BH-A0H6	luminal	0.9	-	nonHRD	1	nonHRD	1	71	good	TN
TCGA-A2-A259	luminal	0.7	-	nonHRD	1	nonHRD	3	90	good	TN
TCGA-E2-A15H	HER2+	0.7	-	nonHRD	1	nonHRD	1	86	good	TN
TCGA-A2-A3KC	luminal	0.6	-	nonHRD	1	nonHRD	2	100	good	TN
TCGA-AO-A0JF	luminal	0.9	-	nonHRD	0	nonHRD	5	95	good	TN
TCGA-BH-A0HK	luminal	0.9	-	nonHRD	0	nonHRD	1	97	good	TN
TCGA-AR-A2LK	luminal	0.8	-	nonHRD	0	nonHRD	0	95	good	TN
TCGA-BH-A0BM	luminal	0.7	-	nonHRD	0	nonHRD	0	100	good	TN
TCGA-BH-A0W5	luminal	0.5	-	nonHRD	0	nonHRD	0	100	good	TN
TCGA-A2-A0EU	luminal	0.4	-	nonHRD	0	nonHRD	0	95	good	TN
TCGA-BH-A0B3	TNBC	0.5	-	HRD	22	nonHRD	7	45	average	TN
TCGA-EW-A1PH	TNBC	0.9	-	nonHRD	13	nonHRD	7	53	average	TN
TCGA-A7-A26F	TNBC	0.5	-	nonHRD	13	nonHRD	4	51	average	TN
TCGA-A8-A08B	HER2+	0.9	-	nonHRD	8	nonHRD	3	66	average	TN
TCGA-A2-A04Q	TNBC	0.3	-	nonHRD	6	nonHRD	0	44	average	TN Low tumor content
TCGA-E2-A109	luminal	0.9	-	nonHRD	4	nonHRD	5	67	average	TN
TCGA-A8-A092	luminal	0.9	-	nonHRD	3	nonHRD	0	84	average	TN
TCGA-A7-A0DC	NA	0.8	-	nonHRD	1	nonHRD	1	86	average	TN
TCGA-A8-A08S	HER2+	0.9	-	nonHRD	0	nonHRD	0	85	average	TN

HRD: Homologous Recombination Deficiency; LST: Large-scale State Transitions; LGA: Large Genomic Alterations; TN: true positive; TP: true negative; FN: false negative; FP: false positive. Color code: green: no problem with the case; light orange: large sWGS segments ( $\geq 10$ Mb) conserved in SNP-arrays  $< 70\%$ , as described in Figure S4; dark orange: “borderline” or cases with inconsistent diagnostic of HRD.

- <sup>1</sup>: Low quality WGS due to high unspecific variation failed to be detected automatically.
- <sup>2</sup>: BRCA2<sup>-/-</sup> can have in some rare cases low number of intra-chromosomal breaks.
- <sup>3</sup>: Highly altered genome with no HRD evidence found (still may be HRD).
- <sup>4</sup>: Ploidy of 4 for this case, accessible with SNParray, helping classifying the case as nonHRD.

## Conclusion: “shallowHRD: detection of homologous recombination deficiency from shallow whole genome sequencing”

shallowHRD has a sensitivity and specificity of ~90% to evaluate HRD in breast and ovarian cancers. sWGS is overall cheap and its processing with shallowHRD is fast and straightforward. It can process FFPE samples and doesn't need a matched normal sequencing. It performs similarly to most state-of-the-art approaches, only being surpassed by more resource-intensive methods<sup>169,224</sup>. shallowHRD can however process any Read Depth tool-based approach with few adaptations to its standard input. HRD being a good surrogate marker of PARPi sensitivity for tumors and the inherent advantages that procures shallowHRD with sWGS bears all the reasoning of its usefulness, not only in research but also in clinic. It would especially fit well with a gene panel approach, accounting for variant of unknown significance (VUS) and gene promotor methylation. Further amelioration of this method, to tackle problematic borderline cases ( $15 \leq \text{LGAs} < 20$ ) and to fit the varying quality of clinical samples and the clinical precision required for patient care, would be a tremendous help to reach routinely applicable good practices in clinic.

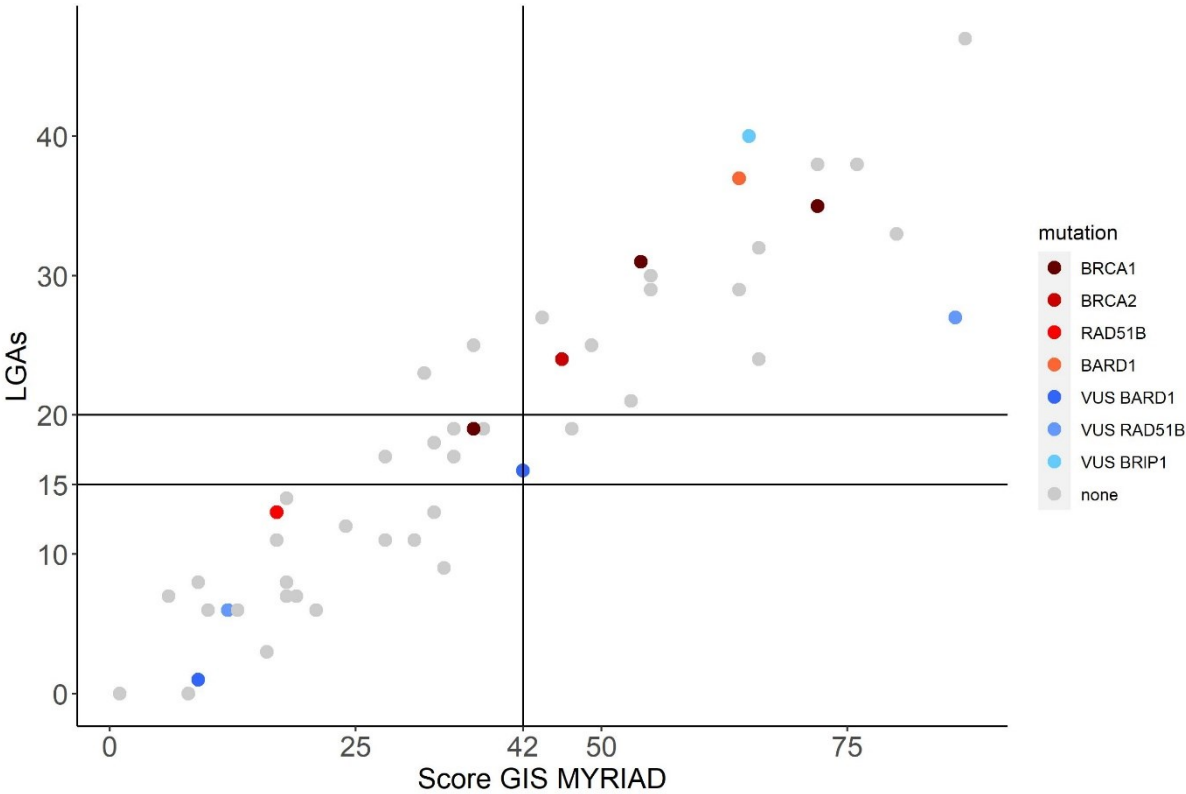
### Additional results of *shallowHRD* in different cohorts

#### Comparison between MYRIAD myChoice® CDx and *shallowHRD*

The commercially available test Myriad myChoice® CDx<sup>210,212</sup> is the only HRD test approved for patient care and PARP prescription without deleterious mutations in *BRCA1* and *BRCA2*. It is also based on large-scale genomic rearrangement, partly like *shallowHRD*. It would therefore be interesting to compare *shallowHRD* with Myriad myChoice® CDx. The combination of the three large-scale genomic rearrangements is reported to be more robust but the actual trade-off between Myriad myChoice® CDx and *shallowHRD* should be investigated.

The preliminary comparison of 60 ovarian cancer tumors for the same sample with *shallowHRD* and Myriad myChoice® CDx showed high correlation (unpublished results, Celine Callens & Adrien Briaux). With a stringent threshold of 20 LGAs, 90% cases classified HRD by Myriad myChoice® CDx were also HRD with shallowHRD (18/20 cases) and 88% classified as HRP were classified HR Proficient (HRP) with our method (29/33). Of note, a flat genomic profile is ambiguous with *shallowHRD* as it can come either from low cellularity of the sample or from the tumor harboring no rearrangement at the Copy-Number level. The latter is however rare for ovarian and breast tumors. A cancer panel approach allowed us to reclassify two tumors with a flat genomic profile, pushing the correspondence for HRP cases to 94% (31/33).

Interestingly, the two discordant cases HRD cases for MYRIAD still presented a high amount of LGAs with 16 and 19 LGAs, underlying the need to further tackle those borderline genomic profiles. Out of the six cases where MYRIAD did not give results, four were not interpretable also with *shallowHRD* because of a low cellularity, whereas two were surprisingly interpretable in sWGS. Nonetheless some cases were intriguing for us. One case harbored a *BRCA1* deleterious mutation with 19 LGAs in sWGS but were classified as HRP with Myriad myChoice® CDx with a Genomic Instability Score (GIS) of 37, below the cut-off of 42. Two other cases harbored characteristic HRD genomic profiles with shallowHRD with 23 and 24 LGAs but were HRP in Myriad myChoice® CDx with a GIS of 32 and 37. The results are presented in Figure 15.

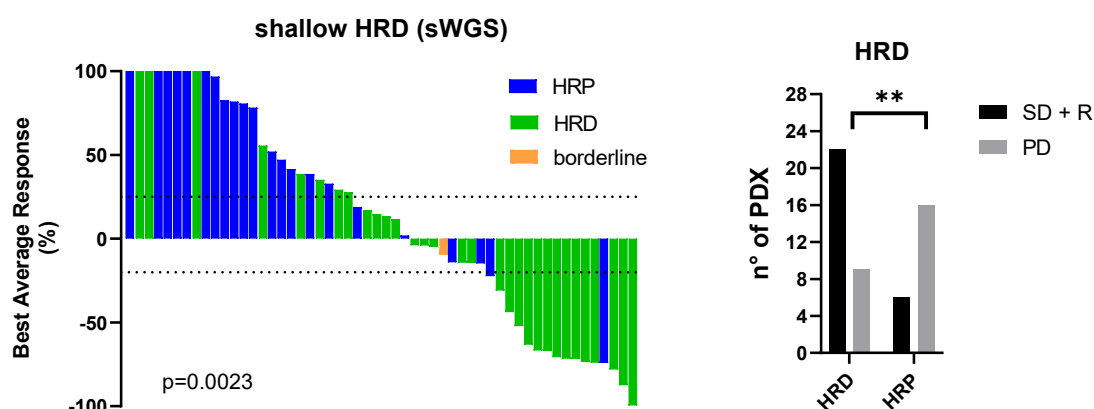


**FIGURE 15: Comparison between the diagnosis of the Myriad myChoice® CDx and *shallowHRD***

x axis: Tumors with a Genomic Instability Score (GIS) below 42 are HRP with the Myriad myChoice® CDx. Tumor with a GIS of 42 and above are HRD. VUS: Variant of Unknown Significance. Genes only indicated by color and their names corresponds to deleterious mutations.

## Cisplatin response according to the HR status of *shallowHRD*

Homologous Recombination is a direct predictor of PARPi sensitivity for tumors. In a cohort of 54 patient derived xenografts (PDX) from TNBC, the correspondence between the response to cisplatin and the status inferred by *shallowHRD* was evaluated (unpublished, Elisabetta Marangoni). 31 HRD cases, 6 borderlines cases and 17 HRP cases was found in the cohort. According to what we observed with the comparison to the Myriad score and the observation of the genomic profiles, 5 borderline cases were manually reclassified as HRP while one remained annotated as borderline. Out of the 31 HRD samples, 47% were explained by the methylation of *BRCA1*, 41% by deleterious mutation in *BRCA1* (9) or *BRCA2* (4) while 12% remained unexplained. 70.97% of HRD cases (22/31 PDX) and 27.27% of HRP cases (6/22 PDX) showed a stable disease, partial response, or complete response. Of note, 4 out 6 initially classified as borderline cases showed a stable disease, partial response, or complete response. The prediction for HR status of *shallowHRD* is therefore highly correlated to the response to platinum treatment in PDX of TNBC. The results are presented in Figure 16.



**FIGURE 16: Response to cisplatin in 54 PDX of TNBC**

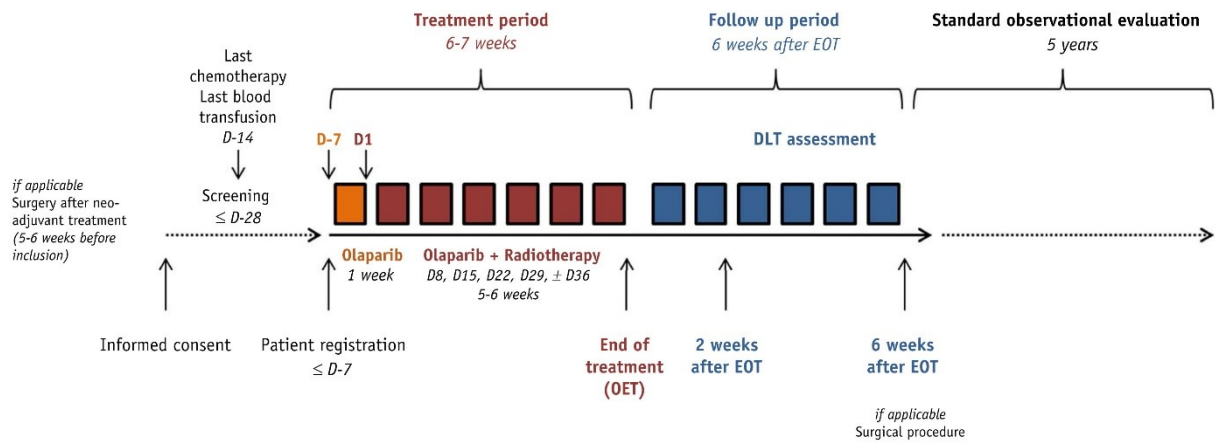
HRP: Homologous Recombination Proficient; HRD: Homologous Recombination Deficient; SD: Stable Disease; R: Response; PD: Progression disease. From Elisabetta Marangoni.

## BRCA1 VUS

The consequences on the protein of the VUS are dubious and the diagnosis is difficult for the clinician to give in this case. A preliminary cohort of 42 breast tumors with *BRCA1* VUS were constituted with two additional tumors harboring deleterious mutations of *BRCA1* as positive controls. Ten cases were not interpretable, either because of the low cellularity for three samples or the low quality of the sequencing for seven samples. The positive controls with *BRCA1* inactivation are correctly classified as HRD. Three other samples were clearly classified as HRD, indicating either that their *BRCA1* VUS is in fact deleterious or that they are



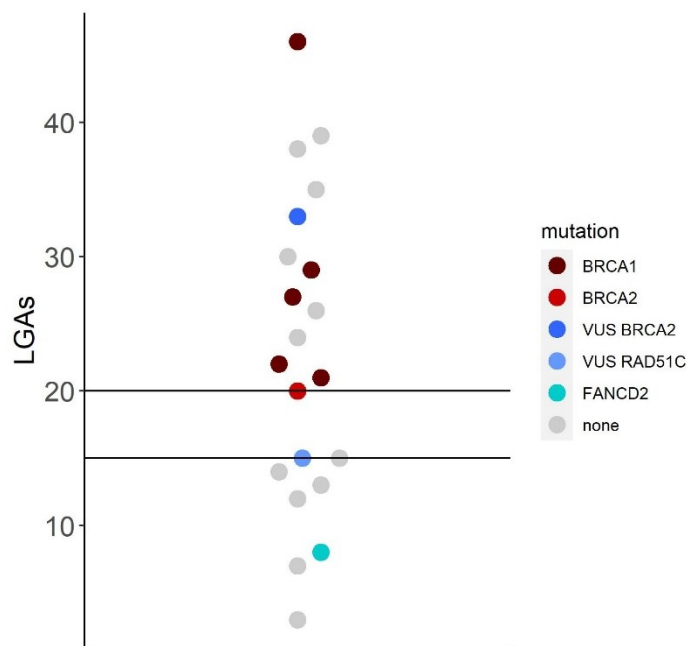




**FIGURE 18: Plan of treatment for the RadioParp study**

D: Day; DLT: Dose-limiting Toxicity. Extracted from Loap et al<sup>259</sup>

22 out of the 24 tumors were sequenced in sWGS. The results for this preliminary cohort are presented in Figure 19. All tumors with identified deleterious mutations in *BRCA1* and *BRCA2* display 20 or more LGAs. One *BRCA2* VUS harbors a high number of LGAs, indicating that this mutation may be deleterious or that either *BRCA1* or *RAD51C* are methylated. It was surprising to see such enrichment of HRD cases within the cohort. Also given that the tumors are TNBC, we also did not anticipate finding several *BRCA2* depleted cases in the cohort.



**FIGURE 19: Results of *shallowHRD* on 22 sWGS of TNBC from the RadioParp study**

## Introduction: “Lack of evidence for CDK12 as an ovarian cancer predisposing gene”

### The cyclin-dependent kinase 12

Cyclin-dependent kinases (CDKs) are proteins that play an important role in cell-cycle, cell-division and transcription. Twenty CDKs has been identified in human, separated in two large families depending on their roles, either in the cell-cycle or in transcription. The cyclin-dependent kinase 12 (CDK12) is a transcription associated CDK that plays numerous roles in genomic stability. In mice models, CDK12 knock-out is embryonic lethal suggesting its essential function for embryonic development<sup>260</sup>. The CDK12 protein is comprised of a central kinase domain, a N-terminal arginine/serine-rich, with two proline rich motifs at the C-terminal and central location<sup>261,262</sup>. Like the other CDKs, CDK12 kinase activity is only functional when it dimerizes to its partner, the Cyclin K (CCNK). It is the only cyclin that was associated with CDK12<sup>263,264</sup>.

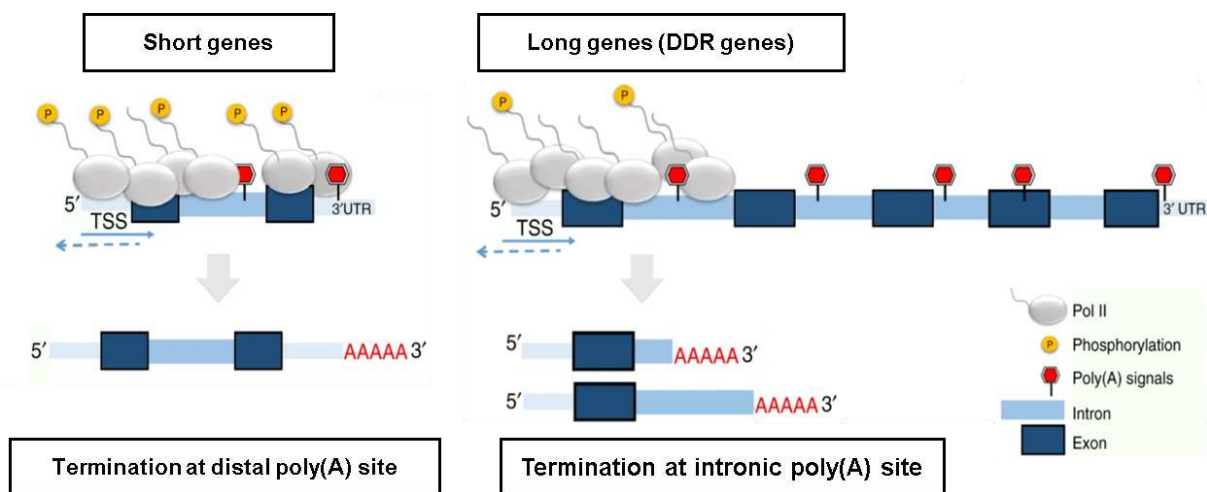
### CDK12 functions

#### **CDK12 regulates the correct transcription of long polyadenylation-site-enriched genes**

In eukaryotes, the RNA polymerase II (RNA pol II) is responsible for the transcription of pre-messenger RNA (mRNA) from DNA in the nucleus. The carboxy-terminal-domain (CTD) of the RNA pol II is composed of numerous repetitions of the same heptapeptide - Tyrosine<sub>1</sub>Serine<sub>2</sub>Proline<sub>3</sub>Threonine<sub>4</sub>Serine<sub>5</sub>Proline<sub>6</sub>Serine<sub>7</sub> (Tyr<sub>1</sub>Ser<sub>2</sub>Pro<sub>3</sub>Thr<sub>4</sub>Ser<sub>5</sub>Pro<sub>6</sub>Ser<sub>7</sub>). The CTD can be modified by the phosphorylation of Ser<sub>2</sub>, Ser<sub>5</sub> and Ser<sub>7</sub>, along with Tyr<sub>1</sub> and Thr<sub>4</sub> and but also by other changes<sup>265,266</sup>. These modifications coincide with the different steps of the transcription and are needed for the regulation of the transcription and co-transcriptional RNA processes like splicing<sup>265,266</sup>.

The principal steps of RNA transcription are the initiation, where RNA polymerase binds to the DNA at the gene promoter, the elongation, where nucleotides are added to the transcript, and the termination where transcription ends. In protein-coding genes, the transcription terminates upstream of the genes where the 3' transcript is cleaved at a specific site and a chain of adenine nucleotides (Poly(A) tail) is synthesized at the 3' of the mRNA. This process is referred to as polyadenylation. Protein-coding genes can have more than one polyadenylation site. The initiation of the elongation is marked by the phosphorylation of Ser<sub>5</sub> while the phosphorylation of Ser<sub>2</sub> helps for actual active elongation post-initiation, splicing and termination<sup>266</sup>. The regulation and post-transcriptional modifications for the CTD of RNA pol II are done by several identified CDKs.

Historically, CDK12 major role was associated with the phosphorylation of Ser2 of the CTD<sup>267,268</sup>. Further *in-vitro* experiments showed that CDK12 can phosphorylate Ser2, Ser5 and Ser7 of the CTD<sup>264,269</sup>. However, the impairment of CDK12 displays only moderate changes in CTD phosphorylation in some *in-vivo* experiments<sup>263,270,271</sup>. Thus, the exact role of CDK12 in CTD phosphorylation and how it interplays with other factors such as CDK13 is not yet fully elucidated. The preferred model is that CDK12 phosphorylates Ser2. Interestingly, the impediment of CDK12 affects the transcription, but only for a specific subset of genes<sup>264,271-273</sup>. The genes affected by CDK12 inactivation are mainly long genes, notably implicated in DNA repair like *BRCA1* and *BARD1* but also genes in DNA replication and cell-cycle<sup>264,274</sup>. The inactivation of CDK12 leads to elongation deficiency of RNA pol II at the 3' of gene targets, in a gene length manner<sup>273</sup>. Indeed, in mouse stem cells CDK12 was shown to suppress intronic polyadenylation events, that are more used in its absence, leading to altered full-length RNA into shorter ones<sup>275</sup>. Those polyadenylation sites are present in many DDR genes. CDK12 deletion effect is not only dependent on the length of the gene but also the actual proportion of intronic polyadenylation sites in the genes compared to its length<sup>274</sup>. Therefore, CDK12 regulates the transcription of long polyadenylation-site-enriched genes by preventing their premature termination through optimal elongation (Figure 20). The mode of action of CDK12 remains unclear but CDK12 was proposed in association with several transcription factors that regulates elongation or the recruitment of CCNK<sup>276-279</sup>.



**FIGURE 20: The inhibition of CDK12 leads to elongation defect by premature cleavage and polyadenylation in a gene-length dependent manner**

CDK12 inhibition leads to a defect in gene transcription by premature cleavage and polyadenylation in long polyadenylation-site-enriched genes. TSS: Transcription Start Site; DDR: DNA Damage repair. Adapted from Krajewska et al<sup>274</sup>

CDK12 has additional roles in specific RNA processing with splicing and 3' end processing. CDK12 is expressed in the nucleus, majorly in the nuclear speckles<sup>261,280</sup> where splicing and 3' end processing factors are located<sup>281</sup>. CDK12 was associated with the indirect regulation of

the RNA splicing of different genes and a splicing factor<sup>154,270,280</sup> and for the 3' end processing of *c-MYC* and *c-FOS*<sup>282,283</sup>. However, the inhibition of CDK12 does not seem to cause a global splicing impediment, only affecting a subset of majorly long genes<sup>273,274</sup>. The splicing changes on long genes could just be the result of CDK12 not repressing premature cleavage and polyadenylation.

### CDK12 helps in translation

CDK12 not only impacts RNA synthesis but it also impacts the translations into proteins of a specific subset of mRNAs. The complex mTORC1 acts as one regulator of the biosynthesis of proteins<sup>284</sup>. CDK12 and mTORC1 phosphorylates a translation repressor called 4E-BP1, which enable the translation for a subset of specific mRNAs of genes implicated in mitosis and DNA damage response<sup>285</sup>. The mitosis actors identified were key parts of centrosome, centromere and kinetochore complexes, which would most probably lead to a defect in mitosis and potential chromosome segregation and misalignment.

### CDK12 partly regulates cell-cycle progression & cell proliferation

CDK12 partly regulates the cell-cycle, which is instrumental for genomic stability. The long-term inactivation of CDK12 or CCNK leads to the accumulation of cell in G2/M<sup>264,286,287</sup>. Moreover, CDK12 regulation of the RNA Pol II as detailed above is needed for important origin recognition and pre-replication complex genes. Its inactivation leads to malformed pre-replication complexes and impedes the recruitment of DNA replication complex on the chromatid, subsequently blocking the cell-cycle in G1/S<sup>271</sup>. The cell cycle regulation has a direct impact on the cell proliferation and the deletion of CDK12 highly decreases the cell proliferation in different cell lines including cancer cell lines<sup>272,287,288</sup>.

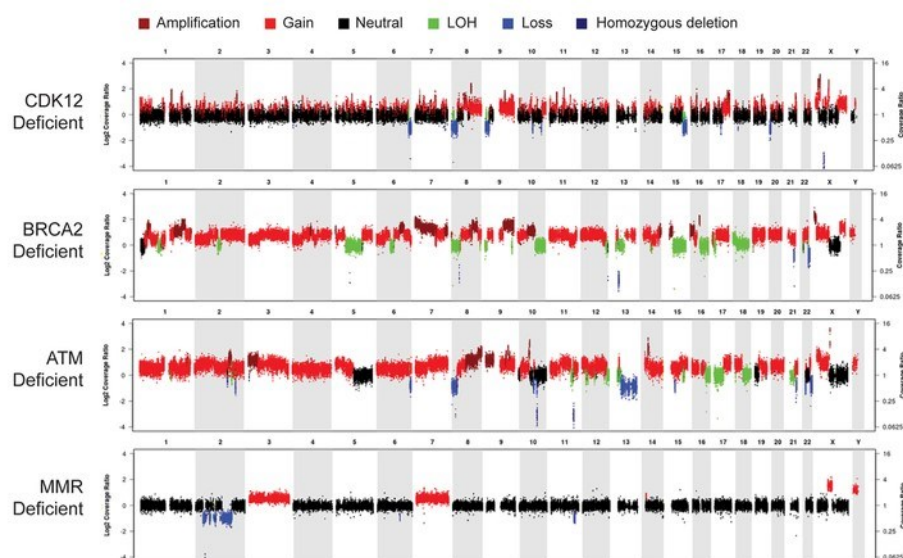
### CDK12 in cancers

*CDK12* can act as a tumor promotor when overexpressed in different cancer types. *CDK12* is located at the chr17q12 localization, close to HER2. In breast cancers, *CDK12* is often associated with HER2 amplification and drives oncogenic events in such context<sup>289,290</sup>. Overexpressed *CDK12* promotes breast cancer invasiveness and migration capacity with alternative last exon splicing of gene isoforms directly linked to those oncogenic abilities and the activation of oncogenic pathways<sup>283,291</sup>. In gastric cancer, tumor overexpressing *CDK12* were associated with worse outcome and a specific oncogenic pathway<sup>292</sup>.

*CDK12* also behaves as a tumor suppressor gene, leading to genomic instability upon bi-allelic deletion through its multiple roles of transcription and translation of DNA repair, replication genes and its impact on mitotic regulation. Bi-allelic *CDK12* inactivation were found in TNBC<sup>293,294</sup>, in EOC<sup>146,295,296</sup> and prostate cancers<sup>297,298</sup>. It is inactivated in 3% of the HGSOE

in the TCGA<sup>146</sup> and approximately 7% in metastatic castration-resistant prostate cancer, a form of advanced prostate cancer<sup>298</sup>.

CDK12<sup>-/-</sup> tumors display a specific genomic phenotype with diploid tumors harboring a high number of gains with focal tandem duplications and with few translocations<sup>296,298</sup>. The CDK12<sup>-/-</sup> associated tandem duplications phenotype (CDK12-TDs) have a bimodal distribution and two peaks at ~0.4 Mb and ~2.4 Mb<sup>296,298</sup>, that may be linked to the aberrant re-replication of DNA in S-phase<sup>298</sup>. To be noted, the TDs observed in *CDK12*<sup>-/-</sup> are unique, different from the one observed in *BRCA1*-deficient tumors or other DNA-repair deficient tumors<sup>296,299,300</sup>. Moreover, the mutation signatures associated with CDK12 inactivation are different, pointing to a unique dysregulation of genomic stability<sup>298</sup>. The distinct genomic profile resulting from the inactivation of CDK12 in prostate tumors is presented in comparison to other classes of driver mutations in Figure 21.



**FIGURE 21: CDK12 inactivated tumors display a distinct genomic profile in metastatic castration-resistant prostate cancers**

Representative genomic profiles of metastatic castration-resistant prostate cancers tumors according to their driver inactivation. CDK12<sup>-/-</sup> tumors have a distinct genomic profile. Extracted from Wu et al<sup>298</sup>

### CDK12 as a potential cancer predisposing gene for epithelial ovarian carcinoma

EOC is associated with known risk factors including age, hormonal history and genetic factors<sup>301</sup>. *BRCA1* and *BRCA2* explain a large majority of hereditary EOC<sup>302</sup>, along with other HR mutated genes<sup>137,303</sup>. Moreover, genes involved in the DNA repair like the MMR family (*MLH1*, *MSH2*, *MSH6*, *PMS2*) account for a smaller part of hereditary EOC<sup>304</sup>. Nonetheless, some hereditary EOC are not yet explained. CDK12 acts as a tumor suppressor, maintains genomic stability, regulates the transcription of DNA repair genes and is among one of the top ten mutated genes in HGSOC<sup>147</sup>. It is therefore a potential predisposition gene for EOC. In a

cohort of breast and ovarian tumors with Tatar ancestry, *CDK12* is indicated as a potential cancer predisposing gene<sup>305</sup>.


In this study published in *Familial Cancer*, we evaluated if *CDK12* is a cancer predisposition gene for EOC by investigating the blood DNA of 416 unrelated and unselected patients that all have an history of EOC and no *BRCA1/2* deleterious germline mutation.

The investigation for those 416 germline DNAs was done with massive parallel sequencing of *CDK12* coding regions in equimolar pools of eight DNAs. Any potentially deleterious *CDK12* mutation found was then checked by Sanger Sequencing. This approach limits the cost of investigation for large cohorts by looking at the same time at relevant mutations in different samples within one sequencing. However, the allelic frequency for each pool is much lower than the 50% expected in for a single germline DNA sample. Mutation calling depends mostly on low error rate and a high coverage of sequencing. The experiment was designed so that at least 320X covered each bases of the coding region for *CDK12* with therefore approximately 20X for each of allele of a pool - 16 allele in one pool of 8 germline DNAs.

The goal of my bioinformatic analysis was to reach the best sensitivity regardless of specificity for mutation calling, to find every possible relevant mutation despite the low allele frequency. Because any variant found is further validated in the eight DNAs of the relevant pool, specificity was of lesser importance. Recent variant callers rely on complex algorithmic and harbor high accuracy such as HaplotypeCaller<sup>306</sup>. I decided to merge the results of several mutation callers<sup>306-308</sup> and added a naive homemade script for indels. A variant was taken into consideration if it was called by at least by one approach. Mutation callers were also selected based on whether it is possible to specify a ploidy to account for an expected frequency of allele in the pool, here of 6.25%. This helps to retain variants at low frequency that might be discarded in their analysis. In the same idea, the variant filtering step was built to be lenient and to reach the best sensitivity at the cost of a lower specificity. Primers of sequencing were removed as they could overlap to amplicon sequences, therefore artificially diluting potential mutations at this kind of genomic location.



# Lack of evidence for *CDK12* as an ovarian cancer predisposing gene

Alexandre Eeckhoutte<sup>1,2</sup> · Mathilde Saint-Ghislain<sup>1,2</sup> · Manon Reverdy<sup>1,2</sup> · Virginie Raynal<sup>2,3</sup> · Sylvain Baulande<sup>3</sup> · Guillaume Bataillon<sup>4</sup> · Lisa Golmard<sup>5</sup> · Dominique Stoppa-Lyonnet<sup>1,5,6</sup> · Tatiana Popova<sup>1,2</sup> · Claude Houdayer<sup>5,7</sup> · Elodie Manié<sup>1,2</sup> · Marc-Henri Stern<sup>1,2,5</sup> 

© Springer Nature B.V. 2020

## Abstract

*CDK12* variants were investigated as a genetic susceptibility to ovarian cancer in a series of 416 unrelated and consecutive patients with ovarian carcinoma and who carry neither germline *BRCA1* nor *BRCA2* pathogenic variant. The presence of *CDK12* variants was searched in germline DNA by massive parallel sequencing on pooled DNAs. The lack of detection of deleterious variants and the observed proportion of missense variants in the series of ovarian carcinoma patients as compared with all human populations strongly suggests that *CDK12* is not an ovarian cancer predisposing gene.

**Keywords** *CDK12* · Cancer susceptibility · Ovarian carcinoma · Pool sequencing · NGS

## Introduction

Epithelial ovarian carcinoma (EOC) is a rare but dreadful disease and represents the 5th cause of death by cancer in women worldwide. The most frequent histology of this carcinoma is High Grade Serous Ovarian Carcinoma (HGSOC). The known risks factors of EOC are age, hormonal history and genetic factors [1]. Genetic factors are involved in approximately 15% of HGSOC, often associated

with the so-called hereditary breast and ovarian cancer syndrome (HBOC). *BRCA1* and *BRCA2*, which code key actors of the homologous recombination (HR) DNA repair pathway, are the two major cancer predisposing genes. Heterozygote germline mutations of *BRCA1* and *BRCA2* lead to the most important increased risk to develop an HGSOC with a Relative Risk (RR) of 40 and 18, respectively [2]. They explain 65% to 75% of hereditary EOC. In addition, mutations of other genes belonging to the HR pathway, such as *RAD51* paralogs, also participate to the HBOC syndrome [3–5]. A second group of predisposition genes belong to the MMR family (*MLH1*, *MSH2*, *MSH6*, *PMS2*), mutated in the Lynch syndrome. Lynch syndrome is associated with a RR of [3.6–13] for EOC, mainly of the endometrioid and clear cell subtypes, and explains ~10% of hereditary EOC [1, 6–8]. Strikingly, all these predisposition genes encode proteins involved in DNA maintenance.

Recently, *CDK12* emerged as an important player in ovarian carcinoma. *CDK12* (cyclin-dependent kinase 12) is one of the ten most frequently mutated genes in HGSOC (3% of the TCGA cohort) and is also mutated in 7% of metastatic castration-resistant prostate cancer [9]. *CDK12* behaves as a classical tumor suppressor gene with bi-allelic somatic inactivation in tumors, with in most cases one deleterious mutation on one allele and one chromosomal partial deletion evidenced by loss of heterozygosity (LOH). We have previously shown that *CDK12*-inactivated tumors are associated with an unusual form of genomic instability named

✉ Marc-Henri Stern  
marc-henri.stern@curie.fr

<sup>1</sup> Inserm U830, DNA Repair and Uveal Melanoma (D.R.U.M.), Equipe Labellisée par la Ligue Nationale Contre le Cancer, Institut Curie, 26 Rue d'Ulm, 75248 Paris, France

<sup>2</sup> Inserm U830, Institut Curie, PSL Research University, 26 Rue d'Ulm, 75248 Paris, France

<sup>3</sup> NGS Platform, Institut Curie, PSL Research University, 26 Rue d'Ulm, 75248 Paris, France

<sup>4</sup> Department of Biopathology, Institut Curie, PSL Research University, 26 rue d'Ulm, 75248 Paris, France

<sup>5</sup> Institut Curie, Hôpital, Service de Génétique, 26 Rue d'Ulm, 75248 Paris, France

<sup>6</sup> University Paris Descartes, Sorbonne Paris Cité, 12 Rue de l'École de Médecine, 75006 Paris, France

<sup>7</sup> Present Address: Department of Genetics, Normandy University, UNIROUEN, Inserm U1245, Normandy Centre for Genomic and Personalized Medicine, Rouen University Hospital, 37 Boulevard Gambetta, 76000 Rouen, France



the TD-plus phenotype and characterized by hundreds of large tandem duplications of up to 10 megabases in size [10]. *CDK12* is an essential gene during development as *Cdk12* inactivation is embryonic lethal in mouse models [11]. *CDK12* is a nuclear serine threonine kinase that dimerizes with Cyclin K (CCNK). Until recently, the only known targets of CCNK/*CDK12* were serines 2 and 5 of the carboxy-terminal-domain (CTD) of the RNA polymerase II, required for elongation and end termination of transcription. In *in-vitro* studies, *CDK12* is required for the expression of a subset of DNA Damage Response (DDR) genes, including *BRCA1*, *FANCI*, *FANCD2* [12], and conversely, *CDK12*-inactivated cell models are highly sensitive to PARP inhibitors [13]. Both *in vitro* studies and analyses of *CDK12*-mutated tumors strongly suggest that *CDK12* plays a role in genomic maintenance. Recent studies have showed that *CDK12* acts by suppressing intronic polyadenylation events, including in DNA repair genes [14–16]. *CDK12* phosphorylates 4E-BP1 to enable mTORC1-dependent translation and maintains mitotic genome stability [17].

*CDK12* as an important tumor suppressor gene in ovarian tumorigenesis pointed it as a potential predisposition gene for EOC. We explored this hypothesis by investigating the germline status of *CDK12* in a series of 416 unselected consecutive and unrelated patients with EOC, negative for *BRCA1* or *BRCA2* mutations.

## Material and methods

### Patients

A series of blood DNA from 416 unselected consecutive and unrelated patients was assembled from the Genetic Department of Institut Curie, initially explored negative for *BRCA1* or *BRCA2* deleterious germline mutations, using the current techniques at time of diagnosis or reanalysis. Five of them were subsequently found to carry deleterious mutations of *BRCA2* (1 case), *RAD51C* (1), *RAD51D* (1), *PMS2* (1) and *TP53* (1). All patients had personal history of EOC (mean age at diagnosis: 56 years-old) and benefited from genetic counseling. One hundred twenty-three of these patients had also developed one or more breast cancers. All patients have signed an informed consent for research of new cancer predisposing genes.

### CDK12 sequencing in pooled DNA and positive control pools

Germline DNAs of the 416 patients were quantified using Qubit dsDNA HS Assay Kit (Thermo Fisher Scientific). The 416 DNAs were pooled into 52 equimolar pools of eight DNAs per pool, with an expected variant frequency

of one mutated allele in 16 alleles (6.25%). Two additional pools were constituted as positive controls, containing seven patients' DNAs plus one tumor DNA with a known *CDK12* mutation, c.137del/p.Lys46SerfsX11 and c.212dup/p.Glu72GlyfsX3 for pools #53 and #54, respectively. *CDK12* coding sequence and flanking introns were sequenced using the TruSeq Custom Amplicon Low Input Kit (Illumina). Briefly, the design included 32 amplicons of 250 bp for a theoretical coverage of 100% on a cumulative target of 4.61 kb. The library was produced by PCR and ligation from 20 ng of pooled genomic DNA, barcoded with 54 indexes, quantified (Bioanalyzer, Agilent) and pooled in an equimolar ratio. The library was then paired-end sequenced (PE250) with a MiSeq v2 Nano flow cell (Illumina).

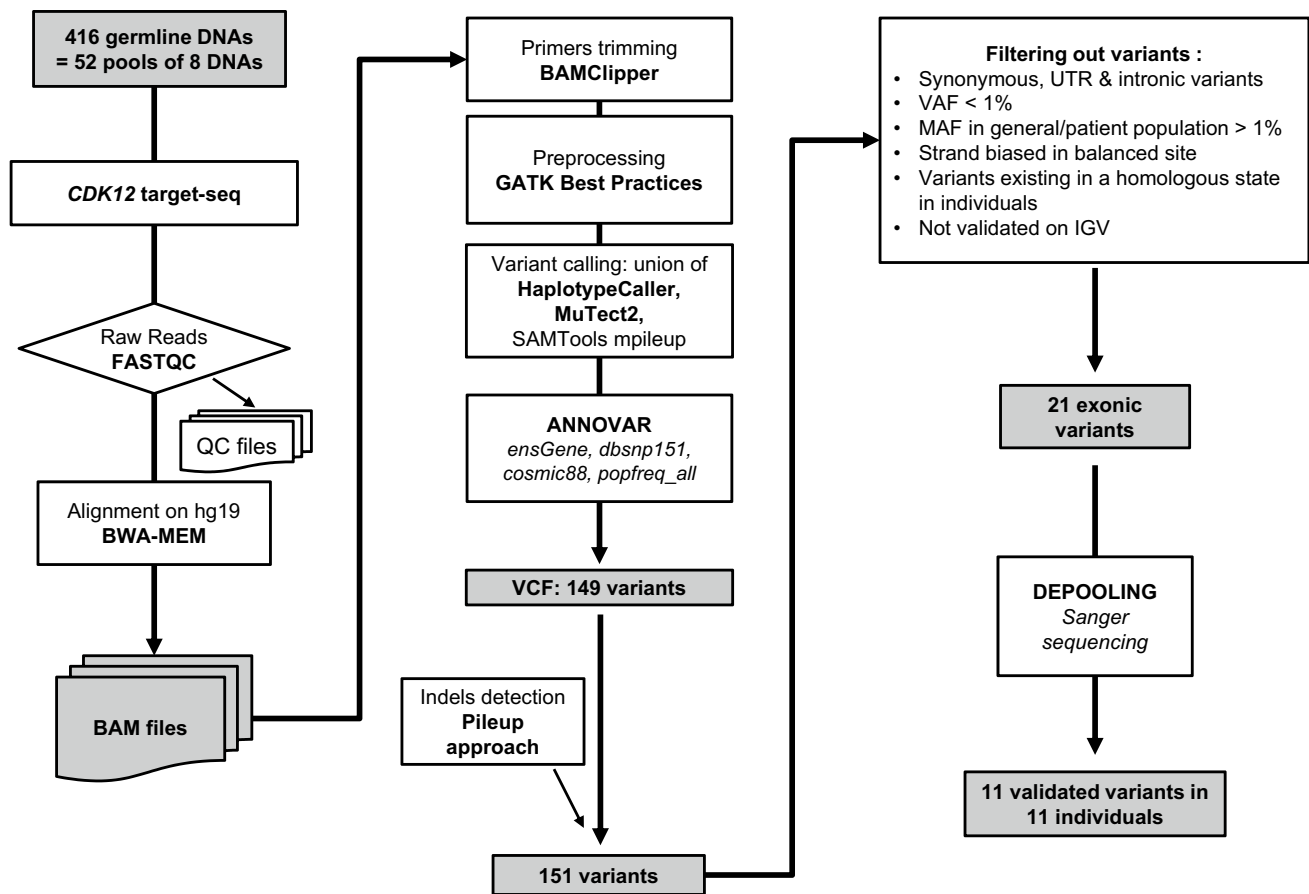
### Bioinformatics pipeline

Quality control was performed using FASTQC. Reads were aligned to the hg19 assembly with BWA MEM (v. 0.7.5a). Primers were soft-clipped with BAMclipper [18]. BAM files were pre-processed with indels realignment and base quality score recalibration according to the GATK Best Practices (v. 3.5) [19]. Variants were detected by Samtools mpileup (v. 1.7) [20], HaplotypeCaller (v. 3.5) and Mutect2 (v. 3.5). The union of all the variants called was annotated with ANNOVAR [21] according to different databases: EnsGene, COSMIC88, dbSNP151 and maximum allele frequency from 1000G, ExAC, ESP6500 and CG46. Variants were filtered out if: (i) synonymous, (ii) intronic and UTR located, (iii) biased for strand direction (outside of [0.4–0.6] ratio in a balanced site), (iv) frequency higher than 1% in any human reference population, (v) frequency higher than 1% in the patients' series, (vi) Variant Allele Frequency (VAF) < 1% and (vii) existing at a homozygous state in any individual. The remaining variants were then manually checked on IGV and five of them were discarded. A pileup approach was also implemented, with Samtools (v. 1.8) and a customized script to retain indels supported by more than five reads (Fig. 1). An *in silico* prediction of splice defects was also performed using the MaxEntScan tool (MaxEnt; [22]).

### Variant validation

The eight DNAs from each positive pool were analyzed independently by Sanger sequencing for the identified variant. Briefly, PCR was performed from 50 ng of DNA using specific primers and Taq Gold using standard protocols (Primers and conditions available on request) and sequenced using Big Dye Terminator kit V1 (3130XL, Applied Biosystems). Quality control of the electropherograms was performed using FinchTV (PerkinElmer) and sequences were analyzed using SeqScape (Applied Biosystems).





**Fig. 1** Variant calling workflow. Workflow of the *CDK12* pool targeted-sequencing, variant calling, filtering and validation, VCF variant calling format, *UTR* untranslated region, *VAF* variant allele fre-

quency in the pool, *MAF* mutation allele frequency in population, *IGV* Interactive Genome Viewer

## Prediction

In-silico predictions of deleterious consequences of the non-synonymous variants were performed using Combined Annotation Dependent Depletion (CADD) Phred score, SIFT (Sorting Intolerant from Tolerant), Polyphen-2 and VEP (Variant Effect Predictor).

## CDK12 defect genomic signature

Tumor DNA extracted from frozen or FFPE blocs from *CDK12*-variant carriers were obtained from the institutional Biobank and the Department of Pathology, respectively. Genomic profiling was obtained by shallow Whole Genome Sequencing (sWGS, approximately 1 read per base) of the tumor DNA on NovaSeq (Illumina). Adapters were trimmed with Cutadapt (v. 1.18). The number of reads in windows of 10 kb was extracted and normalized for GC content and mappability with ControlFREEC [23]. The *CDK12* TD plus pattern characteristic of *CDK12* inactivation was visually checked [10].

## Statistical power of the study

Assuming a distribution in *CDK12* variants following a Poisson law, the theoretical frequency to detect at least one deleterious variant in a cohort of 927 cases with a power of 80 was calculated as the following:

$$P(X \geq 1, \lambda) = 0.80 = 1 - P(X = 0, \lambda)$$

$$\lambda = \frac{-\ln(0.20)}{927} = 1.7e - 3$$

## Results

The goal of this study was to evaluate *CDK12* as a potential EOC predisposing gene. We thus defined the frequency of *CDK12* germline variants in a series of 416 consecutive patients with ovarian carcinomas. The Region Of Interest

(ROI) included 4525 bp corresponding to the coding sequence of *CDK12*, including 14 exons plus 2 base pairs of splicing sites. A pooled DNA sequencing approach was performed. The mean read depth on ROI was 920 X and the majority of ROI (92%) displayed more than 320 X coverage, corresponding to 20 X per allele. The lowest depth on ROI was ranging from 83 to 227 X in the different pools (mean 143 X). The less covered regions were parts of exons 1 and 2, and the whole exon 10.

In addition to the two positive controls, a list of 151 variants was called by at least one of the variant calling methods. 21 different variants in 17 different pools were retained for validation after filtering. Ten predicted variants were not confirmed by Sanger sequencing, and thus were considered as false calls, whereas eleven of these variants were validated in 11 different patients, among which 10 were single nucleotide variants (SNVs) and one an in-frame 3-base deletion (Fig. 1, Table 1). Out of these 11 variants, 9 were previously reported in the dbSNP database v151. All SNVs were missense variants with a predictive deleteriousness ranging from benign to moderate. One yet unreported SNP, c.A2712T, changes a glutamic acid codon conserved in all sequenced vertebrates up to lamprey and located within the kinase domain. However, the consequence of this change for aspartic acid was considered as mild. No Loss of Function (LoF) variant, such as premature stop-gain, frameshift or splicing variant, was found in this series.

As compared with reported frequencies of these variants in the representative non-Finnish European population in the GnomAD database, none of these variants were significantly enriched in our series of EOC patients (Table 1). As *CDK12* deleterious variants are embryonic-lethal at homozygous state, we considered only the missense SNPs reported in dbSNP151 and never found at homozygous state in any human population. The proportion of such missense SNPs was not significantly different in our series from that of the representative non-Finnish European population (10/832 alleles and 823/113,650 alleles respectively; Fisher's exact test:  $p=0.1453$ ), which is in accordance with *CDK12* not being an ovarian cancer predisposing gene.

Although no strong evidence supported the pathogenic effect of the identified variants, we further explored whenever possible the tumor of the corresponding variant carrier. We retrieved four EOC cases for whose tumor material was available. Given that *CDK12* is a tumor suppressor gene, we first assessed the loss of the wild-type allele in tumors, according to the Knudson/two-hit hypothesis. Only one of the four EOC had a loss of the wild allele, and the three others retained the wild-type allele in the tumors. A key feature of *CDK12*-inactivated tumors is a striking genomic profile enriched in numerous and very large tandem duplications, the TD-plus genomic signature [10], which was not found

in the genomic profiles of the four tested EOCs with *CDK12* variants (Table 1 and Fig. 2).

Altogether, we found no evidence of deleterious *CDK12* germline variants in our series of ovarian carcinoma, and no evidence for its role as an ovarian susceptibility gene in the studied population.

## Discussion

*CDK12* recently emerged to play an important role in ovarian and prostatic carcinomas, as a tumor suppressor gene contributing to malignant transformation and genomic instability when inactivated. As such, *CDK12* was a good candidate to also play a role in cancer predisposition. This study evaluated the incidence of *CDK12* germline variants in a series of 416 unrelated and consecutive patients with ovarian carcinoma. A total of eleven *CDK12* exonic variants were identified by massive parallel sequencing and validated by Sanger sequencing. None of the variants was a Loss of Function variant (LoF). However, one was in the kinase domain on a well conserved codon up to lamprey, but with a mild acid to acid change of coded amino-acid. Unfortunately, no tumor sample was available for further investigation of this case. We then compared the proportion of missense variants found in our series from that of the representative non-Finnish European population and found no statistically significant difference. We further mined four variants, for which tumor samples were retrieved and no evidence of *CDK12* inactivation was found in these tumors.

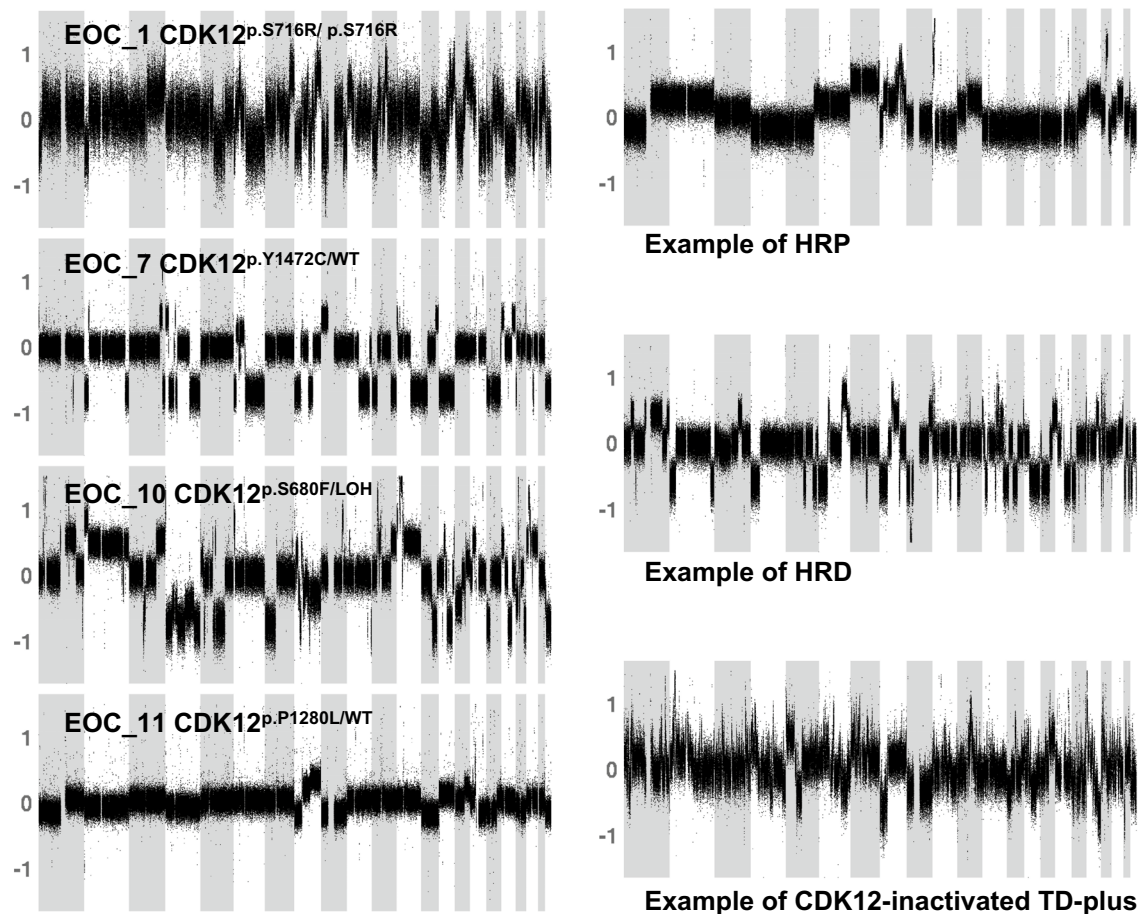
Altogether, we found no evidence for a role of *CDK12* as an ovarian cancer predisposition gene in an unbiased series of 416 *BRCA1/2*-wild-type patients with ovarian carcinoma. Furthermore, the analysis of The Cancer Genome Atlas (TCGA) series of 511 ovarian serous cystadenocarcinomas identified 15 cases with the TD-plus phenotype and bi-allelic inactivation of *CDK12*, none of which carrying any deleterious germline variant [10 and Popova et al., unpublished data]. These analyses of combined series (TCGA and in-house cases) had a power of 80% to detect at least one deleterious germline mutation in the hypothesis of a deleterious variant frequency of  $1.7 \times 10^{-3}$  in EOC patients.

Interestingly, the number of LoF variants in Gnomad ( $n=6$ ) is largely below the expected one, with a probability of being loss-of-function intolerant (pLI) score at 1 [24], and an observed / expected (oe) score of 0.05 (gnomad.broad-institute.org). This suggests that LoF variants are counter-selected in human populations, even at the heterozygous state. Interestingly, a report described the existence of a deleterious c.1047-2A>G germline *CDK12* variant in 8 of the 106 HBOC cases tested (7.6%) in the Tatar population [25], but the association with HBOC was not confirmed in a replication study [26]. If replicated, this would suggest that

**Table 1** Description of *CDK12* variants

Samples	Coordinates	Ref	Alt	Variants	dbsnp151	Pop max freq	Domain	CADD score	VEP	Pph2 status	SIFT_status	LOH	sWGS
EOC_1	chr17:37,649,046–37,649,046	C	A	c.C2148A/p.S716R	rs777401578	0.00006486	–	13.78	Moderate	Benign	Tolerated	<b>ROH</b>	No TD-Plus
EOC_2	chr17:37,687,363–37,687,363	G	A	c.G4267A/p.A1423T	rs201512860	0.0002166	–	20	Moderate	Benign	Tolerated	NA	NA
EOC_3	chr17:37,667,830–37,667,830	A	T	c.A2712T/p.E904D	–	0	Kinase	24.9	Moderate	Probably damaging	Deleterious	NA	NA
EOC_4	chr17:37,682,310–37,682,310	G	T	c.G3498T/p.Q1166H	–	0	–	20.2	Moderate	Benign	Deleterious	NA	NA
EOC_5	chr17:37,627,577–37,627,577	C	G	c.C1489G/p.Q497E	rs766575927	0.00002639	–	19.9	Moderate	Benign	Tolerated	NA	NA
EOC_6	chr17:37,618,415–37,618,417	AAC	–	c.92_94del/p.31_32del	rs780413687	0.00003517	–	-	Moderate	-	-	NA	NA
EOC_7	chr17:37,687,511–37,687,511	A	G	c.A4415G/p.Y1472C	rs373240630	0.00049 (AFR)	–	23.8	Moderate	Benign	Deleterious	<b>ROH</b>	No TD-Plus
EOC_8	chr17:37,682,202–37,682,202	C	G	c.C3390G/p.I1130M	rs376340730	0.00006195	–	14.23	Moderate	Benign	Tolerated	NA	NA
EOC_9	chr17:37,682,501–37,682,501	A	G	c.A3692G/p.N1231S	rs538854021	0.00005544 (EAS)	–	17.99	Moderate	Benign	Tolerated	NA	NA
EOC_10	chr17:37,646,920–37,646,920	C	T	c.C2039T/p.S680F	rs375518105	0.000155	–	28.3	Moderate	Probably damaging	Deleterious	<b>LOH</b>	No TD-Plus
EOC_11	chr17:37,686,935–37,686,935	C	T	c.C3839T/p.P1280L	rs148965508	0.006415 (AFR)	–	23.6	Moderate	Benign	Deleterious	<b>ROH</b>	No TD-Plus

*Ref* reference allele, *Alt* alternative allele, *Pop max freq* maximum allele frequency in any human population (Non-Finish European—NFE—by default), *AFR* African population, *EAS* East-Asian, *EOC* epithelial ovarian carcinoma, *LOH* loss of heterozygosity, *ROH* retention of heterozygosity, *NA* not available, *sWGS* shallow Whole Genome Sequencing, *TD-Plus* tandem duplication-plus genome profile, characteristic of *CDK12* inactivation



**Fig. 2** Genome profiling of epithelial ovarian carcinoma carrying *CDK12* variants. Genome profiling generated using low coverage whole genome sequencing (shallow WGS). Left panel: four epithelial carcinomas in patients carrying germline *CDK12* variants. Right

panel: From top to bottom, examples of tumor genome profiles with Homologous Recombination Proficient (HRP), Homologous Recombination Deficient (HRD) and *CDK12*-inactivated TD-plus profiles, respectively

*CDK12* variants could play a role in HBOC predisposition in some human populations, although strongly counter-selected in most human populations. The reason of this counter-selection is not clear. *Cdk12* inactivation is embryonic lethal in mouse models, but heterozygous *Cdk12*<sup>Δ/wt</sup> pups are viable and born with the expected frequency [11]. Clearly, long-term follow-up of these *Cdk12*<sup>Δ/wt</sup> mice may be instrumental to unravel the mechanism of intolerance of LoF variants in Humans. In a more distant model in *Drosophila*, a decline of courtship learning was observed in *CDK12* heterozygous flies [27]. This could be a plausible mechanism to explain the counter-selection of human heterozygous *CDK12*-mutant carriers, but caution should be taken given the evolutionary distance between flies and mammals.

In conclusion, our data evidenced the absence of deleterious *CDK12* variants in patients with ovarian carcinoma, confirming the rarity of such variants in the general population, and making unlikely the existence of deleterious

variants in more than 0.2% of EOC patients. Thus our data do not support the role of *CDK12* in ovarian carcinoma susceptibility. The origin of the intolerance of deleterious *CDK12* variants in the population has yet to be explained.

**Acknowledgements** Supported by the *Institut National de la Santé et de la Recherche Médicale* (INSERM) and the *Institut Curie*. M. S-G and A. E. are supported by fellowships from the *Université de Rouen* and the *Ligue Nationale Contre le Cancer*, respectively. The Institut Curie ICGex NGS platform is funded by the *EQUIPEX investissements d'avenir* program (ANR-10-EQPX-03) and ANR10-INBS-09-08 from the *Agence Nationale de la Recherche*. We thank the patients. We also acknowledge support from the *Institut Curie* for sample collection, banking and processing: the Biological Resource Center and its members (O. Mariani).

**Author contributions** Conception and design: EM, CH, M-HS. Development of methodology: AE. Acquisition of data: EM, AE, MS-G, MR, VR, SB, TP, GB, LG. Analysis and interpretation of data (e.g., statistical analysis, biostatistics, computational analysis): EM, AE, M-HS. Writing, review, and/or revision of the manuscript: EM, DS-L, AE, M-HS. Study supervision: DS-L, SB, CH, M-HS.

## Compliance with ethical standards

**Conflict of interest** E. Manié, T. Popova and M.-H. Stern are named inventors of a patent licensed to Myriad Genetics. No potential conflicts of interest were disclosed by the other authors.

## References

- Lheureux S, Gourley C, Vergote I, Oza AM (2019) Epithelial ovarian cancer. *Lancet* 393:1240–1253. [https://doi.org/10.1016/S0140-6736\(18\)32552-2](https://doi.org/10.1016/S0140-6736(18)32552-2)
- Kuchenbaecker KB, Hopper JL, Barnes DR et al (2017) Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. *JAMA* 317:2402–2416. <https://doi.org/10.1001/jama.2017.7112>
- Golmard L, Castera L, Krieger S et al (2017) Contribution of germline deleterious variants in the RAD51 paralogs to breast and ovarian cancers. *Eur J Hum Genet* 25:1345–1353. <https://doi.org/10.1038/s41431-017-0021-2>
- Loveday C, Turnbull C, Ramsay E et al (2011) Germline mutations in RAD51D confer susceptibility to ovarian cancer. *Nat Genet* 43:879–882. <https://doi.org/10.1038/ng.893>
- Loveday C, Turnbull C, Ruark E, et al. (2012) Germline RAD51C mutations confer susceptibility to ovarian cancer. *Nat Genet* 44:475–476; author reply 6 <https://doi.org/10.1038/ng.2224>
- Bewtra C, Watson P, Conway T, Read-Hippee C, Lynch HT (1992) Hereditary ovarian cancer: a clinicopathological study. *Int J Gynecol Pathol* 11:180–187
- Helder-Woolderink JM, Blok EA, Vasen HF, Hollema H, Mourits MJ, De Bock GH (2016) Ovarian cancer in Lynch syndrome; a systematic review. *Eur J Cancer* 55:65–73. <https://doi.org/10.1016/j.ejca.2015.12.005>
- Bonadona V, Bonaiti B, Olschwang S et al (2011) Cancer risks associated with germline mutations in MLH1, MSH2, and MSH6 genes in Lynch syndrome. *JAMA* 305:2304–2310. <https://doi.org/10.1001/jama.2011.743>
- Wu YM, Cieslik M, Lonigro RJ et al (2018) Inactivation of CDK12 delineates a distinct immunogenic class of advanced prostate cancer. *Cell* 173(1770–82):e14. <https://doi.org/10.1016/j.cell.2018.04.034>
- Popova T, Manie E, Boeva V et al (2016) Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications. *Cancer Res* 76:1882–1891. <https://doi.org/10.1158/0008-5472.CAN-15-2128>
- Juan HC, Lin Y, Chen HR, Fann MJ (2016) Cdk12 is essential for embryonic development and the maintenance of genomic stability. *Cell Death Differ* 23:1038–1048. <https://doi.org/10.1038/cdd.2015.157>
- Blazek D, Kohoutek J, Bartholomeeusen K et al (2011) The Cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. *Genes Dev* 25:2158–2172. <https://doi.org/10.1101/gad.16962311>
- Bajrami I, Frankum JR, Konde A et al (2014) Genome-wide profiling of genetic synthetic lethality identifies CDK12 as a novel determinant of PARP1/2 inhibitor sensitivity. *Cancer Res* 74:287–297. <https://doi.org/10.1158/0008-5472.CAN-13-2541>
- Tien JF, Mazloomian A, Cheng SG et al (2017) CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. *Nucleic Acids Res* 45:6698–6716. <https://doi.org/10.1093/nar/gkx187>
- Krajewska M, Dries R, Grassetti AV et al (2019) CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. *Nat Commun* 10:1757. <https://doi.org/10.1038/s41467-019-09703-y>
- Dubbury SJ, Boutz PL, Sharp PA (2018) CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. *Nature* 564:141–145. <https://doi.org/10.1038/s41586-018-0758-y>
- Choi SH, Martinez TF, Kim S et al (2019) CDK12 phosphorylates 4E-BP1 to enable mTORC1-dependent translation and mitotic genome stability. *Genes Dev* 33:418–435. <https://doi.org/10.1101/gad.322339.118>
- Au CH, Ho DN, Kwong A, Chan TL, Ma ESK (2017) BAM-Clipper: removing primers from alignments to minimize false-negative mutations in amplicon next-generation sequencing. *Sci Rep* 7:1567. <https://doi.org/10.1038/s41598-017-01703-6>
- DePristo MA, Banks E, Poplin R et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat Genet* 43:491–498. <https://doi.org/10.1038/ng.806>
- Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* 25:2078–2079. <https://doi.org/10.1093/bioinformatics/btp352>
- Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* 38:e164
- Yeo G, Burge CB (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J Comput Biol* 11:377–394. <https://doi.org/10.1089/1066527041410418>
- Boeva V, Popova T, Bleakley K et al (2012) Control-FREEC: a tool for assessing copy number and allelic content using next-generation sequencing data. *Bioinformatics* 28:423–425. <https://doi.org/10.1093/bioinformatics/btr670>
- Lek M, Karczewski KJ, Minikel EV et al (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536:285–291. <https://doi.org/10.1038/nature19057>
- Brovkina OI, Shigapova L, Chudakova DA et al (2018) The ethnic-specific spectrum of germline nucleotide variants in dna damage response and repair genes in hereditary breast and ovarian cancer patients of tatar descent. *Front Oncol* 8:421. <https://doi.org/10.3389/fonc.2018.00421>
- Bogdanova NV, Schurmann P, Valova Y et al (2019) A splice site variant of CDK12 and breast cancer in three Eurasian populations. *Front Oncol* 9:493. <https://doi.org/10.3389/fonc.2019.00493>
- Pan L, Xie W, Li KL et al (2015) Heterochromatin remodeling by CDK12 contributes to learning in *Drosophila*. *Proc Natl Acad Sci USA* 112:13988–13993. <https://doi.org/10.1073/pnas.1502943112>

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

## **Conclusion: “Lack of evidence for CDK12 as an ovarian cancer predisposing gene”**

The analysis of 416 unselected patients with an history of EOC and no *BRCA1/2* genes germline mutation did not display any evidence of deleterious *CDK12* germline variants. Moreover, the proportion of missense SNPs was not significantly different from our cohorts and the representative non-Finnish European population in GnomAD - 10/832 alleles and 823/113650 alleles, respectively (Fisher's exact test: p value = 0.1453). The identified missense variants were further investigated by looking at the available tumors for the loss of the wild-type allele. Only one out of four had an LOH for the wild type allele. The specific *CDK12*<sup>-/-</sup> genomic phenotype comprised of numerous large tandem duplications<sup>296</sup> was not observed on the four different tumor genomic profiles. Moreover, the exploration of the TCGA ovarian cohort did not show any *CDK12* deleterious germline mutations. When combining the TCGA ovarian cases and our in-house cohorts, the study had a power 80% to find at least a deleterious variant with a frequency of 1.7e-3 under the assumption of a Poisson distribution. Altogether, we find no evidence that *CDK12* is a cancer predisposition gene in the cohorts we studied. The fact was we could not investigate all the tumors associated with *CDK12* germline missense mutations and given the statistical power of our study because of the number of patients enrolled do not allow us to completely rule out *CDK12* as a cancer predisposition gene in EOC, but it provides confounding evidence.














## Introduction: “Germline MBD4 Mutations and Predisposition to Uveal Melanoma”

Base Excision Repair (BER) is a DNA reparation pathway that maintains genomic stability at the base level by removing damaged base or inappropriate base, thus preventing potential mutations and fork stalling during replication<sup>309</sup>. BER depends on the initial action of DNA glycosylases that find and remove damaged base. Eleven type of glycosylases have been identified in mammals with preferential base subtract for each<sup>310</sup>. One of those glycosylases is the Methyl-CpG Binding Domain 4 (MBD4) protein. MBD4 removes thymine (T) and uracil (U) followed by a guanine (G). Those substrates emerge generally from the spontaneous deamination of 5-methylcytosine (5mC) residue to a thymine and of a cytosine (C) to an Uracil (U), respectively, provoking mutagenesis with C to T mutation. Importantly, the transition from 5mC to thymine is the most frequent transition in human<sup>311,312</sup> and appears mostly at CG dinucleotide, the so-called CpG sites. MBD4 mostly associates with those abundant methylated-CpG sites<sup>313</sup>, actively maintaining genomic stability at CpG sites.

In 2018, our lab investigated a patient with a metastatic Uveal Melanoma (UM) that showed an exceptional response to immune therapy with anti-Programmed cell death protein 1 (anti-PD1). Strikingly, this patient showed a high Tumor Mutation Burden (TMB) and the characteristic CpG > TpG phenotype associated to the SBS1 signature<sup>4</sup>. Consistently with the function of MBD4 and the mutational phenotype observed, a germline deleterious mutation in the gene and its somatic inactivation was found in the patient. Further investigation of TCGA cohort for the CpG > TpG mutational phenotype harvested two more cases, one in UM and one in glioblastoma, both carrying a deleterious germline mutation of *MBD4* and its somatic inactivation in the tumor<sup>314</sup>. Later study also found an increase mutation rate CpG > TpG, the SBS1 signature, in two UMs with germline loss-of-function of *MBD4*<sup>315</sup>. This motivated us to investigate whether *MBD4* is a cancer predisposition gene in UM.

In this study published in *Journal of the National Cancer Institute*, we evaluated if *MBD4* is a cancer predisposition gene for UM by investigating the blood DNA of 1093 consecutive patients that all have an history of primary UM and 192 UM tumors with monosomy 3 (M3). I participated in this study by doing bioinformatic and statistical analyses for mutations in the germline and tumor cohorts. The bioinformatics pipeline was built in a similar fashion to the one previously established for *CDK12* mutation. More recent versions and a supplementary tool<sup>316</sup> were used that can also be modulate the ploidy expected inside the pool. Tumor DNA pools were tested with different expected ploidy because of potential contamination, without affecting the results. Of importance, this new pipeline validated the relevant mutations found when investigating *CDK12* mutations in EOC while not finding any new mutation.

## Germline MBD4 Mutations and Predisposition to Uveal Melanoma

Anne-Céline Derrien, MSc <sup>1,†</sup> Manuel Rodrigues, MD, PhD,<sup>1,2, †</sup> Alexandre Eeckhoutte, BSc <sup>1</sup>  
Stéphane Dayot, BSc <sup>1</sup> Alexandre Houy, BSc <sup>1</sup> Lenha Mobuchon, PhD <sup>1</sup> Sophie Gardrat, MD,<sup>1,3</sup>  
Delphine Lequin, MD,<sup>3</sup> Stelly Ballet, MSc,<sup>3</sup> Gaëlle Pierron, PhD <sup>3</sup> Samar Alsafadi, PharmD, <sup>1,4</sup>  
Odette Mariani, PhD <sup>5</sup> Ahmed El-Marjou, PhD <sup>6</sup> Alexandre Matet, MD, PhD <sup>7,8</sup>  
Christelle Colas, MD, PhD,<sup>9</sup> Nathalie Cassoux, MD, PhD,<sup>7,8</sup> Marc-Henri Stern, MD, PhD <sup>1,9,\*</sup>

<sup>1</sup>Inserm U830, DNA Repair and Uveal Melanoma (D.R.U.M.), Equipe Labellisée Par la Ligue Nationale Contre le Cancer, Paris, France; <sup>2</sup>Department of Medical Oncology, Institut Curie, PSL Research University, Paris, France; <sup>3</sup>Department of Biopathology, Institut Curie, PSL Research University, Paris, France; <sup>4</sup>Translational Research Department, Institut Curie, PSL Research University, Paris, France; <sup>5</sup>Biological Resource Center, Institut Curie, PSL Research University, Paris, France; <sup>6</sup>Institut Curie, PSL Research University, UMR144, Recombinant Protein Facility, Paris, France; <sup>7</sup>Department of Ocular Oncology, Institut Curie, Paris, France; <sup>8</sup>Faculty of Medicine, University of Paris Descartes, Paris, France and <sup>9</sup>Department of Genetics, Institut Curie, Paris, France

\*Correspondence to: Marc-Henri Stern, MD, PhD, 26 rue d'Ulm, 75248 Paris cedex 05, France (e-mail: marc-henri.stern@curie.fr).

†Authors contributed equally to this work.

### Abstract

**Background:** Uveal melanoma (UM) arises from malignant transformation of melanocytes in the uveal tract of the eye. This rare tumor has a poor outcome with frequent chemo-resistant liver metastases. BAP1 is the only known predisposing gene for UM. UMs are generally characterized by low tumor mutation burden, but some UMs display a high level of CpG>TpG mutations associated with MBD4 inactivation. Here, we explored the incidence of germline MBD4 variants in a consecutive series of 1093 primary UM case patients and a series of 192 UM tumors with monosomy 3 (M3). **Methods:** We performed MBD4 targeted sequencing on pooled germline (n = 1093) and tumor (n = 192) DNA samples of UM patients. MBD4 variants (n = 28) were validated by Sanger sequencing. We performed whole-exome sequencing on available tumor samples harboring MBD4 variants (n = 9). Variants of unknown pathogenicity were further functionally assessed. **Results:** We identified 8 deleterious MBD4 mutations in the consecutive UM series, a 9.15-fold (95% confidence interval = 4.24-fold to 19.73-fold) increased incidence compared with the general population (Fisher exact test,  $P = 2.00 \times 10^{-5}$ , 2-sided), and 4 additional deleterious MBD4 mutations in the M3 cohort, including 3 germline and 1 somatic mutations. Tumors carrying deleterious MBD4 mutations were all associated with high tumor mutation burden and a CpG>TpG hypermutator phenotype. **Conclusions:** We demonstrate that MBD4 is a new predisposing gene for UM associated with hypermutated M3 tumors. The tumor spectrum of this predisposing condition will likely expand with the addition of MBD4 to diagnostic panels. Tumors arising in such a context should be recognized because they may respond to immunotherapy.

Uveal melanoma (UM) is the most frequent primary intraocular tumor in adults with an overall mean incidence of 5.2 per million per year in the United States (1). Metastases arise in more than 30% of case patients, almost invariably in the liver, with a dismal prognosis because of the absence of effective treatment (median survival of 10 months) (2,3). UMs with high risk of developing metastases are characterized by loss of chromosome 3 and by BAP1 (encoded in 3p21) inactivation resulting from loss-of-function (LoF) mutations and loss of the remaining wild-type copy on chromosome 3 (4). Rare familial UMs are associated with germline mutations of BAP1 (Mendelian Inheritance in Man [MIM]: 614327) (5,6), which is the only known highly

penetrant UM predisposition gene. UM mainly affects individuals of European ancestry and is associated with fair skin and light iris color. However, the low tumor mutation burden (TMB) and lack of an ultraviolet-associated mutational signature argue against a role for ultraviolet radiation in UM oncogenesis (7).

Recently, the characterization of a metastatic UM patient with an exceptional response to anti-Programmed cell death protein 1 (anti-PD-1) therapy led us to identify a CpG>TpG mutator phenotype linked to germline protein truncating variants (PTV) in MBD4 (Methyl-CpG Binding Domain Protein 4) and somatic loss of the wild-type allele in tumors in 2 patients with UM and 1 with glioma (8). Another UM patient responding to

Received: December 20, 2019; Revised: March 19, 2020; Accepted: March 26, 2020

© The Author(s) 2020. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)



immune checkpoint inhibitors was subsequently reported with a germline *MBD4* PTV (9). *MBD4* encodes a glycosylase involved in the base excision repair of DNA damage arising from spontaneous deamination of 5-methylcytosine to thymine (10,11), which is consistent with the mCpG>TpG transitions [mutational signature SBS1 (12)] observed in *MBD4*-inactivated tumors (12). *MBD4*, located on chromosome 3, is thought to act as a tumor suppressor gene, following Knudson's 2-hit model with loss of the wild-type allele by monosomy 3 (M3) in UMs (8). Altogether, these germline deleterious *MBD4* variants in UM prompted us to investigate the role of *MBD4* as a predisposing gene for UM. Here, we performed *MBD4* targeted-sequencing in germline DNA of a large consecutive series of 1093 UM patients and in tumor DNA of a second cohort of 192 UM patients with M3 and investigated the TMB and mutational signature in patients harboring *MBD4* mutations.

## Methods

### Study Patients

The 1099 individuals with UM were diagnosed at Institut Curie, France, from 2013 to 2018. The sex proportion of female to male was 52.2 to 47.8 ± 3.0% (95% confidence interval [CI]) and the median age at diagnosis was 64 years old (Q1-Q3 quartile interval = 54-73). All patients provided written informed consent to perform germline genetic analyses and somatic genetic analyses of tumor samples. Six patients were subsequently removed from the study: the UM diagnosis was not confirmed for 5 patients, and the sixth patient had undergone a bone marrow transplantation and his blood sample corresponded to his donor's (Supplementary Figure 1A, available online). The study was conducted in accordance with the declaration of Helsinki and was approved by the ethical committee and institutional review board of the Institut Curie. Germline DNA was extracted from the blood of all patients (DNeasy Blood and Tissue kit, Qiagen). When available, tumor genomic status as part of the prognostication assessment was extracted from the medical record and classified as M3, including isodisomy 3) or disomy 3 (D3). Tumor samples were collected from primary eye tumors. A second series of 192 UM tumor samples with M3 was also accrued at Institut Curie, of which 120 patients were independent from the consecutive germline cohort.

### *MBD4* Targeted Sequencing

Germline DNA of 1099 UM patients from the UM consecutive series (before removal of the 6 aforementioned patients) and DNA of a series of 192 M3 UM tumors were screened for *MBD4* variants by pooled *MBD4* targeted sequencing. Details on the sequencing strategy and bioinformatics pipeline are described in the Supplementary Methods (available online). Deconvolution of the identified pooled DNA samples with an *MBD4* variant was carried out by Sanger sequencing.

Identified *MBD4* variants (Supplementary Table 1, available online) were defined following the recommendations of the Human Genome Variation Society (<http://varnomen.hgvs.org/>) and numbered based on the *MBD4* (MIM: 603574) cDNA and protein sequences (GenBank accession numbers NM\_003925.2 and NP\_003916.1, respectively).

### Glycosylase Activity Assay

Wild-type and mutant *MBD4* were expressed to assess their enzymatic activity by in vitro *MBD4* glycosylase assay as previously described (13,14) (Supplementary Methods, available online).

### Whole-Exome Sequencing (WES) and Mutation Calling

WES was performed on tumor samples from *MBD4* variant carriers who consented to germline studies (Supplementary Table 2, available online). Variant calling and TMB analysis are described in Supplementary Methods (available online).

### Statistical Analysis

Statistical analysis was performed using R software v3.6. Fisher exact test was used to calculate *P* values between our cohort and the general population. Random subsampling 1 000 000 times of 2186 alleles from the GnomAD cohort was also used as statistical "matching" strategy to calculate *P* values between our cohort and the general population. A Mann-Whitney *U* test was used to compare the median ± median absolute deviation for the mutation burden and CpG>TpG proportion between *MBD4*-deficient (*MBD4*<sub>def</sub>) and *MBD4*-proficient patients. To compare age of UM onset, a Wilcoxon Rank-Sum test was used. For survival analysis using Kaplan-Meier curves, statistical analysis was carried out using the log-rank test. All statistical tests were 2-sided, except for 1-sided Wilcoxon Rank-Sum test for early age of onset in *MBD4*<sub>def</sub> patients. Confidence intervals were carried out at 95% confidence level. Confidence intervals for relative risk (RR) measurement was calculated as previously described (15). A *P* value less than .05 was considered to be statistically significant.

## Results

### Mining *MBD4* Germline Variants in Public UM Cohorts

To evaluate the potential predisposing role of *MBD4* in UM, we mined all available public UM cohorts for germline *MBD4* variants. We identified 1 case harboring the germline deleterious *MBD4* PTV c.1443delT (p.Leu482Trpfs\*9) in a first cohort containing 37 UM patients (phs001421.v1.p1) (16). A second cohort of 98 UM patients (phs000823.v1.p1) included a second case with an *MBD4* c.1020delA (p.Asp341Thrfs\*13) PTV. Collectively, 5 *MBD4* germline deleterious variants were found in 268 analyzed UM patients (1.9%) (8,9,16–18). In contrast, such variants are exceedingly rare in an unselected population (88 of approximately 125 000 individuals in GnomAD v2.1.1). Out of these 5 UM case patients with germline *MBD4* variants, 4 had available tumor profiles that all showed M3 and somatic *BAP1* inactivation (8,9,16), presumably because of the localization of both *MBD4* and *BAP1* on chromosome 3.

### Identification of *MBD4* Germline Variants in the In-House Consecutive UM Series

To assess the actual prevalence of *MBD4* germline deleterious variants in UM, we next explored an in-house cohort of 1093 (approximately one-half the annual incidence in United States) consecutive patients diagnosed with UM at Institut Curie between 2013 and 2018. Targeted next-generation sequencing in

pooled patient DNA followed by Sanger sequencing revealed germline *MBD4* PTVs in 7 patients (Table 1): 2 splice site variants (c.1562-1G>T [p.Asp521Profs\*4] in 2 patients and a c.335+1G>A [p.Arg83Profs\*5] variant), 2 frameshift deletion variants (c.1443delT [p.Leu482Trpfs\*9] and c.1384delG [p.Ala462Leufs\*29]), and 1 stop-gain near the end of the last exon of *MBD4* (c.1706G>A [p.Trp569\*] in 2 patients).

We also identified and characterized 9 rare germline *MBD4* variants (frequency <1% in the general population): 7 missense variants in 13 patients and 2 intronic variants in 2 patients (Figure 1A). Out of these, 3 were predicted with possible splicing consequences: 2 missense variants (c.1652A>G [p.Asn551Ser] and c.1400A>G [p.Asn467Ser]), and 1 intronic variant (c.1277-18T>A) (Supplementary Table 1, available online).

### Functional Assessment of *MBD4* Variants

Exon-trapping assays performed on the 3 aforementioned variants demonstrated the use of an alternative acceptor site with c.1277-18T>A, albeit to a lesser extent than the canonical splice site, whereas c.1652A>G and c.1400A>G did not show any measurable effect (Supplementary Figure 2, available online).

All missense variants within the *MBD4* glycosylase domain [aa425-580] were assessed by in vitro glycosylase assay together with p. Trp569\* because of its localization near the end of the protein (Figure 1; Table 1). The assay confirmed that p. Trp569\* results in a catalytically inactive protein and further demonstrated p. Arg468Trp to be a LoF variant (Figure 1). Consistent with this finding, the key role of Arg468 was previously established in binding at the G/T mismatch site, maintaining the T base in a position required for catalysis, and interacting with the orphan G base through hydrogen bonding (22).

We next characterized by WES the 3 available tumors from patients carrying germline *MBD4* LoF variants (UM75: p. Trp569\*, UM605: p. Ala462Leufs\*29, and UM656: p. Leu482Trpfs\*9) (8). We confirmed that the 3 *MBD4* germline LoF mutations were associated with somatic loss of the wild-type allele in the tumors by either M3 (UM75 and UM605) or isodisomy 3 (UM656) (Figure 1C; Supplementary Table 1, available online), consistent with the 3q21.3 location of *MBD4*.

### *MBD4* Mutations in the M3 Tumor Series and Tumor Signature

To further evaluate the incidence of *MBD4* alterations in UM, and assuming that *MBD4*<sub>def</sub> UMs are associated with M3, we accrued a series of 192 UM tumor samples with M3 (of which 120 case patients were independent from the present consecutive UM series) and screened for *MBD4* mutations using the aforementioned strategy (Supplementary Figure 1B, available online). We identified 6 additional *MBD4* variants, including 4 LoF mutations (UMT62: c.1688T>A [p.Leu563\*], UMT45: c.1562-1G>T [p.Asp521Profs\*4], UMT61: c.1002delTTTG [p.Lys335Phefs\*18], UMT162: c.541C>T [p.Arg181\*]) and 2 missense variants (UMT88: c.1402C>T [p.Arg468Trp] and UMT105: c.1073T>C [p.Ile358Thr]) (Table 1; Figure 1; Supplementary Table 1, available online). Of these, 4 patients (UMT45, UMT61, UMT162, and UMT88) had available germline DNA and all consented to germline studies. Characterization by WES of their tumor samples showed that the 3 tested LoF mutations were germline variants, the missense p. Arg468Trp was somatic, and all were associated with Loss Of Heterozygosity (LOH) of the wild-type allele in tumors (Figure 1, A and B; Supplementary Table 1, available online).

*MBD4* inactivation has been associated with a high TMB and a CpG>TpG mutational pattern (8). We confirmed the high TMB in all 3 available *MBD4*<sub>def</sub> UMs from the germline consecutive cohort with 275, 122, and 181 variants per exome in UM75, UM605, and UM656, respectively, compared with  $16 \pm 4.0$  (median  $\pm$  median absolute deviation) variants in *MBD4*-proficient UMs (18) (Figure 1C; Supplementary Table 2, available online). CpG>TpG transitions represented 96.4%, 85.7%, and 92.8% of all single nucleotide variants (SNVs), respectively, compared with  $24.3 \pm 7.6\%$  in *MBD4*-proficient UMs (18) (Figure 1, C and D; Supplementary Table 2, available online). In line with the glycosylase assay, TMB results and somatic chromosome 3 LOH further confirmed the deleterious effect of p. Trp569\* in UM75 (Figure 1). Similarly, within the M3 UM tumor series, the 3 available UMs carrying a germline LoF *MBD4* variant exhibited a high TMB (269, 288, and 86 variants per exome in UMT162, UMT45, and UMT61, respectively) and a predominance of CpG>TpG transitions (85.6%, 94.4%, and 63.9%, respectively) among all SNVs (Figure 1, C and D; Supplementary Table 2, available online). The tumor sample of patient UMT88, carrying a somatic p. Arg468Trp variant identical to that found as a germline variant in UM293, also carried a high TMB (243 variants) and the CpG>TpG mutational pattern (92.5%) (Figure 1, C and D; Supplementary Table 2, available online), thereby confirming the deleterious effect of this missense variant previously demonstrated in the glycosylase assay (Figure 1). Taken together, these 7 patients with *MBD4* deleterious mutations had a 15-fold increase in number of variants per exome (*MBD4*<sub>def</sub>:  $243 \pm 66.7$  variants vs *MBD4*<sub>pro</sub>:  $16 \pm 4.0$ , Mann-Whitney  $P = 8.72 \times 10^{-5}$ ) and a statistically significantly higher CpG>TpG median proportion among SNVs (*MBD4*<sub>def</sub>:  $92.5 \pm 5.7\%$  vs *MBD4*<sub>pro</sub>:  $24.3 \pm 7.5\%$ ,  $P = 9.82 \times 10^{-7}$ ) (18).

In addition, we characterized the available tumor samples from 2 patients harboring missense variants that were not predicted to be deleterious (UM102: c.139G>A [p.Gly47Arg] and UM350: c.1652A>G [p.Asn551Ser]; Figure 1; Supplementary Table 1; Supplementary Figure 2, available online). The low TMB (33 and 40 variants per exome, respectively) and absence of CpG>TpG signature confirmed their neutral effect (Supplementary Table 2; Supplementary Figure 3, available online).

### Incidence of Germline *MBD4* Deleterious Mutations in UM

Taken together, we thus identified 8 LoF germline variants in *MBD4* among the 1093 consecutive UM case patients, including p. Arg468Trp with deleterious effect on *MBD4* glycosylase activity (Table 1). These account for a statistically significant 9.15-fold increase in deleterious variant frequency compared with the general population, even when restricting ourselves to truncating and splicing *MBD4* LoFs as defined by GnomAD (7 LoFs of 2186 observed alleles in UM, representing a variant allele frequency [VAF] of 0.0032 vs 88 LoFs out of a median of 251 450 alleles in the GnomAD v2.1 general population; VAF =  $3.50 \times 10^{-4}$ ; Fisher exact test  $P = 2.00 \times 10^{-5}$ ). To circumvent the imbalanced dataset, a “matching” subsampling approach was used, giving a similar  $P$  value ( $1.60 \times 10^{-5}$ ). Therefore, we demonstrate that the prevalence of *MBD4* germline deleterious variants in UM is approximately 0.7%, close to that of BAP1 germline mutation in UM (1.6%) (23), and that *MBD4* mutations strongly predispose to UM with an RR of 9.15 (95% CI = 4.24 to 19.73). A comparison between *MBD4* germline LoF frequency in

**Table 1.** MBD4 germline deleterious variants in UM and in other malignancies<sup>a</sup>

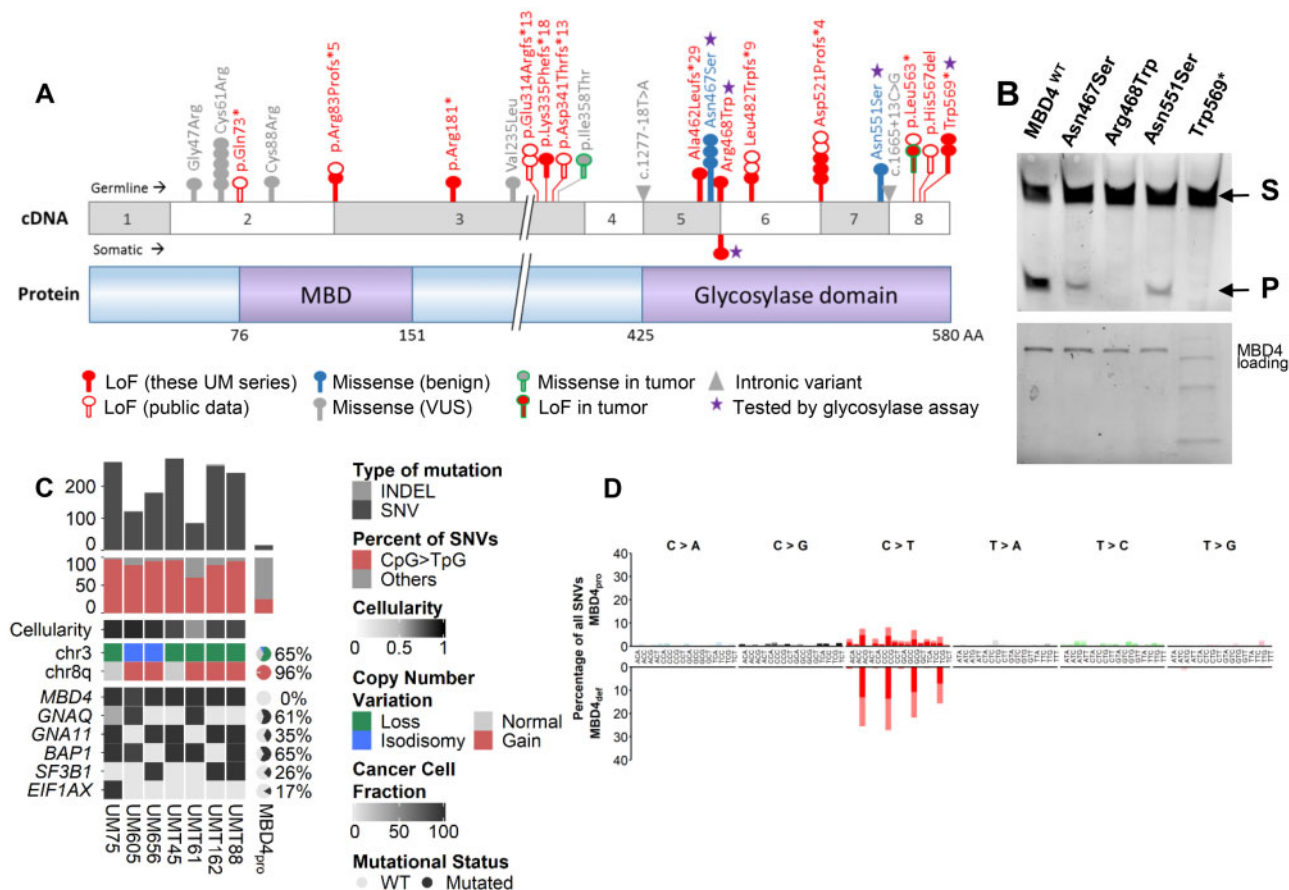
Patient series	Patient	Variant	dbSNP	Mutation type	Glycosylase assay	GnomAD allele frequency (NFE <sup>b</sup> )		
						Allele count	Obs. allele number	Frequency
UM <sup>e</sup> germline consecutive series	UM75	p.Trp569*	rs939751619 <sup>c</sup>	stop_gain	Inactive <sup>d</sup>	2	129 130	1.55 × 10 <sup>-5</sup>
	UM1033							
	UM49	p.Asp521Profs*4	rs778697654 <sup>c</sup>	splice_acceptor	ND	5	113 766	4.39 × 10 <sup>-5</sup>
	UM1088							
	UM656	p.Leu482Trpfs*9	rs769076971 <sup>c</sup>	frameshift_deletion	ND	3	113 752	2.64 × 10 <sup>-5</sup>
	UM293	p.Arg468Trp	rs1380952147	nonsynonymous_SNV	Inactive <sup>d</sup>	0	113 630	0.00
UM M3 tumor series	UM605	p.Ala462Leufs*29	—	frameshift_deletion	ND	—	—	—
	UM436	p.Arg83Profs*5	rs552296498 <sup>c</sup>	splice_donor	ND	3	129 158	2.32 × 10 <sup>-5</sup>
	UMT45	p.Asp521Profs*4	rs778697654 <sup>c</sup>	splice_acceptor	ND	5	113 766	4.39 × 10 <sup>-5</sup>
	UMT61	p.Lys335Phefs*18	rs1443006605	frameshift_deletion	ND	0	113 650	0,00
	UMT162	p.Arg181*	rs1270271346	stop_gain	ND	2	128 972	1.55 × 10 <sup>-5</sup>
	UM (public data)	UM (9)	p.Leu563*	rs200758755	stop_gain	ND	8	113 702
Other malignancies	TCGA_UVM_1 (8)	p.Asp521Profs*4	rs778697654 <sup>c</sup>	splice_acceptor	ND	5	113 766	4.39 × 10 <sup>-5</sup>
	UM <sub>phs001421.v1.p1</sub> (16)	p.Leu482Trpfs*9	rs769076971 <sup>c</sup>	frameshift_deletion	ND	5	113 752	4.40 × 10 <sup>-5</sup>
	UVM_IC (8)							
	UM <sub>phs000823.v1.p1</sub>	p.Asp341Thrfs*13	—	frameshift_deletion	ND	—	—	—
	AML <sub>EMC-AML-1</sub> (13)	p.His567del	rs775848563	inframe_deletion	ND	—	—	—
	AML <sub>WEHI-AML-1/2</sub> (13)	p.Asp521Profs*4	rs778697654 <sup>c</sup>	splice_acceptor	ND	5	113 766	4.39 × 10 <sup>-5</sup>
	AML <sub>WEHI-AML-1/2</sub> (13)	p.Glu314Argfs*13	rs558765093 <sup>c</sup>	frameshift_insertion	ND	—	—	—
	Spiradenocarcinoma (19)							
	TCGA_GBM_4 (8)	p.Arg83Profs*5	rs552296498 <sup>c</sup>	splice_donor	ND	3	129 158	2.32 × 10 <sup>-5</sup>
	Colorectal polyposis (20)	p.Gln73*	rs148098584	stop_gain	ND	0	113 750	0.00
	Pilocytic astrocytoma (21)	NA <sup>g</sup>	NA	NA	ND	NA	NA	NA
	Gastric adenocarcinoma (21)	NA	NA	NA	ND	NA	NA	NA
Pancreatic adenoK (21)	NA	NA	NA	ND	NA	NA	NA	
Pancreatic endocrine tumor (21)	NA	NA	NA	ND	NA	NA	NA	

<sup>a</sup>adenoK = adenocarcinoma; AML = acute myeloid leukemia; GBM = glioblastoma; M3 = monosomy 3; NA = not available; ND = not determined; NFE = non-Finnish European; UM or UVM = uveal melanoma; — = no value given because of the absence of the variant in dbSNP and/or in the GnomAD NFE population.

<sup>b</sup>NFE population of the Genome Aggregation Database (GnomAD v2.1.1).

<sup>c</sup>Variant found in more than 1 nonrelated patient.

<sup>d</sup>Inactive: absence of glycosylase activity of the recombinant protein carrying the variant.



**Figure 1.** Functional consequences and phenotype associated with germline and somatic MBD4 deleterious variants. **A)** Schematic representation of MBD4 cDNA (top) and protein (bottom) sequences. Functional methyl-binding domain (MBD) and glycosylase domain are indicated. The position of all MBD4 variants identified in the 2 uveal melanoma (UM) series (consecutive germline UM series and tumor monosomy 3 [M3] series) is highlighted, with germline and somatic variants above and below the cDNA sequence, respectively, and the 2 variants from the tumor M3 cohort with unknown somatic or germline origin circled in green. These MBD4 variants include loss-of-function (LoF, in red), missense (either benign, in blue-filled circles, or of unknown biological significance [VUS] in gray-filled circles) and intronic (gray triangles) variants. Each circle represents 1 patient harboring the variant. Other MBD4 germline deleterious variants mined on public data are also shown (empty red circles). **B) Top:** Glycosylase activity assay of recombinant wild-type MBD4 (MBD4<sup>WT</sup>) and mutant proteins resulting from missense variants and 1 stop gain variant (purple star in 1A) residing in the MBD4 glycosylase domain. Substrate = S; cleaved product = P. **Bottom:** loading blot for MBD4 wild-type and mutant recombinant proteins corresponding to the glycosylase assay. **C)** Tumor characteristics of MBD4-deficient (MBD4<sub>def</sub>) patients compared with that of MBD4-proficient UM patients (MBD4<sub>pro</sub>) (18). MBD4<sub>def</sub> patients include UM75, UM605, and UM656 from the consecutive germline series and UMT45, UMT61, UMT162, and UMT88 from the M3 UM tumor series. All patients harbor germline MBD4 variants, except for UMT88 with a somatic MBD4 variant. **Top:** tumor mutation burden estimated by number of variants (single nucleotide variants [SNVs] in dark gray, and insertions-deletions [INDELs] in light gray) in the exome; **middle:** proportion of CpG>TpG transitions (red) relative to all SNVs (gray); **bottom:** copy number alterations in chromosomes 3 and 8q, and mutational status of MBD4, GNAQ, GNA11, BAP1, SF3B1, and EIF1AX, represented as percentage for the MBD4<sub>pro</sub> series (18). The clonality or subclonality of these key mutational events is indicated by their cancer cell fraction in black-gray gradation, taking into account the variant allele frequency (VAF), copy number change, and cellularity. A plot of the VAF distribution of all variants in the 7 exomes is available in Supplementary Figure 4 (available online). For each exome in the MBD4<sub>def</sub> group, tumor cellularity is indicated by black-gray shading (and quantified in Supplementary Table 2, available online). **D)** Mutational patterns of the MBD4<sub>def</sub> (top) and MBD4<sub>pro</sub> (bottom) groups based on the relative proportion (y-axis) of each of the 96 types of trinucleotide substitution (x-axis). Dark or bright colors correspond to sense or antisense strands. Individual mutational pattern for all tumor exomes assessed are available in Supplementary Figure 3 (available online).

this UM consecutive series and in different subsets of the GnomAD population (in the general and European populations) is further presented in Table 2.

Within the UM tumor cohort with M3, we identified a total of 5 MBD4 LoF variants, including at least 3 of germline origin and 1 somatic, out of 192 UM patients. These 3 germline LoF variants by themselves account for a VAF of 0.016, more than twice that found in the germline consecutive UM cohort. This was expected given the recurrence of approximately 50% of chromosome 3 loss event among all UM patients (24). This finding confirms that MBD4 deficiency in UM is mainly associated with M3 and that MBD4 germline mutations specifically predispose to hypermutated high-risk M3 UMs.

## Defining the MBD4 Predisposition Syndrome

To further characterize this new cancer predisposition, we investigated the medical records of MBD4 mutation carriers. In contrast with the high RR of 9.15 (and therefore an approximately 9-fold higher risk of developing a UM) conferred by MBD4 LoF germline mutations, none of these individuals had familial or bilateral UM. With a lifetime risk of UM estimated at  $7.69 \times 10^{-5}$  in the general population (25), an RR of 9 would result in a lifetime risk of UM of  $6.92 \times 10^{-4}$ . Such incidence is still too low to observe familial aggregation, which is consistent with our finding. Assuming that all MBD4<sub>def</sub> UMs are associated with M3, we compared their medical records with patients from this cohort with available tumor



**Table 2.** Frequency of *MBD4* germline deleterious variants in the UM series compared with various populations of the GnomAD database<sup>a</sup>

Study population		No. of LoF variants	Allele count <sup>c</sup>	Frequency	RR <sup>d</sup> (95% CI <sup>e</sup> )	Fisher test (P value)
UM consecutive series		7 <sup>f</sup>	2186	0.00320	—	—
GnomAD v2.1.1	NFE <sup>b</sup>	47	113 736	0.00041	7.75 (3.51 to 17.12)	$6.86 \times 10^{-5}$
	Population <sup>b</sup>	88	251 450	0.00035	9.15 (4.24 to 19.73)	$2.00 \times 10^{-5}$
GnomAD v2.1.1 (controls only)	NFE	13	42 768	0.00030	10.53 (4.20 to 26.38)	$2.82 \times 10^{-5}$
	General population	33	109 404	0.00030	10.62 (4.70 to 23.97)	$1.16 \times 10^{-5}$
GnomAD v2.1.1 (noncancer only)	NFE	41	102 730	0.00040	8.02 (3.60 to 17.86)	$5.89 \times 10^{-5}$
	General population	82	236 912	0.00035	9.25 (4.28 to 19.99)	$1.90 \times 10^{-5}$
GnomAD v3	NFE	20	64 571	0.00031	10.34 (4.38 to 24.42)	$2.00 \times 10^{-5}$
	General population	39	143 286	0.00027	11.76 (5.27 to 26.27)	$5.50 \times 10^{-5}$

<sup>a</sup>CI = confidence interval; LoF = loss-of-function (deleterious) variants; NFE = non-Finnish European; RR = relative risk; UM = uveal melanoma; — = no value given here because the relative risk, confidence interval, and statistic tests are presented between the UM consecutive series and each GnomAD subpopulation in the rows below.

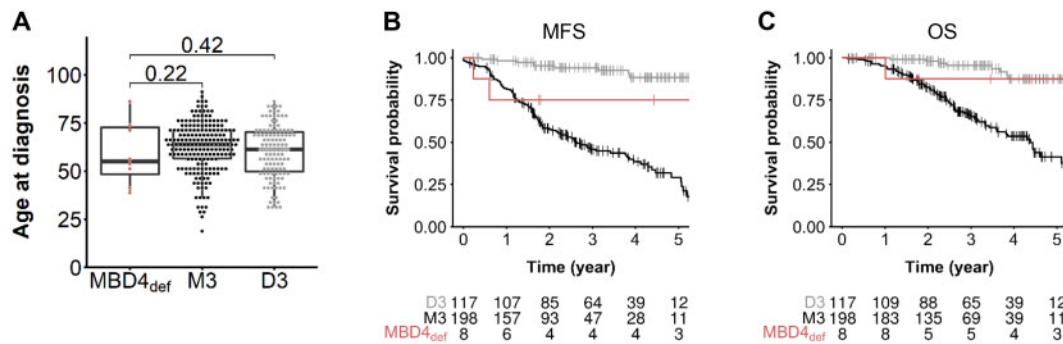
<sup>b</sup>NFE population subset of the Genome Aggregation Database (GnomAD v2.1.1).

<sup>c</sup>For all GnomAD populations described, refers to the median number of allele count.

<sup>d</sup>RR here is calculated by dividing the LoF frequency in the UM consecutive series by the LoF frequency in the corresponding GnomAD population subset.

<sup>e</sup>Confidence interval of the relative risk is calculated as previously described (15).

<sup>f</sup>Seven LoF variants correspond to the 8 deleterious *MBD4* variants identified in this study, with removal of the missense deleterious variant p. Arg468Trp so as to restrict the analysis to LoF variants as defined by GnomAD for accurate comparison.



**Figure 2.** Uveal melanoma (UM) clinical characteristics in an *MBD4*-deficient (*MBD4<sub>def</sub>*) context. **A**) Age of UM onset of *MBD4<sub>def</sub>* patients ( $n = 8$ ) in the germline consecutive UM series compared with disomy 3 (D3,  $n = 117$ ) and monosomy 3 (M3,  $n = 198$ ) *MBD4*-proficient (*MBD4<sub>pro</sub>*) UMs. Wilcoxon test, 1-sided (testing early UM onset in *MBD4<sub>def</sub>* patients): *MBD4<sub>def</sub>* vs M3:  $P = .22$ , *MBD4<sub>def</sub>* vs D3:  $P = .42$ ; no age difference found between D3 and M3 groups, Wilcoxon test, 2-sided  $P = .087$ ; – not shown). **B** and **C**) Metastasis-free survival (MFS, **B**) and overall survival (OS, **C**) of *MBD4<sub>def</sub>* UM patients ( $n = 8$ ) and *MBD4<sub>pro</sub>* UM patients with M3 or D3. Time zero refers to time at primary UM diagnosis. MFS was defined as the interval between the date of primary UM diagnosis and the date of distant metastasis (first imaging) or death from any cause. The number of patients in each group at each time point (year) is indicated. Survival distributions were estimated by the Kaplan-Meier method and compared using the log-rank test: log-rank test, 2-sided, M3 vs D3:  $P = 1.98 \times 10^{-9}$  (OS),  $P = 1.11 \times 10^{-16}$  (MFS); M3 vs *MBD4<sub>def</sub>*:  $P = .11$  (OS),  $P = .06$  (MFS); D3 vs *MBD4<sub>def</sub>*:  $P = .62$  (OS),  $P = .10$  (MFS).

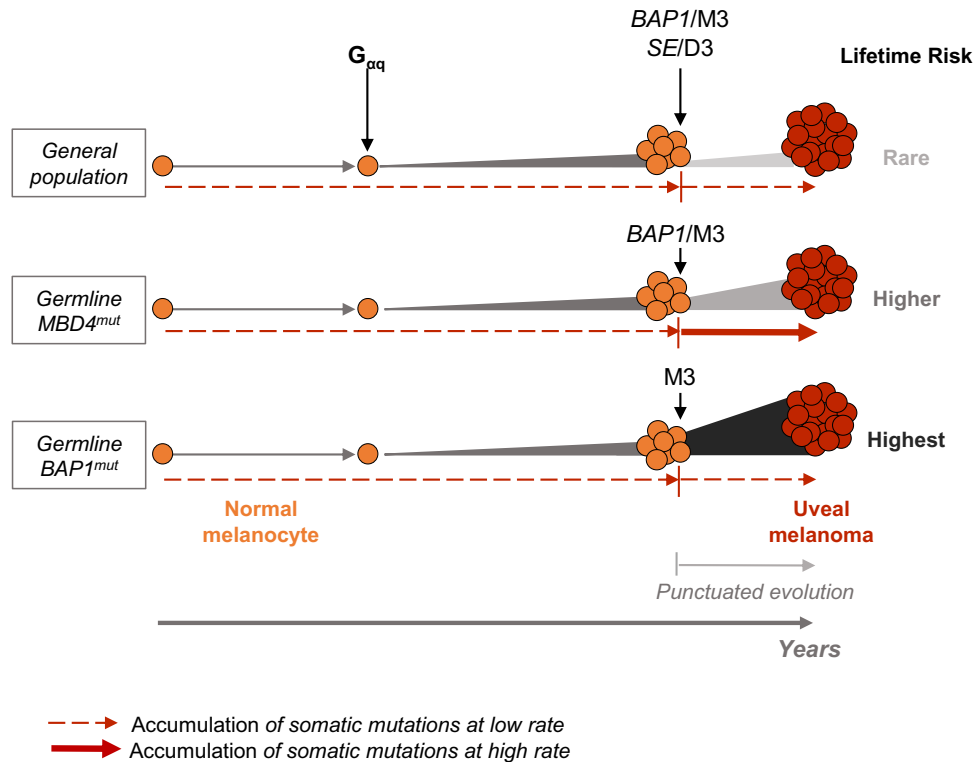
genomic status, that is, 198 *MBD4* wild-type M3 and 117 *MBD4* wild-type D3 UMs. Surprisingly, no early-onset UM was found in *MBD4* carriers compared with noncarriers regardless of their chromosome 3 status (*MBD4<sub>def</sub>*: median age and Q1-Q3 quartile interval = 55.5, 95% CI = 48.4 to 72.8,  $N = 8$ ; D3: 61.3, 95% CI = 49.8 to 70.3,  $N = 117$ ; M3: 63.4, 95% CI = 56.6 to 71.4,  $N = 198$ ; Wilcoxon test, 1-sided: *MBD4<sub>def</sub>* vs M3:  $P = .22$ , *MBD4<sub>def</sub>* vs D3:  $P = .42$ ; no age difference found between D3 and M3 groups, Wilcoxon test, 2-sided  $P = .087$ ; Figure 2A). Although the size of the *MBD4<sub>def</sub>* series prevents any definitive conclusion, we observed no difference in metastatic-free survival or overall survival between *MBD4<sub>def</sub>* and M3 UM patients in contrast with the better outcome in D3 compared with M3 UMs (Figure 2, B and C). Only 1 *MBD4* carrier (UM49) had another cancer, a thyroid papillary carcinoma unrelated to *MBD4* (low TMB without CpG>TpG signature; data not shown).

## Discussion

This study demonstrates that *MBD4* is a predisposing gene for UM, conferring an RR of 9.15 for this dismal disease. We further

demonstrated that *MBD4* deficiency specifically predisposes to high-risk M3 UM. One surprising observation for this new cancer-predisposing condition is the absence of early-onset UM. Interestingly, the same is observed in germline *BAP1*-mutant carriers, even with the high penetrance in that context (6,23). A potential explanation for this paradox is that neither *MBD4* nor *BAP1* predisposing genes can act before a first step in the malignant transformation. This first step, presumably the G $\alpha$ q-initiating event consisting of mutually exclusive activating mutations in *GNAQ*, *GNA11*, *PLCB4*, or *CYSLTR2* (26–29), would be the main determinant of age of onset (Figure 3). The second step in the malignant transformation, composed of mutations in *BAP1*, *SF3B1*, or *E1F1AX* (“BSE” events), leads to a punctuated evolution of UM (16), which would marginally influence age of onset (Figure 3). Whether *MBD4* deficiency favors malignant transformation by increasing driver mutations by modifying the methylation landscape or a distinct mechanism has yet to be determined.

Importantly, germline *MBD4* mutations were recently reported in other types of malignancy, including a polyposis-associated colorectal adenocarcinoma (20), a spiradenocarcinoma



**Figure 3.** Working model for uveal melanoma (UM) malignant transformation process throughout time in different genetic backgrounds. Germline  $MBD4^{mut}/BAP1^{mut}$  population with  $MBD4/BAP1$  germline mutation. Following the  $G_{\alpha q}$  activating mutation, secondary mutational events in each genetic background are indicated, along with their association with either disomy 3 (D3) or monosomy 3 (M3). SE = SF3B1 or EIF1AX mutation. Relative lifetime risk of UM is represented by the expansion size and color from normal melanocytes to UM. Dashed and full red arrows indicate the rate of accumulation of somatic mutations throughout time (low and high, respectively).

(19), a glioblastoma (8), a pilocytic astrocytoma, a gastric adenocarcinoma, a pancreatic adenocarcinoma, and a pancreatic endocrine tumor (21) (Table 1). Furthermore, although the above case patients were all heterozygous in the germline, biallelic germline deleterious  $MBD4$  mutations were reported in 3 individuals who developed acute myeloid leukemias, 2 of which had additional colonic polyposis (13). It is therefore likely that the tumor spectrum associated with  $MBD4$  germline mutations will expand when this gene becomes more systematically explored in clinical diagnosis. It is already clear that this spectrum mostly includes relatively rare tumors and some biological tumor features may underlie their association with  $MBD4$ . To be noticed, both leukemias and UMs associated with  $MBD4$  inactivation share a consistent inactivation of the  $BAP1$ - $ASXL$  complex (8,13).

It should be noticed that we found no other  $MBD4$ -related tumors in our UM series. However, the follow-up and cohort size of this prospective series are limited, and future studies will better characterize the medical history of  $MBD4$  carriers. Larger cohorts will also more precisely define  $MBD4$  mutation frequency in UM patients and the RR conferred by these mutations. Another limitation to the study is the bias for a European population in our cohort, which reflects the higher incidence of the disease in this population (30).

Interestingly, 5 recurrent  $MBD4$  germline deleterious mutations were identified when taking together the LoF variants from our UM cohort and those found in public databases and reports of other cancer types: c.1706G>A [p.Trp569\*] (2 patients), c.1562G>T [p.Asp521Profs\*4] (4 patients), c.1443delT [p.Leu482Trpfs\*9] (3 patients), c.335+1G>A [p.Arg83Profs\*5] (2 patients), and c.939insA [p.Glu314Argfs\*13] (2 patients) (Table 1),

suggesting founder mutations. Furthermore, the observation of different tumor types associated with the same  $MBD4$  germline mutation suggests a more global role of  $MBD4$  in cancer predisposition. The peculiar UM proneness in  $MBD4$ -mutant carriers (13 out of 23 carriers; Table 1) remains unexplained, but the fact that the frequent M3 in UM inactivates wild-type copies of both  $BAP1$  and  $MBD4$  suppressor genes may at least in part explain the frequent inactivation of  $MBD4$  in UM.

In summary, we described here a novel autosomal-dominant syndrome that is caused by germline mutations of  $MBD4$ , characterized by a high RR of developing hypermutated UM and possibly other malignancies. Tumors arising in such a context are associated with a CpG>TpG mutator phenotype and have clinical relevance because they may respond to immune-checkpoint inhibitors.

## Funding

Supported by funding from the European Commission under the Horizon 2020 program and innovation program under the Marie Skłodowska-Curie grant agreement No 666003 (A-CD), the Horizon 2020 program UM Cure (LM; Project number: 667787), the INCa/ITMO/AVIESAN PhD fellowship program "Formation à la recherche translationnelle" (MR), the Ligue Nationale Contre le Cancer (AE), the Institut National de la Santé et de la Recherche Médicale (INSERM), the Institut Curie, the Ligue Nationale Contre le Cancer (Labellisation), the Programme de Recherche Translationnelle en Cancérologie (PRT-K19-51) INCa-DGOS, and the Site de Recherche Intégrée sur le

Cancer (SiRIC) de l'Institut Curie. The Institut Curie ICGex NGS platform is funded by the EQUIPEX "investissements d'avenir" program (ANR-10-EQPX- 03) and ANR10-INBS-09-08 from the Agence Nationale de la Recherche.

## Notes

**Role of the funder:** The funders had no role in the design of the study; the collection, analysis, and interpretation of the data; the writing of the manuscript; and the decision to submit the manuscript for publication.

**Disclosures:** The authors have no conflict of interest to declare except for the following: M.R. received a grant support from Bristol-Myers Squibb and Merck.

**Author contributions:** A.-C.D. and M.R. conceived the study, performed experiments, interpreted the data and wrote the manuscript. A.E. and A.H. performed bioinformatics analyses. L.M., S.A., A.E.-M. and G.P. interpreted the data and provided critical advice. O.M., A.M., S.G. and N.C. provided patients specimens and critical advice. S.D., D.L. and G.P. performed biological analyses. C.C. provided clinical data and critical advice. M.-H.S. conceived and guided the study, interpreted the data and wrote the manuscript. All authors reviewed and approved the final manuscript.

**Acknowledgment:** The authors thank Joshua J. Waterfall for his insightful comments.

## References

- Aronow ME, Topham AK, Singh AD. Uveal melanoma: 5-year update on incidence, treatment, and survival (SEER 1973-2013). *Ocul Oncol Pathol*. 2018;4(3):145-151.
- Khoja L, Atenafu EG, Suci S, et al. Meta-analysis in metastatic uveal melanoma to determine progression free and overall survival benchmarks: an international rare cancers initiative (IRCI) ocular melanoma study. *Ann Oncol*. 2019;30(8):1370-1380.
- Carvajal RD, Schwartz GK, Tezel T, et al. Metastatic disease from uveal melanoma: treatment options and future prospects. *Br J Ophthalmol*. 2017;101(1):38-44.
- Harbour JW, Onken MD, Roberson EDO, et al. Frequent mutation of BAP1 in metastasizing uveal melanomas. *Science*. 2010;330(6009):1410-1413.
- Wiesner T, Murali R, Fried I, et al. A distinct subset of atypical Spitz tumors is characterized by BRAF mutation and loss of BAP1 expression. *Am J Surg Pathol*. 2012;36(6):818-830.
- Walpole S, Pritchard AL, Cebulla CM, et al. Comprehensive study of the clinical phenotype of germline BAP1 variant-carrying families worldwide. *J Natl Cancer Inst*. 2018;110(12):1328-1341.
- Furney SJ, Pedersen M, Gentien D, et al. SF3B1 mutations are associated with alternative splicing in uveal melanoma. *Cancer Discov*. 2013;3(10):1122-1129.
- Rodrigues M, Mobuchon L, Houy A, et al. Outlier response to anti-PD1 in uveal melanoma reveals germline MBD4 mutations in hypermutated tumors. *Nat Commun*. 2018;9(1):1866.
- Johansson PA, Stark A, Palmer JM, et al. Prolonged stable disease in a uveal melanoma patient with germline MBD4 nonsense mutation treated with pembrolizumab and ipilimumab. *Immunogenetics*. 2019;71(5-6):433-436.
- Hendrich B, Hardeland U, Ng HH, et al. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. *Nature*. 1999;401(6750):301-304.
- Yoon JH, Iwai S, O'Connor TR, et al. Human thymine DNA glycosylase (TDG) and methyl-CpG-binding protein 4 (MBD4) excise thymine glycol (Tg) from a Tg: G mismatch. *Nucleic Acids Res*. 2003;31(18):5399-5404.
- Alexandrov LB, Kim J, Haradhvala NJ, et al. PCAWG Mutational Signatures Working Group. The repertoire of mutational signatures in human cancer. *Nature*. 2020;578(7793):94-101.
- Sanders MA, Chew E, Flensburg C, et al. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. *Blood*. 2018;132(14):1526-1534.
- Hashimoto H, Liu Y, Upadhyay AK, et al. Recognition and potential mechanisms for replication and erasure of cytosine hydroxymethylation. *Nucleic Acids Res*. 2012;40(11):4841-4849.
- Rothman KJ. *Epidemiology, an Introduction*. 2nd ed. New York, NY: Oxford University Press; 2002.
- Field MG, Durante MA, Anbunathan H, et al. Punctuated evolution of canonical genomic aberrations in uveal melanoma. *Nat Commun*. 2018;9(1):116.
- Robertson AG, Shih J, Yau C, et al. Integrative analysis identifies four molecular and clinical subsets in uveal melanoma. *Cancer Cell*. 2017;32(2):204-220.e15.
- Rodrigues M, Mobuchon L, Houy A, et al. Evolutionary routes in metastatic uveal melanomas depend on MBD4 alterations. *Clin Cancer Res*. 2019;25(18):5513-5524.
- Davies HR, Hodgson K, Schwalbe E, et al. Epigenetic dysregulation underpins tumorigenesis in a cutaneous tumor syndrome. *bioRxiv* 2019; doi: 10.1101/687459.
- Tanakaya K, Kumamoto K, Tada Y, et al. A germline MBD4 mutation was identified in a patient with colorectal oligopolyposis and early-onset cancer: a case report. *Oncol Rep*. 2019;42(3):1133-1140.
- Waszak SM, Tiao G, Zhu B, et al. Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. *bioRxiv* 2017; doi: 10.1101/208330.
- Morera S, Grin I, Vigouroux A, et al. Biochemical and structural characterization of the glycosylase domain of MBD4 bound to thymine and 5-hydroxymethyluracil-containing DNA. *Nucleic Acids Res*. 2012;40(19):9917-9926.
- Gupta MP, Lane AM, DeAngelis MM, et al. Clinical characteristics of uveal melanoma in patients with germline BAP1 mutations. *JAMA Ophthalmol*. 2015;133(8):881-887.
- Prescher G, Bornfeld N, Hirche H, et al. Prognostic implications of monosomy 3 in uveal melanoma. *Lancet*. 1996;347(9010):1222-1225.
- Singh AD, De Potter P, Fijal BA, et al. Lifetime prevalence of uveal melanoma in white patients with ocular (dermal) melanocytosis. *Ophthalmology*. 1998;105(1):195-198.
- Johansson P, Aoude LG, Wadt K, et al. Deep sequencing of uveal melanoma identifies a recurrent mutation in PLCB4. *Oncotarget*. 2016;7(4):4624-4631.
- Van Raamsdonk CD, Bezrookove V, Green G, et al. Frequent somatic mutations of GNAQ in uveal melanoma and blue naevi. *Nature*. 2009;457(7229):599-602.
- Van Raamsdonk CD, Griewank KG, Crosby MB, et al. Mutations in GNA11 in uveal melanoma. *N Engl J Med*. 2010;363(23):2191-2199.
- Moore AR, Ceraudo E, Sher JJ, et al. Recurrent activating mutations of G-protein-coupled receptor CYSLTR2 in uveal melanoma. *Nat Genet*. 2016;48(6):675-680.
- Mobuchon L, Battistella A, Bardel C, et al. A GWAS in uveal melanoma identifies risk polymorphisms in the CLPTM1L locus. *NPJ Genom Med*. 2017;2(1):1-7.

## Conclusion: “Germline MBD4 Mutations and Predisposition to Uveal Melanoma”

In this study, our team demonstrated that *MBD4* is a cancer predisposition gene for UM, with a Relative-Risk of 9.15 (95% Confidence Interval (CI) = 4.24 to 19.73). It is the only cancer predisposing gene except *BAP1* that was identified in UM until now<sup>317</sup>. *MBD4* deleterious germline mutations are present in 0.7% of our cohort, at a slight less frequent rate than for *BAP1* germline deleterious mutation reported in ~1.5% in patients with UM<sup>318</sup>. Interestingly, no early age of onset was observed for patient with *MBD4* germline mutation with only a vague tendency. Moreover, no clear distinction could be made between metastatic-free survival or overall survival between the *MBD4*def series and M3 UM patients, but also for *MBD4*def and D3 UM patients. The small subset of patient carrying *MBD4* germline mutations however prevents any definitive conclusion. We also validated the specific hypermutator CpG > TpG phenotype that was associated with *MBD4* inactivation<sup>314</sup>.

*MBD4* inactivation is always associated with either monosomy 3 or isodisomy 3 inactivating the wild-type allele, concordantly with the Knudson’s two-hit model. The place of *MBD4* in tumorigenesis and tumor development is however still unclear. *MBD4* deficiency could promote tumorigenesis through the important number of mutations that could affect unspecific or specific cancer driver genes but could also promote malignant transformation through another pathway. The observation of an increased number of CpG > TpG without any tumor development in some *MBD4*-inactivated mice indicate that it may not be enough for tumorigenesis<sup>319</sup>.

The predisposing nature of *MBD4* for UM is of direct clinical importance and *MBD4* should be included in cancer gene panel. No familial UM in our cohort was found for the patient harboring a *MBD4* germline mutation. This must be mitigated again by the size of our cohort and the rarity of UM. We believe however that the identification of *MBD4* as a cancer predisposition gene in UM bears enough rational to investigate the relatives of someone carrying a *MBD4* deleterious germline mutation. Moreover, *MBD4*<sup>-/-</sup> tumors are distinct among UM with a specific genomic phenotype and possible treatments might target this specificity<sup>314</sup>.

Additionally, we reviewed in this paper all the *MBD4* germline deleterious mutations that we could find in the literature. Those germline mutations were present in different UM cases but also in other malignancies<sup>320-323</sup>. The role of *MBD4* as a potential tumor suppressor and cancer predisposing gene could therefore expand beyond UM.



# CONCLUSIONS AND PERSPECTIVES

In this thesis I investigated specific cancer genomic instability and DNA repair deficiency signatures in different types of cancers. Each part of this work has potential direct consequence for patient care and represents the part of efforts in developing the background for future personalized medicine.

The main part of my work is the *shallowHRD* tool that can accurately predict HRD from shallow WGS profile in ovarian and breast tumors. This approach can be applied to FFPE samples, is cheap and simple, with easily storable outputs and is already in use in Institut Curie in routine for ovarian carcinomas and some other cancer types. To the moment almost 850 cases were processed using *shallowHRD* and the results are both encouraging and motivating to further upgrade of the approach.

The performance of the method is comparable to most state-of-the-art methods based on WES or SNP-arrays, such as signature 3<sup>217,220</sup>, WES HRD score<sup>213</sup>, Loss of Heterozygosity HRD score<sup>208</sup>, telomeric Allelic Imbalance<sup>209</sup> and Large-scale State Transition<sup>142,206</sup>. The methods that clearly outperform *shallowHRD* are the classifiers that incorporate all HRD-specific alterations extracted from high-depth WGS, namely HRDetect<sup>224</sup> and CHORD<sup>169</sup>. The latter methods capture exhaustively the tumors with HRD and only their cost and complexity preclude their installation in clinics.

After *shallowHRD* was introduced to clinical research in Institut Curie, most of the time sWGS was accompanied by deep coverage sequencing for the cancer gene panel called DRAGON that incorporates more than 500 genes (including HR genes). Besides mutations in targeted genes, DRAGON sequencing gives variant allele frequency. This helps to say if a genomic flat profile observed in a tumor can be attributed to a low cellularity, which cannot be done for now with only sWGS.

*shallowHRD* is now also in use for initial annotation of VUS in major HR predisposition genes. Several presumably deleterious VUS for *BRCA1*, *RAD51B* and *RAD51C* (LGAs  $\geq 20$ ) have been already identified. Additionally, methylation of *BRCA1* promoter were detected retrospectively in several cancer samples after *shallowHRD* diagnosis. The preliminary results of comparison *shallowHRD* and the clinically approved Myriad myChoice® CDx HRD test shows 92.5% correspondence when using a stringent LGA cut-off of 20. *shallowHRD* is in use for PDX annotation and provide HR status for large retrospective cohorts in clinical research.

In perspective, *shallowHRD* have some room to improve performance to fit clinical precision. The accumulation of hundreds of in-house cases of both retrospective and diagnosis samples expanded the number of cases giving the possibility to build upper-level classification. This mainly concerns borderline cases, the cases where LGA equals or close to the cut-off for HRD

call. In these cases, the error rate (the proportion of cases with no obvious mutation in HR genes) is particularly high. First, the borderline cases need to be thoroughly annotated: cases should be investigated for potential HR gene inactivation and actual genotoxic treatment response. Second, CNA profile classification need to be improved. Several approaches might be used here, such as introducing additional parameters that characterize tumor sWGS profile in terms of CNA dynamics, detailing the source HRD or refining the CNA profile annotation by inferring ploidy or allelic imbalance states. All these improvements could only be tested now, after some 2 years of extensive use, optimization and building adequate data collection.

Another problem that can confuse *shallowHRD* predictions are subclones. HRD and nonHRD breast and ovarian tumors are often characterized by the presence of several clones with slightly different CNA profiles, which introduce extra-breakpoints and might affect HRD diagnostics. Introducing an additional filter for potential subclonal alterations could refine HRD score and make the predictions and profile more precise and reliable in those cases.

Finally, a more thorough investigation toward the quality of the DNA preparation and sequencing would be highly valuable to keep high quality predictions. The sequencing of clinical FFPE samples were mostly well interpretable except for one FFPE cohort, which displayed particularly poor quality with more than 1/3 uninterpretable cases in 80 sWGS (unpublished data). Those FFPE were markedly older, and it was probably the period between tumor sampling and actual DNA extraction that decreased the quality of the genomic profiles. We would like to retrieve more biological information on these samples to describe the material quality conditions where sWSG and *shallowHRD* can be applied.

The second part of my work consisted in the investigation of *CDK12* as a cancer predisposition gene in EOC with a massive parallel pool sequencing approach in a consecutive cohort of 416 unrelated patients with a history of EOC and no germline mutation of *BRCA1/BRCA2*. No evidence towards *CDK12* as a potential role in cancer predisposition gene for EOC was found. Our study has however two weaknesses that prevent us from drawing definitive conclusion on the predisposing nature of *CDK12*: (i) the cohort is not large enough to completely remove *CDK12* as a potential EOC predisposition gene; (ii) the non-synonymous variants that we found in ten patient's germline DNA could not be fully investigated. The investigation of the remaining unavailable tumors would have helped to consolidate our conclusion. Nonetheless, the lack of *CDK12* deleterious mutations in the 511 ovarian cases in TCGA argues in the direction of our conclusion.

Interestingly, Brovkina et al. found 8 *CDK12* germline splice variant mutations c.1047-2A>G in a cohort of 106 breast cancers with Tatar ancestry and 1 out of 238 healthy controls with mixed origin<sup>324</sup>. The study concluded that, in the Tatar ethnicity, the *CDK12* c.1047-2A>G splice

variant strongly associates with hereditary breast cancers. This result was however mitigated by a follow up study which on the one hand found this variant in multiple healthy controls and on the other hand found an alternative splice acceptor site so that the splice variant shortens the transcript by one codon only<sup>325</sup>. This needs to be investigated further, with functional assay of CDK12 for this variant for instance.

Nonetheless, the current literature does not show any strong evidence of *CDK12* being a cancer predisposition gene for any cancer type. A study of 360 metastatic castration-resistant prostate cancer (mCRPC) reported *CDK12* bi-allelic inactivation in 7% of the cases, without any evidence of germline aberrations in the entire cohort<sup>298</sup>. *CDK12* high prevalence in several cancers and its investigation in multiple large cohorts with the absence of a reported clear deleterious germline variant comfort our results<sup>54</sup>.

*CDK12* has a probability of 1 of being intolerant to loss-of-function (LoF) based on an expectation-maximization algorithm developed by Lek et al in 2016<sup>326</sup>. The six LoF variants found in the GnomAD database is largely below the expected number with an observed / expected (oe) score of 0.05 (gnomad.broadinstitute.org). This suggests that LoF variants in *CDK12* are counter-selected in human populations, the reason for this remains to be explained. *CDK12* inactivation is embryonic lethal but heterozygous mice models are viable and born at the correct frequency<sup>260</sup>. However, no long-term follow-up of those mice for their fitness and fertility has been done. It would be instrumental to study *CDK12* intolerance to LoF variants.

The third part of my work was to analyze *MBD4* mutations from the sequencing of pooled germline (n=1093) and tumor (n=192) DNA of UM patients. We found that *MBD4* is an UM predisposing gene with moderate penetrance and a 10-fold increased risk present in 0.7% of UM cases. The CpG > TpG hypermutator phenotype that was observed previously in several cases by Rodrigues et al in 2018 was confirmed in the *MBD4*<sup>-/-</sup> tumors of our cohort, which is in line with the identified role of *MBD4* to repair 5mCpG > TpG mutations<sup>314</sup>. Similar observations regarding the hypermutated CpG > TpG phenotype and the existence of *MBD4* deleterious germline mutations and inactivation in the corresponding tumor was made in two out of 103 cases of another UM cohort, with high tumor mutation burden and the prevalence of the SBS1 signature<sup>315</sup>.

Our study has two limitations: first the small number of cases with *MBD4* inactivation in the tumors prevent us from drawing definitive conclusions for a potential early age of onset, metastatic-free survival, or overall survival. Larger cohort will help clarifying this. The second limitation is a bias toward more patients of European population in our cohort because of the geographic localization of our center and the actual higher frequency of UM in this population<sup>327</sup>.

Heterozygous deleterious germline mutations of *MBD4* were not only observed in UM but also in polyposis-associated colorectal adenocarcinoma<sup>322</sup>, spiroadenocarcinoma<sup>321</sup>, glioblastoma<sup>314</sup>, pilocytic astrocytoma, gastric adenocarcinoma, pancreatic adenocarcinoma and endocrine tumor<sup>323</sup> and in Acute Myeloid Leukemias (AML)<sup>320</sup>. The presence of germline deleterious mutation in *MBD4* in those cancer types and the cancer predisposition role of *MBD4* in UM suggest that it may play a similar role in other cancer types. Our team is collaborating with a team of the hospital La Pitié Salpêtrière to investigate *MBD4* predisposition in gliomas with a large cohort of around 2000 germline DNAs.

Our study of *MBD4* mutations both at the germline and the somatic level motivates the integration of this gene in cancer panel genes. This has already been implemented in Institut Curie for UM and allowed to find several new *MBD4*<sup>-/-</sup> patients. This will be helpful for the research perspective, expanding the size of our cohort of rare mutations in rare tumors. This will be also helpful from a medical perspective to have a more complete description of patient and tumor in view of future application of immune checkpoint inhibitors, already shown to be effective against *MBD4*<sup>-/-</sup> tumor<sup>314</sup>. Immunotherapy was pursued in a larger UM cohort with and without *MBD4*<sup>-/-</sup> and showed a significant responsiveness in those inactivated for *MBD4* with 60% responders (3/5) compared to the overall response that is around 4% (manuscript under preparation). Nonetheless, the first patient that presented an exceptional response to immune checkpoint inhibitors in the publication of Rodrigues et al. in 2018 showed late resistance to the treatment. This might be due to the heterogeneity observed in *MBD4*-null tumors and the continuous mutagenesis induced by *MBD4*<sup>314</sup>. Therefore, other treatments in this context should be explored and we are currently screening more drugs and some already showed a great promising effect.

# References

1. J, F. et al. Global Cancer Observatory: Cancer Today. in *International Agency for Research on Cancer* Vol. 68 (2020).
2. Hanahan, D. & Weinberg, R.A. Hallmarks of cancer: The next generation. in *Cell* Vol. 144 (2011).
3. Hakem, R. DNA-damage repair; the good, the bad, and the ugly. in *EMBO Journal* Vol. 27 (2008).
4. Alexandrov, L.B. et al. The repertoire of mutational signatures in human cancer. in *Nature* Vol. 578 (2020).
5. Li, Y. et al. Patterns of somatic structural variation in human cancer genomes. in *Nature* Vol. 578 (2020).
6. Khanna, K.K. & Jackson, S.P. DNA double-strand breaks: Signaling, repair and the cancer connection. in *Nature Genetics* Vol. 27 (2001).
7. Ge, X.Q., Jackson, D.A. & Blow, J.J. Dormant origins licensed by excess Mcm2-7 are required for human cells to survive replicative stress. in *Genes and Development* Vol. 21 (2007).
8. Woodward, A.M. et al. Excess Mcm2-7 license dormant origins of replication that can be used under conditions of replicative stress. in *Journal of Cell Biology* Vol. 173 (2006).
9. Ciccia, A., Constantinou, A. & West, S.C. Identification and Characterization of the Human Mus81-Emel Endonuclease. in *Journal of Biological Chemistry* Vol. 278 (2003).
10. Mehta, A. & Haber, J.E. Sources of DNA double-strand breaks and models of recombinational DNA repair. in *Cold Spring Harbor Perspectives in Biology* Vol. 6 (2014).
11. Tubbs, A. & Nussenzweig, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. in *Cell* Vol. 168 (2017).
12. Cooke, M.S., Evans, M.D., Dizdaroglu, M. & Lunec, J. Oxidative DNA damage: mechanisms, mutation, and disease. in *The FASEB Journal* Vol. 17 (2003).
13. Kasai, H. & Nishimura, S. Hydroxylation of deoxyguanosine at the C-8 position by ascorbic acid and other reducing agents. in *Nucleic Acids Research* Vol. 12 (1984).
14. Dianov, G.L., O'Neill, P. & Goodhead, D.T. Securing genome stability by orchestrating DNA repair: Removal of radiation-induced clustered lesions in DNA. in *BioEssays* Vol. 23 (2001).
15. Lieber, M.R. The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. in *Annual Review of Biochemistry* Vol. 79 (2010).
16. De Magis, A. et al. DNA damage and genome instability by G-quadruplex ligands are mediated by R loops in human cancer cells. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 116 (2019).
17. Helmrich, A., Ballarino, M. & Tora, L. Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes. in *Molecular Cell* Vol. 44 (2011).
18. Durkin, S.G. & Glover, T.W. Chromosome fragile sites. in *Annual Review of Genetics* Vol. 41 (2007).
19. Teng, G. & Schatz, D.G. Regulation and Evolution of the RAG Recombinase. in *Advances in Immunology* Vol. 128 (2015).
20. Joshi, N., Brown, M.S., Bishop, D.K. & Börner, G.V. Gradual Implementation of the Meiotic Recombination Program via Checkpoint Pathways Controlled by Global DSB Levels. in *Molecular Cell* Vol. 57 (2015).
21. Jackson, S.P. Sensing and repairing DNA double-strand breaks. in *Carcinogenesis* Vol. 23 (2002).
22. Spitz, D.R., Azzam, E.I., Li, J.J. & Gius, D. Metabolic oxidation/reduction reactions and cellular responses to ionizing radiation: A unifying concept in stress response biology. in *Cancer and Metastasis Reviews* Vol. 23 (2004).
23. Margison, G.P. & Santibáñez-Koref, M.F. O6-alkylguanine-DNA alkyltransferase: Role in carcinogenesis and chemotherapy. in *BioEssays* Vol. 24 (2002).
24. Deans, A.J. & West, S.C. DNA interstrand crosslink repair and cancer. in *Nature Reviews Cancer* Vol. 11 (2011).
25. Koster, D.A., Palle, K., Bot, E.S.M., Bjornsti, M.A. & Dekker, N.H. Antitumour drugs impede DNA uncoiling by topoisomerase I. in *Nature* Vol. 448 (2007).
26. Petrini, J.H.J. & Stracker, T.H. The cellular response to DNA double-strand breaks: Defining the sensors and mediators. in *Trends in Cell Biology* Vol. 13 (2003).
27. Koike, M. & Koike, A. Accumulation of Ku80 proteins at DNA double-strand breaks in living cells. in *Experimental Cell Research* Vol. 314 1061-1070 (Academic Press, 2008).
28. Woods, D.S., Sears, C.R. & Turchi, J.J. Recognition of DNA termini by the cterminal region of the Ku80 and the dna-dependent protein kinase catalytic subunit. in *PLoS ONE* Vol. 10 (2015).
29. Ciccia, A. & Elledge, S.J. The DNA Damage Response: Making It Safe to Play with Knives. in *Molecular Cell* Vol. 40 (2010).
30. Falck, J., Coates, J. & Jackson, S.P. Conserved modes of recruitment of ATM, ATR and DNA-PKcs to sites of DNA damage. in *Nature* Vol. 434 (2005).
31. Lavin, M.F. ATM and the Mre11 complex combine to recognize and signal DNA double-strand breaks. in *Oncogene* Vol. 26 (2007).
32. Reynolds, P. et al. The dynamics of Ku70/80 and DNA-PKcs at DSBs induced by ionizing radiation is dependent on the complexity of damage. in *Nucleic Acids Research* Vol. 40 (2012).
33. Dart, D.A., Adams, K.E., Akerman, I. & Lakin, N.D. Recruitment of the Cell Cycle Checkpoint Kinase ATR to Chromatin during S-phase. in *Journal of Biological Chemistry* Vol. 279 (2004).

34. Duursma, A.M., Driscoll, R., Elias, J.E. & Cimprich, K.A. A Role for the MRN Complex in ATR Activation via TOPBP1 Recruitment. in *Molecular Cell* Vol. 50 (2013).
35. Cairns, B.R. et al. RSC, an essential, abundant chromatin-remodeling complex. in *Cell* Vol. 87 (1996).
36. Price, B.D. & D'Andrea, A.D. Chromatin remodeling at DNA double-strand breaks. in *Cell* Vol. 152 (2013).
37. Tang, L., Nogales, E. & Ciferri, C. Structure and function of SWI/SNF chromatin remodeling complexes and mechanistic implications for transcription. in *Progress in Biophysics and Molecular Biology* Vol. 102 (2010).
38. Heo, K. et al. FACT-Mediated Exchange of Histone Variant H2AX Regulated by Phosphorylation of H2AX and ADP-Ribosylation of Spt16. in *Molecular Cell* Vol. 30 (2008).
39. Rogakou, E.P., Pilch, D.R., Orr, A.H., Ivanova, V.S. & Bonner, W.M. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. in *Journal of Biological Chemistry* Vol. 273 (1998).
40. Lou, Z. et al. MDC1 maintains genomic stability by participating in the amplification of ATM-dependent DNA damage signals. in *Molecular Cell* Vol. 21 (2006).
41. Yuan, J., Adamski, R. & Chen, J. Focus on histone variant H2AX: To be or not to be. in *FEBS Letters* Vol. 584 (2010).
42. Burma, S., Chen, B.P., Murphy, M., Kurimasa, A. & Chen, D.J. ATM Phosphorylates Histone H2AX in Response to DNA Double-strand Breaks. in *Journal of Biological Chemistry* Vol. 276 42462-42467 (Elsevier, 2001).
43. Ward, I.M. & Chen, J. Histone H2AX Is Phosphorylated in an ATR-dependent Manner in Response to Replicational Stress. in *Journal of Biological Chemistry* Vol. 276 (2001).
44. An, J. et al. DNA-PKcs plays a dominant role in the regulation of H2AX phosphorylation in response to DNA damage and cell cycle progression. in *BMC Molecular Biology* Vol. 11 (2010).
45. Shen, T. & Huang, S. The Role of Cdc25A in the Regulation of Cell Proliferation and Apoptosis. in *Anti-Cancer Agents in Medicinal Chemistry* Vol. 12 (2012).
46. Tsai, T.C., Huang, H.P., Chang, K.T., Wang, C.J. & Chang, Y.C. Anthocyanins from roselle extract arrest cell cycle G2/M phase transition via ATM/Chk pathway in p53-deficient leukemia HL-60 cells. in *Environmental Toxicology* Vol. 32 (2017).
47. Liu, Q. et al. Chk1 is an essential kinase that is regulated by Atr and required for the G2/M DNA damage checkpoint. in *Genes and Development* Vol. 14 (2000).
48. Green, D.R. Apoptotic pathways: Ten minutes to dead. in *Cell* Vol. 121 (2005).
49. Strasser, A., Harris, A.W., Huang, D.C.S., Krammer, P.H. & Cory, S. Bcl-2 and Fas/APO-1 regulate distinct pathways to lymphocyte apoptosis. in *EMBO Journal* Vol. 14 (1995).
50. Chen, J. The Roles of MDM2 and MDMX Phosphorylation in Stress Signaling to p53. in *Genes and Cancer* Vol. 3 (2012).
51. Bakkenist, C.J. & Kastan, M.B. DNA damage activates ATM through intermolecular autophosphorylation and dimer dissociation. in *Nature* Vol. 421 (2003).
52. Cimprich, K.A. & Cortez, D. ATR: An essential regulator of genome integrity. in *Nature Reviews Molecular Cell Biology* Vol. 9 (2008).
53. Ou, Y.H., Chung, P.H., Sun, T.P. & Shieh, S.Y. p53 C-terminal phosphorylation by CHK1 and CHK2 participates in the regulation of DNA-damage-induced C-terminal acetylation. in *Molecular Biology of the Cell* Vol. 16 (2005).
54. Campbell, P.J. et al. Pan-cancer analysis of whole genomes. in *Nature* Vol. 578 (2020).
55. Langerød, A. et al. TP53 mutation status and gene expression profiles are powerful prognostic markers of breast cancer. in *Breast Cancer Research* Vol. 9 (2007).
56. Wang, Y. et al. TP53 mutations in early-stage ovarian carcinoma, relation to long-term survival. in *British Journal of Cancer* Vol. 90 (2004).
57. Kim, K.P. et al. Sister cohesion and structural axis components mediate homolog bias of meiotic recombination. in *Cell* Vol. 143 (2010).
58. Hanscom, T. & McVey, M. Regulation of Error-Prone DNA Double-Strand Break Repair and Its Impact on Genome Evolution. in *Cells* Vol. 9 (Multidisciplinary Digital Publishing Institute (MDPI), 2020).
59. Mansour, W.Y., Rhein, T. & Dahm-Daphi, J. The alternative end-joining pathway for repair of DNA double-strand breaks requires PARP1 but is not dependent upon microhomologies. in *Nucleic Acids Research* Vol. 38 (2010).
60. Kelso, A.A., Lopezcolorado, F.W., Bhargava, R. & Stark, J.M. Distinct roles of RAD52 and POLQ in chromosomal break repair and replication stress response. in *PLoS Genetics* Vol. 15 (2019).
61. Yousefzadeh, M.J. & Wood, R.D. DNA polymerase POLQ and cellular defense against DNA damage. in *DNA Repair* Vol. 12 (2013).
62. Schimmel, J., Kool, H., Schendel, R.v. & Tijsterman, M. Mutational signatures of non-homologous and polymerase theta-mediated end-joining in embryonic stem cells. in *The EMBO Journal* Vol. 36 3634 (European Molecular Biology Organization, 2017).
63. Wyatt, D.W. et al. Essential Roles for Polymerase  $\theta$ -Mediated End Joining in the Repair of Chromosome Breaks. in *Molecular Cell* Vol. 63 (2016).
64. Bhargava, R., Onyango, D.O. & Stark, J.M. Regulation of Single-Strand Annealing and its Role in Genome Maintenance. in *Trends in Genetics* Vol. 32 (2016).
65. Rothenberg, E., Grimme, J.M., Spies, M. & Ha, T. Human Rad52-mediated homology search and annealing occurs by continuous interactions between overlapping nucleoprotein complexes. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 105 (2008).

66. Al-minawi, A.Z., Saleh-gohari, N. & Helleday, T. The ERCC1/XPF endonuclease is required for efficient single-strand annealing and gene conversion in mammalian cells. in *Nucleic Acids Research* Vol. 36 (2008).
67. Seol, J.H., Shim, E.Y. & Lee, S.E. Microhomology-mediated end joining: Good, bad and ugly. in *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis* Vol. 809 (2018).
68. Ward, I.M., Minn, K., Jorda, K.G. & Chen, J. Accumulation of checkpoint protein 53BP1 at DNA breaks involves its binding to phosphorylated histone H2AX. in *Journal of Biological Chemistry* Vol. 278 (2003).
69. Escribano-Díaz, C. et al. A Cell Cycle-Dependent Regulatory Circuit Composed of 53BP1-RIF1 and BRCA1-CtIP Controls DNA Repair Pathway Choice. in *Molecular Cell* Vol. 49 (2013).
70. Noordermeer, S.M. et al. The shieldin complex mediates 53BP1-dependent DNA repair. in *Nature* Vol. 560 (2018).
71. Zimmermann, M., Lottersberger, F., Buonomo, S.B., Sfeir, A. & De Lange, T. 53BP1 regulates DSB repair using Rif1 to control 5' end resection. in *Science* Vol. 339 (2013).
72. Shao, Z. et al. Persistently bound Ku at DNA ends attenuates DNA end resection and homologous recombination. in *DNA Repair* Vol. 11 (2012).
73. Huertas, P. & Jackason, S.P. Human CtIP mediates cell cycle control of DNA end resection and double strand break repair. in *Journal of Biological Chemistry* Vol. 284 (2009).
74. Yu, X. & Chen, J. DNA Damage-Induced Cell Cycle Checkpoint Control Requires CtIP, a Phosphorylation-Dependent Binding Partner of BRCA1 C-Terminal Domains. in *Molecular and Cellular Biology* Vol. 24 (2004).
75. Brzovic, P.S. et al. Binding and recognition in the assembly of an active BRCA1/BARD1 ubiquitin-ligase complex. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 100 (2003).
76. Bothmer, A. et al. Regulation of DNA End Joining, Resection, and Immunoglobulin Class Switch Recombination by 53BP1. in *Molecular Cell* Vol. 42 (2011).
77. Isono, M. et al. BRCA1 Directs the Repair Pathway to Homologous Recombination by Promoting 53BP1 Dephosphorylation. in *Cell Reports* Vol. 18 (2017).
78. Truong, L.N. et al. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 110 (2013).
79. Cejka, P. DNA end resection: Nucleases team up with the right partners to initiate homologous recombination. in *Journal of Biological Chemistry* Vol. 290 (2015).
80. Nimonkar, A.V., Özsoy, A.Z., Genschel, J., Modrich, P. & Kowalczykowski, S.C. Human exonuclease 1 and BLM helicase interact to resect DNA and initiate DNA repair. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 105 (2008).
81. Nimonkar, A.V. et al. BLM-DNA2-RPA-MRN and EXO1-BLM-RPA-MRN constitute two DNA end resection machineries for human DNA break repair. in *Genes and Development* Vol. 25 (2011).
82. Tran, P.T., Erdeniz, N., Dudley, S. & Liskay, R.M. Characterization of nuclease-dependent functions of Exo1p in *Saccharomyces cerevisiae*. in *DNA Repair* Vol. 1 (2002).
83. Ivanov, E.L., Sugawara, N., Fishman-Lobell, J. & Haber, J.E. Genetic requirements for the single-strand annealing pathway of double-strand break repair in *Saccharomyces cerevisiae*. in *Genetics* Vol. 142 (1996).
84. Ceccaldi, R., Rondinelli, B. & D'Andrea, A.D. Repair Pathway Choices and Consequences at the Double-Strand Break. in *Trends in Cell Biology* Vol. 26 (2016).
85. Cejka, P. et al. DNA end resection by Dna2-Sgs1-RPA and its stimulation by Top3-Rmi1 and Mre11-Rad50-Xrs2. in *Nature* Vol. 467 (2010).
86. Chen, H., Lisby, M. & Symington, L.S. RPA Coordinates DNA End Resection and Prevents Formation of DNA Hairpins. in *Molecular Cell* Vol. 50 (2013).
87. Dray, E. et al. Enhancement of RAD51 recombinase activity by the tumor suppressor PALB2. in *Nature Structural and Molecular Biology* Vol. 17 (2010).
88. Sy, S.M.H., Huen, M.S.Y. & Chen, J. PALB2 is an integral component of the BRCA complex required for homologous recombination repair. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 106 (2009).
89. Zhang, F. et al. PALB2 Links BRCA1 and BRCA2 in the DNA-Damage Response. in *Current Biology* Vol. 19 (2009).
90. Zhao, W. et al. Promotion of BRCA2-Dependent Homologous Recombination by DSS1 via RPA Targeting and DNA Mimicry. in *Molecular Cell* Vol. 59 (2015).
91. Sugiyama, T. & Kowalczykowski, S.C. Rad52 protein associates with replication protein A (RPA)-single-stranded DNA to accelerate Rad51-mediated displacement of RPA and presynaptic complex formation. in *Journal of Biological Chemistry* Vol. 277 (2002).
92. Schay, G. et al. Without Binding ATP, Human Rad51 Does Not Form Helical Filaments on ssDNA. in *Journal of Physical Chemistry B* Vol. 120 (2016).
93. Taylor, M.R.G. et al. Rad51 Paralogs Remodel Pre-synaptic Rad51 Filaments to Stimulate Homologous Recombination. in *Cell* Vol. 162 (2015).
94. Crickard, J.B., Moevus, C.J., Kwon, Y., Sung, P. & Greene, E.C. Rad54 Drives ATP Hydrolysis-Dependent DNA Sequence Alignment during Homologous Recombination. in *Cell* Vol. 181 1380-1394.e18 (Cell Press, 2020).

95. Scully, R., Panday, A., Elango, R. & Willis, N.A. DNA double-strand break repair-pathway choice in somatic mammalian cells. in *Nature Reviews Molecular Cell Biology* Vol. 20 (2019).
96. Pâques, F. & Haber, J.E. Multiple Pathways of Recombination Induced by Double-Strand Breaks in *Saccharomyces cerevisiae* in *Microbiology and Molecular Biology Reviews* Vol. 63 (1999).
97. Bizard, A.H. & Hickson, I.D. The dissolution of double Holliday junctions. in *Cold Spring Harbor Perspectives in Biology* Vol. 6 (2014).
98. Punatar, R.S., Martin, M.J., Wyatt, H.D.M., Chan, Y.W. & West, S.C. Resolution of single and double Holliday junction recombination intermediates by GEN 1. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 114 (2017).
99. Sarbajna, S., Davies, D. & West, S.C. Roles of SLX1-SLX4, MUS81-EME1, and GEN1 in avoiding genome instability and mitotic catastrophe. in *Genes and Development* Vol. 28 (2014).
100. Chi, P., Van Komen, S., Sehorn, M.G., Sigurdsson, S. & Sung, P. Roles of ATP binding and ATP hydrolysis in human Rad51 recombinase function. in *DNA Repair* Vol. 5 (2006).
101. Trenner, A. & Sartori, A.A. Harnessing DNA Double-Strand Break Repair for Cancer Treatment. in *Frontiers in Oncology* Vol. 9 (2019).
102. Kile, A.C. et al. HLTf's Ancient HIRAN Domain Binds 3' DNA Ends to Drive Replication Fork Reversal. in *Molecular Cell* Vol. 58 (2015).
103. Kolinjivadi, A.M. et al. Smarcal1-Mediated Fork Reversal Triggers Mre11-Dependent Degradation of Nascent DNA in the Absence of Brca2 and Stable Rad51 Nucleofilaments. in *Molecular Cell* Vol. 67 (2017).
104. Neelsen, K.J. & Lopes, M. Replication fork reversal in eukaryotes: From dead end to dynamic response. in *Nature Reviews Molecular Cell Biology* Vol. 16 (2015).
105. Vujanovic, M. et al. Replication Fork Slowing and Reversal upon DNA Damage Require PCNA Polyubiquitination and ZRANB3 DNA Translocase Activity. in *Molecular Cell* Vol. 67 (2017).
106. Zeman, M.K. & Cimprich, K.A. Causes and consequences of replication stress. in *Nature Cell Biology* Vol. 16 (2014).
107. Lemaçon, D. et al. MRE11 and EXO1 nucleases degrade reversed forks and elicit MUS81-dependent fork rescue in BRCA2-deficient cells. in *Nature Communications* Vol. 8 (2017).
108. Schlacher, K. et al. Double-strand break repair-independent role for BRCA2 in blocking stalled replication fork degradation by MRE11. in *Cell* Vol. 145 (2011).
109. Schlacher, K., Wu, H. & Jasin, M. A Distinct Replication Fork Protection Pathway Connects Fanconi Anemia Tumor Suppressors to RAD51-BRCA1/2. in *Cancer Cell* Vol. 22 (2012).
110. Tagliatela, A. et al. Restoration of Replication Fork Stability in BRCA1- and BRCA2-Deficient Cells by Inactivation of SNF2-Family Fork Remodelers. in *Molecular Cell* Vol. 68 (2017).
111. Branzei, D. & Szakal, B. DNA damage tolerance by recombination: Molecular pathways and DNA structures. in *DNA Repair* Vol. 44 (2016).
112. Prado, F. Homologous recombination maintenance of genome integrity during DNA damage tolerance. in *Molecular and Cellular Oncology* Vol. 1 (2014).
113. Lambert, S. et al. Homologous recombination restarts blocked replication forks at the expense of genome rearrangements by template exchange. in *Molecular Cell* Vol. 39 (2010).
114. McEachern, M.J. & Haber, J.E. Break-induced replication and recombinational telomere elongation in yeast. in *Annual Review of Biochemistry* Vol. 75 (2006).
115. K, S., H, W. & M, J. A distinct replication fork protection pathway connects Fanconi anemia tumor suppressors to RAD51-BRCA1/2. in *Cancer cell* Vol. 22 106-116 (Cancer Cell, 2012).
116. Schlacher, K. et al. Double-Strand Break Repair Independent Role For BRCA2 In Blocking Stalled Replication Fork Degradation By MRE11. in *Cell* Vol. 145 529 (NIH Public Access, 2011).
117. Ceccaldi, R., Sarangi, P. & D'Andrea, A.D. The Fanconi anaemia pathway: New players and new functions. in *Nature Reviews Molecular Cell Biology* Vol. 17 (2016).
118. Kottmann, M.C. & Smogorzewska, A. Fanconi anaemia and the repair of Watson and Crick DNA crosslinks. in *Nature* Vol. 493 (2013).
119. Palovcak, A., Liu, W., Yuan, F. & Zhang, Y. Maintenance of genome stability by Fanconi anemia proteins. in *Cell and Bioscience* Vol. 7 (2017).
120. Karanja, K.K., Lee, E.H., Hendrickson, E.A. & Campbell, J.L. Preventing over-resection by DNA2 helicase/nuclease suppresses repair defects in Fanconi anemia cells. in *Cell Cycle* Vol. 13 (2014).
121. Murina, O. et al. FANCD2 and CtIP cooperate to repair DNA interstrand crosslinks. in *Cell Reports* Vol. 7 (2014).
122. Kalb, R. et al. Lack of sensitivity of primary Fanconi's anemia fibroblasts to UV and ionizing radiation. in *Radiation Research* Vol. 161 (2004).
123. Nakanishi, K. et al. Homology-directed Fanconi anemia pathway cross-link repair is dependent on DNA replication. in *Nature Structural and Molecular Biology* Vol. 18 (2011).
124. Howlander, N. et al. SEER Cancer Statistics Review, 1975-2017 SEER Cancer Statistics. in *National Cancer Institute* (2017).
125. Antoniou, A. et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: A combined analysis of 22 studies. in *American Journal of Human Genetics* Vol. 72 (2003).
126. Chen, S. & Parmigiani, G. Meta-Analysis of BRCA1 and BRCA2 Penetrance. in *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* Vol. 25 1329 (NIH Public Access, 2007).



127. Tai, Y.C., Domchek, S., Parmigiani, G. & Chen, S. Breast cancer risk among male BRCA1 and BRCA2 mutation carriers. in *Journal of the National Cancer Institute* Vol. 99 (2007).
128. Van Asperen, C.J. et al. Cancer risks in BRCA2 families: Estimates for sites other than breast and ovary. in *Journal of Medical Genetics* Vol. 42 (2005).
129. Thompson, D. & Easton, D.F. Cancer incidence in BRCA1 mutation carriers. in *Journal of the National Cancer Institute* Vol. 94 (2002).
130. Antoniou, A.C. et al. Breast-Cancer Risk in Families with Mutations in PALB2. in *New England Journal of Medicine* Vol. 371 (2014).
131. Yang, X. et al. Cancer risks associated with germline PALB2 pathogenic variants: An international study of 524 families. in *Journal of Clinical Oncology* Vol. 38 (2020).
132. Karczewski, K.J. et al. The mutational constraint spectrum quantified from variation in 141,456 humans. in *Nature* Vol. 581 (2020).
133. Miki, Y. et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. in *Science* Vol. 266 (1994).
134. Ramus, S.J. et al. Germline mutations in the BRIP1, BARD1, PALB2, and NBN genes in women with ovarian cancer. in *Journal of the National Cancer Institute* Vol. 107 (2015).
135. Weber-Lassalle, N. et al. BRIP1 loss-of-function mutations confer high risk for familial ovarian cancer, but not familial breast cancer. in *Breast Cancer Research* Vol. 20 (2018).
136. Ali, M., Delozier, C.D. & Chaudhary, U. BRIP-1 germline mutation and its role in colon cancer: Presentation of two case reports and review of literature. in *BMC Medical Genetics* Vol. 20 (2019).
137. Loveday, C. et al. Germline mutations in RAD51D confer susceptibility to ovarian cancer. in *Nature Genetics* Vol. 43 (2011).
138. Meindl, A. et al. Germline mutations in breast and ovarian cancer pedigrees establish RAD51C as a human cancer susceptibility gene. in *Nature Genetics* Vol. 42 (2010).
139. Golmard, L. et al. Germline mutation in the RAD51B gene confers predisposition to breast cancer. in *BMC Cancer* Vol. 13 (2013).
140. Park, D.J. et al. Rare mutations in XRCC2 increase the risk of breast cancer. in *American Journal of Human Genetics* Vol. 90 (2012).
141. Smith, T.R. et al. Polymorphisms of XRCC1 and XRCC3 genes and susceptibility to breast cancer. in *Cancer Letters* Vol. 190 (2003).
142. Manié, E. et al. Genomic hallmarks of homologous recombination deficiency in invasive breast carcinomas. in *International Journal of Cancer* Vol. 138 (2016).
143. Esteller, M. et al. Promoter hypermethylation and BRCA1 inactivation in sporadic breast and ovarian tumors. in *Journal of the National Cancer Institute* Vol. 92 (2000).
144. Glodzik, D. et al. Comprehensive molecular comparison of BRCA1 hypermethylated and BRCA1 mutated triple negative breast cancers. in *Nature Communications* Vol. 11 (2020).
145. Hansmann, T. et al. Constitutive promoter methylation of BRCA1 and RAD51C in patients with familial ovarian cancer and early-onset sporadic breast cancer. in *Human Molecular Genetics* Vol. 21 4669-4679 (2012).
146. Bell, D. et al. Integrated genomic analyses of ovarian carcinoma. in *Nature* Vol. 474 (2011).
147. Pal, T. et al. BRCA1 and BRCA2 mutations account for a large proportion of ovarian carcinoma cases. in *Cancer* Vol. 104 (2005).
148. Alsop, K. et al. BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: A report from the Australian ovarian cancer study group. in *Journal of Clinical Oncology* Vol. 30 (2012).
149. Koboldt, D.C. et al. Comprehensive molecular portraits of human breast tumours. in *Nature* Vol. 490 (2012).
150. Nik-Zainal, S. et al. Landscape of somatic mutations in 560 breast cancer whole-genome sequences. in *Nature* Vol. 534 (2016).
151. Bane, A.L. et al. BRCA2 mutation-associated breast cancers exhibit a distinguishing phenotype based on morphology and molecular profiles from tissue microarrays. in *American Journal of Surgical Pathology* Vol. 31 (2007).
152. Mavaddat, N. et al. Cancer risks for BRCA1 and BRCA2 mutation carriers: Results from prospective analysis of EMBRACE. in *Journal of the National Cancer Institute* Vol. 105 (2013).
153. Palacios, J., Honrado, E., Cigudosa, J.C. & Benítez, J. ERBB2 and MYC alterations in BRCA1- and BRCA2-associated cancers. in *Genes, Chromosomes and Cancer* Vol. 42 204-205 (John Wiley & Sons, Ltd, 2005).
154. Mateo, J. et al. DNA-Repair Defects and Olaparib in Metastatic Prostate Cancer. in *New England Journal of Medicine* Vol. 373 (2015).
155. Pritchard, C.C. et al. Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. in *New England Journal of Medicine* Vol. 375 (2016).
156. Casolino, R. et al. Homologous Recombination Deficiency in Pancreatic Cancer: A Systematic Review and Prevalence Meta-Analysis. in *Journal of clinical oncology : official journal of the American Society of Clinical Oncology* Vol. 39 (2021).
157. Li, A. et al. Homologous recombination DNA repair defects in PALB2-associated breast cancers. in *npj Breast Cancer* Vol. 5 (2019).

158. Lee, J.E.A. et al. Molecular analysis of PALB2-associated breast cancers. in *Journal of Pathology* Vol. 245 (2018).
159. Schmutte, C. et al. Characterization of the human Rad51 genomic locus and examination of tumors with 15q14-15 loss of heterozygosity (LOH). in *Cancer Research* Vol. 59 (1999).
160. Moynahan, M.E., Chiu, J.W., Koller, B.H. & Jasint, M. Brca1 controls homology-directed DNA repair. in *Molecular Cell* Vol. 4 (1999).
161. Snouwaert, J.N. et al. BRCA1 deficient embryonic stem cells display a decreased homologous recombination frequency and an increased frequency of non-homologous recombination that is corrected by expression of a Brca1 transgene. in *Oncogene* Vol. 18 (1999).
162. Tutt, A. et al. Mutation in Brca2 stimulates error-prone homology-directed repair of DNA double-strand breaks occurring between repeated sequences. in *EMBO Journal* Vol. 20 (2001).
163. Deng, C.X. & Scott, F. Role of the tumor suppressor gene Brca1 in genetic stability and mammary gland tumor formation. in *Oncogene* Vol. 19 (2000).
164. Futaki, M. & Liu, J.M. Chromosomal breakage syndromes and the BRCA1 genome surveillance complex. in *Trends in Molecular Medicine* Vol. 7 (2001).
165. Patel, K.J. et al. Involvement of Brca2 in DNA repair. in *Molecular Cell* Vol. 1 (1998).
166. Welcsh, P.L., Owens, K.N. & King, M.C. Insights into the functions of BRCA1 and BRCA2. in *Trends in Genetics* Vol. 16 (2000).
167. Turner, N., Tutt, A. & Ashworth, A. Hallmarks of 'BRCAness' in sporadic cancers. in *Nature Reviews Cancer* Vol. 4 (2004).
168. Weinstein, J.N. et al. The cancer genome atlas pan-cancer analysis project. in *Nature Genetics* Vol. 45 (2013).
169. Nguyen, L., W. M. Martens, J., Van Hoeck, A. & Cuppen, E. Pan-cancer landscape of homologous recombination deficiency. in *Nature Communications* Vol. 11 (2020).
170. Rothblum-Oviatt, C. et al. Ataxia telangiectasia: A review. in *Orphanet Journal of Rare Diseases* Vol. 11 (2016).
171. Digweed, M., Reis, A. & Sperling, K. Nijmegen Breakage Syndrome: Consequences of defective DNA double strand break repair. in *BioEssays* Vol. 21 (1999).
172. de Bono, J. et al. Olaparib for Metastatic Castration-Resistant Prostate Cancer. in *New England Journal of Medicine* Vol. 382 (2020).
173. Byrski, T. et al. Response to neoadjuvant therapy with cisplatin in BRCA1-positive breast cancer patients. in *Breast Cancer Research and Treatment* Vol. 115 (2009).
174. Tutt, A. et al. Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: The TNT Trial. in *Nature Medicine* Vol. 24 (2018).
175. Bryant, H.E. et al. Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. in *Nature* Vol. 434 (2005).
176. Farmer, H. et al. Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. in *Nature* Vol. 434 (2005).
177. Ashworth, A. A synthetic lethal therapeutic approach: Poly(ADP) ribose polymerase inhibitors for the treatment of cancers deficient in DNA double-strand break repair. in *Journal of Clinical Oncology* Vol. 26 (2008).
178. Mirza, M.R. et al. Niraparib Maintenance Therapy in Platinum-Sensitive, Recurrent Ovarian Cancer. in *New England Journal of Medicine* Vol. 375 (2016).
179. Murai, J. et al. Trapping of PARP1 and PARP2 by clinical PARP inhibitors. in *Cancer Research* Vol. 72 (2012).
180. Murai, J. et al. Rationale for poly(ADP-ribose) polymerase (PARP) inhibitors in combination therapy with camptothecins or temozolomide based on PARP trapping versus catalytic inhibition. in *Journal of Pharmacology and Experimental Therapeutics* Vol. 349 (2014).
181. Patel, A.G., Sarkaria, J.N. & Kaufmann, S.H. Nonhomologous end joining drives poly(ADP-ribose) polymerase (PARP) inhibitor lethality in homologous recombination-deficient cells. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 108 (2011).
182. Li, M. & Yu, X. Function of BRCA1 in the DNA Damage Response Is Mediated by ADP-Ribosylation. in *Cancer Cell* Vol. 23 (2013).
183. Howard, S.M., Yanez, D.A. & Stark, J.M. DNA Damage Response Factors from Diverse Pathways, Including DNA Crosslink Repair, Mediate Alternative End Joining. in *PLoS Genetics* Vol. 11 (2015).
184. Konstantinopoulos, P.A., Ceccaldi, R., Shapiro, G.I. & D'Andrea, A.D. Homologous recombination deficiency: Exploiting the fundamental vulnerability of ovarian cancer. in *Cancer Discovery* Vol. 5 (2015).
185. Ceccaldi, R. et al. Homologous recombination-deficient tumors are hyper-dependent on POLQ-mediated repair. in *Nature* Vol. 518 258 (NIH Public Access, 2015).
186. Hanzlikova, H. et al. The Importance of Poly(ADP-Ribose) Polymerase as a Sensor of Unligated Okazaki Fragments during DNA Replication. in *Molecular Cell* Vol. 71 (2018).
187. Maya-Mendoza, A. et al. High speed of fork progression induces DNA replication stress and genomic instability. in *Nature* Vol. 559 (2018).
188. Mohyuddin, G.R. et al. Similar response rates and survival with PARP inhibitors for patients with solid tumors harboring somatic versus Germline BRCA mutations: A Meta-analysis and systematic review. in *BMC Cancer* Vol. 20 (2020).

189. Vencken, P.M.L.H. et al. Chemosensitivity and outcome of BRCA1- and BRCA2-associated ovarian cancer patients after first-line chemotherapy compared with sporadic ovarian cancer patients. in *Annals of Oncology* Vol. 22 (2011).
190. Ledermann, J.A. et al. Overall survival in patients with platinum-sensitive recurrent serous ovarian cancer receiving olaparib maintenance monotherapy: an updated analysis from a randomised, placebo-controlled, double-blind, phase 2 trial. in *The Lancet Oncology* Vol. 17 (2016).
191. Drost, R. et al. BRCA1 RING function is essential for tumor suppression but dispensable for therapy resistance. in *Cancer Cell* Vol. 20 (2011).
192. Drost, R. et al. BRCA1185delAG tumors may acquire therapy resistance through expression of RING-less BRCA1. in *Journal of Clinical Investigation* Vol. 126 (2016).
193. Labidi-Galy, S.I. et al. Location of mutation in BRCA2 gene and survival in patients with ovarian cancer. in *Clinical Cancer Research* Vol. 24 (2018).
194. Kim, H. et al. Combining PARP with ATR inhibition overcomes PARP inhibitor and platinum resistance in ovarian cancer models. in *Nature Communications* Vol. 11 (2020).
195. Zhou, J. et al. A first-in-class polymerase theta inhibitor selectively targets homologous-recombination-deficient tumors. in *Nature Cancer* Vol. 2 (2021).
196. Zatreanu, D. et al. Polθ inhibitors elicit BRCA-gene synthetic lethality and target PARP inhibitor resistance. in *Nature Communications* Vol. 12 (2021).
197. Sun, C. et al. BRD4 Inhibition Is Synthetic Lethal with PARP Inhibitors through the Induction of Homologous Recombination Deficiency. in *Cancer Cell* Vol. 33 (2018).
198. Fong, P.C. et al. Poly(ADP)-ribose polymerase inhibition: Frequent durable responses in BRCA carrier ovarian cancer correlating with platinum-free interval. in *Journal of Clinical Oncology* Vol. 28 (2010).
199. Audeh, M.W. et al. Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: A proof-of-concept trial. in *The Lancet* Vol. 376 (2010).
200. Mayor, P., Gay, L.M., Lele, S. & Elvin, J.A. BRCA1 reversion mutation acquired after treatment identified by liquid biopsy. in *Gynecologic Oncology Reports* Vol. 21 (2017).
201. Edwards, S.L. et al. Resistance to therapy caused by intragenic deletion in BRCA2. in *Nature* Vol. 451 (2008).
202. Carneiro, B.A. et al. Acquired Resistance to Poly (ADP-ribose) Polymerase Inhibitor Olaparib in BRCA2 - Associated Prostate Cancer Resulting From Biallelic BRCA2 Reversion Mutations Restores Both Germline and Somatic Loss-of-Function Mutations. in *JCO Precision Oncology* (2018).
203. Kondrashova, O. et al. Secondary somatic mutations restoring RAD51C and RAD51D associated with acquired resistance to the PARP inhibitor rucaparib in high-grade ovarian carcinoma. in *Cancer Discovery* Vol. 7 (2017).
204. Jaspers, J.E. et al. Loss of 53BP1 causes PARP inhibitor resistance in BRCA1-mutated mouse mammary tumors. in *Cancer Discovery* Vol. 3 (2013).
205. Hurley, R.M. et al. 53BP1 as a potential predictor of response in PARP inhibitor-treated homologous recombination-deficient ovarian cancer. in *Gynecologic Oncology* Vol. 153 (2019).
206. Popova, T. et al. Ploidy and large-scale genomic instability consistently identify basal-like breast carcinomas with BRCA1/2 inactivation. in *Cancer Research* Vol. 72 (2012).
207. Zack, T.I. et al. Pan-cancer patterns of somatic copy number alteration. in *Nature Genetics* Vol. 45 (2013).
208. Abkevich, V. et al. Patterns of genomic loss of heterozygosity predict homologous recombination repair defects in epithelial ovarian cancer. in *British Journal of Cancer* Vol. 107 (2012).
209. Birkbak, N.J. et al. Telomeric allelic imbalance indicates defective DNA repair and sensitivity to DNA-damaging agents. in *Cancer Discovery* Vol. 2 (2012).
210. Timms, K.M. et al. Association of BRCA1/2 defects with genomic scores predictive of DNA damage repair deficiency among breast cancer subtypes. in *Breast Cancer Research* Vol. 16 (2014).
211. Melinda, L.T. et al. Homologous recombination deficiency (hrd) score predicts response to platinum-containing neoadjuvant chemotherapy in patients with triple-negative breast cancer. in *Clinical Cancer Research* Vol. 22 (2016).
212. Ray-Coquard, I. et al. Olaparib plus Bevacizumab as First-Line Maintenance in Ovarian Cancer. in *New England Journal of Medicine* Vol. 381 (2019).
213. Sztupinszki, Z. et al. Migrating the SNP array-based homologous recombination deficiency measures to next generation sequencing data of breast cancer. in *npj Breast Cancer* Vol. 4 (2018).
214. Lee, D.D. & Seung, H.S. Learning the parts of objects by non-negative matrix factorization. in *Nature* Vol. 401 (1999).
215. Alexandrov, L.B. et al. Signatures of mutational processes in human cancer. in *Nature* Vol. 500 (2013).
216. Waddell, N. et al. Whole genomes redefine the mutational landscape of pancreatic cancer. in *Nature* Vol. 518 (2015).
217. Polak, P. et al. A mutational signature reveals alterations underlying deficient homologous recombination repair in breast cancer. in *Nature Genetics* Vol. 49 (2017).
218. Connor, A.A. et al. Association of distinct mutational signatures with correlates of increased immune activity in pancreatic ductal adenocarcinoma. in *JAMA Oncology* Vol. 3 (2017).
219. Póti, Á. et al. Correlation of homologous recombination deficiency induced mutational signatures with sensitivity to PARP inhibitors and cytotoxic agents. in *Genome Biology* Vol. 20 (2019).
220. Gulhan, D.C., Lee, J.J.K., Melloni, G.E.M., Cortés-Ciriano, I. & Park, P.J. Detecting the mutational signature of homologous recombination deficiency in clinical samples. in *Nature Genetics* Vol. 51 (2019).

221. Jager, M. et al. Deficiency of nucleotide excision repair is associated with mutational signature observed in cancer. in *Genome Research* Vol. 29 (2019).
222. Singh, V.K., Rastogi, A., Hu, X., Wang, Y. & De, S. Mutational signature SBS8 predominantly arises due to late replication errors in cancer. in *Communications Biology* Vol. 3 (2020).
223. Willis, N.A. et al. Mechanism of tandem duplication formation in BRCA1-mutant cells. in *Nature* Vol. 551 (2017).
224. Davies, H. et al. HRDetect is a predictor of BRCA1 and BRCA2 deficiency based on mutational signatures. in *Nature Medicine* Vol. 23 (2017).
225. Staaf, J. et al. Whole-genome sequencing of triple-negative breast cancers in a population-based clinical study. in *Nature Medicine* Vol. 25 (2019).
226. Zhao, E.Y. et al. Homologous recombination deficiency and platinum-based therapy outcomes in advanced breast cancer. in *Clinical Cancer Research* Vol. 23 (2017).
227. Haaf, T., Golub, E.I., Reddy, G., Radding, C.M. & Ward, D.C. Nuclear foci of mammalian Rad51 recombination protein in somatic cells after DNA damage and its localization in synaptonemal complexes. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 92 (1995).
228. Cruz, C. et al. RAD51 foci as a functional biomarker of homologous recombination repair and PARP inhibitor resistance in germline BRCA-mutated breast cancer. in *Annals of Oncology* Vol. 29 (2018).
229. Serra Elizalde, V. et al. 10 Detection of homologous recombination repair deficiency (HRD) in treatment-naïve early triple-negative breast cancer (TNBC) by RAD51 foci and comparison with DNA-based tests. in *Annals of Oncology* Vol. 32 (2021).
230. Stok, C., Kok, Y.P., Van Den Tempel, N. & Van Vugt, M.A.T.M. Shaping the BRCAness mutational landscape by alternative double-strand break repair, replication stress and mitotic aberrancies. in *Nucleic Acids Research* Vol. 49 (2021).
231. Difilippantonio, M.J. et al. Dna repair protein Ku80 suppresses chromosomal aberrations and malignant transformation. in *Nature* Vol. 404 (2000).
232. Ghezraoui, H. et al. Chromosomal Translocations in Human Cells Are Generated by Canonical Nonhomologous End-Joining. in *Molecular Cell* Vol. 55 (2014).
233. Bétermier, M., Bertrand, P. & Lopez, B.S. Is Non-Homologous End-Joining Really an Inherently Error-Prone Process? in *PLoS Genetics* Vol. 10 (2014).
234. McVey, M. & Lee, S.E. MMEJ repair of double-strand breaks (director's cut): deleted sequences and alternative endings. in *Trends in Genetics* Vol. 24 (2008).
235. van Schendel, R., van Heteren, J., Welten, R. & Tijsterman, M. Genomic Scars Generated by Polymerase Theta Reveal the Versatile Mechanism of Alternative End-Joining. in *PLoS Genetics* Vol. 12 (2016).
236. Seki, M. et al. High-efficiency bypass of DNA damage by human DNA polymerase Q. in *EMBO Journal* Vol. 23 (2004).
237. Hwang, T. et al. Defining the mutation signatures of DNA polymerase  $\theta$  in cancer genomes. in *NAR Cancer* Vol. 2 (2020).
238. Kamp, J.A., van Schendel, R., Dilweg, I.W. & Tijsterman, M. BRCA1-associated structural variations are a consequence of polymerase theta-mediated end-joining. in *Nature Communications* Vol. 11 (2020).
239. Anantha, R.W. et al. Functional and mutational landscapes of BRCA1 for homology-directed repair and therapy resistance. in *eLife* Vol. 6 (2017).
240. Reh, W.A., Nairn, R.S., Lowery, M.P. & Vasquez, K.M. The homologous recombination protein RAD51D protects the genome from large deletions. in *Nucleic Acids Research* Vol. 45 (2017).
241. Scheinin, I. et al. DNA copy number analysis of fresh and formalin-fixed specimens by shallow whole-genome sequencing with identification and exclusion of problematic regions in the genome assembly. in *Genome Research* Vol. 24 (2014).
242. Chin, S.F. et al. Shallow whole genome sequencing for robust copy number profiling of formalin-fixed paraffin-embedded breast cancers. in *Experimental and Molecular Pathology* Vol. 104 (2018).
243. Hwang, K.B. et al. Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings. in *Scientific Reports* Vol. 9 (2019).
244. Li, H. [Heng Li - Compares BWA to other long read aligners like CUSHAW2] Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. in *arXiv preprint arXiv* (2013).
245. Musich, R., Cadle-Davidson, L. & Osier, M.V. Comparison of Short-Read Sequence Aligners Indicates Strengths and Weaknesses for Biologists to Consider. in *Frontiers in Plant Science* Vol. 12 (2021).
246. Teo, S.M., Pawitan, Y., Ku, C.S., Chia, K.S. & Salim, A. Statistical challenges associated with detecting copy number variations with next-generation sequencing. in *Bioinformatics* Vol. 28 (2012).
247. Dong, Z. et al. Low-pass whole-genome sequencing in clinical cytogenetics: A validated approach. in *Genetics in Medicine* Vol. 18 (2016).
248. Dong, Z. et al. Copy-number variants detection by low-pass whole-genome sequencing. in *Current Protocols in Human Genetics* Vol. 2017 (2017).
249. Derrien, T. et al. Fast computation and applications of genome mappability. in *PLoS ONE* Vol. 7 (2012).
250. Benjamini, Y. & Speed, T.P. Summarizing and correcting the GC content bias in high-throughput sequencing. in *Nucleic Acids Research* Vol. 40 (2012).
251. Valsesia, A., Macé, A., Jacquemont, S., Beckmann, J.S. & Kutalik, Z. The growing importance of CNVs: New insights for detection and clinical interpretation. in *Frontiers in Genetics* Vol. 4 (2013).
252. Boeva, V. et al. Control-FREEC: A tool for assessing copy number and allelic content using next-generation sequencing data. in *Bioinformatics* Vol. 28 (2012).

253. Adalsteinsson, V.A. et al. Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. in *Nature Communications* Vol. 8 (2017).
254. Dunham, I. et al. An integrated encyclopedia of DNA elements in the human genome. in *Nature* Vol. 489 (2012).
255. Harchaoui, Z. & Lévy-Leduc, C. Catching change-points with Lasso. in *Advances in Neural Information Processing Systems 20 - Proceedings of the 2007 Conference* (2009).
256. Olshen, A.B., Venkatraman, E.S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. in *Biostatistics* Vol. 5 (2004).
257. Venkatraman, E.S. & Olshen, A.B. A faster circular binary segmentation algorithm for the analysis of array CGH data. in *Bioinformatics* Vol. 23 (2007).
258. Goundiam, O. et al. Histo-genomic stratification reveals the frequent amplification/overexpression of CCNE1 and BRD4 genes in non-BRCAness high grade ovarian carcinoma. in *International Journal of Cancer* Vol. 137 (2015).
259. Loap, P. et al. Combination of Olaparib and Radiation Therapy for Triple Negative Breast Cancer: Preliminary Results of the RADIOPARP Phase 1 Trial. *Int J Radiat Oncol Biol Phys* **109**, 436-440 (2021).
260. Juan, H.C., Lin, Y., Chen, H.R. & Fann, M.J. Cdk12 is essential for embryonic development and the maintenance of genomic stability. in *Cell Death and Differentiation* Vol. 23 (2016).
261. Ko, T.K., Kelly, E. & Pines, J. CrkRS: A novel conserved Cdc2-related protein kinase that colocalises with SC35 speckles. in *Journal of Cell Science* Vol. 114 (2001).
262. Kohoutek, J. & Blazek, D. Cyclin K goes with Cdk12 and Cdk13. in *Cell Division* Vol. 7 (2012).
263. Bartkowiak, B. & Greenleaf, A.L. Expression, purification, and identification of associated proteins of the full-length hCDK12/CyclinK complex. in *Journal of Biological Chemistry* Vol. 290 (2015).
264. Blazek, D. et al. The cyclin K/Cdk12 complex maintains genomic stability via regulation of expression of DNA damage response genes. in *Genes and Development* Vol. 25 (2011).
265. Zaborowska, J., Egloff, S. & Murphy, S. The pol II CTD: New twists in the tail. in *Nature Structural and Molecular Biology* Vol. 23 (2016).
266. Harlen, K.M. & Churchman, L.S. The code and beyond: Transcription regulation by the RNA polymerase II carboxy-terminal domain. in *Nature Reviews Molecular Cell Biology* Vol. 18 (2017).
267. Bartkowiak, B. et al. CDK12 is a transcription elongation-associated CTD kinase, the metazoan ortholog of yeast Ctk1. in *Genes and Development* Vol. 24 (2010).
268. Cheng, S.-W.G. et al. Interaction of Cyclin-Dependent Kinase 12/CrkRS with Cyclin K1 Is Required for the Phosphorylation of the C-Terminal Domain of RNA Polymerase II. in *Molecular and Cellular Biology* Vol. 32 (2012).
269. Bösken, C.A. et al. The structure and substrate specificity of human Cdk12/Cyclin K. in *Nature Communications* Vol. 5 (2014).
270. Liang, K. et al. Characterization of Human Cyclin-Dependent Kinase 12 (CDK12) and CDK13 Complexes in C-Terminal Domain Phosphorylation, Gene Transcription, and RNA Processing. in *Molecular and Cellular Biology* Vol. 35 (2015).
271. Chirackal Manavalan, A.P. et al. CDK12 controls G1/S progression by regulating RNAPII processivity at core DNA replication genes. in *EMBO reports* Vol. 20 (2019).
272. Bartkowiak, B., Yan, C. & Greenleaf, A.L. Engineering an analog-sensitive CDK12 cell line using CRISPR/Cas. in *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms* Vol. 1849 (2015).
273. Zhang, T. et al. Covalent targeting of remote cysteine residues to develop CDK12 and CDK13 inhibitors. in *Nature Chemical Biology* Vol. 12 (2016).
274. Krajewska, M. et al. CDK12 loss in cancer cells affects DNA damage response genes through premature cleavage and polyadenylation. in *Nature Communications* Vol. 10 (2019).
275. Dubburly, S.J., Boutz, P.L. & Sharp, P.A. CDK12 regulates DNA repair genes by suppressing intronic polyadenylation. in *Nature* Vol. 564 (2018).
276. Fitz, J., Neumann, T. & Pavri, R. Regulation of RNA polymerase II processivity by Spt5 is restricted to a narrow window during elongation. in *The EMBO Journal* Vol. 37 (2018).
277. Hoshii, T. et al. A Non-catalytic Function of SETD1A Regulates Cyclin K and the DNA Damage Response. in *Cell* Vol. 172 (2018).
278. Hou, L. et al. Paf1C regulates RNA polymerase II progression by modulating elongation rate. in *Proceedings of the National Academy of Sciences of the United States of America* Vol. 116 (2019).
279. Van Oss, S.B., Cucinotta, C.E. & Arndt, K.M. Emerging Insights into the Roles of the Paf1 Complex in Gene Regulation. in *Trends in Biochemical Sciences* Vol. 42 (2017).
280. Chen, H.-H., Wang, Y.-C. & Fann, M.-J. Identification and Characterization of the CDK12/Cyclin L1 Complex Involved in Alternative Splicing Regulation. in *Molecular and Cellular Biology* Vol. 26 (2006).
281. Spector, D.L. & Lamond, A.I. Nuclear speckles. in *Cold Spring Harbor Perspectives in Biology* Vol. 3 (2011).
282. Davidson, L., Muniz, L. & West, S. 3' end formation of pre-mRNA and phosphorylation of Ser2 on the RNA polymerase II CTD are reciprocally coupled in human cells. in *Genes and Development* Vol. 28 (2014).
283. Tien, J.F. et al. CDK12 regulates alternative last exon mRNA splicing and promotes breast cancer cell invasion. in *Nucleic Acids Research* Vol. 45 (2017).
284. Hay, N. & Sonenberg, N. Upstream and downstream of mTOR. in *Genes and Development* Vol. 18 (2004).
285. Choi, S.H. et al. CDK12 phosphorylates 4E-BP1 to enable mTORC1-dependent translation and mitotic genome stability. in *Genes and Development* Vol. 33 (2019).

286. Chen, H.R., Juan, H.C., Wong, Y.H., Tsai, J.W. & Fann, M.J. Cdk12 regulates neurogenesis and late-arising neuronal migration in the developing cerebral cortex. in *Cerebral Cortex* Vol. 27 (2017).
287. Geng, M. et al. Targeting CDK12-mediated transcription regulation in anaplastic thyroid carcinoma. in *Biochemical and Biophysical Research Communications* Vol. 520 (2019).
288. Lei, T. et al. Cyclin K regulates prereplicative complex assembly to promote mammalian cell proliferation. in *Nature Communications* Vol. 9 (2018).
289. Sircoulomb, F. et al. Genome profiling of ERBB2-amplified breast cancers. in *BMC Cancer* Vol. 10 (2010).
290. Mertins, P. et al. Proteogenomics connects somatic mutations to signalling in breast cancer. in *Nature* Vol. 534 (2016).
291. Choi, H.J. et al. CDK12 drives breast tumor initiation and trastuzumab resistance via WNT and IRS1-ErbB-PI3K signaling. in *EMBO Reports* Vol. 20 (European Molecular Biology Organization, 2019).
292. Ji, J. et al. Expression pattern of CDK12 protein in gastric cancer and its positive correlation with CD8+ cell density and CCL12 expression. in *International Journal of Medical Sciences* Vol. 16 (2019).
293. Naidoo, K. et al. Evaluation of CDK12 protein expression as a potential novel biomarker for DNA damage response-targeted therapies in breast cancer. in *Molecular Cancer Therapeutics* Vol. 17 (2018).
294. Quereda, V. et al. Therapeutic Targeting of CDK12/CDK13 in Triple-Negative Breast Cancer. in *Cancer Cell* Vol. 36 (2019).
295. Ekumi, K.M. et al. Ovarian carcinoma CDK12 mutations misregulate expression of DNA repair genes via deficient formation and function of the Cdk12/CycK complex. in *Nucleic Acids Research* Vol. 43 (2015).
296. Popova, T. et al. Ovarian cancers harboring inactivating mutations in CDK12 display a distinct genomic instability pattern characterized by large tandem duplications. in *Cancer Research* Vol. 76 (2016).
297. Reimers, M.A. et al. Clinical Outcomes in Cyclin-dependent Kinase 12 Mutant Advanced Prostate Cancer. in *European Urology* Vol. 77 (2020).
298. Wu, Y.M. et al. Inactivation of CDK12 Delineates a Distinct Immunogenic Class of Advanced Prostate Cancer. in *Cell* Vol. 173 (2018).
299. Menghi, F. et al. The Tandem Duplicator Phenotype Is a Prevalent Genome-Wide Cancer Configuration Driven by Distinct Gene Mutations. in *Cancer Cell* Vol. 34 (2018).
300. Rao, M. & Powers, S. Tandem Duplications May Supply the Missing Genetic Alterations in Many Triple-Negative Breast and Gynecological Cancers. in *Cancer Cell* Vol. 34 (2018).
301. Rimán, T. et al. Risk factors for invasive epithelial ovarian cancer: Results from a Swedish case-control study. in *American Journal of Epidemiology* Vol. 156 (2002).
302. Kuchenbaecker, K.B. et al. Risks of breast, ovarian, and contralateral breast cancer for BRCA1 and BRCA2 mutation carriers. in *JAMA - Journal of the American Medical Association* Vol. 317 (2017).
303. Golmard, L. et al. Contribution of germline deleterious variants in the RAD51 paralogs to breast and ovarian cancers /631/208/68 /631/67/1347 article. in *European Journal of Human Genetics* Vol. 25 (2017).
304. Helder-Woolderink, J.M. et al. Ovarian cancer in Lynch syndrome; A systematic review. in *European Journal of Cancer* Vol. 55 (2016).
305. Shagimardanova, E. et al. CDK12: New breast and ovarian cancer predisposition gene in Tatar population? in *Annals of Oncology* Vol. 28 (2017).
306. Depristo, M.A. et al. A framework for variation discovery and genotyping using next-generation DNA sequencing data. in *Nature Genetics* Vol. 43 (2011).
307. Cibulskis, K. et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. in *Nature Biotechnology* 2013 31:3 Vol. 31 213-219 (Nature Publishing Group, 2013).
308. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. in *Bioinformatics* Vol. 27 (2011).
309. Beard, W.A., Horton, J.K., Prasad, R. & Wilson, S.H. Eukaryotic base excision repair: New approaches shine light on mechanism. in *Annual Review of Biochemistry* Vol. 88 (2019).
310. Jacobs, A.L. & Schär, P. DNA glycosylases: In DNA repair and beyond. in *Chromosoma* Vol. 121 (2012).
311. Cooper, D.N. & Youssoufian, H. The CpG dinucleotide and human genetic disease. in *Human Genetics* Vol. 78 (1988).
312. Duncan, B.K. & Miller, J.H. Mutagenic deamination of cytosine residues in DNA. in *Nature* Vol. 287 (1980).
313. Hendrich, B., Hardeland, U., Ng, H.H., Jiricny, J. & Bird, A. The thymine glycosylase MBD4 can bind to the product of deamination at methylated CpG sites. in *Nature* Vol. 401 (1999).
314. Rodrigues, M. et al. Outlier response to anti-PD1 in uveal melanoma reveals germline MBD4 mutations in hypermutated tumors. in *Nature Communications* Vol. 9 (2018).
315. Johansson, P.A. et al. Whole genome landscapes of uveal melanoma show an ultraviolet radiation signature in iris tumours. in *Nature Communications* Vol. 11 (2020).
316. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. (2012).
317. Harbour, J.W. et al. Frequent mutation of BAP1 in metastasizing uveal melanomas. in *Science* Vol. 330 (2010).
318. Gupta, M.P. et al. Clinical characteristics of uveal melanoma in patients with germline BAP1 mutations. in *JAMA Ophthalmology* Vol. 133 (2015).
319. Sansom, O.J., Bishop, S.M., Bird, A. & Clarke, A.R. MBD4 deficiency does not increase mutation or accelerate tumorigenesis in mice lacking MMR. in *Oncogene* Vol. 23 (2004).
320. Sanders, M.A. et al. MBD4 guards against methylation damage and germ line deficiency predisposes to clonal hematopoiesis and early-onset AML. in *Blood* Vol. 132 (2018).

321. Davies, H.R. et al. Epigenetic dysregulation underpins tumorigenesis in a cutaneous tumor syndrome. in *bioRxiv* 687459 (Cold Spring Harbor Laboratory, 2019).
322. Tanakaya, K. et al. A germline MBD4 mutation was identified in a patient with colorectal oligopolyposis and early-onset cancer: A case report. in *Oncology Reports* Vol. 42 (2019).
323. Waszak, S.M. et al. Germline determinants of the somatic mutation landscape in 2,642 cancer genomes. in *bioRxiv* (2017).
324. Brovkina, O.I. et al. The ethnic-specific spectrum of germline nucleotide variants in DNA damage response and repair genes in hereditary breast and ovarian cancer patients of Tatar descent. in *Frontiers in Oncology* Vol. 8 (2018).
325. Bogdanova, N.V. et al. A splice site variant of CDK12 and breast cancer in three Eurasian populations. in *Frontiers in Oncology* Vol. 9 (2019).
326. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. in *Nature* Vol. 536 (2016).
327. Mobuchon, L. et al. A gwas in uveal melanoma identifies risk polymorphisms in the *clptm1l* locus. in *npj Genomic Medicine* Vol. 2 (2017).