Data analyses and bioinformatic models for genomics

Raphael Mourad

UNIVERSITE PAUL SABATIER

ECOLE DOCTORALE BSB

Biologie Santé Biotechnologies

HABILITATION A DIRIGER DES RECHERCHES

Spécialité : Biologie Computationnelle

Présentée par

Raphaël MOURAD

Analyses de données et modèles bioinformatiques pour la génomique

Data analyses and bioinformatic models for genomics

défendue le 16/05/2022

Jury :

Rapporteurs :	Annick LESNE	DR CNRS	LPTMC
	Benoît BALLESTER	CR HDR INSERM	TAGC
	Romain KOSZUL	DR CNRS	Pasteur
Président :	Pierre NEUVIAL	DR CNRS	IMT
Examinateurs :	Gaëlle LEGUBE	DR CNRS	MCD
	Vera PANCALDI	CR HDR INSERM	CRCT

Acknowledgments

I want to dedicate my HDR thesis to my brother, who was a bright scholar in scholasticism, in particular Thomas Aquinas, and did not have the chance to pursue his exciting researches further. Not only, he was considered the most brilliant among us, but he was also a person with strong idealistic views and who was engaged in politics to help the most needy people. His passing is a great loss for his family and friends.

I also would like to thank my family from Toulouse for their amazing support during the past years, especially Sharon for offering me the best gift I ever received in my life, my little Antoine. I am also very grateful to all my friends for supporting me during these years in postdoc (all my friends from Indianapolis, Chicago and Montpellier) and later (all my friends from Toulouse).

I am very appreciative of my PhD supervisors Christine Sinoquet and Philippe Leray without whom I would not have pursued an academic career, my postdoc supervisors for their help, especially Lang Li, and currently, all Gaelle Legube's team members, and all SaAB team members. I also believe that my high school teachers of biology, chemistry and physics were the best promoters of science I have ever met so far, and I will never forget their incredible lectures. Moreover, I am also very grateful to all the Master and PhD students I supervised, including Alexandre, Vincent, Matthieu, Julien, Elissar, Martin and Nolan, and who greatly contributed to research projects. Lastly, I would like to thank warmly all my university colleagues, especially Patrice, David, Emmanuelle, Noëlie, Sylvia, Maryelle, who helped me to be a better teacher and with whom I spent incredible time.

Summary

Following the sequencing of the human genome in 2001, there has been an explosion of novel high-throughput sequencing projects to interrogate the genome and its functions, opening the so-called postgenomic era. Nowadays, experimental labs generate terabytes of heterogeneous data, necessitating the development of novel statistical and bioinformatic methods and models to process such big data, as well as to make sense of the wide variety of experimental results.

For the last 10 years, I have been investigating on a large number of postgenomic topics, ranging from human genetics in asthma to phylogenetics of HIV virus, transcription, chromatin, DNA secondary structures and DNA repair. This thesis presents my research efforts on both the analysis of biological data, and the development of novel statistical and computational models.

In the first chapter, I introduce the different topics, such as DNA, chromatin, postgenomic methods, human genetics and computational biology. In the second chapter, I then describe my different contributions in data analysis, including the discovery of rare variants associated with increased asthma risk, the role of drug-naïve HIV-positive patients in transmitting antiretroviral resistance, the global 3D genome reorganization due to hormone induction and the link between chromatin loop extrusion and DNA repair. I also present different statistical models to identify genomic factors in 1D that shape the genome in 3D, but also novel models for 3D domain identification, differential analysis and predictions. Moreover, I present machine/deep learning approaches for predicting DNA double-stranded breaks and active G-quadruplexes (G4s).

Finally, in the last chapter, I discuss about my future research projects, focusing on new deep learning models for predicting chromatin data across species, biophysical experiments to characterize G4 SNPs, the identification of non-coding SNPs as drivers of genome instability, and artificial intelligence for personalized medicine.

Keywords: Computational Biology; Artificial Intelligence; Deep Learning; Regulatory Genomics; 3D Genome; DNA Repair, G-quadruplex.

Résumé

Suite au succès du séquençage du génome humain en 2001, une explosion de nouveaux projets de séquençage à haut débit a eu lieu afin d'interroger le génome et ses fonctions, ouvrant la voie à l'ère "postgénomique". De nos jours, les laboratoires génèrent des téraoctets de données hétérogènes, ce qui nécessite le développement de nouvelles méthodes et modèles statistiques et bioinformatiques pour traiter de telles données volumineuses, ainsi que pour donner un sens à la grande variété de résultats expérimentaux.

Au cours des 10 dernières années, j'ai étudié un grand nombre de sujets postgénomiques, allant de la génétique humaine dans l'asthme à la phylogénétique du virus VIH, la transcription, la chromatine, les structures secondaires de l'ADN et la réparation de l'ADN. Cette thèse présente mes efforts de recherche sur l'analyse de données biologiques et le développement de nouveaux modèles statistiques et informatiques.

Dans le premier chapitre, j'introduis les différents sujets, tels que l'ADN, la chromatine, les méthodes postgénomiques, la génétique humaine et la biologie computationnelle. Dans le deuxième chapitre, je décris ensuite mes différentes contributions à l'analyse de données, dont la découverte de variants rares associés à un risque accru à l'asthme, le rôle des patients séropositifs n'ayant jamais été médicamentés dans la transmission de la résistance antirétrovirale, la réorganisation globale du génome en 3D suite à une induction hormonale et le lien entre l'extrusion de la boucle de la chromatine et la réparation de l'ADN. Je présente également différents modèles statistiques pour identifier les facteurs génomiques en 1D qui façonnent le génome en 3D, mais aussi de nouveaux modèles pour l'identification de domaines 3D, leur analyse différentielle et leur prédiction. De plus, je présente des approches d'apprentissage automatique et profond pour prédire les cassures double brin de l'ADN et les G-quadruplexes (G4) actifs.

Enfin, dans le dernier chapitre, je discute de mes futurs projets de recherche, en particulier de nouveaux modèles d'apprentissage en profondeur pour prédire les données de chromatine entre les espèces, des expériences biophysiques pour caractériser les SNP de G4, l'identification de SNP non codants en tant que moteurs de l'instabilité du génome et l'intelligence artificielle pour la médecine personnalisée.

Mots clés : Biologie computationnelle; Intelligence artificielle; Apprentissage profond; Génomique régulatrice ; Génome en 3D ; Réparation de l'ADN, G-quadruplexe.

List of Acronyms

ARV	AntiRetroViral
ATAC-seq	Assay for Transposase-Accessible Chromatin sequencing
ANOVA	ANalysis Of VAriance
BLESS	Breaks Labeling, Enrichment on Streptavidin and next-generation Sequencing
BLISS	Breaks Labeling In Situ and Sequencing
CBI	Centre de Biologie Intégrative
cDNA	complementary DeoxyriboNucleic Acid
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag sequencing
ChIP-seq	Chromatin ImmunoPrecipitation sequencing
CNN	Convolutional Neural Network
DIM	Differential Insulation Model
DNA	DeoxyriboNucleic Acid
DNase-seq	DNase I hypersensitive sites sequencing
DSB	DNA double Strand Break
dsbSNP	DNA double strand break Single Nucleotide Polymorphism
END-seq	DNA end sequencing
ER	Estrogen Receptor
eSNP	expression Single Nucleotide Polymorphism
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Element sequencing
GLM	Generalized Linear Model
GLMI	Generalized Linear Model with Interactions
GPU	Graphics Processing Unit
GRO-seq	Global Run-On sequencing
GWAS	Genome-Wide Association Study
G4	G-quadruplex
G4-seq	G-quadruplex sequencing
G4SNP	G-quadruplex Single Nucleotide Polymorphism
HGP	Human Genome Project
Hi-C	High-throughput chromosome conformation Capture
HIV	Human Immunodeficiency Virus
HR	Homologous Recombination
ICGC	International Cancer Genome Consortium
INRAE	Institut National de Recherche pour l'Agriculture, l'alimentation et l'Environnement
IPBS	Institut de Pharmacologie et de Biologie Structurale
LSTM	Long Short-Term Memory
MIAT	unité Mathématiques et Informatique Appliquées de Toulouse
mRNA	messenger RiboNucleic Acid
MRI	Magnetic Resonance Imaging
NGS	Next-Generation Sequencing
NHEJ	Non-Homologous End Joining
PIM	Prediction Insulation Model
RNA	RiboNucleic Acid

RNA-seq	RiboNucleic Acid sequencing
RNN	Recurrent Neural Network
ROS	Reactive Oxygen Species
SIM	Sparse Insulation Model
SNP	Single Nucleotide Polymorphism
sRNA-seq	small RiboNucleic Acid sequencing
STR	Short Tandem Repeat
TAD	Topologically Associating Domain
TCGA	The Cancer Genome Atlas
TFBS	Transcription Factor Binding Site
TSS	Transcription Starting Site
ZI	Zero-Inflated
3DR	3D Ratio
4C-seq	Circular Chromatin Conformation Capture

Contents

Chapter 1

Curriculum Vitæ

1.1 Professional experience

2021-now **Visiting Prof. at SaAB team, MIAT lab, INRAE**
Research: Deep Learning for genomics.

2018-2021 **Assist. Prof. at Univ. Toulouse III, MCD, CNRS, UMR 5077**
Research: Bioinformatics/Machine Learning of chromatin and DNA repair.
Team: Legube.
Teaching: Biostatistics/bioinformatics of omic data.

2014-2018 **Assist. Prof. at Univ. Toulouse III, LBME, CNRS, UMR 5099**
Research: Computational models and approaches to identify
molecular determinants of gene expression and 3D chromatin.
Teaching: Biostatistics/bioinformatics of omic data.

2013-2014 **Postdoc. at LIRMM, CNRS UMR 5506, Montpellier**
Research: Bioinformatic approach to study drug-resistant HIV viruses.
Supervision: Olivier Gascuel.

2012-2013 **Postdoc. at University of Chicago, USA**
Research: Human genetic study of asthmatic patients (GWAS).
Supervision: Dan Nicolae and Carole Ober.

2011-2012 **Postdoc. at Indiana University, USA**
Research 1: Study of 3D chromatin by analysis of Hi-C data
Research 2: Identification of biological markers using time series gene expression.
Supervision: Lang Li.

2008-2011 **PhD in computer science at University of Nantes**
Supervision: Philippe Leray and Christine Sinoquet.
Award : thèse remarquable de l'Université de Nantes.

2008-2011 **Lecturer at University of Nantes**
Degrees: Ingénieur/Master 2 Bioinfo/Master 1 data mining,
Teaching (113h eq TD): data mining, machine learning, bioinformatics,
biostatistics, omic data.

1.2 Publications

1.2.1 In submission (2 articles)

[1] Coline Arnould, **Vincent Rocher**, Aldo Bader, Emma Lesage, Nadine Puget, Thomas Clouaire, **Raphaël Mourad**, Daan Noordemeer, Martin Bushell and Gaëlle Legube. ATM-dependent formation of a novel chromatin compartment regulates the response to DNA double strand breaks and the biogenesis of translocations.

[2] Cyril Esnault, Encarnacion Garcia-Oliver, Amal Zine El Aabidine, Eugénia Basyuk, Alja Kozulic-Pirher, Magdalena Karpinska, Marie-Cécile Robert, Alexia Pigeot, Yu Luo, Daniele Verga, **Raphael Mourad**, Jean-Louis Mergny, Edouard Bertrand and Jean-Christophe Andrau. G-quadruplexes are promoter elements controlling nucleosomes exclusion and RNA Polymerase II pausing.

1.2.2 As assistant professor (11 articles, 2014-now)

[1] **Raphael Mourad**. TADreg : A versatile regression framework for TAD identification, differential analysis and prediction, 23(1):82, 2022. **BMC Bioinformatics**.

[1] Sarah Cohen, Aude Guenolé, Aline Marnef, Thomas Clouaire, Nadine Puget, **Vincent Rocher**, Coline Arnould, Marion Aguirrebengoa, Matthieu Genais, Dipti Vernekar, **Raphaël Mourad**, Valérie Borde, Gaëlle Legube. A POLD3/BLM dependent pathway handles DSBs in transcribed chromatin upon excessive RNA:DNA hybrid accumulation, 13(1):2012, 2022. **Nature Communications**.

[2] Coline Arnould, **Vincent Rocher**, Thomas Clouaire, Pierre Caron, Philippe. E. Mangeot, Emiliano. P. Ricci, **Raphaël Mourad**, Daan Noordermeer, Gaëlle Legube. Loop extrusion as a mechanism for DNA double-strand breaks repair foci formation, 590(7847):660-665, 2021. **Nature**.

[3] **Vincent Rocher**, Matthieu Genais, Elissar Nassereddine and **Raphaël Mourad**. DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions, 17(8):e1009308, 2021. **PLoS Computational Biology**.

[4] **Raphaël Mourad**. Studying 3D genome evolution using genomic sequence, 36(5):1367-1373, 2019. **Bioinformatics**.

[5] **Raphaël Mourad**, Krzysztof Ginalski, Gaëlle Legube and Olivier Cuvier. Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution, 19:34, 2018. **Genome Biology**.

[6] **Raphaël Mourad** and Olivier Cuvier. TAD-free analysis of architectural proteins and insulators, 46(5):e27, 2018. **Nucleic Acids Research**.

[7] David Umlauf and **Raphaël Mourad**. From fundamental principles to disease and cancer, 90:128-137, 2018. **Seminars in cell & developmental biology**.

[8] **Raphaël Mourad**, Lang Li et Olivier Cuvier. Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach, 13(5):e1005538, 2017. **PLoS Computational Biology**.

[9] **Raphaël Mourad** and Olivier Cuvier. Computational identification of

genomic features that influence 3D chromatin domain formation. **PLoS Computational Biology**, 12(5):e1004908, 2016.

[10] **Raphaël Mourad** and Olivier Cuvier. Predicting the spatial organization of chromosomes using epigenetic data. **Genome Biology**, 16(1):182, 2015.

1.2.3 As postdoc and PhD student (9 articles, 2008-2014)

[1] **Raphaël Mourad**, *et al.*, Olivier Gascuel, Stéphane Hué on behalf of the UK HIV Drug Resistance Database & the Collaborative HIV and Anti-HIV Drug Resistance Network. A phylotype-based analysis highlights the role of drug-naïve HIV positive individuals in the transmission of antiretroviral resistance in the United Kingdom. **AIDS**, 29(15):1917-1925, 2015.

[2] Catherine Iguartua, et al. Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma. **Nature Communications**, 6:5965, 2015.

[3] **Raphaël Mourad**, *et al.* and Lang Li. Estrogen induces global reorganization of chromatin structure in human breast cancer cells. **PLoS ONE**, 9(12):e113354, 2014.

[4] **Raphaël Mourad**, Christine Sinoquet, Nevin L. Zhang, Tengfei Liu and Philippe Leray. A Survey on Latent Tree Models and Applications. **Journal of Artificial Intelligence Research**, 47:157-203, 2013.

[5] **Raphaël Mourad***, Pengyue Zhang, Yang Xiang, Kun Huang, Tim Huang, Kenneth Nephew, Yunlong Liu and Lang Li. A dynamic time order network for time-series gene expression data analysis. **BMC System Biology**, 6(Suppl3):S9, 2012. *Co-first author with Pengyue Zhang

[6] Vittorio Perduca, Christine Sinoquet, **Raphaël Mourad** and Gregory Nuel. Alternative methods for H1 simulations in genome wide association studies. **Human Heredity**, 73(2):95-104, 2012.

[7] **Raphaël Mourad**, Christine Sinoquet and Philippe Leray. Probabilistic graphical models for genetic association studies. **Briefings in Bioinformatics**, 13(1):20-33, 2012.

[8] **Raphaël Mourad**, Christine Sinoquet, Christian Dina and Philippe Leray. Visualization of pairwise and multilocus linkage disequilibrium structure using latent forests. **PLoS ONE**, 6(12): e27320, 2011.

[9] **Raphaël Mourad**, Christine Sinoquet and Philippe Leray. A hierarchical Bayesian network approach for linkage disequilibrium modelling and data-dimensionality reduction prior to genome-wide association studies. **BMC Bioinformatics**, 12:16, 2011.

1.2.4 Scientific book (1 book)

[1] Christine Sinoquet, **Raphaël Mourad**. Probabilistic Graphical Models for Genetics, Genomics and Postgenomics. **Oxford University Press**, 2014.

1.3 Academic activities

1.3.1 Current collaborations

- Ivan Kulakovskiy (Institute of Protein Research RAS, Russia);
- Gaelle Legube (MCD, Toulouse);
- Catherine Tardin (IBPS, Toulouse);
- Jean-Christophe Andrau (IGMM, Montpellier);
- Monsef Benkirane (IGH, Montpellier);
- Lang Li (CCBB, Indianapolis, USA).

1.3.2 Funding

- INRAe (funding to reduce teaching load in 2021): 11k€;
- GSO (co-PI with C. Tardin) : 2.5 k€x 2 = 5 k€;
- 2-year-Master-internship program (apprentice) from CNRS: 36 k€;
- CNRS défi modélisation du vivant : 24 k€x 2 = 48 k€;
- MRT PhD funding : 100 k€;
- CRCT Univ. Toulouse III (funding to reduce teaching load in 2018): 10k€;
- IDEX starting grant Univ. Toulouse III: 14k€;
- ANRS pays du sud: 2 yrs postdoc (declined because hired as Assist. Prof.).

1.3.3 Teaching

I am supervising the teaching unit Biological data analysis (Master Biochemistry), and the teaching unit Bioinformatics for postgenomics (Master Bioinformatics).

Teaching	**Degree**	**Volume** (hours)
Biostatics	Master 1 Biochemistry	90 h per yr
Bioinformatics for NGS	Master 1 Bioinformatics	50 h per yr
Intro. to bioinfo.	Master 1 Biohealth	26 h per yr
GWAS	Master 1 Biohealth	8 h per yr
Intro. to bioinfo.	Bachelor 2/3 Bio	6 h per yr
Statistics	Engineering Master	46 h (PhD)
Bioinfo/Data mining	Master 2 Bioinformatics	37.33 h (PhD)
Probability	Master Erasmus Mundus DMKM	18 h (PhD)

1.3.4 INSERM courses

- Organization committee of Atelier Inserm "Machine Learning from Biology to Health" (2021, Bordeaux).
- Teaching deep learning for genomics in R (1 day and half) (2021, Bordeaux).

1.3.5 Other courses

- Teaching deep learning for genomics in R (1 day and half) for the Platform Biostatistics of Toulouse (2021).

1.3.6 Organization of seminars

In october 2021, I launched a new online seminar on deep learning called DeepBioHealth (https://groupes.renater.fr/sympa/info/deepbiohealth). The seminar aims to facilitate exchanges between scientists interested in deep learning and its applications in the fields of biology, health and agronomy. This is an interdisciplinary Toulouse working group that addresses both new models of deep learning and their recent applications in genomics, medical imaging, oncology, agronomy, etc.

1.3.7 Scientific communication

- DECLICS: Dialogues Entre Chercheurs et Lycéens pour les Intéresser à la Construction des Savoirs (2018).

1.3.8 Supervision

- 1 PhD student: Sébastien Ober (nov 2021-now);
- 1 PhD student: Vincent Rocher (2018-2021);
- 1 PhD student with O. Cuvier: Alexandre Heurteau (2016-2020);
- 2 Master students (2015); 2 Master students (2018); 1 Master student (2020); 1 Master student apprentice (2019-2021); 2 Master students (2021).

1.3.9 Scientific committee

- INSERM workshop in Bordeaux (2021): Introduction to Machine Learning from Biology to Health;
- SeqBIM workshop in Toulouse (2020);
- JOBIM workshop on deep learning in Montpellier (2020);
- Program committee at Intelligent Systems for Molecular Biology conference (rank A conference);
- Scientific Advisory Board of Bioinformatics platform of Centre de Biologie Intégrative;
- PhD thesis jury: 3 students (2018, 2021);
- PhD thesis committee: 2 students (2016).

1.3.10 University committee

- Computers and software for teaching committee of Univ. Toulouse III;

1.3.11 Reviewer

- Nature Structural & Molecular Biology, Genome Research, Genome Biology, Nature Communications, BMC Biology, Bioinformatics, NAR GB, BMC Bioinformatics, BMC Genomics, IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), International Journal of Approximate Reasoning, The Biometrical Journal, PLoS ONE, and many more.
- Junior group leader call at Centre de Biologie Intégrative.

1.3.12 Editor

Book on Probabilistic Graphical Models for Genetics, Genomics, and Postgenomics, Oxford University Press.

1.3.13 Talks

- 10/11/2021: Talk at DeepBioHealth, online.
- 01/10/2021: Talk at MIAT lab, INRAE, Toulouse.
- 14/09/2021: Invited talk at 100 Years of Genome Research 2021, Naples, Italy.
- 30/03/2021: Journal club of bioinfo at CBI, Toulouse.
- 28/06/2020: Symposium of deep learning for genomics, JOBIM, online.
- 06/03/2020: Talk at CBI, Toulouse.
- 11/02/2020: Talk at CNRS, Paris, to present results from funded project of CNRS modelisation du vivant program.
- 17/06/2019: Talk at Chrocogen, INRAe, Toulouse.
- 15/05/2019: Talk at Hi-C days, Toulouse.
- 05/05/2019: Talk at Biopuces, INRAe, Toulouse.
- 02/03/2019: Invited talk at IGFL, Lyon.
- 11/10/2019: Genotoul Biostat Bioinfo Day 2019, Toulouse, France.
- 10/04/2018: Conference Rencontres Scientifiques des Grandes Causses, GDR ADN, Millau, France.
- 18/01/2018: Workshop Biostat Bioinfo 2018, Toulouse, France.
- 11/01/2018: Workshop on Statistical Methods for Post Genomic Data (SMPGD), 2018, Montpellier. Poster.
- 15/11/2018: Invited talk at IGMM, Montpellier.
- 11/12/2017: Young Scientists Workshop - Genome Dynamics and Cancer, 2017, Montauban, France.
- 17/10/2017: Symposium Modelling Pathological Processes: from Molecules to Populations, 2017, Toulouse, France.
- 06/07/2017: Conference JOBIM 2017, Lille, France.
- 15/06/2017: Conference MCEB 2017, Porquerolles, France. Poster.
- 02/12/2016: Genotoul Biostat Bioinfo Day 2016, Toulouse, France.

- 14/10/2016: Conference CARTABLE 2016, Toulouse, France.
- 05/07/2016: Conference ICACG 2016, Toulouse, France.
- 28/06/2016: Conference JOBIM 2016, Lyon, France.
- 07/06/2016: Talk at SaAb team, MIAT, INRA, Toulouse, France.
- 26/02/2016: Talk at MAB team, LIRMM, Montpellier, France.
- 30/11/2015: Talk at CBI, CNRS/Université Paul Sabatier, Toulouse, France.
- 19/11/2015: Talk at IMT, Université Paul Sabatier, Toulouse, France.
- 28/09/2015: Conference Rencontres Scientifiques des Grandes Causses, GDR ADN, Millau, France.
- 03/07/2015: Talk at MIAT, INRA, Toulouse, France.

1.4 Industrial activities

1.4.1 Artificial intelligence for spine surgery

Since 2020, I collaborate with a new start-up in personalized medicine called RemedyLogic, based in New-York USA (https://remedylogic.com/). RemedyLogic is a company that helps insurance companies, self-insured employers, and patients to improve outcomes and reduce the cost of back surgery. With the company, I work on clinical artificial intelligence R&D, in particular for the development of novel machine learning and AI models to recommend spinal surgery and alternative conservative treatments such as physical therapy or medication.

Chapter 2

Introduction

Sommaire

> Science has discovered that, like any work of literature, the human genome is a text in need of commentary, for what Eliot said of poetry is also true of DNA: 'all meanings depend on the key of interpretation.' What makes us human, and what makes each of us his or her own human, is not simply the genes that we have buried into our base pairs, but how our cells, in dialogue with our environment, feed back to our DNA, changing the way we read ourselves. Life is a dialectic.
>
> Jonah Lehrer, Proust Was a Neuroscientist

Computational biology is an interdisciplinary science at the crossroad between biology, computer science and mathematics. It can be defined as the science of using mathematical models, algorithms, and large computing resources together with complex biological experimental data to understand biological systems and relationships, that could be out-of-reach otherwise. Computational biology has many applications in science, including genomics, but also, evolution, biomodeling, neuroscience, structural biology and pharmacology.

This chapter is an attempt to introduce in a concise manner the very diverse concepts useful to understand computational biology applied to genomics. The chapter starts by presenting biological concepts, such as the DNA molecule and its forms, the transcription and chromatin, 3D genome folding, genome stability and DNA repair. Then, genomics and omic technologies and data are introduced, and illustrated with commonly used techniques for the study of transcription and chromatin, *i.e.* RNA-seq, ChIP-seq and Hi-C experiments. Lastly, the chapter presents computational and mathematical fields, including big data, statistics, machine and deep learning, heterogeneous data integration and personalized medicine.

2.1 DNA and G-quadruplex

DNA is a complex molecule carrying the instructions an organism needs to develop, live and reproduce. DNA is composed of two complementary and antiparallel strands (i.e. in opposite directions) facing each other and forming a double helix [Watson & Crick 1953] (Figure 2.1A). Each strand is a polymer (or sequence) of nucleotides. Each nucleotide is made up of 3 molecules: one molecule of phosphoric acid, one molecule of deoxyribose and a nitrogen base. Four different bases exist: adenine (A), guanine (G), cytosine (C) and thymine (T). Thus, the

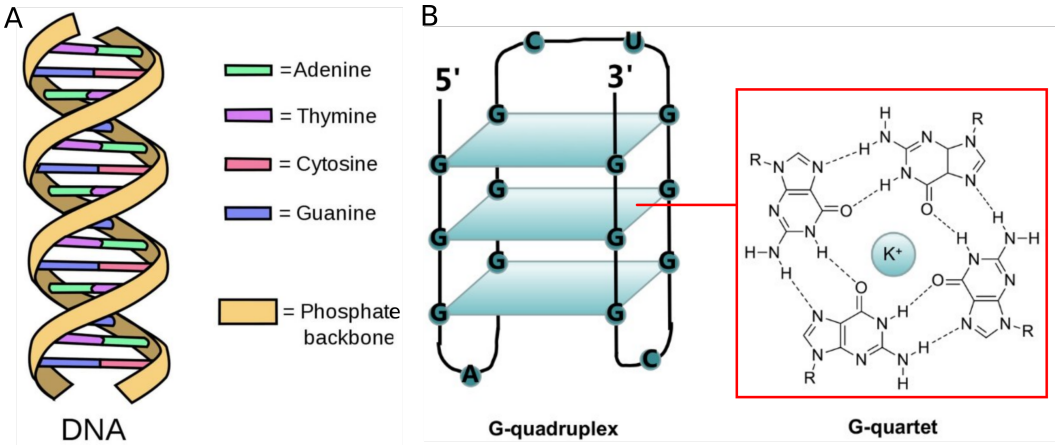


Figure 2.1: Deoxyribonucleic acid molecule (DNA). A) DNA molecule, presented in its most common form (the B form). B) G-quadruplex of DNA, an example of non-B DNA form.

genome constitutes a code formed from a 4-letter alphabet. In human, the size of the genome is large, around 3.4 billion base pairs, which gives it great complexity.

The B form of DNA (B DNA) is believed to predominate in cells [Watson & Crick 1953]. Yet, more than 20 non-B DNA structures have also been reported in the genome [Georgakopoulos-Soares *et al.* 2018]. Among those structures, the G-quadruplex (G4) was discovered in the late 80's [Sen & Gilbert 1988] (Figure 2.1B). G4 sequence contains four continuous stretches of guanines [Chen & Yang 2012]. Four guanines can be held together by Hoogsteen hydrogen bonding to form a square planar structure called a guanine tetrad (G-quartet). Two or more G-quartets can stack to form a G4 [Chen & Yang 2012]. 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 [Bhattacharyya *et al.* 2016]. Numerous works suggest that non-B DNA structures can regulate several essential processes in the cell, such as gene transcription, DNA replication, telomere stability and V(D)J recombination [Spiegel *et al.* 2019]. Moreover, these non-B DNA structures are highly suspected to be implicated in human diseases such as cancers or neurological/psychiatric disorders [Ravichandran *et al.* 2019, Rhodes & Lipps 2015].

2.2 Transcription, chromatin and epigenetics

In the nucleus, DNA is not naked, but is instead associated with proteins, including histones (forming nucleosomes), transcription factors and repair proteins to form a complex structure, the so-called chromatin. Chromatin adopts different levels of compaction to eventually form a chromosome (Figure 2.2A). At the edge of chromosomes, specific regions called telomeres protect chromosomes, while

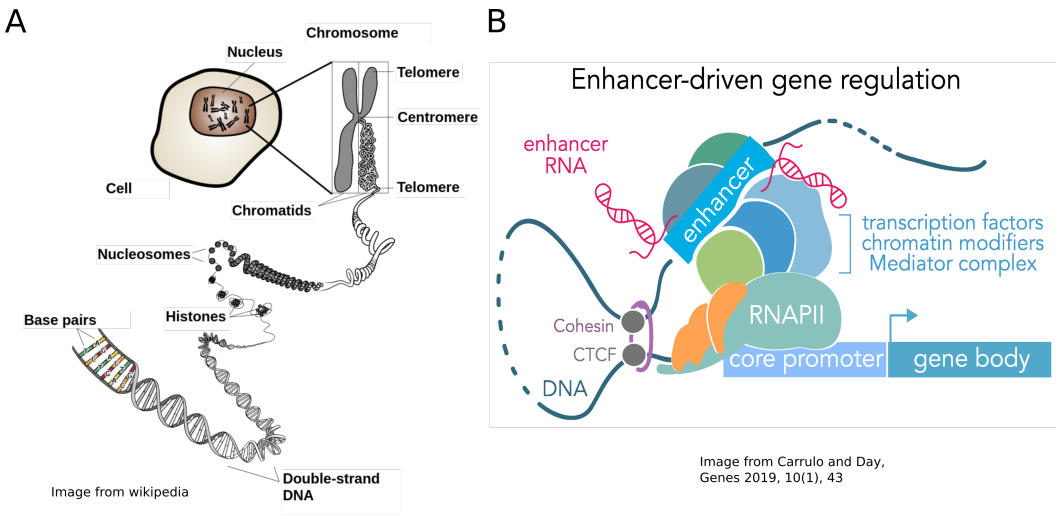


Figure 2.2: Chromosome, chromatin and transcription. A) From DNA to chromosome. B) Transcriptional regulation of genes.

somewhere in the center, centromeres determine kinetochore formation and sister chromatid cohesion. The properties of chromatin vary along the genome and are regulated by epigenetic marks (DNA methylation and histone modifications). Chromatin is composed of euchromatin that is lightly packed, enriched in genes, and is often actively transcribed, and of heterochromatin that is tightly packed, poor in genes, less accessible to ribonucleic acid (RNA) polymerases and therefore less transcribed. Chromatin regulates many cellular processes such as transcription [Hübner & Spector 2010, Ulianov *et al.* 2016], but also DNA replication [Moindrot *et al.* 2012] and DNA repair [Uusküla-Reimand *et al.* 2016].

The genome is comprised of genes that play a central role in the cell and participate in the development of the phenotype. In the human genome, the number of genes is estimated to be between 20000 and 25000 [International Human Genome Sequencing Consortium 2001]. A gene is transcribed by an RNA polymerase yielding to an RNA and eventually to a protein (Figure 2.2B). Genes make up only part of the genome (less than 30%) and gene coding regions, called exons, do not even occupy 3% of the genome. During the last decade, non-coding regions have been extensively studied and were shown to play many important roles including regulation, replication and structure [Khajavinia & Makalowski 2007, modENCODE Consortium *et al.* 2010, The ENCODE Consortium 2012]. A major role of specific non-coding regions, the so-called promoters and enhancers, is to regulate gene expression through the formation of DNA loops that are stabilized by transcription factors, architectural proteins (CTCF and cohesin) and other proteins, and also enhancer RNAs [Marsman & Horsfield 2012, Andersson *et al.* 2014, Carullo & Day 2019].

2.3 The genome in 3D

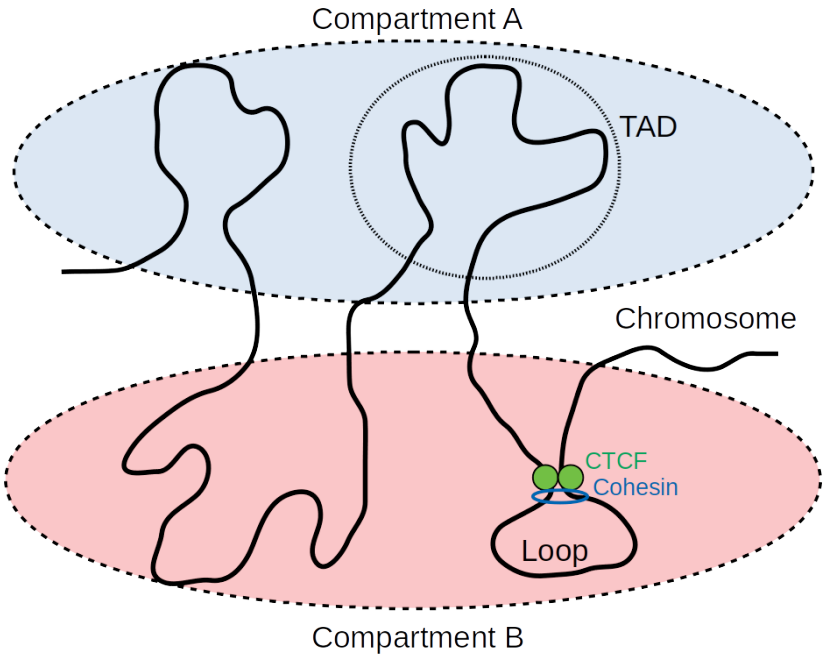


Figure 2.3: The 3D genome.

Chromosomal DNA is highly compacted in 3D, such that about 2 meters of this molecule fits into the microscopic nucleus of a human cell. The compaction of the genome is not random, but is on the contrary structured as recently revealed by mapping chromatin contacts using Hi-C (high-throughput chromosome conformation capture) [Lieberman-Aiden *et al.* 2009, Dixon *et al.* 2012] and ChIA-PET (chromatin interaction analysis by paired-end tag sequencing) [Fullwood *et al.* 2009] (Figure 2.3). In metazoans, compartments A and B were the first revealed structures by chromatin contact mapping. Compartment A tends to be active and gene rich, while compartment B is more inactive and gene poor. Topologically associating domains (TADs) were then discovered with higher mapping resolution. They represent a structural unit shared between cell types and kept between species [Dixon *et al.* 2012]. TADs are essential for many key cellular processes, such as the regulation of gene expression and DNA replication [Dixon *et al.* 2012, Pope *et al.* 2014]. In addition, it has recently been shown that the loss of 3D domains by a mutation can be linked to the onset of genetic diseases and cancer [Lupiáñez *et al.* 2015, Hnisz *et al.* 2016].

Other important 3D structures are chromatin loops between distant loci. They play key roles in gene expression regulation during development [Kadauke & Blobel 2009, Andersson *et al.* 2014, Ghavi-Helm *et al.* 2014]. In particular, in vertebrates, it was shown that loops that demarcate TADs are often

marked by asymmetric CTCF motifs where cohesin is recruited [Rao *et al.* 2014]. Accordingly, depletions of CTCF and cohesin decrease chromatin contacts [Zuin *et al.* 2014]. These results support the extrusion loop model where CTCF and cohesin act together to extrude unknotted loops during interphase [Sanborn *et al.* 2015]. Moreover, loop extrusion mediated by cohesin is a general mechanism that has also been observed in yeast [Dauban *et al.* 2020].

In drosophila, additional proteins shaping the genome in 3D have been identified, including BEAF-32, GAGA factor (GAF), Suppressor of Hairy-Wing (Su(HW)), zeste-white 5 (Zw5) or the drosophila homologue of Brd4, Fs(1)h-L, Pita, and Zinc-finger protein interacting with CP190 (ZIPIC) along with cofactors such as cohesin, CP190 or Lethal (3) malignant brain tumor (L(3)mbt) [Van Bortle *et al.* 2014]. Moreover, long-range contacts are influenced by additional non-architectural factors including transcription or remodeling factors, or more generally by gene density or transcriptional levels [Hou *et al.* 2012, Cubenas-Potts & Corces 2015, Rowley *et al.* 2017]. Additionally, long-range contacts are favored depending on the extent by which the RNA polymerase II (RNAPII) may remain stably "poised" or "paused," which would leave more opportunities for long-range contacts with enhancers [Ghavi-Helm *et al.* 2014].

2.4 Genome stability and DNA repair

Eukaryotic cells are exposed every day to both exogenous (*e.g.* UV and pollutants) and endogenous stresses (*e.g.* metabolic stress and DNA transactions) that can lead to DNA damage [McKinnon & Caldecott 2007]. For instance, ultraviolet (UV) exposure from the sun can induce several DNA damages and eventually lead to mutations and diseases. However, DNA damages are also caused by endogenous stresses which are the by-product of the normal cell activities. In living cells, reactive oxygen species (ROS) are formed continuously as a consequence of metabolic and other biochemical reactions and can lead to several types of DNA damage. A vast amount of DNA damages is also caused by DNA transactions such as DNA replication and transcription.

Among the various types of DNA lesions, DNA double strand breaks (DSBs) are by far the most deleterious, since they can lead to chromosome rearrangements [Mehta & Haber 2014, Kasparek & Humphrey 2011, Marnef *et al.* 2017, Vitor *et al.* 2020]. Chromosome rearrangements are large-scale mutations that include insertions, deletions, translocations, and fusions in the DNA [Zhang *et al.* 2009, Carvalho & Lupski 2016]. Once DNA is broken, DNA repair mechanisms identify and correct damages in the genome. There are two main pathways to repair DSBs: non-homologous end joining (NHEJ) and homologous recombination (HR) [Ceccaldi *et al.* 2016]. NHEJ directly ligates the break ends, whereas HR uses a homologous sequence to guide repair. If DNA repair is successful, the two ends of the same break are rejoined and the original DNA order is restored. But if DNA repair fails, the two ends of different breaks are joined

together, and a chromosomal rearrangement is generated.

An important mechanistic factor of chromosomal rearrangement is the 3D genome organization that can bring two linearly separated loci in physical proximity [Zhang *et al.* 2012]. In fact, DSBs can cluster together to form repair foci that concentrate repair factors [Caron *et al.* 2015]. In particular, DSB clustering mostly occurs in damaged active genes during G1 [Aymard *et al.* 2017, Guénolé & Legube 2017]. Moreover, recent DSB mapping combined with Hi-C experiments revealed that DSBs often occur at loop anchors where CTCF and cohesin bind [Canela *et al.* 2017]. Interestingly, topoisomerase 2B (TOP2B), an enzyme known to mediate DSBs, physically interacts with CTCF and cohesin at TAD borders [Uusküla-Reimand *et al.* 2016]. TOP2B is an enzyme that controls and alters the topological states of DNA. In particular, TOP2B catalyzes the transient breaking and rejoining of two strands of duplex DNA, which allows the strands to pass through one another, and thus the relief of torsional stress during transcription [Pommier *et al.* 2016].

Chromosome rearrangements have the potential to cause cancer, for instance, if they mutate a tumor suppressor gene or activate an oncogene. Rearrangements are also a relatively common cause of developmental disorders, occurring in 1 in 200 individuals, and often involve intellectual disabilities [MacIntyre *et al.* 2003]. Moreover, rearrangements contribute to psychiatric diseases, including schizophrenia and bipolar disorder [Dwyer 2020, Craddock & Owen 1994].

2.5 Genomics and omics

2.5.1 Human Genome Project and the birth of genomics

In 2001, the human genome was sequenced by a large scientific consortium, called the Human Genome Project (HGP) consortium [International Human Genome Sequencing Consortium 2001]. The project involved several countries and costed several billion dollars. This project was seminal for genomics, since it mapped most human genes from the genome as well as intergenic regions, and triggered the development of novel sequencing methods, called next-generation sequencing (NGS) [Goodwin *et al.* 2016]. NGS are high-throughput sequencing technologies that parallelize sequencing, yielding millions of small sequences at once in a fast and cheap manner.

2.5.2 An explosion of omic methods

Following the HGP, there has been an explosion of NGS methods to interrogate the genome and its functions leading to the development of "omic technologies", opening the post-genomic era (Figure 2.4). Numerous methods are currently used to study gene transcription, differential gene expression, and alternative splicing (RNA-seq), but also nascent transcription (GRO-seq), and non-coding small RNAs

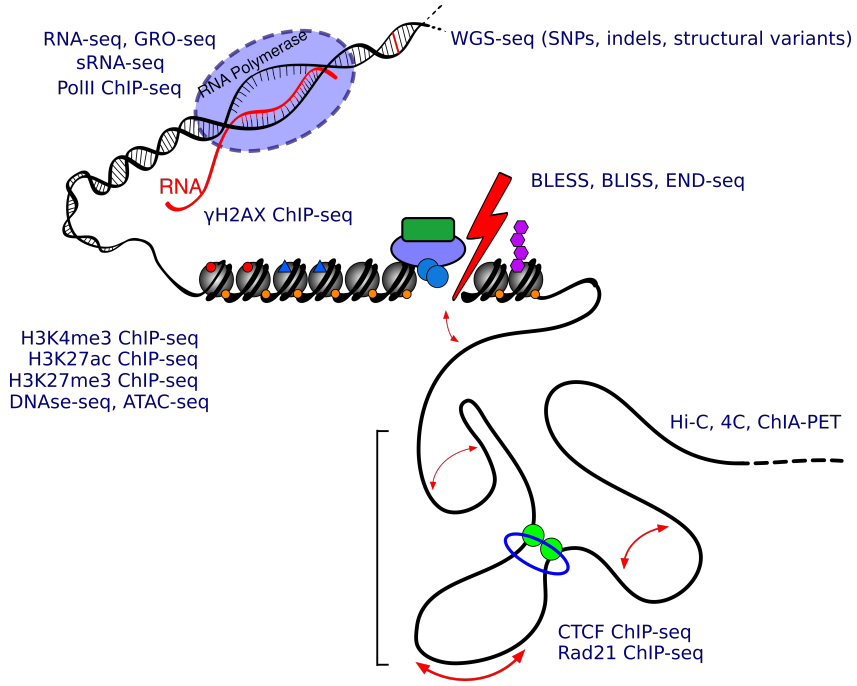


Figure 2.4: The different omic methods for the study of transcription, DNA repair and 3D genome.

(sRNA-seq) [Lowe *et al.* 2017]. To study chromatin, standard techniques are chromatin immunoprecipitation (ChIP-seq) to map transcription factor binding sites and histone modifications, and DNAse-seq [Hesselberth *et al.* 2009] and ATAC-seq (Assay for Transposase-Accessible Chromatin) [Buenrostro *et al.* 2013] to map accessible chromatin. 3D genome organization is mapped by chromatin conformation capture techniques including high-throughput chromatin conformation capture (Hi-C), circular chromatin conformation capture (4C-seq) or chromatin interaction analysis by paired-end tag sequencing (ChIA-PET) [Fullwood *et al.* 2009, Lieberman-Aiden *et al.* 2009, Zhao *et al.* 2006]. DNA damage such as DNA double-strand breaks are currently mapped by BLESS (breaks labeling, enrichment on streptavidin and next-generation sequencing) [Crosetto *et al.* 2013], BLISS (Breaks Labeling In Situ and Sequencing) [Yan *et al.* 2017b] and END-seq (DNA end sequencing) [Canela *et al.* 2016].

2.5.3 Examples of omic experiments and data

There are a wide variety of omic data resulting from diverse experiments. Here, we will focus on widely used NGS experiments for chromatin studies. RNA-seq consists in extracting RNA molecules (for instance mRNA), reverse-transcribing them to cDNAs, fragmenting cDNAs, amplifying fragments and then sequencing to produce reads (Figure 2.5A). Reads are mapped to genes (or any other transcription annotation) and counted. ChIP-seq crosslinks DNA with interacting proteins, frag-

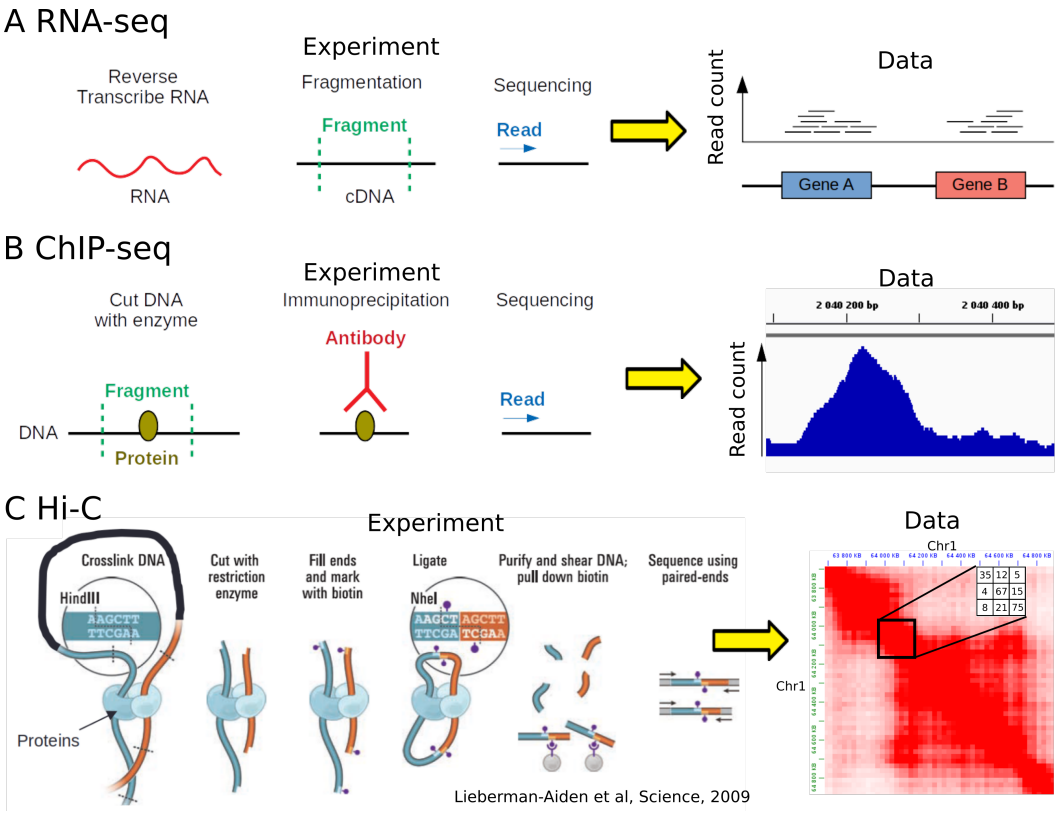


Figure 2.5: Experiments and data used for the study of chromatin. A) RNA-seq for transcription. B) ChIP-seq for protein binding to DNA or histone marks. C) Hi-C for the genome in 3D.

ments DNA with restriction enzymes, immunoprecipitates DNA with antibodies, amplifies fragments and then sequences to produce reads (Figure 2.5B). Reads are mapped to the genome and peaks are identified. Hi-C crosslinks DNA loci that are interacting, although they can be far apart in 1D, fragments DNA with restriction enzyme, fills ends and marks with biotin, ligates ends, immunoprecipitates with antibodies, amplifies and then sequences to produce read pairs (Figure 2.5C). In a pair of reads, the first read maps one DNA fragment (a particular locus), while the other read maps another DNA fragment (another locus). By binning the read pairs, a count matrix is obtained. Binning the read pairs into large bins helps reduce the sparsity of data. In the count matrix, each cell is the count of corresponding read pairs.

2.5.4 Single-cell paradigm

Omic methods were initially developed to analyze cell populations (*e.g.*, millions of cells), since detecting sufficient signal from a single cell represented an impossible challenge. However, recent technological progresses now allow to study omic information from individual cells with optimized NGS techniques, therefore providing a higher resolution of cellular differences and a better understanding of cell-to-cell heterogeneity [Nawy 2014]. For instance, single-cell analysis in the mouse cortex and hippocampus revealed unknown cell types by RNA-seq [Zeisel *et al.* 2015]. Moreover, single cell approaches were also crucial in cancer to reveal tumor heterogeneity due to mutations carried by small populations of cells [Lawson *et al.* 2018].

2.6 GWASs and non-coding SNPs

Complex genetic diseases are caused by the combined effects of multiple mutations with lifestyle and environmental factors [Visscher *et al.* 2017]. These diseases are common in the population and include heart disease, diabetes, schizophrenia and some cancers [Dorn & Cresci 2009, Billings & Florez 2010, Collins & Sullivan 2013, Chung *et al.* 2010]. Over the past decade, genome-wide association studies (GWASs) have successfully identified thousands of single nucleotide polymorphisms (SNPs) associated with complex diseases in an unbiased manner [Visscher *et al.* 2017].

However, GWASs uncovered that over 95% of GWAS associated SNPs are located outside coding sequences, which made it difficult to gain insight into the underlying biological mechanism [Maurano *et al.* 2012]. Interestingly, more than 75% of these SNPs overlap DNase I hypersensitive sites, which suggests a strong association with regulatory elements [Maurano *et al.* 2012]. Thus, a non-coding SNP might influence the expression of the target gene, either by altering its promoter or by affecting an enhancer that is linked to the gene via looping [Cookson *et al.* 2009]. Understanding how SNPs can alter regulatory element activity, as well as, chromatin looping with target genes thus represent a major issue for making sense of GWAS results.

2.7 Computational biology

2.7.1 Big data in genomics

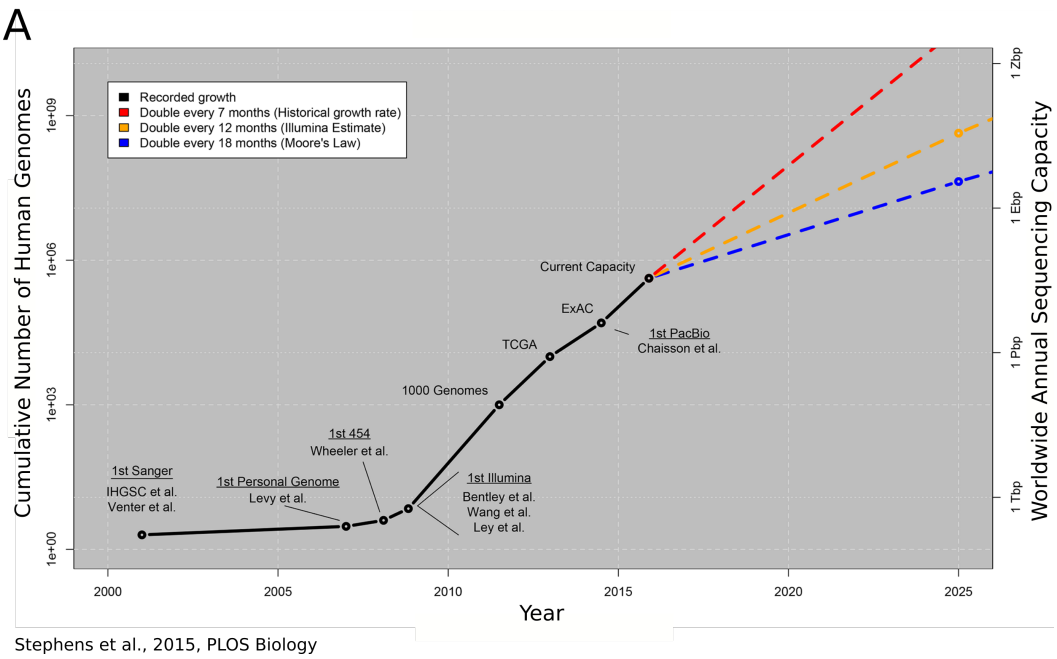


Figure 2.6: Exponential growth of genomic data over the past two decades.

Since the completion of the Human Genome Project in 2001, extraordinary progress has been made in NGS technologies, which has led to a dramatic decrease in sequencing cost and its widespread use in biology, medicine, ecology and evolution. The amount of data has exponentially expanded, and genomics has entered as other fields into the era of ”big data”. Big data refers to data whose characteristics in terms of volume, velocity and variety necessitate the development of novel technologies, algorithms and statistical models in order to extract key information which generally exceed the capacities of a single machine (Figure 2.6). This led to major research efforts in bioinformatics, including genome assembly, sequence alignment, gene identification, protein structure prediction, differential analysis of gene expression, protein-protein interactions, genome-wide association studies, and phylogenetic and evolutionary studies [Lesk 2002, Azuaje & Dopazo 2005, Horner *et al.* 2009, Andreas D. Baxevanis 2020].

2.7.2 Statistics for NGS data

Statistics is the scientific field that collect, analyze, interpret and present sample data. Statistics is at the core of data analysis and thus plays a central role in genomic and omic data. In particular, statistical models are heavily used for analyzing NGS data, especially for differential analysis [Robinson *et al.* 2009]. For instance, NGS data essentially represent count data, since the experimental

measure is often the number of reads that map to a particular region of the genome. For RNA-seq, the number of reads mapping to a gene is counted, whereas for ChIP-seq, the number of reads mapping to a regulatory region is counted, and for Hi-C, the number of read pairs within a bin pair is counted (Figure 2.5).

Since NGS data are counts, statistical models for count distribution are adequate tools for analysis. The most basic distribution for count data is the Poisson distribution:

$$P\left(X = k\right) = \frac{e^{-\lambda}\lambda^{k}}{k!},$$

where $\lambda = E[X] = Var[X]$. A major caveat of the Poisson distribution for modeling NGS data is that the variance is expected to be equal to the mean, whereas it is known that this assumption does not hold for NGS reads [Robinson & Smyth 2007a]. Instead, the negative Binomial distribution is widely used to model NGS counts [Robinson & Smyth 2007b]:

$$P\left(X = k\right) = \frac{\Gamma(k + \phi^{-1})}{\Gamma(\phi^{-1})\Gamma(k+1)} \left(\frac{1}{1+\lambda\phi}\right)^{\phi^{-1}} \left(\frac{\lambda}{\phi^{-1}+\lambda}\right)^{k},$$

because it allows the variance to be independent from the mean:

$$Var[X] = \lambda + \phi\lambda^{2},$$

where $\lambda = E[X]$, and ϕ is called the overdispersion parameter. Note that when $\phi \to 0$, then the negative binomial distribution tends to the Poisson distribution. Biologically speaking, the overdispersion allows to account for the biological variability between samples. More complex distributions were also proposed to model NGS data, in particular the zero-inflated (ZI) distributions (zero-inflated Poisson or zero-inflated negative binomial). ZI distributions are useful when frequent zero-valued observations are present in the data, which is often the case for single-cell NGS data [Risso *et al.* 2018].

The generalized linear model (GLM) implements these count distributions (among others) allowing a flexible generalization of the linear model (regression/ANOVA) useful for NGS data analysis. For instance, for the Poisson and negative binomial distributions, the GLM is:

$$\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{X}\right]\right) = \mathbf{X}\boldsymbol{\beta} \tag{2.1}$$

where $\mathbf{y}$ is the dependent variable, $\mathbf{X}$ the set of independent variables and $\boldsymbol{\beta}$ the model parameters. For differential analysis, the treatment factor is often encoded as a dummy variable with values equal to zero for the control condition and values equal to one for the treatment condition. The associated treatment factor coefficient corresponds to the natural logarithm of the fold-change between the two condition averages (Treatment / Control). The corresponding p-value allows to test if the coefficient is significantly different from zero, therefore assessing the significance of the fold-change between the two 2 conditions.

2.7.3 Machine learning

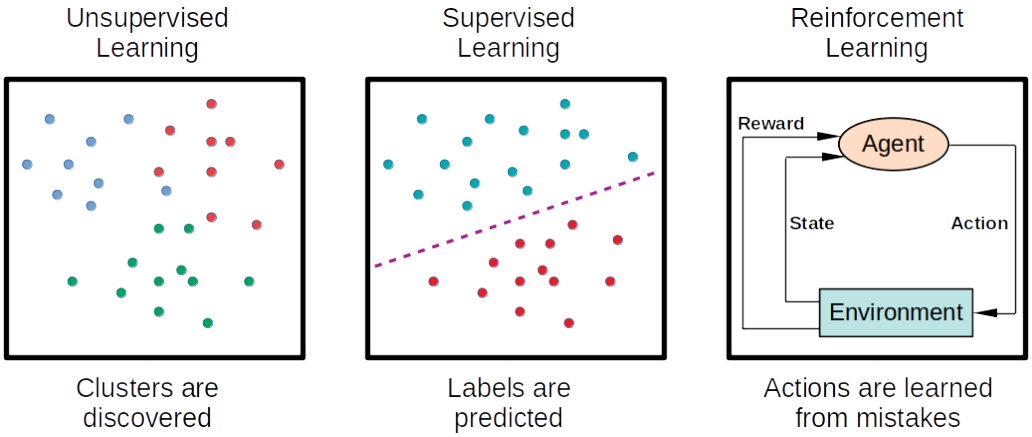


Figure 2.7: The different categories of machine learning methods.

Machine learning methods are increasingly used to analyze omic data. Such methods can be categorized into three approaches: (i) unsupervised learning, (ii) supervised learning, and (iii) reinforcement learning [Hastie *et al.* 2009, Bishop 2007] (Figure 2.7). The first two categories are the most used and developed to date for omic data.

Unsupervised learning looks for previously undetected patterns in a data set with no pre-existing labels and with a minimum of human supervision. Main methods consist in either reducing the dimension to compress data information (*e.g.* principal component analysis or t-distributed stochastic neighbor embedding) or in identifying groups of similar observations, also called clusters (*e.g.* k-means or hierarchical clustering). For instance, principal component analysis helps visualize in a simple manner key information from a large amount of variables and has many applications in omics such as representing ethnic variability from genetic data [Zheng & Weir 2016]. Cluster analysis is used instead to identify groups of individuals such as unknown cell (sub-)types from tissues [Andrews & Hemberg 2018].

Supervised learning considers the task of learning a function $g : \mathcal{X} \to \mathcal{Y}$ that maps an input space $\mathcal{X}$ to an output space $\mathcal{Y}$ based on example input-output pairs. Supervised learning is used to predict the unknown value of a variable (or more) given the values of other variables that are often easier or cheaper to collect. There are many machine learning algorithms that are often used in genomics, including artificial neural networks [Rosenblatt 1958], support vector machines [Boser *et al.* 1992], random forests [Breiman 2001], extreme gradient boosting [Chen & Guestrin 2016] and Bayesian networks [Jensen 1996]. In omics, supervised learning had many successful applications in cancer type predictions [Kourou *et al.* 2015], gene annotation [Mahood *et al.* 2020] or regulatory element mapping [Lee *et al.* 2011].

Reinforcement learning learns which actions to take in a given environment in order to maximize some reward [Sutton & Barto 2018]. Thus, this approach learns from mistake, similarly to humans. Reinforcement learning has tremendous applications in robotics, where a robot has to learn himself how to interact optimally with an environment. To date, the applications of reinforcement learning for omic data are very limited. However, recent preliminary studies suggest that reinforcement learning could improve genome assembly [Xavier *et al.* 2020].

2.7.4 Deep learning

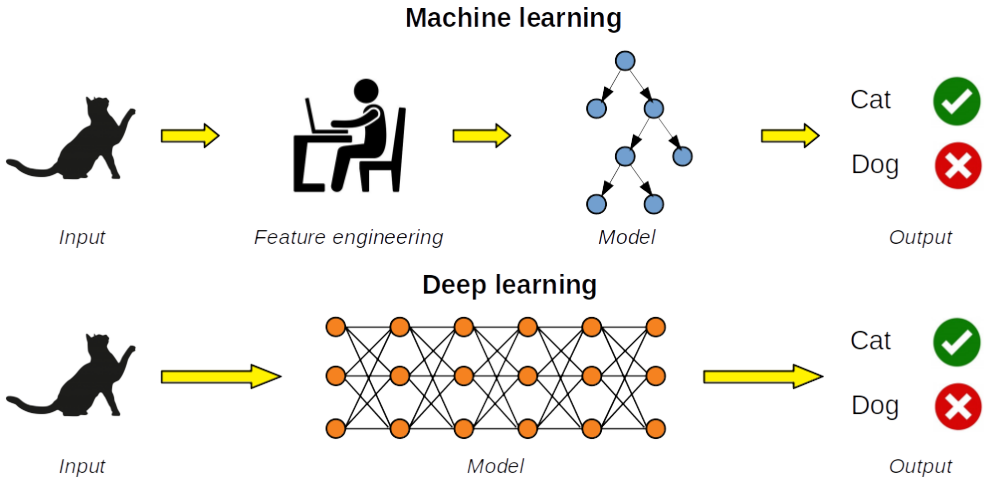


Figure 2.8: Difference between machine and deep learning.

Deep learning is a branch of machine learning that has gained considerable attention during the last years due to tremendous progress in the field [Goodfellow *et al.* 2016]. Deep learning is mostly based on artificial neural networks, but for which multiple layers progressively extract higher-level information from the raw input. The success of deep learning compared to machine learning is linked to the larger amount of data available (big data), new gradient descent algorithms and the use of graphics processing units (GPUs) speeding computations by 100 times.

Nowadays, deep learning achieves the best results for image, textual and audio data problems, for which data is complex and highly organized. Unlike machine learning, deep learning does not necessitate features previously built from expert knowledge, but instead learns directly features from data (Figure 2.8). Among deep learning models, convolutional neural networks (CNNs) were the first successful models [Krizhevsky *et al.* 2012]. CNNs implement a convolutional layer that consists of a set of learnable kernels capturing local patterns. In genomics, CNNs are used to predict regulatory elements from DNA sequence and

to assess in silico the effect of a non-coding SNP on regulatory element activity [Alipanahi *et al.* 2015, Zhou & Troyanskaya 2015]. Recurrent neural networks (RNNs) are another class of neural networks that is used for entire sequences of data [Jain & Medsker 1999]. However, RNNs often fail to process long sequences because of the vanishing gradient problem, and hence long short-term memory (LSTM) were successfully introduced to tackle this issue [Hochreiter & Schmidhuber 1997]. More recently, LSTM was replaced by the Transformer model that implements the attention layer which does not require the sequential data to be processed in a sequential order, allowing much more parallelization than RNNs or LSTMs and therefore considerably reducing training times [Vaswani *et al.* 2017]. Lastly, another recent approach, called transfer learning, consisted in transferring knowledge from a very complex and powerful network trained on a very large dataset to a simple network in order to increase performances when only a few data were available [Tan *et al.* 2018].

2.7.5 Heterogeneous data integration

The study of a biological system is best approached by incorporating knowledge from different perspectives in order to unravel the complexity of biology. Nowadays, genomic and omic technogologies allow to generate data from a wide range of experiments at different levels (mutation, transcription, chromatin modification, protein binding, DNA damage, etc.). Moreover, there are more and more biological databases from which experimental data can be freely and easily queried (Gene Expression Omnibus, https://www.ncbi.nlm.nih.gov/geo/; Expression Atlas, https://www.ebi.ac.uk/gxa/home; TCGA/ICGC, https://dcc.icgc.org/; UCSC Genome Browser, https://genome.ucsc.edu). However, the use of omic data from different experiments, as well as from different techniques, poses major challenges for integrating heterogeneous data.

There are two main approaches for data integration. The first approach heavily relies on expert knowledge from the biologist (hypothesis-driven approach). It consists in combining usually data from 2 or 3 different experiments in such way that this makes sense biologically. Often the biologist does not explore all the data available, but instead makes strong hypotheses for data analysis by focusing on certain candidate regions of the genome or certain candidate genes. While the hypothesis-driven approach is preferred for small research projects from a team, it is not relevant for big projects from a large consortium. Instead, a second approach can be chosen when the amount of data is too large to be exploited using restrictive hypotheses (data-driven approach). In the data-driven approach, statistical and data mining approaches are used. For instance, a wide range of multivariate methods such as principal component analysis, canonical correlation analysis or partial least squares models can summarize key information from data [Rohart *et al.* 2017]. Network-based methods also provide a natural framework for data integration by detecting potential interactions between biological processes or components at different scales [Amar & Shamir 2014, Lee *et al.* 2020]. When data

integration is used for some prediction tasks, such as in precision medicine, then machine learning algorithms provide a nice framework to integrate diverse data [Mobadersany *et al.* 2018].

2.7.6 Personalized medicine

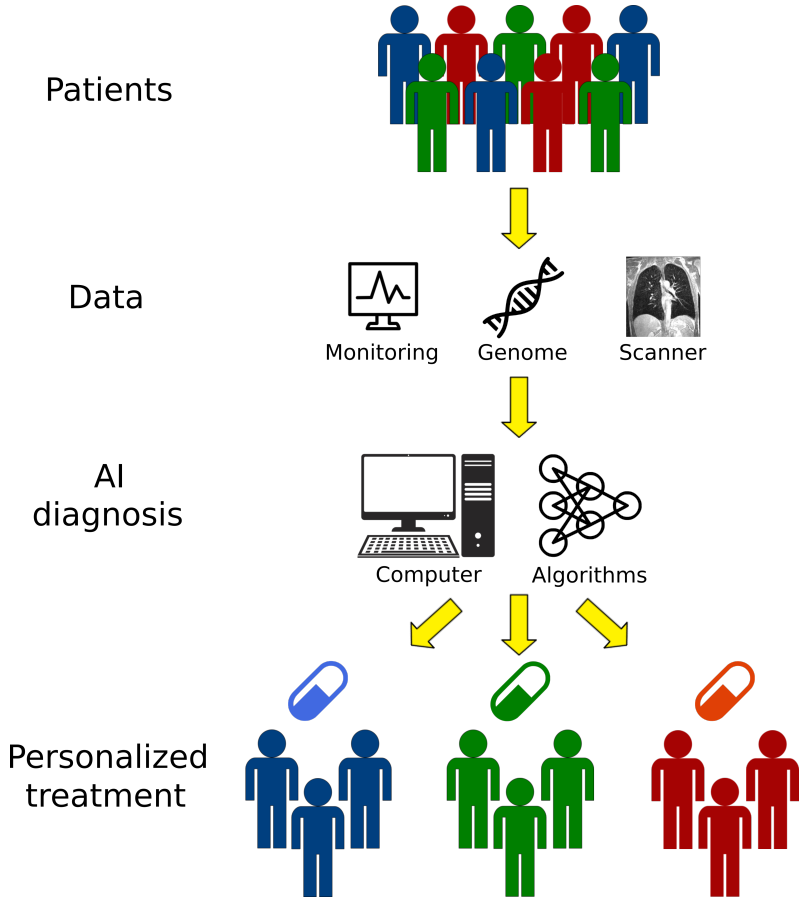


Figure 2.9: How patients can received personalized treatments using patient data combined with artificial intelligence (AI) diagnosis.

For the past ten years, medicine has been at the heart of a technological revolution in the way of considering, diagnosing and treating patients through personalized medicine, also called precision medicine. Patients are less and less considered as individuals from a homogeneous population for which an ideal identical treatment or diagnosis would exist. Conversely, medicine increasingly recognizes the uniqueness of each patient based on their genome, family history, lifestyle, and environment.

Recent advances in genomics and omic sciences in general (*e.g.* transcriptomics, metabolomics and proteomics), but also in medical imaging (*e.g.* MRI scanner and fluorescent labeling), intestinal microbiology and pharmacodynamics, have made

possible to accumulate a large number of genetic and physiological information for patients and their illnesses. Current techniques now generate an ever-increasing amount of medical data on patients, and medicine is considered to have entered like other disciplines into the era of Big Data, where data is immense and must be stored on bigger and bigger servers.

The availability of such large amounts of data at low cost is fueling the development of new approaches for personalized medicine based on computer algorithms, artificial intelligence, computational biology and biostatistics (Figure 2.9). Several types of data are used such as genomics and omics, health monitoring or radiography from MRI for instance. Data are then stored in a database and then processed using machine learning algorithms on supercomputers. Algorithms then decide the best treatment for every patient.

For example, the subtyping of certain cancers, which identifies the best treatment, is considered to be more efficient and more precise with the use of machine learning than traditionally done by physicians. Another successful example is to use a patient's genome to predict the likelihood of later developing a complex genetic disease like heart disease, allergies and asthma, neurological / psychological diseases, as well as, certain cancers of genetic origin.

At the moment, artificial intelligence approaches to personalized medicine are only in their infancy. New computational approaches must be developed in order to improve predictions (i) by automatically integrating more and more heterogeneous data of various kinds (omics, images, questionnaires, publications, etc.), (ii) by analyzing larger volumes of data rapidly, (iii) by exploiting data available in public databases such as ICGC / TGCA for cancer and GWAS Catalog for human genetics, and (iv) by implementing reinforcement learning from patient feedback.

Chapter 3

Contributions to research

Sommaire

> In the longer run and for wide-reaching issues, more creative solutions tend to come from imaginative interdisciplinary collaboration.
>
> Robert J. Shiller

3.1 Introduction

During the last decade, I have been focusing my research efforts on making sense of data and on developing novel computational methods for a variety of biological problems centered on the genome and its functions. During my postdoctorates, I had the chance to work on different topics including chromatin and cancer, human genetics of asthma and phylogenetics of viruses.

After being recruited as an assistant professor (maître de conférences) at University Paul Sabatier, most of my work was focused on the study of the 3D genome. In particular, I worked on the identification of protein binding factors and insulator sequences that could influence the formation of 3D domains, such as topologically associating domains (TADs), and the link with biological processes such as DNA repair and transcription. Another research direction was the development of machine and deep learning models for predicting genomic data, such as endogenous DNA double-strand breaks and active G-quadruplexes, which are related to chromatin, DNA repair and cancer.

But I also always attempted to incorporate into my research projects, concepts and methods borrowed from genetics and evolution I acquired during my postdoctorates. For instance, I investigated the evolution of the 3D genome by inferring CTCF loop characteristics directly from the genome sequence of vertebrate species, and by demonstrating their phylogenetic conservation. Moreover, I studied the impact of SNPs disrupting potential G-quadruplexes, and showed the link with gene expression.

3.2 Human genetics of asthma

Asthma is a complex genetic disease characterized by the inflammation and constriction of the airways. This disease affects more than 300 million people in the world and thus represents a major public health issue. Even though genome-wide association studies of common variants have successfully identified more than one hundred genes linked to asthma, only a fraction of the heritability of the disease could be identified. Among all the hypotheses, the role of rare variants has been proposed as an explanation for this missing heritability characteristic of common genetic diseases. The emergence of new sequencing technologies as well as the constant decrease in their cost currently allows the analysis of rare variants, SNPs and insertions-deletions, of a cohort of several thousand individuals or more.

During my postdoctoral fellowship at University of Chicago (Ober's lab), I analyzed rare variants associated with asthma severity. Our laboratory had sequenced 278 individuals with asthma, including 93 African-Americans, 101 European-Americans and 84 Latin Americans. In order to maximize the detection power of rare variants, I employed an approach based on the accumulation of the effects of rare variants of a gene, the so-called gene-based test [Wu *et al.* 2011]. I have also annotated these variants in order to include in the tests only those predicted to be functional and I only tested a limited number of candidate genes (approximately 300), to further reduce multiple testing issues. Although I have identified rare variants present in the GSDMB gene located in locus 17q12-21 in European Americans and Latin Americans, these results could not be replicated in a different cohort. Failure to identify genetic variants reflected a classical scenario encountered in human genetics: small size of the sample (about 300 individuals), as well as its structure in three populations. In parallel, I was involved in another project on rare

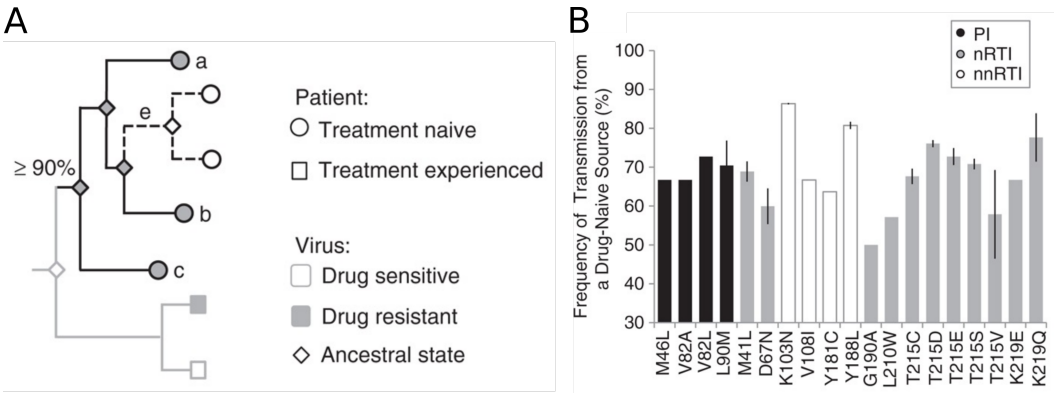


Figure 3.1: Phylogenetic analysis of antiretroviral resistance transmission from treatment-naïve individuals. A) Phylotype analysis to identify viral clusters. B) Frequency of antiretroviral resistance transmission from a drug-naïve source.

variants by Exome Chip for which 10 thousand individuals were available, and with such larger sample size, we could identify rare ethno-specific variants of asthma [Igartua *et al.* 2015].

3.3 Phylogenetics of HIV

Therapy combining antiretroviral (ARV) drugs has been proven highly effective in controlling HIV (human immunodeficiency virus) infections and has significantly improved patients' survival and quality of life. However, resistances to drugs are known to develop in treated individuals. Resistant viruses emerge through the selective pressure induced by antiretrovirals, but can also be transmitted from treated patients to treatment-naïve recipients. Usually the loss of fitness linked to the presence of resistance mutations in the absence of ARV treatment is sufficient to cause the virus to evolve back to its initial form (without resistance). Despite this, the presence of these reservoirs means that, in some cases, the mutant form continues to survive and to be transmitted in the absence of ARV treatment. So the presence of these reservoirs poses a serious threat to the long-term efficacy of the ARV therapy.

During my postdoctoral fellowship at Methods and Algorithms for Bioinformatics (MAB) team (LIRMM, Montpellier), I used a new phylogenetic approach, called Phylotype [Chevenet *et al.* 2013], to identify viral transmission clusters from 24,550 sequences of HIV-1 virus subtype B pol gene (Figure 3.1A). These sequences came from the UK HIV Drug Resistance Data Collection database. Treatment resistance clusters among HIV-positive individuals have been identified as containing at least 3 sequences with at least one shared resistance mutation, intra-clade genetic distance maximum of 4% and basal branch support of at least 90%. The persistence time of transmission chains was estimated using a molecular clock inference approach by least squares. The results showed that at least 70% of resistance to ARVs originated

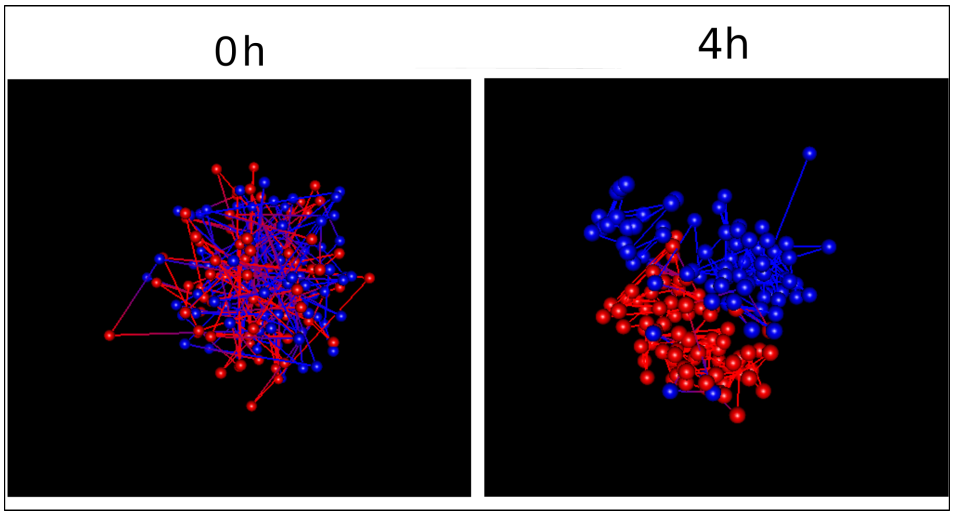


Figure 3.2: Effect of estrogen (E2) on the compartmentalization of chromosome 6. The chromosome was modeled in 3D using Hi-C data before (0h) and after estrogen (4h). Active and inactive chromatin regions are stained red (compartment A) and blue (compartment B), respectively.

from a naïve individual (Figure 3.1B) [Mourad *et al.* 2015].

3.4 The genome in 3D

The 3D genome was one of my research topic when I was a postdoctoral researcher at Indiana University. Back in 2011, studying the 3D genome using NGS was very new and exciting. Only few bioinformaticians were working on the topic, and consequently, we had to develop in-house libraries and scripts for data processing. After I was recruited in a chromatin lab in 2014, the 3D genome became one of my major research topics, since there was already a boom in the field and a lot of room for the development of computational methods and models to analyze Hi-C data. Nowadays, Hi-C experiment has become a standard technique to interrogate the 3D genome, and is routinely applied in research labs.

3.4.1 Estrogen induces global 3D genome reorganization in breast cancer

Estrogen is a class of sex hormone responsible for the development and regulation of the female reproductive system, but whose exposure also increases breast cancer risk. The action of estrogen is mediated by the estrogen receptor (ER), a protein that binds to DNA and controls gene expression. Previous studies showed that gene expression after estrogen stimulation is regulated through DNA looping [Hsu *et al.* 2010, Hsu *et al.* 2013]. Moreover, estrogen is known to alter the large-scale chromatin structure [Nye *et al.* 2002].

During my postdoctoral fellowship at Indiana University, I analyzed Hi-C data before/after estrogen induction [Mourad *et al.* 2014]. I observed that estrogen induces a global change of the 3D conformation of chromosomes in breast cancer cells. The addition of estrogen caused a gradual increase in the spatial compartmentalization of chromatin up to 4 hours (Figure 3.2). By integrating previous results with gene expression and epigenetic data, I demonstrated the link with the global regulation of the gene expression. After estrogen stimulation, gene-rich chromosomes, open and active regions of chromatin are in greater spatial proximity, thus allowing genes to share transcriptional machinery and regulatory elements. At the megabase scale, we also observed that the loci in differential interaction are enriched in genes involved in cancer proliferation and estrogen response. In addition, these loci showed higher estrogen receptor alpha binding and gene expression.

3.4.2 Prediction of 3D genome structure from epigenetic and chromatin data

In a research highlight, we surveyed recent computational methods demonstrating the strong link between 3D genome organization (Hi-C data) and 1D epigenetic and chromatin data (ChIP-seq, DNase-seq, Methylation array) [Mourad & Cuvier 2015]. Such strong link suggests that the 3D genome, which is costly to map experimentally, can be instead predicted using cheaper or publicly available 1D genome data. For instance, 3D compartments A/B are usually inferred from a principal component analysis of the correlation matrix from the Hi-C count matrix. However, the correlation matrix can be predicted using a correlation matrix computed from DNA methylation profiles across patients. Another work showed that machine learning methods such as Bayesian additive regression trees can predict TADs by using epigenetic data from various human cell lines, including tumor cells. Most notably, the localization of histone modifications and CTCF binding sites as observed from ChIP-Seq data provide good predictors of TAD borders.

3.4.3 Generalized linear models for bridging the gap between 1D and 3D genomes

Understanding the biological processes involved to shape the genome in 3D is a major question. One paradigm is to consider that the 1D genome contributes to the formation of 3D chromosomal structures such as 3D domains. In fact, several studies have shown that insulator binding proteins are enriched at 3D domain borders [Phillips-Cremins *et al.* 2013], that CTCF and cohesin proteins are involved in extrusion to form DNA loops [Rao *et al.* 2014, Sanborn *et al.* 2015, Rao *et al.* 2017], and that phase separation of histone marks could explain the formation of compartments [Jost *et al.* 2014]. Moreover, genomic elements, such as repetitive sequences, were also shown to co-localize in 3D [Cournac *et al.* 2015]. Experiments to demonstrate the role of a given protein often consist in depleting the protein. Depletions are very costly and thus cannot be systematically used to study the role of any DNA binding protein. Alternative computational methods are advantageous compared to experimental depletions, since they make it possible without any

cost to study the role of dozens or even hundreds of proteins whose ChIP-seq data or DNA binding motifs are already available in databases such ENCODE (https://www.encodeproject.org/) or JASPAR (http://jaspar.genereg.net/).

I proposed different generalized linear models (GLMs) to integrate and predict the 3D genome from the 1D genome. In a first work, logistic regression was proposed to model TAD border presence / absence depending on protein binding, genomic elements, and DNA motif presence. In a second work, negative binomial regression allowed to model Hi-C counts depending on the interaction between protein binding at different locations. In a third work, negative binomial regression modeled Hi-C counts depending on the blocking effect of protein binding, which did not necessitate any prior TAD identification (TAD-free). In a fourth work, Poisson and negative binomial regressions were used for TAD identification, differential analysis and prediction.

3.4.3.1 TADfeat: identification of protein drivers of TAD borders

A current challenge is to identify the molecular drivers of 3D domains of higher-order chromatin organization. However, few computational tools have been proposed to study the link between insulating proteins or functional elements (genomic factors) and the 3D domains such as TADs. A commonly used approach is to test for genomic factor enrichment at the borders of TADs by Fisher's exact test. However, the enrichment test can only identify the genomic factors that colocalize at TAD borders, but it is unable to determine which genomic factors are more likely to influence the borders. For instance, two genomic features might be both found significantly enriched at domain boundaries, but only one of them might truly influence the domain border establishment or maintenance. This is due to the colocalization (correlation) between the two genomic features. Statistically speaking, correlation does not imply causation. Non-parametric models were also used to predict TAD borders and have identified a subset of predictors. However one factor may accurately predict boundaries without being causative.

I proposed a new approach based on multiple logistic regression to measure the influence of factors on the boundaries of TADs [Mourad & Cuvier 2016]. Unlike the enrichment test, the regression takes into account the conditional independence between the factors and thus better identify the most influential factors (Figure 1, Scenario 1, from the article "Computational Identification of Genomic Features That Influence 3D Chromatin Domain Formation" below). In addition, the regression can account for the interaction between factors, and therefore, can assess the impact of the co-occurrence of factors on borders (Figure 1, Scenario 2, from the article below). In Drosophila, I have shown that, among known architectural proteins, BEAF-32 and CP190 are the main determinants of TADs. In humans, the model identified known proteins CTCF and cohesin, as well as ZNF143 and PRC2 as positive determinant borders. The model also revealed the existence of several factors having a negative effect on borders, including P300, RXRA,

BCL11A and ELK1. Based on the regression results, I proposed a new biological model explaining the formation of 3D domains, where positive driver proteins could favor attraction between loci, while negative driver proteins could instead trigger repulsion (Figure 8, from the article below).

PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Computational Identification of Genomic Features That Influence 3D Chromatin Domain Formation

Raphaël Mourad*, **Olivier Cuvier**

Laboratoire de Biologie Moléculaire Eucaryote (LBME), CNRS, Université Paul Sabatier (UPS), Toulouse, France

* raphael.mourad@ibcg.biotoul.fr

OPEN ACCESS

Citation: Mourad R, Cuvier O (2016) Computational Identification of Genomic Features That Influence 3D Chromatin Domain Formation. PLoS Comput Biol 12 (5): e1004908. doi:10.1371/journal.pcbi.1004908

Editor: Kai Tan, University of Pennsylvania, UNITED STATES

Received: January 14, 2016

Accepted: April 7, 2016

Published: May 20, 2016

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the University of Toulouse IDEX program, the CNRS and by the ANR 'INSULa'. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Abstract

Recent advances in long-range Hi-C contact mapping have revealed the importance of the 3D structure of chromosomes in gene expression. A current challenge is to identify the key molecular drivers of this 3D structure. Several genomic features, such as architectural proteins and functional elements, were shown to be enriched at topological domain borders using classical enrichment tests. Here we propose multiple logistic regression to identify those genomic features that positively or negatively influence domain border establishment or maintenance. The model is flexible, and can account for statistical interactions among multiple genomic features. Using both simulated and real data, we show that our model outperforms enrichment test and non-parametric models, such as random forests, for the identification of genomic features that influence domain borders. Using *Drosophila* Hi-C data at a very high resolution of 1 kb, our model suggests that, among architectural proteins, BEAF-32 and CP190 are the main positive drivers of 3D domain borders. In humans, our model identifies well-known architectural proteins CTCF and cohesin, as well as ZNF143 and Polycomb group proteins as positive drivers of domain borders. The model also reveals the existence of several negative drivers that counteract the presence of domain borders including P300, RXRA, BCL11A and ELK1.

Author Summary

Chromosomal DNA is tightly packed up in 3D such that around 2 meters of this long molecule fits into the microscopic nucleus of every cell. The genome packing is not random, but instead structured in 3D domains that are essential to numerous key processes in the cell, such as for the regulation of gene expression or for the replication of DNA. A current challenge is to identify the key molecular drivers of this higher-order chromosome organization. Here we propose a novel computational integrative approach to identify proteins and DNA elements that positively or negatively influence the establishment or maintenance of 3D domains. Analysis of *Drosophila* data at very high resolution suggests that among architectural proteins, BEAF-32 and CP190 are the main positive drivers of 3D

domains. In humans, our results highlight the roles of CTCF, cohesin, ZNF143 and Polycomb group proteins as positive drivers of 3D domains, in contrast to P300, RXRA, BCL11A and ELK1 that act as negative drivers.

Introduction

High-throughput chromatin conformation capture (Hi-C) has emerged over the past years as an efficient approach to map long-range chromatin contacts [1–3]. This technique has allowed the study of the 3D architecture of chromosomes at an unprecedented resolution for many genomes and cell types [4–7]. Multiple hierarchical levels of genome organization have been revealed: compartments A/B [1], sub-compartments [8], topologically associating domains (TADs) [4, 5] and sub-TADs [7]. Among those domains, TADs represent a pervasive structural feature of the genome organization. TADs are stable across different cell types and highly conserved across species.

A current challenge is to identify the molecular drivers of topological arrangements of higher-order chromatin organization. There is a growing body of evidence that insulator binding proteins (IBPs) such as CTCF, and cofactors such as cohesin, act as mediators of long-range chromatin contacts [5, 6, 9–11]. In human, depletion of cohesin predominantly reduces interactions within TADs, whereas depletion of CTCF not only decreases intradomain contacts but also increases interdomain contacts [12]. The densest Hi-C mapping in human has recently revealed that loops that demarcate domains are often marked by asymmetric CTCF motifs where cohesin is recruited [8]. In *Drosophila*, silencing of cohesin and condensin II have recently demonstrated their roles on long-range contacts [13]. In addition, numerous IBPs, cofactors and functional elements colocalize at TAD borders [11]. However it is unclear if all these proteins and functional elements, or specific combinations of them, play a role in TAD border establishment or maintenance. Computational approaches that integrate protein binding (chromatin immunoprecipitation followed by high-throughput DNA sequencing, ChIP-seq) with Hi-C data may be well-suited to identify the key drivers of chromatin architecture.

Most computational approaches dedicated to chromosome conformation analysis have focused on correcting contact matrices for experimental biases [6, 14–16] in order to assess more precisely the significance of contact counts [17, 18], to identify chromatin compartments [1, 15, 19], or to 3D model chromosome folding [1, 5, 20–22]. However few computational methods have been proposed to study the roles of DNA-binding proteins and functional elements in chromosome folding. A simple yet widely used statistical method consists in assessing enrichment of a genomic feature around 3D domain borders by Fisher's exact or Pearson's chi-squared tests [4, 5, 7]. An important caveat of enrichment test is that it only identifies those genomic features that colocalize at domain borders, but it cannot determine which genomic features influence the domain border establishment or maintenance. For instance, two genomic features might be both found significantly enriched at domain boundaries, but only one of them might truly influence the domain border establishment or maintenance. This is due to the colocalization (correlation) between the two genomic features. Statistically speaking, correlation does not imply causation. Other works focused on the prediction of 3D domain borders using (semi) non-parametric models and identified a subset of genomic features that are the most predictive of TADs [23, 24]. However a genomic feature can efficiently predict 3D domain borders without being influential [25].

In this paper, we propose a multiple logistic regression to assess the influence of genomic features such as DNA-binding proteins and functional elements on topological chromatin

domain borders. Compared to enrichment test and non-parametric models, multiple logistic regression assesses conditional independence and thus can identify most influential proteins with respect to domain borders. Moreover the multiple logistic regression model can easily accommodate interactions between genomic features to assess the impact of co-occurences on domain borders. We illustrate our model using recent *Drosophila* and human Hi-C data allowing to probe TAD borders depending on multiple proteins and functional elements. Using both simulated and real data, we show that our model outperforms enrichment test and non-parametric models such as random forests for the identification of known and suspected architectural proteins. In addition, the proposed method identifies genomic features that positively or negatively impact TAD borders with a very high resolution of 1 kb.

Results

The model

The proposed multiple logistic regression models the influences of p genomic features on 3D domain borders:

$$\ln \frac{Prob(Y=1|\mathbf{X})}{1-Prob(Y=1|\mathbf{X})} = \beta_0 + \boldsymbol{\beta}\mathbf{X} \quad (1)$$

Where $\mathbf{X} = \{X_1, \ldots, X_p\}$ is the set of p genomic features such as DNA-binding proteins and Y is a variable that indicates if the genomic bin belongs to a border ($Y = 1$) or not ($Y = 0$). The set $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ denotes slope parameters, one parameter for each genomic feature. The model can easily accommodate interaction terms between genomic features (see Subsection Materials and Methods, Analysis of interactions). By default, model likelihood is maximized by iteratively reweighted least squares to estimate unbiaised parameters. However, when there are a large number of correlated genomic features in the model, L1-regularization is used instead to reduce instability in parameter estimation [26].

We illustrate the proposed model using two scenarios and compare it with enrichment test (Fig 1). In the first scenario, protein A positively influences 3D domain borders, while protein B colocalizes to protein A. In this scenario, enrichment test will estimate that the parameter associated with protein A $\beta_A > 0$ and the parameter associated with protein B $\beta_B > 0$. In other words, both proteins A and B are enriched at 3D domain borders. Multiple logistic regression will instead estimate that parameters $\beta_A > 0$ and $\beta_B = 0$. This means that protein A positively influences 3D domain borders, while protein B does not. This is because multiple logistic regression can discard spurious associations (here between protein B and 3D domain borders). One would argue that enrichment test can also be used to discard the spurious association if the enrichment of protein B when protein A is absent is tested instead. However such conditional enrichment test becomes intractable when more than 3 proteins colocalize to domain borders, whereas multiple logistic regression is not limited by the numbers of proteins to analyze within the same model.

In the second scenario, the co-occurrence of proteins A and B influences 3D domain borders, but not the proteins alone. Enrichment test will find that each protein alone is enriched at 3D domain borders ($\beta_A > 0$ and $\beta_B > 0$) as well as their interaction ($\beta_{AB} > 0$). The proposed model will instead find that only the interaction between proteins A and B influences 3D domain borders ($\beta_A = 0, \beta_B = 0$ and $\beta_{AB} > 0$).

In addition to these two previous scenarios, another interest of the model is the possibility to study the negative influence of a protein (or of a co-occurence of proteins) on TAD border establishment of maintenance. In other words, its presence counteracts the establishment or

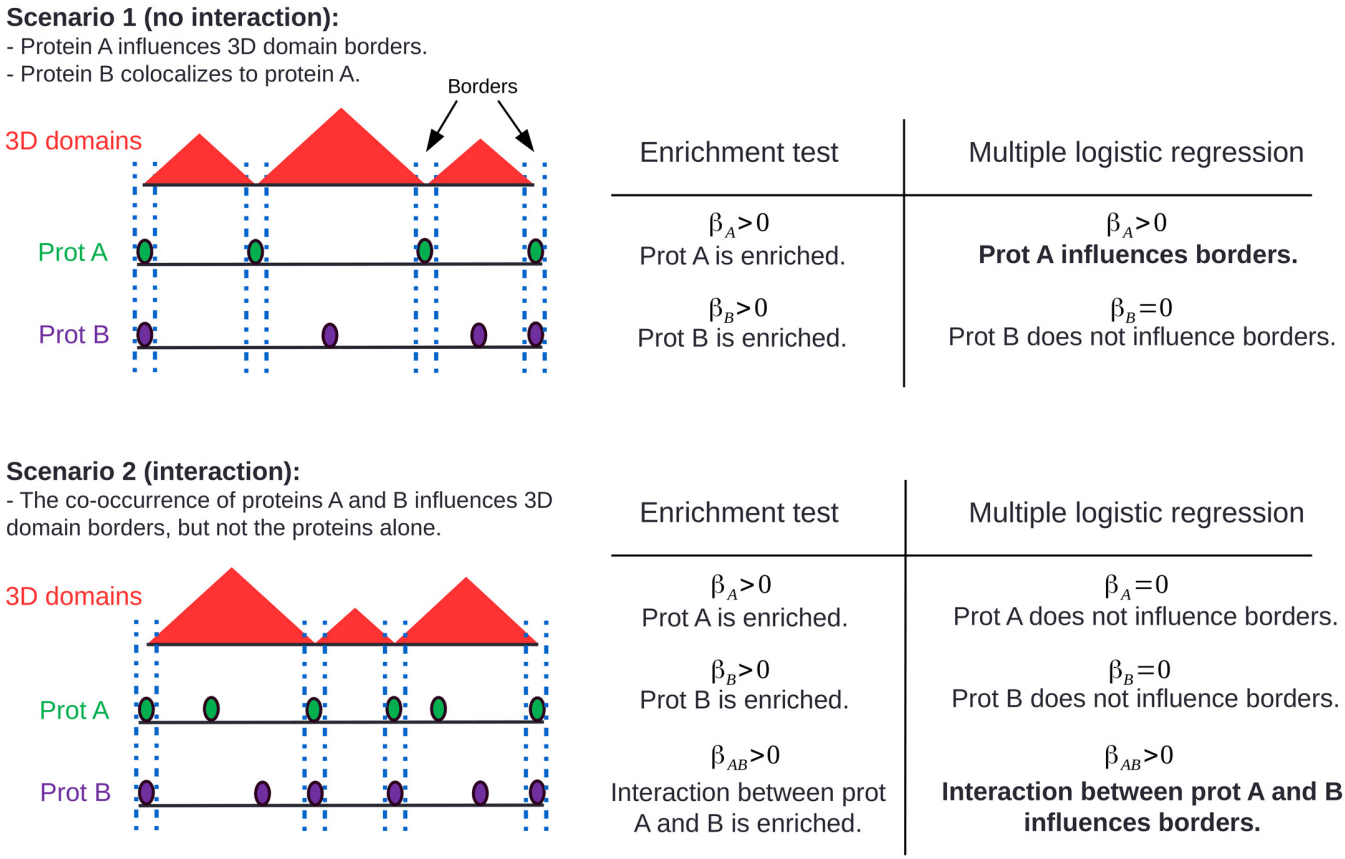


Fig 1. Illustration of the proposed multiple logistic regression to assess the influences of genomic features on 3D domain borders and comparison with enrichment test.

doi:10.1371/journal.pcbi.1004908.g001

maintenance of 3D domain borders. In such scenario, multiple logistic regression will estimate a parameter $\beta < 0$ (see below).

Depending on the parameter estimation algorithm used (likelihood maximization or L1-regularization), results are interpreted differently. If likelihood maximization is used, then a protein beta parameter can be considered as significantly different from zero if the corresponding p-value is lower than the familywise error rate (FWER) computed by Bonferroni procedure. If L1-regularization is used instead, then p-values are not computed. A protein is considered as influential if its beta parameter is different from zero. Using both algorithms, the beta parameter is the only measure used to quantify how strong is the influence of a protein on the 3D domain borders, and the p-value should not be used instead because it depends on the amount of data available. Both algorithms are useful in practice. Likelihood maximization allows to estimate beta parameters without any bias but influential proteins should be known in advance. L1-regularization can be useful to select the influential proteins among a large set of correlated candidates, but estimates will be biased.

Parameter estimation accuracy

Several characteristics of the analyzed ChIP-seq and functional element data might prevent the accurate estimation of multiple logistic regression parameters β. The matrix $\mathbf{X}$ of genomic features is sparse (numerous values equal zero) because genomic features are often absent from a

particular genomic bin. Sparsity of matrix **X** is known to prevent convergence of maximum likelihood maximization for parameter estimation [27]. Moreover some genomic features can be correlated. For instance, different insulator binding proteins might bind to the same genomic regions. For all these reasons, accurate estimation of parameters could fail in theory. Hence we evaluated the accuracy of parameter estimation using simulations.

We simulated data that were similar to real ChIP-seq data (see Subsection Materials and Methods, Data simulation, first paragraph). Both genomic coordinate data (*e.g.*, ChIP-seq peak coordinates) and quantitative data (*e.g.*, ChIP-seq signal intensity $\log\frac{ChIP}{Input}$) were generated. From the simulated data, multiple logistic regression model parameters were then estimated by maximum likelihood. We first simulated 100 genomic coordinate and 100 quantitative datasets that comprised 6 proteins and learned models without considering any interaction terms. In Fig 2a, we plotted true against estimated parameter values. We reported a very good accuracy for parameter estimation for both genomic coordinate and quantitative data with $R^2 = 99.5\%$ ($p < 1 \times 10^{-20}$) and $R^2 > 99.9\%$ ($p < 1 \times 10^{-20}$) between true and estimated parameter values, respectively. Because some proteins might be rare over the genome and only involved in some 3D domain borders, we studied parameter accuracy for simulated proteins with varied ChIP-seq peak numbers. Parameter estimation was highly accurate even for proteins with a low number of peaks over the genome ($R^2 = 97.4\%$ for 50 peaks; S1 Fig). In addition, we sought to assess how parameter estimation is affected by 3D domain border inaccuracy of few kilobases. We observed that with a border inaccuracy equal or lower than 2 kb, parameter estimation was still accurate ($R^2 > 70.9\%$, S2 Fig). We then simulated 100 genomic coordinate and 100 quantitative datasets that comprised the same 6 proteins and learned models with all two-way (*e.g.* $X_1\,X_2$) interaction terms. In Fig 2b, we plotted true against estimated parameter values corresponding to interaction terms only. Parameter estimation accuracy was still high for both genomic coordinate data ($R^2 = 94.6\%$, $p < 1 \times 10^{-20}$) and quantitative data ($R^2 = 99.9\%$, $p < 1 \times 10^{-20}$). We concluded that model parameter estimation was accurate for both marginal and two-way interaction of genomic features.

MLR outperforms enrichment test and random forests to identify drivers of TAD borders

We then sought to assess how multiple logistic regression (MLR) efficiently identifies genomic features that influence TAD borders, comparing with other approaches commonly used to assess the link between TAD borders and genomic features. We compared our model with enrichment test (ET) [4] and non-parametric model [23]. For the non-parametric model, we used random forests (RF) which are very similar to the model used in [23], but for which a scalable implementation allowed high resolution analysis (https://github.com/aloysius-lim/bigrf). For this purpose, we first simulated 100 datasets comprising 11 genomic features $\{X_1, X_2, \ldots, X_{11}\}$ that were similar to real ChIP-seq data (see Subsection Materials and Methods, Data simulation, second paragraph). Among the genomic features, variables X_1 and X_{10} were chosen to be causal with an odds ratio of 4, which was comparable to odds ratios estimated from real data (see below). We compared beta parameters from multiple logistic regression with beta parameters from enrichment test and variable importances from random forests (Fig 3a). Enrichment test correctly identified causal variables X_1 and X_{10} as the most enriched (beta median = 1.3), but also found highly enriched non-causal variables (beta median = 1). Random forests detected X_3 and X_8 as the most influential variables for prediction (variable importance median >2.75), although they were not causal genomic features. In contrast, multiple logistic regression correctly identified X_1 and X_{10} as influential variables (beta median = 0.93) and discarded non-causal variables (beta median = −0.03).

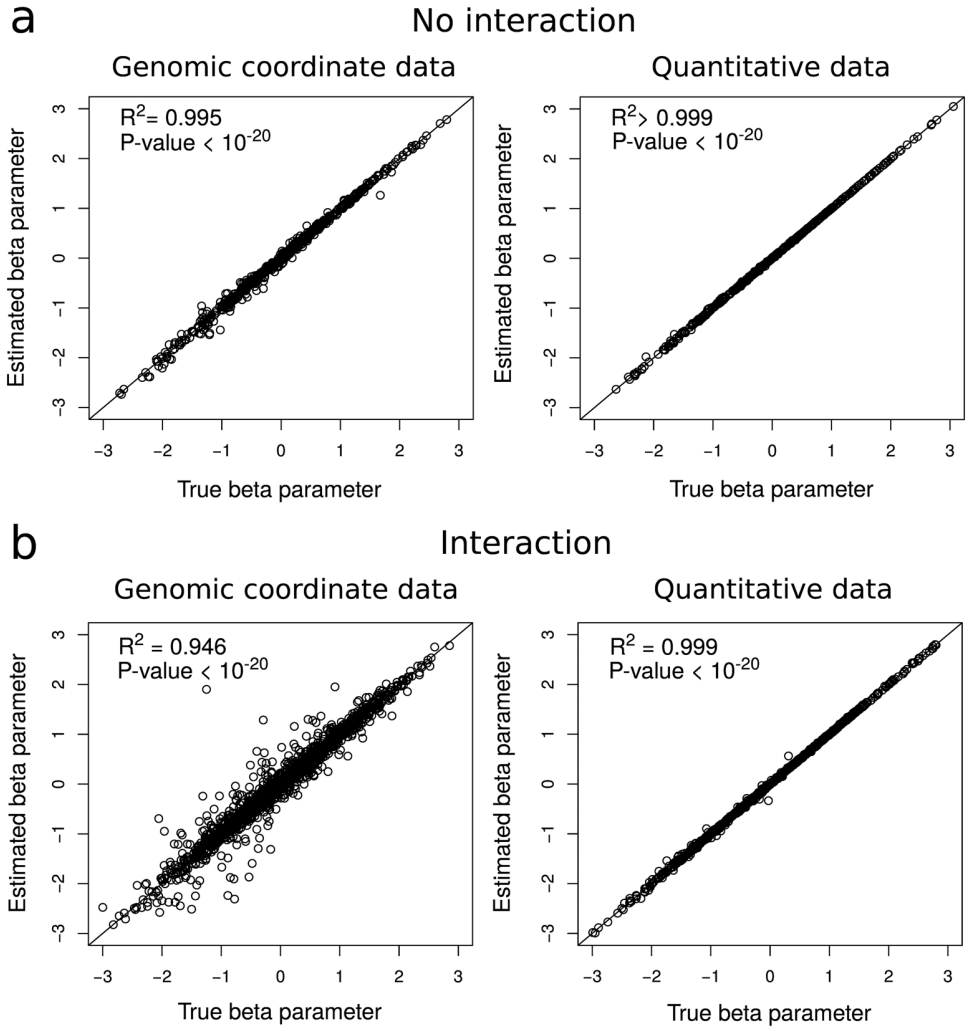


Fig 2. Parameter estimation accuracy of multivariate logistic regression. a) Estimated versus true parameter for marginal genomic features (the model does not include any interaction between genomic features). b) Estimated versus true parameter for two-way interactions between genomic features (*i.e.* for any interaction between two genomic features, see Subsection Materials and Methods, Analysis of interactions). Genomic coordinate data are ChIP-seq peak coordinates. Quantitative data are ChIP-seq signal intensities $log \frac{ChIP}{Input}$.

doi:10.1371/journal.pcbi.1004908.g002

We next simulated more complex scenarios for which the causal variables and their number were randomly chosen for each simulation. In addition, simulations were carried out for different odds ratios to study the influence of effect size. As previously, we compared multiple logistic regression with enrichment test and random forests. For each method, we computed the percentage of models that correctly ranked first the causal variables in terms of beta parameter or variable importance (Fig 3b). We observed that both enrichment test and multiple logistic regression successfully ranked first the causal variables even for a low odds ratio of 2 (93% of models), whereas random forests mostly failed even for the easiest scenario (44% of models for an odds ratio of 8; in the next paragraph, we will see that random forests poorly performed here partly due to high data sparsity). We then compared empirical type I error rate for a significance threshold $\alpha = 10^{-5}$ between enrichment test and multiple logistic regression for which p-values on beta coefficients were available (Fig 3c). Even for a high odds ratio of 8,

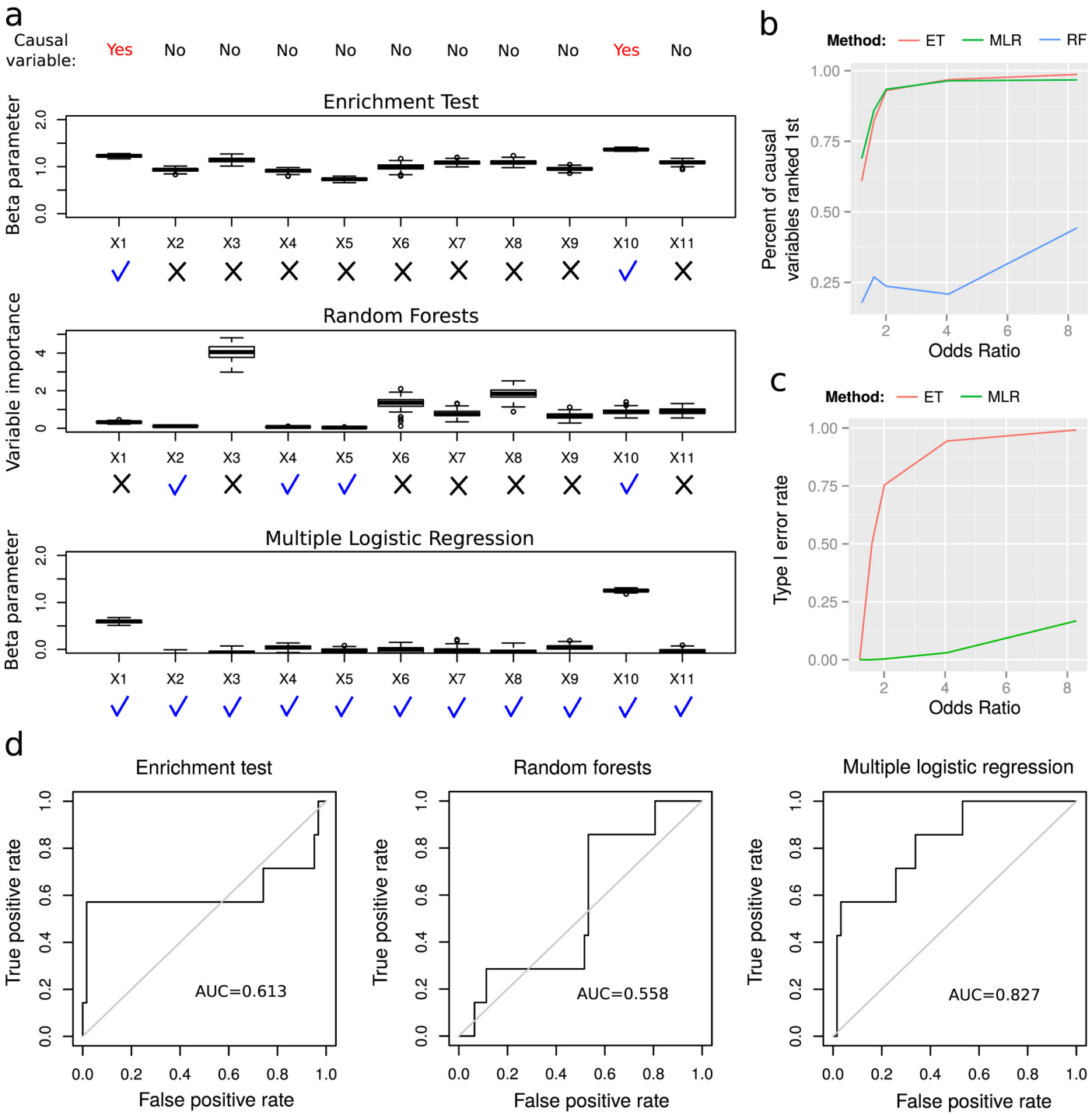


Fig 3. Comparisons between multiple logistic regression (MLR), enrichment test (ET) and random forests (RF) on simulated and real data. a) Comparison of MLR beta parameters with ET beta parameters and RF variable importances obtained from 100 simulated datasets including 11 genomic features. Among the genomic features, variables X_1 and X_{10} were chosen to be causal. For a method, a blue check mark denotes a causal or non-causal variable that was correctly identified as causal (resp. non-causal). A black x mark denotes a causal or non-causal variable that was incorrectly identified as non-causal (resp. causal). b) Percents of causal variables ranked first by ET, MLR and RF computed from 100 simulated datasets and varying odds ratios. Here the causal variables and their number were randomly drawn at each simulation. c) Type I error rates for MLR and ET computed from 100 simulated datasets. RF were not included because no p-values were available. The significance threshold α was set to 10^{-5}. Simulated data were the same as in b). d) Comparison of MLR with ET and RF to detect known or suspected architectural proteins in human using GM12878 cell ChIP-seq data. Receiver operating characteristic (ROC) curves were computed from Wald's statistics for ET, from beta parameters for MLR, and from variable importances for random forests. Computations were carried out at 1 kb resolution.

doi:10.1371/journal.pcbi.1004908.g003

MLR had a low error rate of 16%. Conversely enrichment test showed a high error rate of 75% even for an odds ratio of 2.

We also compared MLR with ET and RF using real data in human. For this purpose, we analyzed new 3D domains detected from recent high resolution Hi-C data at 1 kb for GM12878 cells for which 69 ChIP-seq data were available [8]. Multiple lines of evidence indicate that CTCF and cohesin serve as mediators of long-range contacts [5, 6, 9–11, 28]. However several proteins also colocalize or interact with CTCF, including Yin Yang 1 (YY1), Kaiso, MYC-associated zing-finger protein (MAZ), jun-D proto-oncogene (JUND) and ZNF143 [29]. In addition, recent work has demonstrated the spatial clustering of Polycomb repressive complex proteins [30]. Using the large number of available proteins in GM12878 cells, we could compare MLR with ET and RF to identify known or suspected architectural proteins CTCF, cohesin, YY1, Kaiso, MAZ, JUND, ZNF143 and EZH2. For this purpose, we computed receiver operating characteristic (ROC) curves using Wald's statistics for ET, beta parameters for MLR, and variable importances for RF. We carried out computations at the very high resolution of 1 kb (see Subsection Materials and Methods, Binned data matrix). ROC curves revealed that MLR clearly outperformed ET and RF to identify architectural proteins ($AUC_{MLR} = 0.827$; Fig 3d). Lower performance of ET ($AUC_{ET} = 0.613$) was likely due to its inability to account for correlations among the proteins (average correlation = 0.19). Regarding RF, its low performance ($AUC_{RF} = 0.558$) could be explained by its well-known inefficiency with sparse data (at 1kb, there were 99.4% of zeros in the data matrix **X**). At a lower resolution of 40 kb (88.5% of zeros), RF performed much better ($AUC_{RF} = 0.746$) but still lower than MLR ($AUC_{MLR} = 0.815$; S3 Fig).

To further validate MLR results with real data, we analyzed the impacts of single nucleotide polymorphisms (SNPs) in the consensus CTCF motif in human. SNPs play an important role in common genetic diseases and recent works have uncovered differential long-range contacts due to variations in the CTCF motif [31–33]. SNPs in the consensus CTCF motif are thus expected to affect, and most likely to decrease, the influence of CTCF motif on 3D domain border establishment or maintenance. We then tested if MLR was able to detect the impacts of SNPs on CTCF motif. For this purpose, we included within the same MLR model the wild-type (WT) motif and the three alternative alleles for a given position in the motif. For instance, for the first position, the MLR comprised genomic coordinates of the WT motif CCANNAGNNGGCA and the genomic coordinates of the mutated motifs ACANNAGNNGGCA, GCANNAGNNGGCA and TCANNAGNNGGCA. Over 27 mutated CTCF motifs, 25 showed beta coefficients that were lower than the one of WT CTCF motif, indicating that the corresponding SNPs diminished the influence of CTCF motif on TAD borders as expected (Fig 4). Because correlations among the motif variables were very low (average correlation <0.01), ET performed as efficiently as MLR to detect the influences of SNPs ($AUC_{ET} = 0.926$ and $AUC_{MLR} = 0.926$), but RF was inaccurate ($AUC_{RF} = 0.638$; S4 Fig). For instance, for the first position, we observed that all three alternative alleles (A, G and T) diminished the influence of the motif with respect to 3D domain borders. Some mutations even canceled the influence of CTCF motif (for instance, alleles A and T on position 2). On the last position, allele G had a higher influence than the WT motif. This result was actually consistent with the ambiguity between allele A and G in the motif. Similar results were obtained for consensus BEAF-32 motif CGATA in *Drosophila* (S5 Fig).

Using both simulated and real data, we concluded that multiple logistic regression correctly identified causal variables and discarded spurious associations of non-causal variables with TAD borders while both enrichment test and random forests failed. In addition, multiple logistic regression successfully predicted expected effects of SNPs on CTCF and BEAF-32 motifs known to influence long-range contacts in human and *Drosophila*, respectively. These predicted effects of SNPs could further serve to identify new regulatory variants in the context of genome-wide association studies.

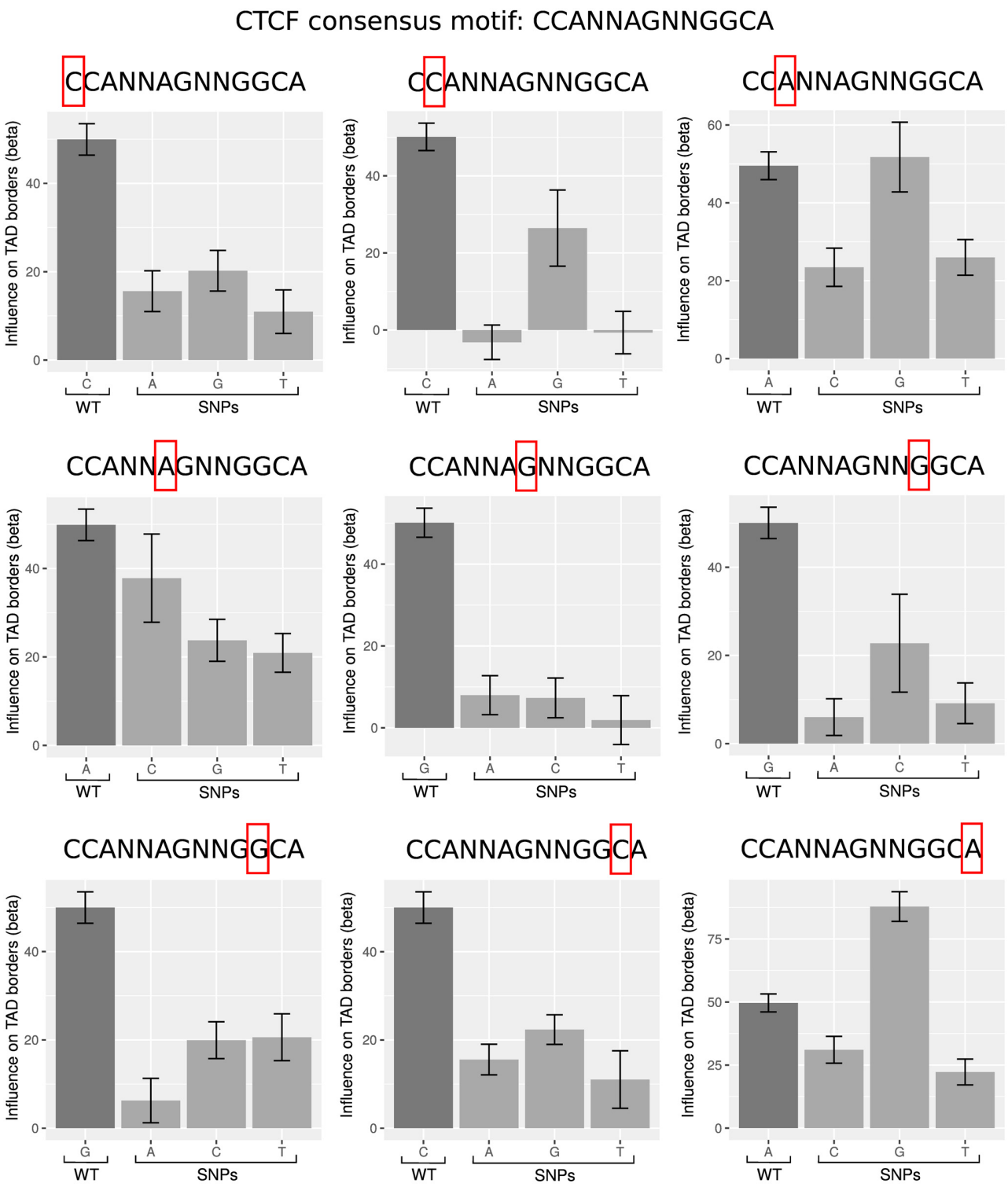


Fig 4. Analysis of the impacts of single nucleotide polymorphisms on the consensus CTCF motif in human GM12878 cells.

doi:10.1371/journal.pcbi.1004908.g004

BEAF-32 influences TAD borders in *Drosophila*

We implemented the proposed model such that it can deal with either genomic coordinate data or quantitative data. However, in the present study, we chose to focus on genomic coordinate data as in [11, 34]. An advantage of this approach was that both DNA-binding proteins and functional elements could be included within the same model. In addition, we observed that logistic regression models built from genomic coordinate data usually outperformed those obtained with quantitative data in terms of deviance ratio and AIC (model deviance ratios and AICs are given in S1 Table).

The influences of genomic features such as DNA-binding proteins or gene transcription on TAD border establishment or maintenance can be estimated by the proposed multiple logistic regression. Using *Drosophila* Kc167 cell Hi-C data at 1 kb resolution, we assessed the effects of insulator binding proteins, cofactors, gene transcription and functional elements on TAD borders. Although TADs were computed from 1 kb resolution Hi-C data, genomic features were binned at an even higher resolution of 50 bp in order to better discriminate between genomic features that influence TAD borders and those that do not, and to reduce standard errors of model parameters (see Subsection Materials and Methods, Binned data matrix). In this subsection, we first focused on the effects of insulator binding proteins in driving TAD borders [35].

In *Drosophila*, there are five subclasses of insulator sequences [36]. Each subclass is bound by a particular type of insulator binding protein (IBP): suppressor of hairy wing (Su(Hw)), *Drosophila* CTCF (dCTCF), boundary-element-associated factor of 32 kDa (BEAF-32), GAGA binding factor (GAF), and Zeste-White 5 (ZW5) [10]. In addition, the general transcription factor dTFIIIC was recently identified as a new IBP [11]. We assessed enrichments of these IBPs within TAD borders (Fig 5). We observed enrichments for all these IBPs (all coefficients $\hat{\beta} > 1.34$ and all p-values $p < 1 \times 10^{-20}$). BEAF-32 was the most enriched IBP with a

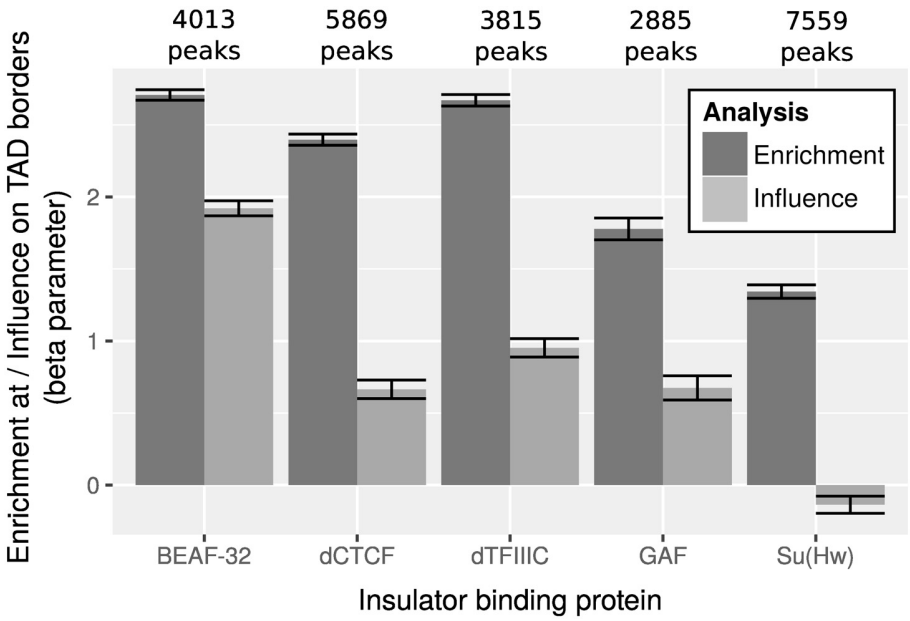


Fig 5. Comparison between enrichments by enrichment tests and influences by multiple logistic regression of insulator binding proteins at topologically associating domain (TAD) borders of wild-type *Drosophila* Kc167 cells. In both enrichment test and multiple logistic regression, beta parameters are computed and displayed. Error bars show 95% confidence intervals of beta parameters.

doi:10.1371/journal.pcbi.1004908.g005

coefficient $\hat{\beta} = 2.71$, corresponding to an odds ratio $\hat{OR} = 15.03$, whereas GAF was the least enriched IBP with a coefficient $\hat{\beta} = 1.34$, corresponding to an odds ratio $\hat{OR} = 3.82$.

Multiple logistic regression yielded different results (Fig 5). All beta coefficients decreased reflecting colocalization among the proteins (average correlation of 0.28). Despite these correlations, the tight 95% confidence intervals reflect that betas were estimated with low standard errors. This is due to the very large number of observations (>1 million) compared to the low number of variables (6 variables) obtained for a binning at 50 bp. There were clear differences of betas among the IBPs compared with enrichment analysis [5, 6]. Only BEAF-32 showed high and significant beta (BEAF-32: $\hat{\beta} = 1.92, p < 1 \times 10^{-20}$). For other IBPs, betas were significant but much lower ($\hat{\beta} < 0.95, p < 1 \times 10^{-20}$). Thus although dCTCF, dTFIIIC, GAF and Su(Hw) were enriched at TAD borders, multiple logistic regression revealed that they weakly influence TAD borders. High enrichments of these proteins are due to their correlations with BEAF-32. For instance, previous work showed that numerous dCTCF sites align tightly with BEAF-32 [37]. These results supported the role of BEAF-32 as most influential IBP of TAD borders.

Architectural proteins impact more TAD-based organization than transcription

There has been an ongoing debate to know whether transcription or architectural proteins are the main cause of TAD border demarcation [6]. Using enrichment test, we observed that active transcription start sites (TSSs) were enriched at TAD borders ($\hat{\beta} = 1.82, p < 1 \times 10^{-20}$), as well as architectural proteins such as BEAF-32 ($\hat{\beta} = 2.72, p < 1 \times 10^{-20}$). Using multiple logistic regression, we then estimated the effects of transcription and of architectural proteins on TAD borders within the same model (S6 Fig). We observed that active TSSs had a significant positive effect in TAD border establishment/maintenance ($\hat{\beta} = 0.42, p < 1 \times 10^{-20}$). This effect was much lower than the one of architectural protein BEAF-32 ($\hat{\beta} = 2.59, p < 1 \times 10^{-20}$). Our model thus reveals that architectural protein BEAF-32 contributes much more to TAD-based organization than transcription. However one might argue that the comparison between active TSSs and BEAF-32 was not straightforward because the latter represented two distinct genomic features, a functional element and a protein, respectively. Hence for a proper comparison between transcription and architectural proteins, we compared within the same multiple logistic regression the effects of the short isoform of *Drosophila* Brd4 homologue (Fs(1)h-S), a major transcriptional factor involved in transcriptional activation, with the long isoform (Fs(1)h-L), a recently identified architectural protein [38]. We observed that Fs(1)h-S had a significant positive effect on TAD borders ($\hat{\beta} = 1.87, p < 1 \times 10^{-20}$), but which was lower than the one of Fs(1)h-L ($\hat{\beta} = 2.60, p < 1 \times 10^{-20}$). Our results thus highlighted the prevalent roles of architectural proteins compared to transcription, which was highly consistent with recent results suggesting a lower impact of transcription [13].

The role of cofactors in *Drosophila*

Recent work supported the idea that IBPs may favor long-range contacts by recruiting cofactors directly involved in stabilizing long-range contacts [8–10]. In *Drosophila*, several cofactors were identified: condensin I, condensin II, Chromator, centrosomal protein of 190 kDa (CP190), cohesin [10, 13, 39, 40] and Fs(1)h-L [38]. We first analyzed by multiple logistic regression all abovementioned cofactors in their own to understand their relative contribution to TAD borders (S7 Fig). Among the cofactors, CP190 had the highest influence on TAD borders in agreement with previous findings [5] ($\hat{\beta} = 1.12, p < 1 \times 10^{-20}$). Because cofactors were

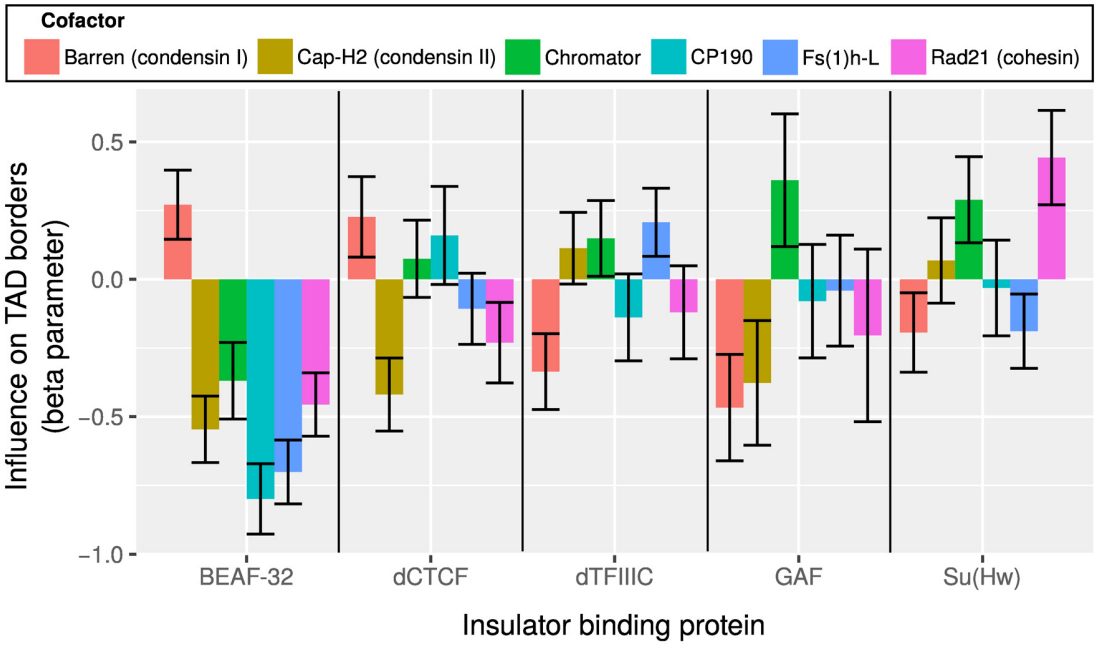


Fig 6. Analysis of interactions between insulator binding proteins (IBPs) and cofactors at topologically associating domain (TAD) borders of wild-type *Drosophila* Kc167 cells. Beta parameter corresponding to each interaction IBP-cofactor from the multiple logistic regression is plotted. Interaction terms are detailed in Subsection Materials and Methods, Analysis of interactions. Error bars show 95% confidence intervals of beta parameters. Barren is a subunit of condensin I, Cap-H2 is a subunit of condensin II and Rad21 is a subunit of cohesin.

doi:10.1371/journal.pcbi.1004908.g006

expected to be recruited by IBPs to the chromatin [8, 9, 39, 40], we then regressed cofactors with all IBPs and all IBP-cofactor interactions (see S2 Table). We observed that CP190 still presented a high beta ($\hat{\beta} = 1.13$, $p < 1 \times 10^{-20}$), which reflect that additional IBPs are able to recruit these cofactors in concordance with recent results [41].

An important question is to know if IBPs demarcate TAD borders depending on the presence of specific cofactors [10]. To answer this question, we assessed if the co-occurence of an IBP with a cofactor could affect TAD borders by estimating the corresponding statistical interaction IBP-cofactor (Fig 6). Among the significant positive interactions, we reported effects for

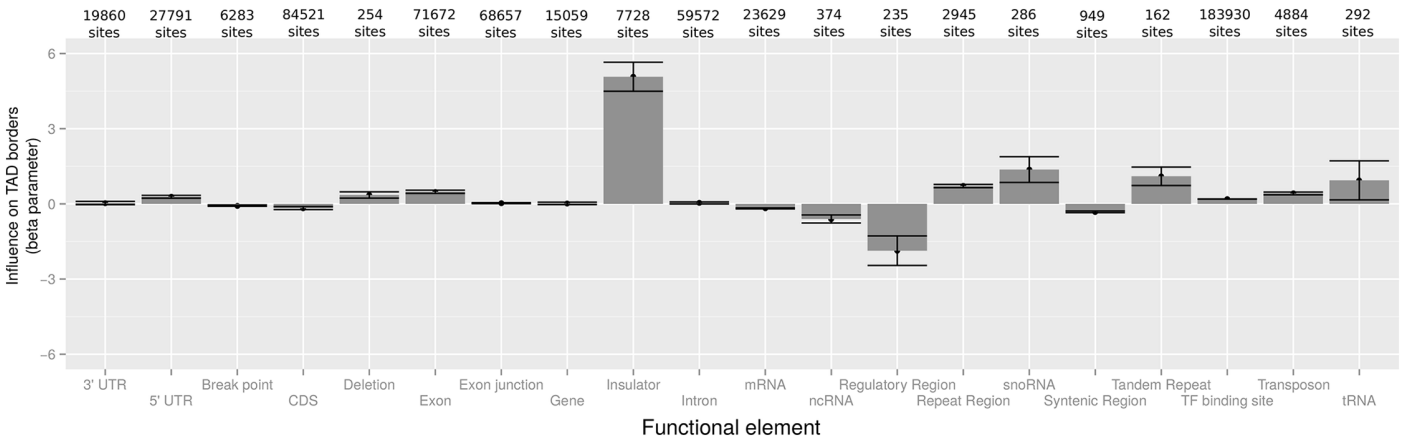


Fig 7. Analysis of functional elements using multiple logistic regression at topologically associating domain (TAD) borders of wild-type *Drosophila* Kc167 cells. Error bars show 95% confidence intervals of beta parameters.

doi:10.1371/journal.pcbi.1004908.g007

Su(Hw) with Rad21 ($\hat{\beta} = 0.44$, $p = 3 \times 10^{-7}$), and lower effects of Su(Hw) with Chromator ($\hat{\beta} = 0.29$, $p = 2 \times 10^{-4}$), BEAF-32 with condensin I (Barren) ($\hat{\beta} = 0.27$, $p = 2 \times 10^{-5}$), dTFIIIC with Fs(1)h-L ($\hat{\beta} = 0.21$, $p = 0.001$), dCTCF with condensin I (Barren) ($\hat{\beta} = 0.23$, $p = 2 \times 10^{-3}$). These positive interactions reflected synergistic effects of IBPs with cofactors. We did not report any significant positive statistical interaction between dCTCF and cohesin as observed in human [8]. In contrast to vertebrates, *Drosophila* CTCF does not appear to rely on cohesin to establish or maintain interactions [42]. Of interest, our method further highlighted strong and significant negative interactions that revealed antagonistic effects at domain borders, in particular for BEAF-32 with cofactor CP190 ($\hat{\beta} = -0.80$, $p < 1 \times 10^{-20}$). As such, our model may allow to retrieve both synergistic and antagonistic influences of co-factors, which may better reflect the complexity behind the establishment or maintenance of TAD borders.

Analysis of functional elements in *Drosophila*

We sought to further investigate a wide variety of functional elements such as insulators and regulatory sequences. Results are reported in Fig 7. Insulators were by far the most influential functional elements with respect to domain borders ($\hat{\beta} = 5.07$, $p < 1 \times 10^{-20}$), as established in human [8, 31]. Regarding other functional elements, we found positive effects for repeat regions ($\hat{\beta} = 0.71$, $p < 1 \times 10^{-20}$), and especially for tandem repeats on TAD borders ($\hat{\beta} = 1.10$, $p = 5 \times 10^{-9}$). Repeat regions were previously reported to spatially cluster together [43]. In addition, snoRNA genes had a positive influence on domain borders ($\hat{\beta} = 1.37$, $p = 1 \times 10^{-7}$), which may reflect their role in higher-order chromatin structure [44]. Furthermore, a negative impact on TAD border was detected for regulatory sequences ($\hat{\beta} = 1.87$, $p = 6 \times 10^{-10}$), strengthening the hypothesis that functional long-range contacts involving regulatory elements could compete with structural contacts [45] (see Discussion).

Positive and negative effects of proteins in human

We next analyzed the effects of DNA-binding proteins on 3D domains of human genome where fewer architectural proteins have been uncovered [29]. To investigate the possible contributions of these proteins, we analyzed new 3D domains detected from recent high resolution Hi-C data at 1 kb for GM12878 cells for which a large number of ChIP-seq data were available [8]. Over the 69 proteins analyzed, 51 proteins presented very high and significant enrichments (all coefficients $\hat{\beta} > 3$ and all p-values $p < 1 \times 10^{-20}$). Multiple logistic regression instead detected 15 proteins with significant positive effects on domain borders (all coefficients $\hat{\beta} > 0.5$ and all p-values $p < 5 \times 10^{-4}$; S3 Table). Our analyses confirmed that, in contrast to *Drosophila*, CTCF and cohesin (subunit Rad21) presented the highest effects among all factors (CTCF: $\hat{\beta} = 1.90$, $p < 1 \times 10^{-20}$; cohesin: $\hat{\beta} = 1.91$, $p < 1 \times 10^{-20}$), in complete agreement with numerous studies showing their important roles in shaping chromosome 3D structure in mammals [8, 9, 12]. ZNF143 had the third highest effect ($\hat{\beta} = 1.85$, $p < 1 \times 10^{-20}$), in total agreement with a very recent study demonstrating its role in long-range contacts [46]. In addition, multiple logistic regression identified EZH2, the catalytic subunit of the Polycomb repressive complex 2 (PRC2), as a protein that significantly impacted TAD borders (4th highest effect: $\hat{\beta} = 1.32$, $p < 5 \times 10^{-11}$). In contrast, multiple logistic regression estimated a null beta for candidate architectural proteins JUND ($\hat{\beta} = 0.04$, $p = 0.85$), Kaiso ($\hat{\beta} = 0.43$, $p = 0.10$) and a very low beta for MAZ ($\hat{\beta} = 0.23$, $p = 3 \times 10^{-4}$). Although these three proteins colocalize or interact with CTCF, our model suggests that they might not impact TAD borders. We also

notably identified several factors associated with transcriptional activation that had significant negative influences on TAD borders. These proteins included RXRA ($\hat{\beta} = -1.37$, $p = 3 \times 10^{-4}$), P300 ($\hat{\beta} = -1.22$, $p = 1 \times 10^{-10}$), BCL11A ($\hat{\beta} = -0.82$, $p = 1 \times 10^{-9}$) and ELK1 ($\hat{\beta} = -0.74$, $p = 4 \times 10^{-9}$), reinforcing the view that transcription could also interfere with TAD borders depending on context.

Large-scale analysis of DNA motifs in human

In the previous subsection, analyses of DNA-binding proteins were limited by available ChIP-seq data. Here we alleviated this limitation by analyzing transcription factor binding site (TFBS) motifs available from the large MotifMap database [47]. Given the large number of TFBS motifs (544 motifs), we used L1-regularization for parameter estimation. We identified 213 positive drivers (all coefficients $\hat{\beta} > 1$) and 75 negative drivers (all coefficients $\hat{\beta} < 1$), meaning that a large number of TFBSs actually play a role in TAD border establishment or maintenance. CTCF motifs ranked first ($\hat{\beta} = 45.34$) in complete agreement with recent studies [8, 31]. But our model also uncovered other TFBSs whose roles in TAD borders are less well known such as EGR-1 ($\hat{\beta} = 34.04$), p53 ($\hat{\beta} = 25.55$), MIZF ($\hat{\beta} = 22.46$), GABP ($\hat{\beta} = 21.94$) and many others (for a complete list, see S4 Table). For instance, p53 is a major tumor suppressor gene and the most frequently mutated gene (>50%) in human cancer [48]. Regarding negative drivers, we identified ALX4 ($\hat{\beta} = -35.82$), EGR4 ($\hat{\beta} = -26.72$), ZNF423 ($\hat{\beta} = -23.97$). All these results highlighted the great potential of TFBS motif analysis allowing the study of a very large number of DNA-binding proteins.

Discussion

Here, we describe a multiple logistic regression (MLR) to assess the roles of genomic features such as DNA-binding proteins and functional elements on TAD border establishment/maintenance. Based on conditional independence, such regression model can identify genomic features that impact TAD borders, unlike enrichment test (ET) and non-parametric models. Using simulations, we demonstrate that model parameters can be accurately estimated for both marginal genomic features (no interaction) and two-way interactions. In addition, we show that our model outperforms enrichment test and random forests for the identification of genomic features that influence domain borders. Using recent experimental Hi-C and ChIP-seq data, the proposed model can identify genomic features that are most influential with respect to TAD borders at a very high resolution of 1 kb in both *Drosophila* and human. The proposed model could thus guide the biologists for the design of most critical Hi-C experiments aiming at unraveling the key molecular determinants of higher-order chromatin organization.

Enrichment test shows slight differences of enrichments among architectural proteins. This could suggest that domain borders are determined by the number and levels of all proteins present at the border rather than the presence of specific proteins [11, 13]. However MLR instead reveals that only some architectural proteins influence the presence of 3D domain borders. Moreover, MLR retrieves both positive and negative contributions among most influencial proteins, depending on contexts such as co-occurence. From these novel results, we propose a biological model for 3D domain border establishment or maintenance (Fig 8). In this model, three kinds of proteins are distinguished: positive drivers ($\beta_{MLR} > 0$), negative drivers ($\beta_{MLR} < 0$), and proteins that are enriched or depleted at borders but are not drivers ($\beta_{ET} > 0$ or $\beta_{ET} < 0$, and $\beta_{MLR} = 0$). Positive drivers favor attraction between domain borders leading to the formation of 3D domains. CTCF and cohesin are well-studied positive drivers in

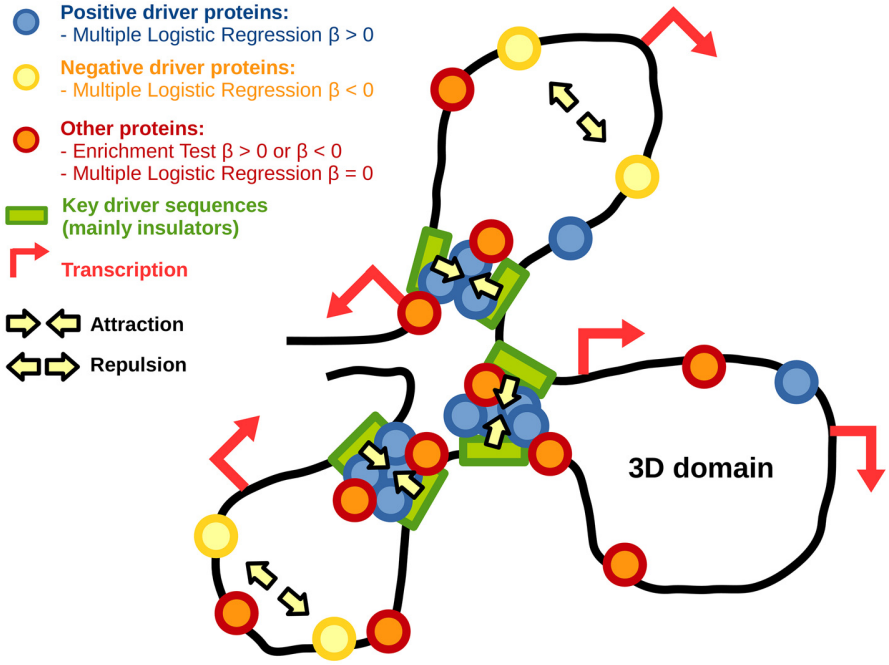


Fig 8. Model for 3D domain border establishment or maintenance.

doi:10.1371/journal.pcbi.1004908.g008

mammals [8, 10]. By contrast little is known about negative drivers of 3D domain borders that could favor repulsion between specific chromatin regions [49]. Repulsion phenomenon could be the result of allosteric effects of loops in chromatin [45]. Negative drivers could also regulate disassembly of protein complex that mediate long-range contacts [50].

In *Drosophila*, MLR identifies BEAF-32, a well-characterized IBP, as a positive driver of TAD borders [51, 52]. Conversely, other IBPs including dCTCF, dTFIIIC, GAF and Su(Hw) are found significantly enriched at TAD borders, but present weak or no influences, in agreement with recent works [53]. Regarding cofactors, CP190 presents a high and significant positive influence on domain demarcation, in agreement with previous findings [5]. Regarding functional elements, although our data highlight that insulators are by far the main positive drivers of TAD borders, they also show that additional elements, that are known to colocalize in 3D [18, 43, 44], play a role including repeat regions. Moreover, MLR suggests that snoRNA genes are novel functional elements that positively influence border demarcation. Recent works suggest that active chromatin and transcription also play a key role in chromosome partitioning in TADs [53]. Here our results reveal that both architectural proteins and transcription contribute to TAD borders. In contrast, regulatory regions are identified as negative drivers of TAD borders. One possible explanation is that such regulatory regions are involved in functional long-range contacts with gene promoters that would compete with the formation of more structural contacts at the origin of TADs [45]. Alternatively, a negative influence may be linked to the transient nature of certain functional contacts [54].

Almost half of dCTCF and cohesin sites are overlapping in *Drosophila*, and knockdown of dCTCF results in a strong decrease of cohesin binding [11]. As such, one might expect synergistic effects of dCTCF with cohesin (also called statistical interaction) in driving TAD borders. However, such conclusion could not be drawn. Following statistical theory, it is not because two variables are correlated (here dCTCF and cohesin colocalize), that it implies a synergistic effect of the two variables on TAD borders. Although dCTCF and cohesin are both enriched at

TAD borders, MLR does not detect a significant interaction of dCTCF with cohesin. Instead we observe a high interaction of Su(Hw) with cohesin. Negative interactions that reflect antagonistic effects between architectural proteins are found between IBP BEAF-32 and cofactor CP190. These antagonistic effects suggest that cofactors might not always help IBPs in stabilizing loops [10]. One explanation is that cofactors could sometimes compete with IBPs for long-range protein-protein interactions.

In human, MLR identifies well-studied architectural proteins CTCF and cohesin as the most influential positive drivers of 3D domains, in complete agreement with their established roles in shaping chromosome 3D structure [8, 9, 12]. MLR also points out the positive influences of ZNF143 and PRC2 proteins whose recent studies have uncovered their roles in controlling spatial organization [30, 46]. In addition, our model reveals the roles of additional factors including RXRA, P300, BCL11A and ELK1 as negative drivers of 3D domain borders. P300 was previously shown to be depleted at domain borders [55]. Here we find that P300 and three other proteins can counteract the establishment or maintenance of domain borders. P300 is a well-known regulator of cell growth and division, and helps prevent the growth of cancerous tumors [56]. Interestingly, the three other proteins RXRA, BCL11A and ELK1 are also related to cancer [57–59]. Furthermore, the analysis of a large number of TFBS motifs confirmed the role of CTCF in TAD border formation [8, 31]. But this analysis also uncovered many other TFBSs, such as p53, a major tumor suppressor gene [48].

The proposed method relies on the accurate identification of 3D domains. To further improve our understanding of the key drivers of 3D domain borders, Hi-C experiments at a higher resolution are needed. In addition, a variety of methods have been recently developed for 3D domain inference, and no consensus has been reached yet to determine which method is the most appropriate. Another important question is to understand the roles of key drivers in chromatin interactions within domains. For instance, it is essential to identify proteins that influence functional interactions between enhancers and promoters that regulate gene expression. Although far more complex, it is of note that similar regression approach may largely help in retrieving positive from negative patterns in these contexts.

Materials and Methods

Hi-C data and topologically associating domains

For *Drosophila* 3D domain analysis, we used publicly available high-throughput chromatin conformation capture (Hi-C) data from Gene Expression Omnibus (GEO) accession GSE63515 [13]. Hi-C experiments were done for wild-type *Drosophila melanogaster* Kc167 cells with DpnII restriction enzyme. Hi-C data were binned at 1 kb resolution. Contact matrices were normalized using ICE method [15] implemented in the R package HiTC (http://www.bioconductor.org/packages//2.11/bioc/html/HiTC.html). From the normalized contact matrices, TAD genomic coordinates were identified using HiCseg method [19].

For human 3D domain analysis, we used publicly available 3D domains of GM12878 cells identified by the Arrowhead algorithm from Gene Expression Omnibus (GEO) accession GSE63525 [8].

ChIP-seq data

For *Drosophila* analysis, we used publicly available binding profiles of chromatin proteins of *Drosophila melanogaster* wild-type embryonic Kc167 cells. ChIP-seq data for CP190, Su(Hw), dCTCF and BEAF-32 were obtained from GEO accession GSE30740 [60]. ChIP-seq data for Barren (condensin I), Cap-H2 (condensin II), Chromator, Rad21 (cohesin), GAF and dTFIIIC were obtained from GEO accession GSE54529 [11]. ChIP-seq data for Fs(1)h-L and Fs(1)h-LS

were obtained from GEO accession GSE42086 [38]. ChIP-seq peaks were called using MACS 1.4.2 (https://github.com/taoliu/MACS). Fs(1)h-S peaks were defined as peaks from Fs(1)h-LS that did not overlap any Fs(1)h-L peak.

For human analysis, we used publicly available ChIP-seq peaks of 69 chromatin proteins (ATF2, ATF3, BATF, BCL11A, BCL3, BCLAF1, BHLHE40, BRCA1, CEBPB, CHD1, CHD2, CTCF, E2F4, EBF1, EGR1, ELF1, ELK1, ETS1, EZH2, FOS, FOXM1, IKZF1, IRF3, IRF4, JUND, MAFK, MAX, MAZ, MEF2A, MEF2C, MTA3, MXI1, MYC, NFATC1, NFE2, NFIC, NFYA, NFYB, NRF1, P300, PAX5, PBX3, PIGG, PML, POU2F2, RAD21, REST, RFX5, RUNX3, RXRA, SIN3A, SIX5, SP1, SRF, STAT1, STAT3, STAT5A, TAF1, TCF12, TCF3, USF1, USF2, YY1, ZBTB33, ZEB1, ZNF143, ZNF274, ZNF384 and ZZZ3) of GM12878 cells from ENCODE [61].

Functional elements

For *Drosophila* analysis, we used RNA-seq data from wild-type Kc167 cells to map active transcription start sites (TSSs) [62]. For all other functional elements, we used flybase reference genome annotation (http://flybase.org/).

DNA motifs

For human analysis, we used transcription factor binding site (TFBS) motifs from the MotifMap database (http://motifmap.ics.uci.edu/).

Binned data matrix

From TAD coordinates, ChIP-seq data and functional element mapping, we constructed 50-base and 1-kb binned data matrices that were further used for multiple logistic regressions with *Drosophila* and human data, respectively. A matrix was composed of a column variable Y that indicated if the genomic bin belonged to a TAD boundary ($Y = 1$) or not ($Y = 0$). To define TAD boundaries, we extracted 1 kb and 20 kb regions that were centered around the positions demarcating two TADs in *Drosophila* and human genomes, respectively. The other column variables $\mathbf{X} = \{X_1, \ldots, X_p\}$ were the set of p genomic feature variables of interest. If genomic coordinate data were used (*e.g.*, ChIP-seq peak or functional element coordinates), variable X_i denoted the presence ($X_i = 1$) or absence ($X_i = 0$) of the genomic feature i within the genomic bin. Note that if a genomic coordinate only overlapped x% of the genomic bin, then $X_i = x$%. If quantitative data were used (*e.g.*, ChIP-seq signal intensity log(*ChIP/Input*)), variable X_i was the average value within the genomic bin.

Enrichment test

Enrichment test assesses the enrichment of a genomic feature within chromatin domain borders. The genomic feature of interest can be protein-DNA binding sites detected from ChIP-seq experiment. Chromatin domain borders can be borders between topologically associating domains identified from Hi-C experiment.

From the contingency table (Table 1), one can test the odds ratio that reflects the magnitude of enrichment ($OR > 1$) or depletion ($OR < 1$) of the genomic feature within the domain borders. The test consists in assessing the following null (H_0) and alternative (H_1) hypotheses about odds ratio OR:

$$H_0 : OR = 1 \tag{2}$$

$$H_1 : OR \neq 1 \tag{3}$$

Table 1. Example of a contingency table to assess enrichment (or depletion) of a genomic feature within the domain borders.

	Presence of the feature	Absence of the feature
Inside border	500	5000
Outside border	2000	200000

doi:10.1371/journal.pcbi.1004908.t001

The odds ratio is the ratio of the inside border odds (500/5000) to the outside border odds (2000/2000000). Here $\hat{OR} = \frac{500/5000}{2000/200000} = 10$.

Previous enrichment test can be reformulated as a simple logistic regression model:

$$\ln \frac{Prob(Y = 1|X_i)}{1 - Prob(Y = 1|X_i)} = \beta_0 + \beta X_i \quad (4)$$

Variables $X_i \in \mathbf{X}$ and Y are described in Subsection Materials and Methods, Binned data matrix. In the simple logistic regression, the slope parameter β is the natural logarithm of the abovementioned odds ratio OR. Thus $\beta > 0$ means enrichment, while $\beta < 0$ reflects depletion. Using logistic regression model, parameter β can be tested by Wald's test. The Wald's statistic is calculated as:

$$W = \frac{\hat{\beta} - \beta_*}{\hat{\sigma}_\beta} = \frac{\hat{\beta} - 0}{\hat{\sigma}_\beta} = \frac{\hat{\beta}}{\hat{\sigma}_\beta} \quad (5)$$

Where β_* is the beta parameter value under H_0 assumption ($\beta_* = 0$) and $\hat{\sigma}_\beta$ denotes the standard error of parameter β. Statistic W follows a normal distribution.

An important drawback of enrichment test relies on the fact that it does not account for potential colocalizations (*i.e.* correlations) among the genomic features of interest. The presence of correlations might prevent the identification of the genomic features that really drive the establishment or maintenance of domain borders. For instance, if two genomic features are significantly enriched, this might not mean that both are involved in the establishment or maintenance of the borders. One feature might truly affect borders while the other feature might only be correlated to the former. There is thus a need for a model that could identify those enriched features that drive the presence of borders.

Multiple logistic regression

The proposed multiple logistic regression is an extension of the simple logistic regression for p genomic features:

$$\ln \frac{Prob(Y = 1|\mathbf{X})}{1 - Prob(Y = 1|\mathbf{X})} = \beta_0 + \beta \mathbf{X} \quad (6)$$

Where $\mathbf{X} = \{X_1, \ldots, X_p\}$ is the set of p genomic features of interest and $\boldsymbol{\beta} = \{\beta_1, \ldots, \beta_p\}$ denotes the set of slope parameters (one parameter for each genomic feature). As for simple logistic regression, each $\beta_i \in \boldsymbol{\beta}$ coefficient can be tested by a Wald's test.

By default, multiple logistic regression β_0 and $\boldsymbol{\beta}$ parameters are estimated by iteratively reweighted least squares. However, when there are a large number of correlated genomic features in the model, L1-regularization is applied and parameters are learned by coordinate descent [26]. The L1-regularization lambda that gives the lowest mean cross-validated error is selected. To assess quality of fit for a model, we use the deviance ratio defined as the ratio of the

fitted model deviance to the saturated model deviance. We also use Akaike information criterion (AIC).

The matrix **X** is sparse and the Wald's test might be biased when data are sparse [27]. Hence likelihood ratio test (LRT) that is not affected by data sparseness can be used instead. To test parameter β_i with LRT, two models are built: a first model $\mathcal{M}_1$ over all variables **X**, and a second model $\mathcal{M}_2$ over all variables except X_i ($\mathbf{X} \setminus X_i$). Then the following D_i statistic is calculated:

$$D_i = -2\ln\left(\frac{L_{\mathcal{M}_1}}{L_{\mathcal{M}_2}}\right) \tag{7}$$

Where $L_{\mathcal{M}_1}$ is the likelihood of $\mathcal{M}_1$ and $L_{\mathcal{M}_2}$ is the likelihood of $\mathcal{M}_2$. Statistic D_i follows a chi-squared distribution with one degree of freedom. The better accuracy of LRT comes at the cost of more intensive computations. In practice, we observe that Wald's test p-values are close to LRT p-values.

In the multiple logistic regression setting, parameter β_i measures the effect of genomic feature X_i on the presence of borders conditional on the other genomic features that belong to $\mathbf{X} \setminus X_i$. A value of $\beta_i > 0$ or $\beta_i < 0$ means that the genomic feature X_i positively or negatively influences the presence of borders, respectively. A value of $\beta_i = 0$ reflects the fact that the genomic feature X_i does not affect the presence of borders. If two genomic features X_1 and X_2 are colocalized and only X_1 drives the establishment or maintenance of domain borders, then only the corresponding β_1 parameter will be significantly different from zero. However the above formulation of the model does not account for potential statistical interactions between genomic features.

Analysis of interactions

Interaction terms can be included in the multiple logistic regression to account for potential interactions between genomic features. For instance, one can include in the model an interaction term between two genomic features X_1 and X_2:

$$\ln\frac{Prob(Y=1|X_1,X_2)}{1-Prob(Y=1|X_1,X_2)} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_{12} X_1 X_2 \tag{8}$$

The product $X_1 X_2$ is the statistical interaction term between the two genomic features X_1 and X_2. Parameter β_{12} measures the effect of interaction $X_1 X_2$ on the presence of borders.

Data simulation

In order to assess the accuracy of multiple logistic regression parameter estimation, we simulated data that were the most similar to the real genomic data using the following procedure. First, for a simulation *s*, a set of observation rows was randomly drawn with resampling from matrix **X** (nonparametric bootstrap). This resampling allowed to keep the original correlation structure among the variables. The bootstrapped data matrix was denoted $\mathbf{X}^s$. Second $\boldsymbol{\beta}^s = \{\beta_1^s, ..., \beta_p^s\}$ parameter values were drawn from a normal distribution $\mathcal{N}(\mu, \sigma)$ with mean $\mu = 0$ and variance $\sigma = 1$. Parameter β_0^s (intercept) value was drawn from a normal distribution with same variance but with mean $\mu = -4.5$. This setting of the mean of β_0^s allowed to control the number of values $Y = 1$ close to the one observed from real data (the number of borders in real data was low). Third a quantitative variable Z^s was calculated using the regression formula: $Z^s = \beta_0^s + \boldsymbol{\beta}^s \mathbf{X}^s$. A probability variable $Prob^s$ was calculated by the inverse logit function: $1/(1 + exp(-Z^s))$. Then each probability value from $Prob^s$ was used to draw a value for Y^s using binomial distribution.

We also used simulated data to compare multiple logistic regression with enrichment test and random forests. As previously, for a simulation s, we used non-parametric bootstrap and kept the correlation structure of original data. Among the variables, a subset of variables $\mathbf{X}_c \in \mathbf{X}$ was chosen to be causal, *i.e.* to influence the presence of borders. We chose a generative model that was non-linear and non-additive not to favor multiple logistic regression over other models. For this purpose, we set a probability p_0 of the presence of a border in a bin if all causal variable values were inferior to 0.5. We also set a probability p_1 (with $p_1 > p_0$) if at least one causal variable had a value superior or equal to 0.5. Values of p_0 and p_1 were chosen according to the number of borders in real data. Then, for each bin, the value for Y^s was drawn using a binomial distribution with either p_0 or p_1 depending on the causal variable values.

Implementation and availability

The multiple logistic regression is implemented in R language. The model is available in the R package "HiCfeat" which can be downloaded from the Comprehensive R Archive Network and from the web page of Raphaël Mourad (https://sites.google.com/site/raphaelmouradeng/home/programs).

Supporting Information

S1 Table. Deviance ratios and Akaike information criteria obtained for multiple logistic regression models in wild-type *Drosophila* Kc167 cells.
(PDF)

S2 Table. Multiple logistic regression including insulator-binding proteins (IBPs), cofactors and IBP-cofactor interactions at topologically associating domain borders of wild-type *Drosophila* Kc167 cells.
(PDF)

S3 Table. Multiple logistic regression including DNA-binding proteins in human GM12878 cells at 3D domain borders. Here 3D domains identified by the Arrowhead algorithm were used.
(PDF)

S4 Table. Multiple logistic regression including transcription factor binding site (TFBS) motifs in human GM12878 cells at 3D domain borders. Here 3D domains identified by the Arrowhead algorithm were used.
(PDF)

S1 Fig. Parameter estimation accuracy of multiple logistic regression for simulated proteins with varied numbers of ChIP-seq peaks.
(PDF)

S2 Fig. Impact of the inaccuracy of topologically associating domain (TAD) borders on multiple logistic regression beta parameters. R squared is computed between beta parameters estimated from TAD borders and beta parameters estimated from TAD borders with random noise. Random noise was drawn from a normal distribution of mean zero and varying standard deviations in kb (x-axis).
(PDF)

S3 Fig. Comparison of multiple logistic regression (MLR) with enrichment test (ET) and random forests (RF) to detect known and suspected architectural proteins in human using GM12878 cell ChIP-seq data binned at 40 kb resolution. Receiver operating characteristic

(ROC) curves were computed from Wald's statistics for ET, beta parameters for MLR, and variable importances for random forests.
(PDF)

S4 Fig. Comparison of MLR with ET and RF to detect the influences of single nucleotide polymorphisms (SNPs) in the CTCF motif on 3D domains in human. Receiver operating characteristic (ROC) curves were computed from Wald's statistics for ET, from beta parameters for MLR, and from variable importances for random forests. Computations were carried out at 1 kb resolution.
(PDF)

S5 Fig. Analysis of the impacts of single nucleotide polymorphisms on the consensus BEAF-32 motif in wild-type *Drosophila* Kc167 cells.
(PDF)

S6 Fig. Comparison of the influences of transcription and of architectural proteins on topologically associating domain borders of wild-type *Drosophila* Kc167 cells. a) Multiple logistic regression of active TSSs and BEAF-32. b) Multiple logistic regression of Fs(1)h-S and Fs(1)h-L.
(PDF)

S7 Fig. Multiple logistic regression of cofactors at topologically associating domain borders of wild-type *Drosophila* Kc167 cells.
(PDF)

Acknowledgments

The authors thank Pascal Martin and Laurent Lacroix for useful discussions. The authors are grateful to Corces lab (Emory University, USA) and Cavalli lab (Institute of Human Genetics, France) for data and for help in processing them.

Author Contributions

Conceived and designed the experiments: RM. Performed the experiments: RM. Analyzed the data: RM. Contributed reagents/materials/analysis tools: RM. Wrote the paper: RM OC.

References

1. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009 Oct; 326(5950):289–293. doi: 10.1126/science.1181369 PMID: 19815776
2. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics. 2013 Jun; 14(6):390–403. doi: 10.1038/nrg3454 PMID: 23657480
3. Hu M, Deng K, Qin Z, Liu JS. Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. Quantitative Biology. 2013 May; 1(2):156–174. doi: 10.1007/s40484-013-0016-0 PMID: 26124977
4. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012 May; 485(7398):376–380. doi: 10.1038/nature11082 PMID: 22495300
5. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell. 2012 Feb; 148(3):458–472. doi: 10.1016/j.cell.2012.01.010 PMID: 22265598
6. Hou C, Li L, Zhaohui SQ, Corces VG. Gene density, transcription, and insulators contribute to the partition of the *Drosophila* genome into physical domains. Molecular Cell. 2012 November; 48(3):471–484. doi: 10.1016/j.molcel.2012.08.031 PMID: 23041285

7. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013 November; 503(7475):290–294. doi: 10.1038/nature12644 PMID: 24141950

8. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2015 Feb; 159 (7):1665–1680. doi: 10.1016/j.cell.2014.11.021

9. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013 Jun; 153 (6):1281–1295. doi: 10.1016/j.cell.2013.04.053 PMID: 23706625

10. Phillips-Cremins JE, Corces VG. Chromatin insulators: Linking genome organization to cellular function. Molecular Cell. 2013 May; 50(4):461–474. doi: 10.1016/j.molcel.2013.04.018 PMID: 23706817

11. Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, Qin ZS, et al. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biology. 2014 June; 15(5): R82+. doi: 10.1186/gb-2014-15-5-r82 PMID: 24981874

12. Zuin J, Dixon JR, van der Reijden MIJA, Ye Z, Kolovos P, Brouwer RWW, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proceedings of the National Academy of Sciences. 2014 October; 111(3):996–1001. Available from: http://www.pnas.org/content/111/3/996.abstract. doi: 10.1073/pnas.1317788111

13. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong CT, et al. Widespread rearrangement of 3D chromatin organization underlies Polycomb-mediated stress-induced silencing. Molecular Cell. 2015 March; (15:):S1097–2765.

14. Yaffe E, Tanay A. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. Nature Genetics. 2011 November; 43(11):1059–1065. doi: 10.1038/ng.947 PMID: 22001755

15. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods. 2012 Oct; 9 (10):999–1003. doi: 10.1038/nmeth.2148 PMID: 22941365

16. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012 Dec; 28(23):3131–3133. doi: 10.1093/bioinformatics/bts570 PMID: 23023982

17. Paulsen J, Lien TG, Sandve GK, Holden L, Borgan Ø, Glad IK, et al. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. Nucleic Acids Research. 2013 May; 41 (10):5164–5174. doi: 10.1093/nar/gkt227 PMID: 23571755

18. Ay F, Bailey TL, Noble WS. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. Genome Research. 2014 Jun; 24(6):999–1011. doi: 10.1101/gr.160374.113 PMID: 24501021

19. Levy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014; 30(17):i386–i392. doi: 10.1093/bioinformatics/btu443 PMID: 25161224

20. Hu M, Deng K, Qin Z, Dixon J, Selvaraj S, Fang J, et al. Bayesian inference of spatial organizations of chromosomes. PLoS Computational Biology. 2013 Jan; 9(1):e1002893+. doi: 10.1371/journal.pcbi.1002893 PMID: 23382666

21. Lesne A, Riposo J, Roger P, Cournac A, Mozziconacci J. 3D genome reconstruction from chromosomal contacts. Nature Methods. 2014 Nov; 11(11):1141–1143. doi: 10.1038/nmeth.3104 PMID: 25240436

22. Jost D, Carrivain P, Cavalli G, Vaillant C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Research. 2014 Aug; 42(15):9553–9561. doi: 10.1093/nar/gku698 PMID: 25092923

23. Huang J, Marco E, Pinello L, Yuan GC. Predicting chromatin organization using histone marks. Genome Biology. 2015; 16(1):162. Available from: http://genomebiology.com/2015/16/1/162. doi: 10.1186/s13059-015-0740-z PMID: 26272203

24. Sefer E, Kingsford C. Semi-nonparametric modeling of topological domain formation from epigenetic data. In: Pop M, Touzet H, editors. Algorithms in Bioinformatics. vol. 9289 of Lecture Notes in Computer Science. Springer Berlin Heidelberg; 2015. p. 148–161.

25. Shmueli G. To Explain or to Predict? Statistical Science. 2010; 25(3):289–310. doi: 10.1214/10-STS330

26. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B (Methodological). 1996 January; 58(1):267–288.

27. Hosmer DW, Lemeshow S. Applied logistic regression (Wiley Series in probability and statistics). 2nd ed. Wiley-Interscience Publication; 2000. Available from: http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0471356328.

28. Botta M, Haider S, Leung IX, Lio P, Mozziconacci J. Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Molecular Systems Biology. 2010 Nov; 6:426. doi: 10.1038/msb.2010.79 PMID: 21045820

29. Cubeñas-Potts C, Corces VG. Architectural proteins, transcription, and the three-dimensional organization of the genome. FEBS Letters. 2015; 589(20PartA):2923–2930. doi: 10.1016/j.febslet.2015.05.025 PMID: 26008126

30. Schoenfelder S, Sugar R, Dimond A, Javierre BM, Armstrong H, Mifsud B, et al. Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. Nature Genetics. 2015 Aug; 47(10):1179–1186. doi: 10.1038/ng.3393 PMID: 26323060

31. Sanborn AL, Rao SSP, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proceedings of the National Academy of Sciences. 2015 November; 112(47):E6456–E6465. Available from: http://www.pnas.org/content/early/2015/10/22/1518552112.abstract. doi: 10.1073/pnas.1518552112

32. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015 Sep; 161(5):1012–1025. doi: 10.1016/j.cell.2015.04.004 PMID: 25959774

33. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2016 Feb; 163(7):1611–1627. doi: 10.1016/j.cell.2015.11.024

34. Welch RP, Lee C, Imbriano PM, Patil S, Weymouth TE, Smith RA, et al. ChIP-Enrich: gene set enrichment testing for ChIP-seq data. Nucleic Acids Research. 2014 May; 42(13):e105. doi: 10.1093/nar/gku463 PMID: 24878920

35. Van Bortle K, Corces VG. The role of chromatin insulators in nuclear architecture and genome function. Current Opinion in Genetics & Development. 2013; 23(2):212–218. doi: 10.1016/j.gde.2012.11.003

36. Gurudatta BV, Corces VG. Chromatin insulators: lessons from the fly. Briefings in Functional Genomics & Proteomics. 2009 July; 8(4):276–282. Available from: http://bfg.oxfordjournals.org/content/8/4/276.abstract. doi: 10.1093/bfgp/elp032

37. Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi JE, Corces VG. *Drosophila* CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. Genome Research. 2012 Nov; 22(11):2176–2187. doi: 10.1101/gr.136788.111 PMID: 22722341

38. Kellner WA, Van Bortle K, Li L, Ramos E, Takenaka N, Corces VG. Distinct isoforms of the *Drosophila* Brd4 homologue are present at enhancers, promoters and insulator sites. Nucleic Acids Research. 2013 Nov; 41(20):9274–9283. doi: 10.1093/nar/gkt722 PMID: 23945939

39. Liang J, Lacroix L, Gamot A, Cuddapah S, Queille S, Lhoumaud P, et al. Chromatin immunoprecipitation indirect peaks highlight functional long-range interactions among insulator proteins and RNAII pausing. Molecular Cell. 2014 February; 53(4):672–681. doi: 10.1016/j.molcel.2013.12.029 PMID: 24486021

40. Vogelmann J, Le Gall A, Dejardin S, Allemand F, Gamot A, Labesse G, et al. Chromatin insulator factors involved in long-range DNA interactions and their role in the folding of the *Drosophila* genome. PLoS Genetics. 2014 august; 10(8):e1004544. doi: 10.1371/journal.pgen.1004544 PMID: 25165871

41. Maksimenko O, Bartkuhn M, Stakhov V, Herold M, Zolotarev N, Jox T, et al. Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin. Genome Research. 2015 January; 25(1):89–99. Available from: http://genome.cshlp.org/content/25/1/89.abstract. doi: 10.1101/gr.174169.114 PMID: 25342723

42. Dorsett D. Cohesin, gene expression and development: lessons from *Drosophila*. Chromosome Research. 2009; 17(2):185–200. doi: 10.1007/s10577-009-9022-5 PMID: 19308700

43. Tang SJ. Chromatin organization by repetitive elements (CORE): A genomic principle for the higher-order structure of chromosomes. Genes. 2011 Aug; 2(3):502–515. doi: 10.3390/genes2030502 PMID: 24710208

44. Schubert T, Pusch MCC, Diermeier S, Benes V, Kremmer E, Imhof A, et al. Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin. Molecular Cell. 2012 Nov; 48(3):434–444. doi: 10.1016/j.molcel.2012.08.021 PMID: 23022379

45. Doyle B, Fudenberg G, Imakaev M, Mirny LA. Chromatin loops as allosteric modulators of enhancer-promoter interactions. PLoS Computational Biology. 2014 Oct; 10(10):e1003867+. doi: 10.1371/journal.pcbi.1003867 PMID: 25340767

46. Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, et al. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. Nature Communications. 2015 February; 2:6186. Available from: http://view.ncbi.nlm.nih.gov/pubmed/25645053. doi: 10.1038/ncomms7186 PMID: 25645053

47. Xie X, Rigor P, Baldi P. MotifMap: a human genome-wide map of candidate regulatory motif sites. Bioinformatics. 2009; 25(2):167–174. Available from: http://bioinformatics.oxfordjournals.org/content/25/2/167.abstract. doi: 10.1093/bioinformatics/btn605 PMID: 19017655

48. Joerger AC, Fersht AR. The p53 pathway: Origins, inactivation in cancer, and emerging therapeutic approaches. Annual Review of Biochemistry. 2016; 85(1). Available from: http://www.annualreviews.org/doi/abs/10.1146/annurev-biochem-060815-014710.

49. Saberi S, Farré P, Cuvier O, Emberly E. Probing long-range interactions by extracting free energies from genome-wide chromosome conformation capture data. BMC Bioinformatics. 2015 May; 16:171. doi: 10.1186/s12859-015-0584-2 PMID: 26001583

50. Neuwald AF, Aravind L, Spouge JL, Koonin EV. AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. Genome Research. 1999; 9(1):27–43. Available from: http://genome.cshlp.org/content/9/1/27.abstract. PMID: 9927482

51. Zhao K, Hart CM, Laemmli UK. Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell. 1995 June; 81(6):879–889. doi: 10.1016/0092-8674(95)90008-X PMID: 7781065

52. Yang J, Ramos E, Corces VG. The BEAF-32 insulator coordinates genome organization and function during the evolution of *Drosophila* species. Genome Research. 2012 Nov; 22(11):2199–2207. doi: 10.1101/gr.142125.112 PMID: 22895281

53. Ulianov SV, Khrameeva EE, Gavrilov AA, Flyamer IM, Kos P, Mikhaleva EA, et al. Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains. Genome Research. 2016 Jan; 26(1):70–84. doi: 10.1101/gr.196006.115 PMID: 26518482

54. Zhang Y, Wong CH, Birnbaum RY, Li G, Favaro R, Ngan CY, et al. Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. Nature. 2013 Nov; 504(7479):306–310. doi: 10.1038/nature12716 PMID: 24213634

55. Barutcu A, Lajoie B, McCord R, Tye C, Hong D, Messier T, et al. Chromatin interaction analysis reveals changes in small chromosome and telomere clustering between epithelial and breast cancer cells. Genome Biology. 2015 September; 16(1):214. Available from: http://genomebiology.com/2015/16/1/214. doi: 10.1186/s13059-015-0768-0 PMID: 26415882

56. Iyer NG, Ozdag H, Caldas C. p300/CBP and cancer. Oncogene. 2004 May; 23(24):4225–4231. doi: 10.1038/sj.onc.1207118 PMID: 15156177

57. Altucci L, Leibowitz MD, Ogilvie KM, de Lera AR, Gronemeyer H. RAR and RXR modulation in cancer and metabolic disease. Nature Reviews Drug Discovery. 2007 October; 6(10):793–810. doi: 10.1038/nrd2397 PMID: 17906642

58. Khaled WT, Choon Lee S, Stingl J, Chen X, Raza Ali H, Rueda OM, et al. BCL11A is a triple-negative breast cancer gene with critical functions in stem and progenitor cells. Nature Communications. 2015 Jan; 6:5987+. doi: 10.1038/ncomms6987 PMID: 25574598

59. Chai Y, Chipitsyna G, Cui J, Liao B, Liu S, Aysola K, et al. c-Fos oncogene regulator Elk-1 interacts with BRCA1 splice variants BRCA1a/1b and enhances BRCA1a/1b-mediated growth suppression in breast cancer cells. Oncogene. 2011 Mars; 20(11):1357–1367. doi: 10.1038/sj.onc.1204256

60. Wood AM, Van Bortle K, Ramos E, Takenaka N, Rohrbaugh M, Jones BC, et al. Regulation of chromatin organization and inducible gene expression by a *Drosophila* insulator. Molecular Cell. 2011 Oct; 44(1):29–38. doi: 10.1016/j.molcel.2011.07.035 PMID: 21981916

61. The ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012 Sep; 489(7414):57–74. doi: 10.1038/nature11247 PMID: 22955616

62. Filion GJ, van Bemmel JG, Braunschweig U, Talhout W, Kind J, Ward LD, et al. Systematic protein location mapping reveals five principal chromatin types in *Drosophila* cells. Cell. 2010 Oct; 143(2):212–224. doi: 10.1016/j.cell.2010.09.009 PMID: 20888037

3.4.3.2 HiCglmi: identification of protein complex mediating looping

DNA loops result from the physical contact of two separated loci brought in 3D proximity. Those loops are essential to numerous key processes in the cell, such as gene expression [Jin *et al.* 2013] and DNA replication [Pope *et al.* 2014]. For instance, the expression of a gene is often regulated by regulatory elements that are far linearly on the genome, but that are in 3D contact with the gene promoter. In addition, several studies have shown that the disruption of DNA loops can lead to genetic diseases and cancers [Lupiáñez *et al.* 2015, Hnisz *et al.* 2016]. Understanding how DNA loops are formed and what are their molecular determinants is thus a fundamental issue.

I proposed a generalized linear model with interactions (GLMI) to identify the molecular determinants of loops, including protein and DNA sequence (Equation 1 and Figure 1, from the article "Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach" below) [Mourad *et al.* 2017]. GLMI has multiple assets over existing approaches such as enrichment test, correlation and random forests. Compared to enrichment test [Dixon *et al.* 2012, Djekidel *et al.* 2015] or correlation [Pancaldi *et al.* 2016] that respectively assesses the protein enrichment or correlation at highly confident loops, GLMI quantitatively links the frequency of all long-range contacts to complex co-occupancies of proteins while accounting for known Hi-C biases and polymer background. Moreover, GLMI accounts for colocalizations among protein binding, a strong issue when analyzing protein binding sites known to largely overlap over the genome. In contrast to random forests [He *et al.* 2014] which are efficient predictive models, but sometimes poor explanatory ones, GLMI allows to identify key chromatin loop driver proteins and motifs. GLMI can also uncover numerous mechanisms behind loop formation using higher-order interaction terms and proper confounding variables. For instance, GLMI can determine if a cofactor is necessary to mediate long-range contacts between distant protein binding sites.

Using real Drosophila Hi-C and ChIP-seq data, we validate numerous GLMI predictions of long-range contacts that involve insulator binding proteins, cofactors and motifs, and which were confirmed by previous microscopy and mutational studies. For instance, our model estimates long-range contacts between distant BEAF-32 motifs, which were previously observed with both fluorescence cross-correlation spectroscopy [Vogelmann *et al.* 2014] and high-resolution microscopy [23]. In addition, our model finds a mediating role of CP190 in bridging long-range contacts between distant BEAF-32 and GAF binding sites, in agreement with mutational experiments [19]. Of interest, GLMI analyses highlight a role of cohesin in stabilizing long-range contacts between CTCF sites in Drosophila, similarly to its role in human [7]. Supporting this role, we show that such influence is reduced upon cohesin subunit Rad21 depletion. It has to be noted that the absence of complete loss of contacts between CTCF sites after Rad21 depletion can be explained by the fast turnover of chromosome-bound cohesin in interphase [56]. Moreover, GLMI

outperforms enrichment test, correlation and random forests in the identification of known architectural proteins and motifs, and in the detection of the effects of mutations in the dCTCF motif.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach

Raphaël Mourad[1]*, Lang Li[2], Olivier Cuvier[1]

1 Laboratoire de Biologie Moléculaire Eucaryote (LBME), CNRS, Université Paul Sabatier (UPS), Toulouse, France, **2** Center for Computational Biology and Bioinformatics (CCBB), Indiana University, Indianapolis, Indiana, United States of America

* raphael.mourad@ibcg.biotoul.fr

OPEN ACCESS

Citation: Mourad R, Li L, Cuvier O (2017) Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach. PLoS Comput Biol 13(5): e1005538. https://doi.org/10.1371/journal.pcbi.1005538

Editor: Alexandre V Morozov, Rutgers University, UNITED STATES

Received: December 17, 2016

Accepted: April 28, 2017

Published: May 23, 2017

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: This work was supported by the University of Toulouse, Fondation pour la Recherche Médicale and the CNRS. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

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

Abstract

Chromosomal organization in 3D plays a central role in regulating cell-type specific transcriptional and DNA replication timing programs. Yet it remains unclear to what extent the resulting long-range contacts depend on specific molecular drivers. Here we propose a model that comprehensively assesses the influence on contacts of DNA-binding proteins, cis-regulatory elements and DNA consensus motifs. Using real data, we validate a large number of predictions for long-range contacts involving known architectural proteins and DNA motifs. Our model outperforms existing approaches including enrichment test, random forests and correlation, and it uncovers numerous novel long-range contacts in *Drosophila* and human. The model uncovers the orientation-dependent specificity for long-range contacts between CTCF motifs in *Drosophila*, highlighting its conserved property in 3D organization of metazoan genomes. Our model further unravels long-range contacts depending on co-factors recruited to DNA indirectly, as illustrated by the influence of cohesin in stabilizing long-range contacts between CTCF sites. It also reveals asymmetric contacts such as enhancer-promoter contacts that highlight opposite influences of the transcription factors EBF1, EGR1 or MEF2C depending on RNA Polymerase II pausing.

Author summary

Chromosomal DNA is tightly packed in three dimensions (3D) such that a 2-meter long human genome can fit into a microscopic nucleus. Recent studies have revealed that such packing of DNA is not random but instead structured into functional DNA loops. Those loops are essential to numerous key processes in the cell, such as genome expression and DNA replication. In addition, disruption of DNA loops can lead to genetic diseases and cancers. Understanding how DNA loops are formed and what are their molecular determinants is thus a fundamental issue. In this work, we propose a computational model to identify the molecular determinants of loops, including protein and DNA sequence. Most notably, the model offers insights in the different mechanistic scenarios behind loop formation. Using this model, we uncover numerous novel DNA loops and underlying

mechanisms in *Drosophila* and human. We find that the orientation-dependent specificity between CTCF motifs is conserved in metazoans. We show how loops between DNA-binding proteins can be mediated by additional cofactors. Our analyses further reveal opposite influences of transcription factors depending on RNA Polymerase II pausing.

Introduction

Chromosomal DNA is tightly packed in three dimensions (3D) such that a 2-meter long human genome can fit into a nucleus of approximately 10 microns in diameter [1]. Such 3D structure of chromosome has recently been explored by chromosome conformation capture combined with high-throughput sequencing technique (Hi-C) at an unprecedented resolution [2–4]. Multiple hierarchical levels of genome organization have been uncovered such as compartments A/B [5] and topologically associating domains (TADs) [2, 3]. In particular, TADs represent a pervasive structural feature of the genome organization and are highly conserved across species. Functional studies revealed that spatial organization of chromosome is essential to numerous key processes such as for the regulation of gene expression by distal enhancers [4] or for the replication-timing program [6].

The comprehensive analysis of 3D chromatin drivers is currently a hot topic [7]. A growing body of evidence supports the role of insulator binding proteins (IBPs) such as CTCF, and cofactors like cohesin, as mediators of long-range chromatin contacts [3, 8, 9]. In human, high-resolution Hi-C mapping has recently revealed that loops that demarcate domains were often marked by asymmetric CTCF motifs where cohesin is recruited [10]. Depletions of CTCF and cohesin decreased chromatin contacts [11]. However the impact of these depletions was limited suggesting that other proteins might be involved in shaping the chromosome in 3D. For instance, numerous IBPs, cofactors and functional elements were shown to colocalize at TAD borders [9, 12]. The identification of 3D chromatin drivers is thus an active avenue of research. Computational approaches that integrate the large amount of available protein binding data (chromatin immunoprecipitation followed by high-throughput DNA sequencing, ChIP-seq), functional elements (promoters and enhancers), and DNA motifs, with Hi-C data may be well-suited to identify novel factors that participate in shaping the chromosome in 3D [13].

In this paper, we propose a model to comprehensively analyze the roles of genomic features, such as DNA-binding proteins or motifs, in establishing or maintaining chromatin contacts. The proposed model offers insights in the different mechanistic scenarios behind loop formation, because of its ability to rigorously assess the effect of protein complex on long-range contact frequency. Using real data, the model successfully predicted numerous long-range interactions involving motifs and proteins as highlighted in previous independent studies. Moreover, our model outperformed current approaches to identify architectural proteins and motifs, and to detect the effects of single nucleotide polymorphisms (SNPs) in the dCTCF motif. In addition, our model is the only approach able to assess the effect of a cofactor in mediating long-range contacts between distant protein binding sites, such as cohesin with CTCF. Using recent *Drosophila* and human Hi-C data at high resolution, combined with a large number of ChIP-seq, RNA-seq, CAGE-seq and DNA motif data, we revealed numerous novel motifs, insulator binding proteins, cofactors and functional elements that positively or negatively impact long-range contacts depending on transcriptional activity or motif orientation.

Results and discussion

The model

We propose to use a generalized linear model with interactions (GLMI) to analyze the effects of genomic features such as architectural protein co-occupancies on chromatin contacts at genome-wide level:

$$\begin{aligned} \log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta\mathbf{X} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \end{aligned} \quad (1)$$

Variable **y** denotes the number of Hi-C contacts for any pair of bins on the same chromosome. Variable set $\mathbf{X} = \{\mathbf{d}, \mathbf{B}, \mathbf{C}, \mathbf{g}\}$ comprises several variable subsets: the log-distance variable **d**, the bias variables **B**, the confounding variable set **C** and the genomic variable of interest **g**. The log-distance variable **d** accounts for the background polymer effect (log-log relation between distance and Hi-C count) [14]. Bias variables $\mathbf{B} = \{\mathbf{len}, \mathbf{GC}, \mathbf{map}\}$ are known Hi-C biases including fragment length (**len**), GC-content (**GC**) and mappability (**map**) that are computed as in [15] (S1 Appendix, Bias variable computation). Including those bias variables into the model allows to correct for biases in Hi-C data. Bias normalization by matrix balancing methods [16] is avoided, because these methods might remove effect of genomic variable of interest. Variable **g** represents the genomic feature of interest, whose associated β_g parameter value reflects its effects on chromatin contacts. Variable set **C** comprises confounding variables included to properly estimate β_g. Model (1) is very general and can be developed in multiple versions depending on the variable **g** of interest. In the following paragraphs, we will see the different kinds of variables **g**. The corresponding models are detailed in Subsection Materials and Methods, The different models.

We illustrate the different model variables in Fig 1. For simplicity, we illustrate our model with protein binding sites, yet the same model is applicable to many other genomic features such as motifs or promoters. Let consider a pair of bins that we call left bin (L) and right bin (R). The attribution for left and right bins is arbitrary. Let also consider 3 genomic features F_i (whose binding is colored in blue in Fig 1), F_j (in red) and F_k (in green) that represent binding sites of 3 different proteins. For the genomic feature F_i, occupancy variables $\mathbf{z}_{iL}$ and $\mathbf{z}_{iR}$ denote the occupancies of F_i on left and right bins, respectively. For an occupancy variable, a value of 0/1 means absence/presence of the corresponding feature on the bin, *e.g.* absence/presence of the protein on the bin (a value between 0 and 1 means partial overlap of the feature). Occupancy variables are used to build 4 main kinds of model variables as follows.

A "homologous interaction" variable $\mathbf{n}_{ii}$ is the product of $\mathbf{z}_{iL}$ and $\mathbf{z}_{iR}$ ($\mathbf{n}_{ii} = \mathbf{z}_{iL} \times \mathbf{z}_{iR}$). The associated $\beta_{n_{ii}}$ parameter reflects the extent by which the genomic feature F_i interacts with itself through chromatin contacts (Fig 1a). For instance, distant CTCF binding sites were shown to form loops in human [10, 17].

A "heterologuous interaction" variable $\mathbf{n}_{ij}$ is the average of the product $\mathbf{z}_{iL} \times \mathbf{z}_{jR}$ and the product $\mathbf{z}_{jL} \times \mathbf{z}_{iR}$ ($\mathbf{n}_{ij} = \frac{1}{2}(\mathbf{z}_{iL} \times \mathbf{z}_{jR} + \mathbf{z}_{jL} \times \mathbf{z}_{iR})$), because both products are identically associated to **y**. The associated $\beta_{n_{ij}}$ parameter reflects the extent by which the genomic feature F_i interacts with another genomic feature F_j through chromatin contacts (Fig 1b). For instance, enhancers are in long-range contacts with promoters to regulate target gene expression [14, 18].

A "homologous interaction cofactor" variable $\mathbf{c}_{iik}$ is the product of an interaction variable $\mathbf{n}_{ii}$ and an interaction variable $\mathbf{n}_{kk}$ ($\mathbf{c}_{iik} = \mathbf{n}_{ii} \times \mathbf{n}_{kk} = \mathbf{z}_{iL} \times \mathbf{z}_{iR} \times \mathbf{z}_{kL} \times \mathbf{z}_{kR}$). Here we consider the cofactor F_k as a protein that does not directly bind to DNA, but which is instead bound by an

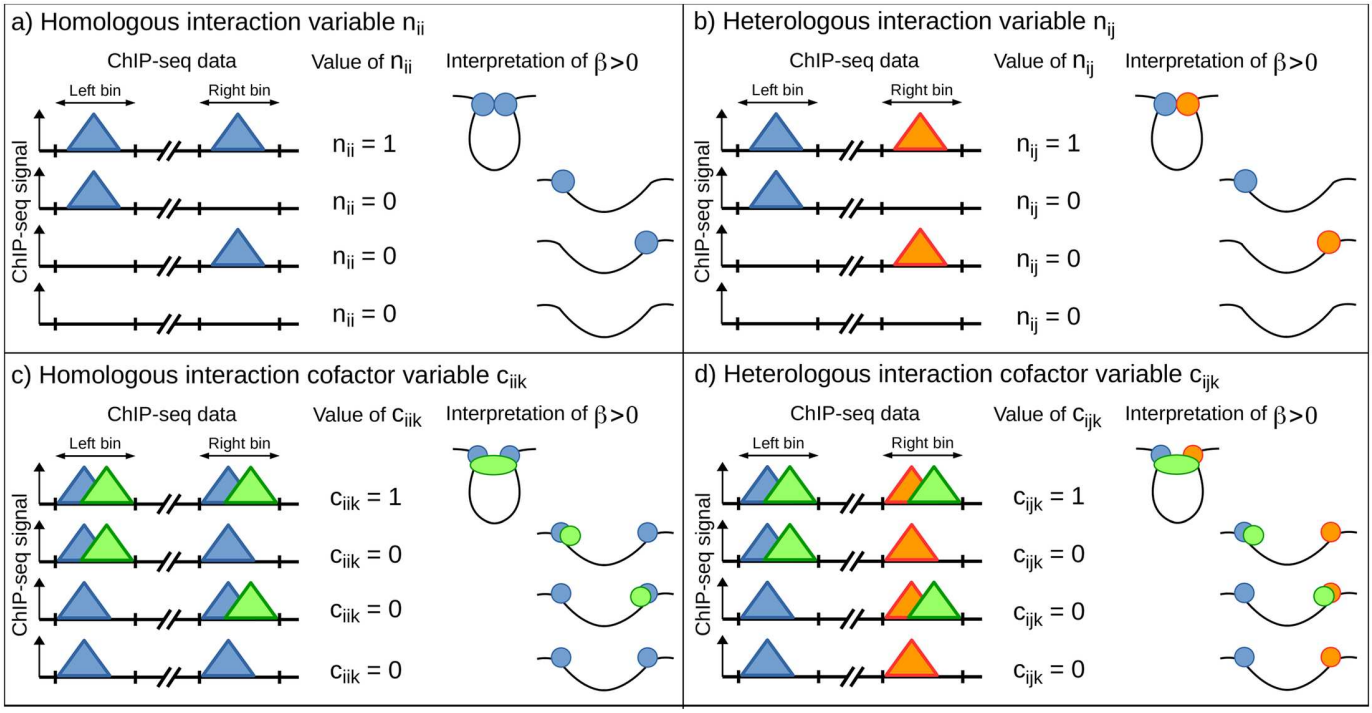


Fig 1. Illustration of the proposed model and variables in the context of protein ChIP-seq data. a) Homologous interaction variable. b) Heterologous interaction variable. c) Homologous interaction cofactor variable. d) Heterologous interaction cofactor variable. The 3 proteins F_i, F_j and F_k are colored in blue, red and green, respectively. Here F_i and F_j are insulator binding proteins (IBPs), and F_k is a cofactor (recruited by IBPs).

https://doi.org/10.1371/journal.pcbi.1005538.g001

insulator binding protein F_i (IBP) to DNA, such as cohesin is recruited by CTCF to DNA. Hence we expect that a cofactor will be found at both bins L and R in contact, *e.g.* cohesin ring entraps both chromatin fibers and is thus observed at both bins [10, 17]. That explains why $\mathbf{c}_{iik}$ is the product of $\mathbf{n}_{ii}$ and $\mathbf{n}_{kk}$. The associated $\beta_{c_{iik}}$ parameter reflects the extent by which chromatin contacts between genomic feature F_i and itself are mediated by a genomic feature F_k, the cofactor (Fig 1c).

A "heterologous interaction cofactor" variable $\mathbf{c}_{ijk}$ is the product of an interaction variable $\mathbf{n}_{ij}$ and an interaction variable $\mathbf{n}_{kk}$ ($\mathbf{c}_{ijk} = \mathbf{n}_{ij} \times \mathbf{n}_{kk} = \frac{1}{2}(\mathbf{z}_{iL} \times \mathbf{z}_{jR} \times \mathbf{z}_{kL} \times \mathbf{z}_{kR} + \mathbf{z}_{jL} \times \mathbf{z}_{iR} \times \mathbf{z}_{kL} \times \mathbf{z}_{kR})$). Here we consider the cofactor F_k as a protein that does not directly bind to DNA, but which is instead bound to two IBPs F_i and F_j. For instance, a loop can be mediated by CP190 that binds to BEAF-32 and GAF sites that are distant [19]. The associated $\beta_{c_{ijk}}$ parameter reflects the extent by which chromatin contacts between genomic features F_i and F_j are mediated by a third genomic feature F_k, the cofactor (Fig 1d).

In the previous paragraphs, we introduced numerous variables that were the products of simpler variables, namely the occupancy variables. In (generalized) linear regression, those product variables are called "interaction" terms. To detect such interaction effects, one usually needs a large number of observations. We will see in the next subsections that the tremendous amount of data provided by Hi-C experiments allows to detect such interaction effects with accuracy. The model and the different variables will be illustrated with real world scenarios in the next subsections.

Prediction of known factors and validation with experimental data

We first sought to validate our model using experimental data. For this purpose, we focused on the *Drosophila* model because several insulator binding proteins (IBPs) that mediate long-range interactions have been well characterized in this organism. *Drosophila* IBPs comprise suppressor of hairy wing (Su(Hw)), *Drosophila* CTCF (dCTCF), boundary-element-associated factor of 32 kDa (BEAF-32), GAGA binding factor (GAF), Zeste-White 5 (ZW5) [20], the general transcription factor dTFIIIC [9] and DNA replication-related element factor (DREF) [7]. We analyzed Kc167 Hi-C data at 10 kb resolution and focused on 20kb-1Mb distances for which contact frequencies were accurately measured experimentally [21]. At this distance range, the log-log relation between Hi-C count and distance was linear ($R^2 = 0.99$, S1 Fig), supporting the use of the log-distance term in the model. The data comprised approximately 1 million of observations, which allowed to detect higher-order interactions with enough precision (tight parameter confidence intervals reflected by low p-values, see below). Because of Hi-C count overdispersion, we used negative binomial regression as the most appropriate specification of the generalized linear model.

It has been shown that BEAF-32 motifs can form long-range interactions with each other using both fluorescence cross-correlation spectroscopy [22] and high-resolution microscopy [23]. Following this observation, we first validated our model by successfully estimating long-range contacts between the BEAF-32 CGATA motifs using model (2) ($\hat{\beta}_{n_{ii}} = 6.7 \times 10^3$, $p < 10^{-20}$; Fig 2a; model (2) and all other models used in the following are described in Subsection Materials and Methods, The different models). This result was confirmed as we observed that the Hi-C count increased with co-occupancy of BEAF-32 motifs (variable $\mathbf{n}_{ii}$) (Fig 2b). We also observed long-range contacts between dCTCF motifs ($\hat{\beta}_{n_{ii}} = 2.4 \times 10^4$, $p = 3 \times 10^{-14}$), highlighting their important roles in loop formation in *Drosophila* as observed in human [10, 17]. Over the 7 known IBPs, the model correctly identified all IBP motifs as involved in long-range contacts among themselves (Fig 2c). Next the same approach was used to evaluate the model's ability to discriminate between the 7 IBP motifs (true positives) and 83 other DNA-binding protein motifs (false positives). This approach obtained good predictions (area under the curve (AUC) = 0.855; Fig 2d). Among the motifs that we considered as false positives, M1BP and Ttk69K motifs presented high and significant interaction effects (M1BP: $\hat{\beta}_{n_{ii}} = 1.7 \times 10^5$; Ttk69K: $\hat{\beta}_{n_{ii}} = 2.3 \times 10^4$, $p < 10^{-12}$, resp.). These results suggested that M1BP and Ttk69K might represent new insulator-binding protein candidates. Accordingly, M1BP protein binds to the promoters of paused genes that were shown to be involved in long-range contacts [18, 24]. Ttk69K protein has a homomeric dimerization BTB/POZ domain that could help bridging two distant proteins through long-range contacts [22].

We then used GLMI to study the role of cofactors that cannot directly bind to DNA, but are instead recruited by IBPs, and are required to mediate or stabilize long-range contacts between two IBP binding sites. In *Drosophila*, well-known cofactors include condensin I, condensin II, Chromator, centrosomal protein of 190 kDa (CP190), cohesin [19–22], Fs(1)h-L [25] and lethal (3) malignant brain tumor (L(3)Mbt) [7]. Most notably, fluorescence cross-correlation spectroscopy (FCCS) experiments have shown that CP190 is required to bridge long-range contacts between two BEAF-32 binding sites [22]. Using ChIP-seq peak data with model (4), we estimated a significant and positive effect of CP190 in mediating long-range contacts between BEAF-32 sites ($\hat{\beta}_{c_{iik}} = 878$, $p < 10^{-20}$; Fig 2e), in complete agreement with recent work [22]. Similar result was obtained for Chromator in mediating long-range contacts between BEAF-32 sites ($\hat{\beta}_{c_{iik}} = 3.4 \times 10^3$, $p < 10^{-20}$) [22]. In addition, previous BEAF-32 mutation by our group has revealed that cofactor CP190 is also required to bridge long-range

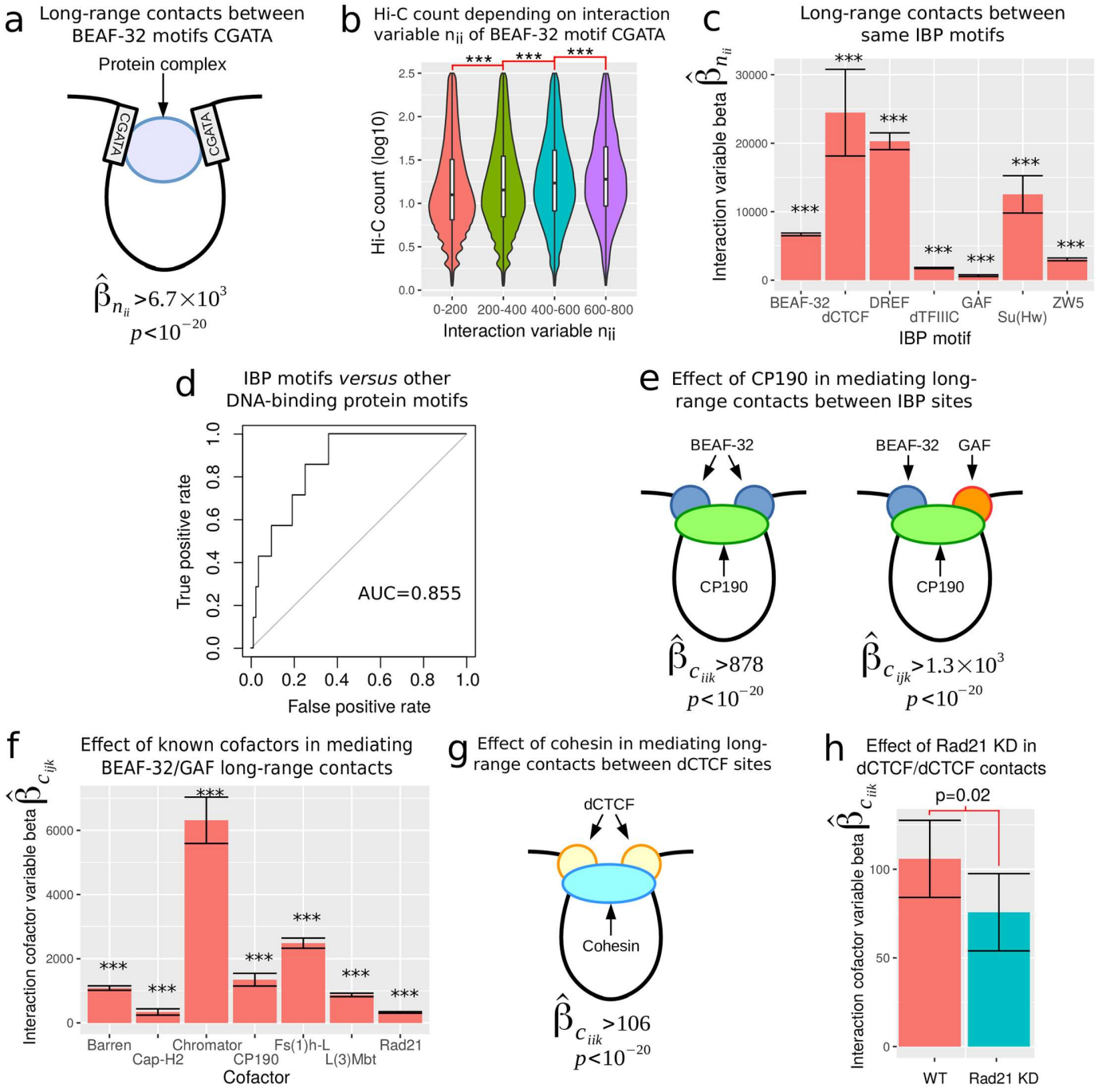


Fig 2. Biological validation of the model. a) Long-range contacts between BEAF-32 motifs. b) Hi-C count as a function of interaction variable $\mathbf{n}_{ii}$ of BEAF-32 motifs. c) Long-range contacts between same insulator binding protein (IBP) motifs. d) Receiver operating characteristic (ROC) curves of long-range contacts between same motifs. Known IBP motifs (true positives) are compared to other protein motifs (false positives). e) Effect of CP190 in mediating long-range contacts between IBP sites. f) Effect of known cofactors in mediating long-range contacts between distant BEAF-32 and GAF binding sites. Barren, Cap-H2 and Rad21 are subunits of condensin I, condensin II and cohesin, respectively. g) Effect of cohesin in mediating long-range contacts between dCTCF sites. h) Effect of cohesin in mediating long-range contacts between distant dCTCF binding sites in wild-type (WT) compared to Rad21 KD cells.

https://doi.org/10.1371/journal.pcbi.1005538.g002

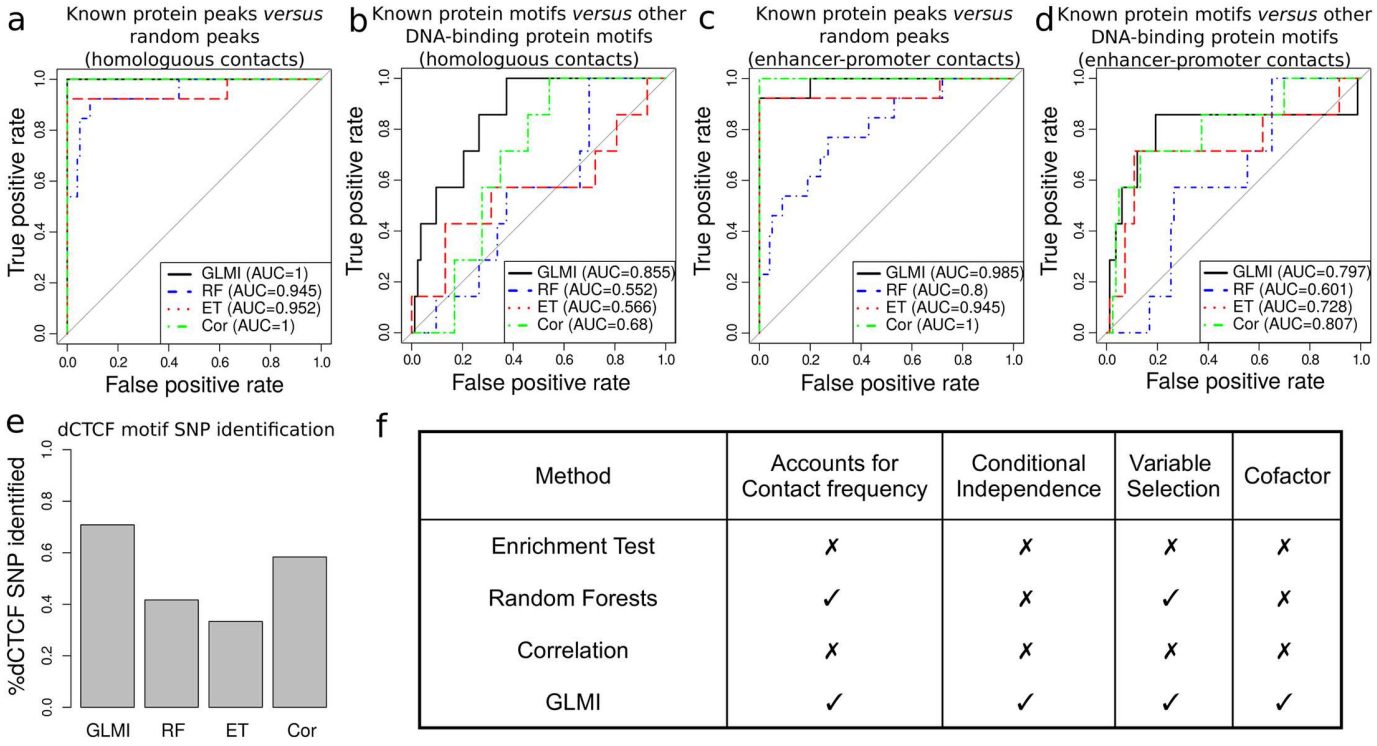
contacts between BEAF-32 and GAF binding sites [19]. Using ChIP-seq peak data with model (5), we estimated a significant and positive effect of CP190 in bridging distant BEAF-32 and GAF sites ($\hat{\beta}_{c_{ijk}} = 1.3 \times 10^3$, $p < 10^{-20}$; Fig 2e) [19]. We applied the same modeling approach to the 6 other known cofactors and found that all were associated with significant positive effects in mediating contacts between BEAF-32 and GAF binding sites (all betas $\hat{\beta}_{c_{ijk}} > 326$, all p-values $p < 10^{-20}$; Fig 2f). Because CP190 was also shown to mediate long-range contacts between BEAF-32 and dCTCF, and between BEAF-32 and Su(Hw) [19], we estimated the corresponding cofactor effects. We again found significant positive effect of CP190 between BEAF-32 and dCTCF ($\hat{\beta}_{c_{ijk}} = 892$, $p < 10^{-20}$), but our method only detected a slightly significant mediating effect of CP190 between BEAF-32 and Su(Hw) ($\hat{\beta}_{c_{ijk}} = 175$, $p = 0.02$). In human, the most studied cofactor is cohesin that is able to entrap two chromatin fibers thereby stabilizing long-range contacts between CTCF sites [10, 17]. Hence we assessed the impact of cohesin in mediating long-range contacts between two dCTCF binding sites in *Drosophila*. We found a significant and positive effect of cohesin ($\hat{\beta}_{c_{iik}} = 105.8$, $p < 10^{-20}$; Fig 2g), thus supporting a conserved function of cohesin in stabilizing long-range contacts between CTCF sites in metazoans.

We further tested our model for cofactor effects using perturbed conditions such as the removal of these cofactors, as obtained through knocking-down (KD) followed by Hi-C experiment. Of note, Hi-C experiments are expensive and complex to carry out, and the possibility to predict long-range contacts upon such KD is of major importance. We compared the impact of cohesin in the context of long-range contacts bridging CTCF sites in WT and Rad21 (cohesin subunit) KD Hi-C data. Our model estimated a significant but lower cofactor effect of cohesin in Rad21 KD ($\hat{\beta}_{c_{iik}} = 75.7$, $p = 9 \times 10^{-12}$), compared to WT ($\hat{\beta}_{c_{iik}} = 105.8$, $p < 10^{-20}$). The difference between WT and Rad21 KD associated coefficients was negative and significant (beta difference = −30.1, $p = 0.027$), corresponding to a beta decrease of 28% (Fig 2h). This result therefore validated the estimated effect of cohesin in mediating distant dCTCF binding sites, which decreased upon cohesin depletion as expected.

Using real data, we concluded that our model successfully predicted the roles of IBP motifs in long-range contacts between distant loci, as well as the roles of known cofactors in bridging distant IBP binding sites. The GLMI predictions were validated in the literature and using protein KD followed by Hi-C experiment.

GLMI outperformed existing methods

We then compared GLMI with existing methods for their ability to identify genomic features known to be involved in long-range contacts. For this purpose, we compared GLMI with (1) enrichment test (ET) on highly confident chromatin interaction pairs as previously [26], (2) correlation (Cor) on highly confident chromatin interaction pairs [27] and (3) random forests (RF) discriminating highly confident chromatin interaction pairs from non-interacting pairs [28]. As a first and simple benchmark, we assessed the different methods to identify long-range contacts between protein binding sites of the same proteins (model (2)). We evaluated the ability to discriminate between architectural proteins known to be involved in long-range contacts (13 true positives including IBPs and cofactors) and random protein peaks (100 false positives) using receiver operating characteristic (ROC) curves. We observed that all four methods were very efficient to detect long-range contacts between known architectural protein binding sites (Fig 3a). In particular, GLMI and Cor showed perfect predictions (*AUC* = 1). RF and ET were also very accurate (*AUC* > 0.94). Previous benchmark was an easy task because it



f

Method	Accounts for Contact frequency	Conditional Independence	Variable Selection	Cofactor
Enrichment Test	✗	✗	✗	✗
Random Forests	✓	✗	✓	✗
Correlation	✗	✗	✗	✗
GLMI	✓	✓	✓	✓

Fig 3. Comparisons between generalized linear regression with interactions (GLMI), highly confident chromatin interaction pair detection followed by pair type enrichment (ET), highly confident chromatin interaction pair detection followed by correlation (Cor) and random forests (RF). a) Receiver operating characteristic (ROC) curves of the four methods to distinguish between known protein peaks (13 true positives) and random peaks (100 false positives). Long-range contacts are assessed between a protein and itself (homologous contacts). b) ROC curves of the four methods to distinguish between known protein motifs (7 true positives) and other DNA-binding protein motifs (83 false positives). Long-range contacts are assessed between a motif and itself (homologous contacts). c) ROC curves of the four methods to distinguish between known protein peaks and random peaks. Long-range contacts are assessed between a protein and promoters (enhancer-promoter contacts). d) ROC curves of the four methods to distinguish between known protein motifs and other DNA-binding protein motifs. Long-range contacts are assessed between a motif and promoters (enhancer-promoter contacts). e) Percent of dCTCF motif SNP that have a homologous interaction variable beta lower than the one of the dCTCF concensus motif. f) Comparison table of the methods.

https://doi.org/10.1371/journal.pcbi.1005538.g003

relied on random protein peaks whose binding was very different from real protein binding. For a more realistic benchmark, we then evaluated the ability to discriminate between motifs whose proteins are known to be involved in long-range contacts (7 true positives) and other DNA-binding protein motifs (83 false positives) using ROC curves. Using this benchmark, all the four methods performed less well (Fig 3b). However we found that GLMI clearly outperformed the three other methods to detect long-range contacts between DNA motifs known to be involved in chromatin interactions ($AUC_{GLMI} = 0.855$).

Another benchmark consisted in identifying long-range contacts between binding sites of a protein and active promoters. Here, as previously, we evaluated the ability to discriminate between architectural proteins known to be involved in enhancer-promoter contacts (13 true positives including IBPs and cofactors) and random protein peaks (100 false positives) using ROC curves. We observed that all four methods were very efficient to detect long-range contacts between known architectural protein binding sites and active promoters (Fig 3c). In particular, GLMI and Cor showed excellent predictions ($AUC_{GLMI} = 0.985$ and $AUC_{Cor} = 1$). We then evaluated the ability to discriminate between motifs whose proteins are known to be involved in enhancer-promoter contacts (7 true positives) and other DNA-binding protein motifs (83 false positives) using ROC curves. Both GLMI and Cor performed

well (AUC_{GLMI} = 0.797 and AUC_{Cor} = 0.807; Fig 3d). Conversely, ET and RF showed lower perfomance (AUC_{ET} = 0.728 and AUC_{RF} = 0.601).

We next analyzed the impacts of mutations in the consensus dCTCF motif. Single nucleotide polymorphisms (SNPs) play an important role in common genetic diseases and recent works have uncovered differential long-range contacts due to variations in the CTCF motif in human [17, 29, 30]. Hence we evaluated the methods to detect the impacts of single nucleotide mutations in the dCTCF motif. For this purpose, we considered the dCTCF consensus motif AGGTGGCG (wild-type motif) [31] and generated dCTCF motifs with single nucleotide mutations for each position (mutated motifs). For instance, for the first position, the mutated motifs were TGGTGGCG, GGGTGGCG and CGGTGGCG. Over the 24 possible mutated motifs (8 positions × 3 alternative nucleotides), GLMI detected 17 motifs (71%; Fig 3e) with homologous interaction variable betas that were lower than the one of the wild-type motif, indicating that the corresponding mutations diminished the ability of dCTCF to bridge long-range contact. Compared to GLMI, other approaches showed lower performance (Cor: 14/24; RF = 10/24; ET = 8/24).

In addition to its better prediction performances, our model presents several theoretical advantages over the three other methods as summarized in Fig 3f. All the methods can assess long-range contacts between protein binding sites. However, GLMI is the only model that, at the same time, (1) accounts for the contact frequency which can vary among highly confident loops, (2) can deal with the presence of colocalization among proteins using conditional independence, (3) allows variable selection using lasso or stepwise, and (4) can assess the effect of cofactors by including higher-order interaction terms.

Analysis of insulator binding protein motifs in *Drosophila*

Given the biological validation of our model, we next sought to address the roles of IBP motifs in establishing or maintaining long-range interactions in *Drosophila*. We first assessed how IBP motifs were coupled to form loops (*i.e.* for all combinations of distant IBP motifs). For this purpose, we estimated homologous and heterologous interaction variable effects for any couple of IBP motifs using models (2) and (3), and using the same Hi-C data, distance range and resolution as above (Fig 4a). The strongest long-range contacts were between dCTCF and DREF motifs ($\hat{\beta}_{n_{ij}} = 2.8 \times 10^4, p < 10^{-20}$), between dCTCF motifs ($\hat{\beta}_{n_{ii}} = 2.4 \times 10^4$, $p < 10^{-20}$) and between DREF motifs ($\hat{\beta}_{n_{ii}} = 2 \times 10^4, p < 10^{-20}$). High levels of long-range contacts were also found between BEAF-32 and DREF motifs ($\hat{\beta}_{n_{ij}} = 1.9 \times 10^4, p < 10^{-20}$) and between BEAF32 and dCTCF motifs ($\hat{\beta}_{n_{ij}} = 1.9 \times 10^4, p < 10^{-20}$). Thus in *Drosophila*, chromatin loops not only involve dCTCF motifs but also DREF and BEAF-32 motifs that all work together. We then explored if these long-range contacts depended on the distance between motifs. At short distance (<100kb), long-range contacts were mainly detected between DREF motifs ($\hat{\beta}_{n_{ii}} = 1.8 \times 10^4, p < 10^{-20}$), whereas at long distance (> 750kb), they were more frequent between dCTCF and DREF motifs ($\hat{\beta}_{n_{ij}} = 3.5 \times 10^4, p = 7 \times 10^{-9}$) (Fig 4b). In addition, long-range contacts between dCTCF motifs peaked at 500 kb. Our results therefore raise the possibility that long-range contacts between IBP motifs could be distant-dependent. This observation might provide a molecular explanation for the observed hierarchical nature of 3D chromatin structure [32, 33], for which loops could be formed at different scales by the interplay of specific proteins.

Next we sought to comprehensively test whether motif orientation could influence long-range contacts, as originally shown for CTCF motifs in human [10] and more generally

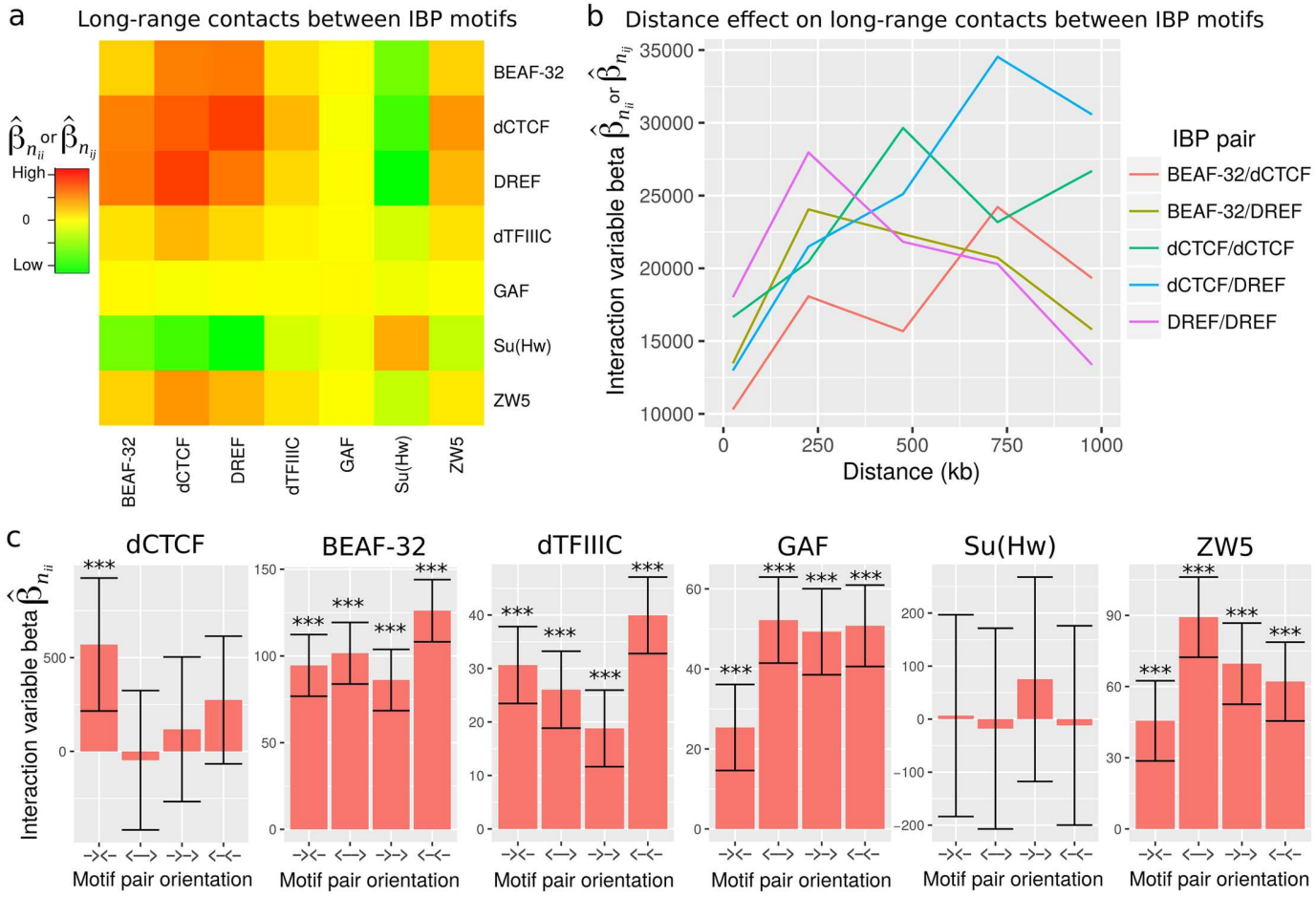


Fig 4. Analysis of long-range contacts between insulator binding protein (IBP) motifs. a) Long-range contacts between IBP motifs, as measured by interaction variable betas estimated using models (2) and (3). b) Long-range contacts between IBP motifs depending on the distance. c) Long-range contacts between IBP motifs depending on the motif pair orientation.

https://doi.org/10.1371/journal.pcbi.1005538.g004

in mammals [34]. We distinguished the motifs that were on the positive DNA strand (denoted +), from those that were on the negative DNA strand (denoted -). Then it was possible to compute four types of homologous interaction variables: $\mathbf{n}_{ii+-} = \mathbf{z}_{iL+} \times \mathbf{z}_{iR-}$ (orientation →←), $\mathbf{n}_{ii-+} = \mathbf{z}_{iL-} \times \mathbf{z}_{iR+}$ (orientation ←→), $\mathbf{n}_{ii--} = \mathbf{z}_{iL-} \times \mathbf{z}_{iR-}$ (orientation ←←), $\mathbf{n}_{ii++} = \mathbf{z}_{iL+} \times \mathbf{z}_{iR+}$ (orientation →→). The corresponding models are detailed in Subsection Materials and Methods, The different models. Here we processed data at 1 kb resolution for better accuracy in distinguishing the different orientations. Similarly to in human and mammals, we found significant long-range contacts for motifs in convergent orientation ($\hat{\beta}_{n_{ii}} = 570$, $p = 2 \times 10^{-3}$), and no significant contacts for the 3 other possible orientations (←→, →→ and ←←; Fig 4c), revealing conservation of convergent CTCF mediated loops in agreement with 4C analyses [35]. We then assessed motif orientation for all other IBP motifs. Of note, the orientation of DREF TATCGATA motifs could not be assessed because of its palindromic property. For BEAF-32, dTFIIIC and Su(Hw) motifs, we could not detect any strong orientation effect (Fig 4c). Conversely, for GAF and ZW5 motifs, we found stronger contacts for motifs in divergent orientation (←→) compared to convergent orientation (→←), suggesting a different mode of binding of the corresponding protein to DNA

or a different constraint depending of its interaction with cofactors. Thus motif orientation in loops depends on the protein involved, and the dependence on convergent orientation of motifs does not apply to all insulator binding proteins.

Analysis of insulator binding protein sites in *Drosophila*

IBP binding sites might significantly vary depending on the cell type and stage. Hence we reanalyzed the roles of IBP binding in Kc167 *Drosophila* cells using available ChIP-seq data (same cell type with Hi-C data; ZW5 data were not available). As in the previous subsection, we estimated interaction effects for any couple of IBP motifs using models (2) and (3). Similarly to the analysis of IBP motifs, we observed high levels of long-range contacts involving DREF and dCTCF (Fig 5a). In particular, we found strong long-range contacts between distant DREF

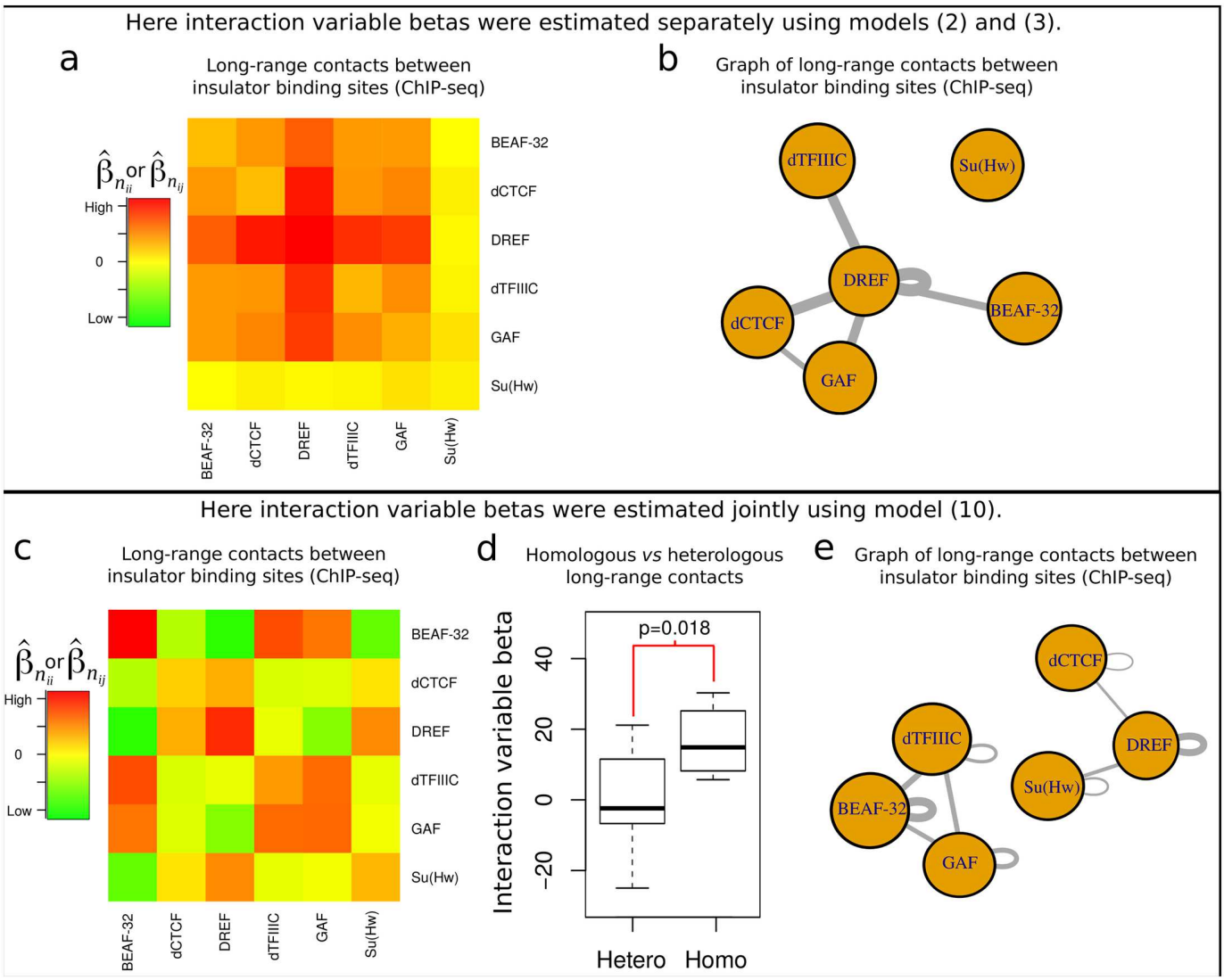


Fig 5. Analysis of long-range contacts between insulator binding protein (IBP) sites. a) Long-range contacts between IBP sites, as measured by interaction variable betas estimated separately (models (2) and (3)). b) Graph of long-range contacts (betas) between IBP sites estimated in a). c) Long-range contacts between insulator binding sites, as measured by interaction variable betas estimated jointly (model (10)). d) Comparison between homologous and heterologous interaction variable betas. e) Graph of long-range contacts (betas) between IBP sites estimated in c).

https://doi.org/10.1371/journal.pcbi.1005538.g005

binding sites ($\hat{\beta}_{n_{ii}} = 147, p < 10^{-20}$) and between dCTCF and DREF binding sites ($\hat{\beta}_{n_{ij}} = 133$, $p < 10^{-20}$). However, we also observed strong long-range contacts between DREF and dTFIIIC ($\hat{\beta}_{n_{ij}} = 119, p < 10^{-20}$), and between DREF and GAF ($\hat{\beta}_{n_{ij}} = 112, p < 10^{-20}$), which could not be detected by previous analysis of IBP motifs. We then built a graph using estimated betas by adding an edge between two proteins F_i and F_j with a weight $\hat{\beta}_{n_{ij}}$, and by adding an edge between a protein F_i and itself with a weight $\hat{\beta}_{n_{ii}}$ (Fig 5b). Analysis of the graph clearly revealed the role of DREF as a hub, *i.e.* DREF was involved in many long-range contacts with other IBPs, such as BEAF-32, DREF, dTFIIIC and GAF. Such DREF-mediated loops might be in apparent contradiction with recent experiments showing that DREF motifs tag proximal activation of housekeeping genes, in contrast to long-range activation of developmental genes [36]. However such DREF-mediated loops can be explained by long-range contacts between promoters ($\hat{\beta}_{n_{ii}} = 203, p < 10^{-20}$).

Previous results should be carrefully interpreted since IBPs often linearly colocalize (*i.e.* correlate) with each other on the chromosome [31]. Such correlations can lead to "indirect" long-range contacts between IBPs. For instance, if a loop is maintained by two distant dCTCF binding sites, and that BEAF-32 colocalizes to dCTCF, then it is likely that we will also observe loops between distant BEAF-32 and dCTCF sites, and even between BEAF-32 sites. The impact of such correlations between proteins in the study of 3D chromatin has been discussed in details [12]. Models (2) and (3) could not account for such correlations between IBPs because only one interaction variable term was included. Instead one should use another model that includes all possible interaction variable terms between IBPs (model (10), see Subsection Materials and methods, The different models). To better discard indirect long-range contacts between the 6 IBPs, we thus re-estimated interaction variable beta parameters using model (10) that included all marginal variables (6 variables, one for each IBP) and all interaction variables (21 variables, one for each combination of IBPs). Using model (10), we obtained rather different results (Fig 5c). We still observed strong long-range contacts between DREF binding sites ($\hat{\beta}_{n_{ii}} = 25, p < 10^{-11}$). However other long-range contacts were observed such as between BEAF-32 sites ($\hat{\beta}_{n_{ii}} = 30, p < 10^{-20}$). In turn, such analysis showed that an IBP tended to interact more with itself (homologous interactions) than with another IBP (heterologous interactions) ($p = 0.018$; Fig 5d), in agreement with insulator bodies observed by microscopy [37]. In addition, the model (10) allowed to infer negative and significant interaction effects, such as between distant DREF and BEAF-32 ($\hat{\beta}_{n_{ij}} = -25, p < 10^{-11}$), which could not be detected before. This negative effect means that BEAF-32 and DREF tend to avoid each other in long-range contacts, *i.e.* they tend to have a repulsive effect. This might reflect the known antagonistic relationship between BEAF-32 and DREF in competing for binding to overlapping binding sites [38, 39]. As previously, we built a graph of betas and could detect groups of IBPs that may cluster together through long-range contacts as found for the two connected components BEAF-32/dTFIIIC/GAF and DREF/Su(Hw)/dCTCF, respectively (Fig 5e). Interestingly, these two classes of IBPs that worked together in 3D were different from the two classes that were previously identified by 1D analysis: dCTCF/BEAF-32 and Su(Hw), respectively [40]. Such observations strenghtened the importance of analyzing protein complexes in 3D in complement to 1D analysis (see Discussion).

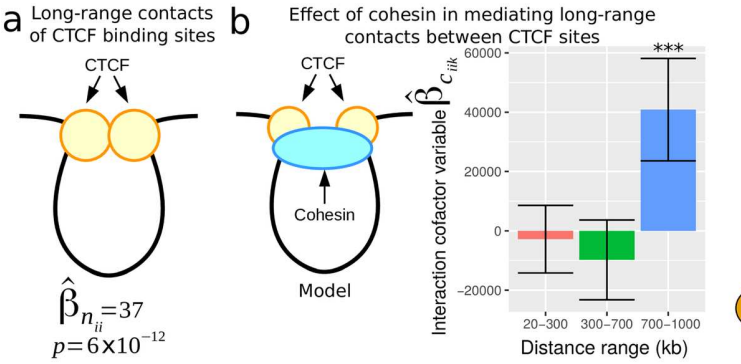
Analysis of DNA-binding protein sites in human

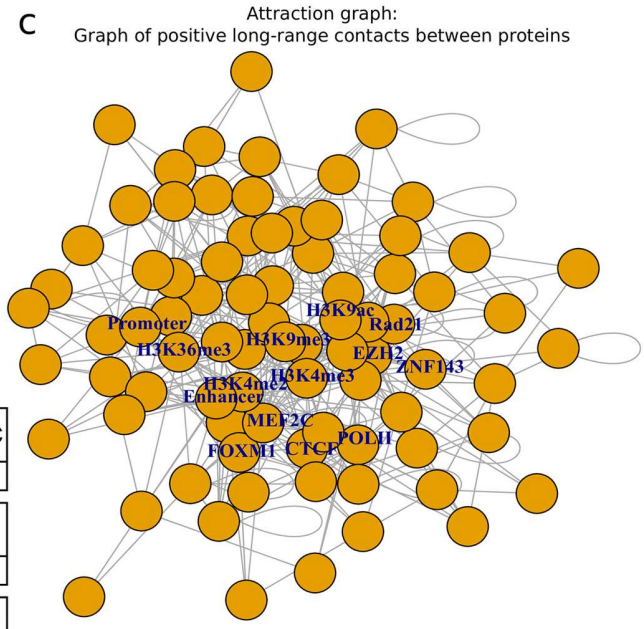
In human and mammals, the main model of loop formation involves CTCF and cohesin [10, 17]. According to this model, a loop may form by the homodimerization of two CTCF proteins

bound to two distant CTCF motifs that are in convergent orientation [10]. The loop also involves cohesin that is recruited by CTCF and that has the ability to entrap the two DNA fibers inside a ring. In addition to CTCF and cohesin, other architectural proteins have been recently uncovered such as ZNF143 [41] and PcG proteins [42]. In order to systematically analyze proteins mediating loops, we considered integrating available protein binding data (73 proteins) together with high-resolution Hi-C data in human GM12878 cells using our GLMI model. As previously done for *Drosophila*, we analyzed Hi-C data at 10 kb resolution and focused on 20kb-1Mb distances [10]. At this distance range, the Hi-C data comprised a very large number of bin pairs (around 22 millions), and hence, its analysis often required subsampling to few million pairs to achieve tractable regression parameter estimation. As for *Drosophila*, the log-log relation between Hi-C count and distance was linear at this distance range (R^2 = 0.992, S2 Fig), supporting the use of the log-distance term in the model.

We first investigated contacts between distant CTCF binding sites using model (2). As expected, we observed strong long-range contacts ($\hat{\beta}_{n_{ii}} = 37, p = 6 \times 10^{-12}$) [10]. Moreover high levels of long-range contacts were detected between cohesin subunit Rad21 binding sites as expected ($\hat{\beta}_{n_{ii}} = 89, p < 10^{-20}$; Fig 6a) [10], as well as between cohesin subunit SMC3 ($\hat{\beta}_{n_{ii}} = 75, p < 10^{-20}$). We then used the same approach to estimate long-range contacts for all 73 proteins available (S1 Table). Among the proteins that significantly interacted among themselves, we found several proteins known to colocalize to CTCF binding sites including YY1 ($\hat{\beta}_{n_{ii}} = 31, p < 10^{-20}$), MAZ ($\hat{\beta}_{n_{ii}} = 16, p < 10^{-20}$) and JUND ($\hat{\beta}_{n_{ii}} = 258, p = 10^{-9}$) [7]. We also found P300, an important transcriptional coactivator [43] ($\hat{\beta}_{n_{ii}} = 264, p < 10^{-20}$). In addition, histone marks including H3K27me3, H3K36me3, H3K4me2, H3K4me3, H3K9ac and H3K9me3 showed homologous long-range contacts, as previously shown by polymer simulations [44] (all $\hat{\beta}_{n_{ii}} > 0.05, p < 10^{-20}$). Curiously, H4K20me1 sites presented repulsive effects with each other ($\hat{\beta}_{n_{ii}} = -0.07, p < 10^{-20}$), indicating that distant H4K20me1 marked sites may avoid each other. We further estimated the well-known influence of cohesin in mediating long-range contacts between distant CTCF binding sites in human using model (4) [8, 10]. Interestingly, we found that the effect of cohesin depended on the distance between CTCF binding sites, with no significant contacts for short distances (20-300kb: $\hat{\beta}_{c_{iik}} = -3 \times 10^3$, p = 0.63; 300-700kb: $\hat{\beta}_{c_{iik}} = -1 \times 10^4$, p = 0.15) and significant contacts for long distances (700-1000kb: $\hat{\beta}_{c_{iik}} = 4 \times 10^4$, $p = 3 \times 10^{-6}$) (Fig 6b). This suggested that cohesin is required for stabilizing CTCF-mediated loops for long distances, but is not necessary for short distances for which homodimerization of CTCF might be sufficient. We also sought for other proteins whose loops could be mediated by cohesin for long distances (S2 Table). Most notably, we found that cohesin positively influences long-range contacts between architectural protein ZNF143 binding sites ($\hat{\beta}_{c_{iik}} = 4.8 \times 10^4$, $p = 2 \times 10^{-9}$), between PolII binding sites ($\hat{\beta}_{c_{iik}} = 446$, $p = 6 \times 10^{-16}$), and between transcriptional factor binding sites (EGR1, ELF1, FOXM1, MAZ, MXI1, NRF1, YY1), which suggests a wider role for cohesin in mediating long-range contacts.

Further analyses of long-range contacts for every couple of proteins were performed using model (10) that included together all possible interaction variables. We considered 73 proteins, 7 histone modifications, active enhancers and active promoters. The model thus comprised $(82 \times 83)/2 = 3403$ interaction variables. To deal with such a large number of interaction variables, we used a Poisson lasso estimation [45]. An interaction variable beta of zero was expected to reflect the absence of direct long-range contact between two proteins. From the estimated betas, we built a first graph that we called "attraction graph" by adding an edge





d

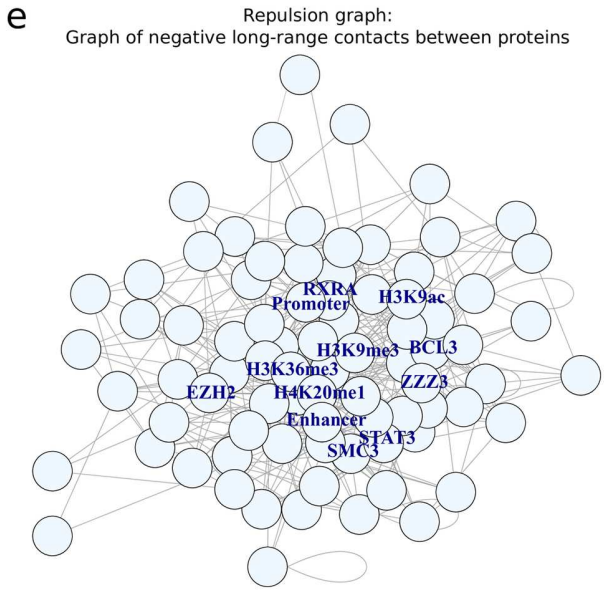
Highest node centrality scores of the attraction graph

Genomic Feature	H3K9ac	H3K9me3	H3K4me3	Enhancer	H3K4me2	MEF2C
Hub score	1,000	0,980	0,948	0,840	0,783	0,725

Genomic Feature	FOXM1	IKZF1	H3K36me3	POL3	PML	WHIP
Hub score	0,692	0,690	0,685	0,668	0,650	0,648

Genomic Feature	CTCF	CHD1	ATF3	STAT1	E2F4	STAT3
Hub score	0,619	0,590	0,582	0,579	0,570	0,556

Genomic Feature	RAD21	EZH2	POL2	RFX5	Promoter	GABP
Hub score	0,555	0,551	0,551	0,540	0,534	0,527



f

Highest node centrality scores of the repulsion graph

Genomic Feature	H3K36me3	H4K20me1	H3K9me3	Enhancer	RXRA	Promoter
Hub score	1,000	0,974	0,798	0,766	0,759	0,697

Genomic Feature	EZH2	BCL3	ZZZ3	SMC3	STAT3	MEF2C
Hub score	0,673	0,664	0,654	0,633	0,591	0,591

Genomic Feature	SPI1	CEBPB	RAD21	EBF1	TCF12	BCLAF1
Hub score	0,589	0,588	0,578	0,575	0,534	0,529

Genomic Feature	BATF	POL3	H3K9ac	CHD1	ZNF274	IKZF1
Hub score	0,528	0,524	0,524	0,519	0,518	0,504

Fig 6. Analysis of long-range contacts between architectural protein binding (IBP) sites in human GM12878 cells. a) Long-range contacts between CTCF sites, and between Rad21 sites, as measured by interaction variable betas estimated using model (2). b) Effect of cohesin in mediating long-range contacts between CTCF sites. c) Attraction graph of long-range contacts between DNA-binding protein sites estimated using positive interaction variable betas from model (10). d) Highest node centrality scores from the attraction graph as measured by eigen decomposition. e) Repulsion graph of long-range contacts between DNA-binding protein sites estimated using negative interaction variable betas from model (10). f) Highest node centrality scores from the repulsion graph as measured by eigen decomposition.

https://doi.org/10.1371/journal.pcbi.1005538.g006

between two proteins F_i and F_j if $\hat{\beta}_{n_{ij}} > 0$, and by adding an edge between a protein F_i and itself if $\hat{\beta}_{n_{ii}} > 0$ (Fig 6c). To identify hubs in the graph, we used eigenvector centrality that reflected how central is a node (Fig 6d). Both active and repressed chromatin marks as well as enhancers were the most central nodes (H3K9ac: score = 1; H3K9me3: score = 0.98; H3K4me3: score = 0.948; Enhancer: score = 0.84). Among DNA-binding proteins, CTCF and Rad21 showed high values (CTCF: score = 0.619; Rad21: score = 0.555). Surprisingly, however, other proteins MEF2C and FOXM1 presented the highest values (MEF2C: score = 0.725; FOXM1: score = 0.692). Previous studies showed that MEF2C is necessary for bone marrow B-lymphopoiesis (GM12878 is a lymphoblastoid cell line) [46], and that FOXM1 has an important role in maintenance of chromosomal segregation [47]. We then looked for cliques in the graph, *i.e.* a group of nodes that were all connected to each other (complete list in S3 Table). As expected, we found a clique composed of CTCF and the cohesin subunits Rad21 and SMC3, that are known to mediate together loops [10]. But we also found novel protein complexes that were specific to lymphocyte B such as the clique IKZF1/RFX5/PolII. IKZF1 plays a role in the development of lymphocytes [48], RFX5 is involved in bare lymphocyte syndrome [49] and polymerase II catalyzes gene transcription. In addition, we found many cliques involving Polymerase III (PolIII) such as the cliques MEF2C/RUNX3/PolIII and MEF2C/WHIP/PolIII, which might reflect the influence of architectural protein RNA polymerase III-associated factor (TFIIIC) at tRNA genes [2, 50].

Very little is known about repulsion effects between distant binding sites. Such repulsive effects could result from allosteric effects of loops [51], or factors that disassociate protein complexes involved in loops [52]. To investigate repulsive effects, we built a second graph that we called "repulsion graph" by adding an edge between two proteins F_i and F_j if $\hat{\beta}_{n_{ij}} < 0$, and by adding an edge between a protein F_i and itself if $\hat{\beta}_{n_{ii}} < 0$ (Fig 6e). The repulsion graph was very different from the attraction graph. Different histone marks were central in the repulsion graph, including H3K36me3 (score: 1) and H4K20me1 (score: 0.974), except histone mark H3K9me3 (score: 0.798) that was central in both the attraction and repulsion graphs (Fig 6f). Interestingly, we found that enhancers presented a high centrality score in the repulsion graph (score: 0.766), as found in the attraction graph. This result highlights the ability of enhancers to specifically interact with distant protein partner binding sites while avoiding others. Supporting this interpretation, we found enhancers to be in attraction with CFOS, NRF1 or POU2F2, and in repulsion with RXRA, NFE2 or P300. We then looked at pairs of proteins that were in repulsion. Most notably, we found CTCF to be in repulsion with EZH2, which might result from steric effects of CTCF-mediated loops [10] with Polycomb-mediated loops [42].

The influence of DNA-binding proteins on enhancer-promoter interactions in human

Enhancer-promoter (EP) interactions play an essential role in the regulation of gene expression [14, 18]. Therefore, we explored the roles of DNA-binding proteins in establishing or maintaining EP interactions. Before assessing the role of proteins, we first measured long-range contacts between active enhancers and promoters depending on gene expression using model (3) (Fig 7a). We observed an attraction effect between active enhancers and highly expressed gene promoters ($\hat{\beta}_{n_{ij}} = 2, p = 3 \times 10^{-5}$), and conversely, a repulsion effect between active enhancers and low expressed gene promoters ($\hat{\beta}_{n_{ij}} = -1.7, p < 1 \times 10^{-20}$), in complete agreement with the established positive influence of long-range contacts on gene expression [53]. To identify the influence of DNA-binding proteins, we then assessed the presence of

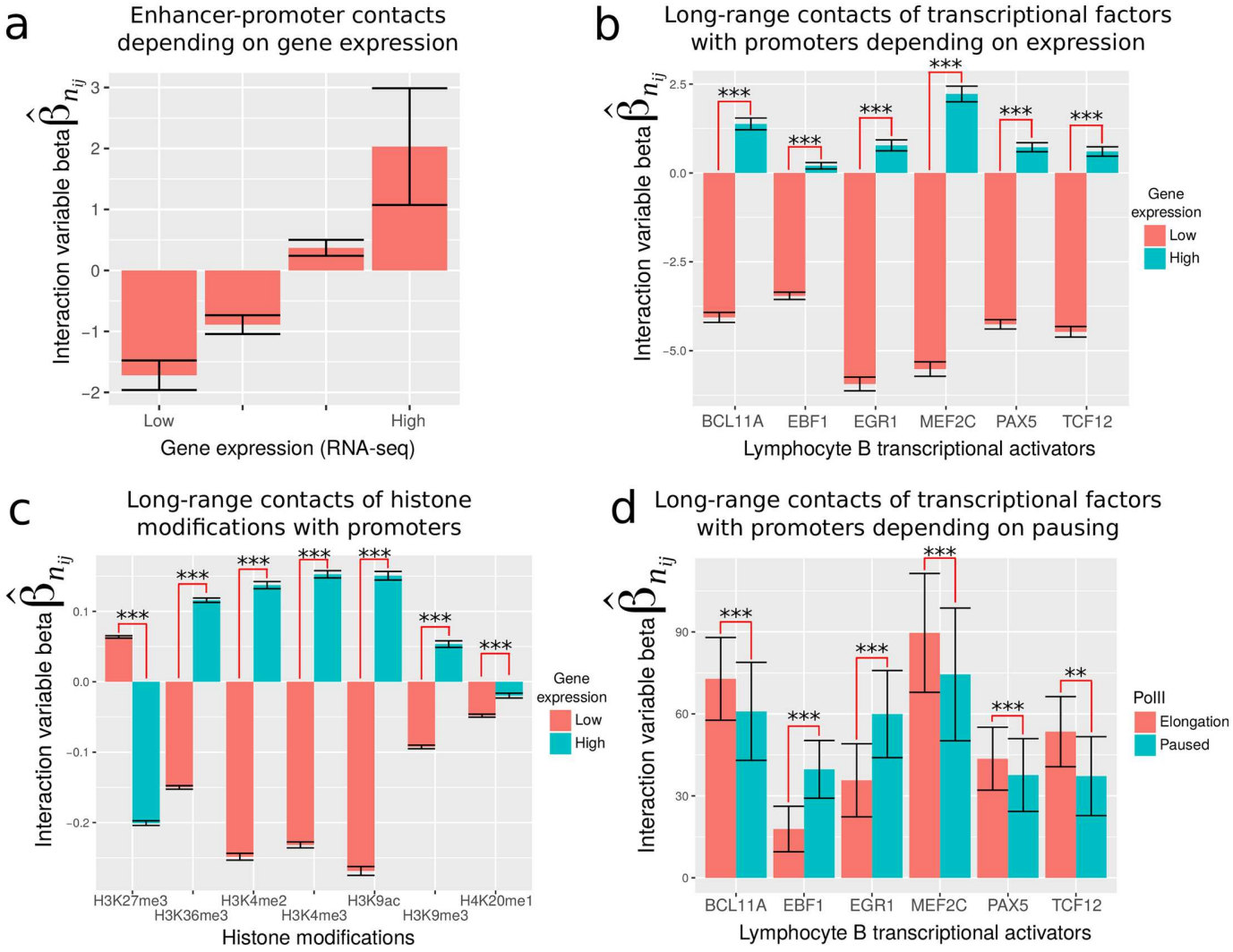


Fig 7. Influence of DNA-binding proteins and histone marks on enhancer-promoter contacts in human GM12878 cells. a) Enhancer-promoter contacts depending on gene expression, as measured by interaction variable betas estimated using model (3). b) Long-range contacts of transcriptional factors with promoters depending on gene expression. c) Long-range contacts of histone modifications with promoters depending on gene expression. d) Long-range contacts of transcriptional factors with promoters depending on PolII pausing or elongation.

https://doi.org/10.1371/journal.pcbi.1005538.g007

long-range contacts between lymphocyte B transcriptional activator binding sites (ChIP-seq data) and promoters using the same model (3). All lymphocyte B transcriptional activators including BCL11A, EBF1, EGR1, MEF2C, PAX5 and TCF12 showed long-range contacts with highly expressed gene promoters, compared to weakly transcribed gene promoters (Fig 7b). This clearly showed that lymphocyte B transcriptional activators regulate expression of target genes through long-range contacts. Among the proteins available, we could not identify any that acted as silencers, *i.e.* proteins whose long-range contacts are high with low expressed gene promoters and low with highly expressed gene promoters. However when we focused on histone modifications, we found that long-range contacts of H3K27me3 mark were stronger to weakly transcribed gene promoters ($\hat{\beta}_{n_{ij}} = 0.06, p < 10^{-20}$), compared to highly expressed gene promoters ($\hat{\beta}_{n_{ij}} = -0.2, p < 10^{-20}$) (Fig 7c). This suggested that H3K27me3 mark not

only acts as a transcriptional silencer in linear proximity [54], but could also repress target genes at distance through loops. Conversely, active marks such as H3K4me3 and H3K9ac interacted more with highly expressed genes. Because enhancer-promoter contacts were previously shown to be associated with Polymerase II pausing [18], we then assessed enhancer-promoter interactions depending on gene transcription pausing. As expected, we found higher EP contacts at paused genes ($\hat{\beta}_{n_{ij}} = 62.2$, $p = 10^{-3}$), compared to genes in elongation ($\hat{\beta}_{n_{ij}} = 49.3$, $p = 2 \times 10^{-3}$). We then looked at the influence of DNA-binding proteins (Fig 7d). For instance, EBF1 sites showed higher long-range contacts with promoters of genes in pause ($\hat{\beta}_{n_{ij}} = 39.7$, $p = 1 \times 10^{-13}$), compared to those in elongation ($\hat{\beta}_{n_{ij}} = 17.8$, $p = 3 \times 10^{-5}$), in agreement with [18]. But, surprisingly, we also found that BCL11A sites showed higher long-range contacts with promoters of genes in elongation ($\hat{\beta}_{n_{ij}} = 72.8$, $p < 10^{-20}$) than with genes in pause ($\hat{\beta}_{n_{ij}} = 60.9$, $p = 2 \times 10^{-11}$). These observations suggest that, depending on the protein involved, long-range contacts with promoters are not always associated with pausing, but could also be linked to elongation.

Conclusion

Here, we propose to use a generalized linear regression with interactions (GLMI) to study the roles of genomic features such as DNA-binding proteins, motifs or promoters to bridge long-range contacts in the genome, depending on transcriptional status or motif orientation. GLMI has multiple assets over existing approaches such as enrichment test, correlation and random forests. Compared to enrichment test [2, 55] or correlation [27] that respectively assesses the protein enrichment or correlation at highly confident loops, GLMI quantitatively links the frequency of all long-range contacts to complex co-occupancies of proteins while accounting for known Hi-C biases and polymer background. Moreover, GLMI accounts for colocalizations among protein binding, a strong issue when analyzing protein binding sites known to largely overlap over the genome. In contrast to random forests [28] which are efficient predictive models but sometimes poor explanatory ones, GLMI allows to identify key chromatin loop driver proteins and motifs. GLMI can also uncover numerous mechanisms behind loop formation using higher-order interaction terms and proper confounding variables. For instance, GLMI can determine if a cofactor is necessary to mediate long-range contacts between distant protein binding sites.

Using real *Drosophila* Hi-C and ChIP-seq data, we validate numerous GLMI predictions of long-range contacts that involve insulator binding proteins, cofactors and motifs, and which were confirmed by previous microscopy and mutational studies. For instance, our model estimates long-range contacts between distant BEAF-32 motifs, which were previously observed with both fluorescence cross-correlation spectroscopy [22] and high-resolution microscopy [23]. In addition, our model finds a mediating role of CP190 in bridging long-range contacts between distant BEAF-32 and GAF binding sites, in agreement with mutational experiments [19]. Of interest, GLMI analyses highlight a role of cohesin in stabilizing long-range contacts between CTCF sites in *Drosophila*, similarly to its role in human [7]. Supporting this role, we show that such influence is reduced upon cohesin subunit Rad21 depletion. It has to be noted that the absence of complete loss of contacts between CTCF sites after Rad21 depletion can be explained by the fast turnover of chromosome-bound cohesin in interphase [56]. Moreover, GLMI outperforms enrichment test, correlation and random forests in the identification of known architectural proteins and motifs, and in the detection of the effects of mutations in the dCTCF motif.

The proposed model also uncovers several novel results. In *Drosophila*, GAF and ZW5 motifs are shown to act in divergent orientation to form loops, in contrast to CTCF motifs that are found in convergent orientation in *Drosophila* and human [10, 17], suggesting a different mode of action of corresponding proteins. In addition, we identify two groups of proteins that act in 3D to form loops. The first group comprises BEAF-32, dTFIIIC and GAF, and the other group includes DREF, Su(Hw) and dCTCF. Those groups are different from the ones observed with 1D analysis only (*i.e.* linear colocalization on the genome) [40], highlighting the importance of 3D analysis using GLMI. In human, we identify numerous long-range contacts between protein binding sites. In addition to the well-known protein complex CTCF/RAD21/SMC3, we uncover new protein complexes that are specific to lymphocyte B such as IKZF1/RFX5. We also found that enhancers could be either in long-range contact or repulsion with certain protein binding sites, highlighting potential specificity in selecting protein partners for long-range contacts. Our observations therefore support the idea that enhancer-promoter contacts are not solely driven by insulators or TAD borders that physically constrain such long-range interactions [29, 36, 57]. Rather, enhancer-promoter contacts may also be encoded by the specificity of protein-protein interactions. In addition, our results suggest that repressive mark H3K27me3 does not only repress genes that are contigous [54], but it could also repress from a distance through the juxtaposition of H3K27me3 with genes in 3D. We also find that, depending on the protein involved, long-range enhancer-promoter contacts are not always favored by PolII pausing [18], which may highlight distinct mechanisms by which proteins can influence transcription-associated long-range contacts.

There are several limitations of the proposed approach. First, the present analysis is restricted to a 10-kb resolution because of the quadratic complexity of Hi-C data. Second, our analysis is limited by the amount of higher-order interaction variable parameters that can be learned within the same model (full model) using current parameter learning programs. Most notably, all possible interaction cofactor variables cannot be included in the same model because of the cubic complexity of such model, and hence they are learned separately instead (using models (4) and (5)). In addition, although generalized linear models can include interactions of any order involving large protein complexes (for instance, complexes of more than 4 proteins), parameter learning is limited by the availability of data and computational resources. Increasing depth of Hi-C data will allow inference of more complex models in the near future. Moreover the development of new big data learning algorithms could be used to process the data at a higher resolution that would allow in-depth analysis of 3D chromatin drivers [58]. An alternative to the exploration of all possible higher-order interactions together might be to guide the search using prior information, such as protein-protein interaction network [55]. Lastly, in order to explore all possible higher-order interaction variables within the same model (full model), one should use a lasso regression model with hierarchically constrained interactions [59].

Materials and methods

Hi-C data

We used publicly available high-throughput chromatin conformation capture (Hi-C) data from Gene Expression Omnibus (GEO) accession GSE62904 [21]. Hi-C experiments have been done for *Drosophila melanogaster* wild-type and Rad21 knock-down Kc167 cells with DpnII restriction enzyme. Hi-C data were binned at 1 and 10 kb resolutions.

For human data analysis, we used publicly available Hi-C data of lymphoblastoid cells GM12878 cells from Gene Expression Omnibus (GEO) accession GSE63525 [10]. We used Hi-C data binned at 10 kb resolution.

ChIP-seq data

For *Drosophila* analysis, we used publicly available binding profiles of chromatin proteins of *Drosophila melanogaster* wild-type embryonic Kc167 cells. ChIP-seq data for CP190, Su(Hw), dCTCF and BEAF-32 were obtained from GEO accession GSE30740 [60]. ChIP-seq data for Barren (condensin I), Cap-H2 (condensin II), Chromator, Rad21 (cohesin), GAF and dTFIIIC were obtained from GEO accession GSE54529 [9]. ChIP-seq data for DREF and L(3)Mbt were obtainted from GEO accession GSE62904 [21]. ChIP-seq data for Fs(1)h-L and Fs(1)h-LS were obtained from GEO accession GSE42086 [25]. Peak calling was done using MACS 2.1.0 (https://github.com/taoliu/MACS).

For human analysis, we used publicly available binding peaks of 73 chromatin proteins (RAD21, CTCF, YY1, ZBTB33, MAZ, JUND, ZNF143, EZH2, ATF2, ATF3, BATF, BCL11A, BCL3, BCLAF1, BHLHE40, BRCA1, CEBPB, CFOS, CHD1, CHD2, CMYC, COREST, E2F4, EBF1, EGR1, ELF1, ELK1, FOXM1, GABP, IKZF1, IRF4, MAX, MEF2C, MTA3, MXI1, NFATC1, NFE2, NFIC, NFKB, NFYA, NFYB, NRF1, NRSF, P300, PAX5, PBX3, PML, POL2, POL3, POU2F2, RFX5, RUNX3, RXRA, SIN3A, SIX5, SMC3, SP1, SPI1, SRF, STAT1, STAT3, STAT5, TBLR1, TBP, TCF12, TCF3, TR4, USF1, USF2, WHIP, ZEB1, ZNF274, ZZZ3) and histone marks (H3K27me3, H3K36me3, H3K4me2, H3K4me3, H3K9ac, H3K9me3, H4K20me1) of GM12878 cells from ENCODE [61]. We downloaded peaks that were uniformly processed (Uniform Peaks).

Functional elements

For human analysis, we divided promoters into quartiles of gene expression using RNA-seq data [61]. We also divided promoters into quartiles of gene pausing and into quartiles of gene elongation using PolII ChIP-seq data [61]. For enhancer mapping, we used lymphocyte of B lineage differentially expressed enhancers identified from the Fantom5 project [62].

DNA motifs

For both *Drosophila* and human analyses, we used transcription factor binding site (TFBS) motifs from the MotifMap database (http://motifmap.ics.uci.edu/).

Power-law distribution testing

The proposed GLMI assumed a linear relation between logarithm of Hi-C counts and the logarithm of distance between bins as previously shown in [5]. This assumption only holds locally, *i.e.* for a specific distance scale. Hence we restricted GLM modeling to a certain range of distances, *e.g.* for 20kb to 1Mb. In addition, we tested this assumption on data before using GLMI. We considered that this assumption holds when the $R^2 > 0.95$.

Occupancy variables z

Before computing variables for the GLMI presented above, intermediate variables from the genomic features such as DNA-binding proteins needed to be calculated. Intermediate "occupancy" variable $\mathbf{z}_i$ denoted the presence ($\mathbf{z}_i = 1$) or absence ($\mathbf{z}_i = 0$) of the protein F_i within the genomic bin. If the protein only overlapped 60% of the genomic bin, then $\mathbf{z}_i = 0.6$.

The different models

Here are described the different models derived from model (1) that we used. In order to assess a homologous interaction variable $\mathbf{n}_{ii} = \mathbf{z}_{iL} \times \mathbf{z}_{iR}$ (here $\mathbf{g} = \mathbf{n}_{ii}$), model (1) becomes:

$$\begin{aligned} \log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{n_{ii}}\mathbf{n}_{ii} \end{aligned} \quad (2)$$

Following the hierarchy principle in (generalized) linear models, the assessment of a statistical interaction variable, such as $\mathbf{n}_{ii} = \mathbf{z}_{iL} \times \mathbf{z}_{iR}$, must include both $\mathbf{z}_{iL}$ and $\mathbf{z}_{iR}$ as confounding variables. Because $\mathbf{z}_{iL}$ and $\mathbf{z}_{iR}$ are identically associated to $\mathbf{y}$ (the attribution for left and right bins is arbitrary), their values are averaged to give $\mathbf{m}_i = \frac{1}{2}(\mathbf{z}_{iL} + \mathbf{z}_{iR})$. Hence $\mathbf{C} = \mathbf{m}_i$ is used as a confounder of $\mathbf{n}_{ii}$.

In order to assess a heterologous interaction variable $\mathbf{n}_{ij} = \frac{1}{2}(\mathbf{z}_{iL} \times \mathbf{z}_{jR} + \mathbf{z}_{jL} \times \mathbf{z}_{iR})$ (here $\mathbf{g} = \mathbf{n}_{ij}$), model (1) becomes:

$$\begin{aligned} \log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{m_j}\mathbf{m}_j + \beta_{n_{ij}}\mathbf{n}_{ij} \end{aligned} \quad (3)$$

Following the hierarchy principle, $\mathbf{z}_{iL}$, $\mathbf{z}_{iR}$, $\mathbf{z}_{jL}$ and $\mathbf{z}_{jR}$ have to be included as confounding variables. As previously, $\mathbf{z}_{iL}$ and $\mathbf{z}_{iR}$ are averaged to give $\mathbf{m}_i = \frac{1}{2}(\mathbf{z}_{iL} + \mathbf{z}_{iR})$. Similarly, $\mathbf{z}_{jL}$ and $\mathbf{z}_{jR}$ are averaged to give $\mathbf{m}_j = \frac{1}{2}(\mathbf{z}_{jL} + \mathbf{z}_{jR})$. Hence $\mathbf{C} = \{\mathbf{m}_i, \mathbf{m}_j\}$ is used as confounder of $\mathbf{n}_{ij}$.

In order to assess a homologous interaction cofactor variable $\mathbf{c}_{iik} = \mathbf{n}_{ii} \times \mathbf{n}_{kk}$ (here $\mathbf{g} = \mathbf{c}_{iik}$), model (1) becomes:

$$\begin{aligned} \log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{m_k}\mathbf{m}_k + \beta_{m_{ik}}\mathbf{m}_{ik} + \beta_{n_{ii}}\mathbf{n}_{ii} + \beta_{n_{kk}}\mathbf{n}_{kk} + \beta_{n_{ik}}\mathbf{n}_{ik} \\ &+ \beta_{n_{ii}\times m_k}(\mathbf{n}_{ii} \times \mathbf{m}_k) + \beta_{n_{kk}\times m_i}(\mathbf{n}_{kk} \times \mathbf{m}_i) + \beta_{c_{iik}}\mathbf{c}_{iik}, \end{aligned} \quad (4)$$

Here variable $\mathbf{c}_{iik}$ is a four-way interaction term and hence there are a large number of confounding variables included in variable set $\mathbf{C} = \{\mathbf{m}_i, \mathbf{m}_k, \mathbf{m}_{ik}, \mathbf{n}_{ii}, \mathbf{n}_{kk}, \mathbf{n}_{ik}, \mathbf{n}_{ii} \times \mathbf{m}_k, \mathbf{n}_{kk} \times \mathbf{m}_i\}$. We need to introduce a new type of variable, noted $\mathbf{m}_{ij}$, the average of product $\mathbf{z}_{iL} \times \mathbf{z}_{jL}$ and product $\mathbf{z}_{iR} \times \mathbf{z}_{jR}$ ($\mathbf{m}_{ij} = \frac{1}{2}(\mathbf{z}_{iL} \times \mathbf{z}_{jL} + \mathbf{z}_{iR} \times \mathbf{z}_{jR})$). For a detailed explanation of the confounder set $\mathbf{C}$, see S1 Appendix, Confounder sets.

In order to assess a heterologous interaction cofactor variable $\mathbf{c}_{ijk} = \mathbf{n}_{ij} \times \mathbf{n}_{kk}$ (here $\mathbf{g} = \mathbf{c}_{ijk}$), model (1) becomes:

$$\begin{aligned} \log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{m_j}\mathbf{m}_j + \beta_{m_k}\mathbf{m}_k + \beta_{m_{ik}}\mathbf{m}_{ik} + \beta_{m_{jk}}\mathbf{m}_{jk} \\ &+ \beta_{n_{ij}}\mathbf{n}_{ij} + \beta_{n_{jk}}\mathbf{n}_{jk} + \beta_{n_{ik}}\mathbf{n}_{ik} + \beta_{n_{kk}}\mathbf{n}_{kk} \\ &+ \beta_{n_{ij}\times m_k}\mathbf{n}_{ij} \times \mathbf{m}_k + \beta_{n_{kk}\times m_i}\mathbf{n}_{kk} \times \mathbf{m}_i + \beta_{n_{kk}\times m_j}\mathbf{n}_{kk} \times \mathbf{m}_j + \beta_{c_{ijk}}\mathbf{c}_{ijk}. \end{aligned} \quad (5)$$

Here variable $\mathbf{c}_{ijk}$ is a four-way interaction term and hence there are a large number of confounding variables included in variable set $\mathbf{C} = \{\mathbf{m}_i, \mathbf{m}_j, \mathbf{m}_k, \mathbf{m}_{ik}, \mathbf{m}_{jk}, \mathbf{n}_{ij}, \mathbf{n}_{jk}, \mathbf{n}_{ik}, \mathbf{n}_{kk}, \mathbf{n}_{ij} \times \mathbf{m}_k, \mathbf{n}_{kk} \times \mathbf{m}_i, \mathbf{n}_{kk} \times \mathbf{m}_j\}$. For a detailed explanation of the confounder set $\mathbf{C}$, see S1 Appendix, Confounder sets.

In addition, we formulated models for homologous interaction variables, depending on motif pair orientation. For a pair of motifs in convergent orientation (→←), model (1)

becomes:

$$\begin{aligned}\log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{z_{iL+}}\mathbf{z}_{iL+} + \beta_{z_{iR-}}\mathbf{z}_{iR-} + \beta_{n_{ii+-}}\mathbf{n}_{ii+-}\end{aligned} \quad (6)$$

with $\mathbf{n}_{ii+-} = \mathbf{z}_{iL+} \times \mathbf{z}_{iR-}$. Symbol "+" denoted motifs that were on the forward DNA strand, while symbol "-" denoted motifs that were on the reverse DNA strand. For instance, variable $\mathbf{z}_{iL+}$ was the occupancy of a motif on the forward DNA strand within genomic bins.

For a pair of motifs in divergent orientation (←→), model (1) becomes:

$$\begin{aligned}\log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{z_{iL-}}\mathbf{z}_{iL-} + \beta_{z_{iR+}}\mathbf{z}_{iR+} + \beta_{n_{ii-+}}\mathbf{n}_{ii-+},\end{aligned} \quad (7)$$

with $\mathbf{n}_{ii-+} = \mathbf{z}_{iL-} \times \mathbf{z}_{iR+}$.

For a pair of motifs in same orientation (→→), model (1) becomes:

$$\begin{aligned}\log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{z_{iL+}}\mathbf{z}_{iL+} + \beta_{z_{iR+}}\mathbf{z}_{iR+} + \beta_{n_{ii++}}\mathbf{n}_{ii++},\end{aligned} \quad (8)$$

with $\mathbf{n}_{ii++} = \mathbf{z}_{iL+} \times \mathbf{z}_{iR+}$.

For a pair of motifs in same orientation (←←), model (1) becomes:

$$\begin{aligned}\log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_g\mathbf{g} \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{z_{iL-}}\mathbf{z}_{iL-} + \beta_{z_{iR-}}\mathbf{z}_{iR-} + \beta_{n_{ii--}}\mathbf{n}_{ii--},\end{aligned} \quad (9)$$

with $\mathbf{n}_{ii--} = \mathbf{z}_{iL-} \times \mathbf{z}_{iR-}$.

Moreover, we formulated an additional "full" model where all possible homologous and heterologous interaction variables were included. For instance, if we study two proteins F_i and F_j that tend to linearly colocalize, then the following "full" model would be:

$$\begin{aligned}\log(\mathrm{E}[\mathbf{y}|\mathbf{X}]) &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_C\mathbf{C} + \beta_G\mathbf{G}, \\ &= \beta_0 + \beta_d\mathbf{d} + \beta_B\mathbf{B} + \beta_{m_i}\mathbf{m}_i + \beta_{m_j}\mathbf{m}_j + \beta_{n_{ii}}\mathbf{n}_{ii} + \beta_{n_{jj}}\mathbf{n}_{jj} + \beta_{n_{ij}}\mathbf{n}_{ij},\end{aligned} \quad (10)$$

where $\mathbf{G}$ is the set of all possible homologous and heterologous interaction variables. Here $\mathbf{G} = \{\mathbf{n}_{ii}, \mathbf{n}_{jj}, \mathbf{n}_{ij}\}$ for two proteins F_i and F_j. The confounder set $\mathbf{C} = \{\mathbf{m}_i, \mathbf{m}_j\}$ includes all marginal variables.

Implementation

The general linear regression with interactions is implemented in R language. The model is available in the R package "HiCglmi" which can be downloaded from the Comprehensive R Archive Network.

Supporting information

S1 Appendix. Bias variable computation and confounder sets.
(PDF)

S1 Fig. Log-log relation between Hi-C count and distance between bins in *Drosophila*.
20 kb resolution for distances comprised between 10kb and 1Mb. *Drosophila* Kc167 cell data.
(PDF)

S2 Fig. Log-log relation between Hi-C count and distance between bins in human. 20 kb resolution for distances comprised between 10kb and 1Mb. Human GM12878 cell data.
(PDF)

S1 Table. Long-range contacts between same genomic feature. Long-range contacts measured by homologous interaction variable betas. GM12878 cell ChIP-seq data.
(PDF)

S2 Table. Mediating effect of cohesin (Rad21 subunit) on long-range contacts between same genomic feature. Mediating effect of cohesin measured by homologous interaction cofactor variable betas. GM12878 cell ChIP-seq data.
(PDF)

S3 Table. Cliques from the attraction graph. GM12878 cell ChIP-seq data.
(PDF)

Acknowledgments

The authors thank Pascal Martin and Laurent Lacroix for useful discussions. The authors are grateful to Corces lab (Emory University, USA) for data and for help in processing them. The authors are also grateful to the genotoul bioinformatics platform Toulouse Midi-Pyrenees for providing computing resources.

Author Contributions

Conceptualization: RM LL OC.

Data curation: RM.

Formal analysis: RM LL.

Funding acquisition: RM OC.

Investigation: RM.

Methodology: RM.

Project administration: RM.

Software: RM.

Supervision: RM.

Visualization: RM.

Writing – original draft: RM LL OC.

References

1. Halverson JD, Smrek J, Kremer K, Grosberg AY. From a melt of rings to chromosome territories: the role of topological constraints in genome folding. Reports on Progress in Physics. 2014; 77(2):022601. https://doi.org/10.1088/0034-4885/77/2/022601 PMID: 24472896
2. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012; 485(7398):376–380. https://doi.org/10.1038/nature11082 PMID: 22495300
3. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, et al. Three-dimensional folding and functional organization principles of the *Drosophila* genome. Cell. 2012; 148(3):458–472. https://doi.org/10.1016/j.cell.2012.01.010 PMID: 22265598

4. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, et al. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013; 503(7475):290–294. https://doi.org/10.1038/nature12644 PMID: 24141950

5. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009; 326(5950):289–293. https://doi.org/10.1126/science.1181369 PMID: 19815776

6. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, et al. Topologically associating domains are stable units of replication-timing regulation. Nature. 2014; 515(7527):402–405. https://doi.org/10.1038/nature13986 PMID: 25409831

7. Cubenas-Potts C, Corces VG. Architectural proteins, transcription, and the three-dimensional organization of the genome. FEBS Letters. 2015; 589(20PartA):2923–2930. https://doi.org/10.1016/j.febslet.2015.05.025 PMID: 26008126

8. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013; 153(6):1281–1295. https://doi.org/10.1016/j.cell.2013.04.053 PMID: 23706625

9. Van Bortle K, Nichols MH, Li L, Ong CT, Takenaka N, Qin ZS, et al. Insulator function and topological domain border strength scale with architectural protein occupancy. Genome Biology. 2014; 15(5):R82 +. https://doi.org/10.1186/gb-2014-15-5-r82 PMID: 24981874

10. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, et al. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2015; 159(7):1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

11. Zuin J, Dixon JR, van der Reijden MIJA, Ye Z, Kolovos P, Brouwer RWW, et al. Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. Proceedings of the National Academy of Sciences. 2014; 111(3):996–1001. https://doi.org/10.1073/pnas.1317788111

12. Mourad R, Cuvier O. Computational identification of genomic features that influence 3D chromatin domain formation. PLoS Computational Biology. 2016; 12(5):e1004908. https://doi.org/10.1371/journal.pcbi.1004908 PMID: 27203237

13. Mourad R, Cuvier O. Predicting the spatial organization of chromosomes using epigenetic data. Genome Biology. 2015; 16(1):1–3. https://doi.org/10.1186/s13059-015-0752-8 PMID: 26319942

14. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nature Reviews Genetics. 2013; 14(6):390–403. https://doi.org/10.1038/nrg3454 PMID: 23657480

15. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012; 28(23):3131–3133. https://doi.org/10.1093/bioinformatics/bts570 PMID: 23023982

16. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, et al. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nature Methods. 2012; 9(10):999–1003. https://doi.org/10.1038/nmeth.2148 PMID: 22941365

17. Sanborn AL, Rao SSP, Huang SC, Durand NC, Huntley MH, Jewett AI, et al. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proceedings of the National Academy of Sciences. 2015; 112(47):E6456–E6465. https://doi.org/10.1073/pnas.1518552112 PMID: 26499245

18. Ghavi-Helm Y, Klein FA, Pakozdi T, Ciglar L, Noordermeer D, Huber W, et al. Enhancer loops appear stable during development and are associated with paused polymerase. Nature. 2014; 512(7512):96–100. PMID: 25043061

19. Liang J, Lacroix L, Gamot A, Cuddapah S, Queille S, Lhoumaud P, et al. Chromatin immunoprecipitation indirect peaks highlight functional long-range interactions among insulator proteins and RNAII pausing. Molecular Cell. 2014; 53(4):672–681. https://doi.org/10.1016/j.molcel.2013.12.029 PMID: 24486021

20. Phillips-Cremins JE, Corces VG. Chromatin insulators: Linking genome organization to cellular function. Molecular Cell. 2013; 50(4):461–474. https://doi.org/10.1016/j.molcel.2013.04.018 PMID: 23706817

21. Li L, Lyu X, Hou C, Takenaka N, Nguyen HQ, Ong CT, et al. Widespread rearrangement of 3D chromatin organization underlies Polycomb-mediated stress-induced silencing. Molecular Cell. 2015;(15):S1097–2765. https://doi.org/10.1016/j.molcel.2015.02.023 PMID: 25818644

22. Vogelmann J, Le Gall A, Dejardin S, Allemand F, Gamot A, Labesse G, et al. Chromatin insulator factors involved in long-range DNA interactions and their role in the folding of the *Drosophila* genome. PLoS Genetics. 2014; 10(8):e1004544. https://doi.org/10.1371/journal.pgen.1004544 PMID: 25165871

23. Georgieva M, Cattoni DI, Fiche JB, Mutin T, Chamousset D, Nollmann M. Nanometer resolved single-molecule colocalization of nuclear factors by two-color super resolution microscopy imaging. Methods. 2016; 105:44–55. http://dx.doi.org/10.1016/j.ymeth.2016.03.029. PMID: 27045944

24. Li J, Gilmour DS. Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. The EMBO Journal. 2013; 32(13):1829–1841. https://doi.org/10.1038/emboj.2013.111 PMID: 23708796

25. Kellner WA, Van Bortle K, Li L, Ramos E, Takenaka N, Corces VG. Distinct isoforms of the *Drosophila* Brd4 homologue are present at enhancers, promoters and insulator sites. Nucleic Acids Research. 2013; 41(20):9274–9283. https://doi.org/10.1093/nar/gkt722 PMID: 23945939

26. Wong KC, Li Y, Peng C. Identification of coupling DNA motif pairs on long-range chromatin interactions in human K562 cells. Bioinformatics. 2015;

27. Pancaldi V, Carrillo-de Santa-Pau E, Javierre BM, Juan D, Fraser P, Spivakov M, et al. Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity. Genome Biology. 2016; 17(1):1–19. https://doi.org/10.1186/s13059-016-1003-3 PMID: 27391817

28. He B, Chen C, Teng L, Tan K. Global view of enhancer-promoter interactome in human cells. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(21):201320308–E2199. https://doi.org/10.1073/pnas.1320308111 PMID: 24821768

29. Lupiáñez DG, Kraft K, Heinrich V, Krawitz P, Brancati F, Klopocki E, et al. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell. 2015; 161(5):1012–1025. https://doi.org/10.1016/j.cell.2015.04.004 PMID: 25959774

30. Tang Z, Luo OJ, Li X, Zheng M, Zhu JJ, Szalaj P, et al. CTCF-mediated human 3D genome architecture reveals chromatin topology for transcription. Cell. 2016; 163(7):1611–1627. https://doi.org/10.1016/j.cell.2015.11.024

31. Van Bortle K, Ramos E, Takenaka N, Yang J, Wahi JE, Corces VG. *Drosophila* CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains. Genome Research. 2012; 22(11):2176–2187. https://doi.org/10.1101/gr.136788.111 PMID: 22722341

32. Gibcus J, Dekker J. The hierarchy of the 3D genome. Molecular Cell. 2013; 49(5):773–782. http://dx.doi.org/10.1016/j.molcel.2013.02.011. PMID: 23473598

33. Filippova D, Patro R, Duggal G, Kingsford C. Identification of alternative topological domains in chromatin. Algorithms for Molecular Biology. 2014; 9(1):1–11. https://doi.org/10.1186/1748-7188-9-14 PMID: 24868242

34. Vietri Rudan M, Barrington C, Henderson S, Ernst C, Odom D, Tanay A, et al. Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. Cell Reports. 2015; 10(8):1297–1309. http://dx.doi.org/10.1016/j.celrep.2015.02.004. PMID: 25732821

35. Gómez-Marín C, Tena JJ, Acemel RD, López-Mayorga M, Naranjo S, de la Calle-Mustienes E, et al. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proceedings of the National Academy of Sciences. 2015; 112(24):7542–7547. https://doi.org/10.1073/pnas.1505463112 PMID: 26034287

36. Zabidi MA, Arnold CD, Schernhuber K, Pagani M, Rath M, Frank O, et al. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. Nature. 2014; 518(7540):556–559. https://doi.org/10.1038/nature13994 PMID: 25517091

37. Buxa MK, Slotman JA, van Royen ME, Paul MW, Houtsmuller AB, Renkawitz R. Insulator speckles associated with long-distance chromatin contacts. Biology Open. 2016; 5(9):1266–1274. https://doi.org/10.1242/bio.019455 PMID: 27464669

38. Hart CM, Cuvier O, Laemmli UK. Evidence for an antagonistic relationship between the boundary element-associated factor BEAF and the transcription factor DREF. Chromosoma. 1999; 108(6):375–383. https://doi.org/10.1007/s004120050389 PMID: 10591997

39. Jiang N, Emberly E, Cuvier O, Hart CM. Genome-wide mapping of Boundary Element-Associated Factor (BEAF) binding sites in *Drosophila melanogaster* links BEAF to transcription. Molecular and Cellular Biology. 2009; 29(13):3556–3568. https://doi.org/10.1128/MCB.01748-08 PMID: 19380483

40. Negre N, Brown CD, Shah PK, Kheradpour P, Morrison CA, Henikoff JG, et al. A comprehensive map of insulator elements for the *Drosophila* genome. PLoS Genetics. 2010; 6(1):e1000814+. https://doi.org/10.1371/journal.pgen.1000814 PMID: 20084099

41. Bailey SD, Zhang X, Desai K, Aid M, Corradin O, Cowper-Sal Lari R, et al. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. Nature Communications. 2015; 2:6186. https://doi.org/10.1038/ncomms7186 PMID: 25645053

42. Wani AH, Boettiger AN, Schorderet P, Ergun A, Munger C, Sadreyev RI, et al. Chromatin topology is coupled to Polycomb group protein subnuclear organization. Nature Communications. 2015; 7:10291. https://doi.org/10.1038/ncomms10291

43. Wang F, Marshall CB, Ikura M. Transcriptional/epigenetic regulator CBP/p300 in tumorigenesis: structural and functional versatility in target recognition. Cellular and Molecular Life Sciences. 2013; 70 (21):3989–4008. https://doi.org/10.1007/s00018-012-1254-4 PMID: 23307074

44. Jost D, Carrivain P, Cavalli G, Vaillant C. Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. Nucleic Acids Research. 2014; 42(15):9553–9561. https://doi.org/10.1093/nar/gku698 PMID: 25092923

45. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. Journal of Statistical Software. 2010; 33(1):1–22. https://doi.org/10.18637/jss.v033.i01 PMID: 20808728

46. Wang W, Org T, Montel-Hagen A, Pioli PD, Duan D, Israely E, et al. MEF2C protects bone marrow B-lymphoid progenitors during stress haematopoiesis. Nature Communications. 2016; 7:12376+. https://doi.org/10.1038/ncomms12376 PMID: 27507714

47. Laoukili J, Kooistra MR, Brás A, Kauw J, Kerkhoven RM, Morrison A, et al. FoxM1 is required for execution of the mitotic programme and chromosome stability. Nature Cell Biology. 2005; 7(2):126–136. https://doi.org/10.1038/ncb1217 PMID: 15654331

48. John LB, Ward AC. The Ikaros gene family: Transcriptional regulators of hematopoiesis and immunity. Molecular Immunology. 2011; 48(9–10):1272–1278. http://dx.doi.org/10.1016/j.molimm.2011.03.006. PMID: 21477865

49. DeSandro AM, Nagarajan UM, Boss JM. Associations and interactions between Bare lymphocyte syndrome factors. Molecular and Cellular Biology. 2000; 20(17):6587–6599. https://doi.org/10.1128/MCB.20.17.6587-6599.2000 PMID: 10938133

50. Moqtaderi Z, Wang J, Raha D, White RJ, Snyder M, Weng Z, et al. Genomic binding profiles of functionally distinct RNA polymerase III transcription complexes in human cells. Nature Structural & Molecular Biology. 2010; 17(5):635–640. https://doi.org/10.1038/nsmb.1794 PMID: 20418883

51. Doyle B, Fudenberg G, Imakaev M, Mirny LA. Chromatin loops as allosteric modulators of enhancer-promoter interactions. PLoS Computational Biology. 2014; 10(10):e1003867+. https://doi.org/10.1371/journal.pcbi.1003867 PMID: 25340767

52. Neuwald AF, Aravind L, Spouge JL, Koonin EV. AAA+: A class of chaperone-like ATPases associated with the assembly, operation, and disassembly of protein complexes. Genome Research. 1999; 9 (1):27–43. PMID: 9927482

53. Marsman J, Horsfield JA. Long distance relationships: Enhancer-promoter communication and dynamic gene transcription. Biochimica et Biophysica Acta (BBA)—Gene Regulatory Mechanisms. 2012; 1819 (11–12):1217–1227. http://dx.doi.org/10.1016/j.bbagrm.2012.10.008. PMID: 23124110

54. Cao R, Wang L, Wang H, Xia L, Erdjument-Bromage H, Tempst P, et al. Role of histone H3 lysine 27 methylation in polycomb-group silencing. Science. 2002; 298(5595):1039–1043. https://doi.org/10.1126/science.1076997 PMID: 12351676

55. Djekidel MN, Liang Z, Wang Q, Hu Z, Li G, Chen Y, et al. 3CPET: finding co-factor complexes from ChIA-PET data using a hierarchical Dirichlet process. Genome Biology. 2015; 16(1):288+. https://doi.org/10.1186/s13059-015-0851-6 PMID: 26694485

56. Schwarzer W, Abdennur N, Goloborodko A, Pekowska A, Fudenberg G, Loe-Mie Y, et al. Two independent modes of chromosome organization are revealed by cohesin removal. bioRxiv. 2016;

57. Hnisz D, Weintraub AS, Day DS, Valton AL, Bak RO, Li CH, et al. Activation of proto-oncogenes by disruption of chromosome neighborhoods. Science. 2016; https://doi.org/10.1126/science.aad9024 PMID: 26940867

58. Facchinei F, Scutari G, Sagratella S. Parallel selective algorithms for nonconvex big data optimization. IEEE Transactions on Signal Processing. 2015; 63(7):1874–1889. https://doi.org/10.1109/TSP.2015.2399858

59. Bien J, Taylor J, Tibshirani R. A Lasso for Hierarchical Interactions. Annals of Statistics. 2012;.

60. Wood AM, Van Bortle K, Ramos E, Takenaka N, Rohrbaugh M, Jones BC, et al. Regulation of chromatin organization and inducible gene expression by a *Drosophila* insulator. Molecular Cell. 2011; 44 (1):29–38. https://doi.org/10.1016/j.molcel.2011.07.035 PMID: 21981916

61. 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 PMID: 22955616

62. Andersson R, Gebhard C, Miguel-Escalada I, Hoof I, Bornholdt J, Boyd M, et al. An atlas of active enhancers across human cell types and tissues. Nature. 2014; 507(7493):455–461. https://doi.org/10.1038/nature12787 PMID: 24670763

3.4.3.3 HiCblock: TAD-free analysis of insulators

Standard approaches to identify architectural proteins involved in TAD formation rely on the previous mapping of TADs. Once TADs are mapped, enrichment tests or multiple logistic regression can be further used to characterize which proteins are more likely to influence the presence of borders [Dixon *et al.* 2012, Mourad & Cuvier 2016]. However, an important drawback of the enrichment test and multiple logistic regression is that they rely on accurate TAD mapping, which is problematic for multiple reasons: (i) TAD mapping strongly depends on the algorithm used [Shin *et al.* 2016], (ii) TADs only capture a fraction of the information from Hi-C data, and other important 3D domains, including A/B compartments [Lieberman-Aiden *et al.* 2009], loop domains [Rao *et al.* 2014] and subTADs [Jin *et al.* 2013] were discovered and (iii) TAD borders are blurry [Van Bortle *et al.* 2014].

I proposed a TAD-free model to directly estimate the blocking effects of architectural proteins, insulators and DNA motifs on long-range contacts, making the model intuitive and biologically meaningful (Equation 1 and Figure 1, from the article "TAD-free analysis of architectural proteins and insulators" below) [Mourad & Cuvier 2018]. The model allows analyzing the whole Hi-C information content (2D information) instead of only focusing on TAD borders (1D information). The model outperformed multiple logistic regression at TAD borders in terms of parameter estimation accuracy and was validated by enhancer-blocking assays. In Drosophila, the results supported the insulating role of simple sequence repeats and suggested that the blocking effects depend on the number of repeats. Motif analysis uncovered the roles of the transcriptional factors pannier and tramtrack in blocking long-range contacts. In human, the results suggested that the blocking effects of the well-known architectural proteins CTCF, cohesin and ZNF143 depend on the distance between loci, where each protein may participate at different scales of the 3D chromatin organization.

Nucleic Acids Research, 2017 **1**
doi: 10.1093/nar/gkx1246

TAD-free analysis of architectural proteins and insulators

Raphaël Mourad* **and Olivier Cuvier**

LBME, Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 31062 Toulouse, France

Received July 19, 2017; Revised November 22, 2017; Editorial Decision November 30, 2017; Accepted December 05, 2017

ABSTRACT

The three-dimensional (3D) organization of the genome is intimately related to numerous key biological functions including gene expression and DNA replication regulations. The mechanisms by which molecular drivers functionally organize the 3D genome, such as topologically associating domains (TADs), remain to be explored. Current approaches consist in assessing the enrichments or influences of proteins at TAD borders. Here, we propose a TAD-free model to directly estimate the blocking effects of architectural proteins, insulators and DNA motifs on long-range contacts, making the model intuitive and biologically meaningful. In addition, the model allows analyzing the whole Hi-C information content (2D information) instead of only focusing on TAD borders (1D information). The model outperforms multiple logistic regression at TAD borders in terms of parameter estimation accuracy and is validated by enhancer-blocking assays. In *Drosophila*, the results support the insulating role of simple sequence repeats and suggest that the blocking effects depend on the number of repeats. Motif analysis uncovered the roles of the transcriptional factors pannier and tramtrack in blocking long-range contacts. In human, the results suggest that the blocking effects of the well-known architectural proteins CTCF, cohesin and ZNF143 depend on the distance between loci, where each protein may participate at different scales of the 3D chromatin organization.

INTRODUCTION

In higher eukaryotes, chromosomes are packed in three dimensions and form complex structures (1). Such three-dimensional (3D) structure has recently been investigated by chromosome conformation capture combined with high-throughput sequencing technique (Hi-C) at an unprecedented resolution (2–4). Hi-C experiments reveal multiple levels of genome organization including compartments A/B (5) and topologically associating domains (TADs) (2,3). Most notably, TADs are relatively constant between different cell types and are highly conserved across species. These TADs play important roles in key cell processes such as long-range regulation of genes by enhancers (4) or replication-timing regulation (6).

The identification of architectural proteins and functional elements involved in shaping the genome in 3D represents an intensive field of research (7). Seminal works using enhancer-blocking assays (EBAs) revealed that functional elements called insulators (or boundary elements) can suppress the activation of a promoter by a distant enhancer when interposed (8,9). Multiple evidence actually supports the role of insulator binding proteins (IBPs) such as CTCF, and co-factors like cohesin, as mediators of long-range chromatin contacts (3,10–13), which may in turn result in blocking enhancers from contacting promoters by forming alternative DNA loops. In mammals, high-resolution mapping of long-range contacts has recently revealed that loops occur at domain boundaries and bind CTCF in a convergent orientation where cohesin is recruited (12,14). Depletion of CTCF and cohesin decreased chromatin contacts (13). However, the impact of those depletions was limited suggesting that other proteins might be involved in shaping the chromosome in 3D. Accordingly, other IBPs, co-factors and functional elements were also shown to colocalize at TAD borders (11,15).

A classical approach to identify proteins involved in shaping the 3D genome structure consists in assessing their enrichments at TAD borders (2,3,12). Among a set of enriched proteins, multiple logistic regression (MLR) can be further used to characterize which proteins are more likely to influence the presence of borders (15). However, an important drawback of the enrichment test and MLR is that they rely on accurate TAD mapping, which is problematic for multiple reasons: (i) TAD mapping strongly depends on the algorithm used (16), (ii) TADs only capture a fraction of the information from Hi-C data, and other important 3D domains including A/B compartments (5), loop domains (12) and subTADs (4) were discovered and (iii) TAD borders are blurry (11).

Here, we propose a model named 'blocking model', to systematically analyze the roles of architectural proteins

*To whom correspondence should be addressed. Tel: +33 561 335 956; Fax: +33 561 335 886; Email: raphael.mourad@ibcg.biotoul.fr

and functional elements in blocking long-range contacts between loci. The proposed model does not rely on TAD mapping from Hi-C data. Thus, the model's outcome is not affected by the blurriness of borders. Instead of testing the enrichment/influence of protein binding at TAD borders, the model directly estimates the blocking effect of proteins on long-range contacts between flanking loci, making the model intuitive and biologically meaningful. The model only depends on a simple biological parameter: the distance between insulated loci. The model directly analyzes the Hi-C contact matrix, thus taking advantage of the whole Hi-C information content (2D information) instead of only focusing on TAD borders (1D information). Moreover, the model successfully predicts *in silico* the outcomes from low-throughput enhancer blocking assays, thus enabling genome-wide analyses. Using recent *Drosophila* and human Hi-C data at high resolution, combined with a large number of ChIP-seq and DNA motif data, we revealed numerous combinations of proteins, functional elements and DNA motifs that block long-range contacts depending on scale and synergistic/antagonistic effects.

MATERIALS AND METHODS

Hi-C data

For *Drosophila* data analysis, we used publicly available high-throughput chromatin conformation capture (Hi-C) data of embryonic Kc167 cells from Gene Expression Omnibus (GEO) accession GSE62904 (17). We also used Kc167 Hi-C data from GEO accession GSE89112 (18). Hi-C data were binned at 1, 2 and 5 kb resolutions.

For human data analysis, we used publicly available Hi-C data of lymphoblastoid GM12878 cells from GEO accession GSE63525 (12). We used Hi-C data binned at 10, 40 and 100 kb resolution.

ChIP-seq data

For *Drosophila* data analysis, we used publicly available protein-binding profiles of Kc167 cells (except for Pnr whose data were from 6–8 h embryos). ChIP-seq data for CP190, Su(Hw), dCTCF and BEAF-32 were obtained from GEO accession GSE30740 (19). ChIP-seq data for Barren (condensin I), Cap-H2 (condensin II), Chromator, Rad21 (cohesin), GAF and dTFIIIC were obtained from GEO accession GSE54529 (11). ChIP-seq data for Fs(1)h-L were obtained from GEO accession GSE42086 (20). ChIP-seq data for Ttk69k were obtained from GEO accession GSE34698 (21). ChIP-seq peak calling was done using MACS 2.1.0 with default parameters for all proteins (https://github.com/taoliu/MACS). ChIP-chip peaks for Pnr were directly downloaded from (22).

For human data analysis, we used publicly available binding peaks of 73 chromatin proteins (Rad21, CTCF, YY1, ZBTB33, MAZ, JUND, ZNF143, EZH2, ATF2, ATF3, BATF, BCL11A, BCL3, BCLAF1, BHLHE40, BRCA1, CEBPB, CFOS, CHD1, CHD2, CMYC, COREST, E2F4, EBF1, EGR1, ELF1, ELK1, FOXM1, GABP, IKZF1, IRF4, MAX, MEF2C, MTA3, MXI1, NFATC1, NFE2, NFIC, NFKB, NFYA, NFYB, NRF1, NRSF, P300, PAX5, PBX3, PML, POL2, POL3, POU2F2, RFX5, RUNX3, RXRA, SIN3A, SIX5, SMC3, SP1, SPI1, SRF, STAT1, STAT3, STAT5, TBLR1, TBP, TCF12, TCF3, TR4, USF1, USF2, WHIP, ZEB1, ZNF274 and ZZZ3) of GM12878 cells from ENCODE (23). We downloaded peaks that were uniformly processed (Uniform Peaks).

DNA motifs

To scan the genome for motif occurrences, we used Find Individual Motif Occurrences (FIMO) with default parameters and with position-specific priors (PSPs) to improve the identification of true motif occurrences (24). GM12878 DNase data from ENCODE were used as PSPs (23). The motif information was taken either from the litterature (using consensus motif) or from JASPAR database (http://jaspar.genereg.net/).

For *Drosophila* data analysis, we used transcription factor-binding site (TFBS) motifs from the JASPAR database. For some proteins, we used instead motif consensuses from the litterature: BEAF-32 (CGATA) (25), dCTCF (AGGTGGCG) (26), Su(Hw) (TGCATATTT) (27), GAF (GAGAGA) (28), ZW5 (GCTGMG) (29), DREF (TATCGATA) (30), M1BP (GGTCACACT) (31), Ttk69k (GGTCCTGC) (32), dTFIIIC A box (TGGN NNAGNNG), Pita (GGTTNNNNNNNNGCT) (29), ZIPIC (AGGGNTG) (29), Ibf (ATGTANAA) (33), Elba (CCAATAAG) (34) and Zelda (CAGGTAG) (35).

For human data analysis, we also used TFBS motifs from the JASPAR database. In human, motifs with <2000 occurrences were removed from the analysis to reduce uncertainty in the β estimation.

The blocking model

To illustrate the blocking model, we first plotted the example of a *Drosophila* genomic region with embryonic Kc167 cell Hi-C heatmap and ChIP-seq peaks of well-known architectural proteins (Figure 1A). We observed that all architectural proteins BEAF-32, dCTCF, dTFIIIC, GAF and Su(Hw) accumulated on a specific locus (green frame) that acted as an insulator of long-range contacts between flanking regions. This observation suggested that the binding of those proteins blocked long-range contacts (Figure 1B), thereby contributing to the formation of 3D domains.

By integrating Hi-C data with ChIP-seq data or DNA motif data, we propose to model the blocking effects of protein bindings with a generalized linear model:

$$\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{d},\mathbf{B},\mathbf{I}\right]\right) = \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} - \boldsymbol{\beta}_I\mathbf{I} \qquad (1)$$

where, variable **y** denotes Hi-C count for any pair of bins on the same chromosome. The log-distance variable **d** accounts for the background polymer effect (power law decay relation between distance and Hi-C count modeled by a log–log linear relation) (36). Bias variables $\mathbf{B} = \{\mathbf{len}, \mathbf{GC}, \mathbf{map}\}$ are known Hi-C biases including fragment length (**len**), GC-content (**GC**) and mappability (**map**) that are computed as in (37). Including those bias variables into the model allows correcting for biases in Hi-C data. Note that bias variables do not need to be included in the model if Hi-C counts were previously normalized by matrix balancing (38). Variable set $\mathbf{I} = \{\mathbf{i}_1, ..., \mathbf{i}_p\}$ represents the p blocking variables

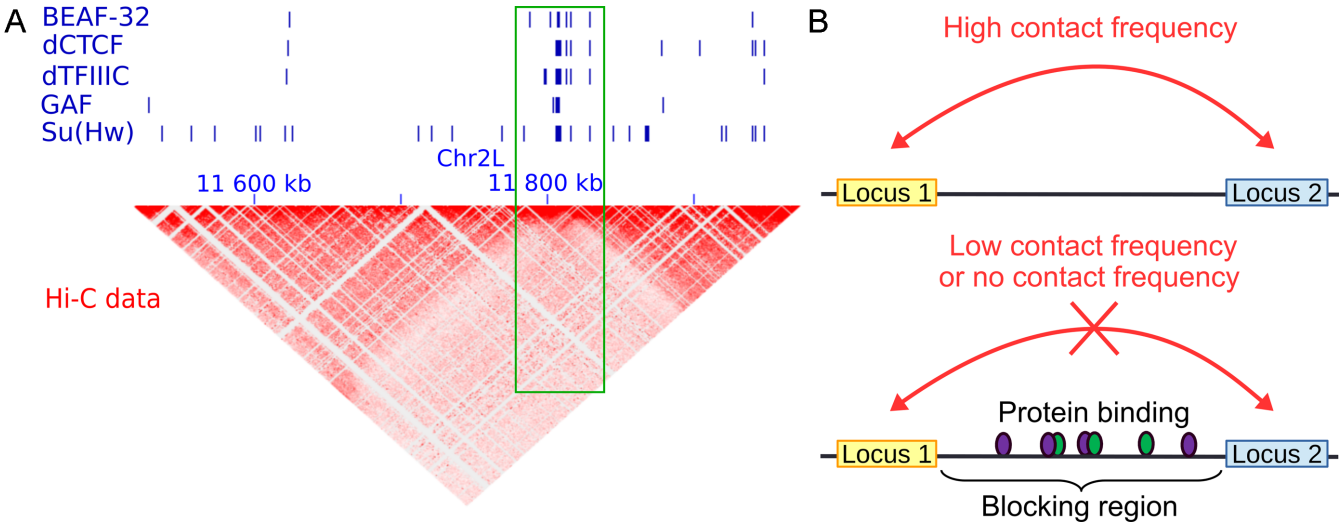


Figure 1. Illustration of the blocking model. (**A**) Example showing that the accumulation of insulator-binding proteins (IBPs) is associated with a blocking effect of long-range contacts between flanking loci in *Drosophila* (see green frame). (**B**) Schema representing the blocking effect of protein binding on long-range contacts between two loci, such as between an enhancer and a promoter.

of interest. A blocking variable stores a value corresponding to a 'blocking region' (Figure 1B), which is the region in-between two bins whose Hi-C contacts are measured. For ChIP-seq data, a blocking variable is defined as the average of the base coverage computed from the $\log_2$ fold-enrichments of peaks found into the blocking region divided by the length of the blocking region. A base within a peak has a coverage value equal to the $\log_2$ fold-enrichment of the peak and a base outside a peak has a coverage value equal to zero. For DNA motif data, a blocking variable is defined as the number of motif occurrences found into the blocking region divided by the length of the blocking region. The corresponding β_i parameter value reflects the blocking effect of the protein on Hi-C counts. A positive value ($\beta_i > 0$) reveals a blocking effect on long-range contacts. Conversely, a negative value ($\beta_i < 0$) shows a facilitating effect on contacts. A null value ($\beta_i = 0$) means that the protein does not have any effect in blocking or facilitating contacts.

Using the model, one can also assess the co-blocking effects of two or more proteins using statistical interaction terms:

$$\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{d},\mathbf{B},\mathbf{i}_1,\mathbf{i}_2\right]\right) = \beta_0 + \beta_d\mathbf{d} + \boldsymbol{\beta}_B\mathbf{B} - \beta_{i_1}\mathbf{i}_1 - \beta_{i_2}\mathbf{i}_2 - \beta_{i_{12}}\mathbf{i}_1\mathbf{i}_2 \qquad (2)$$

where, variables $\mathbf{i}_1$ and $\mathbf{i}_2$ are two blocking variables. The product $\mathbf{i}_1\mathbf{i}_2$ is a second-order statistical interaction. The corresponding parameter $\beta_{i_{12}}$ reflects the co-blocking effect of the two proteins on contacts. A positive value ($\beta_{i_{12}} > 0$) reveals a synergistic effect of the two proteins in blocking contacts. Conversely, a negative value ($\beta_{i_{12}} < 0$) shows an antagonistic effect of the two proteins in blocking contacts. In equation (2), a second-order interaction was included, but higher-order interactions (products of more than two variables) can be included to model co-blocking effects of more than two proteins.

The model only depends on a single parameter: the distance range between insulated loci. This parameter has a strong biological meaning since it reflects the analysis scale of hierarchical 3D genome organization. For instance, in *Drosophila*, we will focus on Hi-C data for 20–50 kb distances which are below the median size of TADs (median size of 60 kb (3)), therefore allowing TAD-scale analyses. But we will also vary the scale of analysis in human (see below).

In some situations, we standardize the blocking variables before computing the model. Standardization allows to reduce the effect of very large differences in the blocking variables between different proteins when estimating the βs and makes the latter more comparable in magnitude. In fact, these blocking variable differences might be due to very large differences in the ChIP-seq signal and the number of peaks that might not be linked to the real blocking activity of proteins. For instance, when analyzing human ChIP-seq data, we found that the highest βs were often associated to proteins with few binding sites when no standardization was used, and that these βs were strongly reduced after standardization (see below).

Because of Hi-C count overdispersion, we use negative binomial regression as the most appropriate specification of the generalized linear model. However, Poisson regression with lasso shrinkage can also be used. We believe that the choice between both depends mainly on the number of variables to analyze. On the one hand, if there are a few candidate variables (<10), it is interesting to estimate β parameters together with corresponding *P*-values to assess significance using negative binomial regression. On the other hand, if there are a large number of variables (10 or more), it is more convenient to use Poisson lasso regression in order to select the key variables and to account for correlations among the variables (frequent in ChIP-seq and motif occurrence data).

The model is available in the R package 'HiCblock' which can be downloaded from the Comprehensive R Archive Network (https://cran.r-project.org/web/packages/HiCblock/index.html). For the negative binomial regression, model βs are learned by iterative weighted least squares (glm.nb function from MASS R package with default parameters). For the Poisson lasso regression, model βs are learned by cyclical coordinate descent and lambda parameter is estimated with 10-fold cross-validation (cv.glmnet function from glmnet R package with default parameters).

Simulation of random protein-binding sites and motif occurrences

For Poisson lasso regression in human, we simulated protein binding sites by randomly drawing genomic regions from the genome whose numbers and fold-enrichments were similar to those observed from real proteins. We then used these random proteins to compute associated β coefficients with the Poisson lasso regression. We expected these βs to be close to zero but with a certain standard deviation $\hat{\sigma}$. We then used this standard deviation to compute a confidence interval as $0 \pm 1.96 \times \hat{\sigma}$ under the null hypothesis that a random protein did not have any blocking or facilitating effect on long-range contacts. For DNA motifs, we used a slightly different approach. We randomly draw 14 base DNA sequences (random motifs) whose number of occurrences over the genome were similar to those of real DNA motifs. We scanned the genome for random motif occurrences. Then, we used these random motif occurrences to compute associated β coefficients with the Poisson lasso regression. As for random proteins, we used these βs to compute a confidence interval under the null hypothesis.

RESULTS

Model validation with enhancer-blocking assays

We first sought to validate our model using EBAs from *Drosophila*. EBA is a classical low-throughput method that can be used to show the ability of an insulator sequence to block the activation of a promoter by a distant enhancer when interposed between them (39) (Figure 2A). We used the model to predict the blocking effect of an insulator region depending on protein binding. For this purpose, we used a compilation of EBA results from (11). It consisted of 32 regions with varying reported insulating activity (15 regions with insulating activity and 17 regions with no insulating activity). In the first benchmark, we selected the 15 regions with insulating activity (positive class). In order to have a large set of regions with no insulating activity, we generated >100 control regions (negative class) by randomly drawing from the *Drosophila* genome with sizes, GC and repeat contents similar to those of the abovementioned 15 regions (40). For each region, we computed blocking variables $\mathbf{I} = \{\mathbf{i}_1, ..., \mathbf{i}_p\}$ using p ChIP-seq data from Kc167 cells. We also used $\hat{\beta}_I = \{\hat{\beta}_{i_1}, ..., \hat{\beta}_{i_p}\}$ model parameters independently learned from Kc167 Hi-C data from Li *et al.* (17) at 2 kb resolution and for 20–50 kb distances, for which Hi-C coverage was high. Model parameters were not learned from EBA assays to prevent overestimation of predictive performance. We predicted insulating activities of the regions by the matrix product $\hat{\beta}_I\mathbf{I}$. We then assessed the accuracy of our model's predictions using receiver operating characteristic (ROC) curve and the area under the ROC curve (AUC). We found that predicted insulating activity was very close to the observed insulator activity from EBA (AUC = 0.981; Figure 2b). In the second benchmark, we did not use generated controls but instead the 17 regions reported to have no insulating activity as negative class. We again predicted insulating activity, and found that predictions were still good (AUC = 0.808; Figure 2C). We found that changing Hi-C data resolution to 1 or 5 kb only slightly affected predictions for the two benchmarks (Supplementary Figure S1). In the third benchmark, we assessed the blocking effect of simple sequence repeats (SSRs) of GATA that were shown to have an insulating activity by EBAs in both *drosophila* and human (41). In *drosophila*, we estimated a blocking effect for SSRs that comprised >4 repeats (Figure 2D and Supplementary Table S1). In particular, we

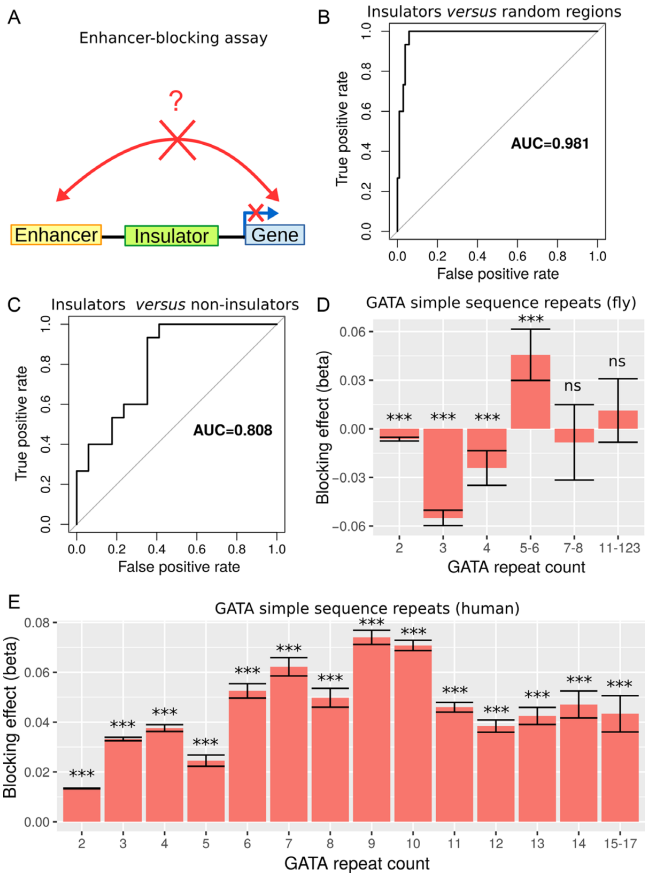


Figure 2. Validation of the model with enhancer-blocking assays (EBAs) from *Drosophila* and human. (**A**) Illustration of the EBAs. (**B**) ROC curves of the prediction of insulating regions (positives) as compared to randomly drawn regions (negatives) in *Drosophila*. Area under the ROC curve (AUC) is plotted. (**C**) ROC curves of the prediction of insulating regions (positives) as compared to non-insulating regions (negatives) in *Drosophila*. (**D**) Blocking effects of GATA SSRs depending on the repeat count in *Drosophila*. (**E**) Blocking effects of GATA SSRs depending on the repeat count in human.

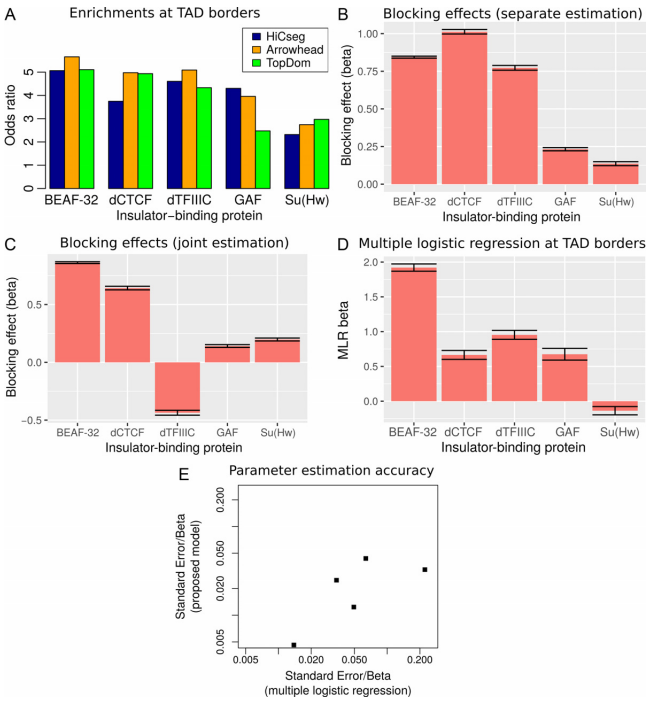


Figure 3. Analysis of IBPs in *Drosophila*. (**A**) Enrichment of IBPs at TAD borders, depending on the TAD mapping algorithm used. (**B**) Blocking effect (β) estimated separately. (**C**) Blocking effect (β) estimated jointly. (**D**) MLR βs estimated from TAD borders (15). (**E**) Parameter estimation accuracy of the proposed model compared to MLR.

found a significant blocking effect for SSRs with five to six repeats ($\hat{\beta} = 0.046$, $P = 2 \times 10^{-8}$). SSRs with >6 repeats were too few to detect any significant blocking effect (only 8 SSRs with 7 to 8 repeats and 9 SSRs with >11 repeats). In human, we detected significant blocking effects for all GATA repeat counts ($P < 10^{-20}$) at short distances (100–250 kb at 10 kb resolution; Figure 2E and Supplementary Table S2). Most notably, we found the highest blocking effects for SSRs with 9 to 10 repeats ($\hat{\beta} > 0.07$, $P < 10^{-20}$), revealing that the blocking effect depends on the number of repeats. For larger distances (950–1000 kb), we could only detect a slight blocking effect for eight repeats, suggesting that SSR blocking effect acted at short distance (Supplementary Figure S2 and Table 3). Using EBAs, we thus concluded that the model was successfully validated.

Analysis of insulator proteins and comparison with current approaches

A major problem of testing protein enrichment at TAD borders is that different algorithms have been developed for TAD mapping which can yield large differences of enrichments for the same protein (42). Accordingly, we observed that the enrichments of BEAF-32, dCTCF, dTFIIIC, GAF and Su(Hw) could greatly vary depending on the TAD algorithm used in *Drosophila* (Figure 3A). For instance, GAF presented an odds ratio (OR) of 4.3 with HiCseg (43), an OR of 4 with Arrowhead (12), whereas it only showed an OR of 2.5 with TopDom TADs (16). Conversely, dCTCF presented an OR of 3.7 with HiCseg, and ORs around 5 with Arrowhead and TopDom.

Instead of testing protein enrichments at TAD borders, we used our model to directly assess the blocking effect of protein binding on long-range contacts. We first estimated separately the blocking effects of IBPs, by including only one IBP in the model at a time. This allowed to compare with previous enrichments. We used Kc167 Hi-C data from Li *et al.* (17) at 2 kb resolution and focused on 20–50 kb distances. Using our model, we found that BEAF-32, dCTCF and dTFIIIC showed the strongest blocking effects (Figure 3B), which was similar to the enrichments observed at TAD borders (Figure 3A) and previously observed by Sexton *et al.* (3). Because the blocking effect might be influenced by the number of protein-binding sites, we sampled different numbers of peaks from BEAF-32 and estimated the corresponding βs. As expected, we found that β accuracy was lower for smaller number of peaks (Supplementary Figure S3). We also observed that the blocking effect was inflated, but such inflation remained reasonable (+63%), even for 1000 sampled peaks which represented only 15% of all BEAF-32 peaks.

Because IBPs often colocalize linearly (e.g. correlate) on the chromosome, one might estimate a blocking effect for a protein, although the protein does not directly impede long-range contacts (15). Hence, we re-estimated blocking effects of IBPs jointly (e.g. by including all IBPs within the same model). BEAF-32 presented the highest blocking effect ($\hat{\beta} = 0.86$, $P < 10^{-20}$) compared to the other proteins (Figure 3C), similarly to previously published MLR analysis at TAD borders (15) (Figure 3D). Our model also estimated a negative β for dTFIIIC, suggesting that the protein could in fact facilitate long-range contacts between flanking regions, contrary to what is found by the separate estimation (previous paragraph). This meant that dTFIIIC blocking effect estimated by separate estimation was in fact due to the colocalization (correlation) of dTFIIIC with other IBPs such as BEAF-32 (correlation between dTFIIIC and BEAF-32 blocking variables equals 0.59, $P < 10^{-20}$). Our model outperformed MLR in terms of parameter estimation accuracy. Standard errors of beta parameters were dramatically lower than the ones from MLR, revealing the higher performance of our model in assessing blocking effects of proteins (Figure 3E). To further compare our new model with MLR, we assessed the ability to discriminate between known architectural proteins (11 true positives including IBPs and co-factors) and random protein peaks (200 false positives) using ROC curves (Supplementary Figure S4). Based on the absolute values of βs, we found that our blocking model was highly accurate (AUC = 0.991) and performed better than MLR (AUC = 0.827). Moreover, we performed the joint analysis of IBPs for different binning resolutions (1 and 5 kb) and found similar results with 2 kb, revealing that the resolution did not have a big impact on the estimation of blocking effects (Supplementary Figure S5). In addition, we analyzed recent Hi-C data with higher coverage from Eagen *et al.* (18) at 1 kb resolution and obtained results that were close to those obtained from Li *et al.* data (Supplementary Figure S6). Thus, by processing the whole Hi-C matrix information, instead of focusing only on

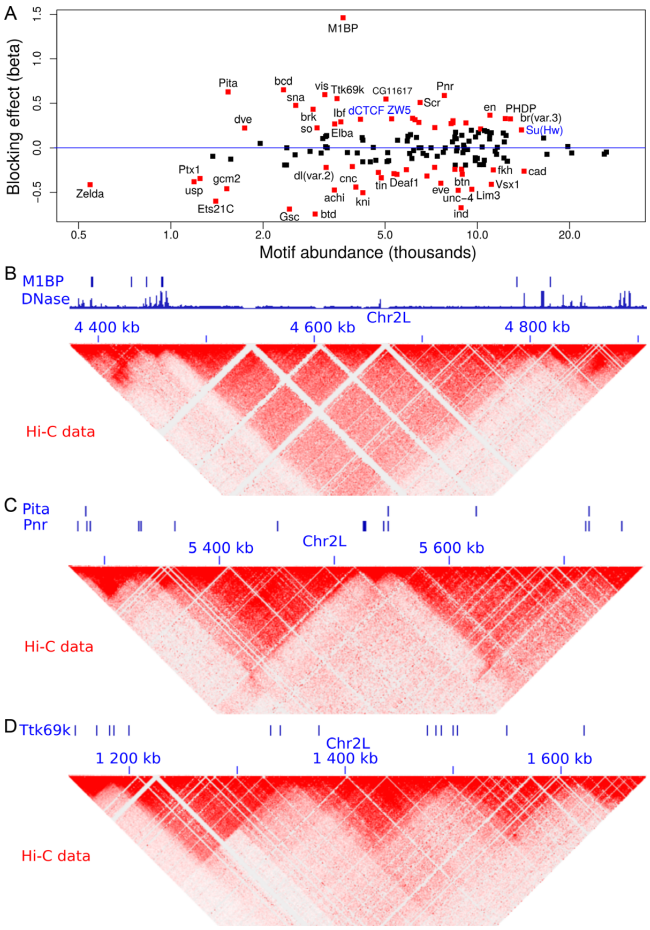


Figure 4. Analysis of protein binding DNA motifs in *Drosophila*. (**A**) Blocking effect (β) in function of motif abundance ($|\hat{\beta}| > 0.2$ are shown in red; known architectural proteins are written in blue). (**B**) Example showing the accumulation of M1BP motifs and DNase I hypersensitive sites between 3D domains. (**C**) Example showing the accumulation of Pita and Pnr motifs between 3D domains. (**D**) Example showing the accumulation of Ttk69k motifs between 3D domains.

TAD borders, the proposed model was more accurate than MLR.

Numerous protein-binding DNA motifs act as blockers

We next sought to analyze the blocking effects of protein-binding DNA-motifs (Figure 4A and Supplementary Table S4). Interestingly, our model found motif 1-binding protein (M1BP) as the motif with the strongest blocking effect ($\hat{\beta} = 1.46$), which was recently found to be enriched at TAD borders during development (35) and was implicated in transcriptional pausing of genes (31). Such transcriptional pausing was recently shown to be involved in long-range contacts (44). When we looked at Hi-C heatmaps, we observed that M1BP motifs accumulated at the borders of 3D domains (Figure 4B; DNase I hypersensitivity is shown to represent the potential activity of the motifs). We also identified other motifs with strong blocking effects including bcd ($\hat{\beta} = 0.65$), Pita ($\hat{\beta} = 0.63$), vis ($\hat{\beta} = 0.60$), Pnr ($\hat{\beta} = 0.59$) and Ttk69k ($\hat{\beta} = 0.55$). Among those proteins, Pita was a recently discovered insulator protein able to target CP190 to chromatin (45) and was found at 3D domain borders (Figure 4C). When we used Ttk69k ChIP-seq and Pnr ChIP-chip data, we found that both Ttk69k and Pnr colocalized at or near architectural protein peaks (Supplementary Figure S7a). For instance, Pnr was enriched at condensin I (Barren), CP190, BEAF-32 and Chromator peaks (Supplementary Figure S7b). Interestingly, Ttk69k was mostly enriched near architectural proteins but did not overlap them, except for condensin I, suggesting that Ttk69k might participate to the formation of 3D domains in a very specific way (Supplementary Figure S7c). Accordingly, we found numerous Pnr and Ttk69k motifs located between 3D domains (Figure 4C and D). We also identified architectural proteins ZW5 ($\hat{\beta} = 0.33$), dCTCF ($\hat{\beta} = 0.32$) and Ibf ($\hat{\beta} = 0.29$). Of note, Ibf was shown to be a novel CP190 interacting protein with insulating activity (33). When we compared with MLR, we also found that M1BP presented a very high positive influence on TAD borders ($\hat{\beta} = 8.65$; Supplementary Table S5). However another motif, Zelda, presented the highest positive influence ($\hat{\beta} = 9.32$), whereas the same motif was identified as a long-range contact facilitator with the blocking model ($\hat{\beta} = -0.41$; Supplementary Table S4). This suggests that the blocking model can capture effects on long-range contacts that could not be assessed by the analysis at the TAD border level. Using the blocking model, we could conclude that many proteins including pannier, a transcriptional regulator involved in several developmental processes (46) and tramtrack 69k, a widely expressed transcriptional factor (TF) related to cell fate specification, cell proliferation and cell-cycle regulation (47), might represent novel candidate architectural proteins in *Drosophila*.

Co-blocking effects of insulator-binding proteins and co-factors

Long-range contacts not only involve IBPs but also co-factors that regulate or stabilize them (11,12,48). Hence, we sought to analyze potential effects of IBPs and co-factors in co-blocking long-range contacts. We first modeled the co-blocking effects of protein pairs using second-order statistical interactions (for every protein pair, we estimated a co-blocking effect). We detected 38/55 significant interactions after Bonferroni correction. Among the significant interactions, the model identified 19 positive co-blocking effects ($\hat{\beta} > 0$), reflecting protein pairs that synergistically blocked long-range contacts (Supplementary Table S6). We represented these synergistic blocking effects by a network of proteins (Figure 5A). In agreement with (49), CP190 co-blocked contacts with BEAF-32 ($\hat{\beta} = 0.76$, $P < 10^{-20}$) and with GAF ($\hat{\beta} = 0.67$, $P < 10^{-20}$). Interestingly, we found that Condensin II (Cap-H2) played a central role in helping other proteins to block contacts, including dCTCF ($\hat{\beta} = 1.33$, $P = 4 \times 10^{-13}$), Barren ($\hat{\beta} = 0.78$, $P < 10^{-20}$), dTFIIIC ($\hat{\beta} = 0.70$, $P = 10^{-6}$) and GAF ($\hat{\beta} = 0.68$, $P = 2 \times 10^{-10}$). dTFIIIC also represented an important protein for co-blocking effects. Conversely, Fs(1)h-L had only one co-blocking partner, dTFIIIC. The model also estimated 19 negative co-blocking effects ($\hat{\beta} < 0$), reflecting protein

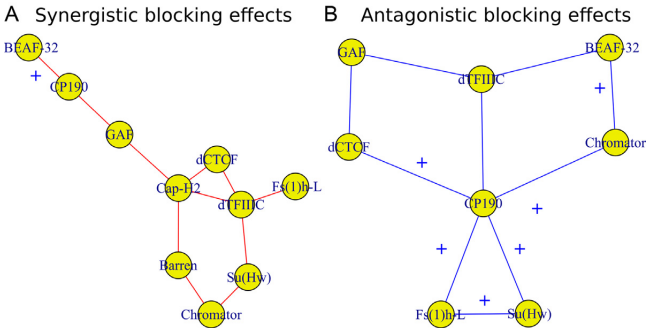


Figure 5. Effects of IBPs and co-factors in co-blocking long-range contacts. (**A**) Synergistic blocking effects estimated by positive second-order interaction βs. An edge between two protein nodes i and j means $\hat{\beta}_{ij} > 0.5$. (**B**) Antagonistic blocking effects estimated by negative second-order interaction βs. An edge between two protein i and j nodes means $\hat{\beta}_{ij} < 0.5$. Blue cross: physical interaction reported in Flybase.

pairs that had antagonistic effects in blocking long-range contacts (Figure 5B and Supplementary Table S6). Most notably, we found numerous antagonistic effects of CP190 in blocking contacts with other proteins, such as dTFIIIC ($\hat{\beta} = -2.33$, $P < 10^{-20}$), Su(Hw) ($\hat{\beta} = -1.78$, $P < 10^{-20}$), Chromator ($\hat{\beta} = -1.68$, $P < 10^{-20}$), dCTCF ($\hat{\beta} = -0.87$, $P < 10^{-20}$) and Fs(1)h-L ($\hat{\beta} = -0.53$, $P = 4 \times 10^{-6}$). Interestingly, Su(Hw) had a slight blocking effect on long-range contacts ($\hat{\beta} = 0.20$, $P < 10^{-20}$; Figure 3C), but when combined with CP190, they presented a strong antagonistic effect which reduced its blocking effect ($\hat{\beta} = -1.78$, $P < 10^{-20}$; Figure 5B). Among the synergistic and antagonistic effects, we found that many corresponded to physical interactions reported in Flybase and previous studies (49), supporting the idea that physical interactions may account for some of them. Analysis of second-order interactions thus revealed the complexity behind the establishment of 3D domains. This may notably depend on numerous synergistic and antagonistic effects of IBPs with key architectural co-factors such as structural maintenance complex (SMC) family of proteins including cohesin and condensin (50,51).

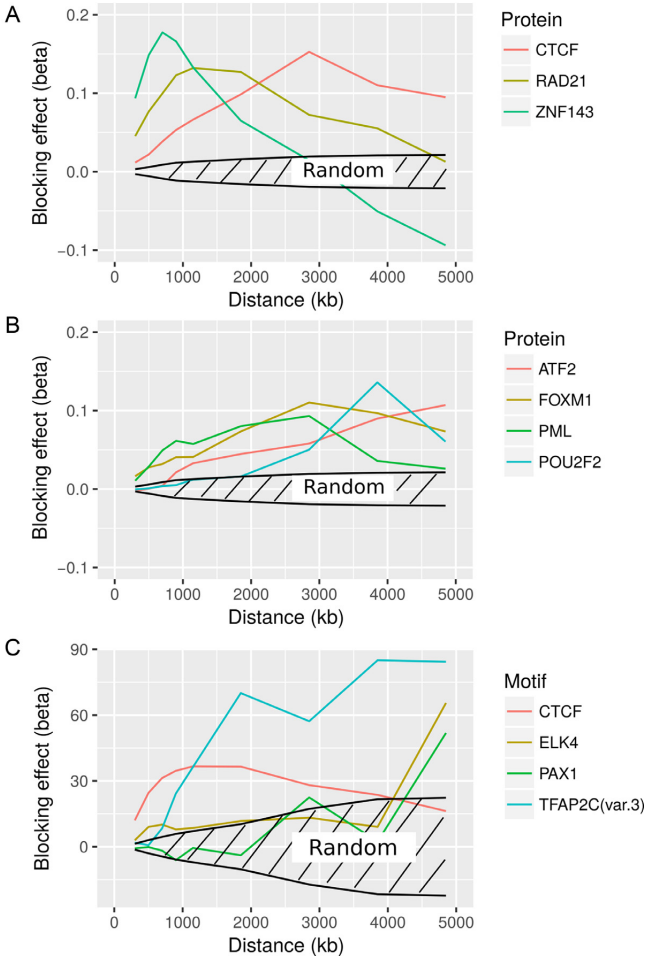


Figure 6. Analysis of protein binding and DNA motif in human. (**A**) Blocking effects of architectural proteins depending on the distance between loci. (**B**) Blocking effects of TFs depending on the distance between loci. (**C**) Blocking effects of protein binding motifs depending on the distance between loci. For all three subfigures, we also plotted confidence intervals under the null hypothesis that a random protein or DNA motif did not have any effect on long-range contacts.

Analysis in human

We then analyzed blocking effects of proteins and DNA motifs in human, depending on the scale of 3D genome organization. For this purpose, we used GM12878 Hi-C data for varying distance ranges: [200–400 kb], [400–600 kb], [600–800 kb], [800–1000 kb], [1000–1300 kb], [1700–2000 kb], [2700–3000 kb], [2700–3000 kb], [3700–4000 kb] and [4700–5000 kb]. We performed analyses at 40 kb resolution to have sufficient coverage at long distance (even though for short distance higher resolution could be used). By varying the distance range, we could assess blocking effects at different scales, thus allowing the analysis of the well-known hierarchical nature of 3D domains (52). Because of the large number of variables (>50), we used Poisson lasso regression. Moreover, for ChIP-seq data analysis, we scaled the blocking variables because the ChIP-seq peak numbers and fold-enrichments greatly varied between proteins and that prevented further comparison of βs. For each analysis, we also computed confidence intervals under the null hypothesis that a protein or DNA motif did not have any blocking or facilitating effect on long-range contacts (see 'Materials and Methods' section, simulation of random protein-binding sites and motif occurrences).

We first focused on known architectural proteins CTCF, Rad21 (cohesin subunit) and ZNF143. Remarkably, we observed that the blocking effects of architectural proteins strongly depended on the distance between loci (Figure 6A and Supplementary Table S7), a question that could not be addressed by previous enrichment or MLR analyses at TAD borders. For instance, CTCF blocking effects peaked around 3 Mb. Interestingly, the main looping partner of CTCF, cohesin, had a blocking effect that peaked at a lower distance, from 1000 to 2000 kb. Another partner of CTCF, ZNF143, also showed a different blocking effect that strikingly peaked at 800–900 kb. This means that although CTCF, cohesin and ZNF143 were known to act to-

gether in establishing chromatin loops (7), they might participate at different scales. We next studied the blocking effects of TFs (Figure 6B and Supplementary Table S7). Compared to architectural proteins, TFs were less abundant over the genome (around few thousands peaks, compared to tens of thousands of peaks for architectural proteins). Among the strongest blockers, we found ATF2, FOXM1, PML and POU2F2, whose effects also depended on distance. POU2F2 effect peaked at 3800 kb, and FOXM1 and PML both peaked at 3 Mb. Interestingly, some TFs, such as ATF2, presented high blocking effects for very large distance (>5 Mb). Thus, although TFs were less frequent over the genome than architectural proteins, they might collectively contribute significantly to the establishment or maintenance of 3D organization. Lastly, we analyzed protein-binding DNA motifs (Figure 6C and Supplementary Table S8). CTCF motif showed a strong blocking effect that peaked from 1000 to 2000 kb, at a shorter distance than found using ChIP-seq data. However, another motif, TFAP2C, presented the strongest blocking effect, especially at long distance. TFAP2C has been implicated in breast cancer oncogenesis, and was previously shown to be a collaborative factor in estrogen-mediated long-range interaction and transcription (53). We also identified ELK4 and PAX1 as strong blockers at long distance. ELK4 is a member of the Ets family of transcription factors, and PAX1, is essential during fetal development. We thus concluded that architectural proteins, but also transcription factors, shaped the 3D human genome at different genomic scales.

DISCUSSION

In this paper, we propose a model to comprehensively study the roles of architectural proteins, insulators and DNA motifs in blocking long-range contacts between flanking loci at different scales, thereby demarcating the genome into functional 3D domains. The proposed approach is TAD-free: it does not rely on any TAD mapping algorithm, it does not focus on TADs but instead on all possible 3D domains at all scales, and it is not affected by the blurriness of TAD borders. The model is validated by numerous EBAs. It outperformed previous MLR of TAD borders (15) in terms of blocking effect estimation accuracy. The model is flexible and can identify both synergistic and antagonistic effects of architectural proteins depending on the presence of specific IBPs and co-factors.

The proposed model also uncovers a number of results. In *Drosophila*, we find that the blocking effect for the GATA SSRs depends of the number of repeats, and in particular, we estimate a significant blocking effect for 5–6 repeats. In human, we find that GATA repeat effect peaks for 9–10 repeats. Moreover, analysis of motifs identifies pannier and tram track as two novel candidate architectural proteins. Interestingly, the protein pannier is a member of the GATA family known to bind to GATA motifs (46), which may explain the insulating activity of GATA repeats by recruiting multiple pannier proteins contiguously to DNA. Moreover, tram track has a homomeric dimerization BTB/POZ domain that could help bridging two distant proteins through long-range contacts (54) and that is known to interact with GAF (55). Analysis of co-blocking effects between architectural proteins further suggests a role for co-factor condensin II in helping other proteins to block contacts. Conversely, CP190 presents numerous antagonistic effects with other proteins, meaning that it reduces their blocking activities. Such co-blocking analyses thus reveal the modulating effects of specific proteins in blocking contacts with other proteins. In human, analyses for varying distance ranges uncover strong distance-dependent blocking effects depending on the protein or DNA motif, that could not be addressed by enrichment test or MLR at TAD borders. For instance, we find that CTCF, cohesin and ZNF143 blocking effects peak at different distances, although the three proteins are known to act together in establishing chromatin loops (7). This suggests that they may participate at different 3D chromatin scales, or alternatively that their mechanisms of action is not always associated with their binding. Supporting this idea, recent results showed that cohesin is recruited at transcription start sites and positioned to CTCF sites by transcription-mediated translocation (56). In addition, we observed changes of the β sign depending on the distance. For instance, ZNF143 presented a blocking effect at short distance (<2500 kb) and a facilitating effect at longer distance. This can be due to ZNF143-mediated loops at short distance that have allosteric effects on long distance interactions (57).

There are different reasons why we restricted our analysis within a limited distance range, e.g. 20–50 kb in *Drosophila* (and not 20–1000 kb, for instance). First, at the high resolution of 2 kb, most of the Hi-C signal is observed within short distance (20–50 kb). Second, our model assumes a power law decay between Hi-C count and distance (equivalent to a log–log linear relation between Hi-C count and distance) which only holds for a limited distance range. Third, not restricting the analysis to a limited distance range can lead to heavy computational burden. One simple way to analyze Hi-C data within a wider distance range would be to analyze data at 10–20 kb resolutions.

There are several limitations of the proposed approach. First, model learning can be computationally demanding in time and memory depending on the distance range or Hi-C data resolution. New big data learning algorithms could be used to process the data at a higher resolution that would allow in-depth analysis of 3D chromatin drivers (58). Second, the model makes the assumption that the accumulation of protein binding blocks long-range contacts, but other scenarios could explain the formation of borders. For instance, attraction/repulsion forces between histone marks can predict the folding of chromatin (59). Third, in human, we observed large changes of βs over distance, for instance for protein ZNF143 and DNA motif TFAP2C(var.3). Because lasso regression is not designed to estimate beta standard deviations, the significance of the difference between two βs obtained for two different distances cannot be tested. Instead, one could use a standard regression with selected variables to assess the significance.

AVAILABILITY

The model is available in the R package 'HiCblock' which can be downloaded from the Comprehensive R

Archive Network (https://cran.r-project.org/web/packages/HiCblock/index.html).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

The authors are grateful to Corces lab (Emory University, USA) for data.

FUNDING

University of Toulouse; CNRS. Funding for open access charge: Fondation pour la Recherche Médicale (DEQ20160334940) to our team (R.M. and O.C.).
Conflict of interest statement. None declared.

REFERENCES

1. Halverson,J.D., Smrek,J., Kremer,K. and Grosberg,A.Y. (2014) From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Rep. Prog. Phys.*, **77**, 022601.
2. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
3. Sexton,T., Yaffe,E., Kenigsberg,E., Bantignies,F., Leblanc,B., Hoichman,M., Parrinello,H., Tanay,A. and Cavalli,G. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.
4. Jin,F., Li,Y., Dixon,J.R., Selvaraj,S., Ye,Z., Lee,A.Y., Yen,C.A., Schmitt,A.D., Espinoza,C.A. and Ren,B. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.
5. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O. *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
6. Pope,B.D., Ryba,T., Dileep,V., Yue,F., Wu,W., Denas,O., Vera,D.L., Wang,Y., Hansen,R.S., Canfield,T.K. *et al.* (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.
7. Cubenas-Potts,C. and Corces,V.G. (2015) Architectural proteins, transcription, and the three-dimensional organization of the genome. *FEBS Lett.*, **589**, 2923–2930.
8. Kellum,R. and Schedl,P. (1991) A position-effect assay for boundaries of higher order chromosomal domains. *Cell*, **64**, 941–950.
9. Kellum,R. and Schedl,P. (1992) A group of scs elements function as domain boundaries in an enhancer-blocking assay. *Mol. Cell. Biol.*, **12**, 2424–2431.
10. Phillips-Cremins,J.E., Sauria,M. E.G., Sanyal,A., Gerasimova,T.I., Lajoie,B.R., Bell,J.S., Ong,C.T., Hookway,T.A., Guo,C., Sun,Y. *et al.* (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.
11. Van Bortle,K., Nichols,M.H., Li,L., Ong,C.-T., Takenaka,N., Qin,Z.S. and Corces,V.G. (2014) Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol.*, **15**, R82.
12. Rao,S.S.P., 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. *et al.* (2015) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
13. Zuin,J., Dixon,J.R., van der Reijden,M.I.J.A., Ye,Z., Kolovos,P., Brouwer,R.W., van de Corput,M.P., van de Werken,H.J., Knoch,T.A., van IJcken,W.F. *et al.* (2014) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 996–1001.
14. Vietri-Rudan,M., Barrington,C., Henderson,S., Ernst,C., Odom,D., Tanay,A. and Hadjur,S. (2015) Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep.*, **10**, 1297–1309.
15. Mourad,R. and Cuvier,O. (2016) Computational identification of genomic features that influence 3D chromatin domain formation. *PLoS Comput. Biol.*, **12**, e1004908.
16. Shin,H., Shi,Y., Dai,C., Tjong,H., Gong,K., Alber,F. and Zhou,X.J. (2016) TopDom: an efficient and deterministic method for identifying topological domains in genomes. *Nucleic Acids Res.*, **44**, e70.
17. Li,L., Lyu,X., Hou,C., Takenaka,N., Nguyen,H.Q., Ong,C.T., Cubeñas-Potts,C., Hu,M., Lei,E.P., Bosco,G. *et al.* (2015) Widespread rearrangement of 3D chromatin organization underlies Polycomb-mediated stress-induced silencing. *Mol. Cell*, **58**, 216–231.
18. Eagen,K.P., Lieberman Aiden,E. and Kornberg,R.D. (2017) Polycomb-mediated chromatin loops revealed by a sub-kilobase resolution chromatin interaction map. *Proc. Natl. Acad. Sci. U.S.A.*, **114**, 8764–8769.
19. Wood,A.M., Van Bortle,K., Ramos,E., Takenaka,N., Rohrbaugh,M., Jones,B.C., Jones,K.C. and Corces,V.G. (2011) Regulation of chromatin organization and inducible gene expression by a *Drosophila* insulator. *Mol. Cell*, **44**, 29–38.
20. Kellner,W.A., Van Bortle,K., Li,L., Ramos,E., Takenaka,N. and Corces,V.G. (2013) Distinct isoforms of the *Drosophila* Brd4 homologue are present at enhancers, promoters and insulator sites. *Nucleic Acids Res.*, **41**, 9274–9283.
21. Negre,N., Brown,C.D., Ma,L., Bristow,C.A.A., Miller,S.W., Wagner,U., Kheradpour,P., Eaton,M.L., Loriaux,P., Sealfon,R. *et al.* (2011) A cis-regulatory map of the *Drosophila* genome. *Nature*, **471**, 527–531.
22. Junion,G., Spivakov,M., Girardot,C., Braun,M., Gustafson,E.H., Birney,E. and Furlong,E.E. (2012) A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, **148**, 473–486.
23. The ENCODE Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
24. Cuellar-Partida,G., Buske,F.A., McLeay,R.C., Whitington,T., Noble,W.S. and Bailey,T.L. (2012) Epigenetic priors for identifying active transcription factor binding sites. *Bioinformatics*, **28**, 56–62.
25. Zhao,K., Hart,C.M. and Laemmli,U.K. (1995) Visualization of chromosomal domains with boundary element-associated factor BEAF-32. *Cell*, **81**, 879–889.
26. Holohan,E.E., Kwong,C., Adryan,B., Bartkuhn,M., Herold,M., Renkawitz,R., Russell,S. and White,R. (2007) CTCF genomic binding sites in *Drosophila* and the organisation of the Bithorax complex. *PLoS Genet.*, **3**, e112.
27. Adryan,B., Woerfel,G., Birch-Machin,I., Gao,S., Quick,M., Meadows,L., Russell,S. and White,R. (2007) Genomic mapping of suppressor of hairy-wing binding sites in *Drosophila*. *Genome Biol.*, **8**, R167.
28. Negre,N., Brown,C.D., Shah,P.K., Kheradpour,P., Morrison,C.A., Henikoff,J.G., Feng,X., Ahmad,K., Russell,S., White,R.A. *et al.* (2010) A comprehensive map of insulator elements for the *Drosophila* genome. *PLoS Genet.*, **6**, e1000814.
29. Zolotarev,N., Fedotova,A., Kyrchanova,O., Bonchuk,A., Penin,A.A., Lando,A.S., Eliseeva,I.A., Kulakovskiy,I.V., Maksimenko,O. and Georgiev,P. (2016) Architectural proteins Pita, Zw5 and ZIPIC contain homodimerization domain and support specific long-range interactions in *Drosophila*. *Nucleic Acids Res.*, **44**, 7228–7241.
30. Hart,C.M., Cuvier,O. and Laemmli,U.K. (1999) Evidence for an antagonistic relationship between the boundary element-associated factor BEAF and the transcription factor DREF. *Chromosoma*, **108**, 375–383.
31. Li,J. and Gilmour,D.S. (2013) Distinct mechanisms of transcriptional pausing orchestrated by GAGA factor and M1BP, a novel transcription factor. *EMBO J.*, **32**, 1829–1841.
32. Read,D. and Manley,J.L. (1992) Alternatively spliced transcripts of the *Drosophila* tramtrack gene encode zinc finger proteins with distinct DNA binding specificities. *EMBO J.*, **11**, 1035–1044.
33. Cuartero,S., Fresan,U., Reina,O., Planet,E. and Espinas,M.L. (2014) Ibf1 and Ibf2 are novel CP190-interacting proteins required for insulator function. *EMBO J.*, **33**, 637–647.

34. Dai,Q., Ren,A., Westholm,J.O., Duan,H., Patel,D.J. and Lai,E.C. (2015) Common and distinct DNA-binding and regulatory activities of the BEN-solo transcription factor family. *Genes Dev.*, **29**, 48–62.
35. Hug,C.B., Grimaldi,A.G., Kruse,K. and Vaquerizas,J.M. (2017) Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell*, **169**, 216–228.
36. Dekker,J., Marti-Renom,M.A. and Mirny,L.A. (2013) Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat. Rev. Genet.*, **14**, 390–403.
37. Hu,M., Deng,K., Selvaraj,S., Qin,Z., Ren,B. and Liu,J.S. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, **28**, 3131–3133.
38. Imakaev,M., Fudenberg,G., McCord,R.P., Naumova,N., Goloborodko,A., Lajoie,B.R., Dekker,J. and Mirny,L.A. (2012) Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat. Methods*, **9**, 999–1003.
39. Gaszner,M. and Felsenfeld,G. (2006) Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.*, **7**, 703–713.
40. Ghandi,M., Mohammad-Noori,M., Ghareghani,N., Lee,D., Garraway,L. and Beer,M.A. (2016) gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics*, **32**, 2205–2207.
41. Kumar,R.P., Krishnan,J., Singh,N.P., Singh,L. and Mishra,R.K. (2013) GATA simple sequence repeats function as enhancer blocker boundaries. *Nat. Commun.*, **4**, 1844.
42. Dali,R. and Blanchette,M. (2017) A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.*, **45**, 2994–3005.
43. Levy-Leduc,C., Delattre,M., Mary-Huard,T. and Robin,S. (2014) Two-dimensional segmentation for analyzing Hi-C data. *Bioinformatics*, **30**, i386–i392.
44. Ghavi-Helm,Y., Klein,F.A., Pakozdi,T., Ciglar,L., Noordermeer,D., Huber,W. and Furlong,E.E. (2014) Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, **512**, 96–100.
45. Maksimenko,O., Bartkuhn,M., Stakhov,V., Herold,M., Zolotarev,N., Jox,T., Buxa,M.K., Kirsch,R., Bonchuk,A., Fedotova,A. *et al.* (2015) Two new insulator proteins, Pita and ZIPIC, target CP190 to chromatin. *Genome Res.*, **25**, 89–99.
46. Herranz,H. and Morata,G. (2001) The functions of pannier during *Drosophila* embryogenesis. *Development*, **128**, 4837–4846.
47. Wang,C. and Xi,R. (2015) Keeping intestinal stem cell differentiation on the Tramtrack. *Fly*, **9**, 110–114.
48. Djekidel,M.N., Liang,Z., Wang,Q., Hu,Z., Li,G., Chen,Y. and Zhang,M.Q. (2015) 3CPET: finding co-factor complexes from ChIA-PET data using a hierarchical Dirichlet process. *Genome Biol.*, **16**, 288.
49. Liang,J., Lacroix,L., Gamot,A., Cuddapah,S., Queille,S., Lhoumaud,P., Lepetit,P., Martin,P.G.P., Vogelmann,J., Court,F. *et al.* (2014) Chromatin immunoprecipitation indirect peaks highlight functional long-range interactions among insulator proteins and RNAII pausing. *Mol. Cell*, **53**, 672–681.
50. Hirano,T. (2005) Condensins: organizing and segregating the genome. *Curr. Biol.*, **15**, R265–R275.
51. Hirano,T. (2006) At the heart of the chromosome: SMC proteins in action. *Nat. Rev. Mol. Cell Biol.*, **7**, 311–322.
52. Gibcus,J. and Dekker,J. (2013) The hierarchy of the 3D genome. *Mol. Cell*, **49**, 773–782.
53. Tan,S.K., Lin,Z.H., Chang,C.W., Varang,V., Chng,K.R., Pan,Y.F., Yong,E.L., Sung,W.K. and Cheung,E. (2011) AP-2γ regulates oestrogen receptor-mediated long-range chromatin interaction and gene transcription. *EMBO J.*, **30**, 2569–2581.
54. Vogelmann,J., Le Gall,A., Dejardin,S., Allemand,F., Gamot,A., Labesse,G., Cuvier,O., Nègre,N., Cohen-Gonsaud,M., Margeat,E. *et al.* (2014) Chromatin insulator factors involved in long-range DNA interactions and their role in the folding of the *Drosophila* genome. *PLoS Genet.*, **10**, e1004544.
55. Pagans,S., Ortiz-Lombardia,M., Espinas,M.L., Bernues,J. and Azorin,F. (2002) The *Drosophila* transcription factor tramtrack (TTK) interacts with Trithorax-like (GAGA) and represses GAGA-mediated activation. *Nucleic Acids Res.*, **30**, 4406–4413.
56. Busslinger,G.A., Stocsits,R.R., van der Lelij,P., Axelsson,E., Tedeschi,A., Galjart,N. and Peters,J.-M. (2017) Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature*, **544**, 503–507.
57. Doyle,B., Fudenberg,G., Imakaev,M. and Mirny,L.A. (2014) Chromatin loops as allosteric modulators of enhancer-promoter interactions. *PLoS Comput. Biol.*, **10**, e1003867.
58. Facchinei,F., Scutari,G. and Sagratella,S. (2015) Parallel selective algorithms for nonconvex big data optimization. *IEEE Trans. Sig. Process.*, **63**, 1874–1889.
59. Jost,D., Carrivain,P., Cavalli,G. and Vaillant,C. (2014) Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains. *Nucleic Acids Res.*, **42**, 9553–9561.

3.4.3.4 TADreg: TAD identification, differential analysis and prediction

Over the past years, tremendous efforts have been made to develop methods for TAD identification from Hi-C data [Zufferey *et al.* 2018]. The methods can be broadly classified into 4 categories: linear score, statistical model, clustering and network features [Zufferey *et al.* 2018]. The first methods split the genome into bins and define a linear score (insulation score) associated with each bin [Dixon *et al.* 2012, Crane *et al.* 2015, Rao *et al.* 2014, Shin *et al.* 2016]. The second methods rely on statistical models of the interaction distributions [Levy-Leduc *et al.* 2014, Weinreb & Raphael 2015, Serra *et al.* 2017]. The third methods cluster regions of the genome [Oluwadare & Cheng 2017, Haddad *et al.* 2017]. The fourth methods consider the Hi-C data as a graph adjacency matrix and TADs as communities to detect [Chen *et al.* 2016, Yan *et al.* 2017a, Norton *et al.* 2018]. However, very few methods were developed to detect differential TADs between experiments [Zaborowski & Wilczynski 2016, Sadowski *et al.* 2019, Cresswell & Dozmorov 2020]. Moreover, few methods were also proposed to predict the impact of chromosomal rearrangement in reshaping TADs, and more generally the 3D genome [Bianco *et al.* 2018, Huynh & Hormozdiari 2019, Sadowski *et al.* 2019, Kaplan 2019, Belokopytova *et al.* 2020].

I proposed a versatile regression framework that generalizes the insulation score by estimating a relative score and adding a sparsity constrain (”Sparse Insulation Model”, SIM), but also allows differential TAD analysis (”Differential Insulation Model”, DIM) and Hi-C data prediction after chromosomal rearrangement (”Prediction Insulation Model”, PIM) (from submitted article ”TADreg : A versatile regression framework for TAD identification, differential analysis and rearranged 3D genome prediction” below). The proposed model provides a rigorous statistical framework for modeling the interaction distribution, where the model parameters represent sparse insulation scores that have an intuitive interpretation and are easy to visualize (Figures 1A and 1B, article below). Our model assumes additivity of insulation parameters as previously proposed by [Rowley *et al.* 2017, Mourad & Cuvier 2018, Huynh & Hormozdiari 2019, Kaplan 2019]. By adding interaction terms in the model, the regression framework can naturally be used for differential TAD border identification between two different Hi-C experiments. Moreover, the regression can predict Hi-C data in the case of structural variants, thereby allowing to explore the deleterious impact of the de novo enhancer-promoter interactions. Using recent high resolution human and mouse Hi-C data, I found that our approach ranked among the top TAD callers, when evaluated using external assessment designed not to favor any tool. Moreover, it identified new features of the genome, we called TAD facilitators, which were demonstrated to be biologically relevant. Our approach could also identify numerous differential TAD borders involved in cortical neuron differentiation. Such borders were depleted in CTCF compared to embryonic stem cells and enriched in a large number of known neuronal transcription factors including NFATC1/3, NEUROD2, HiC1 and Dmbx1. Lastly, my approach outperformed state-of-the-art algorithm PRISMR to predict

Hi-C data after chromosomal rearrangement.

Mourad *BMC Bioinformatics* (2022) 23:82
https://doi.org/10.1186/s12859-022-04614-0

BMC Bioinformatics

RESEARCH Open Access

TADreg: a versatile regression framework for TAD identification, differential analysis and rearranged 3D genome prediction

Raphaël Mourad*

*Correspondence:
raphael.mourad@univ-tlse3.fr
CNRS, UPS, MCD, Centre de Biologie Intégrative (CBI), University of Toulouse, 31062 Toulouse, France

Abstract

Background/Aim: In higher eukaryotes, the three-dimensional (3D) organization of the genome is intimately related to numerous key biological functions including gene expression, DNA repair and DNA replication regulations. Alteration of 3D organization, in particular topologically associating domains (TADs), is detrimental to the organism and can give rise to a broad range of diseases such as cancers.

Methods: Here, we propose a versatile regression framework which not only identifies TADs in a fast and accurate manner, but also detects differential TAD borders across conditions for which few methods exist, and predicts 3D genome reorganization after chromosomal rearrangement. Moreover, the framework is biologically meaningful, has an intuitive interpretation and is easy to visualize.

Result and conclusion: The novel regression ranks among top TAD callers. Moreover, it identifies new features of the genome we called TAD facilitators, and that are enriched with specific transcription factors. It also unveils the importance of cell-type specific transcription factors in establishing novel TAD borders during neuronal differentiation. Lastly, it compares favorably with the state-of-the-art method for predicting rearranged 3D genome.

Keywords: Chromatin interaction, Hi-C, ChIP-seq, Insulator binding protein, Generalized linear model

Introduction

In higher eukaryotes, chromosomes are packed into three dimensions (3Ds) and form complex structures [1]. Such 3D structure of chromosomes has recently been investigated by chromosome conformation capture combined with high-throughput sequencing technique (Hi-C) at an unprecedented resolution [2–4]. Hi-C experiments revealed multiple levels of genome organization including compartments A/B [5] and topologically associating domains (TADs) [2, 3]. Most notably, TADs are relatively constant between different cell types and are highly conserved across species. Those TADs play central roles in key cell processes such as for the long-range regulation of genes by enhancers [4] or for the replication-timing regulation [6].

Over the past years, tremendous efforts have been made to develop methods for TAD identification from Hi-C data [7]. The methods can be broadly classified into 4 categories: linear score, statistical model, clustering and network features [7]. The first methods split the genome into bins and define a linear score (insulation score) associated to each bin [2, 8–10]. The second methods rely on statistical models of the interaction distributions [11–13]. The third methods cluster regions of the genome [14–16]. The fourth methods consider the Hi-C data as a graph adjacency matrix and TADs as communities to detect [17–19]. However, very few methods were developed to detect differential TADs between experiments [20–22]. Moreover, few methods were also proposed to predict the impact of chromosomal rearrangement in reshaping TADs, and more generally the 3D genome [21, 23–26].

We propose a versatile regression framework that generalizes the insulation score by estimating a relative score and adding a sparsity constrain ("Sparse Insulation Model", SIM), but also allows differential TAD analysis ("Differential Insulation Model", DIM) and Hi-C data prediction after chromosomal rearrangement ("Prediction Insulation Model", PIM). The proposed model provides a rigorous statistical framework for modeling the interaction distribution, where model parameters represent sparse insulation scores that have an intuitive interpretation and are easy to visualize. Our model assumes additivity of insulation parameters as previously proposed by [24, 25, 27, 28]. By adding interaction terms into the model, the regression framework can naturally be used for differential TAD border identification between two different Hi-C experiments. Moreover, the regression can predict Hi-C data in the case of chromosomal rearrangements such as deletion and inversion, thereby allowing to explore the deleterious impact of de novo enhancer-promoter interactions on genetic diseases and cancers.

Using recent high resolution human and mouse Hi-C data, we found that our approach ranked among the top TAD callers, when evaluated using external assessment designed not to favor any tool. Moreover, it identified new features of the genome we called TAD facilitators, which were demonstrated to be biologically relevant. Our approach could also identify numerous novel TAD borders emerging during cortical neuron differentiation. Such borders were depleted in CTCF compared to embryonic stem cells and enriched in a large number of known neuronal transcription factors including NFATC1/3, NEUROD2, HiC1 and Dmbx1. Lastly, our approach outperformed state-of-the-art algorithm PRISMR to predict Hi-C data after chromosomal rearrangement.

Materials and methods

Hi-C data

We used publicly available Hi-C data of lymphoblastoid GM12878 and lung IMR90 cells from Gene Expression Omnibus (GEO) accession GSE63525 [9]. We also used publicly available Hi-C data of mouse embryonic stem (ES) and cortical neuron (CN) cells from GEO accession GSE96107 [29]. Hi-C data were binned at 25 and 50 kb resolutions and normalized by matrix balancing [30].

Capture Hi-C data

We used publicly available capture Hi-C data of wild-type (WT) and mutant distal limb buds of E11.5 mice from Gene Expression Omnibus (GEO) accession GSE92294

[23]. Hi-C data were binned at 10 kb resolution and normalized by matrix balancing [30].

ChIP-seq data

We used publicly available binding peaks of 73 chromatin proteins (Rad21, CTCF, YY1, ZBTB33, MAZ, JUND, ZNF143, EZH2, ATF2, ATF3, BATF, BCL11A, BCL3, BCLAF1, BHLHE40, BRCA1, CEBPB, CFOS, CHD1, CHD2, CMYC, COREST, E2F4, EBF1, EGR1, ELF1, ELK1, FOXM1, GABP, IKZF1, IRF4, MAX, MEF2C, MTA3, MXI1, NFATC1, NFE2, NFIC, NFKB, NFYA, NFYB, NRF1, NRSF, P300, PAX5, PBX3, PML, POL2, POL3, POU2F2, RFX5, RUNX3, RXRA, SIN3A, SIX5, SMC3, SP1, SPI1, SRF, STAT1, STAT3, STAT5, TBLR1, TBP, TCF12, TCF3, TR4, USF1, USF2, WHIP, ZEB1, ZNF274, ZZZ3) of GM12878 cells from ENCODE [31]. We downloaded peaks that were uniformly processed (Uniform Peaks).

We also used publicly available CTCF ChIP-seq data of mouse embryonic stem (ES) and cortical neuron (CN) cells from GEO accession GSE96107 [29].

JASPAR motifs

To scan the mouse genome for motif occurrences, we used FIMO with default parameters (meme-suite.org). The motif position weight matrices were downloaded from JASPAR database (http://jaspar.genereg.net/).

TAD manual annotation

We used manual annotation of GM12878 TADs at 50 kb from Dali and Blanchette [32]. As previously described by Dali and Blanchette, TADs were manually traced on GM12878 Hi-C maps from the full data set at 50 kb resolution for regions 40-45 mb of 10 different, randomly chosen, chromosomes (chr2, chr3, chr4, chr5, chr6, chr7, chr12, chr18, chr20 and chr22). Briefly, interaction maps of the regions of interest were plotted using HiCplotter. In Adobe Illustrator, dotted squares were manually traced around visually identifiable TADs on the interaction map plots. Regions annotated as TADs had the following properties: (i) sharp visual contrast between within and across TAD interaction frequencies, over the entire TAD region; (ii) minimum size of 250 kb. To give all tools an equal chance, Dali and Blanchette created a dense set of TAD annotations that included any identifiable TAD structure. For example, if two potential TADs were overlapping, both were retained, irrespective of whether one had stronger visual support than the other. TAD boundaries were allowed to overlap or be nested, as long as there is a clearly traceable square along the diagonal. Bed files with TAD ranges were manually created and used for tool comparison.

Since 29% of genomic bins could be considered as relevant TAD borders using this annotation, we considered as TAD borders those supported by at least two TADs that were manually identified.

Insulation score

For a bin $i \in \{1, ..., p\}$, the insulation score was defined as [8]:

$$IS_i = \log_2\left(\frac{M_i}{\frac{1}{p}\sum_{i=1}^{p} M_i}\right), \tag{1}$$

where M_i was the number of Hi-C counts that occurred across bin i (up to some distance) on the same chromosome.

Sparse insulation model (SIM)

We first removed the distance effect (polymer effect) from the normalized Hi-C counts using a generalized additive model with a negative binomial distribution:

$$\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{d}\right]\right) = \beta_0 + \mathrm{f}(\mathbf{d}) \tag{2}$$

Variable **y** denoted normalized Hi-C count for any pair of bins on the same chromosome. The log-distance variable **d** accounted for the background polymer effect. The local power law decay relation between distance and Hi-C count was modeled by regression spline [33]. We noted that if bias variables such as GC content, mappability and fragment length were added to the model [34], then the model could also handle unnormalized Hi-C data. Regression residuals (noted **z**) were then used as input for a linear model. Using residuals allowed us to then use best subset selection (L0 penalty) for which there is only linear model implementation in R (see as follows).

Then, a linear model called the "sparse insulation model" (SIM) was proposed to estimate the insulating effects of genomic loci on long-range interactions:

$$\mathrm{E}\left[\mathbf{z}|\mathbf{X}\right] = \beta_0 + \mathbf{X}\boldsymbol{\beta}_{\mathrm{X}} \tag{3}$$

Variable set $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_p\}$ represented the p insulation variables, one for each bin of the chromosome. For a bin $i \in \{1, ..., p\}$, the insulation variable $\mathbf{x}_i$ was set to one when the bin lied in-between the two bins whose interaction counts were measured by Hi-C, and was set to zero otherwise. The corresponding β_{x_i} parameter value reflected the effect of the bin i on Hi-C counts. A negative beta value ($\beta_{x_i} < 0$) revealed an insulation effect on long-range contacts. Conversely, a positive beta value ($\beta_{x_i} > 0$) showed a facilitating effect on contacts. A null beta value ($\beta_{x_i} = 0$) meant that the bin had no effect on contacts.

Best subset selection was used to select the best insulation variables when estimating the $\boldsymbol{\beta}_X$ parameters by adding an L0 penalty:

$$\min_{\beta_0, \boldsymbol{\beta}_X} \frac{1}{N}\sum_{j=1}^{N} l(z_j, \beta_0 + X_j\boldsymbol{\beta}_X) + \lambda||\boldsymbol{\beta}_X||_0 \tag{4}$$

as done using the L0Learn R package (https://cran.r-project.org/web/packages/L0Learn). Parameter λ was obtained by 10 fold cross-validation of the mean square error (L0Learn.cvfit function with default parameters).

Often the number of insulation variables was too big for L0Learn R package (>5000) and we had to prefilter the variables. For this purpose, we used lasso regression (glmnet R package, https://cran.r-project.org/web/packages/glmnet/) and kept variables with $|\hat{\beta}_{x_i}| > 0.2$. This allowed to reduce the number of variables to few thousands for L0Learn

to work, while still keeping most relevant variables. We found that prefiltering yielded betas that were similar to the ones obtained without prefiltering (Additional file 1: Figure S1).

Differential insulation model (DIM)

The model could be extended to identify differential TAD borders between two different Hi-C experiment matrices (*e.g.* between two conditions). For this purpose, we first ran SIM for each Hi-C experiment matrix independently. Only the union of bins with $|\hat{\beta}_{x_i}| > 0$ from both SIMs were kept for differential analysis (we noted the new bin set $\mathbf{S} = \{\mathbf{s}_1, ..., \mathbf{s}_q\}$). To prevent bin uncertainty between experiments, only one bin was kept among two consecutive bins. Bins from **S** were then used to build a novel model for differential analysis called the "differential insulation model" (DIM).

The differential insulation model was written as follows:

$$\mathrm{E}\left[\mathbf{z}|\mathbf{S},\mathbf{e}\right] = \beta_0 + \mathbf{S}\boldsymbol{\beta}_{\mathrm{S}} + \beta_{\mathrm{e}}\mathbf{e} + \sum_{\mathrm{j}=1}^{\mathrm{q}} \beta_{\mathrm{s_j e}}\mathbf{s_j e} \tag{5}$$

Variable **e** denoted the experiment from which the Hi-C count is measured. Variable $\mathbf{s}_j\mathbf{e}$ was the interaction term between the insulation variable $\mathbf{s}_j$ and the experiment variable **e**, computed as the product between both variables. For a bin *j*, a negative beta value ($\beta_{s_j e} < 0$) revealed higher insulation effect on long-range contacts for the 2nd experiment compared to the 1st experiment, while a positive value ($\beta_{s_j e} > 0$) meant lower insulation effect. A null value ($\beta_{s_j e} = 0$) showed no differential effect. Because the model used as input only bins previously identified by the sparse insulation model, there was no need to use any penalty for parameter estimation. Moreover, the absence of a penalty term allowed to estimate differential effects without bias.

Prediction insulation model (PIM)

The model could be modified to predict Hi-C data, which we called the "prediction insulation model" (PIM). For this purpose, we modeled the Hi-C count by a generalized linear model (Poisson regression):

$$\log\left(\mathrm{E}\left[\mathbf{y}|\mathbf{d},\mathbf{X}\right]\right) = \beta_0 + \beta_{\mathrm{d}}\mathbf{d} + \mathbf{X}\boldsymbol{\beta}_{\mathrm{X}} \tag{6}$$

Here, since we didn't need to identify sharply the borders with L0 penalty, we could use directly the Poisson regression. PIM could be used to predict Hi-C data after chromosomal rearrangement. For this purpose, PIM was first trained using wild-type Hi-C data (no rearrangement). Then, the distance variable (**d**) and the insulation variables (**X**) were modified in a way to account for the chromosomal rearrangement. In the case of a deletion, the distance variable values were shrunk by the length of the deletion (producing a new distance variable noted $\mathbf{d}'$), and all insulation variables spanning the deletion were set to zero (producing new insulation variables $\mathbf{X}'$). In the case of an inversion, bins spanning the inversion were flipped and the distance variable and insulation variables were recomputed accordingly. The new variables ($\mathbf{d}'$ and $\mathbf{X}'$) together with the trained PIM model (with parameters $\hat{\beta}_0$, $\hat{\beta}_d$ and $\hat{\boldsymbol{\beta}}_X$) were used to predict Hi-C data after rearrangement:

$$\log(\mathrm{E}[\mathbf{y}|\mathbf{d}', \mathbf{X}']) = \hat{\beta}_0 + \hat{\beta}_{\mathrm{d}}\mathbf{d}' + \mathbf{X}'\hat{\boldsymbol{\beta}}_X \tag{7}$$

Results and discussion

Identification of TAD borders and facilitators

We proposed the sparse insulation model (SIM) to estimate the insulating/facilitating effects of genomic loci on long-range interactions (Fig. 1A). SIM required only one parameter, the maximal distance between two bins from the Hi-C matrix, which we set here to bin size $\times 10$ in order to reduce computational burden. We illustrated the model with high-depth Hi-C data at 25 kb resolution from human IMR90 lung cells, whose TADs could be easily visualized. We plotted the example of a 10-Mb-long genomic region of chromosome 1 (Fig. 1B). We first computed the insulation score (IS) to identify loci of high insulation. The insulation score is a standard measure reflecting the aggregate of interactions occurring across each interval. It is often used by experimentalists because of its simple and quantitative interpretation: the lower, the higher the insulation effect of the loci on overlapping contacts [8]. We observed peaks of negative IS, reflecting the presence of TAD borders with varying strengths (Fig. 1C). Alternatively, IS also revealed regions facilitating long-range contacts (score above zero).

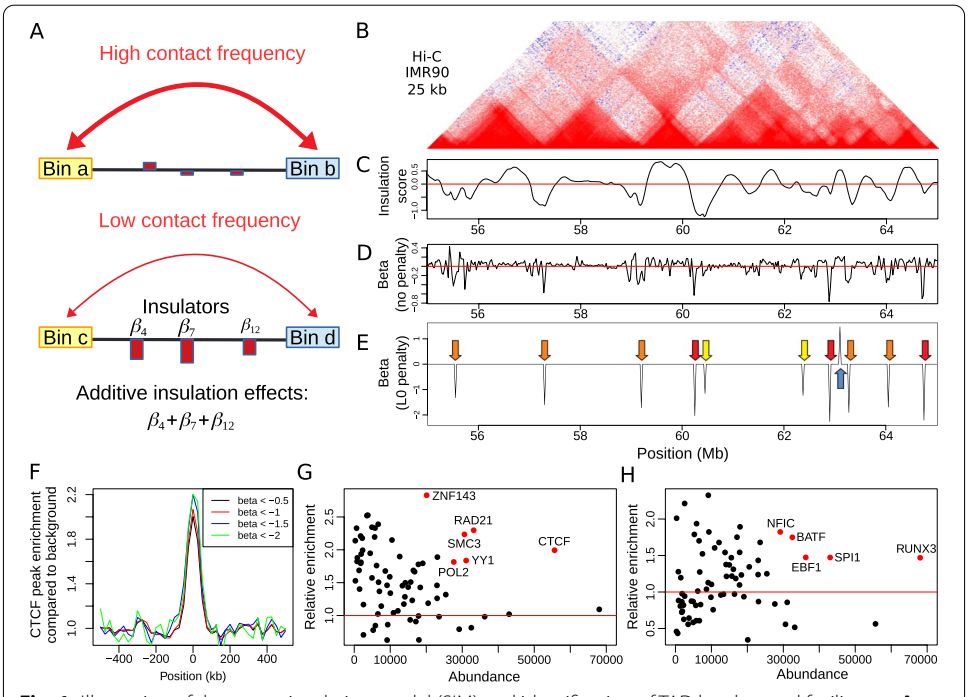


Fig. 1 Illustration of the sparse insulation model (SIM) and identification of TAD borders and facilitators. **A** Schema representing the insulation effects modeled by SIM on long-range contacts between two bins (two loci), such as between an enhancer and a promoter. **B** Hi-C heatmap from IMR90 cells at 25 kb resolution. **C** Insulation score. **D** SIM beta (no penalty). **E** SIM beta (L0 penalty). **F** CTCF enrichment profile depending on border strength (beta). **G** DNA binding protein enrichment compared to background for several proteins at TAD borders identified by our model (negative betas) depending on protein abundance. **H** DNA binding protein enrichment compared to background for several proteins at TAD facilitators identified by our model (positive betas) depending on protein abundance

Using SIM, we estimated instead sparse insulation scores (beta parameters). For a bin i, the β_{x_i} parameter has a nice and intuitive interpretation: it is the insulation score, after accounting for the insulating/facilitating effects of the other bins. If no penalty is used to learn beta parameters, the betas correspond to a relative score (Fig. 1D). Using this relative score, we observed sharp peaks instead of wide valleys with the standard IS which prevented accurate location of TAD borders. Moreover, if an L0 penalty is used, then the regression leads to a sparse estimation of the insulation score. This helped to identify the exact location of bins with insulating/facilitating effects (Fig. 1E), in contrast to IS. In SIM, a negative beta value ($\beta_{x_i} < 0$) reveals an insulation effect on long-range contacts (the bin is an insulator). Conversely, a positive beta value ($\beta_{x_i} > 0$) shows a facilitating effect on contacts (the bin is a facilitator). A null beta value ($\beta_{x_i} = 0$) means that the bin has no effect on contacts.

In the genomic region, SIM could detect ten TAD borders ($\hat{\beta} < 0$). Using SIM, TADs could be simply defined as regions in-between two consecutive TAD borders. Visual inspection of the Hi-C matrix clearly revealed that our TAD identification was relevant (Fig. 1B). Moreover, SIM could identify TAD borders with varying strengths. We found three strong TAD borders ($\hat{\beta} < -2$; red arrows), five moderate TAD borders ($-1.2 < \hat{\beta} < -2$; orange arrows) and two weak TAD borders ($\hat{\beta} \approx -1.1$; yellow arrows). Moreover, the model uncovered one region with facilitating effects ($\hat{\beta} > 0$; blue arrow).

We then looked at the enrichment of the CTCF protein, a major 3D genome organizer, at TAD borders over the whole genome depending on the beta value. Here, we used GM12878 Hi-C data for which there are ChIP-seq data for a very large number of proteins, which helped us to comprehensively assess the role of DNA-binding proteins (see bellow). Overall, we found a strong two-fold enrichment of CTCF at TAD borders (Fig. 1F). Moreover, we observed that stronger TAD borders presented higher CTCF enrichment (2-fold for $\hat{\beta} < -0.5$; 2.2-fold for $\hat{\beta} < -1.5$), meaning that border strength estimated by SIM scaled accordingly with CTCF presence. Then, we evaluated enrichment for all available protein binding ChIP-seq data, and observed as previously shown the highest enrichments for CTCF, RAD21, SMC3, ZNF143, YY1 and POL2 (Fig. 1G) [2, 35, 36]. SIM could also identify regions facilitating contacts *e.g.* regions with $\hat{\beta} > 0$ (we called "TAD facilitators"), unlike most TAD detection tools. IS could also detect facilitators, but without accurate location, thereby preventing enrichment analysis. Using SIM, we found that lymphocyte transcription factors (TFs) BATF, EBF1, NFIC, RUNX3 and SPI1 were enriched at such facilitator regions (Fig. 1H). Such high enrichment revealed that TAD facilitators were indeed biologically meaningful regions.

Thus, we could conclude that SIM had an intuitive interpretation in terms of insulating/facilitating quantitative effects, which could also sharply identify TAD borders unlike the insulation score. Moreover, our model could accurately identify a novel class of 3D elements that we called TAD facilitators, which were highly enriched in cell specific TFs.

Performance and comparison with state-of-the-art tools

SIM was very accurate to identify TAD borders. We compared it to 7 other algorithms including Armatus, Arrowhead, DomainCaller, TADbit, TADtree, TopDom, HiCseg using human GM12878 Hi-C data as from [32] (Fig. 2). At both 25 kb and 50 kb, SIM

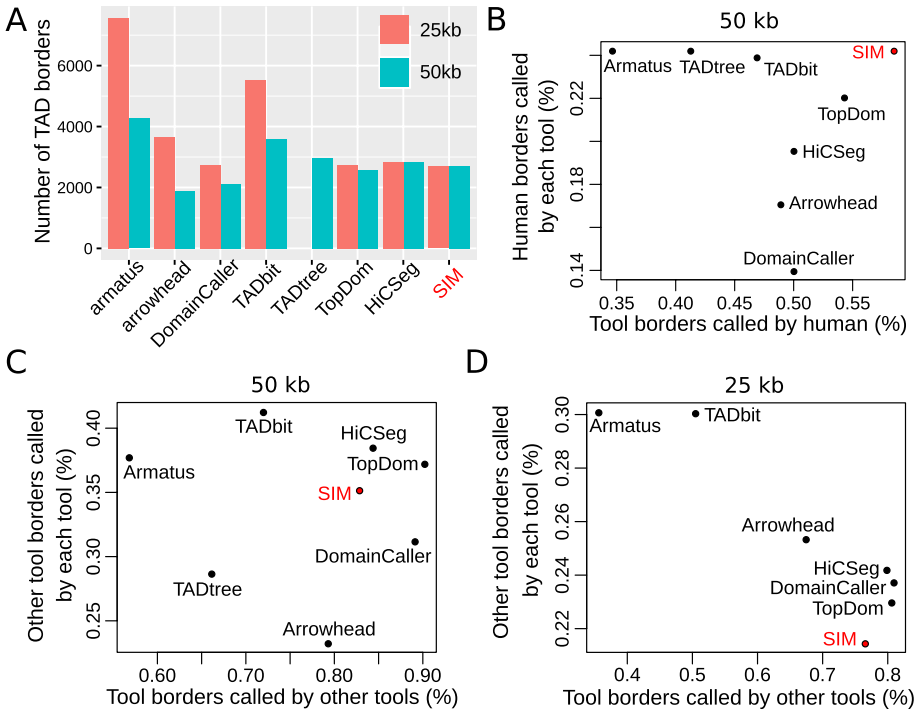


Fig. 2 Comparison with existing TAD calling methods using GM12878 Hi-C dataset. **A** Number of TAD borders called by each method (50 kb resolution). **B** TAD border prediction concordance with manual annotation from [32] (50 kb). **C** TAD border prediction concordance between methods (50 kb). **D** TAD border prediction concordance between methods (25 kb)

identified a small number of TAD borders (2691 and 2711, respectively), such as HiCSeg (2835 and 2835, respectively) and TopDom (2738 and 2568, respectively) (Fig. 2A). Conversely, Armatus identified much more TAD borders (7567 and 4265, respectively) (Fig. 2A). Overall, we found that the number of borders identified by SIM (as well as HiCseg and TopDom) was only slightly impacted by Hi-C data resolution, unlike for the other algorithms. We also compared the TAD borders identified by SIM for different normalizations of the Hi-C data (Knight-Ruiz (KR) [30], iterative correction and eigenvector decomposition (ICE) [37] and square root vanilla coverage (VC SQRT) [38]), and globally found similar results at 50 kb resolution (Additional file 1: Figure S2). We then compared TAD border prediction concordance with manual annotation of TADs at 50 kb from [32] (Fig. 2B). These manually annotated TADs represented an external assessment which was designed not to favor any tool. We found that 58.5% of borders predicted by SIM were also found by manual annotation, which ranked first SIM. Moreover, SIM was able to detect 24.2% of manually annotated borders. In comparison, the large numbers of TAD borders detected by Armatus (>4000 at 50 kb) or TADbit (>3500 at 50 kb) were proportionally less confirmed by manual annotation (34.6% and 46.9%, respectively).

We then assessed TAD border prediction concordance between the different tools. At 50 kb, 82.8% of borders detected by SIM were also identified by the other tools, and 35.2% of other tools' borders were called by SIM, which was similar to the top tools, HiCSeg and TopDom (Fig. 2C). At 25 kb, 76.5% of borders detected by SIM were also

identified by the other tools, and 21.4% of other tools' borders were called by SIM, which was similar to HiCSeg, TopDom and DomainCaller (Fig. 2D). Thus, SIM ranked among the best tools to predict TAD border. Meanwhile, SIM was relatively fast and memory efficient. For chromosome 1 with 25 kb resolution and considering a maximal distance of 250 kb, SIM ran in only 151 seconds for one core and around 6.9 Gb.

Identification of novel borders during cell differentiation

The 3D genome is dynamic, especially during the developmental process, and global reorganization was previously reported during differentiation [29]. However, very few methods were developed for differential analysis of TADs [20, 22]. Using our versatile regression framework, we could easily implement differential TAD analysis in order to identify novel TAD borders, or alternatively depleted TAD borders, during cell differentiation. For this purpose, interaction terms were added in the model to account for differential insulation effects depending on the cell type. We called this model the differential insulation model (DIM). The corresponding interaction betas were then used to assess differential TAD border strength.

To illustrate differential analysis, we studied mouse embryonic stem cells (ESs) differentiation into cortical neurons (CNs) using ultra-deep coverage Hi-C, where novel TAD borders were shown to colocalize with developmental genes that were activated [29]. We first focused on a 5-Mb-long genomic region of chromosome 18 around the developmental gene *Zfp608*. In ES cells, we observed a big TAD in the middle of the Hi-C map (Fig. 3A, C). In CN cells, this big TAD was split into two new TADs separated by a novel border located at 55 Mb overlapping the gene *Zfp608* (Fig. 3B, D). Using the two Hi-C maps, DIM accordingly identified a strong and significant differential TAD border at 55 Mb ($\hat{\beta} \approx -1.8$, $p < 10^{-70}$; blue arrow; Fig. 3E), reflecting TAD split during differentiation. Moreover, DIM could also reveal less obvious differences in border strength. In particular, DIM detected two smaller differential TAD borders ($\hat{\beta} < 1.2$, $p < 10^{-8}$; red arrows), which corresponded to borders present in ES cells and lost in CN cells.

We then ran differential analysis by DIM genome-wide. We observed a higher number of TAD borders after differentiation (fold-change = 1.1; Fig. 3F, left), meaning that new TADs were created after differentiation. If we only considered strong TAD borders, we observed an even larger number of TAD borders after differentiation (fold-change = 1.51 for abs(beta)> 1; fold-change = 2.82 for abs(beta) > 1.5). Moreover, the absolute values of DIM betas in CN were significantly higher than in ES (fold-change = 1.11, p-value = 0.01; Fig. 3G), suggesting that those new TADs were particularly strong and insulated. We then compared CTCF enrichment at CN-specific borders and ES-specific borders (Fig. 3H). We found that although CTCF was very enriched at ES borders (fold-change = 1.64), it was far less enriched at CN borders (fold-change = 1.07), suggesting that the novel TAD borders were maintained by other factors than CTCF. It was previously showed that novel TAD borders located to neural transcription factors Pax6, NeuroD2, and Tbr1 [29]. However, their analysis was limited by available ChIP-seq data. Here, instead, we systematically assessed the enrichment of 579 protein binding DNA motifs at novel CN borders (Fig. 3I). We found a tremendous amount of motifs enriched at novel borders. All enriched motifs were known neural TFs, including Foxd3, NFATC3, NEUROD2, HiC1, Dmbx1, Hmx2 and NFATC1. This result suggested

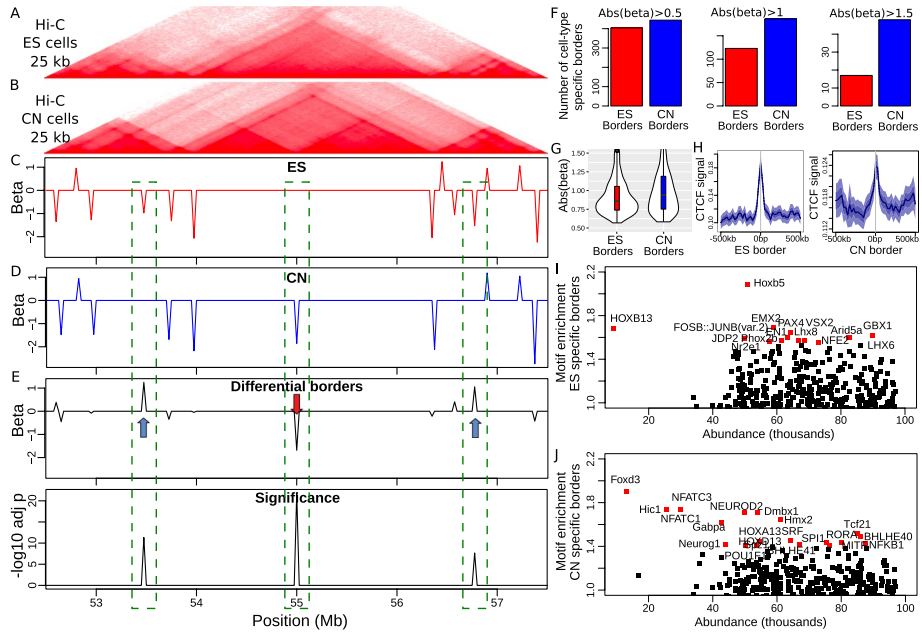


Fig. 3 Differential analysis of TAD borders with the Differential Insulation Model (DIM). **A** Hi-C heatmap in mouse embryonic stem (ES) cells. **B** Corresponding Hi-C heatmap in cortical neuron (CN) cells. **C** Identification of TAD borders in ES cells (spase insulation model beta is plotted). **D** Identification of TAD borders in CN cells (spase insulation model beta is plotted). **E** Identification of differential TAD borders (DIM beta is plotted). For each beta, an adjusted *p*-value is plotted to show significance. **F** Number of cell-type specific borders, for varying differential border strengths. **G** Absolute value of beta between CN specific borders and ES specific borders. **H** CTCF enrichment at CN specific borders compared to ES specific borders. **I** DNA-binding protein motif enrichment (fold-change) at CN specific borders. **J** DNA-binding protein motif enrichment (fold-change) at ES specific borders

that chromatin was reorganized due to not only Pax6, NeuroD2, and Tbr1, but also to numerous other TFs involved in neural differentiation. In comparison, ES borders were strongly enriched in known stem cell TFs, such as Hoxb5, EMX2, PAX4. Thus, we could conclude that cell type specific TFs played a major role in reshaping the genome in 3D during differentiation.

Predictions of Hi-C data after chromosomal rearrangements

Our versatile regression framework could also be used to faithfully model the 3D genome and predict Hi-C data. In particular, predicting the effects of chromosomal rearrangement on 3D genome is an important challenge, since 3D genome alteration can impact essential cellular processes such as enhancer-promoter transcriptional regulation. However, until now, only few methods were developed for this task. Hence, we assessed the ability of the model to predict Hi-C data after chromosomal rearrangement. In this case, we called this model the prediction insulation model (PIM). For this purpose, PIM was trained on wild-type (WT) Hi-C data, producing a model with parameters $\hat{\beta}_0$, $\hat{\beta}_d$ and $\hat{\boldsymbol{\beta}}_X$. Then, in the PIM model, the distance variable ($\mathbf{d}$) and the insulation variables ($\mathbf{X}$) were modified in a way to account for the chromosomal rearrangement. For instance, in the case of a deletion, the distance variable values were shrunk by the length of the deletion (producing a new distance variable noted $\mathbf{d}'$), and all insulation variables spanning

the deletion were set to zero (producing new insulation variables $\mathbf{X}'$). The new variables together with the trained PIM model were used to predict Hi-C after deletion.

PIM prediction accuracy was assessed using 10 kb resolution capture Hi-C experiments performed in E11.5 limb buds from WT and mutant mouses with a deletion or an inversion [23]. For the DelB/DelB mutant (homozygous deletion), we found very accurate Hi-C data predictions as compared to observed data in the mutation mouse (Fig. 4A). Most notably, PIM was able to finely model the distance effect, the numerous TADs, but also the complex hierarchies of TADs. Prediction accuracy was very high as measured by Pearson correlation between log-counts $r = 0.882$ and Spearman correlation between counts $r_s = 0.879$ (Fig. 4A). In comparison, the state-of-the-art model PRIMSR achieved comparable performance in terms of Pearson and Spearson correlations ($r = 0.821$, $r_s = 0.895$; Fig. 4B). But, when distance effect was removed using stratum adjusted correlation in order to only capture the biological variability, PIM performed better than PRISMR (PIM:

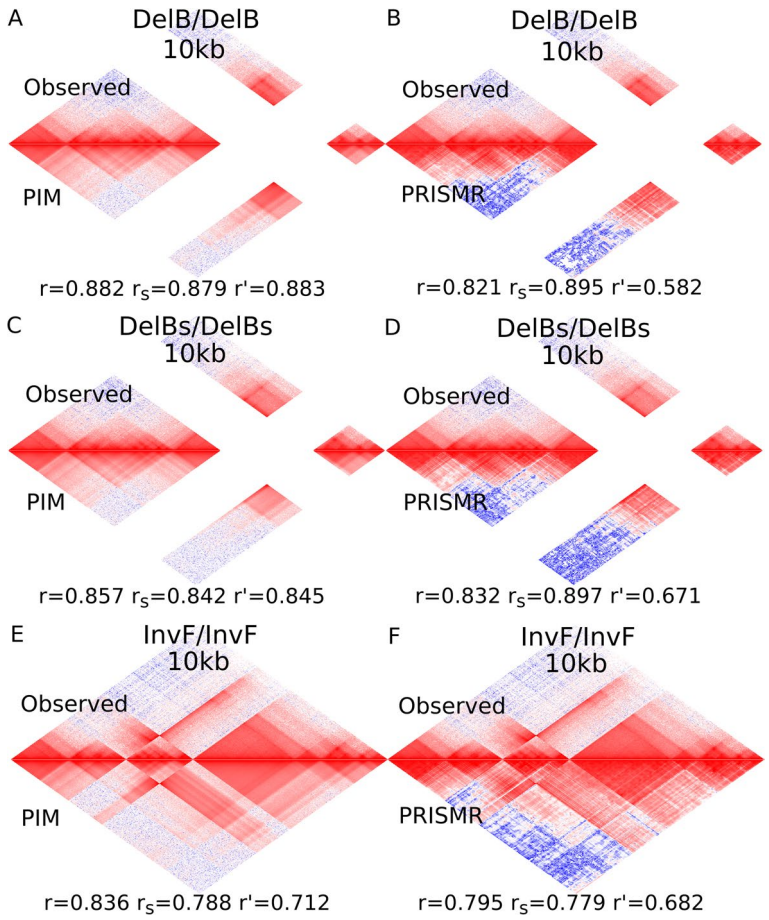


Fig. 4 Prediction insulation model (PIM) predicts rearranged 3D genome with high accurary and comparison with PRISMR using mouse data from [23]. Models were trained using wild-type Hi-C data to predict rearranged Hi-C data (for PRISMR, we used predictions provided by the authors). **A** PIM prediction for DelB/DelB genotype and comparison with observed data. **B** PRISMR prediction for DelB/DelB genotype. **C** PIM prediction for DelBs/DelBs genotype and comparison with observed data. **D** PRISMR prediction for DelBs/DelBs genotype. **E** PIM prediction for InvF/InvF genotype and comparison with observed data. **F** PRISMR prediction for InvF/InvF genotype

$r' = 0.883$ and PRISMR: $r' = 0.582$; Fig. 4A, B), reflecting its better ability to model biological variability underlying TADs and sub-TADs. We next compared PIM and PRIMSR using other mouse mutants. For the DelBs/DelBs mutant, we also found that PIM and PRISMR achieved similar performance in term of r and r_s (PIM: $r = 0.857$, $r_s = 0.842$; PRISMR: $r = 0.832$, $r_s = 0.897$; Fig. 4C, D), but PIM predictions compared favorably in term of biological variability with r' (PIM: $r = 0.845$; PRISMR: $r = 0.671$; Fig. 4C, D). Lastly, we predicted data for an inversion (InvF/InvF). As for deletions, we found that PIM yielded better predictions than PRISMR in term of biological variability with r'.

Conclusion

In this article, we propose a versatile regression framework for Hi-C data analyses. Our framework was designed for TAD identification (SIM model), but also differential analysis (DIM model) and Hi-C data predictions after chromosomal rearrangement (PIM model). First, SIM accurately detected TAD borders in a quantitative manner, and was ranked among the top TAD callers when comparing with state-of-the-art methods on an unbiased dataset. Moreover, SIM also identified a novel class of elements we called facilitators which facilitated long-range contacts as opposed to borders, and were shown to be associated with specific transcription factors. Second, DIM identified novel borders during neuronal differentiation. Such novel borders were particularly enriched for other factors than CTCF, in particular, numerous transcriptional factors specific to neurons including Foxd3, NFATC3, NEUROD2, HiC1, Dmbx1, Hmx2 and NFATC1. In comparison, ES specific borders were enriched in stem cell TFs. Third, PIM accurately predicted rearranged 3D genome in mouse mutants, when trained with wild-type Hi-C data. Such approach is very promising to assess the impact of chromosomal rearrangements on the 3D genome. Moreover, PIM compared favorably with state-of-the-art PRISMR in terms of biological variability captured by Hi-C data.

There are several limitations of the proposed framework. First, the proposed framework is designed for the analysis of bulk Hi-C data, *i.e.* data from a population of cells. However, single-cell experiments are getting widely used in 3D genome studies, and necessitate the development of new tools. The proposed framework must be further extended for data that are too sparse, which is the case for single cell data. The use of an empirical Bayes approach to estimate regression betas across cells might be a elegant solution for this purpose. Second, the same framework can be further extended for other Hi-C data analysis tasks. For instance, the regression can be used to infer frequently interacting regions (FIREs) and differential FIREs from Hi-C data [39]. Third, variable selection for the SIM model is based on best subset selection using L0Learn R package. However, one problem is that L0Learn cannot work with more than 5000 variables on a standard computer, and for the largest chromosomes, prefiltering is done using lasso regression and a threshold of $|\hat{\beta}_{x_i}| > 0.2$ to sufficiently reduce the number of variables for processing. However, this prefiltering might affect best subset selection. Other prefiltering approaches not relying on an arbitrary thresholding can be used instead. For instance, knockoff can be used for removing unnecessary variables while controlling the false discovery rate (FDR) [40]. Alternatively, bootstrap stability investigation can be used [41]. Fourth, SIM is methodologically similar to other TAD callers based on the computation of a linear score such as TopDom [10] or those based on statistical models

of the interaction distributions such as HiCseg [11]. We thus expect SIM to call similar TAD borders (performances between SIM, TopDom and HiCseg were similar, Fig. 2). But SIM is very different from other TAD callers based on clustering [14–16] or graphs [17–19], and thus SIM is more likely to miss those TADs. Fifth, compared to other TAD callers, SIM is conservative for the detection of TAD borders, meaning that fewer but correct TADs were called rather than many TADs including a few false positives. This stringency is related to the use of best subset selection. The use of other variable selection procedures could be investigated to assess if more TAD borders could be identified.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-022-04614-0.

Additional file 1. **Figure S1**. Comparison of betas between SIM with prefiltering by lasso regression and SIM without prefiltering. **Figure S2**. Comparison of TAD borders identified by SIM for different normalizations of the Hi-C data (Knight-Ruiz (KR)), iterative correction and eigenvector decomposition (ICE) and square root vanilla coverage (VC SQRT) at 50 kb resolution.

Acknowledgements
The author is grateful to Nicodemi's lab (INFN Sezione di Napoli, Italy) for Hi-C capture data and for providing PRISMR predictions. The author is also thankful to all the other labs that generated Hi-C and ChIP-seq data used in this article.

Authors' contributions
RM conceived and designed the project. RM implemented the model and analyzed the data. RM wrote the manuscript.

Funding
This work was supported by the University of Toulouse and the CNRS.

Availability of data and materials
An R package called "TADreg" was developed and is available at: https://github.com/raphaelmourad/TADreg.

Declaration

Abbreviations
Not applicable.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
I declare that the authors have no competing interests as defined by BMC, or other interests that might be perceived to influence the results and/or discussion reported in this paper.

Received: 19 October 2021 Accepted: 16 February 2022
Published online: 02 March 2022

References

1. Halverson JD, Smrek J, Kremer K, Grosberg AY. From a melt of rings to chromosome territories: the role of topological constraints in genome folding. Rep Progress Phys. 2014;77(2):022601.
2. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature. 2012;485(7398):376–80.
3. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G. Three-dimensional folding and functional organization principles of the Drosophila genome. Cell. 2012;148(3):458–72.
4. Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen C-A, Schmitt AD, Espinoza CA, Ren B. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature. 2013;503(7475):290–4.
5. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science. 2009;326(5950):289–93.
6. Pope BD, Ryba T, Dileep V, Yue F, Wu W, Denas O, Vera DL, Wang Y, Hansen RS, Canfield TK, Thurman RE, Cheng Y, Gulsoy G, Dennis JH, Snyder MP, Stamatoyannopoulos JA, Taylor J, Hardison RC, Kahveci T, Ren B, Gilbert DM. Topologically associating domains are stable units of replication-timing regulation. Nature. 2014;515(7527):402–5.

7. Zufferey M, Tavernari D, Oricchio E, Ciriello G. Comparison of computational methods for the identification of topologically associating domains. Genome Biol. 2018;19(1):217.
8. Crane E, Bian Q, McCord RP, Lajoie BR, Wheeler BS, Ralston EJ, Uzawa S, Dekker J, Meyer BJ. Condensin-driven remodelling of X chromosome topology during dosage compensation. Nature. 2015;523:240–4.
9. Rao SSP, Huntley MH, Durand NC, Stamenova EK, Bochkov ID, Robinson JT, Sanborn AL, Machol I, Omer AD, Lander ES, Aiden EL. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. Cell. 2014;159(7):1665–80.
10. Shin H, Shi Y, Dai C, Tjong H, Gong K, Alber F, Zhou XJ. TopDom: an efficient and deterministic method for identifying topological domains in genomes. Nucleic Acids Res. 2016;44(7): e70.
11. Levy-Leduc C, Delattre M, Mary-Huard T, Robin S. Two-dimensional segmentation for analyzing Hi-C data. Bioinformatics. 2014;30(17):i386–92.
12. Weinreb C, Raphael BJ. Identification of hierarchical chromatin domains. Bioinformatics. 2015;32(11):1601–9.
13. Serra F, Bau D, Goodstadt M, Castillo D, Filion GJ, Marti-Renom MA. Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors. PLoS Comput Biol. 2017;13(7):1–17.
14. Oluwadare O, Cheng J. ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data. BMC Bioinform. 2017;18(1):480.
15. Haddad N, Vaillant C, Jost D. IC-finder: inferring robustly the hierarchical organization of chromatin folding. Nucleic Acids Res. 2017;45(10):e81–e81.
16. Randriamihamison N, Vialaneix N, Neuvial P. Applicability and interpretability of Ward's hierarchical agglomerative clustering with or without contiguity constraints. J Classif. 2020.
17. Chen J, Hero AOI, Rajapakse I. Spectral identification of topological domains. Bioinformatics. 2016;32(14):2151–8.
18. Yan K-K, Lou S, Gerstein M. MrTADFinder: a network modularity based approach to identify topologically associating domains in multiple resolutions. PLoS Comput Biol. 2017;13(7):1–22.
19. Norton HK, Emerson DJ, Huang H, Kim J, Titus KR, Gu S, Bassett DS, Phillips-Cremins JE. Detecting hierarchical genome folding with network modularity. Nat Methods. 2018;15:119–22.
20. Zaborowski R, Wilczynski B. DiffTAD: detecting Differential contact frequency in topologically associating domains Hi-C experiments between conditions. bioRxiv. 2016.
21. Sadowski M, Kraft A, Szalaj P, Wlasnowolski M, Tang Z, Ruan Y, Plewczynski D. Spatial chromatin architecture alteration by structural variations in human genomes at the population scale. Genome Biol. 2019;20(1):148.
22. Cresswell KG, Dozmorov MG. TADCompare: an R package for differential and temporal analysis of topologically associated domains. Front Genet. 2020;11:158.
23. Bianco S, Lupiáñez DG, Chiariello AM, Annunziatella C, Kraft K, Schöpflin R, Wittler L, Andrey G, Vingron M, Pombo A, Mundlos S, Nicodemi M. Polymer physics predicts the effects of structural variants on chromatin architecture. Nat Genet. 2018;50(5):662–7.
24. Huynh L, Hormozdiari F. TAD fusion score: discovery and ranking the contribution of deletions to genome structure. Genome Biol. 2019;20(1):60.
25. Kaplan N. Explicit probabilistic models for exploiting and explaining the 3D genome. In: Proceedings of statistics for post genomic data (SMPGD 2019); 2019.
26. Belokopytova PS, Nuriddinov MA, Mozheiko EA, Fishman D, Fishman V. Quantitative prediction of enhancer-promoter interactions. Genome Res. 2020;30(1):72–84.
27. Rowley MJ, Nichols MH, Lyu X, Ando-Kuri M, Rivera ISM, Hermetz K, Wang P, Ruan Y, Corces VG. Evolutionarily conserved principles predict 3D chromatin organization. Mol Cell. 2017;67(5):837-852.e7.
28. Mourad R, Cuvier O. TAD-free analysis of architectural proteins and insulators. Nucleic Acids Res. 2018;46(5): e27.
29. Bonev B, MendelsonCohen N, Szabo Q, Fritsch L, Papadopoulos GL, Lubling Y, Xu X, Lv X, Hugnot J-P, Tanay A, Cavalli G. Multiscale 3D genome rewiring during mouse neural development. Cell. 2017;171(3):557–72.
30. Knight PA, Ruiz D. A fast algorithm for matrix balancing. IMA J Numer Anal. 2012.
31. The ENCODE Consortium. An integrated encyclopedia of DNA elements in the human genome. Nature. 2012;489(7414):57–74.
32. Dali R, Blanchette M. A critical assessment of topologically associating domain prediction tools. Nucleic Acids Res. 2017;45(6):2994–3005.
33. Dekker J, Marti-Renom MA, Mirny LA. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. Nat Rev Genet. 2013;14(6):390–403.
34. Hu M, Deng K, Selvaraj S, Qin Z, Ren B, Liu JS. HiCNorm: removing biases in Hi-C data via Poisson regression. Bioinformatics. 2012;28(23):3131–3.
35. Moore B, Aitken S, Semple C. Integrative modeling reveals the principles of multi-scale chromatin boundary formation in human nuclear organization. Genome Biol. 2015;16(1):110.
36. Mourad R, Cuvier O. Computational identification of genomic features that influence 3D chromatin domain formation. PLoS Comput Biol. 2016;12(5): e1004908.
37. Imakaev M, Fudenberg G, McCord RP, Naumova N, Goloborodko A, Lajoie BR, Dekker J, Mirny LA. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. Nat Methods. 2012;9(10):999–1003.
38. Durand NC, Shamim MS, Machol I, Rao SS, Huntley MH, Lander ES, Aiden EL. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. Cell Syst. 2016;3(1):95–8.
39. Crowley C, Yang Y, Qiu Y, Hu B, Abnousi A, Lipiński J, Plewczyński D, Wu D, Won H, Ren B, Hu M, Li Y. FIREcaller: detecting frequently interacting regions from Hi-C data. Comput Struct Biotechnol J. 2021;19:355–62.
40. Barber RF, Candès EJ. Controlling the false discovery rate via knockoffs. Ann Stat. 2015;43(5):2055–85.
41. Royston P, Sauerbrei W. Bootstrap assessment of the stability of multivariable models. Stand Genom Sci. 2009;9(4):547–70.

Publisher's Note

3.4.4 3D genome and evolution

In vertebrates, a large portion of chromatin loops is mediated by CTCF. The loops are often marked by asymmetric CTCF motifs where cohesin is recruited [Rao *et al.* 2014]. These results support the extrusion loop model where CTCF and cohesin act together to extrude unknotted loops during interphase [Sanborn *et al.* 2015].

CTCF is an 11-zinc-finger (ZF) protein that is functionally conserved in vertebrates and *Drosophila* melanogaster [Hore *et al.* 2008, Heger *et al.* 2012]. CTCF-binding sites and Hox gene clusters were shown to be closely correlated throughout the animal kingdom, suggesting the conservation of the Hox-CTCF link across the Bilateria, as principal organizer of bilaterian body plans [Heger *et al.* 2012]. Comparative Hi-C further showed that CTCF motif position and orientation are conserved across species and that divergence of CTCF binding is correlated with divergence of internal 3D domain structure [Vietri-Rudan *et al.* 2015]. These observations suggest that the genome could undergo a continuous flux of local conformation changes by CTCF motif turnover that allow or prevent the de novo enhancer-promoter interactions and misexpression [Gómez-Marín *et al.* 2015]. Thus, the comparative analysis of CTCF-mediated looping across species is crucial to understand how gene expression or other key processes evolve. However, 3D genome analysis relies on complex and costly Hi-C experiments, which currently limits their use for evolutionary studies over a large number of species.

I proposed a novel approach to study the 3D genome evolution in vertebrates using the genome sequence only, e.g. without the need for Hi-C data [Mourad 2019]. The approach is simple and relies on comparing the distances between convergent and divergent CTCF motifs (ratio 3DR, Equation 1 and Figure 1, from the article "Studying 3D genome evolution using genomic sequence" below). I showed that 3DR is a powerful statistic to detect CTCF looping encoded in the human genome sequence, thus reflecting strong evolutionary constraints encoded in DNA and associated with the 3D genome. Moreover, I found that 3DR varies depending on the chromosome region, such as 3D (sub-)compartments, suggesting that 3DR is not homogeneous along the genome and might functionally define 3D chromatin state. When comparing 3DR across vertebrates, the results revealed that the distance between convergent motifs which underlie CTCF looping and TAD organization evolves over time.

To conclude, I showed that the DNA sequence encodes loop extrusion, and that CTCF looping can be studied in species for which no Hi-C data are available, *e.g.* the majority of species. Moreover, I showed that phylogenetic methods such as ancestral character reconstruction can be used to infer CTCF looping in ancestral genomes. Therefore, 3DR makes it possible to study the evolution of CTCF looping across a large number of species, which is impossible with the Hi-C technique.

Bioinformatics, 36(5), 2020, 1367–1373
doi: 10.1093/bioinformatics/btz775
Advance Access Publication Date: 12 October 2019
Original Paper

Genome analysis

Studying 3D genome evolution using genomic sequence

Raphaël Mourad

LBCMCP, Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 31062 Toulouse, France

Associate Editor: John Hancock

Received on June 5, 2019; revised on October 3, 2019; editorial decision on October 4, 2019; accepted on October 8, 2019

Abstract

Motivation: The three dimensions (3D) genome is essential to numerous key processes such as the regulation of gene expression and the replication-timing program. In vertebrates, chromatin looping is often mediated by CTCF, and marked by CTCF motif pairs in convergent orientation. Comparative high-throughput sequencing technique (Hi-C) recently revealed that chromatin looping evolves across species. However, Hi-C experiments are complex and costly, which currently limits their use for evolutionary studies over a large number of species.

Results: Here, we propose a novel approach to study the 3D genome evolution in vertebrates using the genomic sequence only, e.g. without the need for Hi-C data. The approach is simple and relies on comparing the distances between convergent and divergent CTCF motifs by computing a ratio we named the 3D ratio or '3DR'. We show that 3DR is a powerful statistic to detect CTCF looping encoded in the human genome sequence, thus reflecting strong evolutionary constraints encoded in DNA and associated with the 3D genome. When comparing vertebrate genomes, our results reveal that 3DR which underlies CTCF looping and topologically associating domain organization evolves over time and suggest that ancestral character reconstruction can be used to infer 3DR in ancestral genomes.

Availability and implementation: The R code is available at https://github.com/morphos30/PhyloCTCFLooping.

Contact: raphael.mourad@univ-tlse3.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Chromosomes are tightly packed in three dimensions (3D) such that a 2-m long human genome can fit into a nucleus of ~10 microns in diameter (Halverson *et al.*, 2014). Over the past years, the 3D chromosome structure has been comprehensively explored by chromosome conformation capture combined with high-throughput sequencing technique (Hi-C) at an unprecedented resolution (Dixon *et al.*, 2012; Jin *et al.*, 2013; Sexton *et al.*, 2012). Multiple hierarchical levels of genome organization have been uncovered. Among them, topologically associating domains (TADs) (Dixon *et al.*, 2012; Sexton *et al.*, 2012) and chromatin loops (Rao *et al.*, 2014) represent pervasive structural features of the genome organization. Moreover, functional studies revealed that spatial organization of chromosomes is essential to numerous key processes such as for the regulation of gene expression by distal enhancers (Jin *et al.*, 2013; Lupiáñez *et al.*, 2015) or for the replication-timing program (Pope *et al.*, 2014).

A growing body of evidence supports the role of insulator binding proteins such as CTCF, and cofactors like cohesin, as mediators of long-range chromatin contacts (Phillips-Cremins *et al.*, 2013; Sexton *et al.*, 2012; Van Bortle *et al.*, 2014). In mammals, depletions of CTCF and cohesin decreased chromatin contacts (Zuin *et al.*, 2014). Moreover, high-resolution Hi-C mapping has recently revealed that loops that demarcate domains were often marked by asymmetric CTCF motifs where cohesin is recruited (Rao *et al.*, 2014). These results support the extrusion loop model where CTCF and cohesin act together to extrude unknotted loops during interphase (Sanborn *et al.*, 2015).

CTCF is an 11-zinc-finger protein that is functionally conserved in vertebrates and *Drosophila melanogaster* (Heger *et al.*, 2012; Hore *et al.*, 2008). CTCF-binding sites and Hox gene clusters were shown to be closely correlated throughout the animal kingdom suggesting the conservation of the Hox-CTCF link across the Bilateria, as principal organizer of bilaterian body plans (Heger *et al.*, 2012). Comparative Hi-C further showed that CTCF motif position and orientation are conserved across species and that divergence of CTCF binding is correlated with divergence of internal domain structure (Vietri-Rudan *et al.*, 2015). These observations suggest that the genome could undergo a continuous flux of local conformation changes by CTCF motif turnover that allow or prevent *de novo* enhancer–promoter interactions and misexpression (Gómez-Marín *et al.*, 2015). Thus, the comparative analysis of CTCF-mediated looping across species is crucial to understand how gene expression or other key processes evolve. However, 3D genome analysis relies on complex and costly Hi-C experiments, which currently limits their use for evolutionary studies over a large number of species.

Here, we propose a novel approach to study the 3D genome evolution in vertebrates using the genome sequence only, e.g. without

the need for Hi-C data. Therefore, this approach allows a comprehensive analysis of vertebrate 3D genomes whose number is exponentially increasing due to ongoing large sequencing projects such as the Vertebrate Genomes Project (VGP). The approach is simple and relies on comparing the distances between convergent and divergent CTCF motifs (using a ratio we named the 3D ratio or '3DR'). We show that 3DR is a powerful statistic to detect CTCF looping encoded in the human genome sequence, thus reflecting strong evolutionary constraints encoded in DNA and associated with the 3D genome organization. Moreover, we found that 3DR varies depending on the chromosome region, such as 3D (sub-)compartments, suggesting that 3DR is not homogeneous along the genome and might functionally define 3D chromatin state. When comparing 3DR across vertebrates, our results reveal that the distance between convergent motifs which underly CTCF looping and TAD organization evolves over time and suggest that ancestral character reconstruction can be used to infer 3DR in ancestral genomes.

2 Materials and methods

2.1 Hi-C data, compartments, subcompartments and TADs

In human, we computed compartments A/B using Juicer Tools (Durand *et al.*, 2016). For this purpose, we used publicly available Hi-C data from GM12878 cells from Gene Expression Omnibus (GEO) accession GSE63525 (Rao *et al.*, 2014). For subcompartments, we downloaded the genomic coordinates from GEO GSE63525. For TAD borders and loop anchors, we downloaded respectively Arrowhead domains and HiCCUPS loops called from GM12878 Hi-C data from GEO GSE63525.

2.2 Isochores

In human, we called isochores using isoSegmenter program on hg38 assembly (Cozzi *et al.*, 2015).

2.3 Replication timing

In human, we used GM12878 Repli-seq from ENCODE (The ENCODE Consortium, 2012).

2.4 CTCF motif calling

We used the vertebrate CTCF motif position frequency matrix MA0139.1 from the JASPAR database (http://jaspar. genereg.net/). We scanned CTCF binding sites on the following genome assemblies: ailMel1, allMis1, anoCar2, apiMel2, aplCal1, aptMan1, balAcu1, bosTau8, braFlo1, calJac3, calMil1, canFam3, cavPor3, ce11, cerSim1, choHof1, criGri1, danRer10, dipOrd1, dm6, droYak2, echTel2, equCab2, eriEur2, felCat8, fr3, gadMor1, galGal4, gasAcu1, geoFor1, gorGor3, hetGla2, hg38, latCha1, loxAfr3, macEug2, melGal1, melUnd1, micMur2, mm10, monDom5, musFur1, myoLuc2, nomLeu3, ochPri3, oreNil2, ornAna2, oryCun2, oryLat2, otoGar3, oviAri3, panPan1, panTro5, papAnu2, petMar2, ponAbe2, proCap1, pteVam1, rheMac3, rn6, saiBol1, sarHar1, sorAra2, speTri2, strPur2, susScr3, taeGut2, tarSyr2, tetNig2, triMan1, tupBel1, turTru2, vicPac2, xenTro7. For this purpose, we used MEME FIMO program with default parameters (http://meme-suite.org/doc/fimo.html).

2.5 CTCF ChIP-seq peak

In human, we used CTCF ChIP-seq peaks for several cell lines from ENCODE (https://genome.ucsc.edu/encode/).

2.6 Deepbind

To improve binding predictions for CTCF, we used deepbind to predict binding on the 500 base region surrounding motif occurrence (http://tools.genes.toronto.edu/deepbind/). We used the deepbind model trained on CTCF ChIP-seq data, noted D00328.018.

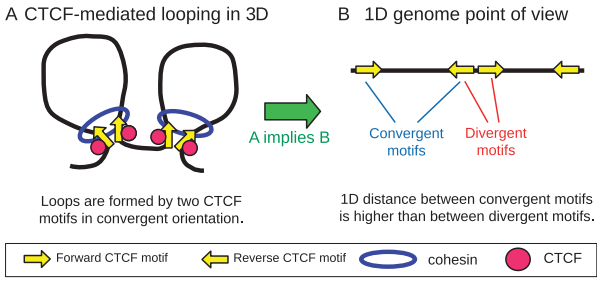


Fig. 1. CTCF-mediated looping in 3D and 1D genome points of view. (**A**) The CTCF-mediated looping in 3D. (**B**) The 1D genome point of view of CTCF-mediated looping

2.7 Conservation score

We computed the average conservation score of the 50 bases surrounding the CTCF binding sites using hg38 phastCons scores from UCSC Genome Browser (https://genome.ucsc.edu/). For other assemblies, we liftovered hg38 phastCons scores.

3 Results and discussion

3.1 CTCF-mediated looping in 3D and 1D genome point of view

In vertebrates, the 3D genome is organized in chromatin loops often mediated by CTCF and cohesin: the CTCF-mediated loops. In particular, CTCF sites at loop anchors occur predominantly (>90%) in a convergent orientation, i.e. with a forward motif on the left anchor and a reverse motif on the right anchor (Rao *et al.*, 2014) (Fig. 1A). From a 1D genome point of view, the CTCF-mediated looping implies that two motifs in convergent orientation should be located farther apart than two motifs in divergent orientation (Fig. 1B). Thus, based on this implication, we sought to compare the distances between contiguous motifs depending on their orientation as a mean to study 3D genome from genomic sequence in species for which Hi-C data were not available.

For this purpose, we estimated the following ratio 3DR:

$$3DR = median(d_{\rightarrow\leftarrow})/median(d_{\leftarrow\rightarrow}), \quad (1)$$

that was the ratio of two medians: the median of the distances between two contiguous motifs in convergent orientation (noted '→←'), and the median of the distances between two contiguous motifs in divergent orientation (noted '←→'). We hypothesized that a 3DR significantly greater than one reflects CTCF looping in the genome. Because 3DR was a ratio of distance medians, it accounted for the genome size effect and could thus allow comparisons between different genomes whose sizes may vary.

Additionally, we estimated another ratio used as a control:

$$3DC = median(d_{\rightarrow\rightarrow})/median(d_{\leftarrow\leftarrow}) \quad (2)$$

that was the ratio of two medians: the median of the distances between two contiguous motifs in the same forward orientation (noted '→→') and the median of the distances between two contiguous motifs in same reverse orientation (noted '←←'). Following the 1D genome point of view, the control ratio was supposed to show no difference between the two orientations. Deviations of 3DC from 1 might reflect biases in the genome that were not related to CTCF looping. To assess the significance of ratio 3DR (and 3DC), we used the Wilcoxon rank-sum test. This test could assess differences of distances even if the distances did not follow a normal distribution.

3.2 Validation of 3DR as a measure of CTCF-mediated looping

We first studied the ratio 3DR using the human genome. For this purpose, the human genome hg38 assembly was used and vertebrate CTCF motifs (JASPAR MA0139.1) were called along the genome.

The distance between any two consecutive motifs was computed. To only keep motifs with a higher chance of binding, motifs whose binding scores were lower than a specific quantile threshold were removed. We found that 3DR strongly increased with the binding score and was maximal for a quantile threshold of 80% (Fig. 2A). However, the confidence interval of 3DR was higher for 80% than for lower quantiles, because too many binding sites were discarded. Thus, as a trade-off, a quantile of 70% was then considered as a threshold for further analyses, because it better allowed comparison of 3DR between species with sufficient statistical power (statistical power depends on the number of binding sites).

We found that the distance between two contiguous motifs in convergent orientation was significantly higher than between two contiguous motifs in divergent orientation, as expected by the 1D genome point of view of CTCF-mediated looping (3DR = 1.28, Wilcoxon test $P < 3 \times 10^{-17}$; Fig. 2B). The 3DR was computed based on 6426 convergent motif pairs and on 6370 divergent motif pairs. In comparison, the distance between two motifs in forward orientation was not significantly different from the distance between two motifs in reverse orientation, as expected by the 1D genome point of view (3DC = 0.97, $P = 0.41$). The bootstrapped distributions of the distance medians were also computed for convergent and divergent motifs, respectively (Supplementary Fig. S1). The two distributions were far apart, reflecting the significant differences of medians. Because the accuracy of the distance between motifs depended on the genome assembly, the ratio was assessed for old and more recent assemblies. As expected, 3DR increased with recent assemblies (Supplementary Fig. S2). However, these improvements were very modest, revealing that the assembly version did not have a big impact on the estimation of 3DR in human.

We then used CTCF GM12878 ChIP-seq data to remove motifs not bound by CTCF *in vivo*. The ratio 3DR was much higher than previously and very significant (3DR = 1.69, $P < 5 \times 10^{-51}$; Fig. 2C), reflecting the important difference in distance between motifs overlapping CTCF peaks depending on orientation. *In vivo* information thus helped us to remove false positive motif occurrences and to estimate 3DR with more power. We next assessed 3DR using CTCF peaks from all ENCODE cell lines (Supplementary Table S1). Interestingly, we found that 3DR varied depending on cell type. Moreover, 3DR was especially low for embryonic stem cells and cancer cells, reflecting lower CTCF looping and thus lower organization of the genome in 3D domains in these cells. However, in practice, only genome assemblies were available for most species and no ChIP-seq data were available. Hence, to circumvent this issue, CTCF ChIP-seq peaks surrounding the motifs were predicted using convolutional neural network learned from human data (Alipanahi *et al.*, 2015). This ratio estimated using predicted peaks was noted $3DR_p$. The ratio $3DR_p$ was higher than the one computed from motifs only ($3DR_p = 1.44$, $P < 2 \times 10^{-34}$; Fig. 2D), revealing the better ratio estimation using peak prediction.

We next filtered motifs located inside 3D domain borders, since those motifs were more likely to influence the 3D genome. For this purpose, we used Arrowhead domains from GM12878 Hi-C data (Rao *et al.*, 2014). We extended domain borders to 20 kb on each side and only kept motifs belonging to borders. Accounting for 3D domain borders strikingly improved 3DR (3DR = 5.67, $P < 7 \times 10^{-27}$; Fig. 2E). We also filtered motifs located at loop anchors (Rao *et al.*, 2014). Again, we extended loop anchors to 20 kb on each side and kept motifs belonging to anchors. Surprisingly, we found a much lower 3DR than for 3D domain borders (3DR = 1.77, $P < 2 \times 10^{-15}$; Supplementary Fig. S3).

CTCF binding sites located at 3D domain borders were previously shown to be evolutionary conserved (Vietri-Rudan *et al.*, 2015). Hence, we sought to improve 3DR computation by discarding non-conserved motifs. This ratio estimated using conservation was noted $3DR_c$. This approach greatly improved the ratio ($3DR_c = 1.64$, $P < 7 \times 10^{-44}$; Fig. 2F). If both conservation and predicted peaks were used together, the ratio was even higher ($3DR_c = 1.80$, $P < 5 \times 10^{-52}$). We also computed 3DR within synteny blocks, but only observed a slight improvement (3DR = 1.30, $P < 3 \times 10^{-9}$; Supplementary Fig. S4). Thus, accounting for conservation score allowed to further improved ratio estimation.

As a control, we computed 3DR for *Drosophila* genomes (*melanogaster* and *yakuba*) and *Caenorhabditis elegans*. In *D.melanogaster*, recent high resolution Hi-C data showed the absence of loops mediated by CTCF motifs in convergent orientation (Eagen *et al.*, 2017). Accordingly, 3DR was computed for *melanogaster* and *yakuba* genomes and were close to one and not significant (dm6: 3DR = 0.93, $P = 0.15$; droYak2: 3DR = 1.02, $P = 0.41$;

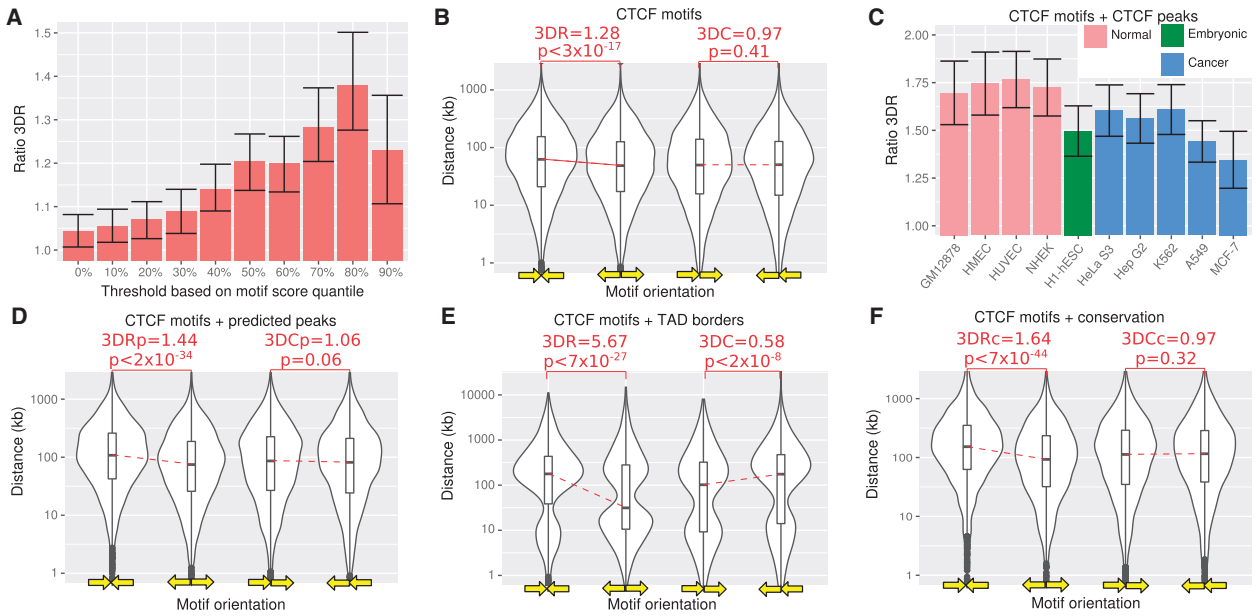


Fig. 2. Ratios 3DR and 3DC computed from the human genome assembly. (**A**) Ratio 3DR for different binding score thresholds. (**B**) Distance between consecutive CTCF motifs depending on motif orientation. (**C**) Ratio 3DR when accounting for CTCF ChIP-seq data for different cell lines. (**D**) Distance between consecutive CTCF motifs depending on motif orientation, when accounting for predicted CTCF ChIP-seq data. (**E**) Distance between consecutive CTCF motifs depending on motif orientation, when accounting for TAD borders. (**F**) Distance between consecutive CTCF motifs depending on motif orientation, when accounting for conservation score

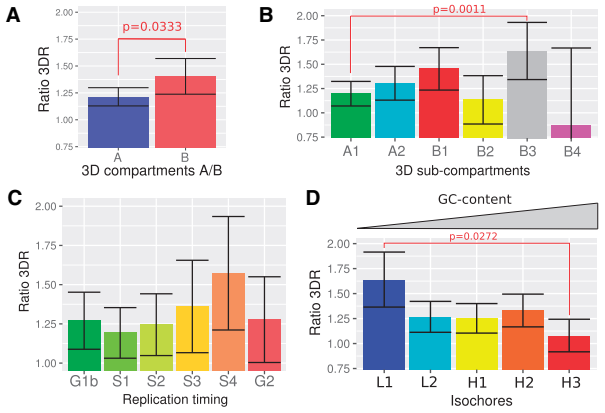


Fig. 3. Ratio 3DR computed for different chromatin regions in human. (**A**) Ratio 3DR estimated for 3D genome compartments A/B. (**B**) Ratio 3DR depending on 3D genome subcompartments. (**C**) Ratio 3DR and replication timing. (**D**) Ratio 3DR depending on GC-content isochores

Supplementary Fig. S5). In addition, in *C.elegans*, CTCF has been lost during nematode evolution (Heger *et al.*, 2009). In agreement, 3DR was also close to one and not significant (ce11: 3DR = 1.02, $P = 0.22$; Supplementary Fig. S5).

Analysis of the human genome thus validated the 1D genome point of view of CTCF-mediated looping. Such looping can be easily estimated from the genomic sequence alone by computing the 3DR ratio of distances depending on motif orientation. Moreover, control results revealed the ability of 3DR to be equal to one for genomes that are known not to harbor CTCF-mediated loops.

3.3 Ratio 3DR varies with 3D compartments and isochores

We then computed 3DR depending on the underlying genomic and chromatin regions in the human genome (Supplementary Table S2). We first investigated if 3DR could differ depending on megabase 3D genome compartments, known as A/B compartments, that were shown to divide the genome into gene rich, active and open chromatin (compartment A) and into gene poor, inactive and close chromatin (compartment B) (Lieberman-Aiden *et al.*, 2009). We found that 3DR was greater in compartment B (3DR = 1.40, $P < 3 \times 10^{-8}$) than in compartment A (3DR = 1.21, $P < 3 \times 10^{-8}$; Fig. 3A), with a slightly significant difference ($P = 0.03$). Accordingly, chromatin loops were larger in compartment B than in compartment A (fold-change = 1.4, $P < 1 \times 10^{-20}$; Supplementary Fig. S6). At high resolution (25 kb), compartments A/B were further shown to be composed of subcompartments A1, A2 (active) and B1, B2, B3, B4 (inactive) (Rao *et al.*, 2014). We found that 3DR varied between subcompartments. Subcompartments A1 and A2 presented 3DR values close to the 3DR computed genome-wide (A1: 3DR = 1.20, $P < 4 \times 10^{-4}$; A2: 3DR = 1.30, $P < 4 \times 10^{-4}$; Fig. 3B). Conversely, B subcompartments showed high variability of 3DR. B1 and B3 showed 3DR values greater than the genome-wide 3DR (B1: 3DR = 1.45, $P < 4 \times 10^{-6}$; B3: 3DR = 1.64, $P < 2 \times 10^{-10}$; Fig. 3B), while B2 and B4 had 3DR values that were lower than the genome-wide 3DR (B2: 3DR = 1.13, $P = 0.27$; B4: 3DR = 0.87, $P = 0.76$; Fig. 3B). When comparing A and B subcompartments, we found a significant difference between A1 and B3 ($P = 0.0011$). We next analyzed 3DR depending on DNA replication timing. We found a 3DR value close to the genome-wide value for early replicating regions (3DR = 1.27, $P < 2 \times 10^{-4}$; Fig. 3C), but a high 3DR value for late S replicating regions (3DR = 1.57, $P < 2 \times 10^{-5}$; Fig. 3C).

Another important feature of the genome is the GC-content that varies considerably along the chromosomes. In particular, the genome was shown to be composed of isochores which are large DNA segments of homogeneous GC-content (Costantini *et al.*, 2006) and that were recently shown to be correlated with subcompartments (Jabbari and Bernardi, 2017). We then computed 3DR depending on isochore class (L1, L2, H1, H2 and H3) and observed differences between classes. In particular, L1 isochores (lowest GC-content) showed the highest 3DR value (3DR = 1.64, $P < 5 \times 10^{-7}$; Fig. 3D), which was considerably larger than the one estimated genome-wide. Interestingly, L1 3DR value was very close to subcompartment B3 3DR value. Conversely, H3 isochores (highest GC-content) showed the lowest 3DR value (3DR = 1.08, $P = 0.15$; Fig. 3D), which was lower than the genome-wide 3DR.

The 3DR ratio thus varied with the underlying genomic and chromatin context. Most notably, we found that 3DR was higher in compartment B, in mid-late replication timing regions and in low GC-content isochores, which were associated with heterochromatin.

3.4 CTCF looping in mammals

We then estimated 3DR for available mammal genomes. Because the accuracy of 3DR estimation depended on the number of motif pairs, we computed 3DR for genomes with a sufficient number of pairs (>8000). We found that all mammals presented a 3DR value that was superior to one and significant (Fig. 4A; Supplementary Table S3). The Tasmanian devil and the pika presented the highest values (3DR > 1.5), whereas the horse and the guinea pig showed the lowest values (3DR close to 1.2). It was very interesting to see that 3DR estimation could be significantly different from one even for assemblies whose qualities were much lower than hg38, such as papAnu2 (scaffold N_{50} = 586 kb, scaffold L_{50} = 1481; 3DR = 1.37, $P < 9 \times 10^{-22}$) and ornAna2 (scaffold N_{50} = 959 kb, scaffold L_{50} = 309; 3DR = 1.44, $P < 7 \times 10^{-18}$).

We also predicted CTCF ChIP-seq peaks surrounding the motifs, and estimated $3DR_p$. The ratio $3DR_p$ was superior to 3DR estimated from motifs only (Fig. 4A; Supplementary Table S4). Interestingly, although the convolutional neural network we used was trained from human data, it could dramatically increase the ratio for most species. For instance, $3DR_p$ was higher than 3DR for the dog (canFam3: 3DR = 1.20, $3DR_p$ = 1.49, 24% increase) and even for the platypus (ornAna2: 3DR = 1.44, $3DR_p$ = 1.68, 17% increase). We also filtered conserved motifs and computed $3DR_c$ (Fig. 4A; Supplementary Table S5). The ratio $3DR_c$ was even higher than $3DR_p$ for most species. For example, $3DR_c$ were higher than 3DR and $3DR_p$ for the dog (canFam3: 3DR = 1.20, $3DR_c$ = 1.79, 49% increase) and the platypus (ornAna2: 3DR = 1.44, $3DR_p$ = 1.99, 38% increase). However, a major drawback of $3DR_c$ and $3DR_p$ was their larger confidence intervals, and that is the reason why we kept 3DR for further analyses.

We next investigated if 3DR was influenced by the genome size, which could explain the observed differences of 3DR between species. No significant correlation was found between the genome size and 3DR (Fig. 4B), confirming that 3DR was not biased by the genome size, and thus allowing 3DR comparison between species. No significant correlation was also found with the median chromosome size (Supplementary Fig. S7). We also assessed if 3DR was influenced by the density of motifs in the genome (number of motifs per Mb), and no significant correlation was found (Fig. 4C). For instance, the platypus and rat genomes presented a 3DR value around 1.45, but contained 4.66 motifs per Mb and 13.33 motifs per Mb, respectively. Moreover, we found no link between 3DR and GC-content between mammals (Supplementary Fig. S8A).

The 3DR ratio can thus be used to study the 3D genome organization in CTCF loops in mammals even for species whose Hi-C data were not available. Moreover, we found important differences of 3DR between mammals. For instance, we found that species that were evolutionary distant, such as the human and the Tasmanian devil, presented an important difference of 3DR.

3.5 Phylogenetic analysis of CTCF looping in vertebrates

We then estimated 3DR for vertebrate species in order to investigate differences between mammals, reptiles, amphibians and fishes. As for mammals, we found no link between 3DR and genome size or motif density among vertebrates (Supplementary Fig. S9). However, we observed a weak but significant link between 3DR and

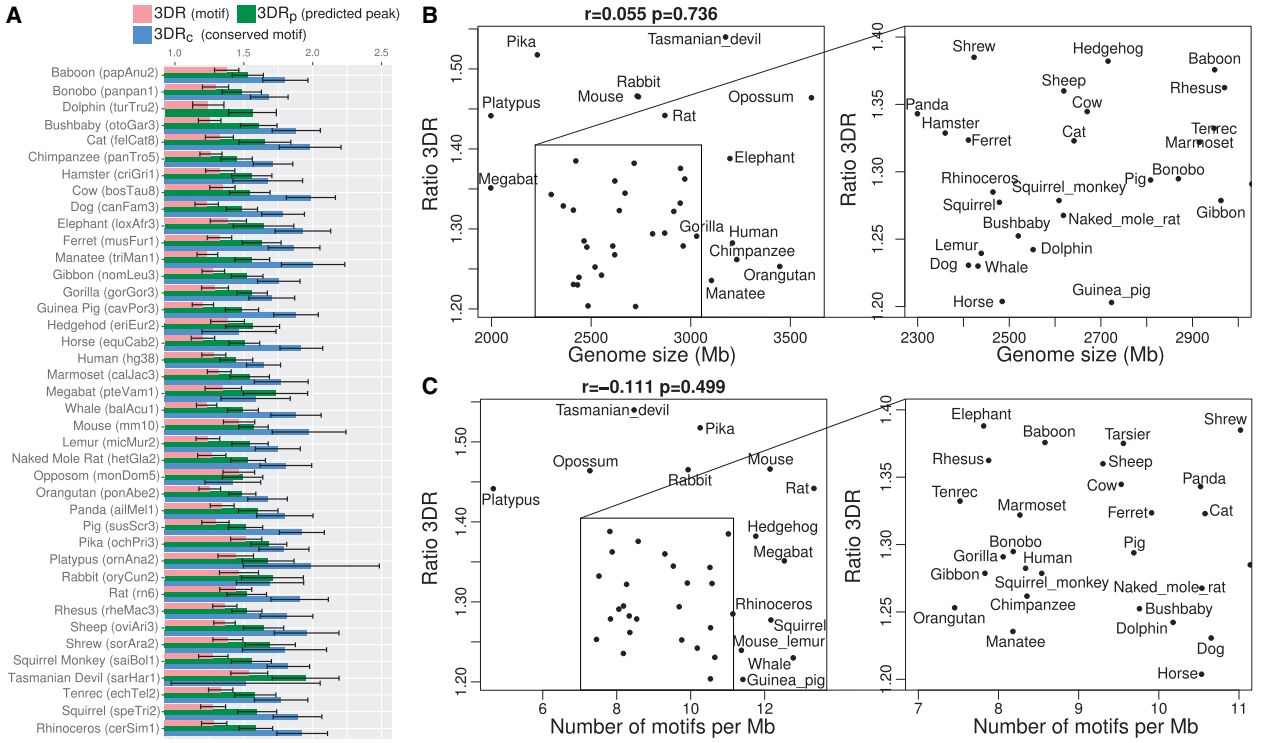


Fig. 4. Ratio 3DR computed from mammal genomes. (**A**) Ratios 3DR, $3DR_p$ and $3DR_c$ computed from all mammal genome assemblies. For each assembly, 3DR is plotted at the top, $3DR_p$ in the middle and $3DR_c$ at the bottom. (**B**) Ratio 3DR versus genome size. (**C**) Ratio 3DR versus motif density (number of motifs per Mb). *Note:* Chinese hamster was not plotted, since it was a very strong outlier (21.3 motifs per Mb)

GC-content ($r = 0.346$, $P = 0.007$; Supplementary Fig. S8B and C). Ratios 3DR were next plotted on the phylogenetic tree to investigate the potential link between CTCF looping and evolution (Fig. 5). Among the vertebrates, most jaw fishes presented very high 3DR values, especially the tetraodon (tetNig2: 3DR = 1.65, $P < 2 \times 10^{-27}$) and fugu (fr3: 3DR = 1.59, $P < 4 \times 10^{-60}$). The zebrafish instead presented a low 3DR = 0.98, which was inconsistent with recent Hi-C results supporting loop formation by CTCF in convergent orientation (Kaaij *et al.*, 2018). This low 3DR in zebrafish could be related to its low genome GC-content, compared to the tetraodon and fugu presenting both high 3DR and GC-content (Supplementary Fig. S8B). In addition, the amphibian *Xenopus* showed a very high 3DR value (xenTro7: 3DR = 1.63, $P < 3 \times 10^{-79}$). Interestingly, using peak prediction models trained on human data, the $3DR_p$ values were even higher: tetraodon ($3DR_p = 1.84$, $P < 8 \times 10^{-90}$) and *Xenopus* ($3DR_p = 1.85$, $P < 2 \times 10^{-24}$). Lampreys which are jawless fishes that diverged from the jawed vertebrate lineage more than 500 million years ago also revealed a significant ratio (3DR = 1.30, $P < 7 \times 10^{-9}$), supporting the ancient establishment of CTCF looping prior to vertebrates (Heger *et al.*, 2012).

The different assemblies did not have the same quality, which thus introduced some inaccuracy in the estimation of 3DR, especially for species that were recently sequenced (those with an assembly number close to one). Despite 3DR inaccuracy due to heterogeneous assembly quality, we found that evolutionary close species tended to have a similar 3DR value (Mantel test $P = 5 \times 10^{-5}$), revealing conservation of 3DR among species (Fig. 5). For instance, two relatively close species in the tree, the rat (rn6: $3DR_p = 1.44$, $P < 1 \times 10^{-27}$) and the mouse (mm10: $3DR_p = 1.47$, $P < 3 \times 10^{-36}$) presented very similar 3DR values ($P = 0.61$). Hence, ancestral 3DR reconstruction could be carried out (Fig. 5). It revealed that a large 3DR value was acquired in the common ancestor of the rat and the mouse (Supplementary Fig. S10). Similar findings were observed for the American pika (ochPri3) and the European rabbit (oryCun2), and also for the Tasmanian devil (sarHar1) and the opossum (monDom5).

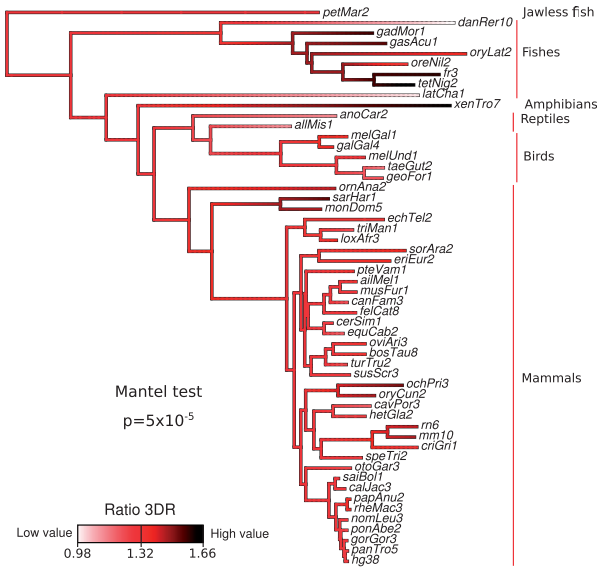


Fig. 5. Phylogenetic analysis of 3DR in vertebrates. Ancestral 3DR reconstruction was done using maximum likelihood inference

Another important parameter contributing to CTCF looping is the CTCF motif density in bilaterian genomes (Heger *et al.*, 2012). Hence, we estimated CTCF motif density in vertebrates and observed a strong conservation (Mantel test $P < 1 \times 10^{-5}$; Supplementary Fig. S11). Jaw fishes showed high motif densities, such as the fugu (fr3: 40.58 motifs/Mb). Conversely, birds showed very low motif densities, such as the chicken (galGal4: 7.05 motifs/Mb). Mammals presented varying densities, for instance 4.66 for the

platypus (ornAna2) and 21.4 for the Chinese hamster (criGri1). Among mammals, we observed very homogeneous clades, such as primates, whose motif density varied from 7.45 to 9.76. Moreover, we found that CTCF motif density was evolutionary conserved (Mantel test $P < 1 \times 10^{-5}$), which suggested that ancestral motif density reconstruction could be done. Inference of ancestral density uncovered interesting results, such as the low density for the primate ancestor as compared to the higher density for the muridae ancestor.

Results revealed the evolutionary conservation of 3DR among vertebrates. 3DR thus represented a useful tool to study 3D genome evolution, in addition to CTCF motif density. The two parameters could be used to study CTCF looping in ancestral genomes by using ancestral character reconstruction.

4 Conclusion

In this article, we propose a novel approach to study the 3D genome evolution in vertebrates using the genomic sequence only, without the need of costly and challenging Hi-C data to produce. Therefore, the approach allows a comprehensive analysis of vertebrates whose genome assemblies are now available and whose number will exponentially increase with large sequencing projects such as the VGP aiming to sequence 66 000 extant vertebrate species. The proposed approach is very simple and makes very few assumptions. It relies on the CTCF motif which is known to be conserved across vertebrates and the CTCF looping model that implies a 1D genome point of view where convergent motifs are expected to be more distant than divergent motifs. The approach can be further improved by using predicted CTCF ChIP-seq peaks or by using the conservation score surrounding the CTCF motif, reflecting strong conservation of the DNA context surrounding CTCF motifs in vertebrates, especially for mammals. Using the human genome as a reference, we validate the 1D genome point of view and demonstrate that the ratio of distances between convergent and divergent motif pairs (ratio 3DR) can assess the presence of CTCF looping. These results reflect strong evolutionary constraints encoded in the genome that are associated with the 3D genome organization.

The proposed approach also uncovers a number of results. We found that 3DR varies with the underlying genomic and chromatin regions, such as 3D compartments and sub-compartments, isochores and replication timing. Moreover, the analysis of 3DR combined with CTCF ChIP-seq peaks showed a lower value for 3DR in cancer and embryonic cells compared to normal cell lines. Thus, depending on the cell state, 3DR can be modulated by CTCF binding *in vivo*, thereby regulating CTCF looping. Regarding 3DR in different species, we show most notably that 3DR is evolutionary conserved among vertebrates. Species that are phylogenetically close tend to have a ratio that is closer than species that are phylogenetically far. Among vertebrates, several fishes and amphibians show the highest ratio, whereas reptiles show low values. In mammals, ancestral character reconstruction reveals that the genome of the ancestor of the rat and mouse likely evolved to have a high 3DR value. A previous study showed the linear divergence of CTCF binding sites with evolutionary distance, and the birth of new genes associated with the birth of new CTCF binding sites (Ni *et al.*, 2012). Here, our approach suggests that the distance between convergent motifs which underlies CTCF looping and TAD organization evolves over time between vertebrates, and thus further reinforces the notion that it represents an important factor contributing to 3D genome evolution.

There are several limitations of the proposed approach. First, we could not identify any 3D genome feature such as TAD or loop size that correlates with 3DR differences observed between species, which might be due to the small number of available Hi-C datasets in different vertebrate species. Thus, 3DR differences between species, such as between the human and the mouse genomes, should be quantitatively interpreted with caution. Second, we find a non-significant 3DR value for the zebrafish (danRer10) which is in contradiction with recent Hi-C data (Kaaij *et al.*, 2018), thus revealing the inadequacy of 3DR for certain species. The positive link of 3DR with GC-content in vertebrates (and more particularly between jaw fishes) suggests that the low 3DR in zebrafish is related to its low genome GC-content, as compared to tetraodon and fugu which present both high 3DR and high GC-content. However, the link with GC-content is not strong and some species such as the Tasmanian devil or the opossum have a high 3DR with a low GC-content. Analysis of only high-quality vertebrate assemblies similarly reveals a positive but weak link (Supplementary Fig. S8C). If we use only CTCF motifs present in synteny blocks common between zebrafish and tetraodon, then a higher 3DR value is found although not significant (3DR = 1.15, $P = 0.22$). There are other reasons why 3DR might not robustly identify the presence of CTCF-mediated loops in some species, such as the zebrafish. For instance, it is possible that the high density of CTCF motifs (40 motifs per Mb) makes the estimation of 3DR less reliable, since most motifs are not used as loop anchors. It might also be difficult to accurately estimate 3DR for genomes with small domains, because it could make the distance difference between convergent and divergent motifs smaller. Another reason might be the contribution of other proteins in mediating loops, for instance YY1 or Polycomb (Schoenfelder *et al.*, 2015; Weintraub *et al.*, 2017). Third, 3DR can be underestimated due to false positive motifs, as the CTCF protein does not bind to all detected motifs *in vivo*. Fourth, the estimation of distances between CTCF motifs depends on the genome assembly quality. Thus, for draft genomes, it is likely that the 3DR ratio will not be accurately estimated, especially when scaffolds are small. Fifth, deep learning models can be used to improve 3DR for species without any available ChIP-seq data, but the models were learned from human data and thus CTCF peak prediction is expected to be less accurate for species that are very distant from human. Sixth, phylogenetic conservation of 3DR can be accurately assessed for species that are within the same clade (as primates or muridaes), or more generally evolutionary close. Conversely, it is difficult to assess phylogenetic transmission of 3DR for the lamprey, since we have only one sequenced genome within the clade. Seventh, the estimation of 3DR is less accurate when we focus on certain genomic regions in human, such as isochores or compartments. For instance, we find a value of 1.08 for H3 isochores, but this does not mean that CTCF-mediated loops are absent from those regions. In fact, the corresponding 95% confidence interval is very large (between 0.87 and 1.30), meaning that 3DR could not be estimated accurately due to a lack of statistical power, and precluding the detection of CTCF-mediated loops by 3DR.

Funding

This work was supported by the University of Toulouse and by the CNRS.

Conflict of Interest: none declared.

References

Alipanahi,B. *et al.* (2015) Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.*, **33**, 831–838.

Costantini,M. *et al.* (2006) An isochore map of human chromosomes. *Genome Res.*, **16**, 536–541.

Cozzi,P. *et al.* (2015) Segmenting the human genome into isochores. *Evol. Bioinformatics Online*, **11**, 253–261.

Dixon,J.R. *et al.* (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.

Durand,N.C. *et al.* (2016) Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.*, **3**, 95–98.

Eagen,K.P. *et al.* (2017) Polycomb-mediated chromatin loops revealed by a sub-kilobase resolution chromatin interaction map. *Proc. Natl. Acad. Sci. USA*, **114**, 8764–8769.

Gómez-Marín,C. *et al.* (2015) Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. *Proc. Natl. Acad. Sci. USA*, **112**, 7542–7547.

Halverson,J.D. *et al.* (2014) From a melt of rings to chromosome territories: the role of topological constraints in genome folding. *Rep. Prog. Phys.*, **77**, 022601.

Heger,P. *et al.* (2009) Loss of the insulator protein CTCF during nematode evolution. *BMC Mol. Biol.*, **10**, 84.

Heger,P. *et al.* (2012) The chromatin insulator CTCF and the emergence of metazoan diversity. *Proc. Natl. Acad. Sci. USA*, **109**, 17507–17512.

Hore,T.A. *et al*. (2008) The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes. *PLoS Genet*., **4**, e1000169–11.
Jabbari,K. and Bernardi,G. (2017) An isochore framework underlies chromatin architecture. *PLoS One*, **12**, e0168023–12.
Jin,F. *et al*. (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, **503**, 290–294.
Kaaij,L.J. *et al*. (2018) Systemic loss and gain of chromatin architecture throughout zebrafish development. *Cell Rep*., **24**, 1–10.e4.
Lieberman-Aiden,E. *et al*. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
Lupiáñez,D.G. *et al*. (2015) Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, **161**, 1012–1025.
Ni,X. *et al*. (2012) Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. *PLoS Biol*., **10**, e1001420–16.
Phillips-Cremins,J.E. *et al*. (2013) Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, **153**, 1281–1295.
Pope,B.D. *et al*. (2014) Topologically associating domains are stable units of replication-timing regulation. *Nature*, **515**, 402–405.
Rao,S.S.P. *et al*. (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
Sanborn,A.L. *et al*. (2015) Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proc. Natl. Acad. Sci. USA*, **112**, E6456–E6465.
Schoenfelder,S. *et al*. (2015) Polycomb repressive complex PRC1 spatially constrains the mouse embryonic stem cell genome. *Nat. Genet*., **47**, 1179–1186.
Sexton,T. *et al*. (2012) Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, **148**, 458–472.
The ENCODE Consortium. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
Van Bortle,K. *et al*. (2014) Insulator function and topological domain border strength scale with architectural protein occupancy. *Genome Biol*., **15**, R82.
Vietri-Rudan,M. *et al*. (2015) Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture. *Cell Rep*., **10**, 1297–1309.
Weintraub,A.S. *et al*. (2017) Yy1 is a structural regulator of enhancer-promoter loops. *Cell*, **171**, 1573–1588.e28.
Zuin,J. *et al*. (2014) Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells. *Proc. Natl. Acad. Sci. USA*, **111**, 996–1001.

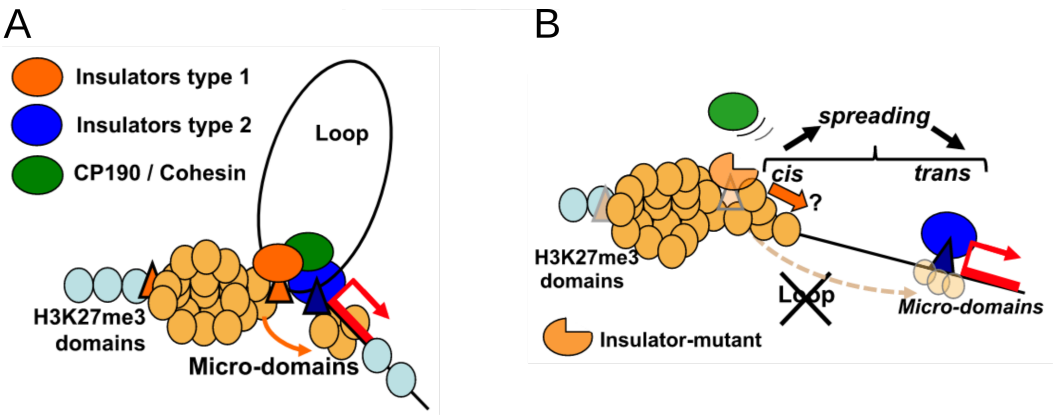


Figure 3.3: Insulator protein mutants impair H3K27me3 micro-domains depending on CP190 recruitment. A) Scheme representing the 3D-based formation of micro-domains involving the indicated molecular players of long-range interactions. B) Scheme representing the impact of BEAF-32 looping mutants on insulator-mediated LRIs by GAF /dCTCF and CP190 co-factors that results in both distant spreading onto micro-domains and (gain) in local spreading at borders.

3.4.5 3D genome and heterochromatin (Alexandre Heurteau)

Trimethylation of lysine 27 on histone H3 (H3K27me3) by the Polycomb 2 repressor complex (PRC2) is a feature of facultative heterochromatin associated with the repression of cell type specific genes [Cao *et al.* 2002, Morey & Helin 2010]. The faithful inheritance of the H3K27me3 chromatin marks by replication ensures the stability of the transcriptional silence mediated by PRC2 over cell generations, thus protecting cellular identities. H3K27me3 marks form repressive domains over the genome, where PRC2 writing and reading activities enable the spreading along the chromosome within domains. Insulators act as chromatin barriers to block the spreading outside repressive domains [modENCODE Consortium *et al.* 2010, Negre *et al.* 2010]. In drosophila, dCTCF, and other insulator-binding proteins such as BEAF-32, GAF and CP190, are specifically enriched at heterochromatin domain borders [Van Bortle *et al.* 2012, Van Bortle *et al.* 2014, Vogelmann *et al.* 2014]. Such proteins are also known to be involved in shaping the genome in 3D, which suggests a strong interplay between the formation of H3K27me3 domains and the genome in 3D.

Heurteau *et al.* analyzed the spreading of heterochromatin H3K27me3 marks depending on insulator-binding proteins and long-range interactions (LRIs) [Heurteau *et al.* 2020]. They showed that removal of insulator proteins BEAF-32 leads to H3K27me3 spreading locally, across borders (Figure 3.3). In addition, BEAF-32 promotes spreading onto distant euchromatin sites named "micro-domains". Systematic measurements of LRIs suggest that H3K27me3 micro-domains do not form due to the weakness of TAD borders. Rather, micro-domains

were visible at sites showing high levels of LRIs, including distant dCTCF and GAF insulator sites bound by the looping co-factor CP190. Also, micro-domain formation appears to depend on such specific insulator-mediated LRIs utilized to spread H3K27me3 to distant sites through looping. Supporting these results, specific synthetic mutants that impair LRIs compromise distant spreading over micro-domains. Distant spreading at micro-domains is further associated with insulator-based control of genes and it influences H3K27me3 throughout developmental stages of Drosophila. The data highlight how specific LRIs encoded by insulator-mediated loops contribute to the regulation of H3K27me3 spreading over the distance. Heurteau *et al.* propose that micro-domains reflect how insulators participate to chromatin folding dynamics in 3D, aside additional factors required to separate heterochromatin nano-compartments from nearby euchromatin domains.

3.4.6 3D genome and DNA double strand break repair

3.4.6.1 Loop extrusion as a mechanism for DSB repair foci formation (Vincent Rocher)

Among DNA damages, DNA double-strand breaks (DSBs) are by far the most deleterious, since they can lead to chromosome rearrangements [Marnef *et al.* 2017, Vitor *et al.* 2020]. There is a strong link between the genomic localization of DSBs and the chromatin environment [Lensing *et al.* 2016]. For instance, the DSB repair pathway choice between the two main pathways, non-homologous end joining (NHEJ) and homologous recombination (HR), depends on the chromatin landscape. HR tends to occur in transcriptionally active genes, as compared to NHEJ. In particular, the trimethylation of histone H3 on lysine 36 (H3K36me3), that correlates with elongating RNA Pol II, acts as a critical determinant for HR. However, little is known about the link between DSB repair and the 3D genome [Arnould & Legube 2020].

Using 4C and Hi-C experiments, Coline Arnould, Vincent Rocher *et al.* found that the histone mark γH2AX, which is induced by DSBs, was spread along the chromatin within domain boundaries that coincide with TAD boundaries [Arnould *et al.* 2021]. This result implied that the TAD is the functional unit of DSB repair (Figure 1 from the article "Loop extrusion as a mechanism for DNA double-strand breaks repair foci formation" below). Moreover, the recruitment of cohesin at the DSB site, and the emergence of stripes at the Hi-C matrix profile, revealed one-sided loop extrusion on both sides of the DSB, where DSB cohesin loading or fixation allowed the DSB locus to act as a loop anchor (Figure 2 from the article below). Coline Arnould, Vincent Rocher *et al.* found that the TAD structure remains globally unchanged, except stronger interactions between the DSB loci and its neighboring sequences. Such interaction increase was abolished in cohesin depleted cells, confirming the role of loop extrusion in this process (Figure 3 from the article below). During this process, the phosphorylated ATM (pATM), the enzyme recruited at the DSB and responsible for the phosphorylation of H2AX, was brought into physical proximity with the neighboring sequences. In the light of

these results, Coline Arnould, Vincent Rocher *et al.* proposed that the loop extrusion is responsible for spreading the gH2AX mark at the neighboring sequences by pATM recruited at the DSB loci (Figure 4 from the article below).

Article

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

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 (γH2AX), which act as seeds for the formation of DNA-damage response foci[1]. 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[2,3]. Here we show that topologically associating domains are functional units of the DNA damage response, and are instrumental for the correct establishment of γH2AX–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 (γH2AX, ubiquitin accumulation and histone H1 depletion) and the accumulation of DDR factors (53BP1 and MDC1)[4–6]. Previous evidence indicated that chromosome architecture may control the spread of γH2AX. Indeed, γH2AX domain boundaries were found in some instances to coincide with topologically associating domain (TAD) boundaries[7]. 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 γH2AX foci[8]. In addition, 53BP1 can form nanodomains that frequently overlap with TADs, as detected by DNA fluorescence in situ hybridization (DNA-FISH)[9]. High-resolution chromatin immunoprecipitation with sequencing (ChIP–seq) mapping after the induction of multiple DSBs at annotated positions (using human DIvA (DSB inducible via AsiSI) cells)[10] revealed that the spreading of these DDR focus components on nearby chromatin follows a highly stereotyped pattern[5] (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.

γH2AX 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 DIvA cells. As viewpoints we selected three genomic locations that are damaged in DIvA 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 γH2AX determined post DSB induction (Fig. 1a, b, Extended Data Fig. 1a), suggesting that initial chromosome architecture dictates γH2AX 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 γH2AX (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.

[1]Molecular, Cellular and Developmental Biology Unit (MCD), Centre de Biologie Intégrative (CBI), UPS, CNRS, Toulouse, France. [2]Rosenstiel Basic Medical Sciences Research Center and Department of Biology, Brandeis University, Waltham, MA, USA. [3]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. [4]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. [5]Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), Gif-sur-Yvette, France. ✉e-mail: gaelle.legube@univ-tlse3.fr

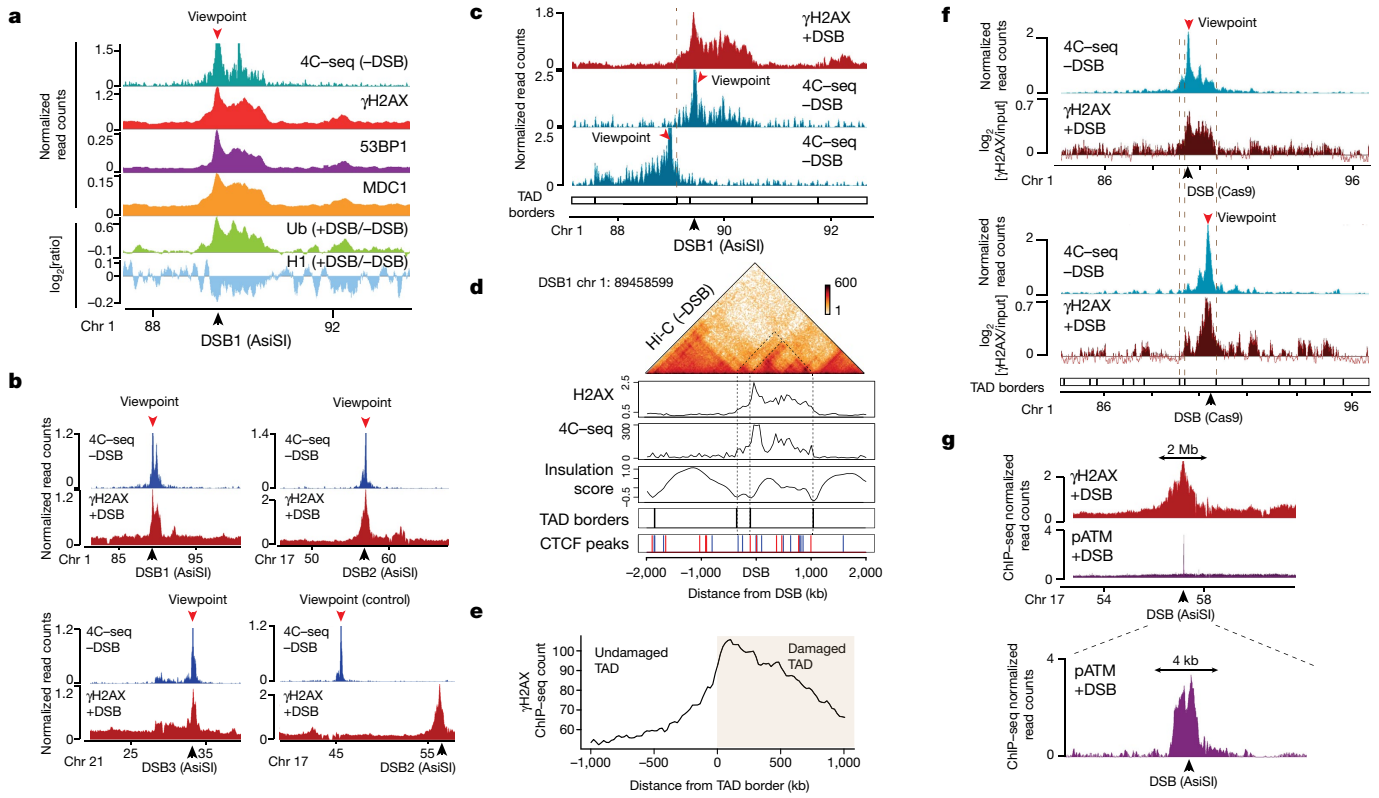


Fig. 1 | TADs are functional units that govern the establishment of DDR chromatin domains. a, 4C–seq track in undamaged cells (–DSB) and ChIP–seq tracks of histone H1 (H1.2) and ubiquitin (Ub; FK2) ($\log_2$(+DSB/–DSB)) as well as γH2AX, MDC1 and 53BP1 (+DSB) as indicated. ChIP–seq and 4C–seq data were smoothed using 50-kb and 10-kb spans, respectively. **b**, 4C–seq tracks before DSB induction (–DSB) and γH2AX 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**, γH2AX 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 DIvA cells before DSB induction. One representative experiment is shown (out of $n = 2$). Below, γH2AX 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 γH2AX 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, γH2AX ChIP–chip tracks ($\log_2$[sample/input], smoothed using 500-probe span) after DSB induction with CRISPR–Cas9. **g**, γH2AX 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 DIvA cells (Extended Data Fig. 1b-d). Notably, computed TAD borders and CTCF-bound genomic loci coincided with a sharp decrease in γH2AX signals (Fig. 1d, e, Extended Data Fig. 1e). Consistent with this, γH2AX, 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[11].

To further investigate whether TADs dictate γH2AX 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 γH2AX distribution. Cas9-induced DSBs recapitulated the γH2AX spreading observed when DSBs were induced at the same genomic locations by AsiSI (Extended Data Fig. 1g), thus confirming that γH2AX 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 γH2AX 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 γH2AX domains. Given that γH2AX 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 γH2AX upon inhibition of ATM[12] (Extended Data Fig. 1h–j) but not of DNAPK[12] or ATR (Extended Data Fig. 1i, j). To gain more insights into the mechanism that mediates the establishment of γH2AX 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 γH2AX (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[13,14], 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[15,16], 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[17–20]. However, cohesin enrichment at DSBs has been identified throughout the cell cycle, which argues against

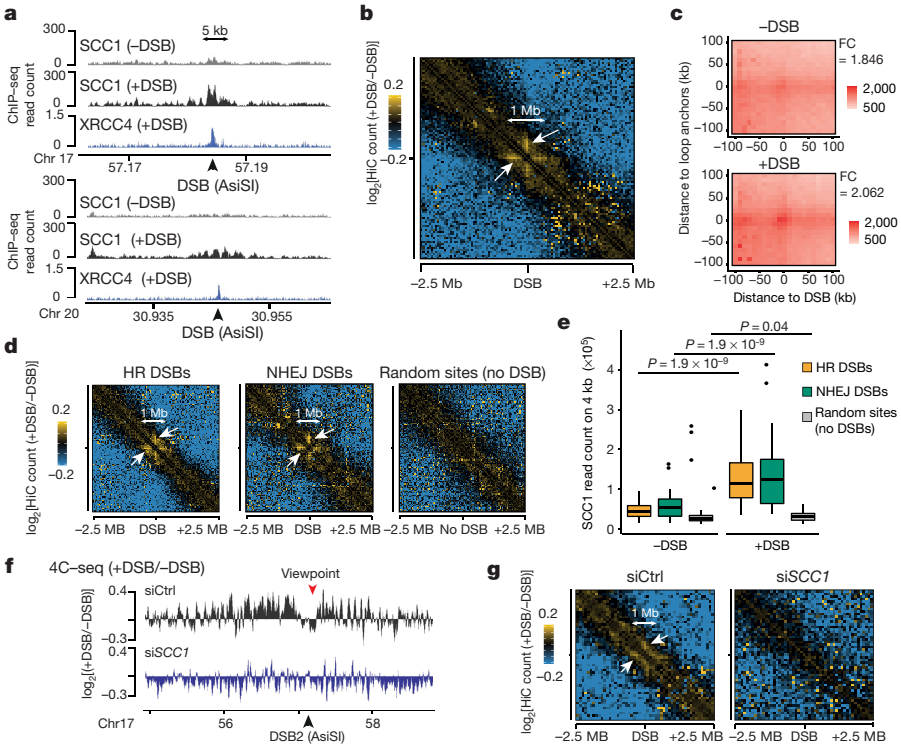


Fig. 2 | DSB-anchored cohesin mediates loop extrusion. a, Genomic tracks of SCC1 and XRCC4 ChIP–seq at two DSBs. **b**, Averaged Hi-C contact matrix of $\log_2$[+DSB/–DSB] ($n = 2$ biological replicates) centred on the 80 best-induced DSBs (50-kb resolution, 5-Mb window). White arrows, stripes. **c**, 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 × 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[7,16]. 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[21–24]. 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 DIvA 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[22,24–27]. 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 DIvA cells[28]. 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 DIvA 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

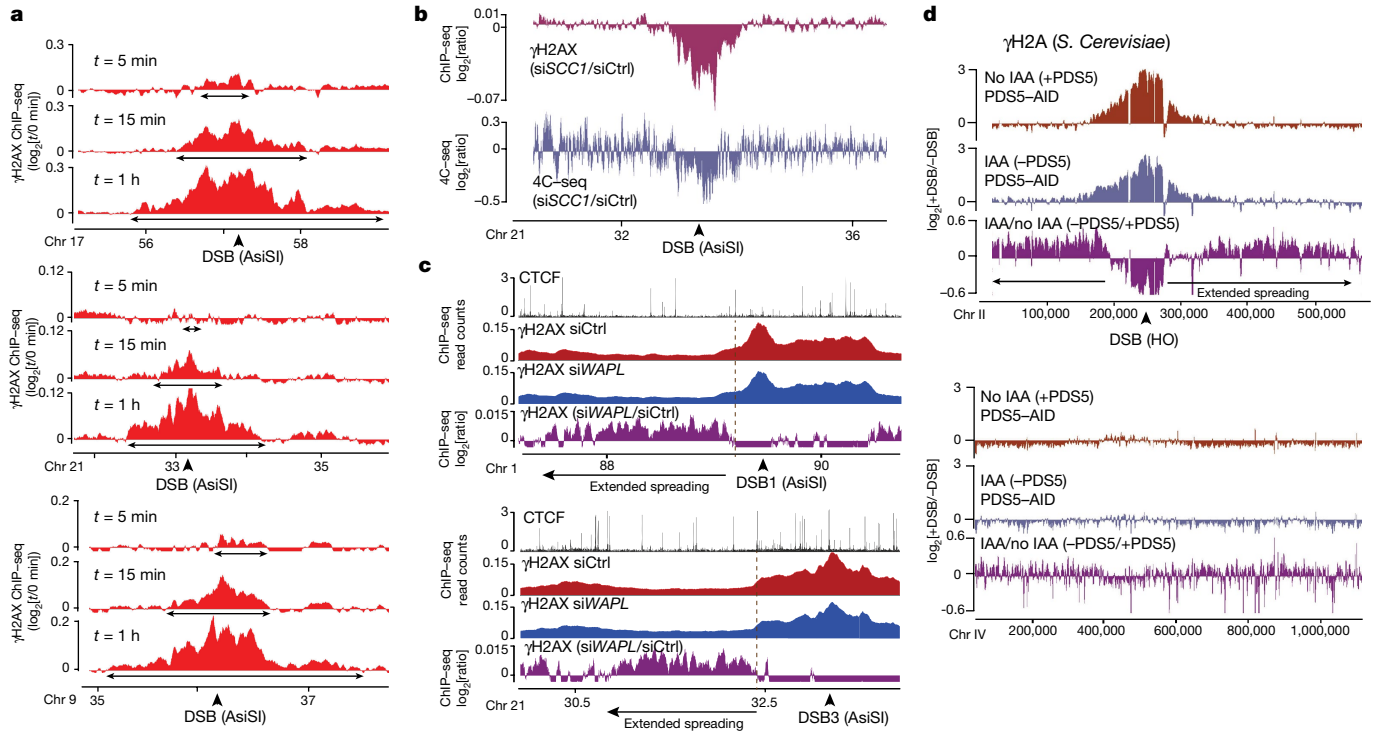


Fig. 3 | DSB-anchored loop extrusion mediates γH2AX spreading. a, γH2AX ChIP–seq tracks at three DSB sites upon DSB induction at different time points after release of ATM inhibition (ATMi) (expressed as $\log_2$[+DSB + ATMi + time after washes/+DSB + ATMi + 0 min after washes]) (20-kb smoothed, $n = 1$). **b**, Top, genomic track showing differential ($\log_2$[si*SCC1*/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 ($\log_2$[si*SCC1*/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 $\log_2$[si*WAPL*/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[14,29], 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 'poising' 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^{-1}, consistent with a loop-extrusion-dependent mechanism[21,23] (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)[7] 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, SMC5/6) 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

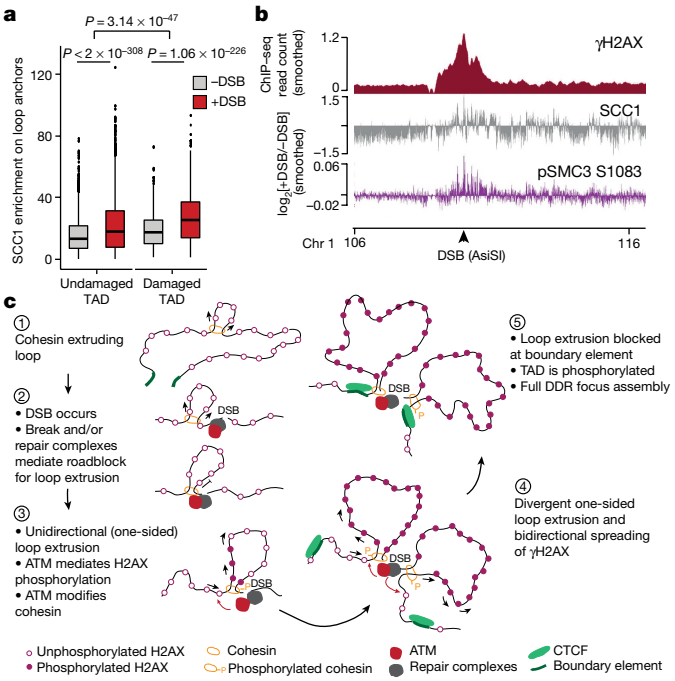


Fig. 4 | DSBs trigger modifications of cohesin biology at a genome-wide scale, accentuated in damaged TADs. a, Quantification of SCC1 recruitment on loop anchors before (grey) and after (red) DSB induction, within damaged (n = 1,456) or undamaged TADs (n = 7,804). Centre line, median; box limits, first 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 γH2AX 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 γH2AX 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 γH2AX. 5, Loop extrusion triggers enlargement of γH2AX-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 0.5–2 kb s^{-1}) ensures that the entire damaged TAD is phosphorylated in 10–30 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[29,30]. Notably, we observed extended spreading of γH2AX in WAPL-depleted cells (Fig. 3c, Extended Data Fig. 5e), which is consistent with the idea that loop extrusion contributes to γH2AX deposition. This was accompanied by a decrease in γH2AX within TADs (Extended Data Fig. 5f). Given that WAPL depletion, while enlarging loops, also decreases intra-TAD chromatin interactions[30], this suggests that intra-TAD chromosome dynamics also contribute to full deposition of γH2AX.

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[31] that carries three HO endonuclease cleavage sites[32]. Consistent with our observations in human WAPL-depleted cells, extended spreading of γH2A occurred following depletion of PDS5 in yeast cells (Fig. 3d). Notably, PDS5 deficiency triggered a decrease in γH2A 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[33,34]. 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[35], and these modifications are essential for reinforcement of cohesin on the genome after irradiation[34]. 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 γH2AX 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[36]), 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 0.5–2 kb s^{-1} in vitro[21,23], such a mechanism would allow rapid assembly of DDR foci, with the entire megabase-sized chromatin domain being modified in about 10–30 min, which fits with the observed rate of assembly of γH2AX foci[9]. 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[37]; and with the recent observation[38], using light-induced activation of Cas9, that γH2AX is established at a speed of about 150 kb min^{-1} and can in some instance reach up to 30 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[39,40], and in DSB occurrence through topoisomerase reactions[41,42]. 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 rejoining 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 PDS5 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).

Methods

Cell culture and treatments

DIvA (AsiSI-ER-U2OS)[10] 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 (DIvA cells) at 37 °C under a humidified atmosphere with 5% CO_2. 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): CAUGUCAUGUGUCACAUCU; and an siRNA targeting *SCC1* (si*SCC1*): GGUGAAAAUGGCAUUACGG; or *WAPL* (si*WAPL*): CGGACUACCC UUAGCACAA; or *NIPBL* (si*NIPBL*): GCUCGGAACAAAGCAAUUA; 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: GGGCCAGTCGCGGCACTCGC) 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[43,44]. 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 PmeI1 and inserted into the *leu2* locus of strain YSCL004[32]. Insertion of *TIR1*[45] 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[33].

Immunofluorescence

DIvA 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 Colombus 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, 10mM Tris pH7.5, 0.5M 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 DIvA cells using the Arima Hi-C kit (Arima Genomics) according to the manufacturer's instructions. Cells (1×10^6) 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[46] with minor modifications. In brief, 15×10^6 DIvA 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[47], 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)[48] 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. [5]. ChIP experiments for pATM, MDC1 and phosphorylated cohesins were performed in DIvA cells as described[10] 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[10]. 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[49]. In brief, cross-linked DIvA 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 DIvA 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[49] and using the following formula: $(input_{ctrl} \times reads_{exp})/(input_{exp} \times reads_{ctrl})$, in which $input_{ctrl}$ is the total number of reads mapped in ES input (mouse) and $input_{exp}$ is the total number of reads in DIvA input. $reads_{ctrl}$ and $reads_{exp}$ 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 DIvA 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[10], 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[10,12] 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 5×10^6 cells per ml and 8×10^6 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 $\log_2$(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 is = 800000 and ids = 100000. TADs were called using Hi-C matrices at 50-kb resolution with TopDom R 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

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 × 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[5], 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 fromdeeptools, 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 $\log_2$ 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 (si*SCC1*) 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/LegubeDNAREPAIR/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. Rispal 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.

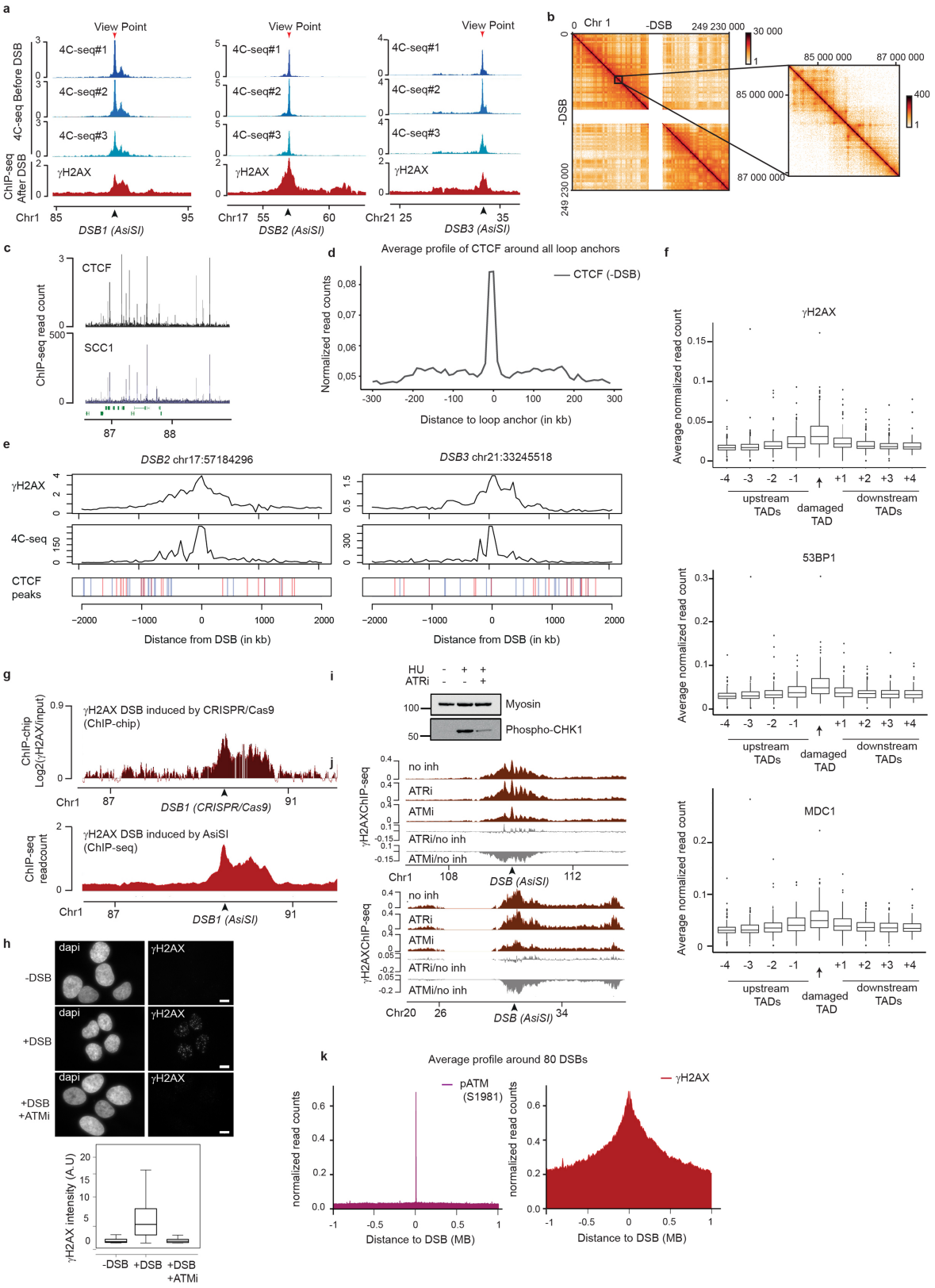
Author contributions C.A. performed 4C–seq, Hi-C, ChIP–seq, ChIP–chip and ChIP–qPCR experiments. A.-L.F. contributed to siRNA experiments and performed CTCF ChIP–seq. K.L. and F.Z. performed yeast strain construction and γH2A ChIP. P.C. performed ChIP–chip in *SCC1* siRNA. V.R. and R.M. performed bioinformatic 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.

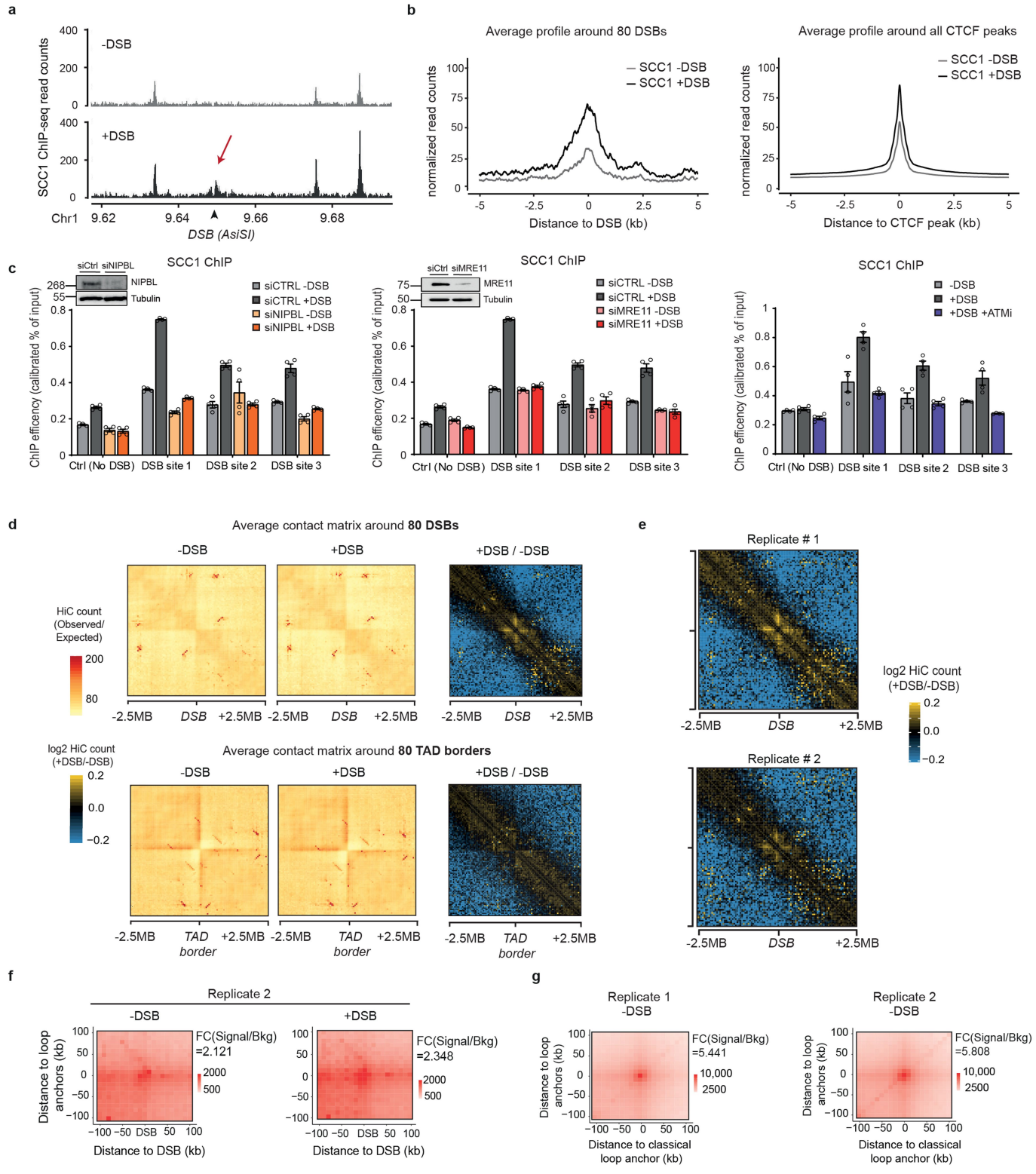
Correspondence and requests for materials should be addressed to G.L.

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.



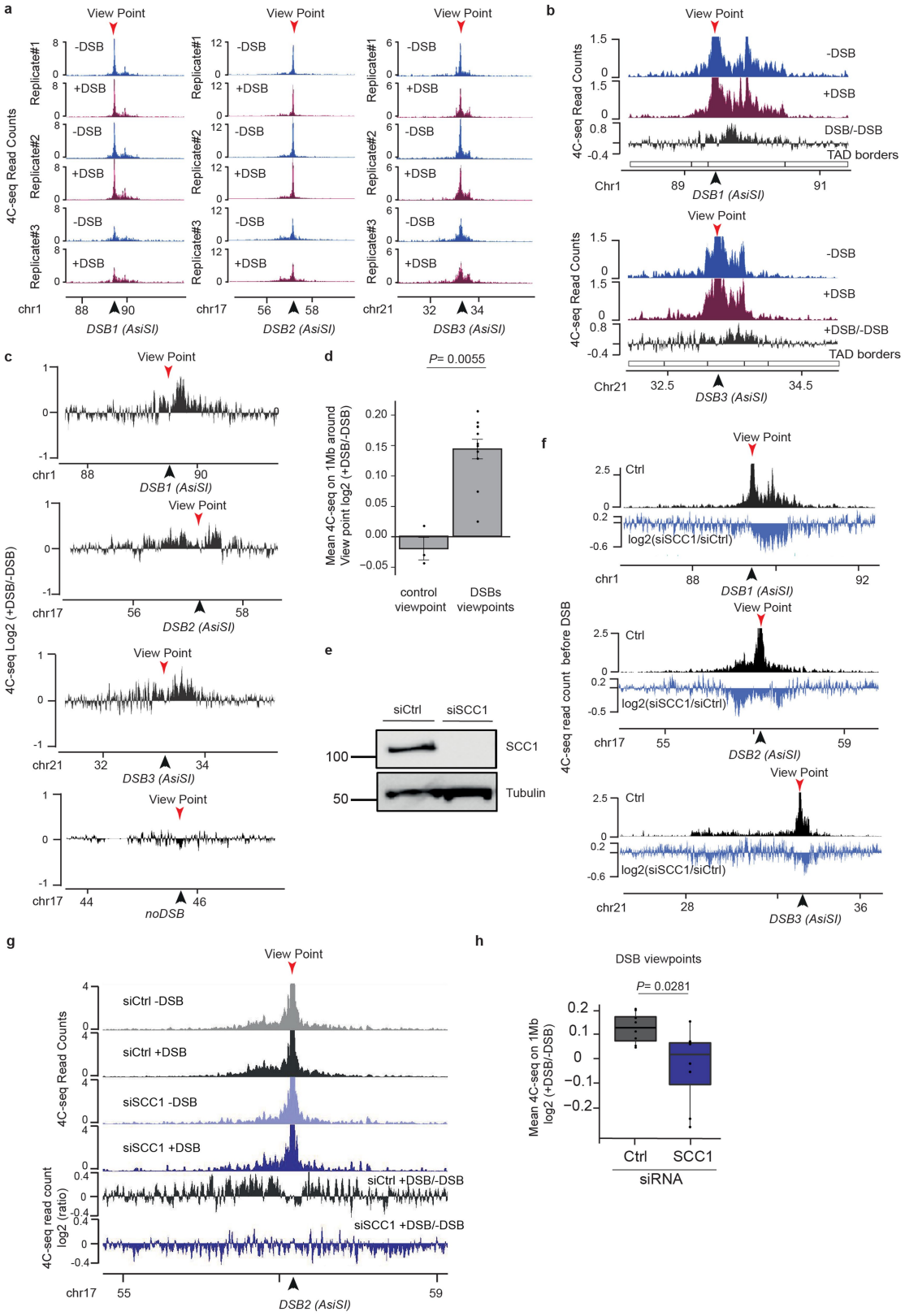
Extended Data Fig. 1 | See next page for caption.

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 AsiSI 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 DIvA 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 $\log_2$[sample/input], smoothed using 100-probe windows) and by AsiSI 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 + ATMi, $n = 95$ 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 ATMi/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 DIvA cells.



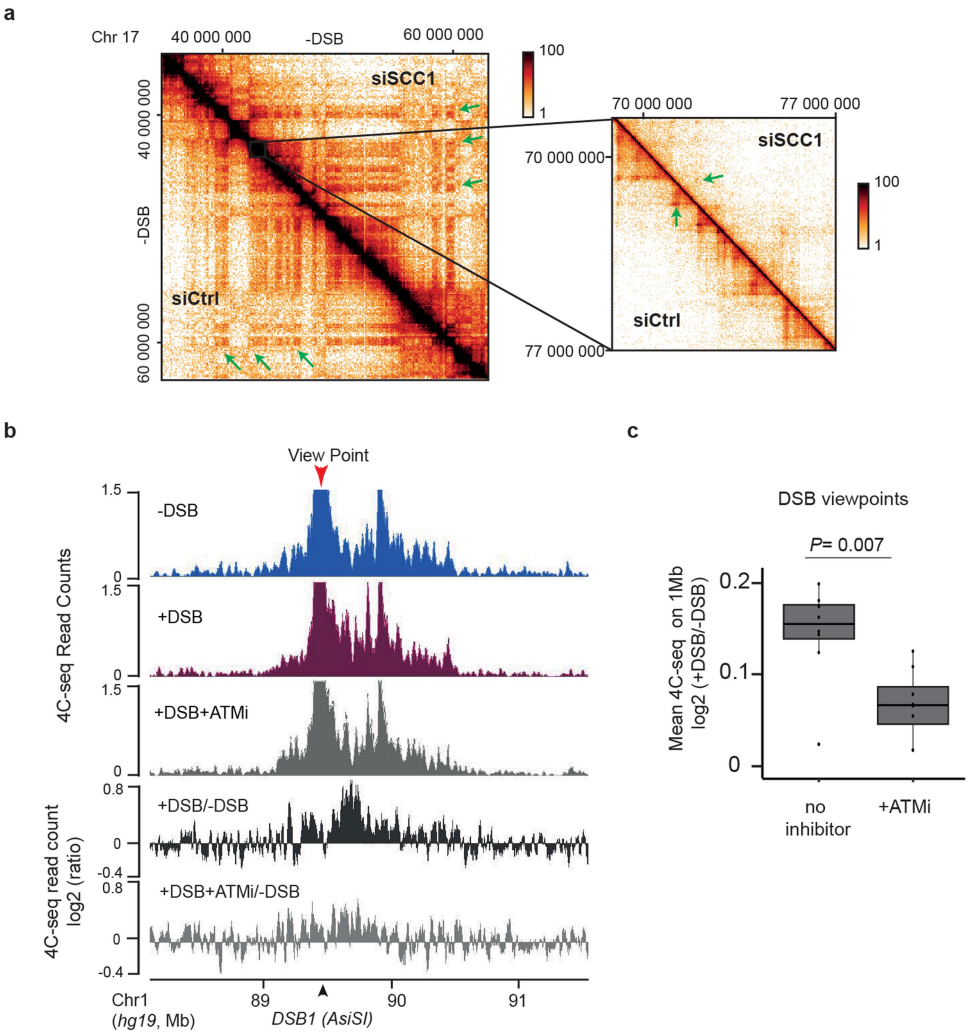
Extended Data Fig. 2 | See next page for caption.

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 ± 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 × 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 × 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).

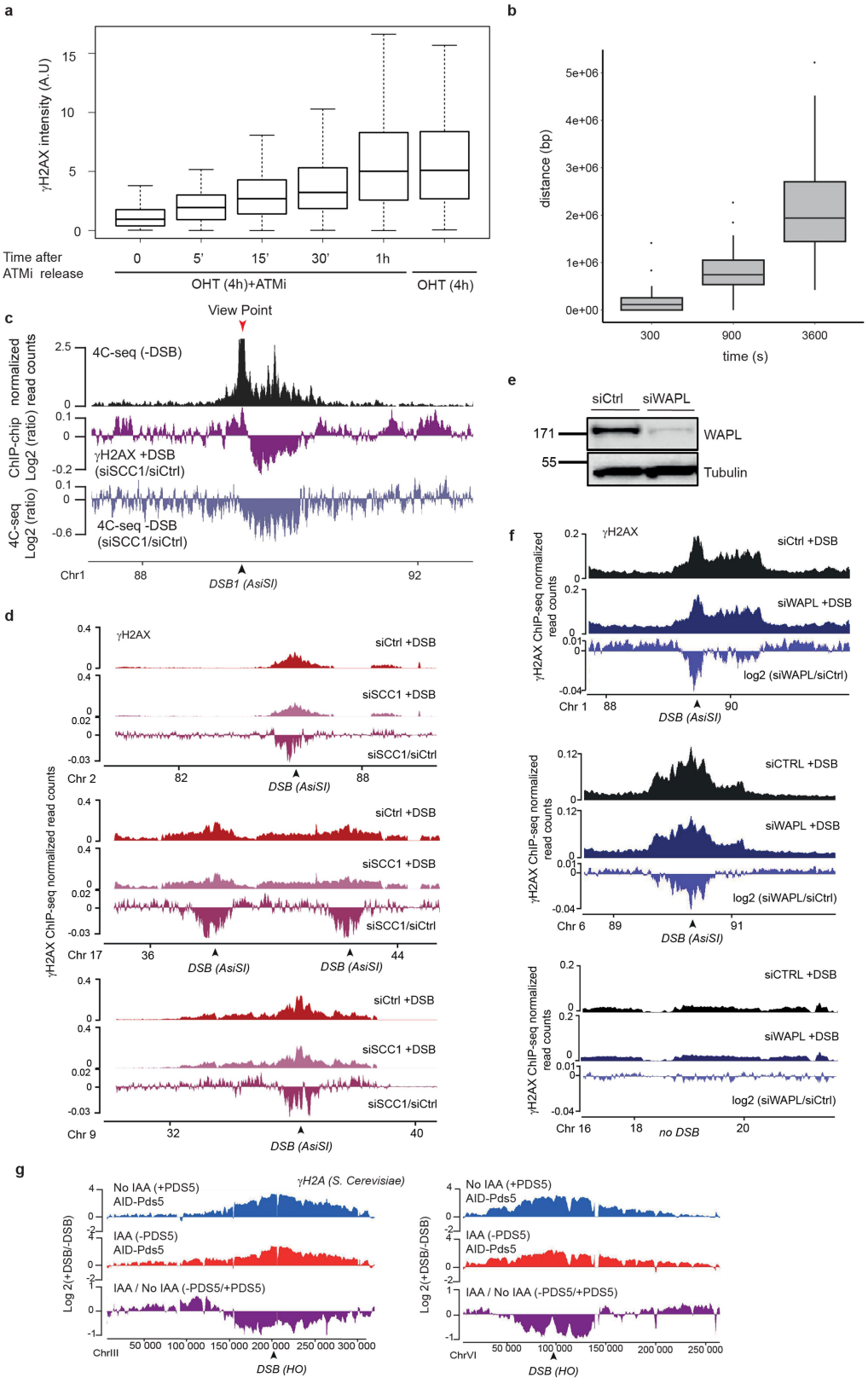


Extended Data Fig. 3 | See next page for caption.

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 ± 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.

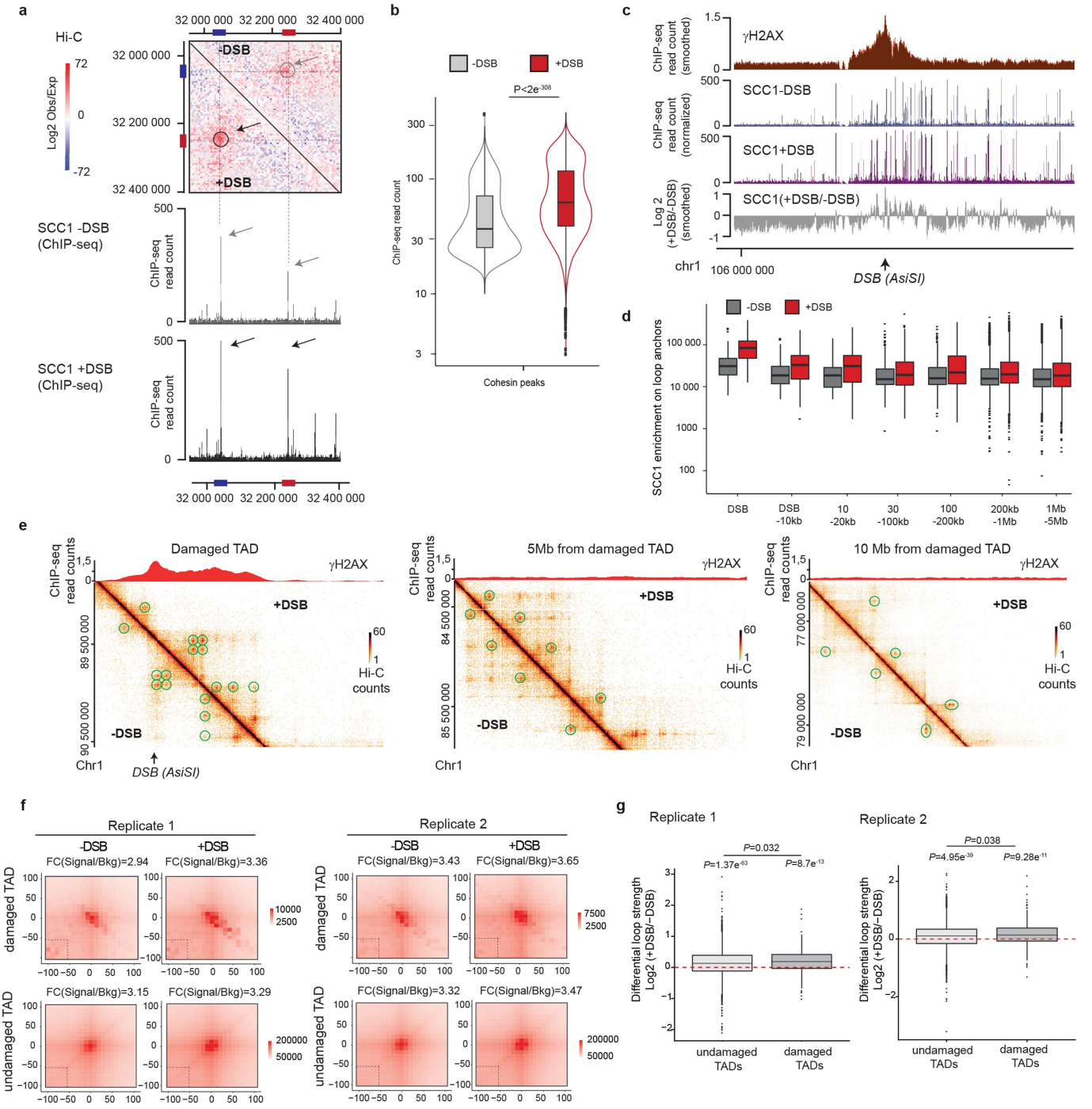


Extended Data Fig. 4 | ATM activity is required for loop extrusion at DSBs. **a**, Hi-C maps before DSB induction of a region of chromosome 17 in control and SCC1-depleted cells. Left, 100-kb resolution; right, 25-kb resolution. **b**, Genomic tracks of 4C–seq before and after DSB induction in untreated 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 ($n = 8$).

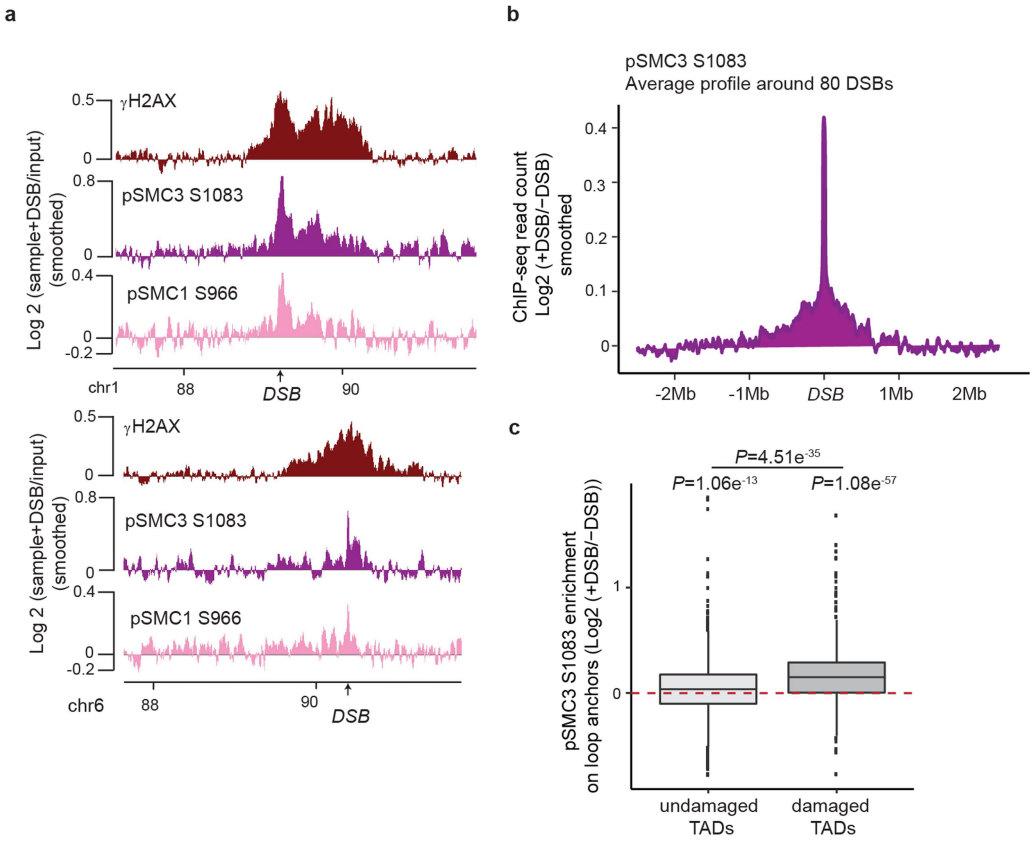


Extended Data Fig. 5 | See next page for caption.

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 $\log_2$[si*SCC1*/siCtrl]). Light blue, differential 4C–seq signal obtained in SCC1-depleted versus control cells before DSB induction ($\log_2$[si*SCC1*/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 ($\log_2$[si*SCC1*/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 ($\log_2$[si*WAPL*/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 ($\log_2$[+DSB/−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 $\log_2$[observed/expected] before or after DSB induction on a region showing a loop on chromosome 20 and devoid of AsiSI 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 γH2AX (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-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**, γH2AX 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 × 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 γH2AX, pSMC3 S1083 and pSMC1 S966 ChIP–chip 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.

Extended Data Table 1 | Antibodies used in this study

Target	Application	Reference	Quantity
γH2AX (S139)	ChIP	Merck Millipore 07-164	2 µg
γH2AX (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
γH2A (yeast)	ChIP	Abcam ab15083	2µg

Extended Data Table 2 | Primers used in this study

Application	Name	Forward primer	Reverse primer
4C-seq	Viewpoint DSB1	AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCTC TTCCGATCTAACCTGGCAACTTATG AATCAGGA	CAAGCAGAAGACGGCATACGAGAT NNNNNNGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTATGTCAAAA GCCAAGGGGACA
4C-seq	Viewpoint DSB2	AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCTC TTCCGATCTTCCTTACGATTATTTGT GAATTTTG	CAAGCAGAAGACGGCATACGAGAT NNNNNNGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTAAGCTAATT CTGAGTTACATACATT
4C-seq	Viewpoint DSB3	AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCTC TTCCGATCTGATTACGTAGAAGGGT GCC	CAAGCAGAAGACGGCATACGAGAT NNNNNNGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTAAGGCAAAT GATAACCCTGT
4C-seq	Viewpoint ctrl region	AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCTC TTCCGATCTTCCTCAGGTTATCATC CCAA	CAAGCAGAAGACGGCATACGAGAT NNNNNNGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTCACCTTCGC TGTACCTTTG
4C-seq	Viewpoint CRISPR site	AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCTC TTCCGATCTTTAAGCACCCTCCTCC TAG	CAAGCAGAAGACGGCATACGAGAT NNNNNNGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTACCTTTACA CCTCAAAACCT
4C-seq	Viewpoint 470 kb upstream (Fig. 1c)	AATGATACGGCGACCACCGAGATC TACACTCTTTCCCTACACGACGCTC TTCCGATCTACAAGGAAGAAGCAG GCATTCA	CAAGCAGAAGACGGCATACGAGAT NNNNNNGTGACTGGAGTTCAGACG TGTGCTCTTCCGATCTTTGAAATGA GTACTCTGCCATCCA
ChIP-qPCR	Ctrl region	AGCACATGGGATTTTGCAGG	TTCCCTCCTTTGTGTCACCA
ChIP-qPCR	DSB site 1	TCCCCTGTTTCTCAGCACTT	CTTCTGCTGTTCTGCGTCCT
ChIP-qPCR	DSB site 2	CCGCCAGAAAGTTTCCTAGA	CTCACCCTTGCAGCACTTG
ChIP-qPCR	DSB site 3	CCTAGCTGAGGTCGGTGCTA	GAAGAGTGAGGAGGGGGAGT
ChIP-qPCR	Cohesin positive site (mouse)	CAGAGATTTGCGGTGTTCGG	TTCACACCTAGAGGAGGGGT

NNN is the position of the optional index.

3.4.6.2 ATM-dependent formation of a novel chromatin compartment (Vincent Rocher)

Using capture Hi-C experiments to study the clustering of induced DSBs at defined loci in the human genome, the team previously demonstrated that DSBs physically cluster, but only when induced within transcriptionally active genes [Aymard *et al.* 2017]. Damaged gene clustering mainly occurs in G1 cell-cycle phase and corresponds to delayed repair. In addition, clustering of DSBs depends on the MRN complex as well as the Formin 2 (FMN2) nuclear actin organizer and the linker of nuclear and cytoplasmic skeleton (LINC) complex, which suggests a role of active mechanisms to promote clustering. However, the role of DSB clustering has remained enigmatic given that the physical proximity of several DSBs can also trigger translocations by illegitimate rejoining of two DNA ends, thus increasing genome instability, questioning the selective advantage of DSB clustering for DNA repair. Moreover, deeper analyses of DSB clustering was limited by the resolution of capture Hi-C data at 100 kb resolution.

Using Hi-C experiments at high resolution (5-10 kb), Coline Arnould, Vincent Rocher *et al.* revealed that the clustering of DSBs involves the formation of a new chromatin sub-compartment (called “D” compartment) driven by ATM and associated with γH2AX and 53BP1 (Figures 1 and 3 from the submitted article ”Loop extrusion as a mechanism for DNA double-strand breaks repair foci formation” below). Formation of “D” compartment mainly occurs during G1 phase, is cohesin independent and is increased by DNA-PK pharmacological inhibition (Figure 2). Most notably, a subset of DNA damage responsive genes upregulated after DSB induction also physically relocate to the D sub-compartment, supporting a role for DSB clustering in activating the DNA Damage Response (Figure 3). However, 3D clustering of DSBs also comes at the expense of an increased translocations rate, which is responsible for genomic instability in cancer (Figure 4).

ATM-dependent formation of a novel chromatin compartment regulates the Response to DNA Double Strand Breaks and the biogenesis of translocations

Coline Arnould[1#], Vincent Rocher[1#], Aldo S. Bader[2], Emma Lesage[1], Nadine Puget[1], Thomas Clouaire[1], Raphael Mourad[1], Daan Noordermeer[3], Martin Bushell[2,4] and Gaëlle Legube[1]*

1. *MCD, Centre de Biologie Intégrative (CBI), CNRS, Université de Toulouse, UT3*
2. *Cancer Research UK Beatson Institute, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1BD, UK*
3. *Université Paris-Saclay, CEA, CNRS, Institute for Integrative Biology of the Cell (I2BC), 91198, Gif-sur-Yvette, France*
4. *Institute of Cancer Sciences, University of Glasgow, Garscube Estate, Switchback Road, Bearsden, Glasgow G61 1QH, UK*

* corresponding author: gaelle.legube@univ-tlse3.fr

# these authors contributed equally

Abstract

DNA Double-Strand Breaks (DSBs) repair is essential to safeguard genome integrity but the contribution of chromosome folding into this process remains elusive. Here we unveiled basic principles of chromosome dynamics upon DSBs in mammalian cells, controlled by key kinases from the DNA Damage Response. We report that ATM is responsible for the reinforcement of topologically associating domains (TAD) that experience a DSB. ATM further drives the formation of a new chromatin sub-compartment ("D" compartment) upon clustering of damaged TADs decorated with γH2AX and 53BP1. "D" compartment formation mostly occurs in G1, is independent of cohesin and is enhanced upon DNA-PK pharmacological inhibition. Importantly, a subset of DNA damage responsive genes that are upregulated following DSBs also physically localize in the D sub-compartment and this ensures their optimal activation, providing a function for DSB clustering in activating the DNA Damage Response. However, these DSB-induced changes in genome organization also come at the expense of an increased translocations rate, which we could also detect on cancer genomes. Overall, our work provides a function for DSB-induced compartmentalization in orchestrating the DNA Damage Response and highlights the critical impact of chromosome architecture in genomic instability.

Main

DNA Double-Strand Breaks (DSBs) are highly toxic lesions that can trigger translocations or gross chromosomal rearrangements, thereby severely challenging genome integrity and cell homeostasis. Chromatin plays a pivotal function during DNA repair, which is achieved by either non-homologous end joining or homologous recombination pathways[1]. Yet, little is known about the contribution of chromosome architecture into these processes. DSBs activate the DNA Damage Response (DDR) that largely relies on PI3K kinases, including ATM and DNA-PK, and on the establishment of megabase-sized, γH2AX-decorated chromatin domains that act as seeds for subsequent signaling events, such as 53BP1 recruitment and DDR foci formation[2,3].

Importantly, γH2AX spreading is largely influenced by the pre-existing chromosome conformation in topologically associating domains (TADs)[4–6] and we recently reported that loop-extrusion, which compacts the chromatin and leads to TADs formation, is instrumental for γH2AX spreading and DDR foci assembly[5]. Moreover, irradiation induces a general chromatin response reinforcing TADs genome wide[7]. At a larger scale, previous work in mammalian cells revealed that DSBs display the ability to "cluster" within the nuclear space (*i.e.,* fuse) forming large microscopically visible repair foci, composed of several individual repair foci[8–10]. DSB clustering depends on the actin network, the LINC (a nuclear envelope embedded complex)[9,11,12], as well as on the liquid-liquid phase separation properties of 53BP1[13,14]. The function of DSB clustering has remained enigmatic given that juxtaposition of several DSBs can elicit translocation (*i.e*: illegitimate rejoining of two DNA ends)[10], questioning the selective advantage of DSB clustering/ repair foci fusion[15].

ATM drives an acute reinforcement of damaged TADs.

In order to get comprehensive insights into chromosome behavior following DSBs, we analyzed 3D genome organization using Hi-C data generated in the human DIvA cell line where multiple DSBs are induced at annotated positions upon hydroxytamoxifen (OHT) addition[16]. Our previous analyses using γH2AX ChIP-seq and direct DSB mapping by BLESS allowed us to identify 80 robustly induced DSBs on the human genome[3]. Using differential Hi-C maps, we found that intra-TAD contacts frequencies were strongly increased within TADs that experience a DSB (*i.e.* damaged TADs, Fig. 1a, right panel red square) compared to undamaged TADs, while contacts with neighboring adjacent domains were significantly decreased (Fig. 1a, right panel blue square, Fig. 1b). Interestingly, in some instances, the DSB itself displayed a particularly strong depletion of contact frequency with adjacent chromatin (Fig. 1c black arrow) indicating that the DSB is kept isolated from the surrounding environment, outside of its own TAD.

We further investigated the contribution of PI3-Kinases involved in response to DSB by performing Hi-C in presence of inhibitors of ATM and DNA-PK, which respectively negatively and positively impact γH2AX accumulation at DSBs (in contrast to ATR inhibition, which does not noticeably alter γH2AX foci formation in DIvA cells)[5,17]. Notably, DNA-PK inhibition exacerbated the increase in intra-TAD contacts following DSB induction, while ATM inhibition abrogated it (Fig. 1d, Fig. S1a). TAD structures visualized on Hi-C maps are believed to arise thanks to cohesin-mediated loop extrusion[18]. Our previous work indicated that a bidirectional, divergent, cohesin-dependent loop-extrusion process takes place at DSBs[5]. This DSB-anchored loop extrusion can be visualized on differential Hi-C maps by a “cross” pattern centered on the DSB (Fig. 1e). Notably, ATM inhibition impaired loop extrusion, while DNA-PK inhibition strongly increased it (Fig. 1e). Moreover, depletion of the cohesin subunit SCC1, which abolishes DSB-induced loop extrusion[5], decreased the reinforcement of intra TAD-contacts in damaged, γH2AX-decorated, chromatin domains (Fig. 1f, Fig. S1b).

Altogether these data indicate that ATM triggers cohesin-mediated loop extrusion arising from the DSB and the insulation of the damaged TADs from the surrounding chromatin.

ATM drives clustering of damaged TADs, in a cell cycle regulated manner

We further analyzed Hi-C data with respect to long-range contacts within the nuclear space. Hi-C data revealed that DSBs cluster together (Fig. 2a, red square away from the diagonal), as previously observed using Capture Hi-C[9]. The higher resolution of this Hi-C dataset now enables us to conclude that DSB clustering takes place between entire γH2AX-decorated TADs and can happen between DSBs induced on the same chromosome (Fig. S2a) as well as on different chromosomes (Fig. S2b). Of interest, some γH2AX domains were able to interact with more than a single other γH2AX domain (Fig. 2b, black arrows). Notably, this ability to form clusters of multiples TADs (also known as TADs cliques[19]) upon DSB induction correlated with several DSB-induced chromatin features that occur at the scale of an entire TAD[3], including γH2AX, 53BP1 and ubiquitin chains levels as well as the depletion of histone H1 around DSB detected by ChIP-seq (Fig. 2c). Moreover, it also correlated with initial RNAPII occupancy prior DSB induction indicating that DSBs prone to cluster and form damaged TAD cliques are those occurring in transcribed loci (Fig. 2c).

We further examined the effect of cohesin depletion on damaged TAD clustering. Inspection of individual DSBs indicated that SCC1 depletion by siRNA did not alter clustering (Fig. 2d). Quantification of *trans* interactions between all DSBs also indicates that SCC1 depletion did not modify the ability of damaged TAD to physically interact together (Fig. S2c). Additionally, we found that inhibition of ATM compromised DSB clustering, whilst inhibiting DNA-PK activity triggered a substantial increase in DSB clustering (Fig. 2e, Fig. S2d).

Given the conflicting data regarding the cell cycle regulation of DSB clustering[8,9,12], we further investigated DSB clustering in synchronized cells. DSB clustering (*i.e*. damaged TAD-TAD interaction) could be readily detected by 4C-seq when using a DSB as a view point, as shown by the increase of 4C-seq signal observed on other DSBs induced on the genome (Fig. 2f). We used five individual view-points: one control view point located on an undamaged locus, and four viewpoints at DSBs sites, three of which being "cluster-prone" DSBs, and one efficiently induced DSB which is unable to cluster with other DSBs. 4C-seq experiments performed before and after DSB induction in synchronized cells indicated that DSB clustering is readily detectable during G1 and is strongly reduced during the other cell cycle stages (see an example Fig. S2e). G1-specific DSB clustering was observed only when using as viewpoints "clustering-prone" DSBs, but not when using the undamaged control locus or the DSB unable to cluster (Fig. 2g).

Taken altogether, our results indicate that upon DSB formation, TADs that carry DSBs are able to physically contact each other in the nuclear space (*i.e*. cluster) in a manner that is entirely dependent on ATM, exacerbated upon DNA-PK inhibition, and mostly independent of the cohesin complex. Damaged TAD clustering mostly takes place in G1 and correlates with TAD-scale DSB-induced chromatin modifications (γH2AX, Ubiquitin accumulation and H1 depletion) as well as 53BP1 accumulation.

A new "D" sub-compartment forms following DSB induction

Previous work identified the existence of two main, spatially distinct, self-segregated, chromatin "compartments" in mammalian nuclei. These chromatin compartments were determined by Principal Component Analysis (PCA) of Hi-C chromosomal contact maps where the first principal component allowed to identify loci that share similar interaction pattern, and

that can be visualized linearly using eigenvectors. Further correlations with epigenomic features revealed that these two spatially segregated compartments correspond to active (the "A" compartment or euchromatin) and inactive chromatin (the "B" compartment or heterochromatin)[20]. The identification of A/B compartment using our Hi-C datasets revealed that DSB induction does not trigger major changes in genome compartmentalization into euchromatin *versus* heterochromatin (Fig. S3a). Saddle plots further confirmed that neither DSB treatment nor the pharmacological inhibition of DNAPK and ATM significantly modified the ability of the genome to segregate into active A and inactive B compartments (Fig. S3b). Moreover, DSB induction did not generally lead to compartment switch of the underlying chromatin domain, except in very few cases: Among the 80 DSBs induced by AsiSI, 58 DSBs were induced in the A compartment and all of them remained in the A compartment following DSB induction (see an example Fig. S3c top panel). Conversely, among the 22 DSBs induced in the B compartment, only 4 showed a shift from B to A (see two examples Fig. S3c middle and bottom panels). We further investigated the relationship between the compartment type and the ability of DSBs to cluster together. Of interest, DSB clustering was detectable mostly for DSBs in the A compartment (Fig. S3d).

Beyond the main classification between A/B compartments, sub-compartments have since been identified using higher resolution Hi-C maps, which correspond to subsets of heterochromatin loci (B1-B4) and of active loci (A1-A2)[21]. Of interest, such sub-compartments also correspond to microscopically visible nuclear structures such as nuclear speckles (A1)[22] or Polycomb bodies (B1)[21] for instance. Given that previous studies have long identified large, microscopically detectable γH2AX bodies following DNA damage and that our Hi-C data revealed clustering of damaged TADs, we postulated that DSBs may also induce a sub-compartment, in particular within the A compartment (*i.e*,: some A compartment, damaged-loci further segregate from the rest of the active compartment). In order to investigate this point,

we applied PCA analysis on differential Hi-C maps (*i.e.* contact matrices of +DSB/-DSB) on each individual chromosome. The first Chromosomal Eigenvector (CEV, PC1) allowed us to identify a DSB-induced chromatin compartment mainly on chromosomes displaying a large number of DSBs (chr1,17 and X) (Fig. S4a, Fig. 3a,). Notably, a similar analysis on Hi-C maps generated upon DNA-PK inhibition, which impairs repair[17] and increases DSB clustering (Fig. 2), allowed to identify this compartment on more chromosomes (such as chr6 for instance, Fig. S4b, bottom track). This sub-compartment displayed a very strong correlation with γH2AX-decorated chromatin following DSB (Fig. 3a, Fig. S4a-d) and was henceforth further named "D" sub-compartment (for DSB-induced compartment). Yet, further inspection revealed that the D sub- compartment is not solely generated through the clustering of damaged chromatin (*i.e.* TADs that carry DSBs and are enriched in γH2AX). Indeed, we could identify chromatin domains, not containing any DSB and not decorated by γH2AX, that associate with the D sub-compartment after damage (blue rectangle Fig. 3b). After exclusion of γH2AX-covered chromatin domains, correlation analysis using chromosomes 1,17 and X, on which the D sub-compartment was readily detected, indicated that non-damaged loci that tend to segregate with the D compartment are enriched in H2AZac, H3K4me3 and H3K79me2 (Fig. S4e, Fig. 3b). Conversely, these loci targeted to the D compartment displayed a negative correlation with repressive marks such as H3K9me3 (Fig. S4e). A similar trend was observed when D sub-compartment was computed from the Hi-C data obtained in presence of the DNA-PK inhibitor and correlation analysis performed on all chromosomes showing D compartmentalization (*i.e,* chr 1,2,6,9,13,17,18,20 and X) (Fig. S4e bottom panel). Altogether our data indicate that upon DSB production on the genome, damaged TADs, covered by γH2AX/53BP1, form a new chromatin compartment that segregates from the rest of the genome and in which some additional undamaged loci that exhibit chromatin marks typical of active transcription can be further targeted.

A subset of DNA damage responsive genes segregates with the D sub-compartment to achieve optimal activation.

In order to decipher the nature of the active genes targeted to the D compartment, we further explored the DNA motifs enriched on "D" genes compared to "non D" genes, *i.e.* genes recruited to the D compartment, *versus* the one that do not display targeting to the D compartment (discarding all genes directly comprised in γH2AX domains). Notably, the top enriched motifs included OSR1, TP73, Nkx3.1 and E2F binding sites, which are tumor suppressor and /or known to be involved in the DNA damage response (Fig. S4f)[23–26], suggesting a direct physical targeting of DNA damage responsive genes to the "D" sub-compartment. In agreement, visual inspection revealed that some known p53 target genes which are upregulated following DSB induction were associated with the D compartment, even when as far as >20MB from the closest DSB (see an example Fig. 3c). To test the hypothesis that DNA damage responsive genes are recruited to the D compartment, we performed RNA-seq before and after DSB induction and retrieved genes that are upregulated following DSB induction. Notably, genes upregulated following DSB induction displayed a higher D compartment signal compared to genes that were either not regulated or downregulated after DSBs (Fig. 3d). Of note, if some of the upregulated genes were indeed targeted to the D compartment, this was not the case for all of them. Importantly, the upregulated genes targeted to the D-compartment were not in average closer to DSBs than the upregulated genes not-targeted to the D compartment (Fig. S4g), ruling out a potential bias due to the genomic distribution of AsiSI DSBs.

In order to determine whether recruitment of those genes to the D sub-compartment contribute to their activation following DNA damage, we investigated the consequence of disrupting DSB

clustering (and hence formation of D compartment) by depleting the SUN2 component of the LINC complex, previously found as a DSB-clustering promoting factor[9,11]. SUN2 depletion altered the transcriptional activation of genes found to be upregulated and targeted to the D sub-compartment upon DSB in DIvA cells (Fig. 3e).

Altogether these data indicate that DSB induction triggers the formation of a novel chromatin sub-compartment that comprises not only damaged TADs, decorated by γH2AX and 53BP1, but also a subset of genes upregulated following DNA damage, for which targeting to D sub-compartment is required for optimal activation. Altogether this suggests a role of the D sub-compartment, and hence DSB clustering, in the activation of the DNA Damage Response.

DSB-induced reorganization of chromosome folding favors translocations.

Importantly, while our above data suggest a beneficial role of DSB clustering in potentiating the DDR, it may also be detrimental, since bringing two DSBs in a close proximity may fosters translocations (illegitimate rejoining of two DSBs), as previously proposed[10]. We therefore assessed by qPCR the frequency of translocations events occurring in DIvA cells post-DSB induction, in conditions where we found altered DSBs clustering and D compartment formation.

Notably, translocations are increased in G1 compared to S/G2-synchronized cells (Fig. 4a), in agreement with an enhanced DSB clustering observed in G1 cells (Fig. 2). Moreover, DNA-PK inhibition, that increased D-compartment formation (Fig. 2e, Fig. S2d, Fig. S4b) also strongly increased translocation frequency (Fig. 4b). On another hand, depletion of 53BP1 (Fig. S5a), previously found to mediate repair foci phase separation[13], as well as a treatment with 1,6-hexanediol, which disrupts phase condensates (Fig. S5b), decreased translocations (Fig. 4c). Similarly, depletion of SUN2, member of the LINC complex and of ARP2, an actin branching

factor (Fig. S5a), reported as mediating DSB clustering[9,11,12], decreased translocations (Fig. 4c). Surprisingly, depletion of the cohesin subunits SMC1 or SCC1 also decreased translocation frequency (Fig. 4d, Fig S5c). This was unexpected since SCC1-depleted cells do not display clustering defects (Fig. 2).

Given that the two translocations assessed by our qPCR assay are both intra-chromosomal translocations (*i.e.*: rejoining of two distant DSBs located on the same chromosome) we hypothesized that translocation frequency at the intra-chromosomal level may also be regulated by the DSB-induced loop extrusion that depends on the cohesin complex. In order to investigate more broadly translocation events between multiple DSBs induced in the DIvA cell line, we designed a novel multiplexed amplification protocol followed by NGS sequencing. In control cells, we could readily detect increased translocation frequency upon induction of DSB compared to control genomic locations (Fig. S5d). Strikingly, depletion of SCC1 decreased the frequency of intra-chromosomal translocations, while leaving inter-chromosomal translocations unaffected (Fig. 4e). In contrast depletion of SUN2 and ARP2 decreased both intra- and inter-chromosomal translocations (Fig. 4f-g). Taken together these data suggest that both the DSB-induced loop extrusion and the formation of the D sub-compartment through clustering of damaged TADs, display the potential to generate translocations.

Given our above finding that a subset of genes upregulated following DSB induction can be physically targeted to the D compartment after break induction (Fig. 3), we further hypothesized that such a physical proximity may account for some of the translocations observed on cancer genomes. We retrieved breakpoint positions of inter-chromosomal translocations of 1493 individuals across 18 different cancers types (from[27]), and assessed their potential overlap with genes targeted to the D sub-compartment (reproducibly detected in the three Hi-C replicates on chr1,17 and X, on which D sub-compartment could be identified accurately). D-targeted genes were further sorted as either upregulated, downregulated or not significantly altered following

DSB induction, and compared to their counterparts not targeted to the D compartment. We found that genes that are upregulated following DSB induction and that are targeted to the D compartment displayed a significant overlap with translocations breakpoints, in contrast to genes that are not targeted to the D compartment (non-D) (Fig. 4h). Altogether these data indicate that the relocalization of upregulated genes during the DNA Damage response in the DSB-induced sub-compartment likely accounts for some of the translocations detected on cancer genomes. Given that DDR genes comprise a number of tumor suppressor genes, such a physical proximity of these genes with DSBs within the D sub-compartment formed in response to DNA damage, may be a key mechanism driving oncogenesis, through fostering the instability of tumor suppressor genes.

Conclusion

Altogether this work shows that DSB-induced changes in chromosome architecture is an integral component of the DNA Damage Response, but also acts as a double-edged sword that can challenge genomic integrity through the formation of translocations.

Our data suggest that a chromatin sub-compartment arises when γH2AX/53BP1-decorated domains, established by ATM-induced loop extrusion post DSB, self-segregate from the rest of chromatin. This may, at least in part, occur thanks to the LLPS properties of 53BP1[13,14,28]. This DSB-induced ("D") sub-compartment further recruits a subset of genes involved in the DNA damage response and contributes to their activation (Fig. S5e). This model is in agreement with previous work which identified 53BP1 as critical for p53 target genes activation[29], with the findings that disrupting 53BP1 droplet formation alters checkpoint activation[13] and with the fact that enhanced 53BP1 phase separation triggers an elevated p53 response[30] as does the loss of TIRR, a protein that regulates 53BP1 association to DSBs[31,32]. We propose that the formation

of the "D" sub-compartment allows to precisely tune the magnitude of the DDR with respect to DSB load and persistency, providing a function for these enigmatically large γH2AX/53BP1-decorated chromatin domains and to DSB clustering. Furthermore, this observation may provide a rationale for why so many transcription factors (including p53) were found recruited at DSBs repair foci[33]. While initially thought to allow chromatin remodeling in order to enhance DSB repair, the recruitment of transcription factors to DSB repair foci may in fact rather reflects the relocalization of DDR genes within the D compartment (hence at physical proximity of the DSB).

Yet, this comes at the expense of potential translocations, as both loop extrusion and coalescence of damaged TAD are able to bring linearly distant DSBs in close physical proximity (Fig. S5e). Importantly, we found that the genes upregulated in response to DSB and relocated to the D compartment displayed significant overlap with translocation breakpoints identified by whole genome sequencing in patient cancer samples. In agreement with an increased occurrence of structural variants on tumor suppressor genes[27], we propose that the physical targeting of DNA damage responsive genes to the D compartment, by bringing DSBs and DDR genes in close spatial proximity, may occasionally trigger deleterious rearrangements on genes involved in the control of cell proliferation and apoptosis upon DNA damage, and may hence act as a critical driver of oncogenesis by disrupting the integrity of tumor suppressor genes.

Methods

Cell culture and treatments

DIvA (AsiSI-ER-U20S)[16] and AID-DIvA (AID-AsiSI-ER-U20S)[34] cells were grown in Dubelcco's modified Eagle's medium (DMEM) supplemented with 10% SVF (Invitrogen), antibiotics and either 1 µg/mL puromycin (DIvA cells) or 800 µg/mL G418 (AID-DIvA cells) at 37 °C under a humidified atmosphere with 5% CO2. To induce DSBs, cells were treated with 300nM 4OHT (Sigma, H7904) for 4 h. For ATM or DNA-PK inhibition, cells were pretreated for 1 h respectively with 20µM KU-55933 (Sigma, SML1109) or 2µM NU-7441 (Selleckchem, S2638) and during subsequent 4OHT treatment. Treatment with 10% 1,6-hexanediol (Sigma, 240117) was performed for 3 min before the end of the 4OHT treatment. For cell synchronization, cells were incubated for 18 h with 2 mM thymidine (Sigma, T1895), then released during 11 h, followed by a second thymidine treatment for 18 hr. S, G2 and G1 cells were then respectively treated with OHT at, 0, 6 or 11 h following thymidine release and harvested 4 h later. siRNA transfections were performed 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. siRNA transfections were performed using a control siRNA (siCTRL): CAUGUCAUGUGUCACAUCU; or using a siRNA targeting *SCC1* (siSCC1): GGUGAAAAUGGCAUUACGG; or SMC1 (siSMC1): UAGGCUUCCUGGAGGUCACAUUUAA; or 53BP1 (si53BP1): GAACGAGGAGACGGUAAUA; or SUN2 (siSUN2): CGAGCCTATTCAGACGTTTCA; or ARP2 (siARP2): GGCACCGGGUUUGUGAAGU.

Translocation assay

Translocation assays after siRNA transfection or 1,6-Hexanediol treatment were performed at least in triplicates in AID-DIvA cells as described in[35]. Translocation assay in synchronized cells was performed in DIvA cells following a 4OHT treatment (n=4 biological replicates). Two different possible translocations between different AsiSI sites were assessed by qPCR using the following primers: Translocation1_Fw: GACTGGCATAAGCGTCTTCG, Translocation1_Rev: TCTGAAGTCTGCGCTTTCCA, Translocation2_ Fw: GGAAGCCGCCCAGAATAAGA, Translocation2_Rev: TCTGAAGTCTGCGCTTTCCA. Results were normalized using two control regions, both far from any AsiSI sites and γH2AX domain using the following primers: Ctrl_chr1_82844750_Fw: AGCACATGGGATTTTGCAGG, Ctrl_chr1_82844992_Rev: TTCCCTCCTTTGTGTCACCA, Ctrl_chr17_9784962_Fw: ACAGTGGGAGACAGAAGAGC, Ctrl_chr17_9785135_Rev: CTCCATCATCGCACCCTTTG. Normalized translocation frequencies were calculated using the Bio-Rad CFX Manager 3.1 software69.

Amplicon –seq

AID-DIvA cells were treated with or without 300nM 4OHT for 4 h followed by treatment with indole-3-acetic acid for 14 h. Cells were then lysed in cytoplasmic lysis buffer (50mM HEPES pH7.9, 10mM KCl2, 1.5mM MgCl2, 0.34M sucrose, 0.5% triton X-100, 10% glycerol, 1mM DTT) for 10 minutes on ice, then washed once in cytoplasmic lysis buffer before lysis in genomic extraction buffer (50mM Tris pH8.0, 5mM EDTA, 1% SDS, 0.5mg/mL proteinase K). Lysate was incubated at 60°C for 1 h. Genomic DNA was then ethanol precipitated on ice for 1h, pelleted at 19,000g for 20 min and washed twice in 75% ethanol. Genomic DNA was then used in a multiplex PCR reaction that amplified 25 target sites; 20 AsiSI cut sites and 5 uncut

control sites (Supplementary Table 1). Amplicons were size selected using SPRIselect beads (Beckman, B23318) and subjected to DNA library preparation via the NEBNext Ultra II kit (NEB, E7645L). Libraries were pooled at equimolar concentrations and sequenced via an Illumina NextSeq 500 system using paired end 150 cycles. The data was analyzed via our custom tool mProfile, available at github.com/aldob/mProfile. This identified the genomic primers used in the original genomic PCR reaction to amplify each read in the pair. Translocated reads were therefore identified as those where each read in a pair was amplified by a different primer set, and this was normalized to the total reads that were correctly amplified by these primer sets.

RT-qPCR

RNA was extracted from fresh DIvA cells before and after DSB induction using the RNeasy kit (Qiagen). RNA was then reverse transcribed to cDNA using the AMV reverse transcriptase (Promega, M510F). qPCR experiments were performed to assess the levels of cDNA using primers targeting RPLP0 (FW: GGCGACCTGGAAGTCCAACT; REV: CCATCAGCACCACAGCCTTC), RNF19B (FW: CATCAAGCCATGCCCACGAT; REV: GAATGTACAGCCAGAGGGGC), PLK3 (FW: GCCTGCCGCCGGTTT; REV: GTCTGACGTCGGTAGCCCG), FAS (FW: ATGCACACTCACCAGCAACA; REV: AAGAAGACAAAGCCACCCCA) or GADD45A (FW: ACGATCACTGTCGGGGTGTA; REV: CCACATCTCTGTCGTCGTCC). cDNA levels were then normalized with RPLP0 cDNA level, then expressed at the percentage of the undamaged condition.

Immunofluorescence

DIvA cells were grown on glass coverslips and fixed with 4% paraformaldehyde during 15 min at room temperature. Permeabilization step was performed by treating cells with 0,5% Triton X-100 in PBS for 10 min then cells were blocked with PBS-BSA 3% for 30min. Primary

antibodies targeting RNA PolI (Santa Cruz sc48385) or PML (Santa Cruz sc-966 (PG-M3)) were diluted 1:500 in PBS-BSA 3% and incubated with cells overnight at 4°C. After washes in 1X PBS, cells were incubated with anti-mouse secondary antibody (conjugated to Alexa 594 or Alexa 488, Invitrogen), diluted 1:1000 in PBS-BSA 3%, for 1h at room temperature. After a DAPI staining, Citifluor (Citifluor, AF-1) was used for coverslips mounting. Images were acquired with the software MetaMorph, using the 100X objective of a wide-field microscope (Leica, DM6000), equipped with a camera (DR-328G-C01-SIL-505, ANDOR Technology).

Western Blot

Western Blot experiments were performed as in[5] using primary antibody targeting SUN2 (Abcam ab124916 1:1000), ARP2 (Abcam ab128934 1:1000), 53BP1 (Novus Biologicals NB100-305 1:1000), SCC1 (Abcam ab992 1:500) or SMC1 (Abcam ab75819 1:1000).

RNA-seq

RNA-seq was performed as described in[35]. RNA-seq were mapped in paired-end to a custom human genome (hg19 merged with ERCC92) using STAR. Count matrices were extracted using htseq-count with union as resolution-mode and reverse strand mode. Differential expression analysis was made on the count matrix using edgeR with two replicates per condition and differential genes were determined with log-ratio test (LRT). Whole genome coverage was computed using deeptools and bamCoverage to generate bigwig using bam files (without PCR duplicate suppression). Using a cutoff of 0.1 for the adjusted p-value and 0.5 log2 fold-change (~41% increase/decrease of expression), we were able to determine 286 up-regulated and 125 down-regulated genes with 11 of them directly damaged by a DSB. Differential coverage between two conditions was performed using BamCompare from deeptools with setting binsize parameter at 50bp. Log2FC was calculated by edgeR in differential expression analysis.

4C-seq

4C-seq experiments performed in synchronized cells, before and after DSB induction were performed as in[5]. Briefly, 10-15×10^6 DIvA cells per condition were cross-linked, lysed and digested with MboI (New England Biolabs). DNA ligation was performed using the T4 DNA ligase (HC) (Promega), and ligated DNA was digested again using NlaIII (New England Biolabs). Digested DNA was religated with the T4 DNA ligase (HC) (Promega) before to proceed to 4C–seq library preparation. 16 individual PCR reactions were performed in order to amplify ~800ng of 4C-seq template, using inverse primers including the Illumina adaptor sequences and a unique index for each condition (Supplementary Table 2). Libraries were pooled and sent to a Nextseq500 platform at the I2BC Next Generation Sequencing Core Facility (Gif-sur-Yvette).

4C-seq data were processed as described in[5]. Briefly, bwa mem was used for mapping and samtools for sorting and indexing. A custom R script (https://github.com/bbcf/bbcfutils/blob/master/R/smoothData.R) 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. Differential 4C-seq data were computed using BamCompare from deeptools with binsize=50bp. Average of total Trans interactions between viewpoints and DSB were then computed using a 1Mb window around the breaks (80 best) and after exclusion of viewpoint-viewpoint (Cis) interactions.

Hi-C

Hi-C data obtained before and after DSB induction and upon CTRL or SCC1 depletion in DIvA cells were retrieved from[5]. Hi-C experiments with or without DSB induction and upon ATM or DNA-PK inhibition were performed in DIvA cells as in[5]. Briefly, 1 million cells were used per condition. Hi-C libraries were generated using the Arima Hi-C kit (Arima Genomics) by

following the manufacturer instructions. DNA was sheared to an average fragment size of 350-400 pb using the Covaris S220 and sequencing libraries were prepared on beads using the NEB Next 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.

Hi-C data analyses

Hi-C heatmaps. Hi-C reads were mapped to hg19 and processed with Juicer using default settings (https://github.com/aidenlab/juicer). Hi-C count matrices were generated using Juicer at multiple resolutions: 100 kb, 50 kb, 25 kb, 10 kb and 5 kb. Hi-C heatmaps screenshots were generated using Juicebox (https://github.com/aidenlab/Juicebox/wiki/Download). Aggregate heatmaps were computed on a set of sub-matrices extracted from originals observed Hi-C matrices at 50kb resolution or 100kb resolution. Region of 5Mb around DSBs (80 best) were extracted and then averaged. Log2 ratio was then computed using Hi-C counts (+DSB/-DSB) and plotted as heatmaps.

Cis Contacts Quantification. For *cis* contact quantification interaction within γH2AX domains (-0.5/+0.5Mb around 80 best DSBs) were extracted from the observed Hi-C matrix at 100kb resolution, and log2 ratio was computed on damaged vs undamaged Hi-C counts (+DSB/-DSB). Adjacent windows (-1.5Mb-0.5Mb and +0.5Mb-1.5Mb around 80 best DSBs) were retrieved to quantify interactions between damaged domains and adjacent undamaged domains. Boxplots: Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers. Significance was calculated using non-parametric Wilcoxon test.

Trans contact quantification. To determine interaction changes in *trans* (inter-chromosomal) we built the whole-genome Hi-C matrix for each experiment by merging together all chr-chr interaction matrices using Juicer and R. The result is a genome matrix with 33kx33k bin

interactions for 100kb resolution. Interactions between bins inside damaged TADs (240X240 for 80 DSBs) were extracted and counted for each condition, log2 ratio was calculated on normalized count (cpm), and plotted as boxplots. Boxplots: Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers.

TAD Cliques. TAD Cliques were computed using the igraph R package on an undirected graph representing DSB clustering. This graph was computed on the differential Hi-C matrix (+DSB/-DSB) counts, at 500 kb resolution, considering a change of ~86% of interaction (0.9 in log2) as between two DSBs as a node on the graph. Averaged signal of ChIP-seq values (53BP1/γH2AX/H1/Ubiquitin FK2) were then computed for each categories of cliques using 500kb windows around DSB. For prior RNAPII occupancy, the signal was computed on 10kb around DSBs.

A/B compartment. To identify the two mains chromosomal compartments (A/B), the extraction of the first eigenvector of the correlation matrix (PC1) was done on the Observed/Expected matrix at 500kb resolution using juicer eigenvector command. The resulting values were then correlated with ATAC-seq signal in order to attributes positives and negatives values to the A and B compartment, respectively, on each chromosomes. The Observed/Expected bins were arranged based on the PC1 values and aggregated into 21 percentiles, to visualize A-B interactions on our experiments (saddle plots).

D compartment. To identify the D compartment, we retrieved the first component (PC1) of a PCA made on the differential observed Hi-C matrix $log2\left(\frac{damaged}{undamaged}\right)$ at 100kb resolution. Each matrix was extracted from the .hic files using Juicer and the ratio was computed bin per bin. Pearson Correlation matrices were then computed for each chromosome, and PCA was applied on each matrix. The first component of each PCA was then extracted and correlated with the positions of DSB. A PC1 showing a positive correlation with DSB was then called D

compartment, and PC1 showing negative correlation with DSBs were multiplied by -1. We were able to extract the D compartment on chromosomes 1,17 and X for +DSB/-DSB and chromosomes 1,2,6,9,13,17,18,20 and X for +DSB/-DSB in DNA-PKi condition. D compartment (first component of the PCA) was converted into a coverage file using rtracklayer R package. Using the same package, D compartment value was computed around DSBs and genes at 100kb resolution, and plotted as boxplot. Boxplots: Centre line, median; box limits, first and third quartiles; whiskers, maximum and minimum without outliers; points, outliers.

Transcription factor motif analysis. TF-binding motifs were extracted on the promoter regions (-500bp/TSS) of genes with positive value of D compartment (2161) *vs* genes with negative value (2112) using motifmatchr and TFBSTools R packages on JASPAR2020 database. Motifs were sorted by significance using fisher exact test and adjusted with Benjamini-Hochberg procedure between motifs found on gene inside the D compartment versus genes outside D compartment.

Translocation breakpoints. For translocation breakpoints, data from[27] were retrieved, and only breakpoints for interchromosomal structural variant selected (N=28051). Genes reproducibly enriched in Compartment D in the three biological replicates, on chr1, 17 and X (N=604) as well as genes not enriched in Compartment D (N=1439) were retrieved. The significance of the overlap between genes and breakpoints was determined using the regioneR package[36] using resampling test with PermTest. Briefly, we selected 1000 times a control set of genes, with same size and on the same chromosome as our original gene set. We tested the overlap between each genes and breakpoints, to determine a distribution of the number of overlaps between control set and breakpoints. We further tested if the overlap between our gene set (D compartment or non D compartment) and breakpoints was significant, by counting the number of times we got more overlap in control than in our gene set.

Acknowledgments

We thank the genomics core facility of EMBL and of the I2BC (Centre de Recherche de Gif) for high-throughput sequencing. M.B. was supported by the CRUK Beatson Institute core grant A29252; A.B was supported by national productivity award from the MRC, MC_ST_U17040. Funding in GL laboratory was provided by grants from the European Research Council (ERC-2014-CoG 647344), Agence Nationale pour la Recherche (ANR-18-CE12-0015) and the Ligue Nationale contre le Cancer (LNCC). C.A was a recipient of a FRM fellowship (FRM FDT201904007941). T.C and N.P are INSERM researchers.

Authors contributions

C.A., E.L., T.C., and N.P. performed and analyzed experiments. V.R., and R.M, performed bioinformatic analyses of all high-throughput sequencing datasets. A.B performed the Amplicon-seq experiment under the supervision of M.B. D.N. helped to realize and analyze 4C-seq experiments. G.L. and T.C. wrote the manuscript. All authors commented and edited the manuscript.

Competing Interest

The authors declare no competing interest

Data Availability

All high-throughput sequencing data (Hi-C, 4C–seq, Amplicon-seq and RNA-seq) have been deposited to Array Express (https://www.ebi.ac.uk/arrayexpress/) under accession number E-MTAB-XXXX.

Code availability

Source codes are available from https://github.com/LegubeDNAREPAIR/

Figures Legends

Figure 1: Cohesin and ATM-dependent TAD reinforcement in response to DSBs.

(a) Hi-C contact matrix of the log2 (+DSB/-DSB) in 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 γH2AX 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 (log2 +DSB/-DSB)) within γH2AX 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. γH2AX 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 DIvA cells transfected with a control siRNA or a siRNA directed against SCC1.

Figure 2: Cell cycle regulated, ATM-dependent but cohesin- and DNA-PK-independent clustering of damaged-TADs.

(a) Hi-C contact matrix of the log2 (+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 log2 (+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 (log2 +DSB/-DSB), Ubiquitin chains detected with the FK2 antibody (log2 +DSB/-DSB) or pre-existing RNAPII (-DSB) within the corresponding domains were computed across each category.

(d) Left panel: Hi-C contact matrix of the log2(+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 log2 (+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) Genomics tracks showing differential 4C-seq (log2 (+DSB/-DSB)) (smoothed with a 10 kb span) obtained using a DSB located on chr20 as a viewpoint (red arrow), γH2AX 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.

(g) *Trans* interactions (log2 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.

Figure 3. Formation of a DSB-specific sub-compartment that ensures optimal activation of the DDR.

(a) Genomic tracks of γH2AX ChIP-seq and first Chromosomal eigenvector (CEV) computed on differential (+DSB/-DSB) Hi-C matrix on chromosome 1 (top panel) and chromosome X (bottom panel). Three biological replicate experiments are shown as well as the CEV obtained upon DNA-PK inhibition.

(b) Genomic tracks of γH2AX (red), H3K79me2 (black) and H3K4me3 (yellow) ChIP-seq, and the first Chromosomal Eigenvector computed on the differential Hi-C (CEV, blue). The brown

rectangles highlight genomic regions present in D sub-compartment that carry a DSB and are enriched in γH2AX. In contrast the blue rectangle shows a genomic region that is devoid in γH2AX and DSB, but is nevertheless found in the D sub-compartment.

(c) As in (a) but with a zoom on an undamaged region of the chromosome 1 that displayed positive D sub-compartment signal. The differential RNA-seq (log2 (+DSB/-DSB)) for this region containing the p53-target gene GADD45A is also shown (green).

(d) Boxplot showing the quantification of the D compartment signal computed from Hi-C data (+DSB+DNA-PKi/-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.

(e) 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.

Figure 4. 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≥3 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 DIvA cell line, upon Ctrl or SCC1 depletion (log2 siSCC1/siCTRL) (n=4 independent replicates). P-values, non-parametric wilcoxon test tested against μ=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 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.

Bibliography

1. Clouaire, T. & Legube, G. A Snapshot on the Cis Chromatin Response to DNA Double-Strand Breaks. *Trends Genet.* **35**, 330–345 (2019).
2. Rogakou, E. P., Pilch, D. R., Orr, A. H., Ivanova, V. S. & Bonner, W. M. DNA Double-stranded Breaks Induce Histone H2AX Phosphorylation on Serine 139. *J. Biol. Chem.* **273**, 5858–5868 (1998).
3. 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).
4. Collins, P. L. *et al.* DNA double-strand breaks induce H2Ax phosphorylation domains in a contact-dependent manner. *Nat. Commun.* **11**, 3158 (2020).
5. Arnould, C. *et al.* Loop extrusion as a mechanism for formation of DNA damage repair foci. *Nature* **590**, 660–665 (2021).
6. Caron, P. *et al.* Cohesin Protects Genes against γH2AX Induced by DNA Double-Strand Breaks. *PLoS Genet.* **8**, e1002460 (2012).
7. Sanders, J. T. *et al.* Radiation-induced DNA damage and repair effects on 3D genome organization. *Nat. Commun.* **11**, 6178 (2020).
8. Aten, J. A. *et al.* Dynamics of DNA Double-Strand Breaks Revealed by Clustering of Damaged Chromosome Domains. *Science* **303**, 92–95 (2004).
9. Aymard, F. *et al.* 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 (2017).
10. Roukos, V. *et al.* Spatial Dynamics of Chromosome Translocations in Living Cells. *Science* **341**, 660–664 (2013).

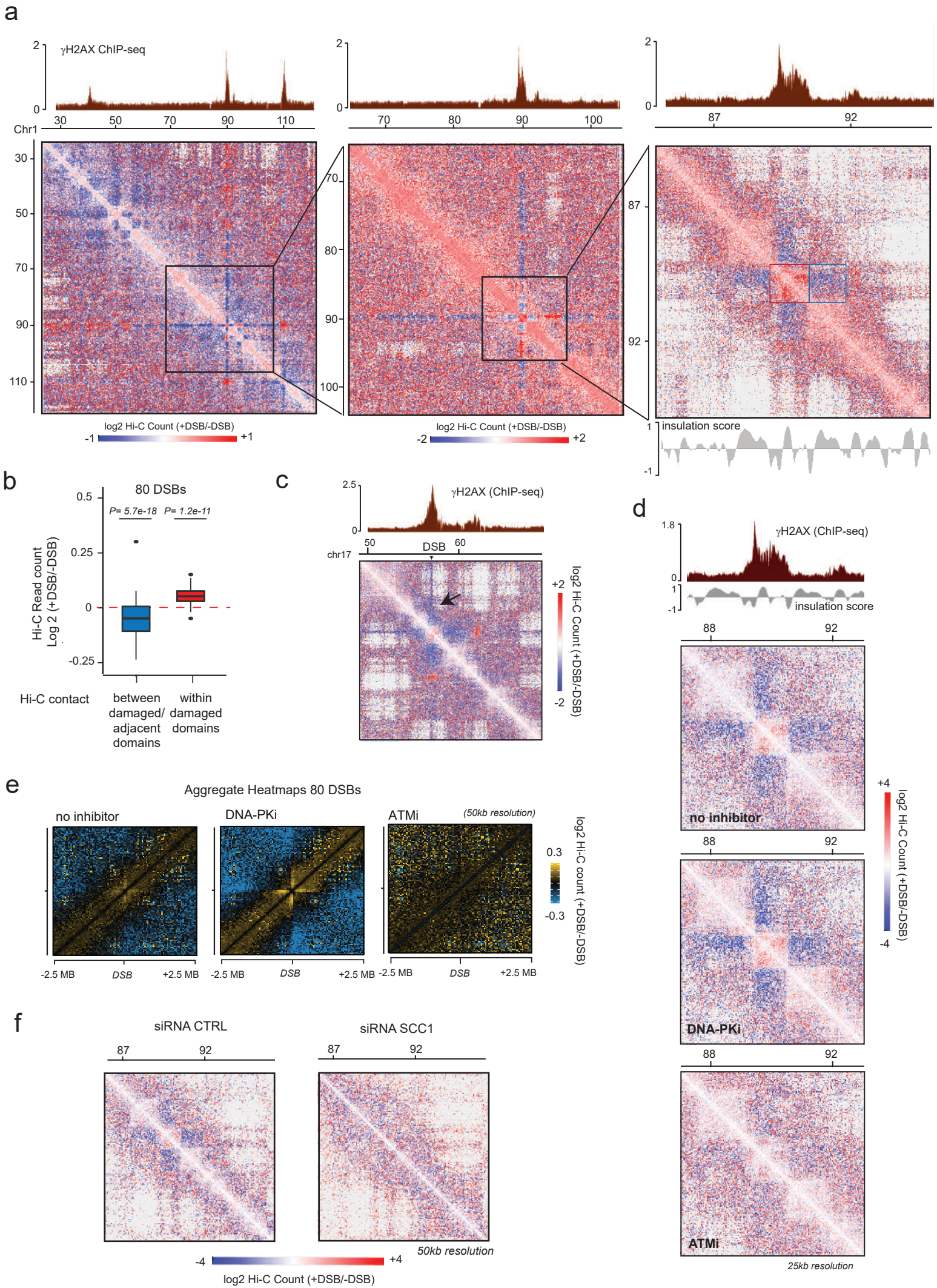
11. Lottersberger, F., Karssemeijer, R. A., Dimitrova, N. & de Lange, T. 53BP1 and the LINC Complex Promote Microtubule-Dependent DSB Mobility and DNA Repair. *Cell* **163**, 880–893 (2015).
12. Schrank, B. R. *et al.* Nuclear ARP2/3 drives DNA break clustering for homology-directed repair. *Nature* **559**, 61–66 (2018).
13. Kilic, S. *et al.* Phase separation of 53 BP 1 determines liquid-like behavior of DNA repair compartments. *EMBO J.* **38**, (2019).
14. Pessina, F. *et al.* Functional transcription promoters at DNA double-strand breaks mediate RNA-driven phase separation of damage-response factors. *Nat. Cell Biol.* **21**, 1286–1299 (2019).
15. Guénolé, A. & Legube, G. A meeting at risk: Unrepaired DSBs go for broke. *Nucleus* **8**, 589–599 (2017).
16. 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).
17. 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).
18. Fudenberg, G. *et al.* Formation of Chromosomal Domains by Loop Extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
19. Paulsen, J. *et al.* Long-range interactions between topologically associating domains shape the four-dimensional genome during differentiation. *Nat. Genet.* **51**, 835–843 (2019).
20. Lieberman-Aiden, E. *et al.* Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science* **326**, 289–293 (2009).
21. Rao, S. S. P. *et al.* A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* **159**, 1665–1680 (2014).

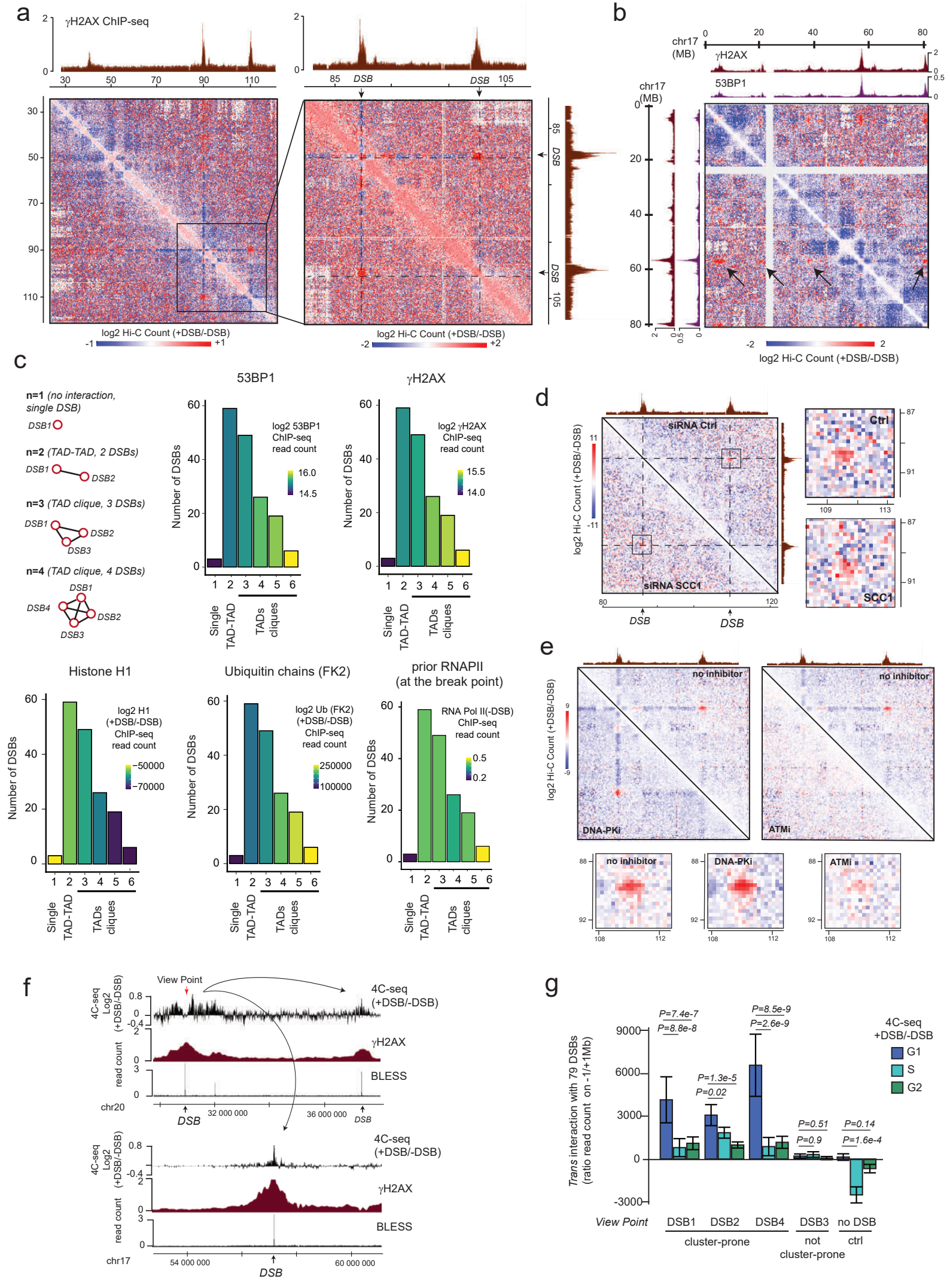
22. Chen, Y. *et al.* Mapping 3D genome organization relative to nuclear compartments using TSA-Seq as a cytological ruler. *J. Cell Biol.* **217**, 4025–4048 (2018).
23. Allocati, N., Di Ilio, C. & De Laurenzi, V. p63/p73 in the control of cell cycle and cell death. *Exp. Cell Res.* **318**, 1285–1290 (2012).
24. Fouad, S., Hauton, D. & D'Angiolella, V. E2F1: Cause and Consequence of DNA Replication Stress. *Front. Mol. Biosci.* **7**, 599332 (2021).
25. Huang, Q. *et al.* Identification of p53 regulators by genome-wide functional analysis. *Proc. Natl. Acad. Sci.* **101**, 3456–3461 (2004).
26. Bowen, C. & Gelmann, E. P. NKX3.1 Activates Cellular Response to DNA Damage. *Cancer Res.* **70**, 3089–3097 (2010).
27. Zhang, Y. *et al.* A Pan-Cancer Compendium of Genes Deregulated by Somatic Genomic Rearrangement across More Than 1,400 Cases. *Cell Rep.* **24**, 515–527 (2018).
28. Spegg, V. & Altmeyer, M. Biomolecular condensates at sites of DNA damage: More than just a phase. *DNA Repair* **106**, 103179 (2021).
29. Cuella-Martin, R. *et al.* 53BP1 Integrates DNA Repair and p53-Dependent Cell Fate Decisions via Distinct Mechanisms. *Mol. Cell* **64**, 51–64 (2016).
30. Ghodke, I. *et al.* AHNAK controls 53BP1-mediated p53 response by restraining 53BP1 oligomerization and phase separation. *Mol. Cell* **81**, 2596-2610.e7 (2021).
31. Drané, P. *et al.* TIRR regulates 53BP1 by masking its histone methyl-lysine binding function. *Nature* **543**, 211–216 (2017).
32. Parnandi, N. *et al.* TIRR inhibits the 53BP1-p53 complex to alter cell-fate programs. *Mol. Cell* **81**, 2583-2595.e6 (2021).
33. Izhar, L. *et al.* A Systematic Analysis of Factors Localized to Damaged Chromatin Reveals PARP-Dependent Recruitment of Transcription Factors. *Cell Rep.* **11**, 1486–1500 (2015).

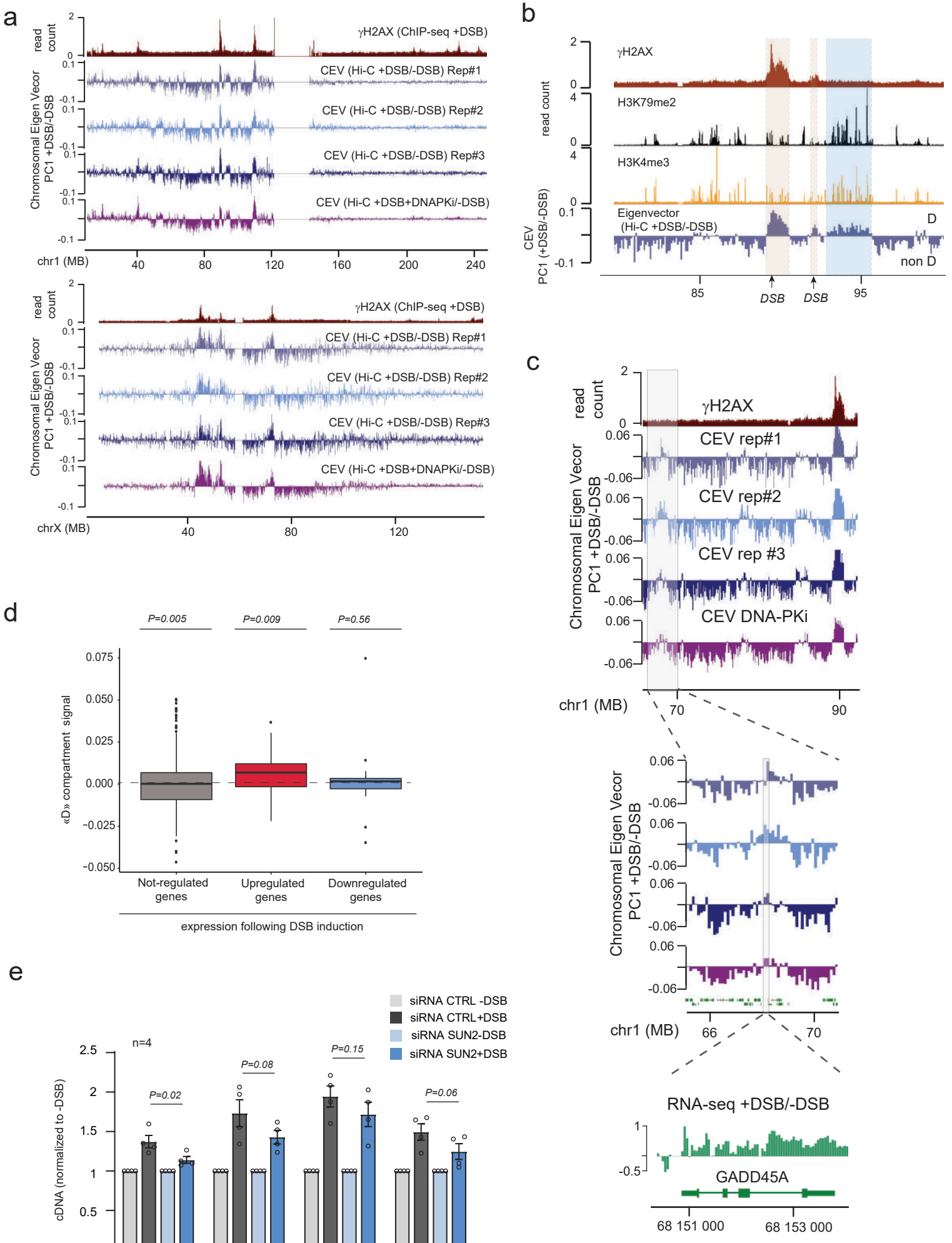
34. Aymard, F. *et al.* Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nat. Struct. Mol. Biol.* **21**, 366–374 (2014).

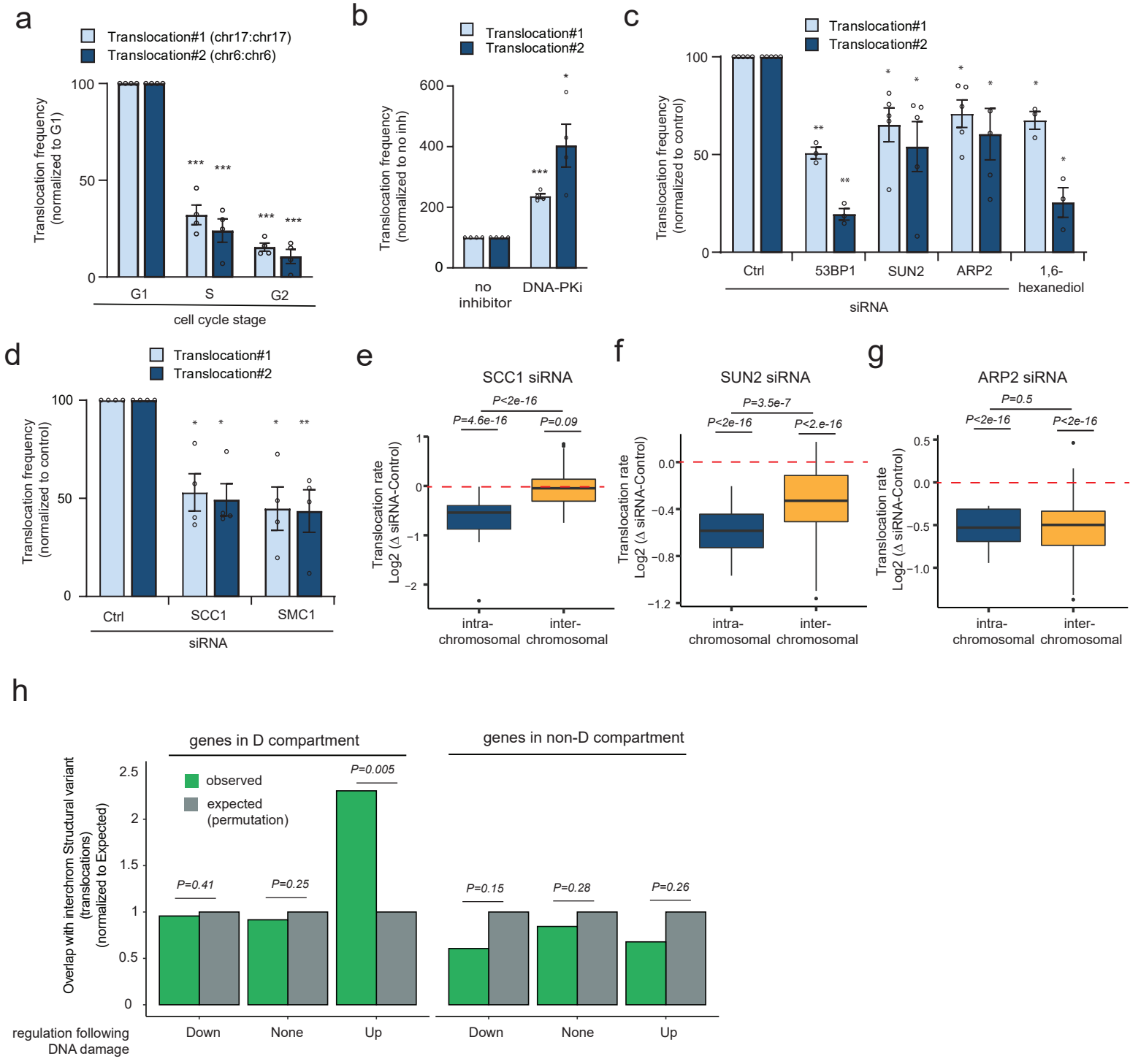
35. Cohen, S. *et al.* Senataxin resolves RNA:DNA hybrids forming at DNA double-strand breaks to prevent translocations. *Nat. Commun.* **9**, 533 (2018).

36. Gel, B. *et al.* regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics* btv562 (2015) doi:10.1093/bioinformatics/btv562.









3.5 G4s as novel promoters and G4 SNPs

Until recently, DNA structures alternative to B-DNA, such as G4s, were mainly studied in vitro. However, recent advances in the genome-wide mapping in vivo of G4s have shown their essential role in processes such as transcription, replication and DNA repair [Hänsel-Hertsch *et al.* 2016, Hänsel-Hertsch *et al.* 2020, Marnef *et al.* 2017, Spiegel *et al.* 2021]. However, the link between the structure of non-B DNA and their function has only started to be revealed.

In a collaboration with Cyril Esnault and Jean-Christophe Andrau, we demonstrated that G-quadruplexes can act as promoter elements and chromatin organizers (article submitted). Most notably, we discovered that G4s are located at the deepest point of nucleosome exclusion at promoters, and we found that they correlate with maximum promoter activity. Moreover, G4s exclude nucleosomes not only at promoters but also at intergenic areas, and are associated with a strong nucleosome positioning potential. Importantly, G4 stabilisation results in global reduction of proximal promoter pausing and +1 nucleosome barrier, suggesting a role in RNA Polymerase II pausing regulation.

In addition, using genetic analyses of SNPs that are known to affect gene expression (eSNPs) from GTEx and TGCA databases, we could assess the influence of predicted G4s in promoting transcriptional activity. We found that SNPs increasing predicted G4 propensities in promoters, as defined upstream of the TSS, did also increase accordingly the expression of the target genes, as compared to SNPs decreasing predicted G4 propensities. Remarkably, this observation did not hold true for SNPs altering G4s downstream the TSS (control SNPs outside promoters). These results thus supported the role of G4 as promoter elements, similarly to the classical definition of promoters as a combination of transcription factor binding sites.

3.6 Machine and deep learning for genomics

3.6.1 PredDSB: Predicting double-strand DNA breaks using epigenome marks or DNA

DNA double-strand breaks (DSBs) result from the attack of both DNA strands by multiple sources, including radiation and chemicals. Recent techniques allow the genome-wide mapping of DSBs at high resolution, enabling the comprehensive study of their origins. Several high-throughput sequencing techniques have been developed, such as BLESS [Crosetto *et al.* 2013], GUIDE-seq [Tsai *et al.* 2015], END-seq [Canela *et al.* 2016] and DSBCapture [Lensing *et al.* 2016]. One of the most recent techniques, DSBCapture, allowed to map more than 80 thousand endogenous DSBs at a resolution lower than 1 kb in human. To date, DSBs have been mapped at high resolution only for a few number of cell lines due to high sequencing costs and experimental difficulties. This has prevented the

comprehensive study of the double-strand break landscape in the human genome across diverse cell lines and tissues.

There is a strong link between DSB occurrence and chromatin landscape. DSBs and associated DNA repair mechanisms are linked to epigenetic marks, including H3K4me1/2/3 histone modifications and chromatin accessibility [Lensing *et al.* 2016], as well as the concentration of repair proteins at the sites of breaks [Kinner *et al.* 2008, Price & D'Andrea 2013]. If there is a strong link between DSBs and chromatin, then the mapping of DSBs along the genome can be computationally predicted using the huge amount of publicly available chromatin data for cell lines [The ENCODE Consortium 2012] and tissues [Consortium 2017]. Moreover, a computational approach would demonstrate the extent to which histone modifications or DNA patterns allow to predict and regulate the cellular response to double-stranded breaks.

Hence, I devised a computational approach based on random forests to predict DSBs using the epigenomic and chromatin context [Mourad *et al.* 2018]. This was the first demonstration that endogenous DSBs can be computationally predicted given the epigenomic and chromatin context. The predictions achieved excellent accuracy (AUROC>0.97) at high resolution (<1kb) using available ChIP-seq and DNase-seq data from public databases (Figure 3a from the article "Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution" below). DNase, CTCF binding and H3K4me1/2/3 were among the best predictors of DSBs, reflecting the importances of chromatin accessibility, activity and long-range contacts in determining DSB sites and subsequent repairing (Figure 3b from the article below). Since CTCF binding and chromatin marks are known to be computationally predictable from the DNA sequence, the proposed model was also used to predict DSB sites directly from the DNA sequence using DNA motif occurrences and DNA shape. The model could predict DSB sites using DNA sequence only (AUROC = 0.838), reflecting the contribution of TFBS motifs, including CTCF but also AP-1 protein complex, tumor proteins p53, p63 and p73, and the contribution of DNA shapes (Figure 7 from the article below).

Mourad *et al. Genome Biology* (2018) 19:34
https://doi.org/10.1186/s13059-018-1411-7

Genome Biology

METHOD Open Access

Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution

Raphaël Mourad[1*], Krzysztof Ginalski[2], Gaëlle Legube[3] and Olivier Cuvier[1]

Abstract

Double-strand breaks (DSBs) result from the attack of both DNA strands by multiple sources, including radiation and chemicals. DSBs can cause the abnormal chromosomal rearrangements associated with cancer. Recent techniques allow the genome-wide mapping of DSBs at high resolution, enabling the comprehensive study of their origins. However, these techniques are costly and challenging. Hence, we devise a computational approach to predict DSBs using the epigenomic and chromatin context, for which public data are readily available from the ENCODE project. We achieve excellent prediction accuracy at high resolution. We identify chromatin accessibility, activity, and long-range contacts as the best predictors.

Keywords: Double-strand breaks, Epigenetics, Chromatin, Machine learning

Background

Double-strand breaks (DSBs) arise when both DNA strands of the double helix are severed. DSBs are caused by the attack of deoxyribose and DNA bases by reactive oxygen species and other electrophilic molecules [1]. DSBs are particularly hazardous to a cell because they can lead to deletions, translocations, and fusions in the DNA, collectively referred to as chromosomal rearrangements [2]. DSBs are most commonly found in cancer cells. Several high-throughput sequencing techniques have been developed for the genome-wide mapping of DSBs in situ such as BLESS [3], GUIDE-seq [4], END-seq [5], and DSBCapture [6]. One of the most recent techniques, DSBCapture, was used to map more than 80 000 endogenous DSBs at a resolution lower than 1 kb in human. To date, DSBs have been mapped at high resolution only for a few cell lines due to the high sequencing costs and experimental difficulties. This has prevented the comprehensive study of the DSB landscape in the human genome across diverse cell lines and tissues.

Chromatin immunoprecipitation followed by high-throughput DNA sequencing (ChIP-seq) and DNase I hypersensitive site sequencing (DNase-seq) data are publicly available for dozens of cell lines and tissues from the ENCODE [7] and Roadmap Epigenomics [8] projects. On the one hand, recent studies have shown that the mapping of regulatory elements such as enhancers and promoters can be accurately predicted using available epigenome and chromatin data [9, 10]. Other studies have shown that the epigenome can be predicted by combinations of DNA motifs and DNA shape [11–14]. On the other hand, DSBs and the resulting DNA repair mechanisms have been shown to be linked to epigenome marks, including H3K4me1/2/3 and chromatin accessibility [6]. Accordingly, PRDM9-mediated trimethylation of H3K4 (H3K4me3) was originally shown to play a critical role in regulating DSBs associated with meiotic recombination hotspots [15–17]. Moreover, the repair of DSBs involves both post-translational modification of histones, in particular γ-H2AX, and concentration of DNA-repair proteins at the site of damage [18, 19]. It remains unclear to what extent DNA motifs or histone modifications predict or regulate the cellular response to DSBs in other developmental stages. Here, we thus sought to test whether publicly available epigenome and chromatin data, or DNA motifs and shape, could be used to predict DSBs.

In this article, we demonstrate, for the first time, that endogenous DSBs can be computationally predicted using

*Correspondence: raphael.mourad@ibcg.biotoul.fr
[1]LBME, Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 118, route de Narbonne, 31062 Toulouse, France
Full list of author information is available at the end of the article

the epigenomic and chromatin context, or using DNA sequence and DNA shape. Our predictions achieve excellent accuracy (area under the receiver operating characteristic curve or AUROC > 0.97) at high resolution (< 1 kb) using available ChIP-seq and DNase-seq data from public databases. Despite the highly imbalanced data when predicting DSBs genome-wide, our approach detects a reasonable number of false positives (area under the precision–recall curve or AUPR = 0.459). DNase, CTCF binding, and H3K4me1/2/3 are among the best predictors of DSBs, reflecting the importance of chromatin accessibility, activity, and long-range contacts in determining DSB sites and subsequent repairing. We also successfully predict DSB sites using DNA motif occurrences only (AUROC = 0.839) and identify the CTCF motif as a strong predictor. In addition, DNA shape analysis further reveals the importance of the structure-based readout in determining DSB sites, complementary to the sequence-based readout (motifs).

Results and discussion

Double-strand break prediction approach

Our computational approach for predicting DSBs is schematically illustrated in Fig. 1. In the first step, we analyzed public DSBCapture data from Lensing el at. [6], which is the most sensitive and accurate genome-wide mapping of DSBs to date (Fig. 1a). DSBCapture captures DSBs in situ and it can directly map them at single-nucleotide resolution. DSBCapture peaks were called with less than 1-kb resolution (median size of 391 bases). The DSBCapture peaks obtained from two biological replicates were intersected to yield more reliable DSB sites. Endogenous breaks were captured for normal human epidermal keratinocytes (NHEKs), for which numerous ChIP-seq and DNase-seq data are publicly available from the ENCODE project [7]. In the second step, we integrated and mapped different types of data within DSB sites and non-DSB sites. To prevent bias effects, non-DSB sites were randomly drawn from the human genome with sizes, GC, and repeat contents similar to those of DSB sites [20] (Fig. 1b). ChIP-seq and DNase-seq peaks in NHEKs, as obtained from the ENCODE project, were mapped to corresponding DSB and non-DSB sites [7]. We also mapped p63 ChIP-seq peaks from keratinocytes [21]. We further searched for potential protein-binding sites at DSB and non-DSB sites using motif position weight matrices from the JASPAR 2016 database [22], and predicted DNA shape at DSB and non-DSB sites using Monte Carlo simulations [23]. In the third step, a random forest classifier was built to discriminate between DSB sites and non-DSB sites based on epigenome marks or DNA (Fig. 1c). Random forest variable importance values were used to estimate the predictive importance of a feature. We also compared random forest predictions with another popular method,

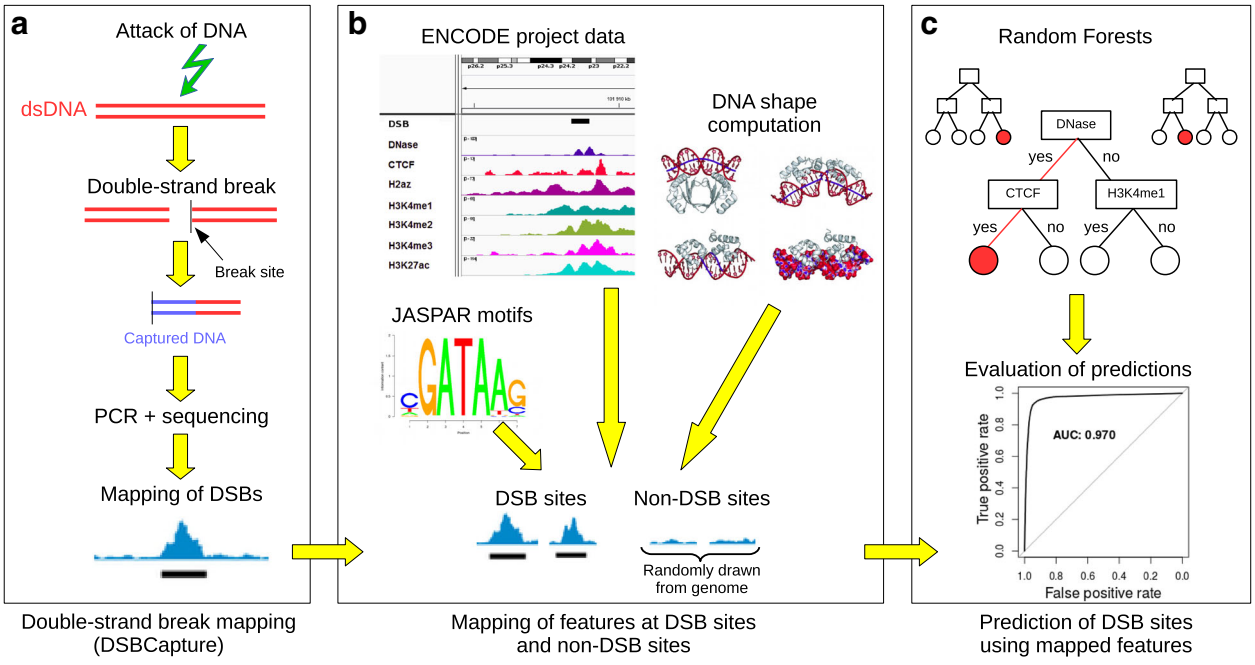


Fig. 1 Double-strand break (DSB) prediction using epigenome marks or DNA. The prediction approach has three steps. **a** Mapping of DSBCapture sequencing data and DSB peak calling. **b** Mapping of features at DSB and non-DSB sites. Features include epigenomic and chromatin data from the ENCODE project, DNA motifs from the JASPAR database, and DNA shape predictions. **c** Prediction of DSB sites using features. AUC area under the curve, ds double strand, DSB double-strand break, PCR polymerase chain reaction

lasso logistic regression [24]. Using lasso regression, we assessed the positive, negative, or null contribution of a feature to DSBs. We then split the DSB dataset into a training set to learn model parameters by cross-validation, and into a testing set to compute the receiver operating characteristic (ROC) and precision–recall (PR) curves, as well as AUROC and AUPR, to evaluate prediction accuracy.

Double-strand breaks are enriched with epigenome marks and DNA motifs

We first sought to assess comprehensively the link between DSBs and epigenome marks or DNA motifs. As previously shown [6, 25], several epigenomic and chromatin marks colocalized at DSBs (Fig. 2a). Among the most enriched marks were DNase I hypersensitive sites, H3H4 methylation, and CTCF (Fig. 2b). For instance, 91% of DSBs colocalized to a DNase site, whereas this percentage dropped to 11% for non-DSB regions. This corresponded to an odds ratio (OR) of 89.3. Similarly, high enrichment was found for H3K4me2 (74% versus 11%; OR = 22.4) and for the insulator protein CTCF (25% versus 2%; OR = 19), which may involve its interactions with the insulator-related cofactor cohesin, which has been shown to protect genes from DSBs [26]. As such, DSBs mostly localized within open and active regions that were often implicated in long-range contacts [27]. Interestingly, DSBs also colocalized with tumor protein p63 binding (19.4% versus 1%; OR = 23.8), a member of the p53 gene family [28, 29]. In addition, we could distinguish DNase and CTCF sites that were enriched at the center of DSBs from histone marks that were found at the edges of DSB sites (Fig. 2c). Therefore, the strong enrichment of epigenomic and chromatin marks at DSB sites suggests that DSB regions could be accurately predicted using available ChIP-seq and DNase-seq data from public databases, including ENCODE and Roadmap Epigenomics.

Previous enrichment analyses of DNA-binding proteins were limited by the ChIP-seq data available. Hence, we sought DNA motifs that may be enriched at DSB sites as a way to obtain a more comprehensive list of candidate DNA-binding proteins. Of the 454 available motifs from the JASPAR 2016 database, 134 were significantly enriched ($p < 0.05$, Bonferroni correction), indicating that DSBs were associated with a large number of protein-binding sites (Fig. 2d). Among the most enriched and frequent motifs, we identified numerous motifs specifically recognized by protein cofactors of the transcription factor complex AP-1. This included JUND (OR = 1.40, 12% of DSBs), JUNB (OR = 1.27, 19% of DSBs), the heterodimer BATF::JUN (OR = 1.31, 10% of DSBs), and also FOS (OR = 1.37, 20% of DSBs), FOSL1 (OR = 1.37, 17% of DSBs), and FOSL2 (OR = 1.27, 18% of DSBs). Among the most enriched but less frequent motifs, we expectedly found CTCF (OR = 1.54, 1.7% of DSBs), as well as members of the tumor protein family p53, i.e., p53 itself (OR = 1.54, 0.2% of DSBs), p63 (OR = 1.49, 0.3% of DSBs), and p73 (OR = 1.54, 0.1% of DSBs) [28, 29]. Such enrichment of DNA motifs at DSB sites, therefore, supports that DNA sequence can alone predict some of the DSBs encountered.

Prediction using epigenomic and chromatin data

Given the strong link between DSBs and epigenomic and chromatin marks, we sought to build a classifier to discriminate DSB sites from non-DSB sites based on the presence or absence of such marks. For this, we used random forests, which are very efficient classifiers for predicting a feature. They can capture non-linear and complex interaction effects [30]. We split the data into a training set to learn model parameters and a testing set to evaluate prediction accuracy. Using this classifier, we obtained excellent predictions of DSBs based on the epigenomic and chromatin marks available (AUROC = 0.970 and AUPR = 0.985; Fig. 3a; Additional file 1: Figure S1). Bootstrap analysis of 2000 replicates revealed that these predictions were very robust (95% confidence interval, CI, of AUROC: [0.968,0.972]). We also computed the variable importance (VI), which reflects the importance of a mark as a predictor (Fig. 3b). Among the marks, DNase showed the highest variable importance (VI = 0.180), reflecting the known higher chromatin accessibility after DNA damage [19] or the involvement of chromatin-remodeling complexes in DSB processing [31]. Other good predictors were CTCF (VI = 0.042), p63 (VI = 0.031), H3K4me1 (VI = 0.028), H3K4me2 (VI = 0.019), H3K4me3 (VI = 0.012), and H3K27ac (VI = 0.010), highlighting the roles of active chromatin, but also long-range contacts and DNA damage response in predicting DSB sites.

A drawback of variable importance lies in its inability to distinguish between the positive or negative contribution of the predictive mark on DSBs. For this reason, we also used lasso logistic regression to predict DSBs [24]. With this second model, we obtained excellent predictions, although slightly less accurate (AUROC = 0.967, $CI_{95\%}$: [0.966,0.971]; AUPR = 0.982; Additional file 1: Figure S2). From lasso regression, we could assess the positive or negative contributions of the predictive marks using beta coefficients (Fig. 3c). We also performed logistic regression without any regularization and obtained very similar coefficients (Additional file 1: Figure S3). This allowed us to compute p values associated with the coefficients. We found that all variables, except H3K79me2, H3K9ac, and H4K20me1, were significantly associated with DSBs (Additional file 1: Table S1). We identified positive predictive contributions of DNase, CTCF, p63, H3K4me1, and H3K4me2 marks, as previously revealed by enrichment analysis. We also uncovered negative predictive contributions of H3K9ac, H3K36me3, and H3K79me2.

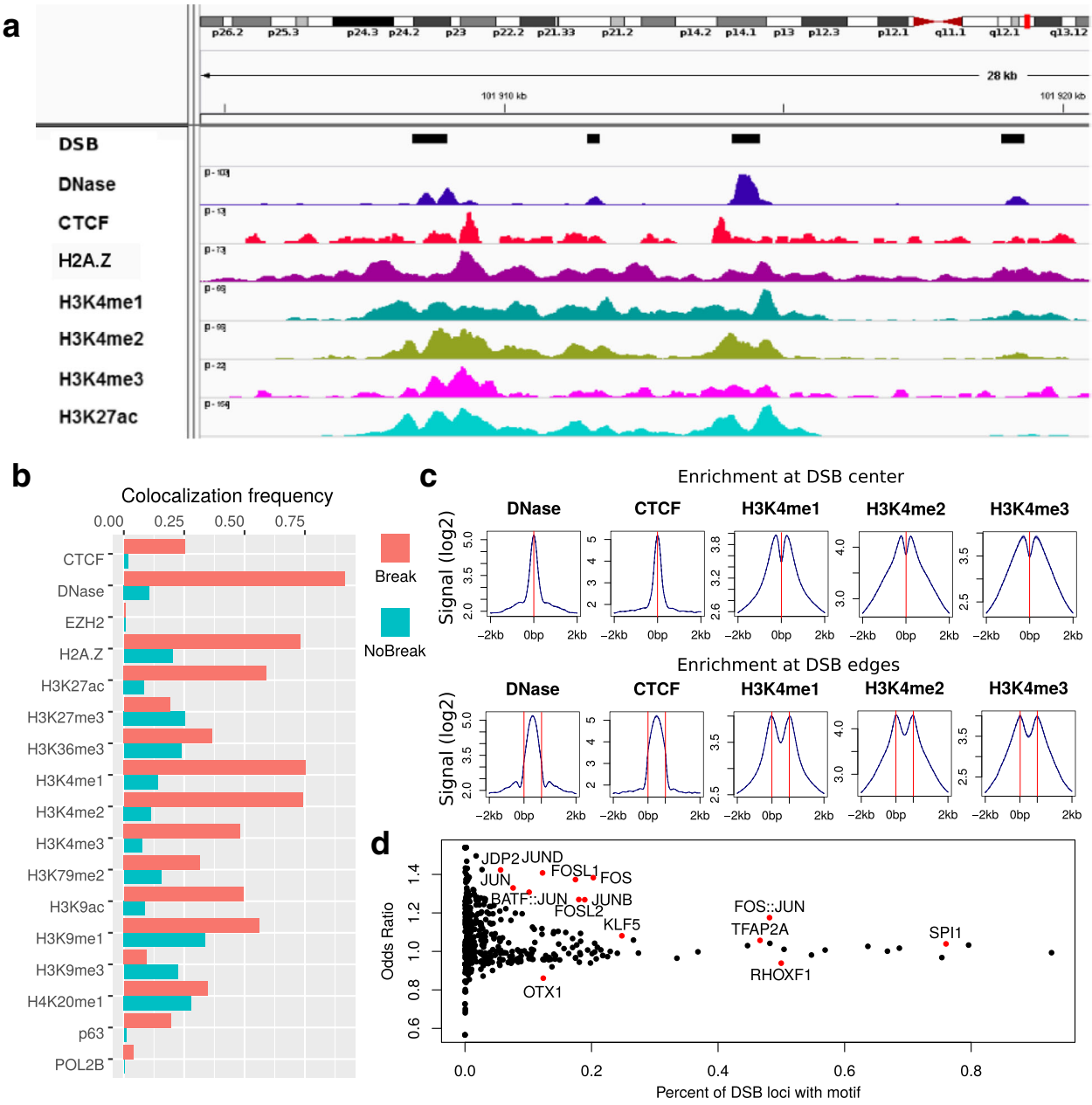


Fig. 2 Epigenomic, chromatin, and DNA motif profiles of double-strand breaks (DSBs). **a** A genome browser view of DSBs with histone marks, chromatin openness (DNase-seq), and DNA-binding proteins. **b** Colocalization frequencies of epigenomic marks and DNA-binding proteins at DSB sites, compared to non-DSB sites. **c** Average profiles of epigenomic marks and DNA-binding proteins at DSB sites. **d** Enrichment of DNA motifs at DSB sites, as measured by the odds ratio and the percentage of DSB loci with a motif. DSB double-strand break

In agreement, H3K9ac was shown to be rapidly and reversibly reduced in response to DNA damage [32]. Moreover, H3K36me3 may negatively impede DSBs by restricting chromatin accessibility through nucleosome positioning [33] or more directly by favoring the repair of DSBs [34].

We next sought to build a classifier using only one or two epigenomic marks, because this may be able to predict DSB sites even for cells for which only a few data points are available. We found that DNase I sites alone were sufficient to achieve good prediction accuracy (AUROC = 0.919 and AUPR = 0.962; Fig. 3d; Additional file 1: Figure S4), whereas H3K4me2 was not sufficient (AUROC = 0.816 and AUPR = 0.907; Fig. 3d; Additional file 1: Figure S4). Combinations of DNase with H2A.Z or H3K4me1 yielded very accurate predictions (AUROC = 0.952 and AUPR = 0.977; AUROC = 0.951 and AUPR = 0.976, respectively; Fig. 3d; Additional file 1: Figure S4), close to the model

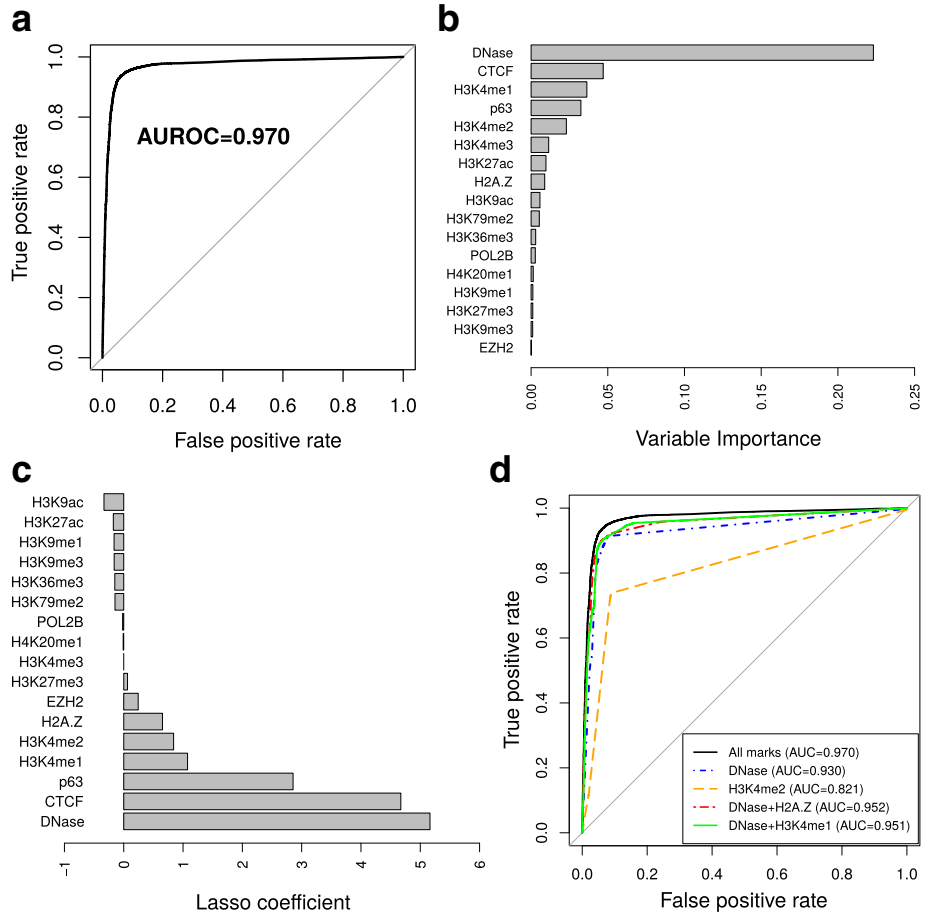


Fig. 3 Prediction of double-strand breaks using epigenomic and chromatin data with random forests. **a** Receiver operating characteristic curve for the prediction of double-strand breaks. Area under the ROC curve (AUROC) is plotted. **b** Variable importance of epigenomic and chromatin variables. **c** Lasso logistic regression coefficients. **d** Different predictive models including all variables, DNase only, H3K4me2 only, DNase+H2A.Z, or DNase+H3K4me1. AUROC area under the receiver operating characteristic curve

including all marks. Because DNase was a strong predictor, we explored where DNase was absent at DSBs to identify other marks that could be predictive here. We thus built a classifier using only DSBs that did not overlap any DNase site. DSB sites were still predicted well (AUROC = 0.869 and AUPR = 0.792; Additional file 1: Figure S5a and S5b), and CTCF and H3K4me1 were the most highly predictive variables (Additional file 1: Figure S5c). This revealed enhancer looping as a major driver of DSBs, in agreement with recent studies showing that DSBs form at loop anchors [35] and that CTCF facilitates DSB repair [36]. These results demonstrate that DSBs can be accurately predicted at less than 1-kb resolution using just a small amount of data.

Comparison with BLESS experiment and validation using an independent dataset

We then compared previous DSB predictions with DSBs identified by BLESS experiments [3, 6]. We also included in the comparison DSBCapture DSBs as the gold standard because of its higher sensitivity compared to BLESS: 84 821 DSBs were found by DSBCapture compared to 18 510 DSBs found by BLESS [6]. We first looked at predicted DSB sites surrounding the two genes MYC and MAP2K3 (Fig. 4a). For MYC, random forests correctly identified the four DSBs that were detected by DSBCapture, but erroneously predicted one DSB (yellow circle), whereas BLESS identified only one DSB out of four. For MAP2K3, random forests successfully predicted all DSBs detected by DSBCapture, whereas BLESS identified only three DSBs out of 11.

We then compared predictions with BLESS at the genome-wide level (Fig. 4b). We observed that random forests correctly predicted 18 084 out of 18 510 DSB sites (97.70%) found by BLESS, while it also successfully identified an additional 63 587 out of 66 591 DSB sites (95.49%) found by DSBCapture that were not detected by BLESS. The model misclassified only 1552 out of 83 225 predicted

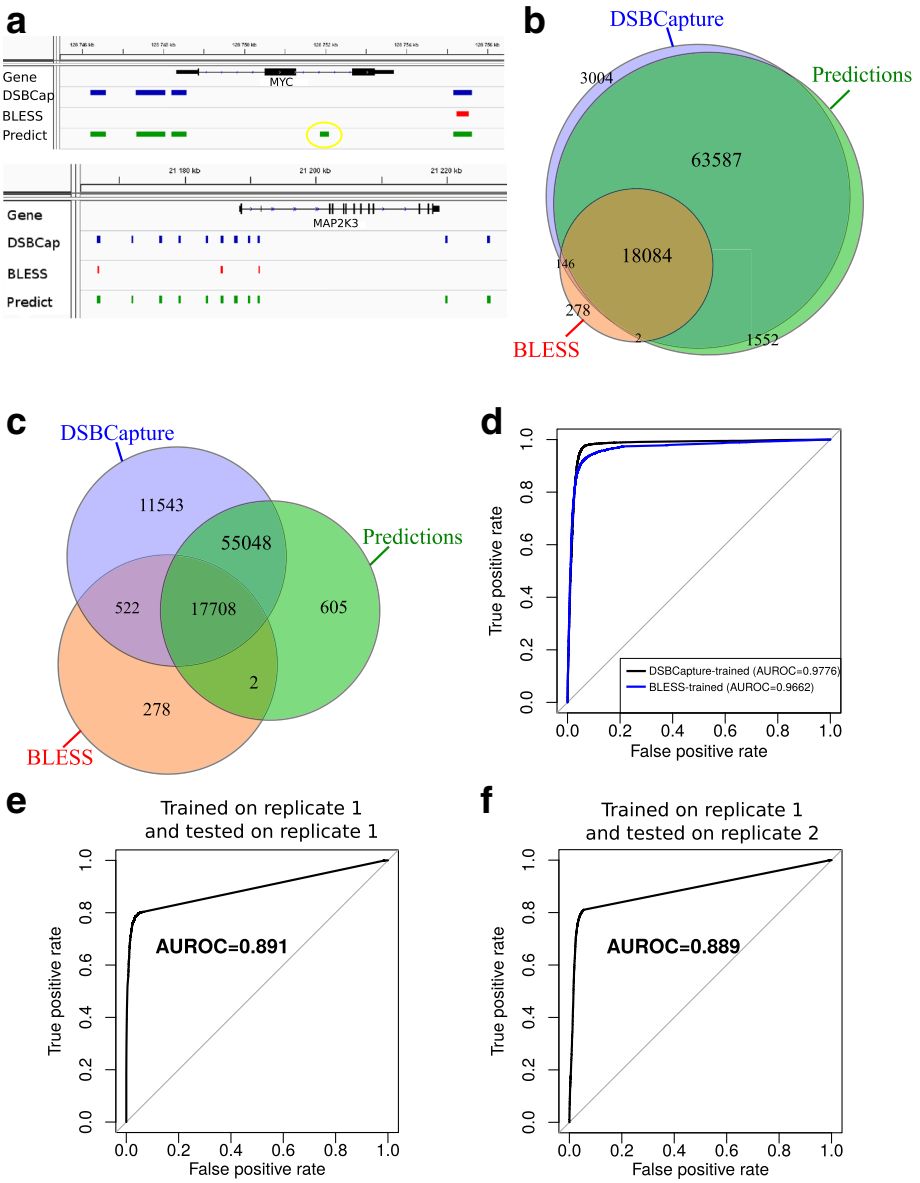


Fig. 4 Comparison of predicted and BLESS double-strand breaks (DSBs) and validation with an independent dataset. **a** Comparison for the MYC and MAP2K3 genes. **b** Venn diagram illustrating the overlaps between DSBCapture, random forest DSBCapture-trained model predictions, and BLESS DSBs. **c** Venn diagram illustrating the overlaps between DSBCapture, random forest BLESS-trained model predictions, and BLESS DSBs. **d** Comparison of receiver operating characteristic (ROC) curves between DSBCapture-trained and BLESS-trained models. Areas under the ROC curves (AUROCs) are plotted. **e** ROC curve for the prediction of DSBs trained on replicate 1 and tested on the same replicate. **f** ROC curve for the prediction of DSBs trained on replicate 1 and tested on replicate 2. AUROC area under the ROC curve, DSB double-strand break, ROC receiver operating characteristic

DSB sites (1.86%). However, this previous prediction comparison should be carefully interpreted, because the model was learned from DSBCapture and then used to predict DSBCapture and BLESS DSBs.

To demonstrate the power of model-based predictions further, we devised another computational experiment, which consisted of training the model with BLESS DSBs and then predicting DSBCapture DSBs to test if the model could predict DSBCapture DSBs that were not detected by BLESS. Very interestingly, we found that the model was able to predict an additional 55 048 out of 84 821 DSBs (64.90%) that were detected by DSBCapture but not by BLESS, and it identified only 605 DSBs out of 73 363 predicted DSBs (0.82%), which may be false positives not detected by DSBCapture and BLESS (Fig. 4c).

We then sought to compare models learned using DSBCapture and BLESS DSBs with a fair benchmark. For this, we devised the following strategy. A first model was

learned from DSBCapture and was used to predict BLESS DSB sites (the DSBCapture-trained model), and a second model was learned from BLESS and was used to predict DSBCapture DSB sites (the BLESS-trained model). We found that both models had very good prediction performance ($AUROC_{model1} = 0.9776$ and $AUPR_{model1} = 0.971$; $AUROC_{model2} = 0.9662$ and $AUPR_{model2} = 0.983$; Fig. 4d; Additional file 1: Figure S6).

In the previous section, we evaluated the accuracy of model predictions using a testing dataset that was from the same data as the training data (DSBs that overlapped between two replicates were split into a training dataset and a testing dataset). Here, we assessed model predictions by training random forests on one biological replicate and by testing prediction accuracy on a second biological replicate. For this, we used the two available DSBCapture biological replicates [6]. Accordingly, we used ENCODE epigenomic and chromatin data for which two biological replicates were available: DNase, CTCF, H3K4me3, H3K27me3, and H3K36me3. The first (respectively, second) replicates of the ENCODE data were associated with the first (respectively, second) DSBCapture replicate. Using only those five DNase-seq and ChIP-seq items, the model that was learned with the first replicate achieved accurate predictions on the testing data from the first replicate (AUROC = 0.891 and AUPR = 0.906; Fig. 4e; Additional file 1: Figure S7a). Note that the observed lower accuracy compared to that in the previous section (Fig. 3a,d) can be explained by the small amount of available epigenomic and chromatin data, and the lower reliability of DSBs identified using only one DSBCapture replicate. To validate the model on an independent dataset, we predicted DSBs from the second replicate using the model trained on the first replicate together with DNase-seq and ChIP-seq data for the second replicate. We obtained accurate predictions close to that obtained for the first replicate (AUROC = 0.889 and AUPR = 0.913; Fig. 4f; Additional file 1: Figure S7b). These accurate predictions demonstrate that using a classifier trained with epigenome and chromatin data is a reliable strategy for predicting DSBs.

The impact of controls on prediction

To assess if the high predictive accuracy of the model was inflated due to the way we selected non-DSB sites (the negative class), we devised different strategies. We first focused on gene promoters and built a random forest classifier to discriminate between promoters with DSBs (16 801 sites) and promoters without (48 838 sites). As previously done, we computed the ROC curve but we also included the PR curve to account for class imbalance. We obtained very good performance for both the ROC curve (AUROC = 0.941; Fig. 5a) and the PR curve (AUPR = 0.860; Fig. 5b). Second, we built a classifier to discriminate between gene bodies with DSBs (2187 sites) and gene bodies without (34 573 sites). We also obtained a very good ROC curve (AUROC = 0.943; Fig. 5c), but with a lower PR curve because of the higher class imbalance in gene bodies (AUPR = 0.538; Fig. 5d). Third, we built a classifier to discriminate between enhancers with DSBs (7373 sites) and enhancers without (38 521 sites). We again observed a very good ROC curve (AUROC = 0.933; Fig. 5e) and good PR (AUPR = 0.705; Fig. 5f). Fourth, we evaluated predictions over the whole genome in an unbiased way. For this, we split the genome into 250-base bins. Then we built a classifier to discriminate between bins with DSBs (189 132 bins) and bins without (11 362 262 bins). Using this approach, we obtained very good ROC accuracy (AUROC = 0.967) but with lower PR accuracy (AUPR = 0.459) due to the high class imbalance, revealing a high number of false positives detected genome-wide by our method. We concluded that the excellent accuracy of model-based predictions was not inflated due to the way non-DSB sites were selected over the genome.

Prediction in another cell type

To validate our model-based predictions further, we used the random forest learned from DSBs in one cell type (NHEK) to predict DSBs in another cell type (U2OS). For this, we used data that were available for both NHEK and U2OS cells: DNA-seq, CTCF, H3K4me1/3, H3K9me3, H3K27ac, H3K27me3, H3K36me3, and POL2B. The validation is illustrated in Additional file 1: Figure S8. In summary, we trained a random forest with DSBCapture DSBs and DNase-seq and ChIP-seq data in NHEKs. We then predicted DSBs in U2OS cells using the NHEK-trained random forest with U2OS DNA-seq and ChIP-seq data. We validated the predictions with U2OS DSB data.

To evaluate prediction accuracy, we used the DSB data (DSBCapture [6] and BLESS [37]) that were generated for a specific cell line called U20S AID-DIvA. These DSB data were the only ones available in U20S. This cell line was a U2OS cell line that expressed the AsiSI restriction enzyme inducing DSBs at targeted sites [38]. To focus on endogenous DSBs, we kept only DSB data that did not overlap AsiSI sites. Most likely, only a fraction of all endogenous DSBs in U2OS could be mapped because DSB read coverage was low outside AsiSI sites.

In the first benchmark, we computed ROC and PR curves to evaluate the accuracy of model-based predictions. We compared our DSB predictions to a list of 2327 DSB sites identified by DSBCapture peak calling and 6443 non-DSB sites that were randomly drawn. Although this endogenous DSB list was far from complete, we obtained good prediction accuracy (AUROC = 0.835; $CI_{95\%}$: [0.824,0.846]; AUPR = 0.881; Fig. 6a; Additional file 1: Figure. S9). In agreement, we found that U2OS DSB prediction using a U2OS-trained random forest

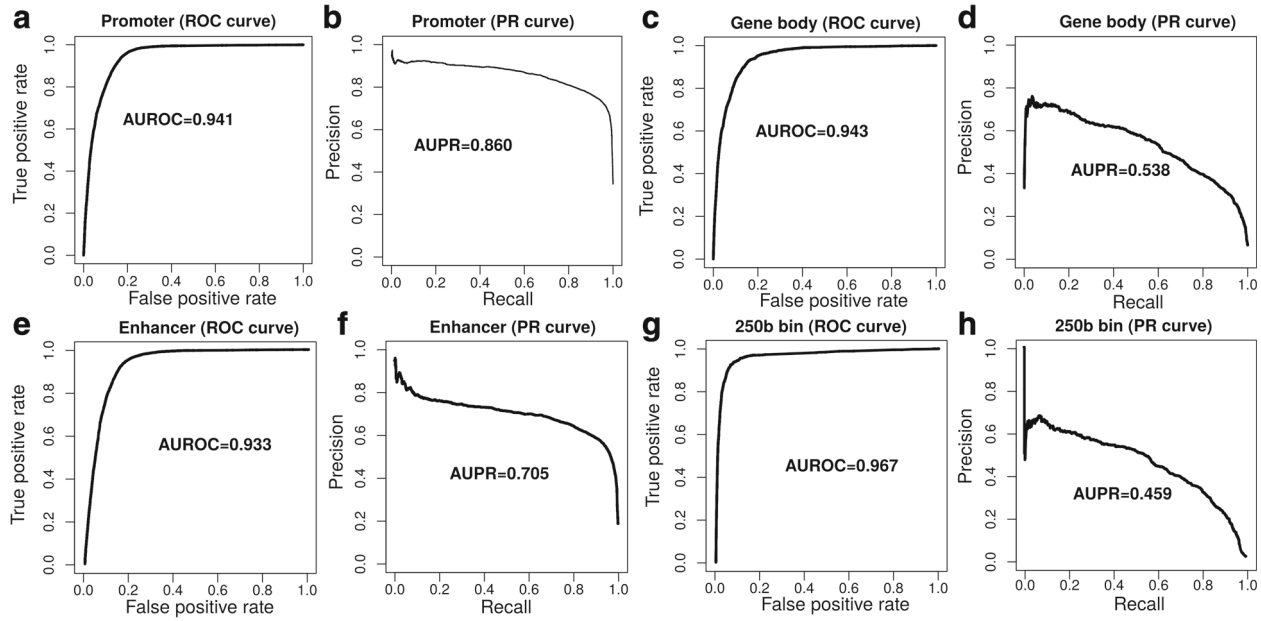


Fig. 5 Prediction of double-strand breaks (DSBs) using different controls. **a** Receiver operating characteristic (ROC) curve of a random forest discriminating between promoters with DSBs and promoters without. Area under the ROC curve (AUROC) is plotted. **b** Precision–recall (PR) curve of the random forest used in (**a**). Area under the PR curve (AUPR) is plotted. **c** ROC curve of a random forest discriminating between gene bodies with DSBs and gene bodies without. **d** Precision–recall curve of the random forest used in (**c**). **e** ROC curve of a random forest discriminating between enhancers with DSBs and enhancers without. **f** Precision–recall curve of the random forest used in (**e**). **g** ROC curve of a random forest discriminating between 250-base bins with DSBs and 250-base bins without. **h** Precision–recall curve of the random forest used in (**g**). AUPR, area under the PR curve, AUROC area under the ROC curve, DSB double-strand break, PR precision–recall, ROC receiver operating characteristic

yielded only slightly better predictions than using a NHEK-trained random forest (AUROC = 0.859; $CI_{95\%}$: [0.849,0.868]; AUPR = 0.904; Additional file 1: Figure S10). Moreover, DNase and CTCF had the highest variable importance, as found in NHEKs (Fig. 6b). Unfortunately, we could not carry out the same ROC and PR curve analyses with the BLESS data because not enough DSB sites were identified by peak calling.

In the second benchmark, we split the genome into 250-base bins and then predicted DSBs genome-wide. The model identified 87 190 bins with a high DSB score (predicted DSBs) and 77 510 bins with a low DSB score (predicted controls). As expected, we found a high enrichment of both DSBCapture and BLESS reads at predicted DSBs compared to predicted controls (Fig. 6c). On average, both DSBCapture and BLESS signals accordingly increased with the predicted DSB signal (Additional file 1: Figure S11a,b). Fortunately, there were also ChIP-seq data available for XRCC4, a DNA repair protein involved in non-homologous end-joining. Hence, we looked at whether XRCC4 was recruited at predicted DSBs. We found a high enrichment of XRCC4 at predicted DSBs compared to predicted controls (Fig. 6c), and an increase of the XRCC4 signal depending on the predicted DSB signal (Additional file 1: Figure S11c). In addition, ChIP-seq data were available for γ-H2AX, a histone mark that is induced at a megabase domain scale after DSBs, but is depleted on the few kilobases surrounding the exact break point [38, 39]. Accordingly, we observed that γ-H2AX was depleted at predicted DSBs compared to predicted controls (Fig. 6c), and we found a decrease of the γ-H2AX signal with the predicted DSB signal (Additional file 1: Figure S11d).

Additionally, we performed genome-wide DSB predictions in two other cell types for which endogenous DSB data were available, namely KBM7 (chronic myelogenous leukemia) and MCF-7 (breast cancer). For KBM7 cells, we used DNase-seq, CTCF, H3K4me1/me3, and H3K9me3 for prediction and BLISS for validation [40]. The model identified 163 113 bins with a high DSB score (predicted DSBs) and 115 204 bins with a low DSB score (predicted controls). We found an enrichment of BLISS reads at predicted DSBs compared to predicted controls (Additional file 1: Figure S12a). On average, the BLISS signal accordingly increased with the predicted DSB signal (Additional file 1: Figure S12b). For MCF-7 cells, we used DNase-seq, CTCF, H3K4me1/me3, H3K9ac/me3, and H3K27me3 for prediction and END-seq for validation [35]. The model identified 54 746 bins with a high DSB score (predicted DSBs) and 84 576 bins with a low DSB score (predicted controls). As expected, we found an enrichment of END-seq reads at predicted DSBs compared to predicted

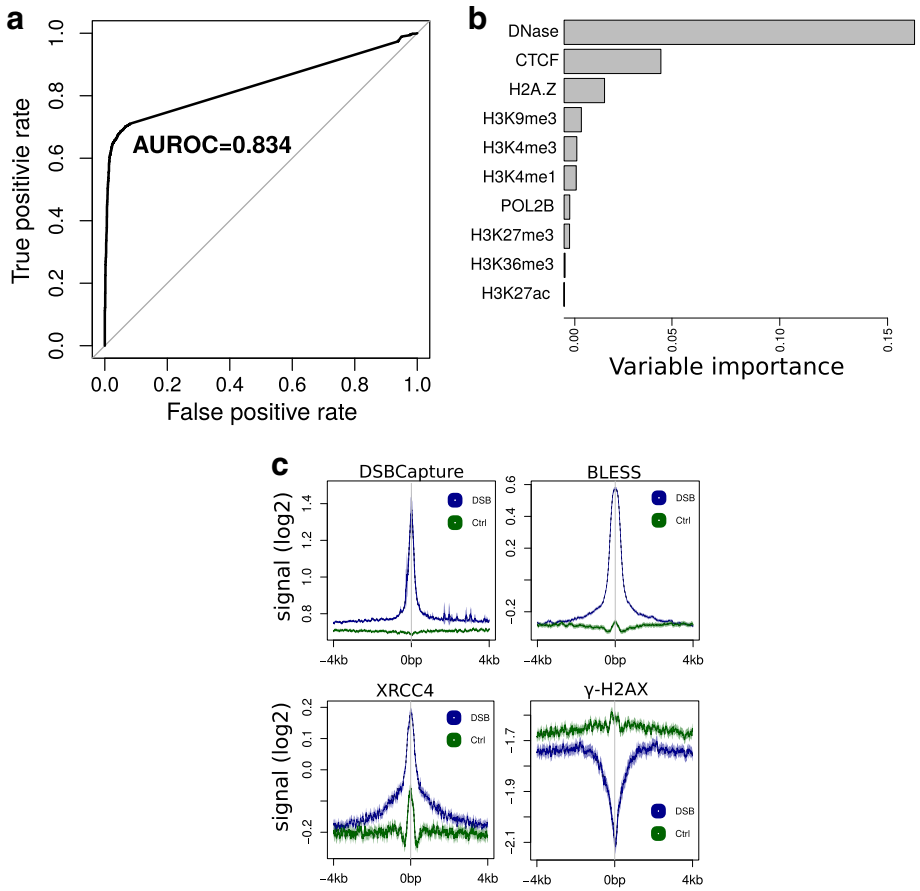


Fig. 6 Prediction of double-strand breaks (DSBs) using a random forest learned from DSBs in one cell type (NHEK) to predict DSBs in another cell type (U2OS). **a** Receiver operating characteristic (ROC) curve to predict U2OS DSBs using the NHEK-learned random forest. Area under the ROC curve (AUROC) is plotted. **b** Variable importance from the prediction of U2OS DSBs using the U2OS-learned random forest. **c** Average profiles of DSBCapture, BLESS, XRCC4, and γ-H2AX at predicted DSB regions compared to non-DSB regions over the whole genome. AUROC area under the ROC curve, DSB double-strand break, ROC receiver operating characteristic

controls (Additional file 1: Figure S12c). On average, the END-seq signal accordingly increased with the predicted DSB signal (Additional file 1: Figure S12d). We also tested whether our predictions in MCF-7 cells overlapped etoposide (ETO) induced DSBs mapped by END-seq. Interestingly, we found a strong enrichment of ETO END-seq reads at predicted DSBs compared to predicted controls (Additional file 1: Figure S12e). On average, the END-seq signal accordingly increased with the predicted DSB signal (Additional file 1: Figure S12f).

All these results revealed that the strongest predictors including DNase and CTCF were the same in two different cell types, and that accordingly, a random forest learned in one cell type can efficiently predict DSBs in another cell type.

Prediction from DNA motifs and shape

We then explored the possibility of predicting DSBs based on DNA sequence using DNA motif occurrences. We built a random forest classifier using 454 available motifs from the JASPAR 2016 database and obtained good prediction accuracy (AUROC = 0.827; $CI_{95\%}$: [0.819,0.831]; AUPR = 0.910; Fig. 7a; Additional file 1: Figure S13a). Several motifs from the transcription factor complex AP-1 were good predictors, such as FOS::JUN (VI = 0.016) and FOS (VI = 0.009) (Fig. 7b), which were previously shown to be enriched at DSB sites (see Section "Results and discussion", DSBs are enriched with epigenome marks and DNA motifs). Using lasso regression, we improved previous predictions (AUROC = 0.839; $CI_{95\%}$: [0.829,0.840]; AUPR = 0.919; Fig. 7a; Additional file 1: Figure S13a). Based on lasso regression, we found that the CTCF motif had the highest beta coefficient ($\beta = 3.22$), corresponding to OR = 25 (Fig. 7c), supporting recent evidence showing that long-range contacts are involved in DNA repair [25, 35, 41]. Furthermore, motifs of tumor proteins p53, p63, and p73 had high coefficients ($\beta > 2.03$, OR > 7.6), in agreement with previous predictions based on

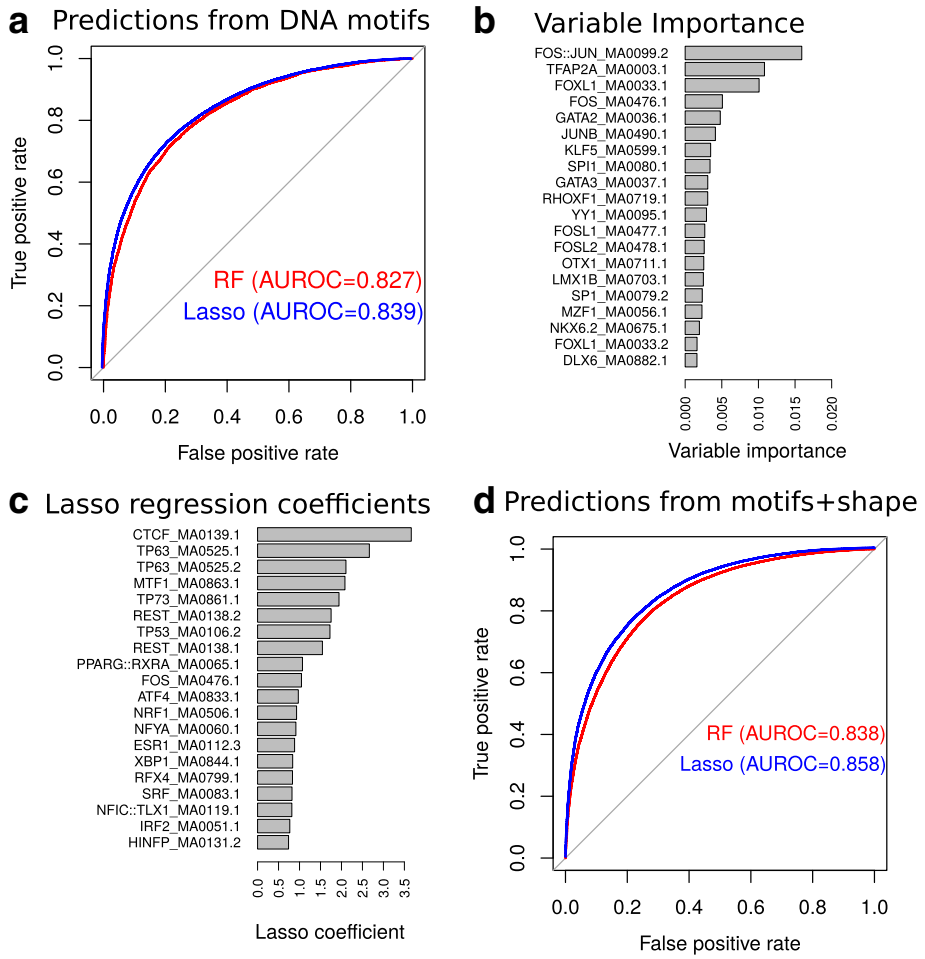


Fig. 7 Prediction of double-strand breaks (DSBs) using DNA motifs and shape. **a** Receiver operating characteristic (ROC) curve for the DSB predictions using DNA motifs from the JASPAR 2016 database. Random forest (RF) and lasso logistic regression were compared. **b** The 20 highest DNA motif variable importance values. **c** The 20 highest DNA motif lasso coefficients. **d** ROC curve for the DSB predictions using DNA motifs with DNA shape. AUROC area under the ROC curve, DSB double-strand break, RF random forest, ROC receiver operating characteristic

ChIP-seq data (see above). We also found motifs recognized by factors involved in heavy metal response (MTF-1: $\beta = 2.08$, OR $= 8$), in oxidative stress response (NRF1: $\beta = 0.93$, OR $= 2.53$; REST: $\beta = 1.75$, OR $= 5.75$), in endoplasmic reticulum stress (ATF4: $\beta = 0.97$, OR $= 2.64$), and in estrogen-induced DNA damage (ESR1: $\beta = 0.88$, OR $= 2.41$). To assess the significance of those motifs, we built a logistic regression model without any regularization including all motifs with $\beta > 0.5$. We found that most motifs (22/29) were significantly associated with DSBs ($p < 0.05$ after false discovery correction; Additional file 1: Table S2). Many of the above mentioned proteins have been shown to interact with each other. For instance, NRF1 associates with Jun proteins of the AP-1 complex [42]. ESR1 associates with AP-1/JUN and FOS to mediate estrogen element response-independent signaling [43].

DNA shape was recently shown to predict transcription factor binding sites and gene expression [14, 44]. Thus, we assessed if DNA shape could similarly serve to predict DSBs together with motifs. For this, we predicted four DNA shape features using simulations: minor groove width (MGW), propeller twist (ProT), roll (Roll), and helix twist (HelT) of DSB sites at base resolution. From each feature, we computed 12 predictors including quantiles (0, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100%) and the variance to describe the distribution of the feature within a DSB site. We used the resulting 48 variables combined with motif occurrences to predict DSBs with random forests and obtained better accuracy (AUROC $= 0.838$ and AUPR $= 0.915$; Fig. 7d; Additional file 1: Figure S13b) compared to using motifs alone (AUROC $= 0.827$ and AUPR $= 0.910$; Fig. 7a; Additional file 1: Figure S13a). Among the DNA shape variables,

ProT median and MGW variance had the highest variable importance (VI = 0.01 and VI = 0.01, respectively). Using lasso regression, we also obtained better predictions (AUROC = 0.858), compared to using motifs only (AUROC = 0.839 and AUPR = 0.928; Fig. 7d; Additional file 1: Figure S13b). These results reflect the importance of DNA shape in determining DSB sites, in agreement with studies showing that narrow minor grooves (created by either sequence context or DNA bending) limit access of reactive oxygen species [45].

Conclusions

DSBs are a major threat to a cell and they are associated with cancer development. Over the past years, new techniques have been developed to map DSBs at high resolution and genome-wide level. However, these techniques are costly and challenging. Here, we show, for the first time, that such DSBs can be computationally predicted using public epigenomic data, even when the availability of data is limited (e.g., DNase I and H3K4me1). By using state-of-the-art computational models, we achieve excellent prediction accuracy, paving the way for a better understanding of DSB formation depending on developmental stage or cell-type specific epigenetic marks. Thus, our computational approach should allow the genome-wide mapping of DSBs in numerous cell lines and tissues using the ENCODE and Roadmap Epigenomics databases.

There are multiple perspectives for this work. Recent developments from deep (convolutional) neural networks [13, 46] can improve model predictions and decrease the number of false positives at the genome level. In addition, our current model did not account for the impact of copy number variation in cancer cells on prediction, and future studies should integrate copy number variation as a quantitative predictor variable in the model to correct for this bias.

Methods

Double-strand breaks

All double-strand DNA break data used are summarized in Table 1. We used double-strand DNA breaks mapped by DSBCapture and BLESS in human epidermal keratinocyte (NHEK) cells from the Gene Expression Omnibus (GEO) accession GSE78172 [6]. DSBCapture and BLESS peaks were called using MACS 2.1.0 on human genome assembly hg19 (https://github.com/taoliu/MACS). The peaks obtained from two biological replicates were intersected to yield more reliable DSB sites for model predictions.

We used double-strand DNA breaks mapped by DSBCapture and BLESS in AID-DIvA cells, a U2OS cell line (human bone osteosarcoma epithelial cells) expressing the AsiSI restriction enzyme fused to a modified estrogen receptor ligand-binding domain [38]. Upon tamoxifen treatment, AsiSI induces sequence-specific DSBs at GCGATCGC sites. DSBCapture data were from tamoxifen-treated cells from GEO accession GSE78172 [6]. DSBCapture peaks were called using MACS 2.1.0 on human genome assembly hg19. BLESS data were from untreated cells arrested in G1 phase from ArrayExpress accession E-MTAB-4846 [37]. Because of the low coverage of BLESS data, a sufficient number of DSB peaks could not be called.

We used double-strand DNA breaks mapped by BLISS in KBM7 cells (human myeloid leukemia) from NCBI Sequence Read Archive at SRP099132 [40]. We also used double-strand DNA breaks mapped by END-seq in untreated and etoposide-treated MCF-7 cells (human breast cancer) from GSE99197 [35].

ChIP-seq and DNase-seq data

All ChIP-seq and DNase-seq data used are summarized in Table 2. We used ChIP-seq uniform peaks (CTCF, POL2B, EZH2, H3K4me1/me2/me3, H3K9me1/me3/ac, H3K27me3/ac, H3K36me3, H3K79me2, H4K20me1, and H2A.Z) and DNase-seq uniform peaks for NHEKs from the ENCODE project [7] (https://genome.ucsc.edu/encode). We also used p63 ChIP-seq of keratinocytes from GEO accession GSE59827 [21].

For U2OS cells, we used DNase-seq and H3K27ac ChIP-seq peaks from GEO accession GSE87831 [47]. We used H3K4me1 and POL2B ChIP-seq peaks from GEO accession GSE73742 [48]. We used H3K4me3 and H3K27me3 ChIP-seq peaks from GSE35573 [49]. We used H3K9me3

Table 1 Double-strand DNA break data summary

Cell line	Treatment	Technique	Number of replicates	Accession
NHEK	No treatment	DSBCapture	2	GSE78172
NHEK	No treatment	BLESS	2	GSE78172
U2OS	4-hydroxytamoxifen	DSBCapture	1	GSE78172
U2OS	No treatment	BLESS	1	E-MTAB-4846
KBM7	No treatment	BLISS	1	SRP099132
MCF-7	No treatment	END-seq	1	GSE99197
MCF-7	Etoposide	END-seq	1	GSE99197

Table 2 ChIP-seq and DNase-seq data summary

Cell line	Treatment	Technique	Number of replicates	Accession
NHEK	No treatment	CTCF, H3K4me3, H3K27me3, H3K36me3 ChIP-seq	2	ENCODE uniform peaks
NHEK	No treatment	EZH2, H3K4me1/me2, H3K9me1/me3/ac, H3K79me2, H4K20me1, H2A.Z, H3K27ac, POL2B ChIP-seq	1	ENCODE uniform peaks
NHEK	No treatment	DNase-seq	2	ENCODE uniform peaks
NHEK	No treatment	p63 ChIP-seq	1	GSE59827
U2OS	No treatment	DNase-seq, H3K27ac ChIP-seq	1	GSE87831
U2OS	No treatment	H3K4me1, POL2B ChIP-seq	1	GSE73742
U2OS	No treatment	H3K4me3, H3K27me3 ChIP-seq	1	GSE35573
U2OS	No treatment	H3K9me3, H3K36me3 ChIP-seq	1	ENCODE
U2OS	No treatment	CTCF ChIP-seq	1	ChIP-Atlas
U2OS	4-hydroxytamoxifen	XRCC4, γ-H2A.X ChIP-seq	1	E-MTAB-1241
KBM7	No treatment	DNase-seq	1	ChIP-Atlas
KBM7	No treatment	H3K9me3 ChIP-seq	1	GSE60056
K562	No treatment	CTCF, H3K4me1/me3 ChIP-seq	1	ENCODE
MCF-7	No treatment	H3K4me1/me3, H3K9ac/me3, H3K27me3 ChIP-seq	1	GSE23701
MCF-7	No treatment	DNase-seq and CTCF ChIP-seq	1	ENCODE

and H3K36me3 ChIP-seq peaks from ENCODE [7]. We used CTCF ChIP-seq peaks from the ChIP-Atlas database (http://chip-atlas.org/). We used XRCC4 and γ-H2A.X ChIP-seq for tamoxifen-treated DIvA cells from ArrayExpress accession E-MTAB-1241 [37].

For KBM7 cells, we used DNase-seq from the ChIP-Atlas database, and H3K9me3 ChIP-seq from GSE60056 [50]. Instead of KBM7, we used K562 (chronic myelogenous leukemia) for CTCF, H3K4me1/me3 ChIP-seq from the ENCODE project [7] (https://genome.ucsc.edu/encode). For MCF-7 cells, we used H3K4me1/me3, H3K9ac/me3, and H3K27me3 ChIP-seq without treatment (DMSO) from GSE23701 [51, 52]. We used DNase-seq and CTCF ChIP-seq from ENCODE [7].

DNA motifs

We used motif position frequency matrices for transcription factor binding sites from the JASPAR 2016 database (http://jaspar.genereg.net). We called transcription factor binding sites over the human genome using the position weight matrices and a minimum matching score of 80%.

DNA shape

We predicted four DNA shape features using Monte Carlo simulations: minor groove width (MGW) and propeller twist (ProT) at base pair resolution and roll (Roll) and helix twist (HelT) at base pair step resolution using R package DNAshapeR (https://bioconductor.org/packages/release/bioc/html/DNAshapeR.html).

Random forest and lasso regression

We used R package ranger (https://cran.r-project.org/web/packages/ranger) to compute the random forest classification efficiently [30]. We used the default package parameters: `num.trees=500` and `mtry` is the square root of the number of variables. Variable importance was computed using the mean decrease in accuracy in the out-of-bag sample. To discriminate between DSB and non-DSB sites, we randomly selected genomic sequences that matched sizes, GC, and repeat contents of DSB sites using R package gkmSVM (https://cran.r-project.org/web/packages/gkmSVM). To learn the model, we mapped epigenomic data, DNA motifs, and DNA shape as follows. For epigenomic data including ChIP-seq and DNase-seq data, we used peak genomic coordinates of a feature (for instance, CTCF binding sites) and considered the presence ($x = 1$) or absence ($x = 0$) of the corresponding feature at the DSB site. If a feature peak overlapped only 60% of the DSB site, then $x = 0.6$. For DNA motifs, we computed the number of motif occurrences within DSB and non-DSB sites. For DNA shape, we computed four features including MGW, ProT, Roll, and HelT of DSB sites at base resolution. For each DNA shape feature, we then computed 12 predictors, including quantiles (0, 10, 20, 30, 40, 50, 60, 70, 80, 90, and 100%) and the variance to describe the distribution of the feature within a DSB site. The DSB data were next split into two sets: the training set used for learning the model and a test set used for assessing prediction

accuracy. We also used R package glmnet (https://cran.r-project.org/web/packages/glmnet/index.html) to compute lasso logistic regression with cross-validation. To assess the prediction accuracy of random forest and lasso regression, we computed the ROC curve and AUROC. To estimate the confidence interval for AUROC, we used the pROC R package (https://cran.r-project.org/web/packages/pROC). We also computed the PR curve and AUPR to assess prediction accuracy when the classes were very imbalanced, especially for genome-wide analyses. For this, we used the PRROC R package (https://cran.r-project.org/web/packages/PRROC).

Additional file

Additional file 1: Additional figures and tables. **Figures S1–13** and **Tables S1, S2**. (PDF 1618 kb)

Acknowledgments
The authors are grateful to the Balasubramanian lab (Babraham Institute, UK), to the Crosetto lab (Karolinska Institutet, Sweden), and to the Nussenzweig lab (National Institutes of Health, USA) for data and for help in processing the data.

Funding
This work was supported by the University of Toulouse and by the CNRS. Funding for open access charge: Fondation pour la Recherche Médicale (DEQ20160334940).

Availability of data and materials
The pipeline was developed in the R language and is available at https://github.com/morphos30/PredDSB [53] under Apache License 2.0. The v1.0 release was deposited at https://zenodo.org/badge/latestdoi/117546880 with DOI 10.5281/zenodo.1174011.
The data used in this study were downloaded using the following accession numbers and databases:

- GSE78172 (NHEK DSBCapture and BLESS) [6]
- GSE78172 (U2OS AID-DIvA DSBCapture) [6]
- E-MTAB-4846 (U2OS AID-DIvA BLESS) [37]
- SRP099132 (KBM7 BLISS) [40]
- GSE99197 (MCF-7 END-seq) [35]
- ENCODE (NHEK ChIP-seq and DNase-seq) [7]
- GSE59827 (NHEK p63 ChIP-seq) [21]
- GSE87831 (U2OS DNase-seq and H3K27ac ChIP-seq) [47]
- GSE73742 (U2OS H3K4me1 and POL2B ChIP-seq) [48]
- GSE35573 (U2OS H3K4me3 and H3K27me3 ChIP-seq) [49]
- ENCODE (U2OS H3K9me3 and H3K36me3 ChIP-seq) [7]
- ChIP-Atlas database (U2OS CTCF ChIP-seq) [54]
- E-MTAB-1241 (U2OS XRCC4 and γ-H2A.X ChIP-seq) [37]
- ChIP-Atlas database (KBM7 DNase-seq) [54]
- GSE60056 (KBM7 H3K9me3 ChIP-seq) [50]
- ENCODE (K562 CTCF and H3K4me1/me3 ChIP-seq) [7]
- GSE23701 (MCF-7H3K4me1/me3, H3K9ac/me3, H3K27me3 ChIP-seq) [51, 52]
- ENCODE (MCF-7 DNase-seq and CTCF ChIP-seq) [7].

Authors' contributions
RM supervised the project, conceived the method, wrote the code, designed the data analysis, and analyzed the data. KG performed the BLESS experiments for U2OS AID-DIvA cells. RM, GL, and OC interpreted the results and wrote the paper. All authors read and approved the final manuscript.

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Publisher's Note

Author details
[1]LBME, Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 118, route de Narbonne, 31062 Toulouse, France. [2]Laboratory of Bioinformatics and Systems Biology, Centre of New Technologies, University of Warsaw, Zwirki i Wigury 93, 02-089 Warsaw, Poland. [3]LBCMCP, Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 118, route de Narbonne, 31062 Toulouse, France.

Received: 30 October 2017 Accepted: 22 February 2018
Published online: 15 March 2018

References

1. McKinnon PJ, Caldecott KW. DNA strand break repair and human genetic disease. Annu Rev Genomics Hum Genet. 2007;8(1):37–55. https://doi.org/10.1146/annurev.genom.7.080505.115648.
2. Mehta A, Haber JE. Sources of DNA double-strand breaks and models of recombinational DNA repair. Cold Spring Harb Perspect Biol. 2014;6(9):016428. https://doi.org/10.1101/cshperspect.a016428. http://cshperspectives.cshlp.org/content/6/9/a016428.full.pdf+html.
3. Crosetto N, Mitra A, Silva MJ, Bienko M, Dojer N, Wang Q, et al. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. Nat Methods. 2013;10(4):361–5. https://doi.org/10.1038/nmeth.2408.
4. Tsai SQ, Zheng Z, Nguyen NT, Liebers M, Topkar VV, Thapar V, et al. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. Nat Biotechnol. 2015;33(2):187–97.
5. Canela A, Sridharan S, Sciascia N, Tubbs A, Meltzer P, Sleckman B, et al. DNA breaks and end resection measured genome-wide by end sequencing. Mol Cell. 2016;63(5):898–911.
6. Lensing SV, Marsico G, Hansel-Hertsch R, Lam EY, Tannahill D, Balasubramanian S. DSBCapture: in situ capture and sequencing of DNA breaks. Nat Methods. 2016;13(10):855–7. https://doi.org/10.1038/nmeth.3960.
7. 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.
8. The Roadmap Epigenomics Consortium, Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317–30. https://doi.org/10.1038/nature14248.
9. Kleftogiannis D, Kalnis P, Bajic VB. DEEP: a general computational framework for predicting enhancers. Nucleic Acids Res. 2014;43(1):6. https://doi.org/10.1093/nar/gku1058.
10. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nat Methods. 2012;9(3):215–6. https://doi.org/10.1038/nmeth.1906.
11. Taverna SD, Li H, Ruthenburg AJ, Allis CD, Patel DJ. How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. Nat Struct Mol Biol. 2007;14(11):1025–40. https://doi.org/10.1038/nsmb1338.
12. Whitaker JW, Chen Z, Wang W. Predicting the human epigenome from DNA motifs. Nat Methods. 2015;12(3):265–72.
13. Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning-based sequence model. Nat Methods. 2015;12(10):931–4. https://doi.org/10.1038/nmeth.3547.
14. Mathelier A, Xin B, Chiu TP, Yang L, Rohs R, Wasserman WW. DNA shape features improve transcription factor binding site predictions in vivo. Cell Syst. 2016;3(3):278–864. https://doi.org/10.1016/j.cels.2016.07.001.
15. Hayashi K, Yoshida K, Matsui Y. A histone H3 methyltransferase controls epigenetic events required for meiotic prophase. Nature. 2005;438(7066):374–8. https://doi.org/10.1038/nature04112.
16. Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science. 2010;327(5967):876–9. https://doi.org/10.1126/

science.1182363. http://science.sciencemag.org/content/327/5967/876.full.pdf.
17. Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010;327(5967):836–40. https://doi.org/10.1126/science.1183439. http://science.sciencemag.org/content/327/5967/836.full.pdf.
18. Kinner A, Wu W, Staudt C, Iliakis G. γ-H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin. Nucleic Acids Res. 2008;36(17):5678–94. https://doi.org/10.1093/nar/gkn550.
19. Price BD, D'Andrea AD. Chromatin remodeling at DNA double-strand breaks Cell. 2013;152(6):1344–54. https://doi.org/10.1016/j.cell.2013.02.011.
20. Ghandi M, Mohammad-Noori M, Ghareghani N, Lee D, Garraway L, Beer MA. gkmSVM: an R package for gapped-kmer SVM. Bioinformatics. 2016;32(14):2205–7. https://doi.org/10.1093/bioinformatics/btw203.
21. Kouwenhoven EN, Oti M, Niehues H, van Heeringen SJ, Schalkwijk J, Stunnenberg HG, et al. Transcription factor p63 bookmarks and regulates dynamic enhancers during epidermal differentiation. EMBO Rep. 2015;16(7):863–78. https://doi.org/10.15252/embr.201439941.
22. Mathelier A, Fornes O, Arenillas DJ, Chen C-Y, Denay G, Lee J, et al. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. 2016;44(D1):110–5. https://doi.org/10.1093/nar/gkv1176.
23. Chiu TP, Comoglio F, Zhou T, Yang L, Paro R, Rohs R. DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. Bioinformatics. 2016;32(8):1211–3. https://doi.org/10.1093/bioinformatics/btv735.
24. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Methodol). 1996;58(1):267–88. https://doi.org/10.2307/2346178.
25. Tchurikov NA, Fedoseeva DM, Sosin DV, Snezhkina AV, Melnikova NV, Kudryavtseva AV, et al. Hot spots of DNA double-strand breaks and genomic contacts of human rDNA units are involved in epigenetic regulation. J Mol Cell Biol. 2015;7(4):366–82. https://doi.org/10.1093/jmcb/mju038.
26. Caron P, Aymard F, Iacovoni JS, Briois S, Canitrot Y, Bugler B, et al. Cohesin protects genes against γ-H2AX induced by DNA double-strand breaks. PLoS Genet. 2012;8(1):10002460. https://doi.org/10.1371/journal.pgen.1002460.
27. Phillips-Cremins JE, Sauria MEG, Sanyal A, Gerasimova TI, Lajoie BR, Bell JSK, et al. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. Cell. 2013;153(6):1281–95. https://doi.org/10.1016/j.cell.2013.04.053.
28. Lin YL, Sengupta S, Gurdziel K, Bell GW, Jacks T, Flores ER. p63 and p73 transcriptionally regulate genes involved in DNA repair. PLOS Genet. 2009;5(10):1000680. https://doi.org/10.1371/journal.pgen.1000680.
29. Williams AB, Schumacher B. p53 in the DNA-damage-repair process. Cold Spring Harb Perspect Med. 2016;6(5):026070. https://doi.org/10.1101/cshperspect.a026070. http://perspectivesinmedicine.cshlp.org/content/6/5/a026070.full.pdf+html.
30. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324.
31. Jacquet K, Fradet-Turcotte A, Avvakumov N, Lambert JP, Roques C, Pandita R, et al. The TIP60 complex regulates bivalent chromatin recognition by 53BP1 through direct H4K20me binding and H2AK15 acetylation. Mol Cell. 2016;62(3):409–21. https://doi.org/10.1016/j.molcel.2016.03.031.
32. Tjeertes JV, Miller KM, Jackson SP. Screen for DNA-damage-responsive histone modifications identifies H3K9Ac and H3K56Ac in human cells. EMBO J. 2009;28(13):1878–89. https://doi.org/10.1038/emboj.2009.119. http://emboj.embopress.org/content/28/13/1878.full.pdf.
33. Lhoumaud P, Hennion M, Gamot A, Cuddapah S, Queille S, Liang J, et al. Insulators recruit histone methyltransferase dMes4 to regulate chromatin of flanking genes. EMBO J. 2014;33(14):1599–613. https://doi.org/10.15252/embj.201385965.
34. Pfister SX, Ahrabi S, Zalmas LP, Sarkar S, Aymard F, Bachrati CZ, et al. SETD2-dependent histone H3K36 trimethylation is required for homologous recombination repair and genome stability. Cell Rep. 2014;7(6):2006–18. https://doi.org/10.1016/j.celrep.2014.05.026.
35. Canela A, Maman Y, Jung S, Wong N, Callen E, Day A, et al. Genome organization drives chromosome fragility. Cell. 2017;170(3):507–2118. https://doi.org/10.1016/j.cell.2017.06.034.
36. Hilmi K, Jangal M, Marques M, Zhao T, Saad A, Zhang C, et al. CTCF facilitates DNA double-strand break repair by enhancing homologous recombination repair. Sci Adv. 2017;3(5):1601898. https://doi.org/10.1126/sciadv.1601898. http://advances.sciencemag.org/content/3/5/e1601898.full.pdf.
37. Aymard F, Aguirrebengoa M, Guillou E, Javierre BM, Bugler B, Arnould C, et al. Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes. Nat Struct Mol Biol. 2017;24(4):353–61. https://doi.org/10.1038/nsmb.3387.
38. Iacovoni JS, Caron P, Lassadi I, Nicolas E, Massip L, Trouche D, et al. High-resolution profiling of γ-H2AX around DNA double strand breaks in the mammalian genome. EMBO J. 2010;29(8):1446–57. https://doi.org/10.1038/emboj.2010.38. http://emboj.embopress.org/content/29/8/1446.full.pdf.
39. Savic V, Yin B, Maas NL, Bredemeyer AL, Carpenter AC, Helmink BA, et al. Formation of dynamic γ-H2AX domains along broken DNA strands is distinctly regulated by ATM and MDC1 and dependent upon H2AX densities in chromatin. Mol Cell. 2009;34(3):298–310. https://doi.org/10.1016/j.molcel.2009.04.012.
40. Yan WX, Mirzazadeh R, Garnerone S, Scott D, Schneider MW, Kallas T, et al. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. Nat Commun. 2017;8:15058. https://doi.org/10.1038/ncomms15058.
41. Bekker-Jensen S, Mailand N. Assembly and function of DNA double-strand break repair foci in mammalian cells. DNA Repair. 2010;9(12):1219–28. https://doi.org/10.1016/j.dnarep.2010.09.010.
42. Venugopal R, Jaiswal AK. Nrf2 and Nrf1 in association with Jun proteins regulate antioxidant response element-mediated expression and coordinated induction of genes encoding detoxifying enzymes. Oncogene. 1998;17(24):3145–56.
43. Kushner PJ, Agard DA, Greene GL, Scanlan TS, Shiau AK, Uht RM, et al. Estrogen receptor pathways to AP-1. J Steroid Biochem Mol Biol. 2000;74(5):311–7.
44. Peng PC, Sinha S. Quantitative modeling of gene expression using DNA shape features of binding sites. Nucleic Acids Res. 2016;44(13):120. https://doi.org/10.1093/nar/gkw446.
45. Cannan WJ, Pederson DS. Mechanisms and consequences of double-strand DNA break formation in chromatin. J Cell Physiol. 2016;231(1):3–14. https://doi.org/10.1002/jcp.25048.
46. Kim SG, Harwani M, Grama A, Chaterji S. EP-DNN: a deep neural network-based global enhancer prediction algorithm. Sci Rep. 2016;6:38433.
47. Ibarra A, Benner C, Tyagi S, Cool J, Hetzer MW. Nucleoporin-mediated regulation of cell identity genes. Gene Dev. 2016;30(20):2253–8. https://doi.org/10.1101/gad.287417.116.
48. Pradhan SK, Su T, Yen L, Jacquet K, Huang C, Cote J, et al. EP400 deposits H3.3 into promoters and enhancers during gene activation. Mol Cell. 2016;61(1):27–38. https://doi.org/10.1016/j.molcel.2015.10.039.
49. Easwaran H, Johnstone SE, Van Neste L, Ohm J, Mosbruger T, Wang Q, et al. A DNA hypermethylation module for the stem/progenitor cell signature of cancer. Genome Res. 2012;22(5):837–49. https://doi.org/10.1101/gr.131169.111.
50. Tchasovnikarova IA, Timms RT, Matheson NJ, Wals K, Antrobus R, Göttgens B. Epigenetic silencing by the HUSH complex mediates position -effect variegation in human cells. Science. 2015;348(6242):1481–5. https://doi.org/10.1126/science.aaa7227.
51. Joseph R, Orlov YL, Huss M, Sun W, Li Kong S, Ukil L. Integrative model of genomic factors for determining binding site selection by estrogen receptor-α. Mol Syst Biol. 2010;6:456. https://doi.org/10.1038/msb.2010.109.
52. Kong SL, Li G, Loh SL, Sung WK, Liu ET. Cellular reprogramming by the conjoint action of ERα, FOXA1, and GATA3 to a ligand-inducible growth state. Mol Syst Biol. 2011;7:526. https://doi.org/10.1038/msb.2011.59.
53. Mourad R. morphos30/preddsb v1.0. GitHub. 2018. https://doi.org/10.5281/zenodo.1174011. https://github.com/morphos30/PredDSB.
54. Oki S, Ohta T, Shioi G, Hatanaka H, Ogasawara O, Okuda Y, et al. Integrative analysis of transcription factor occupancy at enhancers and disease risk loci in noncoding genomic regions. bioRxiv. 2018:262899. https://doi.org/10.1101/262899.

3.6.2 DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions (Vincent Rocher)

G4s are important DNA secondary structures that are known to regulate several essential processes in the cell, such as gene transcription, DNA replication, DNA repair, telomere stability and V(D)J recombination [Spiegel *et al.* 2019]. Moreover, G4s are highly suspected to be implicated in human diseases such as cancer or neurological/psychiatric disorders [Cimino-Reale *et al.* 2016, Asamitsu *et al.* 2019, Hänsel-Hertsch *et al.* 2020]. G4 structures can be predicted from the DNA sequence. The most basic 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) [Huppert & Balasubramanian 2005, Huppert & Balasubramanian 2006]. However, looking for a canonical motif lacked flexibility to capture the wide variety of sequences underlying G4 structures. More flexible algorithms instead assessed G-richness and G-skewness or alternatively sequence features including k-mers, and more recently involved machine/deep learning. However, current algorithms aimed to predict G4s in vitro, but were not designed to assess the ability of G4 sequences to form in vivo (*e.g.* G4 activity). Indeed, many G4s are formed in vitro but not vivo, and thus have no activity in the cell [Hänsel-Hertsch *et al.* 2016].

Vincent Rocher *et al.* proposed 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 [Rocher *et al.* 2021]. 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 [Hänsel-Hertsch *et al.* 2018]. DeepG4 achieved excellent accuracy at predicting cell-type specific active G4 regions (area under the receiver operating characteristic curve or AUROC > 0.98) (Figure 3 from the article "DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions" below). Moreover, DeepG4 identified key DNA motifs that were predictive of active G4 regions (Figure 4 from the article below). Among those motifs, Vincent Rocher *et al.* 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. Lastly, Vincent Rocher *et al.* used our new algorithm to map active G4 regions in multiple tissues and cancers as a comprehensive resource for the G4 community. Such active G4 regions represent novel therapeutic targets of recent G4-ligand drugs that are currently being tested.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

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

Vincent Rocher[1], **Matthieu Genais**[2], **Elissar Nassereddine**[1], **Raphael Mourad**[1]*

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

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

Data Availability Statement: All data and code were deposited on a Github repository. We downloaded G4 ChIP-seq data for HaCaT, K562 and HEKnp 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

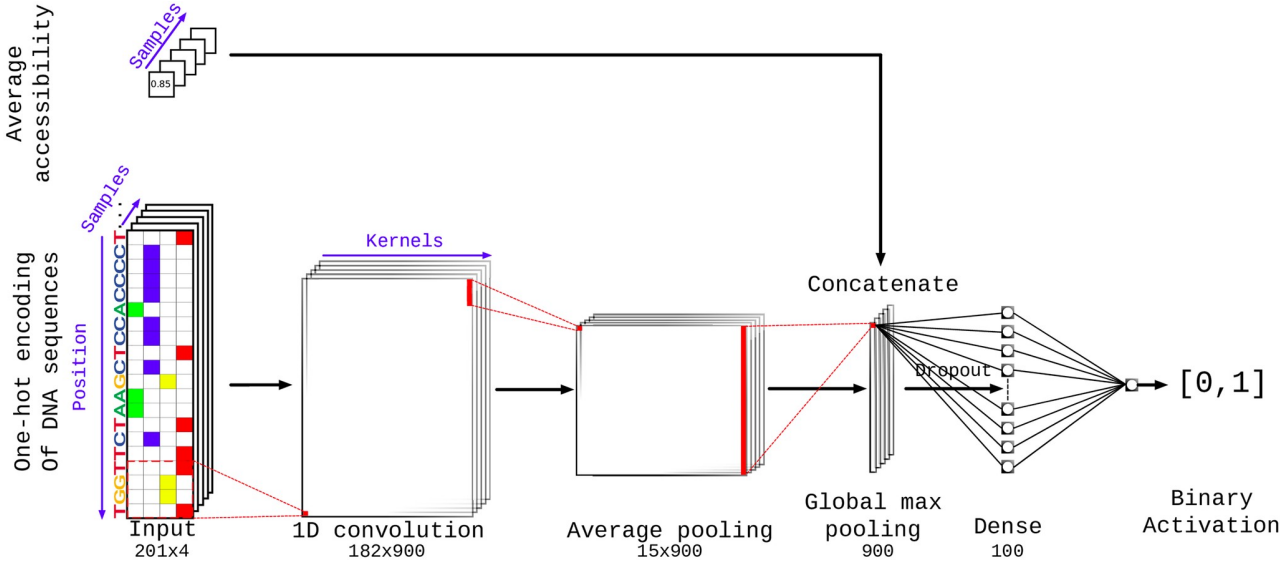


Fig 1. DeepG4 model architecture. Here, one-hot encoding is a numerical encoding of a 201-bp DNA sequence as a 201 × 4 matrix where each column corresponds to a DNA letter (A, C, G or T), and for instance, a value of one in the first column corresponds to a letter A in the sequence at a given position. 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 availabity

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

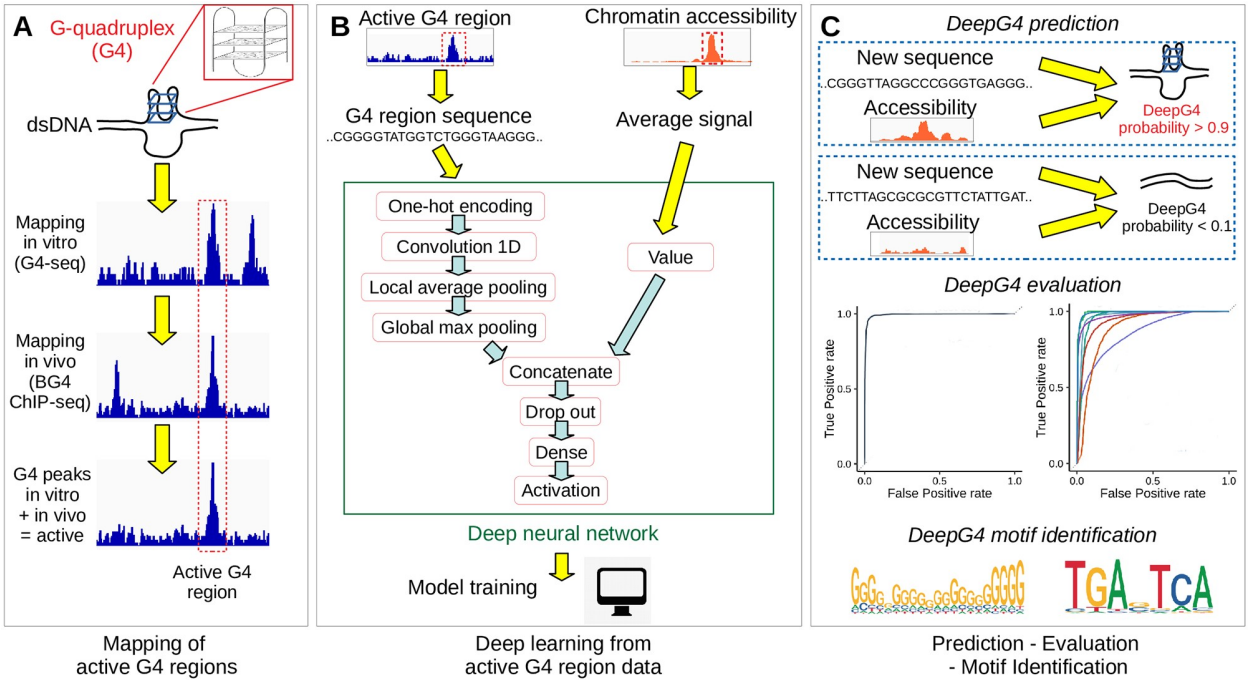


Fig 2. Illustration of DeepG4. A) Mapping of active G4 region sequences both in vitro and in vivo using NGS techniques. B) Deep learning model training using active G4 regions and control sequences. C) G4 activity prediction, evaluation and motif identification.

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 $\approx$ 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; A459: 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

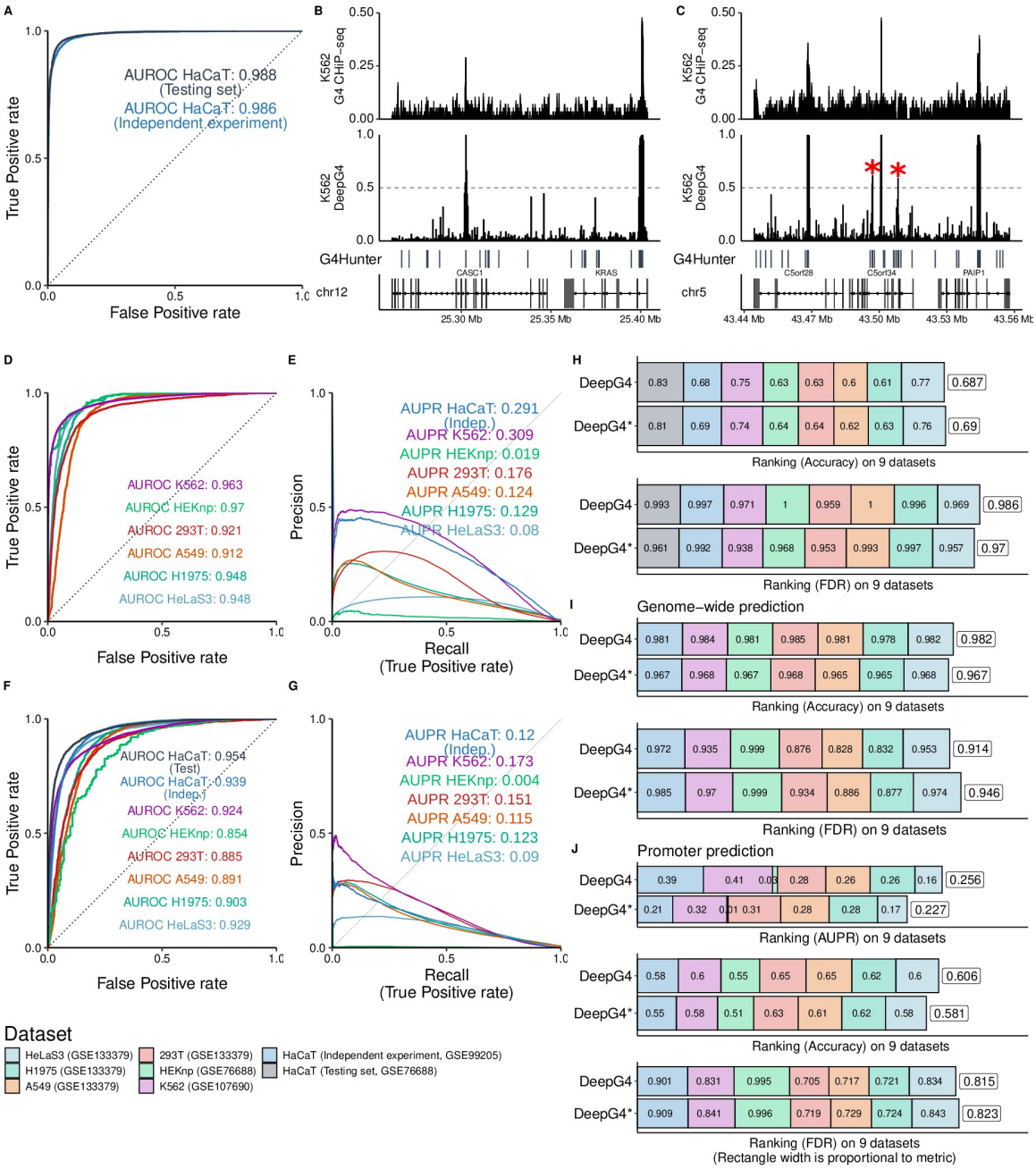


Fig 3. Prediction performance of DeepG4 to predict active G4 regions (regions where G4s form both in vitro and in vivo). A) Prediction performance of DeepG4. The model was trained and evaluated using HaCaT cell data. Predictions were evaluated on the testing set of sequences (same experiment as training 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

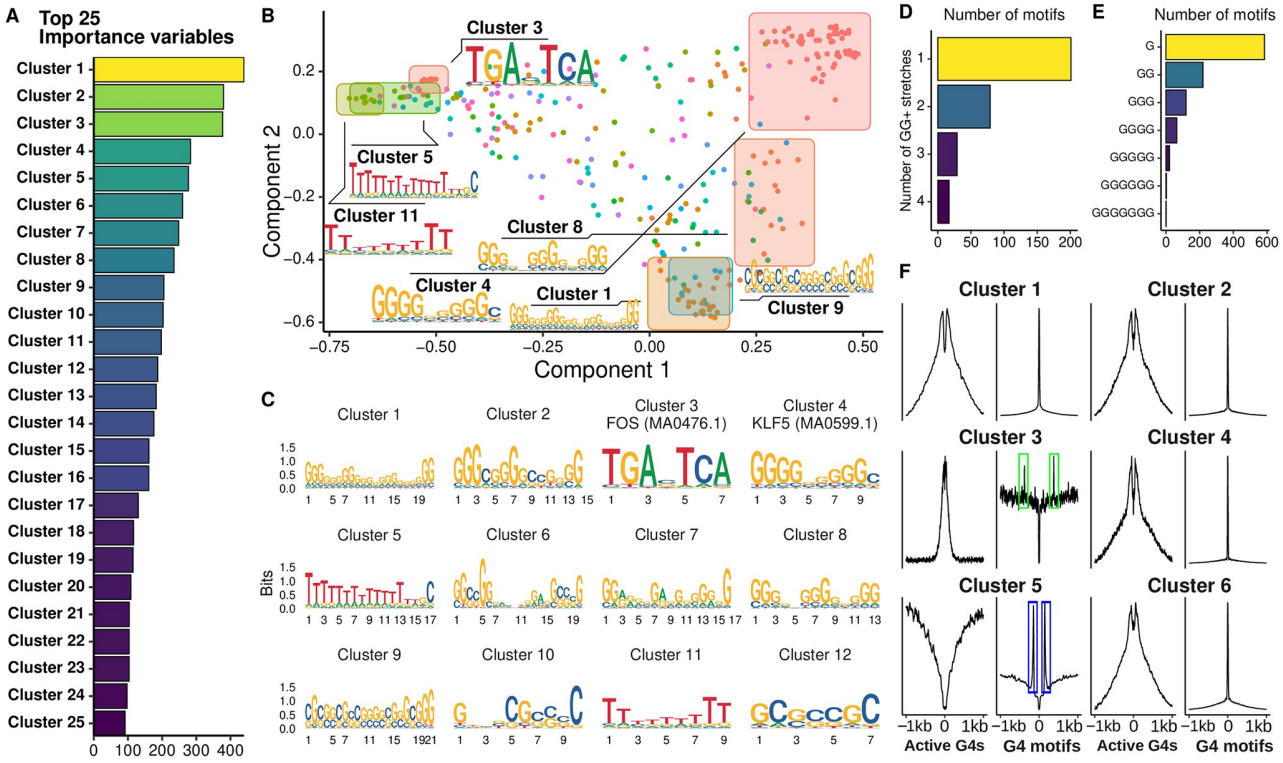


Fig 4. DNA motifs identified by DeepG4. A) Variable importances of DeepG4 cluster motifs, as estimated by random forests. Clustering of DeepG4 kernel motifs was done by RSAT matrix-clustering program to obtain cluster motifs. B) Multidimensional scaling (MDS) of DeepG4 motifs. As an input, matrix-clustering 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 ressembled 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 = 2×10^{-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 ressembling 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.

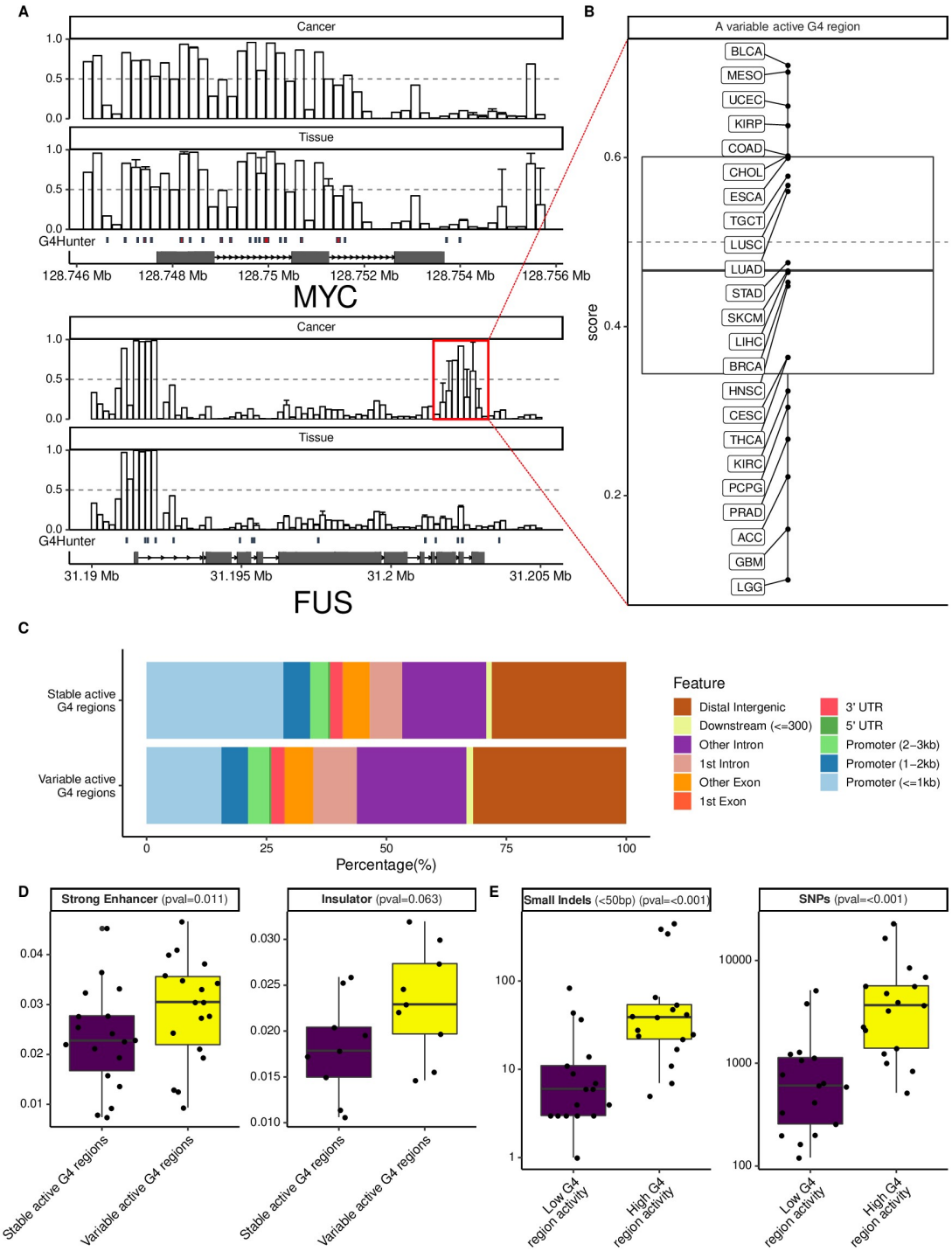


Fig 5. Genome-wide prediction of active G4 regions in tissues and cancers. A) Genome browser of DeepG4 predictions at MYC and FUS genes in tissues and cancers. B) Relationship between DeepG4 predicted G4 activity and the amount of mutations, depending on the mutation class. Cancer cohort abbreviations (*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 developped, 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, Beraldi 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, Martínez Cuesta S, Beraldi 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, Gerthoffert 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

Chapter 4

Future research projects

Sommaire

4.1 Introduction

In this last chapter, I present my future research projects and directions in computational biology. Some projects are in progress, for instance, the development of new models for the prediction of chromatin features such as protein binding sites using genomic sequences across different species, during a visiting professor position (délégation) at MIAT lab of INRAE. In particular, I am working on a novel deep learning model combining convolutional layers with graph neural network layers in order to borrow predictive information from both the target DNA sequence (sequence to be predicted) and also orthologous sequences from other related species, respectively.

I also plan to develop with Legube's team future long-term projects that are novel and risky in the field of DNA repair. For instance, I previously showed that DNA motifs, such as CTCF binding motif, could be strong predictors of endogenous DNA double-strand break (DSB) hotspots. This result suggests that mutations, such as SNPs, can drive genome instability when disrupting a CTCF motif. In the past, research efforts carried out to identify heritable mutations involved in cancer and neuronal/psychiatric diseases found several DSB repair pathway genes, such as BRCA1 and BRCA2. Beside SNPs affecting DSB repair genes (coding SNPs), I propose a novel genetic paradigm where non-coding SNPs could contribute to genome instability, and therefore, act as genetic drivers of cancer predisposition and neuronal/psychiatric diseases. However, to date, there is no experimental evidence supporting this hypothesis and potential SNPs affecting DSB frequency (dsbSNPs) should be mapped.

Beside academic research, I am launching a start-up for research and development in personalized medicine. With the "loi Pacte", I will have more flexibility to further extend such scientific activities. The start-up will aim (i) to scientifically advice industrial partners that are developing novel personalized medicine tools, and (ii) to potentially develop our own personalized medicine services.

4.2 Prediction of chromatin data in other species

The 3DR ratio was proposed to study 3D genome evolution in vertebrates using the genome sequence only [Mourad 2019]. This work suggested that chromatin can be highly conserved across phylogenetically related species. Moreover, the genome sequence could help to predict 3D chromatin information in species for which Hi-C data are not available.

Given the exponentially increasing number of genomes getting sequenced, one ambitious computational project would be to develop novel models for annotating every non-coding functional regions (identified from ChIP-seq, Hi-C, ATAC-seq, ...) using newly available genomic sequences. For instance, one important question is to determine if we can predict transcription factor, histone mark or accessible chro-

matin peaks in one species by training a convolutional neural network (CNN) in other species for which huge amount of data are available. The human and mouse could be considered as training references, given the comprehensive mapping of data in those species. Using the human- and mouse-trained CNNs, one could predict non-coding functional regions of related species such as mammals, and potentially vertebrates. This could be of great interest for species with high agronomic value, such as the pig, the cow and the sheep, that are phylogenetically close to the human or mouse and whose genomic sequences became recently available. Such CNN model should also include the phylogenetic distances between the target species (whose genome annotation is to be predicted) and the reference species used for training, and possibly integrate orthology information among species, and graph neural networks can be adapted for this purpose.

4.3 Biophysical experiments on G4 SNPs

DeepG4 could be used to predict the existence of SNPs altering G4 activity (G4SNPs). The mapping of G4SNPs genome-wide is of great importance for a better characterization of DNA determinants of G4 activity, since SNPs would help to decipher which parts of the G4 canonical motif or any other adjacent DNA motifs, such as transcription factor motifs, are of great impact in vivo. Moreover, mapping G4SNPs could also identify a novel causal molecular mechanism by which SNPs affecting DNA secondary structures can increase disease susceptibility of genetic diseases. However, our previous identification of G4SNPs only relied on in silico DeepG4 predictions and indirect functional studies (*e.g.* gene expression). To date, there are no experimental results that directly support the existence of G4SNPs.

In collaboration with Catherine Tardin, an expert in DNA biophysics from IPBS laboratory (Toulouse), we will carry out biophysical experiments to determine the invitro characteristics of predicted G4SNPs. Moreover, we will assess the in vivo activity of SWI/SNF remodelers [Bashyam *et al.* 2019] on DNA containing precisely positioned nucleosomes and G4SNPs using single molecule tools, called high throughput tethered particle motion, that permits real-time monitoring of the conformational dynamic of hundreds of single DNA [Brunet *et al.* 2015] and also using atomic force microscopy [Rousseau *et al.* 2010].

4.4 Non-coding SNPs as drivers of genome instability

I previously demonstrated that DNase, CTCF binding and motif, and H3K4me1/2/3 could predict DSB occurrence along the genome, reflecting the importance of chromatin and DNA sequence in determining DSB sites and subsequent repairing [Mourad *et al.* 2018]. Other works showed the importance of non-B DNA structures, such as G4s, in causing DSBs [Georgakopoulos-Soares *et al.* 2018]. However, the causal mechanisms linking the genomic and epigenomic determinants (*i.e.* the chromatin) to DSB formation and repair are still poorly understood. In

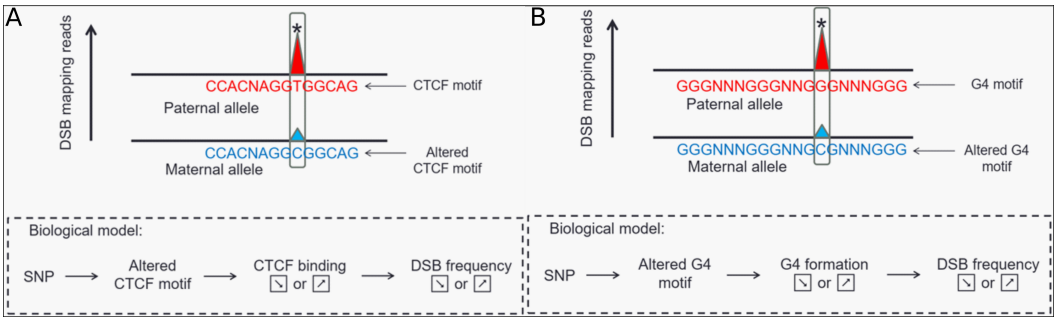


Figure 4.1: SNPs can help decipher causality between chromatin and genome instability. A) Potential scenario where a SNP alters a CTCF motif, leading to CTCF binding perturbation, and consequently to DSB frequency variation. B) Potential scenario where a SNP alters a G4 motif, leading to pertubed G4 formation, and consequently to DSB frequency variation.

particular, it is unknown if chromatin determinants are associated to DSBs because they cause them (causative model) or because they are caused by them (reactive model). Moreover, the impact of DNA mutations on DSB frequency and the link with genetic diseases are still unknown.

One future project will identify DSB allelic imbalance SNPs (named dsbSNPs) from the recently available mapping of endogenous DSBs by the team using phosphorylated ATM (pATM) ChIP-seq. Allelic imbalance will detect significant DSB frequency differences between the two paternal and maternal alleles at heterozygous loci within a single genome. One candidate model considers that a SNP that alters a CTCF motif would affect CTCF binding and therefore impact DSB frequency (Figure 4.1A). Another candidate model assumes that a SNP that disrupts a G4 would affect G4 formation (G4SNP) and consequently DSB frequency (Figure 4.1B). By integrating dsbSNPs with epigenetic (histone marks, DNase and CTCF) data, we will try to decipher the causal mechanism(s) behind DSB frequency. We plan to validate dsbSNPs by CRISPR/Cas9 and by assessing their impacts on DSB frequency and translocation. Moreover, we will validate causal mechanism(s) by assessing CRISPR/Cas9 SNPs' influence on intermediate processes such as CTCF binding, chromatin accessibility and histone marks. Thus, this approach will allow us to understand how genomic, and epigenomic determinants altogether causally affect the formation of endogenous DSBs.

In this project, we will also develop a novel paradigm explaining how SNPs can cause genetic diseases including cancer and neurological/psychiatric disorders by increasing genome instability. Over the past years, non-coding SNPs were shown to affect gene regulation (the so-called eSNPs), contributing to disease susceptibility [Nicolae *et al.* 2010]. In this project, we will seek to demonstrate a novel molecular mechanism by which SNPs can increase disease susceptibility by impacting the frequency of endogenous DSBs. The accumulation of dozens of dsbSNPs or even

more at key positions in the genome is likely to contribute to genetic diseases. The project will therefore contribute to a better understanding of the etiology of complex genetic diseases by uncovering a novel molecular mechanism of SNPs involved in DSBs and consequently in chromosomal rearrangements. Moreover, the project is complementary to the identification of G4SNPs, and the study of their impact on genetic diseases.

4.5 Z-DNA structures as key determinants of endogenous DSBs in neurological disorders

Recently, my Master's students, Elissar Nassereddine and Martin Tournaire, identified (GT)n short tandem repeats (STRs) strongly enriched at newly mapped endogenous DSBs by pATM ChIP-seq developed by the team. Around 15% of pATM peaks overlapped with a (GT)n STR. Such STRs are known to potentially form Z-DNA secondary structures, which are associated with increased genome instability [Georgakopoulos-Soares *et al.* 2018]. Moreover, pATM peaks were associated with genes involved in neuronal processes. However, those pATM peaks were not found within gene promoters as previously found [Canela *et al.* 2017], but instead within enhancers suggesting an unknown mechanism driving such endogenous DSBs in neurons. To test experimentally which factors are driving the formation of endogenous DSBs, we will map pATM peaks in different conditions including curaxin to stabilize Z-DNA, but also DRB to inhibit transcription by RNA polymerase II.

4.6 STR length is associated with high genome instability

Since STRs could undergo expansion or shortening, we then assessed if STR length between different genomic positions was associated with genome instability. We found that longer (GT)n STRs were positively correlated with higher genome instability. Interestingly, STR expansion is a very well-known and important cause of rare neurological genetic diseases such as myotonic dystrophies, Fragile X syndrome and Huntington's disease [Depienne & Mandel 2021]. Moreover, in general, neurological diseases (but also neurons from healthy patients) have been often associated with high genome instability, in particular large numbers of chromosomal rearrangements known to be caused by endogenous DSBs [Lee & Lupski 2006]. Based on our results, we predict that STR expansion will lead to higher pATM binding, reflecting higher risk of endogenous DSBs at the STR loci. Not only rare neurological genetic, but also common neurological genetic diseases identified by GWASs and associated with STR length, could be explained by a higher genomic instability caused by longer STR alleles. Our current genomic demonstration was based on the comparison between STRs from different loci, and thus lacks a proper genetic comparison between short and long alleles. Hence, we plan to carry out novel experiments to compare pATM ChIP-seq data between control patients and affected patients that carry STR expansion.

4.7 Mapping of endogenous DSBs at single-cell level

Currently, endogenous DSBs were mostly mapped at the cell population level, and nothing is known about the heterogeneity of the breakome (DSB pattern of the genome). Moreover, it is unknown if the breakome depends on the cell stage and differentiation. Hence, a novel PhD student, Sébastien Auber, will investigate how pATM binding varies across cells by developing single-cell approaches [Rotem *et al.* 2015, Ramani *et al.* 2017]. Moreover, in order to study neurological disorders, such experiment will be carried out in mini-brain models that reconstitute the arrangement of structural tissues and some of the complex biological functions of the human brain.

4.8 Candidate gene screening using public cancer databases

DNA repair pathways are triggered by the cell to maintain its genome integrity and stability when exposed to DNA damages. If DNA repair genes are deregulated, there is a higher chance of initiation and progression of cancer. Depending on the DNA repair pathway affected (non-homologous end joining, homologous recombination, microhomology-mediated end joining, ...), the genomic signature of cancer (somatic mutation patterns) will vary, since each pathway has specialized in the repair of certain mutations (SNPs, insertion-deletions, tandem repeats, translocations, ...). By integrating gene deregulation with genomic signature in cancers using publicly available databases, such as from TCGA/ICGC consortiums, we can predict novel candidate genes potentially involved in DNA repair pathways. This first gene screening will then lead to further investigation by experimentalists from the team (Nadine Puget and her students), who routinely assess DNA repair associated phenotypes in depleted cells.

4.9 AI for personalized medicine

As in other Western countries, the French population is aging and this is leading to the emergence of an ever-increasing number of diseases, as well as a galloping inflation of health costs for the country. In addition, France is undergoing medical desertification of regions far from urban areas, making access to specialists difficult for a segment of the population. In this context, personalized medicine makes it possible to address these problems by considerably reducing health-related costs as well as the need for specialist doctors by using automatic diagnosis.

I will create a start-up in AI for personalized medicine. In this start-up, I will be the scientific expert (”concours scientifique”) regarding:

- The scientific expertise of personalized medicine for biotechnology companies. Often companies have great resources for engineering, product marketing, sales, but lack a scientific vision and a deep understanding of the latest science advances. For instance, I will provide my expertise in biostatistics for

the use of published longitudinal studies of patients and published genetic studies (GWASs), my expertise of bioinformatics for the use of omic data and associated public databases, and my expertise in machine learning models and approaches for diagnostic predictions.

- The development of new tools for diagnosis. I will develop novel computational biology tools for the diagnosis of human diseases such as cancer, genetic diseases and chronic diseases using omic approaches (structural variants, gene expression profiles with RNA-seq / DNA chips, chromatin data), genetic databases (GWAS catalog, GTEx, ...) or deep learning predicted SNP impacts (DeepBind, DeepG4, ...), but also using other types of data like questionnaires, doctor's reports, medical imaging (MRI), etc.

Bibliography

[Alipanahi *et al.* 2015] Babak Alipanahi, Andrew Delong, Matthew T. Weirauch and Brendan J. Frey. *Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning.* Nature Biotechnology, vol. 33, pages 831–838, July 2015. (Cited on page 23.)

[Amar & Shamir 2014] David Amar and Ron Shamir. *Constructing module maps for integrated analysis of heterogeneous biological networks.* Nucleic Acids Research, vol. 42, no. 7, pages 4208–4219, January 2014. (Cited on page 23.)

[Andersson *et al.* 2014] Robin Andersson, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, Xiaobei Zhao, Christian Schmidl, Takahiro Suzuki, Evgenia Ntini, Erik Arner, Eivind Valen, Kang Li, Lucia Schwarzfischer, Dagmar Glatz, Johanna Raithel, Berit Lilje, Nicolas Rapin, Frederik Otzen O. Bagger, Mette Jørgensen, Peter Refsing R. Andersen, Nicolas Bertin, Owen Rackham, A. Maxwell Burroughs, J. Kenneth Baillie, Yuri Ishizu, Yuri Shimizu, Erina Furuhata, Shiori Maeda, Yutaka Negishi, Christopher J. Mungall, Terrence F. Meehan, Timo Lassmann, Masayoshi Itoh, Hideya Kawaji, Naoto Kondo, Jun Kawai, Andreas Lennartsson, Carsten O. Daub, Peter Heutink, David A. Hume, Torben Heick H. Jensen, Harukazu Suzuki, Yoshihide Hayashizaki, Ferenc Müller, FANTOM Consortium, Alistair R. Forrest, Piero Carninci, Michael Rehli and Albin Sandelin. *An atlas of active enhancers across human cell types and tissues.* Nature, vol. 507, no. 7493, pages 455–461, March 2014. (Cited on pages 12 and 13.)

[Andreas D. Baxevanis 2020] David S. Wishart Andreas D. Baxevanis Gary D. Bader, editor. Bioinformatics. Wiley, 4th édition, 2020. (Cited on page 19.)

[Andrews & Hemberg 2018] Tallulah S. Andrews and Martin Hemberg. *Identifying cell populations with scRNASeq.* Molecular Aspects of Medicine, vol. 59, pages 114 – 122, February 2018. The emerging field of single-cell analysis. (Cited on page 21.)

[Arnould & Legube 2020] Coline Arnould and Gaëlle Legube. *The secret life of chromosome loops upon DNA double-strand break.* Journal of Molecular Biology, vol. 432, no. 3, pages 724 – 736, February 2020. Perspectives on Chromosome Folding. (Cited on page 121.)

[Arnould *et al.* 2021] Coline Arnould, Vincent Rocher, Anne-Laure Finoux, Thomas Clouaire, Kevin Li, Felix Zhou, Pierre Caron, Philippe. E. Mangeot, Emiliano P. Ricci, Raphaël Mourad, James E. Haber, Daan Noordermeer and Gaëlle Legube. *Loop extrusion as a mechanism for formation of DNA damage repair foci.* Nature, vol. 590, no. 7847, pages 660–665, Feb 2021. (Cited on page 121.)

[Asamitsu *et al.* 2019] Sefan Asamitsu, Masayuki Takeuchi, Susumu Ikenoshita, Yoshiki Imai, Hirohito Kashiwagi and Norifumi Shioda. *Perspectives for applying G-quadruplex structures in neurobiology and neuropharmacology.* International Journal of Molecular Sciences, vol. 20, no. 12, June 2019. (Cited on page 198.)

[Aymard *et al.* 2017] François Aymard, Marion Aguirrebengoa, Emmanuelle Guillou, Biola M. Javierre, Beatrix Bugler, Coline Arnould, Vincent Rocher, Jason S. Iacovoni, Anna Biernacka, Magdalena Skrzypczak, Krzysztof Ginalski, Maga Rowicka, Peter Fraser and Gaëlle Legube. *Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes.* Nature Structural & Molecular Biology, vol. 24, no. 4, pages 353–361, March 2017. (Cited on pages 15 and 145.)

[Azuaje & Dopazo 2005] Francisco Azuaje and Joaquin Dopazo. Data analysis and visualization in genomics and proteomics. Wiley-Blackwell, 2005. (Cited on page 19.)

[Bashyam *et al.* 2019] Murali Dharan Bashyam, A Srinivas and B Pratyusha. *Taming the Master: SWI/SNF chromatin remodeller as a therapeutic target in cancer.* Current Science, vol. 116, no. 10, pages 1653–1665, May 2019. (Cited on page 217.)

[Belokopytova *et al.* 2020] Polina S. Belokopytova, Miroslav A. Nuriddinov, Evgeniy A. Mozheiko, Daniil Fishman and Veniamin Fishman. *Quantitative prediction of enhancer-promoter interactions.* Genome Research, vol. 30, no. 1, pages 72–84, January 2020. (Cited on page 96.)

[Bhattacharyya *et al.* 2016] Debmalya Bhattacharyya, Gayan Mirihana Arachchilage and Soumitra Basu. *Metal cations in G-quadruplex folding and stability.* Frontiers in Chemistry, vol. 4, page 38, September 2016. (Cited on page 11.)

[Bianco *et al.* 2018] Simona Bianco, Darío G. Lupiáñez, Andrea M. Chiariello, Carlo Annunziatella, Katerina Kraft, Robert Schöpflin, Lars Wittler, Guillaume Andrey, Martin Vingron, Ana Pombo, Stefan Mundlos and Mario Nicodemi. *Polymer physics predicts the effects of structural variants on chromatin architecture.* Nature Genetics, vol. 50, no. 5, pages 662–667, April 2018. (Cited on page 96.)

[Billings & Florez 2010] Liana K. Billings and Jose C. Florez. *The genetics of type 2 diabetes: What have we learned from GWAS?* Annals of the New York Academy of Sciences, vol. 1212, no. 1, pages 59–77, November 2010. (Cited on page 18.)

[Bishop 2007] Christopher M. Bishop. Pattern recognition and machine learning (information science and statistics). Springer, 1 édition, 2007. (Cited on page 21.)

[Boser *et al.* 1992] Bernhard E. Boser, Isabelle M. Guyon and Vladimir N. Vapnik. *A training algorithm for optimal margin classifiers.* In David Haussler, editor, Proceedings of the 5th Annual Workshop on Computational Learning Theory (COLT92), pages 144–152, Pittsburgh, PA, USA, July 1992. ACM Press. (Cited on page 21.)

[Breiman 2001] Leo Breiman. *Random forests.* Machine Learning, vol. 45, no. 1, pages 5–32, October 2001. (Cited on page 21.)

[Brunet *et al.* 2015] Annaël Brunet, Sébastien Chevalier, Nicolas Destainville, Manoel Manghi, Philippe Rousseau, Maya Salhi, Laurence Salomé and Catherine Tardin. *Probing a label-free local bend in DNA by single molecule tethered particle motion.* Nucleic Acids Research, vol. 43, no. 11, page e72, March 2015. (Cited on page 217.)

[Buenrostro *et al.* 2013] Jason D. Buenrostro, Paul G. Giresi, Lisa C. Zaba, Howard Y. Chang and William J. Greenleaf. *Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position.* Nature Methods, vol. 10, no. 12, pages 1213–1218, December 2013. (Cited on page 16.)

[Canela *et al.* 2016] A Canela, S Sridharan, N Sciascia, A Tubbs, P Meltzer, BP Sleckman and A Nussenzweig. *DNA breaks and end resection measured genome-wide by end sequencing.* Molecular Cell, vol. 63, no. 5, pages 898–911, September 2016. (Cited on pages 16 and 182.)

[Canela *et al.* 2017] Andres Canela, Yaakov Maman, Seolkyoung Jung, Nancy Wong, Elsa Callen, Amanda Day, Kyong-Rim Kieffer-Kwon, Aleksandra Pekowska, Hongliang Zhang, Suhas S.P. Rao, Su-Chen Huang, Peter J. Mckinnon, Peter D. Aplan, Yves Pommier, Erez Lieberman Aiden, Rafael Casellas and André Nussenzweig. *Genome organization drives chromosome fragility.* Cell, vol. 170, no. 3, pages 507–521.e18, July 2017. (Cited on pages 15 and 219.)

[Cao *et al.* 2002] Ru Cao, Liangjun Wang, Hengbin Wang, Li Xia, Hediye Erdjument-Bromage, Paul Tempst, Richard S. Jones and Yi Zhang. *Role of histone H3 lysine 27 methylation in polycomb-group silencing.* Science, vol. 298, no. 5595, pages 1039–1043, 2002. (Cited on page 120.)

[Caron *et al.* 2015] P. Caron, J. Choudjaye, T. Clouaire, B. Bugler, V. Daburon, M. Aguirrebengoa, T. Mangeat, J. S. Iacovoni, A. Alvarez-Quilon, F. Cortes-Ledesma and G. Legube. *Non-redundant functions of ATM and DNA-PKcs in response to DNA double-strand breaks.* Cell Reports, vol. 13, no. 8, pages 1598–1609, November 2015. (Cited on page 15.)

[Carullo & Day 2019] Nancy V. N. Carullo and Jeremy J. Day. *Genomic enhancers in brain health and disease.* Genes, vol. 10, no. 1, 2019. (Cited on page 12.)

[Carvalho & Lupski 2016] Claudia M. B. Carvalho and James R. Lupski. *Mechanisms underlying structural variant formation in genomic disorders.* Nature Reviews Genetics, vol. 17, no. 4, pages 224–238, February 2016. (Cited on page 14.)

[Ceccaldi *et al.* 2016] Raphael Ceccaldi, Beatrice Rondinelli and Alan D. D'Andrea. *Repair pathway choices and consequences at the double-strand break.* Trends in Cell Biology, vol. 26, no. 1, pages 52–64, January 2016. (Cited on page 14.)

[Chen & Guestrin 2016] Tianqi Chen and Carlos Guestrin. *XGBoost: A scalable tree boosting system.* In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD 16, pages 785–794, New York, NY, USA, 2016. Association for Computing Machinery. (Cited on page 21.)

[Chen & Yang 2012] Yuwei Chen and Danzhou Yang. *Sequence, stability, and structure of G-quadruplexes and their interactions with drugs.* Current Protocols in Nucleic Acid Chemistry, vol. 50, no. 1, pages 17.5.1–17.5.17, September 2012. (Cited on page 11.)

[Chen *et al.* 2016] Jie Chen, III Hero Alfred O. and Indika Rajapakse. *Spectral identification of topological domains.* Bioinformatics, vol. 32, no. 14, pages 2151–2158, May 2016. (Cited on page 96.)

[Chevenet *et al.* 2013] François Chevenet, Matthieu Jung, Martine Peeters, Tulio de Oliveira and Olivier Gascuel. *Searching for virus phylotypes.* Bioinformatics, vol. 29, no. 5, pages 561–570, January 2013. (Cited on page 29.)

[Chung *et al.* 2010] Charles C. Chung, Wagner C. S. Magalhaes, Jesus Gonzalez-Bosquet and Stephen J. Chanock. *Genome-wide association studies in cancer: Current and future directions.* Carcinogenesis, vol. 31, no. 1, pages 111–120, January 2010. (Cited on page 18.)

[Cimino-Reale *et al.* 2016] Graziella Cimino-Reale, Nadia Zaffaroni and Marco Folini. *Emerging role of G-quadruplex DNA as target in anticancer therapy.* Current Pharmaceutical Design, vol. 22, no. 44, pages 6612–6624, 2016. (Cited on page 198.)

[Collins & Sullivan 2013] Ann L. Collins and Patrick F. Sullivan. *Genome-wide association studies in psychiatry: What have we learned?* The British Journal of Psychiatry, vol. 202, no. 1, pages 1–4, January 2013. (Cited on page 18.)

[Consortium 2017] GTEx Consortium. *Genetic effects on gene expression across human tissues.* Nature, vol. 550, no. 7675, pages 204–213, October 2017. (Cited on page 183.)

[Cookson *et al.* 2009] William Cookson, Liming Liang, Gonçalo Abecasis, Miriam Moffatt and Mark Lathrop. *Mapping complex disease traits with global gene expression.* Nature Reviews Genetics, vol. 10, pages 184–194, March 2009. (Cited on page 18.)

[Cournac *et al.* 2015] Axel Cournac, Romain Koszul and Julien Mozziconacci. *The 3D folding of metazoan genomes correlates with the association of similar repetitive elements*. Nucleic Acids Research, vol. 44, no. 1, pages 245–255, 11 2015. (Cited on page 31.)

[Craddock & Owen 1994] Nick Craddock and Mike Owen. *Chromosomal aberrations and bipolar affective disorder*. British Journal of Psychiatry, vol. 164, no. 4, page 507–512, April 1994. (Cited on page 15.)

[Crane *et al.* 2015] Emily Crane, Qian Bian, Rachel Patton McCord, Bryan R. Lajoie, Bayly S. Wheeler, Edward J. Ralston, Satoru Uzawa, Job Dekker and Barbara J. Meyer. *Condensin-driven remodelling of X chromosome topology during dosage compensation*. Nature, vol. 523, pages 240–244, June 2015. (Cited on page 96.)

[Cresswell & Dozmorov 2020] Kellen G. Cresswell and Mikhail G. Dozmorov. *TADCompare: An R package for differential and temporal analysis of topologically associated domains*. Frontiers in Genetics, vol. 11, page 158, March 2020. (Cited on page 96.)

[Crosetto *et al.* 2013] Nicola Crosetto, Abhishek Mitra, Maria J. Silva, Magda Bienko, Norbert Dojer, Qi Wang, Elif Karaca, Roberto Chiarle, Magdalena Skrzypczak, Krzysztof Ginalski, Philippe Pasero, Maga Rowicka and Ivan Dikic. *Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing*. Nature Methods, vol. 10, no. 4, pages 361–365, April 2013. (Cited on pages 16 and 182.)

[Cubenas-Potts & Corces 2015] Caelin Cubenas-Potts and Victor G. Corces. *Architectural proteins, transcription, and the three-dimensional organization of the genome*. FEBS Letters, vol. 589, no. 20PartA, pages 2923–2930, October 2015. (Cited on page 14.)

[Dauban *et al.* 2020] Lise Dauban, Rémi Montagne, Agnès Thierry, Luciana Lazar-Stefanita, Nathalie Bastié, Olivier Gadal, Axel Cournac, Romain Koszul and Frédéric Beckouët. *Regulation of Cohesin-Mediated Chromosome Folding by Eco1 and Other Partners*. Molecular Cell, vol. 77, no. 6, pages 1279–1293.e4, 2020. (Cited on page 14.)

[Depienne & Mandel 2021] Christel Depienne and Jean-Louis Mandel. *30 years of repeat expansion disorders: What have we learned and what are the remaining challenges?* The American Journal of Human Genetics, vol. 108, no. 5, pages 764–785, 2021. (Cited on page 219.)

[Dixon *et al.* 2012] Jesse R. Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu and Bing Ren. *Topological domains in mammalian genomes identified by analysis of chromatin interactions*. Nature, vol. 485, no. 7398, pages 376–380, May 2012. (Cited on pages 13, 58, 85 and 96.)

[Djekidel *et al.* 2015] Mohamed N. Djekidel, Zhengyu Liang, Qi Wang, Zhirui Hu, Guipeng Li, Yang Chen and Michael Q. Zhang. *3CPET: finding co-factor complexes from ChIA-PET data using a hierarchical Dirichlet process.* Genome Biology, vol. 16, no. 1, pages 288+, December 2015. (Cited on page 58.)

[Dorn & Cresci 2009] Gerald W. Dorn and Sharon Cresci. *Genome-wide association studies of coronary artery disease and heart failure: where are we going?* Pharmacogenomics, vol. 10, no. 2, pages 213–223, February 2009. PMID: 19207022. (Cited on page 18.)

[Dwyer 2020] D. S. Dwyer. *Genomic chaos begets psychiatric disorder.* Complex Psychiatry, vol. 6, no. 1-2, pages 20–29, February 2020. (Cited on page 15.)

[Fullwood *et al.* 2009] Melissa J. Fullwood, Mei Hui H. Liu, You Fu F. Pan, Jun Liu, Han Xu, Yusoff Bin B. Mohamed, Yuriy L. Orlov, Stoyan Velkov, Andrea Ho, Poh Huay H. Mei, Elaine G. Chew, Phillips Yao Hui Y. Huang, Willem-Jan J. Welboren, Yuyuan Han, Hong Sain S. Ooi, Pramila N. Ariyaratne, Vinsensius B. Vega, Yanquan Luo, Peck Yean Y. Tan, Pei Ye Y. Choy, Senali Abayratna D. Wansa, Bing Zhao, Kar Sian S. Lim, Shi Chi C. Leow, Jit Sin S. Yow, Roy Joseph, Haixia Li, Kartiki V. Desai, Jane S. Thomsen, Yew Kok K. Lee, R. Krishna Murthy Karuturi, Thoreau Herve, Guillaume Bourque, Hendrik G. Stunnenberg, Xiaoan Ruan, Valere Cacheux-Rataboul, Wing-Kin K. Sung, Edison T. Liu, Chia-Lin L. Wei, Edwin Cheung and Yijun Ruan. *An oestrogen-receptor-α-bound human chromatin interactome.* Nature, vol. 462, no. 7269, pages 58–64, November 2009. (Cited on pages 13 and 16.)

[Georgakopoulos-Soares *et al.* 2018] Ilias Georgakopoulos-Soares, Sandro Morganella, Naman Jain, Martin Hemberg and Serena Nik-Zainal. *Noncanonical secondary structures arising from non-B DNA motifs are determinants of mutagenesis.* Genome Research, vol. 28, no. 9, pages 1264–1271, August 2018. (Cited on pages 11, 217 and 219.)

[Ghavi-Helm *et al.* 2014] Yad Ghavi-Helm, Felix A. Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber and Eileen E. Furlong. *Enhancer loops appear stable during development and are associated with paused polymerase.* Nature, vol. 512, no. 7512, pages 96–100, August 2014. (Cited on pages 13 and 14.)

[Gómez-Marín *et al.* 2015] Carlos Gómez-Marín, Juan J. Tena, Rafael D. Acemel, Macarena López-Mayorga, Silvia Naranjo, Elisa de la Calle-Mustienes, Ignacio Maeso, Leonardo Beccari, Ivy Aneas, Erika Vielmas, Paola Bovolenta, Marcelo A. Nobrega, Jaime Carvajal and José L. Gómez-Skarmeta. *Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders.* Proceedings of the National Academy of Sciences, vol. 112, no. 24, pages 7542–7547, June 2015. (Cited on page 112.)

[Goodfellow *et al.* 2016] Ian Goodfellow, Yoshua Bengio and Aaron Courville. Deep Learning. MIT Press, 2016. http://www.deeplearningbook.org. (Cited on page 22.)

[Goodwin *et al.* 2016] Sara Goodwin, John D. McPherson and W. Richard McCombie. *Coming of age: ten years of next-generation sequencing technologies.* Nature Reviews Genetics, vol. 17, no. 6, pages 333–351, Jun 2016. (Cited on page 15.)

[Guénolé & Legube 2017] Aude Guénolé and Gaëlle Legube. *A meeting at risk: Unrepaired DSBs go for broke.* Nucleus, vol. 8, no. 6, pages 589–599, 2017. PMID: 29099269. (Cited on page 15.)

[Haddad *et al.* 2017] Noelle Haddad, Cédric Vaillant and Daniel Jost. *IC-Finder: inferring robustly the hierarchical organization of chromatin folding.* Nucleic Acids Research, vol. 45, no. 10, pages e81–e81, January 2017. (Cited on page 96.)

[Hänsel-Hertsch *et al.* 2016] Robert Hänsel-Hertsch, Dario Beraldi, Stefanie V. Lensing, Giovanni Marsico, Katherine Zyner, Aled Parry, Marco Di Antonio, Jeremy Pike, Hiroshi Kimura, Masashi Narita, David Tannahill and Shankar Balasubramanian. *G-quadruplex structures mark human regulatory chromatin.* Nature Genetics, vol. 48, no. 10, pages 1267–1272, September 2016. (Cited on pages 182 and 198.)

[Hänsel-Hertsch *et al.* 2018] Robert Hänsel-Hertsch, Jochen Spiegel, Giovanni Marsico, David Tannahill and Shankar Balasubramanian. *Genome-wide mapping of endogenous G-quadruplex DNA structures by chromatin immunoprecipitation and high-throughput sequencing.* Nature Protocols, vol. 13, no. 3, pages 551–564, 2018. (Cited on page 198.)

[Hänsel-Hertsch *et al.* 2020] Robert Hänsel-Hertsch, Angela Simeone, Abigail Shea, Winnie W. I. Hui, Katherine G. Zyner, Giovanni Marsico, Oscar M. Rueda, Alejandra Bruna, Alistair Martin, Xiaoyun Zhang, Santosh Adhikari, David Tannahill, Carlos Caldas and Shankar Balasubramanian. *Landscape of G-quadruplex DNA structural regions in breast cancer.* Nature Genetics, vol. 52, no. 9, pages 878–883, September 2020. (Cited on pages 182 and 198.)

[Hastie *et al.* 2009] Trevor Hastie, Robert Tibshirani and Jerome Friedman. The elements of statistical learning: data mining, inference and prediction. Springer, 2 édition, 2009. (Cited on page 21.)

[He *et al.* 2014] Bing He, Changya Chen, Li Teng and Kai Tan. *Global view of enhancer-promoter interactome in human cells.* Proceedings of the National Academy of Sciences of the United States of America, vol. 111, no. 21, pages 201320308–E2199, May 2014. (Cited on page 58.)

[Heger *et al.* 2012] Peter Heger, Birger Marin, Marek Bartkuhn, Einhard Schierenberg and Thomas Wiehe. *The chromatin insulator CTCF and the emergence*

of metazoan diversity. Proceedings of the National Academy of Sciences, vol. 109, no. 43, pages 17507–17512, October 2012. (Cited on page 112.)

[Hesselberth *et al.* 2009] Jay R. Hesselberth, Xiaoyu Chen, Zhihong Zhang, Peter J. Sabo, Richard Sandstrom, Alex P. Reynolds, Robert E. Thurman, Shane Neph, Michael S. Kuehn, William S. Noble, Stanley Fields and John A. Stamatoyannopoulos. *Global mapping of protein-DNA interactions in vivo by digital genomic footprinting.* Nature Methods, vol. 6, no. 4, pages 283–289, March 2009. (Cited on page 16.)

[Heurteau *et al.* 2020] Alexandre Heurteau, Charlène Perrois, David Depierre, Olivier Fosseprez, Jonathan Humbert, Stéphane Schaak and Olivier Cuvier. *Insulator-based loops mediate the spreading of H3K27me3 over distant micro-domains repressing euchromatin genes.* Genome Biology, vol. 21, no. 1, page 193, Aug 2020. (Cited on page 120.)

[Hnisz *et al.* 2016] Denes Hnisz, Abraham S. Weintraub, Daniel S. Day, Anne-Laure Valton, Rasmus O. Bak, Charles H. Li, Johanna Goldmann, Bryan R. Lajoie, Zi Peng Fan, Alla A. Sigova, Jessica Reddy, Diego Borges-Rivera, Tong Ihn Lee, Rudolf Jaenisch, Matthew H. Porteus, Job Dekker and Richard A. Young. *Activation of proto-oncogenes by disruption of chromosome neighborhoods.* Science, 2016. (Cited on pages 13 and 58.)

[Hochreiter & Schmidhuber 1997] Sepp Hochreiter and Jürgen Schmidhuber. *Long short-term memory.* Neural computation, vol. 9, no. 8, pages 1735–1780, November 1997. (Cited on page 23.)

[Hore *et al.* 2008] Timothy A. Hore, Janine E. Deakin and Jennifer A. Marshall Graves. *The evolution of epigenetic regulators CTCF and BORIS/CTCFL in amniotes.* PLOS Genetics, vol. 4, no. 8, pages 1–11, August 2008. (Cited on page 112.)

[Horner *et al.* 2009] David Stephen Horner, Giulio Pavesi, Tiziana Castrignano, Paolo D'Onorio De Meo, Sabino Liuni, Michael Sammeth, Ernesto Picardi and Graziano Pesole. *Bioinformatics approaches for genomics and post genomics applications of next-generation sequencing.* Briefings in Bioinformatics, vol. 11, no. 2, pages 181–197, October 2009. (Cited on page 19.)

[Hou *et al.* 2012] Chunhui Hou, Li Li, S. Qin Zhaohui and Victor G. Corces. *Gene density, transcription, and insulators contribute to the partition of the Drosophila genome into physical domains.* Molecular Cell, vol. 48, no. 3, pages 471–484, November 2012. (Cited on page 14.)

[Hsu *et al.* 2010] Pei-Yin Hsu, Hang-Kai Hsu, Gregory A.C. Singer, Pearlly S. Yan, Benjamin A.T. Rodriguez, Joseph C. Liu, Yu-I Weng, Daniel E. Deatherage, Zhong Chen, Julia S. Pereira, Ricardo Lopez, Jose Russo, Qianben Wang, Coral A. Lamartiniere, Kenneth P. Nephew and Tim H.-M. Huang. *Estrogen-mediated epigenetic repression of large chromosomal regions through DNA*

looping. Genome Research, vol. 20, no. 6, pages 733–744, June 2010. (Cited on page 30.)

[Hsu *et al.* 2013] Pei-Yin Hsu, Hang-Kai Hsu, Xun Lan, Liran Juan, Pearlly S. Yan, Jadwiga Labanowska, Nyla Heerema, Tzu-Hung Hsiao, Yu-Chiao Chiu, Yidong Chen, Yunlong Liu, Lang Li, Rong Li, Ian M. Thompson, Kenneth P. Nephew, Zelton D. Sharp, Nameer B. Kirma, Victor X. Jin and Tim H.-M. Huang. *Amplification of distant estrogen response elements deregulates target genes associated with tamoxifen resistance in breast cancer*. Cancer Cell, vol. 24, no. 2, pages 197 – 212, 2013. (Cited on page 30.)

[Hübner & Spector 2010] Michael R. Hübner and David L. Spector. *Chromatin dynamics*. Annual Review of Biophysics, vol. 39, no. 1, pages 471–489, June 2010. (Cited on page 12.)

[Huppert & Balasubramanian 2005] Julian L. Huppert and Shankar Balasubramanian. *Prevalence of quadruplexes in the human genome*. Nucleic Acids Research, vol. 33, no. 9, pages 2908–2916, January 2005. (Cited on page 198.)

[Huppert & Balasubramanian 2006] Julian L. Huppert and Shankar Balasubramanian. *G-quadruplexes in promoters throughout the human genome*. Nucleic Acids Research, vol. 35, no. 2, pages 406–413, December 2006. (Cited on page 198.)

[Huynh & Hormozdiari 2019] Linh Huynh and Fereydoun Hormozdiari. *TAD fusion score: discovery and ranking the contribution of deletions to genome structure*. Genome Biology, vol. 20, no. 1, page 60, March 2019. (Cited on page 96.)

[Igartua *et al.* 2015] Catherine Igartua, Rachel A. Myers, Rasika A. Mathias, Maria Pino-Yanes, Celeste Eng, Penelope E. Graves, Albert M. Levin, Blanca E. Del-Rio-Navarro, Daniel J. Jackson, Oren E. Livne, Nicholas Rafaels, Christopher K. Edlund, James J. Yang, Scott Huntsman, Muhammad T. Salam, Isabelle Romieu, Raphael Mourad, James E. Gern, Robert F. Lemanske, Annah Wyss, Jane A. Hoppin, Kathleen C. Barnes, Esteban G. Burchard, W. James Gauderman, Fernando D. Martinez, Benjamin A. Raby, Scott T. Weiss, L. Keoki Williams, Stephanie J. London, Frank D. Gilliland, Dan L. Nicolae and Carole Ober. *Ethnic-specific associations of rare and low-frequency DNA sequence variants with asthma*. Nature Communications, vol. 6, no. 1, page 5965, January 2015. (Cited on page 29.)

[International Human Genome Sequencing Consortium 2001] International Human Genome Sequencing Consortium. *Initial sequencing and analysis of the human genome*. Nature, vol. 409, no. 6822, pages 860–921, February 2001. (Cited on pages 12 and 15.)

[Jain & Medsker 1999] L. C. Jain and L. R. Medsker. Recurrent neural networks: Design and applications. CRC Press, Inc., USA, 1st édition, 1999. (Cited on page 23.)

[Jensen 1996] Finn V. Jensen. Introduction to bayesian networks. Springer-Verlag, Berlin, Heidelberg, 1st édition, 1996. (Cited on page 21.)

[Jin *et al.* 2013] Fulai Jin, Yan Li, Jesse R. Dixon, Siddarth. Selvaraj, Zhen. Ye, Ah Young Lee, Chia-An Yen, Anthony D. Schmitt, Celso A. Espinoza and Bing Ren. *A high-resolution map of the three-dimensional chromatin interactome in human cells.* Nature, vol. 503, no. 7475, pages 290–294, November 2013. (Cited on pages 58 and 85.)

[Jost *et al.* 2014] Daniel Jost, Pascal Carrivain, Giacomo Cavalli and Cédric Vaillant. *Modeling epigenome folding: formation and dynamics of topologically associated chromatin domains.* Nucleic Acids Research, vol. 42, no. 15, pages 9553–9561, August 2014. (Cited on page 31.)

[Kadauke & Blobel 2009] Stephan Kadauke and Gerd A. Blobel. *Chromatin loops in gene regulation.* Biochimica et Biophysica Acta, vol. 1789, pages 17–25, January 2009. (Cited on page 13.)

[Kaplan 2019] Noam Kaplan. *Explicit probabilistic models for exploiting and explaining the 3D genome.* In Proceedings of Statistics for Post Genomic Data (SMPGD 2019), January 2019. (Cited on page 96.)

[Kasparek & Humphrey 2011] Torben R. Kasparek and Timothy C. Humphrey. *DNA double-strand break repair pathways, chromosomal rearrangements and cancer.* Seminars in Cell & Developmental Biology, vol. 22, no. 8, pages 886–897, October 2011. (Cited on page 14.)

[Khajavinia & Makalowski 2007] A. Khajavinia and W. Makalowski. *What is "junk" DNA, and what is it worth?* Scientific American, vol. 296, no. 5, page 104, May 2007. (Cited on page 12.)

[Kinner *et al.* 2008] Andrea Kinner, Wenqi Wu, Christian Staudt and George Iliakis. *γ-H2AX in recognition and signaling of DNA double-strand breaks in the context of chromatin.* Nucleic Acids Research, vol. 36, no. 17, pages 5678–5694, October 2008. (Cited on page 183.)

[Kourou *et al.* 2015] Konstantina Kourou, Themis P. Exarchos, Konstantinos P. Exarchos, Michalis V. Karamouzis and Dimitrios I. Fotiadis. *Machine learning applications in cancer prognosis and prediction.* Computational and Structural Biotechnology Journal, vol. 13, pages 8 – 17, November 2015. (Cited on page 21.)

[Krizhevsky *et al.* 2012] Alex Krizhevsky, Ilya Sutskever and Geoffrey E. Hinton. *ImageNet Classification with Deep Convolutional Neural Networks.* In F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 25, pages 1097–1105. Curran Associates, Inc., 2012. (Cited on page 22.)

[Lawson *et al.* 2018] Devon A. Lawson, Kai Kessenbrock, Ryan T. Davis, Nicholas Pervolarakis and Zena Werb. *Tumour heterogeneity and metastasis at single-cell resolution.* Nature Cell Biology, vol. 20, no. 12, pages 1349–1360, December 2018. (Cited on page 18.)

[Lee & Lupski 2006] Jennifer A. Lee and James R. Lupski. *Genomic Rearrangements and Gene Copy-Number Alterations as a Cause of Nervous System Disorders.* Neuron, vol. 52, no. 1, pages 103–121, Oct 2006. (Cited on page 219.)

[Lee *et al.* 2011] Dongwon Lee, Rachel Karchin and Michael A. Beer. *Discriminative prediction of mammalian enhancers from DNA sequence.* Genome Research, vol. 21, no. 12, pages 2167–2180, August 2011. (Cited on page 21.)

[Lee *et al.* 2020] Bohyun Lee, Shuo Zhang, Aleksandar Poleksic and Lei Xie. *Heterogeneous multi-layered network model for omics data integration and analysis.* Frontiers in Genetics, vol. 10, page 1381, January 2020. (Cited on page 23.)

[Lensing *et al.* 2016] Stefanie V. Lensing, Giovanni Marsico, Robert Hansel-Hertsch, Enid Y. Lam, David Tannahill and Shankar Balasubramanian. *DSBCapture: in situ capture and sequencing of DNA breaks.* Nature Methods, vol. 13, no. 10, pages 855–857, October 2016. (Cited on pages 121, 182 and 183.)

[Lesk 2002] Arthur M. Lesk. Introduction to bioinformatics. Oxford University Press, Inc., USA, 2002. (Cited on page 19.)

[Levy-Leduc *et al.* 2014] Celine Levy-Leduc, M. Delattre, T. Mary-Huard and S. Robin. *Two-dimensional segmentation for analyzing Hi-C data.* Bioinformatics, vol. 30, no. 17, pages i386–i392, September 2014. (Cited on page 96.)

[Lieberman-Aiden *et al.* 2009] Erez Lieberman-Aiden, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R. Lajoie, Peter J. Sabo, Michael O. Dorschner, Richard Sandstrom, Bradley Bernstein, M. A. Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A. Mirny, Eric S. Lander and Job Dekker. *Comprehensive mapping of long-range interactions reveals folding principles of the human genome.* Science, vol. 326, no. 5950, pages 289–293, October 2009. (Cited on pages 13, 16 and 85.)

[Lowe *et al.* 2017] Rohan Lowe, Neil Shirley, Mark Bleackley, Stephen Dolan and Thomas Shafee. *Transcriptomics technologies.* PLOS Computational Biology, vol. 13, no. 5, pages 1–23, May 2017. (Cited on page 16.)

[Lupiáñez *et al.* 2015] Darío G. Lupiáñez, Katerina Kraft, Verena Heinrich, Peter Krawitz, Francesco Brancati, Eva Klopocki, Denise Horn, Hülya Kayserili,

John M. Opitz, Renata Laxova, Fernando Santos-Simarro, Brigitte Gilbert-Dussardier, Lars Wittler, Marina Borschiwer, Stefan A. Haas, Marco Osterwalder, Martin Franke, Bernd Timmermann, Jochen Hecht, Malte Spielmann, Axel Visel and Stefan Mundlos. *Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions.* Cell, vol. 161, no. 5, pages 1012–1025, September 2015. (Cited on pages 13 and 58.)

[MacIntyre *et al.* 2003] D. J. MacIntyre, D. H. R. Blackwood, D. J. Porteous, B. S. Pickard and W. J. Muir. *Chromosomal abnormalities and mental illness.* Molecular Psychiatry, vol. 8, no. 3, pages 275–287, March 2003. (Cited on page 15.)

[Mahood *et al.* 2020] Elizabeth H. Mahood, Lars H. Kruse and Gaurav D. Moghe. *Machine learning: A powerful tool for gene function prediction in plants.* Applications in Plant Sciences, vol. 8, no. 7, page e11376, July 2020. (Cited on page 21.)

[Marnef *et al.* 2017] Aline Marnef, Sarah Cohen and Gaëlle Legube. *Transcription-coupled DNA double-strand break repair: Active genes need special care.* Journal of Molecular Biology, vol. 429, no. 9, pages 1277 – 1288, March 2017. (Cited on pages 14, 121 and 182.)

[Marsman & Horsfield 2012] Judith Marsman and Julia A. Horsfield. *Long distance relationships: Enhancer-promoter communication and dynamic gene transcription.* Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms, vol. 1819, no. 11-12, pages 1217–1227, 2012. (Cited on page 12.)

[Maurano *et al.* 2012] Matthew T. Maurano, Richard Humbert, Eric Rynes, Robert E. Thurman, Eric Haugen, Hao Wang, Alex P. Reynolds, Richard Sandstrom, Hongzhu Qu, Jennifer Brody, Anthony Shafer, Fidencio Neri, Kristen Lee, Tanya Kutyavin, Sandra Stehling-Sun, Audra K. Johnson, Theresa K. Canfield, Erika Giste, Morgan Diegel, Daniel Bates, R. Scott Hansen, Shane Neph, Peter J. Sabo, Shelly Heimfeld, Antony Raubitschek, Steven Ziegler, Chris Cotsapas, Nona Sotoodehnia, Ian Glass, Shamil R. Sunyaev, Rajinder Kaul and John A. Stamatoyannopoulos. *Systematic localization of common disease-associated variation in regulatory DNA.* Science, vol. 337, no. 6099, pages 1190–1195, September 2012. (Cited on page 18.)

[McKinnon & Caldecott 2007] Peter J. McKinnon and Keith W. Caldecott. *DNA strand break repair and human genetic disease.* Annual Review of Genomics and Human Genetics, vol. 8, no. 1, pages 37–55, September 2007. (Cited on page 14.)

[Mehta & Haber 2014] Anuja Mehta and James E. Haber. *Sources of DNA double-strand breaks and models of recombinational DNA repair.* Cold Spring Harbor Perspectives in Biology, vol. 6, no. 9, page a016428, August 2014. (Cited on page 14.)

[Mobadersany *et al.* 2018] Pooya Mobadersany, Safoora Yousefi, Mohamed Amgad, David A. Gutman, Jill S. Barnholtz-Sloan, José E. Velázquez Vega, Daniel J. Brat and Lee A. D. Cooper. *Predicting cancer outcomes from histology and genomics using convolutional networks.* Proceedings of the National Academy of Sciences, vol. 115, no. 13, pages E2970–E2979, March 2018. (Cited on page 24.)

[modENCODE Consortium *et al.* 2010] The modENCODE Consortium, Sushmita Roy, Jason Ernst, Peter V. Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L. Eaton, Jane M. Landolin, Christopher A. Bristow, Lijia Ma, Michael F. Lin, Stefan Washietl, Bradley I. Arshinoff, Ferhat Ay, Patrick E. Meyer, Nicolas Robine, Nicole L. Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D. Brown, Rogerio Candeias, Joseph W. Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y. Tolstorukov, Sebastian Will, Artyom A. Alekseyenko, Carlo Artieri, Benjamin W. Booth, Angela N. Brooks, Qi Dai, Carrie A. Davis, Michael O. Duff, Xin Feng, Andrey A. Gorchakov, Tingting Gu, Jorja G. Henikoff, Philipp Kapranov, Renhua Li, Heather K. MacAlpine, John Malone, Aki Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K. Powell, Nicole C. Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E. Sandler, Yuri B. Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H. Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E. Brenner, Michael R. Brent, Lucy Cherbas, Sarah C. R. Elgin, Thomas R. Gingeras, Robert Grossman, Roger A. Hoskins, Thomas C. Kaufman, William Kent, Mitzi I. Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W. Posakony, Bing Ren, Steven Russell, Peter Cherbas, Brenton R. Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J. Park, Susan E. Celniker, Steven Henikoff, Gary H. Karpen, Eric C. Lai, David M. MacAlpine, Lincoln D. Stein, Kevin P. White, Manolis Kellis, David Acevedo, Richard Auburn, Galt Barber, Hugo J. Bellen, Eric P. Bishop, Terri D. Bryson, Aurelien Chateigner, Jia Chen, Hiram Clawson, Charles L. G. Comstock, Sergio Contrino, Leyna C. DeNapoli, Queying Ding, Alex Dobin, Marc H. Domanus, Jorg Drenkow, Sandrine Dudoit, Jackie Dumais, Thomas Eng, Delphine Fagegaltier, Sarah E. Gadel, Srinka Ghosh, Francois Guillier, David Hanley, Gregory J. Hannon, Kasper D. Hansen, Elizabeth Heinz, Angie S. Hinrichs, Martin Hirst, Sonali Jha, Lichun Jiang, Youngsook L. Jung, Helena Kashevsky, Cameron D. Kennedy, Ellen T. Kephart, Laura Langton, Ok-Kyung Lee, Sharon Li, Zirong Li, Wei Lin, Daniela Linder-Basso, Paul Lloyd, Rachel Lyne, Sarah E. Marchetti, Marco Marra, Nicolas R. Mattiuzzo, Sheldon McKay, Folker Meyer, David Miller, Steven W. Miller, Richard A. Moore, Carolyn A. Morrison, Joseph A. Prinz, Michelle Rooks, Richard Moore, Kim M. Rutherford, Peter Ruzanov, Douglas A. Scheftner, Lionel Senderowicz, Parantu K. Shah, Gregory Shanower, Richard Smith, E. O. Stinson, Sarah Suchy, Aaron E.

Tenney, Feng Tian, Koen J. T. Venken, Huaien Wang, Robert White, Jared Wilkening, Aarron T. Willingham, Chris Zaleski, Zheng Zha, Dayu Zhang, Yongjun Zhao and Jennifer Zieba. *Identification of functional elements and regulatory circuits by Drosophila modENCODE.* Science, vol. 330, no. 6012, pages 1787–1797, December 2010. (Cited on pages 12 and 120.)

[Moindrot *et al.* 2012] Benoit Moindrot, Benjamin Audit, Petra Klous, Antoine Baker, Claude Thermes, Wouter de Laat, Philippe Bouvet, Fabien Mongelard and Alain Arneodo. *3D chromatin conformation correlates with replication timing and is conserved in resting cells.* Nucleic Acids Research, October 2012. (Cited on page 12.)

[Morey & Helin 2010] Lluis Morey and Kristian Helin. *Polycomb group protein-mediated repression of transcription.* Trends in Biochemical Sciences, vol. 35, no. 6, pages 323–332, 2010. (Cited on page 120.)

[Mourad & Cuvier 2015] Raphaël Mourad and Olivier Cuvier. *Predicting the spatial organization of chromosomes using epigenetic data.* Genome Biology, vol. 16, no. 1, pages 1–3, August 2015. (Cited on page 31.)

[Mourad & Cuvier 2016] Raphael Mourad and Olivier Cuvier. *Computational identification of genomic features that influence 3D chromatin domain formation.* PLoS Computational Biology, vol. 12, no. 5, page e1004908, May 2016. (Cited on pages 32 and 85.)

[Mourad & Cuvier 2018] Raphael Mourad and Olivier Cuvier. *TAD-free analysis of architectural proteins and insulators.* Nucleic Acids Research, vol. 46, no. 5, page e27, March 2018. (Cited on pages 85 and 96.)

[Mourad *et al.* 2014] Raphaël Mourad, Pei-Yin Hsu, Liran Juan, Changyu Shen, Prasad Koneru, Hai Lin, Yunlong Liu, Kenneth Nephew, Tim H. Huang and Lang Li. *Estrogen induces global reorganization of chromatin structure in human breast cancer cells.* PLOS ONE, vol. 9, no. 12, pages 1–24, December 2014. (Cited on page 31.)

[Mourad *et al.* 2015] Raphaël Mourad, François Chevennet, David T. Dunn, Esther Fearnhill, Valerie Delpech, David Asboe, Olivier Gascuel, Stéphane Hue and Anti-HIV Drug Resistance Network on behalf of the UK HIV Drug Resistance Database & the Collaborative HIV. *A phylotype-based analysis highlights the role of drug-naive HIV-positive individuals in the transmission of antiretroviral resistance in the UK.* AIDS, vol. 29, no. 15, September 2015. (Cited on page 30.)

[Mourad *et al.* 2017] Raphaël Mourad, Lang Li and Olivier Cuvier. *Uncovering direct and indirect molecular determinants of chromatin loops using a computational integrative approach.* PLOS Computational Biology, vol. 13, no. 5, pages 1–25, May 2017. (Cited on page 58.)

[Mourad *et al.* 2018] Raphaël Mourad, Krzysztof Ginalski, Gaëlle Legube and Olivier Cuvier. *Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution.* Genome Biology, vol. 19, no. 1, page 34, March 2018. (Cited on pages 183 and 217.)

[Mourad 2019] Raphaël Mourad. *Studying 3D genome evolution using genomic sequence.* Bioinformatics, vol. 36, no. 5, pages 1367–1373, October 2019. (Cited on pages 112 and 216.)

[Nawy 2014] Tal Nawy. *Single-cell sequencing.* Nature Methods, vol. 11, no. 1, pages 18–18, January 2014. (Cited on page 18.)

[Negre *et al.* 2010] Nicolas Negre, Christopher D. Brown, Parantu K. Shah, Pouya Kheradpour, Carolyn A. Morrison, Jorja G. Henikoff, Xin Feng, Kami Ahmad, Steven Russell, Robert A. H. White, Lincoln Stein, Steven Henikoff, Manolis Kellis and Kevin P. White. *A comprehensive map of insulator elements for the Drosophila genome.* PLoS Genetics, vol. 6, no. 1, pages e1000814+, January 2010. (Cited on page 120.)

[Nicolae *et al.* 2010] Dan L. Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M. Eileen Dolan and Nancy J. Cox. *Trait-associated SNPs are more likely to be eQTLs: Annotation to enhance discovery from GWAS.* PLOS Genetics, vol. 6, no. 4, pages 1–10, April 2010. (Cited on page 218.)

[Norton *et al.* 2018] Heidi K. Norton, Daniel J. Emerson, Harvey Huang, Jesi Kim, Katelyn R. Titus, Shi Gu, Danielle S. Bassett and Jennifer E. Phillips-Cremins. *Detecting hierarchical genome folding with network modularity.* Nature Methods, vol. 15, pages 119–122, January 2018. (Cited on page 96.)

[Nye *et al.* 2002] Anne C. Nye, Ramji R. Rajendran, David L. Stenoien, Michael A. Mancini, Benita S. Katzenellenbogen and Andrew S. Belmont. *Alteration of large-scale chromatin structure by estrogen receptor.* Molecular and Cellular Biology, vol. 22, no. 10, pages 3437–3449, May 2002. (Cited on page 30.)

[Oluwadare & Cheng 2017] Oluwatosin Oluwadare and Jianlin Cheng. *ClusterTAD: an unsupervised machine learning approach to detecting topologically associated domains of chromosomes from Hi-C data.* BMC Bioinformatics, vol. 18, no. 1, page 480, November 2017. (Cited on page 96.)

[Pancaldi *et al.* 2016] Vera Pancaldi, Enrique Carrillo-de Santa-Pau, Biola Maria Javierre, David Juan, Peter Fraser, Mikhail Spivakov, Alfonso Valencia and Daniel Rico. *Integrating epigenomic data and 3D genomic structure with a new measure of chromatin assortativity.* Genome Biology, vol. 17, no. 1, pages 1–19, July 2016. (Cited on page 58.)

[Phillips-Cremins *et al.* 2013] Jennifer E. Phillips-Cremins, Michael E. G. Sauria, Amartya Sanyal, Tatiana I. Gerasimova, Bryan R. Lajoie, Joshua S. K. Bell, Chin-Tong Ong, Tracy A. Hookway, Changying Guo, Yuhua Sun, Michael J. Bland, William Wagstaff, Stephen Dalton, Todd C. McDevitt, Ranjan Sen,

Job Dekker, James Taylor and Victor G. Corces. *Architectural protein subclasses shape 3D organization of genomes during lineage commitment.* Cell, vol. 153, no. 6, pages 1281–1295, June 2013. (Cited on page 31.)

[Pommier *et al.* 2016] Yves Pommier, Yilun Sun, Shar-yin N. Huang and John L. Nitiss. *Roles of eukaryotic topoisomerases in transcription, replication and genomic stability.* Nature Reviews Molecular Cell Biology, vol. 17, no. 11, pages 703–721, September 2016. (Cited on page 15.)

[Pope *et al.* 2014] Benjamin D. Pope, Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas, Daniel L. Vera, Yanli Wang, R. Scott Hansen, Theresa K. Canfield, Robert E. Thurman, Yong Cheng, Gunhan Gulsoy, Jonathan H. Dennis, Michael P. Snyder, John A. Stamatoyannopoulos, James Taylor, Ross C. Hardison, Tamer Kahveci, Bing Ren and David M. Gilbert. *Topologically associating domains are stable units of replication-timing regulation.* Nature, vol. 515, no. 7527, pages 402–405, November 2014. (Cited on pages 13 and 58.)

[Price & D'Andrea 2013] Brendan D. Price and Alan D. D'Andrea. *Chromatin remodeling at DNA double-strand breaks.* Cell, vol. 152, no. 6, pages 1344–1354, March 2013. (Cited on page 183.)

[Ramani *et al.* 2017] Vijay Ramani, Xinxian Deng, Ruolan Qiu, Kevin L. Gunderson, Frank J. Steemers, Christine M. Disteche, William S. Noble, Zhijun Duan and Jay Shendure. *Massively multiplex single-cell Hi-C.* Nature Methods, vol. 14, pages 263–266, January 2017. (Cited on page 220.)

[Rao *et al.* 2014] Suhas S. P. Rao, Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, Ido Machol, Arina D. Omer, Eric S. Lander and Erez L. Aiden. *A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping.* Cell, vol. 159, no. 7, pages 1665–1680, December 2014. (Cited on pages 14, 31, 85, 96 and 112.)

[Rao *et al.* 2017] Suhas S. P. Rao, Su-Chen Huang, Brian Glenn St Hilaire, Jesse M. Engreitz, Elizabeth M. Perez, Kyong-Rim Kieffer-Kwon, Adrian L. Sanborn, Sarah E. Johnstone, Gavin D. Bascom, Ivan D. Bochkov, Xingfan Huang, Muhammad S. Shamim, Jaeweon Shin, Douglass Turner, Ziyi Ye, Arina D. Omer, James T. Robinson, Tamar Schlick, Bradley E. Bernstein, Rafael Casellas, Eric S. Lander and Erez Lieberman Aiden. *Cohesin loss eliminates all loop domains.* Cell, vol. 171, no. 2, pages 305–320.e24, October 2017. (Cited on page 31.)

[Ravichandran *et al.* 2019] Subramaniyam Ravichandran, Vinod Kumar Subramani and Kyeong Kyu Kim. *Z-DNA in the genome: from structure to disease.* Biophysical Reviews, vol. 11, no. 3, pages 383–387, June 2019. (Cited on page 11.)

[Rhodes & Lipps 2015] Daniela Rhodes and Hans J. Lipps. *G-quadruplexes and their regulatory roles in biology.* Nucleic Acids Research, vol. 43, no. 18, pages 8627–8637, 10 2015. (Cited on page 11.)

[Risso *et al.* 2018] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit and Jean-Philippe Vert. *A general and flexible method for signal extraction from single-cell RNA-seq data.* Nature Communications, vol. 9, no. 1, page 284, January 2018. (Cited on page 20.)

[Robinson & Smyth 2007a] Mark D. Robinson and Gordon K. Smyth. *Moderated statistical tests for assessing differences in tag abundance.* Bioinformatics, vol. 23, no. 21, pages 2881–2887, September 2007. (Cited on page 20.)

[Robinson & Smyth 2007b] Mark D. Robinson and Gordon K. Smyth. *Small-sample estimation of negative binomial dispersion, with applications to SAGE data.* Biostatistics, vol. 9, no. 2, pages 321–332, August 2007. (Cited on page 20.)

[Robinson *et al.* 2009] Mark D. Robinson, Davis J. McCarthy and Gordon K. Smyth. *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data.* Bioinformatics, vol. 26, no. 1, pages 139–140, November 2009. (Cited on page 19.)

[Rocher *et al.* 2021] Vincent Rocher, Matthieu Genais, Elissar Nassereddine and Raphael Mourad. *DeepG4: A deep learning approach to predict cell-type specific active G-quadruplex regions.* PLOS Computational Biology, vol. 17, no. 8, pages 1–15, 08 2021. (Cited on page 198.)

[Rohart *et al.* 2017] F Rohart, B Gautier, A Singh and K Cao. *mixOmics: an R package for omics feature selection and multiple data integration.* PLOS Computational Biology, vol. 13, no. 11, page e1005752, November 2017. (Cited on page 23.)

[Rosenblatt 1958] F. Rosenblatt. *The perceptron: A probabilistic model for information storage and organization in the brain.* Psychological Review, vol. 65, no. 6, pages 386–408, 1958. (Cited on page 21.)

[Rotem *et al.* 2015] Assaf Rotem, Oren Ram, Noam Shoresh, Ralph A. Sperling, Alon Goren, David A. Weitz and Bradley E. Bernstein. *Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state.* Nature Biotechnology, vol. 33, pages 1165–1172, Oct 2015. (Cited on page 220.)

[Rousseau *et al.* 2010] Philippe Rousseau, Catherine Tardin, Nathalie Tolou, Laurence Salomé and Mick Chandler. *A model for the molecular organisation of the IS911 transpososome.* Mobile DNA, vol. 1, no. 1, page 16, June 2010. (Cited on page 217.)

[Rowley *et al.* 2017] M. Jordan Rowley, Michael H. Nichols, Xiaowen Lyu, Masami Ando-Kuri, I. Sarahi M. Rivera, Karen Hermetz, Ping Wang, Yijun Ruan

and Victor G. Corces. *Evolutionarily conserved principles predict 3D chromatin organization*. Molecular Cell, vol. 67, no. 5, pages 837–852.e7, September 2017. (Cited on pages 14 and 96.)

[Sadowski *et al.* 2019] Michal Sadowski, Agnieszka Kraft, Przemyslaw Szalaj, Michal Wlasnowolski, Zhonghui Tang, Yijun Ruan and Dariusz Plewczynski. *Spatial chromatin architecture alteration by structural variations in human genomes at the population scale*. Genome Biology, vol. 20, no. 1, page 148, July 2019. (Cited on page 96.)

[Sanborn *et al.* 2015] Adrian L. Sanborn, Suhas S. P. Rao, Su-Chen Huang, Neva C. Durand, Miriam H. Huntley, Andrew I. Jewett, Ivan D. Bochkov, Dharmaraj Chinnappan, Ashok Cutkosky, Jian Li, Kristopher P. Geeting, Andreas Gnirke, Alexandre Melnikov, Doug McKenna, Elena K. Stamenova, Eric S. Lander and Erez Lieberman Aiden. *Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes*. Proceedings of the National Academy of Sciences, vol. 112, no. 47, pages E6456–E6465, November 2015. (Cited on pages 14, 31 and 112.)

[Sen & Gilbert 1988] Dipankar Sen and Walter Gilbert. *Formation of parallel four-stranded complexes by guanine-rich motifs in DNA and its implications for meiosis*. Nature, vol. 334, no. 6180, pages 364–366, July 1988. (Cited on page 11.)

[Serra *et al.* 2017] Francois Serra, Davide Bau, Mike Goodstadt, David Castillo, Guillaume J. Filion and Marc A. Marti-Renom. *Automatic analysis and 3D-modelling of Hi-C data using TADbit reveals structural features of the fly chromatin colors*. PLOS Computational Biology, vol. 13, no. 7, pages 1–17, July 2017. (Cited on page 96.)

[Shin *et al.* 2016] Hanjun Shin, Yi Shi, Chao Dai, Harianto Tjong, Ke Gong, Frank Alber and Xianghong Jasmine Zhou. *TopDom: an efficient and deterministic method for identifying topological domains in genomes*. Nucleic Acids Research, vol. 44, no. 7, page e70, April 2016. (Cited on pages 85 and 96.)

[Spiegel *et al.* 2019] Jochen Spiegel, Santosh Adhikari and Shankar Balasubramanian. *The structure and function of DNA G-quadruplexes*. Trends in Chemistry, January 2019. (Cited on pages 11 and 198.)

[Spiegel *et al.* 2021] Jochen Spiegel, Sergio Martínez Cuesta, Santosh Adhikari, Robert Hänsel-Hertsch, David Tannahill and Shankar Balasubramanian. *G-quadruplexes are transcription factor binding hubs in human chromatin*. Genome Biology, vol. 22, no. 1, page 117, April 2021. (Cited on page 182.)

[Sutton & Barto 2018] Richard S. Sutton and Andrew G. Barto. Reinforcement Learning: An Introduction. The MIT Press, second édition, 2018. (Cited on page 22.)

[Tan *et al.* 2018] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang and Chunfang Liu. *A survey on deep transfer learning*, 2018. (Cited on page 23.)

[The ENCODE Consortium 2012] The ENCODE Consortium. *An integrated encyclopedia of DNA elements in the human genome.* Nature, vol. 489, no. 7414, pages 57–74, September 2012. (Cited on pages 12 and 183.)

[Tsai *et al.* 2015] Shengdar Q Tsai, Zongli Zheng, Nhu T Nguyen, Matthew Liebers, Ved V Topkar, Vishal Thapar, Nicolas Wyvekens, Cyd Khayter, A John Lafrate, Long P Le, Martin J Aryee and J Keith Joung. *GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases.* Nature Biotechnology, vol. 33, no. 2, pages 187–197, February 2015. (Cited on page 182.)

[Ulianov *et al.* 2016] Sergey V. Ulianov, Ekaterina E. Khrameeva, Alexey A. Gavrilov, Ilya M. Flyamer, Pavel Kos, Elena A. Mikhaleva, Aleksey A. Penin, Maria D. Logacheva, Maxim V. Imakaev, Alexander Chertovich, Mikhail S. Gelfand, Yuri Y. Shevelyov and Sergey V. Razin. *Active chromatin and transcription play a key role in chromosome partitioning into topologically associating domains.* Genome Research, vol. 26, no. 1, pages 70–84, January 2016. (Cited on page 12.)

[Uuskü la-Reimand *et al.* 2016] Liis Uuskü la-Reimand, Huayun Hou, Payman Samavarchi-Tehrani, Matteo Vietri Rudan, Minggao Liang, Alejandra Medina-Rivera, Hisham Mohammed, Dominic Schmidt, Petra Schwalie, Edwin J. Young, Jüri Reimand, Suzana Hadjur, Anne-Claude Gingras and Michael D. Wilson. *Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders.* Genome Biology, vol. 17, no. 1, page 182, August 2016. (Cited on pages 12 and 15.)

[Van Bortle *et al.* 2012] Kevin Van Bortle, Edward Ramos, Naomi Takenaka, Jingping Yang, Jessica E. Wahi and Victor G. Corces. *Drosophila CTCF tandemly aligns with other insulator proteins at the borders of H3K27me3 domains.* Genome Research, vol. 22, no. 11, pages 2176–2187, November 2012. (Cited on page 120.)

[Van Bortle *et al.* 2014] Kevin Van Bortle, Michael H. Nichols, Li Li, Chin-Tong Ong, Naomi Takenaka, Zhaohui S. Qin and Victor G. Corces. *Insulator function and topological domain border strength scale with architectural protein occupancy.* Genome Biology, vol. 15, no. 5, pages R82+, June 2014. (Cited on pages 14, 85 and 120.)

[Vaswani *et al.* 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. *Attention is all you need.* Arxiv, June 2017. (Cited on page 23.)

[Vietri-Rudan *et al.* 2015] Matteo Vietri-Rudan, Christopher Barrington, Stephen Henderson, Christina Ernst, Duncan T. Odom, Amos Tanay and Suzana

Hadjur. *Comparative Hi-C reveals that CTCF underlies evolution of chromosomal domain architecture.* Cell Reports, vol. 10, no. 8, pages 1297–1309, March 2015. (Cited on page 112.)

[Visscher *et al.* 2017] Peter M. Visscher, Naomi R. Wray, Qian Zhang, Pamela Sklar, Mark I. McCarthy, Matthew A. Brown and Jian Yang. *10 years of GWAS discovery: Biology, function, and translation.* The American Journal of Human Genetics, vol. 101, no. 1, pages 5–22, July 2017. (Cited on page 18.)

[Vitor *et al.* 2020] Alexandra C. Vitor, Pablo Huertas, Gaëlle Legube and Sérgio F. de Almeida. *Studying DNA double-strand break repair: An ever-growing toolbox.* Frontiers in Molecular Biosciences, vol. 7, page 24, February 2020. (Cited on pages 14 and 121.)

[Vogelmann *et al.* 2014] Jutta Vogelmann, Antoine Le Gall, Stephanie Dejardin, Frederic Allemand, Adrien Gamot, Gilles Labesse, Olivier Cuvier, Nicolas Nègre, Martin Cohen-Gonsaud, Emmanuel Margeat and Marcelo Nollmann. *Chromatin insulator factors involved in long-range DNA interactions and their role in the folding of the Drosophila genome.* PLoS Genetics, vol. 10, no. 8, page e1004544, august 2014. (Cited on pages 58 and 120.)

[Watson & Crick 1953] James D. Watson and Francis H. Crick. *A structure for deoxyribose nucleic acid.* Nature, vol. 171, pages 737–738, April 1953. (Cited on pages 10 and 11.)

[Weinreb & Raphael 2015] Caleb Weinreb and Benjamin J. Raphael. *Identification of hierarchical chromatin domains.* Bioinformatics, vol. 32, no. 11, pages 1601–1609, August 2015. (Cited on page 96.)

[Wu *et al.* 2011] Michael C. Wu, Seunggeun Lee, Tianxi Cai, Yun Li, Michael Boehnke and Xihong Lin. *Rare-variant association testing for sequencing data with the Sequence Kernel Association Test.* The American Journal of Human Genetics, vol. 89, no. 1, pages 82–93, July 2011. (Cited on page 28.)

[Xavier *et al.* 2020] Roberto Xavier, Kleber Padovani de Souza, Annie Chateau and Ronnie Alves. *Genome assembly using reinforcement learning.* In Luis Kowada and Daniel de Oliveira, editors, Advances in Bioinformatics and Computational Biology, pages 16–28. Springer International Publishing, 2020. (Cited on page 22.)

[Yan *et al.* 2017a] Koon-Kiu Yan, Shaoke Lou and Mark Gerstein. *MrTADFinder: A network modularity based approach to identify topologically associating domains in multiple resolutions.* PLOS Computational Biology, vol. 13, no. 7, pages 1–22, July 2017. (Cited on page 96.)

[Yan *et al.* 2017b] Winston X. Yan, Reza Mirzazadeh, Silvano Garnerone, David Scott, Martin W. Schneider, Tomasz Kallas, Joaquin Custodio, Erik Wernersson, Yinqing Li, Linyi Gao, Yana Federova, Bernd Zetsche, Feng Zhang,

Magda Bienko and Nicola Crosetto. *BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks.* Nature Communications, vol. 8, page 15058, May 2017. (Cited on page 16.)

[Zaborowski & Wilczynski 2016] Rafal Zaborowski and Bartek Wilczynski. *DiffTAD: Detecting Differential contact frequency in Topologically Associating Domains Hi-C experiments between conditions.* bioRxiv, December 2016. (Cited on page 96.)

[Zeisel *et al.* 2015] Amit Zeisel, Ana B. Muñoz-Manchado, Simone Codeluppi, Peter Lönnerberg, Gioele La Manno, Anna Juréus, Sueli Marques, Hermany Munguba, Liqun He, Christer Betsholtz, Charlotte Rolny, Gonçalo Castelo-Branco, Jens Hjerling-Leffler and Sten Linnarsson. *Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq.* Science, vol. 347, no. 6226, pages 1138–1142, March 2015. (Cited on page 18.)

[Zhang *et al.* 2009] Feng Zhang, Claudia M.B. Carvalho and James R. Lupski. *Complex human chromosomal and genomic rearrangements.* Trends in Genetics, vol. 25, no. 7, pages 298–307, June 2009. (Cited on page 14.)

[Zhang *et al.* 2012] Yu Zhang, Rachel P. McCord, Yu-Jui Ho, Bryan R. Lajoie, Dominic G. Hildebrand, Aline C. Simon, Michael S. Becker, Frederick W. Alt and Job Dekker. *Spatial organization of the mouse genome and its role in recurrent chromosomal translocations.* Cell, vol. 148, no. 5, pages 908–921, March 2012. (Cited on page 15.)

[Zhao *et al.* 2006] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh S. Sandhu, Umashankar Singh, Vinod Pant, Vijay Tiwari, Sreenivasulu Kurukuti and Rolf Ohlsson. *Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.* Nature Genetics, vol. 38, no. 11, pages 1341–1347, November 2006. (Cited on page 16.)

[Zheng & Weir 2016] Xiuwen Zheng and Bruce S. Weir. *Eigen analysis of SNP data with an identity by descent interpretation.* Theoretical Population Biology, vol. 107, pages 65 – 76, February 2016. (Cited on page 21.)

[Zhou & Troyanskaya 2015] Jian Zhou and Olga G. Troyanskaya. *Predicting effects of noncoding variants with deep learning-based sequence model.* Nature Methods, vol. 12, no. 10, pages 931–934, August 2015. (Cited on page 23.)

[Zufferey *et al.* 2018] Marie Zufferey, Daniele Tavernari, Elisa Oricchio and Giovanni Ciriello. *Comparison of computational methods for the identification of topologically associating domains.* Genome Biology, vol. 19, no. 1, page 217, December 2018. (Cited on page 96.)

[Zuin *et al.* 2014] Jessica Zuin, Jesse R. Dixon, Michael I. J. A. van der Reijden, Zhen Ye, Petros Kolovos, Rutger W. W. Brouwer, Mariette P. C. van de

Corput, Harmen J. G. van de Werken, Tobias A. Knoch, Wilfred F. J. van IJcken, Frank G. Grosveld, Bing Ren and Kerstin S. Wendt. *Cohesin and CTCF differentially affect chromatin architecture and gene expression in human cells.* Proceedings of the National Academy of Sciences, vol. 111, no. 3, pages 996–1001, January 2014. (Cited on page 14.)

Summary

Following the sequencing of the human genome in 2001, there has been an explosion of novel high-throughput sequencing projects to interrogate the genome and its functions, opening the so-called postgenomic era. Nowadays, experimental labs generate terabytes of heterogeneous data, necessitating the development of novel statistical and bioinformatic methods and models to process such big data, as well as to make sense of the wide variety of experimental results.

For the last 10 years, I have been investigating on a large number of postgenomic topics, ranging from human genetics in asthma to phylogenetics of HIV virus, transcription, chromatin, DNA secondary structures and DNA repair. This thesis presents my research efforts on both the analysis of biological data, and the development of novel statistical and computational models.

In the first chapter, I introduce the different topics, such as DNA, chromatin, postgenomic methods, human genetics and computational biology. In the second chapter, I then describe my different contributions in data analysis, including the discovery of rare variants associated with increased asthma risk, the role of drug-naïve HIV-positive patients in transmitting antiretroviral resistance, the global 3D genome reorganization due to hormone induction and the link between chromatin loop extrusion and DNA repair. I also present different statistical models to identify genomic factors in 1D that shape the genome in 3D, but also novel models for 3D domain identification, differential analysis and predictions. Moreover, I present machine/deep learning approaches for predicting DNA double-stranded breaks and active G-quadruplexes (G4s).

Finally, in the last chapter, I discuss about my future research projects, focusing on new deep learning models for predicting chromatin data across species, biophysical experiments to characterize G4 SNPs, the identification of non-coding SNPs as drivers of genome instability, and artificial intelligence for personalized medicine.

Keywords: Computational Biology; Artificial Intelligence; Deep Learning; Regulatory Genomics; 3D Genome; DNA Repair, G-quadruplex.