



# THÈSE

**En vue de l'obtention du  
DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE  
Délivré par l'Université Toulouse 3 - Paul Sabatier**

---

**Présentée et soutenue par  
Vincent ROCHER**

Le 23 novembre 2021

**Analyse de données et modèle pour l'étude de la chromatine, des  
G-quadruplexes et de la réparation de l'ADN**

---

Ecole doctorale : **BSB - Biologie, Santé, Biotechnologies**

Spécialité : **BIO-INFORMATIQUE, GENOMIQUE ET BIOLOGIE DES SYSTEMES**

Unité de recherche :  
**MCD - Molecular, Cellular and Developmental Biology Unit**

Thèse dirigée par  
**Raphael MOURAD**

Jury

M. Cédric Vaillant, Rapporteur  
M. Julien Mozziconacci, Rapporteur  
M. Benoît Ballester, Examinateur  
Mme Nathalie Vialaneix, Examinatrice  
M. Raphael MOURAD, Directeur de thèse  
Mme Gaëlle Legube, Co-encadrante de thèse

*“Je n’ai fait celle-ci plus longue que parce  
que je n’ai pas eu le loisir de la faire plus  
courte.”*

BLAISE PASCAL  
*LES PROVINCIALES, LETTRE 16*

## Remerciements

À l'heure où j'écris ces lignes, il ne reste plus que trois semaines avant ma soutenance, et je me dis que j'aurais sûrement dû écrire cette partie bien plus tôt. Ces travaux n'auraient cependant pas pu aboutir sans l'aide et le soutien d'un grand nombre de personnes, qu'il faut absolument remercier.

Je débute donc par ma cheffe, Gaëlle Legube, car c'est par toi que tout a commencé. Tu m'as fait confiance dès le début, et tu m'as engagé pour deux ans, en apprentissage, dans ton équipe incroyable, et cela fait six ans maintenant, à croire que je ne veux pas partir. Tu m'as apportée tellement de choses que je ne peux tout énumérer, mais je peux au moins te remercier pour la façon très humaine dont tu as su m'encadrer, et la grande qualité des projets scientifiques auxquels j'ai participé, grâce à toi.

Je ne peux parler de mon arrivée au labo sans remercier Marion Aguirrebengoa, qui m'a encadrée pendant mon apprentissage, et qui a su m'apporter toutes les compétences nécessaires pour que je sois fier de mon travail d'ingénieur. Merci à toi, et j'ai bien hâte de recommencer à travailler avec toi !

Ensuite, bien sûr, je remercie mon directeur de thèse, Raphaël Mourad, pour l'extrême enthousiasme dont il a pu faire preuve pendant ma thèse, pour toutes les choses qu'il m'a apprises, et les opportunités de m'améliorer. Je peux dire sans hésiter que si tu n'avais pas été là, je n'aurais pas fait de thèse !

Je remercie également l'équipe Legube tout entière : merci à Thomas, pour tes précieux conseils, ton encadrement pendant mon apprentissage, et ton savoir immense qui m'a permis de comprendre beaucoup de choses. Merci à Coline, pour la qualité exceptionnelle de ton travail, et la facilité avec laquelle on peut travailler avec toi. Merci Manu, pour ton enthousiasme et tes conseils qui remontent le moral ; j'envie ta nouvelle doctorante, qui a vraiment de la chance de tomber sur toi ! Merci au reste de l'équipe, Aline, Nadine, Anne-Laure, Emma, Sarah, Florian, Benjamin, Ikrame et Nathalie, ainsi qu'à ceux qui ne sont plus là : Aude, Béa, François et Sarah. Bon courage à ceux qui commencent leur doctorat : Emanuelle et Sébastien !

Je remercie les membres de mon jury d'avoir accepté l'invitation d'assister et d'évaluer mon travail de thèse. Je remercie particulièrement mes rapporteurs : Cédric Vaillant et Julien Mozziconacci pour leur retour sur mon manuscrit et leurs avis positifs. Merci à mon examinateur Benoît Ballester d'accepter de revenir au CBI un mois à peine après y être venu, et enfin merci énormément à Nathalie Vialaneix d'avoir accepté avec enthousiasme de faire partie de mon jury, malgré les quatre autres jurys de thèse prévus.

Je remercie énormément Marianne, pour m'avoir soutenu, suivi, conseillé, pendant ces trois années de thèse. Mon aventure s'arrête là, et tu dois continuer, mais je ne serai pas loin, et j'espère pouvoir t'apporter le même soutien, quand ton temps sera venu de rédiger.

Je remercie maintenant mes parents, pour leur soutien inconditionnel tout au long de mes études. Merci maman, merci papa, pour m'avoir toujours poussé en avant, à continuer, persévérer, malgré la baisse de motivation. Il est plus que certain que je n'aurais pas fait la moitié du chemin sans vous.

Enfin, merci Eva, car c'est toi qui me soutiens au quotidien, qui me motive, qui me rassure dans mes moments de stress et de fatigue. Merci de m'avoir suivi à Toulouse, de partager ma vie et mon bonheur, merci pour le cadeau que tu me fais, presque au moment où je termine ma thèse !

Pour finir, merci à toi, même si je ne te connais pas encore. Tu me donnes déjà des ailes, et j'ai tellement hâte de te rencontrer ! Peut-être qu'un jour, tu liras ces lignes, et tu seras fier de moi.



# Table des matières

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	L'organisation de l'ADN dans le noyau . . . . .	1
1.1.1	La fibre de chromatine . . . . .	1
1.1.2	Les gènes, la partie transcrète du génome . . . . .	4
1.1.3	Les G-quadruplexes (G4) . . . . .	6
1.1.4	La structure 3D de la chromatine . . . . .	8
1.2	Les cassures double-brin de l'ADN ( <i>DSB</i> ) et leur réparation . . . . .	19
1.2.1	Signalisation des DSB . . . . .	20
1.2.2	La réparation des DSB . . . . .	23
1.2.3	Choix de la voie de réparation . . . . .	24
1.2.4	La réparation dans les gènes actifs . . . . .	25
1.2.5	L'inhibition de la transcription par les DSB . . . . .	25
1.2.6	Le rôle du 3D dans la réparation des DSB . . . . .	25
1.3	La génomique, ou l'étude de l'ensemble du génome . . . . .	30
1.3.1	Le séquençage à haut débit (Next Generation Sequencing, NGS) . . . . .	30
1.3.2	Cartographie des sites de liaison protéine-ADN (ChIP-Seq) . . . . .	39
1.3.3	Analyse de motifs de site de liaison de protéines à l'ADN . . . . .	41
1.3.4	Cartographie de l'accessibilité de la chromatine (ATAC-Seq) . . . . .	46
1.3.5	Quantification de l'expression des gènes (RNA-seq) . . . . .	49
1.3.6	Structure 3D de la chromatine ( <i>Chromosome Conformation Capture, 3C</i> ) . . . . .	56
1.3.7	Bases de données génomiques . . . . .	66
1.4	Prédiction de données génomiques . . . . .	68
1.4.1	Apprentissage Automatique ( <i>Machine Learning</i> ) . . . . .	68
1.4.2	Évaluation d'un modèle de <i>Machine Learning</i> . . . . .	73
1.4.3	Les réseaux de neurones artificiels ( <i>Neural Network / Deep Learning</i> ) . . . . .	75
<b>2</b>	<b>Résultats</b>	<b>85</b>
2.1	DeepG4 : une approche <i>Deep Learning</i> pour prédire les régions riches en G-quadruplex actifs spécifiques à un type cellulaire . . . . .	85
2.1.1	Performances du premier modèle ADN sans accessibilité . . . . .	102
2.1.2	Discussion . . . . .	106
2.2	Réparation de l'ADN et structure tri-dimensionnelle du génome . . . . .	110
2.2.1	Le modèle cellulaire DIvA . . . . .	110
2.2.2	Introduction . . . . .	111
2.2.3	La formation des foyers de réparation autour des cassures double-brin de l'ADN dépend du mécanisme de <i>loop extrusion</i> . . . . .	112
2.2.4	La formation d'un nouveau compartiment de la chromatine dépendant d'ATM régule la réponse aux cassures double-brin de l'ADN et la biogénèse des translocations . . . . .	135
2.2.5	Discussion . . . . .	203
<b>Appendix</b>		<b>211</b>
La cartographie complète des modifications d'histones au niveau des cassures double-brin de l'ADN déchiffre les signatures des voies de réparation . . . . .	211	

---

# Introduction

## 1.1 L'organisation de l'ADN dans le noyau

Au sein des cellules des êtres vivants, le support de l'information permettant aux cellules d'assurer leur fonction est encodé par l'ADN (Acide Désoxyribonucléique). L'ADN est une macromolécule composée généralement de deux brins, tous les deux orientés dans des sens opposés : le brin sens (5' vers 3', ou *forward*) et le brin anti-sens (3' vers 5', ou *reverse*) qui forment une double-hélice (WATSON et CRICK 1953). Cette macro-molécule est composée d'une multitude d'unités de bases, les nucléotides. Les nucléotides sont présents sous quatre formes dans l'ADN : l'Adénine (A), la Cytosine (C), la Guanine (G) et la Thymine (T), reliés entre eux par des liaisons covalentes. L'ADN étant double-brin, ces nucléotides sont complémentaires et forment des paires : le A s'associe au T, et le C au G.

Le génome correspond à l'ensemble du matériel génétique encodé par l'ADN apportant l'information nécessaire au développement d'une cellule et à son bon fonctionnement. Chez l'homme, le génome est composé de plus de 3 milliards de nucléotides ou paires de bases (ou *base pair*, bp), réparti principalement sur 23 paires de chromosomes. Une partie du génome encode les protéines, ce sont les gènes, une autre partie, non-codante, contient des éléments régulateurs et fortement répétés de l'ADN. En outre, l'ADN n'est pas nu dans le noyau, mais est associé à des protéines lui permettant de s'organiser et de se compacter, formant un complexe nucléoprotéique qu'on appelle la chromatine.

### 1.1.1 La fibre de chromatine

#### 1.1.1.1 Le nucléosome

L'unité de base de la chromatine est le nucléosome (voir Figure 1.1). Chaque nucléosome est constitué d'ADN d'une taille d'environ 145 bp qui s'enroule autour d'un octamère de protéines, les histones. Chaque histone est formée de deux parties, le “core”, faisant la liaison avec l'ADN et les autres histones, et les extrémités, qui peuvent subir des modifications post-traductionnelles. Les nucléosomes sont composés de deux dimères d'histones H3-H4 et deux dimères H2A-H2B. De cette façon, l'ADN est 6 fois plus compact qu'un fragment d'ADN nu d'une même longueur.

Les nucléosomes sont connectés entre eux via un segment d'ADN associé à une histone de liaison H1, formant le chromatosome. Cette histone de liaison a un rôle important dans la structure de la chromatine,

et stabilise le nucléosome (FYODOROV et al. 2018). La fibre de chromatine ainsi composée en “collier de perle” forme le premier niveau de compaction de l’ADN.

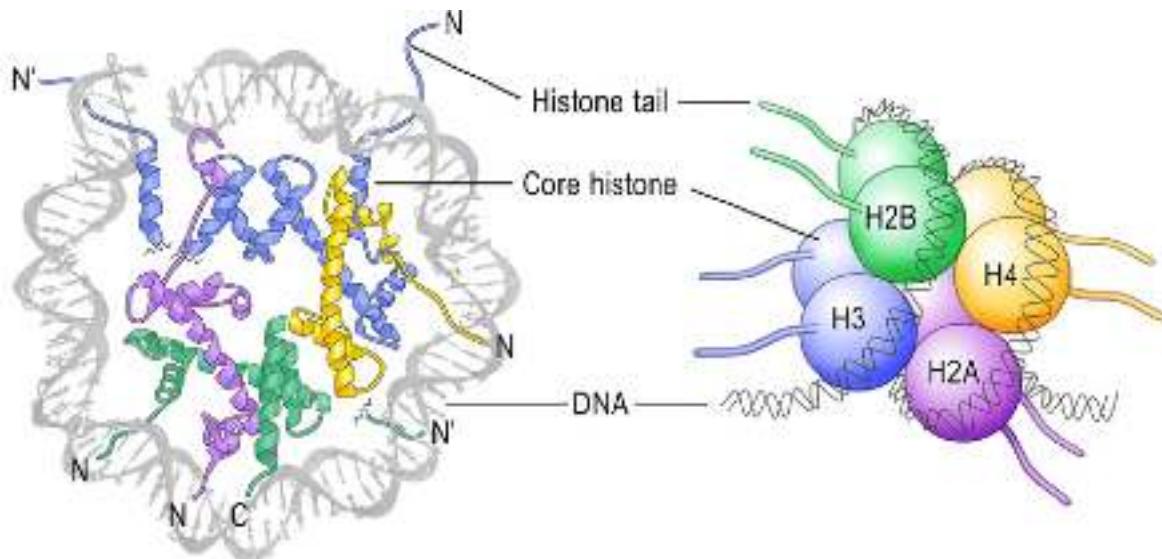


FIGURE 1.1 – Représentation du nucléosome associé à l’ADN. Le nucléosome est composé d’un octamère des histones H2A, H2B, H3 et H4, et d’une molécule d’ADN en double hélice de ~ 145 bp. Adapté de (GRÄFF et MANSUY 2008).

### 1.1.1.2 Niveaux supérieurs de compaction

Le premier niveau de compaction en “collier de perle” forme une fibre de chromatine d’un diamètre d’environ 10 nm. Cette fibre se replie également sur elle-même, lui permettant d’atteindre un niveau de compaction supérieur, d’un diamètre d’environ 30 nm. La formation de cette structure serait assurée par l’association entre nucléosomes via l’intervention de l’histone H1. Cette structure plus large étant difficilement observable, sa présence *in-vivo* est toujours sujet à débat (FYODOROV et al. 2018 ; TREMETHICK 2007).

Les deux états de compactations supérieurs bien connus de la chromatine sont l’hétérochromatine et l’euchromatine. L’hétérochromatine correspond à la chromatine dense ou permissive. En son sein, les gènes sont réprimés par le fait que les nucléosomes sont fortement compactés et repliés sur eux-mêmes. Elle est composée des régions télosomiques et centromériques, en plus des gènes inactifs. L’euchromatine, en revanche, correspond à l’état accessible de la chromatine. En son sein, les nucléosomes sont espacés, et les gènes sont plus facilement accessibles à la machinerie transcriptionnelle (ALLIS et JENUWEIN 2016).

La régulation de ces états de la chromatine dépend de la compaction, plus ou moins dense des nucléosomes, et par leur capacité à interagir entre-eux. Cette régulation de la structure de la chromatine en deux états est assurée par des modifications de l’ADN et des histones qu’on appelle modifications post-traductionnelles ou marques épigénétiques.

### 1.1.1.3 La régulation de la chromatine par les marques épigénétiques

La structure de la chromatine peut être régulée de façon dynamique par diverses modifications épigénétiques, comme la méthylation de l'ADN, les modifications d'histones ou bien les variants d'histones.

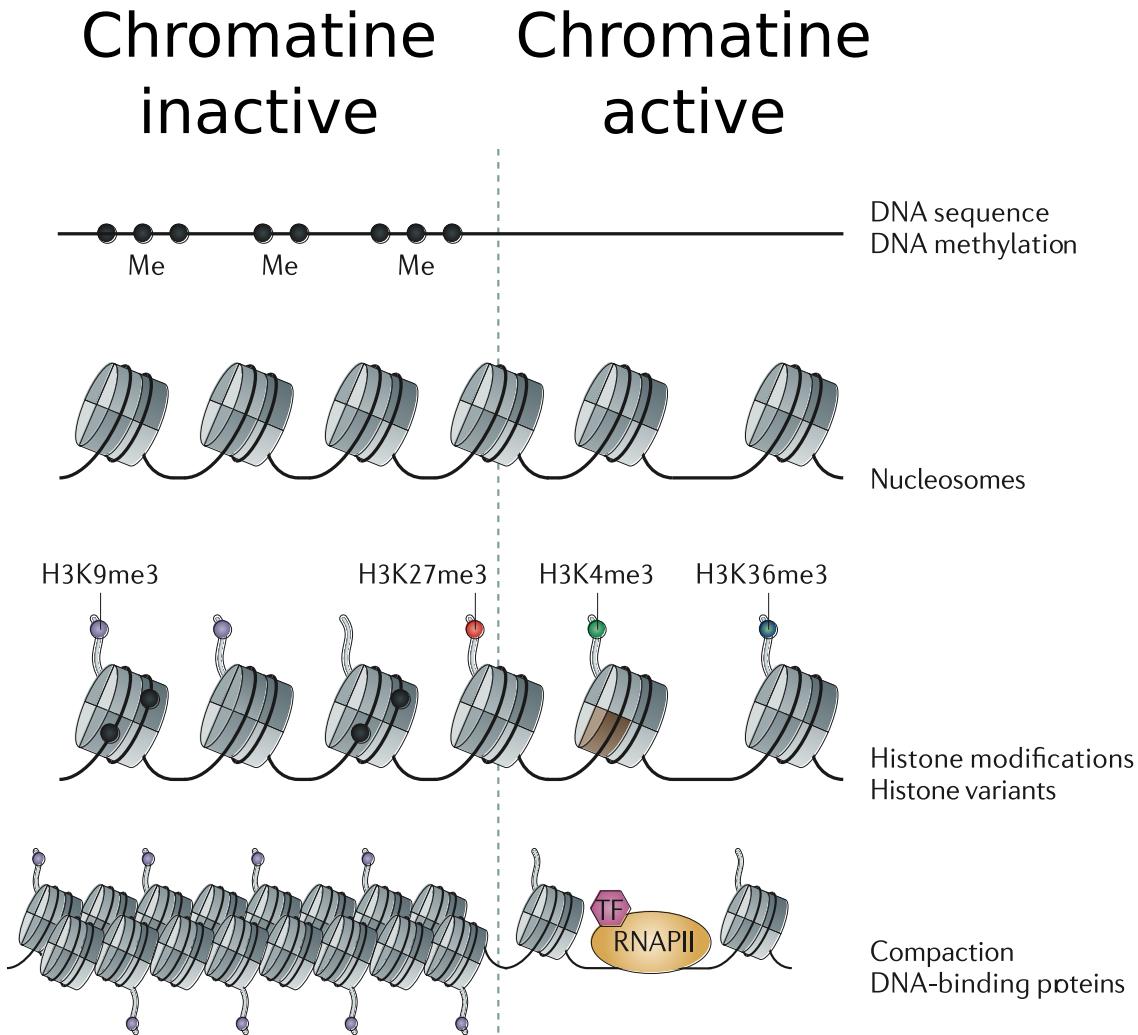


FIGURE 1.2 – Effet des modifications post-traductionnelles sur l'état de compaction de la chromatine. L'ADN méthylique est associé à une chromatine dense. Le premier état de compaction est formé de nucléosomes, qui peuvent être modifiés par des marques épigénétiques. Ces modifications post-traductionnelles vont définir l'état local de compaction de la chromatine. H3K36me3 et H3K4me3 sont des modifications post-traductionnelles associées à la chromatine active ou euchromatine. L'hétérochromatine elle est définie par des marques comme H3K9me3 et H3K27me3. Adapté de (ZHOU, GOREN et BERNSTEIN 2011)

**1.1.1.3.1 La méthylation de l'ADN** Chez les mammifères, la méthylation de l'ADN correspond à l'ajout d'un groupement méthyl  $CH_3$  sur les Cytosines qui précèdent une Guanine, on parle alors de di-nucléotide CpG. Un grand nombre de gènes sont enrichis en CpG, formant des îlots CpG, et leur niveau d'expression corrèle avec le nombre d'îlots observés. Un fort niveau de méthylation de l'ADN au

## 1.1. L'ORGANISATION DE L'ADN DANS LE NOYAU

---

niveau d'un gène correspond souvent à une région inactive. À l'inverse, les régions accessibles ou actives sont moins enrichies en CpG (voir Figure 1.2) (ZHOU, GOREN et BERNSTEIN 2011).

**1.1.1.3.2 Les modifications d'histones** Les histones possèdent une queue N-terminale composée d'une trentaine d'acides aminés accessibles aux modifications post-traductionnelles (voir Figure 1.1). Il existe différentes modifications de cette queue N-terminale, comme l'acétylation, la méthylation, la phosphorylation ou encore l'ubiquitination, qui peuvent agir sur les différents résidus présents. Ces modifications sont notées en fonction du résidu impacté, de leur type, et leur effet peut grandement varier sur l'état local de compaction de la chromatine. Par exemple H3K4me3, qui indique une tri-méthylation de la Lysine en 4ème position de la queue N-terminale de l'histone H3. Elle est un marqueur de l'euchromatine, enrichie au niveau des régions promotrices des gènes actifs. À l'inverse, la tri-méthylation de la 9ème lysine de l'histone H3, elle aussi enrichie au niveau des régions promotrices est une marque associée à la répression de la transcription, et donc de l'hétérochromatine (voir Figure 1.2) (SIMS III, NISHIOKA et REINBERG 2003).

**1.1.1.3.3 Les variants d'histones** Les variants d'histones peuvent aussi agir sur l'état local de la chromatine lorsqu'ils se substituent à leur forme canonique. Ils sont codés par des gènes différents, et peuvent s'incorporer tout au long du cycle cellulaire. Leur action sur l'état de la chromatine peut être directe ou indirecte, et il existe des variants pour chaque type d'histone canoniques (MARTIRE et BANASZYNSKI 2020). Par exemple, le variant d'histone H2AX, lorsqu'il est phosphorylé au niveau de sa queue C-terminale, possède un rôle majeur dans la réparation de l'ADN, notamment dans la signalisation des dommages à l'ADN (YUAN, ADAMSKI et J. CHEN 2010).

L'état d'accessibilité locale de la chromatine, ainsi que ses modifications épigénétiques sont donc représentatives de son niveau de compaction, et de son état accessible (euchromatine) ou au contraire dense (hétérochromatine). En cartographiant l'état d'accessibilité de l'ADN ou différentes protéines et modifications post-traductionnelles, on peut alors mesurer cet état chromatinien en appliquant des méthodes omiques (voir parties 1.3, 1.3.2 et 1.3.4), ce qui permet d'analyser la structure de la chromatine à l'échelle du génome entier.

## 1.1.2 Les gènes, la partie transcrive du génome

Dans le génome, il existe différentes compositions d'ADN qui encodent des fonctions particulières. On peut annoter ces régions selon leurs fonctions, celles ayant un rôle transcriptionnel sont les gènes. La transcription est le procédé permettant de recopier l'ADN sous forme d'ARN messager, le transcript, qui sera ensuite traduit en protéines en dehors du noyau, par les ribosomes.

Les gènes des eucaryotes sont majoritairement structurés par deux types de séquences : les exons, qui contiennent des séquences qui formeront le transcript mature, et les introns, des séquences séparant les exons entre eux, qui seront transcris (présents dans les transcrits pré-matures) mais non traduites (voir Figure 1.3).

Les exons sont subdivisés en deux parties, une partie traduite (*CDS*, pour *Coding DNA Sequence*), qui contient la séquence de la protéine, et une partie régulatrice, la 5' et 3' UTR non traduites mais ayant des propriétés régulatrices de la transcription du gène.

## 1.1. L'ORGANISATION DE L'ADN DANS LE NOYAU

Les introns sont délimités en 5' par un site donneur et en 3' par un site accepteur, qui permettent leur reconnaissance et leur excision par le mécanisme de l'épissage. Ce processus intervient pendant la transcription du gène et permet à celui-ci de coder pour différentes protéines. Pendant ce procédé, certains exons seront conservés et d'autres exclus du produit final, formant l'ARN mature. À partir d'un seul gène, et donc d'un seul ARN pré-mature, l'épissage alternatif permet de produire une multitude de transcrits possibles. Ceci permet par exemple au génome humain de contenir l'information pour encoder plus de 90 000 protéines avec seulement ~ 20 000 gènes.

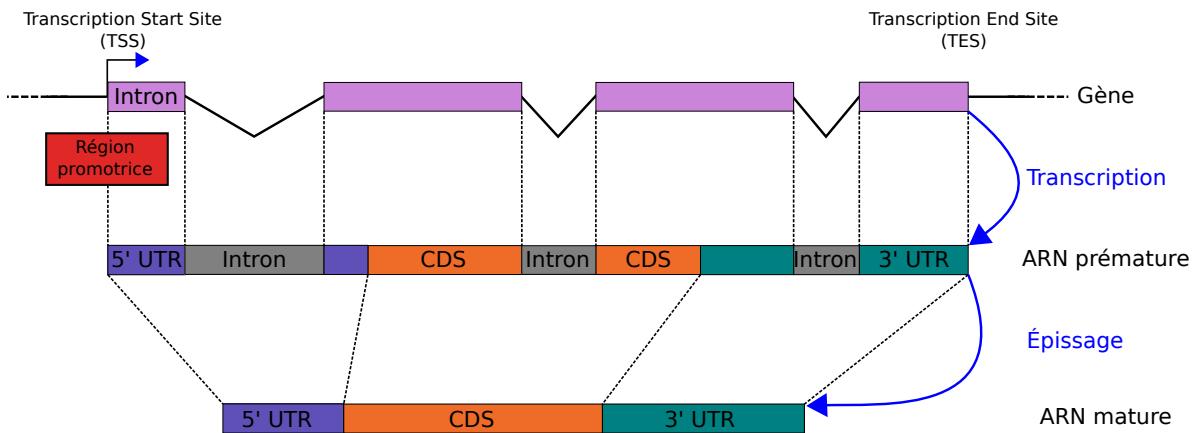


FIGURE 1.3 – Organisation et produits de la transcription et de l'épissage pour un gène. La région génique est divisée en différentes sous régions : le promoteur (en rouge), est une région réduite autour du TSS qui contient des sites de fixation ainsi que différents facteurs de transcription. Le gène est ensuite subdivisé en deux parties : les exons et les introns. Les exons contiennent une partie traduite (CDS) et une partie régulatrice non traduite (5' et 3' UTR).

### 1.1.2.1 La transcription des gènes

Le mécanisme de transcription permet de créer un ARN messager, contenant l'information du gène, pour ensuite être traduit en protéine. Il fait intervenir une ARN polymérase, la *RNA Polymerase II* (*PolII*) qui va se fixer au niveau de la région promotrice du gène (voir Figure 1.3). La transcription sera ensuite initiée à partir du site d'initiation de la transcription ou *Transcription Start Site (TSS)* grâce à l'intervention de protéines, les facteurs de transcription, qui reconnaissent des motifs sur l'ADN dans la région promotrice et qui vont recruter et activer la *PolII* (Patrick CRAMER 2019) (voir Figure 1.4). Après initiation de la transcription au niveau du *TSS*, la *PolII* forme un complexe avec des facteurs d'elongation et l'ADN simple brin permettant ainsi la continuité de la transcription. Celle-ci continue jusqu'au site de fin de transcription ou *Transcription End Site (TES)*. Une queue polyA est ajoutée en 3' du transcript pendant la polyadénylation afin de le stabiliser, empêcher sa dégradation et permettre sa migration en dehors du noyau (SLOMOVIC et al. 2010). Enfin, l'épissage a lieu après la transcription complète de l'ADN en ARN, lorsque les sites donneur et accepteur d'un intron forment une boucle qui sera épissée par le spliceosome (Patrick CRAMER 2019).

Les différents mécanismes de transcription, l'expression des gènes et l'épissage peuvent être étudiés par des méthodes omiques qui permettent leur cartographie (voir partie 1.3.2) et leur quantification (voir partie 1.3.5).

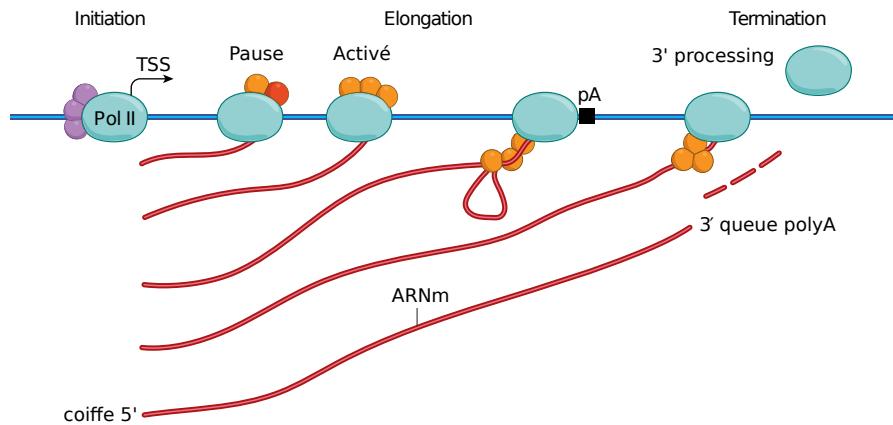


FIGURE 1.4 – Mécanisme de transcription par l'ARN Polymérase II. Les facteurs de transcription (violet) au niveau du promoteur permettent l'initiation de la transcription, mais régulent aussi (jaune et rouge) l'activité de la PolII pendant l'élongation du transcrit. À la fin de la transcription, une queue poly A est ajoutée en 3' du transcrit au niveau du site de polyadénylation (**pA**) pour le stabiliser et empêcher sa dégradation. Adapté de (Patrick CRAMER 2019)

### 1.1.2.2 La région promotrice

L'expression d'un gène est contrôlée par un ensemble de processus complexes, dont une grande partie est localisée dans la région promotrice ou promoteur du gène. Un gène pouvant contenir plusieurs *TSS*, la région promotrice est donc variable et sa composition joue sur l'épissage alternatif (KOLATHUR 2021 ; Paula CRAMER et al. 1997). C'est dans la région promotrice que la *PolII* sera recrutée et la transcription initiée. Elle contient de nombreuses séquences ou motifs (voir partie 1.3.2) reconnus par des facteurs de transcription, qui peuvent réguler la transcription du gène.

En outre, de nombreuses techniques omiques (voir partie 1.3) ont montré que la région promotrice d'un gène actif était particulièrement accessible, avec une densité moindre de nucléosomes (voir partie 1.3.4), ainsi que des marques d'histones favorisant l'expression (voir partie 1.1.1.3.2).

### 1.1.3 Les G-quadruplexes (G4)

Au sein du noyau, l'ADN en double-hélice est majoritairement sous la forme B-DNA (WATSON et CRICK 1953). Cependant, l'ADN en double-hélice peut également prendre des formes alternatives, comme le Z-DNA, et même adopter des structures différentes, tel que l'ADN cruciforme, les triplexes ou bien les G-quadruplexes (MIRKIN et al. 2008). Ces structures contribuent à l'instabilité du génome, et leur distribution dans le génome co-localise souvent avec des dommages associés à des maladies génétiques (ZHAO et al. 2010). Parmi ces structures, les G-quadruplexes, ou G4 sont connues pour être enrichies dans les télomères et dans les régions promotrices des gènes (LAGO et al. 2021), ainsi que dans les origines de réplications (PROROK et al. 2019).

#### 1.1.3.1 La structure du G4

Les G-quadruplexes ou G4 sont des structures formées à partir d'ADN ou d'ARN riches en Guanines (G). La structure du G4 contient une série de quatre G ou plus qui s'enchaînent pour former des

tétrades, ou *G-tetrads*. Ces *G-tetrads* sont liés entre eux par des liaisons hydrogènes (voir Figure 1.5 A), et les séquences entre les *G-tetrads* forment des boucles. La stabilité d'un G4 dépend de l'origine du brin (ADN ou ARN), de la taille des boucles et de leur composition, ainsi que de la composition des *G-tetrads*. L'environnement moléculaire et la composition chimique du milieu possède également une influence sur la stabilité de la structure. Par exemple, des ions potassium ( $K^+$ ) peuvent se positionner entre les *G-tetrads*, et stabiliser la structure (LIGHTFOOT et al. 2019). Les G4 sont formés via différentes configurations, qui dépendent de la façon dont les boucles s'associent pour former la structure (voir Figure 1.5), rendant sa prédiction à partir de la séquence compliquée.

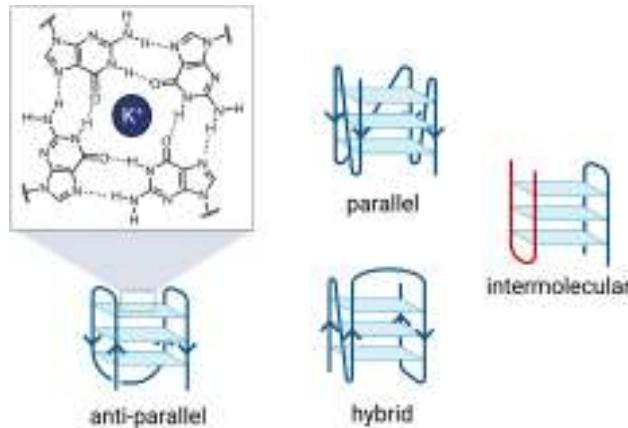


FIGURE 1.5 – Structure d'un G-quadruplex (G4) et ses différentes formations possibles à partir du brin d'ADN. Adapté de (J. ROBINSON et al. 2021).

### 1.1.3.2 Identification et cartographie des G4

Les premières études ayant mis en évidence que l'ADN pouvait se replier pour former des G4 sont des études biophysiques (GELLERT, LIPSETT et DAVIES 1962 ; HOWARD, FRAZIER et MILES 1977). Ces études ont permis d'établir que de nombreuses séquences d'ADN et d'ARN comportant un enchaînement de G pouvaient former un G4, et la mise au point d'une séquence consensus a été faite à partir de ces séquences. Cette séquence consensus ou motif, permet de prédire si une séquence peut former un G4, via l'utilisation d'une expression régulière ( $G \geq 3N_xG \geq 3N_yG \geq 3N_zG \geq 3$ ).

Les G4 n'ont cependant été observés *in vivo* que dans les années 2000, au niveau des télomères (DUQUETTE et al. 2004), mais aussi au niveau de la région promotrice d'un oncogène, *MYC*. Plus intéressant encore, il a été montré que la stabilisation des G4 dans cette région permet de réguler négativement l'expression de *MYC*, proposant un rôle actif de ces structures, de la même façon qu'un élément régulateur (SIDDQUI-JAIN et al. 2002).

C'est suite au séquençage du génome humain (I. H. G. S. CONSORTIUM 2004) que des prédictions ont été faites à l'échelle du génome pour déterminer la position de sites potentiels pouvant former des G4 (*Putative Quadruplex Sequences*, ou *PQS*), en recherchant la présence du motif canonique dans la séquence du génome (HUPPERT et BALASUBRAMANIAN 2005). Ces méthodes ont permis de montrer que les *PQS*, enrichis au niveau des promoteurs, auraient un rôle actif dans la régulation de l'expression des gènes (HUPPERT et BALASUBRAMANIAN 2007). Les méthodes basées uniquement sur une recherche d'une expression régulière manquent de flexibilité lorsqu'une séquence diverge du motif canonique. De

## 1.1. L'ORGANISATION DE L'ADN DANS LE NOYAU

---

nouvelles méthodes ont été développées, basées sur des algorithmes experts, comprenant un ensemble de règles dictées par les propriétés biophysiques pour calculer un score déterminant la propension de la séquence à former un G4 (HON et al. 2017 ; BEDRAT, LACROIX et MERGNY 2016).

Des approches expérimentales utilisant des techniques omiques (voir partie 1.3) ont ensuite été mises au point afin de cartographier les G4 sur le génome. Une première méthode a permis de cartographier les G4 dans des conditions *in vitro*, en mesurant la capacité physique de l'ADN mis à nu à former ces structures (CHAMBERS et al. 2015). Une seconde méthode, via l'utilisation d'un anticorps spécifique (voir partie 1.3.2) a permis de cartographier *in vivo* les G4 (HÄNSEL-HERTSCH, SPIEGEL et al. 2018 ; K.-w. ZHENG et al. 2020).

L'utilisation d'un anticorps spécifique a permis de révolutionner l'analyse des G4 *in vivo*, mettant en évidence leurs rôles actifs dans la structure de la chromatine et sa conformation 3D, dans la stabilité et l'endommagement de l'ADN (J. ROBINSON et al. 2021) (voir Figure 1.5), mais aussi dans la régulation de l'expression des gènes (LAGO et al. 2021). De plus, la cartographie des G4 *in vivo* permet d'avoir accès aux séquences d'ADN correspondant à des sites probables de formation de G4, et donc de disposer de nouvelles sources d'informations pour leur prédiction.

Des modèles de *Machine Learning* ou *Deep Learning* (voir parties 1.4.1 et 1.4.3) peuvent être utilisés pour prédire les sites de liaisons de protéines en utilisant l'ADN (ALIPANAHİ et al. 2015), mais nécessitent de nombreuses données d'entraînement, ce que proposent les méthodes omiques (voir partie 1.3). Des méthodes similaires ont déjà été utilisées pour la prédiction des G-quadruplexes, proposant de bons résultats (SAHAKYAN et al. 2017 ; KLIMENTOVA et al. 2020). Cependant, soit ces modèles se basent encore sur la prédiction des *PQS* (SAHAKYAN et al. 2017), ou alors utilisent les données cartographiant les G4 *in vitro*. Des méthodes permettant la prédiction de régions composées de G4 observés *in vivo* (G4 actifs) selon le type cellulaire sont encore à mettre au point.

### 1.1.4 La structure 3D de la chromatine

La fibre de chromatine est une molécule complexe pouvant adopter différents niveaux de compactations selon les régions du génomes. Elle s'organise en différents niveaux qui régulent les fonctions cellulaires comme l'expression des gènes (BANIGAN et MIRNY 2020), ou la réparation de l'ADN (YATSKEVICH, RHODES et NASMYTH 2019).

#### 1.1.4.1 Les méthodes d'étude de l'organisation 3D de la chromatine

**1.1.4.1.1 Le *DNA-FISH*** L'organisation de l'ADN dans le noyau peut être étudiée grâce à des méthodes de microscopie, comme le *DNA-FISH* (Fluorescence In Situ Hybridization). Cette méthode se base sur l'hybridation de sondes fluorescentes qui ciblent des *loci* d'ADN spécifiques, qui sont visibles par microscopie. Une variante du *FISH*, l'*Oligopaint*, permet de visualiser des régions à plus forte résolution, de l'ordre du *kilobase* (BELIVEAU et al. 2017). Ces méthodes permettent de visualiser la proximité spatiale entre des *loci* d'intérêt, et permettent également d'étudier l'organisation des chromosomes dans le noyau à l'échelle de cellules uniques et de manière dynamique (SZCZEPIŃSKA, RUSEK et PLEWCZYNSKI 2019).

**1.1.4.1.2 La capture de conformation de la chromatine** Afin d'étudier l'organisation 3D de l'ADN dans le noyau à l'échelle du génome entier, des méthodes dites *3C*, pour *Chromosome Conformation Capture*, ont été développées. Elles partent du principe que plus deux *loci* sont proches dans l'espace, plus leur fréquence d'interaction dans la population cellulaire sera importante. Les méthodes de *3C* permettent de mesurer la fréquence de contacts entre *loci* à haut débit dans une population cellulaire et sont ainsi complémentaires des méthodes par imagerie.

Le *3C* (N. NAUMOVA et al. 2012) permet d'étudier les interactions entre deux *loci* d'ADN par *PCR* quantitative, et n'est donc pas une méthode omique, ce qui ne permet pas d'étudier l'organisation 3D dans un contexte pangénomique (voir partie 1.3). C'est pourquoi d'autres méthodes, basées sur le *3C*, mais utilisant le séquençage à haut débit ont vu le jour. Le *4C-seq* (*Circular 3C*) (WERKEN, VREE et al. 2012) et le *Hi-C* (*High-throughput chromosome conformation capture*) (LIEBERMAN-AIDEN et al. 2009 ; S. S. RAO et al. 2014) sont deux méthodes permettant d'étudier l'interaction d'un *locus* avec le reste du génome (*4C-seq*) à très haute résolution, ou d'étudier l'interaction de tous les *loci* du génome, à moindre résolution (*Hi-C*). Ces méthodes sont décrites en détail dans la partie 1.3.6.

La visualisation des interactions tri-dimensionnelles (3D) de la chromatine par *Hi-C* produit des matrices d'interaction, où chaque cellule de la matrice correspond à une paire de *loci*, et sa valeur à la fréquence d'interaction de ces deux *loci* à différentes résolutions possibles (voir Figure 1.7 B et C). Ces matrices d'interaction permettent d'étudier les caractéristiques biophysiques de la chromatine. Notamment que la fréquence d'interaction entre deux *loci* dépend de leur distance génomique selon une loi de puissance (voir Figure 1.6). Dans cet exemple, la fréquence d'interaction peut être modélisée par une loi de puissance ayant un exposant de  $-1$  ("effet" distance). Cet exposant peut varier en fonction du génome, il est par exemple différent entre la *Drosophila* et l'*Homme* (BARBIERI et al. 2012) et varie selon le chromosome et la phase du cycle cellulaire (Natalia NAUMOVA et al. 2013).

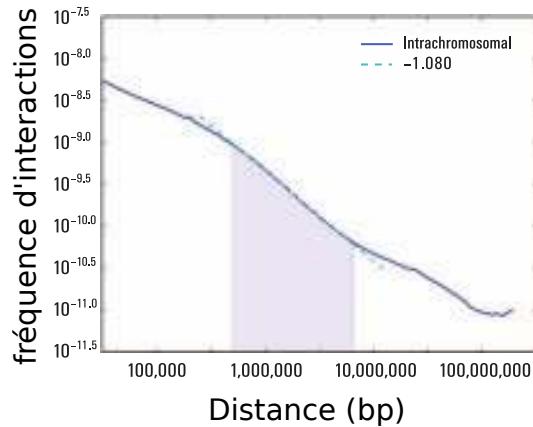


FIGURE 1.6 – La probabilité de contact en fonction de la distance génomique moyenne à travers le génome (bleu) suit une loi de puissance comprise entre 500k et 7Mb (région ombrée) avec une pente de  $-1.08$  (cyan). Adapté de (LIEBERMAN-AIDEN et al. 2009).

On peut observer cet effet distance grâce à la diagonale des données *Hi-C* (voir Figure 1.7 B et C) : l'interaction est forte à faible distance (très rouge), puis diminue à mesure que la distance augmente (pour passer au blanc). En outre, les *loci* interagissent préférentiellement avec d'autres *loci* du même chromosome (intra-chromosomique), plutôt qu'avec ceux du chromosome voisin (inter-chromosomique) (voir partie 1.1.4.6.1).

### 1.1.4.2 Les Domaines Topologiquement Associatifs (*TAD*)

Des expériences utilisant des données Hi-C (DIXON et al. 2012) (voir partie 1.3.6) ont mis en évidence des domaines 3D particuliers d'une taille de 100kb à 1Mb, les Domaines Topologiquement Associatifs, ou *Topologically Associating Domain* (*TAD*). Dans un *TAD*, deux *loci* ont une plus grande fréquence d'interaction qu'entre des *loci* de *TAD* adjacents. Ces interactions 3D formant les *TAD* sont facilement observables en Hi-C, car ils correspondent à des triangles ou des carrés (voir Figure 1.7 A). Les frontières entre les *TAD* sont déterminées par des méthodes bio-informatiques (voir partie 1.3.6.1.5, Figure 1.7 A) et l'identificaton des *TAD* a montré que 90% du génome de la souris étaient recouverts de *TAD*, fortement conservés chez les mammifères.

En outre, il a été montré que les frontières des *TAD* sont délimitées par des séquences insulatrices, qui bloquent les interactions longue distance (DIXON et al. 2012). Les *TAD* ont un rôle fonctionnel : ils favorisent les contacts longue distance entre des régions fonctionnelles, comme les contacts promoteurs/enhanceurs (RON et al. 2017), la recombinaison V(D)J à l'origine de la formation de nouvelles immunoglobulines (J. HU et al. 2015) ou bien la réplication de l'ADN (DEKKER et HEARD 2015).

En augmentant la profondeur de séquençage, et donc la résolution des données Hi-C, on a pu observer sur la carte d'interaction de nouveaux éléments structurels impossible à identifier à plus faible résolution. Par exemple, la première expérience de Hi-C (LIEBERMAN-AIDEN et al. 2009) ne permettait d'observer que des éléments structuraux ayant lieu à grande échelle, à une résolution de 100kb ou moins (voir Figure 1.7 B). Les cartes Hi-C plus résolutives (~1kb) (S. S. RAO et al. 2014) ont mis en évidence de nouveaux domaines plus petits que les *TAD*, des *sous-TAD*, remettant en question la définition de ces domaines 3D (voir Figure 1.7 C). Ces nouveaux domaines, imbriqués de manière hiérarchique dans de plus grands, impliquent que les interactions au sein de la population cellulaire ne sont donc pas homogènes, et donc proviennent de différents états possibles de la chromatine pour un même *locus*. Ceci suggère que la formation d'un *TAD* est issue d'un processus dynamique et variable selon les cellules (S. S. RAO et al. 2014).

### 1.1.4.3 La formation des TAD

La détection des *TAD* à partir d'expériences de Hi-C implique de localiser les frontières d'un domaine 3D. Avec une plus grande résolution, on détecte de nouvelles frontières alors impossibles à observer aux résolutions les plus faibles. Il a été montré que les sous-*TAD* étaient délimités par une boucle de chromatine (S. S. RAO et al. 2014). Toutes les frontières de ces sous-*TAD* sont associées soit à une transition de compartiment ou sous-compartiment (voir partie 1.1.4.5), ou à une boucle de chromatine.

On peut observer les boucles de chromatine sur les expériences de Hi-C à grande résolution, via les interactions entre régions insulatrices délimitant un sous-*TAD*. On peut visualiser ces interactions sous la forme d'un point rouge sur la carte Hi-C (voir Figure 1.7 C). Cette interaction expliquerait comment les *TAD* (ou sous-*TAD*) sont formés, via la création de boucles de chromatine.

**1.1.4.3.1 La protéine insulatrice CTCF** Les séquences insulatrices sont reconnues et fixées par la protéine CTCF (CCCTC-binding), dans 75 à 95% des cas selon le type cellulaire (DIXON et al. 2012). CTCF est une protéine avec une structure en doigts de zinc qui lui permet de se fixer à l'ADN et

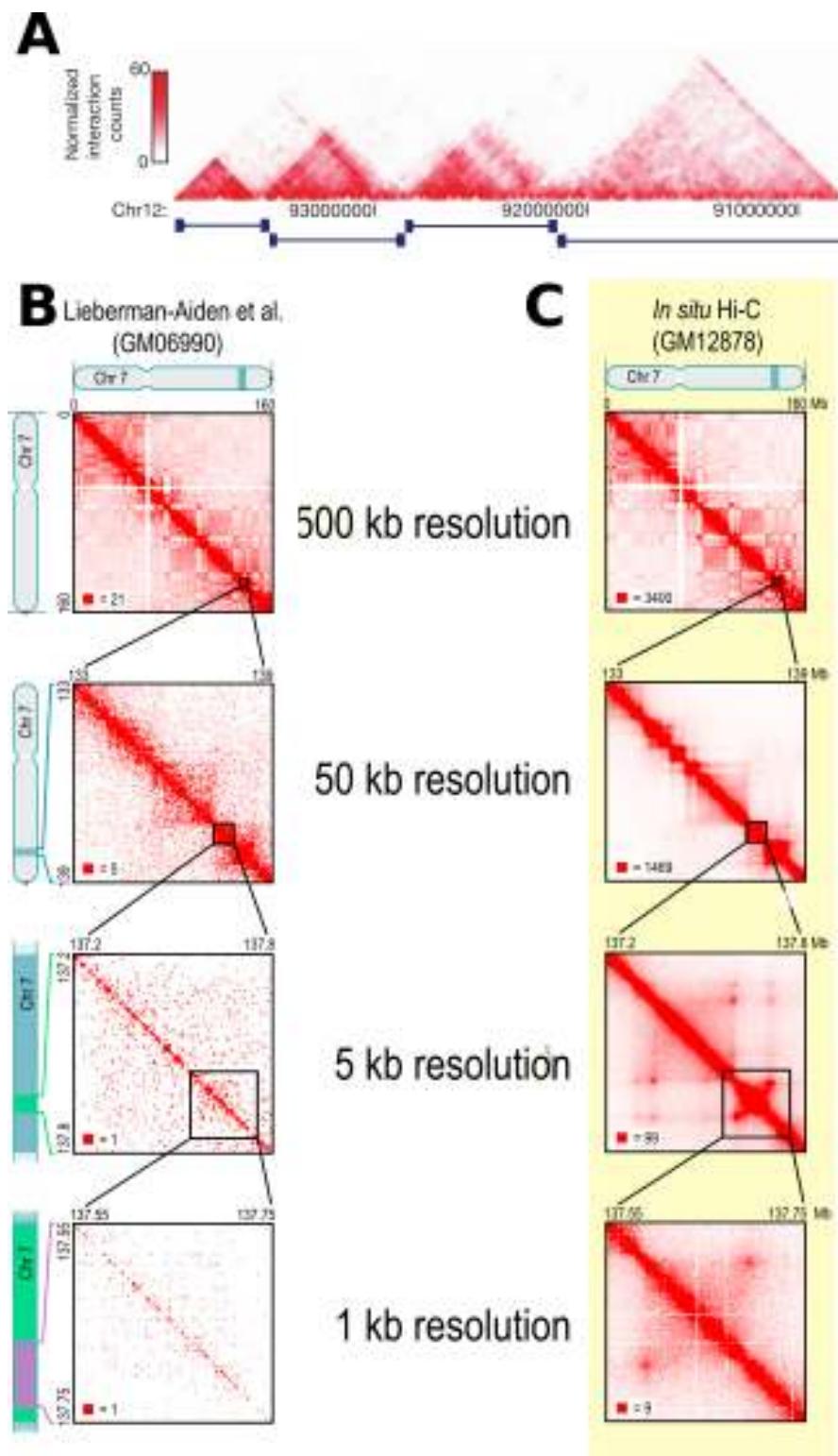


FIGURE 1.7 – Les Domaines Topologiquement associatifs (TAD) et les boucles vues par Hi-C, à différentes résolutions. **A** adapté de (DIXON et al. 2012), **B,C** adapté de (S. S. RAO et al. 2014).

reconnait un motif spécifique orienté. Ces sites sont présents en très grand nombre sur le génome, et tous ne co-localisent pas avec une frontière de *TAD*. Il a cependant été montré que ce sont les sites convergents (donc dans deux sens opposés) qui forment préférentiellement une boucle de chromatine (voir Figure 1.8 A et B) et qu'une inversion d'un des deux sites pouvait amener à une disparition de la boucle de la chromatine (WIT et al. 2015). En outre, à l'échelle du génome, une dégradation de CTCF entraîne un affaiblissement général des frontières des *TAD* (NORA et al. 2017) (voir Figure 1.8 C), ce qui montre que la fixation de CTCF sur des sites convergents est essentiel à la formation des boucles.

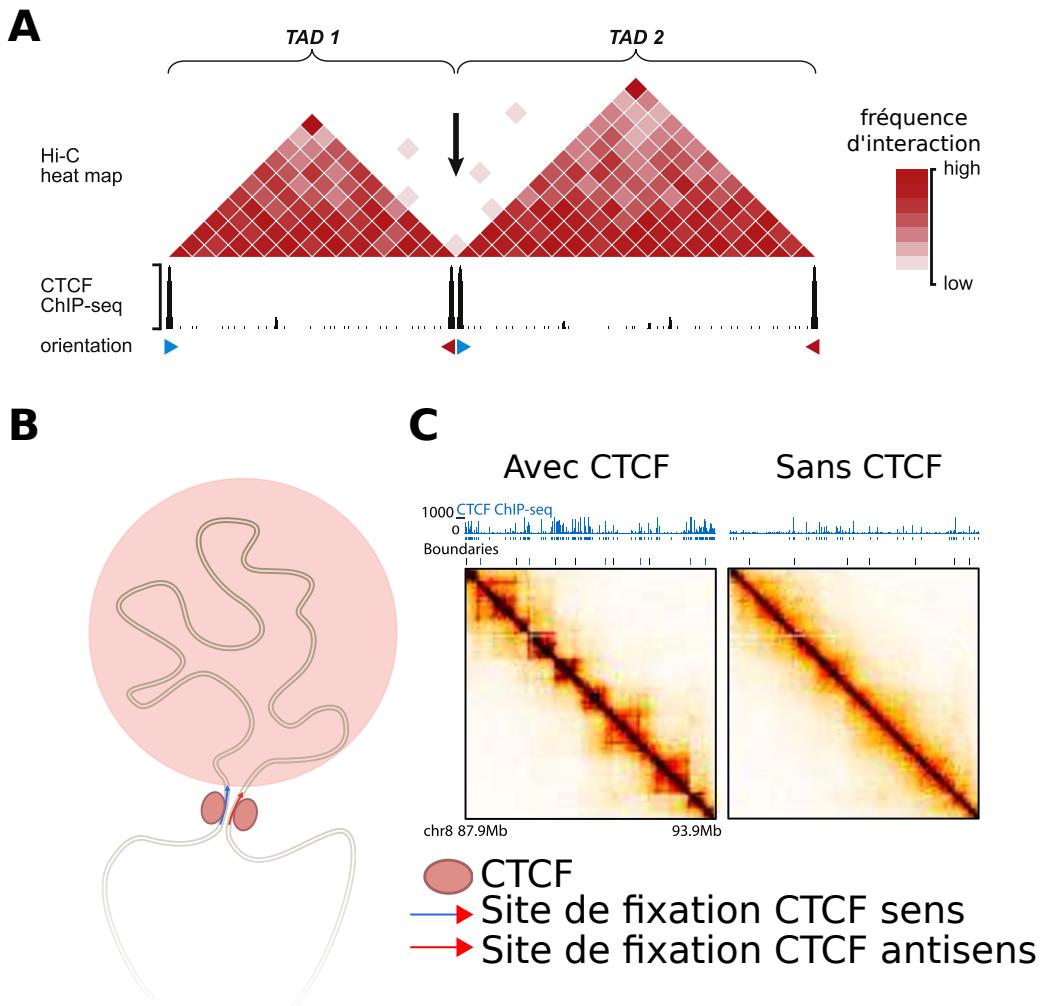


FIGURE 1.8 – Le rôle de CTCF dans les boucles et les frontières de *TAD*. **A** Schéma d'une carte Hi-C montrant la formation de *TAD* par une boucle (point rouge au sommet des triangles) via des frontières insulatrices reconnues par CTCF (orientation représentée par des triangles rouges et bleus), adapté de (CHANG, GHOSH et NOORDERMEER 2020). **B** Schéma de la boucle formant un *TAD* via la reconnaissance de deux sites convergents par CTCF. **C** Carte Hi-C d'une résolution de 20kb avant et après dégradation de la protéine CTCF. La dégradation de CTCF entraîne une baisse d'insulation entraînant la déstructuration des *TAD*. Adapté de (NORA et al. 2017).

**1.1.4.3.2 Le complexe de la cohésine** La cohésine est un complexe de protéines formant un anneau grâce à des protéines de maintenances structurelles ou *Structural Maintenance of Chromosomes* (*SMC*). Les *SMC* font partie d'une grande famille d'ATPase qui interviennent dans l'organisation struc-

## 1.1. L'ORGANISATION DE L'ADN DANS LE NOYAU

turelle de la chromatine au sein du noyau, de part sa présence dans les condensines, et dans les cohésines (YATSKEVICH, RHODES et NASMYTH 2019). En plus de son rôle de cohésion des chromatides sœurs pendant la duplication, elle co-localise souvent avec la protéine CTCF (CHANG, GHOSH et NOORDERMEER 2020), ce qui suggère un rôle dans la formation des *TAD* (voir Figure 1.9 A). En effet, suite à une déplétion d'une des sous unité de la cohésine (SCC1), la totalité des domaines formés par les boucles disparaissent du génome (Suhas SP RAO et al. 2017; WUTZ et al. 2017) (voir Figure 1.9 B).

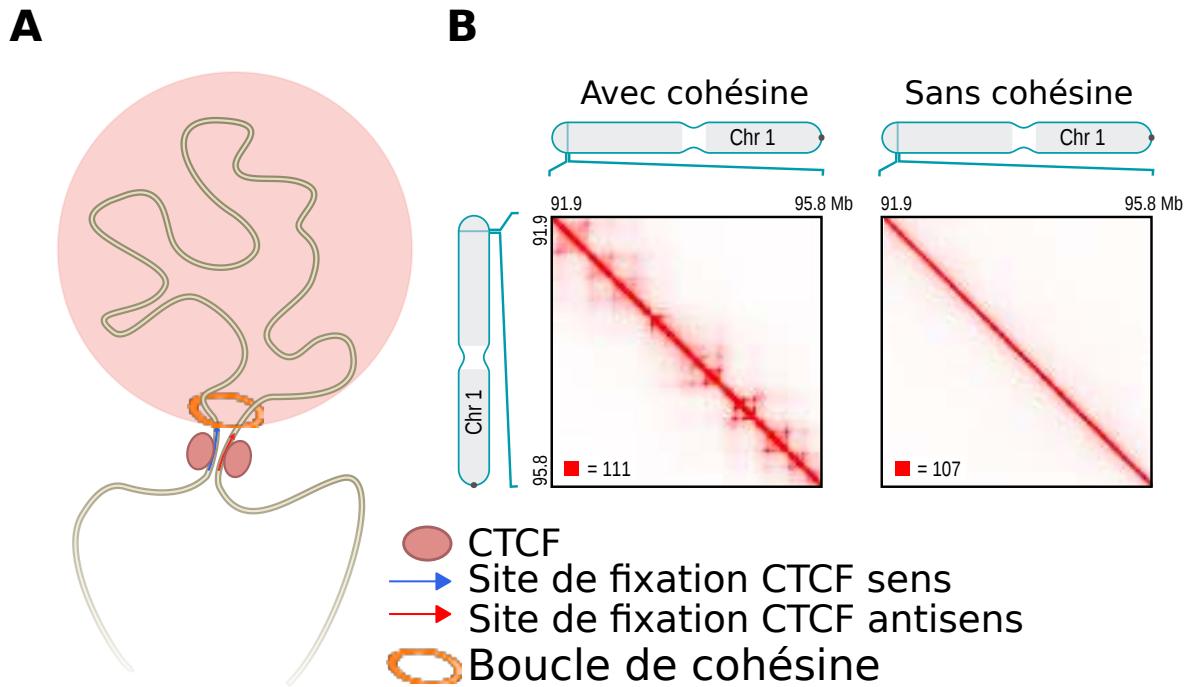


FIGURE 1.9 – Le rôle de la cohésine dans les boucles et les frontières de TAD. **A** Schéma montrant comment la cohésine pourrait maintenir deux sites CTCF convergents proches, formant une frontière de TAD. **B** Carte de Hi-C montrant l'élimination totale des TAD après dégradation de la sous-unité SCC1 de la cohésine, adapté de (Suhas SP RAO et al. 2017).

### 1.1.4.4 Le processus d'extrusion de boucles

L'association entre la boucle de cohésine et CTCF semble avoir un rôle primordial dans la formation des domaines 3D (*TAD* et sous-*TAD*). En se basant sur ce lien, un modèle appelé l'extrusion de boucle, ou *loop extrusion* a été développé *in silico* afin de reproduire les données de Hi-C, et a montré de grandes similitudes avec les données Hi-C expérimentales (FUDENBERG et al. 2016). Dans ce modèle, le complexe de la cohésine s'associe de manière aléatoire avec la fibre de chromatine et forme progressivement une boucle qui s'allonge de manière bidirectionnelle. L'extrusion est bloquée d'un seul côté lorsque le processus rencontre CTCF associé avec l'ADN dans la bonne orientation (voir Figure 1.10). Quand la cohésine a rencontré un site CTCF des deux côtés du *TAD*, la boucle se stabilise, ce qui explique la fréquence d'interaction forte observée au niveau du sommet du *TAD* sur les cartes de Hi-C (voir Figure 1.8 A). Ce mécanisme actif qui “glisse” le long de l'ADN explique également les contacts entre les différents *loci* du *TAD*, puisque ceux-ci entrent en contacts tout le long du processus de *loop extrusion* (CHANG, GHOSH et NOORDERMEER 2020).

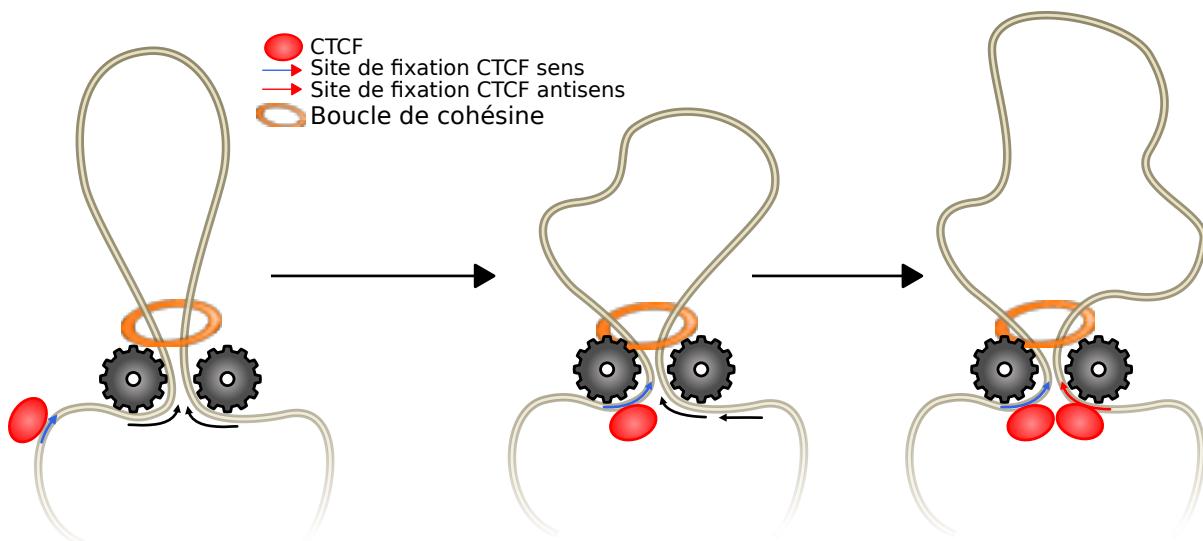


FIGURE 1.10 – Modèle de l'extraction de boucles, faisant intervenir CTCF et l'anneau de cohésine. L'anneau de cohésine agit comme un moteur moléculaire extrudant l'ADN de manière bidirectionnelle, jusqu'à rencontrer CTCF associé à son site de fixation. Adapté de *Loops that mend the genome* par Leonid A. Mirny.

Ce modèle d'extraction de boucle ne s'applique pas uniquement à la création des *TAD*, mais à bien d'autres mécanismes faisant intervenir les complexes SMC, notamment pendant la compaction et ségrégation des chromosomes pendant la mitose, la cohésion des chromatides soeurs pendant la méiose, mais également des processus de contacts entre le promoteur et ses régions régulatrices (enhanceurs) et la réparation de l'ADN (YATSKEVICH, RHODES et NASMYTH 2019). Il a été directement observé une première fois par des expériences *in vitro* sur les condensines de la levure (GANJI et al. 2018), puis sur les cohésines humaines (Y. KIM et al. 2019) (DAVIDSON et al. 2019). Ces expériences ont permis de montrer que ce mécanisme agit comme un moteur moléculaire et extrude très rapidement à une vitesse d'environ 1kb/s (BANIGAN et MIRNY 2020).

Sur les cartes Hi-C très résolutives, on peut également observer des bandes ou *stripes* rouges sur les bordures des *TAD*, jusqu'au sommet, ce qui suggère que le processus d'extraction de boucle est bloqué d'un côté par CTCF, mais continue de progresser de l'autre (voir Figure 1.11) (VIAN et al. 2018). Les cartes Hi-C étant formées par les interactions ADN-ADN dans une population cellulaire, ces *stripes* pourraient résulter d'un ensemble de boucles en train d'extruder d'un seul côté.

De plus, des protéines connues pour s'associer à la cohésine, comme NIPBL, co-localisent avec celle-ci au niveau des frontières insulatrices formant les *TAD*, et notamment sur celles montrant des *stripes*. NIPBL est responsable du chargement de l'anneau de cohésine sur l'ADN, et sa déplétion durant l'extraction de boucle interrompt le mécanisme (DAVIDSON et al. 2019), ce qui suggère un rôle actif de cette protéine dans ce processus. Un autre facteur connu pour s'associer avec la cohésine, WAPL, qui permet notamment la dissolution des anneaux de cohésines semble avoir un rôle important dans la *loop extrusion*, car une déplétion de cette protéine induit la formation de boucles et *stripes* plus longues au niveau des *TAD* (voir Figure 1.12) (HAARHUIS et al. 2017).

L'ensemble de ces éléments suggère un mécanisme actif de la *loop extrusion* à l'origine de la formation des boucles de chromatines et de *TAD*. La visualisation de phénomènes comme les *stripes*, et l'impact

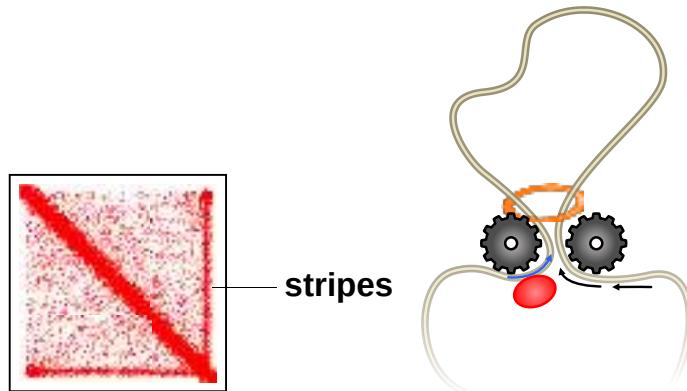


FIGURE 1.11 – La loop extrusion peut former des stripes sur les cartes Hi-C. Ce phénomène peut s'observer dans le cas où un anneau de cohésine est bloqué d'un côté par CTCF et extruderait d'un seul côté. Ce locus entrerait alors en interaction avec l'ensemble du TAD formant une bande rouge d'interaction de forte intensité sur les cartes Hi-C. Adapté de (VIAN et al. 2018)

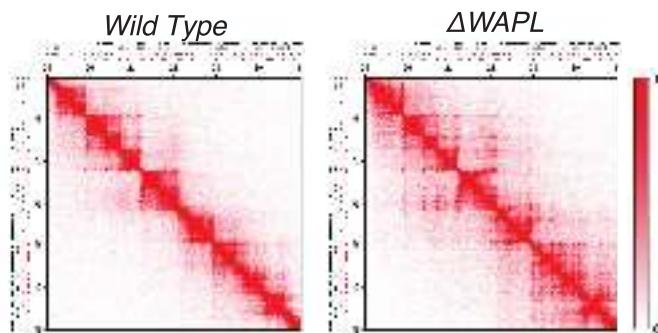


FIGURE 1.12 – Carte de Hi-C à une résolution de 20kb avant et après déletion de WAPL (HAARHUIS et al. 2017). La déplétion de WAPL augmente le temps d'interaction entre l'ADN et les cohésines et forme des boucles plus longues, observables en Hi-C.

de la déplétion des protéines intervenant dans ce modèle sur les cartes Hi-C suggère que ce mécanisme est bien utilisé *in vivo*, et serait responsable d'une grande partie de l'organisation de la structure tridimensionnelle de la chromatine dans le noyau.

#### 1.1.4.5 Les compartiments A et B de la chromatine

Il existe majoritairement deux formes sous laquelle la chromatine est présente à grande échelle, l'euchromatine et l'hétérochromatine (voir partie 1.1.1.2). Cette organisation de la chromatine se reflète en 3D dans un niveau supérieur à celui des TAD. En effet, la première expérience de Hi-C, réalisée en 2009, a mis en évidence une décomposition des interactions des *loci* en deux ensembles, appelés compartiment A et compartiment B (LIEBERMAN-AIDEN et al. 2009). Ces deux compartiments n'interagissent pas ou peu ensemble, et forment des damiers sur la carte d'interaction normalisée (voir Figure 1.13 et partie 1.3.6.1.3.2). Ce quadrillage formé par l'alternance des deux compartiments le long du chromosome montre que deux *loci* d'ADN sont capables d'interagir à longue distance et peuvent être proches dans l'espace. Ceci suggère un partitionnement du génome dans l'espace du noyau (voir Figure 1.14 A). Il est possible, via des méthodes statistiques de réduction de dimensions, de détecter ces compartiments à partir des matrices Hi-C (voir partie 1.3.6.1.4).

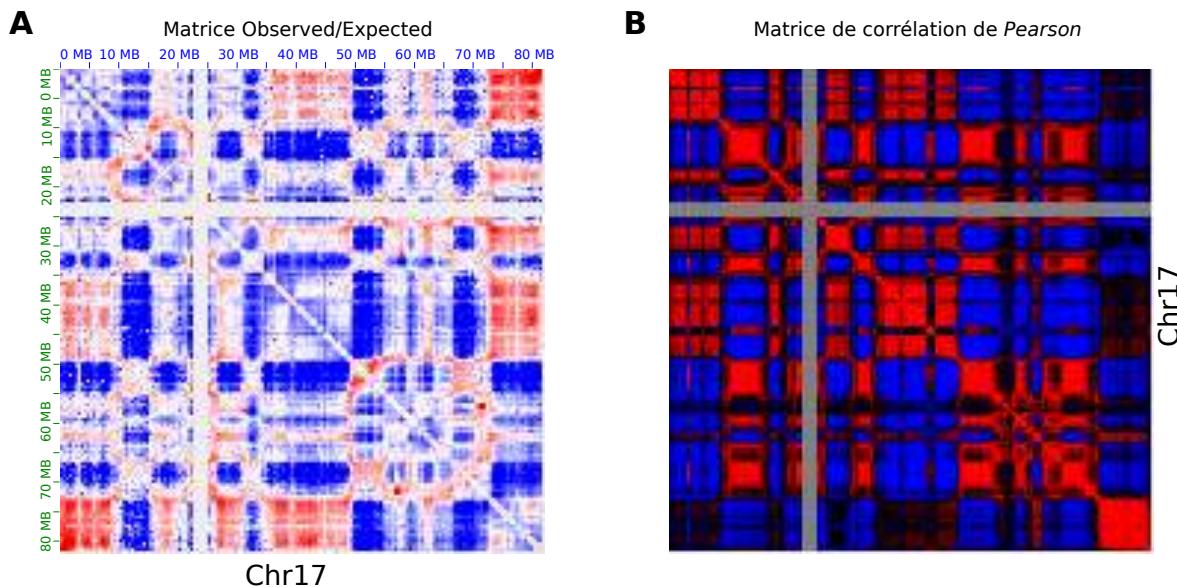


FIGURE 1.13 – Carte d'interaction Hi-C normalisée du chromosome 17 chez l'Homme et sa matrice de corrélation de Pearson correspondante. Sur ces matrices, les *loci* d'ADN sont séparés en deux groupes qui n'interagissent pas ensemble, formant deux compartiments distincts qui s'alternent tout le long du chromosome.

La composition en modifications d'histones au sein des compartiments A et B a également été analysée, ce qui a permis d'associer ces compartiments à l'état ouvert/fermé de la chromatine. Le compartiment A correspond à une chromatine ouverte, transcriptionnellement active, très caractéristique de l'euchromatine. Le compartiment B lui est en revanche associé à des marques caractéristiques de l'hétérochromatine (voir Figure 1.14 A, S. S. RAO et al. (2014)). L'utilisation de données plus résolutives a permis de diviser les compartiments en sous-compartiments, en fonction de leur composition en modifications

## 1.1. L'ORGANISATION DE L'ADN DANS LE NOYAU

post-traductionnelles. Par exemple, le compartiment B peut être divisé en 3 parties : B1, qui possède des marques d'hétérochromatine facultatives (corrélation positive avec H3K27me3), B2, caractéristique d'une hétérochromatine associés aux centromères, et B3 associé à l'hétérochromatine constitutive (voir Figure 1.14 B).

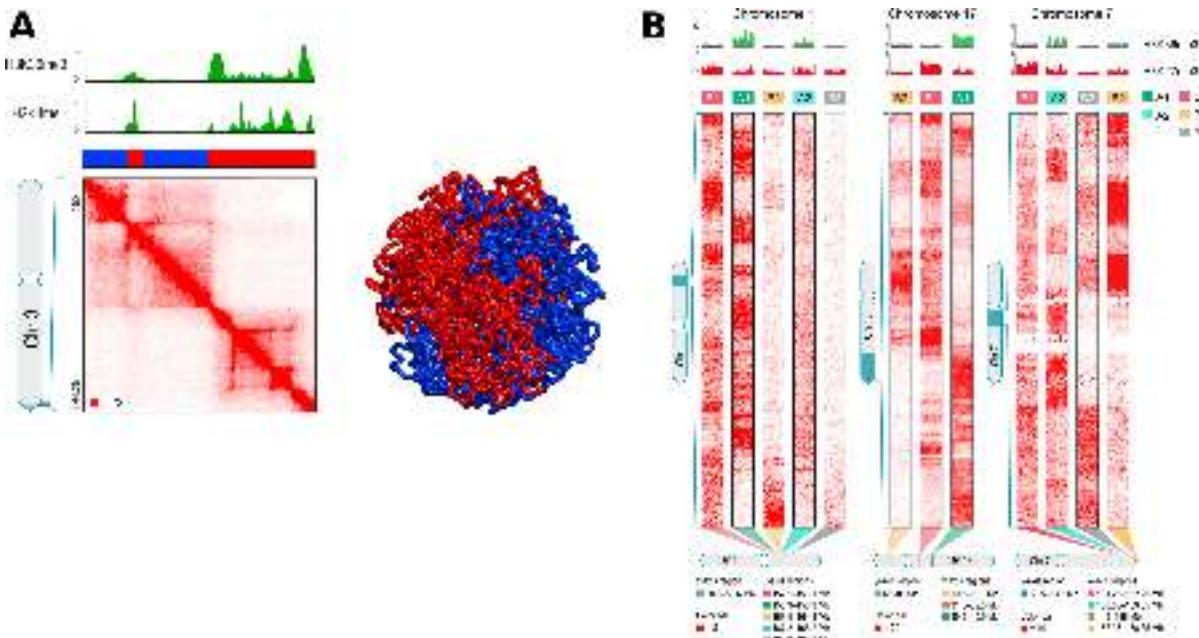


FIGURE 1.14 – Les matrices d’interaction Hi-C s’expliquent par la compartimentation A et B et par la formation des TAD par loop extrusion. **A** Sur une carte de Hi-C on peut visualiser les TAD, qui forment des domaines 3D fermés. Les TAD n’interagissent pas tous entre eux, et forment deux compartiments, A et B, le long du chromosome. Ces compartiments sont enrichis en différentes marques d’histones, ce qui permet de les associer à l’euchromatine (compartiment A) et à l’hétérochromatine (compartiment B). Adapté de (Suhas SP RAO et al. 2017). **B** Les compartiments A et B sont eux-mêmes divisibles en plusieurs sous-compartiments en fonction de leurs caractéristiques chromatiniennes. Adapté de (S. S. RAO et al. 2014)

**1.1.4.5.1 La séparation de phase** La ségrégation des chromosomes en deux compartiments, A et B au sein du noyau est issue d’un phénomène différent de la formation des *TAD*. En effet, la déplétion de différentes protéines associées à l’organisation 3D du génome, comme CTCF et la cohésine (SCC1) affaiblissent les *TAD* mais n’ont au contraire pas d’effet sur les compartiments A et B, voir même les renforce (NORA et al. 2017; HAARHUIS et al. 2017; DAVIDSON et al. 2019; Suhas SP RAO et al. 2017). De plus, quand on renforce l’effet de la *loop extrusion* par l’inhibition de la dégradation de l’anneau de cohésine, les boucles se prolongent plus loin et les compartiments sont fortement inhibés (WUTZ et al. 2017). Ces résultats suggèrent que les compartiments sont formés par un mécanisme différent de celui, qui est responsable de la formation des domaines 3D, et que ces deux entrent en compétition pour organiser le génome (NUEBLER et al. 2018; Suhas SP RAO et al. 2017) (voir Figure 1.15).

Ce mécanisme de compartimentation passif serait dépendant de la composition de la chromatine. En effet, les compartiments riches en chromatine active en transcription (A, euchromatine) et la chromatine répressive (B, hétérochromatine) ont une composition en marques d’histones très différente. Ils se repousseraient donc mutuellement, de manière passive, par un procédé appelé séparation de phase (voir Figure 1.15 A). Ce procédé a été simulé et explique la compartimentation de l’ADN observé sur les

données Hi-C, ainsi que son opposition à la *loop extrusion* qui peut remplacer ce mécanisme à petite échelle (NUEBLER et al. 2018).

Il a été montré que l'hétérochromatine est formé par séparation de phase, par la présence de la protéine HP1, même si son rôle est toujours contesté *in vivo* (STROM et al. 2017; ERDEL et al. 2020). Le compartiment B étant majoritairement constitué d'hétérochromatine, il est possible qu'il soit formé par séparation de phase grâce à HP1.

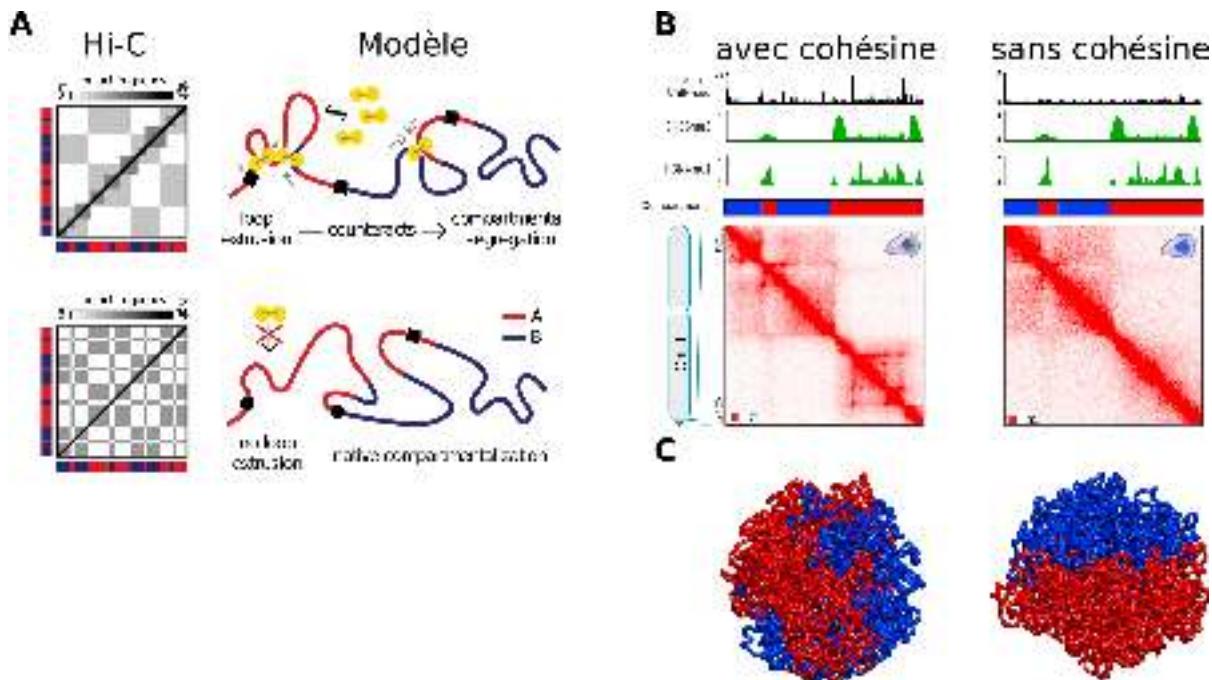


FIGURE 1.15 – La formation des compartiments par séparation de phase, un processus passif en compétition avec la formation des boucles de chromatine. **A** La formation des boucles de chromatine semble s'opposer à la formation passive des compartiments par séparation de phase. L'absence de loop extrusion conduit à l'état natif de la chromatine de par sa composition en modifications d'histones. Adapté de (NUEBLER et al. 2018). **B** Une déplétion de la cohésine conduit à une disparition des boucles, qui amène à une meilleure compartmentation de la chromatine par simulation. Adapté de (Suhas SP RAO et al. 2017)

#### 1.1.4.6 Les niveaux supérieurs d'organisation de la chromatine

**1.1.4.6.1 Les territoires chromosomiques** Les chromosomes au sein du noyaux s'organisent de façon spécifique, particulièrement en interphase. Par des méthodes de microscopie couplée à du *FISH*, il a été montré que chaque région du noyau était sub-divisé en territoires chromosomiques (voir Figure 1.16), dont la position est hétérogène en fonction du tissu (PARADA, MCQUEEN et MISTELI 2004).

Les interactions inter-chromosomiques des cartes Hi-C sont une alternative à la microscopie pour inférer sur la façon dont se comportent et se regroupent les chromosomes dans le noyau (voir partie 1.3.6.1). En effet, les interactions inter-chromosomiques existent, mais sont très largement défavorisées au profit des interactions intra-chromosomiques. Ceci indique que la position des chromosomes dans le noyau est tout sauf aléatoire, et que chaque chromosome occupe une région bien précise qu'on appelle territoire chromosomique (LIEBERMAN-AIDEN et al. 2009).

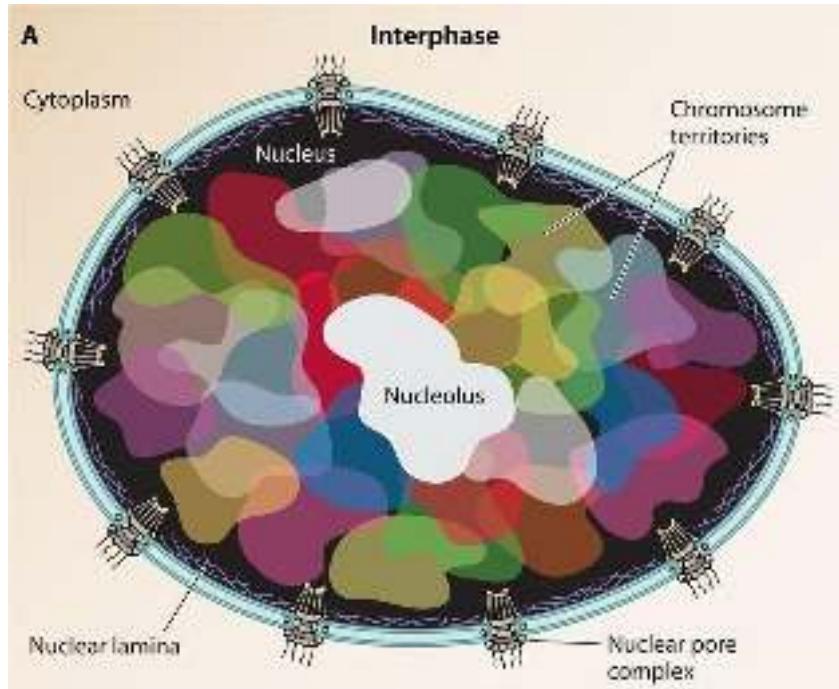


FIGURE 1.16 – Organisation 3D du génome humain dans le noyau. Représentation des territoires chromosomiques pendant l'interphase, du Nucléole (blanc) et du réseau de lamine (filaments violet le long de la membrane nucléaire). Adapté de FRASER et al. 2015

**1.1.4.6.2 Les domaines associés aux lamines** Dans le noyau, un réseau de protéines situées le long de la membrane interne, les lamines, permettent de structurer le noyau. Dans le génome, des régions spécifiques sont connues pour interagir fréquemment avec ses lamines, ce sont les domaines associés aux lamines, ou *Lamina-Associated Domains (LAD)* (voir Figure 1.16). Comme les territoires chromosomiques, les *LAD* sont stables d'une cellule à l'autre. Elles sont majoritairement composées d'hétérochromatine, caractérisées par une répression de la transcription (YOKOCHI et al. 2009). Des études Hi-C ont montré que les *LAD* corrélent fortement avec les *TAD* et le compartiment B (NORA et al. 2017; DIXON et al. 2012).

**1.1.4.6.3 Le nucléole** Dans le noyau, le nucléole est connu pour contenir l'ADN ribosomique. C'est une région nucléaire qui n'est pas séparée du reste par une membrane. L'ADN ribosomique étant réparti sur différents chromosomes, le nucléole contient donc des régions de différents chromosomes qui se regroupent ensemble. C'est une région riche en hétérochromatine, notamment des centromères et des télochromosomes, malgré le fait qu'il soit associé à l'activité de la transcription de l'ADN ribosomique (BERSAGLIERI et SANTORO 2019). Le Nucléole est également associé au compartiment B de la chromatine, et comme lui, a été montré comme capable de faire de la séparation de phase (FERIC et al. 2016).

## 1.2 Les cassures double-brin de l'ADN (*DSB*) et leur réparation

Le génome est soumis à des stress qui provoquent de nombreux dommages à l'ADN. Parmi ces dommages, les *cassures doubles brins* ou *double-strand breaks (DSB)* sont connues pour être les plus délétères, pouvant

causer de multiples réarrangements chromosomiques et mutations.

Ces *DSB* peuvent survenir sur l'ADN suite à une exposition à des facteurs externes, comme la consommation de cigarettes ou l'utilisation de la chimiothérapie, mais également lors du développement normal de la cellule, comme lors de la production des anticorps par le mécanisme de recombinaison V(D)J (SOULAS-SPRAUEL et al. 2007) ou bien du brassage génétique pendant la méiose (GRAY et P. E. COHEN 2016). La cartographie réalisée par séquençage à haut débit a mis en évidence que les *DSB* peuvent également apparaître suite à la collision de la machinerie transcriptionnelle ou réplicative avec des structures secondaires, comme les G-quadruplexes (G4) ou bien les hybrides ADN-ARN, les R-loops (PUGET, MILLER et Gaëlle LEGUBE 2019 ; MARNEF, S. COHEN et Gaëlle LEGUBE 2017). Les frontières des *TAD* sont également sujettes aux *DSB*, par l'action potentielle de la Topoisomérase II (*TOP2*). En effet, la *TOP2* provoque des cassures de l'ADN réversibles, afin de résoudre les contraintes de tensions causées par le déroulement de la fibre de chromatine pendant la transcription. Cette action réversible pourrait être bloquée au niveau des frontières de *TAD*, par la présence des protéines insulatrices et induire des dommages (GOTHE et al. 2019).

L'apparition des *DSB* dans le génome déclenche un ensemble de mécanismes qu'on appelle la *DNA Damage Response*, ou *DDR*. La *DDR* prend en charge la détection, la signalisation, et la réparation des *DSB*. C'est donc un ensemble très complexe, qui fait intervenir de nombreuses protéines, mais aussi met en route la machinerie transcriptionnelle (MARNEF, S. COHEN et Gaëlle LEGUBE 2017), modifie la chromatine au niveau local et global, via des modifications post-traductionnelles (Thomas CLOUAIRO et al. 2018), et de conformation 3D (François AYMAR, AGUIREBENGOA et al. 2017).

### 1.2.1 Signalisation des DSB

Les *DSB* vont être très rapidement détectées et prises en charge par plusieurs mécanismes. L'un des premiers événements intervenant suite aux *DSB* va être l'apparition de foyers visibles dans le noyau en microscopie, formés par la phosphorylation d'un variant d'histone H2AX ( $\gamma$ H2AX) (ROGAKOU et al. 1998). H2AX est phosphorylé par une kinase, ATM, pour *Ataxia Telangiectasia Mutated*, qui est recrutée et activée directement au niveau des *DSB*. Les protéines qui recrutent ATM sont des protéines de reconnaissance des *DSB*, comme le complexe MRN (formé de trois protéines : Rad50, NBS1 et Mre11) (voir Figure 1.17). La détection et la localisation des *DSB* provoque également un arrêt du cycle cellulaire, pour permettre la réparation de l'ADN par la cellule, et empêcher la transmission de l'ADN endommagé aux cellules filles.

#### 1.2.1.1 Les foyers de réparation

Les foyers de réparation sont des domaines de chromatine très larges de l'échelle de la Mégabase (Mb). Ils sont caractérisés par la phosphorylation de H2AX par ATM en  $\gamma$ H2AX, et de protéines de réparation comme 53BP1 et MDC1.

**1.2.1.1.1 Phosphorylation de  $\gamma$ H2AX** La phosphorylation de  $\gamma$ H2AX par ATM intervient quelques secondes seulement après l'endommagement de l'ADN, et va se propager le long du foyer de réparation, donc jusqu'à 1 ou 2 Mb en 15 ou 30 minutes (GEORGULIS et al. 2017 ; IACOVONI et al.

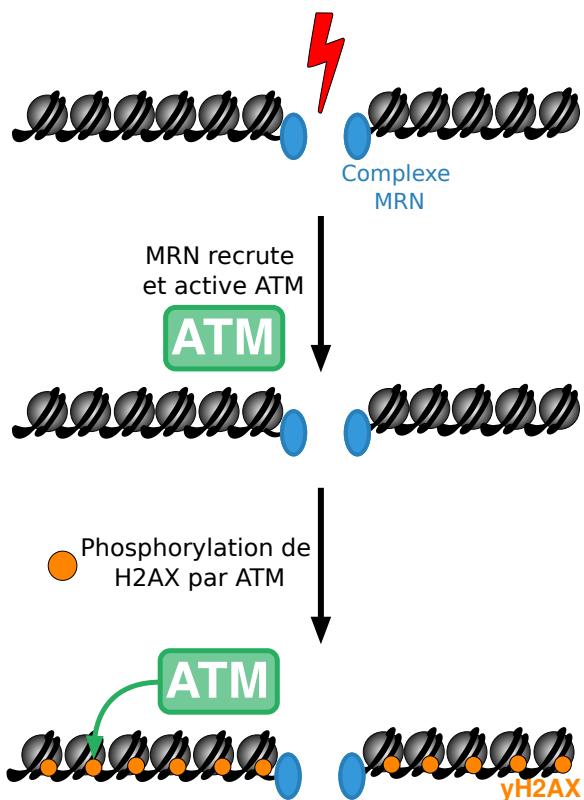


FIGURE 1.17 – Signalisation des DSB par phosphorylation de l'histone H2A par ATM le long de l'ADN endommagé.

2010). En réalisant la cartographie par des méthodes omiques de  $\gamma$ H2AX et ATM (voir partie 1.3.2), une différence flagrante a été observée entre leur localisation (voir Figure 1.18). En effet, la protéine ATM est localisée dans une région très proche de la DSB, alors que  $\gamma$ H2AX lui est diffusé dans un large domaine et de façon asymétrique par rapport à la DSB (CARON, CHOUDJAYE et al. 2015 ; IACOVONI et al. 2010). Comment ATM arrive-t-il à propager à distance  $\gamma$ H2AX est à ce jour très mal compris.

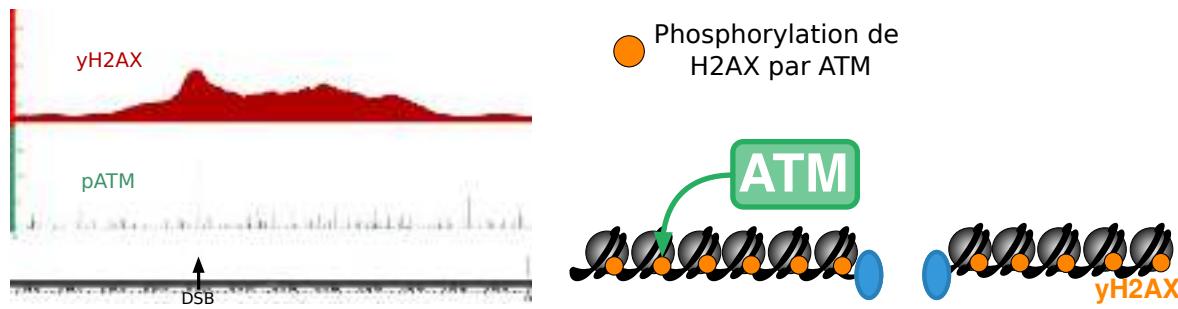


FIGURE 1.18 – Profil centré sur une DSB montrant la propagation de  $\gamma$ H2AX et l'enrichissement très local d'ATM. ATM est responsable de la phosphorylation de H2AX en  $\gamma$ H2AX, mais n'est enrichi qu'à proximité de la DSB, suggérant qu'un autre mécanisme intervient pour propager  $\gamma$ H2AX.

**1.2.1.1.2 Le rôle de  $\gamma$ H2AX dans la signalisation et la réparation** La présence d'aussi larges foyers de réparation (ou domaines gamma), formés par  $\gamma$ H2AX, suggère un rôle important dans la signalisation et la réparation des DSB à l'échelle génomique. Cependant,  $\gamma$ H2AX et ATM ne sont pas indispensables à la prise en charge initiale des DSB, ni indispensables à la réparation (CELESTE et al. 2003), mais favorisent cependant sa fidélité (CARON, CHOUDJAYE et al. 2015). En revanche,  $\gamma$ H2AX possède un rôle actif dans l'accumulation des protéines de réparation au sein du domaine gamma (CELESTE et al. 2003). Il pourrait également influencer certaines voies de réparation plus fidèles (voir partie 1.2.2), et servir de plateforme au recrutement d'acteurs majeurs de la DDR, comme 53BP1 ou MDC1. Tout ceci suggère que la formation d'un domaine gamma pourrait permettre à la chromatine autour de la DSB d'être plus accessible, moins compacte, et mobile au sein du noyau (voir partie 1.2.6.3).

**1.2.1.1.3 53BP1 et MDC1, acteurs majeurs de la DDR** la phosphorylation de  $\gamma$ H2AX permet le recrutement direct de la protéine MDC1, qui elle-même recrute d'autres facteurs (SAVIC et al. 2009). Ce recrutement va permettre de modifier la chromatine au sein du domaine de réparation, notamment par l'ubiquitination de l'histone H1, puis H2A et H2AX, permettant la transmission du signal d'endommagement sur une très grande échelle (Thomas CLOUAIRE et al. 2018).

53BP1 (p53-binding protein 1) est une autre protéine de la DDR ayant un rôle primordial dans la réparation, notamment dans le choix de la voie de réparation (voir partie 1.2.2), ainsi que dans la résection (Figure 1.19). Elle reconnaît certaines modifications d'histones, dont l'ubiquitination de l'histone H2A (HUSTEDT et DUROCHER 2017) ainsi que H4K20me2 (PANIER et BOULTON 2014). Son recrutement au niveau de ses modifications lui permet de se propager sur la totalité d'un domaine gamma (Thomas CLOUAIRE et al. 2018). En outre, c'est également un acteur important de l'arrêt du cycle cellulaire suite à l'apparition de dommages, notamment dans son interaction avec la protéine p53 et son rôle dans la gestion du cycle cellulaire (CUELLA-MARTIN et al. 2016). Enfin, il a été montré que la chromatine enrichie en 53BP1 était capable de réaliser de la séparation de phase (KILIC et al. 2019).

### 1.2.2 La réparation des DSB

Il existe deux principaux mécanismes qui permettent de réparer de façon plus ou moins fidèle les DSBs : la recombinaison homologue (*HR*) et la jonction des extrémités non homologues (*NHEJ*) (Figure 1.19, HARTLERODE et SCULLY (2009)).

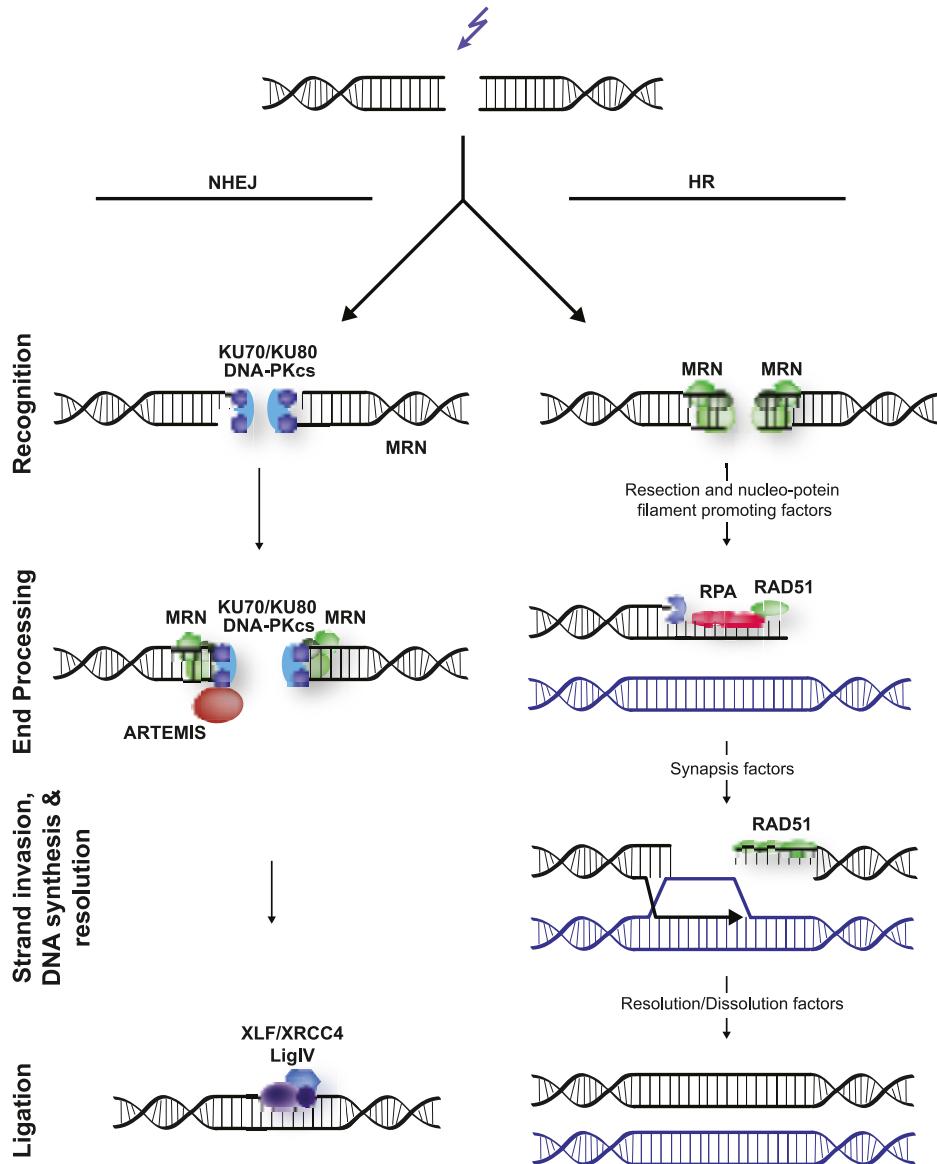


FIGURE 1.19 – Représentation des deux principales voies de réparation d'une DSB. À gauche, la réparation par jonction d'extrémités non homologues (*NHEJ*), qui repose sur la religation directe des extrémités endommagées. À droite, la réparation par recombinaison homologue (*HR*) impliquant la résection des extrémités et la synthèse d'ADN utilisant la chromatide sœur (bleu) comme matrice. (LANS, MARTEIJN et VERMEULEN 2012)

La *NHEJ* est une voie de réparation rapide et efficace. Elle peut intervenir à n'importe quelle phase cellulaire, mais est particulièrement importante en phase G1. C'est une réparation directe via la religation des deux extrémités endommagées, elle est donc simple mais moins fidèle. La *NHEJ* peut induire plusieurs types de mutations (insertion, délétion, substitution), en général localisées proches du point de jonction.

## 1.2. LES CASSURES DOUBLE-BRIN DE L'ADN (DSB) ET LEUR RÉPARATION

Ce type de réparation peut donc aboutir à une perte d'information génétique, et à un décalage du cadre de lecture, si les dommages ont eu lieu dans une partie codante d'un gène.

La *HR* permet une réparation bien plus fidèle de l'ADN endommagé, car elle utilise la chromatide sœur comme matrice de synthèse. C'est un mécanisme bien plus complexe que la *NHEJ*, par résection de l'ADN endommagé nécessitant l'intervention d'un plus grand nombre de protéines. En outre, elle n'est utilisable qu'en phase S et G2 du cycle cellulaire, lorsque la chromatide sœur est disponible. Cette voie de réparation fait intervenir de nombreux acteurs ayant un rôle complexe, comme Rad51 et BRCA1, qui interviennent dans le recrutement de la séquence homologue. Même si ce type de réparation est plus fidèle, il peut tout de même aboutir à des erreurs, comme de larges délétions ou des translocations (sur des régions répétées), ou bien à une perte de l'hétérozygotie (par utilisation du chromosome homologue). *HR* et *NHEJ* ont donc des conséquences variées sur un *locus* endommagé, n'ont pas la même fidélité de réparation et peuvent amener à des erreurs fondamentalement différentes.

### **1.2.3 Choix de la voie de réparation**

Les deux mécanismes principaux de la réparation, *HR* et *NHEJ*, impliquent une fidélité de réparation différente. Le choix de la réparation dépend de la localisation de la *DSB* sur la chromatine, et de son état local. En effet, des protéines comme 53BP1 inhibe le processus de résection, et donc favorisent la mise en place de la *NHEJ*. D'autres éléments entrent en jeu, comme le fait que la *HR* nécessite l'implication de la chromatide sœur disponible uniquement en S et G2 du cycle cellulaire, et la diminution de H4K20me2, marque reconnue par 53BP1, pendant ces phases, permet de favoriser ce type de réparation. Le choix de la réparation dépend donc de la capacité des protéines impliquées dans ces voies à se fixer et reconnaître l'ADN endommagé, capacité définie par l'état local de la chromatine (CLOUAIRE et LEGUBE 2015).

#### **1.2.3.1 L'état chromatinien favorise un type de réparation particulier**

La chromatine est globalement présente sous la forme d'euchromatine ou d'hétérochromatine dans le noyau. L'euchromatine, synonyme de transcription active, est peu dense, et l'hétérochromatine à l'inverse, est très compacte et les gènes sont globalement inactifs. Les marques d'histones sont très différentes selon le type de chromatine (voir Figure 1.2), et certaines de ses marques vont favoriser un type de réparation plutôt qu'un autre. En effet, il a été montré que les régions actives en transcription sont préférentiellement réparées par *HR* (François AYMARD, Beatrix BUGLER et al. 2014; CLOUAIRE et LEGUBE 2015), notamment par la présence de la marque d'histone H3K36me3, enrichie sur tout le corps des gènes actifs, qui favoriseraient le recrutement de Rad51. À l'inverse, les gènes inactifs ou les régions intergéniques de l'hétérochromatine sont plutôt réparés par *NHEJ* (Thomas CLOUAIRE et al. 2018).

Cette réparation définie par l'état de la chromatine peut cependant changer en fonction de la phase, car l'hétérochromatine peut être favorablement réparée par *HR* pendant les phases S et G2 du cycle cellulaire (BEUCHER et al. 2009). Cela s'explique en parti par l'intervention de HP1 qui recrute BRCA1 et favorise *HR* (Y.-H. LEE et al. 2013). En outre, la marque d'histone fortement associé à l'hétérochromatine, H3K9me3, peut recruter des éléments nécessaires à la décompactation de la chromatine, permettant le recrutement de *HR* (AYRAPETOV et al. 2014).

### 1.2.4 La réparation dans les gènes actifs

La réparation dans les gènes actifs est particulièrement compliquée, du fait qu'elle intervient en présence de la machinerie transcriptionnelle. L'expression d'un gène endommagé va être modifiée, souvent caractérisée par une baisse d'expression, mais aussi par des changements de comportement de la machinerie transcriptionnelle pour faciliter la réparation (BADER et al. 2020). Des structures secondaires associées à la transcription d'un gène, comme les hybrides ADN-ARN, ou R-loops, s'accumulent au niveau de la région endommagée du gène, et doivent être pris en charge pour permettre une réparation efficace (S. COHEN et al. 2018 ; MARNEF et Gaëlle LEGUBE 2021).

### 1.2.5 L'inhibition de la transcription par les DSB

Différentes études ont montré que les gènes actifs présents à proximité, en *cis* des *DSB*, pouvaient aussi être inhibés. Une étude omique permettant la quantification de l'expression des gènes (voir partie 1.3.5), montre que la baisse d'expression des gènes serait dépendante de la distance génomique aux *DSB* (IANNELLI et al. 2017). Cependant, il a également été montré que les gènes proches des *DSB*, contenus dans le domaine gamma ne sont pas réprimés suite à l'induction de *DSB* (IACOVONI et al. 2010 ; J. KIM et al. 2016). Ces gènes seraient protégés par des mécanismes d'organisation de la chromatine, et notamment les cohésines (CARON, Francois AYMARD et al. 2012).

L'effet des *DSB* sur l'expression des gènes est un phénomène toujours mal compris, faisant intervenir des protéines associées à l'organisation tri-dimensionnelle de la chromatine. Il est clair que l'état local de la chromatine impacte les mécanismes de réparation des *DSB*, mais la façon dont l'organisation tri-dimensionnelle du génome influe sur ces mécanismes est toujours à étudier.

### 1.2.6 Le rôle du 3D dans la réparation des DSB

#### 1.2.6.1 La conformation initiale de la chromatine impacte la réparation des *DSB*

Il a été montré dans différentes études que la production et la réparation des *DSB* programmées dépendaient de la conformation 3D, notamment celles intervenant pendant le brassage génétique mais aussi pendant le processus de création d'Imunoglobuline (Voir Figure 1.20 B et C) (ARNOULD et Gaëlle LEGUBE 2020). En outre, la proximité spatiale d'un *locus* endommagé par rapport à sa séquence homologue est un élément qui impacte fortement sa capacité à être réparé correctement (C.-S. LEE et al. 2016).

L'étude de  $\gamma$ H2AX par des méthodes omiques (IACOVONI et al. 2010 ; MARNEF et Gaëlle LEGUBE 2017) a également mis en évidence le fait que la propagation de cette modification d'histone sur des régions de ~1 Mb était fortement reproductible selon les expériences. Qu'elle se propageait de façon asymétrique autour de la *DSB* et qu'elle était limitée par des frontières.

Cela suggère que la formation des foyers de réparation pourraient dépendre de la conformation tri-dimensionnelle initiale du génome. En effet, le génome étant formé par une succession de *TAD*, délimités par des éléments insulateurs, on peut très bien imaginer que ce sont ces domaines 3D, qui, par des

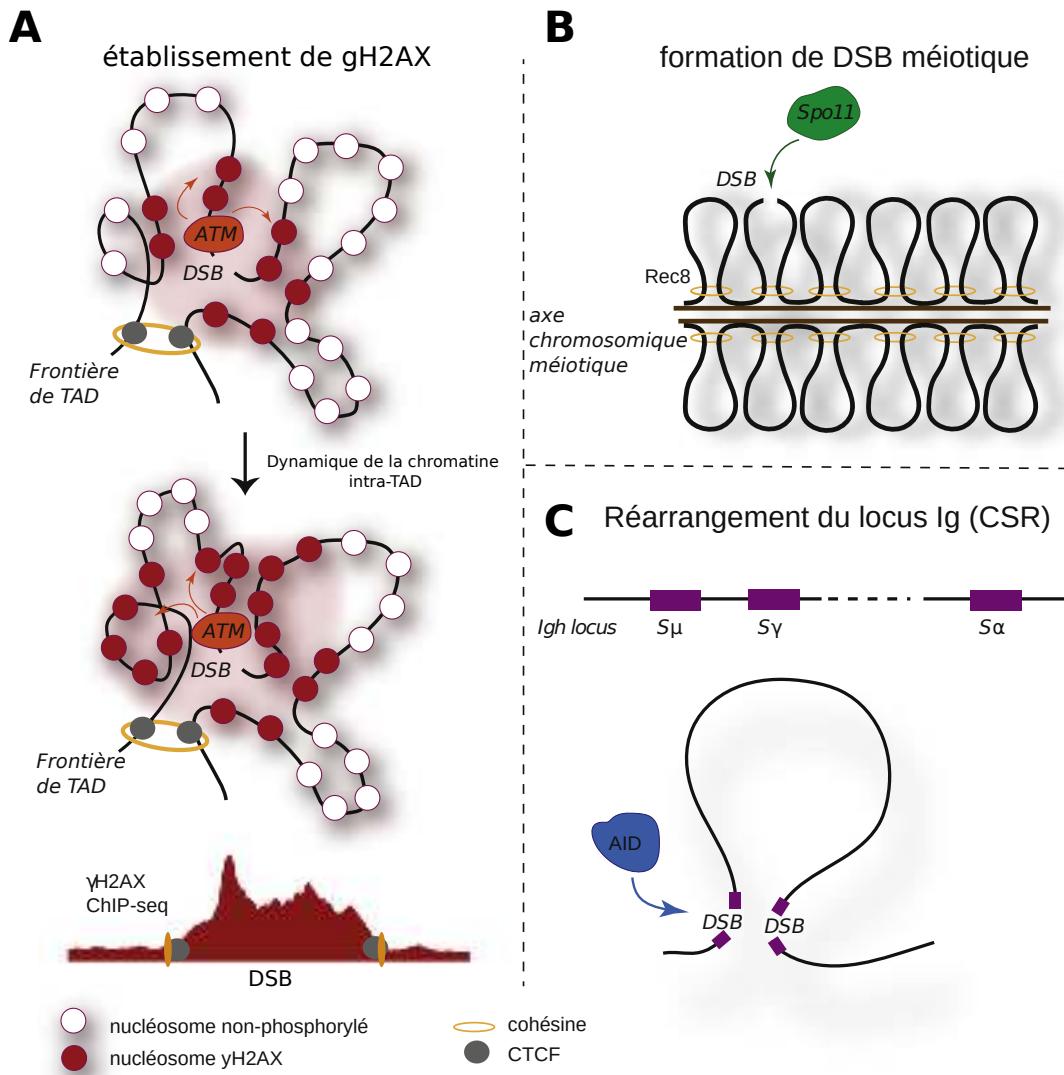


FIGURE 1.20 – Contribution de la chromatine initiale sur la production et la réparation des DSB. **A** La conformation initiale du génome pourrait dicter la création des foyers de réparation. Ici, H2AX est phosphorylé par ATM à l'intérieur du TAD grâce à accès à l'ADN via sa proximité spatiale, et est limité par les frontières insulatrices. **B** La formation des DSB nécessaires au brassage génétique par la protéine Spo11 dépend d'une structuration précise de l'ADN en boucles pendant la prophase. **C** Le renouvellement des anticorps par la création de nouveaux isotypes d'immunoglobulines est induit par l'induction de multiples DSB qui permettent à l'ADN de se réarranger via des interactions longue-distances de la chromatine. Adapté de (ARNOULD et Gaëlle LEGUBE 2020).

interactions intra-TAD permettent la phosphorylation de  $\gamma$ H2AX par ATM sur une aussi grande distance génomique, alors que la protéine reste localisée au niveau de la *DSB* (Voir Figure 1.20 A).

### 1.2.6.2 Le paysage chromatinien est modifié après dommages

Après l'induction de *DSB*, la chromatine subit des modifications à différents niveaux jusqu'à une échelle d'environ 1 Mb, notamment par la phosphorylation du variant H2AX, mais aussi de l'ubiquitinylation des histones H1, puis H2A et H2AX (Thomas CLOUAIRE et al. 2018). Le fait d'établir des modifications sur un si grand domaine de la chromatine et d'agir sur les histones, comme H1, pourrait agir fortement sur l'état local de compaction de la chromatine, et sur sa conformation.

L'organisation tri-dimensionnelle autour des *DSB*, son rôle et la façon dont elle est modifiée suite aux modifications de la chromatine est encore très mal compris. Ces modifications à l'échelle d'un *TAD* pourrait permettre à la chromatine d'être plus accessible aux protéines de la réparation, et réguler sa mobilité au sein du noyau (voir partie 1.2.1.1.3). Une modification de la conformation 3D de la chromatine pourrait également permettre de protéger et d'isoler la *DSB* de son environnement, afin d'empêcher l'apparition de translocations, et donc, de favoriser l'intégrité du génome.

Il existe déjà de nombreuses études montrant que la compaction et la mobilité de la chromatine à proximité des *DSB* est différente (MARNEF et Gaëlle LEGUBE 2017; ZIMMER et FABRE 2019). Cependant, la majorité de ces études se basent sur des expériences réalisées par imagerie, et nos connaissances sur les changements 3D de la chromatine à l'échelle de l'ADN sont peu nombreuses. Une étude réalisée dans notre laboratoire sur l'analyse des données 3D via des méthodes omiques (voir partie 1.3.6), a montré que les interactions physiques des régions endommagées changeaient au profit d'interactions à plus longue distance, comme pour isoler l'ADN endommagé (François AYMARD, AGUIRREBENGOA et al. 2017). Cependant, cette étude n'était pas suffisamment résolutive pour l'étude des changements post-dommages des boucles, qui restent toujours à étudier.

Il est très probable que la cohésine et CTCF aient un rôle important dans la ré-organisation des boucles et de l'architecture 3D en général après induction de dommages. De nombreuses études ont démontrés que ces éléments étaient recrutés à proximité des *DSB* (CARON, Francois AYMARD et al. 2012; MEISENBERG et al. 2019; LANG et al. 2017), cependant, une étude complète sur le rôle de la conformation 3D et sur l'importance des protéines architecturale du génome sur la formation des foyers de réparation permettrait de mieux comprendre la nature et les fonctions du 3D impliquées dans la réparation.

### 1.2.6.3 Les *DSB* sont mobiles dans le noyau et peuvent se regrouper

En plus des changements apportés en *cis* des *DSB*, le génome subit une réorganisation globale de la chromatine après endommagement de l'ADN. En effet, il a été montré que les *DSB* incapables d'être réparées pouvaient migrer vers la périphérie du noyau, chez la levure (OZA et al. 2009; MARNEF et Gaëlle LEGUBE 2017). Les *DSB* sont également capables de se regrouper et de former des *clusters*, de façon visible par imagerie mais également par des méthodes omiques (François AYMARD, AGUIRREBENGOA et al. 2017; KILIC et al. 2019). En effet, après dommages, le nombre d'interactions entre les régions endommagées augmente, suggérant un *clustering* des *DSB* dans le noyau (voir Figure 1.21 A). Les

données de conformation 3D par Hi-C montrent que ce *clustering* est possible, même sur des chromosomes différents.

En outre, les interactions dépendent de l'enrichissement local en  $\gamma$ H2AX (voir Figure 1.21 B), suggérant que ce sont les cassures qui accumulent le plus  $\gamma$ H2AX qui se regroupent le plus. En effet, dans la même étude, il a été montré que ce sont les *DSB* préférentiellement réparées par *HR*, et donc en attente de la chromatide soeur pour la réparation, qui montrent l'augmentation d'interaction la plus forte. Ces *DSB* deviennent persistantes, et  $\gamma$ H2AX s'accumule au sein du domaine endommagé.

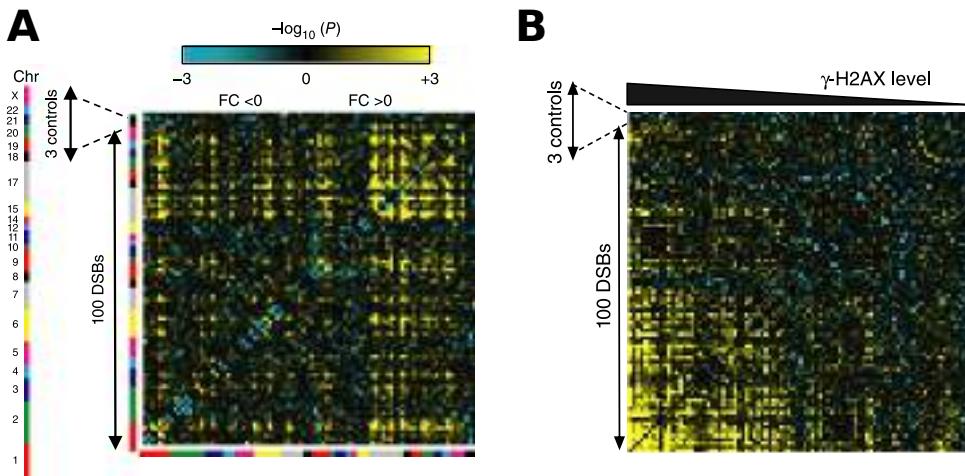


FIGURE 1.21 – Augmentation significative des fréquences d'interaction en Hi-C (capture) après induction de dommages, sur 100 positions endommagées et 3 régions contrôles. **A** Les DSB sont capables d'interagir en *cis*, mais aussi en *trans*, sur des chromosomes différents. **B** Trier les régions endommagées par leur niveau en  $\gamma$ H2AX montre une relation entre l'intensité du signal et les changements d'interaction après dommage. De François AYMARD, AGUIRREBENGOA et al. (2017).

Il est possible que le regroupement des *DSB* induise la création d'un compartiment pour favoriser le recrutement des protéines de la réparation (LISBY, MORTENSEN et ROTHSTEIN 2003). Le regroupement des *DSB* favorisant la réparation par *HR* (François AYMARD, AGUIRREBENGOA et al. 2017), il est également possible que celui-ci favorise la recherche d'homologie (MINÉ-HATTAB et ROTHSTEIN 2012). Cependant, le *clustering* a également été observé en phase G1 du cycle cellulaire (François AYMARD, AGUIRREBENGOA et al. 2017), ce qui suggère un autre rôle que la recherche d'homologie. Les fonctions potentielles du *clustering* sont nombreuses, mais les mécanismes permettant le regroupement de régions endommagées, au risque de causer des translocations (ROUKOS et al. 2013), sont encore mal compris.

#### 1.2.6.4 Les mécanismes potentiels permettant le regroupement des *DSB*

Les mécanismes permettant le regroupement des *DSB* pourraient provenir d'une mécanique active ou directionnelle, via l'intervention par exemple des microtubules et/ou de l'activité du cytosquelette ou du nucléosquelette. En effet, des régulateurs du cytosquelette, WASP et ARP2/3 notamment, sont recrutés pendant la réparation par *HR* des *DSB* (SCHRANK et al. (2018), voir Figure 1.22 A). En outre, il a également été montré que le complexe LINC, connectant le cytosquelette à la chromatine, contribue à la mobilité des *DSB* dans le noyau, et à leur regroupement (LOTTERSBERGER et al. (2015), voir Figure 1.22 B).

Une autre possibilité serait de considérer le regroupement des *DSB* comme un mécanisme passif produit par séparation de phase (voir Figure 1.22 C). En effet, les nombreuses modifications locales de la chromatine, à l'échelle d'un *TAD*, pourraient conduire à la formation d'un nouveau compartiment spécifique regroupant les régions endommagées ayant des propriétés similaires. Il a été montré que 53BP1 permettait à la chromatine enrichie en cette protéine de former un compartiment isolé du reste du noyau (KILIC et al. (2019), voir partie 1.2.1.1.3). 53BP1 étant également recruté à l'échelle du domaine entier, comme  $\gamma$ H2AX, il est possible que son enrichissement local au sein de la chromatine endommagée favorise le regroupement des *DSB*. En outre, l'enrichissement en phase G1 du cycle cellulaire de 53BP1 (François AYMARD, AGUIRREBENGOA et al. 2017 ; Thomas CLOUAIRE et al. 2018), permettrait d'expliquer le *clustering* observé des *DSB* pendant cette phase du cycle cellulaire.

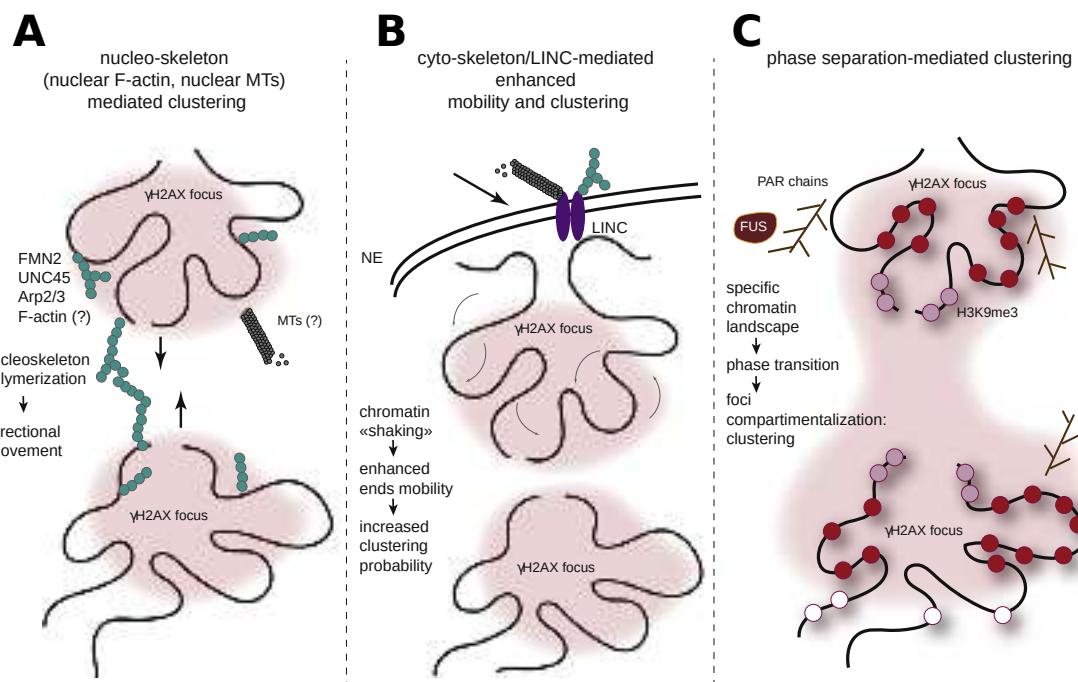


FIGURE 1.22 – Regroupement des DSB (*DSB Clustering*) dans le noyau. Le *clustering* de DSB a été observé par de multiples méthodes, omiques et microscopique. Cependant, les mécanismes mettant en place ce *clustering* sont encore à déterminer. **A** Des éléments du nucléosquelette faisant intervenir de l'actine ou des microtubules pourraient contribuer à la mobilité des DSB et à leur regroupement de manière active et directionnelle. **B** Le cytosquelette pourrait également contribuer à la mobilité des DSB et à leur regroupement par des transmissions de forces du cytosquelette à la chromatine via le complexe LINC. **C** Le *clustering* pourrait également s'expliquer par la formation d'un nouveau compartiment constitué de chromatine endommagée, via séparation de phase. De (ARNOULD et Gaëlle LEGUBE 2020).

La façon dont les *DSB* sont regroupés au sein du noyau est encore mal comprise, et peut provenir de différents mécanismes, actifs et passifs. Ces mécanismes peuvent potentiellement collaborer ou bien entrer en compétition. De manière générale, les connaissances des différentes fonctions de la chromatine agissant sur les *DSB* et sur le choix de leurs réparations sont de mieux en mieux comprises. Cependant le rôle de la conformation 3D de la chromatine sur la réparation des *DSB*, et comment l'endommagement de l'ADN est capable d'induire des changements de conformation sont des phénomènes encore très mal compris.

Pour répondre à ces questions, on peut utiliser de multiples méthodes omiques permettant d'étudier la

conformation 3D et la composition de la chromatine, les différentes protéines de réparation architecturales, et l'expression des gènes. Intégrer ces données dans une seule étude permettrait d'obtenir une vue d'ensemble, au sein d'une population cellulaire et de mieux comprendre les mécanismes qui régulent, structure et réparent le génome.

## **1.3 La génomique, ou l'étude de l'ensemble du génome**

La Génomique est le terme qui désigne l'étude de l'ensemble ( $\Omega$ ) du génome, soit l'ensemble des séquences d'ADN d'un individu. Au moment de la création de la revue du même nom, *Genomics* (McKUSICK et RUDDLE 1987), ce terme définissait la cartographie, le séquençage et la position des gènes sur les chromosomes. À partir du *Human Genome Project* (I. H. G. S. CONSORTIUM 2004), la génomique s'est étendue à l'étude de l'ensemble de l'ADN, codant ou non codant, d'un individu. La génomique est pluridisciplinaire, elle inclue une part liée à la Biologie Moléculaire (par le séquençage et la préparation des librairies) et la Biologie Computationnelle ou Bio-informatique, menant à la création d'équipes associant biologistes, informaticiens, statisticiens et physiciens.

Cette discipline a permis de révolutionner l'étude de la biologie moléculaire en apportant une vision pangénomique ou *genome-wide*. Au cours du temps, de nouvelles méthodes ont été développées pour étudier différents parties du génome : la Transcriptomique, l'étude de l'ensemble des transcrits d'un individu (1.3.5), le Protéome, qui étudie les protéines, l'Épigénomique, pour l'étude globale des modifications épigénétiques de l'ADN et de la chromatine (1.3.2) et son accessibilité (1.3.4), et des méthodes comme le 3C, qui permettent d'étudier la conformation 3D du génome (1.3.6).

Parmi les méthodes qui permettent une approche omique, les techniques de séquençage à haut débit ou *next generation sequencing (NGS)* sont très populaires. Les parties suivantes décriront le traitement et l'analyse des données produites par ces méthodes.

### **1.3.1 Le séquençage à haut débit (Next Generation Sequencing, NGS)**

La technologie des *NGS* s'est rapidement popularisée après le projet *Human Genome Project* (I. H. G. S. CONSORTIUM 2004). Les *NGS* sont définis par une technologie de séquençage en parallèle, très rapide, évolutive et performante. Elle permet de déterminer la séquence de millions de fragments provenant de génomes entiers ou de régions cibles d'ADN ou d'ARN (voir Figure 1.23). Les *NGS* sont des sciences omiques, permettant aux scientifiques d'étudier les systèmes de façon très résolutive pour l'ensemble du génome. À l'inverse du séquençage *Sanger* utilisé pendant le *Human Genome Project*, les *NGS* sont très accessibles ( $\sim 1000 \$$  pour séquencer un génome entier<sup>1</sup>), ce qui leur permet de concurrencer d'autres méthodes d'analyse comme la microscopie.

La méthode grandement majoritaire pour séquencer de l'ADN est développée par Illumina et se base sur le séquençage par synthèse des *short reads* (voir Figure 1.24), permettant de séquencer des *reads* de quelques centaines de paires de bases. Mais il existe aussi la méthode des *long reads*, qui permet de séquencer des molécules d'ADN de plusieurs kilobases.

1. source : <https://www.genome.gov/about-genomics/fact-sheets/DNA-Sequencing-Costs-Data>

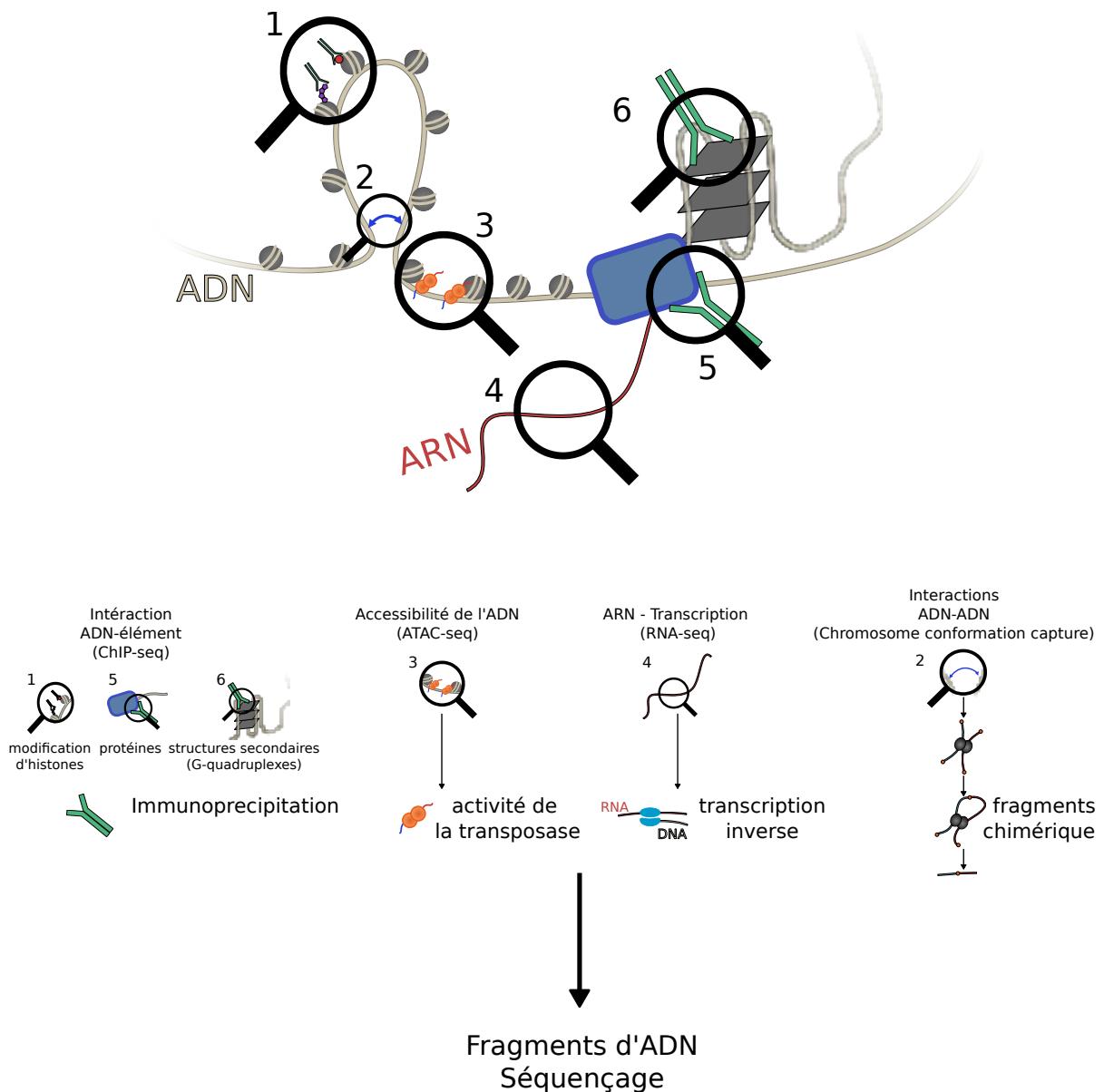


FIGURE 1.23 – Extraction de fragments d'ADN provenant de différentes sources pour le séquençage à haut débit. Le **ChIP-seq** permet de cartographier les interactions ADN-proteines via immunoprécipitation (partie 1.3.2). L'**ATAC-seq** permet de cartographier l'accèsibilité et l'état de compaction de l'ADN par l'activité d'une transposase (voir partie 1.3.4). Le **RNA-seq** permet de quantifier l'expression des gènes par extraction et reverse-transcription des fragments d'ARN matures (voir partie 1.3.5). Les méthodes **3C** dont le **Hi-C** permettent de quantifier les interactions ADN-ADN grâce à la création de chimères d'ADN provenant de deux *loci* (voir partie 1.3.6).

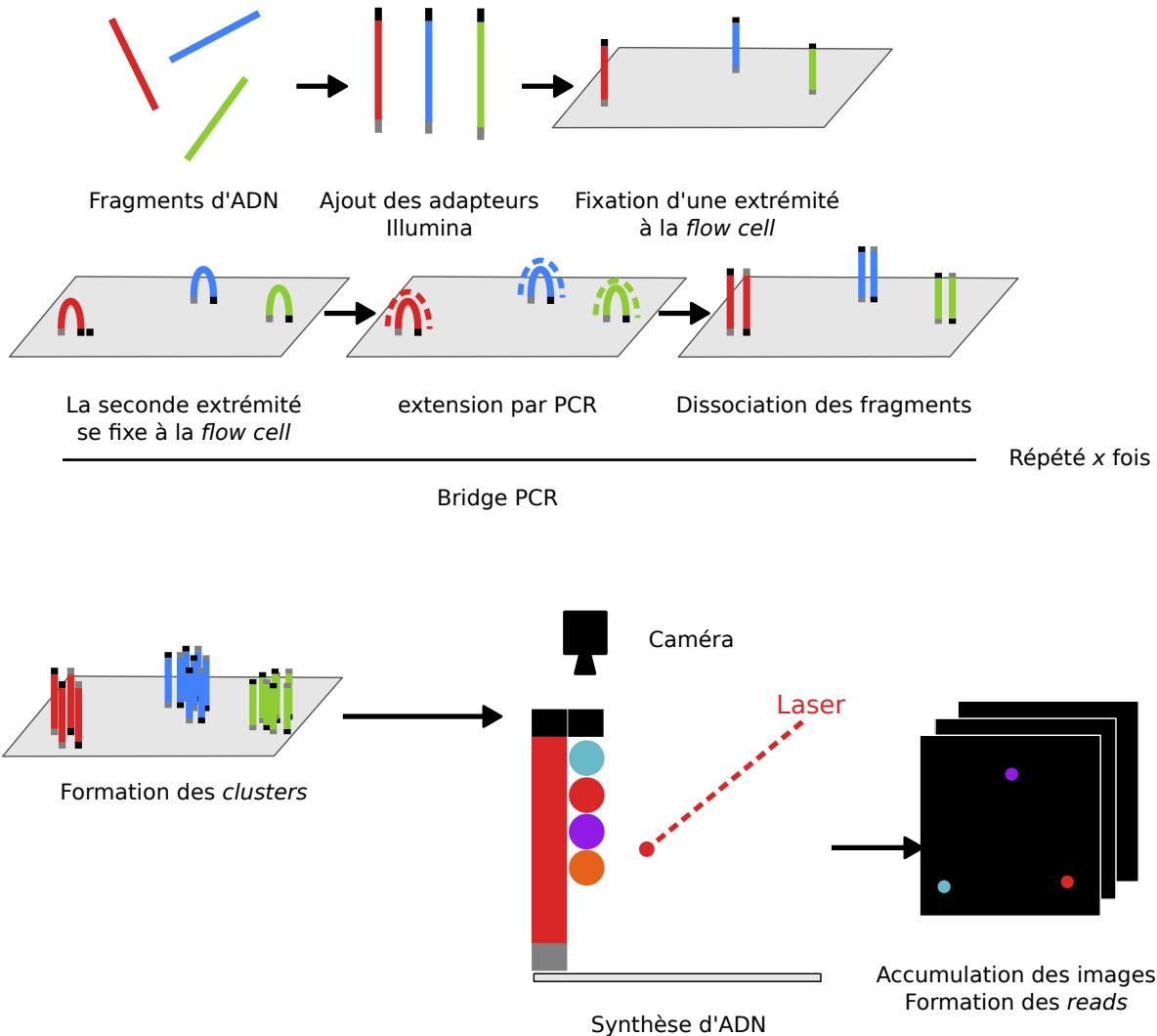


FIGURE 1.24 – Séquençage Illumina par synthèse des *short reads*. Les fragments d'ADN provenant de diverses librairies sont étiquetées afin d'être identifiés et de se fixer sur le support. L'amplification des fragments se fait sur le support par un procédé appelé *bridge PCR*, ce qui permet la formation de *clusters*. Des nucléotides couplés à un fluorophore permettent de capturer par caméra la synthèse du brin complémentaire du fragment. Chaque image représente l'incorporation d'un nucléotide pour chacun des *clusters* sur le support. Ce procédé est répété, ce qui produit les *reads*.

### 1.3.1.1 Préparation des librairies d'ADN

La préparation des librairies d'ADN pour le séquençage est très dépendante de l'expérience à réaliser. Pour séquencer le génome entier d'un individu (*Whole Genome Sequencing*), on fragmente l'ADN contenu dans les noyaux des cellules soit via digestion enzymatique ou par des moyens mécaniques (comme des ultrasons par exemple). Les fragments sont ensuite extraits et on ajoute différentes étiquettes, les *adapters* qui permettent à la fois d'identifier la séquence et de fixer le fragment sur le support. Cette étape est presque toujours suivie d'une amplification de la librairie par réaction en chaîne par Polymérase ou *PCR*. Ceci permet d'utiliser les *primers* fixés aux fragments afin d'amplifier les régions d'intérêts. Selon la méthode, une partie seulement du matériel génétique peut être amplifiée (voir par exemple partie 1.3.6.2). Cette amplification par *PCR* permet d'obtenir suffisamment de matériel génétique pour le séquençage, mais introduit des biais, qu'on appelle *duplicates PCR*, qui devront souvent être supprimés de l'analyse.

La méthode développée par Illumina est la méthode par synthèse (*Séquençage : La technologie d'Illumina* s. d.), qui combine l'incorporation par l'ADN polymerase d'un nucléotide couplé à un fluorophore, et la microscopie par *total internal reflection fluorescence* (TIRF). Selon la machine utilisée, la méthode de séquençage et les fluorophores peuvent varier. Pour la méthode à 4 canaux (HiSeq), chaque nucléotide possède son fluorophore. Pour la méthode à 2 canaux (NextSeq), l'adénine (A) est couplée à deux fluorophores, la thymine (T) et la cytosine (C) à chacune un, et la guanine (G) à aucun. La base est ensuite incorporée à la séquence complémentaire du fragment et le fluorophore bloque l'incorporation d'une autre base. Chaque *cluster* émet ensuite une couleur correspondant à la base et on clive ensuite le fluorophore, ce qui permet à une nouvelle base d'être incorporée. L'étape d'incorporation, d'émission de la couleur, de clivage du fluorophore suivit par une étape de lavage de la *flow cell* correspond à un tour de séquençage et le processus est répété (voir Figure 1.24).

À la fin, on obtient une suite d'images qui forment les *reads*. L'identification des *adapters* permet d'associer un *read* à son expérience. En cas de séquençage *paired-end*, les deux extrémités des fragments sont séquencées et associées entre elles. À l'inverse, le *single-end* ne séquence qu'une seule extrémité du fragment. L'étape séparant les *reads* par expérience s'appelle le *démultiplexage*, et permet d'utiliser le séquenceur au maximum de sa capacité (jusqu'à 500 millions de reads pour le NextSeq) en combinant les expériences de différents fournisseurs.

### 1.3.1.2 Applications des NGS en génomique

La méthode du séquençage *NGS* peut s'appliquer, en plus du génome entier (*Whole genome sequencing*), à des régions plus spécifiques du génome. Par exemple, on parle d'*exome sequencing* lorsque l'on enrichit les séquences d'ADN issues des parties codantes du génome. On peut également utiliser des méthodes de séquençage épigénomique, comme le ChIP-seq, pour identifier les sites de liaisons d'une protéine (ROBERTSON et al. 2007), l'ATAC-seq pour déterminer l'accessibilité de la chromatine (BUENROSTRO et al. 2013) ou encore le méthyl-seq qui mesure le niveau de méthylation de l'ADN (BRUNNER et al. 2009). D'autres méthodes permettent de séquencer les ADN complémentaires des transcrits, comme le RNA-seq, on parle alors de transcriptomique. Enfin, des méthodes ont permis d'identifier des liaisons longues distances entre deux *loci* d'ADN, et d'étudier la conformation 3D du génome, comme le Hi-C (LIEBERMAN-AIDEN et al. 2009 ; S. S. RAO et al. 2014).

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

Dans tous les cas, on obtient des données de comptage de *reads*, qui nécessitent d'utiliser des méthodes statistiques particulières pour les analyser, suivant par exemple une loi de Poisson (pour le *peak calling*, décrit partie 1.3.2.1) ou Binomiale Négative (pour l'analyse différentielle des gènes, décrit partie 1.3.5.2).

#### **1.3.1.3 Le génome de référence**

Les séquences produites par séquençage ou *reads* étant issus de fragments d'ADN, on ne peut les identifier directement. Pour cela, on utilise un génome de référence, lui-même issu du séquençage, mais qui a été reconstruit entièrement par un procédé d'assemblage. On dispose donc d'une séquence continue et globalement identifiée (certaines régions sont difficiles à identifier, notamment les régions répétées), ce qui nous permet, par un processus qu'on appelle l'alignement (voir partie 1.3.1.4.2.1), d'identifier la position des *reads* sur le génome.

Le génome de référence humain est composé des séquences de l'ensemble des 24 chromosomes, séquencées une première fois lors du *Human Genome Project*, puis mis à jour régulièrement. Il est disponible au format fasta sur de nombreuses bases de données tel que l'*University of California Santa Cruz (UCSC)*<sup>2</sup>, le *National Center for Biotechnology Information (NCBI)*<sup>3</sup> ou sur la base de données *Ensembl*<sup>4</sup> de l'*European Bioinformatic Institute (EMBL-EBI)*.

Plusieurs versions d'un même génome de référence peuvent exister. Chez l'Homme, les deux versions les plus utilisées sont *GRCh37* (ou *hg19*) et *GRCh38* (ou *hg38*), assemblés respectivement en 2009 et en 2013.

#### **1.3.1.4 Traitement des données NGS**

**1.3.1.4.1 Le format FastQ** Les *reads* sont construits par une suite d'images directement issues du séquenceur. Cette suite est analysée et chaque *read* est identifié et reconstruit comme une chaîne de caractères dans un fichier texte. Le format le plus répandu permettant de stocker des séquences biologiques est le format fasta :

```
>ID_Séquence  
AGGCTTGGATTGCTTGGATTAGTTA
```

Il n'est cependant pas adapté aux données provenant d'un séquenceur qui nécessitent d'autres informations. Le format FastQ a été développé dans ce but. Il contient un identifiant de séquence unique, la séquence et son contrôle qualité encodé en ASCII de façon à faire correspondre à un caractère une base :

```
@NB502089:1169:HMGTVBGXJ:3:12612:11537:2617 2:N:0:CCAACACT+CAATGTGG  
TGCCAACATATTGTGCTAATGAGTCGCCTCGTCTGTCTTATATTACCGCAAACCCAAAAAG  
+  
EEEEEEEEE6EEAEAEEEEEE</AAE/EEEEE<EAE/EEEEEEEEEEEEE/EEEA//EA
```

2. <https://hgdownload.soe.ucsc.edu/downloads.html>

3. [https://www.ncbi.nlm.nih.gov/assembly/GCF\\_000001405.13/](https://www.ncbi.nlm.nih.gov/assembly/GCF_000001405.13/)

4. <http://grch37.ensembl.org/info/data/ftp/index.html>

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

L'identifiant de la séquence, qui commence toujours par @ contient des informations sur le séquençage, comme l'identifiant de la machine NB502089, le numéro du run HMGTVBGXJ et les coordonnées sur la flow cell. Ici, l'information utile est surtout contenue dans la deuxième partie 2:N:0:CCAACACT+CAATGTGG, où le 2 identifie à quel membre de la paire appartient le *read* (dans le cas de données *paired-end*). On trouvera donc un identifiant identique dans le fichier correspondant au *read 1* de la paire, permettant l'identification et l'association de la paire pendant l'alignement.

La dernière ligne correspond à la qualité de l'alignement, encodée en ASCII +33, correspondant au score  $PHRED - 33$ . Le score PHRED est associé à la probabilité d'erreur de séquençage de la manière suivante (1.1) :

$$Q = -10 \times \log_{10} E \quad (1.1)$$

$$E = 10^{-\frac{Q}{10}} \quad (1.2)$$

On peut donc retrouver la probabilité d'erreur  $E$  de la base en utilisant la deuxième formule (1.2). Par exemple pour la base 1 (1.3) :

$$Q(G) = ASCII(A) - 33 = 32 \quad (1.3)$$

$$A(G) = 1 - 10^{-\frac{32}{10}} = 0.999 \quad (1.4)$$

Ce qui nous permet de déterminer que la précision de séquençage (1.4) de cette base est de 99.9%.

#### 1.3.1.4.2 Alignement de *short reads* sur génome de référence

**1.3.1.4.2.1 Alignement d'ADN** Les algorithmes d'alignement de séquences d'ADN sont des méthodes bio-informatiques historiques, en usage depuis bien avant le séquençage entier du génome humain. Ceux-ci permettent d'aligner deux séquences globalement comme l'algorithme de *Needleman & Wunch* (NEEDLEMAN et WUNSCH 1970), mais aussi de manière locale, comme *Smith & Waterman* (SMITH et WATERMAN 1981) et surtout *BLAST* pour *Basic Local Alignment Search Tool* (ALTSCHUL et al. 1990).

Même si on peut appliquer ces outils sur des séquences provenant d'un séquenceur (*reads*), l'alignement d'un nombre important de séquences sur un génome de référence (3 Gigabases ou Gb chez l'Homme) serait trop chronophage, et d'autres méthodes ont dû être développées. L'objectif de ces méthodes est d'effectuer des requêtes sur un texte (le génome de référence) pour chacune des sous-séquences (*reads*) contenues dans notre FastQ. L'algorithme doit avoir une complexité constante en  $O(|W|)$  où  $W$  représente notre requête de façon à ne pas être trop chronophage, et le moins coûteux en mémoire possible. Les algorithmes les plus utilisés sont *BWA* (H. LI et DURBIN 2009) pour *Burrows-Wheeler Alignment tool* et *bowtie2* (LANGMEAD, TRAPNELL et al. 2009 ; LANGMEAD et SALZBERG 2012).

**1.3.1.4.2.2 Alignement d'ARNs** Les algorithmes classiques d'alignement de *reads* ne fonctionnent pas sur des données provenant d'ARN à cause du mécanisme de l'épissage des transcrits (voir partie

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

1.1.2). Concrètement, le transcript est non-contigu dans le génome, du fait que seul les parties codantes ou *exon* sont conservées dans l'ARN mature (voir partie 1.1.2). Le séquençage par *short reads* fait qu'il est impossible de séquencer la molécule entière d'ARN (ou plutôt son ADN complémentaire), et qu'il est donc nécessaire d'implémenter une méthode qui prend en compte ce mécanisme. Plusieurs outils ont été développés dans ce but, et les plus utilisés sont *TopHat2* (D. KIM et al. 2013) et *STAR* (DOBIN et al. 2013). La principale différence entre les deux étant que *Tophat2* construit un transcriptome de référence en utilisant les positions des jonctions d'épissage des gènes (à partir d'une annotation), puis aligne les *reads* sur ce transcriptome en utilisant *bowtie2* (On retrouve un alignement classique vu partie 1.3.1.4.2.1).

*STAR* lui fonctionne complètement différemment, et utilise une méthode de *Maximal Mappable Prefix (MMP)*. Un *MMP* pour un *read* donné consiste à aligner la sous-chaîne la plus longue possible du *read* sur le génome de référence jusqu'à ce que les caractères ne correspondent plus, puis de mapper à nouveau la partie restante, et ainsi de suite jusqu'à ce que le *read* soit entièrement aligné.

**1.3.1.4.3 Traitement des données d'alignement** Une fois les données alignées sur le génome de référence, l'outil enregistre les *reads* dans un format de fichier **SAM** pour *sequence alignment map* (H. LI, HANDSAKER et al. 2009), ou son équivalent encodé en binaire, le format **BAM** (*Binary Alignment Map*). C'est un format de fichier texte délimité par des tabulations, où chaque *read* est enregistré sur une seule ligne (à l'inverse du **fastQ** qui l'enregistre sur 4 lignes). Chaque enregistrement contient les informations du *read*, comme sa séquence et son identifiant, mais aussi sa position génomique, sa qualité d'alignement, et la présence éventuelle d'erreurs (*mismatches* ou *indels*). Par exemple, le *read* montré précédemment au format **fastQ** dans la partie 1.3.1.4.1, sera enregistré en une seule ligne de cette façon :

```
NB502089:1169:HMGTVBGXJ:3:12612:11537:2617
163 chr2L    4829      40  65M =    5142      378
TGCCAACATATTGTGCTAATGAGTGCCCTCGTCTGTCTTATATTACCGCAAACCCAAAAAG
EEEEEEEEE6EEAEAAAAAAE</AAE/EEEEEE<EAE/EEEEEEEEEAEAA//EA
```

Cette entrée indique qu'il a été mappé sur le chromosome 2L (de la drosophile) à la position 4829, que sa qualité d'alignement (MAPQ) est de 40, ce qui signifie qu'il est aligné à 100% (via l'équation (1.2)). On dispose aussi de la position du deuxième *read* de la paire 5142 et de la taille du fragment 378. On retrouve également le score de qualité de séquençage encodé en *ASCII*.

La valeur 65M correspond au *CIGAR*, qui indique comment le *read* a été aligné sur le génome de référence. Ici, les 65 bases ont toutes été alignées, mais dans certains cas, on peut retrouver la position des *mismatches/indels* à partir du *CIGAR*, par exemple : 12M1I52M indique une insertion à la position 12 dans le *read* de taille 65 par rapport à la référence.

Enfin, la valeur 163 est une combinaison en hexadécimale (base 16) de différentes valeurs *FLAG*. Ici, cette valeur est la combinaison de 4 éléments qui indiquent que le *read* est pairé (correspond à la valeur 0x01), que la paire est correcte (0x02), qu'il s'agit du deuxième élément de la paire (0x80) et que l'autre *read* de la paire est aligné sur le brin *anti-sens* (0x20), celui-ci est donc aligné sur le brin *sens* (voir Table 1.1).

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

TABLE 1.1 – Combinaisons possibles des différents FLAGs dans un fichier SAM et leurs valeurs hexadécimale associée. Les valeurs combinées donnant 163 sont soulignées en rouge.

Décimale	Hexadécimale	Description
1	0x1	<b>Read paired</b>
2	0x2	<b>Read mapped in proper pair</b>
4	0x4	Read unmapped
8	0x8	Mate unmapped
16	0x10	SEQ being reverse complemented
<b>32</b>	<b>0x20</b>	<b>SEQ mate is reverse complemented</b>
64	0x40	The first in pair
<b>128</b>	<b>0x80</b>	<b>The second in pair</b>
256	0x100	Secondary alignment
512	0x200	Not passing filters, such as platform/vendor quality controls
1024	0x400	PCR or optical duplicate
2048	0x800	Supplementary alignment

Les *FLAGs* sont fréquemment utilisés lorsque l'on veut séparer la librairie selon le brin d'origine (voir partie 1.3.5). De manière générale, c'est le brin complémentaire à la molécule d'origine qui est séquencé. Pour retrouver la paire de *reads* originaire du brin *reverse*, on sélectionne les *reads* avec les *FLAGs* 80 (*read reverse strand*, 0x10, *first in pair*, 0x40) et 160 (*mate reverse strand*, 0x20, *second in pair*, 0x80). Les *reads* provenant du brin *forward* sont eux extraits avec les *FLAGs* 96 (*mate reverse strand*, 0x20, *first in pair*, 0x40) et 144 (*read reverse strand*, 0x10, *second in pair*, 0x80).

Ce format est facilement manipulable en ligne de commande via différents outils, dont *Samtools* (H. LI, HANDSAKER et al. 2009) et *Picard* (*Picard toolkit* 2019). *Samtools* permet de convertir un SAM en BAM, de filtrer les *reads* en se basant sur leurs coordonnées génomiques ou bien leur *FLAG* ou *MAPQ* avec `samtools view`. Il permet aussi de trier (numériquement ou par identifiant) avec `samtools sort`, supprimer/marquer les dupliques *PCR* avec la commande `samtools markdup` et même d'identifier des variants (comme des *Single Nucleotide Polymorphism* ou *SNP*) avec la commande `samtools mpileup`.

**1.3.1.4.4 Couverture locale du génome** La couverture locale du génome permet de quantifier le résultat du séquençage et de l'alignement. Pour une région donnée du génome, on peut calculer le nombre de *reads* alignés, ce qui donne une mesure quantitative qu'on peut interpréter différemment selon la méthode utilisée. Pour le ChIP-seq par exemple (voir partie 1.3.2), ce signal sera interprété comme le niveau d'enrichissement local de la protéine d'intérêt dans cette région. On peut soit encoder ce signal à l'échelle de la paire de base, ou diviser le génome en sous-régions (des *bins*) et agréger (par moyenne ou somme) le signal dans ces régions. Différents formats permettent d'encoder le signal et sa position sur le génome.

**1.3.1.4.4.1 Le format *wiggle*** Le format *Wiggle* ou *Wig* et son équivalent binaire le *bigWig* permettent d'enregistrer le nombre de *reads* (ou autre signal) sur le génome de référence. Le *Wig* un format texte où chaque ligne correspond à une donnée quantitative sur le génome, qui peut être encodée de deux façons, en *variableStep* ou en *fixedStep* :

```
variableStep chrom=chr2
300701 11
300702 22
300703 33

fixedStep chrom=chr2 start=300701 step=3
11
22
33
```

Ces lignes sont deux façons différentes d'indiquer qu'aux positions `chr2:300701`, `chr2:300702` et `chr2:300703`, 11, 22 et 33 *reads* ont été alignés.

**1.3.1.4.4.2 Le format *bedGraph*** Le format *bedGraph* est une autre méthode de stockage de données qui ajoute au format *bed* (voir partie 1.3.2.1.2) une colonne numérique correspondant à du signal, par exemple :

```
chr2    300701    300701    11
chr2    300702    300702    22
chr2    300703    300703    33
```

Ces trois lignes enregistrent la même information que pour l'exemple précédent du *Wig*. Cette méthode est conseillée lorsque l'on dispose de beaucoup de données manquantes (*sparse*) et peu de régions génomiques annotées, car c'est un format texte, à l'inverse du *Wig* qui peut être compressé en *bigWig*.

**1.3.1.4.4.3 Visualisation des données génomiques** Pour visualiser les données génomiques, on peut utiliser un explorateur de génome ou *Genome Browser*. Ces applications contiennent de nombreux génomes de références et reconnaissent les différents formats de données habituellement utilisés sur des données omiques. Les *Genome Browsers* sont soit des applications à part entière, comme *Integrated Genome Browser (IGB)* ou *Integrated Genome Viewer (IGV)*, soit instanciées sur les bases de données (comme *Ensembl* ou *UCSC*). Ces applications permettent en outre de visualiser des données publiées sur les dépôts d'archives (voir partie 1.3.7).

**1.3.1.4.5 Normalisation des données de séquençage** La normalisation des données de séquençage est une étape obligatoire lorsque l'on cherche à comparer plusieurs échantillons entre eux. En effet, si le nombre de *reads* totaux (ou taille de la librairie) est différent entre les échantillons, les variations de *reads* observées à un endroit précis du génome (comme un gène) pourront être dues à la seule différence de profondeur des échantillons.

Il faut donc corriger ce biais, et différentes méthodes existent. Une méthode populaire, *Count Per Millions (CPM)*, peut s'appliquer de manière générale sur tout échantillon de séquençage, et *Reads Per Kilo base per Million reads (RPKM)*, peut être utilisé lorsque l'on compare des éléments (des gènes) de tailles

différentes. Les méthodes d'analyses différentielles utilisent leurs propres méthodes de normalisation (voir partie 1.3.5.2).

La méthode *CPM* peut être décrite par cette équation (1.5) :

$$CPM = Y_{ik} \times \frac{1 \times 10^6}{N_k} \quad (1.5)$$

Où  $Y_{ik}$  est le nombre de *reads* présent à la position  $i$  dans l'échantillon  $k$ , et  $N_k$  le nombre total de *reads*.

Une autre méthode simple consiste à échelonner un échantillon par rapport à un autre (1.6) :

$$norm(Y_{ik}) = Y_{ik} \times \frac{N'_k}{N_k} \quad (1.6)$$

D'autres méthodes plus originales introduisent une quantité de matériel génétique d'une autre espèce afin de calibrer l'échantillon. On utilise alors ce nombre de *reads*, aligné sur le génome de l'autre espèce (*spike-in*), pour comme facteur de normalisation.

### 1.3.2 Cartographie des sites de liaison protéine-ADN (ChIP-Seq)

La méthode par ChIP, pour Chromatin Immunoprecipitation est une méthode très populaire permettant de cartographier des sites de liaison protéine-ADN. Elle s'applique très bien au séquençage, même si des méthodes plus récentes existent, comme cut&RUN (SKENE et HENIKOFF 2017) et cut&TAGs (KAYA-OKUR et al. 2019).

Le principe du ChIP appliqué au séquençage est d'extraire et de séquencer l'ADN en contact avec un élément (le plus souvent une protéine) d'intérêt, on parle alors de site de liaison. Les éléments couramment étudiés sont des protéines, comme les facteurs de transcription, les histones, témoins de l'état de la chromatine et d'autres éléments comme les structures secondaires d'ADN, les G-quadruplexes (HÄNSEL-HERTSCH, SPIEGEL et al. 2018) et les R-loops, hybrides ADN-ARN (DRIP-seq, GINNO et al. (2012)) par exemple (voir Figure 1.23).

En pratique, on effectue un *cross-link* réversible de la protéine avec l'ADN de façon à les lier, puis l'ADN est fragmenté. L'élément d'intérêt est ensuite récupéré par immunoprécipitation via l'utilisation d'un anticorps spécifique. Le *cross-link* entre l'ADN et l'élément est ensuite annulé, puis les extrémités des fragments sont séquencés. Les *reads* au format *fastQ* (voir partie (1.3.1.4.1)) sont ensuite alignés de manière classique sur le génome de référence (voir partie 1.3.1.4.2.1).

Le choix du nombre de *reads* va dépendre du nombre de régions génomiques ciblées : pour un facteur de transcription très localisé, on peut utiliser un nombre de *reads* moins important qu'une modification d'histone enrichie sur l'ensemble du génome. En outre, plus on augmentera le nombre de *reads*, plus on sera capable de détecter des événements rares.

Il n'est généralement pas nécessaire de séquencer les deux extrémités du fragment ou *paired-end*, sauf pour conserver l'information du fragment d'origine, ce qui permet de travailler avec deux fois plus de *reads*. Pour détecter des sites de liaisons significatifs, on parle de recherche de pics ou *peak calling*.

La recherche de pics consiste à comparer l'enrichissement de la protéine à un site de liaison par rapport à une expérience contrôle qu'on appelle *input*, où l'expérience de ChIP-seq est reproduite sans anticorps, ce qui permet de prendre en compte les biais expérimentaux. En effet, le séquençage des fragments par ChIP-seq produit des biais expérimentaux. En utilisant un *input*, on retrouve ces mêmes biais expérimentaux et on peut soustraire ce signal aux données ChIP-seq afin d'extraire les régions spécifiquement enrichies par l'utilisation d'un anticorps.

### 1.3.2.1 Recherche des pics dans les données ChIP-Seq

Le *peak calling* est un algorithme de recherche qui permet de détecter des régions enrichies, ou pics, sur le génome entier. C'est également une méthode statistique qui permet de déterminer si une région donnée, qu'on pense être un site de fixation de notre élément d'intérêt, est significativement enrichie par rapport au *background*. Un outil très utilisé pour rechercher des pics est MACS2 (Y. ZHANG et al. 2008). L'algorithme de MACS2 se déplace le long du génome via une fenêtre de taille fixe afin de trouver des pics candidats. La significativité des pics candidats va être testée en utilisant une distribution de Poisson à un paramètre  $\lambda$ , représentant le nombre attendu de *reads* dans cette fenêtre. Au lieu d'utiliser un  $\lambda$  uniforme, MACS2 utilise un  $\lambda$  dynamique et prend la valeur maximum entre le *background* (obtenu via l'*input*) et plusieurs fenêtres autour du pic. Si on dispose de la valeur attendu *lambda*, on peut alors utiliser la distribution de Poisson pour calculer la probabilité d'avoir exactement  $k$  *reads* (observés) (1.7) :

$$P(X = k) = \frac{\lambda^k}{k!} * e^{-\lambda} \quad (1.7)$$

MACS2 ajuste également différents paramètres comme la position des *reads* de façon à obtenir la position la plus précise possible pour le site de liaison. Il normalise également les librairies par rapport au plus petit échantillon. Enfin, il effectue une correction des tests multiples par *Benjamini-Hochberg (BH)* qui permet de réduire le taux de faux positifs.

**1.3.2.1.1 Correction des tests multiples** Le principe de la *p-value* fait que parfois, on rejette l'hypothèse nulle  $H_0$  par pure chance (faux-positifs). Plus on calcule de *p-values*, plus le nombre de faux-positifs sera conséquent. Pour éviter ce problème, on peut diminuer le risque  $\alpha$  de première-espèce de façon à diminuer le % de faux-positifs (en passant à 1% par exemple), ou appliquer des méthodes de corrections de tests multiples, comme la méthode de *Bonferroni* (1.8) :

$$\alpha_{bonferroni} = \frac{\alpha}{n} \quad (1.8)$$

Où  $n$  représente le nombre total d'hypothèses nulles testées. *Bonferroni* est une méthode simple, mais très stringente.

On peut alors appliquer la méthode (*BH*) : pour  $m$  test, déterminer les  $k$  *p-values* où on rejette les hypothèses  $H_0$ , où  $k = \max\{i : p_{(i)} \leq \alpha \cdot \frac{i}{m}\}$  et  $\alpha \cdot \frac{i}{m}$  correspond à la valeur critique *BH*.

**1.3.2.1.2 Le format bed** Le format de sortie de MACS2, et plus généralement d'outils détectant des régions enrichies sur le génome est le format **bed**. Chaque correspond à une ligne indiquant sa

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

position génomique, avec son chromosome, sa position (début et fin) et le brin d'origine, séparé par des tabulations. On peut ensuite charger ce fichier dans un *Genome Browser* (voir partie 1.3.1.4.4.3), pour visualiser les régions génomiques enrichies.

#### 1.3.3 Analyse de motifs de site de liaison de protéines à l'ADN

##### 1.3.3.1 *Position Probability Matrix*

Les protéines qui se fixent à l'ADN reconnaissent généralement un panel plus ou moins large de séquences d'ADN, donnant lieu à une séquence consensus ou “motif” qui décrit la préférence de liaison d'une protéine. Un motif peut être décrit sous la forme d'une *Position Probability Matrix* (**PPM** ou **PFM**), qui calcule l'occurrence des nucléotides (sous la forme de probabilité) à une position donnée du site de liaison (STORMO 2000) (voir Figure 1.25 A,C). On peut représenter ces fréquences sous la forme d'une séquence logo (voir Figure 1.25 B).

À partir de cette matrice, on peut calculer la probabilité (sous réserve d'indépendance statistique des positions) d'une séquence  $S = CCCATTGTTCTC$  étant donné la **PFM M** de la Figure 1.25 D (1.9) :

$$P(S = \mathbf{M}) = 0.48 \times 0.7 \times 0.48 \times 0.82 \times 0.93 \times 0.82 \times 0.93 \times 0.93 \times 0.93 \times 0.59 \times 0.7 \times 0.59 = 0.02 \quad (1.9)$$

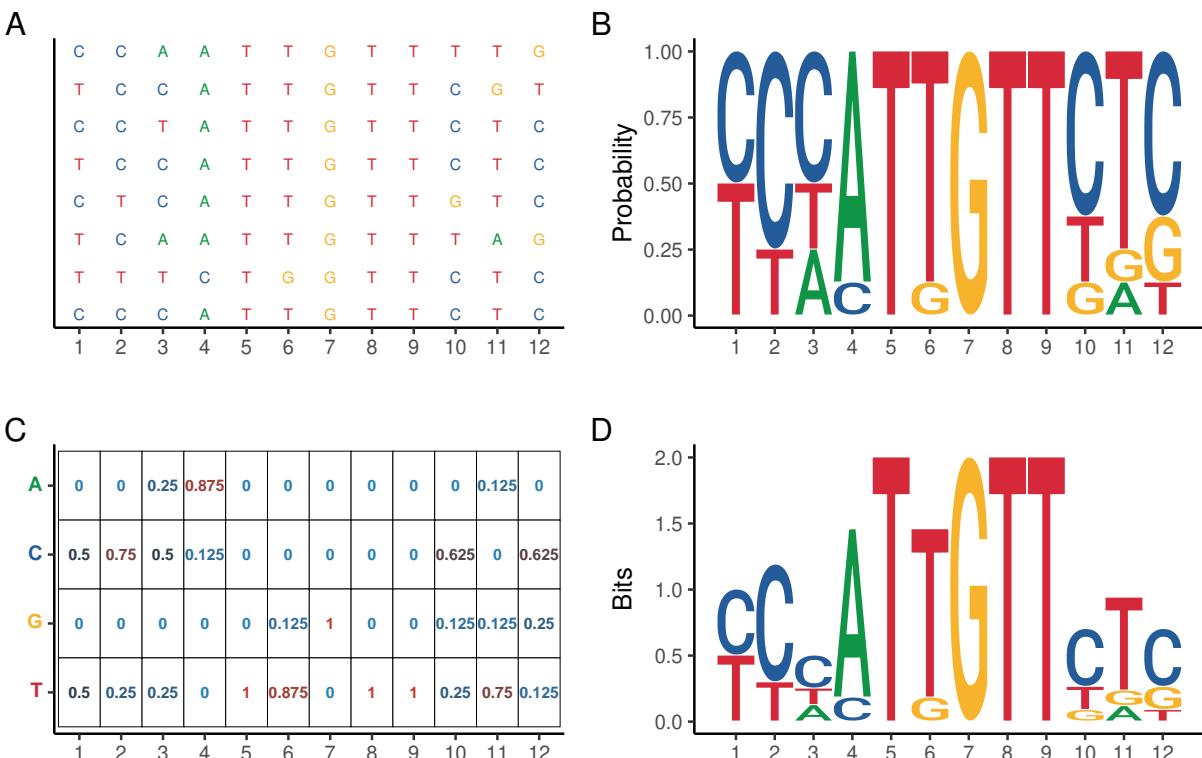


FIGURE 1.25 – Génération d'une *Position Weight Matrix* à partir d'un set de séquence (ROX1) (D'HAESELEER 2006).

### 1.3.3.2 Position Weight Matrix

Les *Position Weight Matrix* (**PWM**) sont en pratique plus souvent utilisées que les **PPM**. Les **PWM** sont calculées en utilisant le ratio de la probabilité par position des nucléotides divisé par sa probabilité attendue (souvent 0.25), puis converties à l'échelle logarithmique<sup>5</sup> (1.10) :

$$S(N_i) = \log_2 \left( \frac{P(N_i)}{B_N} \right) \quad (1.10)$$

Où  $S(N_i)$  est le score du nucléotide  $N$  à la position  $i$  et  $B_N$  la probabilité attendue du nucléotide sur le génome. On obtient ainsi un score logarithmique qu'on peut évaluer comme un rapport des cotes ou **odds ratio** indiquant à quel point la séquence diffère d'une séquence aléatoire. Avec ce score, on détermine la capacité d'une séquence à être un site de liaison potentiel (pour une **PWM** donnée). Plus le score est élevé, plus la séquence correspond au site de liaison représenté par **PWM**.

### 1.3.3.3 Information Content

On peut déterminer la mesure de conservation d'une position dans un motif en calculant l'*Information Content* ou **IC**. Cet **IC** provient de l'*entropie de Shannon* (SHANNON 1948) et s'exprime en *bits*. Avec l'**IC**, on peut déterminer les positions importantes du motif dans la matrice. Dans le cas de *ROX1*, seules les positions 5, 7, 8 et 9 contiennent une information maximale (voir Figure 1.25 D). L'*IC* maximal est basé sur la taille de l'alphabet, ce qui fixe cette valeur à  $IC_{total} = 2$  pour l'ADN. Un nucléotide pour une position donnée ne peut donc dépasser cette valeur. On peut calculer cette entropie ( $U$ ) par position avec la formule suivante (1.11) :

$$U = - \sum_{i=1}^N P(N) \times \log(P(N)) \quad (1.11)$$

Où  $U$  correspond à l'incertitude ou entropie de Shannon, et  $P(N)$  la distribution des probabilités des nucléotides à la position  $i$  de la *PWM*. Si l'entropie est minimale à une position donnée, alors l'importance, l'*IC* sera maximal : (1.12) :

$$IC_{final} = IC_{total} - U \quad (1.12)$$

Avec une valeur minimale ou nulle de  $U$  correspondant à l'information maximale  $IC_{final} = IC_{total} = 2$  pour une position donnée.

### 1.3.3.4 Analyse des motifs dans les séquences biologiques

Un outil bio-informatique développé pour l'analyse des motifs provenant de séquences biologiques est la suite MEME<sup>6</sup> (T. L. BAILEY, BODEN et al. 2009). Cette suite d'outils permet par exemple de détecter

5. Afin d'éviter les petits nombres d'occurrences ou les valeurs nulles, on peut également ajouter un paramètre **pseudocounts**.

6. <https://meme-suite.org/meme/>

des motifs *de novo* (*MEME* (T. L. BAILEY et ELKAN 1994)), de rechercher des motifs connus dans un ensemble de séquences (*FIMO* (GRANT, T. L. BAILEY et W. S. NOBLE 2011)) ou de comparer des motifs (*TOMTOM* (GUPTA et al. 2007)).

Une autre suite d'outil très utile pour l'analyse de *PWM* est *RSAT* (HELDEN 2003), qui permet notamment de convertir les différents formats de *PWM* (format *MEME*, *JASPAR*, *TRANSFAC* par exemple), mais aussi de faire du *clustering* de matrices.

**1.3.3.4.1 Détection de nouveaux motifs** Lorsque l'on souhaite faire de la recherche de nouveaux motifs (*de novo*) ou de motifs non annotés, on peut utiliser différents algorithmes qui permettent de détecter un ou des motifs enrichis dans des séquences biologiques, sans base de données. On peut par exemple utiliser *MEME* (T. L. BAILEY et ELKAN 1994), pour *Multiple EM for Motif Elicitation*. L'objectif de cet outil est d'identifier des sous-séquences enrichies dans un ensemble de séquence biologiques. Sachant qu'une sous-séquence peut varier, celles-ci sont représentées sous la forme de *PWM* (voir partie 1.3.3.2).

*MEME* recherche des motifs en utilisant un modèle de mélange à deux composants : la première composante représente le *background*, et la seconde représente le motif (QUANG et XIE 2014). L'algorithme divise les séquences en sous-séquences chevauchantes de tailles fixes et assume que chaque sous-séquence est soit une instance du motif, soit une instance du *background*. Beaucoup de variantes de cet algorithme existent, améliorant sa vitesse d'exécution comme *EXTREME* (QUANG et XIE 2014) ou bien *CUDA-MEME* (LIU et al. 2010), une implémentation utilisant les cartes graphiques.

**1.3.3.4.2 Recherche de motifs connus** Si on dispose d'une base de données de motifs (voir partie 1.3.7.4), on peut rechercher l'occurrence de chacun de ces motifs dans un ensemble de séquences biologiques. *FIMO* est l'outil de la suite *MEME* qui implémente un algorithme de recherche de motifs. Il calcule un score indiquant la capacité de la séquence à être un site de liaison potentiel (voir partie 1.3.3.2). *FIMO* détermine ensuite la probabilité d'erreur ou *P-value* de ce score en calculant tous les scores potentiels du motif et compare le score de la sous-séquence obtenu, puis ajuste par correction de tests multiples.

**1.3.3.4.3 Enrichissement de motifs connus** Dans R, *motifmatchR* (SCHEP 2019) permet également de rechercher la position des motifs sur des séquences biologiques. Avec ce outil, on peut rechercher les occurrences du motif dans notre ensemble de séquences, et on peut, en comparant avec un ensemble contrôle (généré par permutation ou provenant de régions génomiques contrôles), calculer l'enrichissement de ce motif (voir Table 1.2).

Avec cette table, on peut calculer l'enrichissement du motif en utilisant la formule de *rapport des côtes* ou *odd ratio* (1.13) :

$$Oddratio = \frac{\frac{a}{b}}{\frac{c}{d}} = \frac{\frac{436}{141038}}{\frac{319}{139081}} = 1.35 \quad (1.13)$$

On peut également calculer la probabilité exacte de cet enrichissement, par rapport à un contrôle via un test de *Fisher* unilatéral (1.33). *AME* (MCLEAY et Timothy L BAILEY 2010), de la suite *MEME*,

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

TABLE 1.2 – Exemple d'occurrences d'un motif dans un ensemble de séquences biologiques et contrôles.

	Occurrences dans les autres séquences	Occurrences dans la séquence
Ensemble d'intérêt	141038 (A)	436 (a)
Ensemble contrôle	139081 (B)	319 (b)

permet également de calculer un enrichissement de motifs connus à partir d'un ensemble de séquences positifs et contrôles.

#### 1.3.3.5 Comparaison de motifs

Si on dispose d'une liste de *PWM* (par exemple générés de manière non-supervisée par *MEME*) inconnus, on peut les identifier en les comparant à une base de données (voir partie 1.3.7.4). On peut pour cela utiliser une méthode de similarité entre deux matrices, et calculer la meilleure similarité combinée par somme, moyenne ou moyenne géométrique (GUPTA et al. 2007). Des méthodes comme le calcul de la distance *Euclidienne* (1.14), la corrélation de *Pearson* (1.15), qui mesure la dépendance entre deux vecteurs, ou la divergence de *Kullback-Leibler* (1.20) qui mesure la dissimilarité entre deux distributions sont utilisables sur des matrices :

$$d_{Eucl} = \sqrt{\sum_{i=1}^r \sum_{j=1}^w (n_A(i,j) - n_B(i,j))^2} \quad (1.14)$$

Pour une paire de motifs  $(A, B)$ , avec  $r$  lignes et  $w$  colonnes;  $n_A$  et  $n_B$  les *PWM* (ou matrices de comptages) et  $n_A(i,j)$  et  $n_B(i,j)$  les comptages observés à la  $i^{me}$  ligne et  $j^{me}$  colonne.

$$r_{A,B} = \frac{cov_{A,B}}{\sqrt{v_A v_B}} \quad (1.15)$$

$$cov_{A,B} = \frac{1}{rw} \sum_{i=1}^r \sum_{j=1}^w ((n_A(i,j) - \bar{n}_A)(n_B(i,j) - \bar{n}_B)) \quad (1.16)$$

$$v_A = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^w (n_A(i,j) - \bar{n}_A)^2 \quad (1.17)$$

$$v_B = \frac{1}{n} \sum_{i=1}^r \sum_{j=1}^w (n_B(i,j) - \bar{n}_B)^2 \quad (1.18)$$

Où  $cov_{A,B}$  est la covariance de  $(A, B)$ ,  $v_A$  et  $v_B$  la variance et  $\bar{n}_A$  et  $\bar{n}_B$  sont l'espérance ou moyenne des matrices  $A$  et  $B$ . On peut aussi calculer une mesure de dissimilarité à partir de la corrélation de *Pearson* (compris entre  $[-1; 1]$ ) :

$$dCor = 2 - r_{A,B} \quad (1.19)$$

$$D_{KL}(P|Q) = \sum_i P(i) \log \left( \frac{P(i)}{Q(i)} \right) \quad (1.20)$$

Où  $D_{KL}$  correspond à la divergence de *Kullback-Leibler*. Dans ce cas présent,  $Q(i)$  représente la distribution des nucléotides d'une des deux matrices. Cette équation peut également être utilisée à la place de l'entropie de Shannon (1.11) dans la partie 1.3.3.3, où  $Q(i)$  représente la distribution du *background* des séquences.

Ces méthodes de comparaison de *PWM* sont implémentées dans *matrix-comparison* de l'outil *RSAT* (HELDEN 2003) afin d'identifier un motif.

*TOMTOM* (GUPTA et al. 2007) est un autre outil permettant l'alignement *ungapped* de motifs, et largement utilisé pour identifier un ou plusieurs motifs à une base de données. Comme *compare-matrix*, *TOMTOM* est une méthode utilisant une fonction combinée (ici additive) de comparaison de colonne  $S(Q, T)$  où  $Q$  et  $T$  sont les deux matrices qu'on cherche à comparer. Ici, le score  $S(Q, T)$  est petit si  $Q$  et  $T$  se ressemblent de façon biologiquement pertinente. *TOMTOM* va ensuite calculer de façon dynamique la distribution du score  $S(Q, T)$  sous hypothèse nulle que  $Q$  et  $T$  ne sont pas biologiquement ressemblants. Cette distribution sous hypothèse nulle du score est calculée en utilisant l'ensemble des motifs cibles.

Pour chaque position de départ du motif  $Q$  aligné sur  $T$ , on peut calculer une densité de cette distribution et déterminer une *P-value* du score. *TOMTOM* identifie ensuite un score pour lequel la *P-value*, à une position de départ d'alignement de  $Q$  sur  $T$  et une orientation du brin donnée est minimale. Les *P-values* calculées étant nombreuses, celles-ci sont ensuite corrigées par la méthode *Bonferroni* où chaque *P-value* est multipliée par le nombre de cibles dans la base de données. Le résultat produit une *E-value* correspondant au nombre attendu de fois où le motif  $Q$  correspond à un motif dans la base de données cible par rapport aux correspondances observées dans une base de données aléatoire de la même taille.

### 1.3.3.6 Clustering de motifs

Le regroupement ou *clustering* de motifs permet d'associer une liste de motifs et de créer des groupes afin de réduire le nombre et surtout la redondance de ces motifs. Cette méthode est implémentée dans l'outil *matrix-clustering* (CASTRO-MONDAGON et al. 2017) de la suite *RSAT* (HELDEN 2003). Une méthode de *clustering* générale est de construire un arbre de regroupement hiérarchique ou *hierarchical clustering tree*. Pour construire cet arbre, il faut un algorithme de construction, une métrique de dissimilarité entre les individus à regrouper et une fonction de lien ou *linkage criteria*. Les individus ici sont des *PFM*, et la métrique de dissimilarité peut-être obtenue en comparant les motifs entre-eux (voir partie 1.3.3.5). La méthode de dissimilarité utilisée dans *matrix-clustering* est obtenue par l'outil *matrix-compare* présenté dans la partie 1.3.3.5, ce qui permet de choisir différentes métriques parmi celles proposées par l'outil. Ensuite, l'arbre est construit par méthode descendante ou *agglomerative*.

Chaque motif démarre dans son propre *cluster*, puis chaque *cluster* est combiné avec un autre en utilisant une fonction de lien, jusqu'à former un seul et unique *cluster*. Dans l'arbre, le cluster unique contenant tous les individus représente la racine, et chaque *PFM* les feuilles. Dans *matrix-clustering*, on peut choisir parmi trois méthodes de lien : le saut minimum (*single-linkage clustering*) (1.21), le saut maximum (*complete-linkage clustering*) (1.22) ou le lien moyen (*Unweighted average linkage clustering*) (1.23).

$$\min \{d(a, b) : a \in A, b \in B\} \quad (1.21)$$

$$\max \{d(a, b) : a \in A, b \in B\} \quad (1.22)$$

$$\frac{1}{|A| \cdot |B|} \sum_{a \in A} \sum_{b \in B} d(a, b) \quad (1.23)$$

Pour une paire de cluster  $(A, B)$ ,  $d(a, b)$  étant la mesure de dissimilarité calculée par *matrix-compare*.

Les motifs au sein du même *cluster* sont ensuite alignés ensemble de manière progressive par paires afin de produire un motif *consensus*, représentatif du *cluster* entier. Une fois les *PFM* au sein du cluster alignées, *matrix-clustering* calcule pour chaque noeud du cluster un *PCM* en sommant les fréquences des motifs alignés des branches descendantes, ce qui produit pour chaque cluster de l'arbre un motif consensus et son logo.

### 1.3.4 Cartographie de l'accessibilité de la chromatine (ATAC-Seq)

L'accessibilité de la chromatine définit à quel point des molécules sont capables d'accéder à l'ADN. Elle est déterminée par l'occupation des nucléosomes, leur organisation topologique et leur niveau de compaction (hétéro/eu-chromatine), ainsi que de la présence d'autres facteurs de liaison à la chromatine qui obstruent l'accès à l'ADN. On peut mesurer ce niveau d'accessibilité de l'ADN via l'utilisation de méthodes enzymatiques qui digèrent l'ADN. Plus l'ADN sera accessible, plus les fragments digérés par l'enzyme seront nombreux dans la population cellulaire composant la librairie, et plus le nombre de *reads* issus de ces fragments sera important lors du séquençage. On peut donc cartographier, à l'échelle du génome, l'état local moyen de la chromatine par sa capacité à être digéré ou non (voir Figure 1.26).

Il est également possible de déterminer la composition et les modifications post-traductionnelles des nucléosomes qui reflètent les différents états possibles de la chromatine. On peut alors utiliser des méthodes comme le ChIP-seq pour identifier les états de la chromatine et les sites de liaisons de différents facteurs de transcription.

Il existe différentes méthodes omiques permettant de cartographier l'accessibilité de la chromatine, qui sont basées sur le principe de digestion enzymatique. Les plus connues sont le DNase-seq (A. P. BOYLE et al. 2008), basé sur l'activité de la DNAse I, l'ATAC-seq (BUENROSTRO et al. 2013), pour *Assay for Transposase-Accessible Chromatin* et le MNase-seq (MIECZKOWSKI et al. 2016) utilisant une endonucléase, la Nucléase micrococciale.

Chaque méthode va créer des fragments d'ADN qui vont dépendre de l'activité de l'enzyme (voir Figure 1.27 A). En pratique, le DNase-seq et l'ATAC-seq ont une activité très similaire, et vont cartographier les mêmes endroits. Leur profil de couverture génomique présente donc une forte corrélation (BUENROSTRO et al. 2013). Le MNase-seq est différent, car la Nucléase micrococciale agit à la fois comme une exonucléase et va donc extraire les mêmes fragments que l'ATAC-seq, mais agit aussi comme une endo-nucléase, et va donc pouvoir digérer l'ADN entre les nucléosomes, permettant théoriquement de faire du *nucleosome positioning* de façon précise.

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

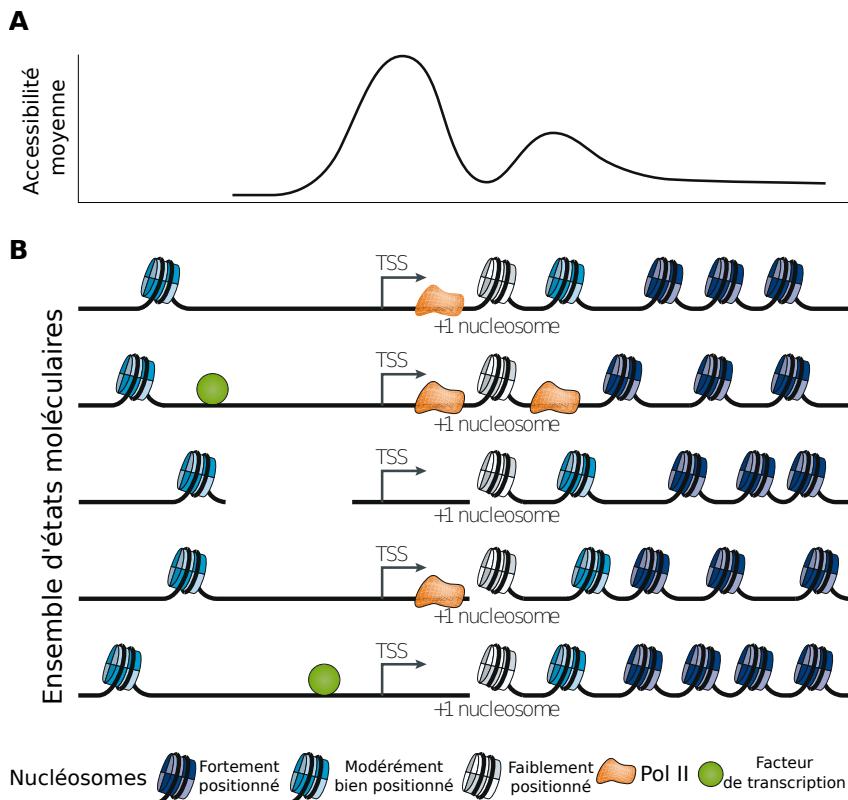


FIGURE 1.26 – Le profil d'accessibilité reflète l'accessibilité moyenne de la population cellulaire. **A** Profil d'accessibilité moyen d'un gène déterminé par séquençage. **B** Représentation possible des états d'accessibilité du gène dans la population cellulaire reflétée par le profil moyen. Adapté de KLEMM, SHIPONY et GREENLEAF 2019

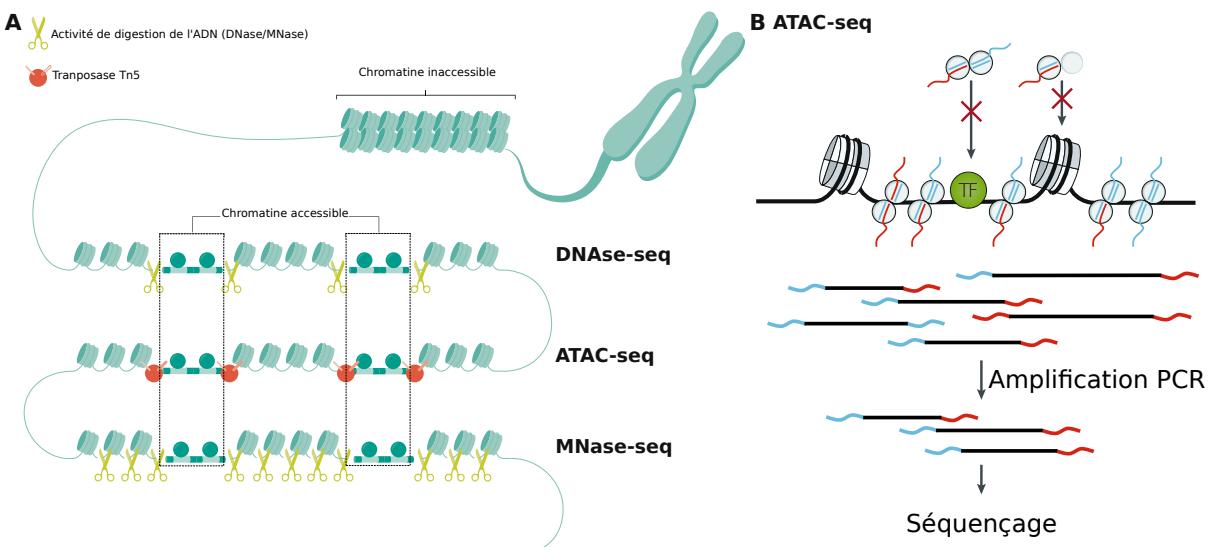


FIGURE 1.27 – Différents protocoles pour identifier l'accessibilité de la chromatine. **A** Différentes méthodes seq pour extraire & amplifier les fragments libres d'ADN, le DNAse-seq se base sur l'activité de la DNAse I, l'ATAC-seq sur l'activité d'une transposase et le MNase-seq sur une endo-nucléase. **B** Méthode de l'ATAC-seq qui utilise la transposase hyperactive Tn5 pour cliver & tagger les fragments libres de la chromatine. Adapté de “Chromatin accessibility profiling methods” 2021 et KLEMM, SHIPONY et GREENLEAF 2019.

### 1.3.4.1 Profil d'accessibilité de la chromatine par ATAC-seq

La méthode de l'ATAC-seq permet de cartographier les régions accessibles de la chromatine. Elle utilise pour cela une transposase Tn5 mutée hyperactive qui est capable à la fois de cliver l'ADN et d'y ajouter les *adapters* Illumina nécessaires au séquençage. Les fragments obtenus sont donc théoriquement directement utilisables, après amplification, pour un séquençage en *paired-end*. Cette méthode ayant peu d'étapes, elle nécessite moins de cellules pour obtenir un profil similaire à la DNase-seq (BUENROSTRO et al. 2013). Une fois les fragments séquencés, on peut utiliser une méthode classique d'alignement (voir partie 1.3.1.4.2.1), comme pour le ChIP-seq (voir partie 1.3.2), à la différence que les deux extrémités du fragment obtenu sont séquencés (*paired-end*). En utilisant les données d'ATAC-seq, on peut détecter les régions actives, en analysant par exemple le niveau d'ATAC-seq au niveau des régions promotrices des gènes (voir Figure 1.28 B), détecter des enhancers actifs, ou bien des éléments insulateurs comme CTCF (KLEMM, SHIPONY et GREENLEAF 2019). On peut également, en utilisant la taille des fragments, inférer la position et l'occupation des nucléosomes d'une région génomique (voir Figure 1.28 B).

### 1.3.4.2 Détermination des fragments associés aux nucléosomes avec l'ATAC-seq

À l'inverse des fragments de ChIP-seq, qui sont obtenus par fragmentation aléatoire de l'ADN immunoprécipité, puis par sélection de taille, les fragments obtenus par ATAC-seq sont directement issus de l'activité de la transposase Tn5, qui ajoute les *adapters* Illumina aux extrémités du fragment (voir Figure 1.27 B). La taille du fragment va alors donner des informations utiles sur l'état "accessible" du fragment. Si l'ADN est "libre", alors la transposase peut y accéder n'importe où, les fragments sont donc petits. Si l'ADN n'est pas libre, du fait qu'une protéine est fixée dessus, ou bien que l'ADN est associé à un nucléosome, alors la taille du fragment dépendra de la taille de la macromolécule associée. On peut, via la distribution de la taille des fragments issus de l'ATAC-seq, identifier quels sont ceux qui sont associés à un, ou plusieurs nucléosomes, et ceux qui sont "libres". Sur la Figure 1.28 A, on peut voir la distribution des fragments d'une expérience d'ATAC-seq, et déterminer quels sont les fragments associés avec les nucléosomes (BUENROSTRO et al. 2013).

### 1.3.4.3 Inférence de l'occupation des nucléosomes avec l'ATAC-seq

Une fois qu'on peut catégoriser les fragments associés aux nucléosomes, on peut calculer un score d'occupation des nucléosomes en se basant sur les fragments associés au mono, et di-nucléosomes, en normalisant par le background (1.24) :

$$Nuc_{occupancy} = \frac{mono_{nuc} + di_{nuc}}{free_{nuc} \times 2} \quad (1.24)$$

où  $Nuc_{occupancy}$  correspond à l'occupation des nucléosomes sur une région précise, et  $di_{nuc}$  sont les fragments d'une taille correspondant aux di-nucléosomes, séparés en 2 fragments. Si on calcule cette occupation sur tout le génome, on peut déterminer la position des nucléosomes et créer un fichier de couverture locale sur l'ensemble du génome (voir Figure 1.28 B). On peut éventuellement ajouter les tri-nucléosomes si leur nombre n'est pas trop faible.

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

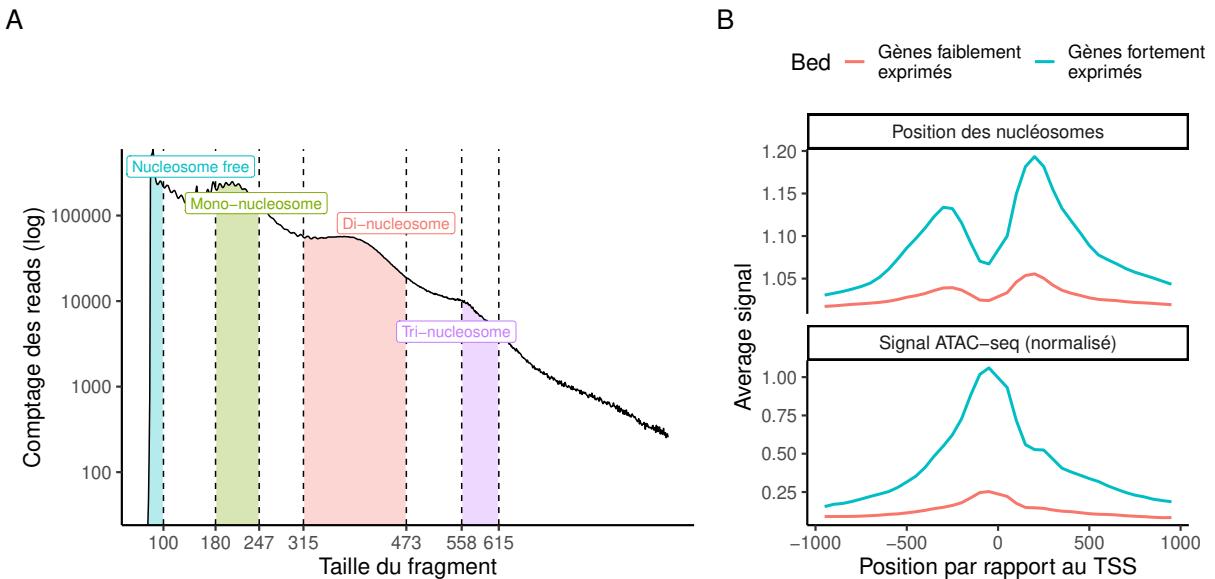


FIGURE 1.28 – Inférence de l'occupation et de la position des nucléosomes en utilisant les données d'ATAC-seq. **A** Détermination des fragments associés aux nucléosomes en fonction de leur taille. **B** Niveau d'accessibilité et d'occupation moyen sur les TSS des gènes les plus et les moins exprimés.

#### 1.3.5 Quantification de l'expression des gènes (RNA-seq)

La quantification de transcrits, ou de gènes, est une méthode quantitative utilisant des techniques de séquençage ou plus anciennement de *microarray* qui utilisent un signal représentant le niveau d'expression d'un gène dans une population cellulaire. Quand on quantifie l'ensemble des gènes d'une population cellulaire à un instant  $t$ , on parle de transcriptome ( $\Omega$ ). L'étude du transcriptome (ou transcriptomique) a pour objectif d'étudier, de cataloguer les types de transcrits (ARN messagers, ARN non-codants, petits ARNs, ...) pour déterminer la structure transcriptionnelle des gènes, leurs événements d'épissage, les variants d'épissage obtenus et bien souvent de quantifier les changements d'expression des gènes sous différentes conditions (voir partie 1.3.5.2 sur l'analyse différentielle).

La méthode de séquençage des transcrits la plus utilisée est le RNA-seq (Z. WANG, GERSTEIN et SNYDER 2009), qui séquence les ARNs matures d'une population cellulaire. Le RNA-seq est très similaire aux méthodes de séquençage classiques, hormis qu'elle séquence l'ADN complémentaire (ADNc) des transcrits. Concrètement, les ARNs sont extraits des noyaux des cellules puis, soit les ARNs messagers sont enrichis (*polyA enrichment*) ou bien les ARN ribosomaux sont déplétés (Ribo-Zero). Les ARNs sont ensuite fragmentés par enzymes (RNase III) ou chimiquement (solutions alcalines), puis des *primers* aléatoires sont couplés aux ARNs pour permettre l'initiation de la transcription inverse par la *transcriptase inverse* de 5' en 3'. L'ARN est ensuite dégradé puis le brin complémentaire de l'ADNc est synthétisé.

À cette étape, on peut utiliser des méthodes qui permettent d'identifier le brin d'origine du transcrit. L'idée est de synthétiser le brin complémentaire de l'ADNc (on obtient donc la même séquence que le transcrit d'origine) en ajoutant de l'Uracile au lieu de la Thymine (méthode dUTP (PARKHOMCHUK et al. 2009)). Des *primers* pour l'amplification et le séquençage sont alors ajoutés à cette étape, puis les brins contenant les Uraciles sont déplétés. Le brin restant, complémentaire de l'ARN d'origine et ses *primers* est défini comme la librairie *brin-spécifique* ou *strand-specific*, et les fragments sont sélectionnés

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

par taille, de manière à correspondre aux standards pour le séquençage Illumina.

Les étapes suivantes sont similaires à celles décrites en partie 1.3.1.1, et on obtient à la fin des données de séquençage au format fastQ (partie 1.3.1.4.1) qui peuvent ensuite être alignées sur le génome de référence en suivant le protocole décrit partie 1.3.1.4.2.2. Des méthodes complémentaires au RNA-seq existent permettant notamment de séquencer les brins d'ARNs naissants (avec les méthodes Bru-Seq (PAULSEN et al. 2013) et GRO-seq (GARDINI 2017)) ou les transcrits primaires (dRNA-seq (SHARMA et VOGEL 2014)) avant épissage.

Les méthodes d'analyses bio-informatiques du RNA-seq sont plus complexes que pour le ChIP-seq (voir partie 1.3.2). Pour une quantification de l'expression des gènes en vue d'une analyse différentielle, Illumina recommande de suivre les standards décrits par ENCODE<sup>7</sup>, typiquement, de 30 à 60 millions de *reads* par échantillon en *paired-end* ou en *single-end*. Si on souhaite analyser la structure des transcrits et identifier de nouveaux isoformes, le nombre de *reads* requis est bien plus important, de l'ordre de 100 à 200 millions de *reads* par échantillon. La méthode décrite ci-après est celle du séquençage de *short reads* d'ADNc pour analyse différentielle. La principale limitation de cette méthode étant que les *reads* ne séquentent que les extrémités du fragment. Il est donc difficile d'identifier la molécule d'ARN d'origine, ce qui ne permet pas de faire de l'analyse correcte d'isoforme, même si elle reste possible (SACOMOTO et al. 2012).

#### **1.3.5.1 Détermination de la couverture en *reads* des gènes**

Une fois les *reads* alignés sur le génome de référence, ceux-ci doivent être associés à leur gène respectif. Pour le RNA-seq, quantifier un nombre de *reads* pour un gène représente son niveau d'expression, en moyenne, dans la population cellulaire séquencée dans notre échantillon. Il existe différentes méthodes pour quantifier les *reads* sur une annotation, tels que *htseq count* (Simon ANDERS, PYL et Wolfgang HUBER 2014), *featureCounts* (LIAO, Gordon K. SMYTH et W. SHI 2013) et *STAR* en utilisant le paramètre *quantMode*. En utilisant une annotation complète contenant les informations sur les gènes, leurs transcrits et les exons du gène, on peut définir un niveau d'annotation sur lesquels les *reads* seront comptés. En général, on indique le nombre de *reads* pour tout le gène, mais on peut également quantifier à l'échelle du transcrit ou même de l'exon. Le format d'annotation utilisé généralement est le format *GTF* pour *General Transfert Format*, un format texte où les colonnes sont séparées par une tabulation. Ici, chaque ligne correspond à une *feature* où est indiqué sa position génomique, son type (gène, transcrit, exon par exemple), son identifiant dans la base de données et son type biologique (comme *protein coding*, *pseudogene* ou *lincRNA*). On peut télécharger une annotation des gènes du génome humain sur différentes bases de données, comme Ensembl<sup>8</sup> (voir partie 1.3.7.1), dont voici un extrait :

```
chr1 ensembl_havana gene 11869 14412 . + . gene_id "ENSG00000223972"; gene_version "4"; gene_name "DDX11L1";
gene_source "ensembl_havana"; gene_biotype "pseudogene";
chr1 havana transcript 11869 14409 . + . gene_id "ENSG00000223972"; gene_version "4"; transcript_id "ENST00000456328";
chr1 havana exon 11869 12227 . + . gene_id "ENSG00000223972"; gene_version "4"; transcript_id "ENST00000456328";
chr1 havana exon 12613 12721 . + . gene_id "ENSG00000223972"; gene_version "4"; transcript_id "ENST00000456328";
chr1 havana exon 13221 14409 . + . gene_id "ENSG00000223972"; gene_version "4"; transcript_id "ENST00000456328";
```

7. <https://www.encodeproject.org/data-standards/rna-seq/long-rnas/>

8. [http://ftp.ensembl.org/pub/grch37/current/gtf/homo\\_sapiens/](http://ftp.ensembl.org/pub/grch37/current/gtf/homo_sapiens/)

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

Ces lignes annotent la position du gène DDX11L1 et de son transcrit ENST00000456328 contenant trois exons. En utilisant les différents paramètres pour déterminer comment compter les *reads* sur un gène (Figure 1.29), *htseq count* permet par exemple de compter les *reads* contenus dans l'intervalle de la position des 3 exons pour quantifier le niveau d'expression du transcrit uniquement, ou bien dans toute l'intervalle du gène.

En cas d'ambiguïté entre deux gènes (les 3 derniers cas de la Figure 1.29), les *reads* peuvent être attribués au hasard, sauf si la librairie RNA-seq préparée est brin-spécifique (voir partie 1.3.5), ce qui permet d'identifier, grâce à l'origine du brin de la paire, l'origine du *read*. On peut pour cela utiliser les *FLAGS* dans le fichier *BAM* pour séparer les *reads* en fonction du brin d'origine (voir Table 1.1).

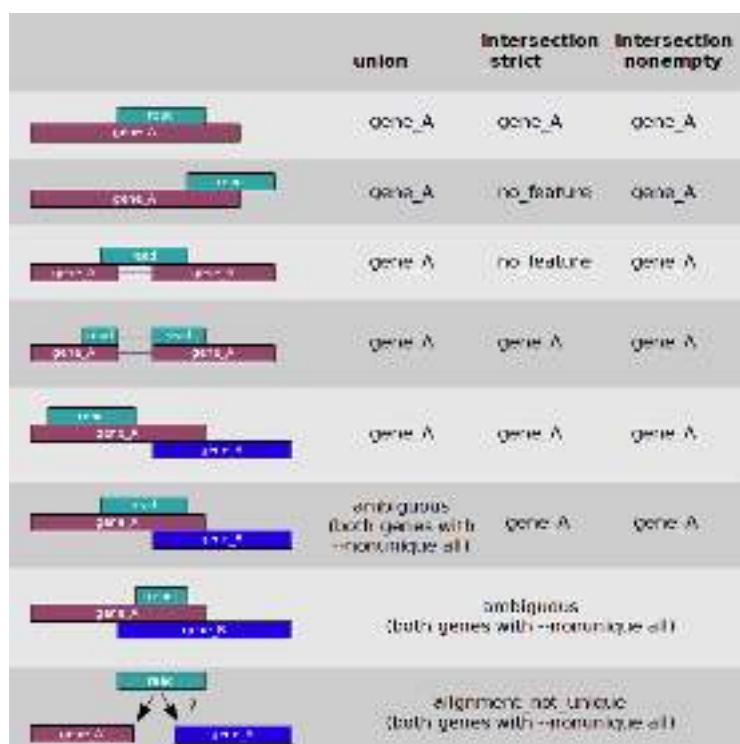


FIGURE 1.29 – HTseq count méthodes. (Simon ANDERS, PYL et Wolfgang HUBER 2014).

#### 1.3.5.2 Analyse d'expression différentielle

L'analyse d'expression différentielle a pour objectif d'identifier des gènes dont l'expression change significativement entre différentes conditions. La principale difficulté étant que les données d'expression de gènes sont des données de comptage, et pour cause de non-normalité, nécessitent d'utiliser des distributions de probabilités discrètes comme la loi de Poisson (voir partie 1.3.2.1) à un seul paramètre  $\lambda$  ou bien la loi Binomiale Négative qui introduit un paramètre de sur-dispersion  $\alpha$  supplémentaire. L'autre difficulté, directement due au coût du séquençage, étant le faible nombre de réplicats, souvent pas plus de deux ou trois, par condition. Les méthodes considérant chaque gène de façon indépendante souffrent donc de l'incertitude due à une grande variance intra-groupe. Dans tous les cas, ces analyses doivent être réalisées au minimum avec 2 réplicats biologiques, ou plus (d'après les standards d'ENCODE<sup>9</sup>).

9. <https://www.encodeproject.org/data-standards/rna-seq/long-rnas/>

**1.3.5.2.1 Méthodes d'analyses** Les outils les plus utilisés pour l'analyse d'expression différentielle pour des données RNA-seq sont *DESeq2* (LOVE, W. HUBER et S. ANDERS 2014) et *edgeR* (M. D. ROBINSON, MCCARTHY et G. K. SMYTH 2010). Ils sont tous les deux basés sur la distribution de probabilité discrète Binomiale Négative permettant d'estimer la sur-dispersion des données RNA-seq et le changement d'expression logarithmique ou *Log Fold Change* (*LFC*) des gènes. Ils testent tous les deux l'hypothèse nulle selon laquelle le *LFC* d'un gène entre le traitement et le contrôle est exactement zéro, indiquant que l'expression du gène n'est pas affectée par le traitement (LOVE, W. HUBER et S. ANDERS 2014). On peut modéliser la relation entre une variable réponse  $Y$  et une autre explicative  $X$  via un modèle linéaire (1.25) :

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (1.25)$$

Où la variable  $Y$  pour le point  $i$  peut être modélisée, avec  $\alpha$  qui correspond à l'origine de la droite  $\beta$  le coefficient directeur de la droite et  $\varepsilon$  les résidus, où la part non expliquée par le modèle. Par exemple, pour le jeu de données *iris* (FISHER 1936), on peut modéliser la relation entre la variable *Petal.Length* et *Petal.Width* en traçant une droite (Figure 1.30 A).

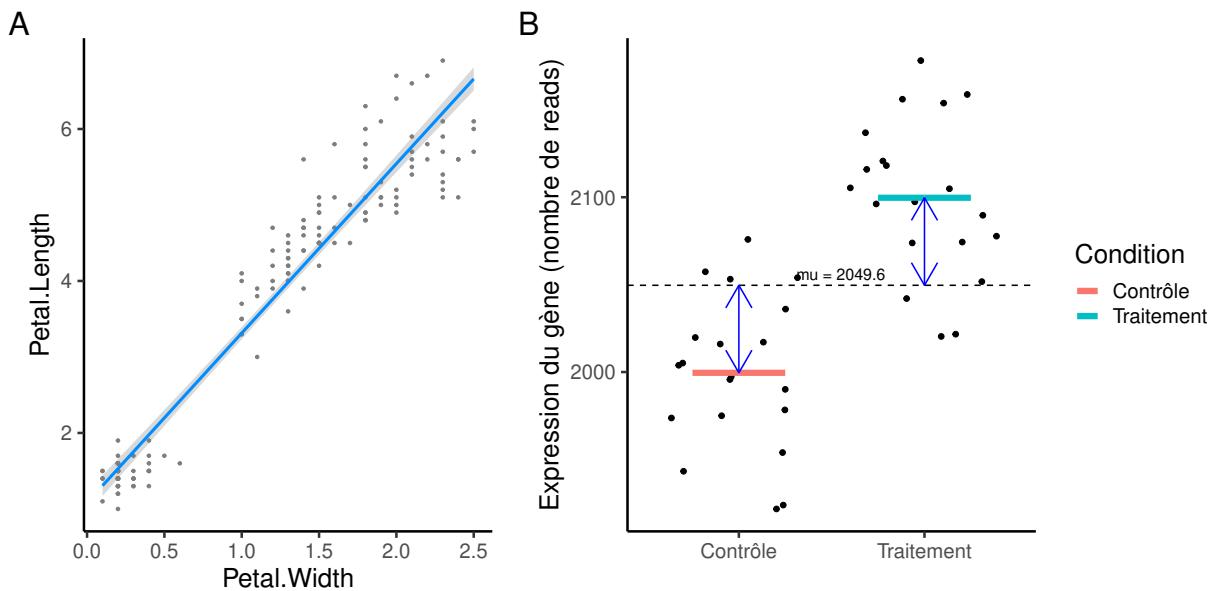


FIGURE 1.30 – Représentation de deux modèles linéaires. (A) Exemple d'une régression linéaire simple avec le jeu de données *iris* (FISHER 1936). (B) Modélisation d'une valeur qualitative (Traitement) sur l'expression d'un gène par ANOVA.

Dans le cas où  $Y$  représente l'expression d'un gène (variable *quantitative*) et  $X$  le traitement (variable *qualitative*), on peut modéliser cette relation par une analyse de la variance, ou *ANOVA* (1.26) :

$$Y_i = \mu + \beta_1 Controle_i + \beta_2 Traitement_i + \varepsilon_i \quad (1.26)$$

Où  $Y_i$  est l'expression d'un gène  $i$ ,  $\mu$  la moyenne générale de l'échantillon,  $Traitement_i$  et  $Controle_i$  l'écart à la moyenne  $\mu$  des groupes *Traitement* et *Contrôle*, respectivement, et  $\varepsilon_i$  le résidu de  $i$  par rapport à la moyenne de son groupe. Ici,  $\beta_1 = 0$ ;  $\beta_2 = 1$  si le point  $i$  fait partie du groupe *Traitement*, et

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

TABLE 1.3 – Table de comptages contenant 57870 annotations et 4 échantillons, avec 2 réplicats biologiques par condition.

	Sample 1	Sample 2	Sample 3	Sample 4
ENSG000000000003	2052	2436	2396	2743
ENSG000000000005	0	0	0	0
ENSG00000000419	3586	4508	3090	3761
ENSG00000000457	583	751	592	687
ENSG00000000460	2769	3493	2490	2422
ENSG00000000938	0	0	0	0
<b>Total</b>	<b>88110765</b>	<b>108113013</b>	<b>86204860</b>	<b>103334299</b>

inversement s'il fait partie du groupe *Contrôle*. On peut ainsi retrouver la valeur du point  $i$  en utilisant cette équation (Figure 1.30 B), et tester l'hypothèse suivante :

- $M0 : Y_i = \mu + \beta_1 Contrôle_i + \varepsilon_i$
- $M1 : Y_i = \mu + \beta_1 Contrôle_i + \beta_2 Traitement_i + \varepsilon_i$

Ce test nous permet de déterminer si le modèle  $M1$  modélise correctement les données ou si le modèle plus simple  $M0$  suffit. Ce qui revient à tester si, en moyenne, la condition *Traitement* est significativement différente de la condition *Contrôle*, et donc d'inférer sur l'effet du traitement sur l'expression du gène.

Les données RNA-seq sont des données de comptage, et ne peuvent donc pas être modélisées par un modèle linéaire classique. *DESeq2* et *edgeR* utilisent un modèle linéaire généralisé ou *GLM* suivant une distribution de probabilité Binomiale Négative. Les *GLM* diffèrent du modèle linéaire classique par l'incorporation d'une fonction de lien dans l'équation linéaire. Dans le cas d'une régression de Poisson ou Binomiale Négative, on peut utiliser le logarithme comme fonction de lien (1.27).

$$\log(E[Y_i|X_i]) = \alpha + \beta X_i \quad (1.27)$$

**1.3.5.2.2 Filtrage & Normalisation des données** Après alignement sur le génome ou transcriptome de référence, on peut compter, pour chacun des gènes son niveau d'expression représenté par le nombre de *reads* alignés sur sa position. On peut alors représenter les données sous la forme d'une table de comptage pour chacun de nos gènes, où chaque ligne représente un gène et chaque colonne un échantillon (voir Table 1.3). L'objectif est alors de regarder pour chacun des gènes si le ou les conditions ont un effet sur l'expression du gène, en moyenne.

On réalise donc un test statistique (voir partie 1.3.5.2.1) sur chacun des gènes, ce qui revient à calculer 57870 tests différents si on conserve toute la table. Lorsque l'on calcule une *P-value*, on calcule la probabilité de rejeter l'hypothèse nulle ( $H_0$ ), alors qu'elle est vraie, ce qui revient à quantifier le risque d'observer une différence d'expression entre les conditions par hasard. Plus on calcule de *P-values*, plus ce risque augmente. On doit donc corriger ces probabilités par des méthodes de correction de tests multiples (voir partie 1.3.2.1). Pour compenser cela, on filtre la table et on supprime les gènes peu ou pas exprimés (voir ligne 2 et 6 de la Table 1.3).

Les données d'expression étant issues de différentes expériences indépendantes, les données doivent être normalisées. Dans la Table 1.3 par exemple, on constate que les tailles de librairie (sur la ligne **Totale** de

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

TABLE 1.4 – Facteurs de normalisation calculés par *TMM* après filtre des gènes peu exprimés.

group	lib.size	norm.factors	samples
Cond1	88110765	0.9782	Sample 1
Cond2	108113013	0.9906	Sample 2
Cond1	86204860	1.0496	Sample 3
Cond2	103334299	0.9833	Sample 4

la Table 1.3) sont différentes. Afin de pouvoir comparer ces échantillons, il faut absolument normaliser les comptages de chaque gène par la taille des librairies.

Les méthodes diffèrent en fonction des outils, *edgeR* va normaliser les données en trouvant un ensemble de *scaling factors* pour les tailles de librairies qui minimise le *LFC* entre les échantillons pour la plupart des gènes. La méthode utilisée est *TMM* pour *Trimmed Mean of M-values* (M. D. ROBINSON et OSHLACK 2010). Elle fait l'hypothèse que la plupart des gènes ne sont pas différentiellement exprimés, et va donc chercher à réduire l'écart du nombre de *reads* entre les échantillons de ces gènes. Pour sa normalisation, *TMM* utilise l'expression absolue du gène  $A_g$  (1.29) et son *LFC*  $M_g$  (1.28) :

$$A_g = \frac{1}{2} \log_2 (Y_{gk}/N_k \cdot Y_{gk'}/N'_k) \quad (1.28)$$

$$M_g = \log_2 \left( \frac{Y_{gk}/N_k}{Y_{gk'}/N'_k} \right) \quad (1.29)$$

ou  $Y_{gk}$   $Y_{gk'}$  sont l'expression d'un gène  $g$  dans les échantillons  $k$  et  $k'$  et  $N_k$  et  $N'_k$  le nombre total de *reads* dans leur librairie respective. Les gènes au dessus et en dessous du seuil de 30% et 5% pour  $M_g$  et  $A_g$  sont supprimés, et une nouvelle moyenne rognée est calculée à partir de ces données, puis pondérée. Un facteur de normalisation peut-être ensuite calculé pour l'échantillon  $k$  en utilisant un échantillon de référence  $r$  (1.30) :

$$\log_2(TMM_k^r) = \frac{\sum_{g \in G} w_{gk}^r \cdot M_{gk}^r}{\sum_{g \in G} w_{gk}^r} \quad (1.30)$$

$$M_{gk}^r = \log_2 \left( \frac{Y_{gk}/N_k}{Y_{gr}/N_r} \right) \quad (1.31)$$

$$w_{gk}^r = \frac{N_k - Y_{gk}}{N_k Y_{gk}} + \frac{N_r - Y_{gr}}{N_r Y_{gr}} \quad (1.32)$$

Où  $\log_2(TMM_k^r)$  est le facteur de normalisation *TMM* et  $w_{gk}^r$  sont les poids utilisés pour la pondération. Les facteurs de normalisation calculés par cette méthode pour la Table 1.3 sont montrés dans la Table 1.4.

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

TABLE 1.5 – Ontologies des termes GO au 08/2021.

Termes	Nombre d'ontologies
Biological process	28532
Molecular function	11167
Cellular component	4179

#### 1.3.5.3 Analyse fonctionnelle des gènes

Les résultats d'une analyse d'expression différentielle produisent une liste de gènes qui sont différentiellement exprimés (*DEG*) dans les conditions de tests. C'est un résultat encore assez "brut" et difficile à interpréter. Pour interpréter les *DEG*, on peut les regrouper en groupes annotés afin de ressortir une fonction particulière affectée par notre traitement, on parle alors d'analyse fonctionnelle des *DEG*.

L'analyse fonctionnelle des gènes et de leurs produits (ARNs / protéines) peut être estimée par l'utilisation de termes biologiques, tels que la *Gene Ontology (GO)* (T. G. O. CONSORTIUM 2008), *Disease Ontology (DO)*, *Kyoto Encyclopedia of Genes and Genomes (KEGG)* et *Medical Subject Headings (MeSH)*.

Les ontologies *GO* ont pour objectif de structurer et contrôler le vocabulaire pour la communauté biologique afin de permettre l'annotation de gènes d'une large variété d'espèces. La structure des termes *GO* peut être représentée par un arbre, ou graphe orienté acyclique. Sur ce graphe, chaque nœud représente une ontologie, et la relation entre chaque terme est représentée par une arête. *GO* étant hiérarchique, chaque "enfant" d'un terme représente un terme plus spécifique que le parent, et un nœud peut avoir plusieurs parents.

Cette hiérarchie est divisée en trois types de terme, qui contiennent eux-mêmes d'autres termes, de plus en plus spécifiques (voir Table 1.5) :

- *Biological Process (BP)* est l'ensemble le plus annoté des termes *GO* et permet de déterminer dans quel processus biologique intervient les produits du gène, tel que *DNA repair*.
- *Molecular Function (MF)* va décrire l'ensemble des termes des activités des produits des gènes, comme une activité catalytique ou bien de transport d'une protéine ou d'un complexe. Cette annotation ne décrira jamais l'entité qui produit l'action ni le lieu où cette action est produite
- *Cellular Component (CC)* est la plus petite annotation qui permet d'annoter la localisation relative (en termes de structure cellulaire) où le produit du gène est actif, comme le ribosome ou bien la mitochondrie.

Par exemple, la protéine *p53* codée par le gène *TP53*, intervenant dans de nombreux processus biologiques est annotée sur UniProt<sup>10</sup> comme ayant une fonction moléculaire de *DNA binding (MF)*, intervient dans le processus de la réparation des *DSBs* (*double-strand break repair, BP*) et est présent dans le *cytoplasme (CC)*.

On peut donc utiliser différentes méthodes s'appliquant aux graphes et aux ensembles sur ces ontologies : afin d'associer des termes par similarité sémantique (J. Z. WANG et al. 2007), représenter ces similarités sémantique par projection sur un espace à faible dimension (En utilisant des algorithmes comme

10. <https://www.uniprot.org/uniprot/P04637>

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

Word2Vec (SMAILI, GAO et HOEHNDORF 2018)), ou pour identifier un processus biologique particulier associé à nos gènes différentiellement exprimés (E. I. BOYLE et al. 2004).

Les analyses de *sur-représentation* ou *Over Representation Analysis (ORA)* sont des méthodes très utilisées pour déterminer si une fonction biologique est enrichie à partir d'un ensemble de *DEG*. Cette méthode se base sur la loi hypergéométrique (E. I. BOYLE et al. 2004) et revient à calculer un test exact de Fisher unilatéral pour calculer la probabilité exacte qu'un ensemble de gènes représentant un processus biologique soit enrichi en *DEG*, par rapport à l'ensemble des gènes annotés (*background distribution*). On peut la calculer avec l'équation (1.33) :

$$p = 1 - \sum_{i=0}^{k-1} \frac{\binom{M}{i} \binom{N-M}{n-i}}{\binom{N}{n}} \quad (1.33)$$

Où  $N$  représente le nombre total de gènes annotés,  $M$  le nombre de gènes annotés dans le processus biologique d'intérêt,  $n$  le nombre de gènes d'intérêt et  $k$  le nombre de gènes dans  $n$  annotés dans  $M$ . De la même façon que pour les motifs, chaque *p-value* calculée doit être corrigée par correction de tests multiples (voir partie 1.3.2.1). Un outil implémentant l'*ORA* et permettant de visualiser les résultats de manière satisfaisante est *ClusterProfiler* (YU et al. 2012).

#### **1.3.6 Structure 3D de la chromatine (*Chromosome Conformation Capture, 3C*)**

La fibre de chromatine se replie sur elle-même et forme des compartiments (voir partie 1.1.4). Son étude par des méthodes comme le ChIP-seq n'est donc pas suffisante pour comprendre la complexité de ses interactions et de ses contraintes (physiques et biologiques) dans le noyau. Pour étudier cette conformation, il existe différentes méthodes dites "C" pour Conformation, comme le *3C* (*Chromosome Conformation Capture*) (N. NAUMOVA et al. 2012), 4C-seq (*Circular 3C*) (WERKEN, VREE et al. 2012), et le Hi-C (*High-throughput chromosome conformation capture*) (LIEBERMAN-AIDEN et al. 2009 ; S. S. RAO et al. 2014). Toutes ces méthodes sont basées sur le même principe de fixation des interactions entre *loci* spatialement proches. L'ADN est ensuite digéré par des enzymes de restriction, comme *DpnII*, qui reconnaît plus de 7 millions de sites de restrictions sur *hg19*. Cette étape de digestion est importante, car elle va définir la résolution des fragments séquencés. Plus le nombre de fragments est important, plus on augmente le nombre de paires d'interaction entre fragments possibles. La chromatine est ensuite religuée puis purifiée pour former une chimère d'ADN contenant les deux régions qui interagissent ensemble.

Le *3C*, permet d'étudier l'interaction de deux *loci* d'ADN particuliers via une amplification par PCR quantitative. Elle n'est donc pas adaptée à une étude globale et de ce fait n'entre pas dans les méthodes d'analyses omiques. Le 4C-seq et le Hi-C en revanche utilisent des méthodes de séquençage à haut débit pour quantifier l'ensemble du génome (*genome wide*) et sont donc des méthodes omiques (voir Figure 1.31). Le 4C-seq permet d'étudier les interactions d'un locus particulier (point de vue ou *viewpoint*) avec le reste du génome, et tout particulièrement de réaliser le profil d'interaction à très haute résolution autour du point de vue (voir Figure 1.32). Pour capturer ces interactions, on utilise une PCR inverse via des amorces spécifiques du *viewpoint* afin d'amplifier spécifiquement l'ADN en interaction avec celui-ci. Pour le Hi-C, on introduit un résidu biotinylé avant ligation afin de purifier de façon sélective les chimères

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

d'ADN, puis on séquence par *paired-end* les fragments produits, chaque élément de la paire étant issue de la séquence d'un des deux *loci* en interaction.

#### 1.3.6.1 Étude pangénomique de la conformation 3D du génome (Hi-C)

La méthode du Hi-C est l'adaptation pangénomique ou *genome wide* du 3C. Elle permet de capturer les interactions physiques de l'ADN le long du même chromosome, en *cis*, ainsi que des interactions longue distance, voir sur des chromosomes différents (en *trans*). Cette méthode a permis de mettre en évidence la présence de territoires chromosomiques, où les chromosomes riches en gènes ont tendance à se regrouper, ainsi que l'identification de deux compartiments (A et B) qui interagissent préférentiellement entre eux, et s'excluent mutuellement (LIEBERMAN-AIDEN et al. 2009). Plus récemment, une expérience de Hi-C plus résolutive a permis de mettre en évidence la formation de domaines 3D spécifiques, les Les Domaines Topologiquement associatifs, ou *Topologically Associating Domain (TAD)* (DIXON et al. 2012), ainsi que les boucles de chromatines à l'origine de leur formation (S. S. RAO et al. 2014) (voir Figure 1.32).

Pour capturer un maximum d'évènements d'interaction et être plus résolutif, il faut augmenter le nombre de *reads* séquencés. Par exemple, pour avoir une résolution à l'échelle du kilobase en Hi-C, soit pouvoir distinguer des interactions entre deux fenêtres, ou *bins* d'une taille de 1kb, il faut au moins séquencer 300 millions de *reads*, selon l'article de Rao et. al. (S. S. RAO et al. 2014).

**1.3.6.1.1 Alignement des données Hi-C** Les données Hi-C consistent en paires d'interaction entre deux *loci* qui forment une chimère d'ADN. Par séquençage, on peut identifier les régions qui interagissent en alignant la paire de *reads* de façon indépendante sur le génome de référence (puisque il s'agit d'une chimère). Les *reads* sont ensuite filtrés selon la façon dont ils ont été alignés sur le génome de référence (voir Figure 1.33). Si la séquence du *read* (et non de la paire) est composée d'un seul *locus* dans le génome, le *read* est considéré comme normal (~75% des *reads* sont normaux, en fonction de l'enzyme utilisée et de la longueur du *read*). Si le *read* est chimérique, c'est-à-dire qu'il s'aligne sur deux *loci* différent du génome, il va être classé "non-ambigu" (~15% des *reads* sont chimériques non-ambigus). Le reste des *reads* sont soit chimériques "ambigus", soit non alignables, et sont supprimés (voir Figure 1.33). Enfin, les duplicats *PCR* sont supprimés de l'analyse, et les cartes de contacts sont générées à partir des *reads* restants.

L'alignement des données HiC par le procédé décrit ci-dessus (S. S. RAO et al. 2014) est implémenté dans l'outil *Juicer* (DURAND, SHAMIM et al. 2016). L'algorithme qui réalise l'alignement doit rapporter les positions des *reads* qui s'alignent sur plusieurs portions du génome, et le deuxième *read* de la paire doit pouvoir contenir un des deux *loci* sur lesquels s'alignent le premier, mais pas les deux. Ce cas s'applique lorsqu'un des deux *reads* est séquencé jusqu'au point de ligation de la chimère. Cette méthode est efficace mais un certain pourcentage des *reads* seront supprimés de l'analyse.

Des méthodes alternatives permettent d'aligner autrement les *reads* sur le génome, en prenant en compte la structure particulière d'une chimère de Hi-C. Notamment, l'alignement itératif (IMAKAEV et al. 2012) permet d'aligner une première partie des *reads*, en tronquant à une certaine longueur, puis en agrégeant les alignements en augmentant la longueur. Cette méthode peut s'avérer très longue, en fonction de la taille utilisé pour tronquer les *reads*, mais s'avère essentielle lorsqu'aucune enzyme de restriction n'est

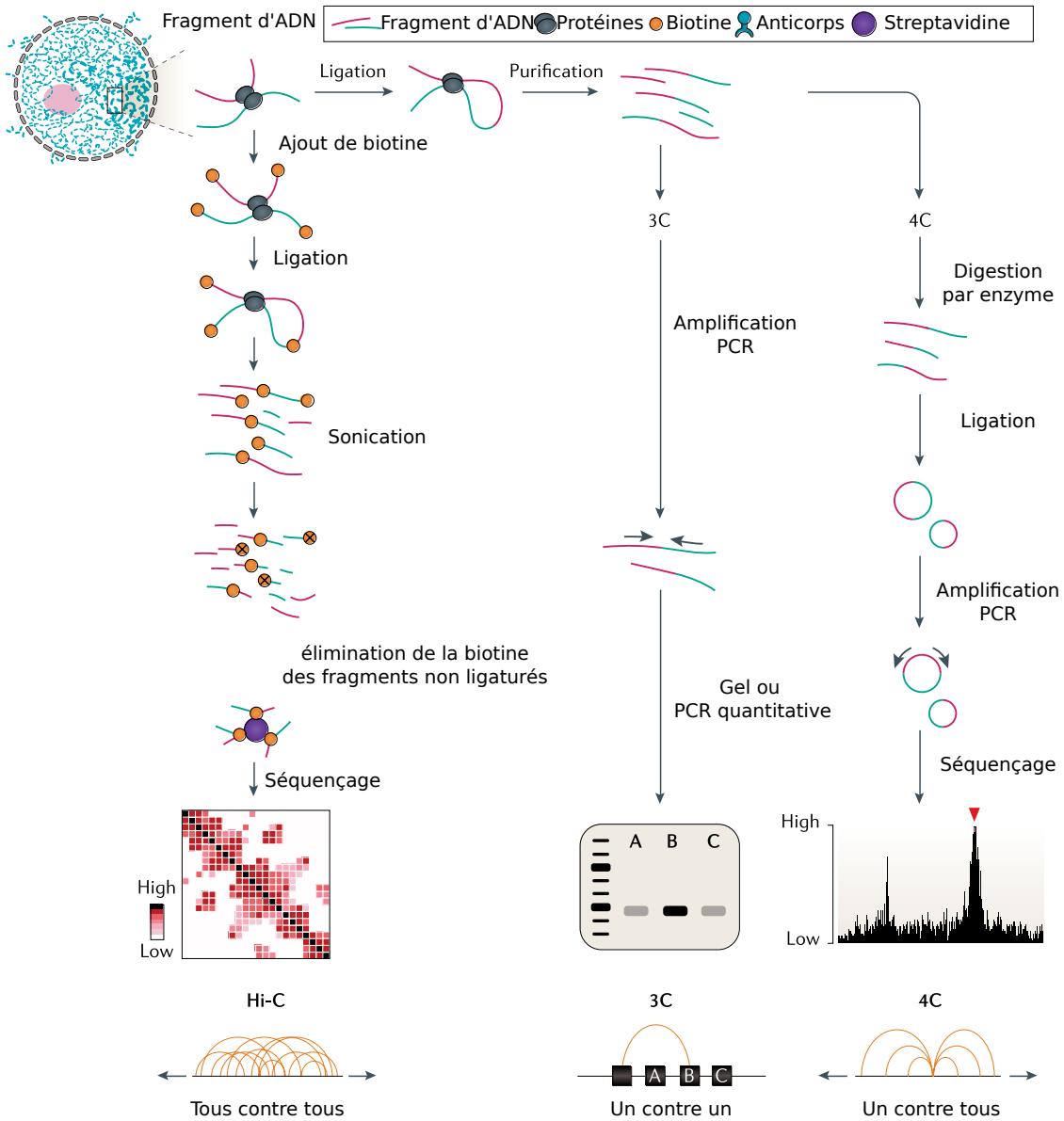


FIGURE 1.31 – Protocoles **3C** et ses dérivés. Le **3C** mesure les fréquences de contacts entre deux *loci* en se basant sur la fixation par *cross-link* et *ligation* de fragments d'ADN à proximité, après fragmentation par des enzymes de restriction. En **3C**, les interactions entre *loci* sont mesurées par **PCR quantitative**, en **4C-seq** et **Hi-C**, celles-ci sont mesurées par **séquençage**. En **Hi-C**, les paires d'interaction sont mesurées entre tous les *loci* du génome, alors qu'en **4C-seq**, on amplifie les interactions d'un *viewpoint* particulier et du reste du génome. Les librairies sont ensuite séquencées par des méthodes classiques du séquençage d'ADN. Adapté de (KEMPFER et POMBO 2020).

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

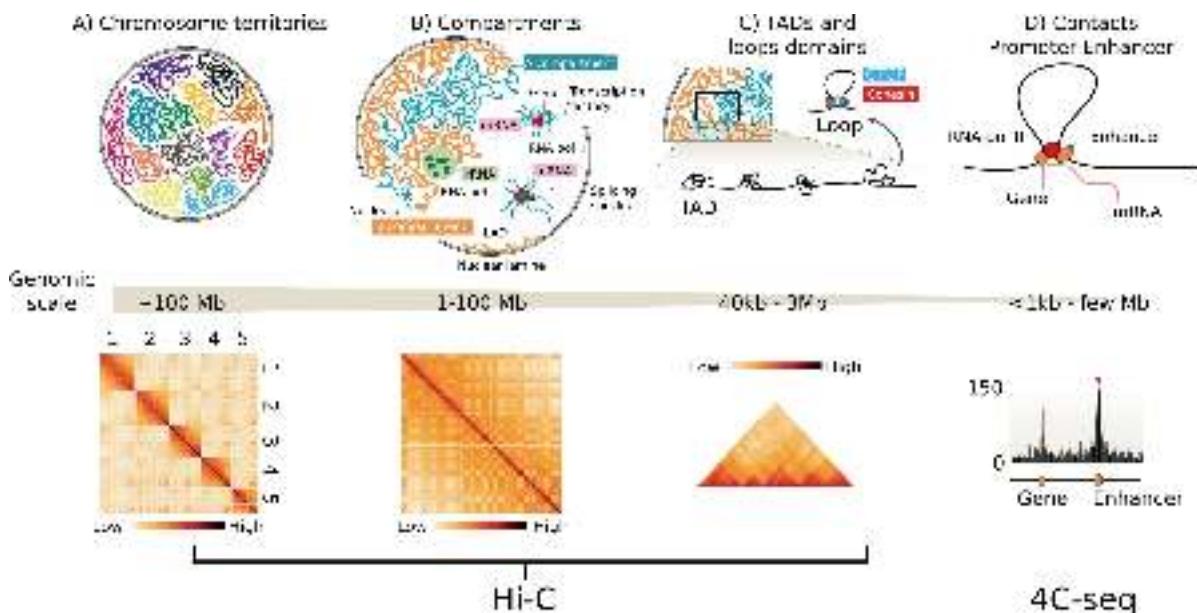


FIGURE 1.32 – Les méthodes 3C omiques comme le Hi-C et le 4C-seq permettent d'étudier la conformation tri-dimensionnelle du génome à différents niveaux d'architecture. Le Hi-C permet d'étudier les interactions des loci d'ADN entre eux à différentes résolutions, mettant en évidence les interactions entre chromosomes, des compartiments au sein d'un même chromosome et des interactions intra-TAD. Le 4C-seq permet lui d'étudier avec précision les interactions très locales à l'échelle du kilobase, comme les interactions promoteurs/enhancers. Adapté de (KEMPFER et POMBO 2020).

utilisée (KRIETENSTEIN et al. 2020). Une approche moins chronophage consiste à pré-tronquer les *reads* qui contiennent un site de ligation et à aligner uniquement la partie la plus longue du *read* (AY et William S NOBLE 2015).

**1.3.6.1.2 Génération des cartes d'interactions Hi-C** Les cartes d'interactions ou de contacts Hi-C sont définis selon une résolution donnée. On regroupe les *reads* en *loci* ayant une taille correspondant à la résolution souhaitée, en divisant le génome de façon linéaire en régions fixes, ou *bins*. Pour chacun de ces *bins*, on compte le nombre de contacts entre chaque paire de *bins*, et on construit la matrice contenant chacune de ces paires. Les étapes d'alignement et de construction des matrices de contacts sont souvent pris en charge par des outils spécialisés, comme **Juicer** (DURAND, SHAMIM et al. 2016). Celui-ci enregistre les matrices Hi-C dans un seul et même fichier compressé **.hic**, qui permet un accès rapide et une utilisation dans un *Genome Browser* spécialisé.

**JuiceBox** (DURAND, J. T. ROBINSON et al. 2016) est un *Genome Browser* spécialisé dans les données Hi-C, et permet de visualiser et de comparer ces données avec d'autres données omiques, comme par exemple des pics ou un profil de ChIP-seq. Il peut également être utilisé comme une base de données en mettant à la disposition des utilisateurs une grande partie des données Hi-C publiées.

**1.3.6.1.3 Normalisation des cartes d'interaction Hi-C** Les contacts observés dans la matrice Hi-C représentent la fréquence d'interaction réelle entre deux *bins*. Cependant, de nombreux biais sont introduits par la méthode elle-même, ainsi que par la difficulté d'obtenir des fragments d'ADNs selon son accessibilité. De même, la fréquence d'interaction entre deux *bins* dépend naturellement de leur

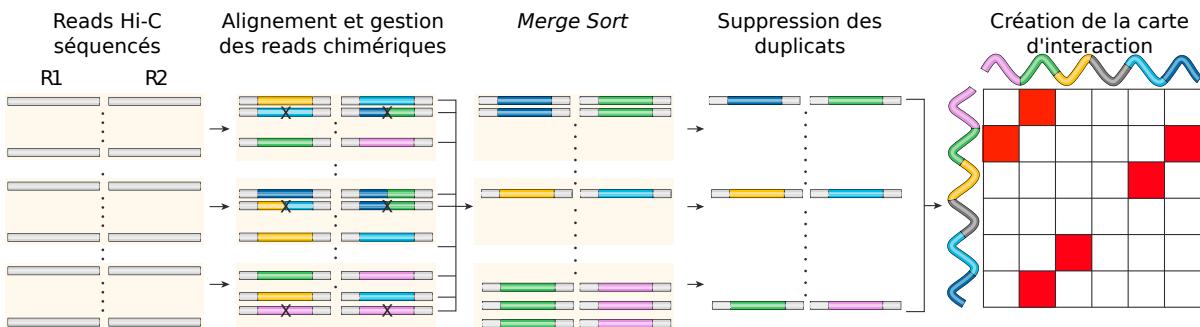


FIGURE 1.33 – Protocole d’alignement des données Hi-C et création de la matrice d’interactions par l’outil Juicer. Adapté de (DURAND, SHAMIM et al. 2016).

distance, et plus ils sont proches, plus le nombre de contacts augmente naturellement, selon la physique des polymères.

**1.3.6.1.3.1 Normalisation des biais** Différentes méthodes ont été introduites afin de corriger les biais liés aux données Hi-C, qui peuvent être divisées en deux groupes : les méthodes **explicites** et **implicites** (LAJOIE, DEKKER et KAPLAN 2015).

Les méthodes explicites modélisent les données Hi-C et utilisent les biais connus en tant que paramètres du modèle pour les supprimer. Par exemple *HiCNorm* (M. HU et al. 2012) modélise les données Hi-C par un modèle de régression de Poisson et inclue dans ses paramètres différents biais connus comme sa capacité à être alignée sur le génome (*mapability*), son taux de *GC* et la taille des fragments, puis considère les résidus comme étant la matrice normalisée Hi-C. L’inconvénient de cette méthode est qu’on doit connaître tous les biais pour normaliser les données correctement.

D’autres méthodes, dites **implicites**, utilisent une autre approche. Les fréquences de contacts entre les régions chromosomiques sont en théorie non-biaisées, c’est-à-dire ne reflétant que l’interaction réelle entre deux polymères d’ADN. On fait donc la supposition que la visibilité de chaque région génomique est identique, et que les biais dans la matrice sont en fait causés par le *locus* lui-même, et impacte l’interaction avec les autres *bins*. En d’autres termes, si un *bin* *i* a un fort biais d’interaction, alors son interaction avec le *bin* *j* sera importante, peu importe que ces deux *bin* interagissent fréquemment ou non. On peut donc normaliser  $M_{ij}$  en prenant en compte le nombre total d’interactions de *i* et de *j*, et définir la matrice normalisée  $M_{ij}^*$  (1.34) :

$$M_{ij}^* = \frac{M_{ij}}{R_i \cdot C_j} \quad (1.34)$$

Où  $R_i$  et  $C_j$  correspondent au nombre total de contacts de dans la ligne (*R*) *i* et la colonne (*C*) *j* respectivement, dans  $M$ . Cette méthode de normalisation, *Vanilla Coverage* a été définie par Lieberman-Aiden et al. (LIEBERMAN-AIDEN et al. 2009).

*Vanilla Coverage* a été améliorée à plusieurs reprises par la méthode Sinkhorn & Knopp (*SK*) (IMAKAEV et al. 2012), et Knight & Ruiz (*KR*) (COURNAC et al. 2012 ; S. S. RAO et al. 2014). Ces deux algorithmes utilisent des méthodes d’équilibrage de matrices ou *matrix balancing*, et sont basés sur l’algorithme de

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

Sinkhorn-Knopp (SINKHORN et KNOPP 1967). L'idée étant de répéter la normalisation de la matrice par la somme des lignes/colonnes jusqu'à convergence d'un critère afin d'obtenir la matrice normalisée.

**1.3.6.1.3.2 Matrice *Observed/Expected* (OE)** La matrice obtenue par Hi-C obéit aux lois de la physique des polymères, ce qui fait que plus deux *loci* sont proches, plus leur probabilité d'interaction est importante. Afin de supprimer cet effet et d'observer uniquement la part d'interaction due à la biologie, on doit normaliser la matrice *Observed* en fonction de la distance linaire entre deux *bins*. Pour calculer le nombre de contacts *Expected* entre deux *bins*,  $i$  et  $j$  à une distance  $|i - j| = d$ , on calcule le nombre total de contacts entre les *bins* étant à la distance  $d$ . On détermine ensuite le nombre de *bins* possibles qui sont séparés par cette distance  $d$ , et on normalise le nombre total par cette valeur. On obtient alors une fonction distance qui dépend du nombre de contacts dans la matrice, et qui peut être utilisée pour faire apparaître sur la matrice la part d'interaction non expliquée par cet effet distance.

**1.3.6.1.4 Déterminer les compartiments A et B de la chromatine** Les compartiments A et B de la chromatine ont été identifiés par Lieberman-Aiden et al. (LIEBERMAN-AIDEN et al. 2009), lors de leur étude des premières données Hi-C. Avec cette résolution, ils ont pu constater que la chromatine est divisée globalement en deux compartiments, qu'ils ont appelés A & B.

Ces compartiments ont été annotés en fonction de l'accessibilité de la chromatine en observant les enrichissements en différentes marques de chromatine par compartiment. Le compartiment A est plutôt corrélé avec des marques de chromatine associées à l'euchromatine (chromatine accessible) et le compartiment B plutôt à de la chromatine permissive, l'hétérochromatine. Ces compartiments dépendent également du type cellulaire, et varient donc entre les cellules.

Les interactions entre *loci* sont contraintes entre régions appartenant au même compartiment, on peut donc facilement visualiser ces compartiments sur une carte de contact Hi-C normalisée (voir partie 1.3.6.1.3.2), où sur la matrice de corrélation calculée à partir de cette matrice (voir Figure 1.34). On peut également les représenter par une analyse en composante principale ou *ACP* (*PCA* en anglais) (voir Figure 1.34). Ces compartiments ont été définis par la première composante de l'ACP sur la matrice *OE*. Spécifiquement, une frontière entre les deux compartiments est définie par un changement de sens de la première composante de l'ACP.

L'*ACP* est une méthode de réduction de dimension basée sur les combinaisons linéaires des variables existantes (ici les interactions entre les bins), afin d'en créer de nouvelles. Le premier axe de l'ACP est construit en modélisant une ligne, ou un axe à partir des données (voir 1.35 A) en minimisant la perte d'inertie. L'axe passe par la moyenne (représentée en rouge sur la figure), et se construit de telle façon que cette droite doit minimiser l'écart orthogonal aux observations en 2 dimensions. Le second axe est construit lui aussi avec les variables existantes, mais en étant totalement dé-corrélaté du premier axe (voir Figure 1.35 A), et explique le restant de variance (sur un problème à deux dimensions). La solution s'applique également à  $n$  dimensions, avec un nombre de composantes égal au nombre de variables initiales, qui peut se calculer de façon analytique en utilisant la décomposition en valeur propre, vecteur propre d'une matrice.

À la fin, on projette les points sur les nouveaux axes afin d'obtenir leurs nouvelles coordonnées (voir 1.35 B), sur un espace réduit qui maximise la variance des données. En pratique, on choisit rarement plus de

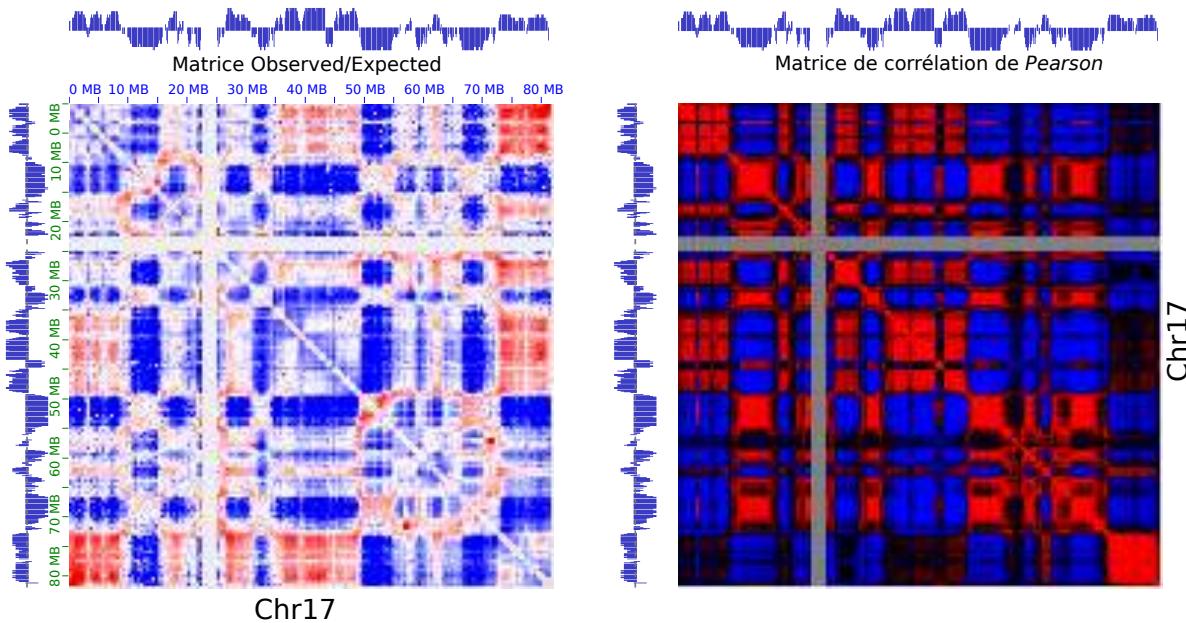


FIGURE 1.34 – Carte de contact *Observed/Expected* du chromosome 17 et matrice de corrélation de Pearson associé. Le profil autour des heatmaps correspond à la première composante de l'ACP calculée à partir de la matrice de contact.

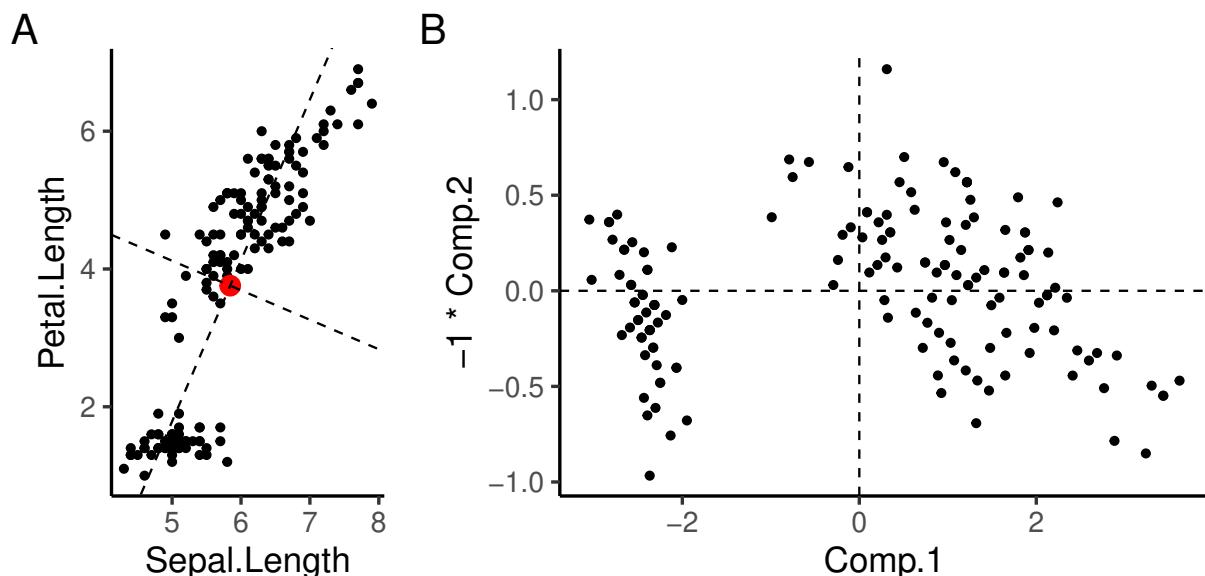


FIGURE 1.35 – Analyse en composante principale du jeu de données *iris* sur deux dimensions (*Petal.Length* et *Sepal.Length*). **A** Les axes de l'ACP sont projetés sur les dimensions d'origine et passent par la moyenne (rouge). **B** Projection des points dans l'espace de l'ACP.

3 dimensions de façon à pouvoir projeter les points sur un graphique, ou dans le cas de la carte de Hi-C, on utilise le premier axe pour déterminer les compartiments A/B.

**1.3.6.1.5 Détermination des TADs** La chromatine est également formée de compartiments à l'échelle plus locale, les *TADs*, qui définissent des régions où les *loci* interagissent préférentiellement entre eux. Pour déterminer ces TADs, plusieurs méthodes existent, dont le *Directionality Index (DI)* (DIXON et al. 2012), l'utilisation des matrices *ArrowHead* (S. S. RAO et al. 2014), le score d'insulation (CRANE et al. 2015) ou *topDom* (SHIN et al. 2016).

**1.3.6.1.5.1 Calcul du score d'insulation** Le score d'insulation calcule le nombre moyen d'interaction d'un bin avec son entourage  $M_i$  (voir les carrés dans la Figure 1.36 D), puis normalise ce score par la moyenne de tous les scores d'insulation calculés le long du chromosome (1.35) :

$$IS = \log_2\left(\frac{M_i}{\frac{1}{p} \sum_{i=1}^p M_i}\right) \quad (1.35)$$

Plus la valeur d'*Insulation Score IS* est grande, plus le *bin* est localisé à l'intérieur du *TAD*. Une frontière entre deux *TAD* est attendue à chaque minimum local de *IS*.

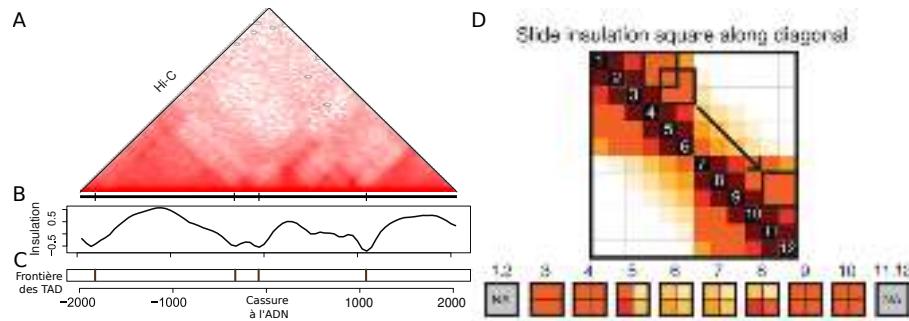


FIGURE 1.36 – Score d'insulation calculé à partir de la carte Hi-C (**A**) centré autour d'une cassure double-brin (DSB). Le score d'insulation (**B**) permet de déterminer les frontières des TADs (**C**) via une méthode illustrée en (**D**). (**D**) Adapté de CRANE et al. 2015

Pour une matrice Hi-C donnée, on peut donc déterminer la position d'un *TAD* en utilisant le score d'insulation calculé le long du chromosome (voir la Figure 1.36 A-C), et éventuellement déterminer des renforcement ou relâchement des frontières de ces *TADs* en comparant les scores d'insulation entre deux conditions.

**1.3.6.1.5.2 Recherche de minimums locaux avec topDom** Topdom (SHIN et al. 2016) se base sur une recherche de minimum locaux, comme le score d'insulation. Cette recherche se base sur une métrique différente, qui assume que la fréquence d'interaction entre les *bins* avant et après une frontière de TAD est moins élevée que ceux à l'intérieur des *TAD*. Cette méthode, *binSignal* se calcule de la façon suivante (1.36) :

$$binSignal_i = \frac{1}{w^2} \sum_{l=1}^w \sum_{m=1}^w \mathbf{M}_{U_i(l), D_i(m)} \quad (1.36)$$

$binSignal_i$  compte les interactions en amont et en aval en utilisant une fenêtre de taille  $w$  ou pour un bin  $i$ ,  $U_i = \{i - w - 1, i - w, \dots, i\}$  et  $D_i = \{i + 1, i + 2, \dots, i + w\}$ , et représente le contact moyen du bin et de son entourage.

Intuitivement, on s'attend à avoir un score important au centre du *TAD* et un score minimum local aux frontières du *TAD*. Pour éviter de détecter des minimums locaux produits par un signal bruité, *TopDom* lisse la courbe de *binSignal* le long du chromosome. Il construit ensuite de façon répétée une fonction linéaire la plus longue possible mais minimisant les résidus, jusqu'à la fin du chromosome. De cette façon, *TopDom* construit de multiples régressions locales qui permettent d'identifier un minimum ou maximum local en identifiant un changement du sens de la dérivée des fonctions. Le minimum local sera ensuite identifié comme ayant la fréquence d'interaction la plus faible.

**1.3.6.1.6 Détermination de la position des boucles de chromatine** Les boucles de chromatine sont formées par le contact physique de deux *loci* appartenant au même chromosome. Celles-ci peuvent être visibles sur les matrices Hi-C lorsque celle-ci est suffisamment résolutive, et lorsque ces deux *loci* sont en contact dans une proportion suffisante de la population cellulaire.

*HiCCUPS* (S. S. RAO et al. 2014) est un algorithme permettant de localiser les boucles à la manière d'un *peak caller* 2D (voir partie 1.3.2). Il va identifier une boucle d'ADN en examinant chaque pixel de la carte de Hi-C et en comparant le nombre de contacts par rapport aux autres pixels dans son entourage proche. Un pixel est considéré enrichi lorsque sa fréquence de contact est plus haute qu'attendue, et qu'il ne fait pas partie d'une plus grande structure. Pour calculer l'enrichissement, le pixel est comparé à 4 voisinages (voir Figure 1.37), et doit être au minimum 50% plus enrichi. Les pixels enrichis sont contigus et forment une région d'une taille comprise entre 5 et 20 pixels. La région considérée comme étant le pic est alors celle avec le plus d'interactions. À cause du grand nombre de pixel à tester, *HiCCUPS* est implémenté pour tourner en parallèle, et peut utiliser les cartes graphiques.

**1.3.6.1.7 APA** Pour tester l'enrichissement des boucles de chromatines (ou autre élément), on peut réaliser une analyse d'agrégation du signal, ou *Aggregate Peak Analysis* (*APA*). Pour cela, on définit un ensemble de sous-matrices carrées dérivées de la matrice Hi-C, centrée sur les régions d'intérêts. On calcule ensuite la somme de toutes ses sous-matrices pour obtenir l'*APA*, qui représente le nombre total de contacts des pics au centre, et de la région avoisinante pour le reste de la matrice. On peut calculer un enrichissement global (comme on détermine l'enrichissement d'un pixel pour les boucles) en comparant le nombre de contacts du pixel central avec des pixels avoisinants.

### 1.3.6.2 Étude des interactions entre un *locus* et le reste du génome (4C-seq)

Le 4C-seq est une méthode efficace et à moindre coût (par rapport au Hi-C) pour étudier la conformation 3D du génome pour une région d'intérêt. En amplifiant les interactions spécifiques entre un *locus* et le reste du génome, on peut, avec une faible profondeur (de 1 à 5 millions de *reads* (WERKEN, LANDAN et al. 2012)), obtenir une résolution importante à l'échelle du kilobase. Cette méthode s'applique particulièrement pour détecter des interactions courtes (en *cis*), mais fonctionne également avec les interactions longue distance, à plus faible résolution. Pour construire une librairie de 4C-seq, on doit définir au minimum un *viewpoint*, qui sera localisé entre deux sites de restrictions. Pour obtenir la

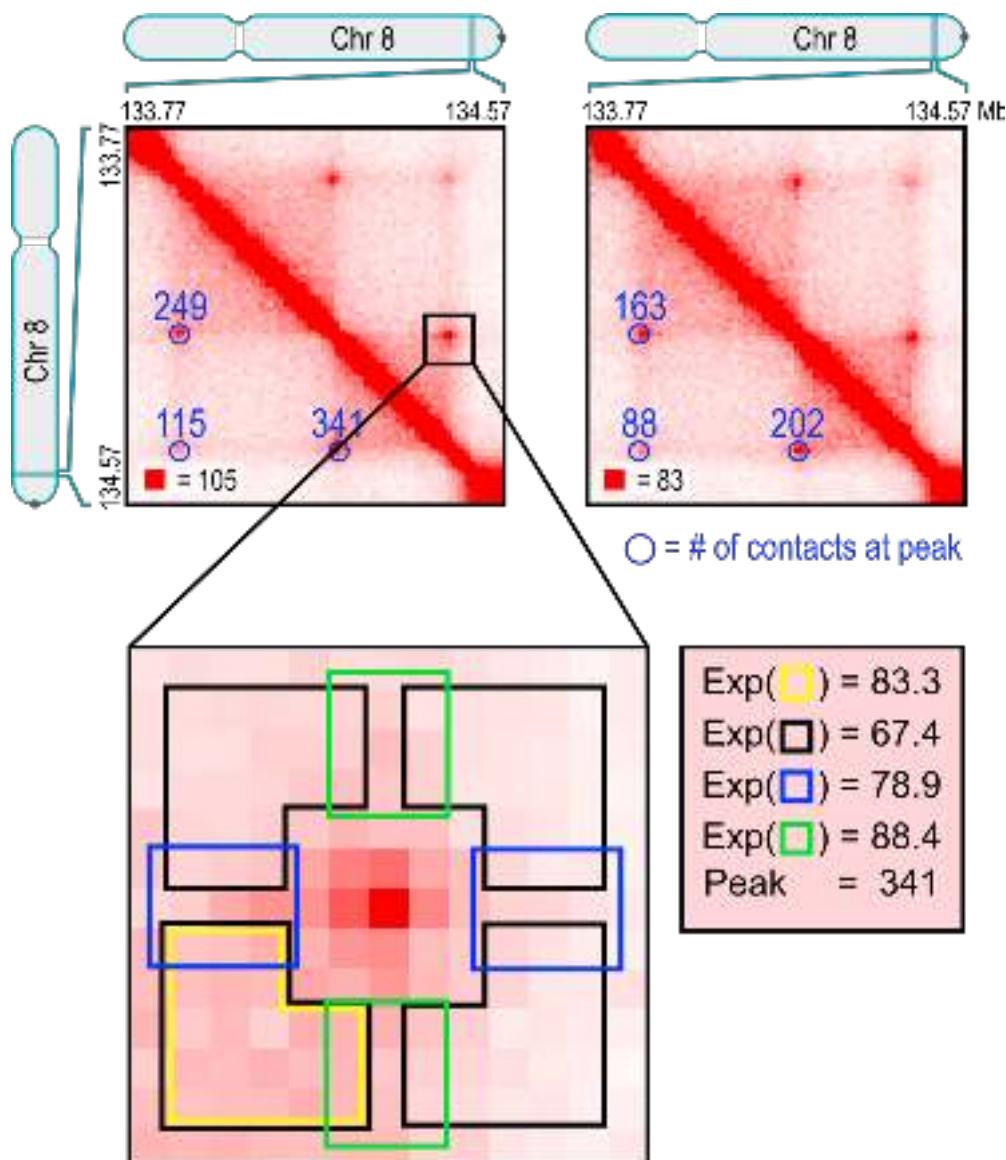


FIGURE 1.37 – Détection des boucles de chromatine en utilisant l'enrichissement local. Pour identifier une boucle, chaque pixel est comparé à 4 régions voisines, définie respectivement en jaune, vert, bleu et noir sur le dessin. Adapté de (S. S. RAO et al. 2014).

### **1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME**

---

meilleure résolution possible, les méthodes de 4C effectuent deux tours de digestion par deux enzymes de restriction (WERKEN, LANDAN et al. 2012).

La librairie construite par 4C-seq contient majoritairement de l'ADN issu des partenaires de ligation du *viewpoint*, plus sa séquence *primers*. On peut séquencer en *single-end* une des extrémités du fragment, puis, par une opération de demultiplexage du *fastQ*, extraire les *reads* qui contiennent la séquence du *primer* spécifique au(x) *viewpoint(s)*. À la fin de cette étape, on obtient un *fastQ* par *viewpoint*, qu'on peut ensuite aligner et traiter avec une méthode classique d'alignement (voir partie 1.3.1.4.2.1).

#### **1.3.7 Bases de données génomiques**

Les méthodes *NGS*, et omiques de manière générale ont révolutionnés l'étude de la biologie en proposant une approche *genome-wide* jusqu'alors impossible à atteindre. Les informations potentielles d'une telle expérience dépasse cependant généralement celles pour laquelle elle a été réalisée. Rendre ces données disponibles peut permettre de les utiliser pour répondre à de nouvelles hypothèses. La création et la maintenance de bases de données génomiques est donc essentielle, à la fois pour des questions de reproductibilité, mais aussi comme formidable outil de recherche pour la communauté scientifique.

Il existe de nombreuses bases de données recensant les données omiques, et on peut les classer en fonction du type de données proposé :

- Les bases de données brutes, contenant uniquement les données sorties du séquenceur, qu'il faut analyser soi-même, comme la *Sequence Read Archive (SRA)*.
- Les bases de données partiellement analysées, qui contiennent des données très hétérogènes, de formats très différents et analysés par des méthodes diverses, comme *Gene Expression Omnibus (GEO)*, du *NCBI* ou bien *ArrayExpress*, de l'*EBI*.
- Enfin les bases de données complètes, qui contiennent un ensemble de données analysées de façon uniforme, avec un format de sortie standardisé. Ces bases de données sont généralement centrées sur un type de données, comme *ExpressionAtlas* (transcriptomique), *ENCODE* (épigénome), ou encore *dbSNP* et *GWAS Catalog* (variants génétiques).

##### **1.3.7.1 Les bases de données d'annotations et de génomes**

Il est nécessaire de disposer d'un génome de référence et d'une annotation de référence lorsque l'on traite des données produites par *NGS*, comme expliqué dans les parties 1.3.1.3 et 1.3.5.1. Les grandes institutions de santé proposent des bases de données complètes (le *NCBI* et l'*EBI* notamment), permettant de télécharger les assemblages et les annotations des génomes de nombreuses espèces, ainsi que différents outils bio-informatiques permettant de faire de la comparaison de séquences, comme *BLAST* (ALTSCHUL et al. 1990).

Ensembl (HOWE et al. 2020) est le projet européen (*EBI*) d'annotation de génomes de vertébrés, il est très utile pour l'analyse des gènes, car il dispose d'une annotation très complète, avec un identifiant unique (*ENSG*) pour chaque gène connu. Les différents transcrits (*ENST*) et exons (*ENSE*) sont annotés, et un ensemble de règles permet de définir un transcript canonique pour un gène donné<sup>11</sup>. L'annotation se fait

---

11. <http://mart.ensembl.org/info/genome/genebuild/canonical.html>

### 1.3. LA GÉNOMIQUE, OU L'ÉTUDE DE L'ENSEMBLE DU GÉNOME

---

de manière automatique, mais se base sur des annotations manuelles et existantes<sup>12</sup>. Ensembl propose une instance de BioMart (SMEDLEY et al. 2009) sur leur site<sup>13</sup> de façon à sélectionner et télécharger des annotations selon les filtres désirés, et ainsi construire celle de son choix. On peut par exemple utiliser l'ensemble des gènes annotés dans Ensembl (67128) ou bien uniquement partir sur les gènes codant une protéine ou *protein coding* (22796), pour la version 104 (Mai 2021). Il existe également d'autres bases de données d'annotation de gènes, notamment *RefSeq*, qui est le projet américain du *NCBI*, qui lui se base sur des curations manuelles et prédictives (O'LEARY et al. 2016).

#### 1.3.7.2 Les bases de données pour les publications scientifiques

*ArrayExpress* (ATHAR et al. 2019) et *GEO* (BARRETT et al. 2012) sont les deux bases de données utilisées pour le stockage de données omiques publiées dans les articles scientifiques. Ces deux bases de données sont primordiales pour la transparence et la reproductibilité de la recherche scientifique, ainsi qu'un outil formidable pour l'utilisation de ces données pour de nouvelles problématiques.

La publication sur ces bases de données se font de manière plus ou moins automatique. Sur *ArrayExpress*, chaque soumission doit être faite dans leur outil spécialisé *Annotare*. Celle-ci doit être décrite précisément, chaque fichier doit être vérifié (via empreinte digitale avec *MD5*), associé à une expérience et à des protocoles détaillés. La soumission est ensuite vérifiée et validée manuellement. Une fois validé, les données deviennent disponibles à la communauté scientifique après publication, sous la forme d'un numéro d'accès.

#### 1.3.7.3 Les catalogues spécialisés

Les bases de données *GEO* ou *ArrayExpress* permet à la communauté scientifique d'utiliser et de reproduire les analyses publiées par leurs pairs. Cependant, la quantité de données est si grande, et la qualité si hétérogène, qu'il est parfois impossible de les utiliser pour ses propres analyses. De ce fait, des projets ont vu le jour afin de proposer une analyse bio-informatique uniforme des données publiques, permettant à la communauté scientifique d'y accéder et de les utiliser.

Ces outils se présentent souvent sous la forme de catalogues spécialisés dans un type de données précis. Par exemple, *ENCODE* (E. P. CONSORTIUM et al. 2012) est un projet spécialisé dans les données omiques associées à l'épigénétique, et contient des données publiées, mais aussi des données spécialement produites pour ce projet. *ENCODE* donne l'accès aux données brutes et analysées de séquençage. Le site met également à disposition des chaînes de traitements afin d'appliquer les mêmes protocoles que leurs analyses, ce qui permet d'avoir une feuille de route fiable pour l'analyse des données omiques.

D'autres catalogues proposent de reprendre les données publiées et de les re-analyser de façon uniforme, comme le ChIP-Atlas (OKI et al. 2018) qui propose la visualisation et le téléchargement de données de ChIP-seq publiques. ReMap (CHÈNEBY et al. 2020) est un projet similaire qui permet de télécharger les sites de liaison de protéines à l'ADN, provenant globalement des données ChIP-seq. L'avantage de ReMap est qu'il propose des positions non-redondantes sur le génome, ce qui facilite grandement l'analyse.

---

12. <https://www.ensembl.org/info/genome/genebuild/index.html>

13. <https://www.ensembl.org/biomart/martview>

Enfin, certains projets regroupent des données de séquençage de cohortes de patients atteints de cancers. *TCGA* est un corpus de données génomiques, transcriptomiques et épigénomique de milliers de patients atteints de cancer, regroupés en cohortes. De même, L'*International Cancer Genome Consortium (ICGC)* propose des données génomiques provenant de plus de 80 cohortes totalisant plus de 2000 patients. Les informations sont déjà analysées, et on a accès à l'expression des gènes et aux études des variants génétiques pour chaque cohorte.

Les catalogues sont des outils formidables pour la bio-informatique, et révolutionnent le travail scientifique en biologie, permettant de faire des découvertes sans faire d'expériences.

### 1.3.7.4 Les bases de données de motifs (*PWM*)

Un grand nombre de motifs reconnus par des facteurs de transcription sont recensés dans les bases de données publiques sous la forme de *PWM*. *JASPAR* (SANDELIN et al. 2004), *TRANSFAC* (WINGENDER et al. 2000), *HOCOMOCO* (KULAKOVSKIY et al. 2018) et *cisBP* (WEIRAUCH et al. 2014) sont parmi les bases de données les plus utilisées. L'avantage de *JASPAR* est qu'on peut l'utiliser directement dans R via un package Bioconductor. On dispose ainsi directement de la base de données ainsi que du choix de la version (de 2014 à 2020 sont disponibles sur Bioconductor) et on peut effectuer des requêtes sur les 746 motifs sur la base de données JASPAR 2020<sup>14</sup>.

## 1.4 Prédiction de données génomiques

### 1.4.1 Apprentissage Automatique (*Machine Learning*)

L'apprentissage automatique ou *Machine Learning (ML)* est une sous branche de l'intelligence artificielle (*IA*) permettant à un automate ou programme d'apprendre des règles sans qu'elles soient explicitement programmées. Le *ML* propose une approche différente de l'*IA* classique qui consiste à programmer un automate selon un ensemble de règles strictes et immuables. À l'inverse, le *ML* va utiliser des données, un jeu d'apprentissage ou d'entraînement, qui permettra au modèle d'apprendre ces règles (voir Figure 1.38).

Le *ML* permet, à l'inverse de l'*IA* classique de proposer une solution simple à un problème, et donne souvent de meilleurs résultats. Il permet également de proposer une solution là où des méthodes traditionnelles n'en trouveraient pas, et peut s'adapter à de nouvelles données de manière dynamique. Enfin, le *ML* s'adapte particulièrement bien aux gros volumes de données, ce qui le rend très utile en génomique.

Cependant, il requiert l'utilisation de données pré-existantes, et plus le problème est complexe, plus le jeu de données d'entraînement devra être conséquent. Dans le cas du *Deep Learning* (voir partie 1.4.3), des milliers voir des millions de données sont parfois nécessaires à la construction d'un modèle prédictif, ce qui restreint son utilisation. Les méthodes paramétriques classiques, telles que les modèles linéaires, nécessitent moins de données mais ne fonctionnent pas sur tous les types de données.

Enfin, il existe principalement deux types d'Apprentissage Automatique : supervisé et non-supervisé. L'apprentissage supervisé résout des problèmes de classifications, et utilise des étiquettes, ou *labels* pour

---

14. JASPAR2020 CORE vertébrés, non redondant

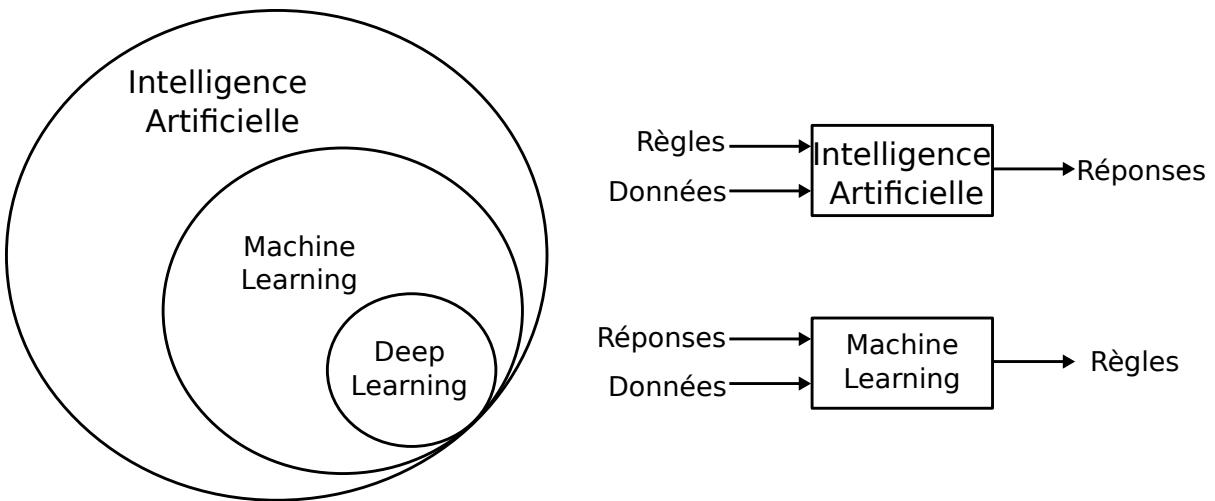


FIGURE 1.38 – L'intelligence artificielle (IA), Le *Machine Learning* (ML) et le *Deep Learning* (DL). l'IA est basé sur des règles expertes inflexibles pour résoudre un problème. Le ML et le DL changent de paradigme et utilisent des données d'entraînements pour déterminer les règles.

ajuster ses paramètres. L'apprentissage non-supervisé est sans étiquettes, et son but est généralement de simplifier des problèmes complexes, par exemple en visualisation de dimensions complexes (comme l'ACP, partie 1.3.6.1.4).

Le *Machine Learning* est très souvent utilisé en génomique, et permet de résoudre de nombreux problèmes, comme la prédiction des régions codantes/non-codantes de l'ADN (en utilisant par exemple des modèles de Markov (BURGE 1997)) ou bien l'impact des variants génétiques à l'échelle du génome (Ho et al. 2019). Dans les parties suivantes, les principales méthodes pour construire un modèle (1.4.1.1), évaluer sa précision (1.4.2), mais aussi comprendre l'importance des variables utilisées pour la prédiction (1.4.1.3) seront décrites. Enfin, la partie 1.4.3 décrira comment fonctionne l'apprentissage profond, ou *Deep Learning* lorsque l'on utilise des séquences d'ADN.

#### 1.4.1.1 Apprentissage d'un modèle de *Machine Learning*

L'objectif d'un algorithme de *Machine Learning* est bien souvent de minimiser une fonction de coût  $C$  pour des prédictions  $\hat{y}$  à partir d'un jeu d'entraînement  $y$  et  $x$  (1.37) :

$$C(y_i, \hat{y}_i) \quad (1.37)$$

$$\hat{y}_i = f(W_1 x_i)) \quad (1.38)$$

Les fonctions de coûts sont multiples, et permettent de répondre à des problèmes très variés. Pour résoudre un problème de régression, on peut chercher à utiliser la méthode des moindres carrés, ou *Mean Squared Error* (*MSE*) (1.40), ou encore la *Mean Absolute Error*. Pour un problème de classification à plus de deux classes (*softmax*), on cherche à minimiser l'entropie croisée (ou *cross entropy*). Pour un problème à deux classes (*logit*), on utilise la *binary cross entropy* ou perte logistique.

Afin de modéliser la relation entre deux variables,  $X$  et  $Y$  sous la forme d'une régression linéaire, on peut écrire la relation suivante comme dans l'équation (1.39) :

$$Y = \alpha + \beta X + \varepsilon \quad (1.39)$$

Les deux paramètres de la droite  $\alpha$  et  $\beta$  doivent être construits de façon à expliquer le mieux cette relation. Pour construire, ou entraîner ces paramètres, on peut utiliser la *MSE* (1.40) :

$$MSE(X, f_{\Theta}) = \frac{1}{m} \sum_{i=1}^m (\hat{y}_i - y_i)^2 \quad (1.40)$$

$$\hat{y}_i = f(x_i) = \Theta^T x_i \quad (1.41)$$

ou  $\hat{y}_i$  représente la prédiction de  $y_i$  en utilisant la fonction de régression linéaire  $f$ . Cette méthode calcule la moyenne des erreurs (à la prédiction) carrés  $(\hat{y}_i - y_i)^2$ , de la droite  $\Theta^T x_i$ .

**1.4.1.1.1 Minimiser MSE par l'équation normale** Pour la régression linéaire, il existe une solution analytique qui permet de trouver  $\hat{\Theta}$  qui minimise  $MSE(X, f_{\Theta})$ , c'est l'équation normale (1.42) :

$$\hat{\Theta} = (X^T X)^{-1} X^T y \quad (1.42)$$

Dans cette équation,  $\hat{\Theta}$  est l'estimateur de  $\Theta$  qui minimise la fonction de coût, et  $y$  à prédire par  $X$ . Dans le cas d'une régression linéaire simple à deux paramètres  $\alpha$  et  $\beta$  :  $\Theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ .

En reprenant la relation du jeu de données `iris` partie 1.3.5.2.1, on peut estimer les paramètres  $\hat{\Theta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1.08 \\ 2.23 \end{pmatrix}$  qui minimisent les erreurs (ou *résidus*) de la droite ( $\hat{y}_i$ ), par rapport à  $y_i$  (voir Figure 1.39 A).

**1.4.1.1.2 Minimiser MSE par descente de gradient** Quand il n'existe pas de solution analytique, ou que le nombre d'observations devient trop grand, on peut utiliser une méthode très générale qui fonctionne avec (presque) toutes les fonctions de coûts et qui permet de trouver une solution optimale à un grand nombre de problèmes. Cette méthode, la *descente de gradient*, va corriger de manière itérative les paramètres jusqu'à trouver le minimum de la fonction de coût.

Pour corriger le paramètre  $\Theta_j$  de façon optimale, on a besoin de connaître dans quelle direction ira cette fonction lorsque l'on le modifie. Pour cela, on peut calculer la dérivée partielle de la fonction *MSE* (1.43) :

$$\frac{\partial}{\partial \Theta_j} MSE(\Theta) = \frac{2}{m} \sum_{i=1}^m (\Theta^T x_i - y_i) \cdot x_{ij} \quad (1.43)$$

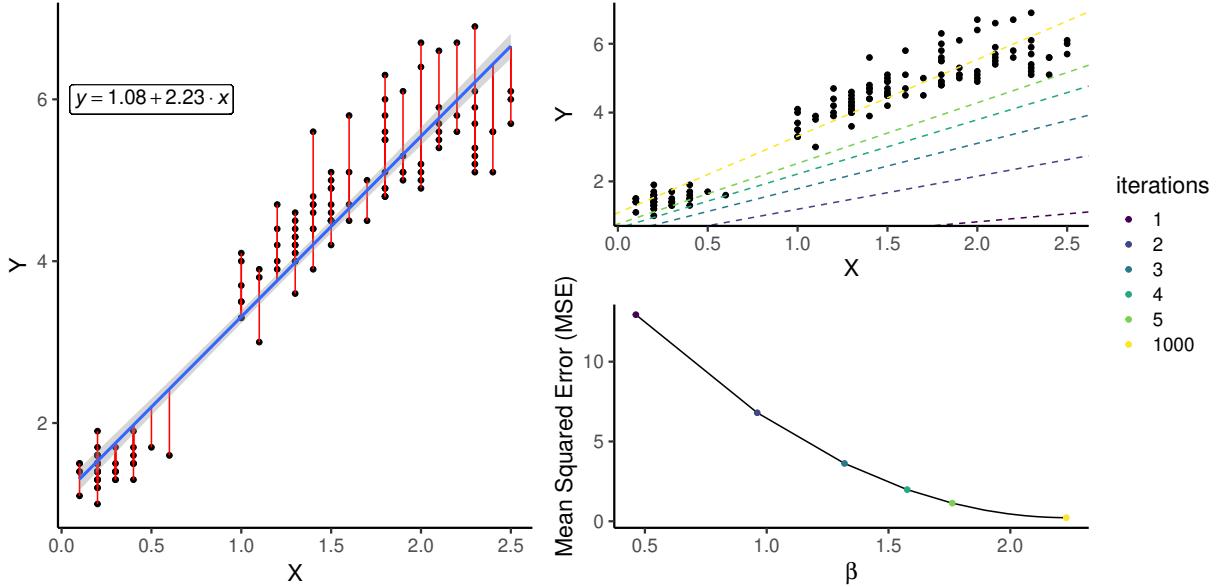


FIGURE 1.39 – Estimation des paramètres d’une régression linéaire par *MSE* en utilisant l’équation normale (**A**), ou par *descente de gradient* (**B,C**)

On peut également calculer le **vecteur gradient**,  $\nabla MSE(\Theta)$ , qui contient l’ensemble des dérivées partielles via une formulation vectorielle (1.44) :

$$\begin{aligned} \nabla MSE(\Theta) &= \left( \begin{array}{l} \frac{\partial}{\partial \alpha} MSE(\Theta) = \frac{2}{m} \sum_{i=1}^m (\Theta^T x_i - y_i) \\ \frac{\partial}{\partial \beta} MSE(\Theta) = \frac{2}{m} \sum_{i=1}^m (\Theta^T x_i - y_i) \cdot x_i \end{array} \right) = \frac{2}{m} X^T (X\Theta - y) \end{aligned} \quad (1.44)$$

Dans le cas où  $\Theta = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}$ .

Une fois qu’on a calculé le vecteur gradient, on peut mettre à jour les  $\Theta$  (1.45) :

$$\Theta = \Theta - \eta \cdot \nabla MSE(\Theta) \quad (1.45)$$

Où  $\eta$  correspond au taux d’apprentissage, ou pas de progression, un *hyper-paramètre* de l’algorithme de descente de gradient qui peut être optimisé (voir partie 1.4.3.9). En appliquant cet algorithme itératif 1000 fois, on peut retrouver la même solution qu’avec l’équation normale :  $\hat{\Theta} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} 1.08 \\ 2.23 \end{pmatrix}$ , et *MSE* diminue à chaque itération (Figure 1.39 B, C). Par défaut, on utilise l’ensemble du jeu de données pour mettre à jour les paramètres, on parle alors de **descente de gradient par batch**. À l’extrême opposé, on peut utiliser une observation au hasard pour mettre à jour les poids, on parle dans ce cas-là de **descente de gradient stochastique**. Cette méthode est très rapide et utilise peu de mémoire, car elle estime les paramètres sur une seule observation, et permet d’éviter que l’algorithme se retrouve coincé dans un *minimum local*. La nature aléatoire de l’observation rend cependant la convergence des paramètres plus difficile. La **descente de gradient par mini-batch** est un bon compromis entre les deux. Elle est moins hasardeuse que la méthode *stochastique* et permet de tirer profit des processeurs graphiques.

### 1.4.1.2 Prédiction par arbre de décision

Les arbres de décisions (BREIMAN et al. 2017) sont des algorithmes de *Machine Learning* très populaires, qui peuvent effectuer à la fois des tâches de régression et de classification. À la différence des modèles linéaires et *GLM*, ils sont non paramétriques et ne font donc pas d'hypothèses sur les résidus, et ne nécessitent pas de fonction de lien entre les données à prédire et les prédicteurs. Les arbres de décisions sont aussi et surtout les composants d'un algorithme très puissant : les forêts aléatoires, ou *random forest* (*RF*) (BREIMAN 2001).

En pratique, on utilise souvent des arbres binaires et l'algorithme *CART* (*Classification And Regression Tree*), pour les construire. L'algorithme sépare le jeu de données en deux sous-ensembles en utilisant une variable  $k$ , et un seuil  $t_k$ . Pour choisir la paire  $(k, t_k)$ , il utilise une fonction de coût qui produit le sous-ensemble le plus pur. Cette pureté, ou impureté, est appelée score de Gini  $G_i$  et se calcule de la manière suivante (1.46) :

$$G_i = 1 - \sum_{k=1}^n (p_{i,k})^2 \quad (1.46)$$

Où  $p_{i,k}$  représente le pourcentage d'observations de la classe  $k$  dans le nœud  $i$ . L'algorithme applique cette logique de séparation des données récursivement jusqu'à ce qu'on atteigne la profondeur maximale (hyper-paramètre de l'algorithme) ou bien jusqu'à ce qu'il n'existe plus de partage qui réduit l'impureté.

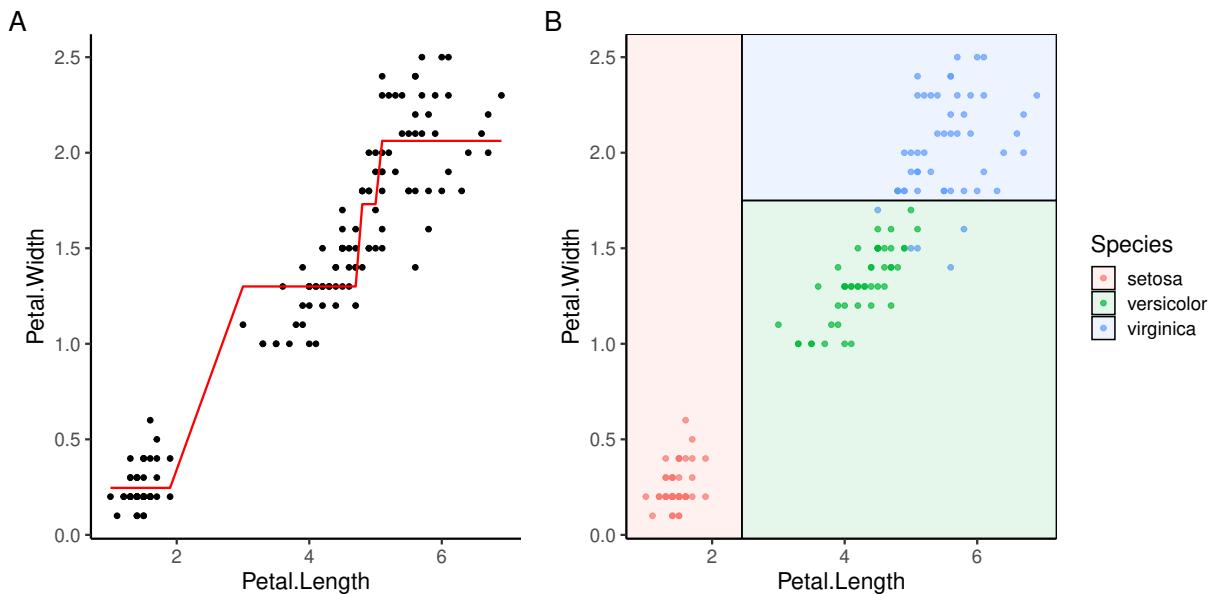


FIGURE 1.40 – Méthode de régression/classification par arbre de décision. **A** Prédiction d'une variable quantitative par arbre de décision, en minimisant *MSE* pour chaque nœud **B** Prédiction d'une variable qualitative par arbre de décision, en minimisant le score *Gini* pour chaque nœud.

On peut également utiliser une autre métrique, l'entropie, comme illustrée dans l'équation (1.11). Pour les tâches de régression, on cherchera à partager le jeu d'entraînement en minimisant la *MSE* entre les observations des deux sous-ensembles.

Pour prédire une nouvelle donnée, on traverse l'arbre de la racine à un nœud terminal qui nous donne

## 1.4. PRÉDICTION DE DONNÉES GÉNOMIQUES

---

TABLE 1.6 – Matrice de confusion pour une regression logistique ayant pour objectif de prédire une des deux espèces de fleurs, versicolor ou virginica.

	versicolor	virginica
versicolor	28	2
virginica	3	17

ensuite la prédiction de notre nouvelle observation.

Les *RF* sont composées d'un ensemble d'arbres de décisions. Elle appartient donc aux méthodes ensemblistes qui consiste à regrouper des prédicteurs pour réaliser un consensus. Ces méthodes donnent souvent de meilleurs résultats qu'avec le meilleur prédicteur pris séparément. Plutôt que d'utiliser des méthodes différentes de prédictions, un algorithme de *RF* va utiliser le même algorithme (*CART*) sur des sous-ensembles différents du jeu d'entraînement. On appelle cette méthode le *bagging* (pour *bootstrap aggregating*) pour un tirage avec remise, le *pasting* pour un tirage des variables sans remise.

### 1.4.1.3 Calcul de l'importance des variables par *Machine Learning*

Une grande qualité des forêts aléatoires est qu'elles permettent de facilement déterminer l'importance prédictive d'une variable. Différentes méthodes existent, comme la *Mean Decrease Accuracy* (*MDA*), qui est calculée sur la précision de prédiction d'un arbre  $t$  avant et après avoir fait une permutation aléatoire d'une variable  $j$ . Ce changement de précision après permutation est ensuite moyenné sur tous les arbres de la forêt et est utilisée comme mesure de l'importance de la variable  $j$ .

Cette mesure est efficace, mais coûteuse en temps de calcul. On peut donc utiliser une autre méthode, la *Mean Decrease Impurity* (*MDI*), qui se base sur la mesure d'impureté du score Gini (voir (1.46)). Pour estimer l'importance d'une variable, on calcule la réduction moyenne pondérée (par le nombre d'observations) d'impureté sur les nœuds utilisant cette variable, pour tous les arbres de la forêt.

### 1.4.2 Évaluation d'un modèle de *Machine Learning*

Si la fonction de perte permet d'optimiser les paramètres du modèle pour donner la meilleure prédiction possible, sa valeur fournie une métrique peu informative, et on choisit généralement un indicateur plus informatif (et pas forcément différentiable), comme la *précision*, l'*exactitude* ou le *rappel*. Pour évaluer ces métriques, on doit séparer le jeu de données à notre disposition en 2 ou même 3 parties.

La première partie est le **jeu d'entraînement** et permet d'entraîner le modèle. La seconde partie contient le **jeu de données de validation** et permet d'évaluer la bonne combinaison d'*hyper-paramètres* du modèle (voir partie 1.4.3.9). Enfin, la dernière partie est appelée **jeu de données de test** sera utilisée une fois notre modèle entraîné et construit, afin d'évaluer sa performance sur un jeu de données indépendant.

On peut calculer toutes ces métriques sur le jeu de données de validation ou de test à partir de la *matrice de confusion* (voir Table 1.6).

Chaque ligne de la matrice de confusion représente une classe réelle, et chaque colonne une classe prédite. La première ligne de cette matrice correspond à la classe positive (versicolor), 28 observations sont

## 1.4. PRÉDICTION DE DONNÉES GÉNOMIQUES

---

correctement classées, ce sont les *vrais positifs* (*TP*), et 2 sont mal classées, ce sont les *faux positifs* (*FP*). La deuxième ligne correspond à la classe négative (*virginica*), 17 observations sont correctement classés, ce sont les *vrais négatifs* (*TN*), et 3 observations sont classés à tort parmi la classe négative, ce sont des *faux négatifs* (*FN*).

À partir de cette table, on peut récupérer les différentes métriques souhaitées : l'*exactitude* (1.47), la *précision* (1.48) ou le *rappel* (1.49) :

$$\text{exactitude} = \frac{TP + TN}{P + N} = \frac{28 + 17}{50} = 0.9 \quad (1.47)$$

$$\text{precision} = \frac{TP}{TP + FP} = \frac{28}{28 + 2} = 0.9333 \quad (1.48)$$

$$\text{rappel} = \frac{TP}{TP + FN} = \frac{28}{28 + 3} = 0.9032 \quad (1.49)$$

l'*exactitude* (*accuracy*) nous donne la proportion d'observations qui sont correctement classés. C'est une métrique qui peut être très mal interprétée lorsque l'on travaille avec des classes déséquilibrées : par exemple, si 95 observations sont de la classe *versicolor*, et 5 *virginica*, un prédicteur classant toutes les observations comme *versicolor* donnerait une exactitude de 95%. Pour cela, on préfère utiliser la paire *précision/rappel*. La précision nous donne le pourcentage de prédictions positives qui sont réellement positives, le rappel nous donne le pourcentage d'observation positives correctement détectés par le classifieur.

Une bonne façon de visualiser les performances d'un modèle est de faire une courbe *ROC*, pour *Receiver Operating Characteristic*. Cette courbe croise le taux de *vrais positifs* (rappel) avec le taux de *faux positifs*, en faisant varier le seuil de décision (voir Figure 1.41 A). Ici, on voit que quand le taux de vrais positifs (rappel) augmente, le taux de faux positifs augmente aussi. Plus le modèle sera bon, plus le niveau de rappel sera important pour un taux de faux positifs faible, peu importe le seuil, ce qui revient à obtenir l'aire sous la courbe la plus importante possible. Ici, le modèle est très bon, car l'AUC est très importante (0.9576).

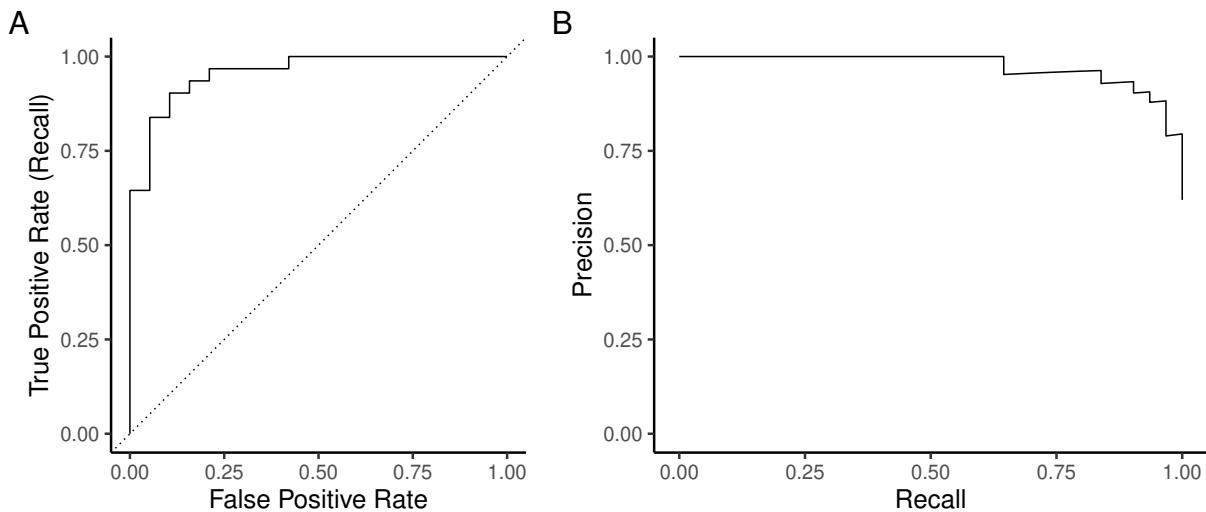


FIGURE 1.41 – ROC curve et Precision/Recall Curve.

On peut également visualiser la courbe Précision/Rappel (voir Figure 1.41 B), qui nous donne une information différente, très utile dans le cas où la classe positive est beaucoup plus rare que la classe négative. Par exemple, si on cherche à prédire des sites de liaison potentiels d'une protéine sur le génome, le nombre de régions génomiques correspondant à un site de liaison de la protéine sera bien plus faible que le nombre de régions n'ayant pas de site de liaison pour cette protéine. Dans ce cas-là, on peut utiliser la courbe Precision/Rappel. Dans tous les cas, l'aire sous la courbe fournit une bonne métrique pour évaluer la performance du modèle (Ici 0.974).

### 1.4.3 Les réseaux de neurones artificiels (*Neural Network / Deep Learning*)

L'*Apprentissage profond* ou *Deep Learning* est un ensemble de modèles faisant partie des réseaux de neurones artificiels, ou *Neural Network*. Ces méthodes permettent de réaliser des tâches complexes, comme la reconnaissance faciale, la traduction, ou encore la conduite automatique. C'est en 2012 pendant un concours d'ImageNet<sup>15</sup>, sur la reconnaissance d'images que le *Deep learning* s'est imposé comme outil de référence pour la résolution de tâches complexes (KRIZHEVSKY, SUTSKEVER et G. E. HINTON 2012). Depuis, ces méthodes n'ont cessé d'apparaître dans les sciences appliquées, et sont aujourd'hui très utilisés en Bio-informatique, pour différentes tâches, comme la prédiction de sites de liaisons de protéines (ALIPANAH et al. 2015) ou l'accessibilité de la chromatine (KELLEY, SNOEK et RINN 2016) à partir de l'ADN, jusqu'à la prédiction de structure 3D des protéines (JUMPER et al. 2021), ou bien l'amélioration du signal de séquençage (LAL et al. 2021) (Yan ZHANG et al. 2018).

Le rôle d'un réseau de neurones est d'approximer une ou plusieurs fonctions  $f$ , afin de transformer  $x$  en  $y$  (1.50) :

$$\hat{y}_i = f^L(W^L f^{L-1}(W^{L-1} \dots f^1(W^1 x_i))) \quad (1.50)$$

Ici, le réseau transforme progressivement  $x_i$  en  $\hat{y}_i$  par un enchaînement de fonctions  $f$ , et apprend les valeurs des paramètres  $W$  qui proposent la meilleure approximation. C'est cet enchaînement de nombreuses fonctions qui lui donne le nom *deep neural network* (voir Figure 1.42).

Chaque fonction est appelée **couche**, et le nombre de couches définit la profondeur du modèle. Seule la première et la dernière couche du modèle sont visibles, et son entraînement se base sur cette dernière pour déterminer sa précision. La perte (voir partie 1.4.1.1) est calculée sur cette couche de sortie. À cause du grand nombre de paramètres à apprendre, les modèles de *Deep Learning* nécessitent souvent un grand nombre de données d'observations pour réaliser une bonne prédiction.

L'avantage d'un tel modèle, et qu'il apprend lui-même les caractéristiques (ou *features*, appelées variables en statistiques) de  $x$  pour produire  $y$ . Le *Deep Learning* permet donc de se passer d'une étape de construction des caractéristiques, à l'inverse du *Machine Learning*. La construction des caractéristiques est une étape très compliquée dans le *Machine Learning*, on parle alors de *feature engineering*, et nécessite une connaissance approfondie des données.

---

15. <https://image-net.org/>

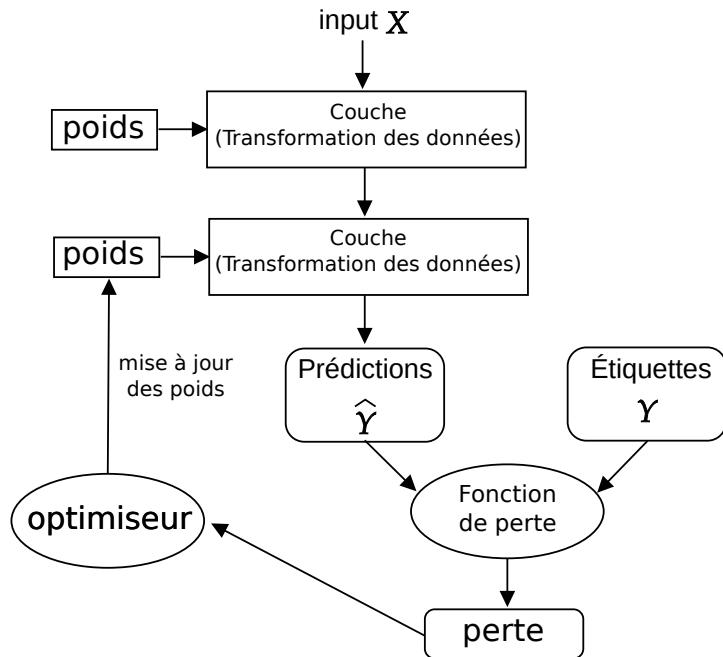


FIGURE 1.42 – Représentation graphique d'un modèle de *Deep Learning* à deux couches, avec les différents algorithmes permettant d'optimiser les poids du modèle.

#### 1.4.3.1 L'unité de base d'un réseau de neurones

Les réseaux de neurones sont basés sur une méthode d'apprentissage supervisée, le *perceptron* (ROSENBLATT 1958), lui-même inspiré du *neurone formel* (MCCULLOCH et PITTS 1943). Il est défini par l'équation suivante (1.51) :

$$\hat{y} = g \left( \sum_{i=1}^m X^T W \right) \quad (1.51)$$

Un perceptron réalise donc la somme des entrées  $X$ , pondérée par  $W$ , puis passe par une fonction d'activation  $g$ , pour produire une sortie  $y$  (voir Figure 1.43). À l'origine, la fonction d'activation produisait une sortie binaire, 1 si la somme pondérée était supérieure à 0, 0 sinon, avec pour but d'imiter l'activation d'un véritable neurone biologique. Dans les réseaux modernes, cette activation a été remplacée par d'autres fonctions d'activations, comme *ReLU*, pour *Rectified Linear Unit*.

#### 1.4.3.2 La fonction d'activation

Si on construit un modèle de *Deep Learning* comme un empilement de fonctions  $\hat{y}_i = f(Wx_i)$ , où  $x$  correspond à la sortie de la couche précédente et  $y$  la sortie de la couche actuelle, alors les transformations apprises par le modèle seront toujours issues d'une combinaison linéaire des  $x$  pour obtenir  $y$ . La présence d'une fonction d'activation est donc essentielle afin de permettre au modèle d'apprendre à transformer non-linéairement les données, pour qu'à la fin, les données soient linéairement séparables (voir Figure 1.44).

En pratique, c'est souvent *ReLU* qui est utilisé (1.52) :

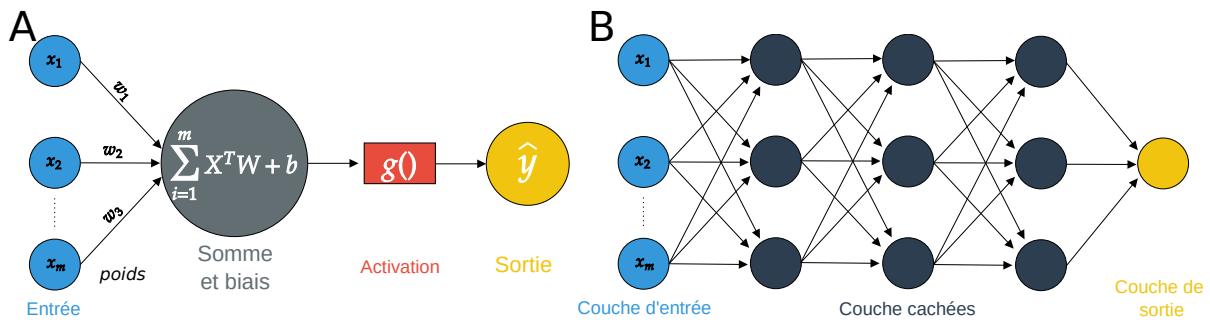


FIGURE 1.43 – L’unité de base d’un réseau de neurones, le perceptron **A**, et son implementation **deep**, le perceptron multicouche **B**.

$$ReLU(x) = \max(0, x) \quad (1.52)$$

ReLU transforme les données de façon à supprimer les valeurs négatives. Elle est très souvent utilisée, car elle permet d’augmenter la vitesse d’entraînement par son gradient simple à calculer (1.53) :

$$\frac{\partial ReLU(x)}{\partial x} = \begin{cases} 0 & \text{if } x < 0 \\ 1 & \text{if } x > 0 \end{cases} \quad (1.53)$$

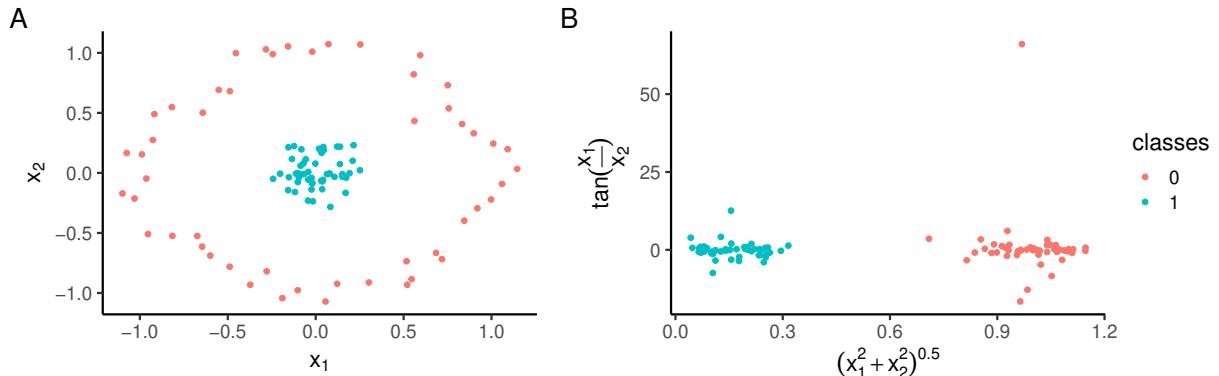


FIGURE 1.44 – Données non linéairement séparables (**A**) et transformation des caractéristiques dans un espace linéairement séparable (**B**).

#### 1.4.3.3 Entrainement du modèle de Deep Learning

Un modèle de *Deep Learning* s’entraîne par un algorithme de descente de gradient (voir partie 1.4.1.1.2). L’algorithme qui permet de calculer le gradient pour chaque neurone, de la dernière couche jusqu’à la première s’appelle la rétropropagation du gradient.

Pour utiliser la rétropropagation du gradient sur chacun des couches/neurones du modèle afin de mettre à jour les poids  $W$ , on utilise un algorithme appelé **optimiseur**. Son but est d’améliorer le calcul de gradient classique en proposant une optimisation par *inertie*. L’algorithme classique faisant de petits pas réguliers dans la bonne direction, l’idée est donc d’ajouter un élément qui lui permet d’accélérer. Pour

cela, elle utilise les gradients **antérieurs** en les gardant en mémoire dans un *vecteur d'inertie*, ce qui permet à l'algorithme d'accélérer le pas, et donc de converger plus vite (1.54) :

$$\mathbf{m} \leftarrow \beta \mathbf{m} - \eta C_{\Theta}(y_i, \hat{y}_i) \quad (1.54)$$

$$\Theta = \Theta + \mathbf{m} \quad (1.55)$$

le vecteur d'inertie **m** est mis à jour avec le gradient local et l'hyper-paramètre  $\beta$  contrôle cette inertie en simulant un frottement plus ou moins élevé (compris entre 0 et 1). La mise à jours des poids est représenté par l'équation (1.55), et prend maintenant en compte les gradients précédents grâce au vecteur d'inertie.

Il existe différentes variantes de l'inertie, qui améliorent les performances, la vitesse et la convergence, par rapport à la méthode classique. Parmi ceux utilisés fréquemment, on compte *AdaGrad* (DUCHI, HAZAN et SINGER 2011), *RMSprop* (G. HINTON s. d.) et *Adam* (KINGMA et BA 2017).

#### 1.4.3.4 Encodage des caractéristiques d'entrée du modèle de *Deep Learning*

A

AGGCGTTTCAACCGCTCCGC

B

A	1								1	1									
C			1					1			1	1		1		1	1	1	1
G		1	1		1								1					1	
T					1	1	1								1				

FIGURE 1.45 – **One Hot Encoding** d'une séquence d'ADN. De manière générale, l'encodage en *one-hot* permet de transformer des chaînes de caractères en valeurs numériques, et est très souvent utilisé pour résoudre des problèmes sémantiques en *Deep Learning*.

Un modèle de *Deep Learning* transforme progressivement  $X$  en  $Y$  (1.50). Afin d'utiliser  $X$  en entrée, celui-ci doit être formaté de manière à être compris par le modèle. Celui-ci étant composé d'un enchaînement de fonctions mathématiques, la seule façon d'encoder une donnée d'entrée est de la transformer en valeur **numérique**. Pour l'ADN par exemple, un moyen simple de l'encoder numériquement est par *one-hot encoding* (voir Figure 1.45). Chaque caractère devient un vecteur à  $i$  bits dont un seul prend la valeur 1. La lettre  $i$  correspond à la taille de l'alphabet ( $i = 4$  pour l'ADN) et l'indice du bit à 1 correspond à l'indice du caractère dans l'alphabet.

De manière générale, on encode toujours les données d'entrée sous forme de **Tenseurs**. Les **Tenseurs** sont des généralisations des **vecteurs** et **matrices** à  $n$  dimensions (voir Table 1.7). Par convention, la première dimension correspond aux observations, et les autres dimensions contiennent leurs caractéristiques. Par exemple, 100 séquences d'ADN de longueur 20 encodés en *one-hot* seront contenues dans

TABLE 1.7 – Exemples de tenseurs encodant différents types de données.

Dimension	Data	Description
0D		
1D	Vector data	2D tensors of shape (samples, features)
2D	Timeseries data	3D tensors of shape (samples, timesteps, features)
3D	Images	Images 4D tensors of shape (samples, height, width, channels)
4D	Videos	Video 5D tensors of shape (samples, frames, height, width, channels)

un tenseur à  $(100, 20, 4)$ . Pour les images, on peut les encoder dans des tenseurs de dimensions 4 : (*observations, hauteur, largeur, canaux*), où *canaux* correspond aux valeurs *RGB*.

#### 1.4.3.5 Combinaisons multiples des variables par couche *dense*

Les couches *dense* ou *dense layer* contiennent un ou plusieurs perceptrons. Cette couche est *intégralement connectée*, car tous les neurones sont connectés avec les neurones de la couche précédente (voir Figure 1.43). On peut représenter cette couche *dense* par l'équation suivante (1.56) :

$$h_{\mathbf{W}, \mathbf{b}}(\mathbf{X}) = g(\mathbf{X}^T \mathbf{W} + \mathbf{b}) \quad (1.56)$$

Où  $\mathbf{X}$  représente la matrice des caractéristiques d'entrée,  $g()$  la fonction d'activation (voir partie 1.4.3.2), et  $\mathbf{W}$  correspond à la matrice des poids, où chaque ligne correspond à une entrée et chaque colonne à un neurone de la couche. Les termes constants sont encodés dans le vecteur  $\mathbf{b}$ .

On peut encoder la couche *dense* pour différentes tâches : on peut simuler des polynômes de degrés supérieurs en enchaînant des couches *denses* de type  $y = x^2 + x$ , ou d'autres fonctions mathématiques plus complexes. En *Machine Learning*, on peut simuler des régressions, linéaires ou logistiques, créer un classifieur non-linéaire, ou bien réduire la dimension de la couche d'entrée (même si on préférera utiliser une couche de plongement ou *embedding*).

En pratique, le nombre de poids d'un réseau de neurones entièrement constitué de couches *denses* est extrêmement élevé, ce qui fait qu'ils sont souvent utilisés dans les dernières couches du modèle ou bien sont délaissées au profit d'autres couches, plus efficaces pour leurs tâches respectives. Parmi les couches alternatives, on peut utiliser les couches convolutions des *Convolutional neural network* pour la reconnaissance d'objets (voir partie 1.4.3.6), ou les couches récurrentes des *Recurrent neural network* pour l'analyse de séries temporelles.

#### 1.4.3.6 La couche de convolution

La couche de convolution est l'élément principal d'un réseau de neurones convolutif, ou *CNN* (*Convolutional Neural Networks*). La principale différence avec la couche *dense* est que les neurones ne sont pas connectés à toutes les caractéristiques d'entrée, mais uniquement à une sous-partie, celles dans leur champ récepteur. Les *CNN* sont très utilisés pour la classification d'images ou détection d'objets. Ils peuvent reconnaître des caractéristiques n'importe où dans une image, et en pratique, on enchaîne très

## 1.4. PRÉDICTION DE DONNÉES GÉNOMIQUES

---

souvent les couches de convolution afin de détecter des caractéristiques de plus en plus haut niveau de l'image (voir Figure 1.46).

Les poids des neurones de convolution sont encodés dans une matrice de la taille du champ récepteur, qu'on peut visualiser sous la forme d'une petite image, dans le cas de la classification d'images (on parle de CNN 2D) (voir Figure 1.46), ou d'un motif dans le cas où on traite des séquences (CNN 1D) (voir Figure 1.47).

Ces poids sont appelés *filtres* ou encore *noyaux de convolutions* (ou *kernels*). L'opération de convolution est définie par l'utilisation de ces *kernels* par la couche de convolution  $h$  sur la séquence d'entrée en glissant d'un pas  $s$  (hyper-paramètre, défini à 1 par défaut), pour produire une carte d'activation ou *response map*  $a$ . L'opération de convolution 1D pour un neurone en position  $i$  utilisant un champ récepteur de taille  $k$  correspond à l'équation suivante (1.57) :

$$h_a = \sum_{j=0}^{k-1} x_{i'} \cdot w_j \quad (1.57)$$

avec  $i' = i \times s + k$ , ce qui correspond à extraire la sous-séquence à une position  $i \times s$  jusqu'à la position  $i \times s + k - 1$ . L'avantage du *CNN* est que tous les neurones d'une *response map* partagent les mêmes poids, le nombre de paramètres du modèle est donc réduit par rapport à une couche *dense*. De plus, cela permet au modèle de reconnaître un motif n'importe où sur la séquence, à l'inverse de la couche *dense* qui ne peut le reconnaître qu'au même endroit.

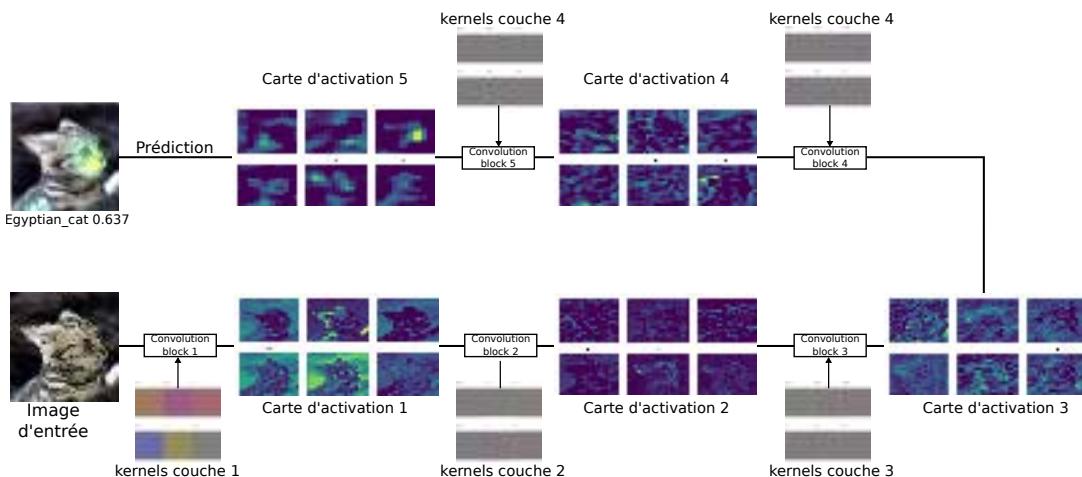


FIGURE 1.46 – Classification d'une image par réseau de neurones convolutif (VGG19). Chaque block de convolution transforme l'image et produit une carte d'activation par kernel dont la résolution diminue au fur et à mesure. Les kernels peuvent être visualisés sous la forme d'une image maximisant sa réponse, par remontée des gradients. L'image de sortie indique l'importance des pixels pour la prédiction.

Les *hyper-paramètres* (voir partie 1.4.3.9) d'une couche de convolution permettent de définir le nombre de *kernels* utilisés, le *pas* effectué par le champ récepteur, si on veut ajouter un *padding* afin de conserver les mêmes dimensions en sortie qu'en entrée.

On peut également, via la remontée des gradients (*gradient ascent*), algorithme qui permet de **maximiser** une fonction, créer une image d'entrée qui donne la meilleure *response map* pour un *kernel* donné. Cette

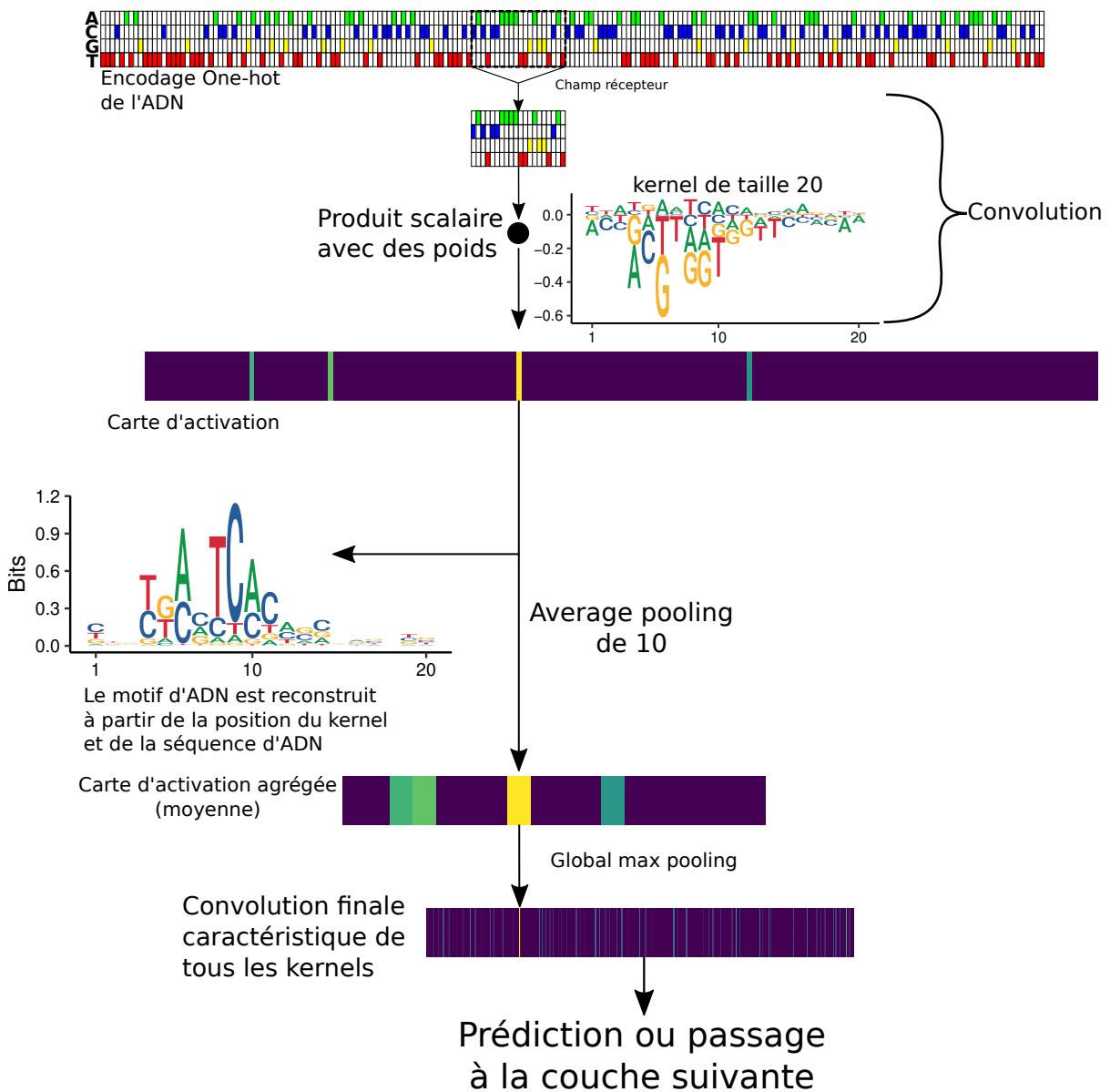


FIGURE 1.47 – Opération de convolution 1D implementée en Deep Learning. l'ADN encodé en one-hot est utilisé comme données d'entrée, par produit scalaire avec un kernel, produit une carte d'activation (response map) indiquant son activation dans la séquence. La carte d'activation passe dans une couche d'*average pooling* qui réduit sa dimension et moyenne le signal sur 10 positions. La valeur maximale pour chaque kernel est ensuite récupérée, et est utilisée comme vecteur sortie.

fois-ci on ne cherche donc pas à trouver les meilleurs paramètres, mais bien l'image qui fait le plus réagir ces paramètres. De même, on peut utiliser un algorithme similaire pour produire une carte de chaleur d'activation de classe, qui permet d'identifier sur l'image l'importance des pixels à la prédiction, pour une classe précise (SELVARAJU et al. 2019) (voir Figure 1.46).

**1.4.3.6.1 Sous échantillonnage par *pooling*** La couche de convolution est rarement seule dans un *CNN*. Elle est très souvent associée à une couche de *pooling*, qui permet de réduire la dimension par sous-échantillonnage afin de réduire les opérations. Le *pooling* ajoute également de l'invariance au modèle envers les petites translations, si une image est légèrement décalée, on obtient toujours la même sortie. On peut utiliser différentes opérations pour réaliser le *pooling*, comme la moyenne ou bien le max d'un certain nombre de positions (hyper-paramètre) de la *response map* (voir Figure 1.47).

La toute dernière couche de convolution est très souvent suivie d'une couche de *global max pooling*, qui récupère l'activation maximale de chaque *response map*, pour produire un vecteur réponse d'une taille correspondant au nombre de *kernels*, qui pourra être utilisée en entrée d'une couche *dense*.

#### 1.4.3.7 Couche de sortie d'un modèle de *Deep Learning*

La couche de sortie d'un modèle de *Deep Learning* correspond à la prédiction attendue du modèle  $\hat{Y}$ . Elle doit donc correspondre aux caractéristiques de  $Y$ . En pratique, pour effectuer une tâche de régression ou de classification binaire, on définit une couche *dense* à un seul neurone dont la sortie correspond à la somme pondérée des vecteurs d'entrée, puis transformée par une fonction d'activation. Pour une tâche de régression, on utilise très souvent une fonction d'activation linéaire (1.58), pour une tâche de classification binaire, la fonction *sigmoïde* est utilisée (1.59).

Dans le cas d'une classification à  $K > 2$  classes, on définit une couche *dense* à  $K$  neurones, et on utilise la fonction *softmax* (1.60). La couche retourne la valeur  $i \in \{1, \dots, K\}$  qui correspond à l'indice de la valeur maximale de  $g_{softmax}(\mathbf{x})$ , pour le vecteur  $\mathbf{x} = (x_1, \dots, x_K)$  (1.61). On dit que cette couche *softmax* est **multi-classes** mais non **multi-sortie**, et ne prédit donc que la classe ayant la probabilité la plus forte.

$$g_{linear}(x) = x \quad (1.58)$$

$$g_{sigmoid}(x) = \frac{1}{1 + e^{-x}} \quad (1.59)$$

$$g_{softmax}(x) = \frac{e^x}{\sum_{k=1}^K e^x} \quad (1.60)$$

$$\hat{y} = argmax(g_{softmax}(\mathbf{x})) \quad (1.61)$$

#### 1.4.3.8 Régularisation du modèle

Les modèles de *Deep Learning* possèdent en général des milliers de paramètres, ils sont donc souvent sujets au sur-ajustement. Pour éviter cela, on peut régulariser notre modèle, en utilisant des méthodes

## 1.4. PRÉDICTION DE DONNÉES GÉNOMIQUES

---

classiques comme la norme  $l_1$  (*lasso*) ou  $l_2$  (*ridge*), ou alors utiliser un arrêt prématué, en stoppant l'apprentissage du modèle quand on constate qu'il sur-apprend sur le jeu de données d'entraînement.

Une méthode originale au *Deep Learning* est le *dropout* (SRIVASTAVA et al. 2014). L'idée du *dropout* est simple : à chaque tour d'entraînement ou *epoch*, chaque neurone a une probabilité  $p$  d'être temporairement éteint, il est donc ignoré pendant cette étape, mais pourra être à nouveau actif à l'étape suivante.  $p$  est un hyper-paramètre, le taux d'extinction, ou *dropout rate*. Après l'entraînement, les neurones ne sont plus jamais éteints, ce qui veut dire qu'un neurone est connecté avec 100% des neurones ayant subit le *dropout*, il faut donc multiplier les poids des connexions d'entrée par  $1 - p$ . Dans le cas contraire, les neurones recevront 50% (si  $p = 0.5$ ) de signal en plus que pendant l'entraînement, et risquent de mal se comporter.

Une couche de *dropout* peut s'intégrer à n'importe quelle couche, sauf la couche de sortie (on peut même appliquer le *dropout* sur la couche d'entrée), ce qui fait qu'on entraîne un modèle légèrement différent à chaque *epoch*, mais qui partage tout de même les mêmes paramètres.

### 1.4.3.9 Optimisation des hyper-paramètres du modèle de *Deep Learning*

Les *hyper-paramètres* des modèles de *Deep Learning* sont des éléments à considérer pendant leurs constructions. Ils permettent de contrôler de nombreux aspects du comportement d'un modèle, comme le temps d'exécution, la mémoire ou même la régularisation. Parmi les *hyper-paramètres* déjà listés dans les parties précédentes, on compte le nombre de neurones d'une couche *dense*, le nombre de *kernels* leur *taille* et le *pas* du champ récepteur d'une couche de convolution, et le *drop rate* d'une couche de *dropout*. Ces éléments ne sont pas entraînables, et doivent être sélectionnés avant d'entraîner le modèle. Cependant, ils jouent un rôle majeur sur la qualité d'un modèle, et choisir un paramètre manuellement sous-entend d'avoir une compréhension exacte de sa fonction.

On peut donc utiliser des méthodes automatiques pour sélectionner les bonnes valeurs d'*hyper-paramètres* pour un modèle. Une pratique simple, lorsqu'on dispose de peu d'*hyper-paramètres* est d'utiliser une recherche par grille. Pour chaque *hyper-paramètre*, on sélectionne un petit nombre fini de valeurs possibles, puis l'algorithme de recherche par grille entraîne un modèle pour chaque combinaison de ces valeurs, et calcule une métrique évaluant la performance du modèle, par exemple la précision sur le jeu de données de validation (voir partie 1.4.2. On constate rapidement qu'avec un grand nombre de combinaisons possibles, le temps de calcul devient vite énorme. On peut alors réduire ce nombre de combinaisons en sélectionnant aléatoirement des combinaisons dans la grille, quitte à lancer une deuxième fois l'algorithme une fois qu'on a une meilleure idée des valeurs "correctes" pour chaque *hyper-paramètre*.

À l'inverse de ces méthodes qui trouvent un ensemble de valeurs pour nos *hyper-paramètres* maximisant la performance en testant toutes les combinaisons, ou par hasard, on peut utiliser une méthode d'optimisation bayésienne. L'idée de cette méthode est de converger vers un ensemble d'*hyper-paramètres* maximisant la performance du modèle en un minimum d'itérations. On construit un ensemble d'observations (qu'on détermine manuellement ou par hasard) pour inférer la probabilité des valeurs qu'on a pas encore observées. Pour cela, on utilise un *processus gaussien* généralisant le concept de la loi Normale aux fonctions, et permet de générer une estimation de la variation de la performance du modèle selon comment on modifie les *hyper-paramètres*.



---

# Résultats

## 2.1 DeepG4 : une approche *Deep Learning* pour prédire les régions riches en G-quadruplex actifs spécifiques à un type cellulaire

Les G-Quadruplex (G4) sont des structures secondaires d'ADN ou d'ARN différentes de la structure en B-DNA. Les G4 sont riches en Guanine, et peuvent former une structure à quatre brins qui s'enchaînent pour former des tétrades, ou *G-tetrads*. Ces structures sont connues pour être enrichies sur les télomères (DUQUETTE et al. 2004), mais également au niveau des promoteurs de gènes actifs (SIDDQUI-JAIN et al. 2002 ; HUPPERT et BALASUBRAMANIAN 2007 ; LAGO et al. 2021). Les méthodes *in-silico* pour détecter et étudier les G4 sont basées principalement sur la détection d'un motif spécifique, mais des méthodes récentes ont été développées pour identifier les G4 à l'échelle du génome en utilisant une approche de séquençage à haut débit, comme G4-seq (*in vitro*, CHAMBERS et al. (2015)) et G4 ChIP-seq (*in vivo*, HÄNSEL-HERTSCH, SPIEGEL et al. (2018)).

L'utilisation de données omiques a permis de mettre en évidence les rôles actifs des G4 dans la structure de la chromatine et sa conformation 3D, dans la stabilité et l'endommagement de l'ADN (J. ROBINSON et al. 2021), mais aussi dans la régulation de l'expression des gènes (LAGO et al. 2021).

Dans cette étude, nous proposons un modèle de *Deep Learning* basé sur la séquence et l'accessibilité de l'ADN G4 *in-vivo*. Le *Deep Learning* est un ensemble d'approches récentes et populaires où le modèle apprend des caractéristiques (ou *features*) à partir des données, permettant d'identifier de nouveaux motifs prédicteurs, y compris des facteurs de transcription connus qui pourraient réguler directement ou indirectement l'activité des G4. Nous avons également cartographié des milliers de régions G4 actives qui peuvent être utilisées pour identifier des cibles potentielles de thérapies anti-cancer récentes à base de ligands de G4 (KOSIOL et al. 2021).

Ce modèle a été programmé avec *keras R*<sup>1</sup> et peut être utilisé sous la forme d'un *package R* que j'ai développé : *DeepG4*<sup>2</sup>. *DeepG4* peut être utilisé sur n'importe quelle séquence d'ADN pour prédire la formation de G4, avec ou sans données d'accessibilité correspondante. Il permet également de récupérer les motifs prédicteurs du modèle, et peut s'appliquer à des séquences de grande taille. Enfin, *DeepG4*

---

1. <https://keras.rstudio.com/>

2. <https://github.com/morphos30/DeepG4>

## 2.1. DEEPG4

permet d'utiliser l'architecture du modèle pour entraîner un nouveau jeu de données de G4 actifs, par exemple, dans le cas où de nouvelles données expérimentales plus précises seront disponibles.

## RESEARCH ARTICLE

# DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions

Vincent Rocher  <sup>1</sup>, Matthieu Genais  <sup>2</sup>, Elissar Nassereddine <sup>1</sup>, Raphael Mourad  <sup>1\*</sup>

**1** Molecular, Cellular and Developmental biology department (MCD), Centre de Biologie Intégrative (CBI), University of Toulouse, CNRS, UPS, Toulouse, France, **2** Centre de Recherches en Cancérologie de Toulouse (CRCT), INSERM U1037, Toulouse, France

\* [raphael.mourad@univ-tlse3.fr](mailto:raphael.mourad@univ-tlse3.fr)



## OPEN ACCESS

**Citation:** Rocher V, Genais M, Nassereddine E, Mourad R (2021) DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions. PLoS Comput Biol 17(8): e1009308. <https://doi.org/10.1371/journal.pcbi.1009308>

**Editor:** Tamar Schlick, New York University, UNITED STATES

**Received:** May 26, 2021

**Accepted:** July 26, 2021

**Published:** August 12, 2021

**Copyright:** © 2021 Rocher et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

**Data Availability Statement:** All data and code were deposited on a Github repository. We downloaded G4 ChIP-seq data for HaCaT, K562 and HEK293 cell lines from Gene Expression Omnibus (GEO) accession numbers GSE76688, GSE99205 and GSE107690. We downloaded G4P ChIP-seq peaks already mapped to hg19 for A549, H1975, 293T and HeLa-S3 cell lines from GEO accession number GSE133379. We downloaded processed G4-seq peaks mapped to hg19 from GEO accession number GSE63874. We downloaded processed DNase-seq bigwig files for

## Abstract

DNA is a complex molecule carrying the instructions an organism needs to develop, live and reproduce. In 1953, Watson and Crick discovered that DNA is composed of two chains forming a double-helix. Later on, other structures of DNA were discovered and shown to play important roles in the cell, in particular G-quadruplex (G4). Following genome sequencing, several bioinformatic algorithms were developed to map G4s *in vitro* based on a canonical sequence motif, G-richness and G-skewness or alternatively sequence features including k-mers, and more recently machine/deep learning. Recently, new sequencing techniques were developed to map G4s *in vitro* (G4-seq) and G4s *in vivo* (G4 ChIP-seq) at few hundred base resolution. Here, we propose a novel convolutional neural network (DeepG4) to map cell-type specific active G4 regions (e.g. regions within which G4s form both *in vitro* and *in vivo*). DeepG4 is very accurate to predict active G4 regions in different cell types. Moreover, DeepG4 identifies key DNA motifs that are predictive of G4 region activity. We found that such motifs do not follow a very flexible sequence pattern as current algorithms seek for. Instead, active G4 regions are determined by numerous specific motifs. Moreover, among those motifs, we identified known transcription factors (TFs) which could play important roles in G4 activity by contributing either directly to G4 structures themselves or indirectly by participating in G4 formation in the vicinity. In addition, we used DeepG4 to predict active G4 regions in a large number of tissues and cancers, thereby providing a comprehensive resource for researchers.

Availability: <https://github.com/morphos30/DeepG4>.

## Author summary

DNA is a molecule carrying genetic information and found in all living cells. In 1953, Watson and Crick found that DNA has a double helix structure. However, other DNA structures were later identified, and most notably, G-quadruplex (G4). In 2000, the Human Genome Project revealed the widespread presence of G4s in the genome using algorithms. To date, all G4 mapping algorithms were developed to map G4s on naked DNA, without knowing if they could be formed in a given cell type. Here, we designed a

different cell lines from ENCODE (<http://hgdownload.soe.ucsc.edu/goldenPath/hg19/encodeDCC/>), and processed ATAC-seq bigwig files for HaCaT cell line from GSE7668. We downloaded processed ATAC-seq bigwig files from ICGC cancer cohorts from <https://gdc.cancer.gov/about-data/publications/ATACseq-AWG>. We downloaded ChromHMM annotations for ENCODE cell lines from <http://hgdownload.cse.ucsc.edu/goldenpath/hg19/encodeDCC/wgEncodeBroadHmm/>. We downloaded breast cancer processed mutation data from ICGC BRCA-US cohort from the portal <https://dcc.icgc.org>. We downloaded position weight matrices for transcription factor binding sites from the JASPAR 2018 database (<http://jaspar.genereg.net>). DeepG4 is available at <https://github.com/morphos30/DeepG4>. All fasta files used for training and predictions were also deposited. Performance analyses of DeepG4 and DeepG4\* presented in this article can be obtained using a pipeline and a docker available at <https://github.com/morphos30/DeepG4ToolsComparison>.

**Funding:** The author(s) received no specific funding for this work.

**Competing interests:** The authors have declared that no competing interests exist.

novel artificial intelligence algorithm that could map G4 regions active in the cell from the DNA sequence and chromatin accessibility. Moreover, we identified key transcriptional factor motifs that could explain G4 activity depending on cell type. Lastly, we used our new algorithm to map active G4 regions in multiple tissues and cancers as a comprehensive resource for the G4 community.

## Introduction

Deoxyribonucleic acid (DNA) is a complex molecule carrying genetic instructions for the development, functioning, growth and reproduction of all known living beings and numerous viruses. In 1953, Watson and Crick discovered that DNA is composed of two chains forming a double-helix [1]. However, other structures of DNA were discovered later and shown to play important roles in the cell. Among those structures, G-quadruplex (G4) was discovered in the late 80's [2]. G4 sequence contains four continuous stretches of guanines [3]. Four guanines can be held together by Hoogsteen hydrogen bonding to form a square planar structure called a guanine tetrad (G-quartets). Two or more G-quartets can stack to form a G4 [3]. The quadruplex structure is further stabilized by the presence of a cation, especially potassium, which sits in a central channel between each pair of tetrads [4]. G4 can be formed of DNA [5] or RNA [6].

G4s were found enriched in gene promoters, DNA replication origins and telomeric sequences [5, 7]. Accordingly, numerous works suggest that G4 structures can regulate several essential processes in the cell, such as gene transcription, DNA replication, DNA repair, telomere stability and V(D)J recombination [5]. For instance, in mammals, telomeric DNA consists of TTAGGG repeats [8]. They can form G4 structures that inhibit telomerase activity responsible for maintaining length of telomeres and are associated with most cancers [9, 10]. G4s can also regulate gene expression such as for MYC oncogene where inhibition of the activity of NM23-H2 molecules, that bind to the G4, silences gene expression [11]. Moreover, G4s are also fragile sites and prone to DNA double-strand breaks [12]. Accordingly, G4s are highly suspected to be implicated in human diseases such as cancer or neurological/psychiatric disorders [13–15].

Following the Human Genome project [16], computational algorithms were developed to predict the location of G4 sequence motifs in the human genome [17, 18]. First algorithms consisted in finding all occurrences of the canonical motif  $G_{3+} N_{1-7} G_{3+} N_{1-7} G_{3+} N_{1-7} G_{3+}$ , or the corresponding C-rich motif (quadparser algorithm) [19, 20]. Using this canonical motif, over 370 thousand G4s were found in the human genome. Nonetheless, such pattern matching algorithms lacked flexibility to accomodate for possible divergences from the canonical pattern. To tackle this issue, novel score-based approaches were developed to compute G4 propensity score by quantifying G-richness and G-skewness (G4Hunter algorithm) [21], or by summing the binding affinities of smaller regions within the G4 and penalizing with the destabilizing effect of loops (pqsfinder algorithm) [22]. Recently, new sequencing techniques were developed to map G4s in vitro (G4-seq) [23], and G4s in vivo (G4 ChIP-seq) [24] as regions of few hundred bases. Machine and deep learning methods were proposed to predict such G4 regions, *i.e.* regions comprising the G4(s) along with flanking sequences. For instance, Quadron—a machine learning approach—was proposed to predict G4s based on sequence features (such as k-mer occurrences) from a region of more than 100 bases, and trained using in vitro G4 regions with G4-seq [25]. By combining with regular expressions, Quadron could predict if a region was found in vitro, but also the exact location and stability value of G4(s) within the

region. Other deep learning approaches had lower resolution for mapping G4s (around 200 bases), but they showed higher prediction performance. PENGUINN, a deep convolutional neural network (CNN), was trained to predict G4 regions in vitro [26]. Another CNN, G4detector, was also designed to predict G4 regions forming in vitro [27]. Thus, all current approaches aimed to predict G4 regions forming in vitro, but were not designed to assess the ability of G4 sequences to form in vivo (*e.g.* G4 activity).

Here, we propose a novel method, named DeepG4, aimed to predict cell-type specific active G4 regions (regions that were mapped both in vitro and in vivo in a given cell type) from DNA sequence and chromatin accessibility. DeepG4 implements a CNN which is trained using a combination of genome-wide in vitro (G4-seq) and in vivo (G4 ChIP-seq) peak DNA sequences, together with chromatin accessibility measures (*e.g.* ATAC-seq). For this purpose, DeepG4 exploits the genomic context (a 201-base region) of a G4, which comprises the potential G4 forming sequence, but also other DNA motifs that may play a role in G4 activity. Moreover, adding chromatin accessibility, which is publicly available for most cell lines, tissues and cancers, into the model allows to predict G4 regions that are active depending on the cell-type, since it was previously shown that in vivo G4 peaks strongly colocalize (98%) with regions identified by either FAIRE-seq or ATAC-seq, or both [28]. DeepG4 achieves excellent accuracy at predicting cell-type specific active G4 regions (area under the receiver operating characteristic curve or AUROC > 0.98). Moreover, DeepG4 identifies key DNA motifs that are predictive of active G4 regions. Among those motifs, we found specific motifs resembling the G4 canonical motif (or parts of G4 canonical motif), but also numerous known transcription factors which could play important roles in enhancing or inhibiting G4 activity directly or indirectly. By mapping active G4 regions that encapsulate one or more potential G4s, DeepG4 represents a complementary approach to existing algorithms based on regular expressions or propensity scores, which can be further used to precisely localize the G4s within the active G4 regions.

## Materials and methods

### G4 data

We downloaded G4 ChIP-seq data for HaCaT, K562 and HEKnp cell lines from Gene Expression Omnibus (GEO) accession numbers GSE76688, GSE99205 and GSE107690 [24, 28, 29]. For every cell line, replicates were mapped to hg19 and merged for peak calling using macs2 with default parameters (<https://pypi.org/project/MACS2/>). We downloaded G4P ChIP-seq (similar to G4 ChIP-seq) peaks already mapped to hg19 for A549, H1975, 293T and HeLa-S3 cell lines from GEO accession number GSE133379 [30]. We used peaks from both replicates (when there were two available replicates). We downloaded processed G4-seq peaks mapped to hg19 from GEO accession number GSE63874 [23]. We used G4-seq from the sodium (Na) and potassium (K) conditions. No filtering step was performed on peak selection.

### Active G4 sequences

We defined positive DNA sequences (active G4 region sequences) as forming both in vitro and in vivo G4s as follows. We only kept G4 ChIP-seq peaks overlapping with G4-seq peaks. We then used the 201-bp DNA sequences centered on the G4 ChIP-seq peak summits.

As negative (control) sequences, we used sequences randomly drawn from the human genome with sizes, GC content (% GC), and repeat content (tandem repeat number from Tandem Repeat Finder mask from hg19 genome) similar to those of positive DNA sequences using genNullSeqs function from gkmSVM R package (<https://cran.r-project.org/web/packages/gkmSVM>).

## Chromatin accessibility

We downloaded processed DNase-seq bigwig files for different cell lines from ENCODE [31], and processed ATAC-seq bigwig files for HaCaT cell line from GSE7668. We downloaded processed ATAC-seq bigwig files from ICGC cancer cohorts from <https://gdc.cancer.gov/about-data/publications/ATACseq-AWG> [32].

## ChromHMM annotations

We downloaded ChromHMM annotations for ENCODE cell lines from <http://hgdownload.cse.ucsc.edu/goldenpath-hg19/encodeDCC/wgEncodeBroadHmm/> [33].

## BRCA cancer mutations

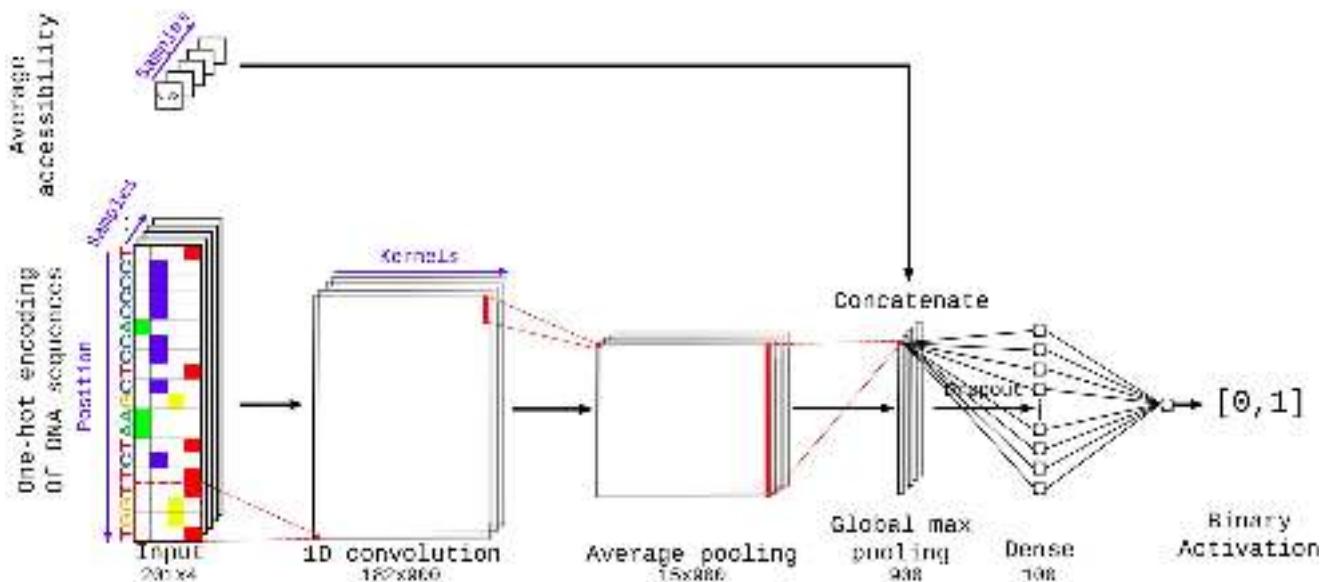
We downloaded breast cancer processed mutation data from ICGC BRCA-US cohort from the portal <https://dcc.icgc.org>.

## JASPAR DNA motifs

We used position weight matrices (PWMs) for transcription factor binding sites from the JASPAR 2018 database (<http://jaspar.genereg.net>).

## DeepG4 model

DeepG4 is a feedforward neural network composed of several layers illustrated in Fig 1. DNA sequence is first encoded as a one-hot encoding layer. Then, a 1-dimension convolutional layer is used with kernels to model DNA motifs. A local average pooling layer is next used. Then, the global max pooling layer extracts the highest signal from the sequence. Dropout is used for regularization. A dense layer then combines the different kernels and the activation sigmoid layer allows to compute the score between 0 and 1 of a sequence to be an active



**Fig 1. DeepG4 model architecture.** Here, one-hot encoding is a numerical encoding of a 201-bp DNA sequence as a  $201 \times 4$  matrix where each column corresponds to a DNA letter (A, C, G or T), and for instance, a value of one in the first column corresponds to a letter A in the sequence at a given position. For one-hot encoding, colored cells indicate ones, while white cells indicate zeroes.

<https://doi.org/10.1371/journal.pcbi.1009308.g001>

G4. The model is described in details in Subsection Results and Discussion, Deep learning approach.

Best hyperparameters including the number of kernels (900), kernel size (20 bp), kernel activation (relu), pool size (12 bp), drop-out (0%), epoch number (20), number of neurons in the dense layer (100) and the optimizer choice (rmsprop) were selected by Bayesian optimization [34]. In [S1 Fig](#), we illustrated how changing the hyper-parameters influenced the accuracy.

### DNA motifs from DeepG4

The first layer of DeepG4 contains kernels capturing specific sequence patterns similar to DNA motifs. In order to obtain DNA motifs from the first layer (convolutional layer) of DeepG4, we proceeded as follows (see [S2 Fig](#)). For a given kernel, we computed activation values for each positive sequence. If a positive sequence contained activation values above 0 (motif hits), we extracted the sub-sequence having the maximum activation value (best motif hit sequence). The set of sub-sequences was then used to obtain a position frequency matrix (PFM) by computing the frequency of each DNA letter at each position for the kernel.

Each kernel PFM was then trimmed by removing low information content positions at each side of the PFM (threshold >0.9). PFMs whose size were lower than 5 bases after trimming were removed. PWMs were next computed from PFMs assuming background probability of 0.25 for each DNA letter as done in JASPAR.

Because many PWMs from DeepG4 were redundant, we used the motif clustering program matrix-clustering from RSAT suite (<http://rsat.sb-roscoff.fr/>) with parameters: median, cor = 0.6, ncor = 0.6. We used PWM cluster centers as DNA motifs for further analyses.

### DeepG4 implementation and sequence availability

DeepG4 was implemented using Keras R library (<https://keras.rstudio.com/>). DeepG4 is available at <https://github.com/morphos30/DeepG4>. All fasta files used for training and predictions were also deposited.

### Performance analyses of DeepG4 and DeepG4\*

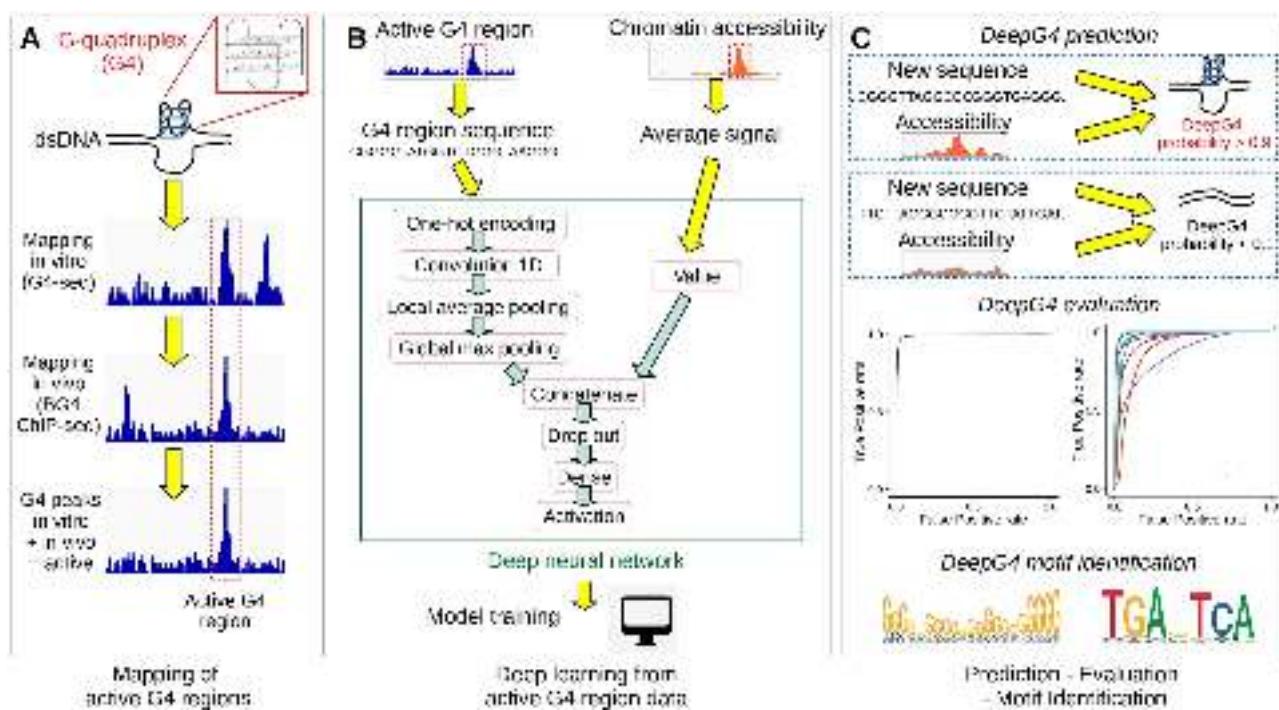
Performance analyses of DeepG4 and DeepG4\* presented in this article can be obtained using a pipeline and a docker available at <https://github.com/morphos30/DeepG4ToolsComparison>.

## Results and discussion

### Deep learning approach

Our computational approach, called DeepG4, for predicting active G4 regions is schematically illustrated in [Fig 2](#). In the first step ([Fig 2A](#)), we retrieved recent genome-wide mapping of in vitro G4 peak human sequences using G4-seq data [23] and of in vivo G4 peak human sequences using G4 ChIP-seq data [24]. Both methods mapped G4 regions at the resolution of few hundred base pairs, within which the exact locations of the G4s are unknown. By overlapping G4 ChIP-seq peaks with G4-seq peaks, we could identify a set of G4 peaks that were formed both in vitro and in vivo, and which we considered as “active G4 regions”. Moreover, we retrieved accessibility mapping data (DNase-seq / ATAC-seq) for the corresponding regions from the same cell line as the G4 ChIP-seq data.

In the second step ([Fig 2B](#)), we extracted the DNA sequences from active G4 regions (positive sequences). As negative sequences, we used sequences randomly drawn from the human genome with sizes, GC, and repeat contents similar to those of positive DNA sequences. For



**Fig 2. Illustration of DeepG4.** A) Mapping of active G4 region sequences both in vitro and in vivo using NGS techniques. B) Deep learning model training using active G4 regions and control sequences. C) G4 activity prediction, evaluation and motif identification.

<https://doi.org/10.1371/journal.pcbi.1009308.g002>

both positive and negative sequences, we computed the corresponding average chromatin accessibilities. Positive and negative sequences, together with average chromatin accessibility values, were then used to train our deep learning classifier called DeepG4. DeepG4 is a feedforward neural network composed of several layers. The DNA sequence (left input) is first encoded as a one-hot encoding layer. Then, a 1-dimension convolutional layer is used with 900 kernels (also called filters) and a kernel size of 20 bp to capture weighted DNA motifs predictive of active G4 regions. The optimal number of kernels and kernel size were determined by Bayesian optimization. A local average pooling layer with a pool size of 12 bp is next used (pool size selected by Bayesian optimization). This layer is important: it allows to aggregate kernel signals that are contiguous along the sequence, such that a G4 sequence can be modeled as multiple contiguous small motifs containing stretches of Gs. For instance, a G4 sequence can be defined by two contiguous motifs GGGNNNGGG separated by 5 bases, yielding the canonical motif GGGNNNGGGNNNNNGGGNNNGGG. Then, the global max pooling layer extracts the highest signal from the sequence for each kernel, and is concatenated with the average chromatin accessibility value (right input). Dropout is used for regularization. A dense layer then combines the different kernel signals. The activation sigmoid layer allows to compute the score between 0 and 1 of a sequence to be an active G4 region.

In the third step ([Fig 2C](#)), we used DeepG4 to predict the G4 region activity (score between 0 and 1) for a novel DNA sequence and its corresponding chromatin accessibility. We split the sequence set (set of positive and negative sequences) from HaCaT cell line (from GEO GSE76688 accession) into a training set to learn model parameters, a validation set to optimize hyper-parameters by Bayesian optimization and a testing set to assess model prediction accuracy. For this purpose, we computed the receiver operating characteristic (ROC) curve and the

area under the ROC (AUROC), as well as the precision-recall (PR) curve and the area under the PR (AUPR). DeepG4 motifs are extracted from the convolutional layer.

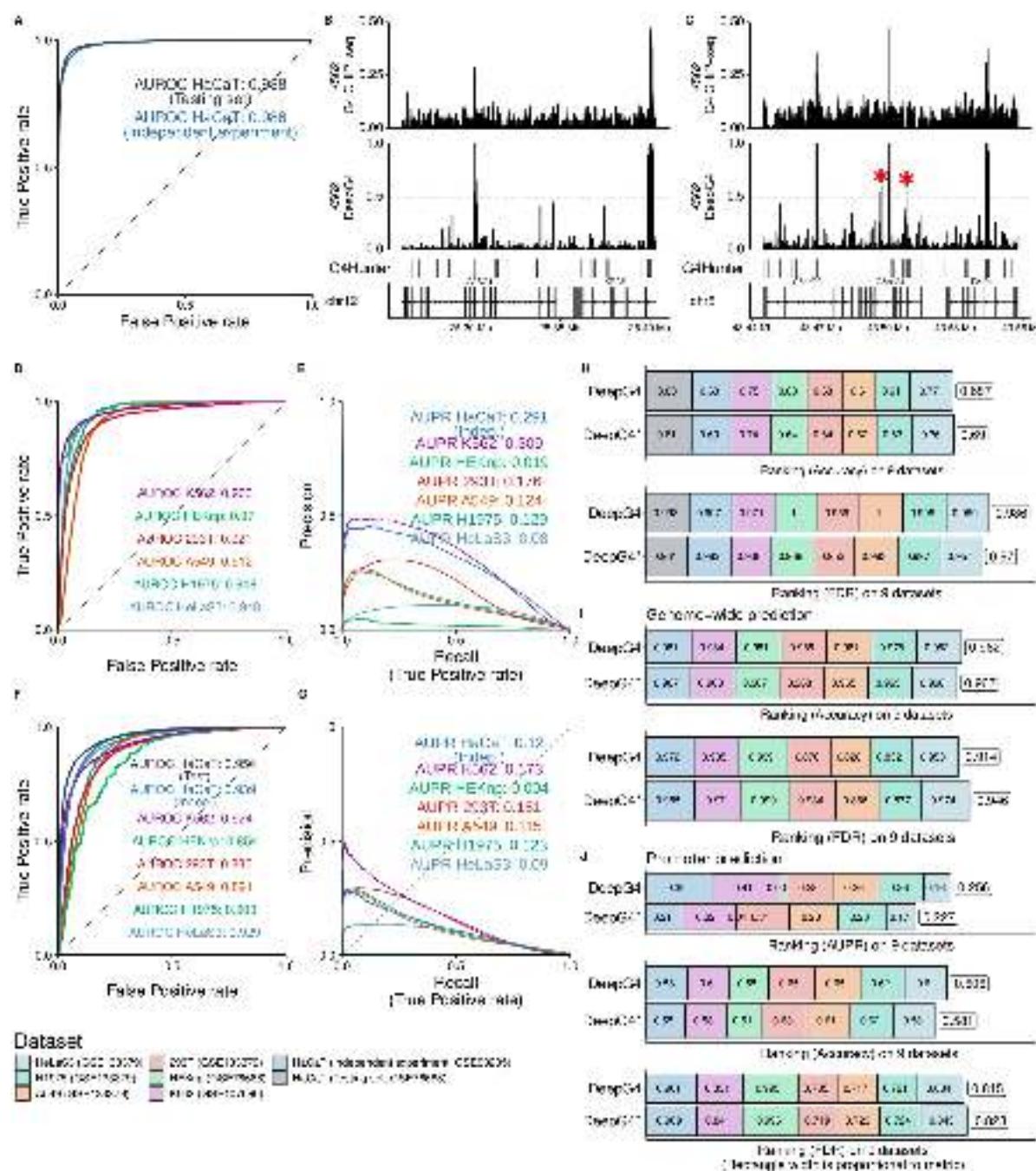
## G4 predictions with DeepG4

We then evaluated the prediction performance of DeepG4. In term of AUROC, DeepG4 obtained excellent predictions of active G4 regions from HaCaT cells on the testing set ([Fig 3A](#); AUROC = 0.988). On an independent ChIP-seq experiment done with the same cell line (from GEO GSE99205 accession), prediction performance of DeepG4 also showed very high accuracy (AUC = 0.986; [Fig 3A](#)). We then evaluated the ability of DeepG4 trained on one cell line (HaCaT) to predict G4s in another cell line (e.g. K562). We first browsed the genome where G4 regions were mapped by ChIP-seq as active in K562. For instance, we looked around the oncogene KRAS known to be regulated by a G4 in its promoter ([Fig 3B](#)). ChIP-seq mapped one active G4 region in the promoter of KRAS, which was also predicted with high score by DeepG4 (score > 0.95). On the left side of KRAS, another active G4 region was mapped experimentally within CASC1 gene and was also predicted by DeepG4. On another locus, ChIP-seq mapped three main active G4 regions, located inside the genes C5orf28 (TMEM267), C5orf34 and PAIP1 ([Fig 3C](#)). These three regions were also predicted as active G4 regions with high score (score > 0.95). DeepG4 also mistakenly predicted with medium score two other regions within C5orf34 (score ≈ 0.6, red stars), which were not mapped by ChIP-seq.

Overall, DeepG4, which was trained using HaCaT cell line data, could well predict in other cell lines. For instance, the AUROC was very high for HEKnp (AUROC = 0.97; [Fig 3D](#)). For K562, HeLaS3 and H1975, AUROCs were also very good (K562: AUROC = 0.963; HeLaS3: AUROC = 0.948; H1975: AUROC = 0.948), except for 293T and A549, which presented good but slightly lower accuracy (293T: AUROC = 0.921; A549: AUROC = 0.912). We then evaluated predictions over the whole genome in an unbiased way. For this purpose, we split the genome into 200-base bins, and evaluated DeepG4 ability to discriminate between bins corresponding to active G4 regions (tens of thousands of bins) and other bins (millions of bins). Despite this highly imbalanced data, DeepG4 showed good prediction accuracy as measured by AUPR for HaCaT (AUPR = 0.291, independent experiment), K562 (AUPR = 0.309), 293T (AUPR = 0.176), A549 (AUPR = 0.124) and H1975 (AUPR = 0.129) ([Fig 3E](#)). For some cell lines, predictions were less good (HEKnp: AUPR = 0.019; HeLaS3: AUPR = 0.08).

We previously hypothesized that chromatin accessibility could help to produce cell-type specific predictions. To verify this assumption, chromatin accessibility was removed from DeepG4 model (yielding an alternative model called DeepG4\*). Removing chromatin accessibility significantly lowered cell-type specific prediction accuracy. For instance, the AUROC of HaCaT (independent) was 0.939 for DeepG4\* as compared to 0.986 for DeepG4, which represented an important difference ([Fig 3F](#)). We also found a large difference for HEKnp (DeepG4\*, AUROC = 0.854; DeepG4, AUROC = 0.970). In terms of accuracy and false discovery rate (FDR) metrics, DeepG4\* performed slightly less well than DeepG4 ([Fig 3H](#)). Regarding genome-wide predictions, removing chromatin accessibility also significantly lowered prediction performance ([Fig 3G](#)). For instance, for HaCaT (independent), we obtained an AUPR of 0.120 with DeepG4\* and an AUPR of 0.291 with DeepG4. Regarding accuracy metric, DeepG4\* performed less well than DeepG4, but slightly better in term of FDR ([Fig 3I](#)). We also assessed predictions on promoters to distinguish the promoters with active G4 regions from the promoters without active G4 regions. DeepG4\* performed less well than DeepG4 in term of AUPR and accuracy, but slightly better in term of FDR ([Fig 3J](#)).

These results thus demonstrated the ability of DeepG4 to accurately predict cell-type specific active G4 regions from DNA sequences and chromatin accessibility. Moreover, results



**Fig 3. Prediction performance of DeepG4 to predict active G4 regions (regions where G4s form both in vitro and in vivo). A)** Prediction performance of DeepG4. The model was trained and evaluated using HaCaT cell data. Predictions were evaluated on the testing set of sequences (same experiment as training set), but also on an independent set of sequences (from a different ChIP-seq experiment). Receiver operating characteristic (ROC) curve and area under the ROC curve (AUROC) were plotted. **B)** Genome browser of HaCaT-trained DeepG4 predictions and G4 ChIP-seq around KRAS gene in K562 cells. **C)** Genome browser of HaCaT-trained DeepG4 predictions and G4 ChIP-seq around C5orf34 gene in K562 cells. **D)** Prediction performance of DeepG4 trained using HaCaT data and evaluated on other cell lines. **E)** Genome-wide prediction performance of DeepG4 trained using HaCaT data and evaluated on other cell lines. Predictions are computed for every 200-b bins of the genome. Area Under the Precision-Recall curve is plotted (AUPR). **F)** Prediction performance of DeepG4\* trained using HaCaT data and evaluated on other cell lines. DeepG4\* is identical to DeepG4 except that chromatin accessibility is not used as input. **G)** Genome-wide prediction performance of DeepG4\* trained using HaCaT data and evaluated on other cell lines. **H)** Comparison of DeepG4 and DeepG4\* prediction performances, in terms of accuracy and false discovery rate (FDR) metrics. **I)** Comparison of DeepG4 and DeepG4\* genome-wide prediction performances, in terms of accuracy and false discovery rate (FDR) metrics. **J)** Comparison of DeepG4 and DeepG4\* promoter prediction performances, in terms of AUPR, accuracy and false discovery rate (FDR) metrics.

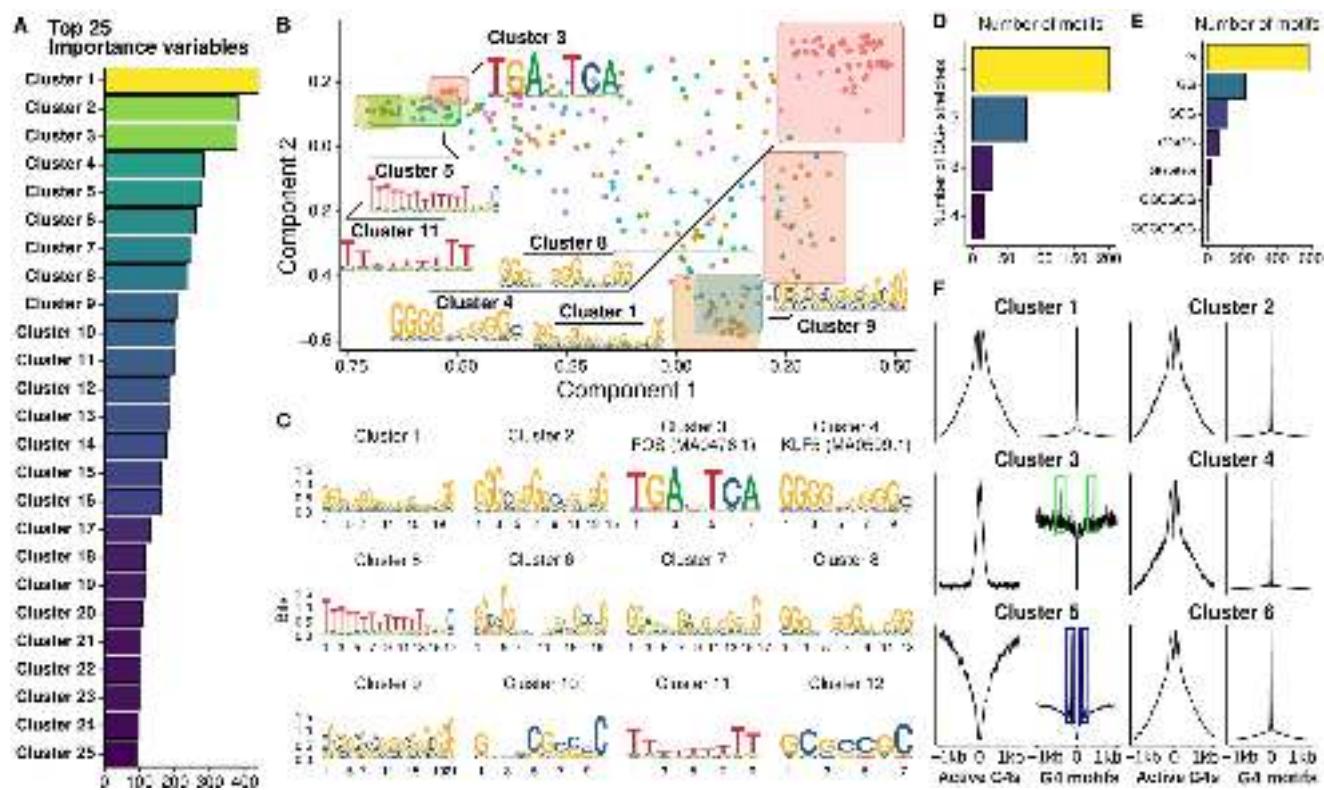
<https://doi.org/10.1371/journal.pcbi.1009308.g003>

also revealed the importance of incorporating chromatin accessibility into DeepG4 for cell-type specific predictions.

### Identification of important motifs from DeepG4

The first layer of DeepG4 convolutional neural network encapsulated kernels that encoded DNA motifs predictive of active G4s. Hence, we extracted from the first layer the kernels and converted them to DNA motif PWMs to better understand which motifs were the best predictors of G4 activity. DeepG4 identified 900 motifs, many of them were redundant. To remove redundancy, we clustered the motifs using RSAT matrix-clustering program and kept the cluster motifs (also called root motifs in the program) for subsequent analyses. Cluster motifs could be divided into two groups: a group of de novo motifs and a group of motifs that resembled known TFBS motifs. To distinguish between these two groups, we used TomTom program (MEME suite) which mapped the cluster motifs to JASPAR database. DeepG4 motifs matching JASPAR were considered as known TFBS motifs, while motifs that did not match were classified as de novo motifs.

We first assessed the ability of DeepG4 motifs to predict active G4 regions. Hence, we computed DeepG4 cluster motif variable importances using random forests and found strong predictors (Fig 4A). In order to visualize the cluster motifs on a map, we used multi-dimensional scaling (MDS), where we also plotted the original kernel motifs used to build the cluster



**Fig 4. DNA motifs identified by DeepG4.** A) Variable importances of DeepG4 cluster motifs, as estimated by random forests. Clustering of DeepG4 kernel motifs was done by RSAT matrix-clustering program to obtain cluster motifs. B) Multidimensional scaling (MDS) of DeepG4 motifs. As an input, matrix-clustering correlation matrix between kernel motifs was used. C) Logos of cluster motifs with highest variable importances. D) Number of kernel motifs containing one or more GG+ stretches. A GG+ stretch is defined as a stretch of 2 or more Gs in the motif consensus sequence. E) Number of kernel motifs containing G stretches depending on stretch length. F) Average profiles measuring the enrichment of cluster motifs centered around active G4 regions or canonical G4 motifs.

<https://doi.org/10.1371/journal.pcbi.1009308.g004>

motifs. We found that the first MDS component reflected the guanine stretch length (higher at the right side), while the second component represented the G content (higher at the bottom) ([Fig 4B](#)).

Many strong predictors were de novo motifs which resembled the G4 canonical motif or parts of the canonical motif. For instance, cluster 1 comprised 4 stretches of GG+, thus almost forming a canonical G4 motif ([Fig 4C](#)). Cluster 2 comprised three stretches of GG+, could thus be considered as three quarters of a canonical G4 motif. We then counted GG+ stretches (stretches of 2 or more guanines) from the kernel motifs and found that many kernel motifs contained more than one GG+ stretch ([Fig 4D](#)). Moreover, the guanine stretches were of varying lengths, ranging from one G up to 5 Gs ([Fig 4E](#)). Among the best predictors, we also found several motifs corresponding to known TFBS motifs ([Fig 4C](#)). For instance, the third best predictor, cluster 3, almost perfectly matched FOS motif MA0476.1 (q-value =  $2 \times 10^{-10}$ ). Other strong predictors, such as cluster 4, matched KLF5 motif MA0599.1 (q-value = 0.09). It was very interesting to observe that such motif corresponding to one half of a canonical G4 motif also matched a known TFBS motif, which supported the complex interplay between G4s and TFBS protein binding [[35](#)].

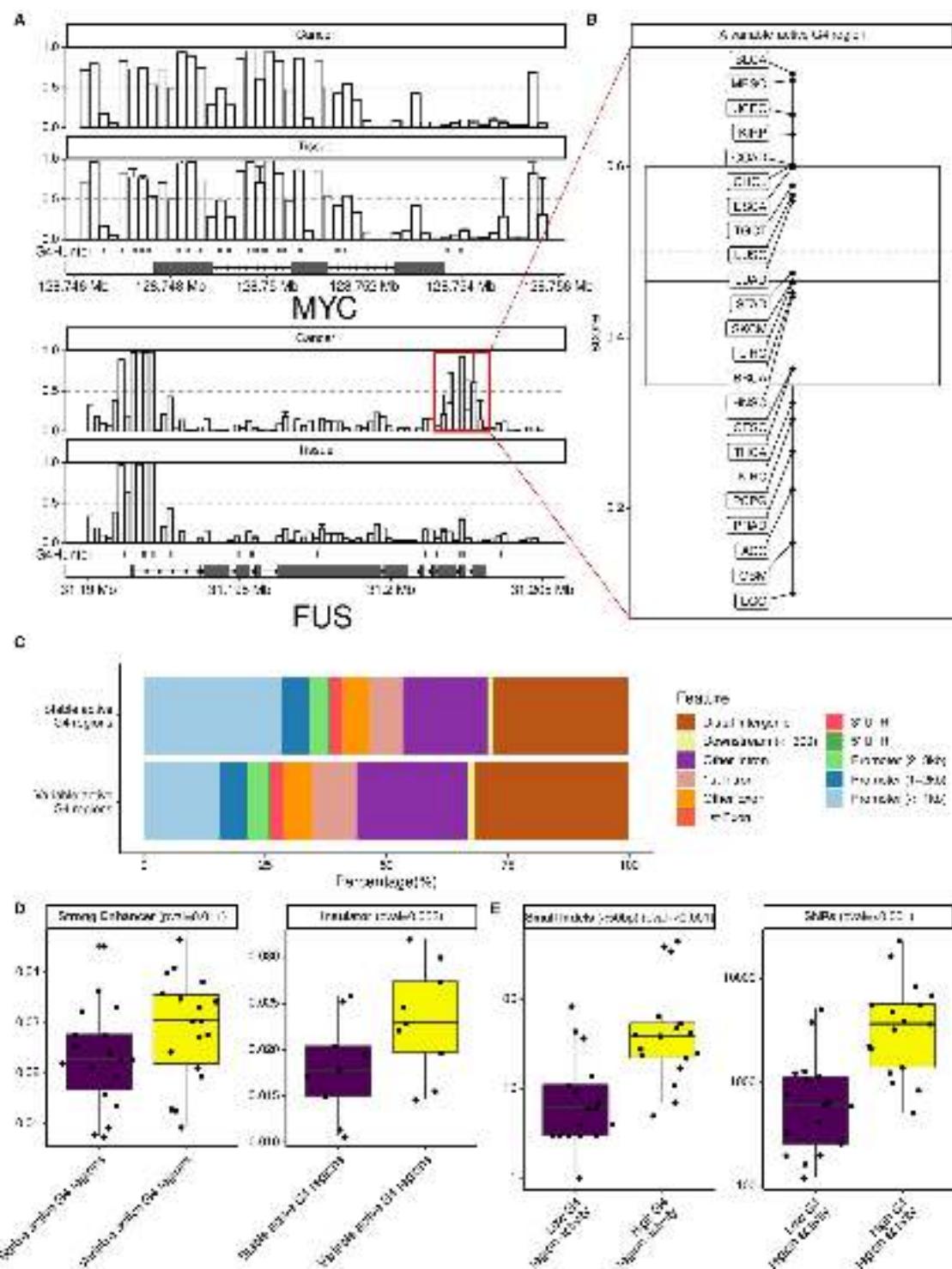
We then assessed the enrichment of DeepG4 cluster motifs around active G4 regions and around canonical G4 motifs ([Fig 4F](#)). Motifs resembling G4 canonical motif or parts of it, such as clusters 1 and 2, were enriched at both active G4 regions and canonical G4 motifs, thus representing actual G4 structures. But other motifs that were very different from the G4 canonical motif, such as cluster 3, were strongly enriched at active G4 regions, but depleted at the exact location of canonical G4 motifs. Interestingly, cluster 3 was enriched close to the canonical G4 motifs (around 300 bp, framed in green), suggesting that cluster 3 (FOS motif MA0476.1) did not participate directly to the G4 structure, but could act in the vicinity to support G4 activity. Conversely, we also found a motif composed mainly of Ts (poly(T) tract), the cluster 5 motif, which was depleted in active G4 regions, but which was at the same time enriched in the vicinity of canonical G4 motifs (framed in blue). This suggests that such poly (T) motif could inhibit the activity of G4 motifs by acting in the vicinity.

These observations revealed the important role of TFBS motifs that could act directly in G4 activity as part of G4 structure, as previously shown for SP1 in vitro [[36](#)], or could participate indirectly to support or inhibit G4 activity in the vicinity of G4s such as FOS motif (AP-1 complex).

### Genome-wide predictions in tissues and cancers

Using DeepG4, we could map active G4 regions genome-wide in many different tissues and cancers for which no G4 ChIP-seq experiments were available, but for which we could find publicly available chromatin accessibility data (ATAC-seq or DNase-seq). Hence, we made the mapping available on the DeepG4 Github repository as a resource for the G4 community.

We first browsed the genome at known oncogenes and looked at predicted active G4 regions ([Fig 5A](#)). In MYC, we predicted many active G4 regions in the promoter but also in the exons and introns. Predicted G4 activity was rather stable and did not vary across the tissues and cancers. In another gene, FUS, we found that the promoter contained an active G4 region that was very stable across tissues and cancer (left side), but we also could identify another G4 region toward the transcription end site (TES, right side) that was not predicted to be active in tissues, but predicted to be active in some cancers (framed in red), in particular in MESO (Mesothelioma), UCEC (Uterine Corpus Endometrial Carcinoma) and BLCA (Bladder Cancer), and inactive in some other cancers including GBM (Brain Cancer) and LGG (Brain Lower Grade Glioma) ([Fig 5B](#)). Thus, DeepG4 could identify regions of variable G4 activity.



**Fig 5. Genome-wide prediction of active G4 regions in tissues and cancers.** A) Genome browser of DeepG4 predictions at MYC and FUS genes in tissues and cancers. B) Relationship between DeepG4 predicted G4 activity and the amount of mutations, depending on the mutation class. Cancer cohort abbreviations (e.g. MESO) are detailed in [S1 Table](#). C) Annotations of predicted stable and variable active G4 regions. D) Mutation rates in BRCA breast cancer depending on predicted G4 region activity.

<https://doi.org/10.1371/journal.pcbi.1009308.g005>

Overall, only a minority of predicted G4 regions varied across the tissues and cancers (around 10%). When we annotated these regions and compared with stable G4 regions, we observed that 29% of stable G4 regions located within promoters, whereas only 16% of variable G4 regions colocalized with promoters (Fig 5C). Instead, we found variable G4 regions in intronic and intergenic regions. We further explored the role of variable G4 regions by using annotations from ENCODE in multiple cell lines from ChromHMM tool [33]. We found that variable G4 regions were enriched at strong enhancers as compared to stable G4 regions ( $p = 0.011$ , Fig 5D), and we also found a near-significant enrichment at insulator regions ( $p = 0.063$ , Fig 5D) in agreement with previous studies showing enrichment near CTCF at 3D domain (topologically associating domain, TAD) borders [37].

Since G4s are known mutagenic regions when unresolved, we then looked at the link between G4 activity and mutation rates in BRCA breast cancer (Fig 5E). We found a strong positive link between high G4 activity and SNP and small indel mutation rates, meaning that when G4s were formed *in vivo* they had a higher chance of yielding mutations and therefore this suggests that the chromatin landscape could greatly influence G4 impact on genome instability at a local scale.

## Conclusion

In this article, we propose a novel deep learning method, named DeepG4, to predict active G4 regions from DNA sequence and chromatin accessibility. The proposed method is designed to predict active G4 regions *i.e.* regions that are detected both *in vitro* and *in vivo*, unlike previous algorithms that were developed to predict G4s forming *in vitro* (naked DNA). For this purpose, our method exploits the genomic context of G4s, which comprises the G4(s) as well as other motifs in the vicinity that may play a role in G4 activity (*i.e.* transcription factor motifs). Moreover, adding chromatin accessibility into the model allows to predict active G4 regions depending on the cell type. Our novel method which maps active G4 regions in a cell-type specific manner at 201-bp resolution is complementary to existing algorithms based on regular expression (*e.g.* quadparser) and scores (*e.g.* G4Hunter), which map the exact location of potential G4 forming sequences and propensities. Moreover, DeepG4 provides a useful tool for mapping active G4 regions for cell lines, tissues and cancers for which no experimental data are available to date. Therefore, DeepG4 comprehensive predictions in tissues and cancers will represent a useful resource for the G4 community.

DeepG4 uncovered numerous specific DNA motifs predictive of active G4s. Many motifs resembled the canonical G4 motif ( $G_{3+} N_{1-7} G_{3+} N_{1-7} G_{3+} N_{1-7} G_{3+}$ ) or even parts of it. Most notably, many motifs corresponded to half or 3/4 of the canonical motif. The combination of these G4 parts, which is captured by DeepG4 as a deep neural network, brings flexibility in G4 modeling. Strikingly, some motifs completely or partly matched known TFBS motifs including KLF5 motif MA0599.1 and FOS (AP-1) motif MA0476.1, suggesting that they could contribute directly to G4 structures themselves or participate indirectly in G4 activity in the vicinity through the binding of transcription factors. In line with this result, it was previously found that G4s are enriched in the vicinity of the architectural protein CTCF at 3D domain (topologically associating domain, TAD) borders [37]. Moreover, it has been shown that SP1 binds to G4s with a comparable affinity as its canonical motif [36], and that G4s are TF hubs [35]. It was also surprising to find a poly(T) motif (cluster 5 motif) depleted in active G4 regions but enriched in the vicinity of canonical G4 motifs, suggesting that such motif could inhibit the activity of canonical G4 motifs in its vicinity.

In addition, we used DeepG4 to predict active G4 regions genome-wide in many tissues and cancers, thereby providing a resource for the chromatin and G4 community. Interestingly,

we identified two types of active G4 regions, those stable across tissues and cancers, and those less frequent that are variable. We found that variable active G4 regions are located within intronic and intergenic regions, and could act as enhancers and insulators, unlike stable G4 regions that are more enriched in promoters.

There are several limitations of the proposed approach. First, one limit of DeepG4 (as well as the other existing machine/deep learning methods) is that it requires a region of several hundred bases, thereby restricting the resolution of G4 mapping. Once an active G4 region is mapped, methods such as G4Hunter or pqsfinder have to be used to identify the exact position of the G4(s) within the region. Our model could be improved by adding novel neural layers in order to find as well the exact location of potential G4 sequences. Second, DeepG4 does not process the DNA sequence in a strand-specific manner, thus a given motif could be redundantly encoded in both strands within the convolutional layer. However, post-processing of DeepG4 motifs using methods such as matrix-clustering alleviates such problem by mapping complementary motifs (same motifs on different strands) to each other to merge them into cluster motifs. Third, the prediction performance of DeepG4 strongly depends on existing datasets that are limited, potentially inaccurate and biased, especially regarding *in vivo* mapping. Once more techniques for *in vivo* G4 mapping will be developed, DeepG4 will need to be retrained in order to improve prediction accuracy. Moreover, since DeepG4 was trained based on human data, predictions on non-mammalian genomes are expected to be less accurate. Fourth, DeepG4 is limited to predict active G4s but a similar approach could be used to predict any active non-B DNA structure using permanganate/S1 nuclease footprinting data [38].

## Supporting information

**S1 Fig. Prediction accuracy estimated from the validation set depending on hyper-parameters, as found from Bayesian optimization.** For each hyper-parameter, the optimum is marked as a red triangle.

(TIF)

**S2 Fig. Extraction and processing of DNA motifs from DeepG4 convolutional layer.**

(TIF)

**S1 Table. Cancer cohort abbreviations from ICGC project.**

(TIF)

## Acknowledgments

The authors are grateful to Balasubramanian lab (University of Cambridge, UK) and to Tan Zheng's group (Chinese Academy of Medicine) for data. The authors are very also thankful to Matthias Zytnicki, Catherine Tardin and the Legube's team for comments.

## Author Contributions

**Conceptualization:** Raphael Mourad.

**Data curation:** Elissar Nassereddine.

**Formal analysis:** Raphael Mourad.

**Investigation:** Vincent Rocher, Matthieu Genais, Elissar Nassereddine, Raphael Mourad.

**Methodology:** Vincent Rocher, Matthieu Genais, Raphael Mourad.

**Resources:** Vincent Rocher.

**Software:** Vincent Rocher.

**Supervision:** Raphael Mourad.

**Validation:** Vincent Rocher.

**Writing – original draft:** Raphael Mourad.

**Writing – review & editing:** Raphael Mourad.

## References

1. Watson JD, Crick FH. A structure for deoxyribose nucleic acid. *Nature*. 1953; 171:737–738. <https://doi.org/10.1038/171737a0>
2. Sen D, Gilbert W. Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis. *Nature*. 1988; 334(6180):364–366. <https://doi.org/10.1038/334364a0>
3. Chen Y, Yang D. Sequence, stability, and structure of G-quadruplexes and their interactions with drugs. *Current Protocols in Nucleic Acid Chemistry*. 2012; 50(1):17.5.1–17.5.17.
4. Bhattacharyya D, Mirihana Arachchilage G, Basu S. Metal cations in G-quadruplex folding and stability. *Frontiers in Chemistry*. 2016; 4:38.
5. Spiegel J, Adhikari S, Balasubramanian S. The structure and function of DNA G-quadruplexes. *Trends in Chemistry*. 2019;
6. Fay MM, Lyons SM, Ivanov P. RNA G-quadruplexes in biology: Principles and molecular mechanisms. *Journal of Molecular Biology*. 2017; 429(14):2127–2147. <https://doi.org/10.1016/j.jmb.2017.05.017>
7. Varshney D, Spiegel J, Zyner K, Tannahill D, Balasubramanian S. The regulation and functions of DNA and RNA G-quadruplexes. *Nature Reviews Molecular Cell Biology*. 2020; 21(8):459–474. <https://doi.org/10.1038/s41580-020-0236-x>
8. Sfeir A. Telomeres at a glance. *Journal of Cell Science*. 2012; 125(18):4173–4178. <https://doi.org/10.1242/jcs.106831>
9. Wang Q, Liu Jq, Chen Z, Zheng Kw, Chen Cy, Hao Yh, et al. G-quadruplex formation at the 3' end of telomere DNA inhibits its extension by telomerase, polymerase and unwinding by helicase. *Nucleic Acids Research*. 2011; 39(14):6229–6237. <https://doi.org/10.1093/nar/gkr164> PMID: 21441540
10. Bryan TM. G-quadruplexes at telomeres: Friend or foe? *Molecules*. 2020; 25(16). <https://doi.org/10.3390/molecules25163686> PMID: 32823549
11. Brooks TA, Hurley LH. Targeting MYC expression through G-quadruplexes. *Genes & Cancer*. 2010; 1(6):641–649. <https://doi.org/10.1177/1947601910377493>
12. Marnef A, Cohen S, Legube G. Transcription-coupled DNA double-strand break repair: Active genes need special care. *Journal of Molecular Biology*. 2017; 429(9):1277–1288. <https://doi.org/10.1016/j.jmb.2017.03.024>
13. Cimino-Reale G, Zaffaroni N, Folini M. Emerging role of G-quadruplex DNA as target in anticancer therapy. *Current Pharmaceutical Design*. 2016; 22(44):6612–6624.
14. Asamitsu S, Takeuchi M, Ikenoshita S, Imai Y, Kashiwagi H, Shioda N. Perspectives for applying G-quadruplex structures in neurobiology and neuropharmacology. *International Journal of Molecular Sciences*. 2019; 20(12). <https://doi.org/10.3390/ijms20122884> PMID: 31200506
15. Hänsel-Hertsch R, Simeone A, Shea A, Hui WWI, Zyner KG, Marsico G, et al. Landscape of G-quadruplex DNA structural regions in breast cancer. *Nature Genetics*. 2020; 52(9):878–883. <https://doi.org/10.1038/s41588-020-0672-8> PMID: 32747825
16. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001; 409(6822):860–921. <https://doi.org/10.1038/35057062>
17. Puig Lombardi E, Londono-Vallejo A. A guide to computational methods for G-quadruplex prediction. *Nucleic Acids Research*. 2019; 48(1):1–15.
18. Miskiewicz J, Sarzynska J, Szachniuk M. How bioinformatics resources work with G4 RNAs. *Briefings in Bioinformatics*. 2020;
19. Huppert JL, Balasubramanian S. Prevalence of quadruplexes in the human genome. *Nucleic Acids Research*. 2005; 33(9):2908–2916. <https://doi.org/10.1093/nar/gki609>
20. Huppert JL, Balasubramanian S. G-quadruplexes in promoters throughout the human genome. *Nucleic Acids Research*. 2006; 35(2):406–413.

21. Bedrat A, Lacroix L, Mergny JL. Re-evaluation of G-quadruplex propensity with G4Hunter. *Nucleic Acids Research*. 2016; 44(4):1746–1759. <https://doi.org/10.1093/nar/gkw006>
22. Hon J, Martinek T, Zendulka J, Lexa M. pqsfinder: an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R. *Bioinformatics*. 2017; 33(21):3373–3379. <https://doi.org/10.1093/bioinformatics/btx413>
23. Chambers VS, Marsico G, Boutell JM, Di Antonio M, Smith GP, Balasubramanian S. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nature Biotechnology*. 2015; 33(8):877–881. <https://doi.org/10.1038/nbt.3295>
24. Hänsel-Hertsch R, Bernaldi D, Lensing SV, Marsico G, Zyner K, Parry A, et al. G-quadruplex structures mark human regulatory chromatin. *Nature Genetics*. 2016; 48(10):1267–1272. <https://doi.org/10.1038/ng.3662> PMID: 27618450
25. Sahakyan AB, Chambers VS, Marsico G, Santner T, Di Antonio M, Balasubramanian S. Machine learning model for sequence-driven DNA G-quadruplex formation. *Scientific Reports*. 2017; 7(1):14535. <https://doi.org/10.1038/s41598-017-14017-4>
26. Klimentova E, Polacek J, Simecek P, Alexiou P. PENGUINN: Precise exploration of nuclear G-quadruplexes using interpretable neural networks. *bioRxiv*. 2020; <https://doi.org/10.3389/fgene.2020.568546> PMID: 33193663
27. Barshai M, Orenstein Y. Predicting G-quadruplexes from DNA sequences using multi-kernel convolutional neural networks. In: Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics. BCB'19. New York, NY, USA: Association for Computing Machinery; 2019. p. 357–365. Available from: <https://doi.org/10.1145/3307339.3342133>.
28. Hänsel-Hertsch R, Spiegel J, Marsico G, Tannahill D, Balasubramanian S. Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing. *Nature Protocols*. 2018; 13(3):551–564. <https://doi.org/10.1038/nprot.2017.150>
29. Mao SQ, Ghanbarian AT, Spiegel J, Martinez Cuesta S, Bernaldi D, Di Antonio M, et al. DNA G-quadruplex structures mold the DNA methylome. *Nature Structural & Molecular Biology*. 2018; 25(10):951–957. <https://doi.org/10.1038/s41594-018-0131-8> PMID: 30275516
30. Zheng Kw, Zhang Jy, He Yd, Gong Jy, Wen Cj, Chen Jn, et al. Detection of genomic G-quadruplexes in living cells using a small artificial protein. *Nucleic Acids Research*. 2020; 48(20):11706–11720. <https://doi.org/10.1093/nar/gkaa841>
31. The ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*. 2012; 489(7414):57–74. <https://doi.org/10.1038/nature11247>
32. Zhang J, Bajari R, Andric D, Gerhoffert F, Lepsa A, Nahal-Bose H, et al. The International Cancer Genome Consortium Data Portal. *Nature Biotechnology*. 2019; 37(4):367–369. <https://doi.org/10.1038/s41587-019-0055-9> PMID: 30877282
33. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*. 2012; 9(3):215–216. <https://doi.org/10.1038/nmeth.1906>
34. Snoek J, Larochelle H, Adams RP. Practical Bayesian optimization of machine learning algorithms. In: Proceedings of the 25th International Conference on Neural Information Processing Systems—Volume 2. NIPS'12. Red Hook, NY, USA: Curran Associates Inc.; 2012. p. 2951–2959.
35. Spiegel J, Cuesta SM, Adhikari S, Hänsel-Hertsch R, Tannahill D, Balasubramanian S. G-quadruplexes are transcription factor binding hubs in human chromatin. *Genome Biology*. 2021; 22(1):117. <https://doi.org/10.1186/s13059-021-02324-z>
36. Raiber EA, Kranaster R, Lam E, Nikan M, Balasubramanian S. A non-canonical DNA structure is a binding motif for the transcription factor SP1 in vitro. *Nucleic Acids Research*. 2011; 40(4):1499–1508.
37. Hou Y, Li F, Zhang R, Li S, Liu H, Qin ZS, et al. Integrative characterization of G-Quadruplexes in the three-dimensional chromatin structure. *Epigenetics*. 2019; 14(9):894–911. <https://doi.org/10.1080/15592294.2019.1621140> PMID: 31177910
38. Kouzine F, Wojtowicz D, Baranello L, Yamane A, Nelson S, Resch W, et al. Permanganate/S1 nuclease footprinting reveals non-B DNA structures with regulatory potential across a mammalian genome. *Cell Systems*. 2017; 4(3):344–356.e7. <https://doi.org/10.1016/j.cels.2017.01.013> PMID: 28237796

### 2.1.1 Performances du premier modèle ADN sans accessibilité

#### 2.1.1.1 Construction du modèle avec des séquences d'ADN

Dans un premier temps, j'ai développé un modèle de *Deep Learning* pour prédire des régions de G4 actifs en utilisant les séquences des pics chevauchant de G4 ChIP-seq (*in vivo*, GSE76688, HÄNSEL-HERTSCH, SPIEGEL et al. (2018)) et de G4-seq (*in vitro* CHAMBERS et al. (2015)). Ce premier modèle possède la même architecture que le modèle avec accessibilité (voir Figure 1 de l'article), mais n'utilise que la séquence d'ADN comme prédicteur. Les hyper-paramètres du modèle ont été déterminés via une recherche aléatoire par grille (voir 1.4.3.9). Ceci nous a permis de trouver un ensemble de valeurs pour nos hyper-paramètres optimisant la prédiction sur le jeu de données de validation : Le nombre de *kernels* a été fixée à 900, leur taille à 20. La couche de *pooling* a été fixée avec un hyper-paramètre de 10bp, et le nombre de neurones de la couche dense à 100.

#### 2.1.1.2 Prédiction de G4 actifs avec DeepG4\*

En utilisant DeepG4\*, nous obtenons d'excellentes prédictions sur le jeu de données de test (HaCaT, GSE76688, AUC=0.958) (voir Figure 2.1). Les prédictions du modèle sur un jeu de données indépendant provenant de la même lignée cellulaire montrent également de très bons résultats (HaCaT, GSE99205, AUC=0.946). Nous avons ensuite évalué la capacité du modèle, entraîné sur un type cellulaire spécifique (HaCaT), à prédire sur d'autres lignées cellulaires. Nous avons constaté que DeepG4\* est capable de prédire correctement les régions actives en G4 sur des données réalisées en G4 ChIP-seq (K562, AUC=0.926 ; HEKnp, AUC=0.871). DeepG4\* est également capable de prédire correctement des jeux de données produits par d'autres méthodes expérimentales : qG4 (similaire au G4 ChIP-seq) dans des cellules de cancer du sein (HÄNSEL-HERTSCH, SIMEONE et al. 2020) (AUC = 0.851), et de G4P (K.-w. ZHENG et al. 2020), dans différents types cellulaires (HeLaS3, AUC=0.931 ; H1975, AUC=0.907 ; A549, AUC=0.897).

#### 2.1.1.3 Comparaison de DeepG4\* avec les outils existants

Dans la littérature, il existe un grand nombre d'outils permettant de prédire la présence d'un ou plusieurs G4 dans des séquences biologiques. Afin d'estimer les performances de DeepG4\* (sans accessibilité), j'ai comparé les résultats d'un grand nombre d'algorithmes différents (voir 2.1) en utilisant 9 jeux de données de G4 actifs à notre disposition (voir Figure 2.2).

Pour une séquence donnée, plusieurs motifs G4 peuvent être détectés. Afin d'obtenir un seul score par séquence, nous calculons la somme des scores des motifs présents dans toute la séquence. Pour G4Hunter, nous avons utilisé les paramètres par défaut (son paramètre de seuil est fixé à 1.5), mais aussi entraîné un *Random Forest* en utilisant différentes valeurs de son seuil de détection comme variables d'entrées, afin d'améliorer ses prédictions. Afin de proposer une comparaison plus juste, nous avons également ré-entraîné les modèles de *Machine* et *Deep Learning* sur le même jeu de données d'entraînement que DeepG4\*.

Après comparaison, nous avons observés que DeepG4\* surpassé les autres algorithmes sur les données G4 ChIP-seq, et propose de bonnes performances sur les données de G4P et de qG4, avec une AUC moyenne

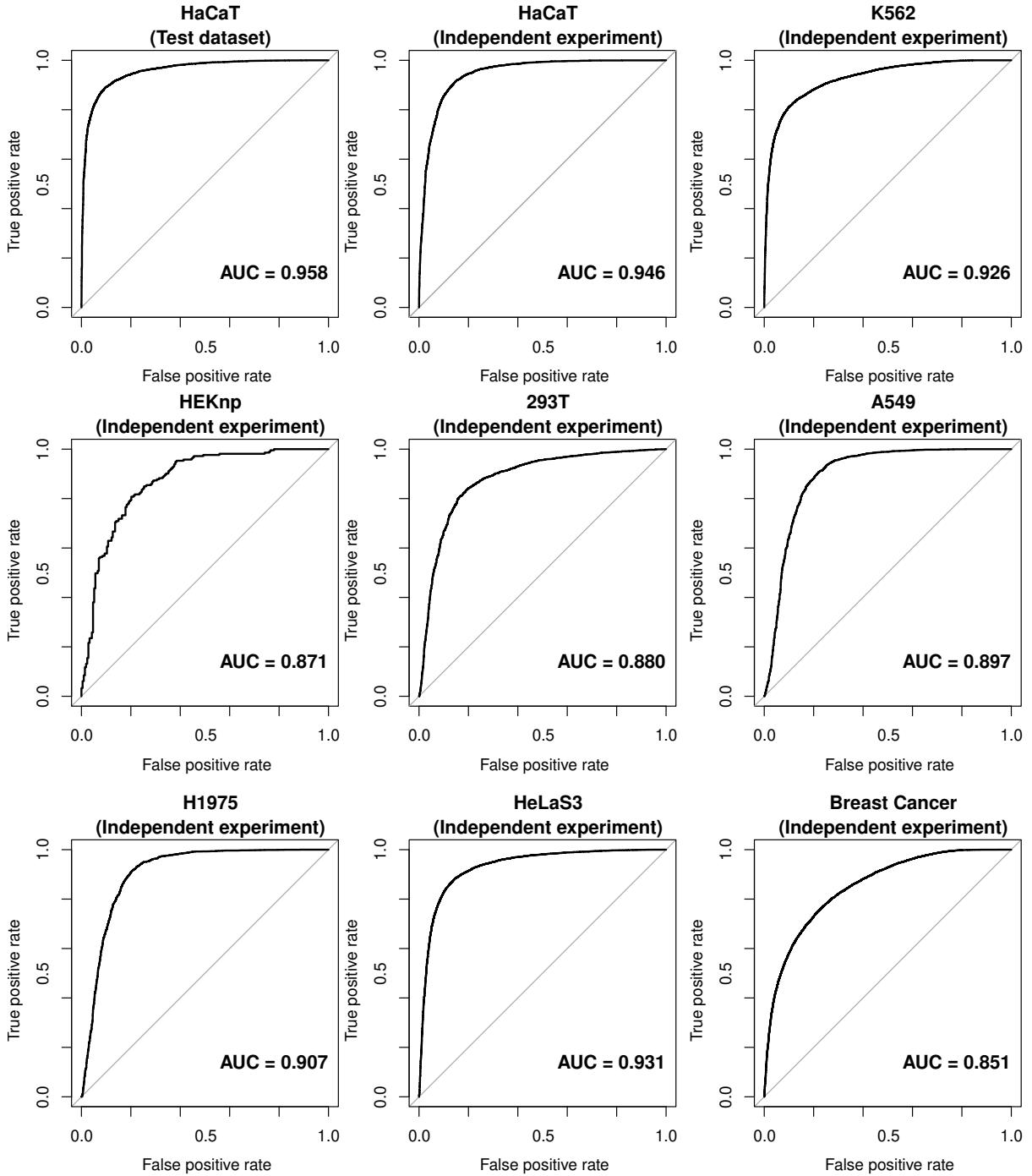


FIGURE 2.1 – Performances du modèle DeepG4\* (sans accessibilité). Le modèle a été entrainé et évalué à l'aide de données de la lignée cellulaire HaCaT (GSE76688). La courbe ROC et son aire sous la courbe (AUC) sont utilisées pour évaluer ses performances. DeepG4\* est également évalué avec différents jeux de données expérimentaux indépendants : HaCaT (Test, GSE76688), HaCaT (indépendant, GSE99205), K562 (GSE107690), HEKnp (GSE76688) provenant de données de G4 ChIP-seq (*in vivo*, HÄNSEL-HERTSCH, SPIEGEL et al. (2018)). Des données de qG4 (similaire au G4 ChIP-seq) dans des cellules de cancer du sein (HÄNSEL-HERTSCH, SIMEONE et al. 2020), et de G4P (K.-w. ZHENG et al. 2020) dans différents types cellulaires : HeLaS3, H1975, A549 (GSE133379) ont également été utilisées.

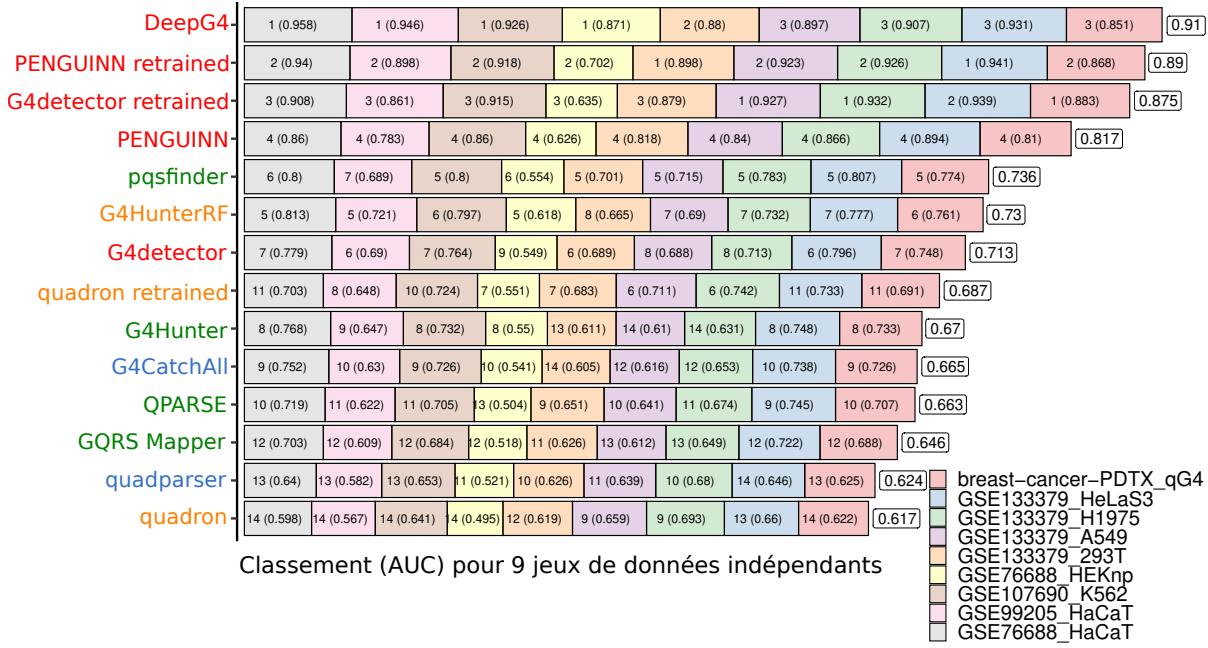


FIGURE 2.2 – Comparaison du modèle DeepG4\* (sans accessibilité) avec les algorithmes existants. Pour chaque ensemble de données, l’AUC d’un algorithme est affichée entre parenthèses avec son classement. Sur la droite, la valeur moyenne de l’AUC est tracée pour chaque algorithme. Chaque algorithme est évalué avec 9 différents jeux de données expérimentaux indépendants. Les différents seuils de détection de G4Hunter ont été utilisés comme variables à une *Random Forest* (G4HunterRF), et les modèles quadron, PENGUINN et G4detector ont été ré-entraînés sur le même jeux de données que DeepG4\* (retrained).

de 0.91. Les autres algorithmes de *Deep Learning*, PENGUINN et G4detector, sont en 2ème et 3ème positions avec une AUC moyenne de 0.89 et 0.875, respectivement, après ré-entraînement des modèles sur le jeu de données HaCaT. En utilisant les modèles d’origine, ils sont classés respectivement 4ème (AUC moyenne = 0.817) et 7ème (AUC moyenne = 0.713). L’outil Quadron, basé sur une approche de *Machine Learning*, entraîné sur des données G4 *in vitro* uniquement, produit les moins bonnes prédictions sur les données des G4 actifs (AUC moyenne = 0.617). Le ré-entraînement de Quadron a donné des résultats considérablement meilleurs que le modèle d’origine (AUC moyenne = 0.687, en 8ème position), mais qui sont quand même beaucoup moins précis que ceux de DeepG4\* et d’autres méthodes de *Deep Learning*.

Les performances sur les modèles de *Deep Learning*, avant et après ré-entraînement montrent que l’architecture et les hyper-paramètres sélectionnés sont ici moins importants que les données utilisées pendant l’entraînement. Les mauvais résultats de Quadron sur ces données nous indiquent que cet outil a été développé pour la prédiction de G4 *in vitro*, et révèle que les variables construites pour modéliser ce type de données ne sont pas suffisantes pour capturer la dynamique des G4 actifs.

## 2.1. DEEPG4

---

TABLE 2.1 – Méthodes existantes pour la prédiction des G-quadruplexes. Les méthodes étiquetées ‘regex’ sont basées sur la recherche d’une expression régulière. Les méthodes à base de score utilisent la composition de la séquence d’ADN pour prédire la position d’un G4. Quadron est une méthode de *Machine learning* utilisant à la fois la recherche du motif G4 et son contexte comme variables. Penguin et G4detector sont des méthodes de *Deep Learning* utilisant les réseaux de convolution afin de prédire un score de G4 dans des régions de 200bp.

Outil	Méthode	Implémentation	Année
<b>quadparser</b>	Regex	<b>Python</b>	2005
<b>gqrs_mapper</b>	Score	<b>Python</b>	2006
<b>G4hunter</b>	Score	<b>Python</b>	2016
<b>pqsfinder</b>	Score	<b>R</b>	2017
<b>quadron</b>	Machine Learning	<b>R xgboost</b>	2017
<b>qparse</b>	Score	<b>Python</b>	2019
<b>G4CatchAll</b>	Regex	<b>Python</b>	2019
<b>G4detector</b>	Deep Learning	<b>Python / Tensorflow</b>	2019
<b>penguinn</b>	Deep Learning	<b>Python / Tensorflow</b>	2020

Les algorithmes basés sur le calcul d’un score, notamment les outils pqsfinder, G4Hunter, QPARSE et GQRS mapper n’ont pas donné de bons résultats avec des AUC moyennes de 0.736 (classé 5), de 0.67 (classé 9), de 0.663 (classé 11) et de 0.646 respectivement (classé 12). Ces performances peuvent s’expliquer par le fait que les paramètres de ces modèles experts ne sont pas déterminés pour la prédiction de G4 actifs. Pour remédier à cela, nous avons entraîné un modèle de *Random Forest* qui se base sur les prédictions de G4Hunter en faisant varier son paramètre de seuil (1 à 2 par incrément de 0.1). De cette façon, nous pouvons déterminer quel seuil est le plus efficace pour prédire les G4 actifs par apprentissage automatique. Ce modèle propose une comparaison plus juste avec notre modèle, et a largement surpassé les prédictions de G4Hunter avec une AUC moyenne de 0.73. Ces prédictions sont cependant toujours moins bonnes que celles de DeepG4\*, et des méthodes de *Deep Learning* en général. Les algorithmes basés sur des motifs, comme quadparser et G4CatchAll ont également de mauvais résultats lors de la prédiction des G4 actifs, avec des AUC moyennes de 0.624 et 0.665.

### 2.1.1.4 Conclusion sur DeepG4\*

Ces différents résultats montrent que DeepG4\* est capable de prédire avec précision des régions enrichies en G4 actifs à partir de séquences d’ADN uniquement. Il montre également de meilleures performances que les outils existants pour prédire ce type de données, et montre l’incapacité des outils experts à prédire précisément les G4 qui se forment *in vivo*.

DeepG4\*, bien que capable de prédire très efficacement sur le même type cellulaire que celui utilisé comme jeu de données d’entraînement, montre des baisses de performances significatives lorsque l’on prédit sur d’autres lignées cellulaires. Afin de pallier ce problème, nous avons proposé dans l’article une version améliorée de ce modèle qui utilise l’accessibilité locale de la chromatine afin d’améliorer les prédictions et proposer de bons résultats sur les différents types cellulaires testés.

### 2.1.2 Discussion

#### 2.1.2.1 La prédiction de régions contenant des G4 actifs en fonction du type cellulaire

Dans ces travaux, j'ai développé un modèle de *Deep Learning* se basant sur des séquences d'ADN provenant de régions détectées par des méthodes omiques. Ces données proviennent du chevauchement entre les pics détectés *in vitro* (CHAMBERS et al. 2015) et ceux détectés *in vivo* (HÄNSEL-HERTSCH, SPIEGEL et al. 2018). Ce chevauchement permet d'associer la capacité physique de la séquence à former un G4 (*in vitro*) à un signal détecté *in vivo*. À cela, nous avons ajouté l'accessibilité locale de la chromatine, permettant à notre modèle de prédire les régions G4 actives selon le type cellulaire.

Notre méthode permet de prédire des régions pour des types cellulaires, des tissus ou des cancers dont les données expérimentales ne sont pas disponibles. Nous avons utilisé notre outil sur le génome entier dans de nombreux tissus et cancers dont les données d'accessibilité étaient disponibles, fournissant une base de données de régions potentielles utile à la communauté scientifique.

En utilisant ces données prédites, nous avons identifié des régions enrichies en G4 actives très stables entre les types de tissus et cancers, et d'autres variables. Nous avons constaté que les régions variables sont préférentiellement localisées dans les régions intergéniques ou introniques, contrairement aux régions stables qui sont préférentiellement localisées dans les gènes. Ces régions variables pourraient être enrichies dans les régions *enhancers* des gènes, et réguler à distance leurs expressions, alors que les régions stables sont plus enrichies au niveau des promoteurs et pourraient réguler leurs expressions de manière plus directe.

*DeepG4* est complémentaire aux algorithmes existants basés sur des méthodes expertes. En effet, notre modèle utilise des régions de 201 bp pour sa prédiction, alors qu'un motif G4 fait plutôt entre 20 et 30 bp. Un ou plusieurs G4 peuvent être localisés dans la séquence, sans qu'on puisse identifier leur position avec précision. Une fois que *DeepG4* a prédit les régions G4 actives, les algorithmes experts existants, comme Quadparser (HUPPERT et BALASUBRAMANIAN 2005) ou G4Hunter (BEDRAT, LACROIX et MERGNY 2016), permettent de localiser la position exacte d'une séquence, son brin d'origine, ainsi que sa proportion à former un G4.

En utilisant des régions de 201 bp, *DeepG4* permet de prendre en compte le contexte pour la prédiction de régions G4 actives. Ainsi, *DeepG4* a identifié des motifs reconnus par des facteurs de transcription (*TF*) comme prédicteurs importants de notre modèle.

*DeepG4* utilise les séquences pour construire les *features* qui lui permettent de prédire la présence d'un G4 actif. La majorité de ces *features* sont des motifs qui reconnaissent le motif G4 canonique ( $G \geq 3N_xG \geq 3N_yG \geq 3N_zG \geq 3$ ) ou bien des morceaux de ce motif, ainsi que son équivalent complémentaire. La combinaison de ces motifs reconnus et construits par notre modèle apporte la flexibilité nécessaire à la prédiction de G4 dans un contexte *in vivo*.

#### 2.1.2.2 Limitations et améliorations possibles du modèle

Notre modèle utilise l'accessibilité locale de l'ADN comme prédicteur. Pour ajouter cette information, il suffit de concaténer le signal moyen d'accessibilité des 201 bp à la couche de *global max pooling*,

ajoutant une variable à un vecteur de taille 900 représentant l'enrichissement des motifs. On encode ainsi l'accessibilité en ajoutant un minimum de complexité dans le modèle. Cependant, la qualité des données omiques nous permettant d'obtenir cette information est déterminante dans la qualité de la prédiction, et rajoute une complexité supplémentaire dans la construction des *features* du modèle, là où auparavant il suffisait d'encoder l'ADN en *one-hot*.

**2.1.2.2.1 Utilisation des données de tous les types cellulaires pendant l'entraînement** Nous pourrions envisager une approche différente en se basant sur l'apprentissage d'un modèle pour chaque type cellulaire à notre disposition, puis récupérer l'information pour construire un nouveau modèle, sans utiliser l'accessibilité locale de l'ADN.

On peut envisager pour cela une méthode d'apprentissage ensembliste (*Ensemble learning*), où l'on utilise plusieurs modèles pour obtenir de meilleures prédictions. De manière générale, les méthodes ensemblistes combinent plusieurs modèles différentes et présentent souvent une meilleure prédiction que le meilleur des prédicteurs pris individuellement. Dans le cas de l'algorithme de *Random Forest* par exemple, l'architecture des modèles est la même, mais ceux-ci sont entraînés sur des sous-échantillons des données différents.

Cette approche est applicable au *Deep Learning*, où chaque modèle peut être entraîné sur un sous-échantillon indépendant, ou sur un échantillonnage avec remise. Ceux-ci auront ainsi une prédiction et un sur-ajustement légèrement différents.

Ces prédictions sont ensuite fournies à un autre modèle (un *meta-learner*) qui combine ces prédictions afin d'obtenir une nouvelle prédiction. Ce *meta learner* peut être un simple vote majoritaire, un *GLM*, comme une régression logistique, ou un modèle de *Deep Learning* lui-même. La prédiction consensus combine les prédictions de ces modèles pour réduire la variance des prédictions et réduire l'erreur de généralisation.

Dans notre cas, on peut entraîner un modèle de *Deep Learning* par type cellulaire, et utiliser la prédiction du *meta-learner* comme prédiction sur n'importe quel type cellulaire, tissue, ou cancer. L'avantage, c'est qu'on peut renforcer notre prédiction en ajoutant celles des algorithmes experts existants, comme G4Hunter (BEDRAT, LACROIX et MERGNY 2016), ou de *Machine Learning*, comme Quadron (SAHAKYAN et al. 2017). Cependant, ce *meta-learner* doit lui aussi être entraîné, ce qui signifie qu'on doit réservé une part du jeu de données d'entraînement à ce *meta learner* ce qui nécessite d'avoir à notre disposition un grand jeu de données.

**2.1.2.2.2 Pré-remplissage des paramètres en utilisant des motifs existants** Sachant que les sites de fixation de *TF* sont utiles à la prédiction, nous pourrions utiliser des bases de données de motifs, comme *JASPAR* pour initier les paramètres du modèle. Cette méthode implique que les *PWM* soient transformées en *kernels* de la couche de convolution (DING et al. 2017). L'intérêt de cette méthode est qu'on aiguille le modèle pour qu'il utilise ces motifs pour la prédiction sans avoir à les construire lui-même. On peut fixer ses paramètres afin que le modèle ne puisse pas les modifier pendant l'optimisation, et ajouter des paramètres supplémentaires qui eux peuvent être appris. On peut également permettre la modification de ces paramètres pendant l'entraînement, afin que l'optimiseur puisse les favoriser ou les dé-favoriser, en fonction de la prédiction. Dans tous les cas, cela pourrait aider la prédiction dans le cas

où le contexte d'un G4 actif est différent des données utilisées pendant l'entraînement, lorsque le type cellulaire ou le tissu est différent par exemple.

#### 2.1.2.2.3 Alternatives à la re-construction des *PWM* à partir des paramètres du modèle

Nous avons reconstruit les motifs en se basant sur la méthode utilisée par *DeepBind* (ALIPANAHY et al. 2015), en partant de la position de l'activation maximale de chacun de nos 900 *kernels* dans nos données pour reconstruire les *PWM*. Ensuite, nous utilisons une méthode de *clustering* de matrices pour réduire leur redondance. Enfin, l'importance de ces motifs *clusters* est calculé en les utilisant comme variables d'un algorithme de *Random Forest*.

Cependant, il existe d'autres méthodes qui utilisent des approches différentes, comme les scores d'importance basés sur les gradients (SHRIKUMAR, GREENSIDE et KUNDAJE 2017a) pour faire de la découverte de motifs *de novo* (SHRIKUMAR, TIAN et al. 2018). Ces méthodes permettent de diminuer la redondance des motifs découverts, et de calculer leurs importances directement, sans avoir besoin d'utiliser un autre modèle. Utiliser cette méthode permettrait d'avoir une vision plus directe de l'importance des couches de notre modèle, plutôt qu'une importance calculée sur des *PWM* reconstruites à partir de la couche de convolution uniquement.

#### 2.1.2.2.4 Prédiction de régions de G4 actives avec *DeepG4*

Notre modèle ne permet également pas de prédire la position exacte d'un G4 actif, mais plutôt une région, pouvant contenir un ou plusieurs G4 actifs. Cette position exacte ne peut être pour l'instant calculée qu'à l'aide d'algorithmes experts existants, comme G4Hunter (BEDRAT, LACROIX et MERGNY 2016). Cependant, nous pouvons déjà déterminer, pour chaque *kernel*, la position d'activation maximale sur la séquence d'entrée à partir de la *response map* (voir partie 1.4.3.6).

En identifiant les *kernels* qui ressemblent le plus à un motif canonique G4, nous pouvons détecter la position exacte de son activation sur la séquence, et de la même manière que pour reconstruire des *PWM* à partir des *kernels* (voir partie 1.4.3.6), inférer sur la position du G4. Une approche différente, serait d'ajouter une couche de sortie correspondant à la position exacte d'un G4 prédict, en s'inspirant des méthodes de *peak calling* en *Deep Learning* (OH et al. 2020). Le principe est similaire à la segmentation utilisée pour l'identification d'objets dans des images. Le principe de la segmentation et d'étiqueter chaque pixel ou position par la classe correspondante. En pratique, on encode un vecteur de la même taille que la séquence d'entrée, et on indique pour chaque position à quelle classe elle appartient. C'est ce vecteur que le modèle prédit pendant l'entraînement. Dans notre cas, la classe "G4" peut correspondre à la position d'un pic provenant de données de séquençage, ou bien de la position d'un motif canonique par des prédictions bio-informatiques.

#### 2.1.2.2.5 Utilisation du brin complémentaire dans la prédiction

En outre, notre modèle ne prend pas en compte l'origine du brin, ce qui fait qu'un motif encodé par les paramètres de la couche de convolution peut être présent de manière redondante. Nous avons résolu en parti ce problème par une méthode de *clustering* de motifs, qui construit des groupes par similarité, en prenant en compte la complémentarité, puis détermine un motif consensus ou racine par *cluster*. On peut cependant prendre en compte l'information du brin pendant l'entraînement, de différentes façons.

Il est possible d'identifier l'origine du brin en utilisant les données *in vitro* qui sont brin-spécifique, puis utiliser la séquence d'ADN complémentaire lorsque le G4 est identifié sur ce brin. Cependant, les motifs complémentaires ne seront plus appris, et cela implique que la séquence d'ADN complémentaire devra être fournie à chaque prédiction. On peut également configurer le modèle pour qu'il puisse prédire le brin probable de formation du ou des G4 dans la région. Ces méthodes impliquent cependant de supprimer les régions ou des G4 sont présents à la fois sur le brin sens et sur le brin anti-sens, ce qui diminue le nombre de données d'entraînement.

Pour considérer l'information contenue dans le brin complémentaire de manière systématique pendant l'entraînement, des couches convolutives spécifiques ont été développées (SHRIKUMAR, GREENSIDE et KUNDAJE 2017b). Ces couches exploitent la propriété de complémentarité de l'ADN en partageant les paramètres entre les représentations sens/anti-sens des séquences, pour obtenir une prédiction identique peu importe le brin utilisé. Leur utilisation dans notre modèle nous permettrait de considérer l'information du brin, sans perte de données ni redondance.

**2.1.2.2.6 Alternatives à la couche de convolution** Enfin, il est également possible de considérer l'utilisation d'autres couches de *Deep Learning* en plus, ou à la place de la convolution. Par exemple, les modèles récents développés pour traiter des problèmes de traitement automatique des langues ou *Natural Language Processing (NLP)* utilisent des couches *Transformer* basées sur des mécanismes d'*Attention* (VASWANI et al. 2017) et de *self attention*.

Avant l'utilisation des *Transformer*, les modèles utilisés pour des tâches de traduction étaient constitués d'une structure *Encoder-Decoder* utilisée avec des couches récurrentes, comme les couches *LSTM* (HOCHREITER et SCHMIDHUBER 1997). Cette structure encode la séquence de la langue à traduire dans une représentation vectorielle, puis la décode à nouveau dans la langue traduite. Le *Transformer* est basé sur des principes similaires d'encodage/décodage mais ajoute un mécanisme d'*Attention*. L'*Attention* mesure à quel point les éléments de deux séquences sont liés, et transmet l'information au reste du modèle. La *self attention* permet de faire la même chose, mais sur une seule séquence, et détermine l'inter-dépendance des éléments au sein de la séquence afin de proposer un *embedding*, une projection pertinente. L'architecture *BERT* de Google (DEVLIN et al. 2018) utilise ces *Transformer* pour résoudre des tâches de *NLP*.

*DNABERT* (JI et al. 2021) implémente cette structure pour les séquences d'ADN, et permet de déterminer des relations sémantiques au sein des séquences. Cependant, *DNABERT* fonctionne avec une approche différente de la notre et encode l'ADN en *k-mers*, ce qui consiste à diviser une séquence en sous séquences de taille *k* avec *k* – 1 mots en commun. Cette approche nécessite l'utilisation d'*embedding* ou couches de plongements (ou projection vectorielle), qui complexifient l'interprétation. Dans *DNABERT* Le mécanisme de *self attention* est utilisé pour déterminer les régions d'importances au sein des séquences d'entrée, à la résolution du nucléotide. On peut extraire les sous-séquences présentant un fort niveau d'attention, et ainsi reconstruire les motifs sous forme de *PWM*.

*DNABERT* propose une alternative solide aux réseaux de convolution pour les séquences biologiques. Implémenter ce type d'architecture nous permettrait d'utiliser les relations sémantiques entre motifs pour la prédiction.

### 2.1.2.3 Applications de DeepG4

La présence des G-quadruplexes dans les promoteurs des oncogènes humains en font des cibles thérapeutiques de choix pour réguler la transcription dans les cellules cancéreuses (KOSIOL et al. 2021). En effet, les G4 sont connus pour être enrichis au niveau des promoteurs de certains oncogènes, dont *MYC*, et des études *in vitro* ont révélé que leur déstabilisation est corrélée avec une réduction de l'expression de ces gènes (SIDDQUI-JAIN et al. 2002) (voir Figure 2.3).

Enfin, les structures G4 peuvent impacter la stabilité du génome, en provoquant des dommages à l'ADN comme les cassures double-brin (*DSB*). Cette instabilité génomique accrue est déterminante dans la thérapie contre le cancer, car elle stimule d'une part la formation de tumeurs, mais peut également être utilisée comme approche thérapeutique, lorsque l'instabilité est telle qu'elle conduit à l'apoptose des cellules cancéreuses (voir Figure 2.3).

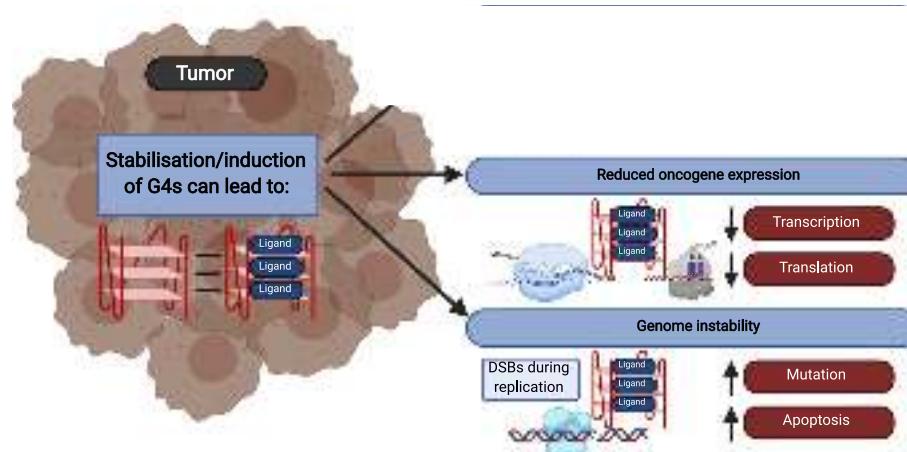


FIGURE 2.3 – Les effets potentiels des ligands de G4 sur les cellules cancéreuses. L'action des ligands sur les G4 enrichis sur les promoteurs des oncogènes pourrait permettre de réguler négativement leurs expressions ainsi que d'augmenter l'instabilité du génome et conduire à l'apoptose (KOSIOL et al. 2021).

Les approches actuelles ont permis de montrer qu'il existait plus de 700 000 régions génomiques potentielles pouvant former des G4 (CHAMBERS et al. 2015; HUPPERT et BALASUBRAMANIAN 2005). Cependant, un motif G4 indique uniquement le potentiel de l'ADN à former un G4 et ne garantit pas sa présence *in vivo*. L'application d'approches prédictives basées sur des données *in vivo* pourrait être déterminante dans le choix de cibles thérapeutiques potentielles basées sur des ligands de G4.

## 2.2 Réparation de l'ADN et structure tri-dimensionnelle du génome

### 2.2.1 Le modèle cellulaire DIvA

Les expériences réalisées dans les articles suivants ont été produits dans la lignée cellulaire *DIvA*, pour *DSB induced by ASI* (voir Figure 2.4). Ce modèle cellulaire a été développé dans notre équipe et se base sur des cellules humaines U2OS modifiées pour exprimer une enzyme de restriction couplée à une

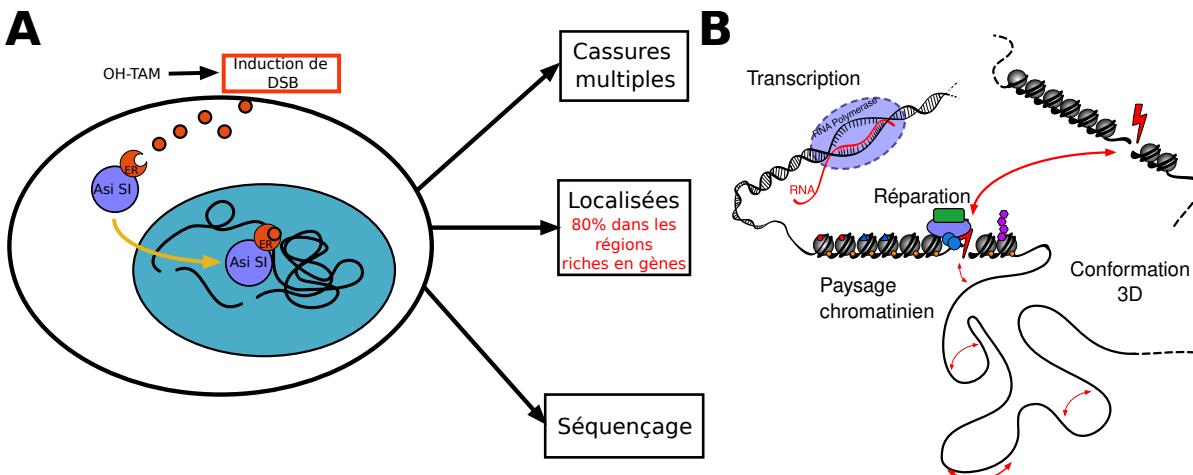


FIGURE 2.4 – Le modèle cellulaire *DIvA* permet d’induire de manière localisée une centaine de cassures sur le génome. **A** Ces cassures sont induites par la relocalisation d’une protéine de fusion dans le noyau après l’ajout d’une drogue 4-Hydroxy-Tamoxifen (4OHT). La protéine de fusion est composée d’une enzyme de restriction, AsiSI, qui provoque une centaine de cassures double-brin (DSB) sur le génome, préférentiellement dans les régions riches en gènes. **B** Grâce au modèle *DIvA*, on peut étudier l’ensemble du paysage chromatinien, son accessibilité, sa conformation 3D ainsi que la transcription à proximité des DSB, avant et après dommages.

version modifiée du récepteur aux œstrogènes qui se lie au 4-Hydroxy-Tamoxifen (4OHT) (IACOVONI et al. 2010). Ce modèle biologique nous permet d’induire des cassures double-brin (*DSB*) multiples (< 100) à différents endroits du génome. L’avantage de ce modèle par rapport à des méthodes telles que l’irradiation est qu’il permet d’induire des cassures à des positions connues sur l’ensemble du génome en utilisant l’enzyme de restriction AsiSI, de manière contrôlée. Les positions des *DSB* induites par *DIvA* étant connues, notre équipe a générée une grande quantité de données omiques afin d’étudier l’ensemble du paysage chromatinien (ChIP-seq), son accessibilité (ATAC-seq), sa conformation 3D (Hi-C et 4C-seq) ainsi que la transcription (RNA-seq) à proximité des *DSB*, avant et après dommages. La technique BLESS (Breaks Labeling, Enrichment on Streptavidin and next-generation Sequencing) (CROSETTO et al. 2013) a été utilisée afin de déterminer l’efficacité de clivage des *DSB* et d’identifier les sites les plus endommagés du génome sur les 1211<sup>3</sup> reconnus par AsiSI.

## 2.2.2 Introduction

Les cassures double-brin (*DSB*) de l’ADN sont des lésions très délétères qui peuvent survenir suite à de nombreux stress, endogènes comme exogènes. Leur réparation est donc essentielle pour préserver l’intégrité du génome. Suite à ces dommages, de larges domaines de l’ordre du Mégabase sont enrichis en variants d’histones H2AX phosphorylés par la kinase ATM. Les mécanismes qui permettent de former de tels foyers de réparation sont encore très mal compris. Les acteurs à l’origine de la formation de grands domaines 3D, les *TAD* pour *Topologically Associating Domains* ont très probablement un rôle dans la formation et la signalisation des *DSB* sur le génome, mais celui-ci n’est pas encore connu. Pourtant, il a été montré que les protéines responsables de l’architecture 3D du génome, comme les cohésines, sont recrutées au niveau des *DSB* (CARON, Francois AYMARD et al. 2012).

3. 1211 sites reconnus par AsiSI sur la version hg19 du génome de référence.

### 2.2.3 La formation des foyers de réparation autour des cassures double-brin de l’ADN dépend du mécanisme de *loop extrusion*

Dans cet article, nous avons étudié le rôle des *TAD* et des acteurs de leur formation, CTCF et la cohésine dans l’établissement des foyers de réparation. Pour cela, nous avons réalisé des expériences omiques de Hi-C à très haute résolution (~500 millions de *reads* par expérience). Des expériences de 4C-seq, centrés sur les *DSB*, ainsi que de ChIP-seq ciblant les protéines architecturales ont également été produites au sein de notre équipe.

Avec ces données, nous avons démontré que les *TAD* sont des unités fonctionnelles indispensables de la réparation qui délimitent la propagation de  $\gamma$ H2AX. Nous avons également montré que le processus déterminant de la formation des *TAD* dépendant de la cohésine, la *loop extrusion*, est également à l’origine de la formation du foyer de réparation. La *loop extrusion* permet, de manière bidirectionnelle, à la kinase ATM de phosphoryler H2AX sur un *TAD* entier de façon très rapide. Nous avons également montré une augmentation du recrutement des cohésines au sein des domaines endommagés, probablement pour renforcer le processus de *loop extrusion*.

# Loop extrusion as a mechanism for formation of DNA damage repair foci

<https://doi.org/10.1038/s41586-021-03193-z>

Received: 7 February 2020

Accepted: 6 January 2021

Published online: 17 February 2021

 Check for updates

Coline Arnould<sup>1</sup>, Vincent Rocher<sup>1</sup>, Anne-Laure Finoux<sup>1</sup>, Thomas Clouaire<sup>1</sup>, Kevin Li<sup>2</sup>, Felix Zhou<sup>2</sup>, Pierre Caron<sup>1</sup>, Philippe E. Mangeot<sup>3</sup>, Emiliano P. Ricci<sup>4</sup>, Raphaël Mourad<sup>1</sup>, James E. Haber<sup>2</sup>, Daan Noordermeer<sup>5</sup> & Gaëlle Legube<sup>1</sup>✉

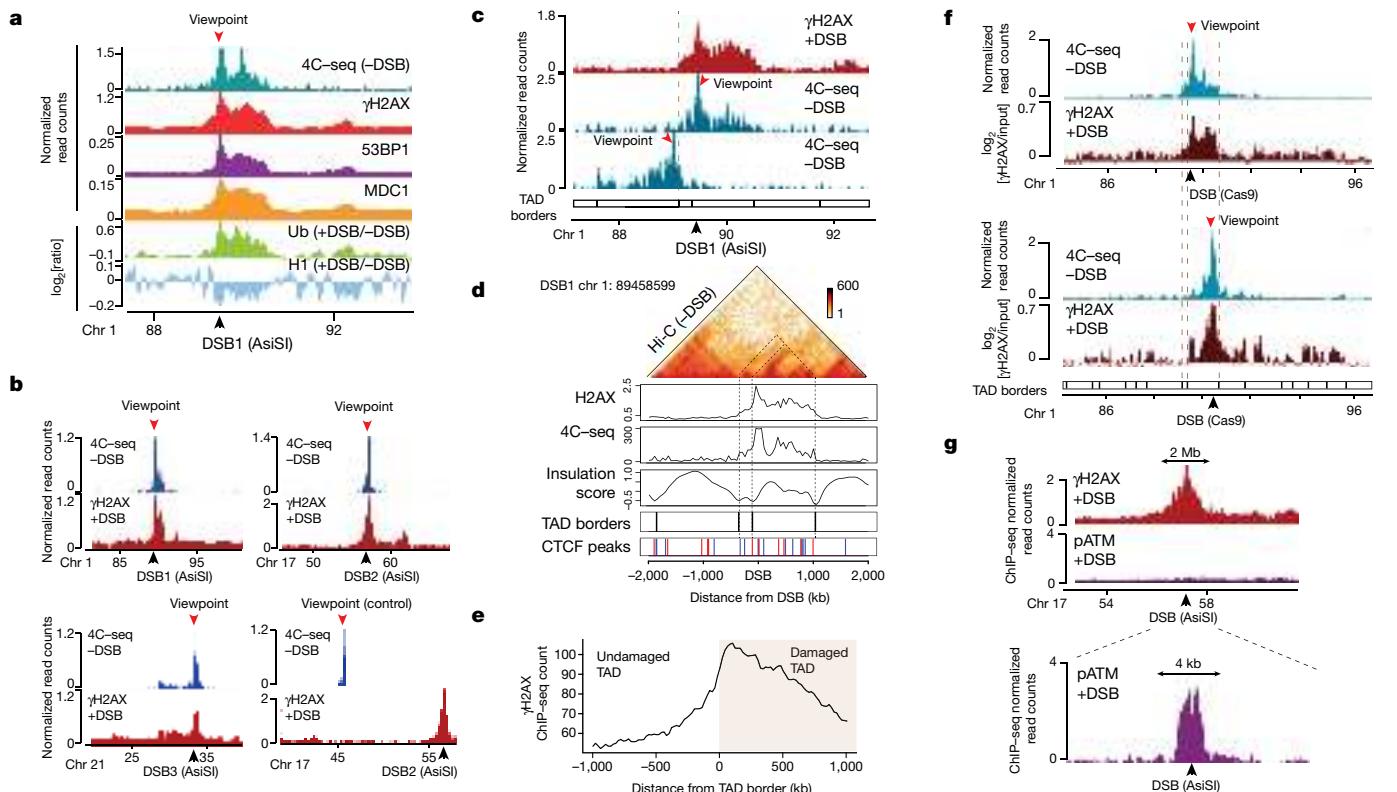
The repair of DNA double-strand breaks (DSBs) is essential for safeguarding genome integrity. When a DSB forms, the PI3K-related ATM kinase rapidly triggers the establishment of megabase-sized, chromatin domains decorated with phosphorylated histone H2AX ( $\gamma$ H2AX), which act as seeds for the formation of DNA-damage response foci<sup>1</sup>. It is unclear how these foci are rapidly assembled to establish a ‘repair-prone’ environment within the nucleus. Topologically associating domains are a key feature of 3D genome organization that compartmentalize transcription and replication, but little is known about their contribution to DNA repair processes<sup>2,3</sup>. Here we show that topologically associating domains are functional units of the DNA damage response, and are instrumental for the correct establishment of  $\gamma$ H2AX–53BP1 chromatin domains in a manner that involves one-sided cohesin-mediated loop extrusion on both sides of the DSB. We propose a model in which H2AX-containing nucleosomes are rapidly phosphorylated as they actively pass by DSB-anchored cohesin. Our work highlights the importance of chromosome conformation in the maintenance of genome integrity and demonstrates the establishment of a chromatin modification by loop extrusion.

DNA DSBs induce the formation of DNA-damage response (DDR) foci, which are microscopically visible and characterized by specific chromatin modifications ( $\gamma$ H2AX, ubiquitin accumulation and histone H1 depletion) and the accumulation of DDR factors (53BP1 and MDC1)<sup>4–6</sup>. Previous evidence indicated that chromosome architecture may control the spread of  $\gamma$ H2AX. Indeed,  $\gamma$ H2AX domain boundaries were found in some instances to coincide with topologically associating domain (TAD) boundaries<sup>7</sup>. Moreover, super-resolution light microscopy revealed that CTCF, which binds at TAD boundaries and thereby constrains the loop-extruding activity of the cohesin complex that shapes these domains in undamaged cells, is juxtaposed to  $\gamma$ H2AX foci<sup>8</sup>. In addition, 53BP1 can form nanodomains that frequently overlap with TADs, as detected by DNA fluorescence *in situ* hybridization (DNA-FISH)<sup>9</sup>. High-resolution chromatin immunoprecipitation with sequencing (ChIP-seq) mapping after the induction of multiple DSBs at annotated positions (using human DlVA (DSB inducible via AsISI) cells)<sup>10</sup> revealed that the spreading of these DDR focus components on nearby chromatin follows a highly stereotyped pattern<sup>5</sup> (one example shown in Fig. 1a). We hypothesized that such patterns could be governed by pre-existing high-order chromatin structure established before DSB induction.

## $\gamma$ H2AX spreads within TADs

To relate the spreading of DDR focus components to chromosome conformation, we performed circular chromosome conformation capture coupled to high-throughput sequencing (4C-seq) experiments in undamaged human DlVA cells. As viewpoints we selected three genomic locations that are damaged in DlVA cells following activation of the AsISI restriction enzyme as well as one undamaged control region. The chromatin conformation around these three viewpoints in undamaged condition was notably similar to the distribution of  $\gamma$ H2AX determined post DSB induction (Fig. 1a, b, Extended Data Fig. 1a), suggesting that initial chromosome architecture dictates  $\gamma$ H2AX spreading and downstream events such as accumulation of MDC1, ubiquitin and 53BP1 following DSB. To prove that DDR domains do not spread into neighbouring self-interacting domains, we focused on a DSB located on chr1, for which spreading of DDR foci components is profoundly asymmetrical (Fig. 1c, red track). 4C-seq performed at two viewpoints separated by 470 kb revealed the existence of two adjacent self-interacting domains with a boundary corresponding to the abrupt drop in  $\gamma$ H2AX (Fig. 1c, blue track; TAD boundary is indicated by the dotted line). This strongly suggests that pre-existing chromatin domains, established before any damage occurs, constrain the spread of DDR foci.

<sup>1</sup>Molecular, Cellular and Developmental Biology Unit (MCD), Centre de Biologie Intégrative (CBI), UPS, CNRS, Toulouse, France. <sup>2</sup>Rosenstiel Basic Medical Sciences Research Center and Department of Biology, Brandeis University, Waltham, MA, USA. <sup>3</sup>CIRI – International Center for Infectiology Research, Inserm U1111, Université Claude Bernard Lyon 1, CNRS, UMR5308, Ecole Normale Supérieure de Lyon, University of Lyon, Lyon, France. <sup>4</sup>Laboratoire de Biologie et Modélisation de la Cellule, Université de Lyon, INSERM U1293, CNRS UMR 5239, Ecole Normale Supérieure de Lyon, Université Claude Bernard Lyon 1, Lyon, France. <sup>5</sup>Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France.  
✉ e-mail: gaelle.legube@univ-tlse3.fr



**Fig. 1 | TADs are functional units that govern the establishment of DDR chromatin domains.** **a**, 4C-seq track in undamaged cells (−DSB) and ChIP-seq tracks of histone H1 (H1.2) and ubiquitin (Ub; FK2) ( $\log_2(+\text{DSB}/-\text{DSB})$ ) as well as  $\gamma$ H2AX, MDC1 and 53BP1 (+DSB) as indicated. ChIP-seq and 4C-seq data were smoothed using 50-kb and 10-kb spans, respectively. **b**, 4C-seq tracks before DSB induction (−DSB) and  $\gamma$ H2AX ChIP-seq tracks after DSB induction (+DSB) (smoothed using a 50-kb span) for viewpoints located at three AsiSI sites or a control region. One representative experiment is shown (out of  $n=3$ ). **c**,  $\gamma$ H2AX ChIP-seq (+DSB) and 4C-seq (−DSB) tracks (10-kb smoothed) for viewpoints at the AsiSI site or 470 kb upstream of the AsiSI site. **d**, Top, Hi-C contact matrix of a region of chromosome 1 in DlVA cells before DSB induction. One

representative experiment is shown (out of  $n=2$ ). Below,  $\gamma$ H2AX ChIP-seq after DSB induction, 4C-seq signal, insulation scores, TAD borders computed from Hi-C data and CTCF ChIP-seq peaks before DSB induction. Peaks in blue and red contain CTCF motifs in the forward and reverse orientations, respectively. **e**, Average profile of  $\gamma$ H2AX ChIP-seq after DSB induction centred on the closest TAD border to the 174 best-induced DSBs (damaged TAD on the right). **f**, Blue, 4C-seq track (10-kb smoothed) before DSB induction (−DSB) using viewpoints as indicated. Red,  $\gamma$ H2AX ChIP-chip tracks ( $\log_2[\text{sample}/\text{input}]$ , smoothed using 500-probe span) after DSB induction with CRISPR–Cas9. **g**,  $\gamma$ H2AX and pATM (S1981) ChIP-seq ( $n=1$ ) tracks after DSB induction on an 8-Mb window (top) and a 15-kb window (bottom) around an AsiSI site.

To generalize this finding, we performed high-throughput chromosome conformation capture (Hi-C) and CTCF ChIP-seq in undamaged DlVA cells (Extended Data Fig. 1b-d). Notably, computed TAD borders and CTCF-bound genomic loci coincided with a sharp decrease in  $\gamma$ H2AX signals (Fig. 1d, e, Extended Data Fig. 1e). Consistent with this,  $\gamma$ H2AX, MDC1 and 53BP1 were substantially more enriched in the damaged TADs than in neighbouring TADs (Extended Data Fig. 1f), although spreading through boundaries was observed to some extent, in agreement with the moderate insulation properties of TAD boundaries<sup>11</sup>.

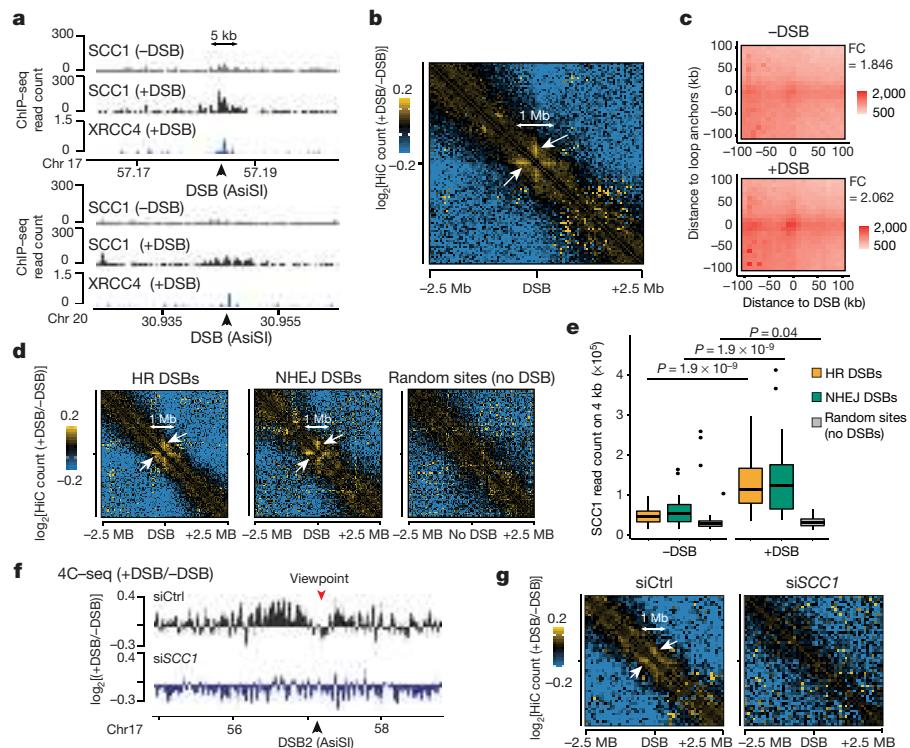
To further investigate whether TADs dictate  $\gamma$ H2AX spreading, we used the CRISPR–Cas9 system to induce a single DSB at designated positions within the same TAD, and investigated both chromosome conformation and  $\gamma$ H2AX distribution. Cas9-induced DSBs recapitulated the  $\gamma$ H2AX spreading observed when DSBs were induced at the same genomic locations by AsiSI (Extended Data Fig. 1g), thus confirming that  $\gamma$ H2AX spreading is independent of the method of DSB induction. Moving the DSB to a further downstream position in the TAD triggered a change in the  $\gamma$ H2AX profile that was notably similar to the 3D interaction pattern of this genomic region, but it remained constrained within the same TAD (Fig. 1f). Together, these data indicate that the mechanisms that govern the spatial organization of chromosomes into self-interacting domains facilitate and demarcate the formation of  $\gamma$ H2AX domains. Given that  $\gamma$ H2AX seeds further signalling events that lead to the stable assembly of DDR foci,

this suggests that genome organization within TADs is critical for the response to DNA damage.

In human cells, ATM is the main DDR kinase that catalyses H2AX phosphorylation upon DSB detection, as indicated by a strong decrease in  $\gamma$ H2AX upon inhibition of ATM<sup>12</sup> (Extended Data Fig. 1h-j) but not of DNAPK<sup>12</sup> or ATR (Extended Data Fig. 1i, j). To gain more insights into the mechanism that mediates the establishment of  $\gamma$ H2AX on entire self-interacting domains, we further profiled ATM. Binding of activated ATM (autophosphorylated on S1981) was restricted to the immediate vicinity of the DSB (less than 5-kb span), in sharp contrast to the pattern observed for  $\gamma$ H2AX (Fig. 1g, Extended Data Fig. 1k). This indicates that phosphorylation of H2AX is not mediated by the linear spreading of the kinase on entire TADs.

## Cohesin-mediated loop extrusion at DSBs

The organization of the genome into TADs is driven by the activity of cohesin<sup>13,14</sup>, a ring-shaped protein complex, which was initially identified for its essential role in sister chromatid cohesion. Notably, there is strong evidence that cohesin helps to maintain genome integrity<sup>15,16</sup>, and cohesin accumulates at sites of damage, which may be consistent with a role in sister chromatid cohesion during homologous recombination in S/G2 phase cells<sup>17–20</sup>. However, cohesin enrichment at DSBs has been identified throughout the cell cycle, which argues against



**Fig. 2 | DSB-anchored cohesin mediates loop extrusion.** **a**, Genomic tracks of SCC1 and XRCC4 ChIP-seq at two DSBs. **b**, Averaged Hi-C contact matrix of  $\log_2[+DSB/-DSB]$  ( $n = 2$  biological replicates) centred on the 80 best-induced DSBs (50-kb resolution, 5-Mb window). White arrows, stripes. **c**, Mean aggregate peak analysis (APA) plotted on a 200-kb window (10-kb resolution) before and after DSB induction, calculated between the DSBs and nearby loop anchors ( $n = 525$  pairs). The fold-change (FC) between the signal (central pixel) and the background (upper left corner  $5 \times 5$  pixels) is indicated. **d**, Averaged differential Hi-C contact matrix (+DSB/-DSB) ( $n = 2$  biological replicates) around 30 homologous recombination-repaired DSBs, 30 NHEJ-repaired DSBs

and 30 random undamaged sites. **e**, Box plot of the SCC1 ChIP-seq enrichment before and after DSB on 4 kb around DSBs repaired by homologous recombination (yellow) or NHEJ (green) and random undamaged sites (grey) ( $n = 30$ ). Paired two-sided Wilcoxon test. Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers. **f**, Differential 4C-seq track in control (black) or *SCC1* siRNA condition (blue) (a representative experiment is shown from  $n = 2$ ). **g**, Averaged  $\log_2[+DSB/-DSB]$  Hi-C matrix upon control or *SCC1* siRNA, around 80 best-induced DSBs (100-kb resolution) ( $n = 1$ ).

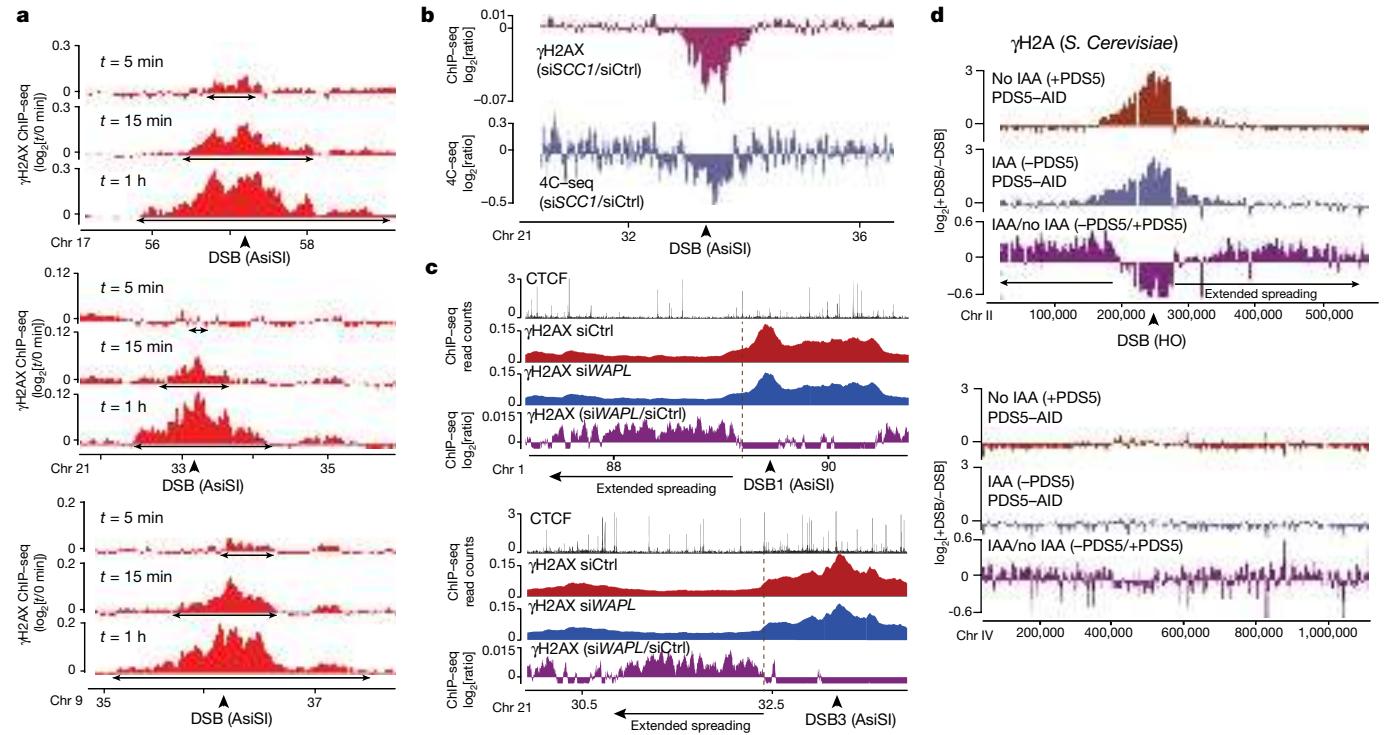
an exclusive role for cohesin in homologous recombination<sup>7,16</sup>. To get insights into cohesin binding at DSBs at high resolution, we performed calibrated ChIP-seq profiling of the *SCC1* cohesin subunit in both undamaged and damaged conditions. Notably, cohesin was enriched at sites of damage spanning 2–5 kb around the DSB (Fig. 2a), leading to the formation of peaks at DSB sites that were nearly as high as pre-existing cohesin peaks at CTCF binding sites (Extended Data Fig. 2a, b). This enrichment depended on the cohesin loader NIPBL, on ATM activity and on the MRN complex subunit MRE11 (Extended Data Fig. 2c).

Cohesins structure TADs by an active, ATP-dependent, loop extrusion mechanism<sup>21–24</sup>. Once loaded onto chromatin, cohesin leads to the formation and enlargement of DNA loops that are eventually arrested at boundary elements. A large fraction of boundary elements is bound by the CTCF insulator protein. Increased cohesin around DSBs could thus indicate locally increased loop extrusion at the site of damage. We analysed 3D genome organization by Hi-C before and after DSB induction in DlVA cells, focusing on the frequency of *cis* interactions around DSBs. Differential (+DSB/-DSB) aggregate Hi-C maps were further computed around DSBs and around TAD borders as a control (Extended Data Fig. 2d). Notably, a pattern of ‘stripes’ appeared on both sides of the DSBs following DSB induction (Fig. 2b (white arrows), Extended Data Fig. 2d, e). These stripes or lines were previously reported to arise from arrested loop extrusion at CTCF-bound loci<sup>22,24–27</sup>. Indeed, our averaged Hi-C contact matrixes around TAD borders revealed, as expected, similar stripes, but these were independent

of DSB induction (Extended Data Fig. 2d). We further performed aggregate plot analysis (APA) to assess looping between the DSB position and neighbouring anchors. Notably, the APA score increased following production of DSBs (Fig. 2c, Extended Data Fig. 2f) indicating that the DSBs themselves display the potential to arrest loop extrusion, although to a lesser extent than classical loop anchors (CTCF-bound loci) (Extended Data Fig. 2g).

It was previously determined which repair pathway (that is, homologous recombination or non-homologous end joining (NHEJ)) is preferentially used at different DSBs induced by AsiSI in DlVA cells<sup>28</sup>. Notably, an equivalent stripe pattern was observed at DSBs repaired by either homologous recombination or NHEJ (Fig. 2d). Consistent with these data, SCC1 accumulates in a 4-kb window around DSBs irrespective of the pathway used for repair (Fig. 2e). Together, these data suggest that cohesin accumulates on either side of a DSB, irrespective of the pathway used for repair, to induce divergent one-sided loop extrusion towards (and thereby to increase contacts with) the surrounding regions on both sides of the break.

To further investigate DSB-anchored loop extrusion, we performed 4C-seq before and after DSB induction, using viewpoints located at the exact positions of three DSBs induced in DlVA cells (same viewpoints as in Fig. 1). Notably, the overall structure and boundaries of TADs were well-maintained after DSB induction (Extended Data Fig. 3a), indicating that chromosome conformation within TADs is not completely reshuffled upon damage induction. Yet, as expected from Hi-C data, we detected increased interactions between viewpoints and surrounding



**Fig. 3 | DSB-anchored loop extrusion mediates γH2AX spreading.** **a**, γH2AX ChIP-seq tracks at three DSB sites upon DSB induction at different time points after release of ATM inhibition (ATMi) (expressed as  $\log_2[+DSB + ATMi + time after washes/+DSB + ATMi + 0 \text{ min after washes}]$ ) (20-kb smoothed,  $n=1$ ). **b**, Top, genomic track showing differential ( $\log_2[\text{si}SCCI/\text{siCtrl}]$ ) γH2AX enrichment obtained after DSB induction (20-kb smoothed). Bottom, differential 4C-seq signal obtained in SCC1-depleted versus control cells before DSB induction ( $\log_2[\text{si}SCCI/\text{siCtrl}]$ ) ( $n=1$ ). **c**, Genomic tracks showing the CTCF signal before DSB induction, the γH2AX ChIP-seq signal after DSB

induction in control or WAPL-depleted cells and the differential γH2AX signal obtained after DSB induction (expressed as  $\log_2[\text{si}WAPL/\text{siCtrl}]$ , 20-kb smoothed) at two DSB sites ( $n=1$ ). **d**, Genomic tracks showing the differential γH2A ChIP-seq signal ( $\log_2[+DSB/-DSB]$ ) before or after PDS5 degradation using auxin (indole-3-acetic acid (IAA)) at one DSB site (HO site) (top) and in a control region (without DSB) (bottom) in *S. cerevisiae* expressing PDS5 fused to an auxin-inducible degron (PDS5-AID). The differential signal between after and before PDS5 degradation (IAA/no IAA) is also shown (purple) ( $n=1$ ). Data are smoothed with a 2-kb span.

loci after DSB induction (Extended Data Fig. 3b–d), which was not the case when using a control undamaged sequence as a viewpoint (Extended Data Fig. 3c, d). If DSB-anchored, cohesin-mediated loop extrusion is responsible for the enhanced interaction frequency of the DSB with neighbouring sequences after DSB induction, such behaviour should be abolished following cohesin depletion. Indeed, 4C-seq experiments revealed that depletion of SCC1 by short interfering RNA (siRNA) (Extended Data Fig. 3e, f) strongly impaired the overall increase in contacts between the DSBs and their neighbouring sequences in damaged TADs (Fig. 2f, Extended Data Fig. 3g, h). We further performed Hi-C in damaged and undamaged conditions following depletion of SCC1. As expected from previous studies<sup>14,29</sup>, depletion of SCC1 led to the dissolution of TADs and to stronger compartmentalization (plaid pattern) on Hi-C maps (Extended Data Fig. 4a). Notably, depletion of SCC1 abolished the stripe pattern induced at DSBs following damage (Fig. 2g). Given that ATM is involved in recruitment of SCC1 at DSBs (Extended Data Fig. 2c), we used 4C-seq to assess the consequences of pharmaceutical inhibition of ATM kinase activity on the interaction frequency after DSB induction. ATM inhibition strongly reduced the ability of the DSB to engage contacts with proximal sequences within damaged TADs (Extended Data Fig. 4b, c), consistent with defective SCC1 recruitment at DSBs under these conditions (Extended Data Fig. 2c).

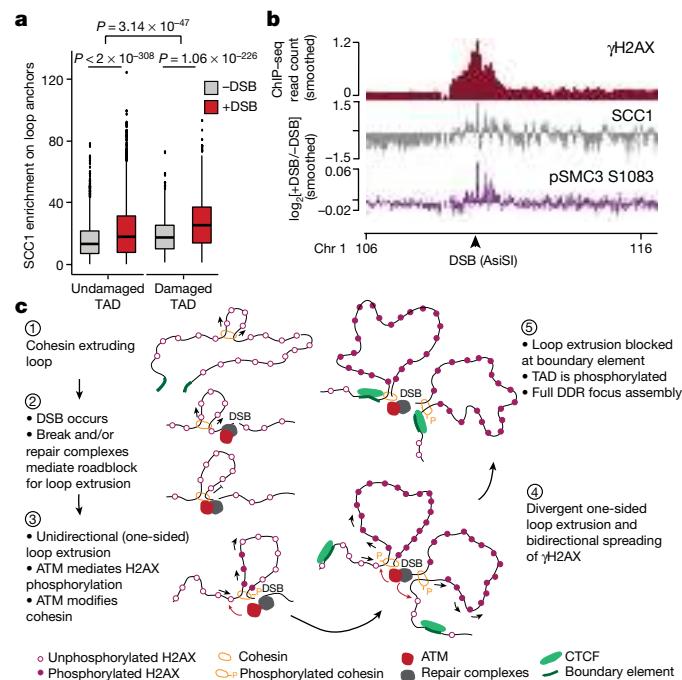
These data indicate that the ability of the DSB to contact neighbouring loci within the damaged TAD is a proper DNA damage response and cannot be explained solely by physical disruption of the DNA. It depends on ATM activity and on the cohesin complex, in agreement with a DSB-anchored loop extrusion mechanism.

## Loop extrusion in γH2AX domain formation

We further investigated whether cohesin-mediated loop extrusion that takes place at DSBs is instrumental for deposition of γH2AX. In this scenario, γH2AX should spread linearly from the DSB site over time. To achieve high synchronization of γH2AX deposition within the cell population, we induced DSBs (by OHT treatment) but concomitantly inhibited ATM activity (using an ATM inhibitor), thereby ‘poisoning’ γH2AX establishment. Relieving ATM inhibition allowed fast and synchronous accumulation of γH2AX (Extended Data Fig. 5a). Using ChIP-seq with this experimental setup, we observed linear and bidirectional spreading of γH2AX from the DSBs that proceeded at a speed of approximately 0.6 kb s<sup>-1</sup>, consistent with a loop-extrusion-dependent mechanism<sup>21,23</sup> (Fig. 3a, Extended Data Fig. 5b).

To investigate whether cohesin-mediated loop extrusion contributes to the formation of DDR foci, we analysed γH2AX profiles in SCC1-deficient cells. Both ChIP with microarray (ChIP-chip)<sup>7</sup> and ChIP-seq showed altered γH2AX spreading in SCC1-deficient cells compared to SCC1-proficient cells (Fig. 3b, Extended Data Fig. 5c, d) that coincided with a loss of *cis* contacts upon cohesin depletion (Fig. 3b, Extended Data Fig. 5c). Of note, the decrease in γH2AX in cohesin-depleted cells was small (about 5–10%) compared to the decrease in 4C-seq signal (30%), which may indicate that other factors (for example, SMCS5/6) could contribute to loop extrusion-mediated γH2AX establishment and/or that intra-TAD chromatin dynamics contribute to γH2AX deposition.

Cohesin is released from chromatin by the accessory WAPL and PDS5 factors. Consequently, depletion of these factors triggers an



**Fig. 4 | DSBs trigger modifications of cohesin biology at a genome-wide scale, accentuated in damaged TADs.** **a**, Quantification of SCC1 recruitment on loop anchors before (grey) and after (red) DSB induction, within damaged ( $n=1,456$ ) or undamaged TADs ( $n=7,804$ ). Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers. Two-sided Wilcoxon test. The increased SCC1 enrichment on loop anchors following DSB is higher in damaged TADs than in undamaged TADs. **b**, Genomics tracks showing the  $\gamma$ H2AX ChIP-seq signal (50-kb smoothed), SCC1 and phosphorylated SMC3 (pSMC3 S1083) ChIP-seq signal expressed as  $\log_2[+DSB/-DSB]$  (20-kb smoothed). **c**, Model. Cohesin-mediated loop extrusion ensures  $\gamma$ H2AX establishment on the entire damaged TAD. 1, Loop extrusion constantly occurs on the genome. 2, The occurrence of a DSB creates a roadblock for cohesin-mediated loop extrusion, leading to accumulation of cohesin at the site of damage. 3, Cohesin blocked at the DSB continues to mediate one-sided loop extrusion (arrows). ATM, recruited at the immediate vicinity of the break, phosphorylates H2AX-containing nucleosomes as they are extruded. Meanwhile, cohesin is also phosphorylated by ATM. 4, The same process takes place on both sides of the DSB, leading to divergent one-sided loop extrusion on either side of the break and ensuring bidirectional spreading of  $\gamma$ H2AX. 5, Loop extrusion triggers enlargement of  $\gamma$ H2AX-modified chromatin and halts at boundary elements such as CTCF-bound loci that demarcate TAD borders. The speed of loop extrusion (measured in vitro as  $0.5\text{--}2\text{ kb s}^{-1}$ ) ensures that the entire damaged TAD is phosphorylated in  $10\text{--}30\text{ min}$ , giving rise to a DDR focus. Cohesin is shown as a ring encircling DNA, but it is not known yet whether or how a cohesin ring entraps DNA during loop extrusion.

increase in the lengths of chromatin loops that is proposed to arise from a more processive, cohesin-mediated loop extrusion<sup>29,30</sup>. Notably, we observed extended spreading of  $\gamma$ H2AX in WAPL-depleted cells (Fig. 3c, Extended Data Fig. 5e), which is consistent with the idea that loop extrusion contributes to  $\gamma$ H2AX deposition. This was accompanied by a decrease in  $\gamma$ H2AX within TADs (Extended Data Fig. 5f). Given that WAPL depletion, while enlarging loops, also decreases intra-TAD chromatin interactions<sup>30</sup>, this suggests that intra-TAD chromosome dynamics also contribute to full deposition of  $\gamma$ H2AX.

To investigate whether such a cohesin-dependent mechanism could account for the establishment of DDR foci in budding yeast, we depleted PDS5 using an auxin-inducible system in a *Saccharomyces cerevisiae* strain<sup>31</sup> that carries three HO endonuclease cleavage sites<sup>32</sup>. Consistent with our observations in human WAPL-depleted cells, extended spreading of  $\gamma$ H2A occurred following depletion of PDS5 in yeast cells

(Fig. 3d). Notably, PDS5 deficiency triggered a decrease in  $\gamma$ H2A levels adjacent to the DSBs (Extended Data Fig. 5g), similarly to WAPL depletion in human cells.

Together, these data suggest that cohesin accumulation at DSBs initiates a one-sided loop extrusion process on either side of the break that helps to establish phosphorylation of H2AX and spreads until it reaches a strong boundary element (that is, a TAD border). This cohesin-dependent mechanism is conserved from yeast to human.

## Cohesin changes in damaged TADs

Previous work has indicated that radiation triggers a genome-wide increase in cohesin and reinforcement of TADs<sup>33,34</sup>. Consistent with this, we found that SCC1 enrichment was increased at cohesin-binding sites after break induction, coinciding with increased loop strength (Extended Data Fig. 6a, b). DSB-induced increases in loop strength and SCC1 accrual were more pronounced in damaged TADs than in undamaged TADs and decreased with the distance to DSBs (Fig. 4a, Extended Data Fig. 6c–g). Thus, our data indicate a generalized increase in SCC1 occupancy and loop strength throughout the genome after DSB production that is weakly exacerbated within TADs that are subjected to DSB. The SMC1 and SMC3 cohesin subunits have been reported to be phosphorylated by ATM following DSB induction<sup>35</sup>, and these modifications are essential for reinforcement of cohesin on the genome after irradiation<sup>34</sup>. ChIP-chip analyses indicated that phosphorylated SMC1 (pSMC1 S966) and SMC3 (pSMC3 S1083) accumulated on entire TADs around DSBs (Extended Data Fig. 7a). ChIP-seq against pSMC3 S1083 confirmed that phosphorylated SMC3 increased at cohesin-bound sites and loop anchors in damaged TADs (Fig. 4b, Extended Data Fig. 7b, c). The accumulation of these DSB-induced, ATM-mediated cohesin modifications around DSBs may regulate cohesin properties, such as loop extrusion velocity or chromatin unloading, which could translate into increased cohesin residence time at boundary elements and may help to isolate DDR domains from adjacent chromatin.

## A model for $\gamma$ H2AX domain formation

In summary, our data show that TADs are the template for the spreading of many DSB repair signalling events, such as the phosphorylation of H2AX (in agreement with a recent report<sup>36</sup>), the eviction of histone H1 and the accrual of 53BP1, MDC1 and ubiquitin, allowing DSB signalling at the megabase scale. Our results suggest a DSB-anchored cohesin-mediated loop extrusion model that would mediate phosphorylation of H2AX (Fig. 4c). In this model, cohesin accumulates rapidly on both sides of a DSB in a manner that is fostered by ATM, NIPBL and the MRN complex. Whether this is due to prior ongoing loop extrusion arresting at DSB or to de novo loading of the cohesin complex still needs to be determined. Divergent one-sided loop extrusion takes place at the DSB, which in turn allows the locally recruited ATM to phosphorylate H2AX containing nucleosomes as the chromatin fibre is pulled by the cohesin ring. Given that current estimates of cohesin-mediated loop extrusion suggest a rate of  $0.5\text{--}2\text{ kb s}^{-1}$  in vitro<sup>21,23</sup>, such a mechanism would allow rapid assembly of DDR foci, with the entire megabase-sized chromatin domain being modified in about  $10\text{--}30\text{ min}$ , which fits with the observed rate of assembly of  $\gamma$ H2AX foci<sup>9</sup>. This model is consistent with the finding that in yeast, the ATM orthologue Tel1 mediates H2A phosphorylation in a manner that agrees with a 1D sliding model rather than a 3D diffusion model<sup>37</sup>; and with the recent observation<sup>38</sup>, using light-induced activation of Cas9, that  $\gamma$ H2AX is established at a speed of about  $150\text{ kb min}^{-1}$  and can in some instance reach up to  $30\text{ Mb}$ . Moreover, our data also indicate that, upon DSB induction, the loop strength is reinforced, cohesin accumulates at loop anchors and the cohesin complex itself is modified by ATM within damaged TADs. We propose that ATM-mediated phosphorylation of the cohesin complex may alter the properties of cohesin, such as loop extrusion velocity or its

capability to load onto or unload from chromatin. These changes may further reinforce H2AX phosphorylation thanks to intra-TAD chromatin dynamics following initial loop-extrusion-dependent establishment of γH2AX.

Recent work supports the key role of TAD borders and loop extrusion in the maintenance of genome architecture and stability, including rearrangements of immunoglobulin loci<sup>39,40</sup>, and in DSB occurrence through topoisomerase reactions<sup>41,42</sup>. Our study shows that genome architecture is also instrumental for the correct establishment of γH2AX and DDR foci, expanding the function of genome organization within TADs to the response to DNA damage. We propose that arresting loop extrusion provides an efficient and rapid way to signal a DSB and assemble a DDR focus, while boundary elements help to constrain DDR signalling to DSB-surrounding, self-interacting chromatin domains. This creates a specific repair-prone chromatin compartment with modified dynamics properties, which may, for example, reduce the search time for DNA end rejoicing and homology search, and/or concentrate repair factors.

## Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at <https://doi.org/10.1038/s41586-021-03193-z>.

1. Clouaire, T., Marnef, A. & Legube, G. Taming tricky DSBs: ATM on duty. *DNA Repair* (Amst.) **56**, 84–91 (2017).
2. McCord, R. P., Kaplan, N. & Giorgetti, L. Chromosome conformation capture and beyond: toward an integrative view of chromosome structure and function. *Mol. Cell* **77**, 688–708 (2020).
3. Arnould, C. & Legube, G. The secret life of chromosome loops upon DNA double-strand break. *J. Mol. Biol.* **432**, 724–736 (2020).
4. Rogakou, E. P., Boon, C., Redon, C. & Bonner, W. M. Megabase chromatin domains involved in DNA double-strand breaks *in vivo*. *J. Cell Biol.* **146**, 905–916 (1999).
5. Clouaire, T. et al. Comprehensive mapping of histone modifications at DNA double-strand breaks deciphers repair pathway chromatin signatures. *Mol. Cell* **72**, 250–262.e6 (2018).
6. Stewart, G. S., Wang, B., Bignell, C. R., Taylor, A. M. R. & Elledge, S. J. MDC1 is a mediator of the mammalian DNA damage checkpoint. *Nature* **421**, 961–966 (2003).
7. Caron, P. et al. Cohesin protects genes against γH2AX induced by DNA double-strand breaks. *PLoS Genet.* **8**, e1002460 (2012).
8. Natale, F. et al. Identification of the elementary structural units of the DNA damage response. *Nat. Commun.* **8**, 15760 (2017).
9. Ochs, F. et al. Stabilization of chromatin topology safeguards genome integrity. *Nature* **574**, 571–574 (2019).
10. Iacovoni, J. S. et al. High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome. *EMBO J.* **29**, 1446–1457 (2010).
11. Chang, L.-H., Ghosh, S. & Noordermeer, D. TADs and their borders: free movement or building a wall? *J. Mol. Biol.* **432**, 643–652 (2020).
12. Caron, P. et al. Non-redundant functions of ATM and DNA-PKcs in response to DNA double-strand breaks. *Cell Rep.* **13**, 1598–1609 (2015).
13. Schwarzer, W. et al. Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551**, 51–56 (2017).
14. Rao, S. S. P. et al. Cohesin loss eliminates all loop domains. *Cell* **171**, 305–320.e24 (2017).
15. Gelot, C. et al. The cohesin complex prevents the end joining of distant DNA double-strand ends. *Mol. Cell* **61**, 15–26 (2016).
16. Meisenberg, C. et al. Repression of transcription at DNA breaks requires cohesin throughout interphase and prevents genome instability. *Mol. Cell* **73**, 212–223.e7 (2019).
17. Potts, P. R., Porteus, M. H. & Yu, H. Human SMC5/6 complex promotes sister chromatid homologous recombination by recruiting the SMC1/3 cohesin complex to double-strand breaks. *EMBO J.* **25**, 3377–3388 (2006).
18. Ström, L., Lindroos, H. B., Shirahige, K. & Sjögren, C. Postreplicative recruitment of cohesin to double-strand breaks is required for DNA repair. *Mol. Cell* **16**, 1003–1015 (2004).
19. Unal, E. et al. DNA damage response pathway uses histone modification to assemble a double-strand break-specific cohesin domain. *Mol. Cell* **16**, 991–1002 (2004).
20. Covo, S., Westmoreland, J. W., Gordenin, D. A. & Resnick, M. A. Cohesin is limiting for the suppression of DNA damage-induced recombination between homologous chromosomes. *PLoS Genet.* **6**, e1001006 (2010).
21. Davidson, I. F. et al. DNA loop extrusion by human cohesin. *Science* **366**, 1338–1345 (2019).
22. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
23. Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J. & Yu, H. Human cohesin compacts DNA by loop extrusion. *Science* **366**, 1345–1349 (2019).
24. Vian, L. et al. The energetics and physiological impact of cohesin extrusion. *Cell* **173**, 1165–1178.e20 (2018).
25. Schmitt, A. D. et al. A compendium of chromatin contact maps reveals spatially active regions in the human genome. *Cell Rep.* **17**, 2042–2059 (2016).
26. Mirny, L. A., Imakaev, M. & Abdennur, N. Two major mechanisms of chromosome organization. *Curr. Opin. Cell Biol.* **58**, 142–152 (2019).
27. Barrington, C. et al. Enhancer accessibility and CTCF occupancy underlie asymmetric TAD architecture and cell type specific genome topology. *Nat. Commun.* **10**, 2908 (2019).
28. Aymard, F. et al. Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nat. Struct. Mol. Biol.* **21**, 366–374 (2014).
29. Wutz, G. et al. Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDSS proteins. *EMBO J.* **36**, 3573–3599 (2017).
30. Haarhuis, J. H. I. et al. The cohesin release factor WAPL restricts chromatin loop extension. *Cell* **169**, 693–707.e14 (2017).
31. Dauban, L. et al. Regulation of cohesin-mediated chromosome folding by Eco1 and other partners. *Mol. Cell* **77**, 1279–1293.e4 (2020).
32. Lee, C.-S., Lee, K., Legube, G. & Haber, J. E. Dynamics of yeast histone H2A and H2B phosphorylation in response to a double-strand break. *Nat. Struct. Mol. Biol.* **21**, 103–109 (2014).
33. Sanders, J. T. et al. Radiation-induced DNA damage and repair effects on 3D genome organization. *Nat. Commun.* **11**, 6178 (2020).
34. Kim, B.-J. et al. Genome-wide reinforcement of cohesin binding at pre-existing cohesin sites in response to ionizing radiation in human cells. *J. Biol. Chem.* **285**, 22784–22792 (2010).
35. Kim, S.-T., Xu, B. & Kastan, M. B. Involvement of the cohesin protein, Smc1, in Atm-dependent and independent responses to DNA damage. *Genes Dev.* **16**, 560–570 (2002).
36. Collins, P. L. et al. DNA double-strand breaks induce H2Ax phosphorylation domains in a contact-dependent manner. *Nat. Commun.* **11**, 3158 (2020).
37. Li, K., Bronk, G., Kondev, J. & Haber, J. E. Yeast ATM and ATR kinases use different mechanisms to spread histone H2A phosphorylation around a DNA double-strand break. *Proc. Natl. Acad. Sci. USA* **117**, 21354–21363 (2020).
38. Liu, Y. et al. Very fast CRISPR on demand. *Science* **368**, 1265–1269 (2020).
39. Zhang, Y. et al. The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. *Nature* **573**, 600–604 (2019).
40. Zhang, X. et al. Fundamental roles of chromatin loop extrusion in antibody class switching. *Nature* **575**, 385–389 (2019).
41. Gothe, H. J. et al. Spatial chromosome folding and active transcription drive DNA fragility and formation of oncogenic MLL translocations. *Mol. Cell* **75**, 267–283.e12 (2019).
42. Canela, A. et al. Topoisomerase II-induced chromosome breakage and translocation is determined by chromosome architecture and transcriptional activity. *Mol. Cell* **75**, 252–266.e8 (2019).

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

© The Author(s), under exclusive licence to Springer Nature Limited 2021

# Article

## Methods

### Cell culture and treatments

DlVA (AsiSI-ER-U2OS)<sup>10</sup> cells generated in our laboratory were grown in Dulbecco's modified Eagle's medium (DMEM) supplemented with 10% SVF (Invitrogen), antibiotics and 1 µg/ml puromycin (DlVA cells) at 37 °C under a humidified atmosphere with 5% CO<sub>2</sub>. Cells were not further authenticated, and were regularly tested and found negative for mycoplasma contamination. For DSB induction, cells were treated with 300 nM 4-hydroxytamoxifen (4OHT) (Sigma, H7904) for 4 h. For ATM inhibition, cells were pretreated for 1 h with 20 µM KU-55933 (Sigma, SML1109) and treatment continued during subsequent 4OHT treatment. For ATR inhibition, cells were pretreated for 1 h with 2 µM ETP-46464 (Sigma, SML1321) and treatment continued during subsequent treatment with 4OHT or hydroxyurea (HU) (1 h at 1 mM (Sigma, H8627)). For kinetics experiment (Fig. 3a), cells were pretreated for 1 h with 20 µM KU-55933 (Sigma, SML1109) and treatment continued during subsequent 4OHT treatment before cells were washed three times with 1× PBS and released after 0 min, 5 min, 15 min or 1 h. siRNA transfections were performed with a control siRNA (siCtrl): CAUGUCAUGUGUCACAUUCU; and an siRNA targeting *SCC1* (si*SCC1*): GGUGAAAAUUGGCAUUACGG; or *WAPL* (si*WAPL*): CGGACUACCC UUAGCACAA; or *NIPBL* (si*NIPBL*): GCUCGGAACAAAGCAAUA; or *MRE11* (si*MRE11*): GCUAAUGACUCUGAUGAUA, using the 4D-Nucleofector and the SE cell line 4D-Nucleofector X kit L (Lonza) according to the manufacturer's instructions, and subsequent treatment(s) were performed 48 h later. For CRISPR–Cas9-mediated DSB induction, sgRNA (AsiSI site position: CGCCGCGATCGCGGAATGGA or position further within the TAD: GGGCCAGTCGCGGCCTCGC) were delivered in U2OS cells using the 'nanoblades' technology, which relies on direct cell transduction with a virus-derived particle containing the Cas9–sgRNA ribonucleoprotein<sup>43,44</sup>. Cells were analysed 24 h after transduction. For calibrated ChIP-seq experiment, mouse chromatin was obtained from E14TG2a ES cells, grown on gelatinized dishes in DMEM (Gibco) supplemented with 10% fetal bovine serum (EmbryoMax ES Cell Qualified FBS, Sigma Aldrich), 1× MEM nonessential amino acids, 1 mM sodium pyruvate, 50 µM 2-mercaptoethanol (Gibco) and 1 U/µl LIF (ESGRO Recombinant Mouse LIF, Sigma Aldrich). mES cells were obtained from A. Bird (WTCCB) and were not further authenticated. They were not tested for mycoplasma contamination.

To make the *S. cerevisiae* strain yFZ014, a linearized *TIR1* gene was obtained through restriction enzyme digestion of plasmid pJH2955 with PmeI and inserted into the *leu2* locus of strain YSCL004<sup>32</sup>. Insertion of *TIR1*<sup>45</sup> was verified by PCR with primers internal to *TIR1* and *leu2*. yFZ016 was made by PCR amplification of plasmid pJH2898 to produce a 9myc-AID::KAN PCR product with homologies at each end to the C terminus of PDS5; this PCR product was inserted using standard yeast transformation protocols to produce a PDS5::9myc-AID fusion protein. A western blot was used to verify the degradation of PDS5::9myc-AID in yFZ014 and yFZ016 after auxin addition. DSBs were induced as described<sup>33</sup>.

### Immunofluorescence

DlVA cells were plated on glass coverslips and fixed with 4% paraformaldehyde for 15 min at room temperature, permeabilized with 0.5% Triton X-100 in PBS for 10 min then blocked with 3% BSA in PBS for 30 min. Cells were then incubated with the primary antibody (Extended Data Table 1) diluted in PBS–BSA overnight at 4 °C, washed with 1× PBS and incubated with the appropriate anti-mouse or anti-rabbit secondary antibodies (conjugated to Alexa 594 or Alexa 488, Invitrogen), diluted 1:1,000 in PBS–BSA, for 1 h at room temperature, followed by DAPI staining. Coverslips were mounted in Citifluor (Citifluor, AF-1). Image acquisition was performed with MetaMorph on a wide-field microscope (Leica, DM6000) equipped with a camera (DR-328G-C01-SIL-505, ANDOR Technology) using 40× or 100× objectives. For quantification,

cells were acquired with a 40× objective and analysed using Columbus software (Perkin Elmer). γH2AX foci were detected using method D in Columbus software.

### Western blot

For detection of SCC1, WAPL, NIPBL and MRE11, cells were incubated in RIPA buffer (50 mM Tris at pH 8, 150 mM NaCl, 0.5% deoxycholate, 1% NP-40, 0.1% SDS) for 20 min on ice and centrifuged at 13,000 rpm for 10 min to remove insoluble material. SDS loading buffer and reducing agent were then added to the supernatant. For detection of pCHK1, cells were resuspended in 100 µl histone extraction buffer (1% SDS, 1% Triton, 10 mM Tris pH7.5, 0.5 M NaCl, phosphatase 0.01× (Sigma, P5726) and complete protease inhibitors 1× (Sigma, 11873580001)) and sonicated twice for 10 s with an amplitude of 30% before addition of SDS loading buffer and reducing agent. All protein extracts were resolved on 3–8% NuPAGE Tris-acetate gels (Invitrogen) and transferred onto PVDF membranes (Invitrogen) according to the manufacturer's instructions. Membranes were blocked in TBS containing 0.1% Tween 20 (Sigma, P1379) and 3% nonfat dry milk for 1 h followed by overnight incubation at 4 °C with primary antibodies (Extended Data Table 1). The appropriate horseradish peroxidase-coupled secondary antibodies were used to reveal the proteins (anti-mouse at 1:10,000 (Sigma, A2554) and anti-rabbit at 1:10,000 (Sigma, A0545)) using a luminol-based enhanced chemiluminescence HRP substrate (Super Signal West Dura Extended Duration Substrate, Thermo Scientific). Pictures of the membranes were acquired with the ChemiDoc Touch Imaging System and were visualized using Image Lab Touch software. Uncropped blots are presented in Supplementary Fig. 1.

### Hi-C

Hi-C experiments were performed in DlVA cells using the Arima Hi-C kit (Arima Genomics) according to the manufacturer's instructions. Cells (1 × 10<sup>6</sup>) were used by condition and experiments were performed in duplicate. In brief, cells were cross-linked with 2% formaldehyde for 10 min at room temperature, lysed, and chromatin was digested with two different restriction enzymes included in the kit. Ends were filled-in in the presence of biotinylated nucleotides, followed by subsequent ligation. Ligated DNA was sonicated using the Covaris S220 to an average fragment size of 350 bp with the following parameters (peak incident power, 140; duty factor, 10%; cycles per burst, 200; treatment time, 70 s). DNA was then subjected to double-size selection to retain DNA fragments between 200 and 600 bp using Ampure XP beads (Beckman Coulter). Biotin-ligated DNA was precipitated with streptavidin-coupled magnetic beads (included in the kit). Hi-C library was prepared on beads using the NEBNext Ultra II DNA Library Prep Kit for Illumina and NEBNext Multiplex Oligos for Illumina (New England Biolabs) following instructions from the Arima Hi-C kit. The final libraries were subjected to 75-bp paired-end sequencing on a Nextseq500 platform at the EMBL Genomics core facility (Heidelberg). Hi-C reads were mapped to hg19 and processed with Juicer using default settings (<https://github.com/aidenlab/juicer>). Matrix-balanced Hi-C count matrices were generated at multiple resolutions (250 kb, 100 kb, 50 kb, 25 kb, 10 kb and 5 kb) and visualized on Juicebox and on Hi-Glass.

### 4C-seq

The 4C-seq experiments were realized as described<sup>46</sup> with minor modifications. In brief, 15 × 10<sup>6</sup> DlVA cells were cross-linked with 2% formaldehyde for 10 min at room temperature, lysed and digested with MboI (New England Biolabs). Two or three rounds of 4 h of digestion with MboI were necessary. Digested DNA was then ligated with a T4 DNA ligase (HC) (Promega), and purified and digested with NlaIII overnight (New England Biolabs). After a second ligation step, DNA was purified before proceeding to library preparation. For DNA purification steps, AMPure XP beads (Beckman Coulter) were used diluted at 1:10 in 20% PEG solution (PEG 8000 (Sigma) 20%, 2.5 M NaCl, Tween

20–20%, Tris pH 8, 10 mM, EDTA 1 mM). For 4C-seq library preparation, 800–900 ng of 4C-seq template was amplified using 16 individual PCR reactions with inverse primers (PAGE-purified) including the Illumina adaptor sequences and a unique index for each condition (Extended Data Table 2). Libraries were purified with the QIAquick PCR Purification Kit (Qiagen), pooled and subjected to 75-bp single-end sequencing on a Nextseq500 platform at the I2BC Next Generation Sequencing Core Facility (Gif-sur-Yvette). Each sample was then demultiplexed using a specific python script from the FourCSeq R package<sup>47</sup>, thus assigning each read to a specific viewpoint based on its primer sequence into separate fastQ files. bwa mem was then used for mapping and samtools for sorting and indexing. A custom R script (<https://github.com/bbcf/bbcfutils/blob/master/R/smoothData.R>)<sup>48</sup> was used to build the coverage file in bedGraph format, to normalize using the average coverage and to exclude the nearest region from each viewpoint (viewpoint-containing restriction fragment and the two adjacent restriction fragments). Then the bedGraph file was converted into a BigWig file using the bedGraphToBigWig program from UCSC.

### ChIP-qPCR, ChIP-seq and ChIP-chip

For Fig. 1a, ubiquitin, H1, γH2AX and 53BP1 ChIP-seq data were retrieved from ref.<sup>5</sup>. ChIP experiments for pATM, MDC1 and phosphorylated cohesins were performed in DlVA cells as described<sup>10</sup> with 200 µg of chromatin per immunoprecipitation. Prior to library preparation, samples from multiple ChIP experiments were pooled and sonicated for 15 cycles (30-s on, 30-s off, high setting) with a Bioruptor (Diagenode) then concentrated with a vacuum concentrator (Eppendorf). CTCF and γH2AX (Fig. 3, Extended Data Figs. 5d, f) ChIP experiments were realized as follows. In brief, cross-linked cells were first lysed for 10 min at 4 °C in 500 µl lysis buffer 1 (10 mM Tris pH 8, 10 mM NaCl, 0.5% NP-40, complete protease inhibitor (Sigma, 11873580001)) then for 10 min at 4 °C in lysis buffer 2 (50 mM Tris pH 8, 10 mM EDTA, 0.5% NP-40, complete protease inhibitor (Sigma)) and subsequently sonicated in 15-ml conical tubes with a Bioruptor Pico (Diagenode) in the presence of 800 mg sonication beads (20 cycles of 30-s on/30-s off) to an average fragment size of 250 pb. Chromatin (200 µg) was then immunoprecipitated as described<sup>10</sup>. The antibodies used are detailed in Extended Data Table 1. Sequencing libraries were prepared by using 10 ng of purified DNA (average size 250–300 bp) with the NEBNext Ultra II Library Prep Kit for Illumina (New England Biolabs) using the application note for ‘Low input ChIP-seq’, and subjected to 75-bp single-end sequencing on a Nextseq500 platform at the EMBL Genomics core facility (Heidelberg).

For the SCC1-calibrated ChIP-seq, we used a spike-in method<sup>49</sup>. In brief, cross-linked DlVA cells or mouse embryonic stem cells (ES cells) were lysed and fragmented as for CTCF and γH2AX. Prior to immunoprecipitation with SCC1 antibody, 20% of chromatin from mouse ES cells (40 µg) was added to chromatin prepared from treated or untreated human DlVA cells (200 µg). Sequencing libraries were prepared from immunoprecipitation and input samples using the NEBNext Ultra II Library Prep Kit for Illumina and subjected to 75-bp single-end sequencing on a Nextseq500 platform at the EMBL Genomics core facility (Heidelberg). First, SCC1 was aligned on the mouse genome (mm10) with bwa to map only the reads used as a reference for the normalization (spike-in). Remaining unmapped reads were re-converted into a fastQ file using bam2fastq and mapped to the human genome (hg19) using bwa. Samtools was used for sorting and indexing, and reads mapped to the mouse genome were used as a normalization factor, as described<sup>49</sup> and using the following formula:  $(\text{input}_{\text{ctrl}} \times \text{reads}_{\text{exp}}) / (\text{input}_{\text{exp}} \times \text{read}_{\text{S,ctrl}})$ , in which  $\text{input}_{\text{ctrl}}$  is the total number of reads mapped in ES input (mouse) and  $\text{input}_{\text{exp}}$  is the total number of reads in DlVA input.  $\text{read}_{\text{S,ctrl}}$  and  $\text{read}_{\text{S,exp}}$  were, respectively, the number of reads from immunoprecipitated samples mapped on the mm10 genome and the hg19 genome.

For calibrated SCC1 ChIP-qPCR, the immunoprecipitated samples from DlVA cells were normalized by the signal of the immunoprecipitated sample from ES cells on a mouse cohesin-positive site (using

primers in Extended Data Table 2). Data were analysed using the Bio-Rad CFX manager software.

For the ChIP-chip experiments, the immunoprecipitated samples of γH2AX, pSMC1 S966, pSMC3 S1083 and input samples were amplified as described<sup>10</sup>, labelled and hybridized on Affymetrix tiling arrays covering human chromosomes 1 and 6 (at the Genotoul GeT-biopuces facility, Toulouse). Scanned array data were normalized using Tiling Affymetrix Software (TAS) (quantile normalization, scale set to 500), analysed as described<sup>10,12</sup> and converted into .wig files using R/Bioconductor software, when necessary, for visualization using the Integrated Genome Browser (<https://www.bioviz.org/>).

For the ChIP experiment in yeast, individual colonies of yFZ014 and yFZ016 were grown in YEP + 3% lactic acid (YEP-Lac) until log phase growth with a final cell concentration between  $5 \times 10^6$  cells per ml and  $8 \times 10^6$  cells per ml. Degradation of Pds5::9myc-AID in yFZ016 was induced by addition of auxin (Sigma Aldrich no. I3750) at a final concentration of 1 mM and confirmed by western blotting. For chromatin immunoprecipitation, 45 ml of culture was fixed and cross-linked with 1% formaldehyde for 10 min, after which 2.5 ml of 2.5 M glycine was added for 5 min to quench the reaction. Cells were pelleted and washed 3 times with 4 °C TBS. Yeast cell walls were disrupted by beating the cells with 425–600 µm glass beads for 1 h in lysis buffer at 4 °C. The lysate was sonicated for 2 min to obtain chromatin fragments of about 500 bp in length. Debris was then pelleted and discarded, and an equal volume of lysate was immunoprecipitated using γ-H2A antibody for 1 h at 4 °C, followed by addition of Protein-A agarose beads (Sigma-Aldrich no. 1719408001) for 1 h at 4 °C. The immunoprecipitate was then washed twice in 140 mM NaCl lysis buffer, once with 0.5 M NaCl lysis buffer, once with 0.25 M LiCl wash buffer and once with TE. Crosslinking was reversed at 65 °C overnight followed by addition of proteinase K and glycogen for 2 h. Protein and nucleic acids were separated by phenol extraction. LiCl was added to a final concentration of 400 mM. DNA was precipitated using 99.5% EtOH. A second precipitation step was carried out using 75% EtOH and the DNA resuspended in TE. Sequencing libraries were prepared and sequenced as for ChIP-seq in human cells.

### Hi-C, 4C-seq and ChIP-seq analyses

**Hi-C heat maps.** Hi-C heat map screenshots were generated using the Juicebox stand-alone program (<https://github.com/aidenlab/Juicebox/wiki/Download>). To build the average heat maps, sub-matrices for *cis* interactions around DSBs were extracted using Juicer, for both observed and observed over expected matrices. We computed  $\log_2(\text{ratio after/before DSB})$  using both Hi-C replicates, and averaged for each bin of the final matrix.

**Insulation score and TAD calling.** Insulation score was computed using Hi-C matrices at 50-kb resolution with matrix2insulation.pl (<https://github.com/dekkerlab/crane-nature-2015>). As parameters, we used  $\text{is} = 800000$  and  $\text{id} = 100000$ . TADs were called using Hi-C matrices at 50-kb resolution with TopDomR package and window size parameter of 10 (<https://github.com/HenrikBengtsson/TopDom>). To filter out very weak TAD borders (corresponding to sub-TAD borders), we filtered TAD borders with an insulation score below a threshold of -0.05. For Extended Data Fig. 2d, 80 TADs were also randomly selected from TopDom output, which did not contain any of the best 80 cleaved DSBs, to be used as controls.

**Loops anchors and APA.** Loops were called using the Juicer Tools HiCCUPS program at 10 kb and 25 kb resolutions (<https://github.com/aidenlab/juicer/wiki/HiCCUPS>). Aggregate peak analysis (APA) was done using the Juicer Tools APA program at 10-kb resolution (<https://github.com/aidenlab/juicer/wiki/APA>). We retrieved 525 loops between the 174 best cleaved DSBs and nearby loop anchors (<1 Mb) for replicate 1 (Fig. 2c), and 552 for replicate 2 (Extended Data Fig. 2f). The fold change between signal (central pixel) and background (upper left corner 5 × 5 pixels) was computed. For Extended Data Fig. 6f, APAs were generated

# Article

for loops filtered on their size (<200 kb) and around the best 80 cleaved DSBs. We retrieved 597 and 17,206 loops in damaged (80 damaged TADs) and undamaged TADs, respectively, in replicate 1, and 645 and 19,150 for replicate 2. The fold change between signal (central pixel) and background (lower left corner  $5 \times 5$  pixels) was computed. APA heat maps were reprocessed using ggplot2 to display counts at the same colour scale between –DSB and +DSB conditions. For Extended Data Fig. 6g, loop strength was extracted from APA files enhancement.txt corresponding to enrichment fold change (peak to mean, P2M). Differential loop strength was the log-ratio of two conditions loop strengths (+DSB/–DSB).

**ChIP-seq analyses.** ChIP-seq data were processed as described<sup>5</sup>, except for yeast ChIP-seq, which was aligned on the *S. cerevisiae* R64-1-1 assembly, and without PCR duplicate removal. SCC1 and CTCF peaks were identified using MACS2 with the callpeak algorithm, with default setting, using input as control and the SCC1 ChIP-seq data before break induction as sample. For SCC1, before breaks, 46,184 peaks were identified, with median and mean sizes of 628 and 742, respectively. For CTCF before breaks, 96,801 peaks were identified, with median and mean sizes of 339 and 500, respectively. Overlap between CTCF peaks and CTCF motifs was then performed, to associate a peak with the orientation of its motif. For representation of genomic tracks, the data were further smoothed using sliding windows as indicated. bamCompare from deeptools, with the parameters –binSize = 50, –operation = log2 and with default normalization (readCount) was used to generate differential tracks. For kinetics analysis (Extended Data Fig. 5b), γH2AX domain boundaries around the best cleaved DSBs were manually retrieved thanks to visualization of the 50-kb smoothed data on a genome browser (IGB) at different time points. The distribution of γH2AX spread is further shown as a box plot ( $n = 71$ ).

**4C-seq.** For differential analyses of the 4C-seq data, the log<sub>2</sub> ratio between two .bam files was computed using bamCompare from deeptools, with the parameters –binSize = 50 and –operation = log2. Extended Data Figure 3d shows the mean and s.e.m of the 4C-seq ratio on 1 Mb around each viewpoint, obtained across four independent experiments (control viewpoints,  $n = 3$ ; DSB viewpoints,  $n = 11$ ). Extended Data Figures 3h, 4c show the distribution (box plots) of the 4C-seq ratio on 1 Mb around DSB viewpoints obtained across two (siSCC1) or three (ATMi) independent experiments ( $n = 8$ ).

## Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

## Data availability

All high-throughput sequencing data (Hi-C, ChIP-seq, 4C-seq) have been deposited to Array Express (<https://www.ebi.ac.uk/arrayexpress/>)

under accession number E-MTAB-8851. ChIP-chip data have been deposited to Array Express under accession number E-MTAB-8793. Uncropped blots are shown in Supplementary Fig. 1. Other data (ChIP-qPCR and raw microscopy data) are available upon request.

## Code availability

Source codes are available from <https://github.com/LegubeDNARE-PAIR/LoopExtrusion>.

43. Mangeot, P. E. et al. Genome editing in primary cells and in vivo using viral-derived Nanoblades loaded with Cas9-sgRNA ribonucleoproteins. *Nat. Commun.* **10**, 45 (2019).
44. Marnef, A. et al. A cohesin/HUSH- and LINC-dependent pathway controls ribosomal DNA double-strand break repair. *Genes Dev.* **33**, 1175–1190 (2019).
45. Morawska, M. & Ulrich, H. D. An expanded tool kit for the auxin-inducible degron system in budding yeast. *Yeast* **30**, 341–351 (2013).
46. Matelot, M. & Noordermeer, D. Determination of high-resolution 3D chromatin organization using circular chromosome conformation capture (4C-seq). *Methods Mol. Biol.* **1480**, 223–241 (2016).
47. Klein, F. A. et al. FourCSeq: analysis of 4C sequencing data. *Bioinformatics* **31**, 3085–3091 (2015).
48. David, F. P. A. et al. HTSstation: a web application and open-access libraries for high-throughput sequencing data analysis. *PLoS ONE* **9**, e85879 (2014).
49. Kojic, A. et al. Distinct roles of cohesin-SA1 and cohesin-SA2 in 3D chromosome organization. *Nat. Struct. Mol. Biol.* **25**, 496–504 (2018).

**Acknowledgements** We thank the genomics core facility of EMBL for high-throughput sequencing; the high-throughput sequencing core facility of the I2BC (Centre de Recherche de Gif) for facilities and expertise; F. Beckouet for advice on yeast work; J. Rispa and N. Firmin for occasional experimental help; and C. Normand for discussions. Work in the Haber laboratory was funded by grant R35 GM127029 from the US National Institutes of Health. F.Z. was supported by the National Institute of General Medical Sciences Training Grant TM32GM007122. E.R. is supported by Labex Ecofect (ANR-11-LABX-0048) of the Université de Lyon, Fondation FINOVI and by the European Research Council (ERC-StG-LS6-805500) under the European Union’s Horizon 2020 research and innovation programmes. Funding in the Legube laboratory was provided by grants from the European Research Council (ERC-2014-CoG 647344), the Agence Nationale pour la Recherche (ANR-14-CE10-0002-01 and ANR-18-CE12-0015), the Institut National Contre le Cancer (INCA), and the Ligue Nationale Contre le Cancer (LNCC). This work was supported by the Fondation pour la Recherche Médicale, grant number FDT201904007941, to C.A.

**Author contributions** C.A. performed 4C-seq, Hi-C, ChIP-seq, ChIP-chip and ChIP-qPCR experiments. A.-L.F. contributed to siRNA experiments and performed CTCF ChIP-seq. K.L. and F.Z. performed yeast strain construction and γH2A ChIP. P.C. performed ChIP-chip in SCC1 siRNA. V.R. and R.M. performed bioinformatic analyses of 4C-seq, Hi-C and ChIP-seq datasets. E.P.R. and P.E.M. provided nanoblades for CRISPR-Cas9 experiments. D.N. helped to realize and analyse 4C-seq experiments. T.C. supervised experiments in human cells and helped with library preparation. J.E.H. conceived and supervised work in yeast. G.L. conceived experiments, supervised the work and wrote the manuscript. All authors commented and edited the manuscript.

**Competing interests** The authors declare no competing interests.

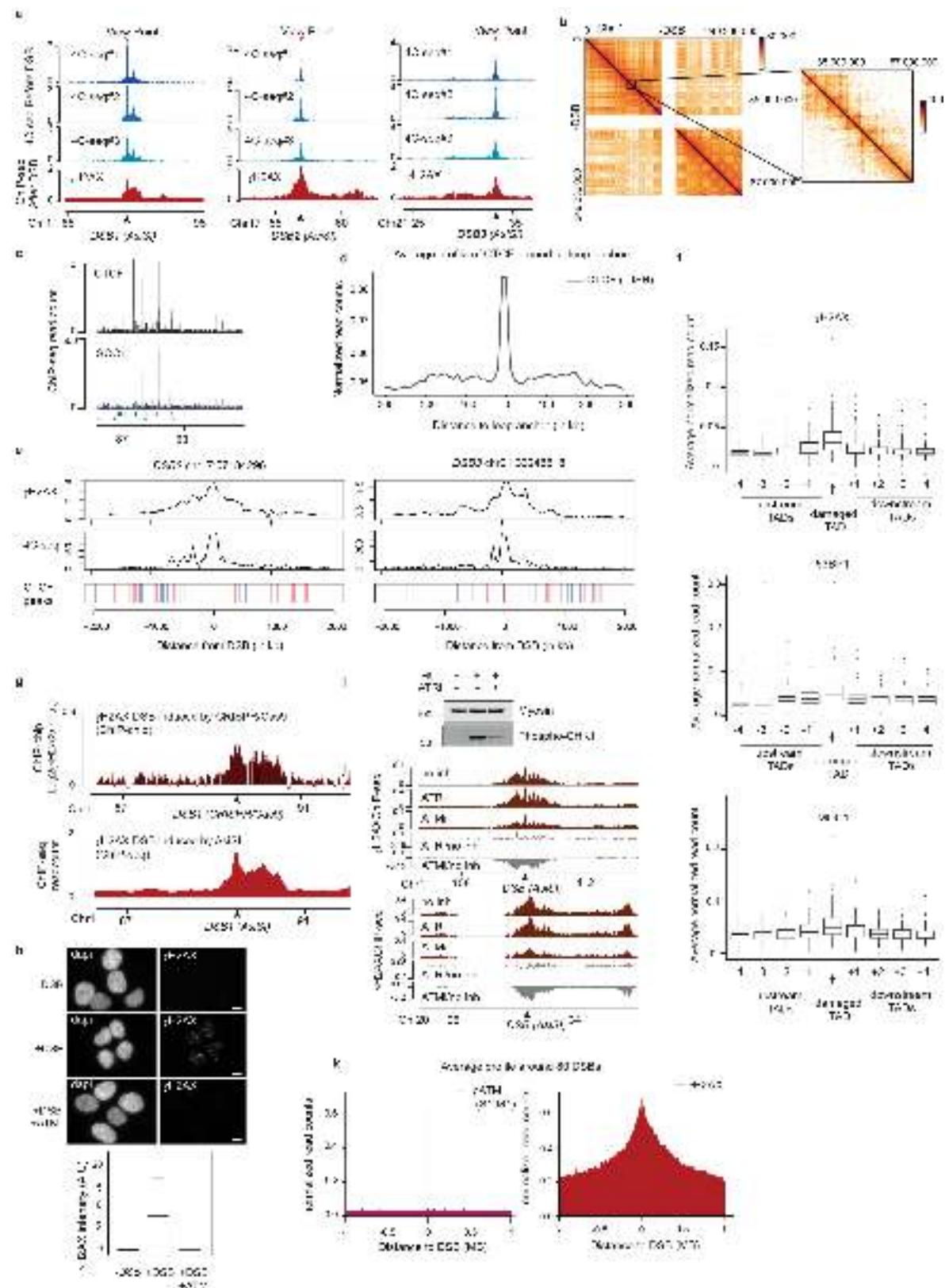
## Additional information

**Supplementary information** The online version contains supplementary material available at <https://doi.org/10.1038/s41586-021-03193-z>.

**Correspondence and requests for materials** should be addressed to G.L.

**Peer review information** *Nature* thanks Leonid Mirny and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

**Reprints and permissions information** is available at <http://www.nature.com/reprints>.

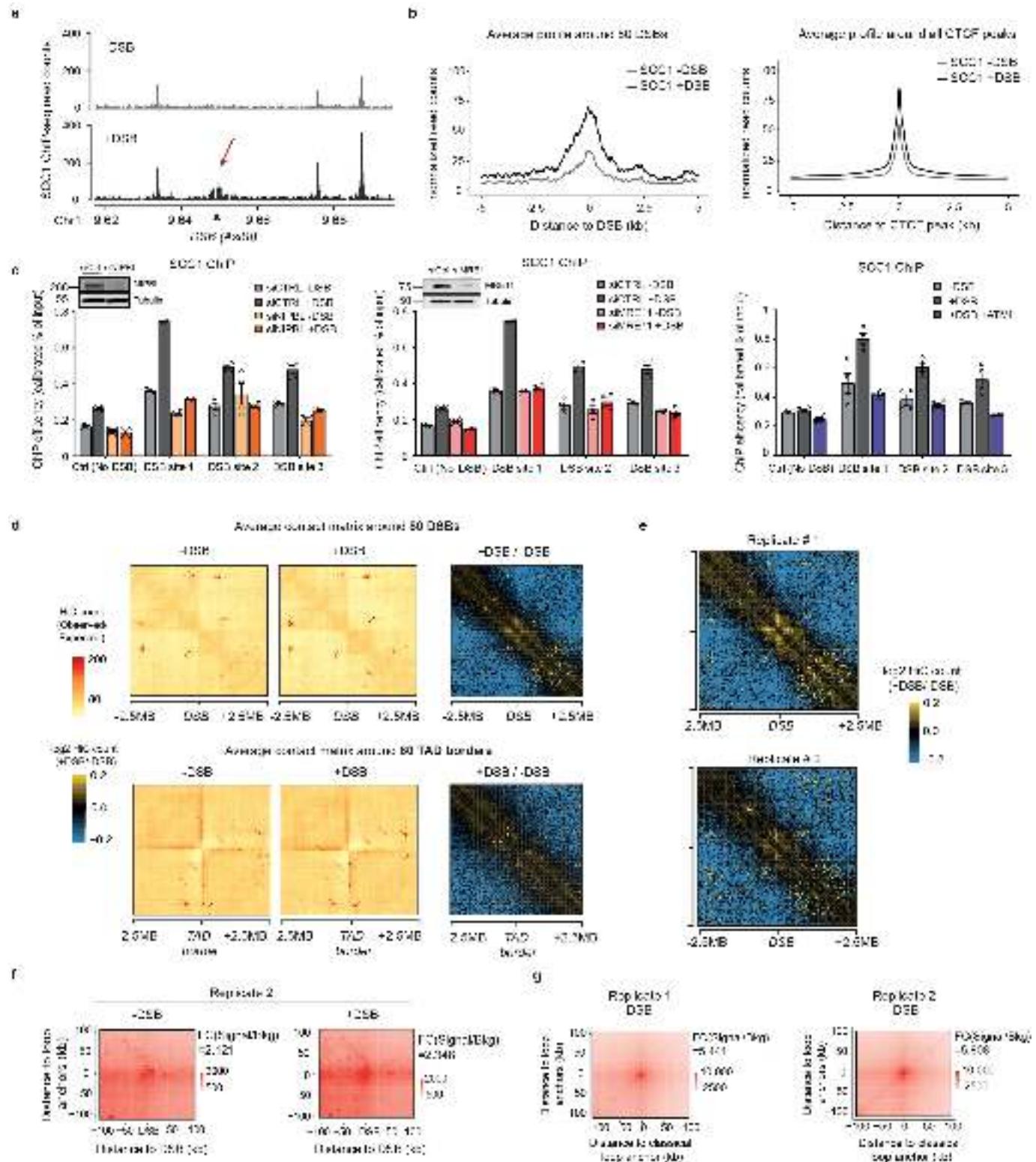


**Extended Data Fig. 1** | See next page for caption.

## Article

**Extended Data Fig. 1 | γH2AX spreads within prior TADs as revealed by 4C-seq.** **a**, 4C-seq tracks before DSB induction obtained for three independent biological replicates and γH2AX ChIP-seq track after DSB induction for different viewpoints (red arrows) localized at three AsI sites (black arrows). ChIP-seq data were smoothed using 100-kb span and 4C-seq data using a 50-kb span. **b**, Example of the Hi-C pattern obtained on chromosome 1 at a 500-kb resolution (left) together with a magnification at a 10-kb resolution (right). **c**, CTCF and calibrated-SCC1 ChIP-seq tracks. **d**, Average profile of CTCF ChIP-seq around all loop anchors on the genome (determined using this Hi-C dataset, Methods), validating both CTCF ChIP-seq and Hi-C datasets. **e**, γH2AX ChIP-seq after DSB induction. 4C-seq and CTCF ChIP-seq peak position before DSB induction are shown (peaks in blue contain a CTCF motif in the forward orientation and peaks in red a CTCF motif in the reverse orientation). **f**, Box plot showing γH2AX (top), 53BP1 (middle) and MDC1 (bottom) ChIP-seq quantification within the damaged TAD and neighbouring TADs for the best cleaved DSBs in DlVA cells (Methods). Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers ( $n=153$ ). **g**, γH2AX tracks around a DSB induced by CRISPR-Cas9

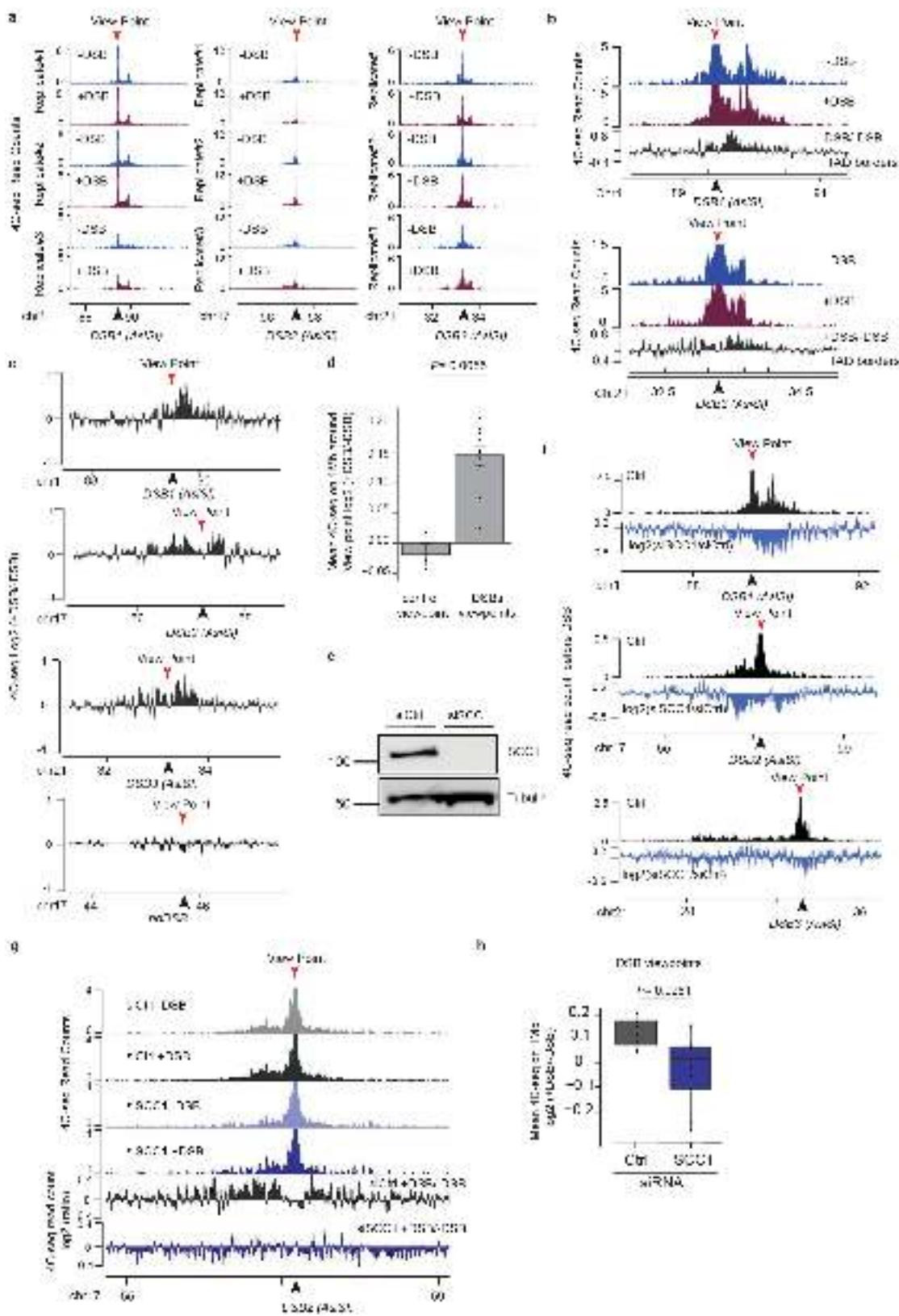
(top, ChIP-chip, expressed as  $\log_2[\text{sample}/\text{input}]$ , smoothed using 100-probe windows) and by AsI at the same position (bottom, ChIP-seq, 50-kb smoothed). **h**, Top, immunofluorescence experiment showing γH2AX and DAPI staining before and after DSB induction with or without ATM inhibitor as indicated (scale bars, 10 μm). Bottom, quantification of γH2AX intensity (expressed in arbitrary units (A.U.)) in the above conditions. One representative experiment is shown (out of  $n=3$  biological replicates). Box plots as in **f**. -DSB,  $n=117$  nuclei; +DSB,  $n=97$  nuclei; +DSB + ATM,  $n=95$  nuclei. **i**, Validation of ATR inhibitor efficiency. Western blot showing the effect of ATRi on the phosphorylation of CHK1 following treatment with hydroxyurea (HU) ( $n=2$ ). For gel source data, see Supplementary Fig. 1. **j**, γH2AX ChIP-seq tracks after DSB induction in untreated cells or in cells treated with an inhibitor of ATM or ATR at two DSB sites (20-kb smoothed). The differential γH2AX signal obtained after DSB induction (expressed as the  $\log_2$  ratio ATM/untreated or ATRi/untreated, grey tracks) is also shown ( $n=1$ ). **k**, Average profile of pATM (S1981) (left) and γH2AX (right) ChIP-seq on a 2-Mb window around the 80 best-cleaved DSBs in DlVA cells.



# Article

**Extended Data Fig. 2 | Cohesin recruitment and loop extrusion occurs at DSBs.** **a**, Calibrated SCC1 ChIP-seq tracks before (grey) and after (black) DSB induction ( $n=1$ ). SCC1 enrichment at DSB site is indicated by a red arrow. **b**, Average profile of SCC1 ChIP-seq signal centred on the 80 best-induced DSBs (left) or centred on all CTCF peaks of the genome (right) on a 10-kb window. **c**, Calibrated ChIP-qPCR of SCC1 in the indicated conditions at three DSB sites or a negative control region. Insets, western blots validating depletion of the proteins NIPBL ( $n=1$ ) and MRE11 ( $n=2$ ) by the corresponding siRNAs. For gel source data, see Supplementary Fig. 1. Mean  $\pm$  s.e.m. for technical replicates ( $n=4$ ) of a representative experiment (out of  $n=2$  biological replicates). **d**, Averaged Hi-C matrix before (−DSB) and after DSB induction (+DSB) (observed/expected) and of the  $\log_2$  ratio between damaged and undamaged cells centred on the 80 best-induced DSBs (top) or centred on eighty random TAD borders (bottom) (50-kb resolution, 5-Mb window; combined replicates). **e**, Averaged Hi-C contact matrix of  $\log_2$ [+DSB/−DSB]

centred on the eighty best-induced DSBs in the two independent biological replicates. **f**, APA plot on a 200-kb window (10-kb resolution) before (−DSB) and after DSB induction (+DSB) in biological replicate no. 2 (replicate no. 1 shown in Fig. 2c). APAs are calculated between the DSBs and loop anchors ( $n=552$  pairs). The fold change between the signal (central pixel) and the background (upper left corner  $5 \times 5$  pixels) is indicated. **g**, For comparison with **f**, APA plot on a 200-kb window (10-kb resolution) before DSB induction computed between classical loop anchors that are near DSB sites (<500 kb;  $n=674$  pairs for replicate 1 and  $n=737$  pairs for replicate 2). The fold change between the signal (central pixel) and the background (upper left corner  $5 \times 5$  pixels) is indicated. The loop strength (quantified by the fold change between signal and background on the APA plot) is higher at loop anchors (**g**, replicate 1 fold-change = 5.4; replicate 2 fold-change = 5.8) than the loop strength observed at DSBs after break induction (Fig. 2c, replicate 1, fold-change = 2; **f**, replicate 2, fold-change = 2.3).



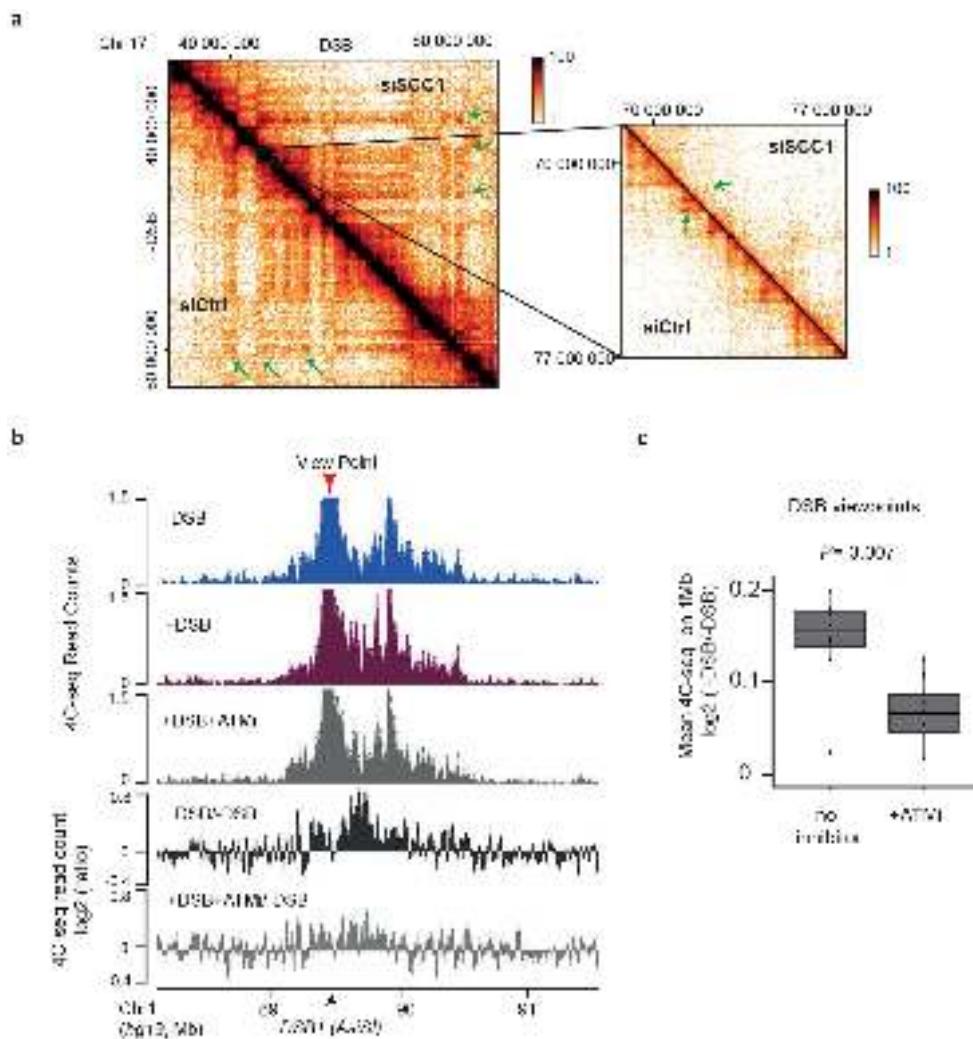
**Extended Data Fig. 3** | See next page for caption.

# Article

## Extended Data Fig. 3 | Loop extrusion at DSBs detected by 4C-seq.

**a**, 4C-seq tracks (10-kb smoothed) before and after DSB induction, obtained for three biological replicates using viewpoints localized at three DSB sites (arrows). **b**, 4C-seq tracks before (blue) and after (purple) DSB induction, at two DSB viewpoints. Differential 4C-seq ( $\log_2[+DSB/-DSB]$ ) is also shown (black). **c**, Differential 4C-seq ( $\log_2[+DSB/-DSB]$ ) for three viewpoints located at DSB sites and on a control region as indicated. **d**, Differential 4C-seq signal ( $\log_2[+DSB/-DSB]$ ) computed on 1 Mb around four independent viewpoints located at DSBs (DSBs viewpoints,  $n=11$ ) and one control region (control viewpoint,  $n=3$ ), across four independent biological experiments (Methods). Two-sided Wilcoxon test; mean  $\pm$  s.e.m. **e**, Western blot showing depletion of

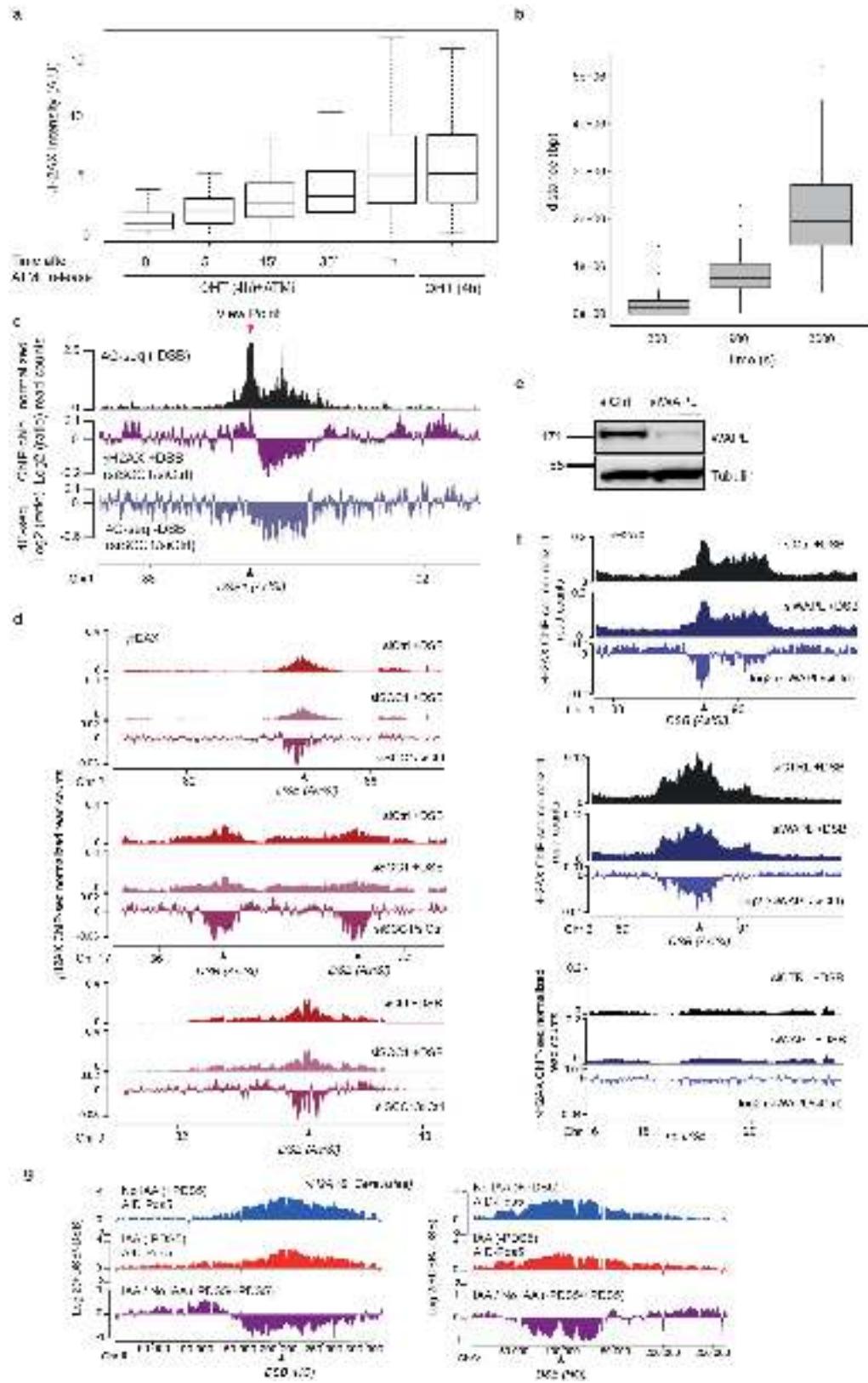
SCC1 by siRNA ( $n=3$ ). For gel source data, see Supplementary Fig. 1. **f**, Differential ( $\log_2$ ) 4C-seq track in si $SCC1$ -treated cells versus control siRNA-treated cells (in undamaged conditions) for three viewpoints. **g**, Genomics tracks showing 4C-seq signals before and after DSB induction in control siRNA- or si $SCC1$ -treated cells and the differential 4C-seq signal in control siRNA- or si $SCC1$ -treated cells ( $\log_2[+DSB/-DSB]$ ; 10-kb smoothed). **h**, Average  $\log_2[+DSB/-DSB]$  4C-seq, on 1 Mb around four DSB viewpoints (two biological experiments) upon treatment with control siRNA or si $SCC1$  (Methods) ( $n=8$ ). Two-sided Wilcoxon test. Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers.



**Extended Data Fig. 4 | ATM activity is required for loop extrusion at DSBs.** **a**, Hi-C maps before DSB induction of a region of chromosome 17 in control and SCC1-depleted cells. Left, 100-kb resolution; right, 25-kb resolution.

**b**, Genomic tracks of 4C-seq before and after DSB induction in untreated or ATM-inhibitor-treated cells and of differential 4C-seq signal ( $\log_2[+DSB/-DSB]$

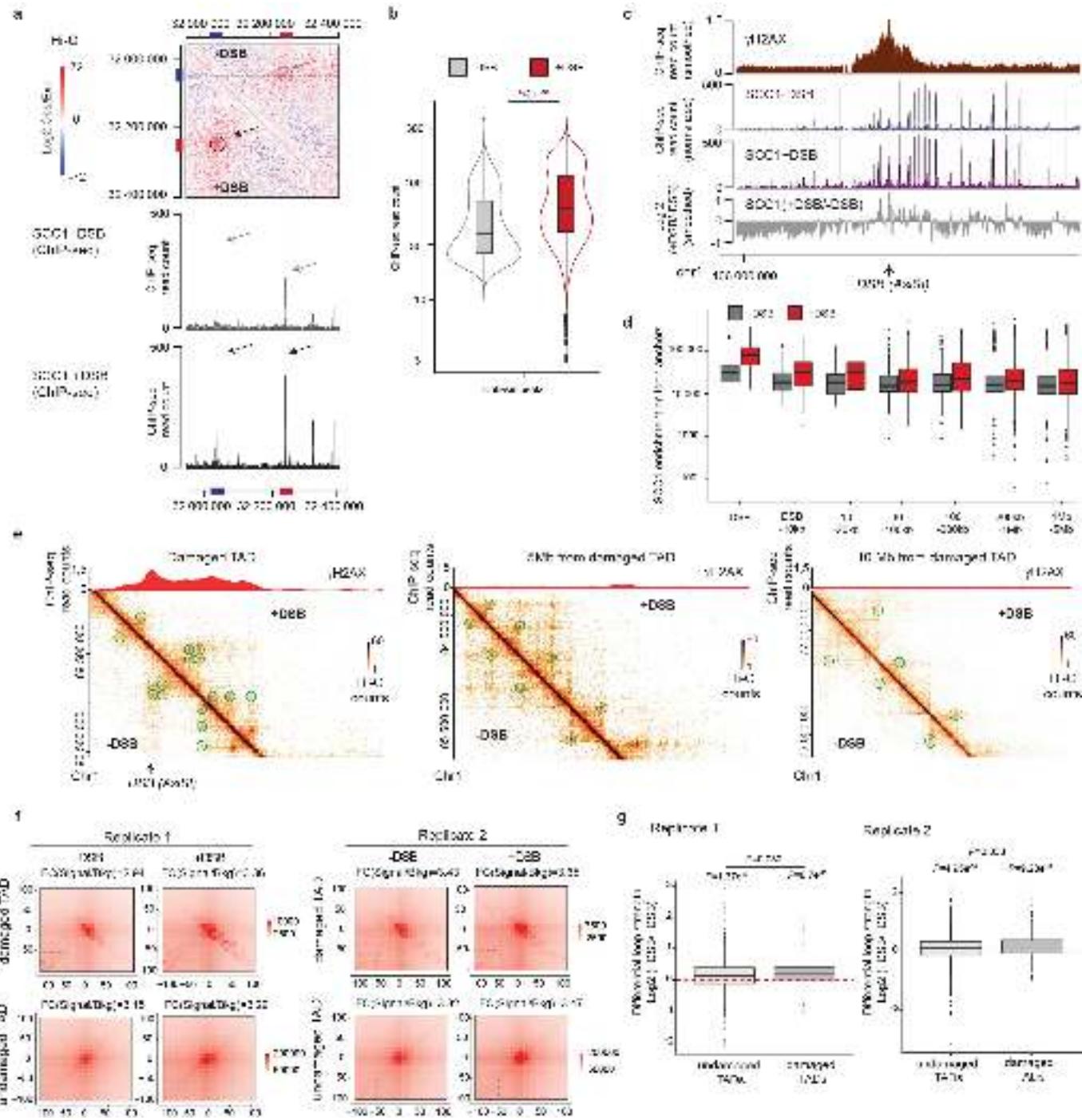
or  $\log_2[+DSB + ATMi/-DSB]$ ; 10-kb smoothed). **c**, *Cis* interactions computed as in Extended Data Fig. 3h for four DSB viewpoints across three biological experiments, in control condition or upon ATM inhibition. Two-sided Wilcoxon test. Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers ( $n=8$ ).



**Extended Data Fig. 5 | Altered loop extrusion modifies γH2AX spreading.**

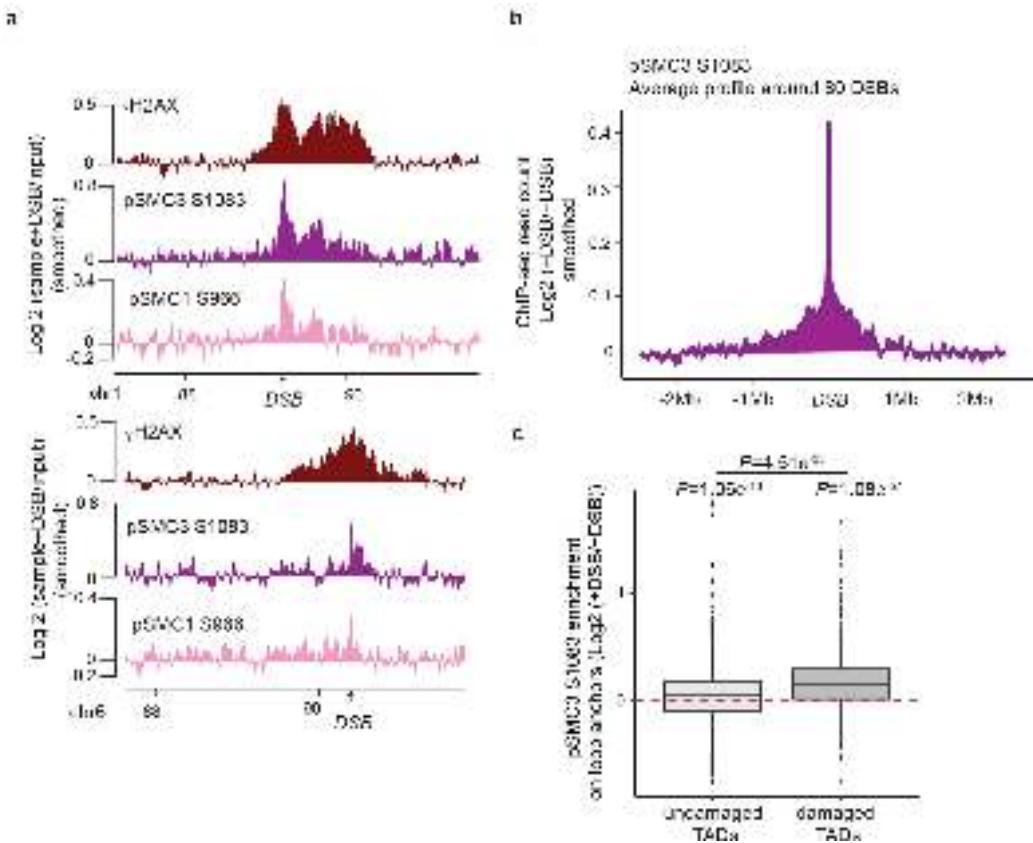
**a**, Quantification of γH2AX intensity after DSB induction (OHT, 4 h) and upon ATM inhibition followed by different times after ATMi release (0 min,  $n=172$  nuclei; 5 min,  $n=183$  nuclei; 15 min,  $n=171$  nuclei; 30 min,  $n=197$  nuclei; 1 h,  $n=189$  nuclei). Treatment with OHT for 4 h without ATMi is also shown ( $n=182$  nuclei). One representative experiment is shown (out of  $n=2$  biological replicates). Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers. **b**, Spread of γH2AX (in bp) at the indicated time points after release from ATMi around the best cleaved DSBs ( $n=71$ ). Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers. **c**, Black, 4C-seq track before DSB induction using a DSB viewpoint. Purple, differential γH2AX signal obtained after DSB induction by ChIP-chip in SCC1-depleted versus control cells (expressed as γH2AX  $\log_2[\text{siSCC1}/\text{siCtrl}]$ ). Light blue, differential

4C-seq signal obtained in SCC1-depleted versus control cells before DSB induction ( $\log_2[\text{siSCC1}/\text{siCtrl}]$ ). **d**, Genomic tracks of γH2AX ChIP-seq signal after DSB induction in control (red) or SCC1-depleted (pink) cells and of the differential γH2AX signal obtained after DSB induction ( $\log_2[\text{siSCC1}/\text{siCtrl}]$ , purple) at two DSB sites. **e**, Western blot validating the effect of the siRNA targeting *WAPL* on the *WAPL* protein level ( $n=2$ ). For gel source data, see Supplementary Fig. 1. **f**, Genomics tracks of γH2AX ChIP-seq after DSB induction in control or *WAPL*-depleted cells and of the differential γH2AX signal obtained after DSB induction ( $\log_2[\text{si}WAPL/\text{siCtrl}]$ ) at two DSB sites and one control (no DSB) genomic locus (20-kb smoothed). **g**, Genomics tracks of the differential γH2A ChIP-seq signal ( $\log_2[+\text{DSB}/-\text{DSB}]$ ) before (no IAA) or after PDS5 degradation (IAA) at two DSB sites (HO sites) in *S. cerevisiae* (SacCer3, coordinates in bp) ( $n=1$ ).



**Extended Data Fig. 6 | Increased genome-wide, DSB-induced, cohesin binding is enhanced within damaged TADs.** **a.** Top, contact matrix (5-kb resolution) showing  $\log_2[\text{observed}/\text{expected}]$  before or after DSB induction on a region showing a loop on chromosome 20 and devoid of AsI site (no DSB). Loops anchors are circled and indicated by red and blue bars. Bottom, genome browser screenshot showing the SCC1-calibrated ChIP-seq on the same region before and after DSB induction. Cohesin enrichment at the loop anchors (blue and red bars) is increased after DSB (black arrows) compared to before DSB (grey arrows), in agreement with increased loop strength (grey and black circles, top). **b.** Violin plots showing SCC1 enrichment at cohesin peaks ( $n=46,194$ ) before and after DSB induction. Paired one-sided Wilcoxon test. **c.** Genomic tracks of  $\gamma$ H2AX (red) and SCC1 ChIP-seq signal before (blue) and after (purple) DSB induction. The ratio between before and after DSB induction (grey) is also shown ( $\log_2[+DSB/-DSB]; 10\text{-kb smoothed}$ ). **d.** Quantification of SCC1 recruitment on loop anchors at different distances from DSB sites as indicated (from left to right,  $n=1,610, 3,161, 1,930, 3,232, 4,786, 25,263$ , 114,461). Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum; points, outliers. **e.**  $\gamma$ H2AX ChIP-seq signal and Hi-C signal at different distances from a damaged TAD on chromosome 1 before (-DSB) and after DSB induction (+DSB). Green circles, chromatin loops. **f.** APA plot on a 200-kb window (10-kb resolution) before (-DSB) and after DSB induction (+DSB) calculated for all loop anchors, in damaged and undamaged TADs. The fold change between the signal (central pixel) and the background (lower left corner  $5 \times 5$  pixels) is indicated. **g.** Differential loop strengths in undamaged or damaged TADs (Methods), computed from Hi-C data obtained before and after DSB, from replicates 1 and 2. *P* values between before and after DSB are indicated (Wilcoxon test,  $\mu=0$ ). The increased loop strength following DSB is significantly higher in damaged TADs than in undamaged TADs (paired two-sided Wilcoxon test) in both Hi-C replicate experiments. Replicate 1: undamaged,  $n=2,936$ ; damaged,  $n=264$ . Replicate 2: undamaged,  $n=3,181$ ; damaged,  $n=302$ . Box plots as in d.

114,461). Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum; points, outliers. **e.**  $\gamma$ H2AX ChIP-seq signal and Hi-C signal at different distances from a damaged TAD on chromosome 1 before (-DSB) and after DSB induction (+DSB). Green circles, chromatin loops. **f.** APA plot on a 200-kb window (10-kb resolution) before (-DSB) and after DSB induction (+DSB) calculated for all loop anchors, in damaged and undamaged TADs. The fold change between the signal (central pixel) and the background (lower left corner  $5 \times 5$  pixels) is indicated. **g.** Differential loop strengths in undamaged or damaged TADs (Methods), computed from Hi-C data obtained before and after DSB, from replicates 1 and 2. *P* values between before and after DSB are indicated (Wilcoxon test,  $\mu=0$ ). The increased loop strength following DSB is significantly higher in damaged TADs than in undamaged TADs (paired two-sided Wilcoxon test) in both Hi-C replicate experiments. Replicate 1: undamaged,  $n=2,936$ ; damaged,  $n=264$ . Replicate 2: undamaged,  $n=3,181$ ; damaged,  $n=302$ . Box plots as in d.



**Extended Data Fig. 7 | DSB-induced phosphorylation of cohesin occurs in damaged TADs.** **a**, Genomic tracks showing  $\gamma$ H2AX, pSMC3 S1083 and pSMC1 S966 ChIP-seq signals expressed as  $\log_2$ [sample/input] after DSB induction. Two damaged genomic locations are shown. **b**, Average profile of pSMC3 S1083 (expressed as  $\log_2$ ([+DSB/-DSB] ChIP-seq signal)) around the 80 best-induced DSBs on a 4-Mb window. **c**, Quantification of pSMC3 S1083 signal on loop anchors in damaged or undamaged TADs. *P* values between before and after

DSB are indicated (paired two-sided Wilcoxon test). The increased pSMC3 S1083 enrichment on loop anchors following DSB is significantly higher in damaged TADs than in undamaged TADs (two-sided Wilcoxon test). Undamaged,  $n = 9,040$ ; damaged,  $n = 1,626$ . Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers.

# Article

Extended Data Table 1 | Antibodies used in this study

Target	Application	Reference	Quantity
$\gamma$ H2AX (S139)	ChIP	Merck Millipore 07-164	2 µg
$\gamma$ H2AX (S139)	IF	Merck Millipore 05-636 (clone JBW301)	1:1000
P-ATM (S1981)	ChIP	Abcam ab81292	2 µg
MDC1	ChIP	Abcam ab11171	3 µg
SCC1	ChIP	Abcam ab992	4 µg
SCC1	Western Blot	Abcam ab992	1:500
WAPL	Western Blot	Santa Cruz sc-365189	1:500
NIPBL	Western Blot	Bethyl Laboratories A301-779A	1:1000
MRE11	Western Blot	GeneTex GTX70212 (clone 12D7)	1:4000
Tubulin	Western Blot	Sigma T6199	1:10000
Myosin	Western Blot	Sigma M3567	1:2000
Phospho-CHK1 (ser345)	Western Blot	Cell Signaling 2348S	1:1000
P-SMC1 (S966)	ChIP	Epitomics EP2858Y	2 µL
P-SMC3 (S1083)	ChIP	Bethyl Laboratories A300-480A	2 µg
CTCF	ChIP	Millipore 07-729	4 µL
$\gamma$ H2A (yeast)	ChIP	Abcam ab15083	2µg

Extended Data Table 2 | Primers used in this study

Application	Name	Forward primer	Reverse primer
4C-seq	Viewpoint DSB1	AATGATAACGGCGACCACCGAGATC TACACTCTTCCTACACGACGCTC TTCCGATCTAACCTGGCAACTTATG AATCAGGA	CAAGCAGAAAGACGGCATACGAGAT <u>NNNNNN</u> GTGACTGGAGTTAGACG TGTGCTCTTCCGATCTATGTCAAAA GCCAAGGGGACA
4C-seq	Viewpoint DSB2	AATGATAACGGCGACCACCGAGATC TACACTCTTCCTACACGACGCTC TTCCGATCTTCCCTACGATTATTGT GAATTTTG	CAAGCAGAAAGACGGCATACGAGAT <u>NNNNNN</u> GTGACTGGAGTTAGACG TGTGCTCTTCCGATCTAAGCTAATT CTGAGTTACATACATT
4C-seq	Viewpoint DSB3	AATGATAACGGCGACCACCGAGATC TACACTCTTCCTACACGACGCTC TTCCGATCTGATTACGTAGAAGGGT GCC	CAAGCAGAAAGACGGCATACGAGAT <u>NNNNNN</u> GTGACTGGAGTTAGACG TGTGCTCTTCCGATCTAAGGCAAAT GATAACCCCTGT
4C-seq	Viewpoint ctrl region	AATGATAACGGCGACCACCGAGATC TACACTCTTCCTACACGACGCTC TTCCGATCTTCCCTCAGGTTATCATC CCA	CAAGCAGAAAGACGGCATACGAGAT <u>NNNNNN</u> GTGACTGGAGTTAGACG TGTGCTCTTCCGATCTCACCTTCGC TGTACCTTG
4C-seq	Viewpoint CRISPR site	AATGATAACGGCGACCACCGAGATC TACACTCTTCCTACACGACGCTC TTCCGATCTTAAGCACCCCTCC TAG	CAAGCAGAAAGACGGCATACGAGAT <u>NNNNNN</u> GTGACTGGAGTTAGACG TGTGCTCTTCCGATCTACCTTACA CCTCAAAACCT
4C-seq	Viewpoint 470 kb upstream (Fig. 1c)	AATGATAACGGCGACCACCGAGATC TACACTCTTCCTACACGACGCTC TTCCGATCTACAAGGAAGAAGCAG GCATTCA	CAAGCAGAAAGACGGCATACGAGAT <u>NNNNNN</u> GTGACTGGAGTTAGACG TGTGCTCTTCCGATCTTGAAATGA GTACTCTGCCATCCA
ChIP-qPCR	Ctrl region	AGCACATGGGATTTGCAGG	TTCCCTCCTTGTGTCACCA
ChIP-qPCR	DSB site 1	TCCCCTGTTCTCAGCACTT	CTTCTGCTGTTCTGCCTCCT
ChIP-qPCR	DSB site 2	CCGCCAGAAAGTTCTAGA	CTCACCCCTGCAGCACTTG
ChIP-qPCR	DSB site 3	CCTAGCTGAGGTGGTGCTA	GAAGAGTGAGGAGGGGAGT
ChIP-qPCR	Cohesin positive site (mouse)	CAGAGATTTGGGTGTTGG	TTCACACCTAGAGGAGGGT

NNN is the position of the optional index.

### 2.2.4 La formation d'un nouveau compartiment de la chromatine dépendant d'ATM régule la réponse aux cassures double-brin de l'ADN et la biogenèse des translocations

Au sein du noyau, l'induction des cassures double-brin (*DSB*) impacte la conformation de la chromatine, à la fois localement mais aussi à l'échelle du génome entier. Cependant, la façon dont l'endommagement de l'ADN et les processus de réparation impactent cette conformation est encore bien mal compris. En outre, les *DSB* sont capables de se regrouper et de former des *clusters*, aux rôles encore méconnus (François AYMARD, AGUIRREBENGOA et al. 2017; KILIC et al. 2019). La manière dont la chromatine dans le noyau est impliquée dans ce phénomène, et les mécanismes à son origine ne sont également pas bien établis. Il a cependant été montré que la chromatine enrichie en 53BP1, acteur majeur de la réparation, se regroupe afin de former un compartiment isolé du reste du génome, par séparation de phase (KILIC et al. 2019).

Dans cet article, nous étudions les principes de la conformation 3D, à la fois locaux et globaux du génome des mammifères suite à l'induction de *DSB*. Nous rapportons que les *TAD* endommagés subissent des changements de conformation 3D dépendants d'ATM. Au sein de ces *TAD*, le mécanisme de *loop extrusion* décrit dans l'article précédent assure la régulation transcriptionnelle. Au sein du génome, les *TAD* endommagés forment un nouveau compartiment génomique, où les gènes acteurs de la réponse aux dommages à l'ADN, ou *DNA Damage Response (DDR)* sont re-localisés physiquement. Ce nouveau compartiment, et ce regroupement des gènes de la *DDR* en son sein suggère un rôle actif du regroupement des *DSB* dans la réparation. Cependant, ceci se fait au détriment d'une augmentation des translocations, risquant de remettre en cause l'intégrité du génome.

1    **ATM-dependent formation of a novel chromatin compartment regulates**  
2    **the Response to DNA Double Strand Breaks and the biogenesis of**  
3    **translocations**

4

5    Coline Arnould<sup>1#</sup>, Vincent Rocher<sup>1#</sup>, Aldo S. Bader<sup>2</sup>, Emma Lesage<sup>1</sup>, Nadine Puget<sup>1</sup>, Thomas  
6    Clouaire<sup>1</sup>, Raphael Mourad<sup>1</sup>, Daan Noordermeer<sup>3</sup>, Martin Bushell<sup>2,4</sup> and Gaëlle Legube<sup>1\*</sup>

7

- 8    1. *MCD, Centre de Biologie Intégrative (CBI), CNRS, Université de Toulouse, UT3*  
9    2. *Cancer Research UK Beatson Institute, Garscube Estate, Switchback Road, Bearsden,*  
10    *Glasgow G61 1BD, UK*  
11    3. *Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell*  
12    *(I2BC), 91198, Gif-sur-Yvette, France*  
13    4. *Institute of Cancer Sciences, University of Glasgow, Garscube Estate, Switchback*  
14    *Road, Bearsden, Glasgow G61 1QH, UK*

15

16    \* corresponding author: [gaelle.legube@univ-tlse3.fr](mailto:gaelle.legube@univ-tlse3.fr)

17    # these authors contributed equally

18

19    **Abstract**

20    **DNA Double-Strand Breaks (DSBs) repair is essential to safeguard genome integrity but**  
21    **the contribution of chromosome folding into this process remains elusive. Here we**  
22    **unveiled basic principles of chromosome dynamics upon DSBs in mammalian cells. We**  
23    **report that topologically associating domains (TAD) that experience a DSB are subjected**  
24    **to acute alterations in an ATM-dependent but DNA-PK-independent manner. Within**  
25    **these damaged TADs, DSB-induced loop extrusion mediates local transcriptional**  
26    **regulation in response to DSBs. We also found that ATM drives damaged, γH2AX-**  
27    **decorated, TADs to coalesce, forming a new chromatin compartment (“D” compartment),**  
28    **where genes upregulated by DNA damage response (DDR) also physically localize,**  
29    **suggesting a function for DSB clustering in activating the DNA damage Response.**  
30    **However, both loop extrusion and compartmentalization following DSB also come at the**  
31    **expense of an increased translocations rate. Our work highlights the critical impact of**  
32    **chromosome conformation in the maintenance of genome integrity.**

33

34

35

36 **Introduction**

37 DNA Double-Strand Breaks (DSBs) are highly toxic lesions that can trigger translocations or  
38 gross chromosomal rearrangements, thereby severely challenging genome integrity and cell  
39 homeostasis. DSBs activate the DNA Damage Response (DDR) that largely relies on PI3K  
40 kinases, including ATM and DNA-PK, and on the establishment of megabase-sized, γH2AX-  
41 decorated chromatin domains that act as seeds for subsequent signaling events, such as 53BP1  
42 recruitment and DDR foci formation<sup>1,2</sup>. Meanwhile DSB repair machineries are recruited at the  
43 break site to restore the genomic sequence, either using Non-Homologous End Joining (NHEJ)  
44 or Homology-Driven Recombination (HDR) pathways<sup>3</sup>.

45 Of importance, repair of DSBs occurs in the context of chromatin, which plays a pivotal  
46 function during repair mechanisms. Past studies revealed that chromatin is selectively modified  
47 on various scales around DSBs ranging from few nucleosomes to megabase domains and that  
48 these modifications depend on the repair pathway used<sup>4</sup>. Moreover, chromatin is also a key  
49 determinant for the repair pathway choice and henceforth directly influences repair outcome<sup>5</sup>.  
50 Histone modifications pre-existing the occurrence of the lesion or established post break  
51 formation regulate end-processing and further act as anchoring platforms for reader DNA repair  
52 proteins, such as the anti-resection factor 53BP1<sup>6,7</sup>. Additionally, DSB also elicit a local (*in cis*)  
53 ATM-dependent transcriptional repression of neighboring genes that depends on DSB-induced  
54 histone modifications<sup>8</sup>. Whether this local inhibition of transcription takes place inside the  
55 entire γH2AX-decorated domains, which would therefore potentially affect tens of genes, is  
56 still under debate<sup>9</sup>.

57 While significant steps have been achieved towards our understanding of the contribution of  
58 chromatin structure during DSB repair, little is known about the contribution of chromosome  
59 architecture into these processes. Importantly, γH2AX spreading is largely influenced by the

60 pre-existing chromosome conformation in topologically associating domains (TADs)<sup>10–12</sup> and  
61 we recently reported that loop-extrusion by the cohesin complex, which compacts the chromatin  
62 to create TADs<sup>13</sup>, is instrumental for γH2AX spreading and DDR foci formation<sup>12</sup>.  
63 Interestingly, the cohesin complex also mediates the local transcriptional repression observed  
64 around DSB through a yet unclear mechanism<sup>14</sup>. At a larger scale, previous work in mammalian  
65 cells revealed that DSBs display the ability to “cluster” within the nuclear space (*i.e.*, fuse)  
66 forming large microscopically visible repair foci, composed of merged individual repair foci<sup>15–</sup>  
67 <sup>17</sup>. DSB clustering depends on the actin network, the LINC (a nuclear envelope embedded  
68 complex)<sup>17–19</sup>, as well as on the liquid-liquid phase separation properties of 53BP1<sup>20,21</sup>. The  
69 function of repair foci fusion has remained enigmatic given that juxtaposition of several DSBs  
70 can elicit translocation (*i.e.*: illegitimate rejoining of two DNA ends)<sup>16</sup>, questioning the selective  
71 advantage of DSB clustering/ repair foci fusion<sup>22</sup>.

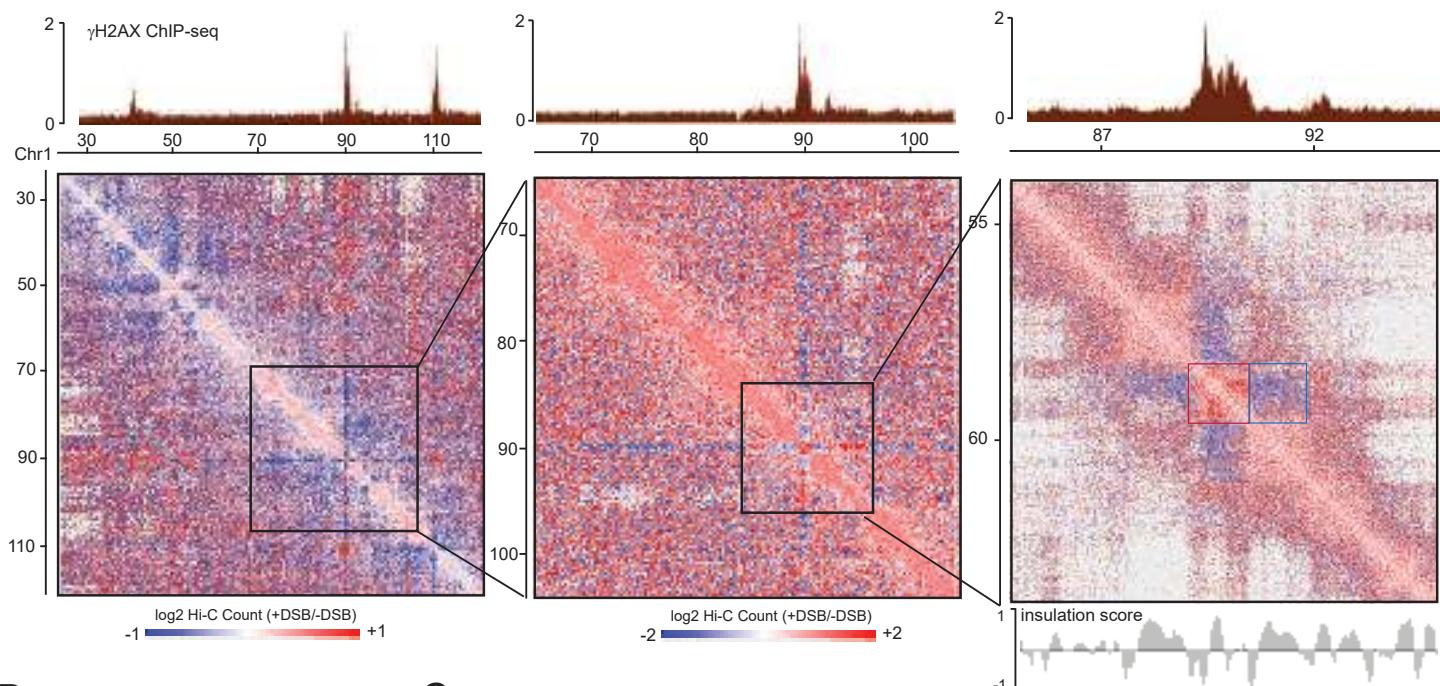
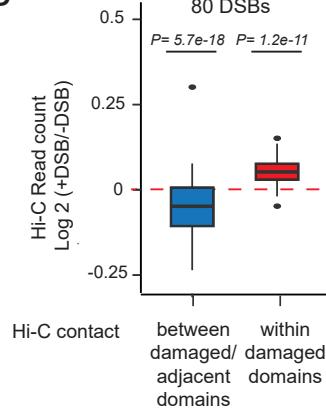
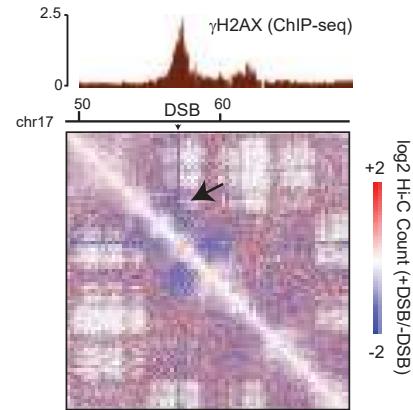
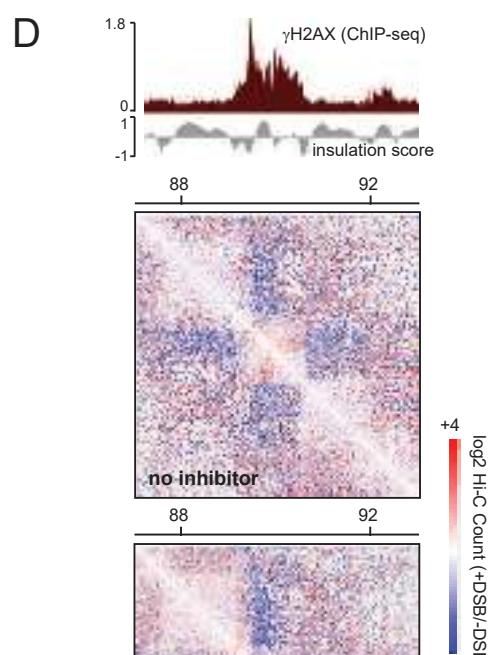
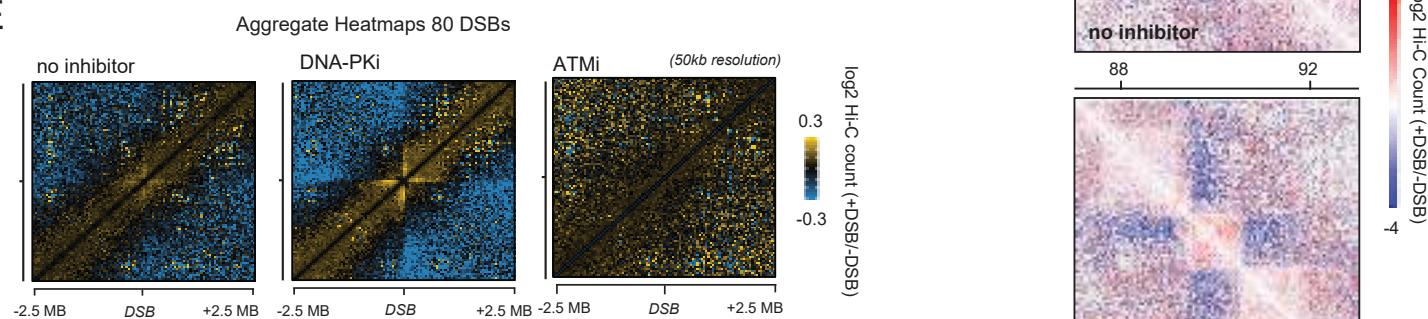
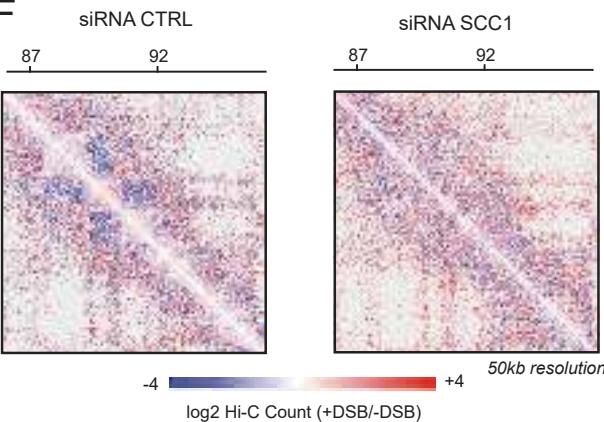
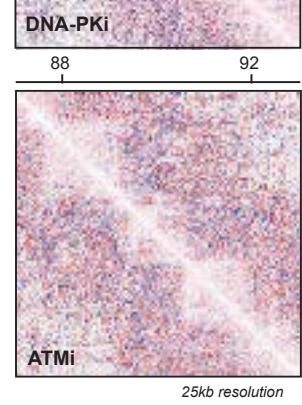
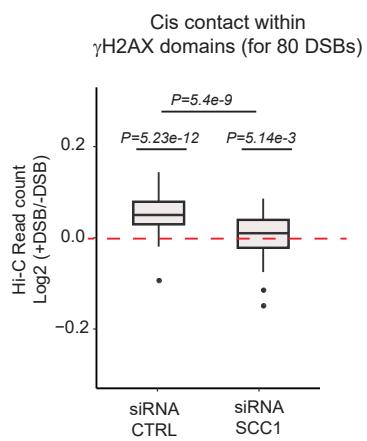
72 Here we used a combination of ChIP-seq and Hi-C to investigate chromosome folding in  
73 response to DSBs. We found that 3D chromosome structure is profoundly altered following  
74 DSBs in an ATM-dependent manner. Within damaged TADs, DSB-induced loop extrusion is  
75 responsible for local transcriptional repression of distant genes. Furthermore, damaged TADs  
76 engage in long distance contacts forming a new DSB-induced chromatin compartment (or “D”  
77 compartment), whose formation is regulated by ATM and cell cycle progression, and where  
78 genes upregulated in response to DNA damage physically localize. Yet, D-compartment and  
79 cohesin-mediated loop-extrusion are decisive for the generation of translocations and DDR  
80 genes targeted to the D compartment are frequently rearranged in cancer genome. Altogether,  
81 our work suggests that the 3D genome organization is strongly remodeled following DSBs and  
82 acts as a key mediator of multiple steps during DSB repair, while also being a critical aspect in  
83 the biogenesis of translocations.

85    **Results**

86    **Upon DSB, ATM drives an acute reinforcement of damaged TADs as well as a mild  
87    strengthening of TADs genome-wide**

88    In order to get comprehensive insights into chromosome behavior following DSBs, we analyzed  
89    3D genome organization using Hi-C data generated in the human DIvA cell line where multiple  
90    DSBs are induced at annotated positions upon hydroxytamoxifen (OHT) addition<sup>12</sup>. Using  
91    differential Hi-C maps, we found that intra-TAD contacts frequencies were strongly increased  
92    within TADs that experience a DSB (*i.e* the damaged TADs) compared to undamaged TADs,  
93    while contacts with neighboring adjacent domains were significantly decreased (Fig. 1A, 1B).  
94    Interestingly, in some instances, the DSB itself displayed a particularly strong depletion of  
95    contact frequency with adjacent chromatin (Fig. 1C black arrow) indicating that the DSB is  
96    kept isolated from the surrounding environment, outside its own TAD.

97    In agreement with previous work showing a general reinforcement of TADs upon irradiation in  
98    mammalian cells<sup>23</sup>, differential Hi-C maps also revealed that intra-TAD contacts frequencies  
99    were increased on entire chromosomal regions devoid of DSBs as visualized by the absence of  
100    $\gamma$ H2AX (Fig. S1A, black square), although to a lesser extent than within the damaged domains  
101   themselves (Fig. S1A, S1B). Furthermore, aggregate differential Hi-C maps computed on  
102   undamaged chromatin (damaged,  $\gamma$ H2AX-covered, chromatin domains excluded) showed that  
103   contacts frequencies below 1Mb range (intra-TAD contacts) are enhanced post damage at the  
104   expense of interactions with neighboring TADs (Fig. S1C). Plotting the contact probability as  
105   a function of genomic distance confirmed that following DSB, contact frequency in the 100-  
106   1000kb range (intra-TAD) are increased (Fig. S1D, top panel) while higher distance contacts  
107   are decreased (Fig. S1D bottom panel). This was visible on entire chromosomes irrespective of  
108   their damage load (16 DSBs annotated on chromosome 1, 2 DSBs on chromosome 16) (Fig.

**A****B****C****D****E****F****G**

**Figure 1: Cohesin and ATM-dependent TAD reinforcement in response to DSBs.**

**(A)** Hi-C contact matrix of the log2 (+DSB/-DSB) in DlVA cells. A region of the chromosome 1 is shown at three different resolutions: 250 kb (left panel), 100 kb (middle panel) and 25 kb (right panel). The  $\gamma$ H2AX ChIP-seq signal following DSB induction is shown on the top panel and indicates the DSBs position. The red square highlights a damaged TAD, within which cis interactions are enhanced, while the blue square highlights decreased interaction between the damaged TAD and its adjacent TAD. One representative experiment is shown. **(B)** Boxplot showing the differential Hi-C read counts (as  $(\log_2 +\text{DSB}/-\text{DSB})$ ) within  $\gamma$ H2AX domains containing the 80 best induced DSBs (red) or between these 80 damaged domains and their adjacent chromatin domains (blue). P-values, non-parametric wilcoxon test tested against  $\mu=0$ . **(C)** Hi-C contact matrix of log2 (+DSB/-DSB) on a region located on chromosome 17 at 50 kb resolution. The contacts engaged by the DSB itself are indicated with a black arrow.  $\gamma$ H2AX ChIP-seq track (+DSB) is shown on the top panel. One representative experiment is shown. **(D)** Hi-C contact matrix of the log2(+DSB/-DSB) without inhibitor (top panel), with DNA-PK inhibitor (middle panel) or with ATM inhibitor (bottom panel). A damaged region of the chromosome 1 is shown at a 25 kb resolution. Grey track represents the insulation score pre-existing to DSB induction (from Hi-C -DSB). **(E)** Averaged Hi-C contact matrix of the log2 (+DSB/-DSB) in untreated cells (left panel), upon DNA-PK inhibition (middle panel) or upon ATM inhibition (right panel), centered on the 80 best-induced DSBs (50 kb resolution on a 5 Mb window). **(F)** Hi-C contact matrix of the log2(+DSB/-DSB) on a region located on chromosome 1 at a 50 kb resolution in DlVA cells transfected with a control siRNA or a siRNA directed against SCC1. **(G)** Boxplot showing the differential Hi-C read counts ( $\log_2 +\text{DSB}/-\text{DSB}$ ) within the 80 best  $\gamma$ H2AX domains in Control or SCC1-depleted conditions. P-values, non-parametric wilcoxon test tested against  $\mu=0$ . siSCC1 vs siCTRL, P=paired wilcoxon test.

109 S1C). In agreement with the above data and with a recent report in yeast<sup>24</sup>, loop strength on  
110 loop anchors (located outside of γH2AX domains) increased following DSB, in a manner that  
111 was independent of the number of DSB per chromosome (Fig. S1E). Taken together these data  
112 indicate that, in response to DSBs, the entire genome undergoes significant changes in  
113 chromosome dynamics, including a mild genome-wide TADs reinforcement and an acute  
114 isolation of the TADs that comprise the DSBs.

115 We further investigated the contribution of PI3-Kinases involved in response to DSB by  
116 performing Hi-C in presence of catalytic inhibitors of ATM and DNA-PK, which respectively  
117 negatively and positively impact γH2AX accumulation at DSBs (in contrast to ATR inhibition,  
118 which did not display any noticeable change on γH2AX foci formation in the DIvA system)<sup>12,25</sup>.  
119 Notably, DNA-PK inhibition exacerbated the increased frequency of intra -TAD contacts  
120 observed following DSB induction, while ATM inhibition abrogated the TAD reinforcement  
121 observed post DSB induction, both on undamaged chromatin (Fig. S2A) and at damaged TADs  
122 (Fig. 1D, Fig. S2B).

123 TAD structures visualized on Hi-C maps are believed to arise thanks to cohesin-mediated loop  
124 extrusion<sup>13</sup>. Our previous work indicated that a bidirectional, divergent, cohesin-dependent  
125 loop-extrusion process takes place at DSBs<sup>12</sup>. This DSB-anchored loop extrusion can be  
126 visualized on differential Hi-C maps by a “cross” pattern centered on the DSB (Fig. 1E).  
127 Notably, ATM inhibition impaired loop extrusion, while DNA-PK inhibition strongly increased  
128 it (Fig. 1E). Moreover, depletion of the cohesin subunit SCC1, which abolishes DSB-induced  
129 loop extrusion<sup>12</sup>, decreased both the insulation of damaged, γH2AX-decorated, chromatin  
130 domains as well as the reinforcement of undamaged TADs (Fig. 1F-G, Fig. S2C)).

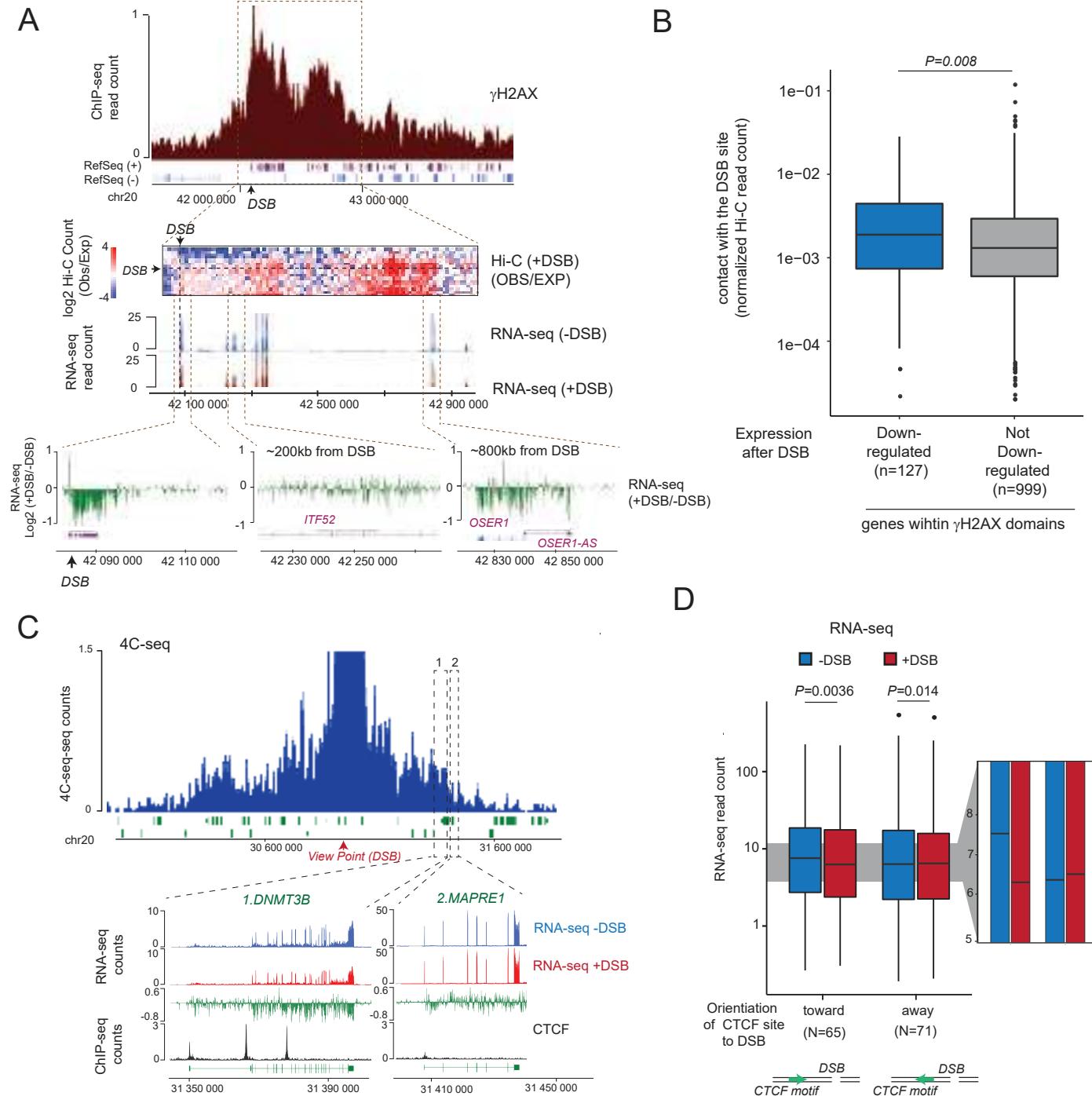
131 Altogether these data indicate that DSBs trigger both a moderate genome-wide reinforcement  
132 of TADs throughout the undamaged genome and an acute insulation of the damaged TADs

133 themselves from the surrounding chromatin, associated with cohesin-mediated loop extrusion  
134 arising from the DSB. Both responses are entirely ATM-dependent and suppressed by DNA-  
135 PK.

136

137 **Cohesin-mediated loop extrusion contributes to local transcriptional repression**

138 Induction of DSBs is known to trigger a local ATM- and DNA-PK-dependent response that  
139 elicits the transcriptional repression of the damaged genes themselves but also that of  
140 neighboring genes<sup>26-30</sup>. However, to which extent this DSB-induced signaling spreads *in cis* to  
141 repress adjacent genes is debated<sup>31,32</sup>. A previous work performed in DIvA cells revealed that  
142 transcriptional repression takes place within entire  $\gamma$ H2AX-decorated megabase chromatin  
143 domains as a function of log(distance) to DSB<sup>32</sup>. We could recapitulate such a general trend  
144 using RNA-seq data that we generated before and after DSB induction and a log scale for the  
145 distance to DSB (Fig. S3A). However, when looking at a distance between 100kb to 1Mb away  
146 from the DSB, the relationship between transcriptional repression and distance to DSB was  
147 only minor (Fig. S3B), showing that linear distance to a DSB is not a strong determinant for  
148 transcriptional shut down. Of interest the cohesin complex has been involved in mediating  
149 DSB-induced *cis* gene repression, suggesting a contribution of chromosome folding<sup>14</sup>. We  
150 therefore examined a potential relationship between transcriptional repression and physical  
151 proximity to the DSB (using our Hi-C data). An example at one of the DSB induced in DIvA  
152 cells is shown on Fig. 2A. Notably, the gene that carries the DSB showed a clear transcriptional  
153 repression (see the left zoom, bottom panel), while a gene located 200kb downstream did not  
154 display such transcriptional repression (middle zoom). However, further downstream (~800kb  
155 from the DSB, still encompassed in the  $\gamma$ H2AX domain, right zoom), another expressed gene  
156 did show a decrease in its expression level post-DSB induction. Interestingly, this 800kb-distant



**Figure 2. Cohesin-mediated loop extrusion contributes to local DSB-induced transcriptional regulation**

**(A)** Top panel: Genomic tracks of  $\gamma$ H2AX ChIP-seq after DSB induction in a region of chromosome 20, DSB position is indicated by a black arrow. Middle panel: Hi-C contact matrix showing the signal obtained after DSB induction ( $\log_2$  observed/expected) and genomic tracks of the RNA-seq signal before (-DSB, blue) and after DSB induction (+DSB, red). The DSB position is indicated with a black arrow and a dashed black line. Bottom panel: Close-up showing differential RNA-seq ( $\log_2$  ratio +DSB/-DSB, green) track for a gene localized at the DSB site (zoom panel, left), at  $\sim$ 200 kb from the DSB site (zoom panel, middle) and at  $\sim$ 800 kb from the DSB site (zoom panel, right). One representative experiment is shown. **(B)** Boxplot showing the quantification of the Hi-C contacts between the genes located in  $\gamma$ H2AX domains and the closest DSB. Genes that did not display transcriptional changes post-DSB induction are shown in grey, genes that showed down-regulation following DSB induction in blue (identified by RNA-seq). P, non parametric wilcoxon test. **(C)** Top panel: Genomic track of the 4C-seq signal after DSB induction using a viewpoint localized at a DSB on chromosome 20 (red arrow). One representative experiment is shown. Bottom panel: Close up of two genomic loci located at  $\sim$ 400 kb from the DSB site (*DNMT3B* and *MAPRE1*), showing RNA-seq read count before (-DSB, blue) and after DSB induction (+DSB, red), differential RNA-seq ( $\log_2$  ratio +DSB/-DSB, green), and CTCF ChIP-seq (-DSB). **(D)** Quantification of the RNA-seq read counts before (-DSB, blue) and after DSB induction (+DSB, red) on genes embedded in  $\gamma$ H2AX domains and displaying a CTCF-bound element facing toward or away from the closest DSB site, as indicated. P, non-parametric paired wilcoxon test.

157 gene had enriched physical contacts with the DSB, while this was not the case for the 200kb-  
158 distant gene (see the above Hi-C heatmap). In general, genes included in  $\gamma$ H2AX domains that  
159 showed downregulation post-DSB induction exhibited a higher contact frequency with the DSB  
160 compared to genes within  $\gamma$ H2AX domains which were not transcriptionally downregulated  
161 following DSB (Fig. 2B). To strengthen these data, we performed 4C-seq, using as a view point  
162 one of the DSB induced on the chromosome 20 in the DIvA system (Fig. 2C). Similarly, we  
163 could observe for instance that *DNMT3B*, located 400kb away from the DSB, displayed  
164 physical contacts with the DSB viewpoint (by 4C-seq) and was downregulated post-DSB  
165 induction (Fig. 2C). However, the adjacent *MAPRE1* gene displaying low 4C-seq signal (hence  
166 a low interaction frequency with the DSB viewpoint), was not transcriptionally affected by the  
167 induction of DSBs (Fig. 2C). Altogether, this suggests that the physical proximity of a gene  
168 with the DSB in 3D, rather than the linear distance, is determining its potential to be  
169 downregulated post-DSB induction. Notably, cohesin depletion impaired the DSB-induced  
170 transcriptional shutdown of spatially proximal, linearly distant, DSB-neighboring genes (Fig.  
171 S3C).

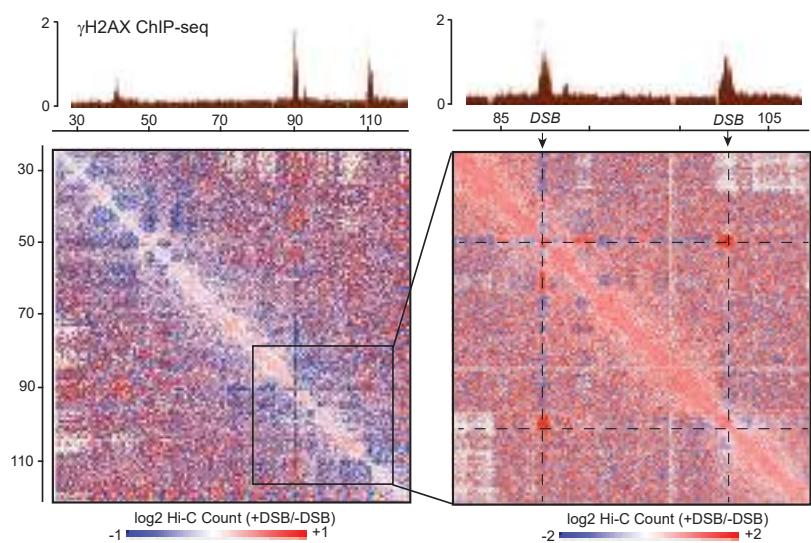
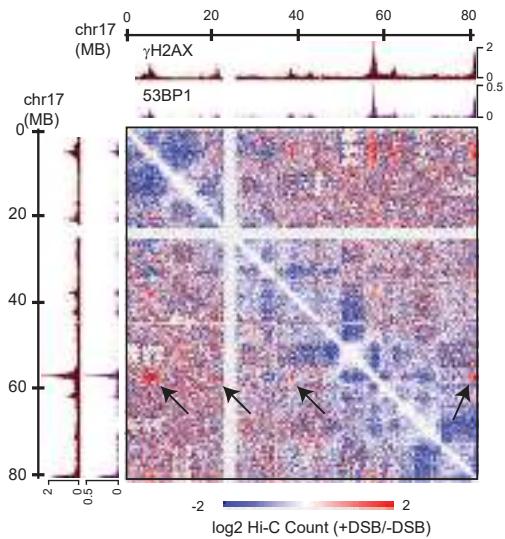
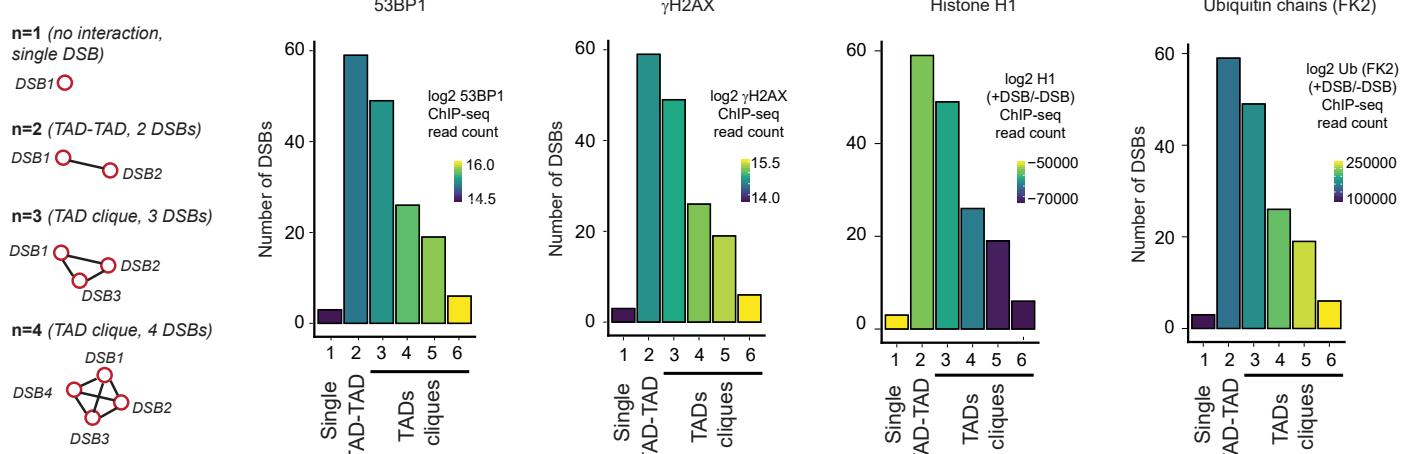
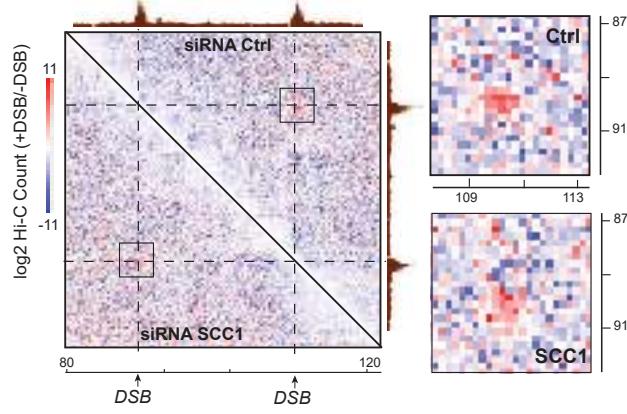
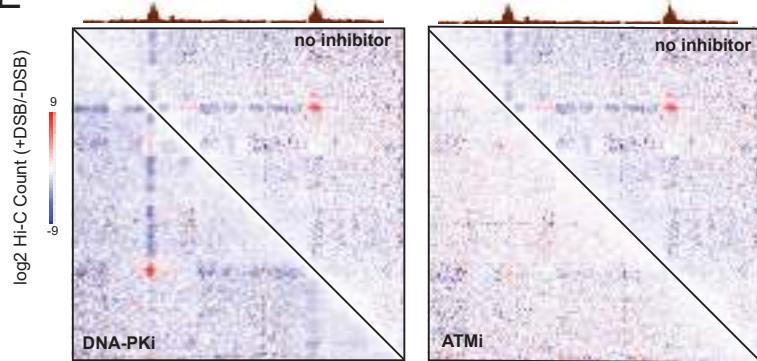
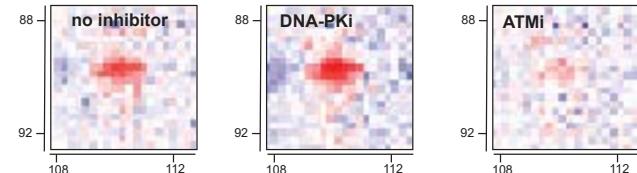
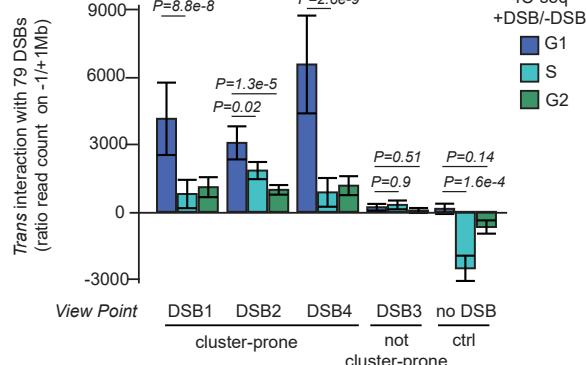
172 The cohesin complex is known to mediate loop extrusion, that pauses at boundaries elements  
173 such as CTCF-bound loci<sup>33</sup>. Importantly, we recently reported (see also Fig. 1) that cohesin-  
174 mediated loop extrusion can be initiated at the DSB itself<sup>12</sup>. We therefore hypothesized that the  
175 presence of a CTCF-bound element could block the DSB-induced loop extrusion, therefore  
176 bringing a DSB-neighboring gene bound by CTCF to the physical proximity of the DSB. To  
177 investigate this further, we compared RNA-seq data with CTCF ChIP-seq performed in DIvA  
178 cells. Notably, within  $\gamma$ H2AX domains, the genes downregulated following DSB showed higher  
179 CTCF binding than unaffected genes (two examples shown Fig. 2C and Fig. S3D). To  
180 efficiently block loop extrusion, CTCF-bound elements must be in a specific orientation relative  
181 to the direction of loop extrusion<sup>34-36</sup>. We therefore retrieved genes located in  $\gamma$ H2AX domains

182 that exhibit CTCF binding by ChIP-seq and sorted them depending on the orientation of the  
183 CTCF binding site, pointing either “towards” or “away” from the DSB. Interestingly, the genes  
184 that encompassed a CTCF-bound element facing towards the DSB displayed transcriptional  
185 repression following DSB, while the genes carrying a CTCF-bound element pointing away  
186 from the DSB were left unaffected (Fig. 2D). Altogether these data support a model in which  
187 loop extrusion that initiates from the DSB itself is able to bring distant genes in the physical  
188 proximity with the DSB. When a CTCF-bound element faces towards the DSB, the pausing of  
189 loop extrusion will ensure sustained physical proximity between the gene carrying the CTCF-  
190 bound locus and the DSB that is bound by ATM, which elicits transcriptional repression.  
191 However, when there are no CTCF-bound element or if this element is facing away from the  
192 DSB, loop extrusion does not stop, expelling the gene outside the immediate vicinity of the  
193 DSB itself, therefore escaping ATM-induced transcriptional shutdown.

194

195 **ATM-dependent clustering of damaged TADs is cell cycle regulated and does not require**  
196 **cohesin or DNA-PK**

197 We further analyzed Hi-C data with respect to long-range contacts within the nuclear space. Hi-  
198 C data revealed that DSBs cluster together (Fig. 3A), as previously observed using Capture Hi-  
199 C<sup>17</sup>. Such increased long range contacts take place between entire chromatin domains covered  
200 by γH2AX and could happen between γH2AX-decorated TADs induced on the same  
201 chromosome (Fig. S4A) as well as on different chromosomes (Fig. S4B). Of interest, some  
202 γH2AX domains were able to interact with more than a single other γH2AX domain (Fig. 3B,  
203 black arrows). Notably, this ability to form clusters of multiples TADs (also known as TADs  
204 cliques<sup>37</sup>) upon DSB induction, correlated with γH2AX, 53BP1 and ubiquitin chains levels as

**A****B****C****D****E****F**

**Figure 3: Cell cycle regulated, ATM-dependent but cohesin- and DNA-PK-independent clustering of damaged-TADs.**

**(A)** Hi-C contact matrix of the log<sub>2</sub> (+DSB/-DSB) on a region of the chromosome 1 at two different resolutions: 250 kb (left panel) and 100 kb (right panel). γH2AX ChIP-seq track following DSB induction is shown on the top panel and on the right. One representative experiment is shown. **(B)** Hi-C contact matrix of the log<sub>2</sub> (+DSB/-DSB) on a region of the chromosome 17 at 250 kb resolution. γH2AX and 53BP1 ChIP-seq tracks following DSB induction are shown on the top panel and on the left. The black arrows indicate clustering of one DSB on the chromosome 17, with several other DSBs on the same chromosome. One representative experiment is shown. **(C)** γH2AX domains were categorized based on their propensity to not interact with any other γH2AX domain (single), with one other γH2AX domain (TAD-TAD) or with multiple other γH2AX domains (TAD cliques containing 3 to 6 DSBs). ChIP-seq levels of γH2AX (+DSB), 53BP1 (+DSB), H1 (log<sub>2</sub> +DSB/-DSB) or Ubiquitin chains detected with the FK2 antibody (log<sub>2</sub> +DSB/-DSB) within the corresponding domains were computed across each category. **(D)** Left panel: Hi-C contact matrix of the log<sub>2</sub>(+DSB/-DSB) upon Ctrl (upper right) or SCC1 depletion (lower left). A region of the chromosome 1 is shown at 250 kb resolution. The γH2AX ChIP-seq track following DSB induction is shown on the top and on the right. Right panel: magnification of the black square, showing Hi-C contacts between the two γH2AX domains. **(E)** Hi-C contact matrix of the log<sub>2</sub> (+DSB/-DSB) without inhibitor, with a DNA-PK inhibitor or with an ATM inhibitor as indicated. A region of the chromosome 1 is shown with a 250 kb resolution. γH2AX ChIP-seq track following DSB induction is shown on the top. Bottom panel: magnification, showing Hi-C contacts between the two γH2AX domains. **(F)** Trans interactions (log<sub>2</sub> ratio +DSB/-DSB) between the view point and the other DSBs (n=79) were computed from 4C-seq experiments in synchronized cells (G1, S and G2 as indicated). Three cluster-prone DSBs, one not cluster-prone and one control undamaged locus were used as viewpoints. P, non-parametric paired wilcoxon test.

205 well as the depletion of histone H1 around DSB (Fig. 3C), which are all DSB-induced  
206 chromatin features that occur at the scale of an entire TAD<sup>2</sup>.

207 We further examined the effect of cohesin depletion on damaged TAD clustering. Inspection  
208 of individual DSBs indicated that SCC1 depletion by siRNA did not alter clustering (Fig. 3D).  
209 Quantification of *trans* interactions between all DSBs also indicated that SCC1 depletion did  
210 not modify the ability of damaged TAD to physically interact together (Fig. S4C). Additionally,  
211 we found that inhibition of ATM compromised DSB clustering, whilst inhibiting DNA-PK  
212 activity triggered an important increase in DSB clustering (Fig. 3E, Fig. S4D).

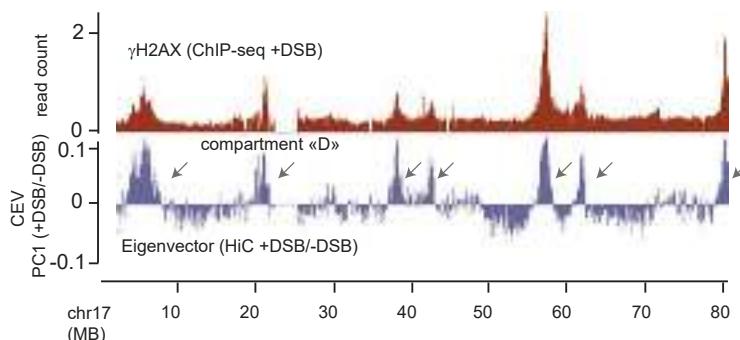
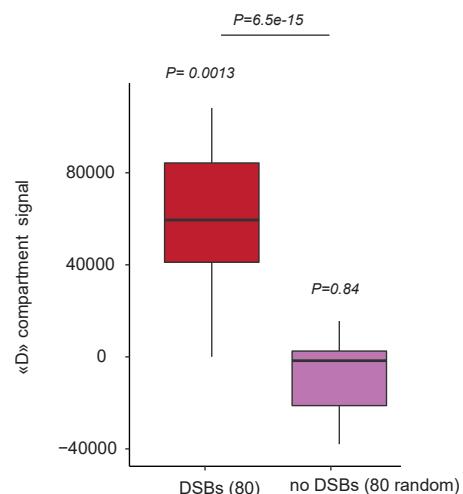
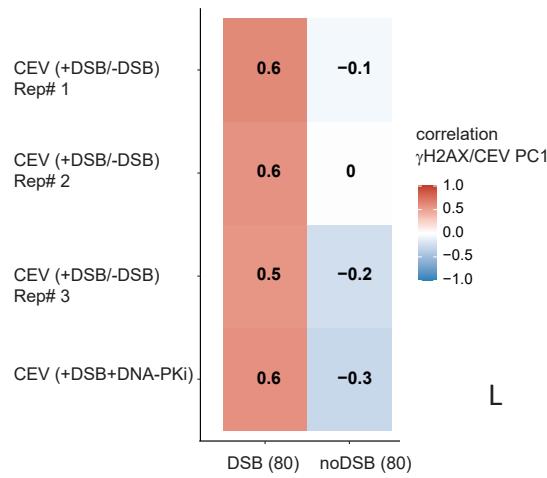
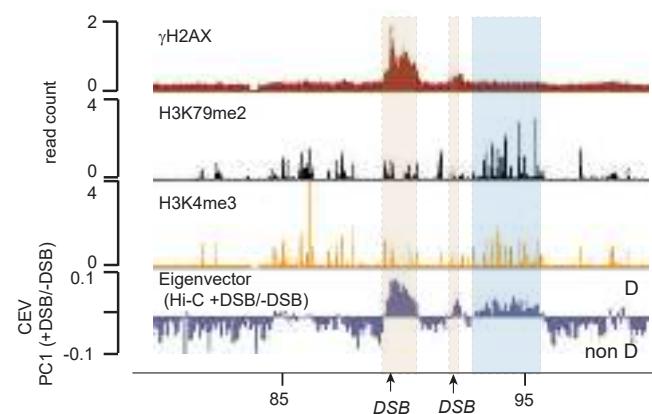
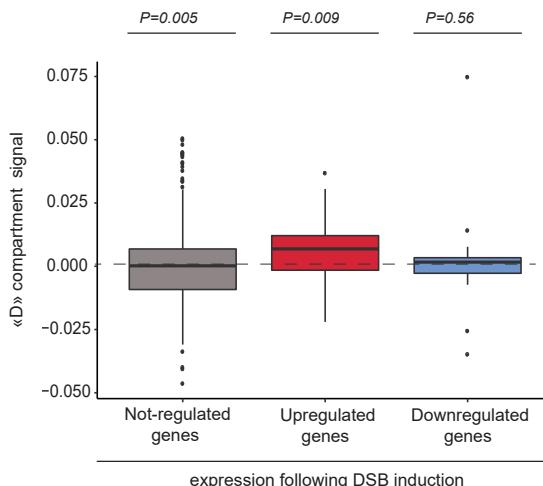
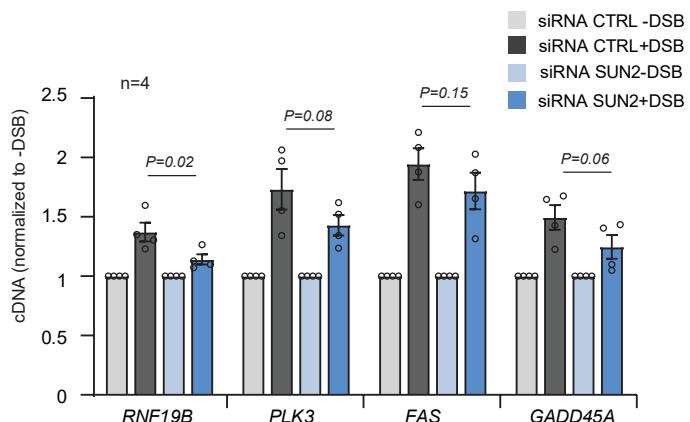
213 We previously reported that DSB clustering occurs preferentially during the G1 phase of the  
214 cell cycle<sup>17</sup>, in agreement with previous work using laser-induced damage coupled with live-  
215 imaging<sup>15</sup>. However, another recent study performed using microscopy in DIvA cells, described  
216 DSB clustering as a process occurring specifically in post-replicative G2 cells and involved in  
217 Homologous Recombination<sup>18</sup>. In order to address the cell cycle regulation of DSB clustering  
218 we turned to 4C-seq experiments. Indeed, we found that DSB clustering (*i.e.* damaged TAD-  
219 TAD interaction) can be readily detected by 4C-seq when using a DSB as a view point, as  
220 shown by the increase of 4C-seq signal observed on other DSBs induced on the genome (Fig.  
221 S4E). Indeed, *trans* 4C-seq signal was significantly detected post-DSB induction on DSBs but  
222 only when the DSB used as a viewpoint was a “clustering-prone” DSB (Fig. S4F)<sup>17</sup>. This was  
223 not observed using an undamaged control locus as a view point (Fig. S4F). Finally, in agreement  
224 with our Hi-C data, 4C-seq performed upon DNA-PK inhibition also showed increased  
225 clustering (Fig. S4G-H) further validating our ability to properly detect clustering by 4C-seq.  
226 Importantly, 4C-seq performed before and after DSB induction using four DSBs and a control  
227 locus as viewpoints in synchronized cells indicated that DSB clustering was readily detectable  
228 during G1 but strongly reduced during the other cell cycle stages (Fig. 3F, Fig. S4I).

229 Taken altogether, our results indicate that upon DSB formation, the TADs that carry DSBs are  
230 able to physically contact each other in the nuclear space (*i.e.* cluster) in a manner that is entirely  
231 dependent on ATM, exacerbated upon DNA-PK inhibition, and mostly independent of the  
232 cohesin complex. Damaged TAD clustering mostly takes place in G1 and correlates with TAD-  
233 scale DSB-induced chromatin modifications ( $\gamma$ H2AX, Ubiquitin accumulation and H1  
234 depletion) as well as 53BP1 accumulation.

235

### 236 **A new “D” compartment is formed following DSB induction**

237 Previous work identified the existence of two main, spatially distinct, chromatin  
238 “compartments” that, in mammalian nuclei, correspond to active and transcriptionally  
239 competent chromatin (the “A” compartment) and repressed heterochromatin (“B”  
240 compartment)<sup>38</sup>. These chromatin compartments were determined by Principal Component  
241 Analysis (PCA) of Hi-C chromosomal contact maps and can be visualized linearly using  
242 eigenvectors. We applied PCA analysis on differential Hi-C maps (*i.e.* contact matrices of  
243 +DSB/-DSB) on each individual chromosome. The first Chromosomal Eigenvector (CEV,  
244 PC1) allowed us to identify a DSB-induced chromatin compartment mainly on chromosomes  
245 displaying a large number of DSBs (chr1,17 and X) (Fig. 4A, Fig. S5A). Notably, a similar  
246 analysis on Hi-C maps generated upon DNA-PK inhibition, which impairs repair<sup>25</sup> and  
247 increases DSB clustering (Fig. 3), allowed to identify this compartment on more chromosomes  
248 (such as chr6 for instance, Fig. S5A bottom panel). This compartment displayed a very strong  
249 correlation with  $\gamma$ H2AX-decorated chromatin following DSB (Fig. 4A-C, Fig. S5A) and was  
250 henceforth further named “D” compartment (for DSB-induced compartment). Yet, additional  
251 analysis revealed that the D compartment is not solely generated through the clustering of  
252 damaged chromatin (*i.e.* TADs that carry DSBs and are enriched in  $\gamma$ H2AX). Indeed, we could

**A****B****C****D****E****F****Figure 4. Formation of a DSB-specific «D compartment».**

(A) Genomic tracks of  $\gamma$ H2AX ChIP-seq and first Chromosomal eigenvector (CEV) computed on differential (+DSB/-DSB) Hi-C matrix of chromosome 17 using a 100kb resolution (blue). Genomic regions displaying a positive CEV signal belong to the DSB-induced «D compartment» (black arrows). (B) Boxplot representing the quantification of the «D compartment» signal on a 1Mb window around the 80 best induced DSBs (red) and around 80 control undamaged regions (random, purple). P, non parametric Wilcoxon test. (C) Pearson correlation between  $\gamma$ H2AX ChIP-seq and D compartment (positive CEV PC1 on differential Hi-C matrices) for three biological Hi-C replicates (Rep#1, Rep#2 and Rep#3) and the Hi-C performed in presence of DNAPK inhibitor as indicated. (D) Genomic tracks of  $\gamma$ H2AX (red), H3K79me2 (black) and H3K4me3 (yellow) ChIP-seq, and the first Chromosomal Eigenvector computed on the differential Hi-C (CEV, blue). The brown squares highlight genomic regions present in D compartment that carry a DSB and are enriched in  $\gamma$ H2AX. In contrast the blue square shows a genomic region that is devoid in  $\gamma$ H2AX and DSB, but is nevertheless found in the D compartment. (E) Boxplot showing the quantification of the «D compartment» signal computed from Hi-C data (+DSB+DNA-PK $\iota$ /-DSB) on genes that are not regulated following DSB induction (Not-regulated genes, grey), genes that are upregulated following DSB induction (Upregulated genes, red) or genes that are downregulated following DSB induction (Downregulated genes, blue), identified by RNA-seq. (F) RT-qPCR quantification of the expression level of four genes (RNF19B, FAS, PLK3 and GADD45A) before and after DSB induction in cells transfected with control or SUN2 siRNA. n=4 independent experiments.

253 identify chromatin domains, not containing any DSB and not decorated by γH2AX, that  
254 associate with the D compartment after damage (Fig. 4D). Correlation analysis using  
255 chromosomes 1,17 and X, on which the D compartment was readily detected, indicated that  
256 non-damaged loci which tend to associate with the D compartment are enriched in H2AZac,  
257 H3K4me3 and H3K79me2 (Fig. S5B, Fig. 4D). Conversely, chromatin domains associated with  
258 repressive marks such as H3K9me3 displayed a negative correlation with D compartment (Fig.  
259 S5B). A similar trend was observed when D compartment was computed from the Hi-C data  
260 obtained in presence of the DNA-PK inhibitor (Fig. S5B bottom panel) and correlation analysis  
261 performed on all chromosomes showing D compartmentalization (*i.e.*, chr 1,2,6,9,13,17,18,20  
262 and X). Altogether our data indicate that upon DSB production on the genome, damaged TADs,  
263 covered by γH2AX/53BP1, form a new chromatin compartment that segregates from the rest  
264 of the genome and further recruits additional chromatin domains enriched in transcribed genes.

265 In order to decipher the nature of such active genes targeted to the D compartment, we further  
266 explored the DNA motifs enriched on “D” genes compared to “non D” genes, *i.e.* genes  
267 recruited to the D compartment, versus the one that do not display targeting to the D  
268 compartment (discarding all genes directly comprised in γH2AX domains). Notably, the top  
269 enriched motifs included OSR1, TP73, Nkx3.1 and E2F binding sites, which are tumor  
270 suppressor and /or known to be involved in the DNA damage response (Fig. S5C)<sup>39–42</sup>,  
271 suggesting a direct physical targeting of DNA damage responsive genes to the “D”  
272 compartment. In agreement, visual inspection revealed that known p53 target genes which are  
273 upregulated in DIVA cells following DSB induction were associated with the D compartment  
274 (Fig. S5D). To test the hypothesis that DNA damage responsive genes are recruited to the D  
275 compartment, we therefore retrieved genes that are upregulated following DSB induction using  
276 our RNA-seq data generated before and after DSB induction. Notably, upregulated genes  
277 displayed a higher D compartment signal than the genes that were either not regulated or

278 downregulated after DSBs (Fig. 4E). Similarly, known direct p53 target genes which are  
279 significantly upregulated after DSB in our experimental conditions displayed higher D  
280 compartment signal compared to those unaffected (Fig. S5E) and this was also the case when  
281 using NF- $\kappa$ B target genes, another major transcription factor regulating DNA damage  
282 response<sup>43</sup> (Fig. S5F). In order to determine whether recruitment of those genes to the D  
283 compartment contribute to their activation following DNA damage, we investigated the  
284 consequence of disrupting DSB clustering (and hence formation of D compartment) by  
285 depleting the SUN2 component of the LINC complex, previously found as a clustering  
286 promoting factor<sup>17,19</sup>. SUN2 depletion altered the transcriptional activation of genes found to  
287 be upregulated and targeted to the D compartment upon DSB in DIvA cells (Fig. 4F).

288 Altogether these data indicate that DSB induction triggers the formation of a novel chromatin  
289 compartment that comprises not only damaged TADs, decorated by  $\gamma$ H2AX and 53BP1, but  
290 also genes upregulated following DNA damage, for which targeting to D compartment is  
291 required for optimal activation. Altogether this suggests a role of the D compartment, and hence  
292 DSB clustering, in the activation of the DNA Damage Response.

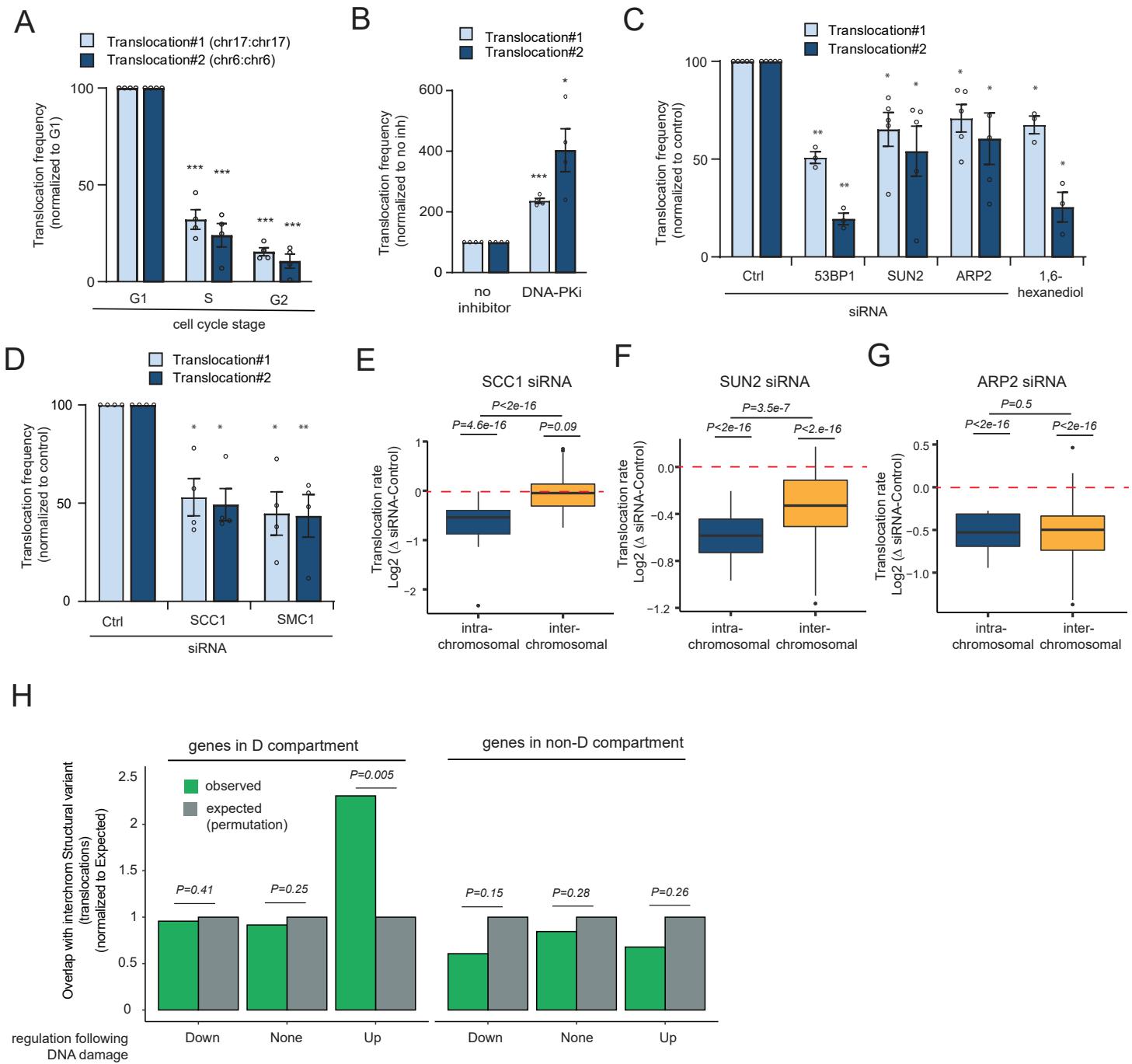
293

294 **DSB-induced reorganization of chromosome folding can favor translocations.**

295 Importantly, while our above data suggest a beneficial role of DSB clustering in potentiating  
296 the DDR by allowing to directly recruit DDR genes to a novel chromatin compartment, it may  
297 also be detrimental, since bringing two DSBs in a close proximity may fosters translocations  
298 (illegitimate rejoining of two DSBs), as previously proposed<sup>16</sup>. We therefore assessed by qPCR  
299 the frequency of translocations events occurring in DIvA cells post-DSB induction, in  
300 conditions where we found altered DSBs clustering and D compartment formation.

301 Notably, translocations are increased in G1 compared to S/G2-synchronized cells (Fig. 5A), in  
302 agreement with an enhanced DSB clustering observed in G1 cells (Fig. 3F). Moreover, DNA-  
303 PK inhibition, that increased D-compartment formation (Fig. 3E, Fig. S4D, Fig. S5A) also  
304 strongly increased translocation frequency (Fig. 5B). On another hand, depletion of 53BP1 (Fig.  
305 S6A), previously found to mediate repair foci phase separation<sup>20</sup>, as well as a treatment with  
306 1,6-hexanediol, which disrupts phase condensates (Fig. S6B), decreased translocations (Fig.  
307 5C). Similarly, depletion of SUN2, a member of the LINC complex and of ARP2, an actin  
308 branching factor (Fig. S6A), that we and others reported as mediating DSB clustering<sup>17-19</sup>,  
309 decreased translocations (Fig. 5C). Surprisingly, depletion of the cohesin subunits SMC1 or  
310 SCC1 also decreased translocation frequency (Fig. 5D, Fig S6C). This was unexpected since  
311 SCC1-depleted cells do not display clustering defects (Fig. 3).

312 Given that the two translocations assessed by our qPCR assay are both intra-chromosomal  
313 translocations (*i.e.* rejoining of two distant DSBs located on the same chromosome) we  
314 hypothesized that translocation frequency at the intra-chromosomal level may also be regulated  
315 by the DSB-induced loop extrusion that depends on the cohesin complex. In order to investigate  
316 more broadly translocation events between multiple DSBs induced in the DIvA cell line, we  
317 designed a novel multiplexed amplification protocol followed by NGS sequencing. In control  
318 cells, we could readily detect increased translocation frequency upon induction of DSB  
319 compared to control genomic locations (Fig. S6D). Strikingly, depletion of SCC1 decreased the  
320 frequency of intra-chromosomal translocations, while leaving inter-chromosomal  
321 translocations unaffected (Fig. 5E). In contrast depletion of SUN2 and ARP2 decreased both  
322 intra- and inter-chromosomal translocations (Fig. 5F-G). Taken altogether these data suggest  
323 that both the DSB-induced loop extrusion and the formation of the D-compartment through  
324 clustering of damaged TADs, display the potential to generate translocations.

**Figure 5. DSB-induced loop extrusion and D-compartment formation drive translocations.**

(A) qPCR quantification of translocations frequency for two independent translocations following DSB induction in cells synchronized in the G1, S or G2 phase ( $n=4$  independent replicates).  $P=$  paired t-test, \*  $P<0.05$ , \*\*  $P<0.001$ , \*\*\* $P<0.0005$ . (B) qPCR quantification of translocations frequency for two independent translocations following DSB induction with or without DNA-PK inhibitor ( $n=4$  independent replicates). (C) qPCR quantification of translocations frequency for two independent translocations following DSB induction in Control, 53BP1, SUN2 or ARP2 depleted cells or upon 1,6-Hexanediol treatment ( $n\geq3$  independent replicates). (D) As in (C) but upon Control, SMC1 or SCC1 depletion ( $n=4$  independent replicates). (E) Intra-chromosomal (blue) or inter-chromosomal translocations (yellow) were quantified using multiplexed amplification followed by high throughput sequencing (amplicon-seq) between 20 different DSBs induced in DlvA cell line, upon Ctrl or SCC1 depletion ( $\log_2 \Delta \text{siSCC1/siC-TRL}$ ) ( $n=4$  independent replicates). P-values, non-parametric wilcoxon test tested against  $\mu=0$ . intra vs inter-chromosomal, P=paired wilcoxon test. (F) As in (E) but the quantification was performed in SUN2 depleted cells ( $n=4$  independent replicates). (G) As in (E) but the quantification was performed in ARP2 depleted cells ( $n=4$  independent replicates). (H) Observed (green) and expected (obtained through 1000 permutations) overlap between breakpoint positions of inter-chromosomal translocations identified on cancer genomes (from (Zhang et al., 2018)) and genes targeted to the D compartment, either upregulated, downregulated or not regulated following DSB induction (identified by RNA-seq) as indicated, compared to their counterparts not targeted to the D compartment.

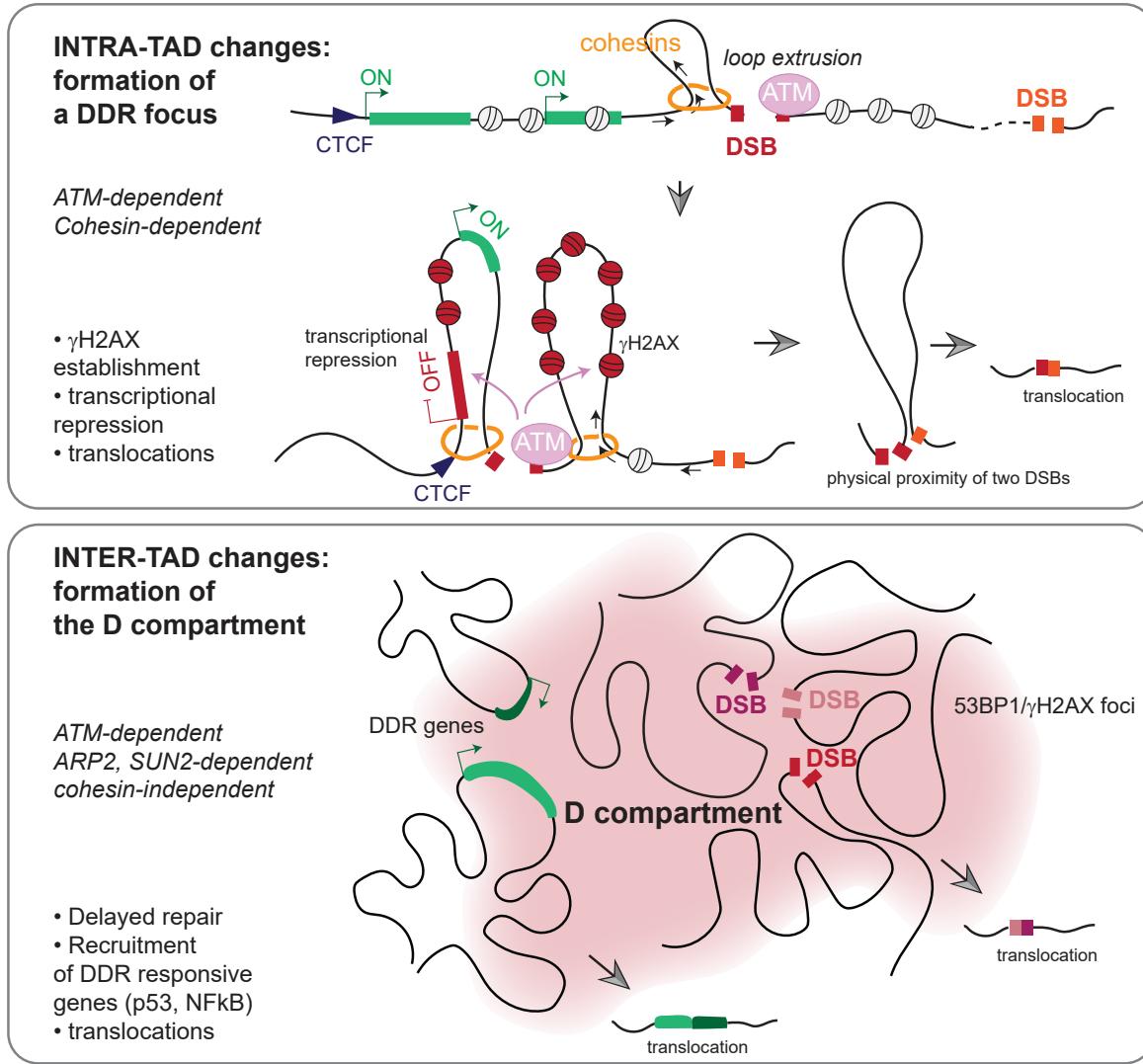
325 Given our finding that additional undamaged loci (including genes upregulated following DSB  
326 induction) can be physically targeted to the D compartment after break induction, we further  
327 hypothesized that such a physical proximity may account for some of the translocations  
328 observed on cancer genomes. We retrieved breakpoint positions of inter-chromosomal  
329 translocations of 1493 individuals across 18 different cancers types (from<sup>44</sup>), and assessed their  
330 potential overlap with genes targeted to the D compartment (reproducibly detected in the three  
331 Hi-C replicates on chr1,17 and X, on which D compartment could be identified accurately). D-  
332 targeted genes were further sorted as either upregulated, downregulated or not significantly  
333 altered following DSB induction, and compared to their counterparts not targeted to the D  
334 compartment. We found that genes that are upregulated following DSB induction and that are  
335 targeted to the D compartment displayed a significant overlap with translocations breakpoints,  
336 in contrast to genes that are not targeted to the D compartment (non-D) (Fig. 5H). Altogether  
337 these data indicate that the relocalization of upregulated genes during the DNA Damage  
338 response in the DSB-induced D compartment likely accounts for some of the translocations  
339 detected on cancer genomes.

340

## 341 **Discussion**

342 Our work allowed us to comprehensively decipher how chromosome folding is dynamically  
343 rearranged following DSB, impacting both the positioning of DSB within the cell nucleus as  
344 well as transcriptional activation and repression events which are mounted following DSB. Our  
345 data suggest that chromosome architecture is an integral component of the DNA Damage  
346 Response, but also acts as a double-edged sword which can challenge genomic integrity.

347 Based on our data, we would like to propose the following model (Fig. 6): DSBs elicit ATM-  
348 induced loop extrusion by cohesin at the break itself, allowing to establish γH2AX on the entire

**Figure 6. Model.**

**(A) INTRA-TAD changes:** Cohesin-dependent loop extrusion arising at DSB ensures ATM-dependent phosphorylation of H2AX on an entire TAD, allowing fast establishment of a TAD-scale DDR focus (Arnould et al., 2021). Genes embedded in the DDR focus that pass through the loop during the ongoing DSB-anchored loop extrusion process are not repressed. However, genes that carry a CTCF-bound site in the proper orientation, acting as a roadblock which stops/pauses loop extrusion, are brought to 3D proximity of the DSB and hence subjected to ATM-dependent down-regulation. On the other hand, DSB-anchored loop extrusion also displays the potential to bring in close proximity two DSBs located on the same chromosome, which can favor the occurrence of intra-chromosomal translocations. **(B) INTER-TAD changes:** Once assembled, the decorated γH2AX/53BP1 damaged TADs can further fuse together presumably by phase separation given previously demonstrated LLPS properties of 53BP1. This creates a new nuclear, DSB-induced compartment, in which the DNA damage responsive genes, such as p53 targeted genes, physically relocate. The ARP2 Actin branching factor, and the SUN2 LINC member also contribute to the formation of this compartment. D-compartment mainly form when DSBs persist, for example in G1, when Homologous Recombination is not available to repair DSB induced in active loci<sup>17</sup> or upon DNA-PK inhibition<sup>25</sup>. On the other hand, induced spatial proximity in D compartment also increases the frequency of translocations between DSBs or between the DDR responsive genes.

349 TAD<sup>12</sup>. Doing so, loop extrusion can trigger transcriptional repression at a distance from DSB  
350 sites, mainly on genes harboring CTCF sites with a specific orientation relative to the DSB.  
351 γH2AX deposition via loop extrusion further facilitates recruitment of downstream effectors  
352 such as MDC1 and 53BP1, which also span entire TADs. Then, 53BP1-decorated TADs can  
353 coalesce due, at least in part, to liquid-liquid phase separation (LLPS) properties of  
354 53BP1<sup>20,21,45</sup>, giving rise to a novel chromatin compartment. This DSB-induced (“D”)  
355 compartment further recruits genes involved in the DNA damage response and contributes to  
356 their activation. We propose that the formation of the “D” may allow to precisely tune the  
357 magnitude of the DDR with respect to DSB load and persistency, providing a function for these  
358 enigmatically large γH2AX/53BP1-decorated chromatin domains and to DSB clustering. Yet,  
359 this comes at the expense of potential translocations, as both loop extrusion and coalescence of  
360 damaged TAD are able to bring linearly distant DSBs in close physical proximity.

361

### 362 **Chromosome folding is reorganized within TAD that experience a DSB**

363 Our data indicate that a TAD that experiences a DSB is profoundly altered, showing enhanced  
364 intra-TAD dynamics and isolation from neighboring TADs. The DSB itself experiences  
365 reduced interaction with surrounding self-interacting domains and is likely to be buried within  
366 the damaged domain. This is in good agreement with the ring-shaped, ordered circular  
367 arrangement of adjacent chromatin surrounding the DSB recently described using super  
368 resolution STED microscopy<sup>46</sup>. Of importance, cohesin-mediated loop extrusion initiated at the  
369 DSB entirely depends on ATM kinase activity. Thus, the presence of the DSB may not by itself  
370 be sufficient to induce loop extrusion at damage sites (*i.e.*: by halting already ongoing loop  
371 extrusion processes by either DNA discontinuity or repair factor hindrance). Cohesin subunits  
372 are themselves phosphorylated by ATM<sup>12,47</sup>, which can modify cohesin retention on the

373 genome<sup>48,49</sup>. It is therefore tempting to speculate that cohesin phosphorylation by ATM  
374 accounts for DSB-induced loop extrusion. Notably, DNA-PK inhibition largely enhanced DSB-  
375 induced loop extrusion and TAD strengthening. Such a strong effect would be in agreement  
376 with the ability of this DNA-PK inhibitor to severely impair DSB repair<sup>25</sup>, leading to persistent  
377 breaks sustaining continuous loop extrusion. Yet, the precise mechanism underlying these  
378 apparently antagonistic roles for the two main DDR kinases on loop extrusion at DSB and  
379 subsequent reorganization of damaged TADs remains to be fully elucidated.

380

### 381 **DSB-induced loop extrusion contributes to local transcriptional repression**

382 Our data also suggest that loop extrusion is instrumental for local transcriptional repression.  
383 Indeed, transcriptional silencing is initiated in *cis* to the break and mediated by RNAPII and  
384 histone modifications, triggered by the concerted action of ATM<sup>26</sup>, DNA-PK<sup>27,50</sup> and PARP<sup>51</sup>.  
385 Yet, the exact span of such transcriptional response has been subjected to debate due to  
386 apparently conflicting results. Genome-wide RNAPII mapping could identify a clear eviction  
387 of RNAPII occupancy for directly damaged genes but not necessarily for DSB-surrounding  
388 genes<sup>28</sup> and RNA-seq experiments (using the I-PpoI system in mouse cells) showed that overall  
389 transcription is maintained in γH2AX domains following DSB induction<sup>52</sup>. However, inducing  
390 DSBs at ~10kb from a reporter gene with FokI indicated that DSB can inhibit transcription at  
391 such a distance<sup>26</sup> and transcriptome analysis by RNAPII mapping and Bru-seq suggested a  
392 correlation between the distance to DSB and the extent of transcriptional shutdown within  
393 γH2AX domains<sup>32</sup>. Here we reconcile these data, by showing that downregulated genes within  
394 γH2AX domains exhibit 3D proximity with the DSB. Our results suggest that during loop  
395 extrusion, adjacent genes are brought within physical proximity of the DSB. The presence of a  
396 CTCF-bound locus in the appropriate orientation relative to the DSB would halt the extrusion

397 process long enough, resulting in sustained proximity of the CTCF-bound gene with ATM (for  
398 which binding is restricted to the immediate vicinity to the DSB<sup>12</sup>). ATM could then elicit  
399 transcriptional silencing through its effect on chromatin by promoting deposition of repressive  
400 histone modifications such as ubiquitylation of H2A<sup>26</sup>, or removal of activating marks such  
401 H3K4me3<sup>53</sup>. Conversely, genes that do not carry a CTCF-bound locus or a CTCF site but in  
402 the wrong orientation, could pass through and be expelled at a physical distance of the DSB,  
403 therefore escaping ATM-mediated transcription silencing. This model is in line with recent data  
404 identifying the cohesin complex, and thus chromosome folding, as a main player in the local  
405 transcriptional silencing near DSB<sup>14</sup>. More broadly, this highlights the fact that chromosome  
406 architecture may regulate transcription in many different ways. Indeed, loop extrusion is largely  
407 considered to favor or restrict contacts between promoters and regulatory elements such as  
408 enhancers<sup>54</sup>. Whether this ability of cohesin to “pull” genes or chromatin segments towards  
409 localized repressive nuclear microenvironment in stress conditions other than DNA damage will  
410 require further studies.

411

#### 412 **DSBs induce the formation of a novel chromatin compartment**

413 Importantly, we found that following DSB induction, γH2AX/53BP1-decorated TADs further  
414 hold the capacity to coalesce, in agreement with previous live imaging studies showing DSB  
415 clustering/repair foci fusion<sup>15,16,18,20,25</sup>. Notably, clustering of DSBs takes place between entire  
416 γH2AX/53BP1-decorated domains and indeed correlates with other DSB-induced chromatin  
417 modifications that encompass the entire TAD, such as the depletion of the histone H1 or the  
418 accumulation of ubiquitin chains<sup>2</sup>. Furthermore, this DSB clustering mostly occurs in G1 (Fig.  
419 3F), in agreement with previous imaging data<sup>15,17</sup>.

420 PCA analysis on Hi-C matrices previously allowed to discover the existence of the two main  
421 chromatin compartments (“A” and “B”). This A/B compartmentalization is giving rise to the  
422 so-called “plaid pattern” on Hi-C maps. The B compartment (heterochromatic foci) has been  
423 suggested to arise from liquid-liquid phase separation (LLPS) promoted by HP1<sup>55,56</sup> and more  
424 generally, chromosome organization at the supermegabase scale is believed to stem from phase  
425 separation<sup>57,58</sup>. Here, using a similar approach, we uncovered a novel “D” compartment  
426 specifically induced following DSB production on the genome. “D” compartment would arise  
427 when γH2AX/53BP1-decorated domains established by loop extrusion self-segregate from the  
428 rest of chromatin at least in part thanks to the LLPS properties of 53BP1<sup>20,21,45</sup>. In the context  
429 of more permanent damage, such as upon DNA-PK inhibition (which triggers a strong repair  
430 defect<sup>25</sup>), or in G1 (where DSB in active genes are left unrepaired<sup>17</sup>), DSB would sustain  
431 continuous loop extrusion and γH2AX/53BP1 accrual, further fostering LLPS-mediated D-  
432 compartment formation and DSB clustering.

433 Since SUN2 (a component of the Nuclear Envelope) and ARP2/3 (actin organizers) were shown  
434 to mediate DSB clustering<sup>17-19</sup>, the contribution of the NE and the nuclear skeleton in the  
435 formation of the D compartment, will require additional investigations. Of note, the interaction  
436 between phase-separated heterochromatic regions (B compartment) and the nuclear lamina was  
437 found to be essential to build the conventional chromosome architecture in nuclei<sup>58</sup>, suggesting  
438 that a similar mechanism might apply for the formation of the DSB-induced chromatin  
439 compartment.

440

#### 441 **The D compartment participates in the DDR**

442 Previous work identified 53BP1 as critical for p53 target genes activation<sup>59</sup> and disrupting  
443 53BP1 droplet formation was found to alter checkpoint activation<sup>20</sup>. Conversely, enhanced

444 53BP1 phase separation led to an elevated p53 response<sup>60</sup> as did the loss of TIRR, a protein that  
445 binds the Tudor domain of 53BP1 and regulates its association to DSBs<sup>61,62</sup>. Altogether these  
446 studies indicated that the formation of 53BP1 foci is both necessary and sufficient to activate  
447 the p53 response. Here we found that p53-responsive genes are targeted to the D compartment  
448 composed of fused 53BP1 foci. Our data suggest that such a physical targeting of DNA damage  
449 responsive genes is instrumental for their regulation, given that impaired DSB clustering  
450 following SUN2 depletion triggered decreased transcriptional activation of DDR genes.  
451 Notably, this was not only the case for p53-responsive genes, but also for a number of other  
452 upregulated genes. This relocation of DDR genes within the same compartment as damaged  
453 chromatin could allow to precisely tune the magnitude of the DDR with respect to the actual  
454 DSB load. Furthermore, this observation may provide a rationale for why so many transcription  
455 factors (including p53) were found recruited at DSBs repair foci<sup>63</sup>. While initially thought to  
456 allow chromatin remodeling in order to enhance DSB repair, we rather suggest that TF  
457 recruitment to DSB repair foci reflects the relocalization of DDR genes within the D  
458 compartment (hence at physical proximity of the DSB). We believe that the same concept may  
459 also applies to 53BP1 bodies, that form in G1 following replicative stress during the previous  
460 S-phase, which were initially described as transcriptional bodies (or OPT bodies for Oct-1, PTF,  
461 transcription)<sup>64,65</sup>.

462

#### 463 **DSB-induced remodeling of chromosome folding contribute to translocations**

464 Numerous reports identified both transcriptional activity and chromosome folding as critical  
465 regulator of translocations, due to their key contribution in the generation of endogenous  
466 DSBs<sup>66–68</sup>. Beyond its influence on DNA fragility, spatial organization of chromosomes also  
467 hold the potential to bring two DSBs in close 3D proximity. In agreement, DSB clustering was

468 found as being a key determinant for the biogenesis of translocations between two I-SceI  
469 endonuclease-induced breaks<sup>16</sup>. In a different context, chromosome folding, and more  
470 specifically loop extrusion, also determines the rejoicing of distant DNA ends during V(D)J  
471 recombination and antibody class switch recombination<sup>69–71</sup>. Our study further indicates a  
472 major influence of DSB-induced chromosomal architecture in the generation of translocations,  
473 with DSB-induced loop extrusion mediating intra-chromosomal translocations and D  
474 compartment increasing both intra- and inter-chromosomal translocations. More importantly,  
475 we found that the genes upregulated in response to DSB and relocated to the D compartment  
476 displayed significant overlap with translocation breakpoints identified by whole genome  
477 sequencing in patient cancer samples. In agreement with an increased occurrence of structural  
478 variants on tumor suppressor genes<sup>44</sup>, we propose that the physical targeting of DNA damage  
479 responsive genes to the D compartment, by bringing DSBs and DDR genes in close spatial  
480 proximity, may occasionally trigger deleterious rearrangements on genes involved in the  
481 control of cell proliferation and apoptosis upon DNA damage, and may hence act as a critical  
482 driver of oncogenesis by disrupting the integrity of tumor suppressor genes.

483

484

485 **Methods**

486

487 **Cell culture and treatments**

488 DIvA (AsiSI-ER-U20S)<sup>28</sup> and AID-DIvA (AID-AsiSI-ER-U20S)<sup>72</sup> cells were grown in  
489 Dubelcco's modified Eagle's medium (DMEM) supplemented with 10% SVF (Invitrogen),  
490 antibiotics and either 1 µg/mL puromycin (DIvA cells) or 800 µg/mL G418 (AID-DIvA cells)  
491 at 37 °C under a humidified atmosphere with 5% CO<sub>2</sub>. To induce DSBs, cells were treated with  
492 300nM 4OHT (Sigma, H7904) for 4 h. For ATM or DNA-PK inhibition, cells were pretreated  
493 for 1 h respectively with 20µM KU-55933 (Sigma, SML1109) or 2µM NU-7441 (Selleckchem,  
494 S2638) and during subsequent 4OHT treatment. Treatment with 10% 1,6-hexanediol (Sigma,  
495 240117) was performed for 3 min before the end of the 4OHT treatment. For cell  
496 synchronization, cells were incubated for 18 h with 2 mM thymidine (Sigma, T1895), then  
497 released during 11 h, followed by a second thymidine treatment for 18 hr. S, G2 and G1 cells  
498 were then respectively treated with OHT at, 0, 6 or 11 h following thymidine release and  
499 harvested 4 h later. siRNA transfections were performed using the 4D-Nucleofector and the SE  
500 cell line 4D-Nucleofector X kit L (Lonza) according to the manufacturer's instructions, and  
501 subsequent treatment(s) were performed 48 h later. siRNA transfections were performed using  
502 a control siRNA (siCTRL): CAUGUCAUGUGUCACAUCAUCU; or using a siRNA targeting  
503 *SCC1* (siSCC1): GGUGAAAAUGGCAUUACGG; or *SMC1* (siSMC1):  
504 UAGGCUUCCUGGAGGUACAUUUAA; or *53BP1* (si53BP1):  
505 GAACGAGGAGACGGUAAUA; or *SUN2* (siSUN2): CGAGCCTATTCAAGACGTTCA; or  
506 *ARP2* (siARP2): GGCACCGGGUUUGUGAAGU.

507 **Translocation assay**

508 Translocation assays after siRNA transfection or 1,6-Hexanediol treatment were performed at  
509 least in triplicates in AID-DIVa cells as described in<sup>73</sup>. Translocation assay in synchronized  
510 cells was performed in DIVa cells following a 4OHT treatment (n=4 biological replicates). Two  
511 different possible translocations between different AsiSI sites were assessed by qPCR using the  
512 following primers: Translocation1\_Fw: GACTGGCATAAGCGTCTTCG,  
513 Translocation1\_Rev: TCTGAAGTCTGCGCTTCCA, Translocation2\_ Fw:  
514 GGAAGCCGCCAGAATAAGA, Translocation2\_Rev: TCTGAAGTCTGCGCTTCCA.  
515 Results were normalized using two control regions, both far from any AsiSI sites and  $\gamma$ H2AX  
516 domain using the following primers: Ctrl\_chr1\_82844750\_Fw:  
517 AGCACATGGGATTTCGCAGG, Ctrl\_chr1\_82844992\_Rev:  
518 TTCCCTCCTTGTGTACCA, Ctrl\_chr17\_9784962\_Fw:  
519 ACAGTGGGAGACAGAAGAGC, Ctrl\_chr17\_9785135\_Rev:  
520 CTCCATCATCGCACCCTTG. Normalized translocation frequencies were calculated using  
521 the Bio-Rad CFX Manager 3.1 software<sup>69</sup>.

## 522 **Amplicon -seq**

523 AID-DIVa cells were treated with or without 300nM 4OHT for 4 h followed by treatment with  
524 indole-3-acetic acid for 14 h. Cells were then lysed in cytoplasmic lysis buffer (50mM HEPES  
525 pH7.9, 10mM KCl<sub>2</sub>, 1.5mM MgCl<sub>2</sub>, 0.34M sucrose, 0.5% triton X-100, 10% glycerol, 1mM  
526 DTT) for 10 minutes on ice, then washed once in cytoplasmic lysis buffer before lysis in  
527 genomic extraction buffer (50mM Tris pH8.0, 5mM EDTA, 1% SDS, 0.5mg/mL proteinase K).  
528 Lysate was incubated at 60°C for 1 h. Genomic DNA was then ethanol precipitated on ice for  
529 1h, pelleted at 19,000g for 20 min and washed twice in 75% ethanol. Genomic DNA was then  
530 used in a multiplex PCR reaction that amplified 25 target sites; 20 AsiSI cut sites and 5 uncut  
531 control sites (Supplementary Table 1). Amplicons were size selected using SPRIselect beads

532 (Beckman, B23318) and subjected to DNA library preparation via the NEBNext Ultra II kit  
533 (NEB, E7645L). Libraries were pooled at equimolar concentrations and sequenced via an  
534 Illumina NextSeq 500 system using paired end 150 cycles. The data was analyzed via our  
535 custom tool mProfile, available at [github.com/aldob/mProfile](https://github.com/aldob/mProfile). This identified the genomic  
536 primers used in the original genomic PCR reaction to amplify each read in the pair. Translocated  
537 reads were therefore identified as those where each read in a pair was amplified by a different  
538 primer set, and this was normalized to the total reads that were correctly amplified by these  
539 primer sets.

540 **RT-qPCR**

541 RNA was extracted from fresh DIvA cells before and after DSB induction using the RNeasy  
542 kit (Qiagen). RNA was then reverse transcribed to cDNA using the AMV reverse transcriptase  
543 (Promega, M510F). qPCR experiments were performed to assess the levels of cDNA using  
544 primers targeting RPLP0 (FW: GGCGACCTGGAAGTCCAAT; REV:  
545 CCATCAGCACACAGCCTTC), GPSM2 (FW: GTAAAGGACTGGATTGGCACA; REV:  
546 CTTCCAAGGCCAGCTCTAGG), DNMT3B (FW: GGCCACCTCAATAAGCTCG; REV:  
547 GTTGCCTGTTGTTGGTTTG), RNF19B (FW: CATCAAGCCATGCCACGAT; REV:  
548 GAATGTACAGCCAGAGGGC), PLK3 (FW: GCCTGCCGCCGGTT; REV:  
549 GTCTGACGTCGGTAGCCCG), FAS (FW: ATGCACACTCACCAGCAACA; REV:  
550 AAGAAGACAAAGCCACCCA) or GADD45A (FW: ACGATCACTGTCGGGTGTA;  
551 REV: CCACATCTCTGTCGTCGTC). cDNA levels were then normalized with RPLP0  
552 cDNA level, then expressed at the percentage of the undamaged condition.

553 **Immunofluorescence**

554 DIvA cells were grown on glass coverslips and fixed with 4% paraformaldehyde during 15 min  
555 at room temperature. Permeabilization step was performed by treating cells with 0,5% Triton

556 X-100 in PBS for 10 min then cells were blocked with PBS-BSA 3% for 30min. Primary  
557 antibodies targeting RNA PolI (Santa Cruz sc48385) or PML (Santa Cruz sc-966 (PG-M3))  
558 were diluted 1:500 in PBS-BSA 3% and incubated with cells overnight at 4°C. After washes in  
559 1X PBS, cells were incubated with anti-mouse secondary antibody (conjugated to Alexa 594 or  
560 Alexa 488, Invitrogen), diluted 1:1000 in PBS-BSA 3%, for 1h at room temperature. After a  
561 DAPI staining, Citifluor (Citifluor, AF-1) was used for coverslips mounting. Images were  
562 acquired with the software MetaMorph, using the 100X objective of a wide-field microscope  
563 (Leica, DM6000), equipped with a camera (DR-328G-C01-SIL-505, ANDOR Technology).

564 **Western Blot**

565 Western Blot experiments were performed as in<sup>12</sup> using primary antibody targeting SUN2  
566 (Abcam ab124916 1:1000), ARP2 (Abcam ab128934 1:1000), 53BP1 (Novus Biologicals  
567 NB100-305 1:1000), SCC1 (Abcam ab992 1:500) or SMC1 (Abcam ab75819 1:1000).

568 **RNA-seq**

569 RNA-seq was performed as described in<sup>73</sup>. RNA-seq were mapped in paired-end to a custom  
570 human genome (hg19 merged with ERCC92) using STAR. Count matrices were extracted using  
571 htseq-count with union as resolution-mode and reverse strand mode. Differential expression  
572 analysis was made on the count matrix using edgeR with two replicates per condition and  
573 differential genes were determined with log-ratio test (LRT). Whole genome coverage was  
574 computed using deeptools and bamCoverage to generate bigwig using bam files (without PCR  
575 duplicate suppression). Using a cutoff of 0.1 for the adjusted p-value and 0.5 log2 fold-change  
576 (~41% increase/decrease of expression), we were able to determine 286 up-regulated and 125  
577 down-regulated genes with 11 of them directly damaged by a DSB. Differential coverage  
578 between two conditions was performed using BamCompare from deeptools with setting binsize  
579 parameter at 50bp. To test the relationship between expression, change after damage (+DSB)

580 and genomic distance we computed the distance with DSB for genes inside damaged TADs  
581 (N=715). Genes were divided into distance category, either 5 (Fig. S3A) to test log-log linear  
582 relation with distance, or 9 (Fig. S3B) to test log-linear relation with distance. Log2FC was  
583 calculated by edgeR in differential expression analysis.

584 **4C-seq**

585 4C-seq experiments performed in synchronized cells or before and after DSB induction upon  
586 DNA-PK inhibition were performed as in<sup>12</sup>. Briefly, 10-15×10<sup>6</sup> DIvA cells per condition were  
587 cross-linked, lysed and digested with MboI (New England Biolabs). DNA ligation was  
588 performed using the T4 DNA ligase (HC) (Promega), and ligated DNA was digested again  
589 using NlaIII (New England Biolabs). Digested DNA was religated with the T4 DNA ligase  
590 (HC) (Promega) before to proceed to 4C-seq library preparation. 16 individual PCR reactions  
591 were performed in order to amplify ~800ng of 4C-seq template, using inverse primers including  
592 the Illumina adaptor sequences and a unique index for each condition (Supplementary Table  
593 2). Libraries were pooled and sent to a Nextseq500 platform at the I2BC Next Generation  
594 Sequencing Core Facility (Gif-sur-Yvette).

595 4C-seq data were processed as described in<sup>12</sup>. Briefly, bwa mem was used for mapping and  
596 samtools for sorting and indexing. A custom R script  
597 (<https://github.com/bbcf/bbcfutils/blob/master/R/smoothData.R>) was used to build the  
598 coverage file in bedGraph format, to normalize using the average coverage and to exclude the  
599 nearest region from each viewpoint. Differential 4C-seq data were computed using  
600 BamCompare from deeptools with binsize=50bp. Average of total Trans interactions between  
601 viewpoints and DSB were then computed using a 1Mb window around the breaks (80 best) and  
602 after exclusion of viewpoint-viewpoint (Cis) interactions.

603 **Hi-C**

604 Hi-C data obtained before and after DSB induction and upon CTRL or SCC1 depletion in DIVA  
605 cells were retrieved from<sup>12</sup>. Hi-C experiments with or without DSB induction and upon ATM  
606 or DNA-PK inhibition were performed in DIVA cells as in<sup>12</sup>. Briefly, 1 million cells were used  
607 per condition. Hi-C libraries were generated using the Arima Hi-C kit (Arima Genomics) by  
608 following the manufacturer instructions. DNA was sheared to an average fragment size of 350-  
609 400 pb using the Covaris S220 and sequencing libraries were prepared on beads using the NEB  
610 Next Ultra II DNA Library Prep Kit for Illumina and NEBNext Multiplex Oligos for Illumina  
611 (New England Biolabs) following instructions from the Arima Hi-C kit.

612 **Hi-C data analyses**

613 *Hi-C heatmaps.* Hi-C reads were mapped to hg19 and processed with Juicer using default  
614 settings (<https://github.com/aidenlab/juicer>). Hi-C count matrices were generated using Juicer  
615 at multiple resolutions: 100 kb, 50 kb, 25 kb, 10 kb and 5 kb. Hi-C heatmaps screenshots were  
616 generated using Juicebox (<https://github.com/aidenlab/Juicebox/wiki/Download>). Aggregate  
617 heatmaps (Fig. 1E, S1C, S2A, S2C) were computed on a set of sub-matrices extracted from  
618 originals observed Hi-C matrices at 50kb resolution or 100kb resolution (Fig. S2C). Region of  
619 5Mb around DSBs (80 best) (Fig. 1E) or 5000 random regions (Fig. S1C, S2A, S2C) were  
620 extracted and then averaged. Log2 ratio was then computed using Hi-C counts (+DSB/-DSB)  
621 and plotted as heatmaps.

622 *Cis Contacts Quantification.* For *cis* contact quantification (Fig. 1B, 1G, S1B, S2B), interaction  
623 within γH2AX domains (-0.5/+0.5Mb around 80 best DSBs) were extracted from the observed  
624 Hi-C matrix at 100kb resolution, and log2 ratio was computed on damaged vs undamaged Hi-  
625 C counts (+DSB/-DSB). Adjacent windows (-1.5Mb-0.5Mb and +0.5Mb-1.5Mb around 80 best  
626 DSBs) were retrieved to quantify interactions in adjacent undamaged domains or between  
627 damaged domains and undamaged domains (for Fig.1B and Fig. S1B). Boxplots: Centre line,

628 median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers;  
629 points, outliers. Significance was calculated using non-parametric Wilcoxon test. For Fig. S1D,  
630 contact probability of Hi-C bins was plotted as a function of distance, using observed Hi-C  
631 matrix, for both conditions (+DSB and -DSB). For each bin on the chromosome 1, interactions  
632 with all the other bins was computed, averaged by distance, and mean interaction was plotted  
633 as a function of distance. Bins inside 1Mb domains around DSBs were removed from the  
634 computation, in order to study the DNA damage response on undamaged loci. For Fig 2B,  
635 interaction between genes inside  $\gamma$ H2AX domains ( $+/- 1\text{Mb}$  around DSB) and the closest DSB  
636 were computed using damaged (+DSB) Hi-C contact matrix at 5kb resolution. Genes were then  
637 sorted between down-regulated and unchanged using RNA-seq count data (from htseq-count).

638 *Insulation score and Loop strength.* Insulation score was computed using Hi-C matrices at 50  
639 kb resolution with matrix2insulation.pl (<https://github.com/dekkerlab/crane-nature-2015>). As  
640 parameters, we used is=800000 and ids=100000. For the differential loop strength (Fig. S1E)  
641 loops were called using Juicer Tools HiCCUPS and loop strength was computed as the P2M  
642 (Peak to Mean) the ratio of the central pixel to the mean of the remaining pixels using Juicer  
643 Tools APA (from enhancement.txt file) as described in<sup>12</sup>. Loops encompassed within -  
644 1Mb/+1Mb around DSBs were removed in order to focus the analysis on undamaged and  
645  $\gamma$ H2AX negative genomic loci (N=2733 loops). Differential loop-strength is the log2 ratio  
646 between loop-strength +DSB/-DSB. Loops were then sorted by number of DSBs inside their  
647 chromosome and plotted independently as boxplots. Boxplots: Centre line, median; box limits,  
648 first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers.

649 *CTCF binding site analysis.* For Fig. 2D, genes encompassed in  $\gamma$ H2AX domains (n=715) and  
650 carrying an annotated CTCF motif located in a 2kb window around their promoter, which  
651 displayed significant CTCF binding in DIVA cells (determined by peak calling using MACS2  
652 with default setting and using an input ChIP-seq) were selected (number of genes in  $\gamma$ H2AX

653 domains with CTCF motif and occupied by CTCF=198). CTCF sites orientation was next  
654 determined based on the relative position to the DSB and its strand. CTCF sites upstream DSB  
655 and with forward orientation (“+”) were classified as “toward”, with reverse orientation (“-”)  
656 classified as “away”. On the opposite, CTCF sites downstream a DSB were classified as “away”  
657 if in forward orientation and “toward” in reverse. RNA-seq average signal on exons was then  
658 extracted for genes overlapping with “toward” and “away” CTCF sites. Ambiguous genes (*i.e.*  
659 carrying CTCF motifs in both toward and away orientation) were removed from the analysis.  
660 Boxplots: Centre line, median; box limits, first and third quartiles; whiskers, maximum and  
661 minimum without outliers; points, outliers.

662 *Trans contact quantification.* To determine interaction changes in *trans* (inter-chromosomal)  
663 (Fig. S4C, S4D), we built the whole-genome Hi-C matrix for each experiment by merging  
664 together all chr-chr interaction matrices using Juicer and R. The result is a genome matrix with  
665 33kx33k bin interactions for 100kb resolution. Interactions between bins inside damaged TADs  
666 (240X240 for 80 DSBs) were extracted and counted for each condition, log2 ratio was  
667 calculated on normalized count (cpm), and plotted as boxplots. Boxplots: Centre line, median;  
668 box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points,  
669 outliers.

670 *TAD Cliques.* TAD Cliques were computed using the igraph R package on an undirected graph  
671 representing DSB clustering. This graph was computed on the differential Hi-C matrix (+DSB/-  
672 DSB) counts, at 500 kb resolution, considering a change of ~86% of interaction (0.9 in log2)  
673 as between two DSBs as a node on the graph. Averaged signal of ChIP-seq values  
674 (53BP1/γH2AX/H1/Ubiquitin FK2) were then computed for each categories of cliques using  
675 500kb windows around DSB.

676 *D compartment.* To identify the D compartment, we retrieved the first component (PC1) of a  
677 PCA made on the differential observed Hi-C matrix  $\log_2 \left( \frac{\text{damaged}}{\text{undamaged}} \right)$  at 100kb resolution.  
678 Each matrix was extracted from the .hic files using Juicer and the ratio was computed bin per  
679 bin. Pearson Correlation matrices were then computed for each chromosome, and PCA was  
680 applied on each matrix. The first component of each PCA was then extracted and correlated  
681 with the positions of DSB. A PC1 showing a positive correlation with DSB was then called D  
682 compartment, and PC1 showing negative correlation with DSBs were multiplied by -1. We  
683 were able to extract the D compartment on chromosomes 1,17 and X for +DSB/-DSB and  
684 chromosomes 1,2,6,9,13,17,18,20 and X for +DSB/-DSB in DNA-PKi condition. D  
685 compartment (first component of the PCA) was converted into a coverage file using rtracklayer  
686 R package. Using the same package, D compartment value was computed around DSBs and  
687 genes at 100kb resolution, and plotted as boxplot. Boxplots: Centre line, median; box limits,  
688 first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers.

689 *Transcription factor motif analysis.* TF-binding motifs were extracted on the promoter regions  
690 (-500bp/TSS) of genes with positive value of D compartment (2161) vs genes with negative  
691 value (2112) using motifmatchr and TFBSTools R packages on JASPAR2020 database. Motifs  
692 were sorted by significance using fisher exact test and adjusted with Benjamini-Hochberg  
693 procedure between motifs found on gene inside the D compartment versus genes outside D  
694 compartment. p53 direct target genes were retrieved from<sup>74</sup> (n=350). p53 target genes  
695 encompassed in -1/+1Mb windows around DSBs were excluded and only genes located on chr  
696 1,2,6,9,13,17,18,21, where D compartment was accurately identified in DNA-PKi conditions  
697 were kept for analysis. These genes were further categorized as upregulated or not regulated  
698 using our RNA-seq dataset.

699     *Translocation breakpoints.* For translocation breakpoints, data from<sup>44</sup> were retrieved, and only  
700     breakpoints for interchromosomal structural variant selected (N=28051). Genes reproducibly  
701     enriched in Compartment D in the three biological replicates, on chr1, 17a and X (N=604) as  
702     well as genes not enriched in Compartment D (N=1439) were retrieved. The significance of the  
703     overlap between genes and breakpoints was determined using the regionR package using  
704     resampling test with PermTest. Briefly, we selected 1000 times a control set of genes, with  
705     same size and on the same chromosome as our original gene set. We tested the overlap between  
706     each genes and breakpoints, to determine a distribution of the number of overlaps between  
707     control set and breakpoints. We further tested if the overlap between our gene set (D  
708     compartment or non D compartment) and breakpoints was significant, by counting the number  
709     of times we got more overlap in control than in our gene set.

710

## 711     **Acknowledgments**

712     We thank the genomics core facility of EMBL and of the I2BC (Centre de Recherche de Gif)  
713     for high-throughput sequencing. M.B. was supported by the CRUK Beatson Institute core grant  
714     A29252; A.B was supported by national productivity award from the MRC, MC\_ST\_U17040.  
715     Funding in GL laboratory was provided by grants from the European Research Council (ERC-  
716     2014-CoG 647344), Agence Nationale pour la Recherche (ANR-18-CE12-0015) and the Ligue  
717     Nationale contre le Cancer (LNCC). C.A was a recipient of a FRM fellowship (FRM  
718     FDT201904007941). T.C and N.P are INSERM researchers.

719

## 720     **Authors contributions**

721 C.A., E.L., T.C., and N.P. performed and analyzed experiments. V.R., and R.M, performed  
722 bioinformatic analyses of all high-throughput sequencing datasets. A.B performed the  
723 Amplicon-seq experiment under the supervision of M.B. D.N. helped to realize and analyze  
724 4C-seq experiments. G.L. and T.C. wrote the manuscript. All authors commented and edited  
725 the manuscript.

726

727 **Competing Interest**

728 The authors declare no competing interest

729

730 **Data Availability**

731 All high-throughput sequencing data (Hi-C, 4C-seq, Amplicon-seq and RNA-seq) have been  
732 deposited to Array Express (<https://www.ebi.ac.uk/arrayexpress/>) under accession number E-  
733 MTAB-XXXX. (Details for reviewer access available in the reporting summary)

734

735 **Code availability**

736

737 Source codes are available from <https://github.com/LegubeDNAREPAIR/>

738

739

740

741 **Figures Legends**

742

743 **Figure 1: Cohesin and ATM-dependent TAD reinforcement in response to DSBs.**

744 (A) Hi-C contact matrix of the log2 (+DSB/-DSB) in DIvA cells. A region of the chromosome  
745 1 is shown at three different resolutions: 250 kb (left panel), 100 kb (middle panel) and 25 kb  
746 (right panel). The  $\gamma$ H2AX ChIP-seq signal following DSB induction is shown on the top panel  
747 and indicates the DSBs position. The red square highlights a damaged TAD, within which *cis*  
748 interactions are enhanced, while the blue square highlights decreased interaction between the  
749 damaged TAD and its adjacent TAD. One representative experiment is shown.

750 (B) Boxplot showing the differential Hi-C read counts (as (log2 +DSB/-DSB)) within  $\gamma$ H2AX  
751 domains containing the 80 best induced DSBs (red) or between these 80 damaged domains and  
752 their adjacent chromatin domains (blue). P-values, non-parametric wilcoxon test tested against  
753  $\mu=0$ .

754 (C) Hi-C contact matrix of log2 (+DSB/-DSB) on a region located on chromosome 17 at 50 kb  
755 resolution. The contacts engaged by the DSB itself are indicated with a black arrow.  $\gamma$ H2AX  
756 ChIP-seq track (+DSB) is shown on the top panel. One representative experiment is shown.

757 (D) Hi-C contact matrix of the log2(+DSB/-DSB) without inhibitor (top panel), with DNA-PK  
758 inhibitor (middle panel) or with ATM inhibitor (bottom panel). A damaged region of the  
759 chromosome 1 is shown at a 25 kb resolution. Grey track represents the insulation score pre-  
760 existing to DSB induction (from Hi-C –DSB)

761 (E) Averaged Hi-C contact matrix of the log2 (+DSB/-DSB) in untreated cells (left panel), upon  
762 DNA-PK inhibition (middle panel) or upon ATM inhibition (right panel), centered on the 80  
763 best-induced DSBs (50 kb resolution on a 5 Mb window).

764 (F) Hi-C contact matrix of the log<sub>2</sub>(+DSB/-DSB) on a region located on chromosome 1 at a 50  
765 kb resolution in DIVA cells transfected with a control siRNA or a siRNA directed against SCC1.

766 (G) Boxplot showing the differential Hi-C read counts (log<sub>2</sub> +DSB/-DSB) within the 80 best  
767 γH2AX domains in Control or SCC1-depleted conditions. P-values, non-parametric wilcoxon  
768 test tested against μ=0. siSCC1 vs siCTRL, P=paired wilcoxon test.

769

770 **Figure 2. Cohesin-mediated loop extrusion contributes to local DSB-induced**  
771 **transcriptional regulation**

772 (A) Top panel: Genomic tracks of γH2AX ChIP-seq after DSB induction in a region of  
773 chromosome 20, DSB position is indicated by a black arrow. Middle panel: Hi-C contact matrix  
774 showing the signal obtained after DSB induction (log<sub>2</sub> observed/expected) and genomic tracks  
775 of the RNA-seq signal before (-DSB, blue) and after DSB induction (+DSB, red). The DSB  
776 position is indicated with a black arrow and a dashed black line. Bottom panel: Close-up  
777 showing differential RNA-seq (log<sub>2</sub> ratio +DSB/-DSB, green) track for a gene localized at the  
778 DSB site (zoom panel, left), at ~200 kb from the DSB site (zoom panel, middle) and at ~800  
779 kb from the DSB site (zoom panel, right). One representative experiment is shown.

780 (B) Boxplot showing the quantification of the Hi-C contacts between the genes located in  
781 γH2AX domains and the closest DSB. Genes that did not display transcriptional changes post-  
782 DSB induction are shown in grey, genes that showed down-regulation following DSB induction  
783 in blue (identified by RNA-seq). *P*, non parametric wilcoxon test.

784 (C) Top panel: Genomic track of the 4C-seq signal after DSB induction using a viewpoint  
785 localized at a DSB on chromosome 20 (red arrow). One representative experiment is shown.  
786 Bottom panel: Close up of two genomic loci located at ~400 kb from the DSB site (*DNMT3B*

787 and *MAPRE1*), showing RNA-seq read count before (-DSB, blue) and after DSB induction  
788 (+DSB, red), differential RNA-seq (log<sub>2</sub> ratio +DSB/-DSB, green), and CTCF ChIP-seq (-  
789 DSB).

790 (D) Quantification of the RNA-seq read counts before (-DSB, blue) and after DSB induction  
791 (+DSB, red) on genes embedded in γH2AX domains and displaying a CTCF-bound element  
792 facing toward or away from the closest DSB site, as indicated. *P*, non-parametric paired  
793 wilcoxon test.

794

795 **Figure 3: Cell cycle regulated, ATM-dependent but cohesin- and DNA-PK-independent**  
796 **clustering of damaged-TADs.**

797 (A) Hi-C contact matrix of the log<sub>2</sub> (+DSB/-DSB) on a region of the chromosome 1 at two  
798 different resolutions: 250 kb (left panel) and 100 kb (right panel). γH2AX ChIP-seq track  
799 following DSB induction is shown on the top panel and on the right. One representative  
800 experiment is shown.

801 (B) Hi-C contact matrix of the log<sub>2</sub> (+DSB/-DSB) on a region of the chromosome 17 at 250 kb  
802 resolution. γH2AX and 53BP1 ChIP-seq tracks following DSB induction are shown on the top  
803 panel and on the left. The black arrows indicate clustering of one DSB on the chromosome 17,  
804 with several other DSBs on the same chromosome. One representative experiment is shown.

805 (C) γH2AX domains were categorized based on their propensity to not interact with any other  
806 γH2AX domain (single), with one other γH2AX domain (TAD-TAD) or with multiple other  
807 γH2AX domains (TAD cliques containing 3 to 6 DSBs). ChIP-seq levels of γH2AX (+DSB),  
808 53BP1 (+DSB), H1 (log<sub>2</sub> +DSB/-DSB) or Ubiquitin chains detected with the FK2 antibody  
809 (log<sub>2</sub> +DSB/-DSB) within the corresponding domains were computed across each category.

810 (D) Left panel: Hi-C contact matrix of the log<sub>2</sub>(+DSB/-DSB) upon Ctrl (upper right) or SCC1  
811 depletion (lower left). A region of the chromosome 1 is shown at 250 kb resolution. The  $\gamma$ H2AX  
812 ChIP-seq track following DSB induction is shown on the top and on the right. Right panel:  
813 magnification of the black square, showing Hi-C contacts between the two  $\gamma$ H2AX domains.

814 (E) Hi-C contact matrix of the log<sub>2</sub> (+DSB/-DSB) without inhibitor, with a DNA-PK inhibitor  
815 or with an ATM inhibitor as indicated. A region of the chromosome 1 is shown with a 250 kb  
816 resolution.  $\gamma$ H2AX ChIP-seq track following DSB induction is shown on the top. Bottom panel:  
817 magnification, showing Hi-C contacts between the two  $\gamma$ H2AX domains.

818 (F) *Trans* interactions (log<sub>2</sub> ratio +DSB/-DSB) between the view point and the other DSBs  
819 (n=79) were computed from 4C-seq experiments in synchronized cells (G1, S and G2 as  
820 indicated). Three cluster-prone DSBs, one not cluster-prone and one control undamaged locus  
821 were used as viewpoints. *P*, non-parametric paired wilcoxon test.

822

823 **Figure 4. Formation of a DSB-specific «D compartment».**

824 (A) Genomic tracks of  $\gamma$ H2AX ChIP-seq and first Chromosomal eigenvector (CEV) computed  
825 on differential (+DSB/-DSB) Hi-C matrix of chromosome 17 using a 100kb resolution (blue).  
826 Genomic regions displaying a positive CEV signal belong to the DSB-induced «D  
827 compartment» (black arrows).

828 (B) Boxplot representing the quantification of the «D compartment» signal on a 1Mb window  
829 around the 80 best induced DSBs (red) and around 80 control undamaged regions (random,  
830 purple). *P*, non parametric Wilcoxon test

831 (C) Pearson correlation between  $\gamma$ H2AX ChIP-seq and D compartment (positive CEV PC1 on  
832 differential Hi-C matrices) for three biological Hi-C replicates (Rep#1, Rep#2 and Rep#3) and  
833 the Hi-C performed in presence of DNAPK inhibitor as indicated.

834 (D) Genomic tracks of  $\gamma$ H2AX (red), H3K79me2 (black) and H3K4me3 (yellow) ChIP-seq,  
835 and the first Chromosomal Eigenvector computed on the differential Hi-C (CEV, blue). The  
836 brown squares highlight genomic regions present in D compartment that carry a DSB and are  
837 enriched in  $\gamma$ H2AX. In contrast the blue square shows a genomic region that is devoid in  $\gamma$ H2AX  
838 and DSB, but is nevertheless found in the D compartment.

839 (E) Boxplot showing the quantification of the «D compartment» signal computed from Hi-C  
840 data (+DSB+DNA-PKi/-DSB) on genes that are not regulated following DSB induction (Not-  
841 regulated genes, grey), genes that are upregulated following DSB induction (Upregulated  
842 genes, red) or genes that are downregulated following DSB induction (Downregulated genes,  
843 blue), identified by RNA-seq.

844 (F) RT-qPCR quantification of the expression level of four genes (*RNF19B*, *FAS*, *PLK3* and  
845 *GADD45A*) before and after DSB induction in cells transfected with control or SUN2 siRNA.  
846 n=4 independent experiments.

847

848 **Figure 5. DSB-induced loop extrusion and D-compartment formation drive  
849 translocations.**

850 (A) qPCR quantification of translocations frequency for two independent translocations  
851 following DSB induction in cells synchronized in the G1, S or G2 phase (n=4 independent  
852 replicates). P= paired t-test, \* P<0.05, \*\* P<0.001, \*\*\*P<0.0005

853 (B) qPCR quantification of translocations frequency for two independent translocations  
854 following DSB induction with or without DNA-PK inhibitor (n=4 independent replicates).

855 (C) qPCR quantification of translocations frequency for two independent translocations  
856 following DSB induction in Control, 53BP1, SUN2 or ARP2 depleted cells or upon 1,6-  
857 Hexanediol treatment (n≥3 independent replicates).

858 (D) As in (C) but upon Control, SMC1 or SCC1 depletion (n=4 independent replicates).

859 (E) Intra-chromosomal (blue) or inter-chromosomal translocations (yellow) were quantified  
860 using multiplexed amplification followed by high throughput sequencing (amplicon-seq)  
861 between 20 different DSBs induced in DIvA cell line, upon Ctrl or SCC1 depletion (log2  
862 siSCC1/siCTRL) (n=4 independent replicates). P-values, non-parametric wilcoxon test tested  
863 against  $\mu=0$ . intra vs inter-chromosomal,  $P$ =paired wilcoxon test.

864 (F) As in (E) but the quantification was performed in SUN2 depleted cells (n=4 independent  
865 replicates).

866 (G) As in (E) but the quantification was performed in ARP2 depleted cells (n=4 independent  
867 replicates).

868 (H) Observed (green) and expected (obtained through 1000 permutations) overlap between  
869 breakpoint positions of inter-chromosomal translocations identified on cancer genomes (from  
870 (Zhang et al., 2018)) and genes targeted to the D compartment, either upregulated,  
871 downregulated or not regulated following DSB induction (identified by RNA-seq) as indicated,  
872 compared to their counterparts not targeted to the D compartment.

873

874 **Figure 6. Model.**

875 (A) INTRA-TAD changes: Cohesin-dependent loop extrusion arising at DSB ensures ATM-  
876 dependent phosphorylation of H2AX on an entire TAD, allowing fast establishment of a TAD-  
877 scale DDR focus (Arnould et al., 2021). Genes embedded in the DDR focus that pass through  
878 the loop during the ongoing DSB-anchored loop extrusion process are not repressed. However,  
879 genes that carry a CTCF-bound site in the proper orientation, acting as a roadblock which  
880 stops/pauses loop extrusion, are brought to 3D proximity of the DSB and hence subjected to  
881 ATM-dependent down-regulation. On the other hand, DSB-anchored loop extrusion also  
882 displays the potential to bring in close proximity two DSBs located on the same chromosome,  
883 which can favor the occurrence of intra-chromosomal translocations.

884 (B) INTER-TAD changes: Once assembled, the decorated  $\gamma$ H2AX/53BP1 damaged TADs can  
885 further fuse together presumably by phase separation given previously demonstrated LLPS  
886 properties of 53BP1. This creates a new nuclear, DSB-induced compartment, in which the DNA  
887 damage responsive genes, such as p53 targeted genes, physically relocate. The ARP2 Actin  
888 branching factor, and the SUN2 LINC member also contribute to the formation of this  
889 compartment. D-compartment mainly form when DSBs persist, for example in G1, when  
890 Homologous Recombination is not available to repair DSB induced in active loci<sup>17</sup> or upon  
891 DNA-PK inhibition<sup>25</sup>. On the other hand, induced spatial proximity in D compartment also  
892 increases the frequency of translocations between DSBs or between the DDR responsive genes.

893

894 **Bibliography**

- 895 1. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA double-  
896 stranded breaks induce histone H2AX phosphorylation on serine 139. *J. Biol. Chem.* **273**,  
897 5858–5868 (1998).
- 898 2. Clouaire, T. *et al.* Comprehensive Mapping of Histone Modifications at DNA Double-  
899 Strand Breaks Deciphers Repair Pathway Chromatin Signatures. *Mol. Cell* **72**, 250–  
900 262.e6 (2018).
- 901 3. Hartlerode, A. J. & Scully, R. Mechanisms of double-strand break repair in somatic  
902 mammalian cells. *Biochem. J.* **423**, 157–168 (2009).
- 903 4. Clouaire, T. & Legube, G. A Snapshot on the Cis Chromatin Response to DNA Double-  
904 Strand Breaks. *Trends Genet.* **35**, 330–345 (2019).
- 905 5. Clouaire, T. & Legube, G. DNA double strand break repair pathway choice: a chromatin  
906 based decision? *Nucleus* **6**, 107–113 (2015).
- 907 6. Panier, S. & Boulton, S. J. Double-strand break repair: 53BP1 comes into focus. *Nat.*  
908 *Rev. Mol. Cell Biol.* **15**, 7–18 (2014).
- 909 7. Wilson, M. D. & Durocher, D. Reading chromatin signatures after DNA double-strand  
910 breaks. *Philos. Trans. R. Soc. Lond. B, Biol. Sci.* **372**, (2017).
- 911 8. Caron, P., van der Linden, J. & van Attikum, H. Bon voyage: A transcriptional journey  
912 around DNA breaks. *DNA Repair (Amst.)* **82**, 102686 (2019).
- 913 9. Lesage, E., Clouaire, T. & Legube, G. Repair of DNA double-strand breaks in RNAPI-  
914 and RNAPII-transcribed loci. *DNA Repair (Amst.)* **104**, 103139 (2021).
- 915 10. Collins, P. L. *et al.* DNA double-strand breaks induce H2Ax phosphorylation domains in  
916 a contact-dependent manner. *Nat. Commun.* **11**, 3158 (2020).
- 917 11. Caron, P. *et al.* Cohesin protects genes against γH2AX Induced by DNA double-strand  
918 breaks. *PLoS Genet.* **8**, e1002460 (2012).
- 919 12. Arnould, C. *et al.* Loop extrusion as a mechanism for formation of DNA damage repair  
920 foci. *Nature* **590**, 660–665 (2021).
- 921 13. Fudenberg, G. *et al.* Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**,  
922 2038–2049 (2016).
- 923 14. Meisenberg, C. *et al.* Repression of Transcription at DNA Breaks Requires Cohesin  
924 throughout Interphase and Prevents Genome Instability. *Mol. Cell* **73**, 212–223.e7  
925 (2019).
- 926 15. Aten, J. A. *et al.* Dynamics of DNA double-strand breaks revealed by clustering of  
927 damaged chromosome domains. *Science* **303**, 92–95 (2004).
- 928 16. Roukos, V. *et al.* Spatial dynamics of chromosome translocations in living cells. *Science*  
929 **341**, 660–664 (2013).

- 930 17. Aymard, F. *et al.* Genome-wide mapping of long-range contacts unveils clustering of  
931 DNA double-strand breaks at damaged active genes. *Nat. Struct. Mol. Biol.* **24**, 353–361  
932 (2017).
- 933 18. Schrank, B. R. *et al.* Nuclear ARP2/3 drives DNA break clustering for homology-  
934 directed repair. *Nature* **559**, 61–66 (2018).
- 935 19. Lottersberger, F., Karssemeijer, R. A., Dimitrova, N. & de Lange, T. 53BP1 and the  
936 LINC Complex Promote Microtubule-Dependent DSB Mobility and DNA Repair. *Cell*  
937 **163**, 880–893 (2015).
- 938 20. Kilic, S. *et al.* Phase separation of 53BP1 determines liquid-like behavior of DNA repair  
939 compartments. *EMBO J.* **38**, e101379 (2019).
- 940 21. Pessina, F. *et al.* Functional transcription promoters at DNA double-strand breaks  
941 mediate RNA-driven phase separation of damage-response factors. *Nat. Cell Biol.* **21**,  
942 1286–1299 (2019).
- 943 22. Guénolé, A. & Legube, G. A meeting at risk: Unrepaired DSBs go for broke. *Nucleus* **8**,  
944 589–599 (2017).
- 945 23. Sanders, J. T. *et al.* Radiation-induced DNA damage and repair effects on 3D genome  
946 organization. *Nat. Commun.* **11**, 6178 (2020).
- 947 24. Piazza, A. *et al.* Cohesin regulates homology search during recombinational DNA repair.  
948 *BioRxiv* (2020). doi:10.1101/2020.12.17.423195
- 949 25. Caron, P. *et al.* Non-redundant Functions of ATM and DNA-PKcs in Response to DNA  
950 Double-Strand Breaks. *Cell Rep.* **13**, 1598–1609 (2015).
- 951 26. Shanbhag, N. M., Rafalska-Metcalf, I. U., Balane-Bolivar, C., Janicki, S. M. &  
952 Greenberg, R. A. ATM-dependent chromatin changes silence transcription in cis to DNA  
953 double-strand breaks. *Cell* **141**, 970–981 (2010).
- 954 27. Pankotai, T., Bonhomme, C., Chen, D. & Soutoglou, E. DNAPKcs-dependent arrest of  
955 RNA polymerase II transcription in the presence of DNA breaks. *Nat. Struct. Mol. Biol.*  
956 **19**, 276–282 (2012).
- 957 28. Iacovoni, J. S. *et al.* High-resolution profiling of gammaH2AX around DNA double  
958 strand breaks in the mammalian genome. *EMBO J.* **29**, 1446–1457 (2010).
- 959 29. Ui, A., Nagaura, Y. & Yasui, A. Transcriptional elongation factor ENL phosphorylated  
960 by ATM recruits polycomb and switches off transcription for DSB repair. *Mol. Cell* **58**,  
961 468–482 (2015).
- 962 30. Kakarougkas, A. *et al.* Requirement for PBAF in transcriptional repression and repair at  
963 DNA breaks in actively transcribed regions of chromatin. *Mol. Cell* **55**, 723–732 (2014).
- 964 31. Puget, N., Miller, K. M. & Legube, G. Non-canonical DNA/RNA structures during  
965 Transcription-Coupled Double-Strand Break Repair: Roadblocks or Bona fide repair  
966 intermediates? *DNA Repair (Amst.)* **81**, 102661 (2019).

- 967 32. Iannelli, F. *et al.* A damaged genome's transcriptional landscape through multilayered  
968 expression profiling around in situ-mapped DNA double-strand breaks. *Nat. Commun.* **8**,  
969 15656 (2017).
- 970 33. Chang, L.-H., Ghosh, S. & Noordermeer, D. Tads and their borders: free movement or  
971 building a wall? *J. Mol. Biol.* **432**, 643–652 (2020).
- 972 34. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals  
973 principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
- 974 35. Guo, Y. *et al.* CRISPR inversion of CTCF sites alters genome topology and  
975 enhancer/promoter function. *Cell* **162**, 900–910 (2015).
- 976 36. de Wit, E. *et al.* CTCF binding polarity determines chromatin looping. *Mol. Cell* **60**,  
977 676–684 (2015).
- 978 37. Paulsen, J. *et al.* Long-range interactions between topologically associating domains  
979 shape the four-dimensional genome during differentiation. *Nat. Genet.* **51**, 835–843  
980 (2019).
- 981 38. Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals  
982 folding principles of the human genome. *Science* **326**, 289–293 (2009).
- 983 39. Huang, Q. *et al.* Identification of p53 regulators by genome-wide functional analysis.  
984 *Proc. Natl. Acad. Sci. USA* **101**, 3456–3461 (2004).
- 985 40. Allocati, N., Di Ilio, C. & De Laurenzi, V. p63/p73 in the control of cell cycle and cell  
986 death. *Exp. Cell Res.* **318**, 1285–1290 (2012).
- 987 41. Fouad, S., Hauton, D. & D'Angiolella, V. E2F1: cause and consequence of DNA  
988 replication stress. *Front. Mol. Biosci.* **7**, 599332 (2020).
- 989 42. Bowen, C. & Gelmann, E. P. NKX3.1 activates cellular response to DNA damage.  
990 *Cancer Res.* **70**, 3089–3097 (2010).
- 991 43. Roos, W. P. & Kaina, B. DNA damage-induced cell death: from specific DNA lesions to  
992 the DNA damage response and apoptosis. *Cancer Lett.* **332**, 237–248 (2013).
- 993 44. Zhang, Y. *et al.* A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic  
994 Rearrangement across More Than 1,400 Cases. *Cell Rep.* **24**, 515–527 (2018).
- 995 45. Spegg, V. & Altmeyer, M. Biomolecular condensates at sites of DNA damage: More  
996 than just a phase. *DNA Repair (Amst.)* 103179 (2021). doi:10.1016/j.dnarep.2021.103179
- 997 46. Ochs, F. *et al.* Stabilization of chromatin topology safeguards genome integrity. *Nature*  
998 **574**, 571–574 (2019).
- 999 47. Kim, S.-T., Xu, B. & Kastan, M. B. Involvement of the cohesin protein, Smc1, in Atm-  
1000 dependent and independent responses to DNA damage. *Genes Dev.* **16**, 560–570 (2002).
- 1001 48. Kim, B.-J. *et al.* Genome-wide reinforcement of cohesin binding at pre-existing cohesin  
1002 sites in response to ionizing radiation in human cells. *J. Biol. Chem.* **285**, 22784–22792  
1003 (2010).

- 1004 49. Bauerschmidt, C. *et al.* Cohesin phosphorylation and mobility of SMC1 at ionizing  
1005 radiation-induced DNA double-strand breaks in human cells. *Exp. Cell Res.* **317**, 330–  
1006 337 (2011).
- 1007 50. Caron, P. *et al.* WWP2 ubiquitylates RNA polymerase II for DNA-PK-dependent  
1008 transcription arrest and repair at DNA breaks. *Genes Dev.* **33**, 684–704 (2019).
- 1009 51. Gong, F. *et al.* Screen identifies bromodomain protein ZMYND8 in chromatin  
1010 recognition of transcription-associated DNA damage that promotes homologous  
1011 recombination. *Genes Dev.* **29**, 197–211 (2015).
- 1012 52. Kim, J., Sturgill, D., Tran, A. D., Sinclair, D. A. & Oberdoerffer, P. Controlled DNA  
1013 double-strand break induction in mice reveals post-damage transcriptome stability.  
1014 *Nucleic Acids Res.* **44**, e64 (2016).
- 1015 53. Gong, F., Clouaire, T., Aguirrebengoa, M., Legube, G. & Miller, K. M. Histone  
1016 demethylase KDM5A regulates the ZMYND8-NuRD chromatin remodeler to promote  
1017 DNA repair. *J. Cell Biol.* **216**, 1959–1974 (2017).
- 1018 54. Chang, L.-H. & Noordermeer, D. Of Dots and Stripes: The Morse Code of Micro-C  
1019 Reveals the Ultrastructure of Transcriptional and Architectural Mammalian 3D Genome  
1020 Organization. *Mol. Cell* **78**, 376–378 (2020).
- 1021 55. Larson, A. G. *et al.* Liquid droplet formation by HP1 $\alpha$  suggests a role for phase  
1022 separation in heterochromatin. *Nature* **547**, 236–240 (2017).
- 1023 56. Strom, A. R. *et al.* Phase separation drives heterochromatin domain formation. *Nature*  
1024 **547**, 241–245 (2017).
- 1025 57. Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. & Mirny, L. A. Chromatin  
1026 organization by an interplay of loop extrusion and compartmental segregation. *Proc.  
1027 Natl. Acad. Sci. USA* **115**, E6697–E6706 (2018).
- 1028 58. Falk, M. *et al.* Heterochromatin drives compartmentalization of inverted and  
1029 conventional nuclei. *Nature* **570**, 395–399 (2019).
- 1030 59. Cuella-Martin, R. *et al.* 53BP1 Integrates DNA Repair and p53-Dependent Cell Fate  
1031 Decisions via Distinct Mechanisms. *Mol. Cell* **64**, 51–64 (2016).
- 1032 60. Ghodke, I. *et al.* AHNAK controls 53BP1-mediated p53 response by restraining 53BP1  
1033 oligomerization and phase separation. *Mol. Cell* **81**, 2596–2610.e7 (2021).
- 1034 61. Drané, P. *et al.* TIRR regulates 53BP1 by masking its histone methyl-lysine binding  
1035 function. *Nature* **543**, 211–216 (2017).
- 1036 62. Parnandi, N. *et al.* TIRR inhibits the 53BP1-p53 complex to alter cell-fate programs.  
1037 *Mol. Cell* **81**, 2583–2595.e6 (2021).
- 1038 63. Izhar, L. *et al.* A Systematic Analysis of Factors Localized to Damaged Chromatin  
1039 Reveals PARP-Dependent Recruitment of Transcription Factors. *Cell Rep.* **11**, 1486–  
1040 1500 (2015).
- 1041 64. Harrigan, J. A. *et al.* Replication stress induces 53BP1-containing OPT domains in G1  
1042 cells. *J. Cell Biol.* **193**, 97–108 (2011).

- 1043 65. Pombo, A. *et al.* Regional and temporal specialization in the nucleus: a transcriptionally-  
1044 active nuclear domain rich in PTF, Oct1 and PIKA antigens associates with specific  
1045 chromosomes early in the cell cycle. *EMBO J.* **17**, 1768–1778 (1998).
- 1046 66. Gothe, H. J. *et al.* Spatial chromosome folding and active transcription drive DNA  
1047 fragility and formation of oncogenic MLL translocations. *Mol. Cell* **75**, 267–283.e12  
1048 (2019).
- 1049 67. Canela, A. *et al.* Topoisomerase II-Induced Chromosome Breakage and Translocation Is  
1050 Determined by Chromosome Architecture and Transcriptional Activity. *Mol. Cell* **75**,  
1051 252–266.e8 (2019).
- 1052 68. Canela, A. *et al.* Genome organization drives chromosome fragility. *Cell* **170**, 507–  
1053 521.e18 (2017).
- 1054 69. Zhang, X. *et al.* Fundamental roles of chromatin loop extrusion in antibody class  
1055 switching. *Nature* **575**, 385–389 (2019).
- 1056 70. Zhang, Y. *et al.* The fundamental role of chromatin loop extrusion in physiological V(D)J  
1057 recombination. *Nature* **573**, 600–604 (2019).
- 1058 71. Dai, H.-Q. *et al.* Loop extrusion mediates physiological Ig locus contraction for RAG  
1059 scanning. *Nature* **590**, 338–343 (2021).
- 1060 72. Aymard, F. *et al.* Transcriptionally active chromatin recruits homologous recombination  
1061 at DNA double-strand breaks. *Nat. Struct. Mol. Biol.* **21**, 366–374 (2014).
- 1062 73. Cohen, S. *et al.* Senataxin resolves RNA:DNA hybrids forming at DNA double-strand  
1063 breaks to prevent translocations. *Nat. Commun.* **9**, 533 (2018).
- 1064 74. Fischer, M. Census and evaluation of p53 target genes. *Oncogene* **36**, 3943–3956 (2017).
- 1065

## **Supplemental file**

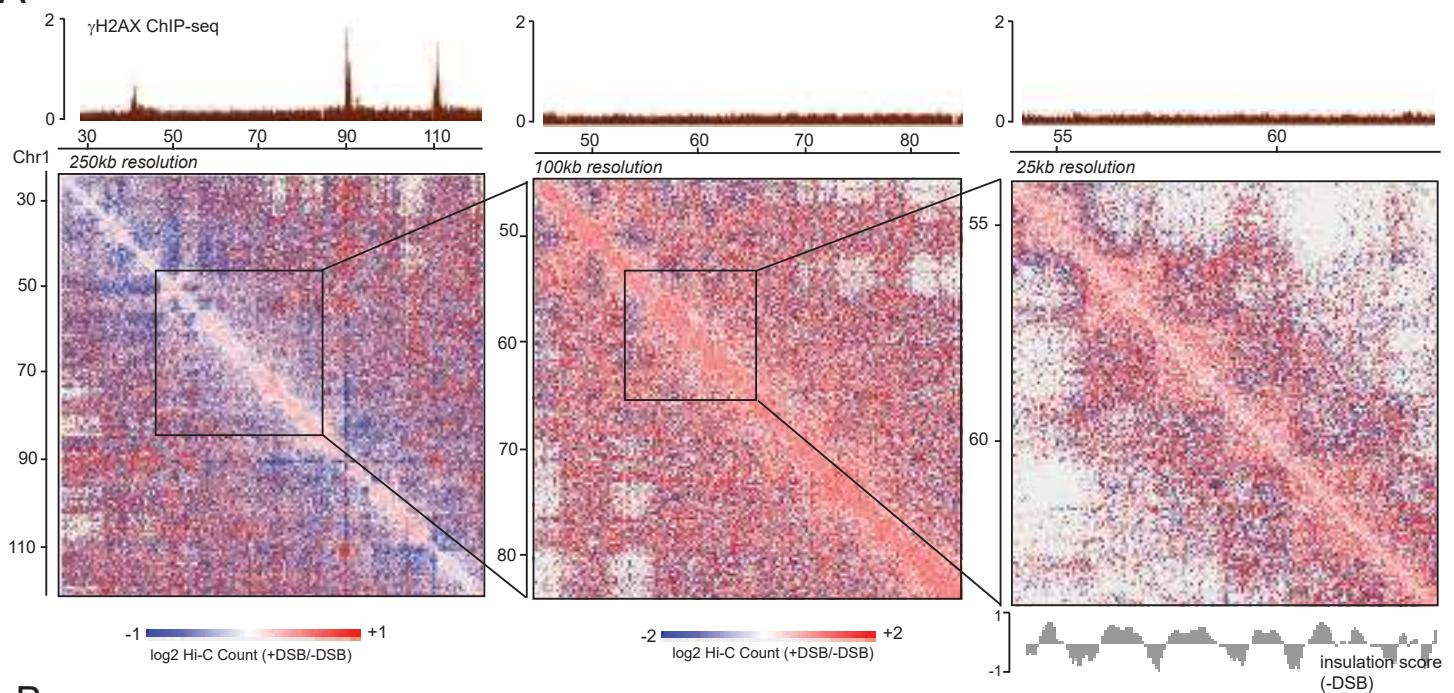
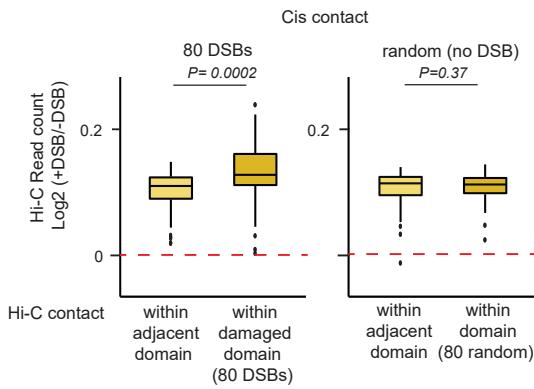
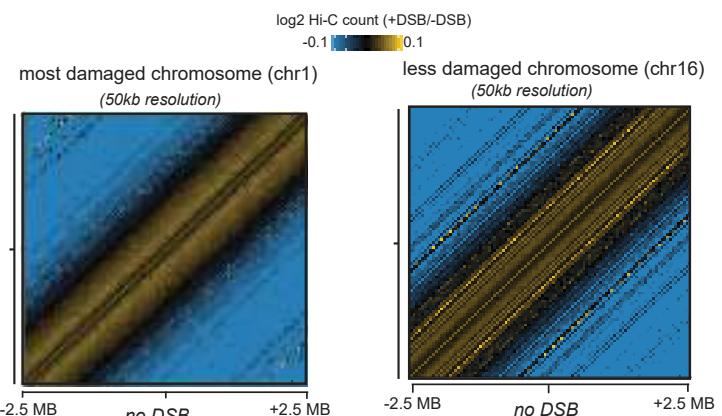
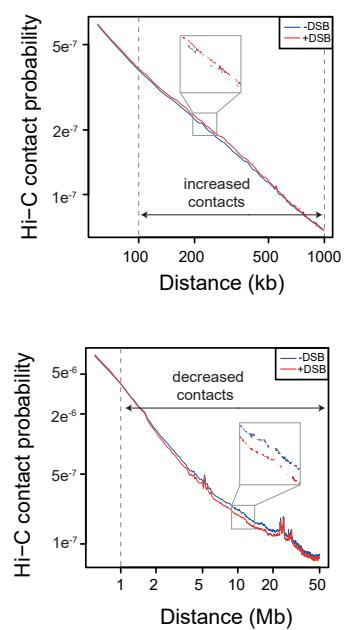
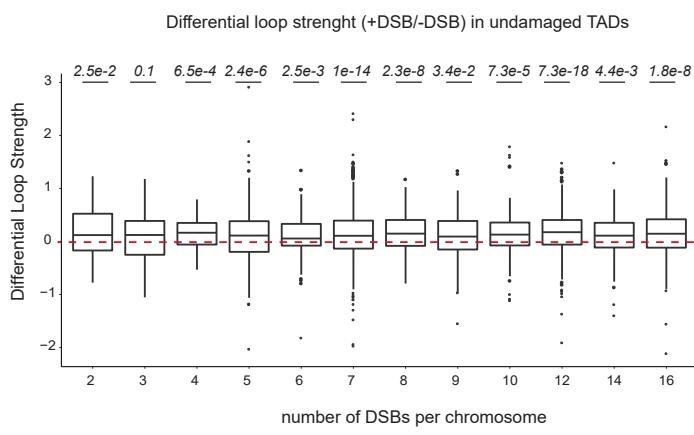
### **ATM-dependent formation of a novel chromatin compartment regulates the Response to DNA Double Strand Breaks and the biogenesis of translocations**

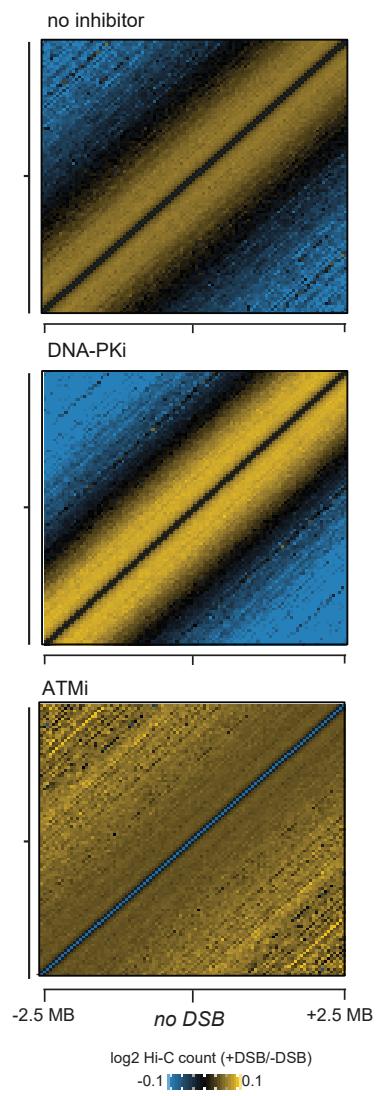
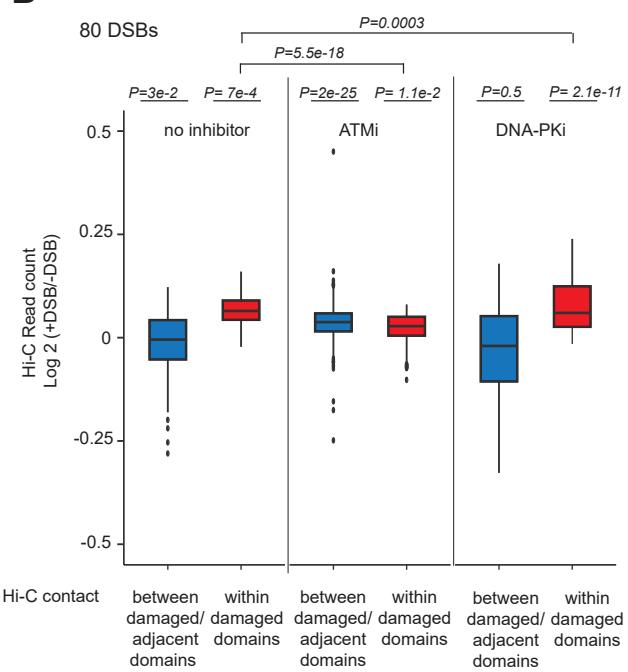
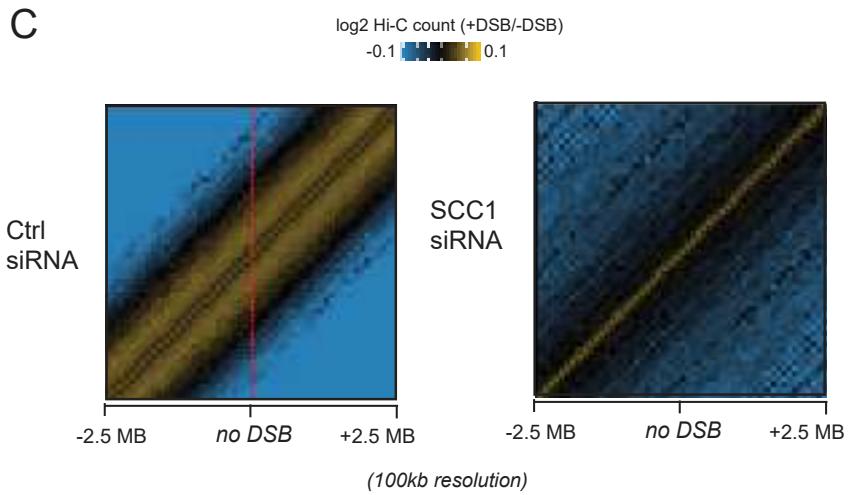
Coline Arnould<sup>1#</sup>, Vincent Rocher<sup>1#</sup>, Aldo Bader<sup>2</sup>, Emma Lesage<sup>1</sup>, Nadine Puget<sup>1</sup>, Thomas Clouaire<sup>1</sup>, Raphael Mourad<sup>1</sup>, Daan Noordemeer<sup>3</sup>, Martin Bushell<sup>2</sup> and Gaëlle Legube<sup>1\*</sup>

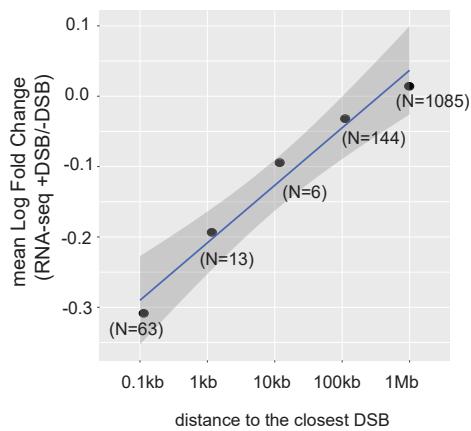
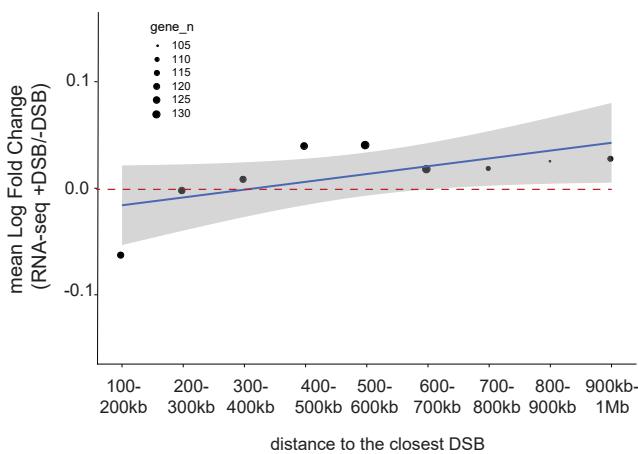
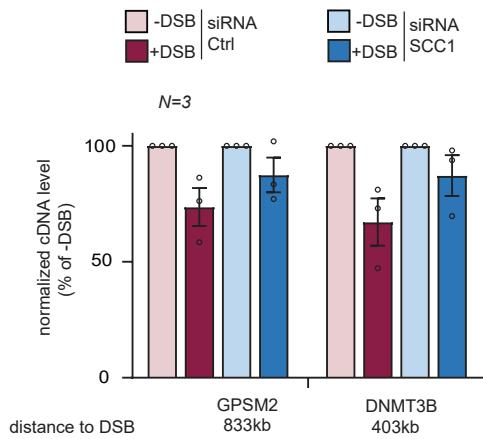
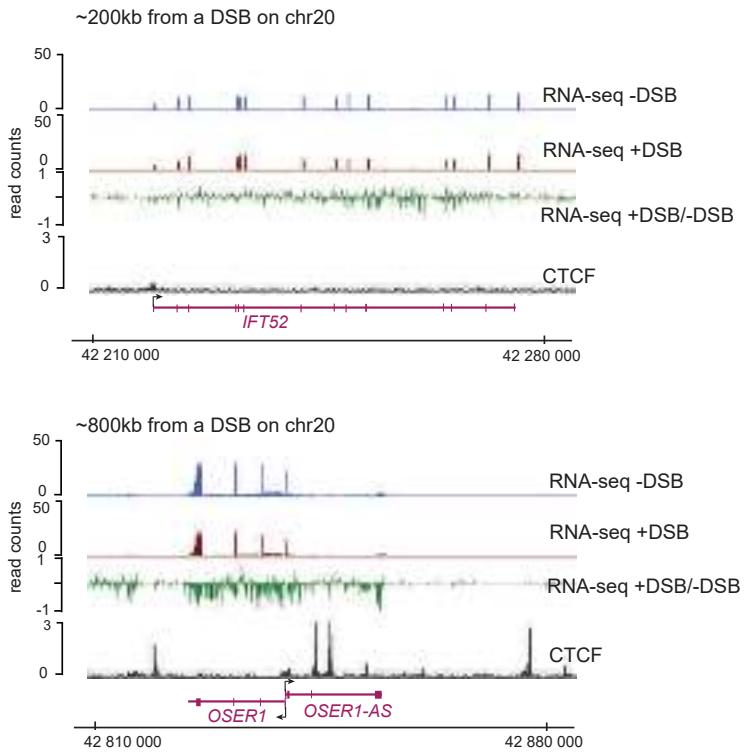
Figure S1-S6

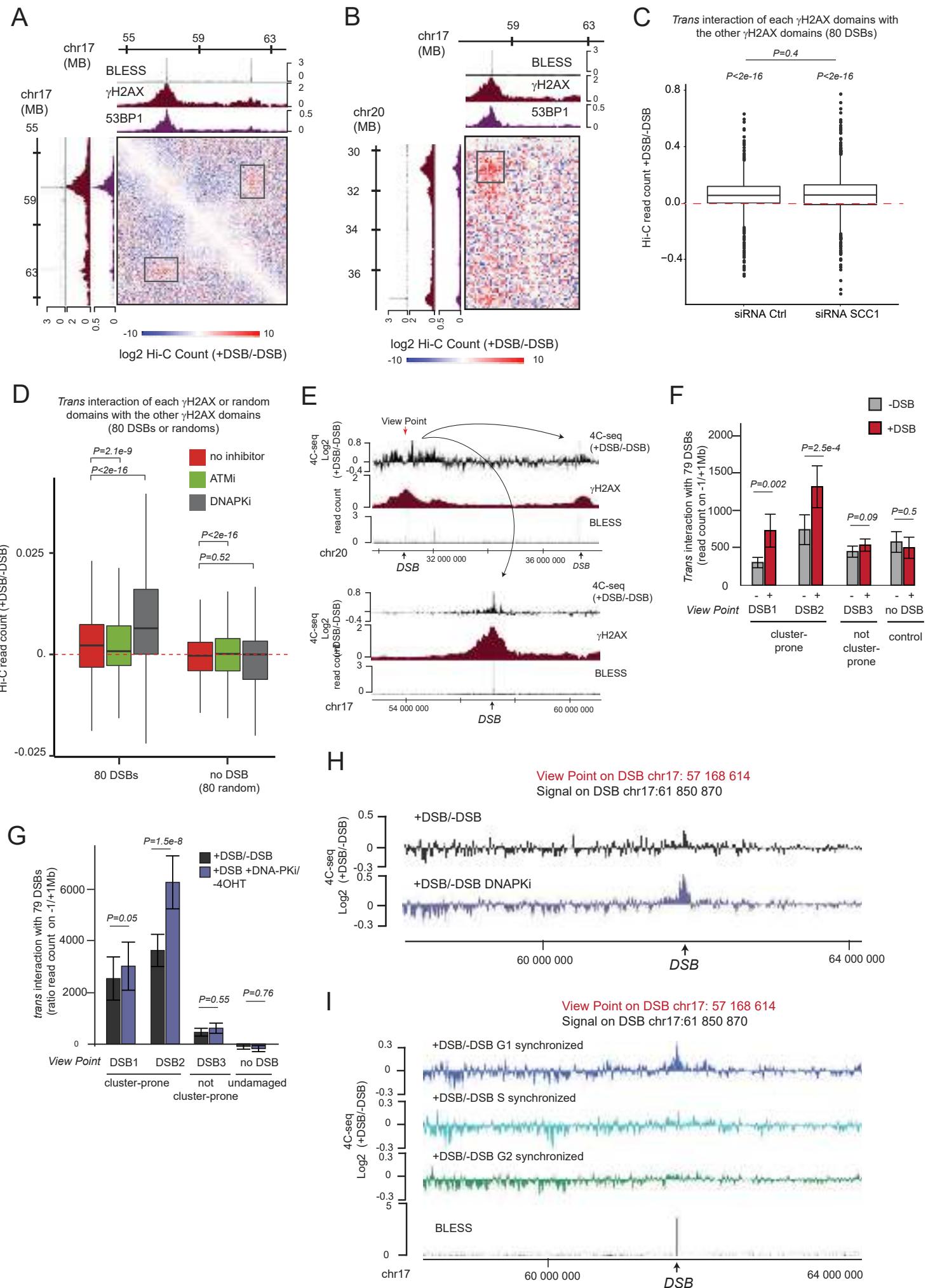
Supplemental Figures legends

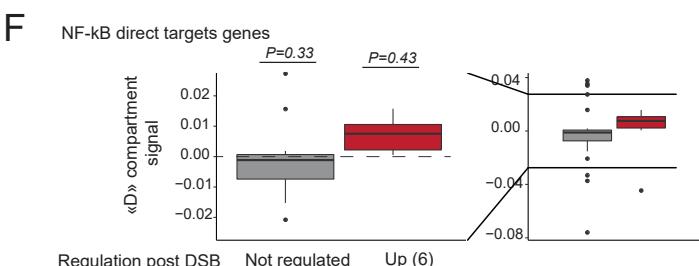
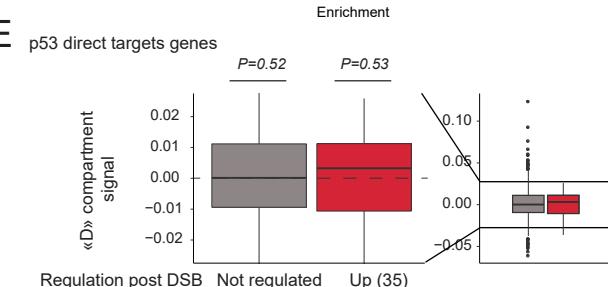
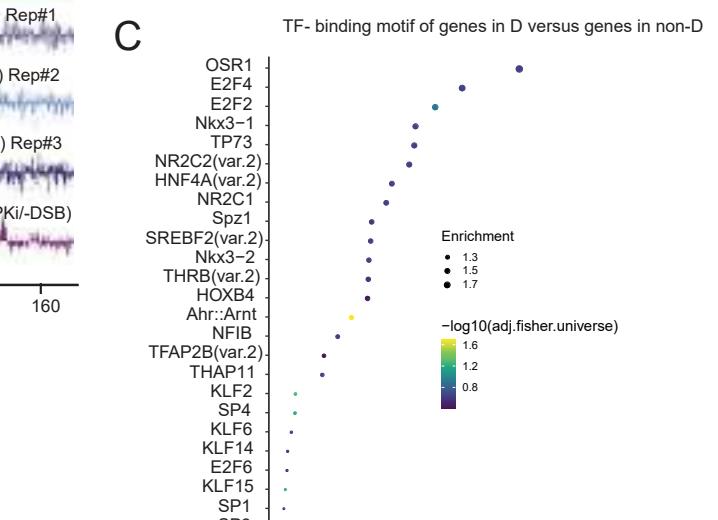
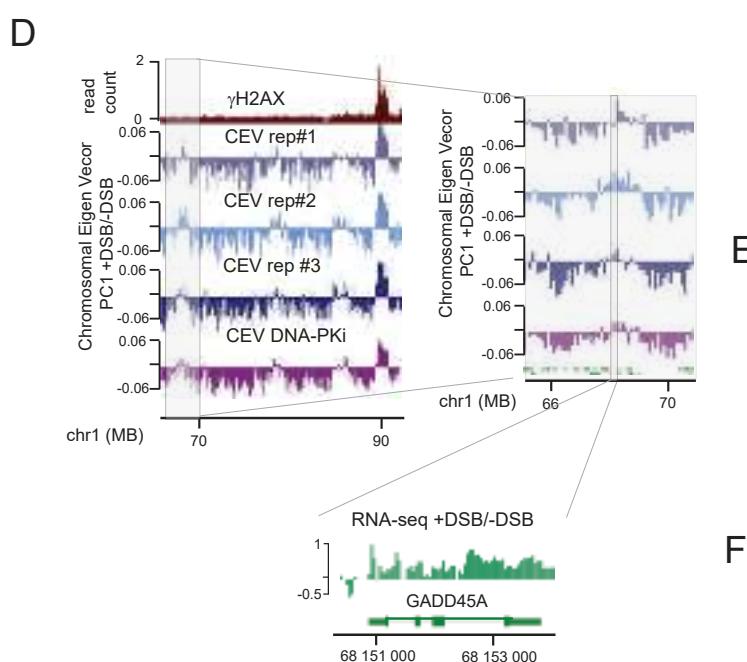
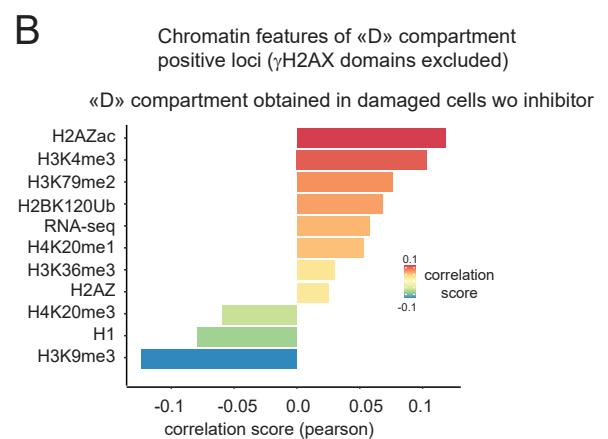
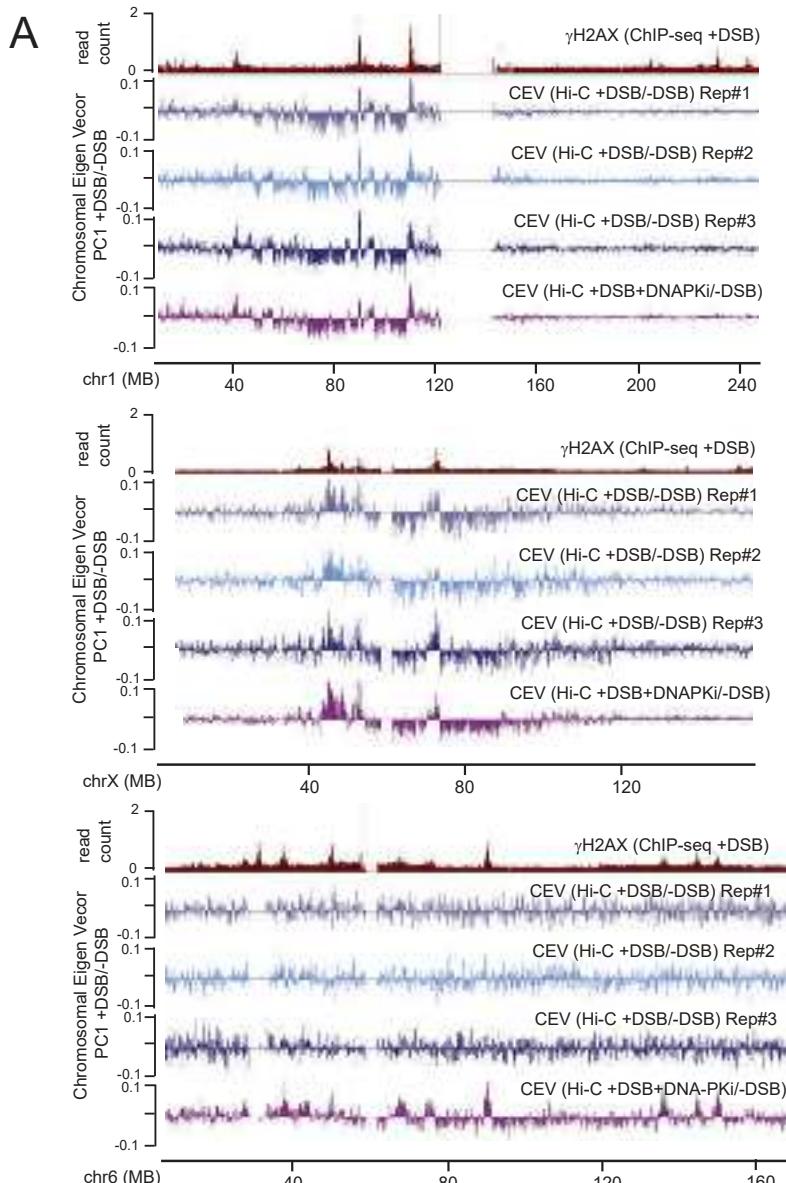
Supplementary Table 1-2

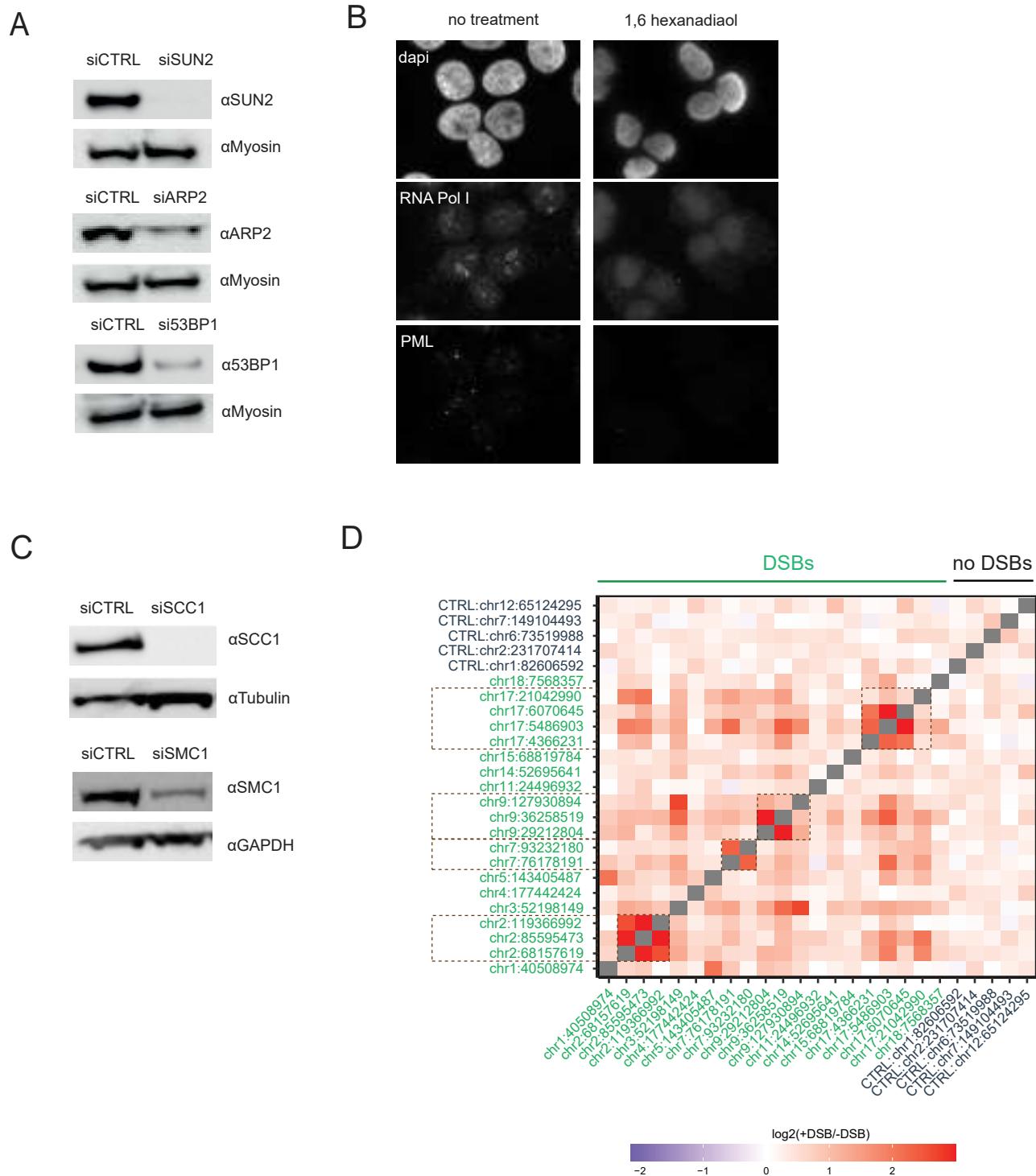
**A****B****C****D****E**

**A****B****C**

**A****B****C****D**







**Figure S1 Related to Figure 1. DSB-induced contact frequency occurring on undamaged chromatin**

- (A) Hi-C contact matrix of the log2 ratio between damaged (+DSB) *versus* undamaged (-DSB) DIvA cells. A region of the chromosome 1 is shown at three different resolutions: 250 kb (left panel), 100 kb (middle panel) and 25 kb (right panel). The  $\gamma$ H2AX ChIP-seq signal following DSB induction is shown on the top panel and indicates the DSBs position. The grey track on the right panel shows the insulation score pre-existing to DSB induction (from Hi-C -DSB). One representative experiment is shown.
- (B) Left panel: Boxplot showing the differential Hi-C read counts (as ( $\log_2 +\text{DSB}/-\text{DSB}$ )) within  $-0.5/+0.5\text{Mb}$  chromatin domains containing the 80 best induced DSBs in the DIvA cell line (80 DSBs) or within their adjacent chromatin domains. DSB-induced increase of *cis* contacts is significantly stronger for the domains that carry the DSB compared to their adjacent domains. Right panel: same as above but for 80 random undamaged regions (random (no DSB)).
- (C) Averaged Hi-C contact matrix of  $\log_2 (+\text{DSB}/-\text{DSB})$  centered on 5000 random undamaged regions of the chromosome 1 (most damaged chromosome, 16 DSBs) or of the chromosome 16 (least damaged chromosome, 2 DSBs) (50 kb resolution on a 5 Mb window). Domains of  $-1/+1\text{Mb}$  around DSBs were excluded from analysis. One representative experiment is shown.
- (D) Contact probability as a function of distance before (-DSB) and after DSB induction (+DSB) on the chromosome 1. Domains of  $-1/+1\text{Mb}$  around DSBs were excluded from analysis. One representative experiment is shown.

(E) Boxplots showing the differential loops strength computed on individual chromosomes from Hi-C data obtained before and after DSB induction and categorized based on the number of DSBs per chromosomes. -1/+1Mb around DSBs were excluded from the analysis.

**Figure S2. Related to Figure 1. ATM and cohesin-dependent local changes within damaged and undamaged TADs**

(A) Averaged Hi-C contact matrix of the differential signal ( $\log_2 +\text{DSB}/-\text{DSB}$ ) upon no inhibition (top panel), DNA-PK inhibition (middle panel) or ATM inhibition (bottom panel), centered on 5000 random undamaged regions (50 kb resolution on a 5 Mb window). Domains of -1/+1Mb around DSBs were excluded from analysis.

(B) Boxplot showing the differential Hi-C read counts ( $\log_2 +\text{DSB}/-\text{DSB}$ ) within  $-0.5/+0.5\text{Mb}$  regions containing the 80 best induced DSBs (Damaged domains, red) or between these 80 damaged domains and their adjacent domains (blue) in untreated condition (no inhibitor), upon ATM inhibition (ATMi) or upon DNA-PK inhibition (DNA-PKi).

(C) Averaged Hi-C contact matrix of the  $\log_2 (+\text{DSB}/-\text{DSB})$  in Control (Ctrl, left panel) or SCC1-depleted cells (right panel), centered on 5000 random undamaged regions (100 kb resolution on a 5 Mb window). Domains of -1/+1Mb around DSBs were excluded from analysis.

**Figure S3. Related to Figure 2. Local transcriptional repression correlates with chromosome conformation**

- (A) Mean differential expression ( $\log_2 +\text{DSB}/-\text{DSB}$ ) upon increasing genomic distance from DSB in a log scale, measured by RNA-seq in damaged and undamaged conditions (from  $n= 2$  independent replicates).
- (B) As in (A) but using a linear scale for distances to DSB between 100kb and 1Mb.
- (C) RT-qPCR quantification of two genes embedded in  $\gamma$ H2AX domains before or after DSB induction and upon Control or SCC1 depletion. Data are normalized to RPLP0 cDNA level and to undamaged condition.
- (D) Genomic tracks showing RNA-seq read count before (-DSB, blue) and after DSB induction (+DSB, red) as well as differential RNA-seq ( $\log_2$  ratio  $+\text{DSB}/-\text{DSB}$ , green) and CTCF ChIP-seq (black) on two genomic locations embedded in a  $\gamma$ H2AX domain and respectively located at ~200 kb and ~800 kb from a DSB site.

**Figure S4. Related to Figure 3. DSB clustering depends on ATM but not on DNAPK and cohesin, and is enhanced in G1**

- (A) Hi-C contact matrix showing  $\log_2(+\text{DSB}/-\text{DSB})$  on a region of the chromosome 17 at 50 kb resolution. Genomic tracks for  $\gamma$ H2AX, 53BP1 ChIP-seq and BLESS following DSB induction are shown on the top panel. One representative experiment is shown.
- (B) Same as in (A) but between a region of chromosome 17 and a region of chromosome 20.
- (C) Boxplot showing the quantification of the differential Hi-C read counts ( $\log_2 (+\text{DSB}/-\text{DSB})$ ) between the 80 most-damaged chromatin domains in control or SCC1-depleted conditions (*cis* contacts were excluded).

(D) Boxplot showing the quantification of the differential Hi-C read counts ( $\log_2(+\text{DSB}/-\text{DSB})$ ) between the 80 most-damaged chromatin domains (80 DSBs) or 80 random undamaged sites (no DSB (80 random)) in untreated cells (red), in cells treated with ATM inhibitor (green) or with DNA-PK inhibitor (grey) (*cis* contacts were excluded).

(E) Genomics tracks showing differential 4C-seq ( $\log_2 (+\text{DSB}/-\text{DSB})$ ) (smoothed with a 10 kb span) obtained using a DSB located on chr20 as a viewpoint (red arrow),  $\gamma$ H2AX ChIP-seq and BLESS, on a ~8 Mb window of chromosome 20 (top panel) and on a ~8 Mb window of chromosome 17 (bottom panel). Black arrows represent interactions between the DSB targeted by the viewpoint and two other DSBs, one located on the same chromosome (chr20) and one located on another chromosome (chr17). One representative experiment is shown.

(F) *Trans* interactions between the view point and the other DSBs (n=79) were computed from 4C-seq experiments. Two cluster-prone DSB, one not cluster-prone and one control undamaged locus were used as viewpoints in 4C-seq experiments. n=4 independent replicates.

(G) *Trans* interactions ( $\log_2 (+\text{DSB}/-\text{DSB})$ ) between the view point and the other DSBs (n=79) were computed from 4C-seq experiments in untreated cells (black) or in cells treated with DNA-PK inhibitor (purple). Two cluster-prone DSB, one not cluster-prone and one control undamaged locus were used as viewpoints. n=2 independent replicates.

(H) Genomic tracks showing the differential 4C-seq signal ( $\log_2 (+\text{DSB}/-\text{DSB})$ ) at the DSB located on chr17: 61 850 870 upon no inhibition (black) or upon DNA-PK inhibition (purple), smoothed using a 10kb span. 4C-seq was performed using as a view point a DSB located on chr17:57 168 614. One representative experiment is shown.

(I) same as in (H) but showing 4C-seq signal ( $\log_2(+\text{DSB}/-\text{DSB})$ ) in G1 (blue), S phase (green) and G2 (turquoise). One representative experiment is shown. The BLESS signal after DSB induction is also shown.

**Figure S5. Related to Figure 4. D compartment recruits upregulated DNA damage responsive genes.**

(A) Genomic tracks of  $\gamma$ H2AX ChIP-seq and first Chromosomal eigenvector (CEV) computed on differential (+DSB/-DSB) Hi-C matrix on chromosome 1 (top panel), chromosome X (middle panel) and chromosome 6 (bottom panel)). Biological replicate experiments are shown as well as the CEV obtained upon DNA-PK inhibition.

(B) Pearson correlation score was calculated on 100kb bins between various histone modifications ChIP-seq data or RNA-seq obtained in DIvA cells and the D compartment (CEV) signal computed from +DSB/-DSB (top panel)) or upon DNA-PK inhibition (+DSB+DNA-PKi/-DSB (bottom panel)).

(C) Enrichment for transcription factor motifs was analyzed for genes that displayed a positive D compartment value, as compared with genes that displayed a negative D compartment value.

(D) As in (A) but with a zoom on an undamaged region of the chromosome 1 that displayed positive D compartment signal. The differential RNA-seq ( $\log_2 (+\text{DSB}/-\text{DSB})$ ) for this region containing the p53-target gene GADD45A is also shown (green).

(E) Boxplot showing D compartment value (PC1 calculated from Hi-C data (+DSB+DNA-PKi/-DSB) on genes that are known direct targets of the p53 transcription factor and either not regulated, or upregulated following DSB induction in DIvA cells (identified by RNA-seq).

(F) same as in (E) but for known NF- $\kappa$ B target genes.

**Figure S6. Related to Figure 5.**

(A) Western Blot showing the depletion of SUN2, ARP2 or 53BP1 by siRNA.

(B) Immunofluorescence experiment showing RNA Pol I, PML and DAPI staining before and after a 1,6-Hexanediol 10% treatment as indicated.

(C) Western Blot showing the depletion of SCC1 or SMC1 by siRNA.

(D) Heatmaps showing the rate of translocations sequenced between five control undamaged regions and twenty DSB sites in the AID-DIVA cell line (expressed as the log2 (+DSB / -DSB)).  
n=4 independent experiments

**Supplementary table 1: Primers sequences for Amplicon-Seq**

Name	Chr	coordinate of AsiSI site (hg19)	Primer coordinate (hg19)	sequence FW	sequence REV
<b>CTRL_CHR1</b>	chr1		83072163	GCACATGGGATTTGCAAGAACATGTCTGTCACTGG	CCATCTGGCTTAGTCCACTTCATACTCAGC
<b>CTRL_CHR12</b>	chr12		65517948	CTAGCATATGATAAAAATTGAAATAATGTGCTAGGC	GGCTATGATCATGTTTATTGTACTATAATTCCAGGTC
<b>CTRL_EEF1A1</b>	chr6		74229585	CCTTCTGGTATTAACATACCTCAGCAGCC	GATCTTGGTTCATTCTCAAGCCTCAGACAGTGG
<b>CTRL_PTMap</b>	chr2		232572004	CGACAGCTAAGTGGGTGGCGATAAGGCCAGCG	CAGAGTCTCCTGGGAAAATCATGTCCTGGAGGCC
<b>CTRL_ZNF425</b>	chr7		148801459	CGCCGCAAGGAAACTCCTCTGCCTGCTGTTGAC	CGAGTGTAAACAAAGTTCCGCCTCAAGAGAAAGCCTG
<b>SITE516</b>	chr1	40974643	40974513	GCGTCCGGGAGCAGCTCGAGGCCGCGGCG	GAAGGAGTCCCAGGCCAACAGCACTGTGTTCCGA
<b>SITE709</b>	chr2	68384748	68384627	CGGCAACCTGAAACAGACTCAAACATTGGC	GGTCTTGGGCTACTCCAGCCGCAGCAACTCCAG
<b>SITE716</b>	chr2	85822593	85822445	TCCGGAAGCGCCATGGCCAACGCTCGCA	CCCTCTCATGGCTTGAAACAAGTCAGGTGACC
<b>SITE722</b>	chr2	120124565	120124434	GCTAGCGCCGCGGGGGCTGGCACGC	GAGCCCCAACGGGCCTCCCGCCCGCTGCACC
<b>SITE765</b>	chr3	52232162	52232026	TCCGCAAGGCCAAGAAGTGGGCTCTGC	GCCGCACCTACCGCGCGGCCCTCACCTGCGCTG
<b>SITE871</b>	chr4	178363575	178363436	TGACGACCAAGGGCAGAGGGCTGGAGCAGC	CGTGACACAACCTCTCCCGGGGCCAGGGACG
<b>SITE919</b>	chr5	142785049	142784920	CGGCGGCTCCCTGCTGACATCTGAAGACG	CCTCTGAGGCGGCCCGTAGATGTCCTCCGG
<b>SITE1025</b>	chr7	75807506	75807383	GCGCCCAGGGCGAAGGTACTAGACAGC	AGCCTAGCCTCTGGCTCCTCTGCAGGCAG
<b>SITE1031</b>	chr7	92861490	92861341	TCACACTCTAGGGGACATCGCTCAATCTGGC	GCTGGGAACTGTAGTCTCTGGACGCCGGTAGCAACG
<b>SITE1123</b>	chr9	29212799	29212659	TCAACCTCCACAGAGAACCCAGTCCGAATCG	GCCTCAAGGTAAGAGGCAGCAAAGCTGCTGCTT
<b>SITE1129</b>	chr9	36258513	36258383	GCCACGAAGCAGGCAGAGCGCGAGC	CATGGAAGATGGTCCGCTGGTCAGC
<b>SITE1159</b>	chr9	130693170	130693031	TCGCCCCCGCAGCAGCTCCCGGGCTTGGCCG	CAGGCGGCCCTAGCGACCCGAGTCCCCACGCCG
<b>SITE72</b>	chr11	24518475	24518337	CAATTATCGGGCAGATTAGGTTCTACTCTCGT	CAGGAGTGTAGTCACCGCGAGGGCGAGCCCTGCC
<b>SITE219</b>	chr14	54955825	53162224	CACGCTTGGGAGGCCAGTCCGACGCTCGGTG	GCGGGGCCGAGCCGGAAAGCTGGTCGGTGC
<b>SITE270</b>	chr15	69112120	69111977	CTGCCGACAAGCCCAGCAGTGGAGAAAGACGGG	TCCACCCACCTTCGGGGCAGATGCTGATGTTCT
<b>SITE341</b>	chr17	4269523	4269398	CGGGCGAGAACGGCTGGGCCGGCCGGACCG	AGCCCCGGCGCCAGGCCGGCCCTCGC
<b>SITE343</b>	chr17	5390220	5390088	CGGCCCTGTGGGTCGGCCTACCCGGCCCTCG	CCGCCTCATCTCTCTGCTAGACTAGAGTTCTG
<b>SITE344</b>	chr17	5973962	5973831	GTCCGCTTGGCCGGCCGGAGACGTGC	CGCCGCGCTCCAAACCGCGCTGAC
<b>SITE360</b>	chr17	20946300	20946179	CGGCAGGGAGGGAGGCCAGGACGACGCCAG	GTGACGTAATCTCCGTCGGCCGGCG
<b>SITE396</b>	chr18	7566712	7568200	GAGTCCCTGGCCGGCTGCAAAGAACAGCAG	AGCCTCCGCAGACGCTGACGCCGCG

**Supplementary table 2: Primer sequences for 4C-seq**

Name	Forward primer	Reverse primer
Viewpoint USE1 (chr1, cluster-prone)	AATGATAACCCCGGGACCGGGAGATCTAACACTCTTTCCT CTACAGGAAAGCTTCTCGATCTAACCTGGACCTTA TGATTCAGGGA	CAGGCAAGAGACGGCGGATACGGAGATNNNNNNTGAC TGAAATTGAGAAGCTGTCTCTGGATTTATGCTAA AAGGCAAGGGGGACA
Viewpoint USE2 (chr17, cluster-prone)	AATGATAACCCCGGGACCGGGAGATCTAACACTCTTTCCT CTACAGGAAAGCTTCTCGATCTAACCTGGATTT TGAAATTGATG	CAAGCAAAAGACGGCGGATACGGAGATNNNNNNTGAC TGAAATTGAGAAGCTGTCTCTGGATTTATGCTAA TCTGAGCTTAAAGCTT
Viewpoint USE3 (chr21, not cluster-prone)	AATGATAACCCCGGGACCGGGAGATCTAACACTCTTTCCT CTACAGGAAAGCTTCTCGATCTATTAACTTAAAGCT TGCGG	CAGGCAAGAGACGGCGGATACGGAGATNNNNNNTGAC TGAAATTGAGAAGCTGTCTCTGGATTTATGCTAA AATGAAACCGCTG
Viewpoint USE4 (chr20, cluster-prone)	AATGATAACCCCGGGACCGGGAGATCTAACACTCTTTCCT CTACAGGAAAGCTTCTCGATCTGGTTTATTAAGAT TGCGG TCGT	CAAGCAAAAGACGGCGGATACGGAGATNNNNNNTGAC TGAAATTGAGAAGCTGTCTCTGGATTTATGCTAA CGCTGGTTTGAT
Viewpoint end region (chr17)	AATGATAACCCCGGGACCGGGAGATCTAACACTCTTTCCT CTACAGGAAAGCTTCTCGATCTTGCGG TCGT TGCGG	CAGGCAAGAGACGGCGGATACGGAGATNNNNNNTGAC TGAAATTGAGAAGCTGTCTCTGGATTTATGCTAA CGCTGAGCTT

NNN is the position of the optional index

### 2.2.5 Discussion

Notre travail a permis de déterminer l'impact des *DSB* sur la conformation 3D de la chromatine, en *cis* (à proximité), mais également en *trans* (à l'échelle du génome). Nos données suggèrent que la structure tri-dimensionnelle de la chromatine joue un rôle primordial dans la réponse aux dommages à l'ADN, mais augmente le risque de translocations (voir Figure 2.5).

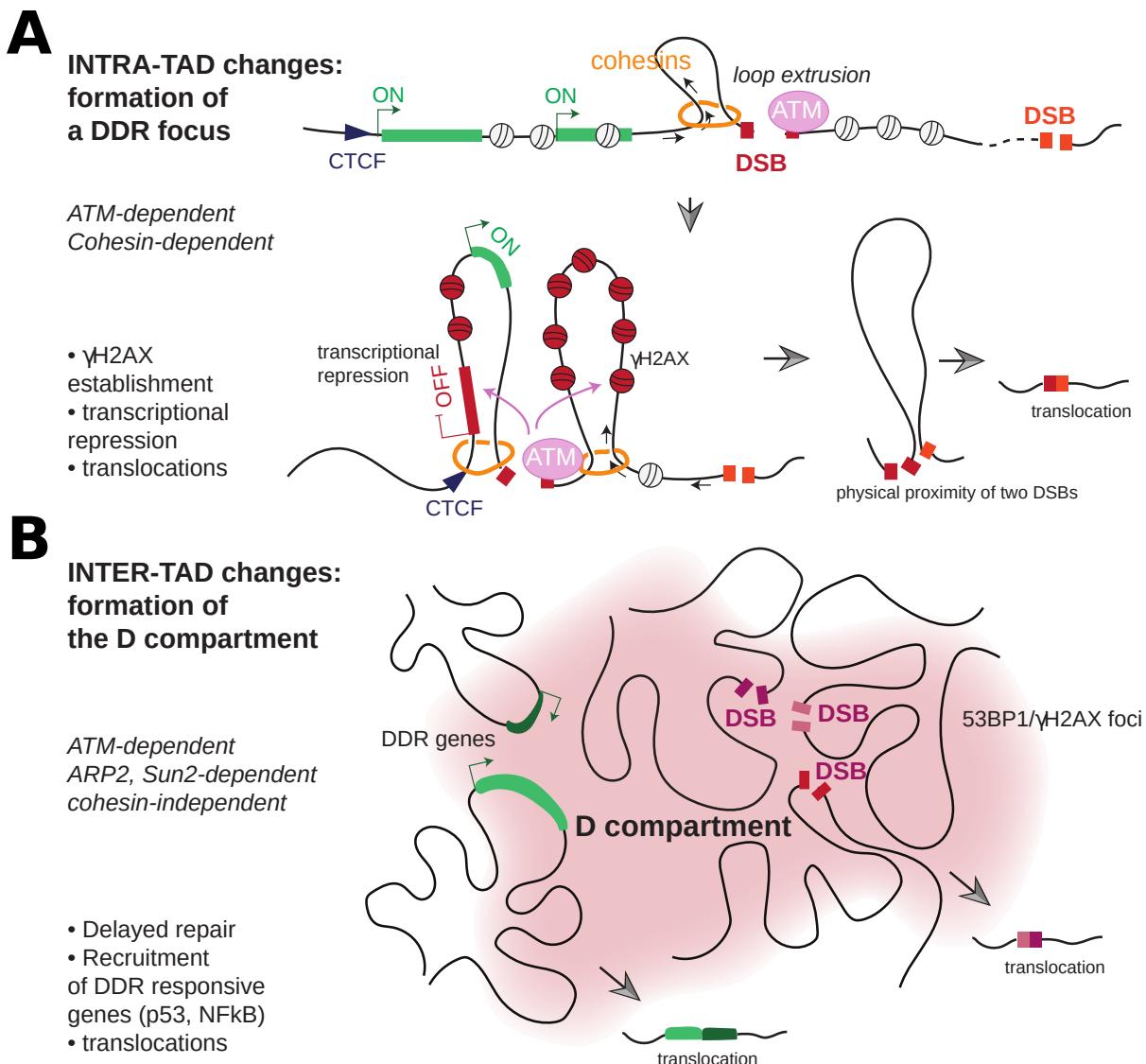


FIGURE 2.5 – Modèle proposé expliquant les changements d’interaction observés après dommages. **A** Le mécanisme de *loop extrusion* permet la phosphorylation du variant H2AX par ATM. Ce phénomène peut réguler négativement l’expression d’un gène lorsqu’il est mis en contact prolongé avec la *DSB*. En outre, le phénomène de *loop extrusion* peut mettre en contact deux *DSB* du même chromosome, augmentant le risque de translocations. **B** Les domaines gH2AX/53BP1 peuvent former des *clusters*, potentiellement via la séparation de phase, et forment un nouveau compartiment *DSB*-dépendant. Ce compartiment relocalise les gènes de la *DNA Damage Response (DDR)* afin d’affiner leurs réponses.

### 2.2.5.1 Le rôle de la *loop extrusion* dans la réparation des *DSB*

Les résultats sur l'étude en *cis* de la conformation 3D du génome à l'aide de données Hi-C ont permis de montrer que ce sont les *TAD* qui favorisent et qui délimitent la signalisation des *DSB*, telle que la phosphorylation de H2AX. Nos données suggèrent que la formation de domaines de réparation de l'échelle du Mégabase sont formés par un mécanisme de *loop extrusion* dépendant de la cohésine, déclenché de chaque côté de la *DSB* (voir Figure 2.6) .

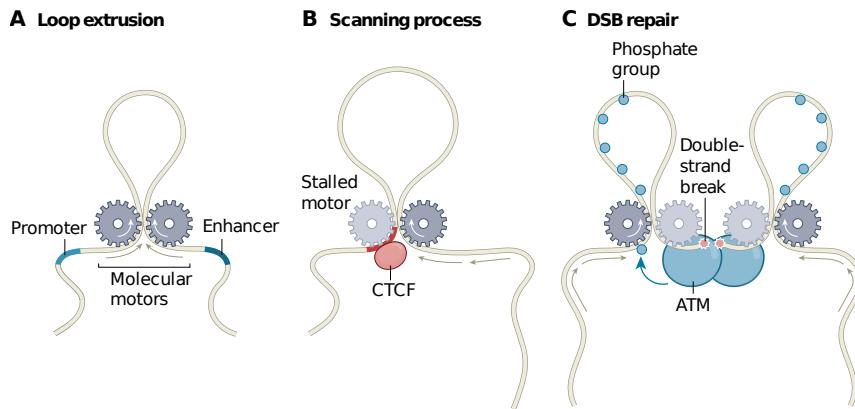


FIGURE 2.6 – Le phénomène de *loop extrusion* à l'origine de la formation des *TAD*. La *loop extrusion* permet de mettre en contact des éléments distants régulant l'expression de gènes (**A**). Ce mécanisme est dépendant de la cohésine et délimité par des éléments insulateurs comme CTCF (**B**). Ce mécanisme est également impliqué dans la réparation des *DSB*, via la phosphorylation de H2AX par ATM à l'échelle du *TAD* entier, par la présence d'évenements de *loop extrusion* des deux côtés de la *DSB* (**C**). De *Loops that mend the genome* par Leonid A. Mirny.

Dans ce modèle, la formation du domaine de réparation dépend de la présence des cohésines qui s'accumulent au niveau des *DSB*. L'origine de cette accumulation n'est cependant pas connue. Il est possible que celle-ci soit due à un recrutement spécifique aux *DSB*, ou alors au fait que la cohésine se retrouve bloquée à proximité des *DSB*. Dans tous les cas, cette accumulation conduit à la formation de boucles de chromatine des deux côtés des *DSB*, via une extrusion unidirectionnelle, permettant à ATM de phosphoryler les variants d'histones H2AX à proximité. Les régions insulatrices enrichies en CTCF vont délimiter et réguler la progression de ces boucles, expliquant la similarité entre le profil ChIP-seq de  $\gamma$ H2AX et le profil d'interaction par 4C-seq centré sur la région endommagée. Cette régulation par CTCF explique également pourquoi la propagation de  $\gamma$ H2AX se limite aux frontières des *TAD*.

Ce phénomène est également complètement dépendant de la présence d'ATM. En effet, la déplétion de cette protéine empêche la formation du processus de *loop extrusion* autour de la *DSB*. La présence de la *DSB* n'est donc a priori pas suffisante en soi pour initier le mécanisme de *loop extrusion*. De plus, il a été montré que les sous-unités de la cohésine étaient elles-mêmes modifiées par ATM (S.-T. KIM, XU et KASTAN 2002), il est possible que cette phosphorylation permette la mise en place de la *loop extrusion* par la cohésine à proximité de la *DSB*.

### 2.2.5.2 Renforcement des *TAD* endommagés

La comparaison des cartes Hi-C avant/après dommages dans nos études a mis en avant le fait que le comportement des *TAD* endommagés est fortement modifié suite à l'induction de *DSB*. En effet, nous

avons constaté que les interactions au sein des *TAD* endommagés (intra-TAD) augmentent fortement au détriment des interactions avec les *TAD* voisins (inter-TAD). Ce renforcement des *TAD* endommagés semble dépendant de l'augmentation du recrutement des cohésines à leurs frontières. Notre étude montre également une augmentation du recrutement des cohésines aux frontières des *TAD* de tout le génome, y compris ceux n'étant pas endommagés. Cette augmentation est cependant moins forte que celles des *TAD* endommagés.

Comme pour la *loop extrusion*, ce renforcement des *TAD* est dépendant d'ATM, et renforcé par l'inhibition de la kinase DNA-PK. L'inhibition de DNA-PK empêche la réparation et rend les *DSB* persistantes. Le processus de *loop extrusion* induite par la *DSB*, ainsi que le renforcement des *TAD* endommagés est donc maintenu. Le fait que DNA-PK, ait un rôle antagoniste à la kinase ATM sur les effets de *loop extrusion* et de renforcement des *TAD* reste à élucider. En outre, une déplétion des cohésines, en plus d'inhiber la *loop extrusion* au niveau des *DSB*, impacte négativement l'isolation des *TAD* endommagés, ainsi que leur renforcement.

Dans l'ensemble, nos données Hi-C montrent que les *DSB* déclenchent à la fois un renforcement des *TAD* dans tout le génome non endommagé, ainsi qu'une isolation des *TAD* endommagés, associé à un processus de *loop extrusion* dépendant de la cohésine et d'ATM.

### 2.2.5.3 Rôle de la *loop extrusion* sur la régulation de la transcription

Au sein des foyers de réparation, dans un domaine d'environ 1 Mb autour des *DSB*, certains gènes subissent une inhibition de leur transcription. Selon une étude du transcriptome (voir partie 1.2.5, IANNELLI et al. (2017)), la diminution de l'expression des gènes corrèle linéairement avec la distance aux *DSB*. Cependant, notre équipe avait mis en évidence le fait que certains gènes proches des *DSB* sont protégés par la cohésine (CARON, Francois AYMARD et al. 2012), mais les facteurs déterminant le choix des gènes réprimés ou protégés n'avaient pas été déterminés. Nos derniers résultats suggèrent que c'est la proximité 3D entre les gènes et les *DSB* qui détermine si le gène est réprimé ou non. En effet, lors du processus de *loop extrusion*, les gènes sont amenés au contact physique de la *DSB*. Le temps de contact entre le gène et la *DSB* serait alors dépendant de la présence ou non d'un site de fixation de CTCF dans une orientation "face" à la *DSB*. La présence de CTCF permettrait l'arrêt du processus de *loop extrusion* suffisamment longtemps pour entraîner une répression du gène, alors que les gènes ne présentant pas de site CTCF, ou dans la mauvaise orientation, ne seraient pas réprimés.

### 2.2.5.4 L'induction de *DSB* mène à la formation d'un nouveau compartiment "D"

Nous avons constaté que suite à l'induction de *DSB*, les *TAD* endommagés ont la capacité de se regrouper et de former des *clusters*. Ce *clustering* est en lien avec les changements induits à l'échelle du TAD, comme la phosphorylation de γH2AX, mais également l'accumulation de modifications ubiquitine permettant entre autres le recrutement de 53BP1. Ce regroupement se produit principalement en phase G1, entre les *DSB* induites dans des régions actives en transcription, et préférentiellement réparés par *HR* (François AYMARD, AGUIRREBENGOA et al. 2017). Nous avons également constaté ce regroupement accru de *DSB* en phase G1 via nos données 4C-seq, en accord avec nos travaux précédents.

Les premières études réalisées sur les données Hi-C (LIEBERMAN-AIDEN et al. 2009) ont permis de mettre en évidence l'existence de deux principaux compartiments de la chromatine, appelés A et B. Cette découverte a été faite grâce à l'utilisation d'une ACP sur les cartes de Hi-C, puis par l'extraction de sa première composante, qui indiquait dans quel compartiment se situait la chromatine. La réalisation d'une ACP sur les données Hi-C se fait sur la matrice normalisée *Observed/Expected* (OE) (voir partie 1.3.6.1.3.2), à une résolution assez faible de 100kb à 1Mb. On calcule sur cette matrice la corrélation de *Pearson*, et on effectue la décomposition en éléments propres qui nous fournit les axes de l'ACP. L'axe qui explique la plus grande proportion de variance représente les compartiments génomiques A et B. Les valeurs positives et négatives de cet axe attribueront un compartiment aux *loci* du chromosome.

Le *clustering* des *DSB* étant visible sur les matrices différentielles, nous avons réalisé une approche similaire à LIEBERMAN-AIDEN et al. (2009), mais cette fois sur la matrice différentielle avant/après dommage  $\log_2(\frac{\text{damaged}}{\text{undamaged}})$ . Ceci a permis de découvrir un nouveau compartiment "D", spécifiquement induit après l'induction de *DSB* sur le génome. Ce compartiment apparaît lorsque les domaines endommagés, enrichis en  $\gamma$ H2AX/53BP1 par *loop extrusion* se regroupent et s'isolent du reste de la chromatine. Ce regroupement pourrait s'expliquer grâce aux propriétés de séparation de phase de 53BP1 (KILIC et al. 2019).

Dans le cas de l'inhibition de DNA-PK, ou en phase G1, Les *DSB* observées sont persistantes, et apparaissent avec une plus grande fréquence dans la population cellulaire. Le processus de *loop extrusion* serait alors maintenu, accumulant  $\gamma$ H2AX/53BP1. Ces procédés favorisent donc le *clustering*, et seraient à l'origine du signal renforcé du compartiment D observé après inhibition de DNA-PK.

Dans des travaux précédents, il a été montré que 53BP1 est essentiel à l'activation des gènes cibles de p53 (CUELLA-MARTIN et al. 2016). De plus, la perturbation de la séparation de phase de 53BP1 a montré une altération de l'activation de ces gènes (KILIC et al. 2019). À l'inverse, une amélioration de cette séparation de phase conduit à une réponse transcriptionnelle améliorée (GHODKE et al. 2021). Dans l'ensemble, ces études montrent que la formation d'un compartiment isolé enrichi en 53BP1 permet l'activation de la réponse de p53. Ici, nos résultats montrent que les gènes ciblés par p53 sont enrichis en compartiment D. Des gènes connus pour intervenir dans la *DNA Damage Response (DDR)* et qui présentent une augmentation de leur expression après dommages sont également enrichis. La formation du compartiment D jouerait un rôle important dans l'activation de la *DDR*, en relocalisant ces gènes à proximité des *DSB* afin d'ajuster avec précision son amplitude de réponse par rapport au niveau d'endommagement.

### 2.2.5.5 Les changements de conformation de la chromatine après *DSB* contribuent aux translocations

Il a été montré que regrouper des *DSB* pouvait conduire à des translocations (ROUKOS et al. 2013). Ce phénomène est également observé sur des mécanismes biologiques, notamment la biogénèse des anticorps par le mécanisme de recombinaison V(D)J (SOULAS-SPRAUEL et al. 2007). Le regroupement des *DSB* et leur rapprochement en *cis* par *loop extrusion* pourrait conduire à des translocations à la fois intra et inter-chromosomiques, et être délétère pour la cellule.

En utilisant des données publiées de patients atteints de cancers, nous avons pu récupérer les positions des translocations observées dans cette population. Ces données suggèrent que la relocalisation des

gènes de la *DDR* vers le compartiment D peut potentiellement déclencher des réarrangements qui sont délétères à la cellule, et, en perturbant ces gènes suppresseurs de tumeurs, conduire à l'apparition de cancers. L'induction de *DSB* étant utilisée dans certaines thérapies anti-cancéreuses, il est possible que l'apparition de cancers secondaires suite à ces thérapies soit en partie due à ce phénomène (EZOE 2012).

Nous avons pu conclure cela en localisant les positions des cassures à l'origine de la translocation, et en les comparant avec les positions des régions qui se regroupent en compartiment D. Nous avons constaté que les gènes de la *DDR* relocalisés montrent un chevauchement significatif avec des événements de translocations observés dans les cancers. Les bases de données telles que *TCGA* et *ICGC* (voir partie 1.3.7.1) mettent à disposition des données génomiques du cancer, qui sont régulièrement mises à jour. Parmi les données disponibles, on peut accéder au transcriptome des patients mais également à un catalogue des anomalies génétiques observées dans différentes cohortes de cancers. Ces données publiques nous permettraient de disposer d'informations supplémentaires sur les mutations génétiques, et notamment les translocations, afin d'étudier les conséquences de la formation de ce compartiment sur l'intégrité du génome.

#### **2.2.5.6 Le compartiment D, un effet inter-chromosomique visible uniquement en *cis***

L'inconvénient d'utiliser la même méthode d'ACP définie dans LIEBERMAN-AIDEN et al. (2009), est que celle-ci est construite uniquement en se basant sur les interactions intra-chromosomiques. En effet, cette ACP est construite sur une matrice qui contient l'interaction du chromosome avec lui-même, et ne prend pas en compte les interactions inter-chromosomiques.

Cependant, le *clustering* de *DSB* s'observe également en *trans*, grâce aux interactions inter-chromosomiques, qui ne sont pas prises en compte. Ceci, couplé avec le fait que les domaines formés autour des *DSB* sont très grands, réduit fortement l'origine des données que nous pouvons étudier. Pour résoudre ce problème, il faudrait pouvoir réaliser l'extraction du compartiment D, soit le premier axe de l'ACP, sur une matrice Hi-C pangénomique.

La construction d'une telle matrice est problématique, car même à résolution réduite (100kb), elle consiste à compter les fréquences d'interaction entre plus de 30 000 *loci* d'ADN, ce qui revient à construire une matrice de près d'un milliard de cellules, et consomme beaucoup de ressources. Pour résoudre ce problème, on peut utiliser des méthodes alternatives qui s'appliquent bien à de grands jeux de données, voir utiliser des techniques associées au *big data*. Par exemple, on peut utiliser des méthodes de calculs distribués pour faire une ACP en utilisant des librairies comme Apache Spark. On peut également construire un modèle de *Deep Learning* appelé *Autoencoder*, qui déconstruit et reconstruit les données d'origine de manière non-supervisée. Cependant, les méthodes implémentées en *Machine Learning* ont souvent des termes de régularisation, qui suppriment les variables non informatives, et n'ont parfois aucune garantie sur l'orthogonalité des axes trouvés, produisant un résultat différent d'une ACP classique.

On peut également se représenter la formation d'un nouveau compartiment comme un changement de compartimentation entre deux conditions et appliquer des analyses différentielles applicables aux données Hi-C comme *diffHiC* (LUN et Gordon K SMYTH 2015). *diffHiC* applique une méthode similaire à l'analyse différentielle pour les données RNA-seq, et utilise un *GLM* pour modéliser les comptages entre les paires de *bins* d'une matrice Hi-C. L'analyse différentielle est donc directement appliquée dans le modèle, et on n'a donc pas besoin de calculer directement cette matrice. *diffHiC* permet d'identifier des paires de

*bins* qui présentent un changement significatif d'interaction. Les *bins* ainsi détectés pourraient permettre d'identifier les régions qui se regroupent pour former le compartiment D.

Enfin, on peut se représenter le problème de manière supervisée, et séparer le génome en deux états, "D" et "Non D", et utiliser d'autres données omiques comme les marques et l'accessibilité de la chromatine pour déterminer ce compartiment (FORTIN et HANSEN 2015). Une autre étude a également modélisé les interactions Hi-C par des marques épigénétiques (NICHOLS et CORCES 2021). Grâce à ce modèle, ils ont pu identifier quelles étaient les marques de chromatine importantes à la prédiction des compartiments génomiques, et proposent une approche plus "fine" que les méthodes classiques par ACP. En effet, les compartiments A et B retrouvés par ACP montrent des compositions en modifications d'histones qui varient selon le type cellulaire. Leur méthode s'éloigne de cette approche "binaire" A/B afin de mieux comprendre les différences de compartiments observées entre différents types cellulaires et d'étudier le rôle des marques épigénétiques dans leur formation.

Leur algorithme modélise un phénomène d'attraction/répulsion utilisant les modifications d'histones, et montre que les régions génomiques contenant des niveaux élevés d'une modification d'histone donnée interagissent plus fréquemment avec d'autres régions ayant la même modification. À l'inverse, il apprend que les interactions seront faibles entre régions ayant une composition différente.

Le phénomène conduisant à la formation du *clustering*, et donc du compartiment D étant très dépendant de la composition de la chromatine, il serait intéressant d'utiliser ce type d'approche pour modéliser ce compartiment et ainsi déterminer par une méthode différente à l'ACP les régions qui le forment.

### 2.2.5.7 Conclusion générale

Dans un premier projet, j'ai développé un modèle de *Deep Learning* capable de prédire des régions contenant des G-quadruplexes (G4) actifs, sur n'importe quel type cellulaire en utilisant la séquence et l'accessibilité de l'ADN. Avec ce modèle, nous avons cartographié des milliers de régions G4 actives qui peuvent être utilisées comme cibles potentielles dans les traitements anti-cancer. En effet, la présence de G4 dans les promoteurs des oncogènes en font des cibles thérapeutiques de choix pour l'utilisation de thérapies à base de ligands de G4. En plus de cela, les G4 peuvent également entrer en collision avec la machinerie de réplication et causer des dommages à l'ADN. Elles sont donc source d'instabilité génomique (KOSIOL et al. 2021).

Dans un second projet, j'ai traité et analysé des données de Hi-C afin de mettre en évidence l'importance de l'organisation 3D du génome dans la signalisation des cassures double-brin (*DSB*). Nous avons découverts le rôle fonctionnel de *TAD* préexistants dans la signalisation et l'isolement des foyers de réparation. De plus, nous avons mis en évidence que le processus qui permet la formation des foyers de réparation par ATM dépend de la *loop extrusion*. Nous avons également montré que les *DSB* peuvent former des *clusters* et se regrouper pour former un nouveau compartiment génomique. Ce compartiment, composé de chromatine décorée par  $\gamma$ H2AX, est également composé de gènes de la réponse aux dommages à l'ADN (DDR). Ce nouveau mécanisme permettrait de réguler la réponse de la DDR par rapport au niveau d'endommagement, au détriment de translocations et d'instabilité génomique.

Les G4 sont capables de provoquer des cassures endogènes lorsque les machineries cellulaires entrent en collision avec ces structures. De la même manière que certains traitements anti-cancer, comme l'étoposide,

## 2.2. RÉPARATION DE L'ADN ET STRUCTURE TRIDIMENSIONNELLE DU GÉNOME

les traitements à base de ligands de G4 pourraient provoquer des *DSB* par l'action de la topoisomérase 2 (PIPIER et al. 2020), contribuer à l'instabilité génomique, et ainsi être utilisée dans les thérapies anti-cancéreuses (KOSIOL et al. 2021). La façon dont la conformation 3D de la chromatine prend en charge les *DSB* endogènes est encore peu étudiée, et il serait intéressant d'appliquer les mêmes types d'analyses que sur notre modèle cellulaire DIvA. Dans cette optique, notre outil DeepG4 nous permettrait d'identifier les régions les plus susceptibles d'être endommagées. En outre, il a été montré que les G4 pouvaient former des boucles de chromatine (L. LI et al. 2021), et la relation entre le *clustering* des *DSB* et celui des G4 est à étudier.



# Appendix

Au cours de ma première année de thèse, j'ai également eu l'opportunité de travailler sur l'article de Clouaire et al. 2018 (Molecular Cell). Pendant ce projet, je me suis occupé du traitement et de l'analyse bio-informatique des données de séquençage. J'ai également généré une grande partie des figures permettant la visualisation de nombreuses données sous la forme de profils moyens et de *boxplots*. Enfin, je me suis occupé de la soumission de ces données sur *ArrayExpress*, afin de les mettre à disposition pour la communauté scientifique.

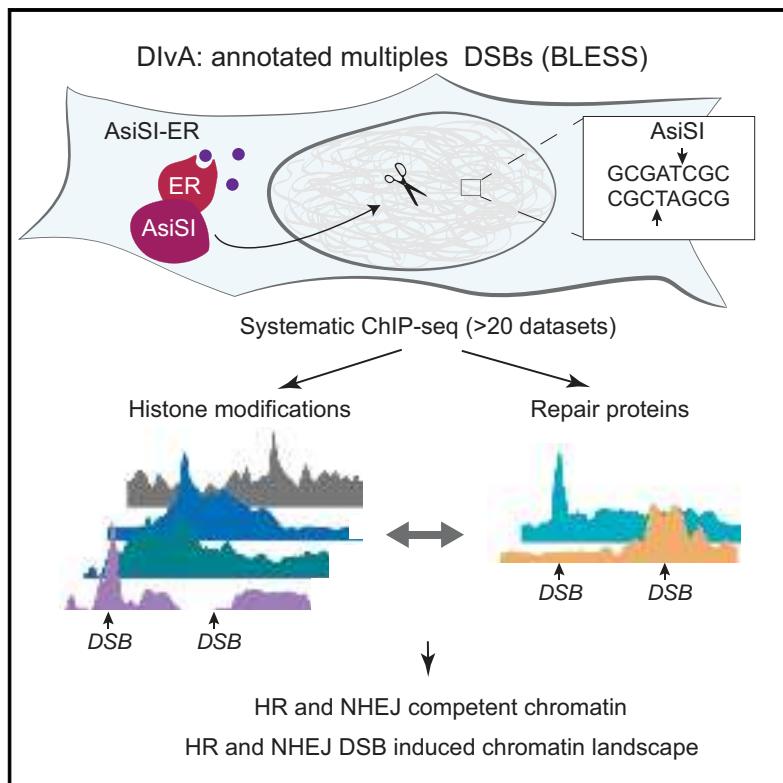
Ce projet a permis de générer de nombreuses données ChIP-seq dans notre lignée cellulaire DIvA qui fournissent des informations utiles sur le contexte chromatinien autour des *DSB*, et sur les changements observés après endommagement de l'ADN, en fonction de la voie de réparation. Ces données sont très souvent utilisées dans les nouveaux projets de l'équipe, dont les articles présentés dans mes résultats.

**La cartographie complète des modifications d'histones au niveau des cassures double-brin de l'ADN déchiffre les signatures des voies de réparation**

# Molecular Cell

## Comprehensive Mapping of Histone Modifications at DNA Double-Strand Breaks Deciphers Repair Pathway Chromatin Signatures

### Graphical Abstract



### Authors

Thomas Clouaire, Vincent Rocher,  
Anahita Lashgari, ...,  
Krzysztof Ginalski, Jacques Côté,  
Gaëlle Legube

### Correspondence

thomas.clouaire@univ-tlse3.fr (T.C.),  
gaelle.legube@univ-tlse3.fr (G.L.)

### In Brief

Using ChIP-seq in a cell line where multiple annotated DNA double-strand breaks can be induced on the human genome, Clouaire et al. report a comprehensive view of the chromatin landscape set up at DSBs and decipher the chromatin signature associated with HR and NHEJ repair.

### Highlights

- DSB-chromatin landscape and HR/NHEJ chromatin signatures uncovered by ChIP-seq
- H2BK120 undergoes a switch from ubiquitination to acetylation at a local scale
- H1 is removed and ubiquitin accumulates on entire γH2AX domains, mainly at HR DSB
- 53BP1 spreads over megabase-sized domains, mostly in G1 at HR-prone DSBs



# Comprehensive Mapping of Histone Modifications at DNA Double-Strand Breaks Deciphers Repair Pathway Chromatin Signatures

Thomas Clouaire,<sup>1,\*</sup> Vincent Rocher,<sup>1</sup> Anahita Lashgari,<sup>2</sup> Coline Arnould,<sup>1</sup> Marion Aguirrebengoa,<sup>1</sup> Anna Biernacka,<sup>3</sup> Magdalena Skrzypczak,<sup>3</sup> François Aymard,<sup>1</sup> Bernard Fongang,<sup>4</sup> Norbert Dojer,<sup>4,5</sup> Jason S. Iacoboni,<sup>6</sup> Maga Rowicka,<sup>4</sup> Krzysztof Ginalski,<sup>3</sup> Jacques Côté,<sup>2</sup> and Gaëlle Legube<sup>1,7,8,\*</sup>

<sup>1</sup>LBCMCP, Centre de Biologie Intégrative (CBI), CNRS, Université de Toulouse, UT3, Toulouse 31062, France

<sup>2</sup>St-Patrick Research Group in Basic Oncology, Laval University Cancer Research Center, Oncology Axis-CHU de Québec-Université Laval Research Center, Quebec City, QC G1R 3S3, Canada

<sup>3</sup>Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Zwirki i Wigury Warsaw 93, 02-089, Poland

<sup>4</sup>Department of Biochemistry and Molecular Biology, University of Texas Medical Branch, Galveston, TX 77555-0615, USA

<sup>5</sup>Institute of Informatics, University of Warsaw, Banacha 2, 02-097 Warsaw, Poland

<sup>6</sup>Bioinformatic Plateau I2MC, INSERM and University of Toulouse, Toulouse 31062, France

<sup>7</sup>Senior author

<sup>8</sup>Lead Contact

\*Correspondence: thomas.clouaire@univ-tlse3.fr (T.C.), gaelle.legube@univ-tlse3.fr (G.L.)

<https://doi.org/10.1016/j.molcel.2018.08.020>

## SUMMARY

Double-strand breaks (DSBs) are extremely detrimental DNA lesions that can lead to cancer-driving mutations and translocations. Non-homologous end joining (NHEJ) and homologous recombination (HR) represent the two main repair pathways operating in the context of chromatin to ensure genome stability. Despite extensive efforts, our knowledge of DSB-induced chromatin still remains fragmented. Here, we describe the distribution of 20 chromatin features at multiple DSBs spread throughout the human genome using ChIP-seq. We provide the most comprehensive picture of the chromatin landscape set up at DSBs and identify NHEJ- and HR-specific chromatin events. This study revealed the existence of a DSB-induced monoubiquitination-to-acetylation switch on histone H2B lysine 120, likely mediated by the SAGA complex, as well as higher-order signaling at HR-repaired DSBs whereby histone H1 is evicted while ubiquitin and 53BP1 accumulate over the entire γH2AX domains.

## INTRODUCTION

DNA double-strand breaks (DSBs) are extremely detrimental since they can lead to mutations and chromosomal rearrangements. DSBs arise from various environmental stresses and upon developmentally scheduled activation of endonucleases, but they can also arise physiologically during replication and transcription. Anomalies in the DSB repair apparatus are responsible for premature aging and neurodegenerative syn-

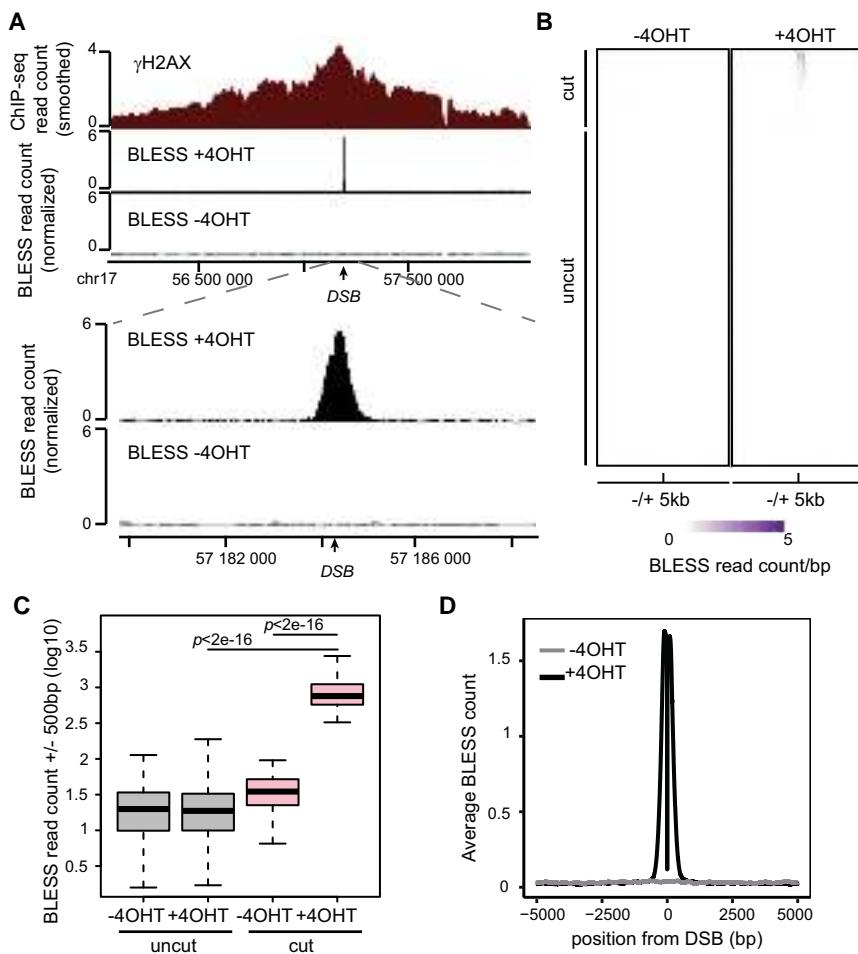
dromes and are strongly implicated in cancer onset and progression.

DSBs are mainly repaired by two partially redundant, yet profoundly different, pathways: homologous recombination (HR) and non-homologous end joining (NHEJ) (for review, see [Mladenov et al., 2016](#)). HR uses an intact copy of the damaged locus as a template and involves many factors for DSB detection, 5' end resection, homology search, strand invasion, and resolution. In contrast, NHEJ repair machineries require no or limited resection and can join the two broken ends with no or minimal homology. Inaccuracy, failure, or misuse of each of these pathways can trigger very different consequences on the genome. DSB repair pathway choice can be influenced by cell cycle phase ([Husted and Durocher, 2016](#)), DNA end complexity ([Schipler and Iliakis, 2013](#)), and the type of damaged locus ([Clouaire and Legube, 2015; Engel et al., 2018](#)).

In eukaryotes, DSB repair occurs in the context of chromatin. Chromatin is a highly dynamic structure, affected by histone post-translational modifications, DNA methylation, or incorporation of histone variants (for review, see [Soshnev et al., 2016; Talbert and Henikoff, 2017](#)). Chromatin modifications can alter the stability of the histone octamer onto DNA but also be specifically recognized by “reader” modules found in chromosomal proteins as well as subunits of DNA transaction machineries. All together, nucleosome modifications regulate DNA accessibility; the stiffness, flexibility, and mobility of chromatin within the nucleus; and the recruitment of molecular machines ensuring transcription, replication, and repair.

Key aspects in the interplay between DSB repair and chromatin environment have already emerged ([Table S2](#)). H2AX is rapidly phosphorylated by ATM (and named γH2AX) over several megabases surrounding the break ([Caron et al., 2015; Iacoboni et al., 2010; Rogakou et al., 1998; Savic et al., 2009](#)). In parallel, histone acetyltransferases and deacetylases tightly control acetylation levels of several residues of H3, H4, and H2A ([Dobbin et al.,](#)





**Figure 1. BLESS Mapping of AsiSI-Induced DSBs**

(A) Genome browser screenshot representing γH2AX ChIP-seq and BLESS signal at a single DSB located on human chromosome 17.

(B) Heatmap representation of the BLESS profile for the 1,211 predicted AsiSI sites in the human genome in DlVA cells untreated (-4OHT) or treated for 4 hr (+4OHT). DSBs are ordered based on decreasing read count in a 1 kb window.

(C) Boxplot representing BLESS signals (1 kb window) for the 80 cut (pink) and the 1,131 uncut (gray) AsiSI sites in untreated (-4OHT) or treated (+4OHT) DlVA cells. p values were calculated using two-sample Wilcoxon tests.

(D) Average BLESS count profile for the 80 induced DSBs before (-4OHT, gray line) and after induction (+4OHT, black line).

Here, using high-throughput genomic approaches in a standardized system in which multiple DSBs are induced at defined positions across the human genome, we report the first comprehensive picture of the chromatin landscape induced around DSBs and its relationship with individual repair pathways.

## RESULTS

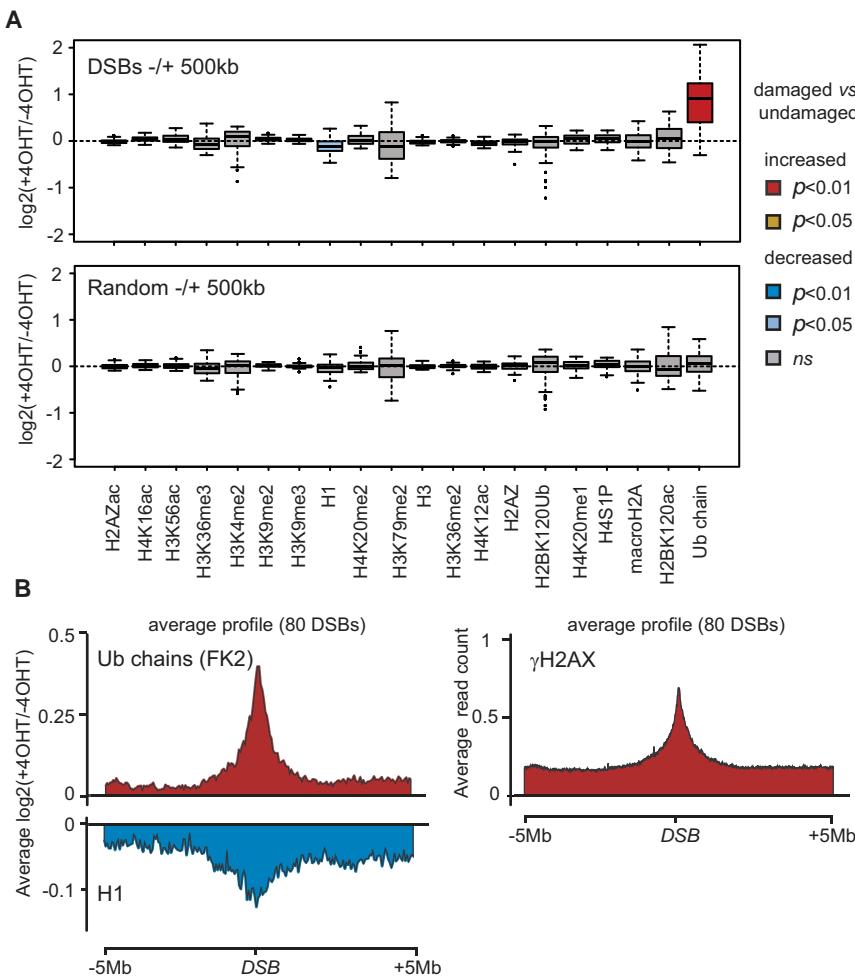
### High-Resolution Mapping of AsiSI-Induced DSBs

To characterize the chromatin landscape at DSBs, we used the DlVA (DSB inducible via AsiSI) cell line, which enables the controlled induction

(using 4-hydroxytamoxifen, or 4OHT) of multiple well-annotated DSBs scattered throughout the human genome (Aymard et al., 2014; Iacovoni et al., 2010). We previously showed that *in vivo*, AsiSI does not produce a DSB at every target site on the genome and that γH2AX accumulation is strongly affected by genomic and epigenomic features (Caron et al., 2012; Iacovoni et al., 2010). To unambiguously identify the position of AsiSI-induced DSBs and obtain a quantitative measurement of the DSB generation rate, we used the recently developed BLESS method (Crosetto et al., 2013; Mitra et al., 2015). We detected sharp BLESS peaks located precisely at predicted AsiSI sites following a 4 hr treatment of DlVA cells with 4OHT and minimal signal in untreated conditions (Figures 1A and S1A). Further analysis revealed that out of the 1,211 predicted AsiSI sites in the human genome, 174 showed a signal significantly higher than background (Figures S1B and 1B). This is in relatively good agreement with our previous estimates of the number of induced DSBs inferred from γH2AX accumulation and with BLISS data in DlVA cells (Aymard et al., 2014; Iannelli et al., 2017). However, we could detect clear

2013; Gong et al., 2015; Jacquet et al., 2016; Lee et al., 2010; Miller et al., 2010; Ogiwara et al., 2011; Toiber et al., 2013) to regulate chromatin relaxation near DSBs. Acetylated histones also participate in the recruitment of nucleosome remodeling factors, enhancing DSB accessibility and facilitating resection (Bennett and Peterson, 2015; Lee et al., 2010; Toiber et al., 2013). Similarly, histone methyltransferases and demethylases can regulate recruitment and/or stabilization of repair proteins (Ctp1, 53BP1, BRCA1 ...) (Table S2). Moreover, ubiquitination and sumoylation pathways contribute to DSB-induced chromatin reorganization (for review, see Schwertman et al., 2016). All together, these specific chromatin modifications generate a chromatin state permissive for repair but also directly contribute to the recruitment of DSB repair machineries, repair pathway choice, and the activation of the DNA damage checkpoint.

Yet the definite map of DSB-induced chromatin modifications and their respective involvement in DSB repair remain largely unknown. Furthermore, NHEJ and HR repair pathways conceivably require very different chromatin settings. Since chromatin structure plays a central role in DNA accessibility and flexibility, an in-depth characterization of the chromatin that assembles at DSBs represents a critical step in understanding how DSB repair machineries operate in the whole nucleus to restore the original DNA sequence and avoid deleterious genome rearrangements.



**Figure 2. Large-Scale DSB-Induced Chromatin Changes**

(A) Boxplot representing the ChIP-seq enrichment ratio between 4OHT-treated and untreated DlvA cells (expressed as a  $\log_2$  ratio) for 80 DSBs (upper panel) or 80 randomly picked genomic regions (lower panel) over a 1 Mb window. Boxes are colored according to p values (two-sample Wilcoxon tests) and the nature of the change.

(B) Left: Average profile of the enrichment between 4OHT-treated and untreated DlvA cells for ubiquitin and H1 over 80 DSBs in a 10 Mb window. Values are expressed as  $\log_2$  ratios. Right: Average profile of  $\gamma$ H2AX in 4OHT-treated DlvA cells.

differences between  $\gamma$ H2AX and BLESS signals, confirming that  $\gamma$ H2AX does not strictly reflect DSB induction rate (Figure S1A, compare left and right panel; Figure S1C). To remove any site that may only be partially cleaved, we focused on a robust set of 80 DSBs defined by BLESS analysis that are significantly induced by 4OHT treatment (Figures 1B–1D, Table S1). These DSBs are characterized by low levels in heterochromatin-associated features such as DNA methylation and H3K9me3 (Figure S1D), and about half are located near active promoters (Figure S1E). Yet analyzing ChIP-seq enrichment for the elongating form of RNA polymerase II (S2P) revealed that not all sites necessarily reside within actively transcribed regions (Figure S1D). We retained this validated set of DSBs, representing both active and inactive euchromatic regions, for further analysis.

#### Mapping of Histone Modifications at AsiSI-Induced DSBs

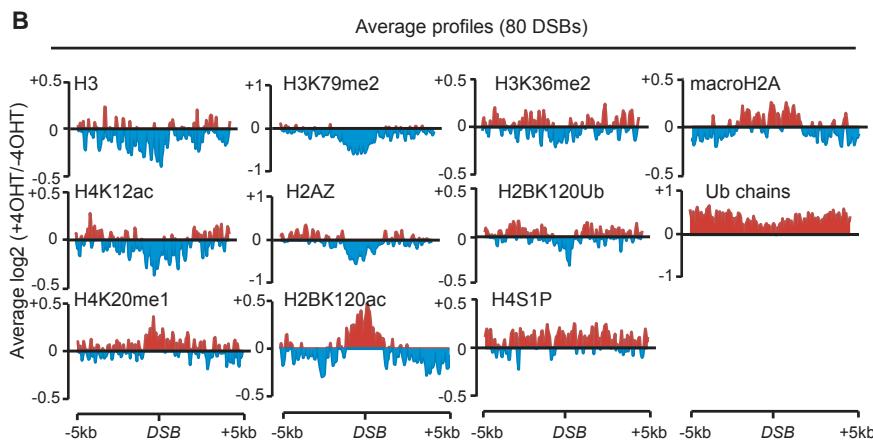
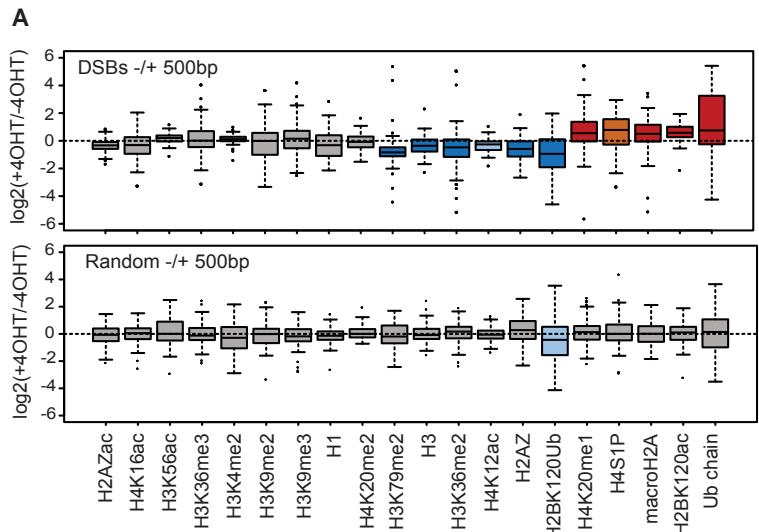
To obtain a comprehensive picture of the chromatin landscape at DSBs, we generated 20 ChIP-seq profiles, in both damaged (4OHT-treated) and undamaged cells. These included linker and core histones, histone variants, and post-translational modifications that were proposed to play a role in DSB repair

(Table S2). We also generated a genome-wide map for chromatin-associated ubiquitin using the FK2 antibody that recognizes a wide variety of ubiquitin conjugates. ChIP-qPCR revealed that each antibody efficiently immunoprecipitated chromatin (Figure S2A). More importantly, we could recapitulate previously reported profiles for each feature over transcription units, categorized by transcriptional activity (Figure S2B). Our H4S1P mapping, which was lacking in mammalian cells, is in agreement with its proposed function in transcriptional regulation and accumulation over gene bodies in yeast (Utley et al., 2005).

Given that  $\gamma$ H2AX accumulation can spread up to two megabases (Mb) from the actual breakpoint (Iacoboni et al., 2010), we first examined DSB-induced

chromatin modifications over a 1 Mb window. At this scale, ubiquitination is largely induced ( $p < 0.01$ , two-sample Wilcoxon test) and histone H1 occupancy significantly reduced ( $p < 0.05$ , two-sample Wilcoxon test) compared to randomly selected genomic regions (Figure 2A). We could not detect significant DSB-induced changes for any other histone modifications within these megabase-sized genomic windows surrounding breaks. Average profiles revealed that ubiquitin conjugates accumulated over approximately 2 Mb, similarly to  $\gamma$ H2AX spreading (Figures 2B and S2C). DSB-induced H1 depletion is maximal over a similar region but remains detectable over a slightly more extended area (Figure 2B, bottom panel, and Figure S2C). All together, this shows that DSBs induce few but large-scale (megabase-sized) chromatin reshufflings, such as phosphorylation of H2AX, accumulation of ubiquitin, and depletion of the linker histone H1.

We next considered that DSB could cause chromatin modifications much closer to the break and examined 1 kb windows surrounding AsiSI-induced DSBs. Interestingly, we found that among the 20 chromatin modifications analyzed, 6 were significantly decreased (H3K79me2, H3, H3K36me2, H4K12ac, H2AZ, and H2BK120ub) and 5 significantly increased (H4S1P,



H4K20me1, macroH2A, H2BK120ac, and ubiquitin) near DSBs (Figure 3A, top panel) when compared to random sites (Figure 3A, bottom panel). Average profiles showed that these proximal DSB-induced chromatin alterations spread over different scales, ranging from 1 to 10 kb (Figure 3B). Decreases in H3, H2AZ, H4K12ac, H2BK120ub, and H3K79me2 and the increases in H4K20me1 and H2BK120ac took place over 2–4 kb windows. MacroH2A and H4S1P showed a DSB induction spanning 5–6 kb and approximately 10 kb, respectively. In most cases, normalizing to histone H3 did not alter the observed changes, with the exception of H4K12ac and H3K36me2, which may thus only reflect changes in nucleosome occupancy (Figure S2D). All together, our systematic ChIP-seq mapping indicates that DSBs trigger many alterations to chromatin structure occurring over many different scales, ranging from 1 kb up to several megabases.

More specifically, our data suggest that in response to DSBs, H2BK120 undergoes a switch from ubiquitination to acetylation (Figure 3A). Time-course experiments confirmed the progressive loss of H2BK120ub upon DSB induction (Figure S3A). Interestingly, SAGA, a prominent chromatin-modifying complex, acetylates histones H3 and H2B and also displays

**Figure 3. Narrow-Scale DSB-Induced Chromatin Changes**

(A) Boxplot representing the ChIP-seq enrichment ratio between treated and untreated DlVA cells (expressed as a  $\log_2$  ratio) for 80 DSBs (upper panel) or 80 randomly picked genomic regions (lower panel) over a 1 kb window. Boxes are colored according to p values (two-sample Wilcoxon tests) and the nature of the change.

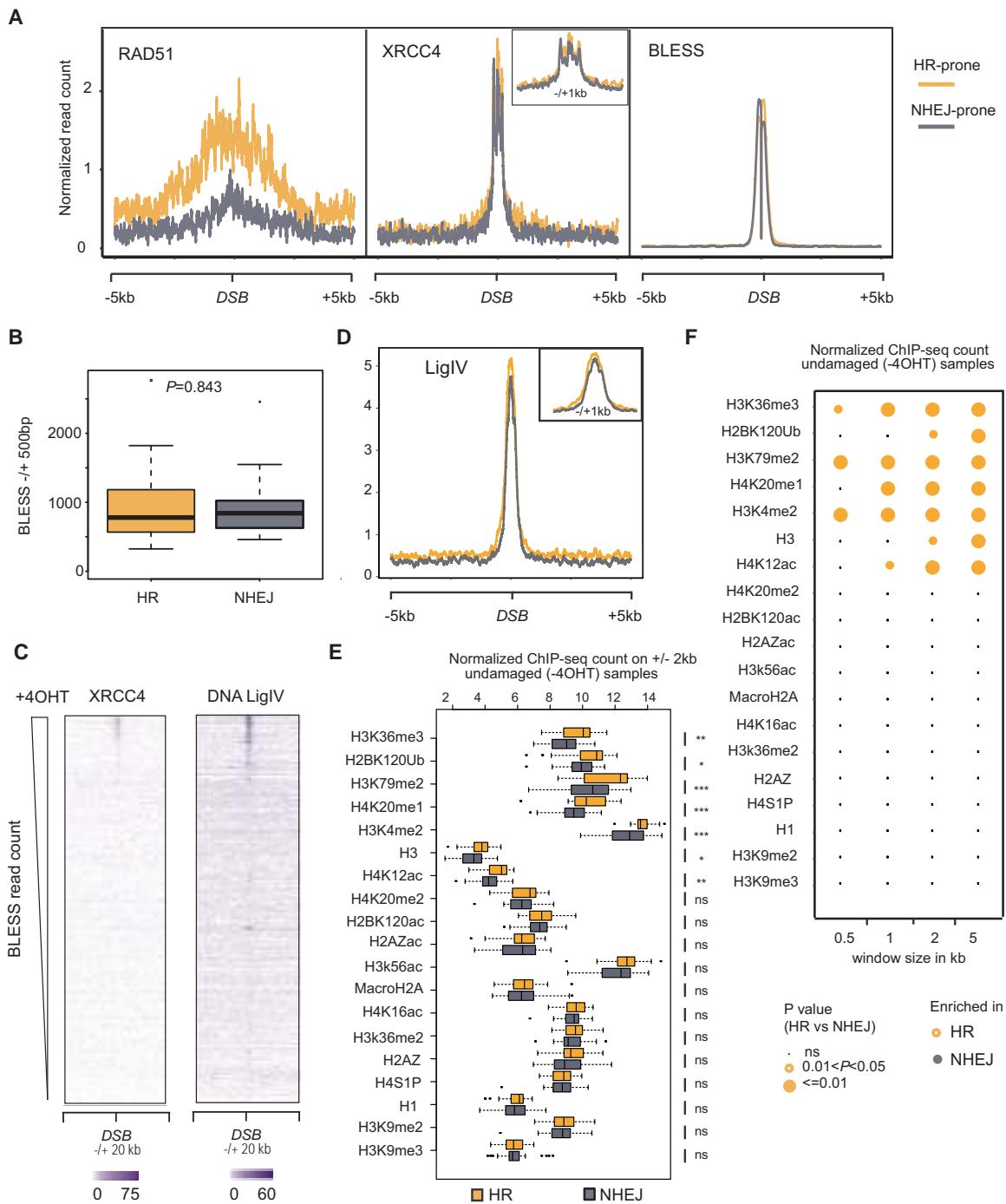
(B) Average profile of the enrichment between treated and untreated DlVA cells for 12 modifications, showing significant differences between each condition over 80 DSBs in a 1 kb window. Values are expressed as  $\log_2$  ratios.

deubiquitinase activity toward H2BK120 (Helmlinger and Tora, 2017). SAGA has also been previously implicated in class switch recombination and response to ionizing radiation (Ramachandran et al., 2016). We show that *in vitro*, the native human SAGA complex (affinity-purified through its specific SUPT7L subunit; Figure S3B) displays both H2BK120 deubiquitinase and acetyltransferase activities, making it a primary candidate for this switch (Figure S3C). Furthermore, depletion of SUPT7L as well as PCAF, one HAT paralog present in the SAGA complex (Figures S3D and S3E), triggered decreases in both HR at an endogenous locus (*LMNA*) following CRISPR/Cas9 breakage (Pauty et al., 2017) and NHEJ in a cell reporter system (Jacquet et al., 2016) (Figures S3F

and S3G), indicating that SAGA indeed contributes to DSB repair in human cells.

#### Damage in Active Chromatin Undergoes Preferential Repair via HR

It is now well established that the local chromatin structure can influence how a given DSB is handled and subsequently repaired (for review, see Clouaire and Legube, 2015). We thus set out to understand the contribution of chromatin to repair pathway choice by defining chromatin states favorable to HR and NHEJ. We defined subsets of BLESS-validated DSBs that are preferentially repaired by HR or NHEJ (30 in each category) by sorting them according to the binding ratio for RAD51 and XRCC4 using ChIP-seq data obtained in 4OHT-treated DlVA cells (Aymard et al., 2014). Average profiles confirmed the prominent accumulation and spreading of RAD51 over 5 kb at DSBs preferentially repaired by HR (Figure 4A). On the other hand, XRCC4 recruitment appeared similar for both sets of DSBs, with a multimodal distribution over 500 bp (Figures 4A and S4A), likely related to XRCC4/XLF filaments at DSBs (Ropars et al., 2011). Importantly, BLESS intensity was comparable between HR and NHEJ DSBs (Figures 4A and 4B), suggesting



**Figure 4. Chromatin Features Associated with HR-Prone DSBs**

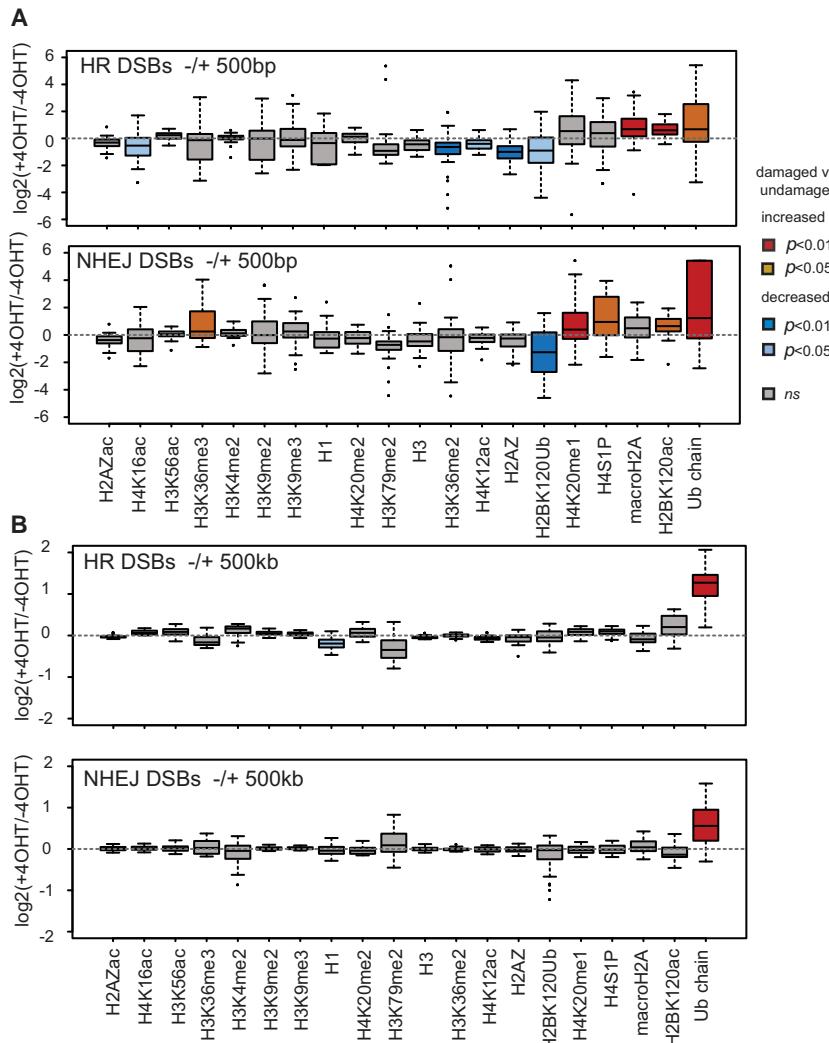
(A) Average profile for RAD51 ChIP-seq, XRCC4 ChIP-seq, and BLESS for 30 HR (yellow) or 30 NHEJ (gray) DSBs. HR and NHEJ DSBs were defined based upon the RAD51/XRCC4 binding ratio (see [STAR Methods](#)).

(B) Boxplot representing BLESS read count (1 kb window) around 30 HR or 30 NHEJ DSBs. p value was calculated using two-sample Wilcoxon test.

(C) Heatmap representing the XRCC4 and DNA Ligase IV ChIP-seq signals on a 40 kb window centered around all AsiSI sites, ordered based on the BLESS level.  
(D) Same as (A) for DNA Ligase IV ChIP-seq.

(E) Boxplot representing the ChIP-seq read count (4 kb window) for each histone modification in untreated cells for 30 HR (yellow) and 30 NHEJ (gray) DSBs. p values were calculated using two-sample Wilcoxon test. \* $p < 0.05$ , \*\* $p < 0.01$ ;  $p > 0.05$  is not significant (ns).

(F) Circle plot representing p values (from two-sample Wilcoxon test) when comparing ChIP-seq signal for HR and NHEJ DSB using increasing window size.



**Figure 5. HR and NHEJ-Induced Chromatin Changes**

(A) Boxplot representing the ChIP-seq enrichment ratio between treated and untreated DlvA cells (expressed as a  $\log_2$  ratio) for 30 HR (upper panel) or 30 NHEJ (lower panel) over a 1 kb window. Boxes are colored according to p values (two-sample Wilcoxon tests) and the nature of the change.

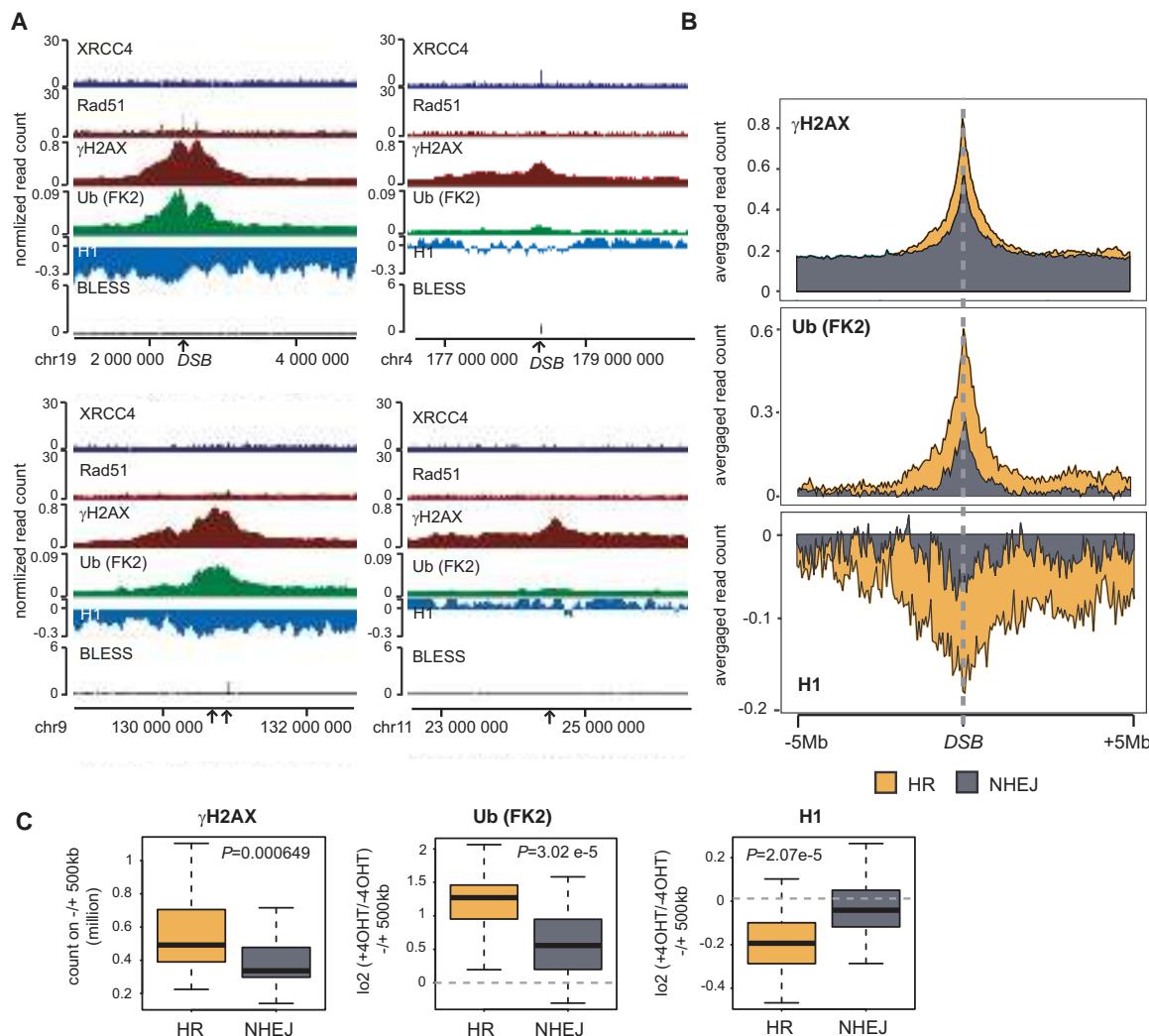
(B) Same as (A) for a 1 Mb window.

repaired by HR or NHEJ. In agreement with our previous finding (Aymard et al., 2014), H3K36me3 was significantly more abundant proximal to DSBs repaired by HR (Figure 4E). HR-competent chromatin also contained elevated levels of H3K79me2, H4K20me1, H2BK120ub, H3K4me2, H4K12ac, and core histone H3 (Figure 4E). This result was confirmed when comparing chromatin signatures over increasing window sizes (Figure 4F). Consistently, we found a similar trend when using DNA ligase IV instead of XRCC4 to determine HR and NHEJ categories (Figure S5D). On the other hand, we found no evidence of a specific signature that may actively favor repair by NHEJ (Figures 4E, 4F, and S5D). All together, these data identified a specific chromatin structure that is competent for HR repair. The HR-specific signature primarily consists of features linked to active transcription (Figure S2B). In agreement, HR-prone DSBs were found to be significantly enriched in the nuclear compartment A1, which was defined by its specific long-range interaction pattern by Hi-C analysis and shown to be enriched in transcriptionally active regions (Rao et al., 2014) (Figure S4B). Hence, our data strongly favor the hypothesis that DSBs induced in active genes are biased toward HR.

that differences in RAD51 binding do not solely depend on the DSB induction rate. Since NHEJ is a relatively fast process compared to HR, it remained possible that some sites could be cleaved, repaired, and mutated with a very fast kinetics, therefore not appearing in our set of NHEJ-prone DSBs. To address this, we performed XRCC4 ChIP-seq at various times of 4OHT treatment (1 hr, 4 hr, and 24 hr) and found that XRCC4 binding was indeed very similar at DSB sites at each time point (Figures S4C–S4E). More importantly, we could not detect any site showing XRCC4 enrichment at 1 hr or 24 hr and not at 4 hr, which strongly validated our HR and NHEJ categories (Figures S4C–S4E). To confirm that XRCC4 binding was indeed indicative of NHEJ, we also performed ChIP-seq using an antibody directed against DNA ligase IV, another key component in this pathway. As expected, DNA ligase IV showed a very similar enrichment when compared to XRCC4 at all AsiSI sites (Figures 4C, 4D, and S5A–S5C). All together, these analyses allowed us to define a robust set of DSBs either prone to engage HR (HR-prone) or not (NHEJ-prone).

Next, we compared basal enrichment (before DSB induction) for each modification at the vicinity of DSBs (2 kb) preferentially

DSB repair by HR and NHEJ depends on two separate mechanisms and operates in distinct chromatin contexts (see above). Hence, it is possible that each pathway involves different remodeling events compatible with its particular repair mechanism. We therefore considered the possibility that DSB-induced alterations in chromatin could be specific for HR- and NHEJ-prone DSBs. Notably, focusing on events occurring proximal to the break point (1 kb windows), we found that while some chromatin marks significantly changed after break induction only at DSBs repaired by HR or by NHEJ, others occurred irrespective of the repair mechanism (Figure 5A). We also examined DSB-induced chromatin changes at HR- and NHEJ-prone sites at various distances from the DSBs (Figure S5E). MacroH2A deposition and H4S1 phosphorylation occurred upon repair by both HR and NHEJ (Figures 5A and S5E). Similarly, the switch from



**Figure 6. Acute Signaling at HR-Prone DSBs**

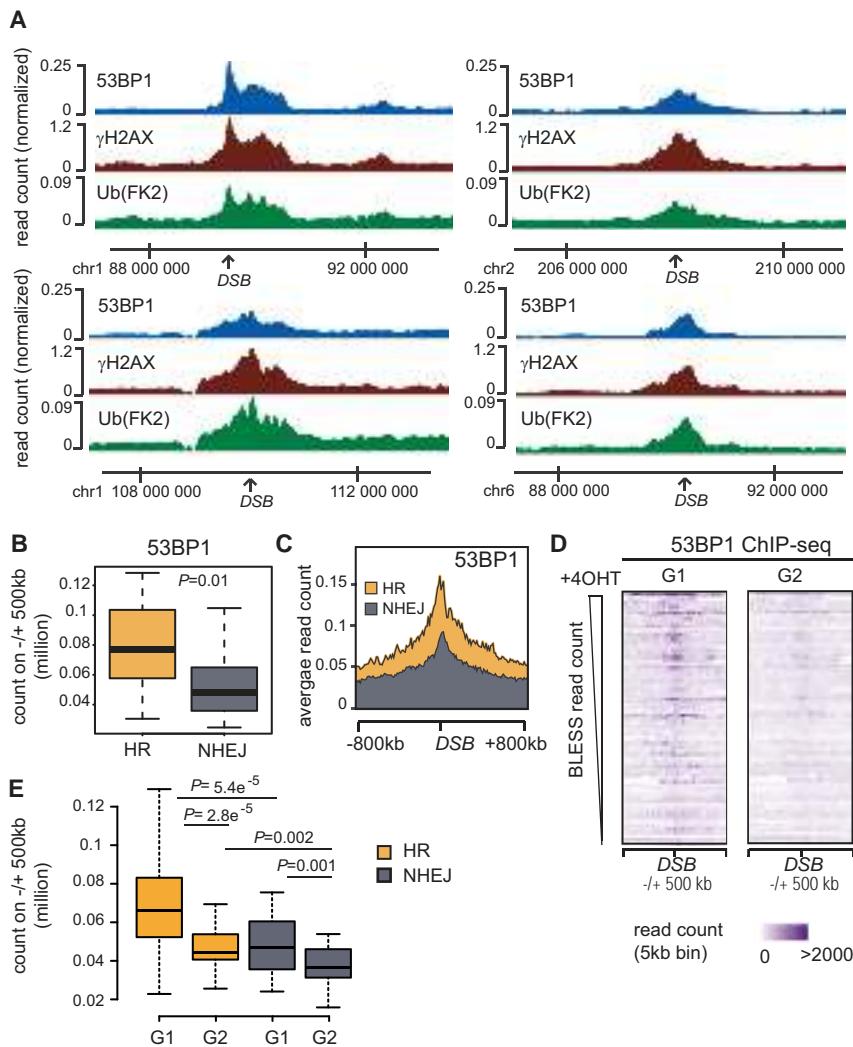
(A) Genome Browser screenshots representing ChIP-seq signals for XRCC4, RAD51,  $\gamma$ H2AX, ubiquitin, H1, and BLESS for two HR-DSBs (left) and two NHEJ DSBs (right). Data are expressed as read count (from 4OHT-treated samples) for XRCC4, RAD51, and  $\gamma$ H2AX ChIP-seq and BLESS. Data for ubiquitin (FK2) and H1 are expressed as a  $\log_2$  ratio between treated and untreated cells.

(B) Average profile for  $\gamma$ H2AX ChIP-seq (read count in OHT treated samples), ubiquitin, and H1 (as a  $\log_2$  ratio between treated and untreated cells) for 30 HR (yellow) and 30 NHEJ (gray) DSBs in a 10 Mb window.

(C) Boxplot representing ChIP-seq read count from treated cells ( $\gamma$ H2AX) or  $\log_2$  ratio between treated and untreated cells (ubiquitin [FK2] and H1) for 30 HR (yellow) or 30 NHEJ (gray) DSBs in a 1 Mb window. p values were calculated using two-sample Wilcoxon tests.

H2BK120 ubiquitination to acetylation also occurred at all DSBs (Figures 5A and S5E), in agreement with our above finding that depletion of the SAGA complex affects both HR and NHEJ (Figures S3F and S3G). In contrast, reduction in H2AZ and H3 occupancy and decrease in H3K36/79 dimethylation and in H4K12/K16 acetylation following DSB induction were significant only at sites prone to HR (Figures 5A and S5E). Conversely, DSB induction at sites repaired by NHEJ was accompanied by a significant increase in H3K36me3 and H4K20me1, modifications not found at HR-repaired DSBs (Figures 5A and S5E). Hence, we were able to define a local histone modification signature, which occurs proximal to DSBs, associated with the two main repair pathways.

We also interrogated our data for pathway-specific events occurring within 1 Mb from the break. Linker histone H1 depletion and ubiquitin accumulation were significantly more pronounced within 1 Mb of HR-repaired sites compared to NHEJ-prone DSBs (Figure 5B). Examination of individual (Figure 6A) or averaged (Figure 6B) profiles confirmed that DSB-induced changes in H1 and ubiquitin were accentuated at HR sites, with H1 depletion being barely detectable at NHEJ sites. We also detected increased megabase-wide  $\gamma$ H2AX levels for HR sites compared to NHEJ sites (Figures 6A and 6B). Accumulation of ubiquitin conjugate, H2AX phosphorylation, and H1 depletion were indeed significantly reinforced surrounding DSBs repaired by HR compared to those repaired by NHEJ (Figure 6C). This finding suggests that damage



**Figure 7. 53BP1 Preferentially Accumulates at HR-Prone DSBs Specifically in G1**

(A) Genome Browser screenshots representing ChIP-seq signals (from 4OHT-treated samples) for 53BP1, γH2AX, and ubiquitin.

(B) Boxplot representing 53BP1 enrichment in a 1 Mb window for 30 HR (yellow) and 30 NHEJ (gray) DSBs. p value was calculated using two-sample Wilcoxon test.

(C) Average profile for 53BP1 ChIP-seq (read count in 4OHT-treated cells) for 30 HR (yellow) and 30 NHEJ (gray) DSBs in a 1.6 Mb window (upper panel).

(D) Heatmap representing 53BP1 ChIP-seq count obtained in G1- and G2-synchronized cells as indicated, on a 1 Mb window surrounding 80 DSBs induced by AsiSI. DSBs are sorted by BLESS read counts.

(E) Boxplot representing 53BP1 enrichment in a 1 Mb window for 30 HR (yellow) and 30 NHEJ (gray) DSBs in G1- and G2-synchronized cells as indicated. p values were calculated using two-sample Wilcoxon test.

occurring in active chromatin and repaired by HR can trigger acute and broad-scale remodeling within high-order chromatin structure that is likely to drastically alter its properties.

#### Acute High-Order Chromatin Modifications Correlate with Enhanced 53BP1 Recruitment *In Vivo*

An obvious role for chromatin-mediated DSB signaling would be to directly attract effectors of the DDR. 53BP1 plays key roles in DSB repair by inhibiting end resection, favoring distal end synapsis, and promoting damaged chromatin mobility (for review, see Panier and Boulton, 2014). 53BP1 interaction with damaged chromatin involves multivalent engagement of several histone modifications such as H2A (or H2AX) lysine 15 ubiquitination and methylation of lysine 20 on histone H4 (Fradet-Turcotte et al., 2013; Wilson et al., 2016). Furthermore, stable 53BP1 association with DSBs may involve a direct interaction with γH2AX (Kleiner et al., 2015; Ward et al., 2003) and can be regulated by acetylation on H4 and H2A (Jacquet et al., 2016; Tang et al., 2013). We generated a genome-wide ChIP-seq map of 53BP1 binding to damaged chromatin *in vivo*. Inspection of 53BP1 distribution around individual DSBs (Figure 7A) or averaged over

the 80 DSBs (Figure S6A) revealed a striking ability for 53BP1 to spread within megabase-wide domains. Furthermore, 53BP1 binding profiles appear almost indistinguishable from those of γH2AX and ubiquitin (Figures 7A and S6A), and we observed a very strong correlation between the accumulation of γH2AX, ubiquitin, and 53BP1 within 1 Mb domains surrounding DSBs (Figure S6B). We also observed that 53BP1 is significantly more enriched within a megabase from sites preferentially repaired by HR compared with sites repaired by NHEJ (Figures 7B, 7C, and S6C). Thus, our data suggest that DSBs occurring in active chromatin, which preferentially undergo HR repair, can trigger an acute high-order chromatin signaling that may favor subsequent recruitment of 53BP1 over megabase-sized regions encompassing the break site.

*In vivo*, 53BP1 was shown to inhibit 5' end resection in various contexts, including class switch recombination and dysfunctional telomeres fusion (for review, see Panier and Boulton, 2014). Yet our results suggest that 53BP1 binding is stronger at sites that are more prone to be repaired by HR, a process that requires end resection. To clarify this apparent discrepancy, we generated 53BP1 ChIP-seq data in damaged G1- and G2-synchronized cells. Overall, we found that 53BP1 accumulation was much stronger during G1 compared to G2, (Figures 7D and S6D–S6F). Because γH2AX appeared rather similar between both phases of the cell cycle (Figure S6F), this suggests that the striking difference in 53BP1 spreading over megabase-wide domains observed when comparing G1 and G2 cells is unlikely to be due to differences in DSB induction rates. Finally, we confirmed that 53BP1 accumulation is more prominent at DSBs repaired by HR in G1 compared to G2, while signal for NHEJ DSB remained consistently lower regardless of the cell

cycle phase (Figure 7E). Thus, our data revealed that 53BP1 binding is favored at HR-prone DSBs, but mostly in G1, when HR usage is restricted. This suggests that 53BP1 exerts its anti-resection function in G1 only at sites prone to undergo HR, while it may be less critical at DSBs usually repaired by NHEJ.

## DISCUSSION

### A Standardized Approach to Investigate DSB-Induced Chromatin Changes

Here, we provide the most comprehensive view to date of the chromatin landscape assembled at DSBs in human cells. By analyzing 20 chromatin features that were previously shown to be involved in DSB repair, we were able to substantiate several earlier findings such as H3, H2AZ, and H1 removal, macroH2A incorporation, H4S1 phosphorylation, H4K16 deacetylation, H4K20 monomethylation, and the accumulation of ubiquitin conjugates at sites of damage (Table S2). Interestingly, DSB-induced phosphorylation of histone H4 has only been previously reported in budding yeast and was proposed to promote NHEJ (Cheung et al., 2005; Utley et al., 2005). Break-induced H4 phosphorylation is therefore conserved in higher eukaryotes and spreads over large regions spanning up to 10 kb from the DSB, independently of the repair pathway. This ability to propagate up to 50 nucleosomes away from the break suggests that H4 phosphorylation functions in DDR signaling, potentially as a recruitment platform for DDR proteins.

Here we failed to detect any change reported by others (Table S2) in H3K9me2, H3K9me3, H4K20me2, or H3K56ac. Conversely, we observed a clear loss of H3K79me2 and H2BK120ub, two histone modifications that were found, respectively, unchanged and increased at sites of breaks (Huyen et al., 2004; Moyal et al., 2011; Nakamura et al., 2011). Such discrepancies could arise from the use of different DSB induction methods that can produce chemically different DNA ends, concomitantly trigger other forms of DNA damage, or create DSBs at a given cell cycle stage. More importantly, DSB-inducing agents do not necessarily damage the same regions across the genome, which will strongly influence the chromatin outcome. Indeed, X- and  $\gamma$ -rays produce DSBs randomly throughout the genome, mainly in untranscribed and potentially heterochromatic regions, while AsiSI mostly introduces DSBs within or near genes, either active or inactive, leaving heterochromatin undamaged. This could explain why neocarzinostatin and ionizing radiation, unlike AsiSI-induced DSBs, trigger an increase in H2BK120ub levels (Moyal et al., 2011; Nakamura et al., 2011). Kinetic aspects should also be considered. We performed our ChIP 4 hr after AsiSI activation, a time point representing an “equilibrium” in which events ranging from early cleavage to late repair cohabit within the cell population. However, this may preclude the identification of very transient and/or very late events. Finally, in order to identify the most prominent DSB-induced chromatin features, we purposely limited our large-scale analysis to an averaged snapshot on a large population of asynchronous cells, which may hinder the identification of events occurring at specific cell cycle stages. Yet despite a few discrepancies with the published literature, our study allowed us to

highlight major chromatin changes during DSB repair and to associate these changes with a preferred repair pathway. Additional studies in repair-deficient cells should now be conducted in order to establish whether these histone modification changes do indeed depend on HR or NHEJ repair or only on the initial chromatin status. Moreover, systematic mapping of chromatin-modifying enzymes following DNA damage will be required to better understand the establishment of this multiscale chromatin landscape. Investigating if, when, and how this DSB-induced chromatin landscape is reverted following repair in order to maintain epigenome stability also represent exciting follow-up studies.

### Histone Crosstalks Involved in Transcription Regulation Are Mobilized in the DSB Response

Importantly, our data fully recapitulate several crosstalks previously identified in the context of chromatin organization and transcriptional regulation. Indeed, our study reveals that upon DSB induction, macroH2A is incorporated (Khurana et al., 2014), and H2BK120 undergoes a switch from ubiquitination to acetylation. MacroH2A can regulate gene expression by stimulating CBP-dependent H2BK120 acetylation (Chen et al., 2014), which precludes H2BK120 ubiquitination. Our findings suggest that this macroH2A-H2BK120ac crosstalk goes beyond transcriptional regulation and may also function following DNA damage and repair. Our work revealed that the transition from ubiquitination to acetylation on H2BK120 may be mediated by the SAGA complex, a well-known complex involved in transcriptional regulation, and that depletion of DUB and HAT subunits of SAGA leads to defective HR and NHEJ, in agreement with a previous study (Ramachandran et al., 2016). Notably, these macroH2A-H2BK120 changes occur independently of the DSB considered (i.e., prone to be repaired by HR or NHEJ), suggesting that they take place before repair reaction is engaged. Given the function of H2BK120ub in nucleosome stabilization (Batta et al., 2011; Fleming et al., 2008), deubiquitinating H2B may loosen nucleosome stability onto DNA, thereby favoring subsequent remodeling and downstream repair events.

We also identified seven additional chromatin changes after DSB induction, including decreases in H3K79me2, H2AZ, and H4K12/K16ac, which specifically occurred at DSBs prone to HR (Figure S7). We also recently demonstrated that H3K4me3 decreases at AsiSI-induced DSBs (Gong et al., 2017). Notably, all these modifications were already found to be coordinated and linked to H2B ubiquitination during transcriptional regulation. For example, H2BK120ub is required for H3K79 methylation (McGinty et al., 2008; Sun and Allis, 2002) and counteracts INO80-dependent H2AZ removal (Segala et al., 2016). Finally, H2B ubiquitination stimulates H3K4 trimethylation (Sun and Allis, 2002; Zhu et al., 2005). Our data indicate that this well-established crosstalk likely also functions at DSBs and that the switch from H2BK120 ubiquitination to acetylation at DSBs may subsequently induce H3K79 and H3K4me3 demethylation, H4 deacetylation, and H2AZ removal by INO80, which itself associates with DSB (Downs et al., 2004; Morrison et al., 2004; van Attikum et al., 2004). Since H2AZ occupancy was found to regulate end resection (Alatwi and Downs, 2015; Gursoy-Yuzugullu et al., 2015; Xu et al., 2012), it is possible that these coordinated

modifications of H2BK120ub-H3K79me2-H2AZ-H4 at HR-prone DSBs contribute to the establishment of a chromatin state competent for resection and/or Rad51 nucleofilament assembly.

### Megabase-Scale Chromatin Signaling: A Central Unit in the DSB Response

Chromatin flanking DSBs also undergoes extensive large-scale remodeling, such as  $\gamma$ H2AX spreading over megabase-wide domains (Iacovoni et al., 2010; Rogakou et al., 1998; Savic et al., 2009). Here we show that this large-scale remodeling is also accompanied by the accumulation of ubiquitin conjugates, 53BP1 accrual, and the removal of histone H1 from the entire  $\gamma$ H2AX domain. 53BP1 is known to be recruited to damaged chromatin by engaging multiple interactions with H2AK15ub and H4K20me2 but also with  $\gamma$ H2AX (for review, see Panier and Boulton, 2014). Our data support this proposed mode of recruitment given that: (1) 53BP1 strongly parallels the pattern observed with  $\gamma$ H2AX and the FK2 antibody (although the broad specificity of this antibody toward many ubiquitinated substrates precludes a definitive conclusion regarding the nature of the modified protein) and (2) 53BP1 association with DSB is minimal in G2, in agreement with the proposed dilution of H4K20me2 on post-replicative chromatin (Pellegrino et al., 2017). H1, on the other hand, was recently shown to be displaced from sites of DNA damage (Sellou et al., 2016; Strickfaden et al., 2016). Interestingly, RNF8-mediated H1 ubiquitination loosens its interaction with chromatin (Thorslund et al., 2015), providing potential mechanisms for H1 eviction from megabase-wide  $\gamma$ H2AX domains. These megabase-wide chromatin modules display well-defined boundaries that may correlate with topologically associated domains (TADs) (Caron et al., 2012; Marnef and Legube, 2017), a basic unit of chromosome folding in 3D. The establishment of these large chromatin changes might therefore be determined by the initial chromosome structure (for review, see Aymard and Legube, 2016; Marnef and Legube, 2017). Such profound changes in entire TADs could regulate the chromatin fiber physical properties such as the mobility of the damaged DNA within the nucleus. In agreement, linker histones strongly affect the global chromatin structure (Fan et al., 2005), and 53BP1 favors long-range motions of DNA ends (Dimitrova et al., 2008).

### The Complex Fate of DSBs in Active Transcription Units

DSB repair varies across the genome. Damage in heterochromatin is largely repaired by HR in G2 by a specific pathway requiring 53BP1-dependent chromatin relaxation, followed by its relocation at the periphery of IR-induced foci (Kakarougkas et al., 2013; Noon et al., 2010). In euchromatin, only transcriptionally active genes, when damaged, are prone to HR repair in G2 (Aymard et al., 2014; this study). Importantly, damaged active genes are refractory to repair in G1, where they instead persist and cluster together within foci (Aymard et al., 2017). Here we found that megabase-size chromatin modifications are more prominent at sites prone to be repaired by HR, e.g., DSBs occurring within active chromatin domains. Such an acute signaling could be indicative of a specific DDR pathway mobilized upon damage in active regions of the genome. Additionally, we found that these DSBs also display enhanced binding of 53BP1, specifically during G1. We did not observe specific 53BP1 accumu-

lation or redistribution at HR-prone DSBs during G2 in our study, suggesting that 53BP1 is likely dispensable to promote HR in the euchromatic, relaxed fraction of the genome. 53BP1 may interact with DSBs to inhibit extensive end processing in G1 and avoid the use of deleterious pathways such as alt-NHEJ during times when canonical HR is not available (Biehs et al., 2017).

In summary, we provide here the first comprehensive depiction of the set of histone modifications induced upon DSB associating with HR and NHEJ repair and report a thorough description of HR-competent chromatin, which is strongly linked with transcriptional activity. Active genes emerge as particularly fragile loci that frequently experience DNA double-strand breakage (for review, see Marnef et al., 2017) and represent translocation hotspots. Characterizing the local and large-scale chromatin structures that assemble at these specific damaged loci represents the first step to understanding how chromatin mobility may change following damage to promote DSB clustering and homology search and, in some instance, lead to translocations.

### STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- CONTACT FOR REAGENT AND RESOURCE SHARING
- EXPERIMENTAL MODEL AND SUBJECT DETAILS
  - Chemicals
  - Cell Culture and Cell Lines
- METHOD DETAILS
  - Cell treatments
  - Establishment of cell line expressing SUPT7L from the AAVS1 safe harbor and TAP Purification of native human SAGA HAT/DUB complex
  - *In vitro* histone acetylation assay
  - *In vitro* histone de-ubiquitination assay
  - RT-qPCR
  - Cas9/mClover-LMNA1 homologous recombination assay
  - NHEJ assay
  - Cell cycle analysis
  - ChIP
  - BLESS
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - ChIP-seq data processing
  - BLESS data processing
  - Descriptive statistics using box-plot representation
  - Statistical analysis and representations
  - Averaged ChIP-seq profiles
  - Generation of random positions
  - Determination of HR-prone and NHEJ-prone DSBs
- DATA AND SOFTWARE AVAILABILITY

### SUPPLEMENTAL INFORMATION

Supplemental Information includes seven figures and two tables and can be found with this article online at <https://doi.org/10.1016/j.molcel.2018.08.020>.

## ACKNOWLEDGMENTS

We thank the GeneCore facility of EMBL for high-throughput sequencing. We thank K.M. Miller (University of Texas) for critical reading of the manuscript and J.-Y. Masson/G. Dellaire for reagents and advice for the Clover/HR assay. Funding was provided by Polish National Science Centre (2011/02/A/NZ2/00014 to K.G., 2016/21/B/ST6/01471 to N.D., and 2015/17/D/NZ2/03711 to M.S.) and Foundation for Polish Science (TEAM to K.G.). Funding to M.R., B.F., and N.D. was provided by NIH (5 R01 GM 112131). J.C. acknowledges funding support from the Canadian Institutes of Health Research (FND-143314) and holds the Canada Research Chair in Chromatin Biology and Molecular Epigenetics. M.A. was supported by the Fondation pour la Recherche Médicale (FRM). Funding in G.L.'s laboratory was provided by grants from the European Research Council (ERC-2014-CoG 647344), Agence Nationale pour la Recherche (ANR-14-CE10-0002-01 and ANR-13-BSV8-0013), the Institut National contre le Cancer (INCA), and the Ligue Nationale contre le Cancer (LNCC). T.C. is an INSERM researcher.

## AUTHOR CONTRIBUTIONS

T.C., C.A., and F.A. performed ChIP experiments. A.L. performed the SAGA complex purification, enzymatic assays, and DSB repair assays under the supervision of J.C. V.R., M.A., and J.S.I. performed bioinformatic analyses of ChIP and BLESS datasets. A.B., M.S., and K.G. performed BLESS experiments. M.R., B.F. and N.D. performed the comparison with Hi-C data. T.C. and G.L. conceived, supervised, and analyzed experiments. T.C. and G.L. wrote the manuscript. All authors commented on and edited the manuscript.

## DECLARATION OF INTERESTS

The authors declare no competing interests.

Received: June 26, 2017

Revised: July 13, 2018

Accepted: August 13, 2018

Published: September 27, 2018

## REFERENCES

- Alawi, H.E., and Downs, J.A. (2015). Removal of H2A.Z by INO80 promotes homologous recombination. *EMBO Rep.* **16**, 986–994.
- Aymard, F., and Legube, G. (2016). A TAD closer to ATM. *Mol. Cell. Oncol.* **3**, e1134411.
- Aymard, F., Bugler, B., Schmidt, C.K., Guillou, E., Caron, P., Briois, S., Iacovoni, J.S., Daburon, V., Miller, K.M., Jackson, S.P., and Legube, G. (2014). Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nat. Struct. Mol. Biol.* **21**, 366–374.
- Aymard, F., Aguirrebengoa, M., Guillou, E., Javierre, B.M., Bugler, B., Arnould, C., Rocher, V., Iacovoni, J.S., Biernacka, A., Skrzypczak, M., et al. (2017). Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes. *Nat. Struct. Mol. Biol.* **24**, 353–361.
- Batta, K., Zhang, Z., Yen, K., Goffman, D.B., and Pugh, B.F. (2011). Genome-wide function of H2B ubiquitylation in promoter and genic regions. *Genes Dev.* **25**, 2254–2265.
- Bennett, G., and Peterson, C.L. (2015). SWI/SNF recruitment to a DNA double-strand break by the NuA4 and Gcn5 histone acetyltransferases. *DNA Repair (Amst.)* **30**, 38–45.
- Biehs, R., Steinlage, M., Barton, O., Juhasz, S., Kunzel, J., Spies, J., Shibata, A., Jeggo, P.A., and Lobrich, M. (2017). DNA Double-Strand Break Resection Occurs during Non-homologous End Joining in G1 but Is Distinct from Resection during Homologous Recombination. *Mol. Cell* **65**, 671–684.e675.
- Caron, P., Aymard, F., Iacovoni, J.S., Briois, S., Canitrot, Y., Bugler, B., Massip, L., Losada, A., and Legube, G. (2012). Cohesin protects genes against γH2AX Induced by DNA double-strand breaks. *PLoS Genet.* **8**, e1002460.
- Caron, P., Choudjaye, J., Clouaire, T., Bugler, B., Daburon, V., Aguirrebengoa, M., Mangeat, T., Iacovoni, J.S., Álvarez-Quilón, A., Cortés-Ledesma, F., and Legube, G. (2015). Non-redundant Functions of ATM and DNA-PKcs in Response to DNA Double-Strand Breaks. *Cell Rep.* **13**, 1598–1609.
- Chen, H., Ruiz, P.D., Novikov, L., Casill, A.D., Park, J.W., and Gamble, M.J. (2014). MacroH2A1.1 and PARP-1 cooperate to regulate transcription by promoting CBP-mediated H2B acetylation. *Nat. Struct. Mol. Biol.* **21**, 981–989.
- Cheung, W.L., Turner, F.B., Krishnamoorthy, T., Wolner, B., Ahn, S.H., Foley, M., Dorsey, J.A., Peterson, C.L., Berger, S.L., and Allis, C.D. (2005). Phosphorylation of histone H4 serine 1 during DNA damage requires casein kinase II in *S. cerevisiae*. *Curr. Biol.* **15**, 656–660.
- Clouaire, T., and Legube, G. (2015). DNA double strand break repair pathway choice: a chromatin based decision? *Nucleus* **6**, 107–113.
- Cohen, S., Puget, N., Lin, Y.L., Clouaire, T., Aguirrebengoa, M., Rocher, V., Pasero, P., Canitrot, Y., and Legube, G. (2018). Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nat. Commun.* **9**, 533.
- Crosetto, N., Mitra, A., Silva, M.J., Bienko, M., Dojer, N., Wang, Q., Karaca, E., Chiarle, R., Skrzypczak, M., Ginalski, K., et al. (2013). Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat. Methods* **10**, 361–365.
- Dalvai, M., Loehr, J., Jacquet, K., Huard, C.C., Roques, C., Herst, P., Côté, J., and Doyon, Y. (2015). A Scalable Genome-Editing-Based Approach for Mapping Multiprotein Complexes in Human Cells. *Cell Rep.* **13**, 621–633.
- Deplus, R., Blanchon, L., Rajavelu, A., Boukaba, A., Defrance, M., Luciani, J., Rothé, F., Dedeurwaerd, S., Denis, H., Brinkman, A.B., et al. (2014). Regulation of DNA methylation patterns by CK2-mediated phosphorylation of Dnmt3a. *Cell Rep.* **8**, 743–753.
- Dimitrova, N., Chen, Y.C., Spector, D.L., and de Lange, T. (2008). 53BP1 promotes non-homologous end joining of telomeres by increasing chromatin mobility. *Nature* **456**, 524–528.
- Dobbin, M.M., Madabhushi, R., Pan, L., Chen, Y., Kim, D., Gao, J., Ahanonu, B., Pao, P.C., Qiu, Y., Zhao, Y., and Tsai, L.H. (2013). SIRT1 collaborates with ATM and HDAC1 to maintain genomic stability in neurons. *Nat. Neurosci.* **16**, 1008–1015.
- Downs, J.A., Allard, S., Jobin-Robitaille, O., Javaheri, A., Auger, A., Bouchard, N., Kron, S.J., Jackson, S.P., and Côté, J. (2004). Binding of chromatin-modifying activities to phosphorylated histone H2A at DNA damage sites. *Mol. Cell* **16**, 979–990.
- Doyon, Y., and Côté, J. (2016). Preparation and Analysis of Native Chromatin-Modifying Complexes. *Methods Enzymol.* **573**, 303–318.
- Engel, M., Eggert, C., Kaplick, P.M., Eder, M., Röh, S., Tietze, L., Namendorf, C., Arloth, J., Weber, P., Rex-Haffner, M., et al. (2018). The Role of m6A/mRNA Methylation in Stress Response Regulation. *Neuron* **99**, 389–403.e9.
- Fan, Y., Nikitina, T., Zhao, J., Fleury, T.J., Bhattacharyya, R., Bouhassira, E.E., Stein, A., Woodcock, C.L., and Skoultschi, A.I. (2005). Histone H1 depletion in mammals alters global chromatin structure but causes specific changes in gene regulation. *Cell* **123**, 1199–1212.
- Fleming, A.B., Kao, C.F., Hillyer, C., Piakaart, M., and Osley, M.A. (2008). H2B ubiquitylation plays a role in nucleosome dynamics during transcription elongation. *Mol. Cell* **31**, 57–66.
- Fradet-Turcotte, A., Canny, M.D., Escribano-Díaz, C., Orthwein, A., Leung, C.C., Huang, H., Landry, M.C., Kitevski-LeBlanc, J., Noordermeer, S.M., Sicheri, F., and Durocher, D. (2013). 53BP1 is a reader of the DNA-damage-induced H2A Lys 15 ubiquitin mark. *Nature* **499**, 50–54.
- Gong, F., Chiu, L.Y., Cox, B., Aymard, F., Clouaire, T., Leung, J.W., Cammarata, M., Perez, M., Agarwal, P., Brodbelt, J.S., et al. (2015). Screen identifies bromodomain protein ZMYND8 in chromatin recognition of transcription-associated DNA damage that promotes homologous recombination. *Genes Dev.* **29**, 197–211.
- Gong, F., Clouaire, T., Aguirrebengoa, M., Legube, G., and Miller, K.M. (2017). Histone demethylase KDM5A regulates the ZMYND8-NuRD chromatin remodeler to promote DNA repair. *J. Cell Biol.* **216**, 1959–1974.

- Gursoy-Yuzugullu, O., Ayrapetov, M.K., and Price, B.D. (2015). Histone chaperone Anp32e removes H2A.Z from DNA double-strand breaks and promotes nucleosome reorganization and DNA repair. *Proc. Natl. Acad. Sci. USA* **112**, 7507–7512.
- Helmlinger, D., and Tora, L. (2017). Sharing the SAGA. *Trends Biochem. Sci.* **42**, 850–861.
- Hustedt, N., and Durocher, D. (2016). The control of DNA repair by the cell cycle. *Nat. Cell Biol.* **19**, 1–9.
- Huyen, Y., Zgheib, O., Ditullio, R.A., Jr., Gorgoulis, V.G., Zacharatos, P., Petty, T.J., Sheston, E.A., Mellert, H.S., Stavridi, E.S., and Halazonetis, T.D. (2004). Methylated lysine 79 of histone H3 targets 53BP1 to DNA double-strand breaks. *Nature* **432**, 406–411.
- Iacovoni, J.S., Caron, P., Lassadi, I., Nicolas, E., Massip, L., Trouche, D., and Legube, G. (2010). High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. *EMBO J.* **29**, 1446–1457.
- Iannelli, F., Galbiati, A., Capozzo, I., Nguyen, Q., Magnuson, B., Michelini, F., D'Alessandro, G., Cabrini, M., Roncador, M., Francia, S., et al. (2017). A damaged genome's transcriptional landscape through multilayered expression profiling around *in situ*-mapped DNA double-strand breaks. *Nat. Commun.* **8**, 15656.
- Jacquet, K., Fradet-Turcotte, A., Avvakumov, N., Lambert, J.P., Roques, C., Pandita, R.K., Paquet, E., Herst, P., Gingras, A.C., Pandita, T.K., et al. (2016). The TIP60 Complex Regulates Bivalent Chromatin Recognition by 53BP1 through Direct H4K20me Binding and H2AK15 Acetylation. *Mol. Cell* **62**, 409–421.
- Kakaroukas, A., Ismail, A., Klement, K., Goodarzi, A.A., Conrad, S., Freire, R., Shibata, A., Lobrich, M., and Jeggo, P.A. (2013). Opposing roles for 53BP1 during homologous recombination. *Nucleic Acids Res.* **41**, 9719–9731.
- Khurana, S., Kruhlak, M.J., Kim, J., Tran, A.D., Liu, J., Nyswaner, K., Shi, L., Jailwala, P., Sung, M.H., Hakim, O., and Oberdoerffer, P. (2014). A macrohistone variant links dynamic chromatin compaction to BRCA1-dependent genome maintenance. *Cell Rep.* **8**, 1049–1062.
- Kleiner, R.E., Verma, P., Molloy, K.R., Chait, B.T., and Kapoor, T.M. (2015). Chemical proteomics reveals a γH2AX–53BP1 interaction in the DNA damage response. *Nat. Chem. Biol.* **11**, 807–814.
- Lee, H.S., Park, J.H., Kim, S.J., Kwon, S.J., and Kwon, J. (2010). A cooperative activation loop among SWI/SNF, gamma-H2AX and H3 acetylation for DNA double-strand break repair. *EMBO J.* **29**, 1434–1445.
- Marnef, A., and Legube, G. (2017). Organizing DNA repair in the nucleus: DSBs hit the road. *Curr. Opin. Cell Biol.* **46**, 1–8.
- Marnef, A., Cohen, S., and Legube, G. (2017). Transcription-Coupled DNA Double-Strand Break Repair: Active Genes Need Special Care. *J. Mol. Biol.* **429**, 1277–1288.
- McGinty, R.K., Kim, J., Chatterjee, C., Roeder, R.G., and Muir, T.W. (2008). Chemically ubiquitylated histone H2B stimulates hDot1L-mediated intranucleosomal methylation. *Nature* **453**, 812–816.
- Miller, K.M., Tjeertes, J.V., Coates, J., Legube, G., Polo, S.E., Britton, S., and Jackson, S.P. (2010). Human HDAC1 and HDAC2 function in the DNA-damage response to promote DNA nonhomologous end-joining. *Nat. Struct. Mol. Biol.* **17**, 1144–1151.
- Mitra, A., Skrzypczak, M., Ginalski, K., and Rowicka, M. (2015). Strategies for achieving high sequencing accuracy for low diversity samples and avoiding sample bleeding using illumina platform. *PLoS ONE* **10**, e0120520.
- Mladenov, E., Magin, S., Soni, A., and Iliakis, G. (2016). DNA double-strand-break repair in higher eukaryotes and its role in genomic instability and cancer: Cell cycle and proliferation-dependent regulation. *Semin. Cancer Biol.* **37–38**, 51–64.
- Morrison, A.J., Highland, J., Krogan, N.J., Arbel-Eden, A., Greenblatt, J.F., Haber, J.E., and Shen, X. (2004). INO80 and gamma-H2AX interaction links ATP-dependent chromatin remodeling to DNA damage repair. *Cell* **119**, 767–775.
- Moyal, L., Lerenthal, Y., Gana-Weisz, M., Mass, G., So, S., Wang, S.Y., Eppink, B., Chung, Y.M., Shalev, G., Shema, E., et al. (2011). Requirement of ATM-dependent monoubiquitylation of histone H2B for timely repair of DNA double-strand breaks. *Mol. Cell* **41**, 529–542.
- Nakamura, K., Kato, A., Kobayashi, J., Yanagihara, H., Sakamoto, S., Oliveira, D.V., Shimada, M., Tauchi, H., Suzuki, H., Tashiro, S., et al. (2011). Regulation of homologous recombination by RNF20-dependent H2B ubiquitination. *Mol. Cell* **41**, 515–528.
- Noon, A.T., Shibata, A., Rief, N., Löbrich, M., Stewart, G.S., Jeggo, P.A., and Goodarzi, A.A. (2010). 53BP1-dependent robust localized KAP-1 phosphorylation is essential for heterochromatic DNA double-strand break repair. *Nat. Cell Biol.* **12**, 177–184.
- Ogiwara, H., Uji, A., Otsuka, A., Satoh, H., Yokomi, I., Nakajima, S., Yasui, A., Yokota, J., and Kohno, T. (2011). Histone acetylation by CBP and p300 at double-strand break sites facilitates SWI/SNF chromatin remodeling and the recruitment of non-homologous end joining factors. *Oncogene* **30**, 2135–2146.
- Panier, S., and Boulton, S.J. (2014). Double-strand break repair: 53BP1 comes into focus. *Nat. Rev. Mol. Cell Biol.* **15**, 7–18.
- Pauty, J., Couturier, A.M., Rodrigue, A., Caron, M.C., Coulombe, Y., Dellaire, G., and Masson, J.Y. (2017). Cancer-causing mutations in the tumor suppressor PALB2 reveal a novel cancer mechanism using a hidden nuclear export signal in the WD40 repeat motif. *Nucleic Acids Res.* **45**, 2644–2657.
- Pellegrino, S., Michelena, J., Teloni, F., Imhof, R., and Altmyer, M. (2017). Replication-Coupled Dilution of H4K20me2 Guides 53BP1 to Pre-replicative Chromatin. *Cell Rep.* **19**, 1819–1831.
- Pinder, J., Salsman, J., and Dellaire, G. (2015). Nuclear domain ‘knock-in’ screen for the evaluation and identification of small molecule enhancers of CRISPR-based genome editing. *Nucleic Acids Res.* **43**, 9379–9392.
- Ramachandran, S., Haddad, D., Li, C., Le, M.X., Ling, A.K., So, C.C., Nepal, R.M., Gommerman, J.L., Yu, K., Ketela, T., et al. (2016). The SAGA Deubiquitination Module Promotes DNA Repair and Class Switch Recombination through ATM and DNAPK-Mediated γH2AX Formation. *Cell Rep.* **15**, 1554–1565.
- Rao, S.S., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., and Aiden, E.L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680.
- Rogakou, E.P., Pilch, D.R., Orr, A.H., Ivanova, V.S., and Bonner, W.M. (1998). DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J. Biol. Chem.* **273**, 5858–5868.
- Ropars, V., Drevet, P., Legrand, P., Baconnais, S., Amram, J., Faure, G., Márquez, J.A., Piétrement, O., Guerois, R., Callebaut, I., et al. (2011). Structural characterization of filaments formed by human Xrc4–Cernunnos/XLF complex involved in nonhomologous DNA end-joining. *Proc. Natl. Acad. Sci. USA* **108**, 12663–12668.
- Savic, V., Yin, B., Maas, N.L., Bredemeyer, A.L., Carpenter, A.C., Helmink, B.A., Yang-Iott, K.S., Sleckman, B.P., and Bassing, C.H. (2009). Formation of dynamic gamma-H2AX domains along broken DNA strands is distinctly regulated by ATM and MDC1 and dependent upon H2AX densities in chromatin. *Mol. Cell* **34**, 298–310.
- Schippler, A., and Iliakis, G. (2013). DNA double-strand-break complexity levels and their possible contributions to the probability for error-prone processing and repair pathway choice. *Nucleic Acids Res.* **41**, 7589–7605.
- Schwertman, P., Bekker-Jensen, S., and Mailand, N. (2016). Regulation of DNA double-strand break repair by ubiquitin and ubiquitin-like modifiers. *Nat. Rev. Mol. Cell Biol.* **17**, 379–394.
- Segala, G., Bennesch, M.A., Pandey, D.P., Hulo, N., and Picard, D. (2016). Monoubiquitination of Histone H2B Blocks Eviction of Histone Variant H2A.Z from Inducible Enhancers. *Mol. Cell* **64**, 334–346.
- Sellou, H., Lebeaupin, T., Chapuis, C., Smith, R., Hegele, A., Singh, H.R., Kozlowski, M., Bultmann, S., Ladurner, A.G., Timinszky, G., and Huet, S. (2016). The poly(ADP-ribose)-dependent chromatin remodeler Alc1 induces local chromatin relaxation upon DNA damage. *Mol. Biol. Cell* **27**, 3791–3799.
- Soshnev, A.A., Josefowicz, S.Z., and Allis, C.D. (2016). Greater Than the Sum of Parts: Complexity of the Dynamic Epigenome. *Mol. Cell* **62**, 681–694.

- Strickfaden, H., McDonald, D., Kruhlak, M.J., Haince, J.F., Th'ng, J.P., Rouleau, M., Ishibashi, T., Corry, G.N., Ausio, J., Underhill, D.A., et al. (2016). Poly(ADP-ribosylation)-dependent Transient Chromatin Decondensation and Histone Displacement following Laser Microirradiation. *J. Biol. Chem.* 291, 1789–1802.
- Sun, Z.W., and Allis, C.D. (2002). Ubiquitination of histone H2B regulates H3 methylation and gene silencing in yeast. *Nature* 418, 104–108.
- Talbert, P.B., and Henikoff, S. (2017). Histone variants on the move: substrates for chromatin dynamics. *Nat. Rev. Mol. Cell Biol.* 18, 115–126.
- Tang, J., Cho, N.W., Cui, G., Manion, E.M., Shanbhag, N.M., Botuyan, M.V., Mer, G., and Greenberg, R.A. (2013). Acetylation limits 53BP1 association with damaged chromatin to promote homologous recombination. *Nat. Struct. Mol. Biol.* 20, 317–325.
- Thorslund, T., Ripplinger, A., Hoffmann, S., Wild, T., Uckelmann, M., Villumsen, B., Narita, T., Sixma, T.K., Choudhary, C., Bekker-Jensen, S., and Mailand, N. (2015). Histone H1 couples initiation and amplification of ubiquitin signalling after DNA damage. *Nature* 527, 389–393.
- Toiber, D., Erdel, F., Bouazoune, K., Silberman, D.M., Zhong, L., Mulligan, P., Sebastian, C., Cosentino, C., Martinez-Pastor, B., Giacosa, S., et al. (2013). SIRT6 recruits SNF2H to DNA break sites, preventing genomic instability through chromatin remodeling. *Mol. Cell* 51, 454–468.
- Utley, R.T., Lacoste, N., Jobin-Robitaille, O., Allard, S., and Côté, J. (2005). Regulation of NuA4 histone acetyltransferase activity in transcription and DNA repair by phosphorylation of histone H4. *Mol. Cell. Biol.* 25, 8179–8190.
- van Attikum, H., Fritsch, O., Hohn, B., and Gasser, S.M. (2004). Recruitment of the INO80 complex by H2A phosphorylation links ATP-dependent chromatin remodeling with DNA double-strand break repair. *Cell* 119, 777–788.
- Ward, I.M., Minn, K., Jorda, K.G., and Chen, J. (2003). Accumulation of check-point protein 53BP1 at DNA breaks involves its binding to phosphorylated histone H2AX. *J. Biol. Chem.* 278, 19579–19582.
- Wilson, M.D., Benlekbir, S., Fradet-Turcotte, A., Sherker, A., Julien, J.P., McEwan, A., Noordermeer, S.M., Sicheri, F., Rubinstein, J.L., and Durocher, D. (2016). The structural basis of modified nucleosome recognition by 53BP1. *Nature* 536, 100–103.
- Xu, Y., Ayrapetov, M.K., Xu, C., Gursoy-Yuzugullu, O., Hu, Y., and Price, B.D. (2012). Histone H2A.Z controls a critical chromatin remodeling step required for DNA double-strand break repair. *Mol. Cell* 48, 723–733.
- Zhu, B., Zheng, Y., Pham, A.D., Mandal, S.S., Erdjument-Bromage, H., Tempst, P., and Reinberg, D. (2005). Monoubiquitination of human histone H2B: the factors involved and their roles in HOX gene regulation. *Mol. Cell* 20, 601–611.

**STAR★METHODS****KEY RESOURCES TABLE**

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Antibodies		
Anti-H4	Abcam	Cat# ab7311, RRID:AB_305837
Anti-H2B	Abcam	Cat# ab1790, RRID:AB_302612
Anti-H3	Abcam	Cat# ab1791, RRID:AB_302613
Anti-H2AZ	Abcam	Cat# ab4174, RRID:AB_304345
Anti-H1	Abcam	Cat# ab17677, RRID:AB_2117984
Anti-macroH2A1	Millipore	Cat# 07-219, RRID:AB_310439
Anti-H2AZac	Abcam	Cat# ab18262, RRID:AB_873820
Anti-H3K79me2	Active Motif	Cat# 39143, RRID:AB_2561018
Anti-H4K20me1	Active Motif	Cat# 39727, RRID:AB_2615074
Anti-H3K9me2	Abcam	Cat# ab1220, RRID:AB_449854
Anti-H3K9me3	Abcam	Cat# ab8898, RRID:AB_306848
Anti-H3K4me2	Millipore	Cat# 07-030, RRID:AB_10099880
Anti-H3K36me2	Abcam	Cat# ab9049, RRID:AB_1280939
Anti-H3K36me3	Abcam	Cat# ab9050, RRID:AB_306966
Anti-H4K12ac	Abcam	Cat# ab46983, RRID:AB_873859
Anti-H4K16ac	Millipore	Cat# 07-329, RRID:AB_310525
Anti-H3K56ac	Abcam	Cat# ab76307, RRID:AB_1523762
Anti-H4S1P	Novus	Cat# NB21-2000, RRID:AB_11019163
Anti-H4K20me2	Abcam	Cat# ab9052, RRID:AB_1951942
Anti-H2BK120Ub	Cell Signaling	Cat# 5546, RRID:AB_106934
Anti-H2BK120ac	Millipore	Cat# 07-564, RRID:AB_11213734
Anti-Ubiquitylated proteins	Millipore	Cat# 04-263, RRID:AB_612093
Anti-γH2AX	Abcam	Cat# ab81299, RRID:AB_1640564
Anti-XRCC4	Abcam	Cat# ab145, RRID:AB_301278
Anti-RAD51	Santa Cruz	Cat# H-92, RRID:AB_2253533
Anti-53BP1	Novus	Cat# NB100-305, RRID:AB_10001695
Anti-DNA Ligase IV	Genetex	Cat# GTX55592
Chemicals, Peptides, and Recombinant Proteins		
(Z)-4-Hydroxytamoxifen	Sigma	Cat# H7904
Thymidine	Sigma	Cat# T1895
Critical Commercial Assays		
Quick-RNA MicroPrep kit	Zymo Research	Cat# R1050
qScript cDNA synthesis kit	Quanta Bio	Cat# 95047-100
Lipofectamine RNAiMAX	Invitrogen	Cat# 13778075
SE. Cell Line 4D-Nucleofector X Kit	Lonza	Cat# V4XC1012
Deposited Data		
Raw data (ChIP-seq and BLESS)	This paper	ArrayExpress: E-MTAB-5817
RAD51 and XRCC4 ChIP-seq	Aymard et al., 2014	ArrayExpress E-MTAB-1241
RNA polymerase II S2P ChIP-seq	Cohen et al., 2018	ArrayExpress E-MTAB-6318
MethylCap-seq	Deplus et al., 2014	GEO GSE26810

(Continued on next page)

***Continued***

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Experimental Models: Cell Lines		
DlVA cell	Iacovoni et al., 2010	N/A
U2OS-1Scel GFP-RFP (NHEJ)	Jacquet et al., 2016	N/A
3xFlag-Twin-strep-tagged SUPT7L K562	Dalvai et al., 2015	N/A
U2OS	N/A	ATCC HTB-96, RRID:CVCL_0042
Oligonucleotides		
HR-DSB1 for ChIP-qPCR FW GATTGGCTATGGGTGTGGAC REV CATCCTTGCAAACCAGTCCT	Aymard et al., 2014	N/A
HR-DSB2 for ChIP-qPCR FW CCGCCAGAAAGTTCCCTAGA REV CTCACCCTTGCAGCACTTG	Aymard et al., 2014	N/A
NHEJ-DSB for ChIP-qPCR FW TGCCGGTCTCCTAGAAGTTG REV GCGCTTGATTCCTGAGT	Aymard et al., 2014	N/A
ACTB for ChIP-qPCR FW AGCCGGGCTCTTGCCAAT REV AGTTAGCGCCCCAAAGGACCA	This paper	N/A
TAF12 for ChIP-qPCR FW GCTGAGACGAACGCTTCACT REV CCTTCGAACACTGACCCACT	This paper	N/A
siRNA siSUPT7L-46 CUACUAGACCCAACAGAAA [dT] [dT] UUCUGUUGGGGUUAGUAG[dT] [dT]	This paper	N/A
siRNA siSUPT7L-47 CUAUCACAGUUACAUGC UA[dT] [dT] UAGCAUGUAACUGUGAUAG[dT] [dT]	This paper	N/A
siRNA siKAT2B (PCAF) CUCUAUCCUCACUCA UUU[dT] [dT] AAAUGAGUGAGGAUUAAGAG[dT] [dT]	This paper	N/A
siRNA siKAT2A (GCN5) GCUACUACGUGACCC GGAA[dT] [dT] UUCCGGGUCACGUAGUAGC[dT] [dT]	This paper	N/A
Recombinant DNA		
pX330-LMNAgRNA1	Pauty et al., 2017	N/A
pCR2.1-CloverLMNAdonor	Pinder et al., 2015	N/A
piRFP670-N1	Pinder et al., 2015	N/A
AVS1_Puro_PKG1_3xFLAG_Twin_Strep	Dalvai et al., 2015	Addgene #68375
Software and Algorithms		
FastQC	N/A	<a href="https://www.bioinformatics.babraham.ac.uk/projects/fastqc">https://www.bioinformatics.babraham.ac.uk/projects/fastqc</a>
Bwa	N/A	<a href="http://bio-bwa.sourceforge.net">http://bio-bwa.sourceforge.net</a>
Samtools	N/A	<a href="http://www.htslib.org/">http://www.htslib.org/</a>
deepTools	N/A	<a href="http://deftools.readthedocs.io">http://deftools.readthedocs.io</a>
R	N/A	<a href="https://www.r-project.org">https://www.r-project.org</a>

**CONTACT FOR REAGENT AND RESOURCE SHARING**

Further information and requests for resources and reagents should be directed to and will be fulfilled by the Lead Contact, Gaëlle Legube ([gaelle.legube@univ-tlse3.fr](mailto:gaelle.legube@univ-tlse3.fr)). DlVA cells are subjected to an MTA, to be signed with the CNRS.

## EXPERIMENTAL MODEL AND SUBJECT DETAILS

### Chemicals

4-hydroxytamoxifen (4OHT) was purchased from Sigma (Sigma; H7904). 4OHT is reconstituted in DMSO at a final concentration of 10 mM, and stored at -20°C. 4OHT was further used at a final concentration of 300 nM to induce DSB in DlVA cells. Thymidine is purchased from Sigma (T1895), reconstituted in PBS at 100 mM and filter sterilized immediately before use.

### Cell Culture and Cell Lines

DlVA (AsiSI-ER-U20S) cells were cultured in Dulbecco's modified Eagle's medium (DMEM) supplemented with antibiotics, 10% FCS (Invitrogen) with 1 µg/mL puromycin at 37°C under a humidified atmosphere with 5% CO<sub>2</sub>. K562 cells stably expressing near physiological levels of 3xFLAG-Twin-Strep-tagged SUPT7L (Dalvai et al., 2015; Doyon and Côté, 2016) were cultured in RPMI medium supplemented with 0.5 µg/mL puromycin.

## METHOD DETAILS

### Cell treatments

For AsiSI-dependent DSB induction, cells were treated with 300 nM 4OHT for 4 hr. For synchronization in G1 and G2, cells were incubated with 2 mM thymidine for 18 hr, released for 12 hr and subjected to the second thymidine treatment for 18 hr. G1 and G2 cells were treated with 4OHT, respectively, at 11 and 6 hr following thymidine release and harvested 4 hr later.

### Establishment of cell line expressing SUPT7L from the AAVS1 safe harbor and TAP Purification of native human SAGA HAT/DUB complex

K562 cells stably expressing near physiological levels of 3xFLAG-Twin-Strep-tagged SUPT7L was established as described previously (Dalvai et al., 2015; Doyon and Côté, 2016). Briefly, SUPT7L cDNA was cloned into AAVS1\_Puro\_PGK1\_3xFLAG\_Twin\_Strep (Addgene #68375). The cassette is integrated at the AAVS1 locus after DSB induction and recombination targeted by co-transfection with ZFN expression plasmid. Two hundred thousand cells were transfected with 400 ng of ZFN expression vector and 4 µg of donor constructs. Selection and cloning were performed in RPMI medium supplemented with 0.5 µg/mL puromycin starting at 2 to 3 days post transfection. Clones were obtained by limiting dilution and expanded before harvest for western blot analysis.

Native SAGA complex was purified from 3L of the 3xFlag-Twin-strep-tagged SUPT7L K562 cells as described in (Doyon and Côté, 2016). Nuclear extracts were prepared following standard procedures and pre-cleared with CL6B Sepharose beads. FLAG immunoprecipitations with anti-FLAG agarose affinity gel (Sigma M2) were performed followed by elution with 3xFLAG peptide (200 µg/mL from Sigma in the following buffer: 20 mM HEPES pH 7.5, 150 mM KCl, 0.1 mM EDTA, 10% glycerol, 0.1% Tween20, 1 mM DTT and supplemented with proteases, deacetylases, and phosphatase inhibitors), followed by Strep immunoprecipitation with Strep-Tactin Sepharose beads (IBA) and elution with 5 mM D-biotin in the same buffer used for Flag elution. Mass spec analysis confirmed the co-purification of all bona fide subunits of the human SAGA complex.

### In vitro histone acetylation assay

Histone acetyltransferase assays were performed as described previously with minor modifications using the purified complex (Jacquet et al., 2016). Briefly, 100 ng of the recombinant H2B was incubated in a 15 µL reaction containing 50 mM Tris-HCl pH 8.0, 10% glycerol, 1 mM EDTA, 1 mM DTT, 1 mM PMSF, 10 mM sodium butyrate 0.15 mM unlabeled Acetyl-CoA (Sigma) with or without the purified complex for 30 min at 30°C. Samples were analyzed by western blot with indicated antibodies.

### In vitro histone de-ubiquitination assay

DUB enzymatic activity was assayed by pre-incubating native nucleosomes purified from HeLa cells in buffer containing 50 mM Tris-HCl pH 7.5, 125 mM NaCl, 1 mM DTT, 1 mM MgCl<sub>2</sub>, 1 mM EDTA, and protease inhibitor cocktail at 30°C for 20 min before adding purified SAGA complex. Reaction were incubated at 30°C throughout the time course and terminated by adding SDS-PAGE sample loading buffer following by western blot analysis using indicated antibodies.

### RT-qPCR

For measuring siRNA-mediated depletion total RNA was extracted using the Quick-RNA MicroPrep kit (Zymo Research). One microgram of RNA was reverse transcribed using qScript cDNA synthesis kit (Quanta biosciences) according to the manufacturers' protocols. Samples were then subjected to quantitative PCR (qPCR) using Lightcycler (Roche). The relative abundance of target mRNA was calculated according to the  $\Delta\Delta$  cycle threshold method ( $\Delta\Delta Ct$ ). mRNA expression levels of the housekeeping gene 36B4 gene (also called ribosomal phosphoprotein P0, RPLP0) were used as an internal control to normalize each qPCR reaction. The relative expression levels were calculated as fold enrichment of treated cells over the control cells. Experiments were performed as independent biological triplicates and data are presented as mean ± SD.

### Cas9/mClover-LMNA1 homologous recombination assay

The pCR2.1-CloverLMNAdonor, pX330-LMNAgRNA1 and piRFP670-N1 plasmids were previously described (Pinder et al., 2015). The pX330-LMNAgRNA1 plasmid used in this study was modified from the original plasmid to remove the 3XFlag tag from the Cas9 endonuclease previously described (Pauty et al., 2017). U2OS cells were seeded and transfected with indicated siRNA using Lipofectamine RNAiMAX (Invitrogen). Twenty-four hours post-transfection, 1.5 million cells were transfected on the 4D-Nucleofector X-unit (program DG-130), using complete nucleofector solution (SE. Cell Line 4D-Nucleofector X Kit, Lonza) containing 1 µg of pCR2.1-CloverLMNAdonor, 1 µg pX330-LMNAgRNA1, 0.1 µg of piRFP670-N1 (used as transfection control) and 200pmol of siRNA and, immediately resuspended in culture media and transferred to a 10 cm dish. After 48 hr, cells were trypsinized and 0.25 million cells plated into glass coverslips and the rest analyzed on BD Accuri C6 Plus Flow Cytometer. Cloverexpression was assayed by fluorescence microscopy the next day that is 72 hr post-nucleofection. Data represent the mean percentages ( $\pm$ SD) of Clover-positive cells (structured nuclear GFP signal) over the iRFP670-positive population from independent experiments performed in triplicates (n > 800 cells per condition).

### NHEJ assay

The U2OS cell line with I-Sce1 reporter GFP-RFP cassette to measure NHEJ has been described (Jacquet et al., 2016). The cells were transfected with 200 nmol of the indicated siRNA using Lipofectamine RNAimax (Invitrogen) for 36 hr and infected with I-Sce1 adenovirus for 1 hr. Cells are harvested 48 hr after DSB induction and analyzed by FACS for GFP and RFP expression on a BD Accuri C6 Plus Flow Cytometer. Control cells produce 10% of RFP positive/GFP negative cells in the NHEJ reporter. The data presented are from biological triplicate experiments.

### Cell cycle analysis

For cell cycle studies, DlvA cells were harvested by trypsinization, fixed with cold 70% ethanol, treated with ribonuclease A and propidium iodide. FACS analysis was performed using a BD Accuri C6 Plus Flow Cytometer.

### ChIP

ChIP assays were carried out according to the protocol described in Aymard et al., 2014 and Iacovoni et al., 2010. The amount of chromatin and antibodies used are detailed in Table S2. For quantitative PCR analysis (Figure S2), both input and IP samples were analyzed using the primers FW GATTGGCTATGGGTGTTGAC and REV CATCCTGCAAACCAGTCCT (HR DSB1). For Figure S4, the following primers were also used: FW CCGCCAGAAAGTTCTAGA and REV CTCACCCCTGCAGCACTTG (HR-DSB2), FW TGCCGGTCTCTAGAACGTTG and REV GCGCTTGATTCCTGAGT (NHEJ-DSB), FW AGCCGGGCTTGCCAAT and REV AGTAGCGCCAAAGGACCA (ACTB), FW GCTGAGACGAACGCTTCACT and REV CCTTCGAACACTGACCCACT (TAF12). ChIP efficiencies were calculated as the percent of input DNA immunoprecipitated. Prior to next-generation sequencing library preparation, samples from multiple ChIP experiments were pooled and sonicated for 5 cycles (30 s on, 30 s off, high setting) with a Bioruptor (Diagenode) then concentrated with a vacuum concentrator (Eppendorf). Sequencing libraries were prepared by using 10 ng of purified DNA (averaged size 250–300 bp), and subjected to high throughput sequencing.

### BLESS

BLESS was performed as described in Crosetto et al., 2013. Briefly, cells were fixed with 2% formaldehyde to stabilize chromatin and prevent artificial DSBs, plasma membranes were lysed, and intact nuclei were recovered by centrifugation. Nuclei were deproteinated using Proteinase K. DSBs were then blunted and 5' phosphorylated using the Quick Blunting Kit (NEB) and ligated to a biotinylated linker (proximal) using T4 ligase (NEB). Total DNA was extracted by precipitation with isopropanol and fragmented by sonication using a Covaris S220 to create ~400 bp fragments. Labeled fragments were captured by streptavidin beads (Invitrogen) and once again blunted and phosphorylated using the Quick Blunting Kit (NEB), and ligated to a second linker (distal). The resulting circular DNA was linearized by I-Sce1 (NEB) digestion and amplified by PCR. Sequencing libraries were prepared using TruSeq Nano DNA LT Library Preparation Kit (Illumina). Quality and quantity of libraries were assessed on 2100 Bioanalyzer (Agilent) using High Sensitivity DNA Kit (Agilent), and on Qubit 2.0 using Qubit dsDNA HS Assay Kit (ThermoFisher).

## QUANTIFICATION AND STATISTICAL ANALYSIS

### ChIP-seq data processing

H3K4me2, H4K20me1, H3K36me3, H3K79me2, H2Bub, H3K56ac, H4K16ac, H3K9me2, H4S1P, H3K36me2, and  $\gamma$ H2AX samples were sequenced using Illumina HiSeq 2500 (single-end, 50 bp reads) at BGI (Beijing Genomics Institute, Hong Kong). FK2 samples were sequenced using Illumina HiSeq 2500 (single-end, 50 bp reads) at GATC biotech (Konstanz, Germany). XRCC4 at 1, 4, and 24 hr post 4OHT treatment, DNA ligase IV, H2BK120ac, H2AZ, H1, MacroH2A, H3K9me3, H4K20me2, 53BP1, and  $\gamma$ H2AX G1 and G2 samples were sequenced using Illumina NextSeq 500 (single-end, 75 bp reads) at EMBL Genomics core facilities (Heidelberg, Germany). H4K12ac, H2AZac, and H3 were sequenced using Illumina HiSeq 2500 (single-end, 50 bp reads) at EMBL Genomics

core facilities (Heidelberg, Germany). Previously published data include RAD51 and XRCC4 ChIP-seq in OHT-treated DlvA cells (Aymard et al., 2014), RNA Polymerase II (S2P) in untreated DlvA cells (Cohen et al., 2018) and MethylCap-Seq data in U2OS cells (Deplus et al., 2014)

The quality of each raw sequencing file (fastq) was verified with FastQC (<https://www.bioinformatics.babraham.ac.uk/projects/fastqc/>). All files were aligned to the reference human genome (hg19) and processed using a classical ChIP-seq pipeline: bwa (<http://bio-bwa.sourceforge.net/>) for mapping and samtools (<http://www.htslib.org/>) for duplicate removal (rmdup), sorting (sort) and indexing (index). Coverage for each aligned ChIP-seq dataset (.bam) were computed with the rtracklayer R package and normalized using total read count for each sample. Coverage data were exported as bigwig (file format) for further processing

### **BLESS data processing**

Samples were sequenced with HiSeq 2500 (Illumina) according to our custom protocol for low-diversity samples (Mitra et al., 2015), generating 61-bp paired-end reads containing BLESS barcodes (distal and proximal). Both barcodes were removed using cutadapt, and BLESS data were aligned on hg19 using bwa in paired-end mode (bwa aln and bwa sampe). In order to prevent reads that represent bona fide DSB signal, but start exactly at cleaved AsiSI sites from being marked as duplicates and improperly discarded, fragments were reconstituted from paired reads using Rsamtools and GenomicAlignments (R packages). Fragments with lengths > 500 bp were dropped as aberrant and the remaining fragments were de-duplicated (fragments with exact same start and end position on genome were considered as duplicates and only kept once). To determine which sites among the 1211 AsiSI sites were indeed cleaved, we used the total count for BLESS signal in a 1 kb windows around all AsiSI sites positions on the genome. Outliers (values > third quartile + 1.5 x interquartile range) were considered to be cleaved. Of the 174 sites defined as outliers, we only focused on 80 DSBs (showing the highest BLESS signal) that were significantly induced (Figures 1B–1D, Table S1).

### **Descriptive statistics using box-plot representation**

Each box-plot representation was generated with R-base. The center line represents the median, box ends represent respectively the first and third quartiles, and whiskers represent the minimum and maximum values without outliers. Outliers were defined as first quartile – (1.5 × interquartile range) and above third quartile + (1.5 × interquartile range).

### **Statistical analysis and representations**

Statistical hypothesis testing was performed using nonparametric unpaired Mann-Whitney-Wilcoxon (wilcoxon.test() function in R) to tests distribution differences between two populations. For damaged versus undamaged chromatin features comparisons, boxplots represent the ChIP-seq enrichment ratio between 4OHT treated and untreated DlvA cells (expressed as a log2 ratio) for DSBs or control regions. Significant differences were determined using two-sample Wilcoxon tests. Significant increases are colored in red or orange (for p value < 0.01 or p value < 0.05, respectively), significant decreases in dark or light blue (for p value < 0.01 or p value < 0.05, respectively) and non-significant differences (ns, p value > 0.05) in gray.

In the circle plot representation (Figures 4F and S5E), each circle radius represents the p value of a nonparametric two-sided Mann-Whitney-Wilcoxon test comparing ChIP-seq counts in a given window size. In Figures 4E and S5D, total ChIP-seq count for each dataset (untreated condition) was compared for HR and NHEJ DSBs. Individual modifications were determined as enriched in HR or NHEJ using one-sided Mann-Whitney-Wilcoxon tests (if p value < 0.05). In Figure S5E, total ChIP-seq count for each dataset was compared in treated and untreated conditions, separately for HR or NHEJ DSBs. Individual modifications were determined as increased or decreased following DSB induction using one-sided Mann-Whitney-Wilcoxon tests (p value < 0.05). In Figure S6B, Spearman correlation coefficients were determined using the cor() function in R and the correlation matrix was generated using the corplot package. For Figure S4B, to check if DSBs repaired by HR or NHEJ exhibit a non-random distribution among the main 3D genomic compartments (A1, A2, B1, B2, B3, B4 and NA (non-assigned)) (Rao et al., 2014)), we compared the distribution of all genomic loops among these compartments with the distributions for loops containing either HR-repaired or NHEJ-repaired DSBs (no loop contained both HR and NHEJ sites). The p values were calculated using hypergeometric probability distribution, with the null hypothesis that HR and NHEJ loops are distributed among the 3D compartments in the same manner as all genomic loops. Conservative Bonferroni correction for multiple hypothesis testing was applied to correct p values to reflect testing all 3D compartments for both possible enrichments and depletions.

### **Averaged ChIP-seq profiles**

Averaged ChIP-seq profiles were generated using the R package ggplot2. The x axis represents genomic position relative to DSB and the y axis represents the mean coverage at each bp, except for larger windows (1 Mb scale) where data were smoothed using a 50 kb span. Log2 ratio was computed using the bamCompare tool from deepTools (<http://deeptools.readthedocs.io>) with two bam files (before and after damage) as inputs. Positive and negative values for log2 ratio are respectively represented in red and blue. For γH2AX, 53BP1, RAD51, and XRCC4, ChIP-seq data are only available for damaged cells. In this case, the y axis only represents the mean ChIP-seq coverage in the damaged condition.

**Generation of random positions**

In order to generate a control set of non-DSB regions, we first computed a thousand random positions on the entire genome (excluding chromosome Y) using R. These random sites were filtered for being at least 1 Mb away from the top 150 cleaved positions with `IRanges::findOverlaps()` function. We also excluded regions with a null ChIP-seq count (in any of our datasets) in a 1 kb window in order to avoid regions that are systematically underrepresented in ChIP-seq experiments. Finally, 80 sites were randomly picked from the remaining list.

**Determination of HR-prone and NHEJ-prone DSBs**

We computed a ChIP-seq coverage ratio between RAD51 (4 kb window) and XRCC4 (1 kb window) for each of the 80 induced DSBs. Sites with the highest ratio were designated as HR-prone and sites with the lowest ratio as NHEJ-prone (30 DSBs in each category). For the analysis performed Figure S5D, HR-prone and NHEJ-prone DSBs were obtained by computing the ratio between RAD51 (4 kb window) and Ligase IV (1 kb window) for each of the 80 induced DSBs and selecting sites with highest and lowest ratios (30 DSBs in each category).

**DATA AND SOFTWARE AVAILABILITY**

The accession number for the high-throughput sequencing data reported in this paper is ArrayExpress: E-MTAB-5817.

Source code for generating boxplots, heatmaps, and average profiles is available on the GitHub repository page: <https://github.com/LegubeDNAREPAIR/HistoneMapping>.



# Bibliographie

- ALIPANAH, Babak et al. (août 2015). "Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning". In : *Nature Biotechnology* 33.8, p. 831-838. ISSN : 1546-1696. DOI : 10.1038/nbt.3300. URL : <https://doi.org/10.1038/nbt.3300>.
- ALLIS, C David et Thomas JENUWEIN (2016). "The molecular hallmarks of epigenetic control". In : *Nature Reviews Genetics* 17.8, p. 487-500.
- ALTSCHUL, S. F. et al. (oct. 1990). "Basic local alignment search tool". In : *J Mol Biol* 215.3, p. 403-410.
- ANDERS, Simon, Paul Theodor PYL et Wolfgang HUBER (sept. 2014). "HTSeq—a Python framework to work with high-throughput sequencing data". In : *Bioinformatics* 31.2, p. 166-169. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btu638. eprint : <https://academic.oup.com/bioinformatics/article-pdf/31/2/166/7000027/btu638.pdf>. URL : <https://doi.org/10.1093/bioinformatics/btu638>.
- ARNOULD, Coline et Gaëlle LEGUBE (2020). "The secret life of chromosome loops upon DNA double-strand break". In : *Journal of molecular biology* 432.3, p. 724-736.
- ATHAR, Awais et al. (2019). "ArrayExpress update—from bulk to single-cell expression data". In : *Nucleic acids research* 47.D1, p. D711-D715.
- AY, Ferhat et William S NOBLE (2015). "Analysis methods for studying the 3D architecture of the genome". In : *Genome biology* 16.1, p. 1-15.
- AYMARD, François, Marion AGUIRREBENGOA et al. (2017). "Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes". In : *Nature structural & molecular biology* 24.4, p. 353-361.
- AYMARD, François, Beatrix BUGLER et al. (2014). "Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks". In : *Nature structural & molecular biology* 21.4, p. 366-374.
- AYRAPETOV, Marina K et al. (2014). "DNA double-strand breaks promote methylation of histone H3 on lysine 9 and transient formation of repressive chromatin". In : *Proceedings of the National Academy of Sciences* 111.25, p. 9169-9174.
- BADER, Aldo S et al. (2020). "The roles of RNA in DNA double-strand break repair". In : *British journal of cancer* 122.5, p. 613-623.
- BAILEY, T. L., M. BODEN et al. (juill. 2009). "MEME SUITE : tools for motif discovery and searching". In : *Nucleic Acids Res* 37.Web Server issue, W202-208.
- BAILEY, T. L. et C. ELKAN (1994). "Fitting a mixture model by expectation maximization to discover motifs in biopolymers". In : *Proc Int Conf Intell Syst Mol Biol* 2, p. 28-36.
- BANIGAN, Edward J et Leonid A MIRNY (2020). "Loop extrusion : theory meets single-molecule experiments". In : *Current opinion in cell biology* 64, p. 124-138.

- BARBIERI, Mariano et al. (2012). "Complexity of chromatin folding is captured by the strings and binders switch model". In : *Proceedings of the National Academy of Sciences* 109.40, p. 16173-16178.
- BARRETT, Tanya et al. (2012). "NCBI GEO : archive for functional genomics data sets—update". In : *Nucleic acids research* 41.D1, p. D991-D995.
- BEDRAT, Amina, Laurent LACROIX et Jean-Louis MERGNY (2016). "Re-evaluation of G-quadruplex propensity with G4Hunter". In : *Nucleic acids research* 44.4, p. 1746-1759.
- BELIVEAU, Brian J et al. (2017). "In situ super-resolution imaging of genomic DNA with OligoSTORM and OligoDNA-PAINT". In : *Super-Resolution Microscopy*. Springer, p. 231-252.
- BERSAGLIERI, Cristiana et Raffaella SANTORO (2019). "Genome organization in and around the nucleolus". In : *Cells* 8.6, p. 579.
- BEUCHER, Andrea et al. (2009). "ATM and Artemis promote homologous recombination of radiation-induced DNA double-strand breaks in G2". In : *The EMBO journal* 28.21, p. 3413-3427.
- BOYLE, A. P. et al. (jan. 2008). "High-resolution mapping and characterization of open chromatin across the genome". In : *Cell* 132.2, p. 311-322.
- BOYLE, Elizabeth I. et al. (août 2004). "GO ::TermFinder—open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes". In : *Bioinformatics* 20.18, p. 3710-3715. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bth456. eprint : <https://academic.oup.com/bioinformatics/article-pdf/20/18/3710/522506/bth456.pdf>. URL : <https://doi.org/10.1093/bioinformatics/bth456>.
- BREIMAN, Leo (oct. 2001). "Random Forests". In : *Machine Learning* 45.1, p. 5-32. ISSN : 1573-0565. DOI : 10.1023/A:1010933404324. URL : <https://doi.org/10.1023/A:1010933404324>.
- BREIMAN, Leo et al. (oct. 2017). *Classification And Regression Trees*. Routledge. DOI : 10.1201/9781315139470. URL : <https://doi.org/10.1201/9781315139470>.
- BRUNNER, A. L. et al. (juin 2009). "Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver". In : *Genome Res* 19.6, p. 1044-1056.
- BUENROSTRO, J. D. et al. (déc. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". In : *Nat Methods* 10.12, p. 1213-1218.
- BURGE, Christopher Boyce (1997). *Identification of genes in human genomic DNA*. Stanford University.
- CARON, Pierre, Francois AYMARD et al. (2012). "Cohesin protects genes against γH2AX induced by DNA double-strand breaks". In : *PLoS genetics* 8.1, e1002460.
- CARON, Pierre, Jonathan CHOUDJAYE et al. (2015). "Non-redundant functions of ATM and DNA-PKcs in response to DNA double-strand breaks". In : *Cell reports* 13.8, p. 1598-1609.
- CASTRO-MONDAGON, Jaime Abraham et al. (juin 2017). "RSAT matrix-clustering : dynamic exploration and redundancy reduction of transcription factor binding motif collections". In : *Nucleic Acids Research* 45.13, e119-e119. ISSN : 0305-1048. DOI : 10.1093/nar/gkx314. eprint : <https://academic.oup.com/nar/article-pdf/45/13/e119/25367647/gkx314.pdf>. URL : <https://doi.org/10.1093/nar/gkx314>.
- CELESTE, Arkady et al. (2003). "Histone H2AX phosphorylation is dispensable for the initial recognition of DNA breaks". In : *Nature cell biology* 5.7, p. 675-679.
- CHAMBERS, Vicki S et al. (2015). "High-throughput sequencing of DNA G-quadruplex structures in the human genome". In : *Nature biotechnology* 33.8, p. 877-881.
- CHANG, Li-Hsin, Sourav GHOSH et Daan NOORDERMEER (2020). "TADs and their borders : free movement or building a wall ?" In : *Journal of molecular biology* 432.3, p. 643-652.

- CHÈNEBY, Jeanne et al. (2020). "ReMap 2020 : a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments". In : *Nucleic acids research* 48.D1, p. D180-D188.
- "Chromatin accessibility profiling methods" (jan. 2021). In : *Nature Reviews Methods Primers* 1.1, p. 11. ISSN : 2662-8449. DOI : 10.1038/s43586-020-00010-1. URL : <https://doi.org/10.1038/s43586-020-00010-1>.
- CLOUaire, T et G LEGUBE (2015). "DNA double strand break repair pathway choice : a chromatin based decision ?" In : *Nucleus* 6.2, p. 107-113.
- CLOUaire, Thomas et al. (2018). "Comprehensive mapping of histone modifications at DNA double-strand breaks deciphers repair pathway chromatin signatures". In : *Molecular cell* 72.2, p. 250-262.
- COHEN, Sarah et al. (2018). "Senataxin resolves RNA : DNA hybrids forming at DNA double-strand breaks to prevent translocations". In : *Nature communications* 9.1, p. 1-14.
- CONSORTIUM, ENCODE Project et al. (2012). "An integrated encyclopedia of DNA elements in the human genome". In : *Nature* 489.7414, p. 57.
- CONSORTIUM, International Human Genome Sequencing (oct. 2004). "Finishing the euchromatic sequence of the human genome". In : *Nature* 431.7011, p. 931-945.
- CONSORTIUM, The Gene Ontology (jan. 2008). "The Gene Ontology project in 2008". In : *Nucleic Acids Res* 36.Database issue, p. D440-444.
- COURNAC, Axel et al. (2012). "Normalization of a chromosomal contact map". In : *BMC genomics* 13.1, p. 1-13.
- CRAMER, Patrick (2019). "Organization and regulation of gene transcription". In : *Nature* 573.7772, p. 45-54.
- CRAMER, Paula et al. (1997). "Functional association between promoter structure and transcript alternative splicing". In : *Proceedings of the National Academy of Sciences* 94.21, p. 11456-11460.
- CRANE, E. et al. (juill. 2015). "Condensin-driven remodelling of X chromosome topology during dosage compensation". In : *Nature* 523.7559, p. 240-244.
- CROSETTO, Nicola et al. (2013). "Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing". In : *Nature methods* 10.4, p. 361-365.
- CUELLA-MARTIN, Raquel et al. (2016). "53BP1 integrates DNA repair and p53-dependent cell fate decisions via distinct mechanisms". In : *Molecular cell* 64.1, p. 51-64.
- D'HAESELEER, P. (avr. 2006). "What are DNA sequence motifs ?" In : *Nat Biotechnol* 24.4, p. 423-425.
- DAVIDSON, Iain F et al. (2019). "DNA loop extrusion by human cohesin". In : *Science* 366.6471, p. 1338-1345.
- DEKKER, Job et Edith HEARD (2015). "Structural and functional diversity of Topologically Associating Domains". In : *FEBS letters* 589.20, p. 2877-2884.
- DEVLIN, Jacob et al. (2018). "Bert : Pre-training of deep bidirectional transformers for language understanding". In : *arXiv preprint arXiv :1810.04805*.
- DING, Yang et al. (2017). "An exact transformation of convolutional kernels applied directly to DNA/RNA sequences". In : *bioRxiv*, p. 163220.
- DIXON, J. R. et al. (avr. 2012). "Topological domains in mammalian genomes identified by analysis of chromatin interactions". In : *Nature* 485.7398, p. 376-380.
- DOBIN, A. et al. (jan. 2013). "STAR : ultrafast universal RNA-seq aligner". In : *Bioinformatics* 29.1, p. 15-21.

- DUCHI, John, Elad HAZAN et Yoram SINGER (2011). "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In : *Journal of Machine Learning Research* 12.61, p. 2121-2159.  
URL : <http://jmlr.org/papers/v12/duchi11a.html>.
- DUQUETTE, Michelle L et al. (2004). "Intracellular transcription of G-rich DNAs induces formation of G-loops, novel structures containing G4 DNA". In : *Genes & development* 18.13, p. 1618-1629.
- DURAND, N. C., J. T. ROBINSON et al. (juill. 2016). "Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom". In : *Cell Syst* 3.1, p. 99-101.
- DURAND, N. C., M. S. SHAMIM et al. (juill. 2016). "Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments". In : *Cell Syst* 3.1, p. 95-98.
- ERDEL, Fabian et al. (2020). "Mouse heterochromatin adopts digital compaction states without showing hallmarks of HP1-driven liquid-liquid phase separation". In : *Molecular cell* 78.2, p. 236-249.
- EZOE, Sachiko (2012). "Secondary leukemia associated with the anti-cancer agent, etoposide, a topoisomerase II inhibitor". In : *International journal of environmental research and public health* 9.7, p. 2444-2453.
- FERIC, Marina et al. (2016). "Coexisting liquid phases underlie nucleolar subcompartments". In : *Cell* 165.7, p. 1686-1697.
- FISHER, R. A. (1936). "THE USE OF MULTIPLE MEASUREMENTS IN TAXONOMIC PROBLEMS". In : *Annals of Eugenics* 7.2, p. 179-188. DOI : <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>.  
eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1936.tb02137.x>.  
URL : <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-1809.1936.tb02137.x>.
- FORTIN, Jean-Philippe et Kasper D HANSEN (2015). "Reconstructing A/B compartments as revealed by Hi-C using long-range correlations in epigenetic data". In : *Genome biology* 16.1, p. 1-23.
- FRASER, James et al. (2015). "An overview of genome organization and how we got there : from FISH to Hi-C". In : *Microbiology and Molecular Biology Reviews* 79.3, p. 347-372.
- FUDENBERG, Geoffrey et al. (2016). "Formation of chromosomal domains by loop extrusion". In : *Cell reports* 15.9, p. 2038-2049.
- FYODOROV, Dmitry V et al. (2018). "Emerging roles of linker histones in regulating chromatin structure and function". In : *Nature reviews Molecular cell biology* 19.3, p. 192-206.
- GANJI, Mahipal et al. (2018). "Real-time imaging of DNA loop extrusion by condensin". In : *Science* 360.6384, p. 102-105.
- GARDINI, A. (2017). "Global Run-On Sequencing (GRO-Seq)". In : *Methods Mol Biol* 1468, p. 111-120.
- SELLERT, Martin, Marie N LIPSETT et David R DAVIES (1962). "Helix formation by guanylic acid". In : *Proceedings of the National Academy of Sciences of the United States of America* 48.12, p. 2013.
- GEORGULIS, Anastasios et al. (2017). "Genome instability and  $\gamma$ H2AX". In : *International journal of molecular sciences* 18.9, p. 1979.
- GHODKE, Indrajeet et al. (2021). "AHNAK controls 53BP1-mediated p53 response by restraining 53BP1 oligomerization and phase separation". In : *Molecular Cell*.
- GINNO, Paul A et al. (2012). "R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters". In : *Molecular cell* 45.6, p. 814-825.
- GOTHE, Henrike Johanna et al. (2019). "Spatial chromosome folding and active transcription drive DNA fragility and formation of oncogenic MLL translocations". In : *Molecular cell* 75.2, p. 267-283.
- GRÄFF, Johannes et Isabelle M MANSUY (2008). "Epigenetic codes in cognition and behaviour". In : *Behavioural brain research* 192.1, p. 70-87.

- GRANT, C. E., T. L. BAILEY et W. S. NOBLE (avr. 2011). "FIMO : scanning for occurrences of a given motif". In : *Bioinformatics* 27.7, p. 1017-1018.
- GRAY, Stephen et Paula E COHEN (2016). "Control of meiotic crossovers : from double-strand break formation to designation". In : *Annual review of genetics* 50, p. 175-210.
- GUPTA, S. et al. (2007). "Quantifying similarity between motifs". In : *Genome Biol* 8.2, R24.
- HAARHUIS, Judith HI et al. (2017). "The cohesin release factor WAPL restricts chromatin loop extension". In : *Cell* 169.4, p. 693-707.
- HÄNSEL-HERTSCH, Robert, Angela SIMEONE et al. (2020). "Landscape of G-quadruplex DNA structural regions in breast cancer". In : *Nature Genetics* 52.9, p. 878-883.
- HÄNSEL-HERTSCH, Robert, Jochen SPIEGEL et al. (2018). "Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing". In : *Nature protocols* 13.3, p. 551-564.
- HARTLERODE, Andrea J et Ralph SCULLY (2009). "Mechanisms of double-strand break repair in somatic mammalian cells". In : *Biochemical Journal* 423.2, p. 157-168.
- HELDEN, J. van (juill. 2003). "Regulatory sequence analysis tools". In : *Nucleic Acids Res* 31.13, p. 3593-3596.
- HINTON, Geoffrey (s. d.). *Neural Networks for Machine Learning*. [https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture\\_slides\\_lec6.pdf](https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf). (Accessed on 09/03/2021).
- HO, Daniel Sik Wai et al. (2019). "Machine learning SNP based prediction for precision medicine". In : *Frontiers in genetics* 10, p. 267.
- HOCHREITER, Sepp et Jürgen SCHMIDHUBER (1997). "Long short-term memory". In : *Neural computation* 9.8, p. 1735-1780.
- HON, Jiří et al. (2017). "pqsfinder : an exhaustive and imperfection-tolerant search tool for potential quadruplex-forming sequences in R". In : *Bioinformatics* 33.21, p. 3373-3379.
- HOWARD, FB, J FRAZIER et H Todd MILES (1977). "Stable and metastable forms of poly (G)". In : *Biopolymers : Original Research on Biomolecules* 16.4, p. 791-809.
- HOWE, Kevin L et al. (nov. 2020). "Ensembl 2021". In : *Nucleic Acids Research* 49.D1, p. D884-D891. ISSN : 0305-1048. DOI : 10.1093/nar/gkaa942. eprint : <https://academic.oup.com/nar/article-pdf/49/D1/D884/35364073/gkaa942.pdf>. URL : <https://doi.org/10.1093/nar/gkaa942>.
- HU, Jiazhi et al. (2015). "Chromosomal loop domains direct the recombination of antigen receptor genes". In : *Cell* 163.4, p. 947-959.
- HU, Ming et al. (sept. 2012). "HiCNorm : removing biases in Hi-C data via Poisson regression". In : *Bioinformatics* 28.23, p. 3131-3133. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bts570. eprint : <https://academic.oup.com/bioinformatics/article-pdf/28/23/3131/18529988/bts570.pdf>. URL : <https://doi.org/10.1093/bioinformatics/bts570>.
- HUPPERT, Julian L et Shankar BALASUBRAMANIAN (2005). "Prevalence of quadruplexes in the human genome". In : *Nucleic acids research* 33.9, p. 2908-2916.
- (2007). "G-quadruplexes in promoters throughout the human genome". In : *Nucleic acids research* 35.2, p. 406-413.
- HUSTEDT, Nicole et Daniel DUROCHER (2017). "The control of DNA repair by the cell cycle". In : *Nature cell biology* 19.1, p. 1-9.
- IACOVONI, Jason S et al. (2010). "High-resolution profiling of  $\gamma$ H2AX around DNA double strand breaks in the mammalian genome". In : *The EMBO journal* 29.8, p. 1446-1457.

- IANNELLI, Fabio et al. (2017). "A damaged genome's transcriptional landscape through multilayered expression profiling around in situ-mapped DNA double-strand breaks". In : *Nature communications* 8.1, p. 1-12.
- IMAKAEV, Maxim et al. (oct. 2012). "Iterative correction of Hi-C data reveals hallmarks of chromosome organization". In : *Nat Methods* 9.10, p. 999-1003.
- JI, Yanrong et al. (2021). "DNABERT : pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome". In : *Bioinformatics* 37.15, p. 2112-2120.
- JUMPER, John et al. (août 2021). "Highly accurate protein structure prediction with AlphaFold". In : *Nature* 596.7873, p. 583-589. ISSN : 1476-4687. DOI : 10.1038/s41586-021-03819-2. URL : <https://doi.org/10.1038/s41586-021-03819-2>.
- KAYA-OKUR, Hatice S et al. (2019). "CUT&Tag for efficient epigenomic profiling of small samples and single cells". In : *Nature communications* 10.1, p. 1-10.
- KELLEY, D. R., J. SNOEK et J. L. RINN (juill. 2016). "Basset : learning the regulatory code of the accessible genome with deep convolutional neural networks". In : *Genome Res* 26.7, p. 990-999.
- KEMPFER, R. et A. POMBO (avr. 2020). "Methods for mapping 3D chromosome architecture". In : *Nat Rev Genet* 21.4, p. 207-226.
- KILIC, Sinan et al. (2019). "Phase separation of 53 BP 1 determines liquid-like behavior of DNA repair compartments". In : *The EMBO journal* 38.16, e101379.
- KIM, D. et al. (avr. 2013). "TopHat2 : accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions". In : *Genome Biol* 14.4, R36.
- KIM, Jeongkyu et al. (2016). "Controlled DNA double-strand break induction in mice reveals post-damage transcriptome stability". In : *Nucleic acids research* 44.7, e64-e64.
- KIM, Seong-Tae, Bo XU et Michael B KASTAN (2002). "Involvement of the cohesin protein, Smc1, in Atm-dependent and independent responses to DNA damage". In : *Genes & Development* 16.5, p. 560-570.
- KIM, Yoori et al. (2019). "Human cohesin compacts DNA by loop extrusion". In : *Science* 366.6471, p. 1345-1349.
- KINGMA, Diederik P. et Jimmy BA (2017). *Adam : A Method for Stochastic Optimization*. arXiv : 1412.6980 [cs.LG].
- KLEMM, S. L., Z. SHIPONY et W. J. GREENLEAF (avr. 2019). "Chromatin accessibility and the regulatory epigenome". In : *Nat Rev Genet* 20.4, p. 207-220.
- KLIMENTOVA, Eva et al. (2020). "PENGUINN : Precise exploration of nuclear G-quadruplexes using interpretable neural networks". In : *Frontiers in Genetics* 11, p. 1287.
- KOLATHUR, Kiran Kumar (2021). "Role of promoters in regulating alternative splicing". In : *Gene*, p. 145523.
- KOSIOL, Nils et al. (2021). "G-quadruplexes : A promising target for cancer therapy". In : *Molecular Cancer* 20.1, p. 1-18.
- KRIETENSTEIN, Nils et al. (2020). "Ultrastructural details of mammalian chromosome architecture". In : *Molecular cell* 78.3, p. 554-565.
- KRIZHEVSKY, Alex, Ilya SUTSKEVER et Geoffrey E HINTON (2012). "ImageNet Classification with Deep Convolutional Neural Networks". In : *Advances in Neural Information Processing Systems*. Sous la dir. de F. PEREIRA et al. T. 25. Curran Associates, Inc. URL : <https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf>.

- KULAKOVSKIY, I. V. et al. (jan. 2018). "HOCOMOCO : towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis". In : *Nucleic Acids Res* 46.D1, p. D252-D259.
- LAGO, Sara et al. (2021). "Promoter G-quadruplexes and transcription factors cooperate to shape the cell type-specific transcriptome". In : *Nature Communications* 12.1, p. 1-13.
- LAJOIE, Bryan R, Job DEKKER et Noam KAPLAN (2015). "The Hitchhiker's guide to Hi-C analysis : practical guidelines". In : *Methods* 72, p. 65-75.
- LAL, Avantika et al. (mars 2021). "Deep learning-based enhancement of epigenomics data with Atac-Works". In : *Nature Communications* 12.1, p. 1507. ISSN : 2041-1723. DOI : 10.1038/s41467-021-21765-5. URL : <https://doi.org/10.1038/s41467-021-21765-5>.
- LANG, Fengchao et al. (2017). "CTCF prevents genomic instability by promoting homologous recombination-directed DNA double-strand break repair". In : *Proceedings of the National Academy of Sciences* 114.41, p. 10912-10917.
- LANGMEAD, B. et S. L. SALZBERG (mars 2012). "Fast gapped-read alignment with Bowtie 2". In : *Nat Methods* 9.4, p. 357-359.
- LANGMEAD, B., C. TRAPNELL et al. (2009). "Ultrafast and memory-efficient alignment of short DNA sequences to the human genome". In : *Genome Biol* 10.3, R25.
- LANS, Hannes, Jurgen A MARTEIJN et Wim VERMEULEN (2012). "ATP-dependent chromatin remodeling in the DNA-damage response". In : *Epigenetics & chromatin* 5.1, p. 1-14.
- LEE, Cheng-Sheng et al. (2016). "Chromosome position determines the success of double-strand break repair". In : *Proceedings of the National Academy of Sciences* 113.2, E146-E154.
- LEE, Young-Ho et al. (2013). "HP1 promotes tumor suppressor BRCA1 functions during the DNA damage response". In : *Nucleic acids research* 41.11, p. 5784-5798.
- LI, H. et R. DURBIN (juill. 2009). "Fast and accurate short read alignment with Burrows-Wheeler transform". In : *Bioinformatics* 25.14, p. 1754-1760.
- LI, H., B. HANDSAKER et al. (août 2009). "The Sequence Alignment/Map format and SAMtools". In : *Bioinformatics* 25.16, p. 2078-2079.
- LI, Lin et al. (2021). "YY1 interacts with guanine quadruplexes to regulate DNA looping and gene expression". In : *Nature chemical biology* 17.2, p. 161-168.
- LIAO, Yang, Gordon K. SMYTH et Wei SHI (nov. 2013). "featureCounts : an efficient general purpose program for assigning sequence reads to genomic features". In : *Bioinformatics* 30.7, p. 923-930. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btt656. eprint : <https://academic.oup.com/bioinformatics/article-pdf/30/7/923/633148/btt656.pdf>. URL : <https://doi.org/10.1093/bioinformatics/btt656>.
- LIEBERMAN-AIDEN, E. et al. (oct. 2009). "Comprehensive mapping of long-range interactions reveals folding principles of the human genome". In : *Science* 326.5950, p. 289-293.
- LIGHTFOOT, Helen L et al. (2019). "The diverse structural landscape of quadruplexes". In : *FEBS letters* 593.16, p. 2083-2102.
- LISBY, Michael, Uffe H MORTENSEN et Rodney ROTHSTEIN (2003). "Colocalization of multiple DNA double-strand breaks at a single Rad52 repair centre". In : *Nature cell biology* 5.6, p. 572-577.
- LIU, Yongchao et al. (2010). "CUDA-MEME : Accelerating motif discovery in biological sequences using CUDA-enabled graphics processing units". In : *Pattern Recognition Letters* 31.14, p. 2170-2177. ISSN : 0167-8655. DOI : <https://doi.org/10.1016/j.patrec.2009.10.009>. URL : <https://www.sciencedirect.com/science/article/pii/S0167865509002840>.

- LOTTERSBERGER, Francisca et al. (2015). "53BP1 and the LINC complex promote microtubule-dependent DSB mobility and DNA repair". In : *Cell* 163.4, p. 880-893.
- LOVE, M. I., W. HUBER et S. ANDERS (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". In : *Genome Biol* 15.12, p. 550.
- LUN, Aaron TL et Gordon K SMYTH (2015). "diffHic : a Bioconductor package to detect differential genomic interactions in Hi-C data". In : *BMC bioinformatics* 16.1, p. 1-11.
- MARNEF, Aline, Sarah COHEN et Gaëlle LEGUBE (2017). "Transcription-coupled DNA double-strand break repair : active genes need special care". In : *Journal of molecular biology* 429.9, p. 1277-1288.
- MARNEF, Aline et Gaëlle LEGUBE (2017). "Organizing DNA repair in the nucleus : DSBs hit the road". In : *Current opinion in cell biology* 46, p. 1-8.
- (2021). "R-loops as Janus-faced modulators of DNA repair". In : *Nature Cell Biology* 23.4, p. 305-313.
- MARTIRE, Sara et Laura A BANASZYNSKI (2020). "The roles of histone variants in fine-tuning chromatin organization and function". In : *Nature reviews Molecular cell biology* 21.9, p. 522-541.
- MCCULLOCH, Warren S et Walter PITTS (1943). "A logical calculus of the ideas immanent in nervous activity". In : *The bulletin of mathematical biophysics* 5.4, p. 115-133.
- McKUSICK, Victor A et Frank H RUDDLE (1987). *A new discipline, a new name, a new journal*.
- MCLEAY, Robert C et Timothy L BAILEY (2010). "Motif Enrichment Analysis : a unified framework and an evaluation on ChIP data". In : *BMC bioinformatics* 11.1, p. 1-11.
- MEISENBERG, Cornelia et al. (2019). "Repression of transcription at DNA breaks requires cohesin throughout interphase and prevents genome instability". In : *Molecular cell* 73.2, p. 212-223.
- MIECZKOWSKI, J. et al. (mai 2016). "MNase titration reveals differences between nucleosome occupancy and chromatin accessibility". In : *Nat Commun* 7, p. 11485.
- MINÉ-HATTAB, Judith et Rodney ROTHSTEIN (2012). "Increased chromosome mobility facilitates homology search during recombination". In : *Nature cell biology* 14.5, p. 510-517.
- MIRKIN, Sergei M et al. (2008). "Discovery of alternative DNA structures : a heroic decade (1979–1989)". In : *Front. Biosci* 13, p. 1064-1071.
- NAUMOVA, N. et al. (nov. 2012). "Analysis of long-range chromatin interactions using Chromosome Conformation Capture". In : *Methods* 58.3, p. 192-203.
- NAUMOVA, Natalia et al. (2013). "Organization of the mitotic chromosome". In : *Science* 342.6161, p. 948-953.
- NEEDLEMAN, S. B. et C. D. WUNSCH (mars 1970). "A general method applicable to the search for similarities in the amino acid sequence of two proteins". In : *J Mol Biol* 48.3, p. 443-453.
- NICHOLS, Michael H et Victor G CORCES (2021). "Principles of 3D compartmentalization of the human genome". In : *Cell reports* 35.13, p. 109330.
- NORA, Elphège P et al. (2017). "Targeted degradation of CTCF decouples local insulation of chromosome domains from genomic compartmentalization". In : *Cell* 169.5, p. 930-944.
- NUEBLER, Johannes et al. (2018). "Chromatin organization by an interplay of loop extrusion and compartmental segregation". In : *Proceedings of the National Academy of Sciences* 115.29, E6697-E6706.
- O'LEARY, N. A. et al. (jan. 2016). "Reference sequence (RefSeq) database at NCBI : current status, taxonomic expansion, and functional annotation". In : *Nucleic Acids Res* 44.D1, p. D733-745.
- OH, Dongpin et al. (2020). "CNN-Peaks : ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection". In : *Scientific reports* 10.1, p. 1-12.

- OKI, Shinya et al. (2018). "Ch IP-Atlas : a data-mining suite powered by full integration of public Ch IP-seq data". In : *EMBO reports* 19.12, e46255.
- OZA, Pranav et al. (2009). "Mechanisms that regulate localization of a DNA double-strand break to the nuclear periphery". In : *Genes & development* 23.8, p. 912-927.
- PANIER, Stephanie et Simon J BOULTON (2014). "Double-strand break repair : 53BP1 comes into focus". In : *Nature reviews Molecular cell biology* 15.1, p. 7-18.
- PARADA, Luis A, Philip G MCQUEEN et Tom MISTELI (2004). "Tissue-specific spatial organization of genomes". In : *Genome biology* 5.7, p. 1-9.
- PARKHOMCHUK, D. et al. (oct. 2009). "Transcriptome analysis by strand-specific sequencing of complementary DNA". In : *Nucleic Acids Res* 37.18, e123.
- PAULSEN, M. T. et al. (fév. 2013). "Coordinated regulation of synthesis and stability of RNA during the acute TNF-induced proinflammatory response". In : *Proc Natl Acad Sci U S A* 110.6, p. 2240-2245.
- Picard toolkit* (2019). <http://broadinstitute.github.io/picard/>.
- PIPIER, A et al. (2020). "Transcription-associated topoisomerase activities control DNA-breaks production by G-quadruplex ligands". In.
- PROROK, Paulina et al. (2019). "Involvement of G-quadruplex regions in mammalian replication origin activity". In : *Nature communications* 10.1, p. 1-16.
- PUGET, Nadine, Kyle M MILLER et Gaëlle LEGUBE (2019). "Non-canonical DNA/RNA structures during transcription-coupled double-strand break repair : roadblocks or bona fide repair intermediates ?" In : *DNA repair* 81, p. 102661.
- QUANG, Daniel et Xiaohui XIE (fév. 2014). "EXTREME : an online EM algorithm for motif discovery". In : *Bioinformatics* 30.12, p. 1667-1673. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btu093. eprint : <https://academic.oup.com/bioinformatics/article-pdf/30/12/1667/17345824/btu093.pdf>. URL : <https://doi.org/10.1093/bioinformatics/btu093>.
- RAO, S. S. et al. (déc. 2014). "A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping". In : *Cell* 159.7, p. 1665-1680.
- RAO, Suhas SP et al. (2017). "Cohesin loss eliminates all loop domains". In : *Cell* 171.2, p. 305-320.
- ROBERTSON, G. et al. (août 2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing". In : *Nat Methods* 4.8, p. 651-657.
- ROBINSON, Jenna et al. (2021). "DNA G-quadruplex structures : more than simple roadblocks to transcription ?" In : *Nucleic Acids Research*.
- ROBINSON, M. D., D. J. McCARTHY et G. K. SMYTH (jan. 2010). "edgeR : a Bioconductor package for differential expression analysis of digital gene expression data". In : *Bioinformatics* 26.1, p. 139-140.
- ROBINSON, M. D. et A. OSHLACK (2010). "A scaling normalization method for differential expression analysis of RNA-seq data". In : *Genome Biol* 11.3, R25.
- ROGAKOU, Emmy P et al. (1998). "DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139". In : *Journal of biological chemistry* 273.10, p. 5858-5868.
- RON, Gil et al. (2017). "Promoter-enhancer interactions identified from Hi-C data using probabilistic models and hierarchical topological domains". In : *Nature communications* 8.1, p. 1-12.
- ROSENBLATT, F. (1958). "The perceptron : A probabilistic model for information storage and organization in the brain." In : *Psychological Review* 65.6, p. 386-408. DOI : 10.1037/h0042519. URL : <https://doi.org/10.1037/h0042519>.

- ROUKOS, Vassilis et al. (2013). "Spatial dynamics of chromosome translocations in living cells". In : *Science* 341.6146, p. 660-664.
- SACOMOTO, G. A. et al. (avr. 2012). "KISSPLICE : de-novo calling alternative splicing events from RNA-seq data". In : *BMC Bioinformatics* 13 Suppl 6, S5.
- SAHAKYAN, Aleksandr B et al. (2017). "Machine learning model for sequence-driven DNA G-quadruplex formation". In : *Scientific reports* 7.1, p. 1-11.
- SANDELIN, A. et al. (jan. 2004). "JASPAR : an open-access database for eukaryotic transcription factor binding profiles". In : *Nucleic Acids Res* 32.Database issue, p. D91-94.
- SAVIC, Velibor et al. (2009). "Formation of dynamic  $\gamma$ -H2AX domains along broken DNA strands is distinctly regulated by ATM and MDC1 and dependent upon H2AX densities in chromatin". In : *Molecular cell* 34.3, p. 298-310.
- SCHEP, Alicia (2019). *motifmatchr : Fast Motif Matching in R*. R package version 1.8.0.
- SCHRANK, Benjamin R et al. (2018). "Nuclear ARP2/3 drives DNA break clustering for homology-directed repair". In : *Nature* 559.7712, p. 61-66.
- SELVARAJU, Ramprasaath R. et al. (oct. 2019). "Grad-CAM : Visual Explanations from Deep Networks via Gradient-Based Localization". In : *International Journal of Computer Vision* 128.2, p. 336-359. ISSN : 1573-1405. DOI : 10.1007/s11263-019-01228-7. URL : <http://dx.doi.org/10.1007/s11263-019-01228-7>.
- Séquençage : La technologie d'Illumina* (s. d.). [https://support.illumina.com/content/dam/illumina-support/courses/sequencing-illumina-technology-wbt-frca/story\\_html5.html](https://support.illumina.com/content/dam/illumina-support/courses/sequencing-illumina-technology-wbt-frca/story_html5.html)?iframe.
- SHANNON, C. E. (1948). "A Mathematical Theory of Communication". In : *Bell System Technical Journal* 27.3, p. 379-423. DOI : <https://doi.org/10.1002/j.1538-7305.1948.tb01338.x>. eprint : <https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.1538-7305.1948.tb01338.x>. URL : <https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1948.tb01338.x>.
- SHARMA, Cynthia M et Jörg VOGEL (2014). "Differential RNA-seq : the approach behind and the biological insight gained". In : *Current Opinion in Microbiology* 19. Ecology and industrial microbiology • Special Section : Novel technologies in microbiology, p. 97-105. ISSN : 1369-5274. DOI : <https://doi.org/10.1016/j.mib.2014.06.010>. URL : <https://www.sciencedirect.com/science/article/pii/S1369527414000800>.
- SHIN, Hanjun et al. (2016). "TopDom : an efficient and deterministic method for identifying topological domains in genomes". In : *Nucleic acids research* 44.7, e70-e70.
- SHRIKUMAR, Avanti, Peyton GREENSIDE et Anshul KUNDAJE (2017a). "Learning important features through propagating activation differences". In : *International Conference on Machine Learning*. PMLR, p. 3145-3153.
- (2017b). "Reverse-complement parameter sharing improves deep learning models for genomics". In : *bioRxiv*, p. 103663.
- SHRIKUMAR, Avanti, Katherine TIAN et al. (2018). "Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5. 6.5". In : *arXiv preprint arXiv :1811.00416*.
- SIDDIQUI-JAIN, Adam et al. (2002). "Direct evidence for a G-quadruplex in a promoter region and its targeting with a small molecule to repress c-MYC transcription". In : *Proceedings of the National Academy of Sciences* 99.18, p. 11593-11598.

- SIMS III, Robert J, Kenichi NISHIOKA et Danny REINBERG (2003). "Histone lysine methylation : a signature for chromatin function". In : *TRENDS in Genetics* 19.11, p. 629-639.
- SINKHORN, Richard et Paul KNOPP (1967). "Concerning nonnegative matrices and doubly stochastic matrices". In : *Pacific Journal of Mathematics* 21.2, p. 343-348.
- SKENE, Peter J et Steven HENIKOFF (2017). "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites". In : *Elife* 6, e21856.
- SLOMOVIC, Shimyn et al. (2010). "Addition of poly (A) and poly (A)-rich tails during RNA degradation in the cytoplasm of human cells". In : *Proceedings of the National Academy of Sciences* 107.16, p. 7407-7412.
- SMAILI, Fatima Zohra, Xin GAO et Robert HOEHNDORF (juin 2018). "Onto2Vec : joint vector-based representation of biological entities and their ontology-based annotations". In : *Bioinformatics* 34.13, p. i52-i60. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bty259. eprint : [https://academic.oup.com/bioinformatics/article-pdf/34/13/i52/25098469/bty259\\_\\\_hoehndorf.30.sup.1.pdf](https://academic.oup.com/bioinformatics/article-pdf/34/13/i52/25098469/bty259_\_hoehndorf.30.sup.1.pdf). URL : <https://doi.org/10.1093/bioinformatics/bty259>.
- SMEDLEY, Damian et al. (2009). "BioMart—biological queries made easy". In : *BMC genomics* 10.1, p. 1-12.
- SMITH, T. F. et M. S. WATERMAN (mars 1981). "Identification of common molecular subsequences". In : *J Mol Biol* 147.1, p. 195-197.
- SOULAS-SPRAUEL, P. et al. (déc. 2007). "V(D)J and immunoglobulin class switch recombinations : a paradigm to study the regulation of DNA end-joining". eng. In : *Oncogene* 26.56, p. 7780-7791. ISSN : 1476-5594. DOI : 10.1038/sj.onc.1210875.
- SRIVASTAVA, Nitish et al. (2014). "Dropout : A Simple Way to Prevent Neural Networks from Overfitting". In : *Journal of Machine Learning Research* 15.56, p. 1929-1958. URL : <http://jmlr.org/papers/v15/srivastava14a.html>.
- STORMO, G. D. (jan. 2000). "DNA binding sites : representation and discovery". In : *Bioinformatics* 16.1, p. 16-23.
- STROM, Amy R et al. (2017). "Phase separation drives heterochromatin domain formation". In : *Nature* 547.7662, p. 241-245.
- SZCZEPIŃSKA, Teresa, Anna Maria RUSEK et Dariusz PLEWCZYNSKI (2019). "Intermingling of chromosome territories". In : *Genes, Chromosomes and Cancer* 58.7, p. 500-506.
- TREMETHICK, David J (2007). "Higher-order structures of chromatin : the elusive 30 nm fiber". In : *Cell* 128.4, p. 651-654.
- VASWANI, Ashish et al. (2017). "Attention is all you need". In : *Advances in neural information processing systems*, p. 5998-6008.
- VIAN, Laura et al. (2018). "The energetics and physiological impact of cohesin extrusion". In : *Cell* 173.5, p. 1165-1178.
- WANG, James Z. et al. (mars 2007). "A new method to measure the semantic similarity of GO terms". In : *Bioinformatics* 23.10, p. 1274-1281. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btm087. eprint : <https://academic.oup.com/bioinformatics/article-pdf/23/10/1274/497100/btm087.pdf>. URL : <https://doi.org/10.1093/bioinformatics/btm087>.
- WANG, Z., M. GERSTEIN et M. SNYDER (jan. 2009). "RNA-Seq : a revolutionary tool for transcriptomics". In : *Nat Rev Genet* 10.1, p. 57-63.

- WATSON, James D et Francis HC CRICK (1953). "Molecular structure of nucleic acids : a structure for deoxyribose nucleic acid". In : *Nature* 171.4356, p. 737-738.
- WEIRAUCH, M. T. et al. (sept. 2014). "Determination and inference of eukaryotic transcription factor sequence specificity". In : *Cell* 158.6, p. 1431-1443.
- WERKEN, H. J. van de, G. LANDAN et al. (oct. 2012). "Robust 4C-seq data analysis to screen for regulatory DNA interactions". In : *Nat Methods* 9.10, p. 969-972.
- WERKEN, H. J. van de, P. J. de VREE et al. (2012). "4C technology : protocols and data analysis". In : *Methods Enzymol* 513, p. 89-112.
- WINGENDER, E. et al. (jan. 2000). "TRANSFAC : an integrated system for gene expression regulation". In : *Nucleic Acids Res* 28.1, p. 316-319.
- WIT, Elzo de et al. (2015). "CTCF binding polarity determines chromatin looping". In : *Molecular cell* 60.4, p. 676-684.
- WUTZ, Gordana et al. (2017). "Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins". In : *The EMBO journal* 36.24, p. 3573-3599.
- YATSKEVICH, Stanislau, James RHODES et Kim NASMYTH (2019). "Organization of chromosomal DNA by SMC complexes". In : *Annual Review of Genetics* 53, p. 445-482.
- YOKOCHI, Tomoki et al. (2009). "G9a selectively represses a class of late-replicating genes at the nuclear periphery". In : *Proceedings of the National Academy of Sciences* 106.46, p. 19363-19368.
- YU, Guangchuang et al. (2012). "clusterProfiler : an R package for comparing biological themes among gene clusters". In : *OMICS : A Journal of Integrative Biology* 16.5, p. 284-287. DOI : 10.1089/omi.2011.0118.
- YUAN, Jingsong, Rachel ADAMSKI et Junjie CHEN (2010). "Focus on histone variant H2AX : to be or not to be". In : *FEBS letters* 584.17, p. 3717-3724.
- ZHANG, Y. et al. (2008). "Model-based analysis of ChIP-Seq (MACS)". In : *Genome Biol* 9.9, R137.
- ZHANG, Yan et al. (fév. 2018). "Enhancing Hi-C data resolution with deep convolutional neural network HiCPlus". In : *Nature Communications* 9.1, p. 750. ISSN : 2041-1723. DOI : 10.1038/s41467-018-03113-2. URL : <https://doi.org/10.1038/s41467-018-03113-2>.
- ZHAO, Junhua et al. (2010). "Non-B DNA structure-induced genetic instability and evolution". In : *Cellular and Molecular Life Sciences* 67.1, p. 43-62.
- ZHENG, Ke-wei et al. (2020). "Detection of genomic G-quadruplexes in living cells using a small artificial protein". In : *Nucleic acids research* 48.20, p. 11706-11720.
- ZHOU, Vicky W, Alon GOREN et Bradley E BERNSTEIN (2011). "Charting histone modifications and the functional organization of mammalian genomes". In : *Nature Reviews Genetics* 12.1, p. 7-18.
- ZIMMER, Christophe et Emmanuelle FABRE (2019). "Chromatin mobility upon DNA damage : state of the art and remaining questions". In : *Current genetics* 65.1, p. 1-9.

**Résumé :** Les cassures double brin de l'ADN (DSB) sont des lésions délétères qui peuvent survenir sur le génome suite à une exposition à des agents génotoxiques, mais aussi de façon endogène, parmi lesquelles la formation de structures secondaires de l'ADN, telles que les G-quadruplexes (G4). Des méthodes computationnelles antérieures ont été développées pour prédire les G4 en fonction de motifs spécifiques, mais des approches récentes, basées sur le séquençage à haut débit ont permis d'identifier les G4 à l'échelle du génome. J'ai développé un nouveau modèle de Deep Learning pour prédire les régions G4 actives en utilisant les séquences d'ADN et l'accessibilité de la chromatine. En utilisant ce modèle, nous avons identifié de nouveaux motifs prédicteurs, y compris des facteurs de transcription connus qui pourraient réguler directement ou indirectement l'activité des G4. Nous avons également cartographié des milliers de régions G4 actives qui peuvent être utilisées dans le traitement du cancer pour identifier des cibles potentielles de médicaments récents à base de ligand-G4.

De plus, une fois induites sur le génome, les DSB déclenchent des modifications locales de la chromatine, comme la phosphorylation du variant d'histone H2AX ( $\gamma$ H2AX) par la kinase ATM, pour former des foyers de réparation à l'échelle du mégabase. Comment ces domaines sont formés pour permettre la signalisation rapide des DSB, et comment ces changements locaux de la chromatine sont gérés par la cellule n'est pas encore bien connu. Nous avons découvert que le recrutement des composants de réparation et la phosphorylation de H2AX sont régis par des domaines topologiques associatifs (TAD) préexistants. De plus, nous avons mis en évidence un processus d'extrusion de boucle unidirectionnel médié par le complexe de cohésine des deux côtés des DSB, ce qui permet la formation de foyers de réparation par ATM. Nous avons également découvert qu'à grande échelle, les DSB peuvent former un nouveau compartiment "D" de chromatine, composé de domaines de chromatine décorés par gH2AX, mais aussi de gènes de la réponse aux dommages à l'ADN (DDR), suggérant un rôle du regroupement des DSB dans l'activation de la DDR.

**Abstract :** DNA Double-strand breaks (DSBs) are harmful lesions that can occur on the genome following exposure to genotoxic agents but also due to endogenous causes, among which the formation of DNA secondary structures, such as G-quadruplexes (G4). Previous methods were developed to computationally predict G4s based on specific motifs, and recent Next Generation Sequencing approaches identified G4 distribution genome-wide. I developed a novel Deep learning model to predict active G4 regions using the DNA sequences and chromatin accessibility. Using this model, we found new motifs predictors including known transcription factors that could regulate directly or indirectly G4s activity. We also mapped thousand of active G4s regions that can be used in cancer therapy to identify potential targets of recent G4-ligand drugs.

Moreover, once induced on the genome, DSBs trigger local chromatin modifications including the phosphorylation of the H2AX histone variant ( $\gamma$ H2AX) by the ATM kinase, to form megabase-sized repair foci. How these domains are formed to enable rapid signaling of DSBs, and how these local chromatin changes are handled by the cell is still unclear. We found, that the recruitment of repair components and the phosphorylation of H2AX is governed by pre-existing Topologically Associating Domain (TADs). Moreover we discovered that an unidirectional loop-extrusion process mediated by the cohesin complex takes place on both side of the DSBs, which allow repair foci formation by ATM. We also found, at a global scale, that DSBs can form a novel "D" chromatin compartment, composed of gH2AX-decorated chromatin domains, but also of DNA damage responsive genes, suggesting a role of DSB clustering in activating the DNA Damage Response.