Étude du repliement de la chromatine à l'aide d'approches de biologie synthétique

Léa Meneu

Sorbonne Université

ED515 École doctorale Complexité du Vivant

Unité de Régulation Spatiale des Génomes

Institut Pasteur

Thèse de doctorat en Génétique et Génomique

Étude du repliement de la chromatine à l'aide d'approches de biologie synthétique

Léa Meneu

Dirigée par Romain Koszul

Devant un jury composé de :

Pr. Frédéric Devaux	Président du jury
Dr. Armelle Lengronne	Rapportrice
Dr. Gianni Liti	Rapporteur
Dr. Gwenaël Badis-Breard	Examinatrice
Pr. Romain Koszul	Directeur de thèse

Remerciements

Je souhaite tout d'abord remercier les membres du jury d'avoir accepté de lire mon projet de thèse. Je remercie Armelle Lengronne et Gianni Liti d'avoir accepté d'être rapporteurs de ce manuscrit, ainsi que Frédéric Devaux et Gwenaël Badis-Breard pour leur participation à mon jury de thèse.

Merci à mon directeur de thèse, Romain, de m'avoir encadrée durant presque cinq ans. Tu m'as offert une grande liberté de recherche et j'ai eu la chance d'évoluer dans un environnement exceptionnel en tant que chercheuse. Tous les membres du laboratoire, anciens comme nouveaux, ont joué un rôle important durant ma thèse. Merci aux membres du laboratoire : Gaël, Axel, Sébastien et Martial, ainsi qu'aux anciens Cyril, Christophe, Aurèle, Théo et tous ceux qui ont passé quelques mois au labo.

Devon, bon courage pour cette fin de thèse, tu seras incroyable. J'espère qu'on continuera à se voir toutes les deux. Manon et Pauline, merci à vous deux d'avoir été si présentes pour moi cette dernière année. Jacques, ton aide en bio-informatique a été indispensable : je reste la première fan de TinyMapper, et maintenant, j'apprécie un peu plus les bio-informaticiens. Tu as été un super collègue pour boire des coups, discuter ou jouer à la pétanque, et tu m'as aussi beaucoup appris scientifiquement.

Mes co-thésards, vous m'avez manqué cette dernière année. Fabien, tu as été présent à la paillasse pour m'aider à réussir mes expériences et me soutenir quand ça ne marchait pas. Amaury, j'ai passé d'excellentes années à boire des verres au PP et à nous soutenir mutuellement. Tu as été super dans ton rôle de "happiness manager", tu es le meilleur des co-thésards !

Agnès, sans toi je pense que ma thèse aurait été bien différente, j'ai appris tellement de choses. Je ne suis pas encore une experte en levure mais je compte bien le devenir. Tu as également été un soutien émotionnel et m'as beaucoup apporté, merci.

Hélène, merci pour tout : tu m'as encadrée tout au long de ma thèse. Tu as été à l'écoute, bienveillante et pédagogue. Tu es une véritable source d'inspiration, et je ne te remercierai jamais assez pour les heures passées à discuter de sciences et à me former ! Tu vas être une maman merveilleuse et une scientifique incroyable.

Je remercie Angela et Myriam : sachez que vous avez eu un grand impact sur ma thèse et sur la suite de mon parcours.

J’ai pu vivre de si belles années grâce aux personnes rencontrées à Pasteur : Maël, Emma, Marine, Henri, Esteban, Viktoriia, Alicia, Agnès, Charlotte et Alexis. Merci aussi aux membres du théâtre : les mardis soirs étaient vraiment sympas grâce à vous tous. À tous mes amis rencontrés pendant le master : Julie, Vincent, Antonin, Tiphaine et Marvin. Bravo à vous, les (futurs) docteurs ! À mes amis du lycée, Salomé, Marion, Michaël, Clément, Quentin, Kévin et Antoine : on ne se voit pas souvent, mais merci d’être toujours là. Marion, Johan, Laurine, Léandre, je suis heureuse que vous ayez été à Paris en même temps que moi. Ma team Montpellier : Maxime, Camille, Tamara et Léo, bravo à tous les docteurs. On a fait du chemin ensemble, je vous aime.

Léo, merci tout particulièrement de m’avoir soutenue et encouragée pendant cette aventure. Tu as été un pilier tout au long de ma thèse et tu as su me supporter malgré mon caractère difficile. Merci infiniment. J’en profite pour remercier ta famille : Jacques, Pascale, Pablo, Caroline et Gauthier pour leur présence ces dernières années.

Merci énormément à ma famille. À mes grands-parents, qui m’ont toujours soutenue et chez qui j’ai pu rédiger ce manuscrit en Bretagne. Merci Viviane, Gildas et Sophie. Et merci à mes parents, Ronan et Anissa, ainsi qu’à mon frère et ma sœur, Merouan et Nora. Vous êtes la meilleure famille.

Résumé

L'ADN, support de l'information génétique, voit sa composition réguler plusieurs niveaux de l'organisation spatiale des chromosomes, ainsi que des processus métaboliques essentiels tels que l'expression génique, la réplication, la réparation et la ségrégation de l'ADN. La coévolution des activités et du repliement des séquences chromatiniennes sur des millions d'années rend complexe la distinction des liens de causalité entre composition de la séquence, structure chromatinienne et activité chromosomique. De ce fait, les séquences d'ADN exogènes et aléatoires insérées dans des génomes hôtes sont de plus en plus exploitées pour déchiffrer les mécanismes évolutifs et les déterminants séquentiels influençant le repliement et l'activité de la chromatine.

J'ai d'abord exploré comment des chromosomes bactériens exogènes intégrés dans le génome de *Saccharomyces cerevisiae* adoptent des compositions et des activités chromatiniennes distinctes en fonction de leur contenu en GC. Remarquablement, le Hi-C et la microscopie révèlent la formation spontanée de compartiments chromosomiques actifs et inactifs rappelant les archétypes de l'euchromatine et l'hétérochromatine. Ensuite, dans un second projet, nous avons intégré une séquence synthétique aléatoire de 100 kilobases (kb) dans un chromosome de levure. La caractérisation de la chromatine, du transcriptome et de l'organisation 3D de cette région révèle que ces éléments varient selon la source de carbone impactant directement l'organisation 3D. L'ingénierie de cette séquence mène à la mise en place de boucles cohésines dépendantes stables et reproductibles.

Ces travaux soulignent l'importance de l'exploitation de séquences exogènes dans le repliement et la régulation de la chromatine chez les eucaryotes. Ils représentent une avancée dans la compréhension de l'interprétation des séquences étrangères par la machinerie nucléaire de l'hôte, lors des transferts de gènes horizontaux, ainsi que dans les projets de génomique synthétique.

Abstract

DNA, the carrier of genetic information, plays a crucial role in regulating multiple levels of chromosomal spatial organization, as well as essential metabolic processes such as gene expression, replication, repair, and segregation of DNA. The co-evolution of activities and the folding of chromatin sequences over millions of years makes it challenging to distinguish causal links between sequence composition, chromatin structure, and chromosomal activity. Consequently, exogenous and random DNA sequences inserted into host genomes are increasingly being exploited to decipher the evolutionary mechanisms and sequential determinants influencing chromatin folding and activity.

I first explored how exogenous bacterial chromosomes integrated into the genome of *Saccharomyces cerevisiae* adopt distinct chromatin compositions and activities depending on their GC content. Remarkably, Hi-C and microscopy reveal the spontaneous formation of active and inactive chromosomal compartments reminiscent of euchromatin and heterochromatin archetypes. In a second project, we integrated a synthetic random sequence of 100 kb into a yeast chromosome. The characterization of the chromatin, transcriptome, and 3D organization of this region reveals that these elements vary depending on the carbon source, directly impacting the 3D organization. Engineering this sequence led to the establishment of stable and reproducible cohesin-dependent loops.

These studies highlight the importance of using exogenous sequences in chromatin folding and regulation in eukaryotes. They represent an advancement in understanding how host nuclear machinery interprets foreign sequences during horizontal gene transfers, as well as in synthetic genomics projects.

Sommaire

Table des figures

Liste des abréviations

3C : Chromosome Conformation Capture

ChIP : Chromatin Immunoprecipitation

CID : Chromosome Interaction Domain

CRISPR-Cas9 : Clustered Regularly Interspaced Short Palindromic Repeats-associated protein 9

CTD : Carboxy-Terminal Domain

CTCF : CCCTC-Binding Factor

CUT : Cryptic Unstable Transcript

FISH : Fluorescence In Situ Hybridization

GFP : Green Fluorescent Protein

Hi-C : High-throughput Chromosome Conformation Capture

MNase : Micrococcal Nuclease

NGS : Next-Generation Sequencing

NNS : Nrd1-Nab3-Sen1

NMD : Nonsense-Mediated Decay

ORFs : Open Reading Frames

P-E : Promoteur-Enhancer

PCR : Polymerase Chain Reaction

SMC : Structural Maintenance of Chromosomes

SPB : Spindle Pole Body

TADs : Topologically Associating Domains

TSS : Transcription Starting Site

Introduction

L'ADN est le support de l'information génétique et sa composition (i.e. les successions de bases qui la compose) influencent et/ou organisent la liaison des facteurs de chromatinisation, tels que les nucléosomes, et d'autres processus métaboliques multiples liés à l'ADN comme l'expression des gènes, la réplication, la réparation ou la ségrégation de l'ADN. À son tour, l'épigénome contribue à la maintenance et à la stabilité du génome.

Chez les eucaryotes, la composition de la séquence corrèle avec : 1) la composition de la chromatine, qui comprend la formation des nucléosomes et la liaison des protéines structurelles et fonctionnelles à l'ADN 2) l'activité de la chromatine, telle que la transcription et la réplication , et 3) l'organisation fonctionnelle en 3D du génome en boucles et en compartiments. Ces relations entre les activités, le repliement des séquences et de la composition de la chromatine reflètent leurs coévolutions continues sur des millions d'années. De ce fait, il est parfois difficile de désenchevêtrer des liens de causalité entre composition de la séquence, structure chromatinienne, et activité chromosomique.

Ainsi, la manière dont un hôte eucaryote peut emballer, replier et réguler avec succès l'activité de séquences d'ADN exogène ou aléatoire, ainsi que l'importance de la composition des séquences, restent largement inconnues. La réponse à ces questions améliorerait notre compréhension des processus évolutifs et des déterminants de séquence impliqués dans le repliement et l'activité de la chromatine. L'avènement récent de la biologie synthétique permet aujourd'hui d'introduire artificiellement des séquences d'ADN, exogènes ou synthétiques, dans des souches microbiennes ou des lignées cellulaires. Ces approches permettent l'intégration à long terme d'ADN étranger dont la composition des séquences diverge fortement du génome de leur hôte. L'utilisation de ces séquences permet d'explorer les processus biologiques sous-jacents en l'absence de contraintes évolutives.

Au cours de ma thèse, j'ai participé à deux projets portant sur la caractérisation de séquences exogènes ou aléatoires intégrées dans le génome de la levure *S. cerevisiae*.

Projet n°1 : Caractérisation d'ADN exogène dans un noyau eucaryote

Dans cette première étude, nous avons caractérisé le comportement de deux chromosomes bactériens introduits artificiellement dans le génome de la levure en tant que chromosome surnuméraire à l'aide d'approches génomiques. Nous avons établi le profil des nucléosomes, de la polymérase II et des cohésines, ainsi que l'organisation 3D des chromosomes au cours du cycle cellulaire. Premièrement, nous montrons que des chromosomes bactériens avec différents contenus en GC présentent des compositions et des activités chromatiniennes différentes. J'ai ensuite, à l'aide de

l’ingénierie CRISPR-Cas9, modifié la séquence des chromosomes pour obtenir des chromosomes chimériques, composées de régions bactériennes et de levures. L’étude de l'organisation 3D par Hi-C de ces chromosomes chimériques révèle finalement la formation spontanée de compartiments chromosomiques actifs et inactifs, similaires à ceux observés chez les mammifères.

Projet n°2 : Caractérisation d'ADN aléatoire dans un noyau eucaryote

En parallèle du premier projet, nous avons intégré et caractérisé une région synthétique aléatoire de 100 kb, appelée Syn100, dans le chromosome IV de la levure. Nous avons établi le profil chromatinien ainsi que l'organisation 3D de la région aléatoire dans plusieurs sources de carbone (glucose, galactose et lactate). Nous montrons que l'assemblage des nucléosomes ainsi que l'activité transcriptionnelle de cette région dépendent de la source de carbone, ce qui se traduit par des structures différentes de la région Syn100 pendant la métaphase. Nous avons ensuite modifié cette séquence pour mettre en place des boucles stables et reproductibles médiées par la cohésine. Ce système montre que nous sommes désormais en mesure de concevoir une grande structure chromatinienne dans la levure afin d'explorer les relations structure-fonction.

Dans l’introduction de ce manuscrit, je présenterai comment la microscopie et les technologies “3C” ont pu mettre en évidence une organisation hiérarchique et hautement organisée. Puis, je développerai le rôle clé de la transcription, un mécanisme bien connu, dans la structuration de la chromatine. Pour finir je présenterai les avancées en génomique synthétique ouvrant des perspectives pour étudier des processus fondamentaux.

1. Organisation des chromosomes eucaryotes

Des décennies de travail permettent d'avoir aujourd'hui une compréhension relativement avancée des principes organisationnels des génomes des espèces eucaryotes, et ont permis d'appréhender les rôles et influences de la structure tridimensionnelle des chromosomes sur des processus biologiques liés à l'ADN. Des avancées majeures récentes dans les approches d'imagerie, ainsi que le développement d'approches génomiques telles que la capture de conformation chromosomique (3C, Hi-C) ont intensifié ces recherches en permettant la description et l'étude, de différentes structures de l'organisation du génome dans de nombreuses espèces.

La cytologie a ainsi mis en évidence dès la fin du 19ème siècle de nombreux éléments structurants des chromosomes et la nature dynamique de leur organisation (condensation mitotique et méiotique, bouquet méiotique, chiasma méiotiques, …). Depuis, les avancées technologiques ont renforcé ces connaissances, montrant que les chromosomes des procaryotes et des eucaryotes présentent plusieurs niveaux d'organisation hiérarchiquement imbriqués. Ainsi, dans toutes les espèces étudiées jusqu'à présent, les chromosomes présentent des structures diverses, comme des boucles de chromatine, des compartiments regroupant des régions spécifiques, ou encore des domaines variant en taille de quelques dizaines à plusieurs centaines de kilobases (kb). Des changements dynamiques de cette organisation influencent ou régulent de nombreux processus chromosomiques, tels que l'expression des gènes, ou la réplication et la réparation de l'ADN. La compréhension de la structure tridimensionnelle (3D) des chromosomes est ainsi devenu un aspect important de l'étude de très nombreux processus biologiques, comme par exemple la différentiation cellulaire, la division mitotique, ou encore la cancérogenèse.

La facilité avec laquelle l'approche génomique Hi-C permet de décrire une organisation chromosomique a également permis de généraliser quelques observations faites sur un sous-ensemble d'espèces "modèles". Ainsi, on peut définir, pour le moment, 2 grands types d'organisation. Le premier type (type I) comprend les trois caractéristiques de type Rabl retrouvés par exemple chez la *S. cerevisiae* : regroupement des centromères, regroupement des télomères et axe entre les télomères et le centromère. Le second (type II) ne comprend que les territoires chromosomiques. Les humains présentent une architecture génomique de type II, avec des territoires chromosomiques forts.

Dans ce premier chapitre, je soulignerai comment les approches multidisciplinaires et en particulier la méthode 3C permettent de décrire et d'étudier l'organisation des génomes. Je présenterai principalement l'organisation de l'espèce que j'ai manipulé expérimentalement au cours de ma thèse, *S. cerevisiae*. Cependant, je mentionnerai et ferai régulièrement des parallèles entre cette organisation et celles trouvées au sein d'une autre espèce très étudiée, en l'occurrence l'homme.

1.1. Approches expérimentales de l'étude de l'organisation 3D des génomes

1.1.1 Approches d'imageries

Historique

En observant des cellules de salamandre au microscope optique après coloration à l'aniline, Walther Flemming a décrit et dessiné la partition égale des chromosomes au cours de la division cellulaire, tandis que Carl Rabl a décrit l'organisation en V des chromosomes durant l'anaphase, avec les centromères et les télomères regroupés à des pôles opposés du noyau (Flemming, 1882 ; Rabl, 1885). En 1928, Emil Heitz améliore une méthode de coloration cytologique et démontre que certaines régions des chromosomes mitotiques sont plus densément colorées que d'autres, définissant ainsi l'hétérochromatine et l'euchromatine (Passarge, 1979). À la fin des années 1950, Joe Hin Tjio et Albert Levan déterminent le nombre correct de chromosomes par cellules humaines en traitant les cellules à la colchicine, ce qui permet l'accumulation des cellules en mitose (Tjio & Levan, 1956). Marthe Gautier et ses collègues mirent ensuite en évidence la première aberration chromosomique en observant qu'un chromosome supplémentaire est présent dans des cultures in vitro de cellules de patient atteint de trisomie 21 (Lejeune et al., 1959). Ces travaux cytologiques pionniers mettent en évidence la diversité des structures des chromosomes au cours du temps et l'importance pour la cellule de maintenir une stabilité chromosomique.

L'hybridation in situ

Au début des années 1980, les sondes radioactives ont été remplacées par des sondes fluorescentes, marquant ainsi le développement de la méthode d'hybridation *in situ* en fluorescence (FISH). Cette technique consiste à marquer de l'ADN par l'hybridation d'une sonde nucléotidique complémentaire contenant des marqueurs fluorescents. L'utilisation de sondes marquées avec des fluorophores différents a permis d'identifier chaque chromosome individuellement au sein d'une cellule, on parle de "chromosome painting". En l'appliquant sur des cellules de poulet, la méthode FISH a révélé que les chromosomes de ce vertébré occupent des territoires distincts dans le compartiment nucléaire (Cremer & Cremer, 2001). De la même manière, cette méthode a été utilisée pour discriminer les 24 chromosomes humains et confirment les résultats observés précédemment chez la poule (**Figure 1. A,B**), (Bolzer et al., 2005). Ces observations ont indiqué que les régions chromosomiques riches en GC et les petits chromosomes se trouvent majoritairement à l'intérieur du noyau, tandis que les régions chromosomiques riches en AT, et les chromosomes longs se situent principalement à la périphérie du noyau (Bolzer et al., 2005). Le marquage de l'ADN avec des

analogues de nucléotides, comme le BrdU ou le Cy3-dUTP, permet de discriminer les régions de l'ADN qui se répliquent tôt (bandes R) et tard (bandes G/C) dans le cycle cellulaire. Les bandes R, “GC riche”, se répliquent précocement lors de la phase S, et se trouvent à l'intérieur du noyau, tandis que les bandes G/C, AT-riche, se répliquent tardivement et sont localisées à la périphérie nucléaire.

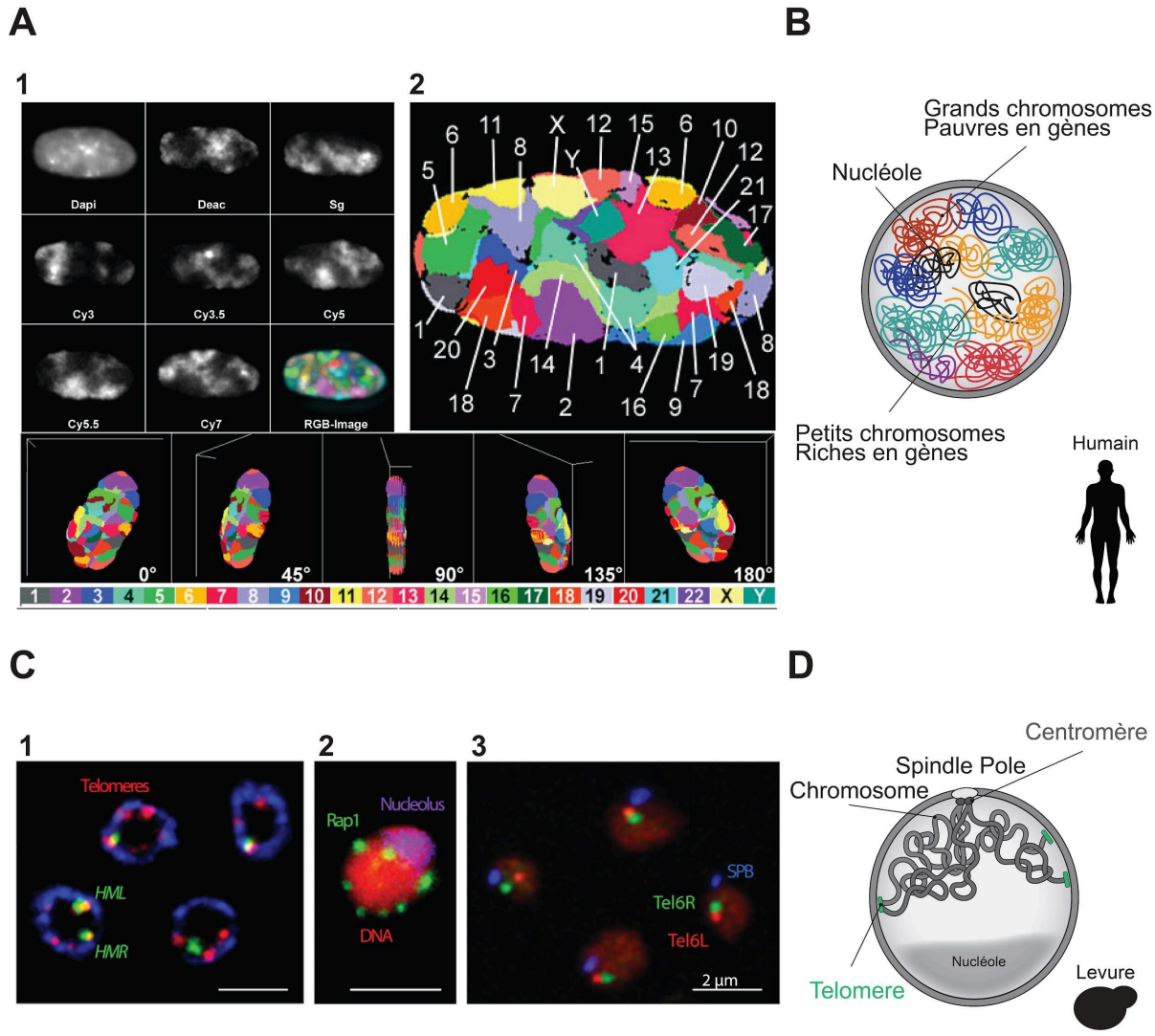


Figure 1 : Preuve directe des territoires chromosomiques par microscopie chez l’humain et chez la levure.

A) 1) Image d'une section nucléaire obtenue par microscopie utilisant huit canaux : un canal pour le DAPI (contre-coloration de l'ADN) et sept canaux pour les fluorochromes. Chaque canal représente un sous-ensemble de territoires chromosomiques avec le fluorochrome respectif. 2) Représentation en fausses couleurs de tous les territoires chromosomiques visibles dans cette section après classification (Bolzer et al., 2005). **B)** Schéma de l’organisation des chromosomes en territoires chromosomiques, chez l’humain. **C) 1)** Images fluorescentes confocales montrant le positionnement périphérique des télomères de levure (en rouge), des loci HML et HMR (en vert) et des pores nucléaires (en bleu) (Gotta et al., 1996). 2) Regroupement des télomères dans un noyau diploïde visualisé par immunofluorescence avec un anticorps anti-Rap1 (en vert), tandis que le nucléole est coloré en violet. La coloration des acides nucléiques par le bromure d'éthidium apparaît en rouge. 3) Visualisation du SPB (en bleu) et des télomères droit et gauche du chromosome VI (en rouge et vert). **D)** Schéma de l’organisation de type Rabl de la levure. Il est démontré qu'elle persiste dans le noyau de la levure en interphase.

En complément du FISH, les méthodes d'immunofluorescence (qui utilisent des anticorps au lieu de sondes nucléotidiques) ont également dépeint les compartiments chez la levure. Elle présente une organisation de type Rabl (en référence aux observations de ce dernier sur les figures anaphasiques des chromosomes) consistant en des centromères qui se regroupent au niveau du corps du pôle du fuseau (SPB pour Spindle Pole Body, centre d'organisation des microtubules de *S. cerevisiae*) et des télomères s'attachant à l'enveloppe nucléaire. Enfin, un nucléole où l'ADN ribosomique (rDNA) est séquestré à l'opposé du SPB (**Figure 1. C,D**) (Gotta et al., 1996; Taddei & Gasser, 2012).

La méthode de FISH a également dépeint l'association de certaines régions d'ADN avec la membrane nucléaire. Il a été possible d'identifier un sous-ensemble de protéines du pore nucléaire jouant un rôle essentiel dans l'induction de la transcription de gènes spécifiques au cours du développement de la drosophile (Capelson et al., 2010). Dans le laboratoire de Susan Gasser, Taddei et al. ont développé un système permettant de visualiser par microscopie la position d'un locus spécifique dans le noyau. Ce système repose sur deux éléments, 1) des protéines d'intérêt fusionnées à la protéine LexA, et 2) une construction contenant des sites de liaison de LexA et des sites LacO, où une protéine chimérique, LacI-GFP, est associée pour visualiser la position du locus dans le noyau. Cette approche a révélé que plusieurs protéines fusionnées à LexA peuvent déplacer le locus vers la périphérie du noyau (Taddei & Gasser, 2004). Ces deux travaux suggèrent une relation fonctionnelle entre l'ancrage périnucléaire et la répression transcriptionnelle et montrent une fonction directe des protéines du pore nucléaire dans la régulation de l'expression génique.

Observation dynamique des loci

Au-delà de l'organisation globale, il est intéressant de visualiser spécifiquement une séquence donnée, et d'obtenir des données quantitatives de l'organisation dans des cellules uniques. En compilant informatiquement la position de milliers de cellules, il est ainsi possible de générer des cartes probabilistes à haute résolution de la localisation des gènes, dans un organisme tel que la levure où il est possible d'orienter les noyaux (Berger et al., 2008). En utilisant cette technique, Berger et al. ont pu caractériser la ségrégation spatiale des gènes de galactose, la localisation nucléolaire de plusieurs gènes impliqués dans la biogenèse des ribosomes chez *S. cerevisiae* (Berger et al., 2008). De nombreuses équipes ont développé ou utilisent des méthodes d'imagerie permettant de visualiser l'organisation de la chromatine à une résolution à l'échelle du kb (Boettiger & Murphy, 2020). C'est le cas de Bintu et al. qui marquent et suivent des régions de 30 kb, et peuvent ainsi mesurer simultanément les positions de paires d'enhancers et de promoteurs ainsi que leurs activités transcriptionnelles (Bintu et al., 2018). Il est également possible d'éditer le génome pour marquer des sites précis de l'ADN, tels que les sites CTCF, afin de mesurer la fréquence de contact entre ces

éléments (Gabriele et al., 2022). L'équipe de Thomas Gregor à l'Institut Pasteur a par exemple systématiquement étudié la séparation de deux loci d'ADN (la partie enhancer du locus eve et le promoteur) pour étudier comment leur interaction est modifiée par la distance qui les sépare sur le chromosome en révélant des mécanismes régulant la transcription en fonction de leur proximité génomique (Brückner et al., 2023). L'apport de ces travaux sera discuté dans la partie 2.3.2, "régulation longue distance"). Pour finir, l'utilisation d'un analogue nucléotidique permet de visualiser la structure fine des chromatides sœurs dans les cellules humaines au cours du cycle cellulaire (Batty et al., 2023).

La puissance du FISH et d'autres méthodes de microscopie réside dans leur capacité à effectuer des analyses unicellulaires, de la position précise d'un locus à l'organisation globale des chromosomes. Cependant, à l'échelle du génome et de la population cellulaire, elles sont limitées en termes de débit et de multiplexage, c'est-à-dire, la capacité à suivre plusieurs loci simultanément.

1.1.2 Capture de conformation de chromosome

Les méthodes "3C"

En 2002, le laboratoire de Nancy Kleckner décrit une méthode de capture de la conformation des chromosomes (3C) inspirée de l'immunoprécipitation de chromatine, mais sans sélection de protéines (Dekker et al., 2002). Cette technique consiste à fixer les contacts chromosomiques au sein du noyau grâce à un agent chimique pontant, tel que le formaldéhyde (les étapes sont présentées dans la **Figure 2. A**). L'ADN est ensuite digéré par des enzymes de restriction, puis les fragments de restriction sont ligués grâce à une ADN ligase, qui raboute préférentiellement les fragments capturés au sein des mêmes complexes ADN/protéine. L'ADN est ensuite purifié. On obtient ainsi des fragments d'ADN hybrides par rapport au génome de référence, correspondant à des séquences physiquement proches dans le noyau, même si elles sont éloignées sur la séquence génomique. L'ensemble des fragments hybrides générés constitue une librairie des produits de ligation, qui peut être analysée par PCR. Dans ce cas on peut détecter si une interaction est possible entre deux fragments d'ADN d'intérêt grâce à des amorces choisies, on parle de technique 1 *versus* 1 (**Figure 2. A**). Cette approche est à l'origine de la première analyse multiéchelle de la conformation 3D du chromosome III de levure (Dekker et al., 2002), sur la base de 13 oligonucléotides répartis tous les 20 kb. L'arrivée des techniques de séquençage à haut débit a, quelques années plus tard, permis d'étendre cette approche "ciblée" (où les oligonucléotides étaient choisis manuellement à des positions d'intérêt) à une approche globale (i.e. Hi-C), permettant de séquencer des contacts entre des milliers de positions au sein du génome et d'obtenir ainsi des résultats sans a priori sur les régions d'intérêt.

La technique Hi-C, dont l'usage a grandi exponentiellement depuis 15 ans, permet en effet de détecter toutes les interactions au sein du génome (tous *versus*. tous). En schématisant, cela est rendu possible par le séquençage à haut débit des extrémités de millions de molécules "chimériques" contenues dans la libraire 3C. Cependant, des étapes ont également été introduites pour optimiser le protocole. Après la digestion par les enzymes de restriction, des nucléotides marqués à la biotine sont incorporés dans l'ADN au site de ligation. Ensuite, une capture (pull-down) par la streptavidine des molécules ainsi biotinylées permet d'enrichir les éléments de re-ligation (**Figure 2. B**) (Lieberman-Aiden et al., 2009). Ces séquences d'ADN purifiées sont ensuite préparées pour le séquençage par l'ajout d'adaptateurs. La librairie est ensuite séquencée en "paired-end" avec des systèmes NGS comme le NextSeq 2000. Le séquençage "paired-end" consiste à séquencer les deux extrémités d'un fragment.

Le traitement des données issues du séquençage NGS "paired-end" nécessite un traitement informatique spécifique. Les fragments séquencés, appelés « reads », sont alignés sur un génome de référence permettant d'identifier les paires de loci en contact. Les contacts identifiés peuvent être visualisés sous forme de carte de fréquences relatives, reflétant la fréquence de contacts de paires de loci. La carte de contact est une matrice à double entrée représentant la fréquence à laquelle deux fragments de restriction sont identifiés comme étant capturés ensemble. Si deux fragments sont souvent associés dans le noyau, cela se traduira par une fréquence de contact plus élevée sur la carte Hi-C (**Figure 2. C,D**). La chromatine étant un polymère, cette molécule présente des propriétés qui reflètent cette nature. Notamment, la fréquence de contact entre deux loci dépend de leur distance génomique : deux loci proches interagissent plus fréquemment que deux loci éloignés. Cette caractéristique se reflète au niveau de la diagonale des cartes Hi-C qui représente une fréquence de contact très élevée pour les loci proches. Cette fréquence de contact diminue à mesure qu'on s'éloigne de la diagonale, donc pour les loci plus éloignés sur la séquence génomique (**Figure 2. CD**). On appelle la courbe reflétant cette fréquence de contact en fonction de la distance génomique "p(s)", une mesure quantitative très utilisée pour mesurer l'effet de mutation sur des acteurs du repliement chromosomique.

La résolution des cartes de contacts dépend 1) de la taille des fragments d'ADN générés, les petits fragments augmentent la résolution, 2) de la profondeur de séquençage, plus on séquence de séquences hybrides plus la résolution est élevée et enfin 3) de la taille des génomes, plus un génome est grand plus il faut séquencer un grand nombre de fragments pour avoir une bonne résolution, et enfin 4) de la distance à laquelle on se positionne par rapport à la diagonale, des contacts entre régions plus proches étant plus couvertes et donc plus statistiquement "résolutives" que des contacts entre régions éloignées (Muller et al., 2018). Par approximation et surtout par facilité, on considère souvent que la résolution d'une matrice Hi-C correspond à la taille des vecteurs qui la composent. De nos

jours, la “résolution” d’une carte de contact Hi-C est de l’ordre du kb pour les cellules de mammifères et les cellules de levure (Dauban et al., 2020; Rao et al., 2014).

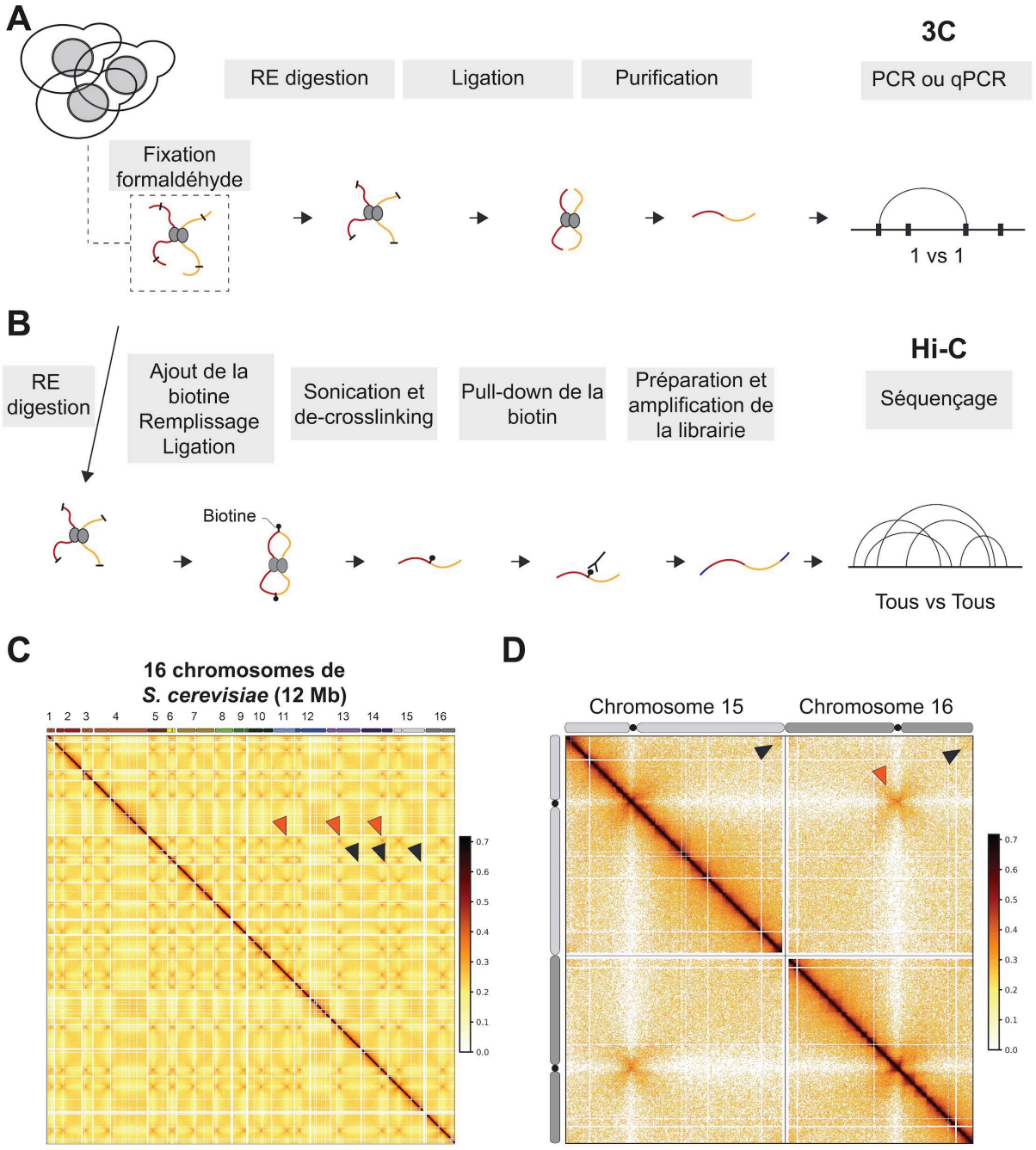


Figure 2 : Les méthodes 3C.

A) Représentation schématique de la méthode 3C. RE = Enzyme de restriction **B)** Représentation schématique de la méthode Hi-C. RE = Enzyme de restriction **C)** Carte de probabilité de contact Hi-C pour tous les chromosomes de levure. Plus la couleur est intense (ici, noire) plus la probabilité d’interaction entre loci est élevée. Pointes de flèches noires : contacts inter-télomères. Flèches jaunes : contacts inter-centromères. **D)** Carte de probabilité de contact Hi-C de deux chromosomes de levure.

Les variantes

De nombreuses variantes de la méthode 3C ont été développées (Davies et al., 2017; Wit & Laat, 2012). Le 3C circulaire (4C) mesure les fréquences d'interaction d'un locus avec de nombreux autres loci (Z. Zhao et al., 2006). Dans le 5C (Chromosome Conformation Capture Carbon Copy), des sondes d'oligos spécifiques sont conçues pour cibler les régions d'intérêt dans le génome. Les complexes ADN-sonde hybridés sont amplifiés par PCR (Polymerase Chain Reaction) et permettent d'étudier tous les contacts d'un locus particulier (Dostie et al., 2006). La Capture-C combine la préparation de librairies 3C ou Hi-C avec une capture ciblée à l'aide d'oligonucléotides. Les librairies 3C/Hi-C sont enrichies en fragments d'intérêt à l'aide de sondes de capture biotinylées conçues pour chaque point de vue, après quoi ces fragments sont amplifiés et séquencés (Jäger et al., 2015). Les contacts Hi-C ont également été déterminés à partir de cellules uniques en isolant et en séquençant des noyaux intacts uniques lors de la préparation de la librairie Hi-C, révélant une variabilité considérable entre les cellules individuelles, mais les données ne sont pas informatives (Nagano et al., 2013). Enfin, il est difficile d'étudier les relations entre les chromatides sœurs à l'aide de techniques génomiques car les méthodes basées sur le séquençage ne permettent pas de distinguer les chromatides sœurs identiques. Pour répondre à cette problématique, l'utilisation d'analogues de nucléotides comme le BrdU ou le Cy3-dUTP permet de discriminer les séquences nouvellement répliquées et, par extension, discriminer les intéractions entre les chromatides A et B (SisterC, Oomen et al., 2020 ; scsHi‑C, Mitter et al., 2020).

L'ensemble de ces technologies utilise des enzymes de restriction, dont le choix, en fonction des organismes, permet d'améliorer la résolution des matrices. Le Micro-C, un autre dérivé du 3C, est l'approche la plus puissante développée par le laboratoire d'Oliver Rando. L'approche utilise la MNase, et non pas une ou plusieurs enzymes de restriction, pour générer une fragmentation uniforme (i.e. de la taille d'un nucléosome, environ 150 pdb), ce qui augmente la résolution locale (Hsieh et al., 2015). La première carte Micro-C révélant principalement des contacts très proches entre nucléosomes, et la position des régions dépourvues de nucléosomes, l'ajout d'un deuxième agent pontant, comme l'EGS (Ethylene Glycol-bis(succinimidyl Succinate)), a été nécessaire pour détecter des signaux à plus longue distance (T.-H. S. Hsieh et al., 2016). Bien que cette technique constitue une amélioration prometteuse pour l'étude de la topologie locale de la chromatine, elle reste relativement complexe à mettre en œuvre et, est moins efficace que le Hi-C pour capturer les contacts à longue distance et inter-chromosomiques. De plus, elle ne fonctionne que sur les génomes contenant des nucléosomes, donc principalement eucaryotes (et n'a jamais été testée avec succès sur les archées portant des histones).

Les limites

La principale limite des méthodes 3C est l'utilisation d'agents fixateurs, qui peuvent générer des artefacts ou amplifier certains signaux (Scolari et al., 2018). De plus, ces méthodes sont basées sur la fixation d'une population de cellules, alors que des approches cellules uniques ont été développées mais restent extrêmement coûteuses, d'autant plus que leur intérêt reste relativement limité si le nombre de cellules n'est pas important (Nagano et al., 2013). Pour limiter au maximum les variations intercellulaires, il est donc intéressant de débuter avec des populations les plus homogènes possibles. Par exemple, pour *S. cerevisiae*, la plupart des études récentes sont aujourd'hui réalisées sur des cellules parfaitement synchronisées à différents stades du cycle cellulaire (Lazar-Stefanita et al., 2017; Schalbetter et al., 2017). L'utilisation de degrons contrôlés permettant par ailleurs de dégrader spécifiquement des protéines d'intérêt et d'en étudier rapidement les effets (Dauban et al., 2020; Lazar-Stefanita et al., 2017; Schalbetter et al., 2017). Une autre limitation consiste dans le fait que ces méthodes se basent sur le séquençage, ce qui ne permet pas, comme dans toutes les études génomiques, de discriminer les régions répétées les unes des autres lors de l'alignement des séquences ni de distinguer les chromosomes homologues ou les chromatides soeurs. Certaines méthodes comme le SisterC et le scsHi‑C, évoquées précédemment, essaient de répondre à ces limitations (Mitter et al., 2020; Oomen et al., 2020).

Néanmoins, même avec une méthodologie imparfaite, les approches "Hi-C" ont connu un essor sans précédent et ont apporté une dimension nouvelle dans notre compréhension du repliement fonctionnel des chromosomes dans toutes les branches du vivant puisque ne nécessitant pas forcément de manipulation génétique ou d'intégrations de transgènes pour donner une photographie de l'organisation 3D de génomes d'espèces très variées.

1.1.3 Complémentarité des méthodes

Les techniques 3C ont révolutionné le domaine de l'étude de l'organisation nucléaire, remplaçant en partie l'ADN FISH comme méthode de choix pour étudier l'architecture tridimensionnelle des chromosomes. Cependant, les limites des méthodes 3C peuvent être complémentées par les approches de microscopie qui, malgré une résolution souvent moins fine, permettent d'intégrer aux études la variabilité des cellules uniques et confirmer les résultats obtenus par 3C (Giorgetti & Heard, 2016; Jerkovic´ & Cavalli, 2021). Il est donc courant, et souvent conseillé, en fonction des questions, de comparer les approches de microscopie et de 3C, en tenant compte des limites et des biais de chaque technique. Par exemple, Bintu, Mateo et leurs collègues ont démontré, avec la technique OligoSTORM, des domaines locaux de contact, bien séparés les uns des autres,

rappelant les domaines topologiquement associés (TADs ; décrits dans la partie 1.2.2) identifiés par Hi-C. C'est une première démonstration alternative au Hi-C que les TADs et boucles sont identifiables via des approches de microscopie (Bintu et al., 2018). Dans ces expériences, l'accord qualitatif et quantitatif entre les résultats indirects de Hi-C basés sur le séquençage et ceux de la microscopie fournit une validation croisée des deux méthodes. Bien que la méthode Hi-C soit supérieure en termes de couverture, la microscopie ouvre des perspectives fondamentales sur les structures physiques 3D d'ordre supérieur et les variations intercellulaires.

Les outils informatiques sont évidemment indispensables pour générer et analyser les matrices de type Hi-C. De nombreux outils ont ainsi été développés pour manipuler les données séquencées (Abdennur & Mirny, 2020; Matthey-Doret et al., 2022; Serizay et al., 2024). Des dizaines d'outils plus ou moins faciles d'utilisation permettent de détecter et d'identifier les motifs directement observés sur les matrices, comme par exemple Chromosight développé dans mon laboratoire de thèse (Matthey-Doret et al., 2020). Enfin, certains outils tentent d'améliorer l'alignement des séquences, notamment pour les éléments répétés (Gradit et al., in prep).

Les simulations de polymères d'ADN permettent de soutenir, corroborer ou explorer plus en avant les caractéristiques observées par imagerie ou Hi-C. Par exemple, c'est le cas des TADs. Ces derniers sont visualisés par Hi-C, mais l'absence de techniques expérimentales ne permet pas de visualiser de manière fiable et complète leur dynamique. Au cours de la dernière décennie, par exemple, des approches informatiques ont été développées permettant de modéliser un mécanisme de formation de boucle via l'extrusion de la chromatine à travers un complexe protéique, et ses effets attendus sur les données Hi-C mesurant la conformation des chromosomes. Les conformations de polymères résultantes sont ensuite utilisées pour générer des cartes de contact in silico qui, moyennées sur des centaines ou des milliers de simulations, peuvent à leur tour être comparées aux données Hi-C expérimentales (Corsi et al., 2023; Fudenberg et al., 2016). Dans un autre exemple, Gibcus et al. combinent Hi-C et microscopie pour visualiser, minute par minute, la compaction des chromosomes au cours de l'interphase à la métaphase dans des cellules DT40 (poulet) synchrones (Gibcus et al., 2018). Ils ont mis en évidence une structure hélicoïdale avec les données de Hi-C, ensuite confirmée par microscopie. Les simulations de polymères ont ensuite mis en évidence le rôle probable et plausible de l'extrusion de boucle dans l'arrangement hélicoïdale des chromosomes conduisant aux matrices observées (Gibcus et al., 2018).

Grâce aux avancées technologiques combinées à la multidisciplinarité des approches, il n'est pas surprenant que la dernière décennie ait fourni des révélations majeures sur l'organisation et la fonction du génome 3D.

1.2. Une organisation chromosomique à plusieurs niveaux

1.2.1 Le nucléosome, l'unité de base de la fibre de chromatine

Composition du nucléosome

L'unité de base de la chromatine est le nucléosome (Oudet et al., 1975). Les histones sont de petites protéines basiques d'environ 15 kDa, très conservées chez les eucaryotes. La plus petite unité de chromatine est la particule d'ADN enroulée autour d'un octamère d'histone (147 pdb) reliée par de courts "linkers" d'ADN (Cutter & Hayes, 2015; Jiang & Pugh, 2009). Les histones sont des protéines basiques chargées positivement qui s'associent avec l'ADN chargé négativement. Il existe 4 histones canoniques : H2A, H2B, H3 et H4. Lors de la formation des nucléosomes, deux hétérodimères H2A/H2B s'associent à deux hétérodimères H3/H4 pour former un octamère (**Figure 3. A**). Les histones sont composées de deux parties : une partie centrale, appelée "core", par laquelle elles s'associent à l'ADN et entre elles, et une extrémité appelée "queue".

L'histone H1

Une cinquième histone, l'histone H1, ou linker, peut s'associer avec le nucléosome pour former ce que l'on appelle le chromatosome, contribuant ainsi à la stabilisation du nucléosome (**Figure 3. A,B**) (Fyodorov et al., 2018). Chez les mammifères on compte onze sous-types de l'histone H1, exprimés dans différents tissus ou impliqués dans divers processus tels que la réparation de l'ADN, la compaction des chromosomes en mitose ou l'expression génique (Fyodorov et al., 2018). La liaison H1 stabilise le nucléosome et est généralement associée à une fibre d'ADN compactée et réprimée transcriptionnellement (**Figure 3. A,B**), (Fyodorov et al., 2018). Beaucoup moins conservée que ses homologues canoniques, la levure *S. cerevisiae* ne compte qu'une seule histone linker, Hho1, et sa délétion n'entraîne pas d'effet notable sur la croissance de la souche (Downs et al., 2003). Hho1 peut agir comme une barrière à certaines marques épigénétiques, inhibe la recombinaison homologue (HR), contribue au maintien des télomères et réprime l'activité des transposons (Downs et al., 2003; Veron et al., 2006).

Modifications post-traductionnelles

Les histones sont décorées par une multitude de modifications post-traductionnelles et sont souvent appelées marques épigénétiques car elles régulent la structure de la chromatine et les processus associés à l'ADN (Hauer & Gasser, 2017; Millán-Zambrano et al., 2022). Ces modifications sont présentes à la fois dans les queues terminales des histones et dans leurs domaines globulaires

centraux. Elles jouent un rôle structurel en influençant les interactions entre histones, les interactions inter-nucléosomes et facilitent le recrutement de protéines qui se lient aux histones (Hauer & Gasser, 2017). En conséquence, les modifications des queues d'histones peuvent non seulement moduler la stabilité des histones sur l'ADN, mais aussi modifier l'organisation de la chromatine en ordres supérieurs et de façon dynamique (Millán-Zambrano et al., 2022). Le Hi-CO, une méthode similaire au Micro-C (T.-H. S. Hsieh et al., 2016), permet de révéler la position et l'orientation des nucléosomes en 3D dans les chromosomes (Ohno et al., 2021). Après la digestion à la MNase, les fragments sont re-ligués, et l'ajout d'un adaptateur d'ADN permet de distinguer l'orientation des nucléosomes. Les données de Hi-CO montrent plusieurs niveaux de structurations des nucléosomes avec une structure minimale : des structures composées de quatre nucléosomes appelées α et β (**Figure 3. C**), (Ohno et al., 2021). Le choix de ces structures est principalement contrôlé par les modifications d'histones enrichies au niveau des promoteurs et corps des gènes et par la liaison à l'ARN polymérase. Cependant, ces motifs α/β ont été identifiés par modélisation in silico, des expériences supplémentaires sont nécessaires pour vérifier leur présence dans les cellules. Enfin, en utilisant des mutants, ils ont observé que les distances entre nucléosomes étaient en moyenne plus longues que les loci portant des modifications d'histones enrichies dans les corps de gènes, telles que H3K36me2 et H3K36me3, renforçant le rôle des marques épigénétiques dans l'organisation de la chromatine (Ohno et al., 2021).

Le nucléosome est le premier degré de compaction de l'ADN. Les nucléosomes s'assemblent ensuite pour former une fibre de chromatine et coexistent, in vivo, avec des structures d'ordres supérieurs telles que les boucles et les compartiments nucléaires.

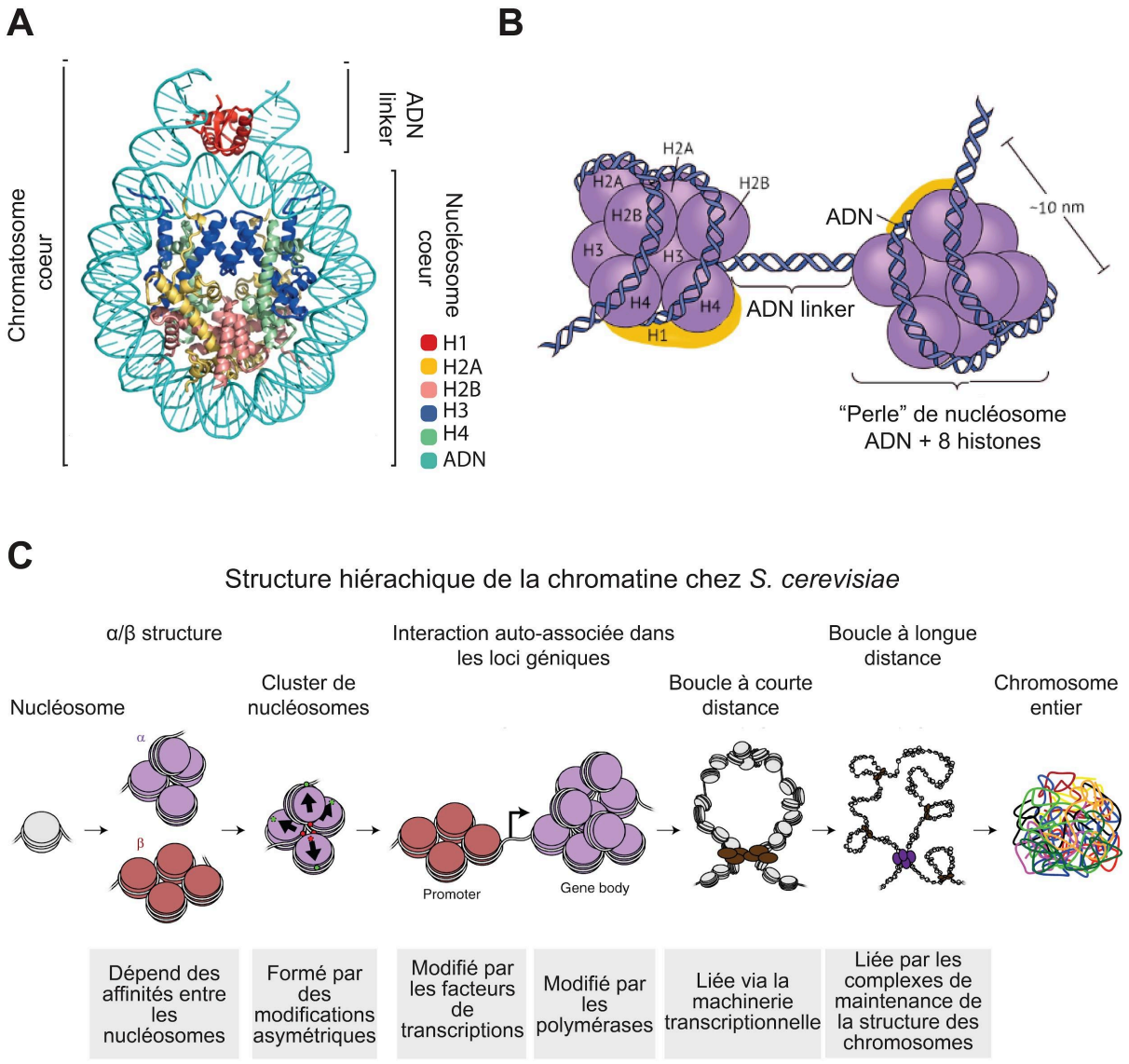


Figure 3 : Structure du nucléosome et son rôle dans l'organisation du génome.

A) Structure cristalline du noyau du chromatosome contenant le domaine globulaire de la protéine H1 de poule (en rouge) et les histones centrales (H2A, H2B, H3 et H4 ; codées en différentes couleurs). **B)** Représentation schématique de deux nucléosomes adjacents séparés par un linker d'ADN. **C)** Modèle d'architecture hiérarchique de la chromatine de *S. cerevisiae*, du nucléosome aux chromosomes entiers (Ohno et al., 2021).

1.2.2 Les structures secondaires et les boucles

Les domaines topologiquement associés (TADs)

Certaines des premières études Hi-C (et autres variantes) ont suggéré une organisation des chromosomes eucaryotes en domaines discrets, inférieurs à la Mb, correspondants à des régions où les contacts sont plus prononcés en intra qu'avec des régions voisines. Ces structures apparaissent comme des triangles dans les matrices Hi-C (**Figure 4. A,B**). Chez les mammifères, et la drosophile, ils ont été appelés domaines topologiquement associés (TADs) (J. R. Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). Chez la bactérie, on a appelé ces structures triangulaires des chromosomes interaction domains (CID) (Le et al., 2013). Par ailleurs, toutes ces structures, bien que ressemblantes sur les matrices de Hi-C, correspondent en fait à des activités ou des propriétés de la chromatine bien différentes d'un phylum à un autre. Je parlerai ici principalement des mammifères puis de la levure.

Chez les mammifères, les "TADs" sont maintenant relativement bien caractérisés. Leur position corrèle et reflète étroitement la démarcation fonctionnelle des régions chromatiniennes en fonction de l'activité transcriptionnelle, des modifications des histones et de la réplication. Ces corrélations ont immédiatement conduit à suggérer dès ces premiers travaux que la structure et la fonction du génome peuvent être mécaniquement couplées (J. R. Dixon et al., 2012; Nora et al., 2012; Sexton et al., 2012). De nombreuses équipes ont disséqué la nature, le mécanisme de formation, et la régulation des TADs pour comprendre comment la perturbation de ces structures par des réarrangements génomiques peut entraîner une mauvaise expression des gènes, et conduire notamment à des maladies ou des pathologies du développement (Lupiáñez et al., 2016). Par exemple, une délétion de 58 kb englobant la limite entre deux TADs sur le chromosome X humain génère des contacts ectopiques entre des séquences normalement isolées (Nora et al., 2012). Cette perte d'organisation entraîne une mauvaise régulation transcriptionnelle à longue distance soulignant le rôle que la partition des chromosomes en TAD peut jouer dans le contrôle transcriptionnel à longue distance (Nora et al., 2012). Une délétion au sein d'un TADs peut entraîner une expression ectopique de Pax3 ce qui se traduit par un raccourcissement des premier et deuxième doigts chez la souris (Lupiáñez et al., 2015). Ces observations ont ainsi suggéré très tôt que les TADs reflètent des structures fonctionnelles, et surtout que les démarcations au sein du génome étaient associées à des séquences spécifiques. Plusieurs équipes se sont alors attelées à caractériser ces séquences, et comment elles conduisent à la formation de ces structures triangulaires.

Caractérisation des boucles chez les mammifères

La résolution et la profondeur de séquençage croissante des données Hi-C (Rao et al., 2014) a mis en évidence qu'un nombre significatif de domaines était délimité par des foyers de contacts entre les extrémités des TADs apparaissant sous forme de points discret à la pointe des TADs (**Figure 4. A**), (Rao et al., 2014). Communément appelés « boucles » car ils correspondent à des contacts discrets entre deux points distants d'un polymère, ces enrichissements sont maintenant interprétés comme reflétant une fraction stable des boucles dynamiques couvrant la région couverte par un TAD (Finn et al., 2019; Luppino et al., 2020; Su et al., 2020)**, (Figure 4. B)**. Les données Hi-C, très séquencées, de Rao et al. ont identifié environ 10 000 boucles à partir d'une carte Hi-C à « résolution de 1 kb » des cellules lymphoblastoïdes humaines GM12878 (Rao et al., 2014). Parmi ces interactions, les auteurs ont montré que ~40% contiennent un motif de fixation du facteur de transcription CCCTC-Binding Factor (CTCF) sur les deux ancres de la boucle (Rao et al., 2014), en accord avec une analyse ayant montré que les sites CTCF étaient enrichis en contacts dans les données Hi-C originales (Botta et al., 2010). Comme la protéine CTCF et la cohésine (un large complexe protéique en forme d'anneau, qui sera décrit dans la partie 1.3.1), colocalisent sur la chromatine (Wendt et al., 2008), ce complexe a été cherché aux bases des boucles, montrant qu'effectivement ces régions sont enrichies en cohésine. Un autre motif fréquent, proche de la diagonale, est une « bande » ou une « ligne » qui émane d'un site CTCF (Fudenberg et al., 2016; Schalbetter et al., 2017), et qui reflète un enrichissement des contacts entre le site CTCF et son voisinage, s'étendant parfois jusqu'à ~1-3 Mb **(Figure 4. B)**. Les données Micro-C et Hi-C les plus récentes et de plus haute résolution ont révélé l'abondance de ces motifs de « points » et de « bandes », et que de nombreuses bandes semblent inclure une série de points (Y.-Y. P. Hsieh et al., 2020; Krietenstein et al., 2020). Il a donc été proposé que la cohésine et les protéines CTCF jouent un rôle dans la formation des TADs.

En utilisant une approche par degron auxine, un système de dégradation inducible, la cohésine a été dégradée dans une lignée cellulaire de cancer du côlon humain (HCT-116) synchronisée en interphase. Après la perte de la cohésine, les boucles sont perdues dans les matrices Hi-C (**Figure 4. A**), (Rao et al., 2017). Suite à l'ajout de l'auxine (donc en arrêtant la dégradation de la cohésine), les TADs sont à nouveau détectés par Hi-C au bout de 15 minutes (**Figure 4. A**), (Rao et al., 2017). Ces résultats indiquent que la formation des domaines en boucle nécessite la cohésine, que les domaines en boucle disparaissent rapidement après sa dégradation, et que sa restauration permet de récupérer ces domaines.

La nature moléculaire des TADs reste débattue: ces structures seraient principalement observées par Hi-C et non par d'autres approches comme la microscopie (Sikorska & Sexton, 2020). En effet, le processus de formation des boucles semble être dynamique, et c'est la fixation d'une population de cellules qui génère ces patterns **(Figure 4. B)**. La microscopie a permis d'observer une

grande variation des distances d'une cellule à l'autre pour des paires de loci à l'intérieur d'un même TAD ou dans des TADs différents (Bintu et al., 2018; Finn et al., 2019; Su et al., 2020). En éditant les sites CTCF gauche et droit du TAD pour mesurer la fréquence de contact entre ces deux éléments, Gabriele et al. mesurent que ~92% du temps, le TAD est partiellement extrudé, avec ~57 à 61% de la région capturée dans une à trois boucles de cohésine extrudées, tandis que ~39 à 43% restent non extrudés. Ainsi, ils révèlent que l'état de boucle médié par le CTCF et la cohésine qui maintient ensemble les limites du CTCF des TADs semble être un évènement relativement rare et transitoire (Gabriele et al., 2022).

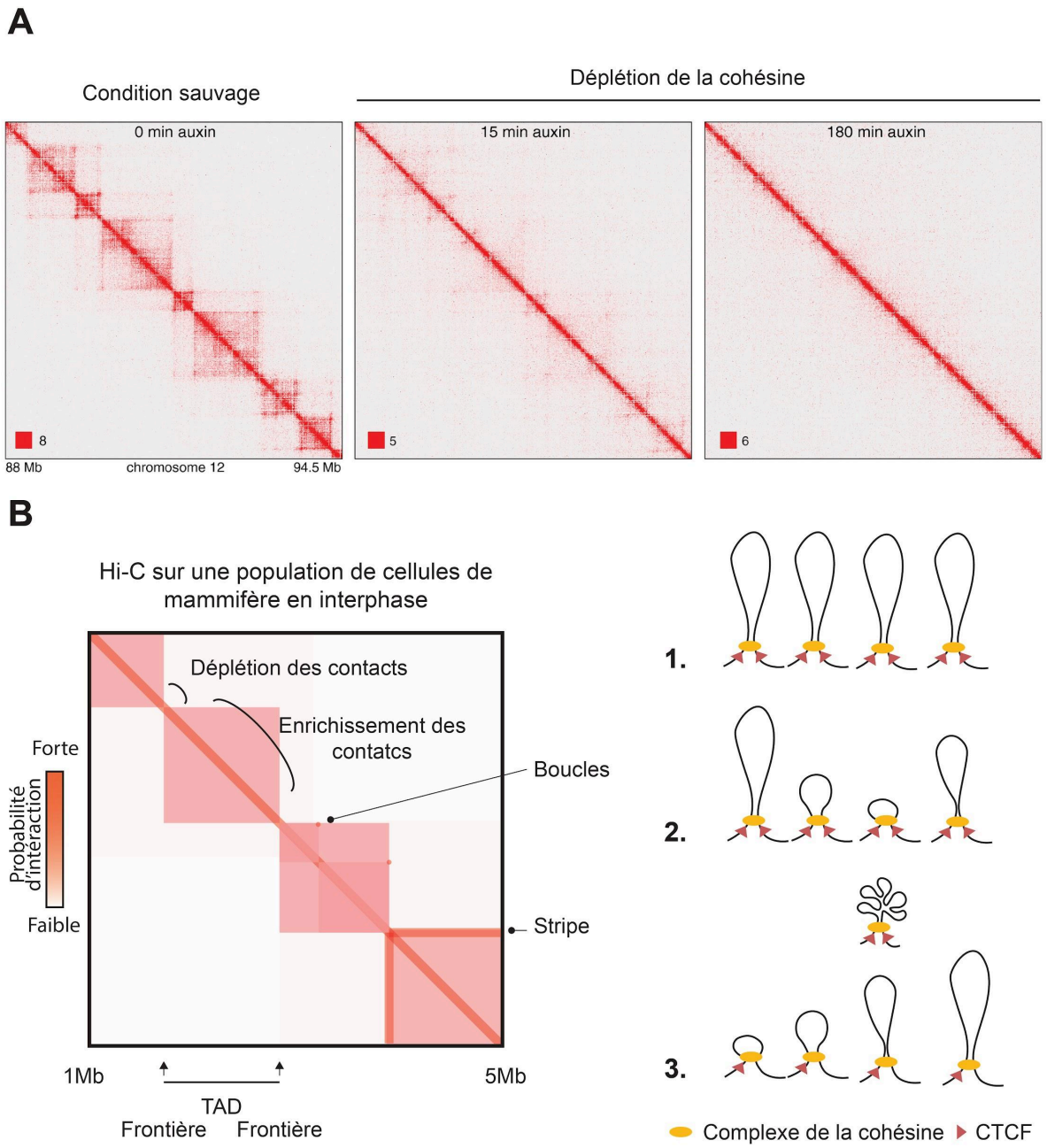


Figure 4 : Les domaines topologiquement associés (TADs) chez les mammifères.

A) Cartes de contact Hi-C de cellules HeLa en G1 en condition sauvage et en l'absence de cohésines (Wutz et al., 2017). **B)** Schéma d'une carte de contact Hi-C avec des TADs **1.** Les boucles d'ADN formées par les interactions récurrentes entre les mêmes séquences dans plusieurs cellules apparaissent sous forme de points sur les cartes Hi-C. **2.** Les TADs, qui apparaissent sous forme de triangles sur les cartes Hi-C, sont des domaines chromosomiques contenant des séquences qui interagissent préférentiellement entre elles et moins avec des séquences situées en dehors du TAD. Les TADs peuvent représenter la moyenne de nombreuses boucles se produisant à différentes positions au sein du TAD dans plusieurs cellules (en haut) ou des boucles interagissant à plusieurs endroits le long de leur longueur (en bas). **3.** Les « stripes » ou bandes se forment lorsqu'une séquence particulière interagit avec plusieurs séquences dans une population de cellules. Cela peut se produire si les boucles sont fréquemment contraintes d'un côté par un bord du TAD. Adapté de (Davidson & Peters, 2021)

Les micro-domaines et les boucles de chromatine chez *S. cerevisiae*

On peut noter qu'en 2009, le premier article décrivant la méthode Hi-C utilisait des bins de 1Mb (soit 1 pixel = 1 Mb) pour étudier le génome humain (Lieberman-Aiden et al., 2009). Une telle résolution chez *S. cerevisiae* conduirait à des matrices (inutilisables) de 12x12. La première étude de type 3C de levure utilisait donc un protocole assez complexe, conduisant à des cartes de 20 kb de résolution (Duan et al., 2010). Cette première analyse n'a souligné que l'organisation Rabl (colocalisation forte des centromères et tendance au regroupement des sub-télomères), et une relative homogénéité des contacts le long des bras chromosomiques (Cournac et al., 2012; Duan et al., 2010). Graduellement l'approche Hi-C classique a été adaptée à *S. cerevisiae* pour atteindre 10 kb, généralisant cette organisation Rabl à d'autres espèces (Marie-Nelly et al., 2014). Enfin, la génération de cartes de contact à plus haute résolution ainsi que le développement du Micro-C a révélé la présence de petits domaines d'auto-interactions (**Figure 5. A**), (T.-H. S. Hsieh et al., 2015, 2016; Lazar-Stefanita et al., 2017; Schalbetter et al., 2017). Ces "domaines", détectés principalement en G1, ont été initialement appelés domaines d'interaction chromosomique (CIDs) par analogie avec les domaines bactériens, tout en établissant un parallèle avec les TADs mammifères en se basant sur le nombre moyen de gènes par domaine (un à cinq). Ce parallèle suggérait que la formation de frontières par le recrutement de protéines régulatrices et structurales était un déterminant clé de l'organisation des chromosomes chez les eucaryotes (T.-H. S. Hsieh et al., 2015). Cependant, tous ces parallèles sont maintenant largement caduques. En effet, les frontières entre ces petits domaines de levure sont enrichies en promoteurs de gènes fortement exprimés, bien que tous les promoteurs ne forment pas de frontières. Contrairement aux TADs, les frontières ne sont pas enrichies en cohésine. Par ailleurs, les CIDs bactériens ne correspondent pas forcément à des structures topologiques distinctes et il est vraisemblable que ces petits domaines correspondent, comme dans la bactérie, à des régions transcrites se distinguant dans les données Hi-C (Bignaud et al., 2024).

En réalisant une analyse complète de l'organisation du génome de *S. cerevisiae* tout au long du cycle cellulaire, deux équipes ont montré que les chromosomes se comportent différemment au cours de la division cellulaire. Particulièrement en G2/M où la p(s) révèle un enrichissement en contacts des chromosomes entre des loci séparés par environ 10-30 kb (Lazar-Stefanita et al., 2017; Schalbetter et al., 2017). Suite à ces travaux, les chromosomes de la levure *S. cerevisiae* ont été montrés comme étant organisés en boucles de chromatine et en domaines dépendants des cohésines (**Figure 5. A,B**), (Dauban et al., 2020; Garcia-Luis et al., 2019). Cette observation a été confirmée par une étude du laboratoire de Doug Koshland utilisant la technique de Micro-C (**Figure 5. A,B**), (Costantino et al., 2020). Les aspects mécanistiques de la cohésine sont abordés dans la partie 1.3.2.

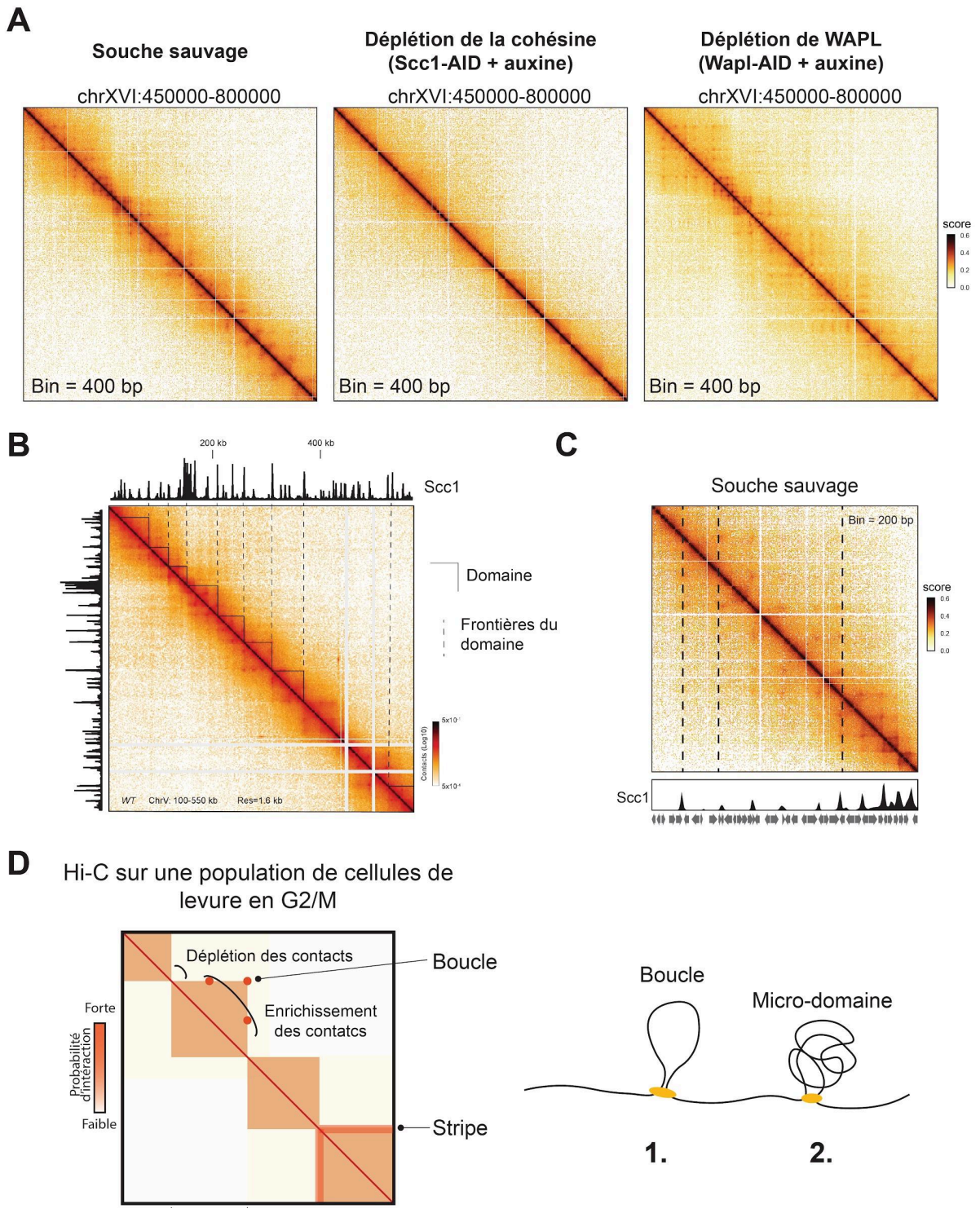


Figure 5 : Domaines et boucles dépendantes de la cohésine chez la levure.

A) Cartes de contact Hi-C d'une partie du chromosome XVI : sauvage, après déplétion de la cohésine, après déplétion de Wapl. (Costantino et al, 2020) **B)** Cartes de contact Hi-C d'une partie du chromosome XVI avec la ChIP contre Scc1. **C)** Les enrichissement en cohésine correspondent aux boucles. **D)** Schéma d'une carte de contact Hi-C de levure en G2/M avec des boucles 1. Les boucles d'ADN formées par les interactions récurrentes entre les mêmes séquences dans plusieurs cellules apparaissent sous forme de points sur les cartes Hi-C. 2. Les CIDs, qui apparaissent sous forme de triangles sur les cartes Hi-C, sont des domaines chromosomiques contenant des séquences qui interagissent préférentiellement entre elles et moins avec des séquences situées en dehors.

1.2.3 Compartimentation

Les compartiments

Dans le noyau, les compartiments sont identifiés comme des agrégats, composés de nombreuses protéines et d'ARN, qui se localisent dans le nucléoplasme et se regroupent dans des corps nucléaires plus ou moins distincts, mais toujours concentrés dans l'espace. Plusieurs éléments peuvent être considérés comme des compartiments. On peut citer le nucléole, compartiment de la biosynthèse des ribosomes et la région génomique ayant la plus forte activité transcriptionnelle (Warner, 1999), l'espace périphérique du noyau avec les subtélomères et les protéines SIR (Taddei & Gasser, 2012), et d'autres éléments identifiés par microscopie comme les corps de Cajal et les " speckles nucléaires" (Belmont, 2022). La méthode FISH a également démontré l'existence de territoires chromosomiques et d'un enchevêtrement de chromosomes à la périphérie de ces territoires (Cremer & Cremer, 2010). De plus, de nombreuses espèces présentent une configuration chromosomique interphasique variante "Rabl". De récentes analyses Hi-C et moléculaires ont révélé que cette configuration Rabl est probablement apparue chez de multiples espèces par le biais d'une évolution convergente induite par la réduction de l'activité de la condensine II dûe à des mutations dans ses sous-unités (Hoencamp et al., 2021).

L'euchromatine et l'hétérochromatine sont observées et décrites dès la fin du XIXe et au début du XXe siècle (Passarge, 1979). Emil Heitz divise la chromatine en euchromatine, décondensée et transcriptionnellement active, et en hétérochromatine, dense, faiblement accessible et transcriptionnellement silencieuse. Il décrira ensuite une hétérochromatine facultative et constitutive. Des travaux plus récents montrent que l'organisation de l'ADN chez la plupart des organismes n'est pas simplement dichotomique, mais comprend plusieurs catégories, avec des régions euchromatiques silencieuses et des régions transcrites alors qu'elles portent des marques typiques de l'hétérochromatine (Hediger & Gasser, 2006). Par exemple, chez la drosophile, on peut distinguer cinq types de chromatine avec des marques épigénétiques spécifiques (Filion et al., 2010). Pour plus de simplicité, on peut approximativement séparer la chromatine en "active" et "inactive".

Une chromatine active et inactive

La chromatine active est ici désignée comme la fraction de la chromatine qui est dans un état "actif" pour l'expression des gènes et l'initiation de la réplication de l'ADN. La chromatine active représente un environnement très accessible, présentant une densité accrue de sites hypersensibles à la DNase I par rapport à d'autres types de chromatine (Thurman et al., 2012). Elle présente des sites de liaison pour de nombreux facteurs chromatiniens et elle est ornée d'une pléthore de modifications d'histones telles que les méthylations H3 sur la lysine 4, la lysine 36 et la lysine 79 et l'acétylation de

plusieurs lysines sur les queues N-terminales de H3 et H4 (Filion et al., 2010).

La chromatine dite “silencieuse” (ou hétérochromatine) est souvent associée à une structure qui entrave l'accès à l'ADN sous-jacent. Chez les mammifères, elle se divise en deux types principaux : facultative et constitutive. L'hétérochromatine facultative est réversible et est souvent marquée par H3K27me3, associée aux complexes Polycomb et aux îlots CpG. L'hétérochromatine constitutive est stable, présente autour des centromères et des télomères, marquée par H3K9me2/3 et impliquant les protéines HP1 et SU(VAR)3-9. Chez *S. cerevisiae*, la chromatine inactive, bien plus limitée, se trouve sur plusieurs sites génomiques, notamment les loci silencieux de type accouplement (HML et HMR), les télomères et l'ADN ribosomique (ADNr) (Gartenberg & Smith, 2016). Les régions hétérochromatinisées par les protéines SIR chez *S. cerevisiae* montrent similairement une répression transcriptionnelle, une réplication tardive, un enrichissement en séquences répétées et s'accumulent en clusters à la périphérie du noyau et autour du nucléole (Gotta et al., 1996; Taddei et al., 2010). Ces différentes formes de chromatines doivent être séquestrées et sont localisées à des endroits spécifiques pour limiter des effets non désirés, ainsi, les différents types de chromatine vont être partitionnés dans le noyau. En mutant les acteurs impliqués dans l’ancrage des télomères à la périphérie nucléaire, les protéines SIR perdent leur capacité à réprimer, ce qui conduit à l’expression ectopique de gènes (Taddei et al., 2009).

Compartiments A et B

Le Hi-C a également permis de révéler une segmentation des contacts de l’ADN en deux ensembles distincts, appelés arbitrairement compartiments A et B. Ces compartiments sont visualisés sur une carte de probabilité de contacts comme un damier nommé “checkboard pattern” ou “plaid pattern” et représentent des contacts inter- et intra- chromosomiques (**Figure 6. A**), (Lieberman-Aiden et al., 2009). L’analyse de la composition des histones, via des analyses de type ChIP-seq, a permis de corréler le compartiment A à la chromatine ouverte, transcriptionnellement active et semblable aux caractéristiques de l’euchromatine. Le compartiment B quant à lui est associé à des marques d’hétérochromatine et est transcriptionnellement inactif (Lieberman-Aiden et al., 2009; Rao et al., 2014). Les régions riches en répétitions de l'hétérochromatine, typiquement péricentromériques et péritélomériques, ne sont pas visibles en Hi-C et ne peuvent donc pas être classées dans le compartiment B. Les états interphasique et mitotique représentent deux organisations 3D du génome fonctionnellement distinctes : les chromosomes mitotiques conservent peu, voire aucune, des caractéristiques structurelles qui définissent les chromosomes en interphase (Naumova et al., 2013). Pendant la mitose, on observe une perte des compartiments et des TADs, qui réapparaissent après la ségrégation mitotique (**Figure 6. C),** (Gibcus et al., 2018).

Il existe peu d'exceptions à cette organisation suivant les types cellulaires de mammifères. La plus notable, pour le moment, consiste en l'organisation de la chromatine dans les cellules photoréceptrices à bâtonnets des mammifères nocturnes. Après coloration au DAPI, ces cellules présentent une architecture nucléaire inversée avec un seul chromocentre alors que les autres types cellulaires présentent plusieurs chromocentres adjacents à la périphérie nucléaire ou au nucléole (Solovei et al., 2009). En combinant une analyse Hi-C des noyaux de bâtonnets inversés avec des approches de microscopie, et une simulation de polymère, Falk et al. mettent en évidence que les attractions entre les régions hétérochromatiques sont cruciales pour établir à la fois la compartimentation et les enveloppes concentriques de l'hétérochromatine péricentromérique, de l'hétérochromatine facultative et de l'euchromatine dans le noyau inversé (Falk et al., 2019). Les expériences et les modélisations suggèrent que ces attractions sont essentielles pour la séparation des phases du génome actif et inactif dans les noyaux inversés et conventionnels.

De manière intéressante la régulation de la formation des TADs et des compartiments A et B semblent antagonistes. En effet une déplétion des cohésines ou de CTCF conduit à la perte des boucles mais à une meilleure ségrégation des compartiments A et B et à l'inverse un plus grand nombre de boucles conduit à une diminution de la compartimentation (Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017, 2020). Ces résultats suggèrent que la dégradation de la cohésine met en évidence une compartimentation "innée" qui est partiellement estompée par l'extrusion de boucle. La contribution de la transcription à cette organisation est actuellement fortement étudiée.

Et chez la levure ?

La levure *S. cerevisiae* ne présente pas de compartiments actifs et inactifs comme trouvés chez les mammifères. Comme mentionné précédemment, elle possède cependant des compartiments comme le nucléole et l'espace périphérique du noyau, incluant les subtélomères. Cela peut s'expliquer par un génome majoritairement « actif », avec très peu d'éléments répétés. La chromatine inactive chez *S. cerevisiae* se trouve sur quelques sites génomiques, notamment les télomères, l'ADNr et les cassettes HML et HMR (Gartenberg & Smith, 2016).

Même si ces régions ne forment pas, à proprement parler, des compartiments A et B, les télomères par exemple interagissent entre eux via les complexes SIR. Les compartiments de chromatine actifs et inactifs semblent donc refléter une propriété de génomes composés d'hétérochromatine canonique de type H3K9me3, et d'éléments génétiques répétés. Cela concerne notamment des grands génomes comme ceux des métazoaires, des insectes ou des plantes, mais pas dans les génomes plus petits et surtout denses en gènes et pauvres en éléments répétés d'eucaryotes tels que la levure.

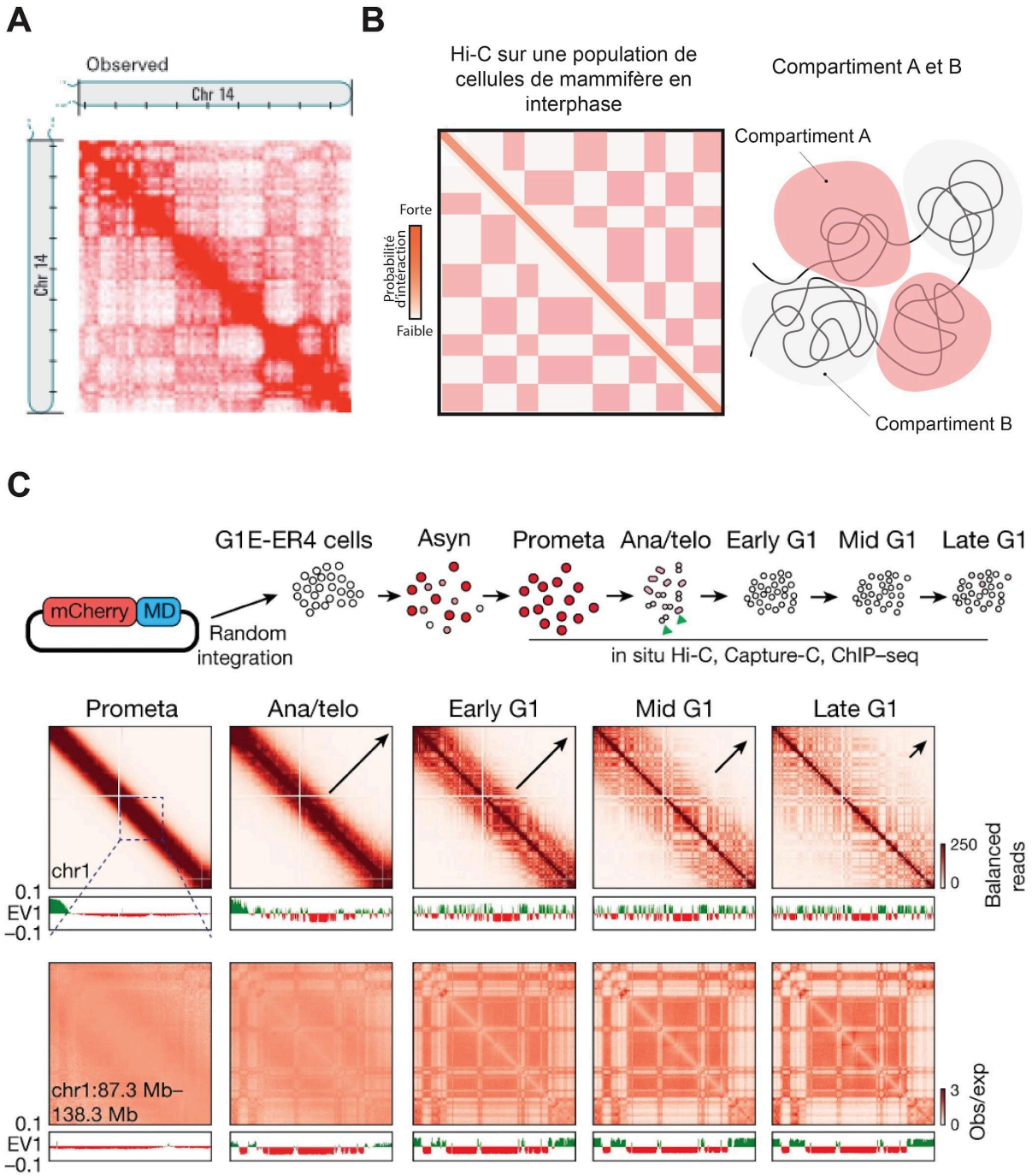


Figure 6 : Les compartiments A et B chez les mammifères détectés par Hi-C.

A) Première évidence des compartiments par Hi-C (Lieberman-Aiden et al., 2009)**. B)** Schéma représentant les compartiments A et B. **C)** Dynamique des compartiments au cours du temps (Y. Zhang, Zhang, et al., 2019). **1)** Schéma montrant le gène rapporteur fusionné au domaine de dégradation mitotique de la cycline B de souris et le signal attendu à chaque stade du cycle cellulaire. Les pointes de flèches vertes indiquent le tri des cellules en anaphase ou en télophase (ana/télo). Asyn, asynchrone ; prometa, prométaphase. **2)** Cartes de contact Hi-C montrant la restauration des compartiments A/B de la chromatine du chromosome 1 (chr1) après la mitose. Taille des matrices : 250 kb. Les flèches indiquent l'expansion des compartiments. **3)** Vue agrandie (chr1 : 87,3 Mb-138,3 Mb) de la carte de b, révélant clairement la mise en place des compartiments en ana/télophase.

1.3. Mécanismes et acteurs principaux de l'organisation 3D

Dans la partie 1.2, nous avons exploré plusieurs structures mises en évidence par microscopie et par les techniques 3C. Ces structures dépendent de plusieurs acteurs, à la fois moléculaires et physiques, directs et indirects. J'aborderai brièvement les processus physiques qui contribuent vraisemblablement à cette organisation, notamment la LLPS (séparation de phase liquide-liquide), qui pourrait être impliquée dans la formation de compartiments au sein du noyau eucaryote. Ensuite, je me concentrerai sur deux principaux acteurs moléculaires actifs de l'organisation 3D mis en évidence ces dernières années : les complexes SMC et la transcription.

1.3.1 La séparation de phase

La séparation de phase est un processus physique où des composants non miscibles, tels que l'eau et l'huile, se séparent en phases distinctes. Ce mécanisme jouerait un rôle crucial dans l'organisation du noyau eucaryote, en contribuant potentiellement à la formation de structures de plus haut ordre telles que les compartiments d'hétérochromatine périphérique, les nucléoles, et d'autres corps nucléaires (Nuebler et al., 2018). Ce processus biophysique peut expliquer le regroupement de types spécifiques de chromatine et l'agrégation d'ensembles protéiques par le biais d'interactions multivalentes faibles. Les hypothèses théoriques conduisent maintenant à des prédictions testables sur la façon dont les perturbations expérimentales modifieraient la compartimentation. Les marques d'histones des deux compartiments étant très différentes, il est possible qu'elles se repoussent mutuellement entraînant ainsi un phénomène de séparation de phase (Falk et al., 2019). La ségrégation spatiale observée dans les compartiments de chromatine peut être comparée à la séparation de phase dans les polymères composés de monomères de types A et B, avec une attraction A-A et/ou B-B. Des simulations de polymères ont montré que les motifs en damier observés en Hi-C peuvent être reproduits lorsque les régions de même type s'attirent mutuellement (Barbieri et al., 2012; Di Pierro et al., 2016).

Les complexités des relations entre le mouvement de la chromatine, la séparation de phases et la fonction du génome ne sont actuellement pas claires, mais des travaux récents ont montré que la séparation de phases des protéines est capable de ségréger les loci en excluant l'ADN des gouttelettes protéiques denses et peut aussi les rapprocher en les attachant à la même gouttelette; ces travaux suggèrent que les protéines qui se séparent en phases peuvent être capables de générer des structures stables à partir d'interactions dynamiques. Il est intéressant de noter que la ségrégation spatiale de la chromatine active et inactive ne dépend pas de l'ancrage de l'hétérochromatine et se produit également dans les noyaux inversés dans lesquels cet ancrage est absent et l'hétérochromatine est regroupée au

centre du noyau (Falk et al., 2019). La séparation spatiale de la chromatine active et inactive (compartimentation) et la formation de corps subnucléaires peuvent toutes deux être comprises comme le résultat d'une séparation de phase (Falk et al., 2019; Hildebrand & Dekker, 2020).

1.3.2 Les complexes de maintenance de la structure des chromosomes

Structure et rôle des complexes de maintenance de la structure des chromosomes

Les protéines de maintien structurel des chromosomes (SMC), conservées au cours de l'évolution, se lient aux chromosomes et modifient leur structure de manière régulée dans l'espace et dans le temps au cours du cycle cellulaire (Aragon et al., 2013; Uhlmann, 2016). Des travaux récents montrent que ces larges complexes ont acquis au cours de l'évolution des rôles très variés, tous en lien avec l'ADN, allant par exemple de la maintenance de la cohésion des chromatides soeur à la défense contre des virus dans certaines bactéries. En 1985, Larionov et al. identifient un gène impliqué dans la stabilité des minichromosomes centromériques chez la levure *S. cerevisiae* (Larionov et al., 1985). Ils nomment cette protéine Smc1 pour "Stability of minichromosomes". Parallèlement à cette découverte d'autres chercheurs identifient d'autres protéines Smc (Smc1 et Smc3) grâce à des cribles génétiques visant à identifier des facteurs impliqués dans la séparation précoce des chromatides sœurs. Impliquées dans la cohésion des chromatides sœurs, ces protéines sont donc nommées cohésines (Guacci et al., 1997; Michaelis et al., 1997). Au milieu des années 1990, la purification de protéines à partir d'extraits d'œufs de xénope par Hirano et Mitchison permet l'identification de deux protéines Smc (Smc2 et Smc4) essentielles au maintien de la structure des chromosomes mitotiques in vitro (Hirano & Mitchison, 1994). La suite de leurs travaux met en évidence trois autres protéines essentielles au maintien de cette structure. La déplétion de ces protéines in vitro conduit à une désorganisation massive des chromosomes qui n'apparaissent plus comme des fibres distinctes mais comme un amas diffus en microscopie à fluorescence (Hirano et al., 1997). Ces protéines impliquées dans la condensation des chromosomes sont nommées condensines. Dans un même temps, la protéine rad18 identifiée par une approche de radiations, est révélée comme étant une protéine associée à un troisième sous-groupe de la famille des protéines SMC, appelé complexe Smc5/6. Ce complexe est principalement impliqué dans la réparation des cassures de l'ADN (Aragón, 2018).

Les membres de la famille des SMC partagent une architecture commune, caractérisée par un anneau tripartite composé de deux sous-unités Smc et d'une sous-unité kleisine. Les protéines Smc ont une conformation spécifique : elles sont composées d'un domaine charnière central (aussi appelé « hinge »), séparé des deux domaines globulaires N- et C-terminaux par de longues répétitions coiled-coil (**Figure 7. A**). Ces dernières interagissent entre elles de manière antiparallèle, ce qui donne naissance à une protéine repliée dans l'espace composée du domaine charnière et d'une « tête »

générée par l'interaction des domaines N- et C-terminaux, de part et d'autre des coiled-coil. La sous-unité kleisin se lie aux deux têtes Smc, ce qui permet la fermeture du V et la formation d'un anneau. Elle sert également de plateforme de recrutement pour diverses protéines régulatrices.

Chez *S. cerevisiae*, le complexe annulaire de la cohésine est composé de trois sous-unités principales : Smc1, Smc3 et Scc1/Mcd1 (Scc1 est remplacé par Rec8 lors de la méiose). Les domaines "head" de Smc1 et Smc3 sont reliés entre eux via la sous-unité Scc1 (**Figure 7. A**). L'activité du complexe de la cohésine est également régulée par plusieurs protéines qui forment un holocomplexe : Scc2/Scc4, Wapl, Scc3, Pds5 et Eco1. La condensine quant à elle est constituée de deux sous-unités SMC (Smc2 et Smc4) reliées par Brn1. Elle interagit avec des protéines régulatrices Ycs4 et Ycg1. Les mammifères possèdent deux complexes de condensines, les condensines I et II, qui s'associent aux chromosomes mitotiques. Enfin, le complexe Smc5/6 est composé de deux sous-unités principales, Smc5 et Smc6 et inclut également plusieurs sous-unités associées (Nse1, Nse2/Mms21, Nse3, Nse4) (**Figure 7. A**).

Bien qu'ils partagent une architecture similaire, les trois complexes SMC ont chacun une ou des activités spécifiques. Les cohésines assurent la cohésion des chromatides soeurs pendant la réplication de l'ADN chez tous les eucaryotes étudiés (Glynn et al., 2004) et sont clivées au niveau de la kleisin par la Séparase lors de la transition métaphase-anaphase pour libérer les chromatides soeurs (Uhlmann et al., 1999). Chez *S. cerevisiae*, et chez plusieurs organismes unicellulaires, la cohésine et la condensine jouent également un rôle dans la compaction lors du stage G2/M, et contribuent à la ségrégation chromosomique (Guérin et al., 2019; Lazar-Stefanita et al., 2017; Schalbetter et al., 2017). La condensine quant à elle semble être impliquée dans la ségrégation des chromatides lors de l'anaphase, et contribue notamment à la structuration du locus de l'ADNr (**Figure 7. B**), (Guérin et al., 2019; Lazar-Stefanita et al., 2017; Schalbetter et al., 2017). Chez les mammifères, les rôles des complexes diffèrent. La cohésine est impliquée dans la structuration des TADs en interphase, et il a été proposé qu'elle agisse via la formation de boucles le long des chromosomes (**Figure 7. C**), (Wutz et al., 2017). Parallèlement, les condensines telles que Smc2 sont chargées sur les chromatides soeurs pour faciliter leur ségrégation (Stephens et al., 2013). Des expériences de microscopie, de Hi-C et de modélisation ont indiqué que la condensine II forme initialement de grandes boucles et que la condensine I forme ensuite de plus petites boucles imbriquées à l'intérieur de ces boucles (**Figure 7. C**), (Gibcus et al., 2018). Le complexe Smc5/6, quant à lui, joue un rôle dans la réparation des cassures d'ADN. D'autres rôles sont actuellement explorés, notamment au niveau des régions sur-enroulées de l'ADN (Jeppsson et al., 2024). Il est important de noter que l'activité de ces complexes est fondamentale puisqu'ils sont tous essentiels pour la survie cellulaire (Nasmyth & Haering, 2005).

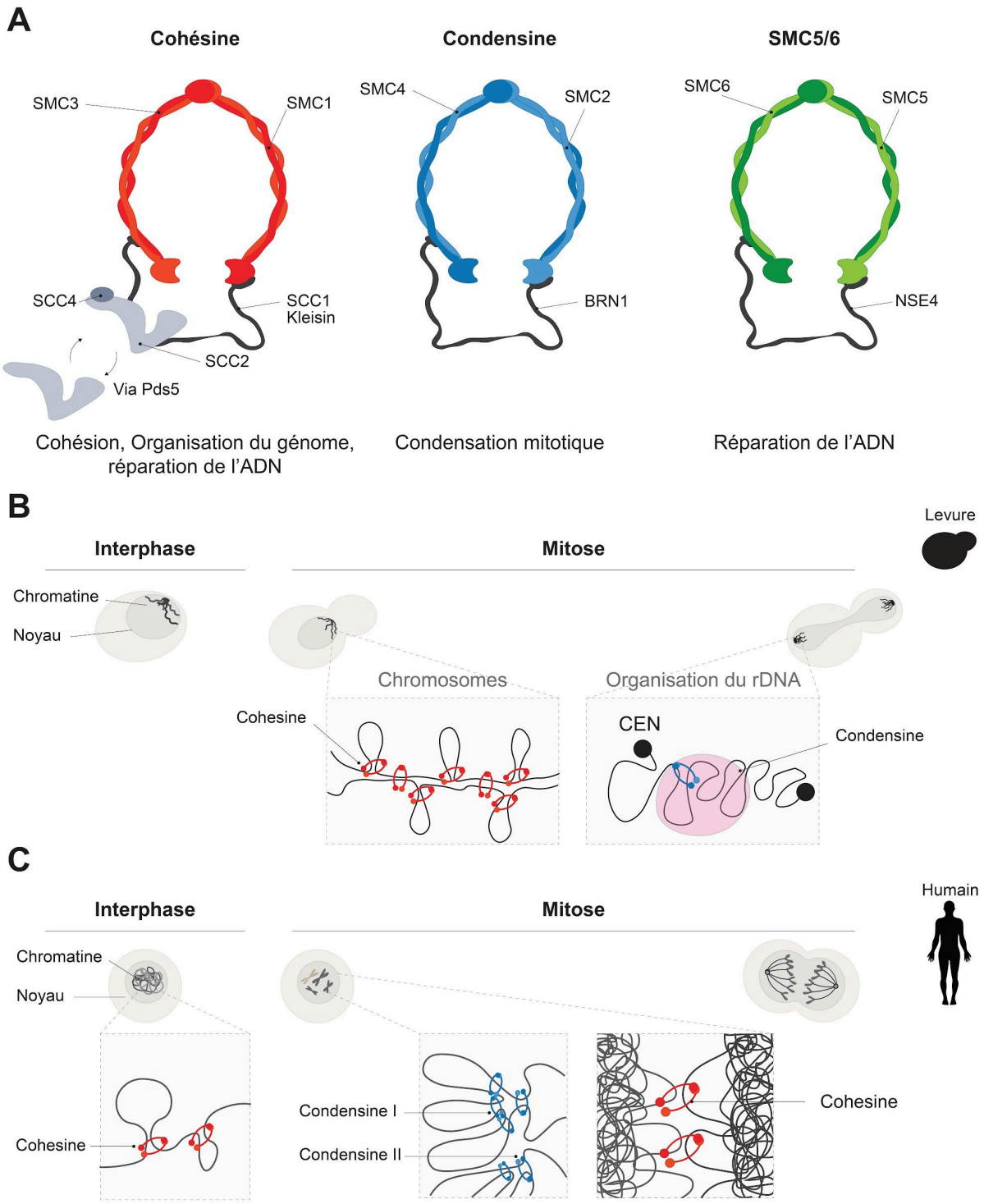


Figure 7 : Les complexes de la famille SMC.

A) Schéma des complexes cohésine, condensine, et SMC5/6 associés à leurs rôles respectifs. **B)** Chez la levure, en mitose, la cohésine forme des boucles de chromatine et maintient les chromatides sœurs ensemble. La condensine structure l'ADNr. **C)** Chez l'humain, la cohésine façonne le génome en interphase en formant des boucles de chromatine et maintient les chromatides sœurs ensemble après la réplication de l'ADN jusqu'au début de l'anaphase. Les complexes de condensine organisent les chromosomes mitotiques en formant des boucles imbriquées. Adapté de (Hoencamp & Rowland, 2023).

L’extrusion de boucle chez les mammifères

Il y a une vingtaine d'années, il a été proposé que les complexes SMC façonnent la chromatine en formant des boucles d'ADN et en les élargissant de manière progressive. Il est proposé dans la littérature que les TADs se forment par un mécanisme appelé "extrusion de boucle", dans lequel les cohésines produisent et élargissent progressivement de petites boucles d'ADN le long des chromosomes jusqu'à ce qu'elles rencontrent une « barrière » (**Figure 8. A**), (Davidson & Peters, 2021; Hassler et al., 2018). L'extrusion de boucles a été récemment observée lors d'expériences montrant que la condensine de levure peut rapidement s'associer à des molécules d'ADN linéaires pour former des boucles in vitro (Ganji et al., 2018). De plus, les laboratoires de Peters et de Hongtao ont démontré que la cohésine humaine purifiée est également capable d'extruder des boucles in vitro (**Figure 8. B**), (Davidson et al., 2019; Y. Kim et al., 2019).

Ce mécanisme d’extrusion de boucle peut expliquer comment la cohésine forme des boucles chromatiniennes en interphase qui sont ancrées sur des sites CTCF : la cohésine formerait initialement une petite boucle qui grandirait jusqu'à ce que le complexe de cohésine extrudé rencontre des protéines CTCF liées à leurs sites de reconnaissance de part et d'autre de la cohésine. Sur ces sites, la protéine CTCF arrête le processus d'extrusion (**Figure. 8 A**), (Davidson & Peters, 2021; Fudenberg et al., 2016; Hassler et al., 2018). Des expériences Hi-C utilisant des cellules de mammifères, dans lesquelles une sous-unité de cohésine ou CTCF peut être dégradée, ont confirmé ces prédictions (Nora et al., 2017; Rao et al., 2017; Schwarzer et al., 2017; Wutz et al., 2017). Ce modèle suggère que la taille des boucles formées par les cohésines dépend de leur capacité d’extrusion et de leur temps de résidence sur l’ADN.

Dynamique de l’extrusion de boucle chez la levure

Chez la levure, et chez beaucoup d'autres espèces, les boucles médiées par la cohésine ne sont pas détectées durant l'interphase. Toutefois, la cohésine joue un rôle crucial dans la formation de boucles chromosomiques pendant la mitose, notamment lors de la métaphase (Dauban et al., 2020). Chez *S. cerevisiae*, deux populations distinctes de cohésines ont été identifiées sur les chromosomes mitotiques : les cohésines cohésives, qui maintiennent la cohésion des chromatides sœurs, et les cohésines impliquées dans l'extrusion de boucles, un processus essentiel pour l'organisation des chromosomes en mitose (Dauban et al., 2020).

Chez les métazoaires, les cohésines sont chargées sur l’ADN dès la télophase alors que chez la *S. cerevisiae* elles sont chargées au moment de la transition G1/S. Cela s’explique par un faible niveau d’expression de la sous-unité Scc1 ainsi que par la persistance de la Séparase capable de la cliver durant la phase G1. Le chargement des cohésines sur l'ADN est ensuite initié par le complexe

formé par Scc2 et Scc4 (Ciosk et al., 2000). L'établissement de la cohésion au cours de la phase S va permettre de faire passer la cohésine d'un état dynamique à un état stable grâce à l'acétylation de deux lysines conservées K112 et K113 au niveau de la tête ATPase de Smc3 par l'acétyltransférase Eco1 (**Figure 8. C**), (Beckouët et al., 2010).

Une fois les cohésines chargées, l'expansion de ces boucles est régulée par deux mécanismes principaux : une voie dépendante de l'activité de dissociation de Wapl et une autre dépendante de Eco1 (Dauban et al., 2020), (**Figure 8. D**). Selon le modèle d'extrusion de boucles, la taille des boucles dépend de deux paramètres : le temps de résidence des cohésines sur l'ADN et leur vitesse de translocation. En l'absence de Pds5, on observe une extension des interactions intrachromosomiques sur de plus grandes distances, et une diminution du nombre de boucles d'ADN, suggérant que Pds5 joue un rôle crucial dans la limitation de l'expansion des boucles. De plus, les bases des boucles se concentrent principalement au niveau des centromères, ce qui laisse supposer que les centromères agissent comme des barrières à l'extrusion des boucles (Costantino et al., 2020; Dauban et al., 2020). L'acétylation de certaines lysines conservées de Smc3 par Eco1 bloque l'extrusion des boucles et stabilise les cohésines sur l'ADN. Scc2 stimule l'hydrolyse d'ATP nécessaire à l'expansion des boucles, mais ce rôle est contrebalancé par le recrutement de Pds5, qui stabilise la cohésine et inhibe l'expansion des boucles via une interaction compétitive avec Scc1(Bastié et al., 2022).

Chez la levure (et de nombreuses espèces) la base des boucles d'ADN dépendantes de la cohésine est préférentiellement positionnée à proximité des gènes convergents, pour des raisons qui encore mal comprises. Le supercoiling, le R-loop, ou tout simplement l'absence de transcription à ces endroits là, sont des paramètres qui pourraient jouer un rôle. Par ailleurs, les polymérases peuvent également représenter des obstacles temporaires à la translocation des SMC en cas de collisions (Glynn et al., 2004; Lengronne et al., 2004).

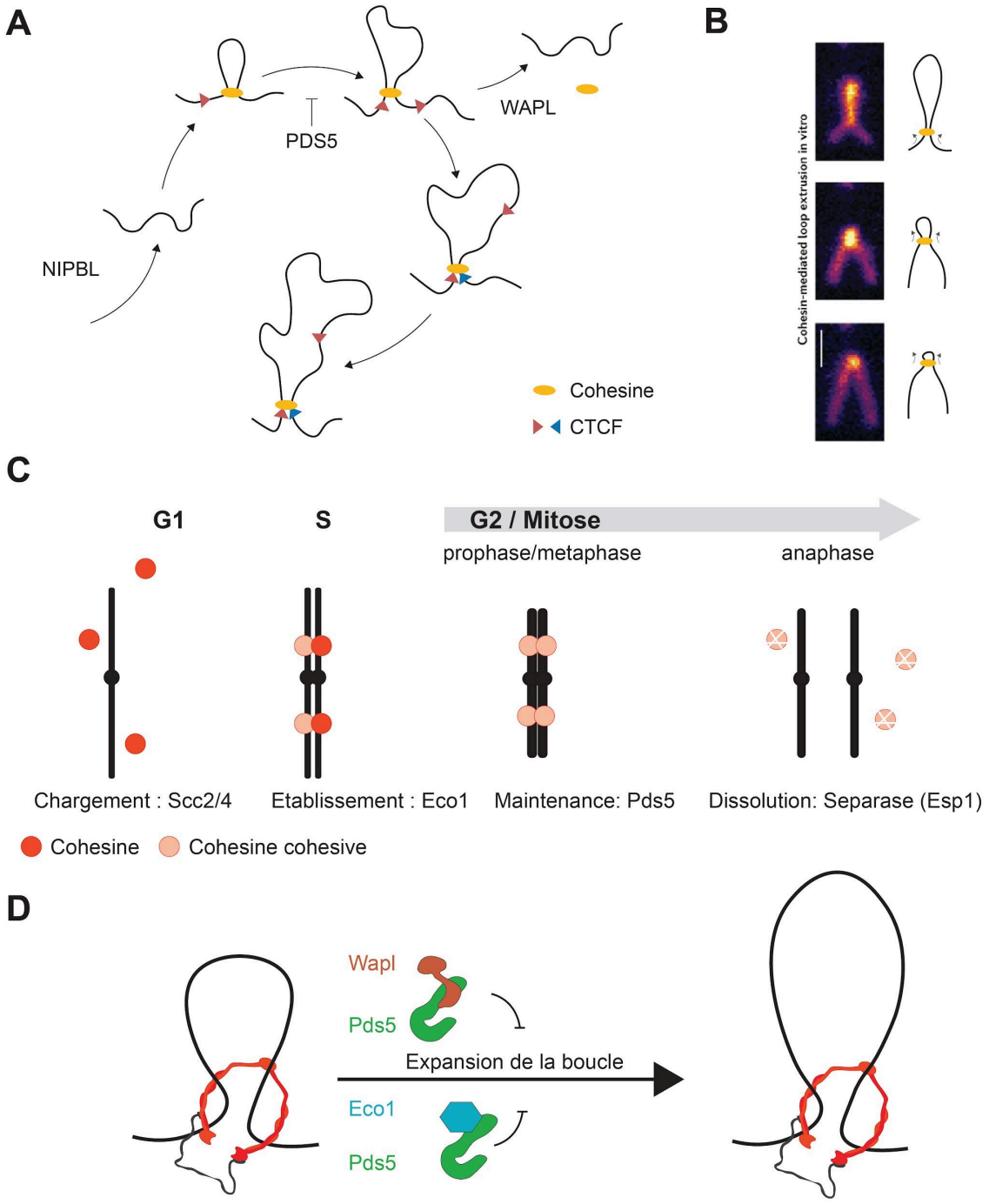


Figure 8 : Dynamique de la cohésine dans l'extrusion de boucles.

A) Modèle d'extrusion de boucles chez les mammifères. **B)** Boucle visualisée in vitro (Davidson et al., 2019) **C)** Dans *S. cerevisiae*, Scc1 est resynthétisé à la fin de G1. La cohésine est chargée sur les chromosomes à la fin de G1/début de la phase S par le complexe Scc2/Scc4. Eco1 acétyle Smc3 sur la lysine 113, pendant la réplication de l'ADN. Cette acétylation convertit la cohésine liée au chromosome en un état cohésif, établissant ainsi la cohésion. Les interactions entre le complexe de cohésine et Pds5 assurent le maintien stable de la cohésion. Au début de l'anaphase, Scc1 subit une dégradation protéolytique par l'activité de la séparase. **D)** Modèle de régulation de la taille des boucles de chromatine par les cohésines proposé par Dauban et al. (2020). Pds5 régule l'expansion des boucles via deux voies : une voie dépendante d'Eco1, qui inhibe l'activité translocase de la cohésine, et une voie dépendante de Wapl, qui dissocie les cohésines de l'ADN. Adapté de (Dauban et al. 2020).

1.3.3 Relation entre la transcription et la structure 3D

Interaction entre la cohésine et la transcription

La machinerie de transcription de l'ARN et le processus d'extrusion de boucle, via les cohésines, interagissent directement le long des chromosomes. Par exemple, les premières études réalisées chez la levure ont suggéré que l'ARN polymérase II pousse les anneaux de cohésine vers les sites de transcription convergente (Glynn et al., 2004; Lengronne et al., 2004). Chez les mammifères, la déplétion de CTCF ou de WAPL entraîne l'accumulation de la cohésine au niveau des gènes convergents (Banigan et al., 2023; Busslinger et al., 2017). Le modèle d'extrusion de boucle suggère que l'ARN polymérase module la position de la cohésine, par exemple en agissant comme une barrière transitoire qui freine la progression de l'extrusion. La cohésine pourrait extruder l'ADN en utilisant l'énergie d'hydrolyse de l'ATP (Davidson et al., 2019), mais la transcription des ARN polymérases pourrait également pousser la cohésine le long de l'ADN (Glynn et al., 2004; Guérin et al., 2023; Lengronne et al., 2004). L'inhibition de la transcription grâce à la thiolutine supprime les barrières d'extrusion des boucles de cohésine, déclenchant la formation de nouvelles interactions cis à longue portée (Jeppsson et al., 2022). Dans l'ensemble, ces études mettent en évidence le rôle de la transcription active et de l'expansion de la boucle médiée par la cohésine dans l'organisation fonctionnelle du génome eucaryote. Cependant, l'interaction entre ces processus n'est toujours pas claire.

Organisation indépendante de la cohésine

Les TADs sont des structures de la chromatine en interphase souvent associées à la cohésine chez les mammifères, formées par l'extrusion de boucles d'ADN. Cependant, des études montrent que les TADs peuvent se former indépendamment de la cohésine dans les cellules individuelles, bien que la cohésine aide à établir des frontières visibles de ces TADs (Bintu et al., 2018). Chez la levure, les micro-domaines détectés par Micro-C sont retrouvés au niveau des gènes transcrits, confortant l'idée que ces motifs correspondent à des régions activement transcrites, similaires à celles détaillées expérimentalement dans les bactéries (Bignaud et al., 2024). De plus, la déplétion de la cohésine empêche la formation de boucles chromatiniennes, mais les micro-domaines persistent, suggérant que l'extrusion de boucles médiée par la cohésine n'est pas essentielle pour la formation des TADs dans ce contexte (Costantino et al., 2020; Dauban et al., 2020; Guérin et al., 2023). Ces observations suggèrent que la cohésine, et par déduction l'extrusion de boucle médiée par la cohésine, n'est pas essentielle pour la formation du TAD dans la levure. On distingue ainsi deux mécanismes importants dans l'organisation 3D des génomes à grande échelle : un mécanisme lié à la transcription qui sépare

la chromatine en compartiments selon les états épigénétiques, et qui peut également être associé à des boucles, et un mécanisme superposé dépendant des SMCs (cohésine, condensine, et probablement Smc5/6) qui génère de manière dynamique des structures en boucle de grande taille à différents stades du cycle cellulaire suivant les espèces.

2. Expression transcriptionnelle eucaryote

L'expression de l'information génétique d'une cellule commence par la transcription. Ce processus biologique fondamental, conservé et extrêmement complexe, est étroitement régulé pour garantir une expression génétique adaptée aux besoins cellulaires. La transcription de l'ADN en ARN messager (ARNm) dépend du complexe ARN polymérase de type II (Pol II) qui transcrit les gènes codant les protéines, un des trois complexes ARN polymérase eucaryotes.

La transcription de l'ADN par la Pol II est un processus orchestré, soumis à une régulation à plusieurs niveaux : 1) la liaison de la Pol II au promoteur, 2) l'initiation de la transcription, et 3) l'élongation. Ces étapes nécessitent l'action concertée de nombreux complexes protéiques et s'accompagnent de modifications locales de la structure chromatinienne. La structure et la composition de l'ADN jouent un rôle prépondérant dans l'association et la régulation de la polymérase : comprendre son organisation et ses interactions avec l'ADN permet d'appréhender la complexité de ce processus.

Au cours des dernières années, la mise en place d'approches transcriptomiques a permis l'étude des transcriptomes et a révélé une organisation du transcriptome fortement entrelacée, impliquant des centaines d'ARNnc. Dans de nombreux organismes eucaryotes, la transcription est omniprésente à l'échelle du génome, étant détectée non seulement au niveau des séquences codantes, mais aussi dans des régions dites "non codantes". Cela soulève des questions sur le rôle de la transcription pervasive au sein du génome mais aussi des régulations post-transcriptionnelles.

Ce second chapitre abordera les aspects fondamentaux de la Pol II, le devenir des transcrits et la régulation transcriptionnelle à longue-distance.

2.1. La synthèse transcriptionnelle

2.1.1 Mise en place du complexe de pré-initiation

La transcription du génome eucaryote est effectuée par l'ARN polymérase I (Pol I), Pol II et Pol III. La Pol I transcrit les précurseurs de l'ARN ribosomal (ARNr) et la Pol III transcrit les petits ARN non codants (ARNnc) tels que les ARN de transferts (ARNt) et les petits ARNs nucléaires. Chez les eucaryotes, la plupart des gènes codants des protéines sont transcrits par l'ARN polymérase II (Pol II) et nous nous intéresserons spécifiquement à cette dernière et ses facteurs.

Un élément majeur de la transcription est le complexe de pré-initiation (PIC), un complexe protéique d'environ 100 protéines qui s'assemble sur les séquences promotrices et permet de positionner l'ARN polymérase II. Les principaux composants de ce PIC sont six facteurs de transcription généraux, la Pol II et des co-activateurs comme le médiateur ou SAGA. Les facteurs de transcription généraux comprennent plusieurs facteurs essentiels pour la transcription, notamment les facteurs d'initiation de la transcription : TFIIA, TFIIB, TFIID, TFIIE, TFIIF et TFIIH qui vont s'assembler progressivement pour former les premiers éléments du PIC (**Figure 9**) (Sainsbury et al., 2015). La Pol II, composée de 12 sous-unités, s'assemble ensuite avec les facteurs de transcription généraux. Le domaine carboxy-terminal (CTD) de la Pol II fait partie de la plus grande sous-unité de l'ARN polymérase II (**Figure 9**). Il contient de multiples répétitions de la séquence heptapeptidique "YSPTSPS", qui sont sujettes à la phosphorylation et à d'autres modifications post-traductionnelles (**Figure 9**), (Buratowski, 2009).

Au sein de ce PIC, l'élément central est la TATA-binding protein (TBP). La TBP est un facteur de transcription universel, nécessaire à l'initiation des trois ARN polymérases et joue un rôle essentiel dans le mécanisme d'activation de la transcription en interagissant avec la TATA box. Cette protéine est un monomère de 27 kDa. En criblant une librairies de mutants de la TBP, il a été démontré que cette protéine a un rôle prépondérant dans l'activation de gènes inductibles comme Gcn4, Gal4, et Ace1 (Lee & Struhl, 1995). Remarquablement, les mutations identifiées sont principalement des mutations liées à l'interaction protéine-ADN permettant de mettre en évidence le contact direct entre la TBP et la séquence d'ADN appelée TATA box (Lee & Struhl, 1995). La TBP interagit ensuite avec TFIID, un des facteurs de transcription généraux spécifique de l'ARN Pol II. TFIIB entre dans le PIC après la TBP et est la condition préalable permettant le recrutement de l'ARN pol II (**Figure 9. A**).

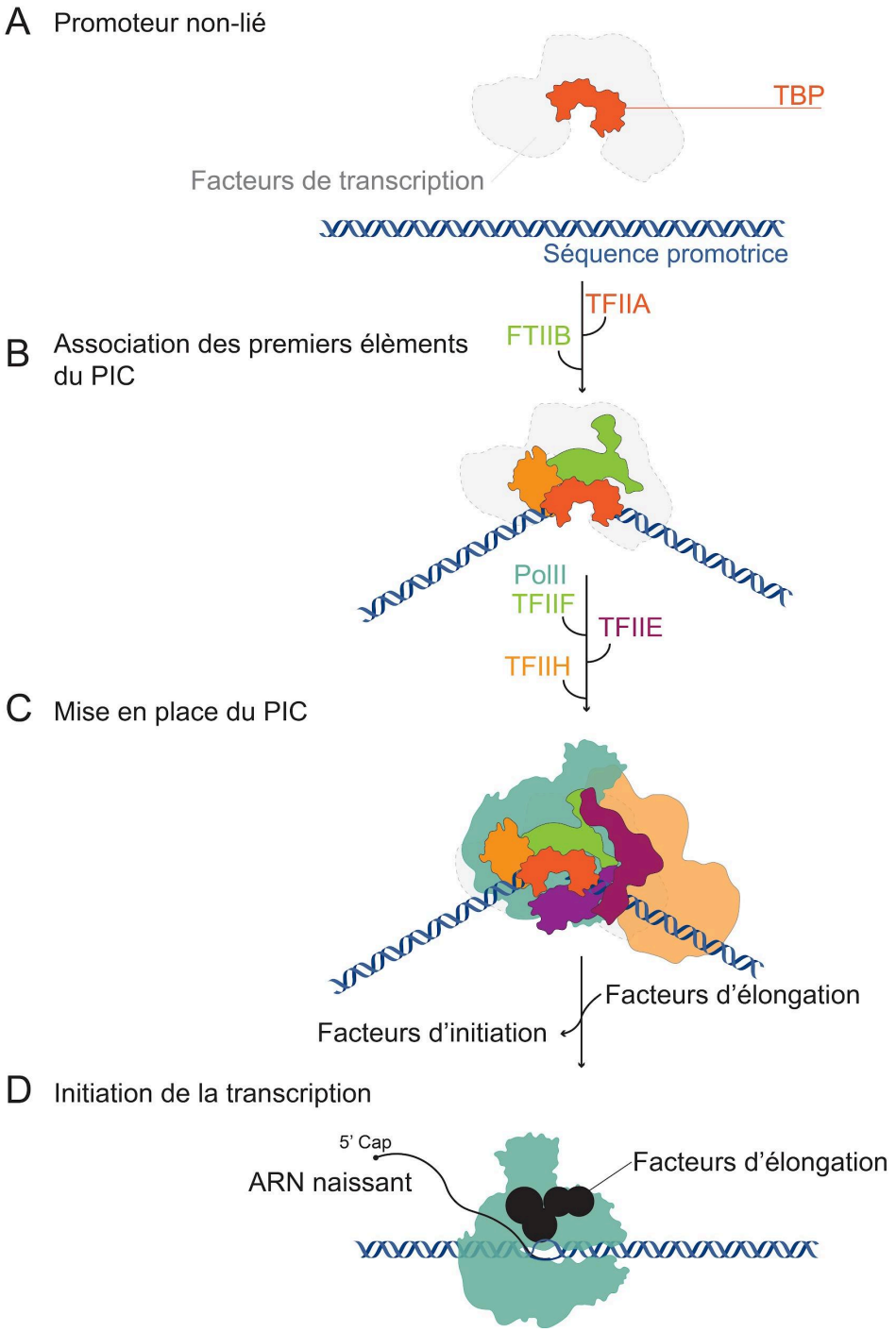


Figure 9 : Schéma de l'assemblage progressif du complexe de pré-initiation.

A partir des facteurs généraux de transcription, présentés sous différentes couleurs, et de l'ARN Pol II sur la séquence promotrice. Les noms des complexes intermédiaires qui se forment pendant la formation du PIC sont indiqués à gauche des images. **A)** TFIID ou sa sous-unité TATA box-binding protein (TBP) se lie à l'ADN promoteur, induisant une courbure. **B)** Le complexe TBP-ADN est ensuite stabilisé par TFIIB et TFIIA, qui encadrent TBP des deux côtés. **C)** Le complexe promoteur amont qui en résulte est rejoint par le complexe Pol II-TFIIF, ce qui conduit à la formation du PIC central. La liaison ultérieure de TFIIE et TFIIH complète le PIC. **D)** Finalement, la transcription est initiée puis les facteurs d'élongation permettent la maturation du transcrit. Figure simplifiée de Sainsbury et al., 2015.

Les coactivateurs transcriptionnels sont indispensables à l'activation de la transcription. Les coactivateurs se distinguent des facteurs de transcription généraux par le fait qu'ils ne sont pas nécessaires à la transcription de base in vitro (Hampsey, 1998). Par exemple, le complexe Médiateur est un complexe multi-sous-unités qui peut interagir avec diverses protéines régulatrices et moduler les signaux transcriptionnels. Il sert de pont entre les facteurs de transcription liés à l'ADN et l'ARN polymérase II, facilitant l'assemblage et la stabilité du PIC (Soutourina, 2018). L'architecture globale du Médiateur visualisée par Cryo-EM est conservée entre les complexes de la levure et de l'homme (Sainsbury et al., 2015; Soutourina, 2018). Le complexe SAGA (Spt-Ada-Gcn5 Acetyltransferase) est également un co-activateur de transcription essentiel qui modifie la structure de la chromatine en acétylant et déubiquitinant les histones, facilitant ainsi l'accessibilité de l'ADN pour la machinerie de transcription (Baptista et al., 2017).

Nous avons passé en revue les acteurs clés de la transcription de l'ARN Pol II chez la levure. La liaison de TBP/TFIID à la séquence d'ADN est une étape critique dans l'assemblage stable du PIC. Quelles sont les particularités de l'ADN qui permettent le recrutement et l'assemblage du PIC ?

2.1.2 Éléments séquentiels du promoteur

Composition de la séquence

La composition de la séquence (GC%, fréquences polynucléotidiques, motifs d'ADN, répétitions) joue un rôle crucial dans la diversité génomique et varie considérablement d'une espèce à l'autre ainsi qu'au sein d'un même génome. Ces derniers présentent une grande diversité dans la composition des séquences, avec un contenu en GC allant de 20 à 65 % (Wang, 2018). Chez *S. cerevisiae*, le GC% calculé sur des fenêtres de 100 kb qui ne se chevauchent pas atteint environ 38% avec une faible variation (+/- 2%). En revanche, le GC% du génome humain est en moyenne de 41%, mais varie de 35 à 60% dans des fenêtres de 100 kb. Les régions présentant des valeurs similaires de contenu en GC sont appelées isochores et leur évolution et leur fonction restent débattues. La composition locale en GC est liée à l'activité transcriptionnelle, les régions à forte teneur en GC étant en moyenne enrichies en séquences codantes, donc plus susceptibles d'être activement transcrites et répliquées rapidement (Holmquist, 1989). De plus, un contenu en GC élevé dans les régions codantes est souvent associé à une stabilité accrue de l'ARNm et à des niveaux d'expression génique plus élevés (Sémon & Duret, 2004). Il existe donc un lien clair entre le GC% et les régions codantes tandis que les régions riches en AT sont plutôt associées à des fonctions régulatrices.

Éléments canoniques des promoteurs

Les promoteurs humains sont connus pour avoir une forte teneur en GC autour du site de

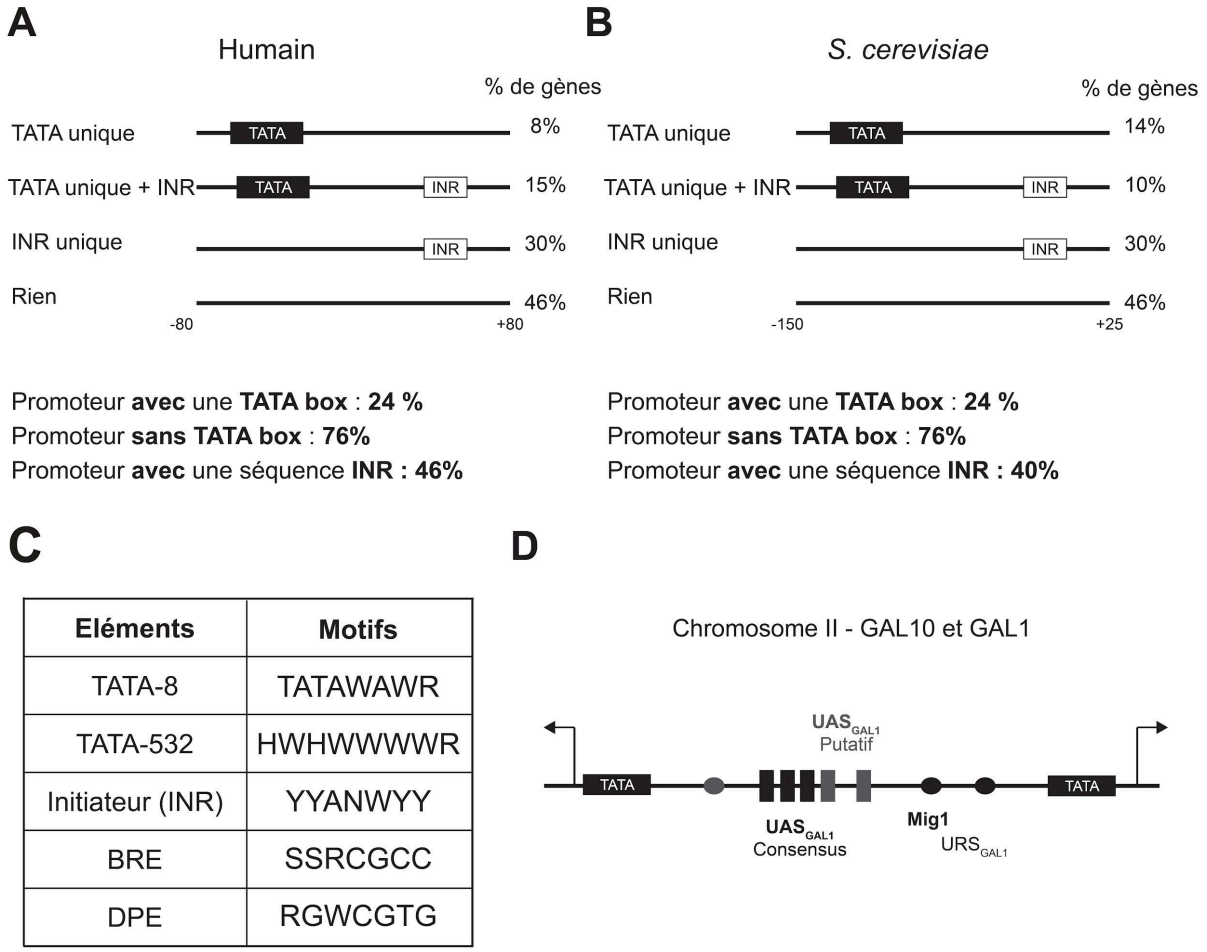
départ de transcription (TSS), en accord avec le fait que la plupart des promoteurs de mammifères sont situés dans des îlots CpG (Yang et al., 2007). Les îlots CpG sont de courtes régions avec une haute densité de dinucléotides CpG et un contenu en GC élevé, souvent situées près des promoteurs de gènes. Cela contraste fortement avec le contenu plus élevé en AT des régions intergéniques et promotrices de la levure (Goffeau et al., 1996). De plus, les promoteurs eucaryotes se composent d'éléments centraux, comme la TATA box, de séquences d'ADN qui définissent les sites de démarrage de la transcription et d'éléments régulateurs qui renforcent ou répriment la transcription d'une manière spécifique au gène. Le promoteur central, dont la longueur est généralement inférieure à 150 pdb, est défini comme nécessaire mais non suffisant pour conduire à l'expression du gène dans un contexte chromatinien in vivo.

Pour définir le promoteur central, on peut commencer par caractériser le site de départ de la transcription (TSS), qui correspond au début du transcrit, en mappant les cDNA (David et al., 2006; Z. Zhang & Dietrich, 2005). Il est majoritairement défini par le premier codon ATG et donne des indications primordiales sur l'emplacement probable des séquences codantes et non codantes. Chez la levure les sites de départ de transcription sont bien définis et permettent de caractériser les autres éléments du promoteur (David et al., 2006).

L'élément majeur de la séquence promotrice est la TATA box à laquelle la TBP peut se lier. En 2004 une séquence consensus de la TATAbox est clairement définie chez *S. cerevisiae* et permet de constater que six des huit lettres de la séquence consensus (TATAWAWR) servent d'éléments majeurs à sa définition et 20 % des gènes possèdent cette séquence (Basehoar et al., 2004). Cela suggère que l'initiation spécifique de la transcription au niveau de la plupart des promoteurs de levure pourrait reposer sur d'autres éléments du promoteur central. De plus, les gènes contenant une TATA box sont caractérisés par leur propension à être subtélomériques, à être exprimés à des niveaux extrêmement élevés ou faibles, induits par le stress et soumis à une pression sélective évolutive (Basehoar et al., 2004). En revanche, les promoteurs constitutifs sont souvent dépourvus de TATA box (Basehoar et al., 2004). En reproduisant l'analyse sur le génome humain, la grande majorité (~ 76%) des promoteurs centraux sont dépourvus de la séquence consensus TATA (Yang et al., 2007).

Un second élément des promoteurs eucaryotes est l'initiateur (INR), qui permet le recrutement de TFIID (Danino et al., 2015). Sa séquence consensus, définie chez les mammifères, est YYANWYY. Avec la même approche bio-informatique à l'échelle du génome, ~ 46 % des promoteurs centraux humains contiennent l'élément INR consensuel et ~ 30 % sont des gènes sans TATA box contenant l'élément INR. Il est surprenant de constater que des séquences INR de type mammifère sont présentes dans la région du TSS de ~ 40 % des promoteurs centraux de levure (Yang et al., 2007). Celà suggère que, tout comme la boîte TATA, l'élément INR est hautement conservé au cours de l'évolution et peut contrôler la transcription d'une fraction significative de gènes chez la plupart des

eucaryotes, de la levure à l'homme (Yang et al., 2007). En plus de ces deux éléments, la TATAbox et l'initiateur, d'autres éléments peuvent jouer un rôle dans la mise en place du PIC. Les motifs BRE (SSRCGCC) et DPE (RGWCGTG) sont identifiés chez les mammifères mais sont largement absents chez la levure (Yang et al., 2007). L'ensemble de ces analyses bio-informatiques permet de résumer la structure des séquences promotrices chez la levure et les mammifères (**Figure 10. A,B**).



Eléments	Motifs
TATA-8	TATAWAWR
TATA-532	HWHWWWWR
Initiateur (INR)	YYANWYY
BRE	SSRCGCC
DPE	RGWCGTG

Figure 10 : Structure du promoteur central des gènes eucaryotes.

A) Fréquences des différentes catégories de promoteurs centraux dans les gènes humains. gènes avec une séquence TATA unique; gènes avec une séquence TATA mais sans séquence INR dans la région -80 à +80 ; gènes ayant à la fois une séquence TATA et une séquence INR dans n'importe quelle orientation et espacement dans la région -80 à +80 ; gènes avec un élément INR mais pas de séquence TATA. Le nombre (No.) et le pourcentage (%) de gènes de chaque catégorie sont indiqués (Yang et al., 2007) **B)** Même description chez *S. cerevisiae* (Yang et al., 2007) **C)** Séquences consensuelles des éléments promoteurs centraux utilisés dans cette étude. Nomenclature IUPAC. **D)** Structure des promoteurs des gènes GAL10 et GAL1.

En plus du promoteur central, des éléments cis-régulateurs distaux jouent généralement un rôle dans le contrôle transcriptionnel à longue distance et ne sont pas nécessairement impliqués dans l'activation des gènes : ils ne font qu'augmenter ou renforcer un autre mode de régulation. Chez la levure, les éléments cis-régulateurs sont appelés Upstream Activating Sequence (UAS) et sont reconnus par des activateurs transcriptionnels tel que Gal4 (**Figure 10. D**). Ils facilitent l'assemblage du PIC, soit par contact direct avec les facteurs de transcription généraux, soit indirectement par l'intermédiaire de coactivateurs (Hampsey, 1998).

Il est désormais clair que les activateurs transcriptionnels se coordonnent avec la structure locale de la chromatine, recrutent la protéine de liaison TATA (TBP) sur les promoteurs, puis interagissent avec TFIID et les autres coactivateurs. Bien que les éléments INR et la TATA box interviennent dans la mise en place du PIC, la plus grande catégorie de promoteurs centraux humains (~ 46%) semble manquer à la fois de séquences TATA et INR (Yang et al., 2007). Quels éléments d'ADN recrutent et/ou positionnent la machinerie basale sur ces promoteurs ? En l'absence d'éléments promoteurs centraux, comment la machinerie de transcription est-elle positionnée ?

La structure de la chromatine, c'est-à-dire l'ADN et les nucléosomes, ajoute un niveau supplémentaire à la mise en place du PIC.

2.1.3 Organisation des nucléosomes au niveau du promoteur

Le nucléosome, l'unité de base de la fibre de chromatine, joue un rôle central dans la régulation des gènes. En 2004, seule la position génomique de quelques centaines de nucléosomes était connue. Le développement des puces à ADN et du séquençage à haut débit a permis de cartographier la position des nucléosomes à l'échelle du génome entier. La caractéristique la plus remarquable est le contraste entre la densité des nucléosomes dans les régions régulatrices et celle dans les séquences transcrites. Chez la levure, plus de 90 % des promoteurs sont très faiblement occupés par des nucléosomes (Mavrich et al., 2008; Yuan et al., 2005). Ces régions sont appelées régions appauvries en nucléosomes ou Nucleosome Depleted Regions (NDR, en anglais).

Chez la levure, la structure des nucléosomes autour des promoteurs des gènes montre une organisation caractéristique : une NDR, environ 150 pdb, près du site de départ de la transcription (TSS), entourée de rangées régulières de nucléosomes (**Figure 11. A**). La position des nucléosomes en aval (+1) et en amont (-1) de ces régions NDR est hautement régulée et liée respectivement à l'initiation de la transcription et à l'élongation (**Figure 11. A**). Cette organisation typique facilite l'accès des facteurs de transcription au promoteur, permettant ainsi une régulation fine de l'expression génique. Cette configuration stéréotypée des nucléosomes est, dans une certaine mesure, conservée

dans toutes les espèces eucaryotes (**Figure 11. B),** (Bai & Morozov, 2010; Jiang & Pugh, 2009; Mavrich et al., 2008).

Les promoteurs inductibles sont plus susceptibles de contenir une boîte TATA et, contrairement à la configuration canonique des nucléosomes, les régions situées immédiatement en amont du TSS sur ces promoteurs ont tendance à être riches en nucléosomes dans des conditions répressives. De cette manière, la boîte TATA et les TSS sont couverts par des nucléosomes, ce qui rend la boîte TATA inaccessible à la protéine de liaison TATA (TBP). C'est le cas de PHO5, de la séquence HO ou des gènes GAL (Bai & Morozov, 2010). Sur le promoteur GAL1-10, quatre sites de liaison de l'activateur Gal4 (5′-CGG-N11-CCG-3) sont situés dans un nucléosome instable, partiellement déroulé et lié par l'enzyme de remodelage des nucléosomes. Dans des conditions d'activation, les nucléosomes flanquant la séquence d'activation en amont sont rapidement retirés du promoteur et les gènes GAL sont activés (**Figure 11. C**). L'organisation des nucléosomes au niveau de l'UAS semble donc jouer un rôle crucial dans la régulation transcriptionnelle.

Le niveau d'expression transcriptionnel de certains promoteurs constitutifs (pour rappel, souvent dépourvu de TATA box) varie lorsqu'une certaine condition environnementale est modifiée. Cependant, en examinant le repositionnement global des nucléosomes avant et après la reprogrammation transcriptionnelle induite par le glucose ou par choc thermique, peu de changement dans l'organisation des nucléosomes sont observés alors que plus de la moitié de tous les gènes changent significativement d'expression (Pelechano & Steinmetz, 2013). Celà suggère que la position des NDR et le positionnement du nucléosome +1 ne sont pas suffisants pour favoriser l'initiation de la transcription. D'autres changements peuvent se produire au niveau des promoteurs, tels que les modifications d' histones, l'incorporation des variants d'histones ou d'autres éléments qui n'ont pas été caractérisés dans ces études.

Plusieurs questions clés demeurent concernant la dynamique des nucléosomes et leur influence sur la transcription (Bai & Morozov, 2010) :

1. Pourquoi certains promoteurs dépourvus de nucléosomes restent-ils réprimés au niveau de la transcription ?
2. Quels sont les liens de causalité entre le positionnement des nucléosomes et l'activation de la transcription ?
3. Comment le positionnement des nucléosomes est-il établi et maintenu in vivo ?
4. Existe-t-il une variation de cellule à cellule dans le positionnement des nucléosomes et comment cela affecte-t-il la variation de l'activité transcriptionnelle ?
5. Comment le positionnement des nucléosomes évolue-t-il entre les espèces ?

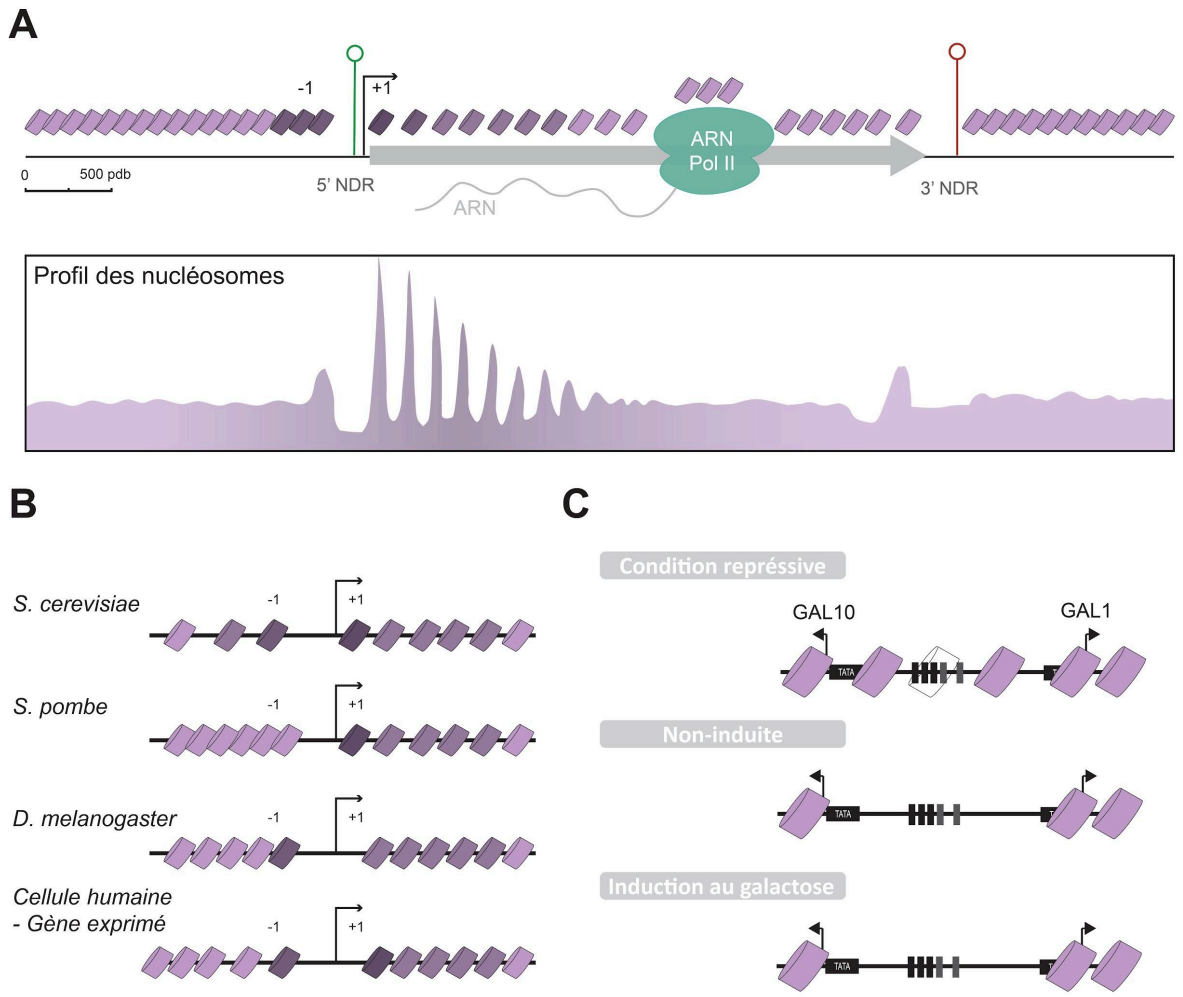


Figure 11 : Positionnement stéréotypé des nucléosomes à proximité des sites de départ de transcription.

A) Représentation schématique des positions typiques des nucléosomes in vivo chez *S. cerevisiae* Flèche : TSS. Cylindre violet : nucléosome. Plus le nucléosome est foncé, plus il est positionné précisément par rapport au TSS. Les nucléosomes gris qui se chevauchent représentent des nucléosomes sans phasage précis. **B)** Organisation des promoteurs de plusieurs espèces. Le nucléosome +1 est situé plus en aval chez la drosophile et l'homme que la levure. La longueur des répétitions nucléosomiques (distance moyenne entre les nucléosomes voisins) peut également varier d'une espèce à l'autre. **C)** Les régions promotrices des gènes GAL1 et GAL10 contient un complexe protéiques couvrant les quatre sites de liaison Gal4. Les nucléosomes sont rapidement éliminés lors de son activation. Les différentes figures sont adaptées de Bai & Morozov, 2010 et Jiang & Pugh, 2009.

2.1.4 Orientation de la transcription

Initiation de la transcription

Les données structurales disponibles montrent comment la Pol II coopère avec les facteurs généraux de transcription pour se lier à la séquence promotrice et l'ouvrir, pour diriger la synthèse de l'ARN et initier la transcription (Sainsbury et al., 2015). Le domaine carboxy-terminal (CTD) de la Pol II joue un rôle crucial dans cette initiation via des modifications post-traductionnelles. Le CTD peut être phosphorylé de manière différentielle au cours des différentes phases de l'élongation de la transcription. La phosphorylation du CTD et des facteurs généraux de transcription permettent la dissociation des co-activateurs, l'élongation de la pol II et le recrutement des enzymes liées à la maturation des ARN (Buratowski, 2009).

Bi-directionnalité des promoteurs

Des analyses transcriptionnelles du transcriptome de *S. cerevisiae*, sur puces à ADN ont révélé l'accumulation de transcrits courts nommés Stable Unannotated Transcripts (SUTs) (Z. Xu et al., 2009). Puis, des mutants rrp6Δ de levure ont mis en évidence l'accumulation d'ARNs nommés "CUTs" (cryptic unstable transcripts) révélant que plusieurs régions intergéniques supposées silencieuses sont en fait transcrites par l'ARN polymérase II (Wyers et al., 2005; Z. Xu et al., 2009). Cette transcription, dite pervasive, serait due à la bi-directionnalité des séquences promotrices.

En parallèle de la levure, la caractérisation d'ARN dans des cellules souches embryonnaires murines a mis en évidence la présence d'ARN de ~20 nucléotides situés près du TSS des gènes codant des protéines (Seila et al., 2008). Chez la levure et les mammifères, environ 80 % des promoteurs actifs présentent un arrangement bidirectionnel de l'initiation (Seila et al., 2008; Z. Xu et al., 2009). Par exemple, chez les cellules souches embryonnaires murines, des ARN de ~20 nucléotides situés près du TSS des gènes codant des protéines ont été identifiés (Seila et al., 2008). Des études en ChIP ont montré une accumulation bidirectionnelle des transcrits. En examinant l'organisation structurelle des PIC et leur spécificité à l'échelle génomique en utilisant ChIP-exo, les données révèlent que deux PIC, en orientation inversée, peuvent occuper les bordures flanquantes de régions sans nucléosomes (Rhee & Pugh, 2012).

Contrairement aux deux pics de la Pol II et de H3K4me3 entourant les TSS, H3K79me2, une marque chromatinienne trouvée sur les régions d'élongation de la Pol II , est uniquement enrichie dans la direction de la transcription "sens" (Seila et al., 2008). Ceci suggère l'existence d'un mécanisme qui discrimine entre la polymérase sens et antisens. Il est intéressant de noter que la Pol II est souvent bloquée dans les directions sens et antisens à environ +50 pdb (sens) et -250 pdb (antisens) par rapport au TSS. Ces deux endroits sont proches du bord des nucléosomes -1 et +1, respectivement. On peut

supposer que les nucléosomes jouent un rôle dans le sens de la transcription (Seila et al., 2008).

Pour étudier la nature des régions promotrices, une stratégie a été d'intégrer des séquences d'environ 150 kb de *K. lactis* et *D. hansenii* dans *S. cerevisiae*, en d'autres termes, intégrer des séquences qui n'ont pas évolué de la même manière dans une autre espèce (Jin et al., 2017). Les équipes de Struhl et Churchman ont ensuite utilisé le NET-seq (permettant le séquençage des transcrits natifs uniquement) afin de cartographier précisément les complexes Pol II lors de l'élongation. Premièrement, certains promoteurs exogènes ont été transcrits de manière bidirectionnelle alors que, dans l'espèce endogène, le promoteur est unidirectionnel. Deuxièmement, ces séquences exogènes ont vu apparaître des régions promotrices inattendues et aussi bidirectionnelles qui corrèlent avec les NDR (Jin et al., 2017). Ces résultats suggèrent fortement que 1) les régions promotrices sont intrinsèquement bidirectionnelle mais que 2) la transcription directionnelle implique la coévolution et la sélection de séquences d'ADN et de facteurs de transcription.

A quoi peut servir cette bi-directionnalité? La transcription antisens pourrait jouer un rôle dans le maintien d'une structure chromatinienne ouverte au niveau des promoteurs ou servir à la formation *de novo* de gènes.

2.2. Voies de dégradations des ARNs

2.2.1 Transcription pervasive

L'essor des puces à ADN et des techniques de séquençage de nouvelle génération a rendu possible l'exploration du transcriptome. La majorité du génome humain est transcrit (75%), même si seule une petite fraction est annotée en tant qu'ARN mature (Djebali et al., 2012). Chez la levure *S. cerevisiae*, 85% du génome est transcrit en milieu riche (David et al., 2006). Le paysage de la transcription chez les eucaryotes supérieurs est plus complexe que prévu, avec une forte proportion de transcrits provenant de régions intergéniques, auparavant considérées comme silencieuses. Ces résultats ont conduit à la notion de transcription pervasive, désignant le fait que la transcription n'est pas limitée à des caractéristiques fonctionnelles bien définies, contrairement aux gènes.

La majorité de cette transcription "pervasive" n'est pas détectée dans les analyses standard du transcriptome. À l'état stable, ces transcrits ne sont pas abondants, et ce sont les mutants impliqués dans l'exosome qui ont permis leur détection (les voies de dégradation nucléaire sont abordées dans la partie 2.2.2). Deux études ont caractérisé le transcriptome de souches mutées pour des composants d'exosomes et de TRAMP : l'analyse transcriptomique du génome a révélé l'existence de CUTs et d' ARNnc appelés transcrits stables non annotés (SUT) (Seila et al., 2008; Z. Xu et al., 2009). Les analyses transcriptomiques ont confirmé que la principale source des transcrits cryptiques provient des promoteurs bidirectionnels et ont indiqué que les CUT proviennent presque exclusivement de régions dépourvues de nucléosomes (Seila et al., 2008; Z. Xu et al., 2009).

Est-ce que ces transcrits indétectables représentent une activité véritablement fonctionnelle ou un "bruit" aléatoire et omniprésent ? Autrement dit, quel est le rôle de la transcription pervasive ? Il est suggéré que ~90% des événements d'initiation de la Pol II chez la levure représentent un bruit transcriptionnel et que la spécificité de l'initiation est comparable à celle des protéines de liaison à l'ADN et d'autres processus biologiques (Struhl, 2007). En introduisant des séquences de *K. lactis* dans *S. cerevisiae*, les séquences exogènes sont spontanément transcrites dès lors qu'une NDR existe. Les équipes de Strulh et Churchman suggèrent que, dans *S. cerevisiae*, un grand nombre des transcrits antisens non codants provenant de régions promotrices bidirectionnelles résultent des caractéristiques intrinsèques de la Pol II et n'auraient pas forcément d'utilité biologique (Jin et al., 2017).

Cependant, l'interférence transcriptionnelle médiée par les antisens serait un mécanisme plus fréquemment utilisé que prévu. En effet, lorsque l'expression de l'ARNm doit être étroitement réprimée dans des conditions spécifiques, l'interférence transcriptionnelle médiée par les antisens pourrait contribuer à la répression de 30 % des gènes les moins exprimés (Nevers et al., 2018). Cette

transcription pervasive aurait donc une utilité à court terme pour la régulation des gènes, avant d'être rapidement dégradée, ce qui la rend indétectable par les approches de RNA-seq.

En conclusion, la transcription pervasive découle en grande partie de la transcription bidirectionnelle de l'ARN polymérase Pol II (Seila et al., 2008; Z. Xu et al., 2009) qui, chez la levure, provient de deux PIC adjacents situés dans une NDR (Rhee & Pugh, 2012). Par conséquent, elle peut interférer avec d'autres événements physiologiques et doit être soigneusement régulée au niveau post-transcriptionnel.

2.2.2 Voies de dégradation

Le contrôle de la transcription pervasive s'effectue à deux niveaux : premièrement, la Pol II qui initie la transcription de manière erronée doit être rapidement interrompue, deuxièmement, les transcrits résultants doivent être efficacement dégradés pour éviter l'accumulation d'ARN non conformes. La caractérisation du transcriptome de mutants de voies de dégradation des ARN chez la levure a mis en évidence plusieurs classes d'ARNnc liés à des voies spécifiques de dégradation (Z. Xu et al., 2009). Ces voies jouent des rôles essentiels dans le noyau et dans le cytoplasme.

Nrd1-Nab3-Sen1 (NNS)

La voie nucléaire Nrd1-Nab3-Sen1 (NNS) est impliquée dans l'arrêt de la transcription et la dégradation des petits ARN nucléaires (snRNA), des petits ARN nucléolaires (snoRNA) ainsi que d'autres ARNnc non fonctionnels (Wyers et al., 2005; Z. Xu et al., 2009). La terminaison de ces précurseurs ARNnc non fonctionnels ne dépend pas de la machinerie de clivage et de polyadénylation utilisée pour la terminaison des ARNm.

Dans le cas de ces ARN cryptiques, le complexe contenant les protéines de liaison à l'ARN Nrd1, Nab3 et l'hélicase ARN Sen1 s'associe au CTD de la Pol II. La terminaison de la transcription se produit en aval des motifs tétranucléotidiques (GUAA/G, UCUUG) qui forment des sites de liaison avec Nrd1 et Nab3 sur l'ARN naissant. Le complexe interagit ensuite directement avec l'exosome nucléaire, couplant ainsi la terminaison de la transcription à la dégradation de l'ARNnc par le complexe TRAMP-exosome, comprenant notamment Rrp6 (**Figure 12**), (Jacquier, 2009).

Il a été proposé que la terminaison de la transcription antisens, dépendante de Nrd1, puisse générer la directionnalité des promoteurs chez *S. cerevisiae* (Jacquier, 2009). En déplétant Nrd1, Schulz et al. ont mis en évidence que la majorité des ARNnc issus des promoteurs bi-directionnel sont associés à Nrd1 et que ce facteur assure la directionnalité du promoteur en supprimant la transcription antisens (Schulz et al., 2013).

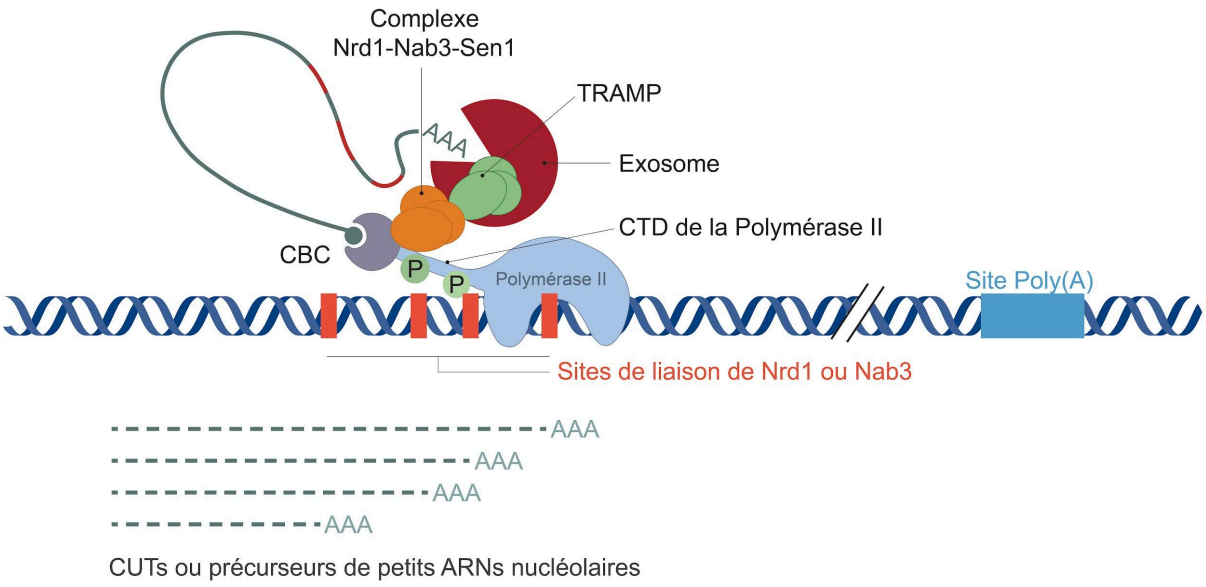


Figure 12 : Reconnaissance et dégradation des transcrits instables cryptiques (CUT) par la voie nucléaire NNS (Nrd1-Nab3-Sen1).

Au cours des premières centaines de nucléotides de la transcription (<1 kb), le domaine carboxy-terminal (CTD) de l'ARN polymérase II (RNAPII) est phosphorylé sur la sérine 5 (cercles verts marqués par P). Le complexe Nrd1-Nab3-Sen1 (orange) s'assemble sur le CTD de RNAPII après que des séquences tétranucléotidiques reconnues par Nrd1 (GUAR) ou Nab3 (UCUU) ont été transcrites. Ces séquences sont représentées par des boîtes rouges sur l'ADN et des sections rouges sur l'ARN (ligne grise incurvée). Ce complexe induit la terminaison de la transcription. Le complexe Nrd1-Nab3-Sen1 interagit également physiquement avec l'exosome nucléaire (représenté en rouge foncé), ce qui permet le couplage entre la terminaison de la transcription des CUTs et leur dégradation par l'exosome. Ce processus génère des transcrits de petite taille (<1 kb) (lignes grises en pointillés en bas de la figure), qui sont hétérogènes. Figure issue de Jacquier, 2009.

Non-mediated decay (NMD)

Décrite il y a plus de vingt ans comme une voie de contrôle qualité, la voie non-mediated decay (NMD) cible la dégradation des ARNm contenant un codon stop prématuré dans le cytoplasme (Kervestin & Jacobson, 2012). Elle reconnaît d'abord l'ARNm portant un codon de terminaison de traduction prématuré, puis déclenche la dégradation, empêchant ainsi la production de protéines tronquées potentiellement délétères (Maquat, 2004). En plus de ce rôle sur les ARN messagers, une proportion significative des transcrits issus de la transcription pervasive peut également échapper à la voie de dégradation nucléaire et être exportée vers le cytoplasme, où ils seront reconnus par la voie NMD (Malabat et al., 2015).

2.3. Régulation transcriptionnelle à longue distance

2.3.1 Régulation longue distance chez les mammifères

Importance de l'organisation 3D dans les intéractions promoteur-enhancer (P-E)

Les enhancers sont des séquences spécifiques permettant d'activer les promoteurs des gènes à longue distance en se rapprochant physiquement de leurs cibles. Chez les métazoaires, ces interactions peuvent s'étendre sur plusieurs mégabases (Mb) en termes de séparation génomique. L'organisation spatiale de plusieurs locus contenant les gènes et les éléments régulateurs associés ont été disséqués par des approches comme le Hi-C. On peut citer le domaine Shh (Paliou et al., 2019) ou le domaine Sox9 (Despang et al., 2019). Dans chaque cas, le TAD contient tous les enhancers connus jusqu'à 1 Mb du TSS et est identifiable dans divers types cellulaires.

Il a donc été proposé que la formation des boucles par la cohésine puisse réguler la régulation transcriptionnelle en rapprochant les enhancers distants des promoteurs (Kagey et al., 2010). La dégradation des cohésines, impliquées dans la formation des boucles en interphase chez les mammifères, peut affecter des programmes d'expression génique spécifiques en réponse à l'activation des macrophages (Cuartero et al., 2018) ainsi que l'expression des gènes clés nécessaires à la maturation des neurones corticaux chez la souris (Canzio et al., 2019). De plus, la modification des sites CTCF ou la modification des barrières des TADs peut entraîner des défauts de développement tels que des polydactylies (**Figure 13. A**), (Lupiáñez et al., 2015) ou des malformations des poumons chez les embryons de souris (Rajderkar et al., 2023). Ces mutations provoquent l'activation ectopique de gènes par des enhancers distants, ce qui indique que les frontières des TADs dépendantes de CTCF peuvent aussi isoler les enhanceurs des promoteurs (Lupiáñez et al., 2015; Rajderkar et al., 2023).

Plusieurs études soutiennent l'idée que l'extrusion de boucle joue un rôle dans le rapprochement P-E pour permettre leur activation (Kane et al., 2022; Zuin et al., 2022). L'une de ces études rapporte que la dégradation de la cohésine empêche l'activation ectopique du gène Shh dans les cellules embryonnaires de souris lorsque son enhancer est situé à plus de 400 kb (Kane et al., 2022). Dans l'étude de Zuin et al. en 2022, le transposon PiggyBac a été utilisé pour insérer l'enhancer de Sox2 à des endroits aléatoires. Cette méthode permet de générer des centaines de lignées cellulaires avec diverses intégrations de la séquence régulatrice. Ils ont ainsi observé que la variabilité de la transcription dépend de la distance génomique entre l'enhancer et le promoteur (Zuin et al., 2022).

Un tableau contrasté

Le rôle fonctionnel de l'extrusion de boucle et les contacts qu'elle génère pendant l'interphase restent sujets à débat. C'est un mécanisme attrayant pour la médiation des interactions P-E, car elle facilite les contacts et permet de contrôler la direction et l'étendue de ces contacts via CTCF et d'autres barrières d'extrusion. Cependant, et de manière étonnante, la dégradation des cohésines et des CTCF a peu d'impact sur la transcription des gènes dans les cellules de mammifères (Nora et al., 2017; Rao et al., 2017; Wutz et al., 2017). Une autre étude, par Micro-C, sur des cellules embryonnaires de souris a conclu que les facteurs principaux de l'extrusion de boucle (CTCF, cohésine ou WAPL) ne sont pas nécessaires au maintien à court terme de la plupart des interactions E-P (T.-H. S. Hsieh et al., 2022).

De plus, les données d'imagerie montrent une image dynamique : en marquant des sites précis de l'ADN dans des cellules souches embryonnaires de souris, notamment deux sites CTCF, Gabriele et al. ont pu mesurer la fréquence de contact entre ces deux éléments (Gabriele et al., 2022). Ils ont constaté que ce contact (CTCF-CTCF) n'est formé que 6 % du temps (Gabriele et al., 2022). Brückner et al. ont étudié l'activité transcriptionnelle simultanément à la visualisation par microscopie d'une paire de P-E plus ou moins séparés le long du chromosome. Ils ont ainsi observé les mêmes caractéristiques que Gabriele et al. et que le temps de rencontre P-E dépend peu de la séparation génomique (Brückner et al., 2023). Un autre exemple de régulation génique montre qu'une diminution de la proximité entre l'enhancer et le promoteur peut même être associée à l'activation de l'enhancer, ce qui suggère que le mécanisme de régulation transcriptionnelle par l'enhancer est très complexe et pluriel (Benabdallah et al., 2019).

Pour expliquer ces résultats parfois contradictoires, de nouveaux modèles suggèrent que les promoteurs peuvent mémoriser les informations émises par les enhancers sous forme d'une empreinte moléculaire. Comme la communication stable P-E ne peut pas être réalisée par les cohésines, pour des raisons d'incompatibilité de dynamique, on pourrait envisager que l'initiation de la communication P-E soit faite par les cohésines puis prise en charge par d'autres facteurs (**Figure 13. B**). La latence qui en découle découplerait la proximité enhancer-promoteur de l'activité transcriptionnelle (**Figure 13. B**) (de Wit & Nora, 2023).

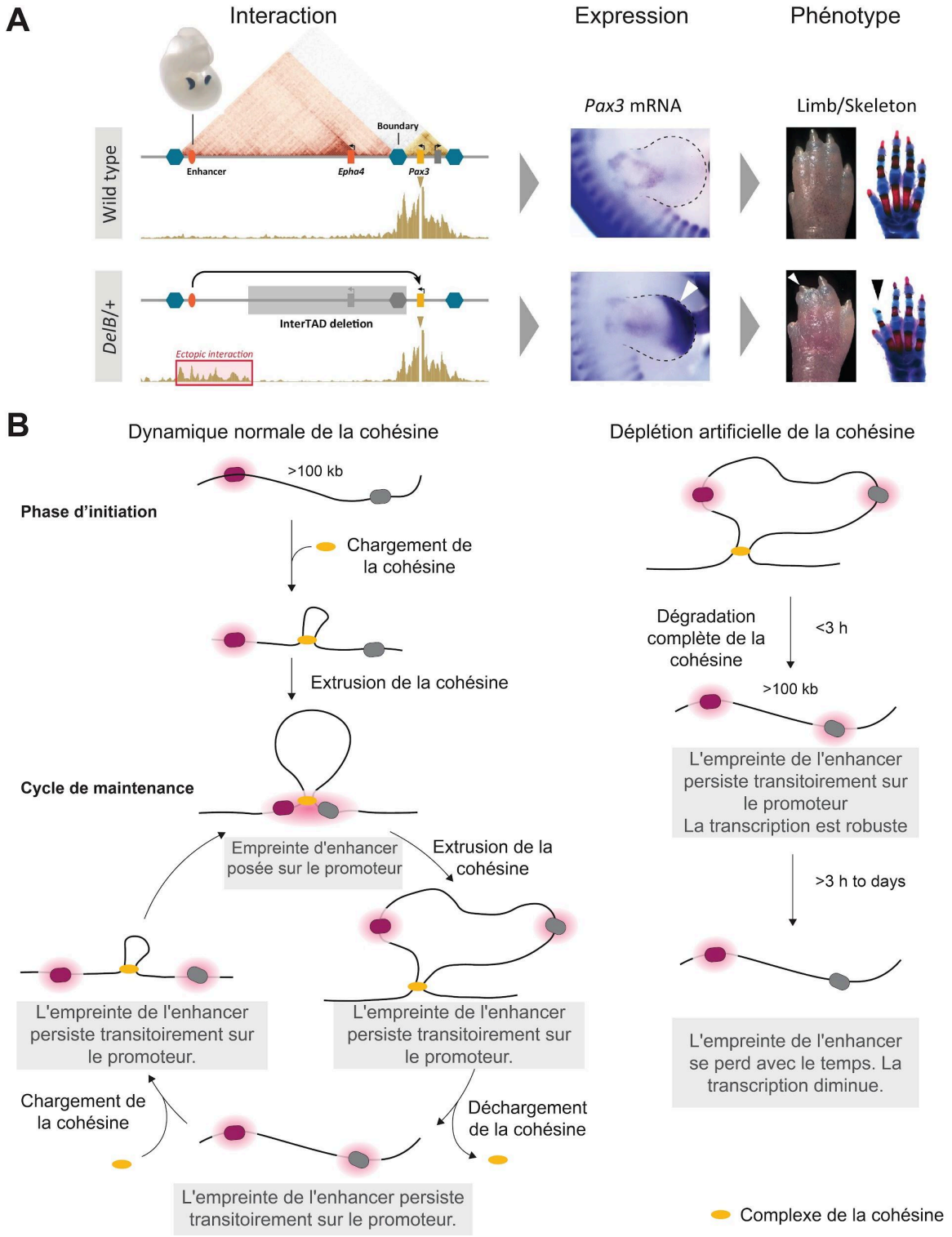


Figure 13 : Rôle de l'organisation 3D dans la régulation à longue distance chez les mammifères.

A) Une dérégulation des frontières des TADs conduit à une activation non spécifique des gènes. Ces dérégulations peuvent entraîner des anomalies de développement (Lupianez et al., 2015). **B)** Les promoteurs enregistrent les interactions avec les enhancers via une empreinte moléculaire, augmentant l'activité transcriptionnelle. Si cette empreinte persiste plus longtemps que l'association enhancer-promoteur, un décalage temporel amortit la transcription contre les variations de l'organisation 3D du génome, expliquant la robustesse transitoire face à la dégradation de la cohésine. Figure issue de (de Wit & Nora, 2023).

2.3.2 Régulation en *cis* chez *S. cerevisiae*

De nombreux aspects de l'initiation de la transcription sont conservés entre *S. cerevisiae* et les autres eucaryotes. Outre son génome dense, une autre différence importante entre la levure et les mammifères, ainsi que d'autres métazoaires comme la drosophile, concerne les séquences d'activation en amont (UAS) et les enhancers. Les deux servent de sites de liaison pour les activateurs spécifiques des gènes, mais tandis que les enhancers sont situés à plusieurs kb du promoteur, les UAS sont généralement positionnés à quelques centaines de paires de bases du promoteur central. Pour tester la distance maximale entre l'UAS et le promoteur, Dobi et Winston 2007 ont placé la séquence UAS à différentes distances du gène rapporteur HIS sous contrôle du core promoteur de GAL1. A une distance de 670 pdb, le rapporteur n'est plus transcrit. On peut supposer que, de par son génome dense, la levure limite la régulation longue distance pour limiter une induction transcriptionnelle non souhaitée (Dobi & Winston, 2007). La régulation transcriptionnelle en *cis* serait donc de courte portée, ce qui pourrait être corrélé à l'absence de boucles en interphase.

La régulation transcriptionnelle à longue distance chez la levure est-elle envisageable ? Plusieurs études ont démontré que, dans certaines circonstances, la formation de boucles se produit chez *S. cerevisiae*. En effet, des boucles ont été détectées avec le 3C entre les promoteurs et les terminateurs de longs gènes, et sont corrélées à une transcription active (O'Sullivan et al., 2004). De plus, la régulation artificielle de gènes distants a déjà été démontrée chez la levure. Dans un premier cas, De Bruin et al. (2001), ont tiré parti de l'organisation des télomères pour rapprocher spatialement l'UAS et le promoteur central pour exprimer le gène rapporteur situé dans cette région (de Bruin et al., 2001). La même construction insérée dans une autre région du génome n'entraîne pas l'expression du rapporteur. Enfin, avec des systèmes de rapprochement artificiels Petrascheck et al. (2005) ont forcé l'interaction entre l'UAS et le promoteur (Petrascheck et al., 2005). Dans ces exemples, les intéractions sont réalisées à faible distance (< 5 kb) et sans tester la contribution moléculaire des facteurs SMC.

Les méthodes 3C ont aussi permis de détecter de nombreuses interactions chromosomiques à longue distance dans le génome de la levure. Par exemple, l'élément RE (Recombination Enhancer) impliqué dans le changement de mating type. Sa délétion impacte l'organisation 3D et limite l'interaction de la séquence HMR avec la séquence HO (Belton et al., 2015). Dans un autre exemple, après un choc thermique de 5 minutes, les gènes de protéines de choc thermique (HSP) dans la levure se regroupent, bien que certains gènes comme UBI4, HSP104, et SSA soient situés à plus de 25 kb et 33 kb (Chowdhary et al., 2019). Ainsi, même dans le génome dense de la levure, des éléments sont impliqués dans la régulation longue-distance mais la fonction de ces interactions dans la régulation de l'expression des gènes n'est pas encore claire.

3. Apport de la génomique synthétique

Grâce au séquençage, nous sommes capables de “lire” l’information contenue dans des milliers de génomes avec une précision sans précédent. En parallèle, une autre technologie permet maintenant d’exploiter cette information de manière originale, ouvrant de nombreuses perspectives de recherche fondamentale ou d’applications. Celle-ci consiste en la synthèse d’ADN, et en particulier de la synthèse de molécules de longue taille (plusieurs dizaines de kb). Cette branche de la biologie synthétique a donné l'essor à la génomique synthétique, un domaine interdisciplinaire combinant les principes biologiques, l’ingénierie des génomes et l'informatique pour concevoir et construire de nouvelles fonctions et systèmes biologiques. De nombreuses équipes travaillant dans ce domaine affichent pour objectifs principaux la reprogrammation d’organismes pour synthétiser des produits d’intérêts ou acquérir de nouvelles fonctions. Cependant, cette approche peut également être mobilisée pour explorer des processus chromosomiques et des fonctions biologiques de manière originale, et faire ainsi progresser la connaissance fondamentale du vivant. La génomique synthétique se concentre sur l’ingénierie génétique, la construction de chromosomes et de génomes. Actuellement, les progrès de la génomique synthétique sont limités par la capacité de synthèse des molécules d'ADN rendant nécessaire le développement de méthodes pour assembler de petits fragments d'ADN en grandes molécules. Ainsi, *S. cerevisiae* a joué un rôle essentiel dans le développement de la génomique synthétique grâce à ses capacités de recombinaison homologue permettant d’assembler des molécules d'ADN. Ayant contribué à l'assemblage de génomes entiers ou partiels de divers organismes au cours de la dernière décennie, *S. cerevisiae* joue un rôle clé dans les développements futurs de la génomique synthétique.

Dans ce dernier chapitre, nous retraçons l’évolution de la génomique synthétique depuis la première synthèse d’un gène en 1970, en passant par la première construction d’un génome viral en 2000, d’une bactérie complète en 2010, jusqu'à la synthèse de génomes eucaryotes. L'avènement des néo-chromosomes ouvre de nombreuses perspectives dans ce domaine, et leur caractérisation joue un rôle novateur dans l’étude des processus biologiques sous-jacents.

3.1. Ingénierie génétique

3.1.1 Synthèse contrôlée de fragments d'ADN

En essayant de disséquer la relation entre la structure et la fonction de l'ARNt d'Alanine, l'équipe du professeur Har Gobind Khorana a introduit la synthèse chimique de molécules d'ADN double brin. Leurs efforts se sont concentrés sur la synthèse chimique de 17 oligonucléotides, codant le gène d'un ARNt d'alanine de 77 nucléotides (Khorana et al., 1972). Grâce à la synthèse d'ADN, ils ont pu réaliser des substitutions, des suppressions et des ajouts délibérés de bases. En 1976, deux équipes ont synthétisé les 21 pdb de l'opérateur LacO et l'ont cloné dans un plasmide. Ils ont ainsi caractérisé in vitro et in vivo le rôle de cette séquence (Heyneker et al., 1976). Sept ans plus tard, le premier gène synthétique codant la somatostatine, une hormone peptidique de 14 acides aminés, a été exprimé dans *Escherichia coli* (Itakura et al., 1977). Ce gène, synthétisé chimiquement, a été intégré au gène de la β-galactosidase sur le plasmide pBR322. La transformation dans *E. coli* a produit un polypeptide incluant la somatostatine. Ces travaux représentent le premier succès dans l'expression d'un gène synthétisé chimiquement.

Pendant quatre décennies, des méthodes chimiques ont été développées pour synthétiser de courtes chaînes d'ADN de moins de 200 nucléotides, appelées oligonucléotides. L'introduction des premiers synthétiseurs d'ADN automatisés a facilité la production d'oligonucléotides la rendant moins coûteuse, plus rapide et plus fiable mais n'a pas permis d'augmenter la longueur des oligos (Hughes & Ellington, 2017). Pour pallier ce problème, Stemmer et al. proposent d'utiliser la PCR comme méthode de synthèse de longues séquences d'ADN à partir d'un grand nombre d'oligos (Stemmer et al., 1995). En une seule étape, ils synthétisent un plasmide de 3000 pdb à partir d'un grand nombre d'oligos de 40 nt. Aujourd'hui, la plupart des fournisseurs de synthèse d'ADN peuvent fournir des séquences allant jusqu'à environ 5 kb (0.8€/bp, Twist) mais propose aussi la synthèse de fragments de 200 kb à des prix plus élevés (0.45€/bp, GenScript).

Bien que le coût du séquençage ait chuté drastiquement au fil du temps grâce aux technologies de séquençage de nouvelle génération (NGS) capables de générer environ 15 pétabases de données de séquences par an dans le monde, le coût de la synthèse des gènes et des oligonucléotides en général n'a pas suivi (**Figure 14)**, (Hoose et al., 2023). Cependant, les capacités de synthèse d'ADN restent impressionnantes et offrent des possibilités d'ingénierie génétique équivalentes.

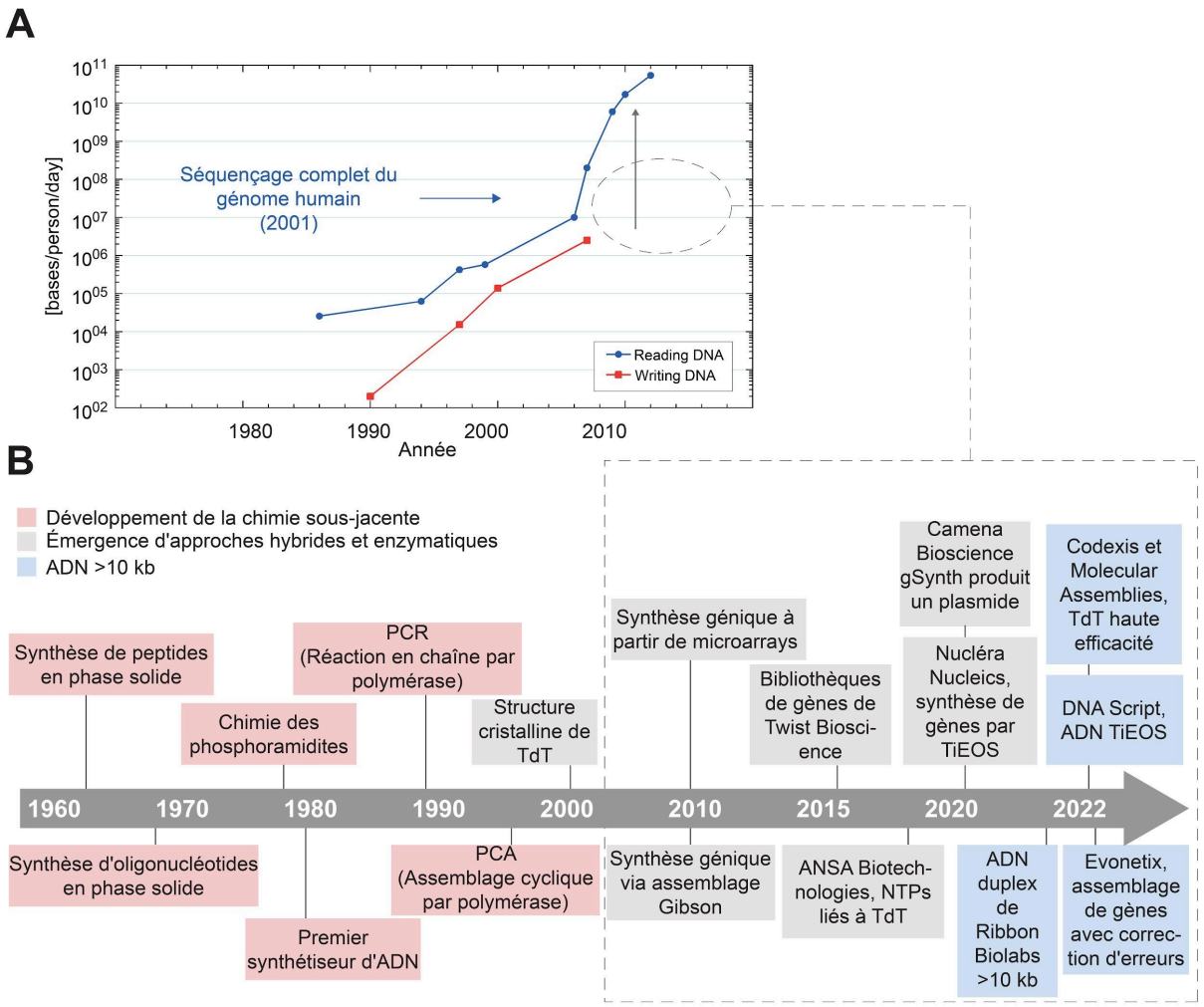


Figure 14 : Etat de l'art de la synthèse de l'ADN.

A) Productivité estimée de la lecture (séquençage) et de l'écriture de l'ADN (synthèse), mesurée en nombre de nucléotides en fonction du temps. La flèche grise indique l'écart actuel de productivité entre la lecture et l'écriture de l'ADN. L'ovale en pointillés souligne la période où l'industrie de la synthèse de l'ADN a franchi la plupart des étapes importantes pour réduire cet écart. Figure issue de (Hoose et al., 2023) **B)** Chronologie des étapes importantes des technologies de synthèse de l'ADN. Détaillée dans (Hoose et al., 2023).

3.1.2 Standardisation et optimisation des fragments synthétiques

Les progrès réalisés au cours des dernières décennies dans la lecture (séquençage) et l'écriture (synthèse) des séquences d'ADN ont considérablement amélioré notre capacité à comprendre et à concevoir des systèmes biologiques. Ainsi, malgré les limites évoqués sur la longueur des fragments d'ADN synthétisés, on peut désormais commander des séquences précises d'ADN ou des pools aléatoires d'oligos-nucléotides (<200 pdb) et ainsi tester de nombreuses combinaisons.

Utilisation de pools d'oligonucléotides pour le criblage

La possibilité de commander des pools de plusieurs millions d'oligos permet de tester rapidement un composant d'intérêt. Les éléments promoteurs sont des éléments indispensables de la biologie synthétique et, sans surprise, ont été parmi les premiers à être annotés et développés pour de nouveaux hôtes. Grâce à une technique de criblage basée sur une librairies de pools d'oligos entre 10 et 30 pdb, neuf séquences promotrices robustes et minimales sont identifiées à partir d'un pool de 15 millions de candidats (Redden & Alper, 2015). Les promoteurs de levure nécessitent des élèments cis-régulateur en plus du promoteur central pour atteindre une capacité de transcription élevée. Ici, ils identifient des UAS minimaux (10 pdb) qui peuvent être assemblés de manière hybride avec ces éléments centraux minimaux pour établir des promoteurs courts (120 pdb) pouvant être aussi puissants que le promoteur GPD (TDH3), long de 655 pdb. Remarquablement, cette étude met aussi en évidence que 30 pdb pourrait être l'espacement minimal requis entre la boîte TATA et le TSS pour *S. cerevisiae* (Redden & Alper, 2015). 5 ans après, le laboratoire de Aviv Regev mesure l'expression de >100 millions de séquences de promoteurs synthétiques de levure qui sont entièrement aléatoires 80 pdb (de Boer et al., 2020), et testent ces constructions dans des conditions de croissance en milieu riche avec différentes sources de carbone. Ces deux exemples illustrent le potentiel de la synthèse aléatoire pour optimiser et minimiser des composants génétiques spécifiques mais contribuent aussi aux approches fondamentales.

Normalisation et standardisation des "composants génétiques"

En plus de l'optimisation des séquences, la normalisation des parties génétiques est devenue un sujet d'intérêt croissant au cours des dernières décennies. L'objectif est de simplifier les procédures de clonage moléculaire, tout en les rendant plus reproductibles. Ces contraintes ont conduit à la conception de plusieurs normes comme les Biobricks sur iGEM, le clonage modulaire (MoClo) et la mise en commun de plasmides (Addgene). Ces outils optimisent le processus de clonage et d'assemblage. Le MoClo, en particulier, est un système qui définit différents composants comme "promoteur", "terminateur", "gène" et permet de les moduler. Le MoClo est basé sur le Golden Gate : c'est grâce aux enzymes de restriction de type II qui permettent d'orienter les fragments sans laisser de cicatrices. La standardisation des composants et la mise en place de niveau de clonage rendent le système modulaire (Otto et al., 2021; Shaw et al., 2023).

Le développement de la synthèse d'oligonucléotides, de la méthode PCR et les améliorations (normalisation, minimisation) permettent à la biologie synthétique des avancées importantes. La synthèse de courts fragments d'ADN est maintenant bien maîtrisée, mais la création de longs

fragments reste cher et complexe dû aux séquences répétées et/ou aux structures secondaires (Hoose et al., 2023).

3.1.3 Synthèse chimique de génomes complets

A plus grande échelle qu'un gène ou un promoteur, la synthèse de génome complet permet 1) d'étudier des organismes qu'on ne peut pas faire pousser en laboratoire et 2) modifier spécifiquement une partie du génome pour en étudier les processus biologiques associés.

Synthèse de génomes viraux

En 2002, la première synthèse *de novo* d'un génome fonctionnel, celui du poliovirus, est réalisée (Cello et al., 2002). La séquence, la carte génétique et la structure cristalline tridimensionnelle du virion avaient été déterminées vingt ans auparavant. Les segments du génome, totalisant 7741 bases, ont été synthétisés en assemblant des oligonucléotides purifiés d'une longueur moyenne de 69 nucléotides. Ces travaux pionniers ont démontré la faisabilité de la synthèse chimique complète d'un génome viral fonctionnel, ouvrant la voie à de nouvelles possibilités dans la biologie synthétique et la virologie (Cello et al., 2002). En 2003, Smith et al., assemblèrent le génome du bactériophage φ X174 (5386 pdb) (Smith et al., 2003) et en 2005, Chan et al. ont partiellement réécrit les 39'937 pdb du génome du bactériophage T7 (Chan et al., 2005).

Synthèse de génomes bactériens

Ces travaux ont naturellement conduit à la synthèse de génomes plus grands. En 2008, une équipe de J. Craig Venter ont assemblé de novo, à l'identique ou presque, le premier plus petit génome bactérien connu : *Mycoplasma genitalium* (583 kb) (Gibson et al., 2008). D'abord, des « cassettes » superposées de 5 à 7 kb sont assemblées à partir d'oligonucléotides synthétisés chimiquement (**Figure 15. A**). Ensuite, elles sont jointes par recombinaison in vitro pour produire des assemblages intermédiaires d'environ 24 kb, 72 kb (« 1/8 du génome ») et 144 kb (« 1/4 du génome ») **(Figure 15. B**). Ces trois premières étapes sont clonées in vitro puis clonées dans *E. coli* (**Figure 15. B**). Cependant, avec cette stratégie, l'équipe n'obtient pas de clones quand ils essaient de cloner les 2 fragments dans *E. coli* (« 1/2 du génome ») (Gibson et al., 2008). Pour surmonter les limites technologiques d'assemblage in vitro de long fragments d'ADN, Gibson et al prennent avantage de la recombinaison homologue de la levure pour assembler les derniers fragments du génome synthétique (**Figure 15. C**) (Gibson et al., 2008). Les derniers fragments servent de matrice à la machinerie de

recombinaison homologue qui les rassemble en un nouveau chromosome (**Figure 15. C**). Ainsi, le génome synthétique complet de *Mycoplasma genitalium*, de 582 970 pdb est cultivé de manière stable sous forme de plasmide centromérique de levure. Ici, les auteur.es ont d'abord essayé de construire le génome in vitro et de le cloner dans *E. coli*. C'est suite aux limitations qu'ils se sont tourné vers la levure pour finaliser l'assemblage des deux derniers fragments. Ils posent d'ailleurs la question : peut-on directement faire l'assemblage complet dans la levure et combien de fragments peuvent être assemblés en une étape ?

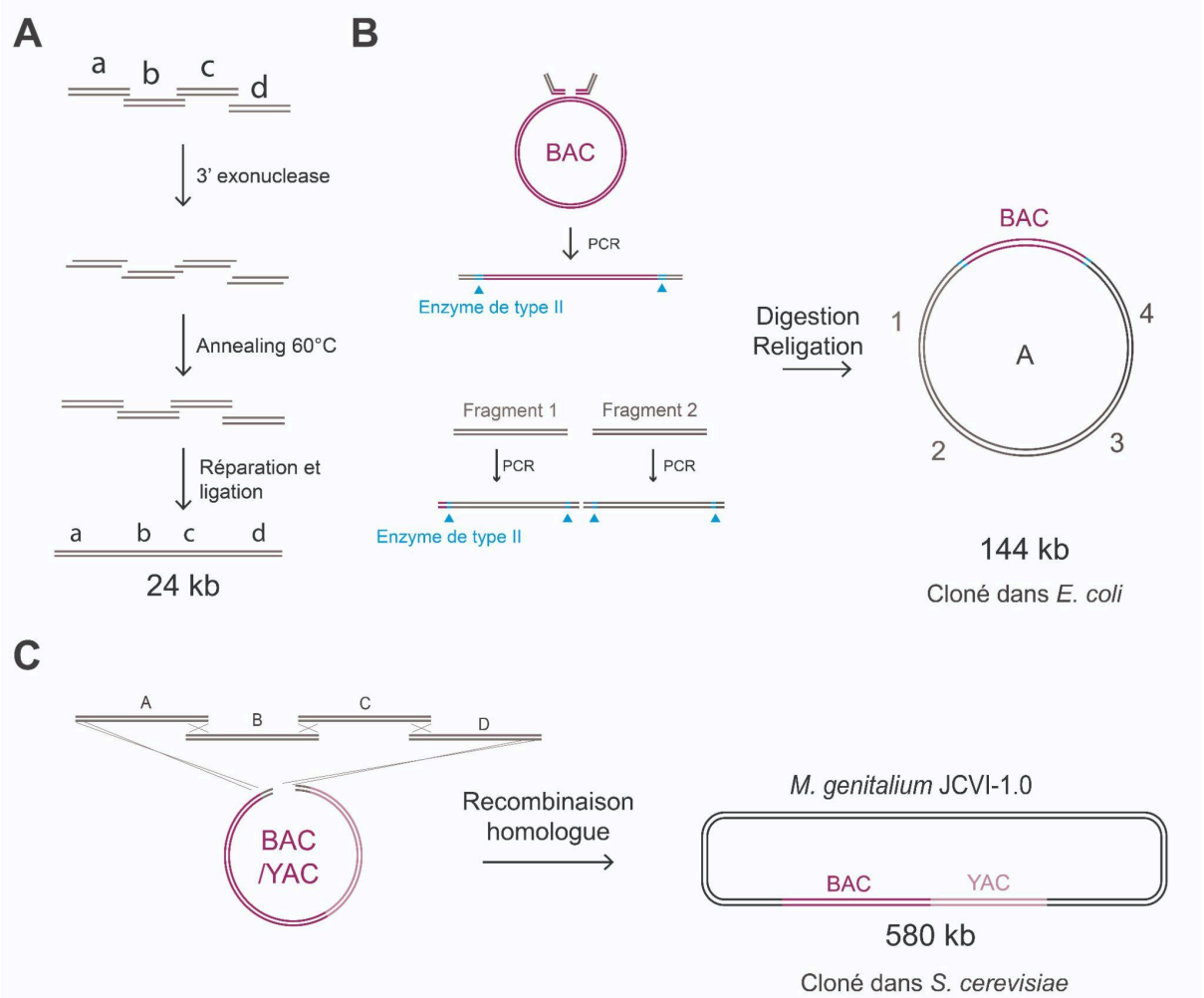


Figure 15 : **Assemblage du génome de *Mycoplasma genitalium* par recombinaison in vitro et in vivo**.
A) Schéma des étapes de la réaction de recombinaison in vitro des oligonucléotides **B)** Le vecteur BAC est linéarisé par PCR à l'aide des amorces illustrées. Après purification sur gel, le vecteur est inclus dans la réaction d'assemblage avec les fragments issus de A), de sorte que l'assemblage souhaité est un ADN circulaire contenant les quatre cassettes et l'ADN BAC. Les différents BAC sont ensuite clonés dans *E. coli* **C)** Les ¼ de génome de *M. genitalium* JCVI-1.0 ont été purifiés à partir d'*E. coli*, digérés par NotI, et co-transformé avec un vecteur BAC/YAC dans *S. cerevisiae.* Le vecteur BAC/YAC contient à la fois des séquences BAC (en rose foncée) et YAC (en rose clair). Une origine de réplication de la levure (ARS) et un marqueur de sélection (HIS) permettent la propagation et la sélection du plasmide dans la levure.

3.2. *S. cerevisiae*, une plateforme d'assemblage de novo et d'édition de génome

3.2.1 Edition de génome

Le modèle *S. cerevisiae* a naturellement été pionnier dans l'édition et l'introduction de fragments et est devenu un outil de choix pour la biologie synthétique (T. A. Dixon et al., 2023). Une caractéristique essentielle de *S. cerevisiae* est sa préférence pour la recombinaison homologue afin de réparer les cassures double brin de l'ADN (Kunes et al., 1985). La recombinaison homologue de l'ADN chez la levure est très efficace et cet organisme n'a besoin que de 20 pdb de séquence d'ADN homologue à proximité des extrémités d'un fragment d'ADN pour être intégré avec précision dans le génome (Hua et al., 1997).

Le développement du système CRISPR-Cas9 (Clustered Regularly Interspaced Short Palindromic Repeats- associated protein 9) a stimulé le domaine de l'ingénierie des génomes et permet de générer des modifications génétiques sans marqueur de sélection (Doudna & Charpentier, 2014). Dirigée par un ARN guide unique spécifique d'une séquence (ARNg), la nucléase Cas9 crée une cassure double brin de l'ADN. L'ARNg peut être facilement conçu pour cibler n'importe quel locus génomique situé à proximité d'un motif adjacent à la séquence PAM : "NGG" (**Figure 16. A**). Cette technologie est aujourd'hui couramment utilisée pour induire des cassures double brin ciblées dans les génomes d'une grande variété d'espèces. Cette cassure peut être réparée par recombinaison homologue avec un ADN donneur cotransformé contenant des polymorphismes empêchant la liaison et le clivage ultérieur par la Cas9 afin d'obtenir une édition réussie (**Figure 16. C**). Par rapport aux cellules de mammifères, *S. cerevisiae* réalise la jonction d'extrémités non homologues (NHEJ) avec une grande fidélité, le principal produit de réparation étant une simple religation. Par conséquent, l'ADN réparé reforme le site de clivage original de Cas9, à moins que des erreurs n'empêchent la reconnaissance ultérieure de la Cas9. On entre dans un cycle de clivage/religation qui se termine soit par l'introduction d'une erreur, ce qui est plutôt rare, soit à l'engagement vers la recombinaison homologue avec l'initiation de la résection qui n'aboutit pas en absence de séquences homologues et entraîne la mort cellulaire par résection de gènes essentiels (**Figure 16. B**).

Cette technologie a été optimisée pour la levure, en facilitant le clonage de l'ARNg (DiCarlo et al., 2013; Laughery et al., 2015) et permettant d'introduire une cassette à un locus précis ou induire une ou plusieurs translocations ciblées avec une précision d'un seul nucléotide (Fleiss et al., 2019). En combinant la standardisation des composants et la puissance de CRISPR-Cas9, les méthodes de clonage dans la levure sont désormais simples et rapides. Par exemple, le Modular Cloning (MoClo), évoqué précédemment, permet de cloner, en un seul coup, des plasmides CRISPR-Cas9 en variant les

marqueurs de sélection. En plus, des équipes ont aussi désigné des plasmides composés de deux séquences homologues de 500 pdb permettant l'intégration de séquences d'intérêt au niveau du site reconnue par l'ARNg. Ainsi, il est possible de modifier le plasmide en ajoutant différents composants (promoteurs, séquence codante, terminateurs) dans le plasmide d'intégration avant de le co-transformer avec le plasmide CRISPR-Cas9 (Otto et al., 2021; Shaw et al., 2023).

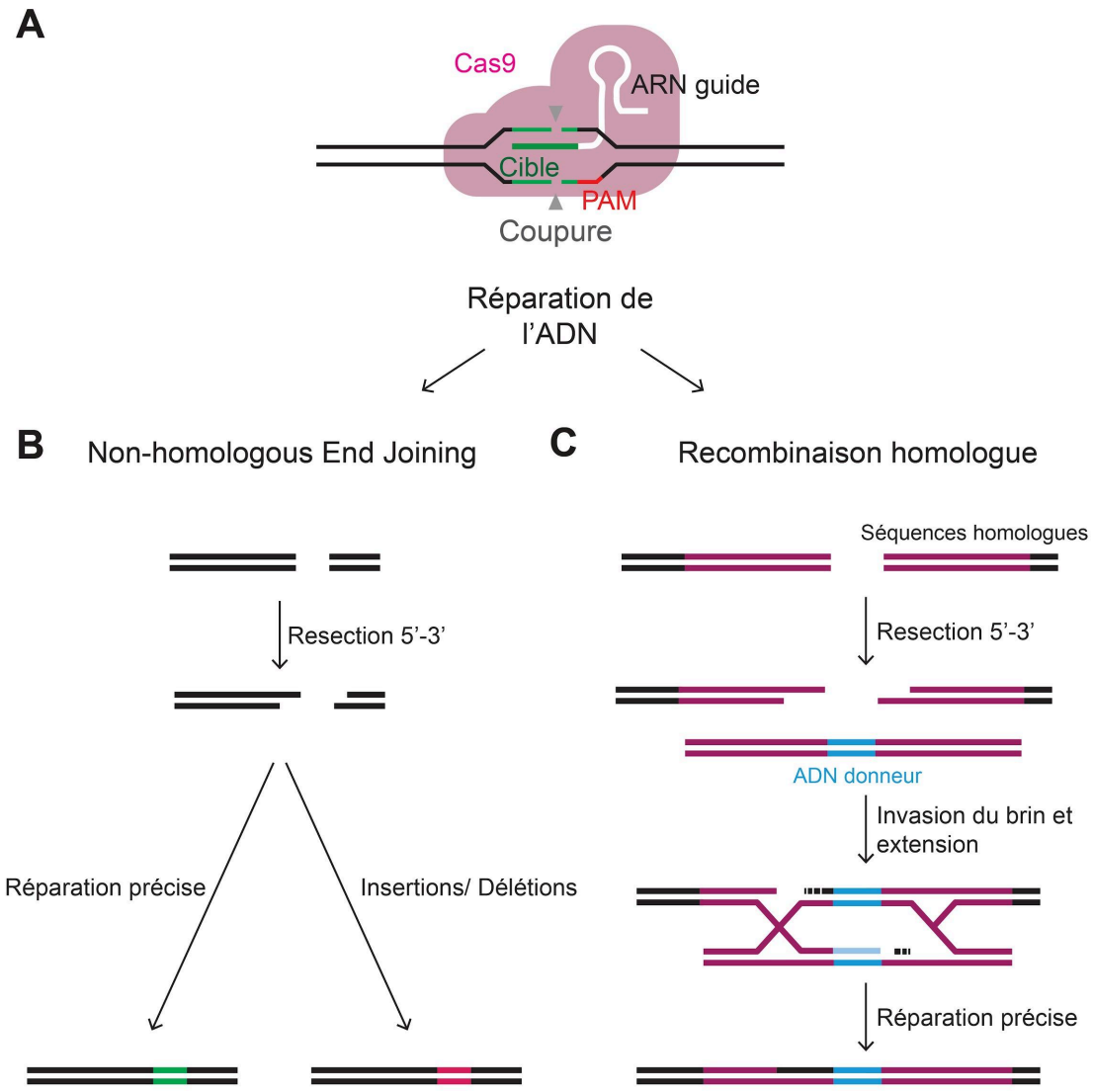


Figure 16 : Schéma du mécanisme CRISPR-Cas9 chez la levure.

A) L'ARNg unique (ARNg) contient deux composants : la partie structurelle de l'ARNg (blanc) et un ARN de 20 nt (vert) permettant de cibler la séquence d'ADN. La protéine Cas9 se lie à l'ARNg, qui sera acheminé aux sites cibles. La Cas 9 induit une cassure double brin au niveau du site cible, 3 pdb en amont du site PAM (triangle gris, le site PAM est en rouge). Une fois qu'une cassure double brin est induite, il y a deux choix de réparation. **B)** La voie NHEJ va efficacement réparer mais être reconnue par le système Cas9 jusqu'à l'insertion de délétions, substitutions ou insertions. **C)** La réparation homologue répare la cassure double brin à l'aide de l'ADN donneur. L'ADN donneur contient des bras d'homologie 5' et 3' et une construction d'intérêt. Après la recombinaison homologue, la DSB est réparée et l'ADN donneur intégré précisément.

L'ère de la biologie synthétique et de l'automatisation exige que les outils d'édition du génome soient multiplexables et efficaces, simples à utiliser et peu coûteux (Malcı et al., 2020). L'optimisation de l'édition multiple de génome, en particulier avec CRISPR-Cas9, joue un rôle important dans ce domaine. Les stratégies d'édition multiples peuvent être basées sur les caractéristiques existantes de la levure comme les éléments mobiles (transposons) ou le rDNA. Chez *S. cerevisiae*, les éléments mobiles ou transposons peuvent dépasser 150 copies, dispersées sur différents chromosomes. Par exemple, les transposons Ty sont composés de deux séquences codantes flanqué par des séquences répétées appelées δ. Ainsi, le ciblage de ces séquences, disséminées dans le génome, est une méthode efficace pour l'intégration multiple de fragments d'intérêts. Couplée à CRISPR-Cas9, le plasmide appelé Di-CRISPR permet cible les séquences δ. Cette approche permet l'intégration simultanée de 18 copies d'une voie de production de (R,R)-2,3-butanediol (Shi et al., 2016). L'intégration de 18 copies de la cassette de 24 kb a été réalisée en une seule étape, même si l'efficacité diminue lorsque la taille de la cassette passe de 8 à 24 kb (Shi et al., 2016). Une autre approche, connue sous le nom de "CRISPR/Transposon gene integration" (CRITGI) cible directement les sites de la séquence du rétrotransposon Ty1 plutôt que les séquences δ, une séquence plus conservée parmis les 31 Ty1 (Hanasaki & Masumoto, 2019).

D'autres stratégies, plus précises, vont combiner plusieurs ARNg en utilisant un array de gRNA-tRNA pour CRISPR-Cas9 (GTR-CRISPR), (Y. Zhang, Wang, et al., 2019). Une séquence de tRNA est insérée entre chaque gRNA et deux promoteurs SNR52 sont utilisés pour transcrire chacun quatre des huits gRNA. Les ARNt endogènes sont ensuite clivés par deux enzymes, la RNase P et la RNase Z (**Figure 17**). Cette optimisation permet de transformer la levure avec un seul plasmide possédant les différents ARNg (Y. Zhang, Wang, et al., 2019). En ré-utilisant ce système, le laboratoire de Tom Ellis a pu optimiser un plasmide pour faire jusqu'à 10 intégrations, en une seule fois et sans marqueur de sélection (Shaw et al., 2023). Plusieurs équipes continuent d'améliorer le système en mutant la Cas9 (c'est le cas du plasmide Di-CRISPR) ou en augmentant le taux de transcriptions des ARNg ou de la protéine.

Le développement de l'ingénierie génétique, avec des nouveaux systèmes plus modulaires et des librairies de plasmides CRISPR-Cas9, a étoffé la boîte à outils déjà sophistiquée de cet organisme. Ainsi, la levure du boulanger est devenue un outil de choix pour la génomique synthétique permettant la construction de long fragment, voire l'assemblage d'un génome complet et son ingénierie.

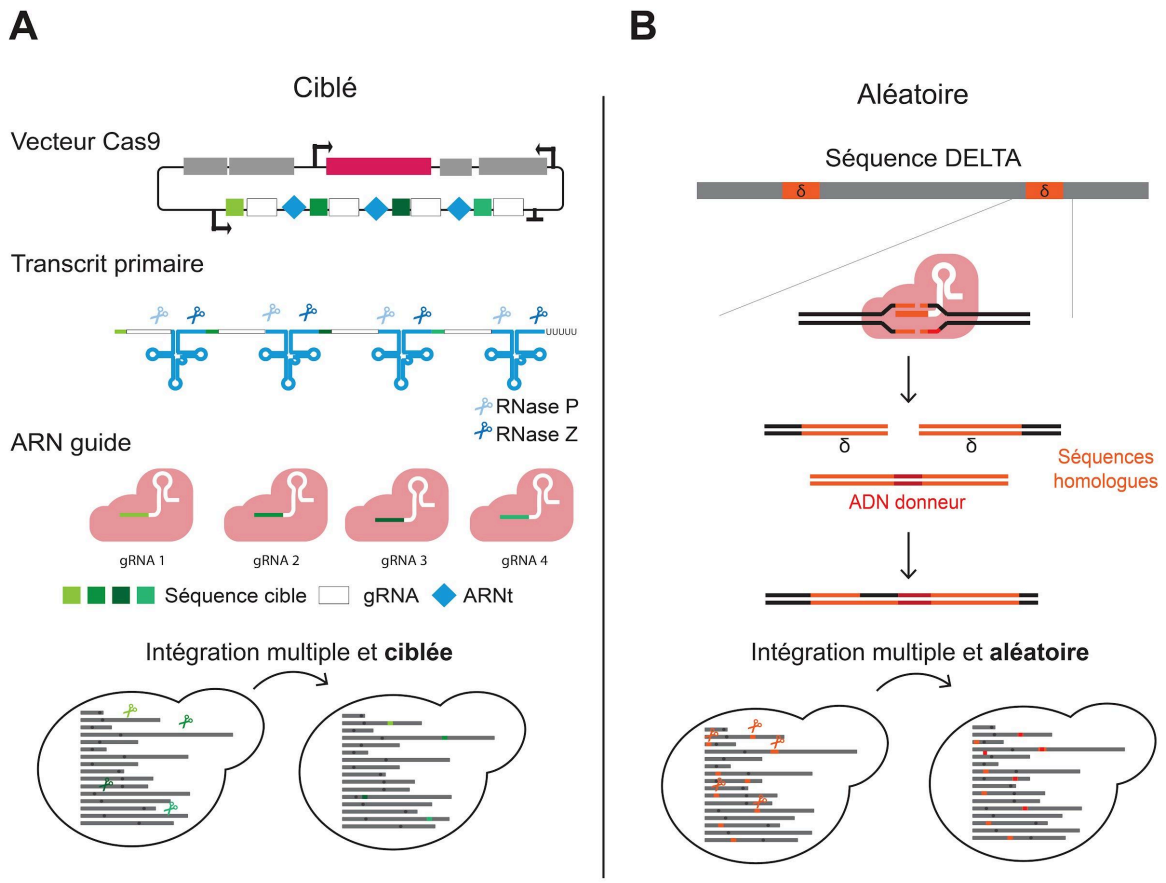


Figure 17 : Vue d'ensemble des systèmes de multi-intégration dans le génome de la levure à l'aide de CRISPR-Cas9.

A) Plusieurs ARNg peuvent être intégrés dans un unique plasmide, espacé par des séquences de tRNA. Après transcription, les séquences tRNA sont reconnues et clivées. Les ARNg, maintenant libérés, sont pris en charge par la Cas9 et pourront cibler plusieurs régions différentes du génome de la levure. **B)** La seconde stratégie utilise un seul ARNg qui cible les séquences delta (δ). Ces séquences répétées dans le génome peuvent être ciblées en même temps et permettre une intégration multi-copies.

3.2.2 Assemblage de génome complet dans la levure

En 2008, Gibson et al. décrivent la synthèse du génome de *Mycoplasma genitalium* de la synthèse à son assemblage final dans la levure *S. cerevisiae* (détaillé dans la section 3.2.2). En parallèle de ces travaux de synthèse du génome de mycoplasme, Carole Lartigue a réalisé des travaux pionniers sur l'intégration de chromosomes de *Mycoplasma* dans la levure. Elle a développé des méthodes pour cloner des chromosomes bactériens entiers en tant que plasmides centromériques dans la levure (Lartigue et al., 2009). L'objectif était de tirer parti de la levure pour assembler et modifier le chromosome bactérien dans la levure puis le transplanter dans la souche bactérienne receveuse

(**Figure 18. A**), (Lartigue et al., 2009). L'ADN génomique intact de *Mycoplasma mycoides* avait été transplanté dans *Mycoplasma capricolum* (Lartigue et al., 2007) mais en essayant de transplanter directement le génome provenant de la levure, Carole Lartigue et ses collègues n'ont pas obtenu de clones. En fait, les génomes bactériens de mycoplasme dans la levure ne sont pas méthylés et, par conséquent, ne sont plus protégés contre le système de restriction de la cellule receveuse. Une étape supplémentaire est introduite pour méthyler le chromosome avant de transplanter le génome dans la cellule endogène (Lartigue et al., 2009).

En 2010, Gibson et al. ont rapporté la synthèse, l'assemblage, le clonage et la transplantation réussie du génome synthétique *M. mycoides* JCVI-syn1.0, composé de 1,08 Mb (Gibson et al., 2010). Le processus d'assemblage s'est déroulé en plusieurs étapes : d'abord dans *E. coli*, puis dans la levure *S. cerevisiae* (**Figure 18. B**). Lorsqu'ils ont construit le génome *in silico*, les membres de l'équipe se sont basés sur le premier génome de référence de *M. mycoides* publié à ce moment-là. Pendant l'assemblage, la mise à jour du génome de référence a mis en évidence 19 différences polymorphiques entre le génome synthétique (JCVI-syn1.0) et le génome naturel (non synthétique) (YCpMmyc1.1), (Gibson et al., 2010). Leur méthodologie a été limitée pendant de nombreuses semaines par une délétion d'une seule paire de bases dans le gène essentiel dnaA. Notablement, une base erronée sur plus d'un million dans un gène essentiel a rendu le génome inactif (Gibson et al., 2010). Les travaux sur les génomes de Mycoplasme ouvrent la voie de la création d'un organisme complètement synthétique, d'abord construit *in silico,* sur ordinateur, jusqu'au clonage d'un tout nouvel organisme.

En 2019, l'équipe de Jason Chin au MRC a reconstitué le génome complet d'*E. coli*, réduisant le nombre de codons utilisés par cet organisme. Pour concevoir ce génome synthétique de 4 Mb, ils ont d'abord assemblés plusieurs fragments d'ADN synthétique dans la levure par recombinaison homologue. Ces fragments ont ensuite été progressivement insérés dans le génome d'*E. coli* (Fredens et al., 2019).

Suite à ces travaux, des génomes de virus et procaryotes ont été assemblés entièrement *(*listés dans Koster et al., 2022). Quelles sont les perspectives pour la synthèse d'un génome eucaryote complet ?

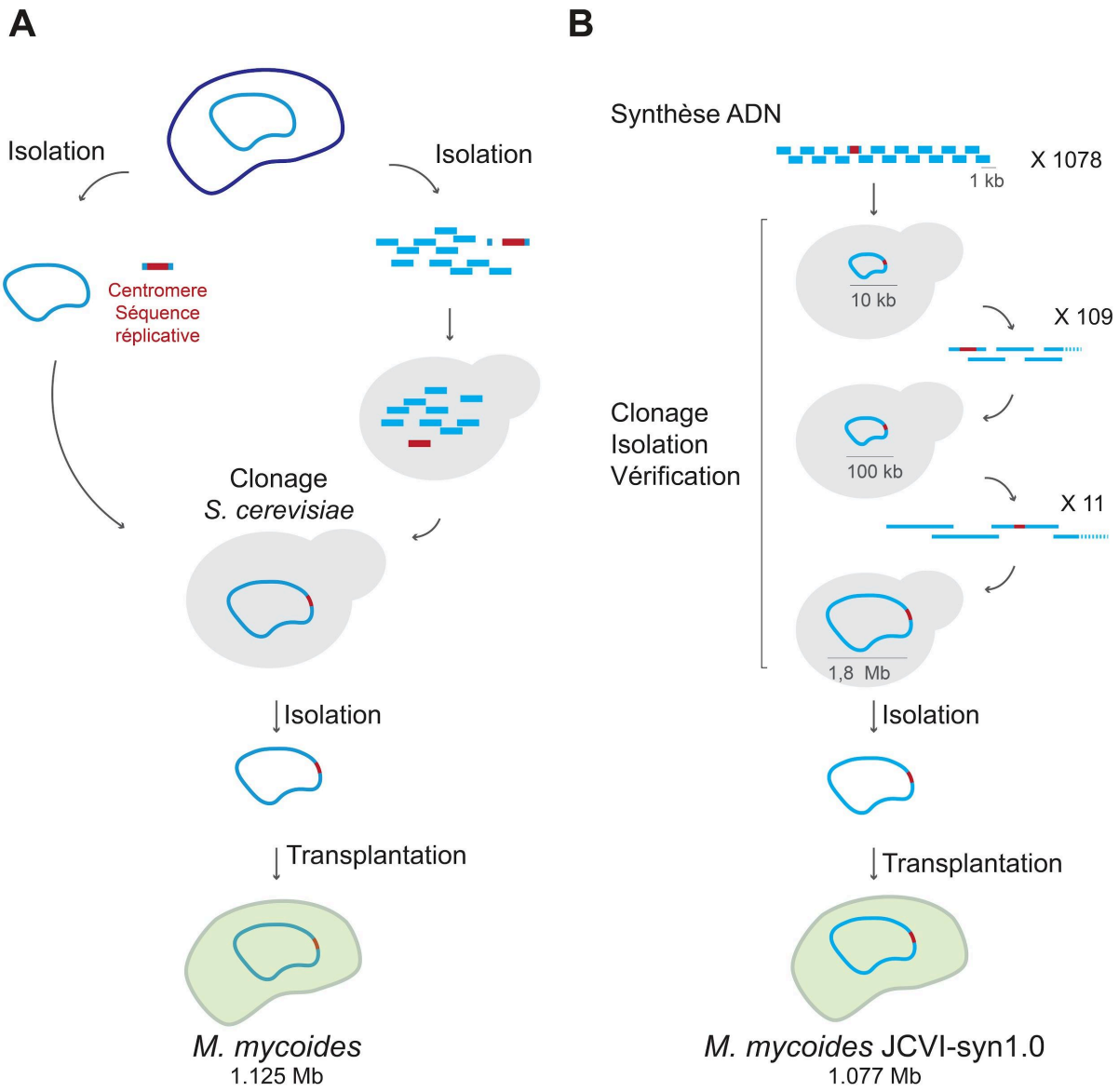


Figure 18 : Travaux sur les génomes de Mycoplasmes.

A) Cette stratégie consiste à transférer un génome bactérien dans la levure, le modifier et le transplanter dans une bactérie. Pour cela, le génome de l'organisme d'intérêt est isolé soit de manière intact, soit fragmenté puis transformé dans la levure avec une séquence réplicatif et un marqueur de sélection. Ce génome sera ensuite modifié, isolé et transplanté dans une cellule réceptrice pour générer une bactérie modifiée (Lartigue et al., 2007, 2009). **B)** Le génome synthétique de *Mycoplasma mycoides* a été assemblé dans la levure à partir de cassettes d'ADN de 1 kb, assemblé progressivement dans la levure en fragments plus grands (10 kb et 100 kb), et finalement en un génome de 1,08 Mb. Le génome est ensuite extrait et transplanté dans *M. capricolum* pour générer une nouvelle souche de *M. mycoides* contrôlée par le génome synthétique (JCVI-syn1.0) (Gibson et al., 2010).

3.2.3 Projet de synthèse du génome de *S. cerevisiae* (Sc2.0)

Un consortium mondial dirigé par Jef Boeke poursuit un projet, Sc2.0, qui vise à construire le premier génome synthétique eucaryote, celui de *S. cerevisiae*. C'est le troisième plus petit chromosome, et premier séquencé, de la levure, le chromosome III (316 617 pdb) qui est réécrit, synthétisé et ré-assemblé pour construire synIII (272 871 pdb) (**Figure 19**), (Annaluru et al., 2014). Les modifications apportées au synIII comprennent les remplacements de codons stop TAG/TAA, la suppression de régions subtélomériques, d'introns, d'ARN de transfert, de transposons, ainsi que l'insertion de sites loxPsym, 3 pdb en aval de tous les gènes non-essentiels. En 2018, huit chromosomes (synI, synII, synIII, synV, synVI, synIX-R, synX et synXII) ont été entièrement assemblés séparément.

L'analyse Hi-C des différents chromosomes issus du projet Sc2.0 a permis de mettre en lumière la conservation de la structure de Rabl de *S. cerevisiae*, guidée principalement par l'ancrage des centromères au SPB et le regroupement des télomères en périphérie du noyau (Mercy et al., 2017). Les fonctions des caractéristiques génomiques telles que la nature répétitive des séquences codantes de l'ARN ribosomique, la présence omniprésente de transposons et l'existence universelle d'introns et de machines d'épissage dans les systèmes eucaryotes ont été délibérément testées dans le cadre du projet de génome synthétique de la levure (Sc2.0) (Mercy et al., 2017). Il reste à déterminer si la souche finale Sc2.0 avec tous les chromosomes synthétiques dans une seule cellule aura un phénotype de type sauvage.

Un des atouts du design Sc2.0 est le système SCRaMbLE. Il peut générer une diversité significative dans la structure, l'ordre et le contenu des génomes, voire modifier la composition de la variance des gènes essentiels et non-essentiels au sein de ces modèles. Lors de l'induction du système par une Cre-recombinase active, un nombre effectivement infini de réarrangements du génome peut être généré, y compris des délétions, des duplications, des inversions et des translocations de gènes entre deux sites loxPsym quelconques. Il devient plus facile de créer un système d'expression personnalisé, qui se débarrasse des entraves des séquences de type sauvage et peut constituer un tournant important dans le domaine de la génomique synthétique, qui passe de l'imitation à la création. Dans une série d'articles récents, la puissance de ce système a été démontrée pour l'optimisation des souches et pour stimuler la production de composés tels que les caroténoïdes, la pénicilline et la violacéine.

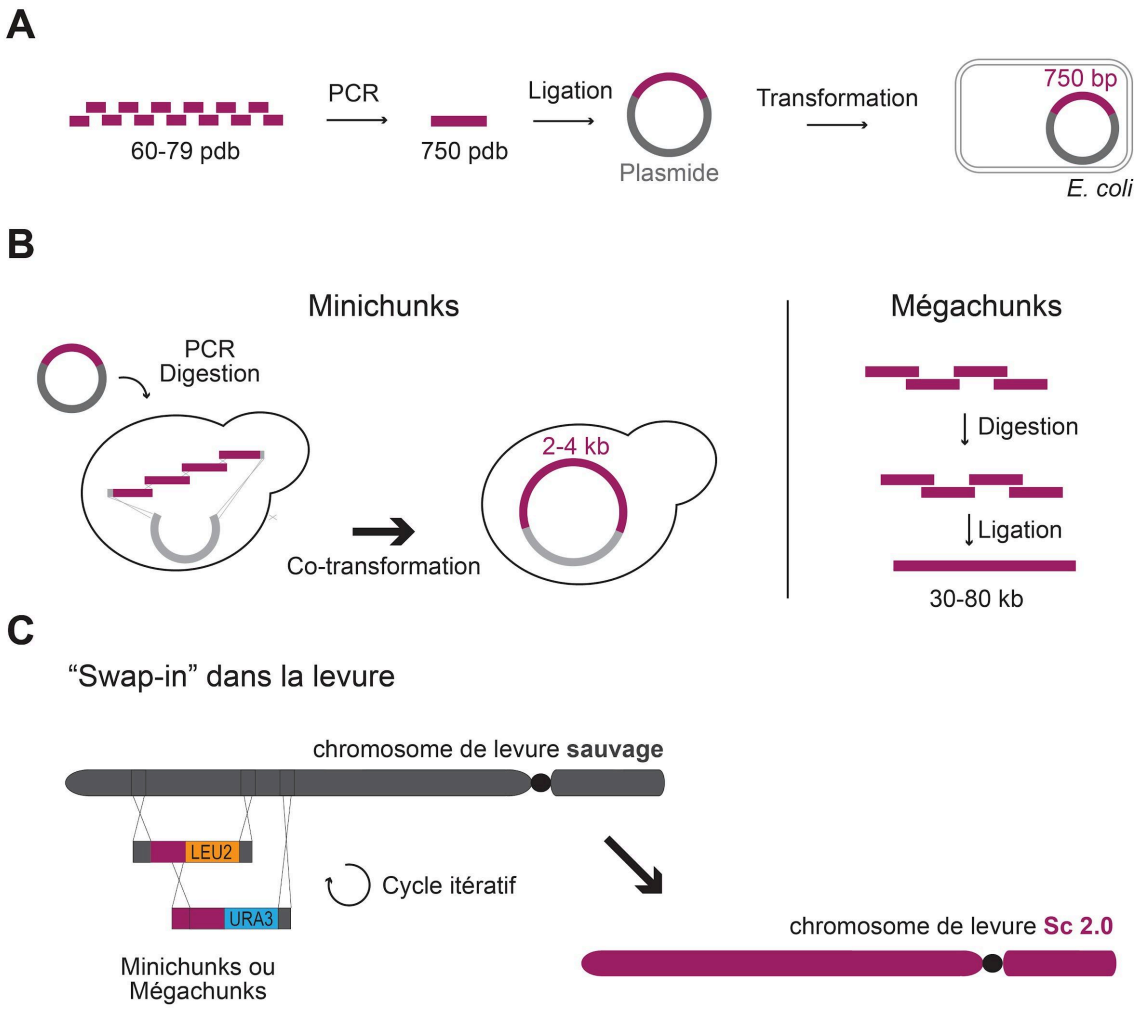


Figure 19 : Construction des chromosomes Sc2.0.

Les premiers chromosomes du projet Sc2.0 ont été initialement construits à partir de blocs de construction de 750 pb, par PCR et clonage dans la levure en mini chunks de 2-4 kb. Les chromosomes Sc2.0 plus récents sont assemblés à partir de chunks de 6-10 kb synthétisés en mégachunks de 30-50 kb par digestion et ligature in vitro. Les minichunks ou mégachunks ont ensuite été transformés dans des cellules de levure pour remplacer le génome natif de manière itérative.

3.2.4 Minimiser et re-fonctionnaliser les génomes

Les organismes évoluant naturellement ont généralement des génomes de grande taille qui leur permettent de survivre et se développer dans diverses conditions. Cette complexité empêche souvent de les comprendre complètement. L'un des moyens d'identifier un ensemble minimal de gènes suffisant pour le développement d'un organisme consisterait à créer et à tester un chromosome artificiel à base de cassettes (X. Xu et al., 2023). En 2016, l'équipe de Craig Venter a minimisé le génome synthétique de Mycoplasma, JCVISyn1.0 (1079 kb) (Gibson et al., 2010) pour obtenir un génome considérablement réduit, JCVI-Syn3.0 (531 kb) (Hutchison et al., 2016). D'autres génomes

d'organismes comme *E. Coli, Bacillus subtilis, Streptomyces avermitilis* ou *Schizosaccharomyces pombe* ont été minimisés (X. Xu et al., 2023).

Comme mentionné précédemment, le système SCRaMbLE pourrait être adapté pour minimiser le génome Sc2.0. Celà pourrait être utilisé comme stratégie pour générer des génomes de levure minimaux dans des conditions données. L'un des principaux défis liés à l'application du système Sc2.0 SCRaMbLE pour la minimisation du génome est la perte de viabilité des cellules après SCRaMbLE, qui diminue les chances d'obtenir des délétions de longues séquences. En effet, bien que les sites loxPsym aient été insérés 3 pdb en aval du codon stop des gènes non essentiels, dans de nombreux cas, des gènes essentiels et des gènes non essentiels sont présents entre deux sites loxPsym. Dans les prochains designs, les sites loxPsym pourraient être insérés plus stratégiquement et permettre des délétions plus larges.

Chez *S. cerevisiae*, deux groupes ont essayé de réduire le génome à un unique chromosome. Un groupe a réussi à fusionner les 16 chromosomes et à construire une souche de levure avec un chromosome unique et l'autre groupe à générer une souche viable contenant deux chromosomes (**Figure 20**), (Luo et al., 2018; Shao et al., 2018). Ces études ont démontré que le nombre de chromosomes dans une cellule eucaryote n'est pas fixe, du moins dans la levure, et que son génome est hautement modulable. Dans le cas du chromosome unique, le groupe est parti d'une cellule haploïde contenant seize chromosomes linéaires, qui, par fusions successives de chromosomes bout à bout, suppressions de centromères et de régions répétées, ont permis la formation d'un seul chromosome (**Figure 20**). Il est intéressant de noter qu'en dépit de multiples tentatives, les deux chromosomes des souches générées par Luo et al. n'ont pas pu être fusionnés, ce qui suggère l'existence de composants essentiels inconnus pour la survie d'une cellule à chromosome unique.

Tous ces résultats de réductions de génomes représentent des sources d'informations considérables pour la conception future de génomes synthétiques, ils permettent de déterminer quelles séquences d'ADN garder ou éliminer en fonctions du phénotype recherché. Les génomes minimisés sont très prometteurs, tant pour la compréhension de la biologie que pour l'ingénierie de souches industrielles supérieures. Cependant, l'assemblage ou l'ingénierie d'un génome minimal prend généralement beaucoup de temps, suivi d'un long processus d'essais et d'erreurs (X. Xu et al., 2023). En outre, un génome minimal est affecté par les conditions dans lesquelles il est construit. Actuellement, la plupart des projets de génomes minimaux sont construits dans un milieu riche en nutriments et les éléments génétiques codant pour la réponse et la tolérance au stress sont souvent écartés, ce qui peut entraîner une diminution de la croissance dans des conditions de fermentation industrielle. Une autre approche consiste à construire des néochromosomes ou des génomes minimaux entiers sur mesure pour répondre à différentes exigences, par exemple pour différentes conditions de fermentation ou produire différents composés.

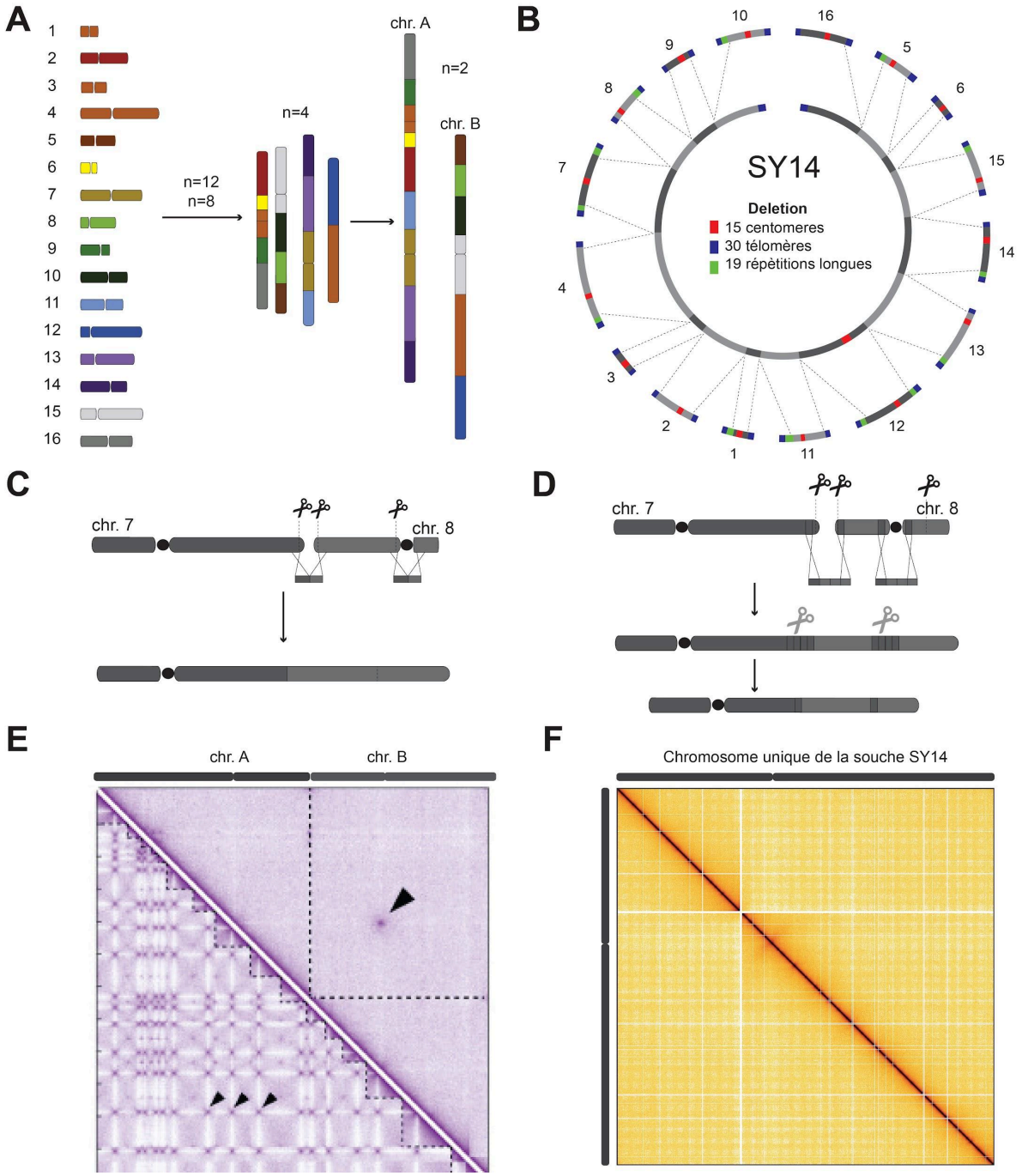


Figure 20 : Stratégie des chromosomes de fusion.

A) Schéma représentant comment les 16 chromosomes de type sauvage (n = 16) sont passés à n = 12, n = 8, n = 4 et enfin n = 2 dans (Luo et al., 2018). Les 16 chromosomes sont colorés de manière unique et classés par numéro. **B)** Seize chromosomes natifs (I-XVI) de BY4742 (type sauvage) sont alignés dans l'anneau extérieur. L'unique chromosome de SY14 aligné dans l'anneau intérieur a subi quinze séries séquentielles de fusions chromosomiques bout à bout, indiquées par des lignes en pointillés dans (Shao et al., 2018). **C)** Schéma de la méthode CRISPR-Cas9 pour fusionner deux chromosomes dans (Luo et al., 2018) **D)** Schéma de la méthode CRISPR-Cas9 pour fusionner deux chromosomes dans (Shao et al., 2018). **E)** Matrice Hi-C de la souche n=2, asynchrone **F)** Matrice Hi-C de la souche SY14.

3.3. Les néochromosomes

3.3.1 “Recomposer” le génome de la levure

Contrairement aux chromosomes synthétiques existants du projet Yeast 2.0, les néochromosomes ne sont généralement pas basés sur un modèle naturel et, en tant que tels, les approches de bio-conception *in silico* jouent un rôle central en raison de leur nature *de novo*. Contrairement aux approches classiques de la biologie synthétique qui considèrent l'ADN comme un "code" à écrire, ces efforts de bio-conception vont jusqu'à traiter les néochromosomes comme une entité physique et nécessitent donc des approches plus proches de l'ingénierie structurelle réalisée à l'échelle moléculaire.

Je présente ici deux cas où des séquences endogènes de la levure ont été dupliquées et/ou ré-organisées : les ARNt et l’ADNr. Dans le premier cas, un néochromosome contenant les 275 gènes ARNt nucléaires de *S. cerevisiae* est construit et caractérisé. Pour maximiser la stabilité, la conception incorpore 89% de séquences provenant d'espèces de levures autres que *S. cerevisiae* pour limiter la recombinaison avec le génome de l’hôte (Schindler et al., 2023). Le néochromosome ARNt de ~190-kb révèle une plasticité génomique inattendue. Cependant, ils n'ont pas testé ce néochromosome dans un génome dépourvu d’ARNt. Similairement, Luciana Lazar-Stefanita déploie une approche innovante d'ingénierie des génomes pour couper et coller à un emplacement chromosomique ectopique - le locus ADNr d'environ 1,5 méga-base - en une seule étape à l'aide de la technologie CRISPR-Cas9 (Lazar-Stefanita et al., 2023) . Après plusieurs croisements permettant d’obtenir une souche comportant deux séquences de l’ARNr, le HiC et la microscopie identifient que les deux séquences interagissent entre elles. Il serait intéressant de voir, si la levure, au cours de la réaction de SCRaMbLE, regroupe spontanément des gènes codant pour une même voie métabolique.

Kutyna et al. ont conçu un néochromosome de 211 409 pdb synthétique pan-génome (PGNC), qui intègre 75 ORFs prédits provenant d'isolats industriels, pathogènes humains et naturels de levure (Kutyna et al., 2022). Grâce à la présence du PGNC, la souche résultante a pu utiliser une gamme plus large de sources de carbone que la souche parentale Sc2.0. L'utilisation différentielle des sources de carbone a fourni les exemples les plus clairs de la croissance sélective de la souche avec le néochromosome, le mélibiose, le palatinose et l'acide butyrique n'étant utilisés qu'en présence du néochromosome (Kutyna et al., 2022). Il s'agit d'un exemple clair de la manière dont la construction d'un néo-chromosome peut améliorer les caractéristiques favorables à l'industrie et étendre les applications de la souche synthétique. Ingénieusement, ce chromosome est compatible avec le cadre de conception et de Sc2.0 (remplacer les codons stop TAG par des TAA, d'introduire des filigranes oligonucléotidiques dans 36 ORF et 63 sites de reconnaissance bidirectionnelle de la Cre-recombinase

(loxPsym)) et permet de générer une diversité significative dans la structure, l'ordre et le contenu de ce néochromosome.

3.3.2 Intégration de séquences d'autres espèces

Comme nous l'avons vu, la capacité d'assembler de grands morceaux d'ADN dans la levure a beaucoup d'applications en biologie. L'identification des éléments nécessaires pour la stabilité d'un chromosome a permis la mise en place des Yeast Artificial Chromosomes (YACs), (Blackburn, 1985). Ces YACs ont été utilisés pour séquencer le génome humain (GC 41%)(Hudson et al., 1995) et aussi divers organismes eucaryotes, comme la souris (41% GC) (Nusbaum et al., 1999), le poisson zèbre (37%) (Postlethwait et al., 1998), *Arabidopsis thaliana (*36% GC*)* (Schmidt et al., 1992)*, Plasmodium falciparum* (19% GC) (Gardner et al., 1998) pour donner quelques exemples. Plusieurs systèmes ont été développés pour optimiser l'intégration de ces fragments (Kouprina & Larionov, 2008). Ainsi, des séquences allant jusqu'à 500 kb d'une grande variété d'organismes sont intégrées dans la levure. Mais le clonage de grands morceaux d'ADN à GC élevé peut s'avérer difficile. Pour cloner *Synechococcus elongatus*, qui a une teneur en GC de 55%, et contourner l'absence de séquence ARS de levure (riche en AT) dans l'assemblage de 454 kb, une origine de réplication de levure est insérée dans plusieurs fragments (Noskov et al., 2012).

Plus récemment, les yeast assembly vectors (YAV) ou vecteur d'assemblage de levure, peuvent être clonés et transférés dans des bactéries pour être isolées avant d'être livrées à des cellules de mammifères. Après l'assemblage initial, le YAV peut être édité dans la levure pour générer des panels de variants de conception. C'est le CAS du système CREEPY optimisé pour l'édition épisomale, qui permet l'ingénierie d'un YAV contenant le gène Sox2 (143 kb) de la souris et les régions régulatrices (Y. Zhao et al., 2023). L'utilisation de la levure a aussi permis l'assemblage du gène humain HPRT1 de 101 kb dans la levure à partir de blocs de construction de 3 kb, puis son intégration dans des cellules souches embryonnaires de souris permettant son expression (Mitchell et al., 2021). Les améliorations ont permis d'introduire de nouvelles caractéristiques et de construire des voies de biosynthèse humaine. En une seule fois, Agmon et al. transplantent 7 gènes humains, constituant la voie de l'adénine *de novo*, dans des cellules de levure (Agmon et al., 2020). La voie complète de biosynthèse de l'adénine peut ensuite être extraites et transformées chimiquement dans un nouveau châssis hôte à volonté, et réintégrées dans le génome humain.

Nous disposons d'une collection de séquences variées intégrées dans la levure, permettant leur édition et réinsertion dans leurs organismes d'origine. Mais ces matrices d'ADN peuvent aussi être utilisées pour étudier leurs comportements dans un organisme qui n'a pas évolué avec cette séquence.

3.3.3 Caractérisation des néo-chromosomes

Jusqu'à présent, il n'y a pas eu beaucoup de caractérisation contrôlée des nouveaux locus d'ADN exogènes dans les génomes hôtes. Ces séquences nommées exogènes ou naïves permettent d'explorer les principes de la régulation transcriptionnelle en l'absence de contraintes évolutives. Zhou et al. ont dressé une carte complète des caractéristiques génétiques, épigénétiques et transcriptionnelles d'un chromosome additionnel, "dChr" (dataStorage, ~254 kb, 50% GC), (Zhou, 2022). Ils ont constaté que cet ADN formait une chromatine active avec une grande accessibilité chromatinienne (mesurée par ATAC-seq) et des niveaux élevés de tri-méthylation H3K4. L'analyse transcriptomique du dChr révèle que cette séquence est transcrite à des niveaux similaires à ceux des ARN issus des portions génomiques de levure (Luthra et al., 2024; Zhou, 2022).

L'équipe de Kevin Strulh a ensuite étudié la structure de la chromatine et la transcription d'une région d'ADN de 18 kb ("chrXVII", 50% GC) dont la séquence a été générée aléatoirement (Gvozdenov et al., 2023). Ils démontrent que les nucléosomes occupent entièrement l'ADN de la séquence aléatoire, mais que les régions appauvries en nucléosomes (NDR) sont beaucoup moins fréquentes avec moins de nucléosomes bien positionnés. Les niveaux d'ARN détectés sur la séquence aléatoire sont comparables à ceux des ARN de la levure. L'initiation de la transcription à partir d'ADN à séquence aléatoire se produit à de nombreux endroits, indiquant une très faible spécificité intrinsèque de la machinerie de l'ARN Pol II. Les ARN provenant de ce chromosome additionnel présentent une plus grande variabilité d'une cellule à l'autre que les ARNm de levure, ce qui suggère que des éléments fonctionnels limitent cette variabilité. Un des éléments limitants de cette étude est la longueur restreinte de la région étudiée pour une exploration approfondie de la régulation de la transcription, comme la recherche de motifs ou d'éléments régulateurs. Le rôle des voies de dégradation des ARN NNS et NMD n'a pas non plus été abordé.

Au lieu d'utiliser une séquence aléatoire, l'équipe de Jef Boeke utilise le gène humain HPRT1 comme matrice exogène. Pour obtenir un nouveau fragment d'ADN, ils ont utilisé la séquence inversée du gène et se sont concentrés sur la caractérisation du paysage chromatinien des deux séquences (gène HPRT1 et HPRT1**R**) dans deux types cellulaires différents : la levure et la souris (Camellato et al., 2024). Dans la levure, les résultats indiquent que le bruit transcriptionnel se produit à des niveaux élevés chez la levure, en accord avec les autres études. En revanche, chez la souris, la séquence d'ADN est transcriptionnellement silencieuse.

Ces exemples renforcent l'hypothèse selon laquelle, chez la levure, les fragments synthétiques ou d'organismes étrangers sont spontanément transcrits. Cette propension à être transcrit chez la levure est surprenante. Une des explications proposées par l'équipe de Jef Boeke est que la levure est

largement isolée par une paroi cellulaire épaisse pendant la majeure partie de son cycle de vie et ne possède pas de virus transmis de manière conventionnelle (Camellato et al., 2024). La levure pourrait donc être un cas isolé, capable de tolérer un état par défaut ouvert et actif dans une plus large mesure que les autres cellules eucaryotes.

En conclusion, il est clair que *S. cerevisiae* est un acteur clé dans la génomique synthétique grâce à la combinaison des avancées en synthèse d'ADN et du développement d'outils d'ingénierie génétique. La puissance d'assemblage permet maintenant d'introduire et construire une variété de longues séquences, synthétiques ou d'organismes étrangers qui peuvent être caractérisées dans divers types de cellules eucaryotes différentes. On peut alors caractériser et étudier comment s'organisent ces segments d'ADN afin de comprendre les processus biologiques sous-jacents.

4. Projet de thèse

Dans cette introduction, nous avons vu que les approches génomiques ont fourni des révélations majeures sur l'organisation et la fonction du génome 3D et révèlent une structure hiérarchique, organisée par divers complexes moléculaires, allant des nucléosomes aux boucles d'ADN médiées par la cohésine, jusqu'aux compartiments de chromatine à grande échelle. L'ADN joue un rôle crucial dans cette organisation 3D où les séquences génomiques et les protéines associées évoluent ensemble pour permettre le repliement des longues molécules d'ADN en chromosomes fonctionnels. La dynamique de l'organisation spatiale des chromosomes revêtent une importance fonctionnelle pour l'expression des gènes, la réplication de l'ADN et la ségrégation. Il est essentiel d'en déterminer les causes et les conséquences, et de comprendre comment l'organisation 3D du génome influence ce processus.

Au cours de ma thèse, j'ai étudié la manière dont un hôte eucaryote, *S. cerevisiae*, peut replier et réguler l'activité de séquences d'ADN bactériennes exogènes longues comme des chromosomes ou des séquences aléatoires. Ces stratégies originales permettent d'explorer les principes de la régulation transcriptionnelle en l'absence de contraintes évolutives.

Avec ces approches, les objectifs sont :

- d'étudier l'importance de la composition des séquences, encore largement inconnues.
- Identifier les déterminants fondamentaux et les règles qui lient la séquence d'un génome à l'organisation fonctionnelle d'ordre supérieur des chromosomes eucaryotes.
- explorer la dynamique et l'impact fonctionnel des grandes molécules d'ADN exogènes qui s'intègrent dans les génomes eucaryotes.

La première partie des résultats portera sur la caractérisation d'ADN bactériens d'environ 1 Mb dans un noyau eucaryote. Nous avons caractérisé le comportement de deux chromosomes bactériens introduits artificiellement dans le génome de la levure en tant que chromosome surnuméraire à l'aide d'approches génomiques. La seconde partie portera sur la caractérisation et l'ingénierie d'une séquence de 100 kb d'ADN aléatoire, intégrée dans le chromosome IV de la levure. Ensemble, ces deux projets exploitent plusieurs méthodes génomiques appliquées à des séquences d'ADN qui n'ont pas évolué avec l'organisme hôte.

Résultats

1. Caractérisation d'ADN bactériens dans un noyau eucaryote

1.1. Article 1 : La composition des séquences détermine l'activité, le repliement et la compartimentation de l'ADN dans un noyau eucaryote.

L'ADN joue un rôle crucial dans l'organisation du génome, où les séquences génomiques et les protéines associées évoluent ensemble pour permettre le repliement des longues molécules d'ADN en chromosomes fonctionnels. Chez les eucaryotes, cette structure hiérarchique est organisée par divers complexes moléculaires, allant des nucléosomes aux boucles d'ADN médiées par la cohésine, jusqu'aux compartiments de chromatine à grande échelle.

Dans ce premier projet, nous avons caractérisé l'assemblage et l'activité de la chromatine dans deux souches de levures portant respectivement un chromosome bactérien exogène : *M. mycoides* et *M. pneumoniae*. Ces derniers ont divergé des séquences eucaryotes il y a plus de 1,5 milliard d'années. Ces souches ont été générées par Carole Lartigue et le projet a été mené en collaboration avec l'expertise de plusieurs laboratoires. Avec Jacques Serizay et Christophe Chapard, nous sommes trois co-auteurs sur ce projet.

Nous avons démontré que l'assemblage des nucléosomes, l'activité transcriptionnelle, les boucles médiées par la cohésine et la compartimentation de la chromatine peuvent se produire dans les chromosomes bactériens d'un hôte eucaryote (**Figure. 1, 2,3**). Remarquablement, nous avons observé la formation de deux archétypes de chromatine différents pour ces deux génomes très divergents.

Pour tester si et comment ces deux types de chromatines puissent coexister sur un seul chromosome, nous avons fusionné le chromosome Mmyco avec le chromosome XVI de la levure et induit des translocations pour générer deux souches supplémentaires avec des chromosomes hébergeant des régions de chromatine U et Y alternées (**Figure. 4A ; Figure. S8A-B Méthodes**). Les translocations ont révélé une compartimentation en *cis* et en *trans* de ces chromosomes mosaïques et confirmé le fait que la chromatine U inactive peut s'étendre sur plusieurs chromosomes. Enfin, nous avons appliqué des méthodes de machine learning pour comprendre les origines de ces observations, ouvrant ainsi de nouvelles perspectives sur l'organisation et l'activité des génomes exogènes au sein des cellules eucaryotes.

Ce projet rassemble différentes expertises. Au cours de ma thèse, j'ai principalement participé à la partie expérimentale. Avec l'aide d'Agnès Thierry, j'ai poursuivi la mise au point du MNase-seq en réalisant des cinétiques de digestion pour limiter les biais de digestion liés au taux de GC. J'ai également mis au point l'ATAC-seq au sein du laboratoire. Ces deux méthodes, appliquées sur nos souches d'intérêt, nous ont permis d'avoir une vue d'ensemble assez complète de la chromatine des chromosomes bactériens. J'ai ensuite réalisé les différentes fusions et translocations des chromosomes bactériens au chromosome 16 via CRISPR-Cas9, ainsi que les synchronisations et Hi-C associées. J'ai participé à la mise en place et réalisé certaines expériences, avec Agnès Thierry et Manon Perrot, des différents tests fonctionnels, afin de commencer à comprendre les aspects mécanistiques associés à la compartimentation. Enfin, j'ai participé à l'interprétation des données et à l'écriture du manuscrit.

Ces résultats constituent une avancée significative dans la compréhension de l'interprétation des séquences étrangères par la machinerie nucléaire de l'hôte lors des transferts horizontaux naturels de gènes, ainsi que dans les projets de génomique synthétique. Dans cette première partie de résultats, je présente d'abord le papier puis des résultats supplémentaires sur les chromosomes bactériens.

Sequence-dependent activity and compartmentalization of foreign DNA in a eukaryotic nucleus

Authors: Léa Meneu[1,2,Ψ], Christophe Chapard[1,Ψ,&], Jacques Serizay[1,Ψ,*], Alex Westbrook[2,3], Etienne Routhier[2,3,4], Myriam Ruault[5], Manon Perrot[1,2], Alexandros Minakakis[8], Fabien Girard[1], Amaury Bignaud[1,2], Antoine Even[5], Agnès Thierry[1], Géraldine Gourgues[6], Domenico Libri[8], Carole Lartigue[6], Aurèle Piazza[1,#], Angela Taddei[5], Frédéric Beckouët[7], Julien Mozziconacci[3,4,9*] and Romain Koszul[1,*]

Affiliations:
[1] Institut Pasteur, CNRS UMR 3525, Université Paris Cité, Unité Régulation Spatiale des Génomes, 75015 Paris, France
[2] Sorbonne Université, College Doctoral
[3] Laboratoire Structure et Instabilité des génomes UMR 7196, Muséum National d'Histoire Naturelle, Paris 75005, France
[4] Laboratoire de Physique Théorique de la Matière Condensée, Sorbonne Université, CNRS, 75005 Paris, France
[5] Institut Curie, PSL University, Sorbonne Université, CNRS, Nuclear Dynamics, 75005 Paris, France
[6] Univ. Bordeaux, INRAE, Biologie du Fruit et Pathologie, UMR 1332, F-33140 Villenave d'Ornon, France
[7] Molecular, Cellular and Developmental biology department (MCD), Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 31062, Toulouse, France
[8] Institut de Génétique Moléculaire de Montpellier (IGMM), 34090 Montpellier, France
[9] UAR 2700 2AD, Muséum National d'Histoire Naturelle, Paris 75005, France
[#] Present address: Univ Lyon, ENS, UCBL, CNRS, INSERM, Laboratory of Biology and Modelling of the Cell, UMR5239, U 1210, F-69364, Lyon, France
[&] Present address: Molecular, Cellular and Developmental biology department (MCD), Centre de Biologie Intégrative (CBI), Université de Toulouse, CNRS, UPS, 31062, Toulouse, France

Ψ These authors contributed equally

* Corresponding authors: romain.koszul@pasteur.fr, julien.mozziconacci@mnhn.fr, jacques.serizay@pasteur.fr

One sentence summary: Exogenous bacterial chromosomes transferred in a eukaryotic nucleus spontaneously form active and inactive chromatin compartments.

Abstract

In eukaryotes, DNA-associated protein complexes co-evolve with genomic sequence to orchestrate chromatin folding into functional chromosomes. Here, we investigate the relationship between DNA sequence and the spontaneous loading and activity of chromatin components in the absence of co-evolutions. Using bacterial genomes integrated into *S. cerevisiae*, which diverged from yeast up to 1.5 billion years ago, we show that nucleosomes, cohesins and the transcriptional machinery can lead to the formation of two different chromatin archetypes, one being transcribed and the other silent. These two archetypes also form on eukaryotic exogenous sequences, and depend on sequence composition. They do not mix in the nucleus, leading to a bipartite nuclear compartmentalisation reminiscent of the organization of vertebrate nuclei. Our findings represent a significant advance in understanding the primary molecular mechanisms that govern the co-option or silencing of DNA sequences integrated into foreign genomes during natural horizontal gene transfers, or in synthetic genomics projects.

Introduction

Genome sequence composition, broadly defined by its GC%, polynucleotide frequencies, DNA motifs and repeats, varies widely between species as well as within individual genomes (*1*, *2*). In eukaryotes, the sequence composition is known to correlate with: i) chromatin composition, which includes nucleosome formation and binding of structural and functional proteins to DNA (*3*), ii) chromatin activity, such as transcription and replication (*4*), and iii) functional 3D organization of the genome into loops and compartments (*5*, *6*). For instance in mammals, GC-rich regions are enriched in actively transcribed sequences, in chromatin loops mediated by the structural maintenance of chromosomes (SMC) cohesin, and coalesce into a specific compartment (*7*). These relationships between sequence and chromatin composition activities and folding reflect their continuous coevolution over millions of years.

Disruptive variations in sequence composition can emerge naturally during evolution, e.g. through transfer of genetic material across species by horizontal gene transfers or introgression, in viral infections (*8–11*), or even artificially, e.g. by introducing chromosome-long DNA molecules into chassis microbial strains or cell lines (*12–14*). Such transfer can lead to the long-term integration of foreign DNA whose sequence composition strongly diverges from their host's genome (e.g. the introgression in *Lachancea kluyveri* of a 1 Mb sequence with a GC content 12% higher that the rest of the genome (*15*)). Once integrated, these sequences are organized and processed by chromatin-associated proteins of the host genome, obeying new rules under which they have not coevolved. How a eukaryotic host packages, regulates the activity and folds long exogenous DNA sequences, and the importance of the sequence composition in this process, remain largely unknown.

Here, we investigate the behavior of natural, chromosome-sized bacterial and eukaryotic sequences, with different sequence composition, artificially introduced into the *S. cerevisiae* genome (*16*, *17*). We profile nucleosome, RNA polymerase and cohesin landscapes, transcriptional activity and 3D organization of these supernumerary chromosomes during the cell cycle. We show that highly divergent bacterial or eukaryotic chromosomes, with different GC content, present different chromatin composition and activities. This eventually leads to the spontaneous formation of two chromatin archetypes, one active and one inactive, each displaying different physical properties and segregating into distinct chromosomal compartments, similar to those observed in complex multicellular organisms such as mammals. This partitioning occurs independently of heterochromatin formation but is driven by transcriptional activity. Sequence determinants, computationally learnt on yeast sequences

are sufficient to predict the chromatin composition and activity of exogenous chromosomes integrated in the yeast, suggesting s that the fate of any DNA molecule introduced into a given cellular context, from nucleosome positioning up to its 3D folding and transcriptional activity, is governed by rules that are both deterministic and predictable.

Adaptation of supernumerary bacteria chromosomes integrated into yeast

To investigate large sequences which have not evolved in a eukaryotic context, we exploited *S. cerevisiae* strains carrying an extra 17th circular chromosome made either from the *Mycoplasma mycoides* subspecies *mycoides* (referred to as "Mmyco") or *Mycoplasma pneumoniae* ("Mpneumo") genomes containing a yeast centromeric sequence and an autonomous replication sequence (ARS) (*18*) (**Methods**; **Table S1**). While the GC content of *S. cerevisiae* is 38%, the GC content of the Mmyco and Mpneumo chromosomes are 24% (GC-poor) and 40% (GC-neutral), respectively (**Fig. 1A**). Dinucleotide composition (*19*) also largely differs between yeast and bacterial chromosomes (**Fig. S1A**). We linearized these chromosomes, added yeast telomeres at the extremities (**Fig. 1A; Fig. S1B**; **Methods**), and investigated their replication and cohesion using split-dot assay and marker frequency analysis (**Fig. S1C,D** and **Supplementary Text**; **Methods**). Overall, these chromosomes do not impose a significant fitness cost on their eukaryotic host and have a segregation rate similar to that of a centromeric plasmid **(Fig. S1E**, **S1F**) making them powerful tools for studying chromatin assembly on foreign sequences.

Spontaneous chromatinization of bacterial chromosomes in a eukaryotic context

While Mmyco and Mpneumo chromosomes evolved in a non-eukaryotic context, both sequences lead to the formation of nucleosome arrays and local nucleosome-depleted regions (NDR) as assessed by time-course MNase-seq and H3 and H2A ChIP-seq experiments (**Fig. 1B, C; Fig. S2A-B**). We annotated nucleosome positions following an approach adapted from Brogaard et al. (*20*), and found that nucleosome arrays over the Mpneumo chromosome are similar to yeast's ones, with a linker length of ~14-18 bp and a nucleosome repeat length (NRL) of 160 bp (**Fig. 1D,E)**. In contrast, nucleosome arrays in the Mmyco chromosome have a longer ~ 25 bp linker and an NRL of 174 bp (**Fig. 1D,E, S2D)**. Sequence features favoring nucleosome organization (short poly(dA) or poly(dT) tracks and strong WW 10 bp periodicity) are enriched in Mpneumo, while sequence features limiting nucleosome assembly (long poly(dA) or poly(dT) tracks and weak WW 10 bp periodicity) are enriched in Mmyco chromosome (**Fig. S2C,D**).

Chromatin accessibility profiling by ATAC-seq confirmed that the Mpneumo chromosome has a similar density and breadth of accessible loci to endogenous yeast chromosomes and that these accessible sites are devoid of nucleosomes, effectively forming nucleosome-depleted regions (NDRs) (**Fig. 1B, Fig. S3A,B, S3D-E**). In comparison, we only detected 16 weaker ATAC peaks over the Mmyco chromosome, which were weakly covered but not completely depleted in nucleosomes (**Fig. 1C, Fig. S3C-E**).

The cohesin (Scc1) and RNA polymerase II (Pol. II) profiles along the Mpneumo chromosome also appear similar to wild-type (WT) yeast chromosomes, with discrete Scc1 peaks preferentially located in nucleosome-depleted and Pol. II-enriched regions (**Fig. 1F, Fig. S2F-H**). In contrast, Scc1 and Pol. II binding profiles in the Mmyco strain show significant differences. On the one hand, Scc1 is strongly enriched over the whole Mmyco chromosome, but does not form the distinct peaks observed on endogenous chromosomes (**Fig. 1F, Fig. S2F,G**). This broad enrichment is correlated with a strongly reduced Scc1 occupancy at centromeres of endogenous *S. cerevisiae* chromosomes (**Fig. S2E**), suggesting that the Mmyco chromosome titrates the dynamic cohesin pool enriched over yeast centromeres (*21*, *22*). The Pol. II occupancy is on the other hand greatly reduced along the Mmyco chromosome compared to yeast chromosomes (**Fig. 1F**).

These results show that large exogenous bacterial chromosomes placed in a eukaryotic context spontaneously adopt features of eukaryotic chromatin, as histones, Pol II and cohesins can all bind bacterial DNA. However, the chromatin landscapes over the two bacterial chromosomes delineate two different chromatin archetypes: (1) “Y” (yeast-like) chromatin landscape, found over Mpneumo (whose 40% GC content is close to the native *S. cerevisiae* GC content), and (2) “U” (for Unconventional) chromatin, found over Mmyco (with a low 24% GC content), featuring less packed nucleosomes, a reduced Pol II coverage, and a broad binding of cohesins across the entire chromosome.

Transcriptional activity of bacterial genomes in a yeast context

Consistent with Pol. II ChIP-seq profiles (**Fig. 1F**), we find that the Y chromatin type on Mpneumo is transcribed to levels similar to those of endogenous yeast chromosomes (**Fig. 2A**). Mpneumo transcription tracks are significantly longer than yeast genes (4.9 kb vs 3.4 kb, p-value $< 2e^{-4}$, two-sided Student's t-test) and do not systematically display clear boundaries, with more loci where both strands are transcribed compared to yeast (**Fig. 2B**, black triangle). They do not preferentially occur over bacterial gene bodies nor seem to initiate at bacterial promoters, consistent with an absence of Pol. II enrichment at these loci

(**Fig. S4A-C**). In sharp contrast, the Mmyco U chromatin type is only sparsely and lowly transcribed (**Fig. 2A,B**), again in good agreement with the reduced levels of Pol. II deposition (**Fig. 1F**).

Remarkably, for both bacterial genomes, the steady-state orientation of transcription tracks preferentially follow those of the bacterial genes annotated along these sequences, despite the fact that the transcription machinery evolved independently for billions of years (**Fig. 2C**, see the orientation of genes and stranded tracks in **Fig. 2A,B**). This preference correlates with the over-representation of A vs. T and G vs. C on the coding strand of both yeast and bacterial genes (**Fig. S4D**, see **Discussion**). To test whether the steady-state unidirectional orientation of transcription is the outcome of active degradation of overlapping antisense transcripts, we profiled nascent transcripts by CRAC-seq, a technique that quantifies RNA molecules still bound to RNA Pol II (**Fig. S5A**). We found a strong correlation between RNA-seq and CRAC-seq (**Fig. S5B**) which confirmed that transcription can initiate at the hundreds of accessible NDRs along the Mpneumo chromosome (**Fig. 2D, E**) and that transcription is extremely sparse over the Mmyco chromosome (**Fig. S5A**). We further validated these observations by performing RNA-seq in *Δupf1*, *Δrrp6* or *Δupf1/rrp6* Mpneumo strains, respectively incompetent for nonsense-mediated RNA decay (NMD), nuclear RNA surveillance by the exosome, or both. This showed that although some overlapping antisense transcripts sporadically emerge in the mutant strains, eukaryotic transcription over a bacterial chromosome preferentially follows the direction of the bacterial genes (**Fig. S5C**).

Inactive U chromatin forms a distinct compartment in the interphase nucleus

Chromatin conformation capture (Hi-C) contact maps show that, as expected in a Rabl configuration (*23*, *24*), yeast centromeres added to both Mmyco and Mpneumo chromosomes cluster with native ones. The two bacterial chromosomes exhibit nonetheless very distinct structural characteristics. The Y-type Mpneumo chromosome behaves similarly to an endogenous *S. cerevisiae* chromosome (**Fig. 3A**), with comparable trans contacts (**Fig. S6A**) and a slope of the contact probability (p) as a function of the genomic distance (*s)* (or genomic contact decay curve) close to -1.5, a value corresponding to a typical random coil structure observed in simulations (**Fig. 3C**) (*25*). DNA-FISH labeling of Mpneumo DNA also revealed an extended and contorted structure within the intranuclear space, confirming the intermixing of the chromosome with the actively transcribed native chromosomes (**Fig. 3E, Fig. S6C, D**).

In contrast, few contacts were detected between the U-type Mmyco and the endogenous chromosomes (**Fig. 3A, Fig. S6A**), predominantly with the 32 yeast subtelomeric regions (**Fig. 3A**, dotted rectangles; **Fig. S6B**). The slope of the p(s) curve also differs dramatically with a value of -1 at shorter distances, corresponding to a crumpled globule (**Fig. 3C**)(*26*, *27*). DNA-FISH labeling of Mmyco confirmed the globular conformation of the chromosome and revealed its preferential position at the nuclear periphery, reflecting its proximity with sub/telomeres (**Fig. 3E; Fig. S6C, D**). These results show that inactive, U chromatin spontaneously forms a distinct compartment in the nuclear space segregated from the active endogenous chromosomal set.

Cohesin compacts both exogenous bacterial chromosomes in G2/M

In yeast G2/M-arrested cells, chromatin is folded into arrays of cohesin-anchored loops (*28*, *29*). This compaction results in a decrease of inter-chromosomal contacts and in a shifted p(s) curve, with an inflection point around the average loop length (*23*, *30*). These structural features are observed for both bacterial chromosomes, indicating a similar mitotic compaction of both Y and U chromatin types (**Fig. 3B, 3D, Fig. S6A,E,F**). We identified 59 loops across the Mpneumo chromosome in G2/M, spanning over slightly longer (28 kb) genomic distances than the loops found along yeast sequences (22 kb) (**Fig. S6G**; **Methods**). Strong Scc1 peaks were positioned close (within 500bp) to these loop anchors (**Fig. 3F,H**), at sites of convergent transcription (**Fig. 2B**, green dotted lines, **Fig. S6H**), with strong cohesin peaks associated with sharper transcriptional convergence (**Fig. 2B**, arrows; **Fig. S6I,J**). The dotted grid pattern in Mpneumo is reminiscent of multiple DNA loops observed along mammalian interphase chromosomes (*31*), suggesting that some loop anchors can be involved in loops of various sizes, a phenomenon not observed across native *S. cerevisiae* chromosomes (**Fig. 3F**). On the other hand, no visible discrete loops were observed along the Mmyco chromosome where cohesins uniformly cover inactive U chromatin regions without clear peaks, probably reflecting the absence of (convergent) transcription (**Fig. 3G**, and also **Fig. 1F** and **Fig. 2B**). Cohesin-mediated loops along yeast chromosomes are formed through loop extrusion, a process by which cohesin complexes organize DNA by capturing and gradually enlarging small loops (*32*, *33*). In absence of Wpl1, which impairs cohesin removal and results in longer loops (*28*, *33*), we observe an increase of long-range contacts for all chromosomes, including Mmyco and Mpneumo (**Fig. 3I; Fig. S6K**), consistent with active cohesin-mediated loop extrusion proceeding on both Y and U chromatin.

Mosaic chromosomes display spontaneous, chromatin type-dependent DNA compartmentalization

The structural and functional features of the Y- and U-type chromatin are reminiscent of the euchromatin and heterochromatin compartments described along metazoan chromosomes (*34*). To test whether and how they can coexist on a single chromosome, we fused the Mmyco chromosome with the native yeast chromosome XVI (XVIfMmyco chromosome) and induced translocations to generate two additional strains with chromosomes harboring alternating U and Y chromatin regions over hundreds of kb (XVIfMmycot1), 50 kb (XVIfMmycot2) or as small as 15 kb (XVIfMmycot2') (**Fig. 4A**; **Fig. S7A-B** Methods). The overall Hi-C contact maps of XVIfMmyco strain synchronized in G1 showed little differences compared to the parental Mmyco strain (but for the deleted and fused regions, **Fig. 4B**, compare panel **i** with **ii**). The Y-type chr. XVI arm intermixes with the other 15 yeast chromosomes while the U-type Mmyco arm remains isolated from the rest of the genome, contacting subtelomeric regions (**Fig. 4B,** panel **ii**). In the translocated strains, the alternating chromatin type regions resulted in striking checkerboard contact patterns within XVIfMmycot1 and XVIfMmycot2 chromosomes (**Fig. 4B**, panels **iii** and **iv**). U-type regions can thus make specific contacts over long distances, bypassing the Y-type regions found in between (**Fig. 4C, D**; **Fig. S7E,** G1). Similarly, Y-type regions of chromosome XVI are also involved in specific Y-Y contacts over longer distances (**Fig. 4C, D**). Intra-U and Y regions contact decay is different in the two chromatin types, U chromatin being prone to longer range contacts in *cis* (**Fig. S7E**). Strikingly, translocation of a U-chromatin segment at the end of chromosome XIII also led to a strong increase of inter-chromosomal contacts with U-chromatin segments in *trans* (**Fig. 4E, F**), illustrating that inactive U chromatin can span multiple chromosomes, a finding that broadens our understanding of chromatin compartmentalization and that further relates Y and U-chromatin with eu- and heterochromatin compartments found in higher eukaryotes (*34*).

Pol. II ChIP-seq in XVIfMmycot2 strain revealed that 50 kb-long translocated Y or U chromatin segments did not exhibit any change in Pol. II occupancy profiles, demonstrating that Pol. II binding is independent of the broader chromatin context (**Fig. S7C**). We further performed RNA-seq profiling in the XVIfMycot2' strain, which harbors a 15 kb segment of Mmyco translocated within the euchromatic yeast chromosome XVI and vice versa. Compared to the XVIfMmycot2 strain, the right end of this 15kb Mmyco segment is directly adjacent to yeast chromatin (**Fig. S7D**). In this context, we observed that transcription immediately at the junction with yeast chromatin can progress over ~1 kb into the Mmyco

segment (**Fig. S7D**, black arrow). In contrast, transcription of yeast genes translocated close to Mmyco chromatin is unaffected. This suggests that although active transcription can spread from Y chromatin to Mmyco U chromatin, the U-type chromatin locally exerts a strong inhibitory effect on the transcription initiation machinery.

In larger genomes, euchromatin and heterochromatin compartments are abolished following loop extrusion-mediated metaphase chromosome compaction (*32*, *35*, *36*). Similarly, in our chimeric strains, Hi-C maps of the mosaic chromosomes in G2/M reveal that intra- and inter-chromosomal U-type compartments all disappear upon cohesin-mediated compaction (**Fig. 4E, Fig. S7F**), with cohesin being over-enriched along inactive Mmyco regions independently of the proximity with the centromere (**Fig. S7C**). At this stage, all the chimeric chromosome regions have the same distance-dependent contact frequency (**Fig. S7E**, G2/M), and loops can span over the Y/U chromatin junction, bridging the nearest cohesin enrichment sites on each side (**Fig. S7G**, green arrows) (*37*, *38*).

Compartmentalization of U-type chromatin depends on transcriptional activity of the yeast genome

We investigated the molecular mechanisms that might be responsible for the formation of this U-type heterochromatin-like chromatin compartment. In *S. cerevisiae*, heterochromatin formation does not rely on the canonical H3K9me3 modification found in most eukaryotes (*39*). Instead, heterochromatin is formed and maintained at telomeres, mating-type and rDNA loci by the SIR (Silent Information Regulator) complex, which consists of Sir2 histone deacetylase (HDAC class III), Sir3, a structural chromatin-binding protein stabilizing deacetylated histones and Sir4, another structural protein bridging Sir2 and Sir3 and interacting with nuclear envelope associated proteins (doi: 10.1534/genetics.112.145243). We found that 1) H4K16ac levels are reduced overall on Mmyco chromosome (**Fig. S8A**), 2) that Sir2 inhibitor nicotinamide (NAM) only increases H4 acetylation levels locally at telomeric regions of both yeast and Mmyco chromosomes (**Fig. S8B**) and 3) that Sir3 is not enriched along the Mmyco chromosome (**Fig. S8C**). These observations reveal that U chromatin is hypo-acetylated independently of SIR-driven telomere heterochromatinization. We further investigated whether widespread deacetylation by other histone deacetylases could lead to heterochromatinization of Mmyco. Following treatment with Trichostatin A (TSA, an HDAC I/II inhibitor), we found that H4 acetylation levels increase globally over yeast segments but not over Mmyco segments (**Fig. S8D**) and that the Mmyco U compartment and the p(s) contact decay curves were not affected (**Fig. S8F-H**). This indicates that neither

hypo-acetylation of U-chromatin nor its compartmentalization rely on histone deacetylation activity.

The unusual 10bp-longer NRL measured in the Mmyco chromosome (**Fig. 1D,E**) suggests that Histone 1 (H1) could bind the DNA linkers between consecutive nucleosomes leading to the formation of U-type chromatin. We tested this hypothesis in Hho1 (yeast H1) deleted mutant and found that the longer NRL, the distance-dependent contact frequency and the Mmyco compartment remained unchanged (**Fig. S8E, S8I-K**), showing that H1 is not required for U-type chromatin formation.

Finally, we assessed whether active transcription of yeast chromatin could be responsible for the segregation of the inactive Mmyco chromosome into a separate compartment. We treated cells with thiolutin, a RNA polymerase inhibitor leading to its rapid dissociation from chromatin. Pol. II ChIP-seq profiling confirmed its unloading from yeast chromatin upon treatment (**Fig. S8L**) and calibrated RNA-seq revealed that transcription of yeast genes was reduced overall (**Fig. S8M,** 23-34% reduction in steady-state transcription). Concomitantly, Hi-C revealed that long-range interactions within the yeast genome increased, and that Y/U compartments were strongly affected (**Fig. 4G-I**), suggesting that chromatin compartmentalization is dependent on yeast transcriptional activity. To test this under physiological conditions, we performed Hi-C in quiescent yeast cells, where transcription is largely silenced (*40*). Hi-C revealed that the Mmyco U-type compartment disappears upon entry into quiescence (**Fig. 4J-L**), and that the Mmyco chromatin segregates away from the yeast telomere hypercluster appearing in quiescence (*41*) (**Fig. S8N**), supporting the independence between U-type chromatin and the telomeric compartment.

Taken together, these results show that formation and maintenance of the Mmyco compartment are independent of histone deacetylation or linker histone H1, and that transcription is necessary to segregate U and Y chromatin compartments.

Chromatin features of bacterial, eukaryotic and random exogenous chromosomes can be predicted by their sequence composition

We explored the extent to which intrinsic DNA sequence composition can be predictive of specific chromatin features of foreign DNA integrated in yeast. We trained convolutional neural networks (CNNs) using yeast DNA sequences from chromosomes I to XV as independent variables to predict nucleosome, cohesin and Pol II coverage tracks along these sequences (see **Methods, Fig. S9A**). These models were then validated on the held-out sequence of yeast chromosome XVI (correlation with experimental signals of 0.63, 0.82 and

0.68 for nucleosome, Scc1 ChIP and Pol. II ChIP predictions respectively, **Fig. 5A,** see **Methods**), confirming that the genomic sequence is predictive of chromatin composition and activity in yeast (*42–44*).

Using our CNRR models, we then predicted coverages over Mpneumo and Mmyco chromosomes, and observed that the rules learned from yeast sequences were sufficient to accurately predict many of the features experimentally characterized on these chromosomes. The predictions recapitulate the features of Y chromatin on Mpneumo, but also the U chromatin features on Mmyco sequences, including the increased linker length (**Fig. 5A insert i, 5B-C**), (2) Scc1 over-enrichment with no discrete peaks (**Fig. 5A insert ii**), and (3) the lower unstructured Pol. II coverage, comparable to the background signal over yeast chromosomes (**Fig. 5A insert iii**).

To broaden the spectrum of exogenous DNA origin, we next experimentally characterized the nucleosome coverage and spatial organization of two other YACs integrated in yeast, containing sequences from the *Plasmodium falciparum* eukaryotic genome (91 kb and 58 kb, 18% GC content; Methods) and the other containing a 284 kb-long, 21% GC YAC from *Candidatus Phytoplasma vitis*, a non-mycoides bacterial species (**Fig. S10**). Both bacterial and eukaryotic AT-rich exogenous chromosomes display features of U-type chromatin, including (1) a NRL longer than that of endogenous yeast chromosomes (~182 bp) and (2) segregation into a distinct compartment, with reduced trans-chromosomal contacts and increased contacts with yeast telomeres (**Fig. S10A-E**). Of note, trans-chromosomal contacts between the two *P. falciparum* YACs were frequent (**Fig. S10A**), reminiscent of trans-chromosomal contacts between Mmyco translocated segments (**Fig. 4E**). We found that the increased NRL in these two foreign chromosomes is correctly predicted by the CNN model (**Fig. S10F, G**). The CNN model also accurately predicts nucleosome coverage over an artificial 18kb-long GC-rich sequence with 50% GC for which MNase-seq data in yeast was already generated (*45*).

To further validate the accuracy of the CNN model, we compared Pol. II coverage predictions over several natural and artificial sequences that have been integrated and characterized in yeast (**Fig. S10H**). In particular, the human HPRT1 sequence or its reverse sequence (HPRT1r) (~ 41% GC), and a 18kb-long artificial sequence (~ 50% GC) have been shown to be transcriptionally more active than yeast endogenous chromosomes, when integrated in yeast (*46*). Over these sequences, we predict a higher Pol II occupancy compared to yeast chromosomes, consistent with published experimental results (*45*, *46*).

GC content, dinucleotide content and more complex sequence features influence chromatin composition

To better characterize the relationship between GC content and chromatin composition, we predicted the average nucleosome, Scc1 and Pol. II coverage over thousands of 1kb artificial random sequences with variable GC content (**Fig. 5D**; **Methods**). We observed different behaviors as a function of GC content: (1) the nucleosome signal was higher for intermediate GC % (i.e. between 30% and 50%) and was strongly decreased outside of this range; (2) the cohesin signal continuously decreased with increasing GC%; (3) the Pol. II signal was minimal up to 25% GC, constant up to 45% GC and then increased up to 85% GC. Importantly, these predictions on random sequences accurately recapitulate experimental measurements over 1kb segments from yeast, Mmyco and Mpneumo sequences (**Fig. 5D)**, suggesting that these predicted variations indeed reflect intrinsic properties of the chromatin assembled on DNA with different GC content.

When comparing CNNs performances with two simpler linear regression models, either based on the GC% or on the dinucleotide composition as predictor variables, we found that all models accurately capture dinucleotide signatures for nucleosome, cohesin or Pol. II coverage (e.g. AG, AC, CT, CA, GT, TG in nucleosome) (**Fig. S9B**) but also that the CNN approach performs better in predicting the actual tracks (**Fig. S9C**). A motif analysis based on the saliency computed from the trained networks also allowed us to identify some of the DNA motifs involved in the sequence-dependent chromatin composition (**Fig. S9D**). These results show that besides GC% and dinucleotide signatures, more complex sequence features influence chromatin composition, regardless of their evolutionary origin.

Discussion

Chromatin composition and activity of exogenous chromosomes is based on their underlying DNA sequence

Several studies have shown that exogenous or random DNA segments in the yeast nucleus with a GC% relatively similar to that of yeast chromosomes are actively transcribed (*45–48*). Here we show that this is not always the case, and that DNA sequences with different GC content adopt one of two archetypes of chromatin that we called Y and U for yeast-like and unconventional. Y chromatin displays nucleosome arrays with a canonical 160bp NRL, recruits Pol. II and is transcriptionally active, whereas U chromatin is

characterized by a longer 174 bp NRL 174bp, and a lower Pol. II occupancy and transcriptional activity (**Fig. 5E**).

Using machine learning models solely trained on yeast chromosomes, we show that the chromatin composition experimentally measured over these exogenous chromosomes can be estimated from their DNA sequence only. This demonstrates that the fate of any DNA molecule introduced into a given cellular context, including its nucleosomal packaging and its transcriptional activity, is influenced by sequence-based rules that are both deterministic and predictable (**Fig. 5A, S11**). CNN-based methods can thus be useful to predict the behavior of exogenous DNA in natural gene transfer events or in synthetic genome engineering.

Biased orientation of RNA-seq signal along prokaryotic sequences in eukaryotic context

Transcriptomics signal on bacterial chromosomes in yeast follows, on average, bacteria genes orientation (**Fig. 2B, C**). Sequence determinants, including GC and AT skews, predate the divergence of eucaryotes and procaryotes, and are notably found in both Mpneumo and Mmyco chromosomes. Such determinants are known to influence polymerase directionality (*49*, *50*). Eventually, the resulting conserved orientation could facilitate the domestication of exogenous sequences during horizontal transfer/introgression events between distant species. Whether these RNA molecules are translated, and peptides of bacterial origins exist in the yeast cell, remains to be determined. If so, they could provide a source of diversity and adaptation.

Cohesin-mediated chromatin folding along non-transcribed DNA templates

Cohesins have been proposed to actively extrude DNA loops until they encounter an obstacle and/or a release signal (*6*, *32*). In yeast and other species, these anchors are determined by a combination of convergent transcription, replication fork progression during S phase and/or the presence at these positions of stably bound cohesin promoting sister chromatid cohesion (*51–53*). Here, we show that the barely transcribed Mmyco chromosome is compacted by cohesin at G2/M, without focal loop anchoring, suggesting that transcription is neither necessary for loading nor a primary driver for translocation. We propose that cohesins can move freely along this template without encountering significant obstacles, making it a suitable model for studying potentially blocking sequences or molecules.

Compartmentalization of transcriptionally inactive foreign DNA in host's nucleus

We show that the introduction of a foreign DNA with a lower GC content spontaneously promotes the formation of transcriptionally inactive chromatin which folds into an isolated crumpled globule compartment at the periphery of the budding yeast nucleus (**Fig. 3E; Fig. S6H**), a behavior that mirrors the metazoan compartment "B" formed by inactive, H3K9me3/HP1-mediated heterochromatin, which is absent in yeast. In contrast, we find that chromatin assembled on sequences with a composition similar to that of the host and transcriptionally active adopts a random coil structure and intermixes with transcribed yeast chromosomes (**Fig. 5E**).

We further show that (1) high or low Pol. II occupancy can be accurately predicted in both chromatin archetypes based on their sequence alone, and that (2) thiolutin treatment, a drug that inhibits transcription by dissociating Pol. II from chromatin, reduces the segregation of U and Y chromatin. These findings suggest that the recruitment of the transcriptional machinery, which correlates to the underlying DNA sequence, directly contributes to the physical segregation of adjacent Pol. II-depleted chromatin into a distinct, globular, inactive compartment. This behavior may be explained by the different chromatin composition (e.g., acetylated histone tail residues or intrinsically disordered proteins composing the transcription machinery) in each compartment

These different fates for two foreign chromosomal sequences raise interesting evolutionary considerations in the context of the invasion of the genome by exogenous mobile elements. This could lead either to the spontaneous isolation of inactive foreign DNA or, on the contrary, to the co-option of a set of active sequences that could represent a reservoir of genetic innovations. We show that an active chromatin region as small as 15 kb is not sensitive to the surrounding silenced chromatin environment, suggesting that the harnessing of small genome-attuned regions could readily occur during introgressions or HGT events. This sequence-dependent mechanism may have contributed to the heterochromatinization of AT-rich transposable elements integrated in mammalian genomes.

Acknowledgements

We are very grateful to Michael Lanzer (Heidelberg University) and Cecilia Sanchez for providing us with *P. falciparum* YACs, and to Sebastian Baumgarten for long-read sequencing of these YACs. We also thank Bernard Dujon, Micheline Fromont-Racine, Alain Jacquier, Gianni Liti, Bertrand Llorente, Marcelo Nollmann, Cosmin Saveanu, Benoit le Tallec and all members of the laboratory Régulation Spatiale des Génomes for fruitful comments on the work and the manuscript. We thank Cyril Matthey-Doret and Guillaume Mercy for help during the earlier steps of the project, the Biomics the PICT-IBiSA@Pasteur Imaging Facility of the Institut Curie, a member of the France Bioimaging National Infrastructure (ANR-10-INBS-04), and particularly Mickael Garnier for his help on FISH quantification. **Funding:** This work was supported by the European Research Council under the Horizon 2020 Program (ERC grant agreement 771813) and Agence Nationale pour la Recherche (ANR-19-CE13-0027-02) to RK. RK, FB, JM and AnT also received support from Agence Nationale pour la Recherche (ANR-22-CE12-0013-01). CC was supported by a Pasteur-Roux-Cantarini fellowship. JS was supported by an ARC fellowship. E. Turc and L. Lemée at Biomics Platform, C2RT, Institut Pasteur, Paris, France, are supported by France Génomique (ANR-10-INBS-09-09) and IBISA for processing and sequencing RNA samples.

Author contributions: Conceptualization: CC, LM, JS, JM and RK. Methodology: LM, CC, JS, CL, JM, RK. Software: JS. Validation: CC, JS, LM. Investigation: LM, CC, with contributions from MP, FG, AP, AB, AE, AgT, MR and FB. Formal analysis: JS (all data processing and integration), AW and ER (CNN models), with contributions from CC, LM, FB, MR, AnT. Data Curation: JS, with contributions from CC, LM. Resources: GG, CL. Visualization: JS. Writing - original draft preparation: CC, JS, LM, JM and RK. Writing – Editing: all authors. Writing – revisions: JS, LM, AW, JM, RK. Supervision: JM, RK. CC co-supervised a student. Funding acquisition: RK. Project Administration: RK. **Competing interests:** The authors declare no competing interests. **Data and materials availability:** Sample description and raw sequences for all figures are accessible on GEO database through the following accession number: GSE217022 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE217022, token *arepcamenvafhuz*). All custom-made code of the analysis of sequencing data is available online https://github.com/koszullab/. Open-access versions of the programs and pipeline used (Hicstuff) are available online on the github account of the Koszul lab: Hicstuff (https://github.com/koszullab/hicstuff) version 3.1.2, Chromosight (version 1.4.1 available online at https://github.com/koszullab/chromosight/), Bowtie2 (version 2.4.5 available online

at http://bowtie-bio.sourceforge.net/bowtie2/), SAMtools (version 1.9 available online at http://www.htslib.org/), Bedtools86 (version 2.29.1 available online at https://bedtools.readthedocs.io/en/latest/content/installation.html) and Cooler (versions 0.8.7–0.8.11 available online at https://cooler.readthedocs.io/en/latest/).

Fig. 1. Chromatin composition of bacterial chromosomes integrated in yeast.

A, Schematic representation of the conversion from circular to linear chromosomes integrated in yeast. The purple and blue colors represent the *M. pneumoniae* (Mpneumo) and *M. mycoides* (Mmyco) bacterial sequence in all figures, respectively. Right: distribution of GC% for 1kb windows over yeast chromosome XVI, *M. pneumoniae* and *M. mycoides* chromosomes.

B, ATAC-seq (orange, CPM), nucleosomal track (gray, see **Methods**) and H2A and H3 ChIP-seq (shades of blue, IP vs input, log2) profiles obtained in the Mpneumo strain (*S. cerevisiae* + *M. pneumoniae*). 10kb-long genomic windows from the chromosome XVI (left: 250-260kb) and the Mpneumo chromosome (right: 710-720kb) are shown at the same scale. Nucleosome-depleted regions are highlighted in yellow.

C, Same as **B** in the Mmyco strain (*S. cerevisiae* + *M. mycoides*). 10kb-long genomic windows from chromosome XVI (left: 250-260kb) and the Mmyco chromosome (right: 989-999kb) are shown at the same scale.

D, Frequency of nucleosome linker DNA length in Mpneumo and Mmyco strains. For each strain, the distribution is calculated for each chromosome separately. The dashed line indicates 14 bp and the dotted line indicates 25 bp. The x axis only shows linker DNA lengths in the 0-100 bp range.

E, Nucleosomal track centered on nucleosome peaks for *S. cerevisiae*, Mpneumo and Mmyco. The Y axis represents the average nucleosomal track (130-165 bp MNase-seq fragments, resized to 40bp, piled-up and normalized to sequencing depth).

F, Scc1 (red) and RNA Pol II (green) ChIP-seq profiles (IP vs input, log2) obtained in the Mpneumo (left) or in the Mmyco (right) strains. For each strain, 60kb-long genomic windows from the bacterial chromosome and yeast chromosome XVI are shown at the same scale. GC% in sliding 1kb windows is shown below the ChIP-seq profiles.

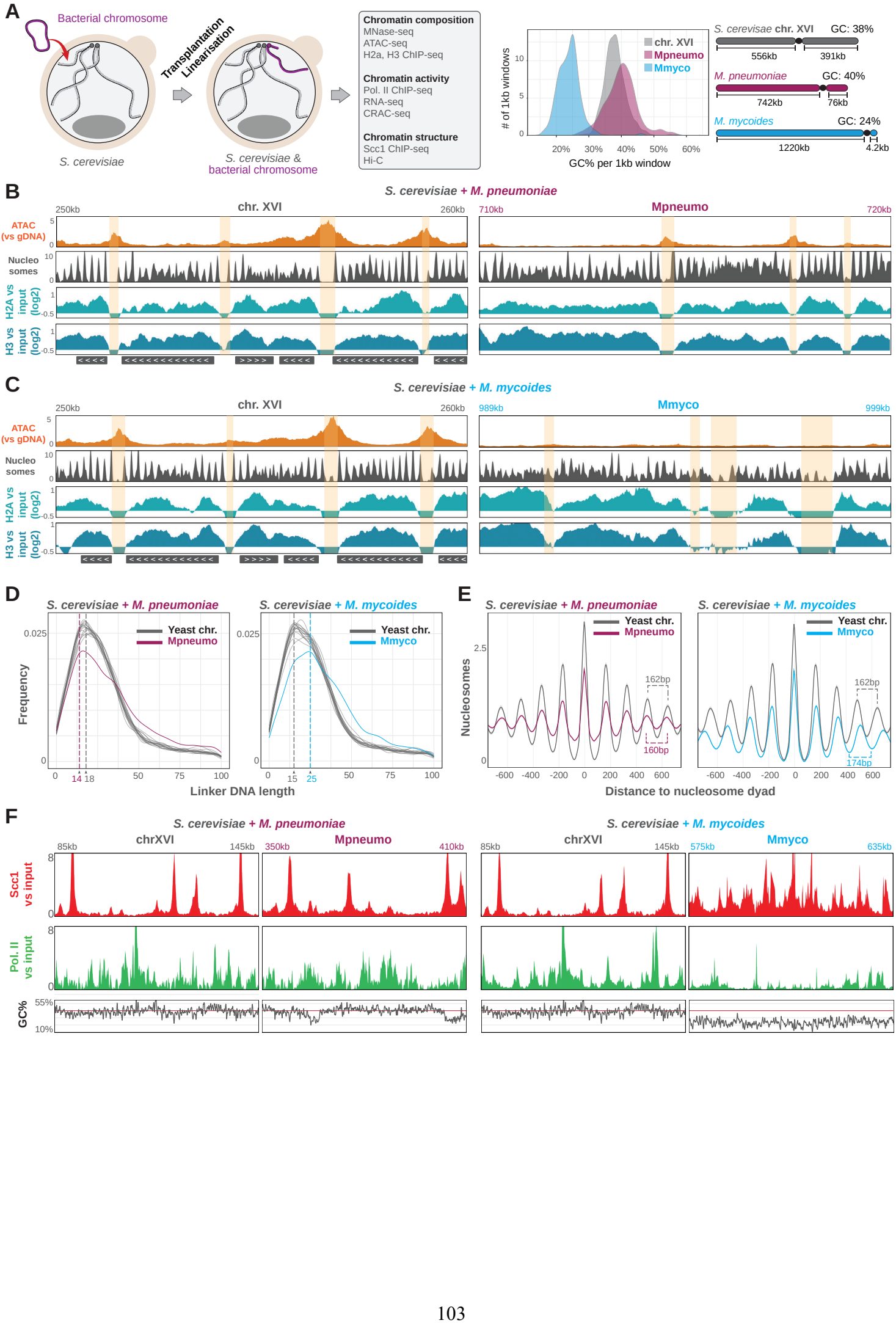

A
Bacterial chromosome
Transplantation
Linearisation
S. cerevisiae
S. cerevisiae &
bacterial chromosome
Chromatin composition
MNase-seq
ATAC-seq
H2a, H3 ChIP-seq
Chromatin activity
Pol. II ChIP-seq
RNA-seq
CRAC-seq
Chromatin structure
Scc1 ChIP-seq
Hi-C
# of 1kb windows
chr. XVI
Mpneumo
Mmyco
GC% per 1kb window
S. cerevisae chr. XVI GC: 38%
556kb
391kb
M. pneumoniae GC: 40%
742kb
76kb
M. mycoides GC: 24%
1220kb
4.2kb
B
S. cerevisiae + M. pneumoniae
chr. XVI
Mpneumo
ATAC (vs gDNA)
Nucleo somes
H2A vs input (log2)
H3 vs input (log2)
C
S. cerevisiae + M. mycoides
chr. XVI
Mmyco
D
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
Yeast chr.
Mpneumo
Mmyco
Frequency
Linker DNA length
E
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
Nucleosomes
162bp
160bp
174bp
Distance to nucleosome dyad
F
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
chrXVI
Mpneumo
Mmyco
Scc1 vs input
Pol. II vs input
GC%

Fig. 2. Transcriptional activity of exogenic bacterial sequences in budding yeast.

A, Stranded RNA-seq profiles along either yeast chromosome XVI or bacterial chromosomes Mpneumo and Mmyco. Forward (pink) and reverse (turquoise) genes along yeast or bacterial sequences are indicated as transparent segments under the tracks. Pink and turquoise represent forward and reverse transcription, respectively.

B, Top: stranded RNA-seq profiles along a 60kb window along yeast chromosome XVI, Mpneumo or Mmyco. Bottom: Scc1 (cohesin) deposition profiles of the corresponding loci. Green dotted lines indicate identified loci of convergent transcription (see Methods). Black triangle: overlapping bidirectional transcription.

C, Forward and reverse RNA-seq coverage of forward- and reverse-oriented yeast or bacterial genes. Scores are normalized by each genomic feature length.

D, Top: stranded CRAC-seq profiles along the region shown in **B**. Bottom: corresponding ATAC-seq profile. Dotted lines point at sites of bidirectional initiation along Mpneumo.

E, Metaplot of 3kb regions centered on ATAC peaks over yeast or Mpneumo chromosomes, of stranded RNA-seq (left panels) or CRAC-seq (right panels) profiles.

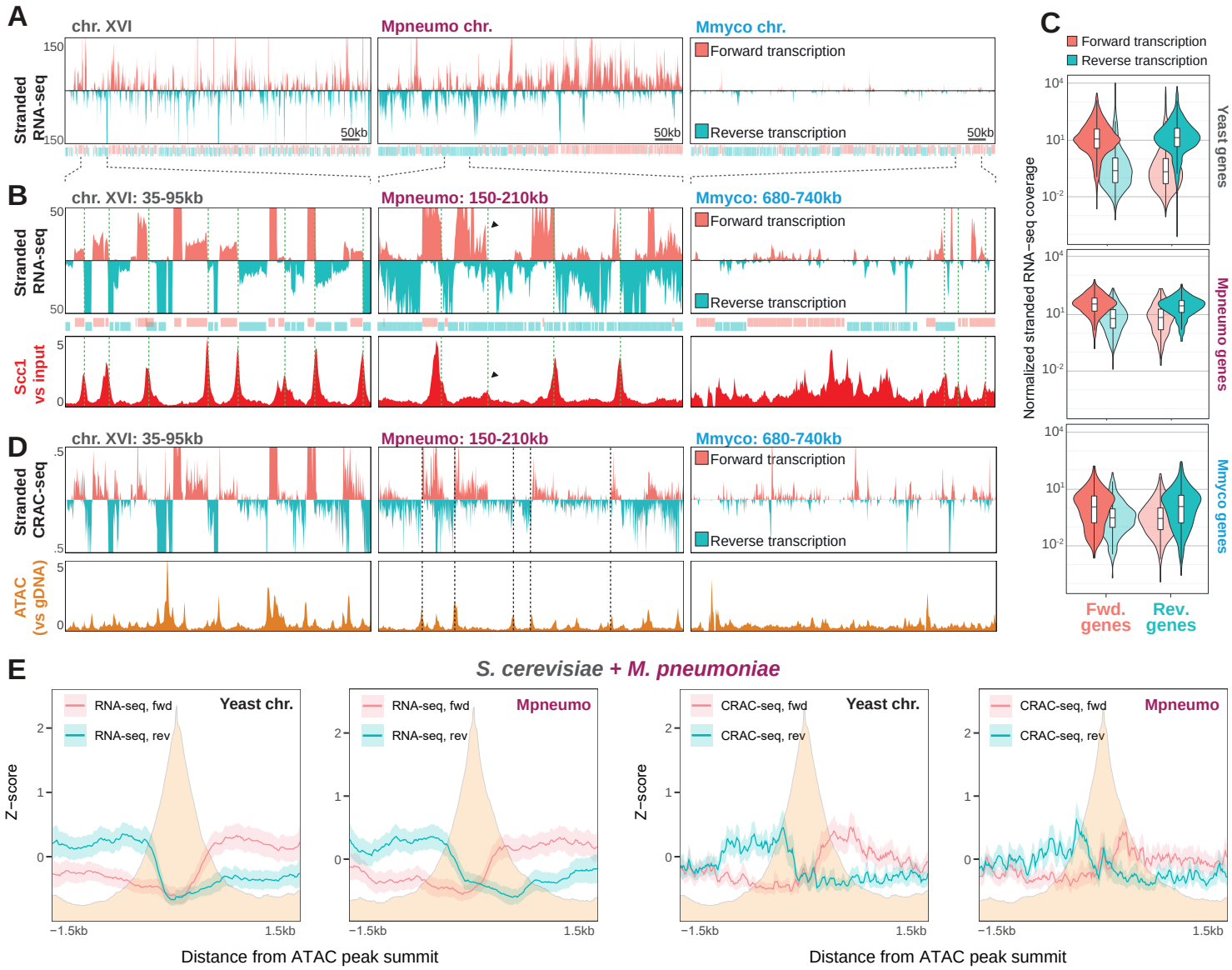

A
chr. XVI
Mpneumo chr.
Mmyco chr.
Stranded RNA-seq
Forward transcription
Reverse transcription
50kb
B
chr. XVI: 35-95kb
Mpneumo: 150-210kb
Mmyco: 680-740kb
Scc1 vs input
C
Normalized stranded RNA−seq coverage
Yeast genes
Mpneumo genes
Mmyco genes
Fwd. genes
Rev. genes
D
Stranded CRAC-seq
ATAC (vs gDNA)
E
S. cerevisiae + M. pneumoniae
RNA-seq, fwd
RNA-seq, rev
CRAC-seq, fwd
CRAC-seq, rev
Yeast chr.
Mpneumo
Z−score
−1.5kb
1.5kb
Distance from ATAC peak summit

Fig. 3. Spatial folding of exogenic bacterial sequences within the yeast nucleus.

A, B, Hi-C contact maps of representative endogenous and of Mmyco and Mpneumo bacterial chromosomes in G1 (**A**) and G2/M (**B**) (4kb resolution).

C, D, Contact frequency (p) as a function of genomic distance (s) plots of endogenous yeast chromosomes (long arms) and of Mmyco and Mpneumo bacterial chromosomes in G1 (**C**) and G2/M (**D**).

E, Left: FISH imaging. Representative field of either (top) Mpneumo or (bottom) Mmyco fixed cells labeled with DAPI (left panel) and hybridized with a fluorescent probe generated from either the Mpneumo or Mmyco chromosome, respectively. Right: For each probe, number of patches detected per nucleus and surface occupied by these patches relative to the whole nucleus surface (Methods).

F, G, Top: magnification of 150kb windows from Hi-C contact maps in G2/M from either an endogenous or the bacterial chromosome in Mpneumo (**F**) and Mmyco (**G**) strains (1kb resolution). Bottom: Scc1 ChIP-seq deposition profile. Black arrowheads: loops. Cyan diamonds in Mpneumo Hi-C contact map: Scc1 peaks positions reported on the contact map diagonal. Inset in each map: *Chromosight* pileup of contacts between cohesin enrichment peaks along either S. cerevisiae or bacterial chromosomes (see **Methods**).

H, Left: distance between chromatin loop anchors and the nearest Scc1 peak in Mpneumo (with and without a random shuffle of peak positions). Right: Scc1 peak strengths in yeast or Mpneumo chromosome, near (< 1kb) or outside loop anchors (p-values from two-sided Student's t-test).

I, Distance-dependent contact frequency in endogenous yeast and in bacterial chromosomes (left, Mpneumo; right, Mmyco), in WT (dashed) and in Δwpl1 mutants.

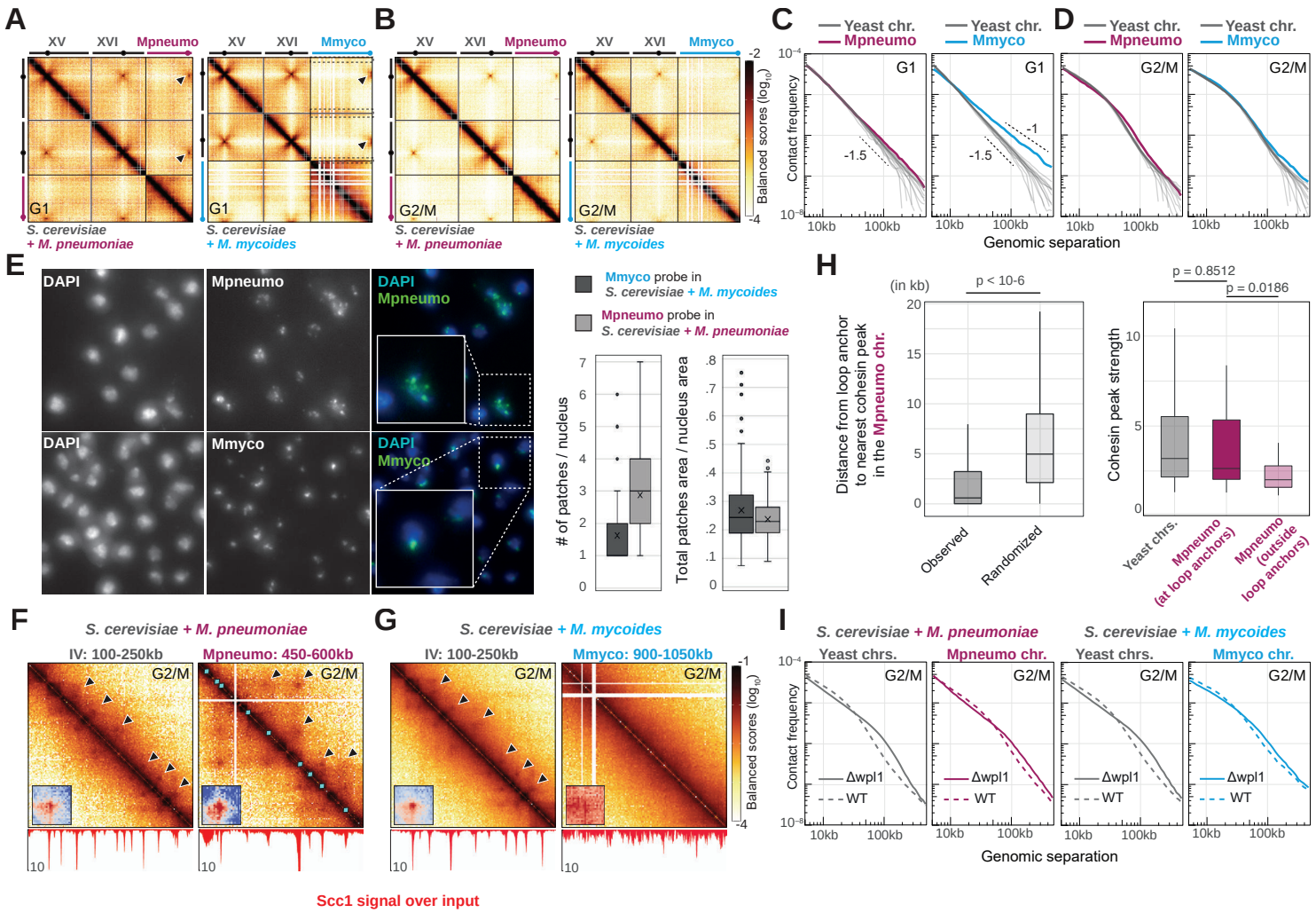

A
B
C
D
XV
XVI
Mpneumo
Mmyco
G1
G2/M
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
Balanced scores (log10)
Yeast chr.
Contact frequency
-1.5
-1
10kb
100kb
Genomic separation
E
DAPI
Mpneumo
Mmyco
Mmyco probe in S. cerevisiae + M. mycoides
Mpneumo probe in S. cerevisiae + M. pneumoniae
# of patches / nucleus
Total patches area / nucleus area
H
(in kb)
p < 10-6
p = 0.8512
p = 0.0186
Distance from loop anchor to nearest cohesin peak in the Mpneumo chr.
Cohesin peak strength
Observed
Randomized
Yeast chrs.
Mpneumo (at loop anchors)
Mpneumo (outside loop anchors)
F
G
I
IV: 100-250kb
Mpneumo: 450-600kb
Mmyco: 900-1050kb
Yeast chrs.
Mpneumo chr.
Mmyco chr.
Δwpl1
WT
Scc1 signal over input

Fig. 4. Compartmentalization of mosaic chromosomes composed of Y and U-type chromatin.

A, Schematic representation of the CRISPR strategy used to generate the Mmyco mosaic chromosomes with alternating Y and U chromatin regions.

B, Top: G1 Hi-C contact maps of chr. XV, XVI, and bacterial chromosomes in the Mmyco strain, and in translocated derivatives from panel **A**) (4kb resolution). Bottom: correlation matrices of the corresponding contact maps.

C, Virtual 4C profiles of viewpoints (indicated as gray/blue arrows) located within yeast segments (in gray) or Mmyco chromosome segments (in blue) of the chimeric chromosome XVI, in XVIfMmyco (top), XVIfMmycot1 (middle) and XVIfMmycot2 (bottom) strains.

D, Quantification of long-range interactions (800kb-1.5Mb) between pairs of distant Mmyco segments (blue), pairs of distant yeast segments (gray) and interactions between a Mmyco segment and a yeast segment (light gray), in double-translocation XVIfMmyco (top), XVIfMmycot1 (middle) and XVIfMmycot2 (bottom) strains.

E, Left: Schematic representation of the CRISPR strategy used to generate the mosaic XVIfMmycot3 strain with alternating yeast and Mmyco segments in chromosomes XIII and XVI. Middle: G1 Hi-C maps of mosaic chromosomes XIII and XVI in XVIfMmycot3 strain. Same color scales as in **B**.

F, Virtual 4C profiles of viewpoints located within yeast segments (gray arrows) or Mmyco segments (blue arrows) of the chimeric chromosomes XIII and XVI, in the XVIfMmycot3 strain.

G, Correlation matrices of the contacts in chr. XV and XVIfMmycot1 in G1, after the addition of DMSO (left) or thiolutin (right). Same color scale as in **B**.

H, Contact frequency (p) as a function of genomic distance (s), for contacts in yeast segments (gray) or Mmmyco segments (blue) of the chimeric chromosome XVIfMmycot1, in G1 after the addition of DMSO (solid) or thiolutin (dotted).

I, Derivatives of curves from **H**.

J, Chr. XV and XVIfMmycot1 correlation maps in G1 or in quiescence. Same color scale as in **B.**

K, Same p(s) as in H, but in G1 and quiescence.

L, Derivatives of curves from **K**.

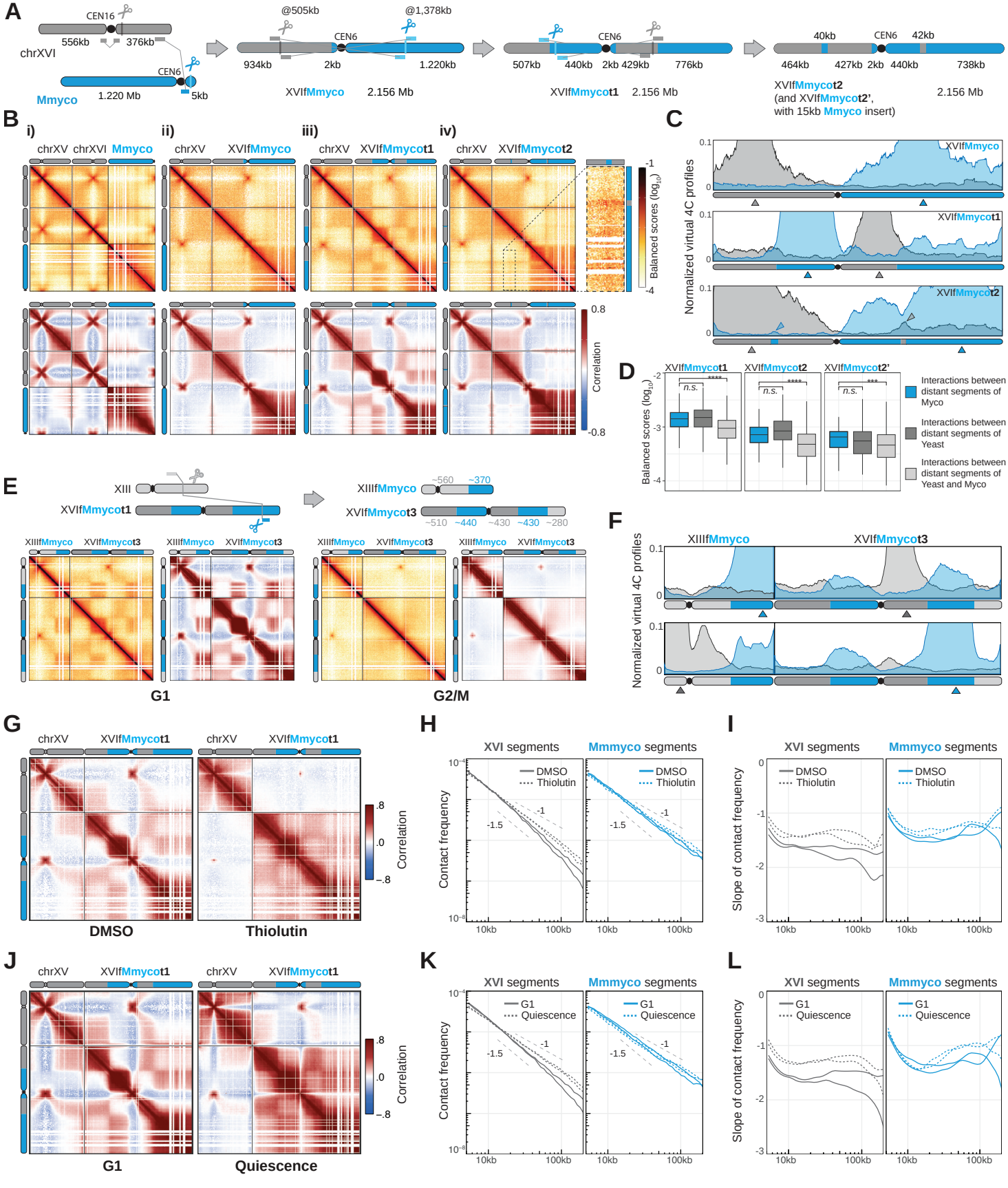

A
CEN16
556kb
376kb
chrXVI
CEN6
1.220 Mb
5kb
Mmyco
@505kb
@1,378kb
CEN6
934kb
2kb
1.220kb
XVIfMmyco 2.156 Mb
CEN6
507kb
440kb
2kb 429kb
776kb
XVIfMmycot1 2.156 Mb
40kb
CEN6
42kb
464kb
427kb 2kb 440kb
738kb
XVIfMmycot2
(and XVIfMmycot2',
with 15kb Mmyco insert)
2.156 Mb
B
i)
chrXV chrXVI Mmyco
ii)
chrXV XVIfMmyco
iii)
chrXV XVIfMmycot1
iv)
chrXV XVIfMmycot2
Balanced scores (log10)
Correlation
C
Normalized virtual 4C profiles
XVIfMmyco
XVIfMmycot1
XVIfMmycot2
D
XVIfMmycot1
XVIfMmycot2
XVIfMmycot2'
Balanced scores (log10)
n.s.
****
***
Interactions between distant segments of Myco
Interactions between distant segments of Yeast
Interactions between distant segments of Yeast and Myco
E
XIII
XVIfMmycot1
XIIIfMmyco
~560 ~370
XVIfMmycot3
~510 ~440 ~430 ~430 ~280
XIIIfMmyco XVIfMmycot3
G1
G2/M
F
Normalized virtual 4C profiles
XIIIfMmyco
XVIfMmycot3
G
chrXV XVIfMmycot1
Correlation
DMSO
Thiolutin
H
XVI segments
Mmmyco segments
DMSO
Thiolutin
Contact frequency
-1
-1.5
10kb
100kb
I
XVI segments
Mmmyco segments
Slope of contact frequency
J
chrXV XVIfMmycot1
Correlation
G1
Quiescence
K
XVI segments
Mmmyco segments
G1
Quiescence
Contact frequency
L
XVI segments
Mmmyco segments
Slope of contact frequency

Fig. 5. DNA sequence is sufficient to predict Y/U chromatin composition

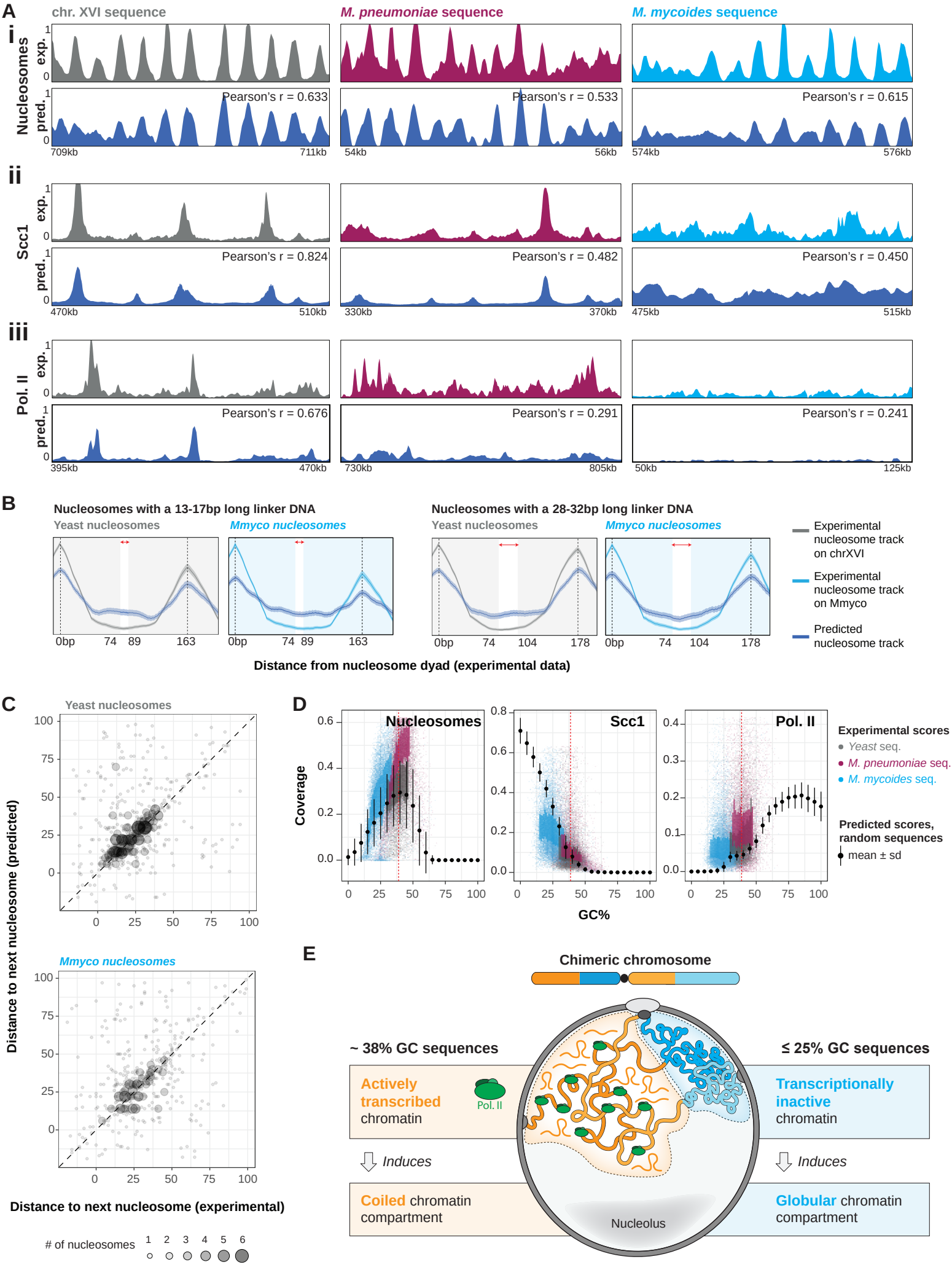
A, Experimental (top) and NN prediction (bottom, dark blue) of nucleosome tracks or Scc1 and Pol. II ChIP-seq coverage tracks, over yeast chromosome XVI or over Mmyco and Mpneumo bacterial chromosomes in yeast. Signals have been smoothed using a sliding genomic window of 10 bp (nucleosome tracks) or 500 bp (ChIP-seq). For each chromosome, the Pearson correlation score was computed between experimental and predicted scores, using averaged scores over non-overlapping 10 bp bins for nucleosomes or 500 bp bins for ChIP-seq. Bins with an average score lower than 0.01 were excluded.

B, Nucleosome tracks (gray or light blue: experimental; dark blue: predicted), aligned at experimental dyads of yeast or Mmmcyo nucleosomes (position 0 on the X axis). Nucleosomes are grouped by the length of their linker DNA (experimentally computed, and represented by a red double-helix separating consecutive nucleosomes). Dotted lines indicate the position of nucleosome dyads (centered at 0 bp and 163 or 178 bp), and shaded areas represent nucleosome-occupied regions (± 74 bp around dyads).

C, Correlation between predicted and experimental linker DNA lengths, for yeast (left) and Mmmcyo (right) nucleosomes. Only nucleosomes whose dyads are aligned ≤ ± 2 bp between experimental and predicted nucleosome track are considered.

D, Nucleosome, Scc1 and Pol. II ChIP-seq average predicted scores in 2kb (nucleosome) or 30kb-long (ChIP-seq) sequences. Scores predicted from random sequences with varying GC content are shown as mean ± sd in black, and average scores predicted from chromosome sequences of individual genomes are shown as colored points. Average experimental scores in 100 bp sequences along yeast, Mmyco and Mpneumo chromosomes are also shown as gray, blue or purple dots (scores of the middle half for each GC% unit are in bold).

E, Schematic of a chimeric chromosome composed of alternating Y- and U-type chromatin.


A
chr. XVI sequence
M. pneumoniae sequence
M. mycoides sequence
i
Nucleosomes
exp.
pred.
Pearson's r = 0.633
Pearson's r = 0.533
Pearson's r = 0.615
709kb
711kb
54kb
56kb
574kb
576kb
ii
Scc1
Pearson's r = 0.824
Pearson's r = 0.482
Pearson's r = 0.450
470kb
510kb
330kb
370kb
475kb
515kb
iii
Pol. II
Pearson's r = 0.676
Pearson's r = 0.291
Pearson's r = 0.241
395kb
470kb
730kb
805kb
50kb
125kb
B
Nucleosomes with a 13-17bp long linker DNA
Nucleosomes with a 28-32bp long linker DNA
Yeast nucleosomes
Mmyco nucleosomes
0bp
74
89
104
163
178
Experimental nucleosome track on chrXVI
Experimental nucleosome track on Mmyco
Predicted nucleosome track
Distance from nucleosome dyad (experimental data)
C
Yeast nucleosomes
Mmyco nucleosomes
Distance to next nucleosome (predicted)
Distance to next nucleosome (experimental)
# of nucleosomes
1
2
3
4
5
6
D
Nucleosomes
Scc1
Pol. II
Coverage
GC%
Experimental scores
Yeast seq.
M. pneumoniae seq.
M. mycoides seq.
Predicted scores, random sequences
mean ± sd
E
Chimeric chromosome
~ 38% GC sequences
≤ 25% GC sequences
Actively transcribed chromatin
Pol. II
Transcriptionally inactive chromatin
Induces
Induces
Coiled chromatin compartment
Globular chromatin compartment
Nucleolus

Supplementary Materials for

Sequence-dependent activity and compartmentalization of foreign DNA in a eukaryotic nucleus

Léa Meneu[1,2,Ψ], Christophe Chapard[1,Ψ,&], Jacques Serizay[1,Ψ,*], Alex Westbrook[2,3], Etienne Routhier[2,3,4], Myriam Ruault[5], Manon Perrot[1,2], Alexandros Minakakis[8], Fabien Girard[1], Amaury Bignaud[1,2], Antoine Even[5], Agnès Thierry[1], Géraldine Gourgues[6], Domenico Libri[8], Carole Lartigue[6], Aurèle Piazza[1,#], Angela Taddei[5], Frédéric Beckouët[7], Julien Mozziconacci[3,4,9*] and Romain Koszul[1,*]

Correspondence to: romain.koszul@pasteur.fr, julien.mozziconacci@mnhn.fr, jacques.serizay@pasteur.fr

This PDF file includes:

Material and Methods

Strains and medium culture conditions

All yeast strains used in this study are derivatives of W303 or VL6-48N and are listed in the Table "Strain list" (**Table S1**). The original Mpneumo and Mmyco yeast strains carry the circular genomes of *Mycoplasma mycoides* subsp. *mycoides* strain PG1 (*16*) and *Mycoplasma pneumoniae* strain M129 (*17*), respectively. Yeast strains were grown overnight at 30°C in 150mL of suitable media to attain $4{,}2x10^8$ cells. Cells were grown in a synthetic complete medium deprived of histidine (SC -His) (0.67% yeast nitrogen base without amino acids) (Difco), supplemented with a mix of amino-acids, 2% glucose) or in in rich medium (YPD): 1% bacto peptone (Difco), 1% bacto yeast extract (Difco), and 2% glucose. Yeast cells were synchronized in G1 by adding α-factor (Proteogenix, WY-13) in the media every 30 min during 2h30 (1 µg/mL final). To arrest cells in metaphase, cells were washed twice in fresh YPD after G1 arrest and released in rich medium (YPD) containing Nocodazole (Sigma-Aldrich, M1404-10MG) during 1h30. Cell synchronization was verified by Flow-cytometry.

Isolation of quiescent cells from stationary phase cultures

Quiescent cells were obtained after isolation from density gradients as described in (*54*) with a few modifications. Cells were first pre-cultured in YPD overnight and diluted the next day at 0.1 OD600nm /ml in an aerated flask with a ratio 1/10 of the flask volume. Cells were grown for 7 days at 30°C shaking at 250rpm. To establish the density gradients, 4,5 mls of a pre-mixed 150 mM NaCl / 90% Percoll (Sigma-Aldrich #P1644) solution were added to a 5 mls open-top thinwall ultra-clear Beckman Coulter tube and centrifuge in the MSL-50 rotor at 25000g for 18 minutes (Accel 4, Decel 9). To isolate the Q cells, 500 ODs of stationary phase cells were washed in 10 mls of 50mM Tris buffer pH7.5, resuspended in 500 µl before loading on 4 pre-established density gradients. The gradients were then run at 1000g for 30 minutes. The lower fraction (750 µl) was collected for each gradient, pooled and washed twice with 10 mls of 50mM Tris buffer pH7.5 before resuspension in a 1 ml final volume. 100 ODs Q cells when then fixed for Hi-C experiments.

Drug treatments

For HDAC inhibition, nicotinamide (NAM) (Sigma-Aldrich, N3376) or trichostatin A (Sigma-Aldrich, TSA) (T8552) were added to cell cultures at a final concentration of 5 mM final and 10 µM final during 4 hours. We then synchronized the cells in G1, and cultures were fixed for Hi-C and ChIP-seq.

For transcription inhibition, after synchronization in G1, thiolutin (Abcam, ab143556) was added to cell cultures at a final concentration of 20 µg ml-1 for 30 min in the RSGY1136 strain. We collected 5 mL of culture for the RNA-seq and the rest was fixed for Hi-C. *Candida glabrata* strain is used for spike-in normalization in RNA-seq.

CRISPR-Cas9 engineering

We used a CRISPR–Cas9 strategy to linearize the circular bacterial chromosome present in the parental yeast strains. Note that the highly acrocentric structure of Mmyco corresponds to the only position that was cleaved among several tested to linearize the chromosome. We suspect that due to the high AT content the CRISPR-Cas9 targeting remains relatively poorly efficient. Plasmids pML107 (*55*) or PAEF5 (*56*) carrying gRNA were co-transformed with 200 bp DNA repair recombinant donor sequences carrying telomeric repeats seeds (Genscript Biotech Netherlands) (**Fig. S1B**).

We also applied a CRISPR-Cas9 approach to concomitantly fuse chromosome XVI with the linear bacterial chromosomes and remove chromosome XVI centromere, as described in Luo et al. (*57*). The resulting chromosome (XVIfMmyco) has one bacterial DNA arm carrying U chromatin and one yeast DNA arm made of Y chromatin (**Fig. 4A**). Strain XVIfMmyco's karyotype was verified using pulsed-field gel electrophoresis (**Fig. S6A**), and that it grows normally was verified by growth assay (**Fig. S6B, C**). Briefly, we used three gRNAs inserted in pGZ110 and pAEF5 (*56*) co-transformed with recombinant DNA donors sequences (Twist Biosciences). These donor sequences are designed so that upon recombination 1) centromere XVI is removed and 2) fusion takes place between subtelomeres of chromosome XVI and either Mmyco and Mpneumo right arms (**Fig. 4A**). The same strategy was applied to generate reciprocal translocations between chromosome XVI and Mmyco sequences of the XVIfMmyco chromosome (**Fig. 4A**). We generated alternating domains of Y and U chromatin along the XVIfMmyco chromosome using CRISPR-induced reciprocal translocations between the two arms, resulting in strains XVIfMmycot1 and t2 carrying alternating regions of U and Y type chromatin along a chromosome (**Fig. 4A**). The chromatin composition of these chimeric chromosome consists of [TEL - 500kb Y - 700kb U - CEN - 400kb Y - 500kb U - TEL] for XVIfMmycot1 and of [TEL - 500kb Y - 50kb U - 350kb Y - CEN - 650kb U - 50kb Y - 500kb U - TEL] for XVIfMmycot2.

gRNA sites were chosen to optimize CRISPR targeting specificity using chopchop.cbu.uib.no (*58*) or CRISPOR (*59*). For all experiments, 100 ng of gRNA expression plasmids and ~400 ng of recombinant donors were co-transformed. After co-transformation, yeast cells were immediately plated on the corresponding selective media. Yeast cells were allowed to recover in fresh YPD for 2h before plating.

Linearization, fusion and translocations of chromosomes were verified by PCR, pulse-field gel electrophoresis and Hi-C.

All gRNA sequences, DNA donor sequences and PCR primers used are provided in **Tables S2-4.**

Liquid growth assays and segregational plasmid stability

Growth rate of parental, Mmyco and Mpneumo strains carrying linear or fused chromosomes were assessed for independent clones in triplicates. Briefly, independent clones were grown

overnight in selective medium (SC -HIS) to saturation at 30°C. The following morning the cultures were diluted to OD600 = 0.01 and inoculated in 96 well plates containing 100µl fresh SC -HIS. The cells were grown under agitation in a Tecan Sunrise plate reader at 30°C (Tecan). optical densities at 600 nm were recorded every 10 min to generate the growth curves for individual wells. The average of triplicates was plotted to compute the doubling time.

Segregational plasmid stability was measured as described in (*60*). Three individual transformants for each strain were inoculated into SC-HIS. P1 is the percentage of bacterial chromosome-carrying cells in selective media. It was determined from the ratio of viable colonies obtained by plating on YPD plates then replicated in selective medium (SC -HIS). To measure the chromosome stability, cultures were diluted into non-selective media and grown for X generations (g). After approximately 12 generations, the percentage of cells containing the bacterial chromosome (P2) was also determined from the ratio of viable colonies on rich medium (YPD) then replicated in selective medium. We then used P1, P2 and (g) number to calculate the segregation rate (m), which is defined as the % of plasmid-free segregants appearing in the final population after a single doubling.

Pulsed-field Gel Electrophoresis and Southern blot hybridization

Agarose plugs containing yeast chromosomes were prepared as described (*61*) and separated by clamped homogeneous electric field gel electrophoresis Rotaphor (Biometra) using the following parameters. Gel: 1% (SeaKemGTG); t=12 °C; buffer: 0.25 × Tris-Borate-EDTA; Program: [140 V, switch time: 300 s to 100 s, run time: 70 h].

DNA from the PFGE was transferred on a membrane. Southern blot was performed using digoxigenin-labeled DNA probes as described (*62*). We used the HIS3 gene (YOR202w) present near the Mmyco and Mpneumo centromere was used as target to validate the linearization and the fusion. Digoxigenin-labeled DNA probes were synthesized from the pRS413 plasmid with the following oligos: 3'CTACATAAGAACACCTTTGG5' and 3'ATGACAGAGCAGAAAGCCCT5'.
3'ATGACAGAGCAGAAAGCCCT5'.

Hi-C procedure and sequencing

Cell fixation with 3% formaldehyde (Sigma-Aldrich, Cat. F8775) was performed as described in Dauban et al. (*63*). Quenching of formaldehyde with 300 mM glycine was performed at 4°C for 20 min. Hi-C experiments were performed with a Hi-C kit (Arima Genomics) with a double DpnII + HinfI restriction digestion following manufacturer instructions. Samples were purified using AMPure XP beads (Beckman A63882), recovered in 120ul H2O and sonicated using Covaris (DNA 300bp) in Covaris microTUBE (Covaris, 520045). Biotinylated DNA was loaded on Dynabeads™ Streptavidin C1 (FISHER SCIENTIFIC, 10202333). Preparation of the samples for paired-end sequencing on an Illumina NextSeq500 (2x35 bp) was performed using Invitrogen TM Collibri TM PS DNA Library Prep Kit for Illumina and following manufacturer instructions.

ChIP-seq

The ChIP-seq protocol is described in (*64*). Experimental replicates (x3) were made for each condition. Briefly, cells of either *S. cerevisiae* or *Candida glabrata* were grown exponentially to OD600 = 0.5. 15 OD600 units of *S. cerevisiae* cells were mixed with 3 OD600 units of *C. glabrata* cells to a total volume of 45 mL for Scc1 calibration only. Cells were fixed using 4mL of fixative solution (50 mM Tris-HCl, pH 8.0; 100 mM NaCl; 0.5 mM EGTA; 1 mM EDTA; 30% (v/v) formaldehyde) for 30 min at room temperature (RT) with rotation. The fixative was quenched with 2mL of 2.5M glycine (RT, 5 min with rotation). The cells were then harvested by centrifugation at 3,500 rpm for 3 min and washed with ice-cold PBS. The cells were then resuspended in 300 mL of ChIP lysis buffer (50 mM HEPES KOH, pH 8.0; 140 mM NaCl; 1 mM EDTA; 1% (v/v) Triton X-100; 0.1% (w/v) sodium deoxycholate; 1 mM PMSF; 2X Complete protease inhibitor cocktail (Roche)) and transferred in 2mL tubes containing glass beads (ozyme, P000913-LYSK0-A.0) before mechanical cells lysis. The soluble fraction was isolated by centrifugation at 2,000 rpm for 3min then transferred to sonication tubes (Covaris milliTUBE 1ml, 520135) and samples were sonicated to produce sheared chromatin with a size range of 200-1,000bp using a Covaris sonicator. After sonication the samples were centrifuged at 13,200 rpm at 4°C for 20min and the supernatant was transferred into 700 µL of ChIP lysis buffer. 80 µL (27 µl of each sample) of the supernatant was removed (termed 'whole cell extract [WCE] sample') and store at -80°C. For Scc1-PK, 5ug of antibody (ab27671, Abcam) was added to the remaining supernatant. For Pol. II, 5ug (05-952-I, Merck Millipore), H2A 2uL (AB_2687477, Active motif), H4K16ac 2uL (07-352, Sigma-Aldrich), and 4ug of Sir3 (a polyclonal ab gift from L. Pillus, University of California, San Diego). The supernatant is then incubated overnight at 4°C (wheel cold room). 50 µL of protein G Dynabeads or protein A Dynabeads was then added and incubated at 4°C for 2h. Beads were washed 2 times with ChIP lysis buffer, 3 times with high salt ChIP lysis buffer (50mMHEPES-KOH, pH 8.0; 500 mM NaCl; 1 mM EDTA; 1% (v/v) Triton X-100; 0.1% (w/v) sodium deoxycholate;1 mM PMSF), 2 times with ChIP wash buffer (10 mM Tris-HCl, pH 8.0; 0.25MLiCl; 0.5% NP-40; 0.5% sodium deoxycholate; 1mM EDTA;1 mMPMSF) and 1 time with TE pH7.5. The immunoprecipitated chromatin was then eluted by incubation in 120 µL TES buffer (50 mMTris-HCl, pH 8.0; 10 mM EDTA; 1% SDS) for 15min at 65°C and the supernatant is collected termed 'IP sample'. The WCE samples were mixed with 40µL of TES3 buffer (50 mM Tris-HCl, pH 8.0; 10 mM EDTA; 3% SDS). All (IP and WCE) samples were de-cross-linked by incubation at 65°C overnight. RNA was degraded by incubation with 2µL RNase A (10 mg/mL) for 1h at 37°C. Proteins were removed by incubation with 10µL of proteinase K (18 mg/mL) for 2h at 65°C. DNA was purified by a phenol/Chloroform extraction. The triplicate IP samples were mixed in 1 tube and libraries for IP and WCE samples were prepared using Invitrogen TM Collibri TM PS DNA Library Prep Kit for Illumina and following manufacturer instructions. Paired-end sequencing on an Illumina NextSeq500 (2x35 bp) was performed. Libraries were performed in one or two biological replicates. When available, the duplicates were averaged for visualization. We calculated pairwise Pearson correlation scores between replicates and all showed high concordance.

RNA-seq

RNA was extracted using MN Nucleospin RNA kit and following manufacturer instructions. Directional mRNA library (rRNA removal) was prepared by the Biomics platform of Institut Pasteur, Paris and Paired-end sequencing (PE150) was performed by Novogene. Libraries were performed in three biological replicates and the triplicates were averaged for visualization. We calculated pairwise Pearson correlation scores between replicates and all showed high concordance.

CRAC-seq

CRAC-seq was performed essentially as described (*65*, *66*). Briefly, 2 L of yeast cells containing Rpb1 tagged with a his6-TEV-proteinA (HTP) tag were grown in CSM-Trp medium to an OD600 of 0.6. 5% of *S. pombe* cells also containing an HTP tagged version of Rpb1 were added to the culture before crosslinking. Cells were exposed to UV light using a W5 UV crosslinking unit (UVO3 Ltd) for 50 seconds, harvested and resuspended in TN150 buffer (50 mM Tris pH 7.8, 150 mM NaCl, 0.1% NP-40 and 5 mM beta mercaptoethanol, 2.4 ml/g of cells) with protease inhibitors. Cells were broken using a Mixer Mill MM 400. Extracts were treated for one hour at 25°C with DNase I (165 U/g of cells) to solubilize chromatin and then clarified by centrifugation (20 min at 20000g at 4°C). Rpb1-RNA complexes were purified by a two-step procedure using the protein A and the his-6 portion of the tag, with TEV cleavage for eluting complexes at the proteinA-dependent purification step. Adaptors were added during the purification steps. The protein-RNA adduct was further purified by denaturing PAGE and the RNA recovered by proteinase K degradation and phenol extraction. The library was amplified by 10 cycles of PCR after reverse transcription and sequenced using Illumina technology

MNase-seq

Each strain was grown to 10^7 cells (OD600 of 0,8~) in 150 mL of SC -his medium at 30°C. Cells were fixed with 4 mL of formaldehyde 37% (1% final) for 20 minutes at room temperature. Cross-linking was stopped by adding 8 mL of 2.5M glycine (125mM final) for 30 minutes. The fixed cells were centrifuged, washed twice with cold phosphate-buffered saline (PBS 1X) and stored at -80°C. Once thawed, 900 µL of cells lysate obtained with a Precellys (Bertin Technologies) were recovered. 100 µL of 1X Micrococcal Nuclease Reaction Buffer (NEB, M0247S) and 5 µL of Bovine Serum Albumin (BSA, 20mg/mL) were added. The mix was divided into 10 tubes x 100 µL. 1 µL of Mnase enzyme (NEB, M0247S) at a concentration of 2,000,000 gel units/ml was added (2,000 units/sample final) in each tube. Samples were incubated at 37°C for varying times (0, 1, 2, 3, 5, 10, 15, 20, 40, and 60 minutes). 300 µL of Stop solution (10 µL of 0.5M EGTA pH 8.0, 30 µL of 20% SDS, 240 µL of H2O, and 20 µL Proteinase K 20mg/mL) was added to each tube. The tubes were then gently mixed and incubated at 65°C overnight. Samples were extracted with phenol/chloroform, and DNA was ethanol precipitated treated with DNase-free RNase. To evaluate the Mnase digestion kinetics, DNA samples were analyzed using an Agilent TapeStation. DNA was purified using 2.2X volume of AmpureXP beads and sequencing

libraries prepared with Invitrogen TM Collibri TM PS DNA Library Prep Kit for Illumina following manufacturer instructions. Paired-end sequencing on an Illumina NextSeq2000 (2 x 50 bp) was performed**.**

ATAC-seq

Independent clones for each strain were inoculated into SC–HIS for overnight culture at 30 °C. Saturated overnight cultures were diluted to an OD600 of 0.1 and cultured for 6 h at 30 °C, until OD600 reached ~0.6. Around 1-5 × 10^6 cells were taken from each culture, pelleted at 3,000g for 5 min, washed twice with spheroplasting buffer (1 M sorbitol, 40 mM HEPES-KOH pH 7.5, 10 mM MgCl2), resuspended in 200 µl spheroplasting buffer with 10 µL of Zymolyase (10 mg/mL), then incubated for 30 min at 37 °C. Spheroplasts were washed twice with 500 µl spheroplasting buffer then resuspended in 50 µl 1× TD buffer with 2.5 µL of TDE1 (Illumina 20034197). Tagmentation was performed for 30 min at 37 °C, 800 rpm with a thermomixer, and DNA was purified using the DNA Clean and Concentrator 5 kit (Zymo Research D4004). PCR was performed using Phusion Master Mix, 11 total cycles. Paired-end sequencing on an Illumina NextSeq2000 (2 x 50 bp) was performed.

Replication MFA experiment

Genomic DNA was prepared from asynchronous and G1 arrested cells in triplicates using Qiagen DNeasy Kit. Pellets were recovered, washed with cold 70% ethanol, air dried and dissolved in 50µl 1xTE. 100 ng of soluble gDNA was transferred to sonication tubes and samples were sonicated to produce sheared chromatin with a size of about 300bp using a Covaris sonicator. Samples were purified using AMPure XP beads (Beckman A63882). Preparation of the samples for paired-end sequencing on an Illumina NextSeq500 (2x35 bp) was performed using Invitrogen TM Collibri TM PS DNA Library Prep Kit for Illumina and following manufacturer instructions.

Processing of reads

Hi-C processing

Reads were aligned and contact maps generated and processed using Hicstuff (https://github.com/koszullab/hicstuff). Briefly, pairs of reads were aligned iteratively and independently using Bowtie2 (*67*) in its most sensitive mode against their reference genome. Each uniquely mapped read was assigned to a restriction fragment. Quantification of pairwise contacts between restriction fragments was performed with default parameters: uncuts, loops and circularization events were filtered as described in (*68*). PCR duplicates (defined as multiple pairs of reads positioned at the exact same position) were discarded. Pairs were binned at 1kb resolution and multi-resolution balanced contact maps (in mcool format) were generated using cooler (*69*)**.**

MNase-seq processing

For each timepoint, bowtie2 was used to align paired-end MNase-seq data on the appropriate genome reference. Only concordant pairs were retained, and fragments with a mapping quality lower than 10 were discarded. PCR duplicates were removed using samtools. Fragment coverage was normalized by library depth (CPM) and converted into a genomic track (bigwig) using deepTools (*70*).

RNA-seq processing

Bowtie2 was used to align paired-end RNA-seq data on the appropriate genome reference. Only fragments with an insert size shorter than 1kb were retained. Only concordant pairs were retained, and fragments with a mapping quality lower than 10 were discarded. PCR duplicates were removed using samtools (*71*). Fragment coverage was normalized by library depth (CPM) and converted into stranded or unstranded genomic tracks (bigwig) using deepTools.

CRAC-seq processing

CRAC-seq datasets were processed and aligned on the appropriate genome reference using the pyCRAC script as described (*66*). Only reads longer than 20 nt were mapped on the *M. mycoides* and *M. pneumoniae* chimeric genomes and only reads longer than 40 nt were mapped to the *S. pombe* genome for the spike-in to avoid inter-species mapping. Signals mapping to the chimeric genomes were normalized using the *S.pombe* spike-in. Homopolymers (fragments with 6+ identical nt) and poly-adenylated reads were discarded.

ATAC-seq processing

Bowtie2 was used to align paired-end ATAC-seq data on the appropriate genome reference. Only fragments with an insert size shorter than 1kb were retained. Only concordant pairs were retained, and fragments with a mapping quality lower than 10 were discarded. PCR duplicates were removed using samtools. Fragment coverage was normalized by library depth (CPM) and converted into a genomic track (bigwig) using deepTools. MACS2 2.2.7.1 was used to call peaks from ATAC-seq data.

ChIP-seq processing

For standard ChIP-seq, bowtie2 was used to align paired-end data on the appropriate genome reference. Only fragments with an insert size shorter than 1kb were retained. Only concordant pairs were retained, and fragments with a mapping quality lower than 10 were discarded. PCR duplicates were removed using samtools. When available, the input was similarly processed. Fragment coverage was normalized by library depth (CPM) and converted into a genomic track (bigwig) using deepTools. Input-normalized fragment coverage was also generated, when possible, using bamCompare from deepTools with "--scaleFactorsMethod readCount". For histone ChIP-seq experiments, IP tracks were divided by input and log2-scaled. For calibrated Scc1 ChIP-seq, CPM-normalized fragment coverages were also multiplied by the ORi factor for calibration ($WCE_{glabrata}$ x $IP_{cerevisiae}$ / $WCE_{cerevisiae}$ x $IP_{glabrata}$, in which $WCE_{glabrata}$ and $IP_{glabrata}$ correspond to the number of paired reads that mapped uniquely

on *C. glabrata* genome and same for *S. cerevisiae* reads). MACS2 2.2.7.1 was used to call peaks using the input alignment files as control.

Analysis of genome-wide assays

All downstream analysis steps were performed in R ≥ 4.1 / Bioconductor ≥ 3.16 (REF), using in-house scripts, unless mentioned otherwise.

Data visualization

Linear tracks were plotted in R using tidyCoverage (*72*).

Coverage heatmaps and aggregated profiles (average +/- 95% CI and heatmap) over windows centered at genomic features (e.g. ChIP-seq peaks or TSSs) were plotted after averaging coverage over 1bp-moving 200bp-wide rolling windows, using tidyCoverage.

All Hi-C contact, ratio and correlated maps were plotted with a log10, a log2 or a linear scaling respectively, using the "plotMatrix" function from HiContacts (*73*).

RNA-seq analysis

To estimate changes in transcript abundance of yeast genes (or bacterial Mmmyco genes) between two yeast strains with double translocations (eg XVIfMycot2 and XVIfMycot2'), *featureCounts* was first used to count raw number of stranded RNA-seq reads mapping to these genomic features of interest, using the corresponding rearranged gene annotation files (*74*). *DESeq2* was then used to identify yeast/Mmmyco genes that were differentially expressed between two chimeric strains (*75*).

MNase-seq analysis

To generate nucleosome tracks, "nucleosomal" fragments (between 130 and 165 bp) from the different MNase-seq timepoints were merged together then resized to a fixed 40bp length centered at the dyad. The resulting coverages were normalized to account for different numbers of nucleosomal fragments sequenced in the different MNase-seq timecourse experiments.

Nucleosome tracks were subsequently used to identify positioned nucleosomes using an approach adapted from (*20*). Briefly, we used a greedy algorithm to identify nucleosome dyads sequentially based on the magnitude of the nucleosome track scores. Each chromosome was segmented into 50 bp-overlapping, 147 bp-long bins. For each bin, the position that had the greatest score from the nucleosome track (see above) was annotated as a nucleosome dyad. The greatest score from the nucleosome track was then iteratively re-calculated within genomic bins re-centered to that position. The algorithm stopped when the local maximum was found for every genomic bin. All the nucleosome dyads were then recovered as the central position of the set of unique genomic bins. The linker DNA length was inferred as the distance between consecutive nucleosome dyads - 147bp.

ChIP-seq analysis

Median Scc1 coverage was calculated over 20 kb windows centered at yeast centromeres or over 300 bp non-overlapping windows tiling the entire yeast chromosomes (excluding centromeres or Scc1 peaks).

Correlation scores between ChIP-seq replicates were calculated from the average coverage scores over 100bp tiled windows of yeast chromosomes.

Sequence biases

10-bp periodicity of dA, dT or dW dinucleotides in sequences up to 160 bp was computed over windows centered at yeast TSSs or tiling bacterial chromosome sequences, using "getPeriodicity" function from periodicDNA (*76*), using ushuffle (*77*) to maintain a constant dinucleotide frequency in shuffled control sequences. K-mer occurrences were estimated over 147 bp windows centered at yeast TSSs or tiling bacterial chromosome sequences. AT (and GC) skews were calculated using the sequence of each DNA strand, in 10bp-wide bins.

Transcription convergence and Scc1 binding

At every position *i* along the genome (every 10 bp), a Dir_i directionality score was calculated as follows:

$$Dir_i = \frac{\sum_{k=i}^{i+200} (RNA_{fwd,k} - RNA_{rev,k})}{\sum_{k=i-200}^{i} (RNA_{fwd,k} - RNA_{rev,k})}$$

Genomic positions *j* for which $Dir_{j-101} > 0$ & $Dir_{j-100} < 0$ were then recovered and correspond to convergent or divergent transcription positions. For every transcription switch position *k*, a convergence score $Conv_k$ was subsequently computed as follows:

$$Conv_k = \left[\sum_{l=k-2000}^{k} (RNA_{fwd,l} > 20) - \sum_{l=k-2000}^{k} (RNA_{rev,l} > 20) + \sum_{l=k}^{k+2000} (RNA_{rev,l} > 20) \right.$$

Thus, $Conv_k$ scores range between -50 and 50 and a positive (negative) $Conv_k$ corresponds to a local genomic position k of convergent (divergent) transcription. The relationship between $Conv$ convergence scores and average Scc1 coverage scores (over a 500bp window centered at the convergent position) was then computed.

Replication MFA profile

Replication profiles were analyzed using Repliscope version 1.1.1 as in (*78*).

Loop detection

Chromosight 1.3.1 (*79*) was used to call loops de novo from contact maps binned at 1 kb and balanced with Cooler (*69*). Matrices were subsampled to contain the same total number of

contacts. De novo loop calling was computed using the "detect" mode of Chromosight, with minimum loop length set at 2kb, percentage undetected set at 25 and pearson correlation threshold set at 0.315. Loop strength was quantified for each loop using the quantify mode of Chromosight and the mean loop score was calculated for each condition. Loop pile-ups of averaged 17kb windows were generated with Chromosight.

Cis-trans ratio and P(s) analysis

Cis-trans ratios were calculated using the "cisTransRatio" function from HiContacts. Contact probability as a function of genomic distance P(s) were determined using pairs files (generated by hicstuff) and the "distanceLaw" function from HiContacts with default parameters, averaging the contact data by individual chromosomes. Briefly, the P(s) were computed by binning interactions according to their genomic distance, using logarithmic-sized distance bins. The number of interactions within each bin was then divided by the bin width and by the total number of interactions, to normalize for varying bin widths and for sequencing depth.

Virtual 4C profiles

Virtual 4C profiles for 20kb-wide viewpoints were computed from 2kb-binned contact maps, using the "virtual4C" function from HiContacts (*73*). Contacts of entire chromosomes across the entire genome were manually computed using subsetting functions from HiContacts.

Deep-Learning analysis

Models architectures and training

Three different models were trained to learn either on the nucleosome, Pol. II and Scc1 profiles from the underlying genomic sequence. The signals were pretreated as follows. For ChIP profiles we first filtered outlier values, converted the signal in count per million and aggregated the different replicates. We computed the ratio between IP and INPUT and discarded all regions for which either profile was equal to zero. Then for all data, we truncated the experimental profiles to a threshold corresponding to the 99th percentile of the profile distribution. We then divided all values in the truncated profile by the maximum value to get a normalized signal between 0 and 1. For each position along the genome, a DNA sequence of length W was associated with a subset of n_out values from the corresponding profile to make the three datasets used for our deep learning framework. For each of these datasets, we used the yeast chromosomes I to XIII for training, XIV and XV for validation and XVI for test.

We implemented all the CNNs using the Keras library (*80*) and Tensorflow (*81*) as back-end. A RTX 2080 Ti GPU was used to improve the training speed. We used the adaptive moment estimation (ADAM) optimization method to compute adaptive learning rates for each parameter. The batch size was set to 512 (Scc1) or 1024 (nucleosome, Pol. II).

For the nucleosome prediction task, our CNN architecture was similar to the one used in (*82*). It consists of three convolutional layers with respectively 64, 16 and 8 kernels of shape

(3x1x4), (8x1x 64) and (80x1x16). A max pooling layer of size 2 and a ReLu activation function was applied after each of these three convolutions. Batch normalization and a dropout of 0.2 was applied after each convolution and the convolution stride was set to 1. Our model takes inputs of shape (2001, 1, 4), the last dimension representing the four nucleotides, and outputs a single value (i.e. n_out =1) corresponding to the value of the nucleosome profile in the middle of the input sequence.

For the Pol. II and Scc1 tasks, the architecture was modified as follows to take into account longer range influences (*83*). It consists of three convolutional layers all with kernels of shape (12x1x4), (8x1x64) and (80x1x16). Max Pooling layers of respective lengths of 8, 4 and 4 followed by ReLu activation were applied after each convolution. Four dilated convolution layers were then applied using 16 kernels of length 5 and dilatation values of respectively 2,4,8 and 16. For Pol. II, the input sequence length was set to 2048 bp and the output length n_out to 16, corresponding to the 16 values of the signal found every 128 bp over the 2,048 bp input sequence. For Scc1, the input sequence length was set to 32768 bp and the output length n_out was 256, corresponding to the 256 values of the signal found every 128bp over the 32,768 bp input sequence.

The loss function used is the sum of the Pearson's dissimilarity (1-correlation) between the prediction x and the target y and the mean absolute error (MAE) between them (loss = MAE(x, y) + 1 - corr(x, y)). This loss function has been previously shown to enable both a faster and a more accurate convergence (*82*).

An early stopping procedure was applied during training to prevent models from overfitting. The loss function was calculated on the validation set at every epoch to evaluate the generalisability of the model. The training procedure was stopped if the validation loss did not decrease at all for 6 epochs and the model parameters were set back to their best performing value. The training procedure usually lasted for 15 to 20 epochs.

Predictions

Predictions were done on all chromosomes, including the chromosomes from the training, validation and test sets as well as exogenous bacterial chromosomes.

In order to characterize the influence of the GC content of the sequences on the overall profile heights, we generated 10000 random 2 kb (nucleosome, Pol II) or 32 kb (Scc1) sequences and predicted each of the three profiles on these sequences.

Comparison with experimental data

For each chromosome, the Pearson correlation score was calculated between experimental measurements and predictions, using averaged scores over non-overlapping 10 bp bins for nucleosomes or 500 bp bins for ChIP-seq. Bins with an average experimental or predicted score lower than 0.01 were excluded from the correlation analysis.

When comparing nucleosome repeat length (NRL) obtained from experimental or predicted nucleosomes tracks, only the nucleosomes whose dyads are aligned $\leq \pm$ 2 bp between experimental and predicted nucleosome track were considered

Identification of nucleotide motifs influencing CNN-based predictions

To identify relevant nucleotide motifs for prediction, we computed the gradients of the model outputs with respect to the input (i.e. the saliency). In the case of models with multiple outputs, corresponding to different locations in a genomic window, we computed the gradient only for one output, corresponding to the center of the genomic window. We used the gradient correction method from Majdanzic et al. to compute attribution maps for all possible windows of the *S. cerevisiae* genome (*84*). Then, for each bp we averaged the absolute values of their attributions computed in all the genomic windows that contain this bp. This resulted in a genome-wide signal of unsigned attribution score for every bp. We computed z-scores for the attribution score and selected bases with a z-score of 4 or higher, to identify regions of high importance in the genome. We then grouped these bases into regions when the genomic distance between these bp was strictly smaller than 20, discarded any remaining solitary bases and extended selected regions by +/- 10bp on their side. As such there were no overlaps between regions, which varied in length from 22bp to 200bp-1000bp (depending on the model). We got 12125, 4446 and 11295 regions of interest respectively for nucleosome, polymerase and cohesin models. We then extracted the sequences in each region and used them as input to the meme suite tools (v5.5.5) (*85*). We ran XSTREME with default parameters, against 5 databases (Yeastract, SwissRegulon S.cer, macisaac, SCPD and uniprobe yeast) using the dinucleotide distribution of the W303 genome as background. This resulted in 6 to 8 motifs discovered per model, most of them matching existing motifs in yeast databases.

Imaging and analysis

Two-dots assay (SCC)

Strains yLD126-36c, FB176 and FB200 were inoculated and grown overnight in SC-MET medium. The next day cultures were diluted to OD600=0.2 in SC-MET. After 3h of exponential growth, alpha factor (10 ul at 5mg/ml) was added every 30 min for 2 h. G1-arrested cells were released into YPD (plus 2 mM methionine), and aliquots sampled for FACS and imaging analysis.

FISH experiments

FISH experiments were performed as described in Gotta et al. (*86*), with some modifications (*87*). The probes were obtained by direct labeling of the bacterial DNA (1.5 µg) using the Nick Translation kit from Jena Bioscience (Atto488 NT Labeling Kit), the labeling reaction was performed at 15°C for 90 min. The labeled DNA was purified using the Qiaquick PCR purification kit from Qiagen, eluted in 30 µl of water. The purified probe was then diluted in the probe mix buffer (50% formamide, 10% dextran sulfate, 2× SSC final). 20 OD of cells (1 OD corresponding to 107 cells) were grown to mid–logarithmic phase (1–2 × 107 cells/ml) and harvested at 1,200 g for 5 min at RT. Cells were fixed in 20 ml of 4% paraformaldehyde for 20 min at RT, washed twice in water, and resuspended in 2 ml of 0.1 M EDTA-KOH pH 8.0, 10 mM DTT for 10 min at 30°C with gentle agitation. Cells were then collected at 800 g,

and the pellet was carefully resuspended in 2 ml YPD - 1.2 M sorbitol. Next, cells were spheroplasted at 30°C for 10 minutes with Zymolyase (60 µg/ml Zymolyase-100T to 1 ml YPD-sorbitol cell suspension). Spheroplasting was stopped by the addition of 40 ml YPD - 1.2 M sorbitol. Cells were washed twice in YPD - 1.2 M sorbitol, and the pellet was resuspended in 1 ml YPD. Cells were put on diagnostic microscope slides and superficially air dried for 2 min. The slides were plunged in methanol at -20°C for 6 min, transferred to acetone at -20°C for 30 s, and air dried for 3 min. After an overnight incubation at RT in 4× SSC, 0.1% Tween, and 20 µg/ml RNase, the slides were washed in H20 and dehydrated in ethanol 70%, 80%, 90%, and 100% consecutively at -20°C for 1 min in each bath. Slides were air dried, and a solution of 2× SSC and 70% formamide was added for 5 min at 72°C. After a second step of dehydration, the denatured probe was added to the slides for 10 min at 72°C followed by a 37°C incubation for 24h in a humid chamber. The slides were then washed twice in 0.05× SSC at 40°C for 5 min and incubated twice in BT buffer (0.15 M NaHCO3, 0.1% Tween, 0.05% BSA) for 30 min at 37°C. For the DAPI staining, the slides were incubated in a DAPI solution (1µg/ml in 1× PBS) for 5 minutes and then washed twice in 1× PBS without DAPI.

FISH image analysis

Images were acquired on a wide-field microscopy system based on an inverted microscope (Nikon TE2000) equipped with a 100×/1.4 NA immersion oil objective, a C-mos camera and a Spectra X light engine lamp for illumination (Lumencor, Inc). The system is driven by the MetaMorph software (Molecular Devices). The axial (z) step is 200 nm and images shown are maximum intensity projection of z-stack images (MIP). Quantifications were done on the MIP images using ilastik for segmentation and Fiji for analyses of the particles.

Data availability

Sample description and raw sequences for all figures are accessible on GEO database through the following accession number: GSE217022. Go to:

https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE217022

Enter token *arepcamenvafhuz* into the box.

Rearranged genome reference sequences for all strains generated in this study are provided as a Zenodo archive (https://doi.org/10.5281/zenodo.7198985).

Source data are provided with this paper.

Code availability

All custom-made code of the analysis of sequencing data is available online https://github.com/koszullab/.

Open-access versions of the programs and pipelines are available online on the github account of the Koszul lab: Hicstuff (https://github.com/koszullab/hicstuff, version 3.1.2) and Chromosight (version 1.4.1, https://github.com/koszullab/chromosight/) (*79*). Bowtie2 (version 2.4.5) is available online at http://bowtie-bio.sourceforge.net/bowtie2/, SAMtools (version 1.9) is available online at http://www.htslib.org/, Bedtools86 (version 2.29.1) is available online at https://bedtools.readthedocs.io/en/latest/content/installation.html and Cooler (versions 0.8.7–0.8.11) is available online at https://cooler.readthedocs.io/en/latest/) (*69*). tidyCoverage (version 1.1.2) is available at https://github.com/js2264/tidyCoverage, HiContacts (version 1.5.0) is available at https://github.com/js2264/HiContacts (*73*).

All deep learning codes are available at:

https://github.com/Alexwestbrook/bacterial_genome

Any additional information, including custom-made code required to reanalyze the data reported in this paper, is available from the lead contact upon request.

Supplementary Text

Replication and pairing of artificial chromosomes

DNA sequence composition and chromatin factors, including nucleosomes, are essential drivers of the replication timing, prompting us to investigate how this timing is established in Y and U chromatin **(*88*)**. In yeast, replication initiates at the level of discrete, small autonomous replication sequences (ARS) positioned along all chromosomes. To investigate the replication timing profile of both bacterial chromosomes, which did not spontaneously evolve to contain these sequences, we used Repliscore (Methods)**(*78*)**. Mpneumo Repliscore profile exhibits DNA copy number variation comparable to that observed along chrXII, indicative of early or mid-S phase firing of multiple replication origins (**Fig. S1C**). This copy number is anti-correlated with GC%, consistent with AT-rich regions replicating earlier. In contrast, DNA copy number variation along Mmyco unveils the early firing of only the centromere-proximal ARS which was artificially integrated during chromosome assembly (**Fig. S1C**). The average copy number of the rest of the Mmyco chromosome appears flat and is not anti-correlated with the GC%. Thus, the replication of U-type chromatin occurs later during the S phase despite its AT-rich sequence composition. These observations confirm the predominant role of chromatin composition in the replication timing, and indicate a heterochromatin-like effect exerted by the U-type chromatin. This pattern is reminiscent of the random and late-replication pattern displayed by the human inactive chromosome X **(*89*)**.

DNA replication and SCC are closely related as cohesion is established during S phase through entrapment of sister DNA molecules by cohesin rings as the replication fork progresses **(*90*)**. We therefore investigated SCC by performing image analysis of chromosome pairing (**Methods**; **Fig. S1D**). Despite the strong enrichment in Scc1 deposition, SCC in Mmyco appears significantly reduced, in agreement with the flat replication pattern suggesting a late and random initiation of the replication process. On the other hand, the endogenous yeast chromosomes in Mmyco strain did not display a significant decrease in SCC (**Fig. S1D**), despite the apparent loss of cohesin at centromeres.

Figure S1. Characterization of Mpneumo and Mmyco strains metabolism.

A, Dinucleotide over-representation (ρ*(XY)) in yeast chromosome XVI, Mpneumo and Mmyco chromosomes.

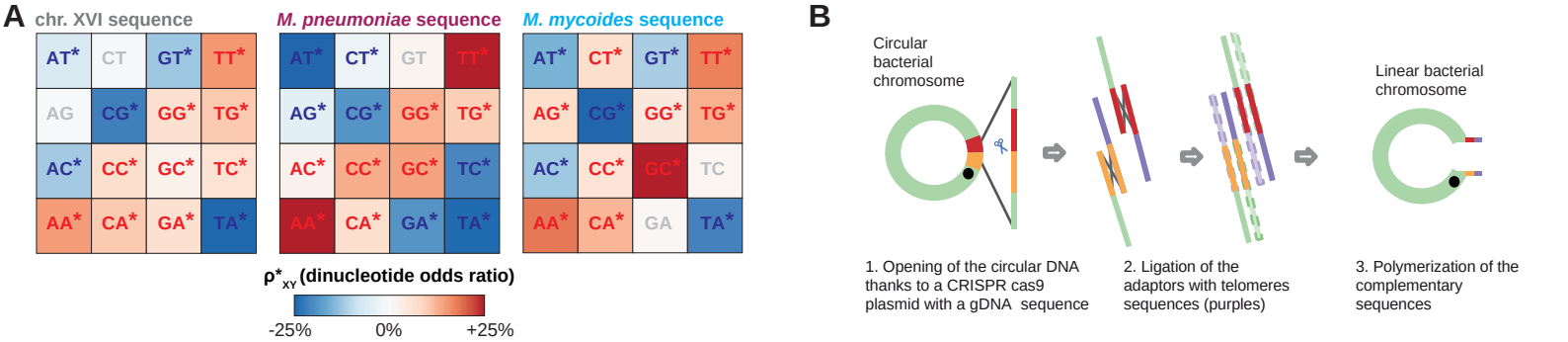
B, Schematic representation of the CRISPR-Cas9 strategy applied to linearize and telomerize the circular bacterial genomes.

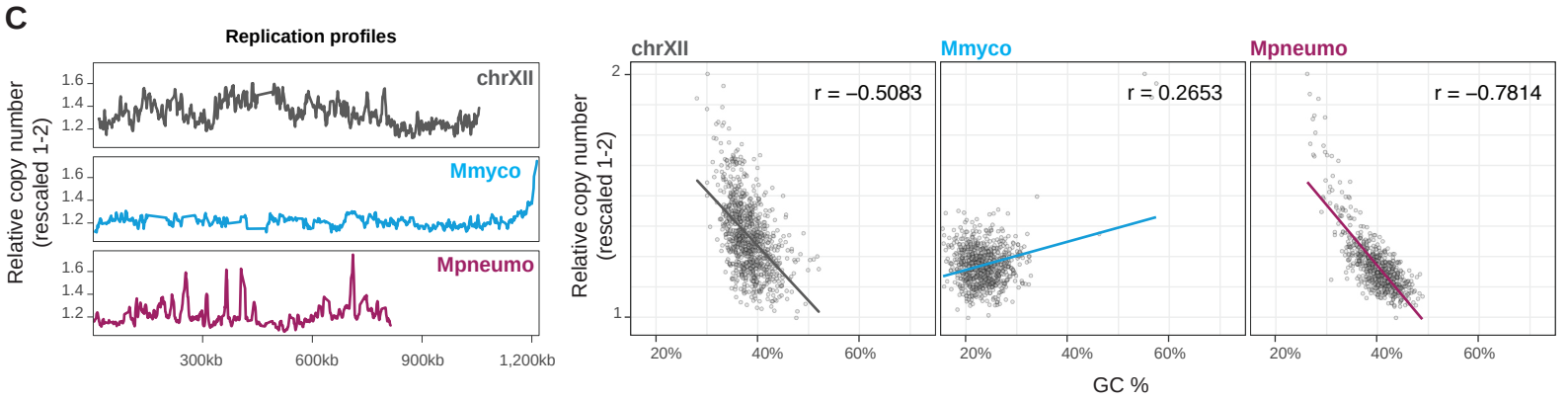
C, MFA (replication profile) analysis of a representative WT yeast chromosome, and of Mmcyo and Mpneumo chromosomes.

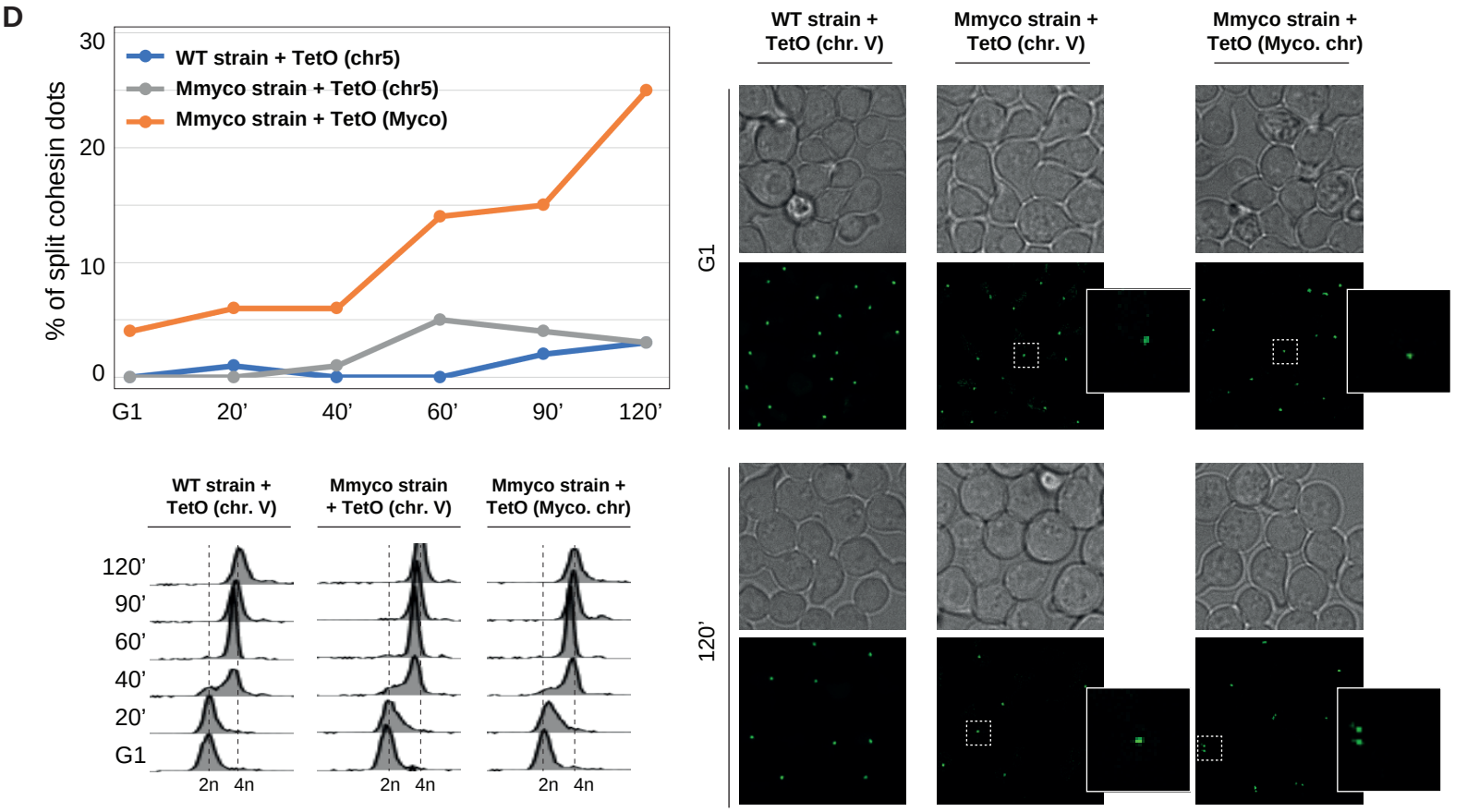
D, Cohesin split dot assays in WT strain (TetO array in chr. V), or in Myco strain (TetO array in chr. V or Myco. chromosome). Top left: % of split cohesin dots in each strain upon G1-release; bottom left: cell ploidy upon G1-release; right: representative cohesin immuofluorescence imaging in each strain following G1-release.

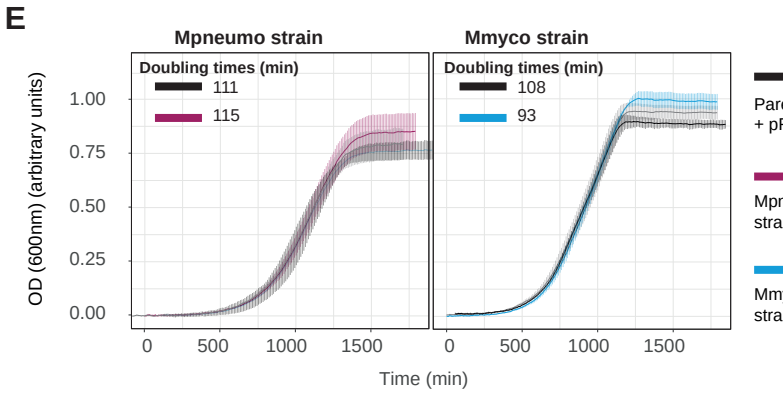
E, Growth curves of WT, Mpneumo and Mmyco strains. For each strain, 3 independent cultures were performed. A pRS413 centromeric plasmid similar to the one onto which bacterial chromosomes were originally cloned is included in the WT strain as a control.

F, The chromosome stability and the segregation rate were measured as described in (*60*) (Methods). Yeast strains used are RSG_Y712 (Mmyco linear) and RSG_Y960 (Mpneumo linear). **P1**: % of bacterial chromosome-carrying cells in selective media. **g**: number of doubling. **P2**: of bacterial chromosome-carrying cells after 12 generations in non-selective media. **m**: segregation rate, i.e. % of plasmid-free segregants appearing in the final population after a single doubling.









F

	Percent (*P*) chromosome-carrying cells			
	After 10 doublings selective	After *g* doublings nonselective		Segregation rate
	P_1	*g*	P_2	*m*
Mmyco.	91.07	11.72	74.8	1.7
Mpneumo.	76.8	11.70	50.7	3.5

Figure S2. Replication of bacterial genomes in yeast.

A, Pearson correlation between replicates of MNase and ChIP experiments. Bin size: 100bp.

B, Coverage of sub-nucleosomal (smaller than 130bp) or nucleosomal (between 130 and 165bp) fragments over 2kb-long genomic loci from chr. II, Mpneumo and Mmyco chromosomes, over an MNase digestion timecourse. All the tracks are displayed at the same scale (0-20 CPM).

C, Number of poly-dA, poly-dT and poly-dW (dA/dT) stretches of various lengths in the yeast genome and over the Mpneumo and Mmyco chromosomes. Stretch numbers are scaled to 147bp.

D, Power spectral density (PSD) of WW 10-bp periodicity, in 300bp-long sequences centered over yeast TSSs or along the Mpneumo or Mmyco chromosomes. Random sequences were generated by shuffling actual sequences while preserving dinucleotide frequency.

E, Cohesin (Scc1) enrichment over yeast centromeres and yeast arms in WT, Mpneumo and Mmyco strains. Scc1 signal over arms was calculated outside of any Scc1 peak.

F, Representative 80kb window of Scc1 ChIP-seq deposition signal over yeast chr. IV in the WT, Mpneumo, and Mmyco strains.

G, Aggregated profile of Scc1 deposition centered at Scc1 peaks (+/- 2kb) called over all endogenous (left panels) or only chr. II (right panels) yeast chromosomes in the WT, Mpneumo and Mmyco strains.

H, Aggregated profile of MNase-seq (blue) or Pol. II (green) deposition centered at Scc1 peaks (+/- 2kb) called over the yeast chromosome II (left) or the Mpneumo chromosome (right)

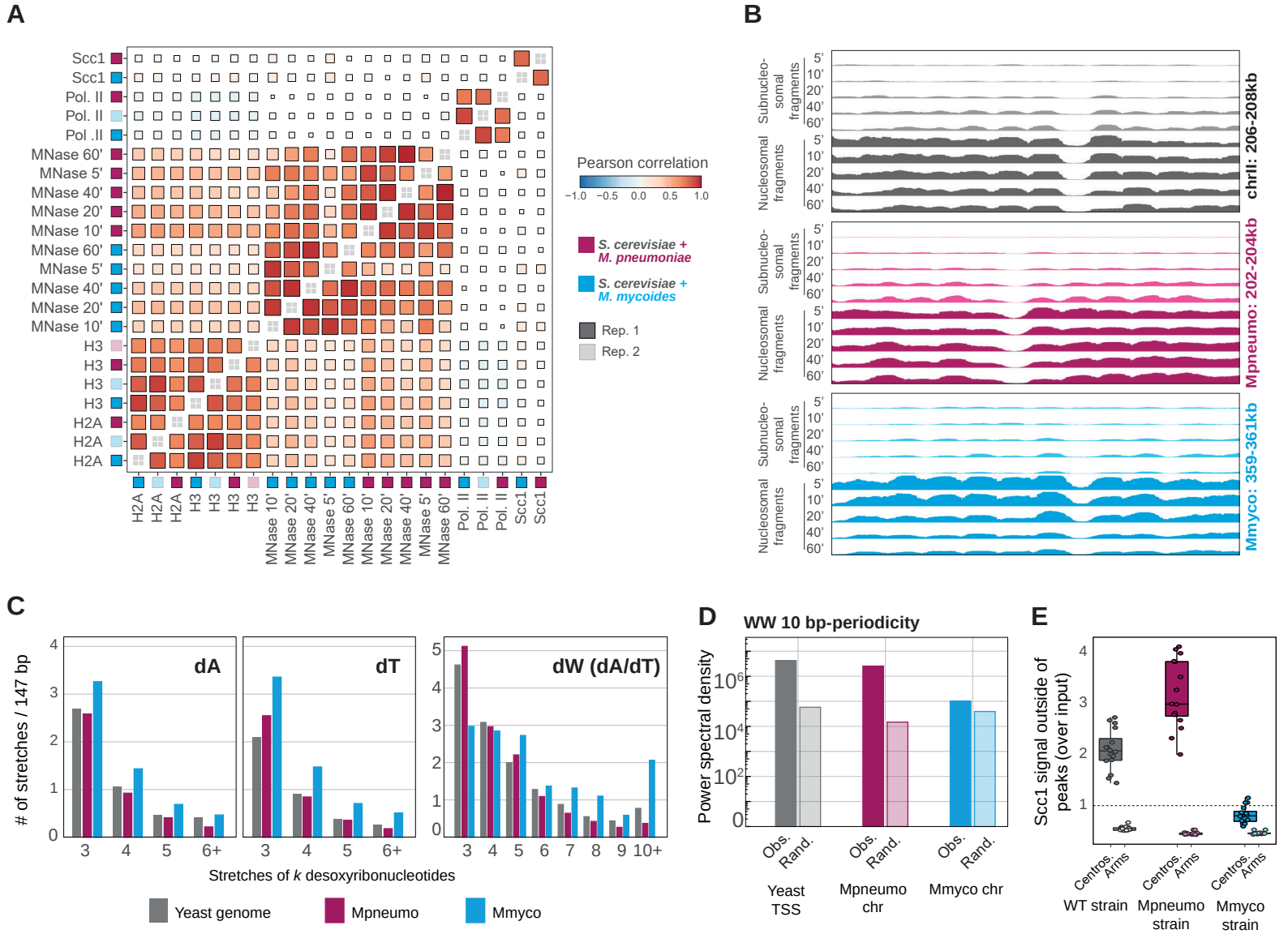

A
Scc1
Scc1
Pol. II
Pol. II
Pol .II
MNase 60'
MNase 5'
MNase 40'
MNase 20'
MNase 10'
MNase 60'
MNase 5'
MNase 40'
MNase 20'
MNase 10'
H3
H3
H3
H3
H2A
H2A
H2A
H2A H2A H2A H3 H3 H3 MNase 10' MNase 20' MNase 40' MNase 5' MNase 60' MNase 10' MNase 20' MNase 40' MNase 5' MNase 60' Pol. II Pol. II Pol. II Scc1 Scc1
Pearson correlation
−1.0 −0.5 0.0 0.5 1.0
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
Rep. 1
Rep. 2
B
Subnucleosomal fragments
Nucleosomal fragments
5' 10' 20' 40' 60'
chrII: 206-208kb
Mpneumo: 202-204kb
Mmyco: 359-361kb
C
dA
dT
dW (dA/dT)
# of stretches / 147 bp
Stretches of k desoxyribonucleotides
Yeast genome
Mpneumo
Mmyco
D
WW 10 bp-periodicity
Power spectral density
Obs.
Rand.
Yeast TSS
Mpneumo chr
Mmyco chr
E
Scc1 signal outside of peaks (over input)
Centros.
Arms
WT strain
Mpneumo strain
Mmyco strain
F
S. cerevisiae
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
Scc1 signal over input
190kb
270kb
chr. IV
G
Scc1 signal at Scc1 peaks (over input)
S. cerevisiae
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
-2kb
2kb
All yeast Scc1 peaks
chrII Scc1 peaks (n=72)
Mpneumo Scc1 peaks (n=52)
Mmyco Scc1 peaks (n=190)
H
S. cerevisiae + M. pneumoniae
MNase-seq
Pol. II ChIP-seq
chrII Scc1 peaks (n=72)
Mpneumo Scc1 peaks (n=52)

Figure S3. Chromatin accessibility of bacterial chromosomes in yeast.

A, ATAC-seq profiles (chromatin/gDNA) obtained in the Mpneumo strain (*S. cerevisiae* + *M. pneumoniae*) and Mmyco strain (*S. cerevisiae* + *M. mycoides*) (10kb-long genomic windows over the chromosome IV, Mpneumo and Mmyco chromosomes).

B-C, ATAC-seq analysis comparing *S. cerevisiae* with *M. pneumoniae* (**B**) or with *M. mycoides* (**C**). Boxplots display ATAC peak widths and barplots summarize the total % of each chromosome covered by ATAC-seq peaks.

D, ATAC-seq fragment length distribution for *S. cerevisiae* combined with *M. pneumoniae* (left) and *M. mycoides* (right). ATAC-seq fragments obtained from nucleosome-free regions (<100 bp), mononucleosomal (~200 bp), and dinucleosomal (~350 bp) fragments are highlighted.

E, Average signal of nucleosome and histone ChIP-seq tracks over ATAC-seq peaks.

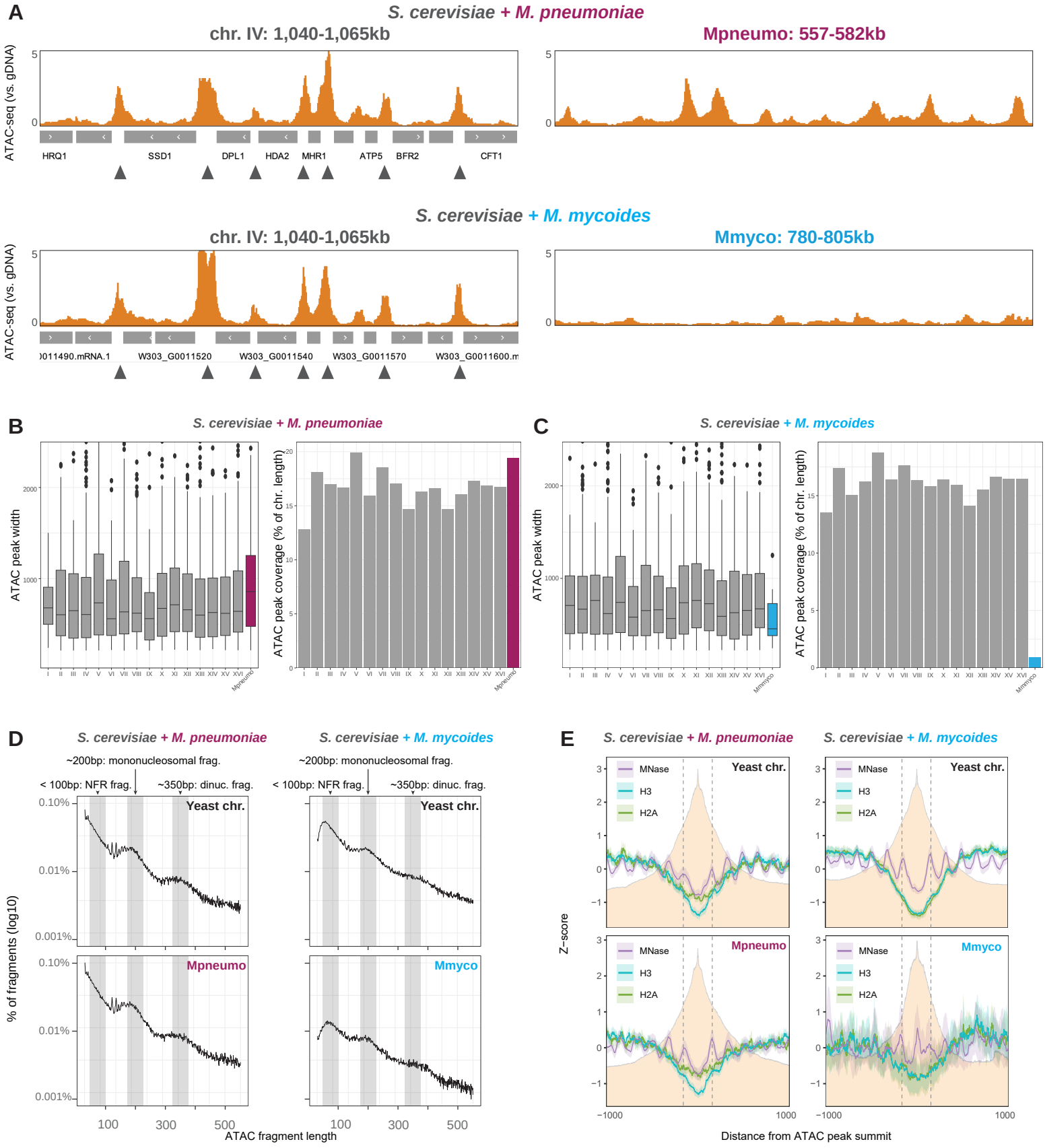

A
S. cerevisiae + M. pneumoniae
chr. IV: 1,040-1,065kb
Mpneumo: 557-582kb
ATAC-seq (vs. gDNA)
HRQ1
SSD1
DPL1
HDA2
MHR1
ATP5
BFR2
CFT1
S. cerevisiae + M. mycoides
chr. IV: 1,040-1,065kb
Mmyco: 780-805kb
W303_G0011520
W303_G0011540
W303_G0011570
W303_G0011600.m
B
S. cerevisiae + M. pneumoniae
ATAC peak width
ATAC peak coverage (% of chr. length)
C
S. cerevisiae + M. mycoides
D
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
~200bp: mononucleosomal frag.
< 100bp: NFR frag.
~350bp: dinuc. frag.
Yeast chr.
Mpneumo
Mmyco
% of fragments (log10)
ATAC fragment length
E
S. cerevisiae + M. pneumoniae
S. cerevisiae + M. mycoides
MNase
H3
H2A
Z−score
Distance from ATAC peak summit

Figure S4. Transcription orientation of bacterial chromosomes in yeast.

A, RNA-seq average signal over yeast or bacterial gene bodies and intergenic regions, in Mpneumo and Mmyco strains. Scores were normalized by the length of each genomic feature (CPM: counts per million of sequenced fragments).

B, Stranded analysis of RNA-seq data in Pneumo. strain. Pile-up of 1kb windows centered on transcription start sites (TSS) of genes either in the forward (Top) or reverse (Bottom) orientation. Left: endogenous yeast genes. Right: TSS of the first gene of annotated operons along the *M. pneumoniae* sequence.

C, Pol. II ChIP-seq coverage analysis in Pneumo strain. Pile-up of 2kb windows centered on transcription start sites (TSS). Grey: endogenous yeast genes. Purple: TSS of the first gene of annotated operons along the *M. pneumoniae* sequence.

D, GC content (left), average AT (middle) and GC (right) skews (computed on Watson strand) over yeast, Myco and Mpneumo forward or reverse genes.

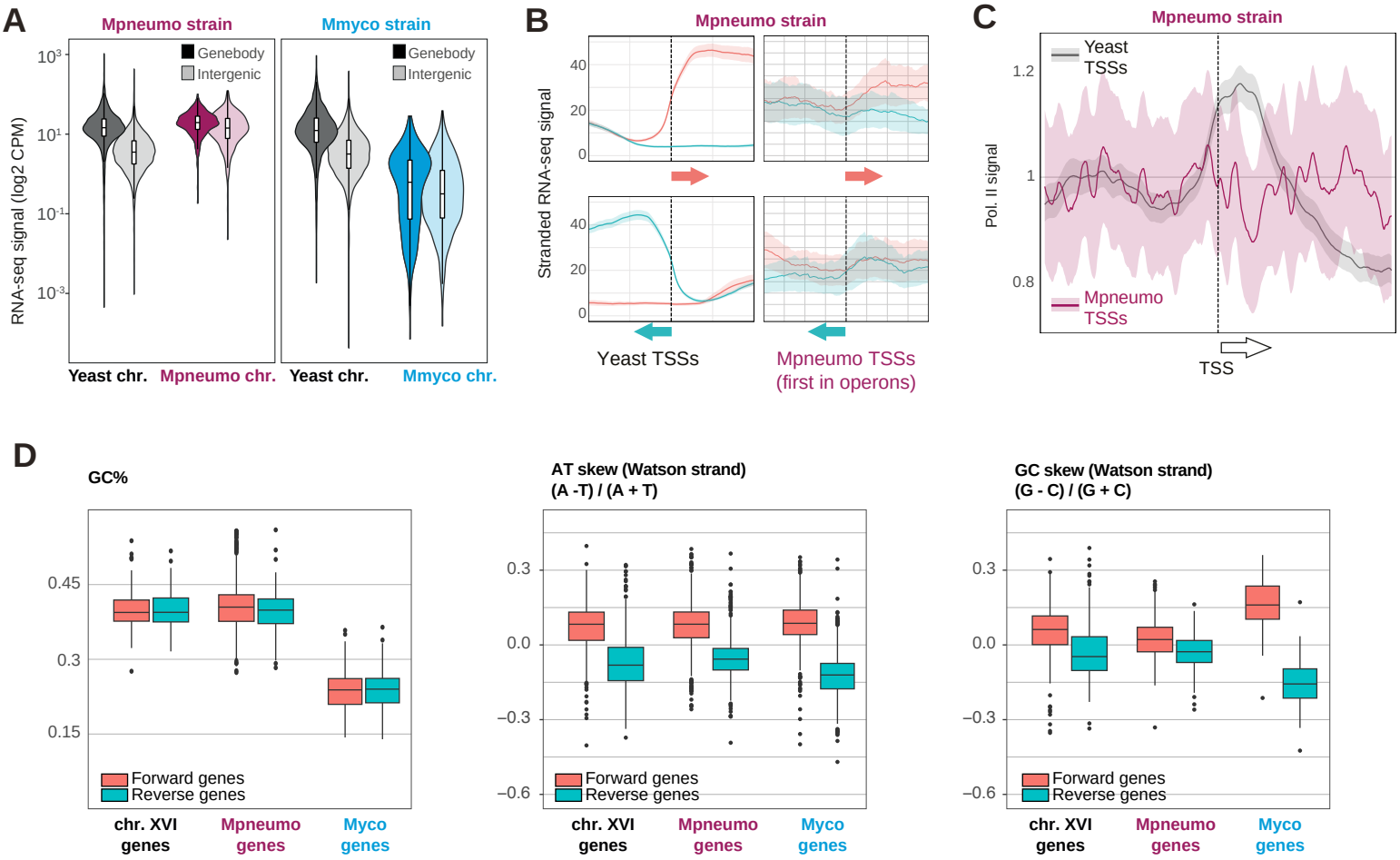

A
Mpneumo strain
Mmyco strain
Genebody
Intergenic
RNA-seq signal (log2 CPM)
Yeast chr.
Mpneumo chr.
Yeast chr.
Mmyco chr.
B
Mpneumo strain
Stranded RNA-seq signal
Yeast TSSs
Mpneumo TSSs
(first in operons)
C
Mpneumo strain
Yeast
TSSs
Pol. II signal
Mpneumo
TSSs
TSS
D
GC%
AT skew (Watson strand)
(A -T) / (A + T)
GC skew (Watson strand)
(G - C) / (G + C)
Forward genes
Reverse genes
chr. XVI
genes
Mpneumo
genes
Myco
genes

Figure S5. Nascent transcription of bacterial chromosomes in yeast.

A. Stranded CRAC-seq and RNA-seq profiles along chr. IV, Mpneumo and Mmyco chromosomes (CPM). Forward transcription profiles are shown in pink, and reverse transcription profiles are shown in turquoise.
B, Relationship between CRAC and RNA coverage, in *S. cerevisiae* strains with *M. pneumoniae* (left) or with *M. mycoides* chromosome (right).
C, Stranded RNA-seq profiles of Mpneumo strains in WT or Δupf1, Δrrp6 and Δupf1/Δrrp6 mutants, along chr. X or Mpneumo chromosomes. Forward transcription profiles are shown in pink, and reverse transcription profiles are shown in turquoise.

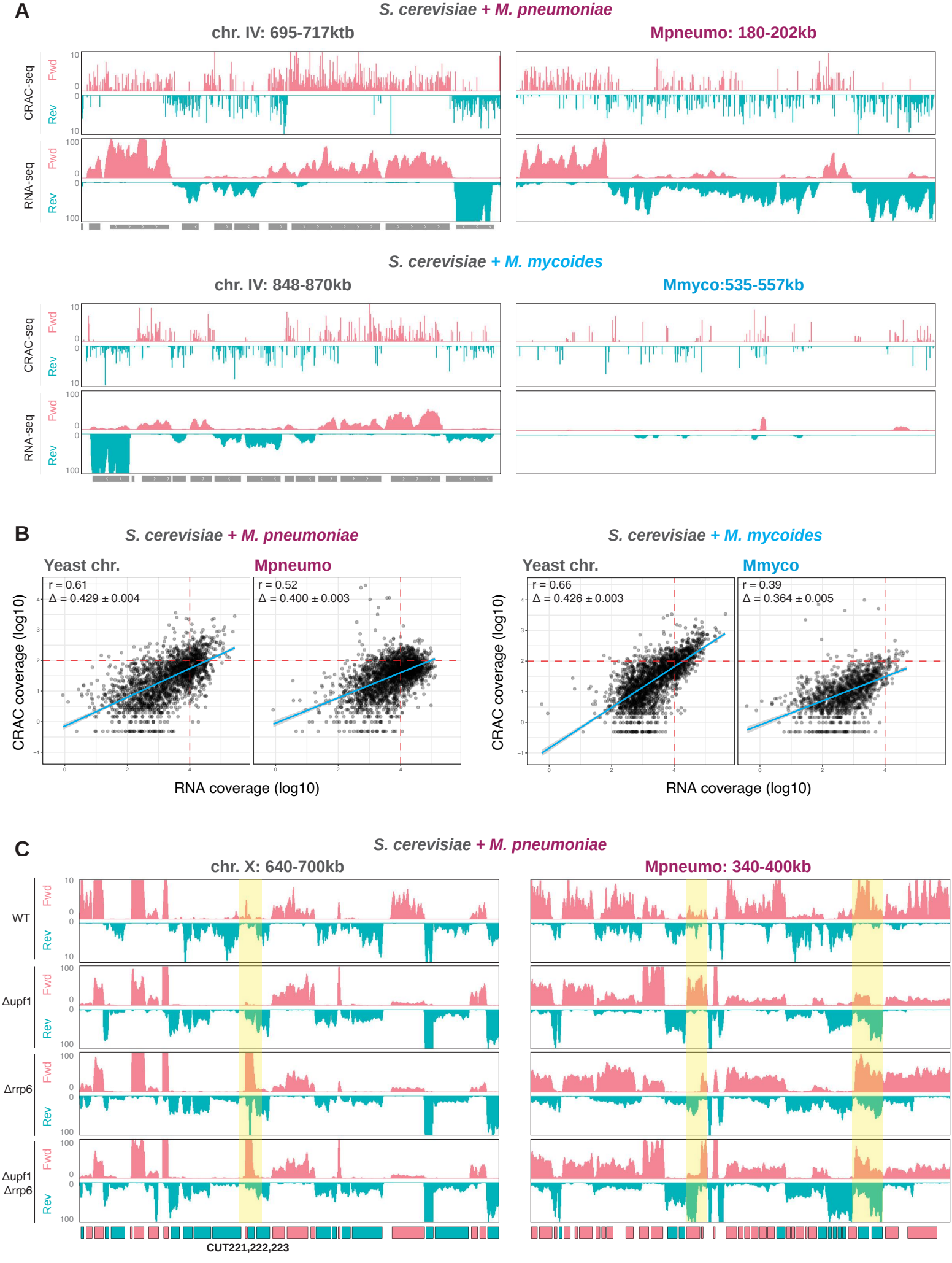

A
S. cerevisiae + M. pneumoniae
chr. IV: 695-717ktb
Mpneumo: 180-202kb
CRAC-seq
RNA-seq
Fwd
Rev
S. cerevisiae + M. mycoides
chr. IV: 848-870kb
Mmyco:535-557kb
B
S. cerevisiae + M. pneumoniae
Yeast chr.
Mpneumo
r = 0.61
Δ = 0.429 ± 0.004
r = 0.52
Δ = 0.400 ± 0.003
CRAC coverage (log10)
RNA coverage (log10)
S. cerevisiae + M. mycoides
Yeast chr.
Mmyco
r = 0.66
Δ = 0.426 ± 0.003
r = 0.39
Δ = 0.364 ± 0.005
C
S. cerevisiae + M. pneumoniae
chr. X: 640-700kb
Mpneumo: 340-400kb
WT
Δupf1
Δrrp6
Δupf1 Δrrp6
CUT221,222,223

Figure S6. Folding of exogenic bacterial sequences within the yeast nucleus.

A, % of total trans-contacts made by endogenous yeast chromosomes, Mmyco (top) or Mpneumo (bottom) bacterial chromosomes, in G1 or G2/M.

B, Left: quantification of contacts between the entire Mmyco chromosome (blue) and endogenous yeast chr. X (grey). Right: similar analysis for the other 15 endogenous yeast chromosomes. Note the increase of Myco. contacts at yeast telomeres.

C, Series of nuclei from Mpneumo or Mmyco fixed cells labeled with DAPI and hybridized with a fluorescent probe generated from either purified Mmyco (left) or Mpneumo (right) chromosome (top row: probe signal; bottom row: DAPI signal).

D, Distance between the patch of fluorescent signal from either the Mmyco or Mpneumo chromosomes and the nucleus border. Note that the Mmyco patches are located closer to the nucleus border than Mpneumo ones.

E, F, Slope of distance-dependent contact frequency of endogenous yeast chromosomes and bacterial chromosomes in G2/M Mpneumo (**C**) and Mmyco (**D**) strains.

G, Distance-dependent loop scores (computed using Chromosight [13]) for loops along either endogenous (grey) or Mpneumo (purple) chromosomes.

H, Distance measured between cohesin peaks and their nearest convergent transcription locus, for peaks located in chr. II or in Mpneumo chromosome. Expected distances, measured after randomly shuffling the position of the cohesin peaks 100 times, are also shown.

I, Left: aggregated profile of Scc1 deposition centered at Scc1 peaks (+/- 4kb) called over chr. II (top) or Mpneumo (bottom), with peaks ordered by peak strength. Middle: corresponding stranded transcription tracks, colored according to their forward or reverse orientation. Right: for chr. II or Mpneumo chromosome, average forward and reverse transcription over the 20% strongest or 20% weakest Scc1 peaks.

J, Correlation between Scc1 (cohesin) binding and convergent transcription strength (see Methods) in chr. II and in Mpneumo chromosome.

K, For yeast chromosome IV, Mpneumo and Mmyco: Left, Hi-C contact maps of the endogenous yeast chromosome IV of the Mpneumo strain synchronized in G2/M in either WT and Wpl1 depleted cells (Δwpl1); Right, corresponding chr. IV ratio map (Δwpl1 over WT). Red (or blue) indicate enriched (or depleted) contacts in Δwpl1.

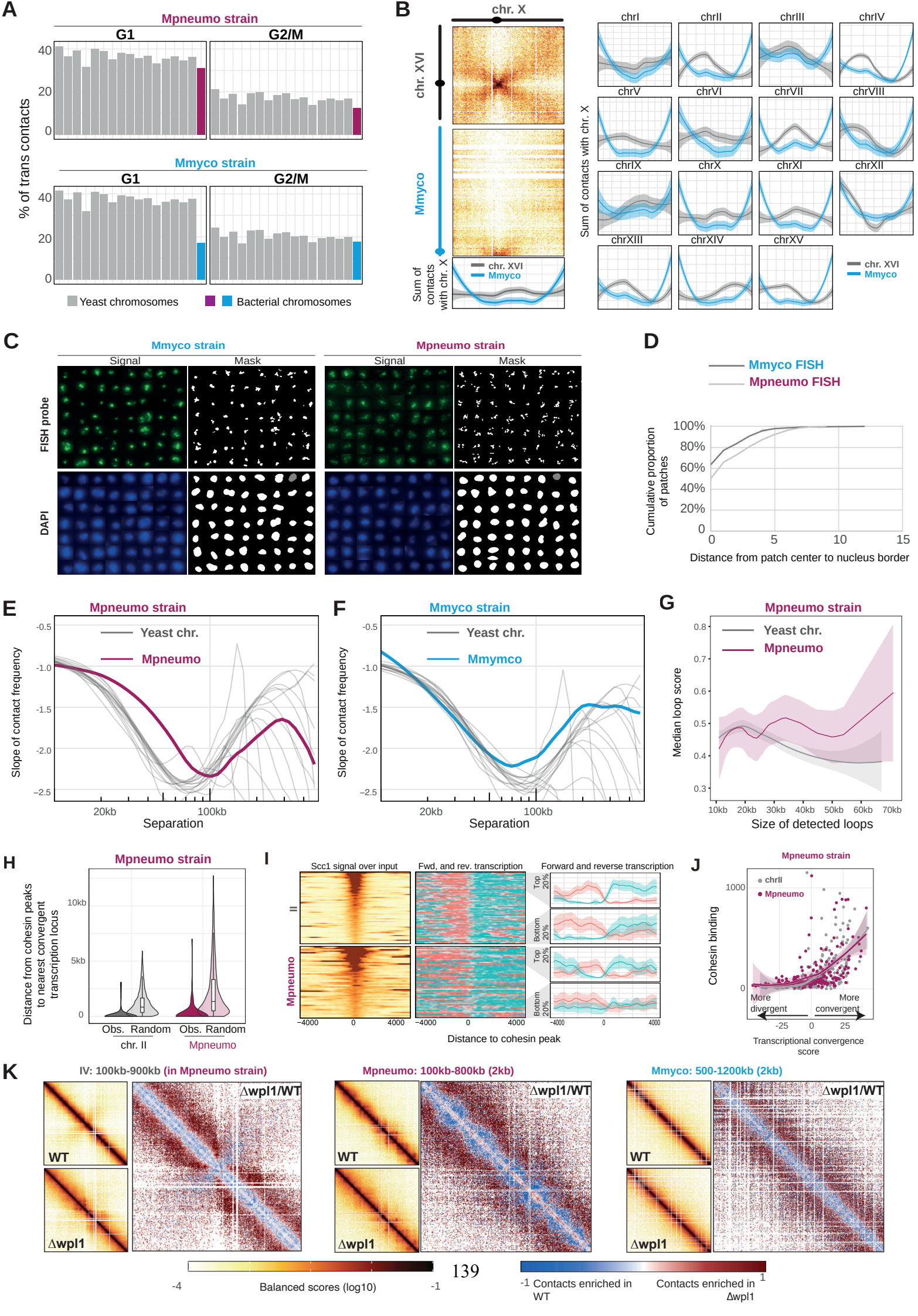

A
Mpneumo strain
G1
G2/M
% of trans contacts
Mmyco strain
Yeast chromosomes
Bacterial chromosomes
B
chr. X
chr. XVI
Mmyco
Sum of contacts with chr. X
chrI
chrII
chrIII
chrIV
chrV
chrVI
chrVII
chrVIII
chrIX
chrX
chrXI
chrXII
chrXIII
chrXIV
chrXV
C
Mmyco strain
Mpneumo strain
Signal
Mask
FISH probe
DAPI
D
Mmyco FISH
Mpneumo FISH
Cumulative proportion of patches
Distance from patch center to nucleus border
E
Mpneumo strain
Yeast chr.
Mpneumo
Slope of contact frequency
Separation
F
Mmyco strain
Mmymco
G
Median loop score
Size of detected loops
H
Distance from cohesin peaks to nearest convergent transcription locus
Obs.
Random
chr. II
Mpneumo
I
Scc1 signal over input
Fwd, and rev. transcription
Forward and reverse transcription
Top 20%
Bottom 20%
Distance to cohesin peak
J
chrII
Mpneumo
Cohesin binding
More divergent
More convergent
Transcriptional convergence score
K
IV: 100kb-900kb (in Mpneumo strain)
Mpneumo: 100kb-800kb (2kb)
Mmyco: 500-1200kb (2kb)
WT
Δwpl1
Δwpl1/WT
Balanced scores (log10)
Contacts enriched in WT
Contacts enriched in Δwpl1

Figure S7. Compartmentalization of mosaic chromosomes composed of Y and U-type chromatin.

A, Left: pulsed-field gel electrophoresis of chromosomes from yeast strains containing the Mmyco chromosome, either linear or fused with endogenous yeast chr. XVI (XVIfMmyco strain). Right: Southern blot hybridization using a *his-3* probe present on the Mmyco chromosome sequence (note that *his-3* is also present on the endogenous chr. XV in the parental yeast strain).

B, Growth curves of WT, Mmyco and XVIfMmyco strains.

C, Pol. II (top) and Scc1 (bottom) ChIP-seq deposition profiles along three representative yeast chromosomes and Mmyco chromosome (left) and mosaic chromosomes in XVIfMmycot1 (center) and XVIfMmycot2 (right) strains. Bin size: 10kb.

D, Expression fold-change (log2) against average expression, for genes located in yeast segments or Mmmcyo segments, between the XVIfMmycot2 and the XVIfMmycot2' strains. Genomic tracks illustrating stranded RNA-seq coverage over yeast (gray) and Mmmcyo (blue) genes, and their surrounding sequences (yeast: gray; Mmmcyo: blue).

E, Average distance of interaction along the fused and mosaic Mmyco chromosomes: XVIfMmyco (left), XVIfMmycot1 (center) and XVIfMmycot2 (right). The shaded ribbon represents the interval between the 25% and 75% quantile of distance of interactions.

F, Top: G2/M Hi-C contact maps of chr. XV, XVI, and bacterial chromosomes in the Mmyco strain, and in its derivatives (i.e. the chr. XVI and Mmyco fusion, and the two strains with translocations resulting in alternating Y and U chromatin segments; Methods). Bottom: correlation contact matrices in wt and mosaic chromosomes strains. The color scales are the same as in **Fig. 4B**.

G, Hi-C contact maps in G2/M of 100 kb window of either the XVIfMmyco (left and center) or XVIfMmycot1 (right) chromosomes, centered on the translocation position of the XVIfMmycot1 chromosome. In XVIfMmycot1, this window is effectively centered at the junction between the yeast chr. XVI segment (upstream) and the Myco chromosome segment (downstream). The Scc1 deposition profile measured by ChIP-seq in each strain is shown underneath each contact map. Green arrows: cohesin peaks flanking the junction between the yeast and the Myco segments in the XVIfMmycot1 chromosome.

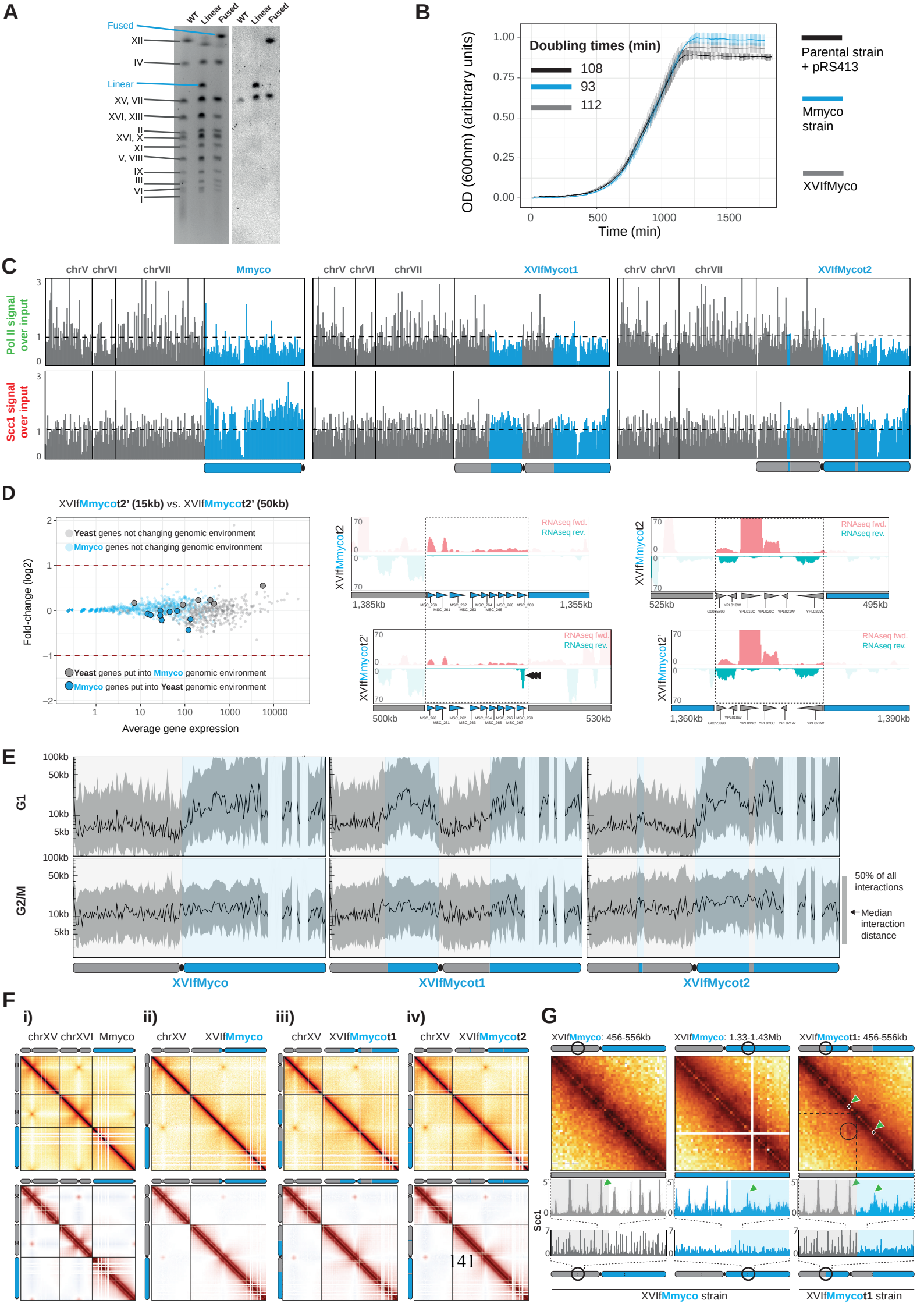
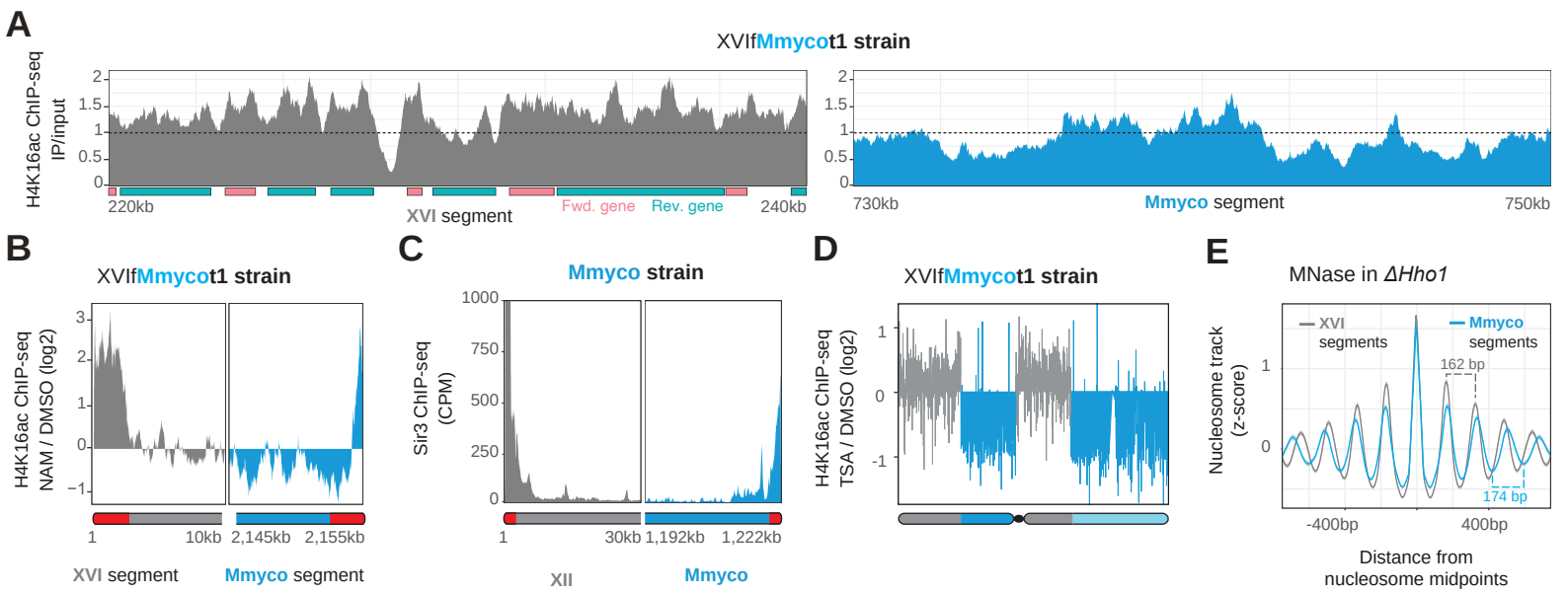
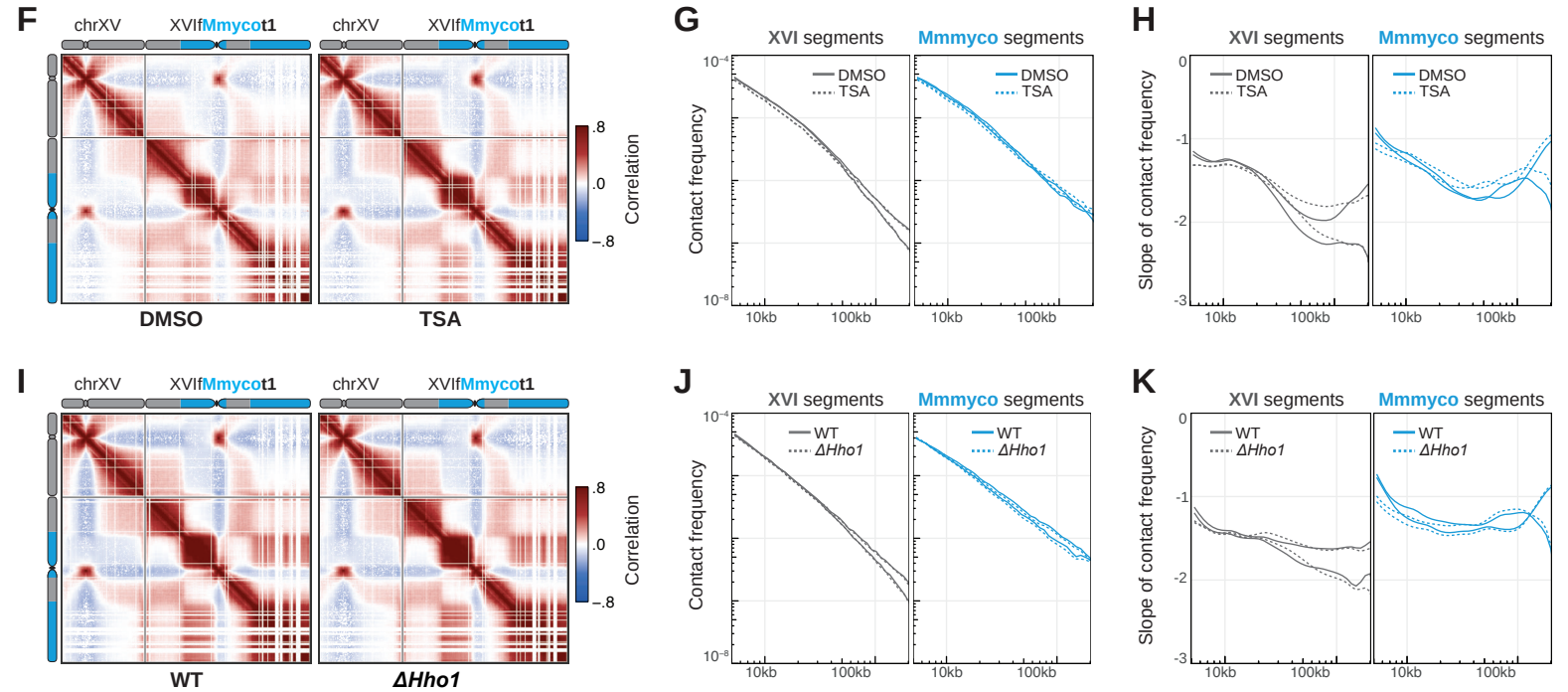
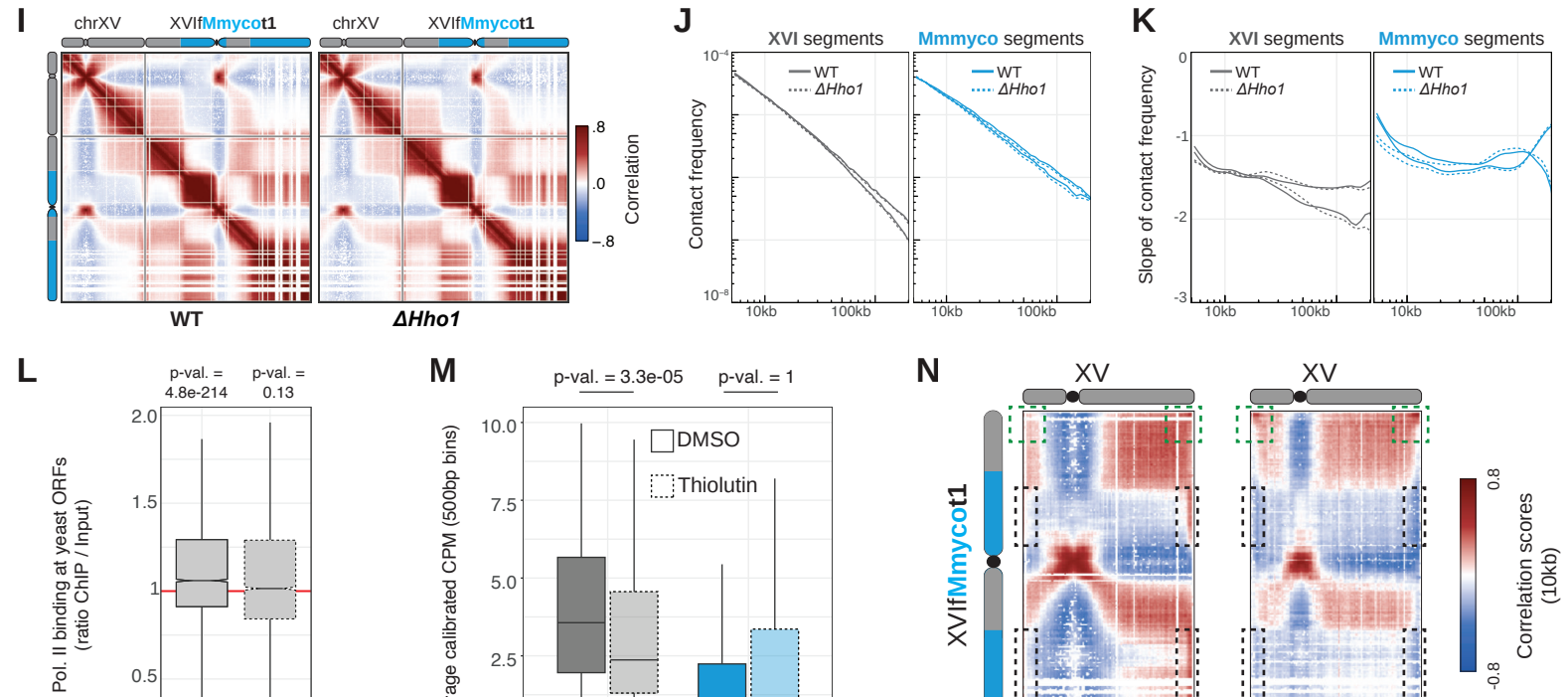

A
WT
Linear
Fused
WT
Linear
Fused
Fused
XII
IV
Linear
XV, VII
XVI, XIII
II
XVI, X
XI
V, VIII
IX
III
VI
I
B
Doubling times (min)
108
93
112
OD (600nm) (aribtrary units)
1.00
0.75
0.50
0.25
0.00
0
500
1000
1500
Time (min)
Parental strain + pRS413
Mmyco strain
XVIfMyco
C
chrV chrVI chrVII Mmyco
chrV chrVI chrVII XVIfMycot1
chrV chrVI chrVII XVIfMycot2
Pol II signal over input
Scc1 signal over input
3
1
0
D
XVIfMmycot2' (15kb) vs. XVIfMmycot2' (50kb)
Yeast genes not changing genomic environment
Mmyco genes not changing genomic environment
Yeast genes put into Mmyco genomic environment
Mmyco genes put into Yeast genomic environment
Fold-change (log2)
Average gene expression
XVIfMmycot2
XVIfMmycot2'
RNAseq fwd.
RNAseq rev.
1,385kb
1,355kb
525kb
495kb
500kb
530kb
1,360kb
1,390kb
E
G1
G2/M
100kb
50kb
10kb
5kb
50% of all interactions
Median interaction distance
XVIfMyco
XVIfMycot1
XVIfMycot2
F
i)
ii)
iii)
iv)
chrXV chrXVI Mmyco
chrXV XVIfMmyco
chrXV XVIfMmycot1
chrXV XVIfMmycot2
G
XVIfMmyco: 456-556kb
XVIfMmyco: 1.33-1.43Mb
XVIfMmycot1: 456-556kb
Scc1
XVIfMmyco strain
XVIfMmycot1 strain
141

Figure S8. Functional investigation of heterochromatinization mechanisms of U-type compartment

A, H4K16ac ChIP-seq profile (ratio IP/input) in the XVIfMmycot1 strain, over a yeast chromosome segment (left) or a Myco chromosome segment (right).
B, Comparison of H4K16ac ChIP-seq profiles in the XVIfMmycot1 strain, in cells treated by Nicotinamide (NAM) vs. DMSO (log2 scale). Telomeres and subtelomeric domains (2.5kb) are shown in red.
C, ChIP-seq profiles of Sir3 in the XVIfMmycot1 strain. Telomeres and subtelomeric domains (2.5kb) are shown in red.
D, Comparison of H4K16ac ChIP-seq profiles in the XVIfMmycot1 strain, in cells treated by Trichostatin A (TSA) vs. DMSO (log2 scale).
E, Nucleosome track over yeast (gray) or Mmyco nucleosomes (blue), in *ΔHho1* yeast strain.
F, Correlation matrices of the contacts in chr. XV and XVIfMmycot1 in G1, after the addition of DMSO (left) or TSA (right).
G, Contact frequency (p) as a function of genomic distance (s), for contacts in yeast segments (gray) or Mmmyco segments (blue) of the chimeric chromosome XVIfMmycot1, in G1 after the addition of DMSO (solid) or TSA (dotted).
H, Derivatives of curves from **F**.
I, Correlation matrices of the contacts in chr. XV and XVIfMmycot1 in G1, after the addition of DMSO (left) or TSA (right).
J, Contact frequency (p) as a function of genomic distance (s), for contacts in yeast segments (gray) or Mmmyco segments (blue) of the chimeric chromosome XVIfMmycot1, in WT (solid) or *ΔHho1* mutant (dotted).
K, Derivatives of curves from **I**.
L, Average Pol. II ChIP-seq coverage (ratio IP/input) at every yeast ORF, in DMSO (solid) or after thiolutin treatment (dotted). P-values from one-sample two-tailed Student t-test.
M, Average RNA-seq coverage over 500bp bins from XVI (gray) or Mmmcyo (blue) segments, in DMSO (solid) or after thiolutin treatment (dotted). P-values from one-tailed Student's t-test.
N, Correlation matrices of the trans-chromosomal contacts between chr. XV and XVIfMmycot1 in G1, in G1 (left) or quiescence (right).


A
XVIfMmycot1 strain
H4K16ac ChIP-seq IP/input
220kb
XVI segment
Fwd. gene
Rev. gene
240kb
730kb
Mmyco segment
750kb
B
XVIfMmycot1 strain
H4K16ac ChIP-seq NAM / DMSO (log2)
XVI segment
Mmyco segment
C
Mmyco strain
Sir3 ChIP-seq (CPM)
XII
Mmyco
D
XVIfMmycot1 strain
H4K16ac ChIP-seq TSA / DMSO (log2)
E
MNase in ΔHho1
Nucleosome track (z-score)
XVI segments
Mmyco segments
162 bp
174 bp
-400bp
400bp
Distance from nucleosome midpoints


F
chrXV
XVIfMmycot1
Correlation
DMSO
TSA
G
XVI segments
Mmmyco segments
DMSO
TSA
Contact frequency
H
Slope of contact frequency


I
chrXV
XVIfMmycot1
Correlation
WT
ΔHho1
J
XVI segments
Mmmyco segments
WT
ΔHho1
Contact frequency
K
Slope of contact frequency

L
p-val. = 4.8e-214
p-val. = 0.13
Pol. II binding at yeast ORFs (ratio ChIP / Input)
DMSO
Thiolutin
M
p-val. = 3.3e-05
p-val. = 1
DMSO
Thiolutin
Average calibrated CPM (500bp bins)
XVI segments
Mmmyco segments
N
XV
XVIfMmycot1
Correlation scores (10kb)
G1
Quiescence

Figure S9. Sequence-based prediction of chromatin composition of exogenous bacterial chromosomes in yeast

A, Left: Schematic representation of the convolution neural network used in this study. The input of the network is a 2 to 30kb DNA sequence. The output is the corresponding value of the nucleosome, Scc1 or Pol. II signal. Details about the size of the input/output and the architecture used are discussed in the Material and Methods. **Right:** Overall training/prediction strategy. We trained a CNN on sequences of chr. I-XV to predict (i) genome-wide nucleosome, Scc1 or Pol. II ChIP-seq coverage tracks over chrXVI and bacterial chromosomes, and (ii) the average nucleosome, Scc1 or Pol. II ChIP-seq signal over 10kb random sequences with varying GC%.

B, Dinucleotide enrichment in genomic loci with 10% greatest nucleosome (left), Scc1 (middle) or Pol. II (right) ChIP-seq coverage (100 bp bins), extracted from experimental or predicted tracks over chromosome XVI, Mpneumo or Mmyco. The dinucleotide composition in these loci is compared to the chromosome-wide dinucleotide composition ($\rho^*(XY)_{[chromosome]}$ / $\rho^*(XY)_{[chromosome]}$).

C, Correlation scores between experimental and predicted nucleosome, Scc1 or Pol. II ChIP-seq data, using a linear model accounting for GC% only (green) or dinucleotide composition (kaki), or using CNNs (blue).

D, Motifs identified de novo as relevant for predicting nucleosome, Scc1 or Pol. II profiles. Matches to similar motifs in yeast databases are shown on the left.

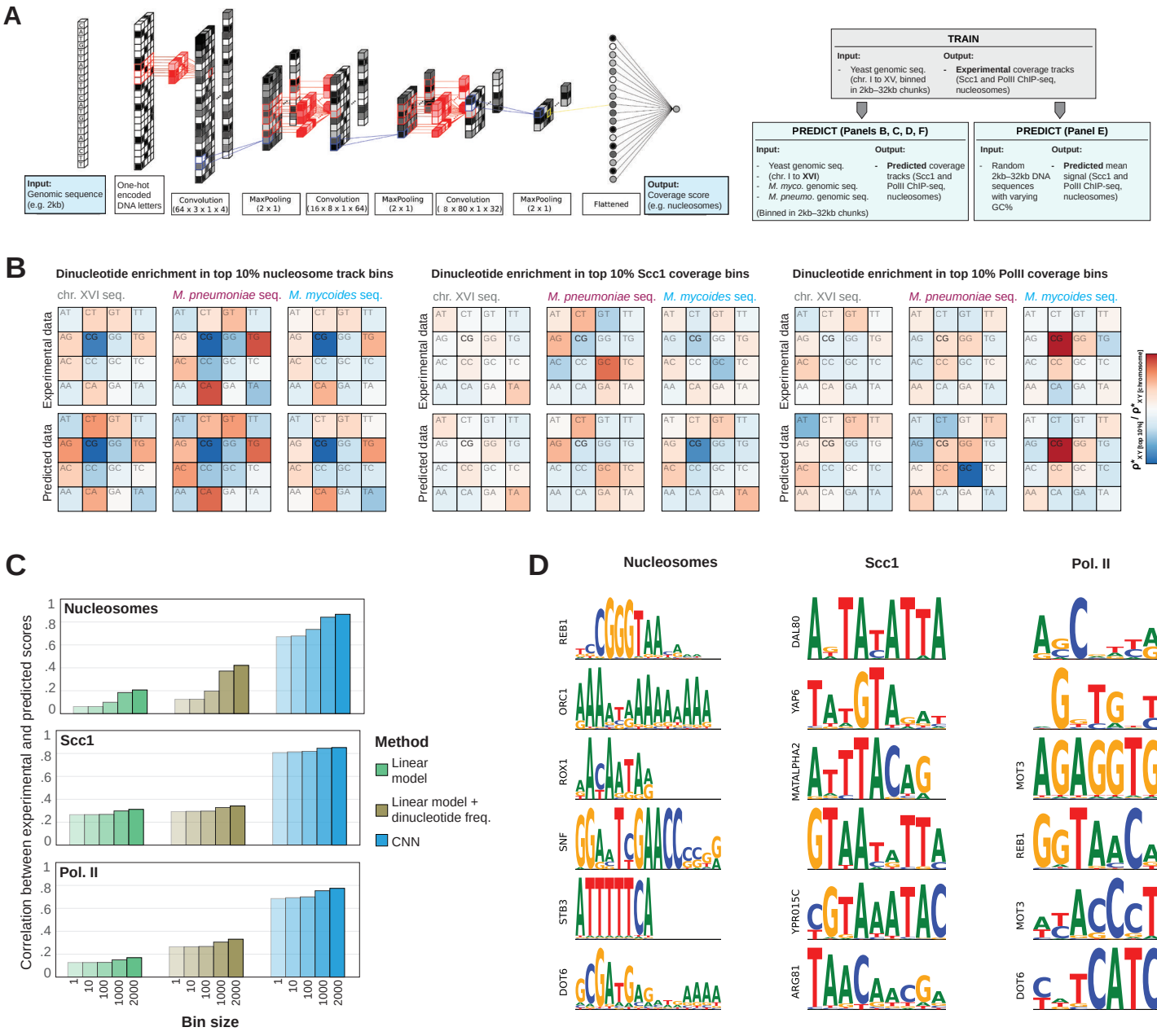

A
Input: Genomic sequence (e.g. 2kb)
One-hot encoded DNA letters
Convolution (64 x 3 x 1 x 4)
MaxPooling (2 x 1)
Convolution (16 x 8 x 1 x 64)
MaxPooling (2 x 1)
Convolution (8 x 80 x 1 x 32)
MaxPooling (2 x 1)
Flattened
Output: Coverage score (e.g. nucleosomes)
TRAIN
Input:
- Yeast genomic seq. (chr. I to XV, binned in 2kb–32kb chunks)
Output:
- Experimental coverage tracks (Scc1 and PolII ChIP-seq, nucleosomes)
PREDICT (Panels B, C, D, F)
Input:
- Yeast genomic seq. (chr. I to XVI)
- M. myco. genomic seq.
- M. pneumo. genomic seq.
(Binned in 2kb–32kb chunks)
Output:
- Predicted coverage tracks (Scc1 and PolII ChIP-seq, nucleosomes)
PREDICT (Panel E)
Input:
- Random 2kb–32kb DNA sequences with varying GC%
Output:
- Predicted mean signal (Scc1 and PolII ChIP-seq, nucleosomes)
B
Dinucleotide enrichment in top 10% nucleosome track bins
Dinucleotide enrichment in top 10% Scc1 coverage bins
Dinucleotide enrichment in top 10% PolII coverage bins
chr. XVI seq.
M. pneumoniae seq.
M. mycoides seq.
Experimental data
Predicted data
+15%
0%
-15%
C
Correlation between experimental and predicted scores
Nucleosomes
Scc1
Pol. II
Bin size
Method
Linear model
Linear model + dinucleotide freq.
CNN
D
Nucleosomes
Scc1
Pol. II
REB1
ORC1
ROX1
SNF
STB3
DOT6
DAL80
YAP6
MATALPHA2
YPR015C
ARG81
MOT3
REB1
MOT3
DOT6

Figure S10. U-type chromatin compartmentalization of AT-rich prokaryotic and eukaryotic YACs

A, Hi-C maps showing chromosomal interactions between *S. cerevisiae* chromosomes (XIV, XV, XVI) and artificial yeast chromosomes (YACs) from *P. falciparum* (top) and *P. vitis* (bottom). The inset for *P. falciparum* highlights the increased trans-chromosomal contacts between the two *P. falciparum* YACs in comparison with the decreased trans-chromosomal contacts between each of the YACs and the yeast chromosome XVI.

B, Percentage of interactions that each YAC has with yeast chromosomes.

C, Percentage of trans interactions between YACs and specific chromosomal regions (chromosome arms, centromere and telomeres), for *P. falciparum* (top) and *P. vitis* (bottom).

D, Average nucleosome signal centered at nucleosomes annotated over *P. falciparum* YAC (top, green) and *P. vitis* YAC (bottom, orange), compared to yeast nucleosomes (gray).

E, Nucleosome track over 3kb-wide segments of the *P. falciparum* YAC (left, green) or *P. vitis* YAC (right, orange).

F, Comparison of experimental (exp.) and predicted (pred.) nucleosome tracks for *P. falciparum* YAC (green), *P. vitis* YAC (orange), and a short random YAC (pink)(*45*).

G, Comparison of experimental (green or orange) and predicted nucleosome signals (dark blue), centered at nucleosomes annotated over *P. falciparum* YAC (top) and *P. vitis* YAC (bottom).

H, Predicted nucleosome, Scc1 and Pol. II coverages in chromosome sequences from different genomes. Coverage values are averaged using 1kb bins tiling each chromosome sequence.

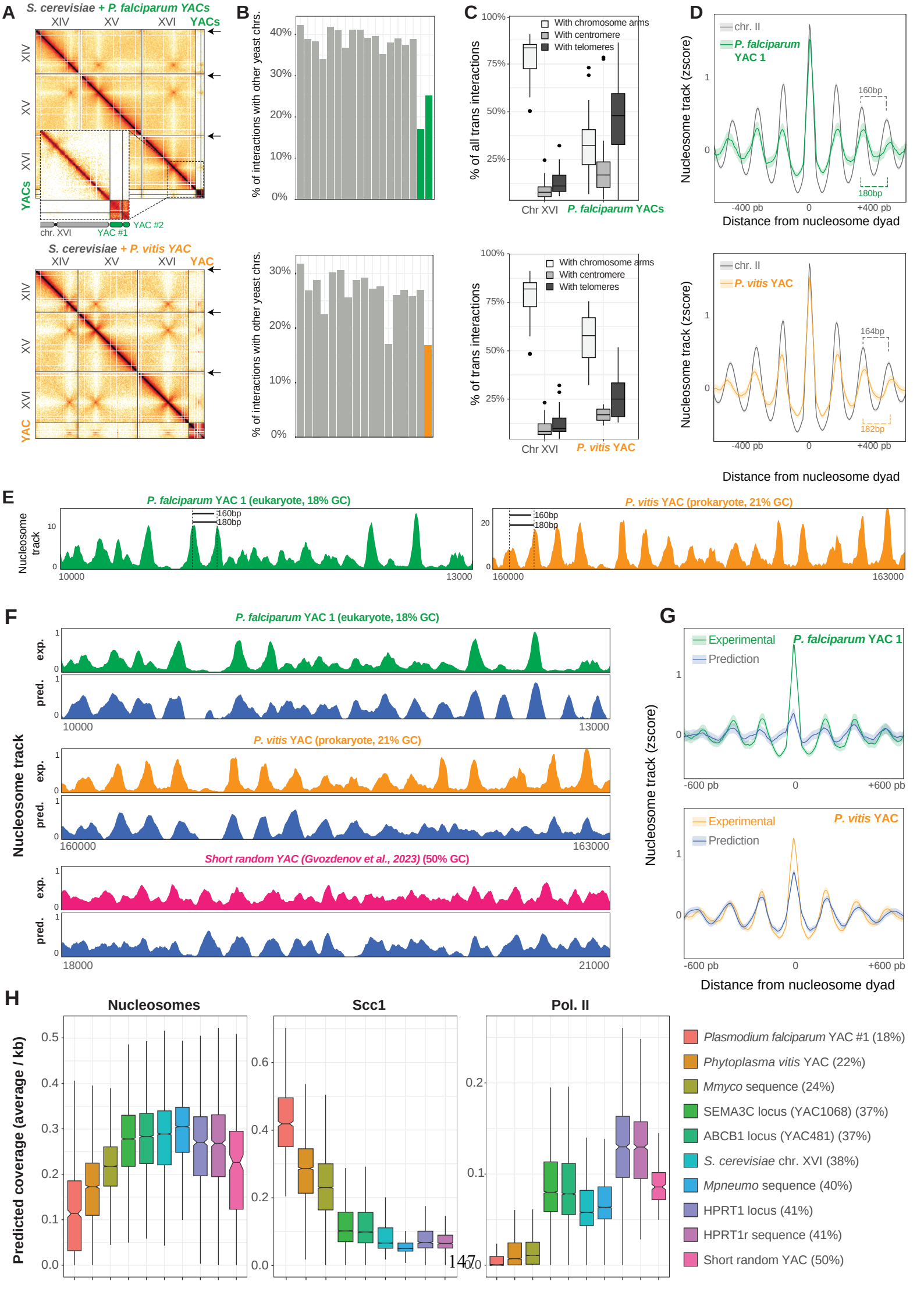
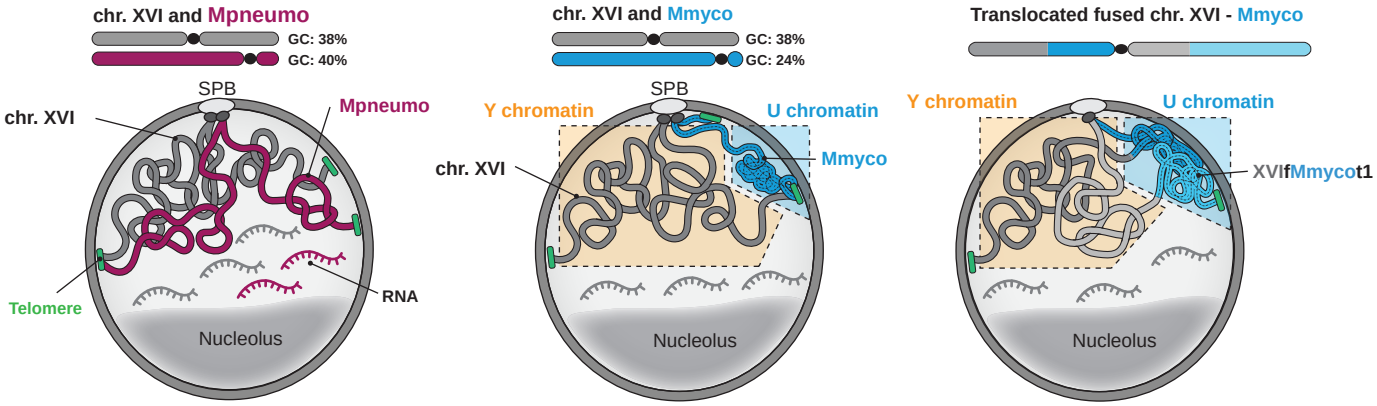
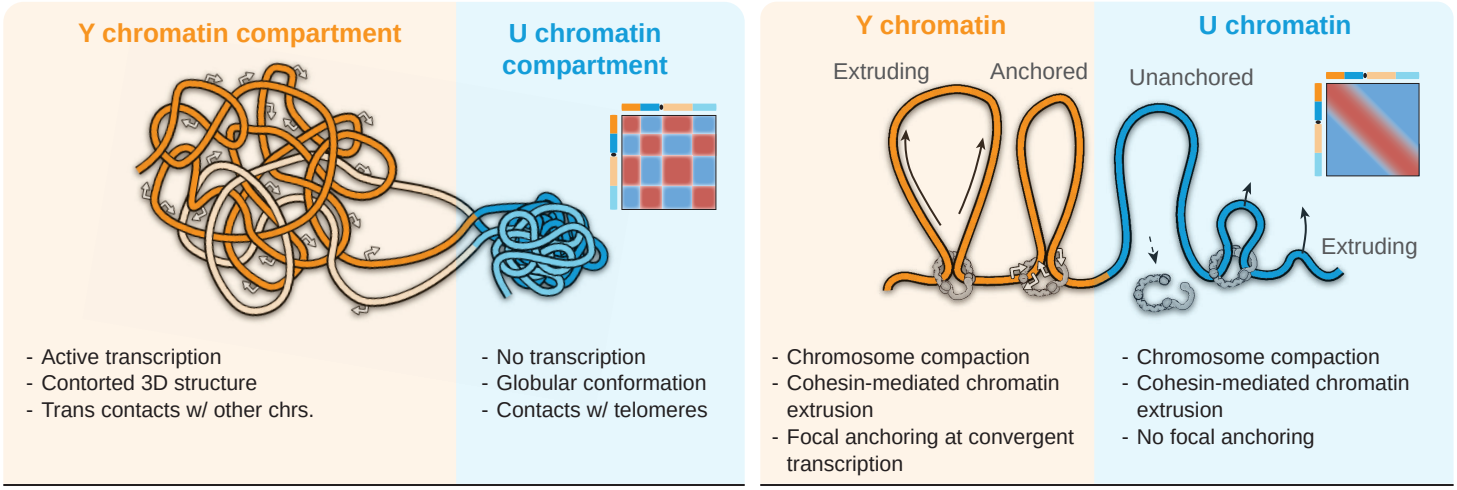

A
S. cerevisiae + P. falciparum YACs
XIV
XV
XVI
YACs
chr. XVI
YAC #1
YAC #2
S. cerevisiae + P. vitis YAC
YAC
B
% of interactions with other yeast chrs.
C
% of all trans interactions
With chromosome arms
With centromere
With telomeres
Chr XVI
P. falciparum YACs
P. vitis YAC
D
Nucleosome track (zscore)
chr. II
P. falciparum YAC 1
160bp
180bp
P. vitis YAC
164bp
182bp
-400 pb
+400 pb
Distance from nucleosome dyad
E
P. falciparum YAC 1 (eukaryote, 18% GC)
P. vitis YAC (prokaryote, 21% GC)
Nucleosome track
10000
13000
160000
163000
F
exp.
pred.
Short random YAC (Gvozdenov et al., 2023) (50% GC)
18000
21000
G
Experimental
Prediction
-600 pb
+600 pb
H
Nucleosomes
Scc1
Pol. II
Predicted coverage (average / kb)
Plasmodium falciparum YAC #1 (18%)
Phytoplasma vitis YAC (22%)
Mmyco sequence (24%)
SEMA3C locus (YAC1068) (37%)
ABCB1 locus (YAC481) (37%)
S. cerevisiae chr. XVI (38%)
Mpneumo sequence (40%)
HPRT1 locus (41%)
HPRT1r sequence (41%)
Short random YAC (50%)

Figure S11.
Illustration of the behaviors and activity of Mpneumo (left panel) and Mmyco (middle and right) chromosome sequences integrated in the yeast genome. Note that yeast and Mpneumo chromosomes intermingle in a single nuclear compartment, whereas the Mmyco chromosome (independently or in a chimeric state within chr. XVI) is condensed and segregated at the nuclear periphery, thereby defining the inactive U-type chromatin compartment.


chr. XVI and Mpneumo
GC: 38%
GC: 40%
SPB
chr. XVI
Mpneumo
Telomere
RNA
Nucleolus
chr. XVI and Mmyco
GC: 38%
GC: 24%
SPB
Y chromatin
U chromatin
chr. XVI
Mmyco
Nucleolus
Translocated fused chr. XVI - Mmyco
Y chromatin
U chromatin
XVIfMmycot1
Nucleolus


Y chromatin compartment
U chromatin compartment
- Active transcription
- Contorted 3D structure
- Trans contacts w/ other chrs.
- No transcription
- Globular conformation
- Contacts w/ telomeres
G1
Y chromatin
U chromatin
Extruding
Anchored
Unanchored
Extruding
- Chromosome compaction
- Cohesin-mediated chromatin extrusion
- Focal anchoring at convergent transcription
- Chromosome compaction
- Cohesin-mediated chromatin extrusion
- No focal anchoring
G2/M

References

1. J. Romiguier, V. Ranwez, E. J. P. Douzery, N. Galtier, Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* **20**, 1001–1009 (2010).
2. J. Tajbakhsh, Spatial Distribution of GC- and AT-Rich DNA Sequences within Human Chromosome Territories. *Experimental Cell Research* **255**, 229–237 (2000).
3. A. Valouev, S. M. Johnson, S. D. Boyd, C. L. Smith, A. Z. Fire, A. Sidow, Determinants of nucleosome organization in primary human cells. *Nature* **474**, 516–520 (2011).
4. G. P. Holmquist, Evolution of chromosome bands: Molecular ecology of noncoding DNA. *J Mol Evol* **28**, 469–486 (1989).
5. L. Mirny, J. Dekker, Mechanisms of Chromosome Folding and Nuclear Organization: Their Interplay and Open Questions. *Cold Spring Harb Perspect Biol* **14**, a040147 (2022).
6. I. F. Davidson, J.-M. Peters, Genome folding through loop extrusion by SMC complexes. *Nat Rev Mol Cell Biol* **22**, 445–464 (2021).
7. A. S. Belmont, Nuclear Compartments: An Incomplete Primer to Nuclear Compartments, Bodies, and Genome Organization Relative to Nuclear Architecture. *Cold Spring Harb Perspect Biol* **14**, a041268 (2022).
8. A. Crisp, C. Boschetti, M. Perry, A. Tunnacliffe, G. Micklem, Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biol* **16**, 50 (2015).
9. N. B. Edelman, J. Mallet, Prevalence and Adaptive Impact of Introgression. *Annu Rev Genet* **55**, 265–283 (2021).
10. J. Van Etten, D. Bhattacharya, Horizontal Gene Transfer in Eukaryotes: Not if, but How Much? *Trends Genet* **36**, 915–925 (2020).
11. J. Peter, M. De Chiara, A. Friedrich, J.-X. Yue, D. Pflieger, A. Bergström, A. Sigwalt, B. Barre, K. Freel, A. Llored, C. Cruaud, K. Labadie, J.-M. Aury, B. Istace, K. Lebrigand, P. Barbry, S. Engelen, A. Lemainque, P. Wincker, G. Liti, J. Schacherer, Genome evolution across 1,011 Saccharomyces cerevisiae isolates. *Nature* **556**, 339–344 (2018).
12. V. Baby, Cloning and Transplantation of the Mesoplasma florum Genome. *ACS Synth Biol* **7**, 209–217 (2018).
13. A. L. V. Coradini, C. B. Hull, I. M. Ehrenreich, Building genomes to understand biology. *Nat Commun* **11**, 6177 (2020).
14. A. Currin, S. Parker, C. J. Robinson, E. Takano, N. S. Scrutton, R. Breitling, The evolving art of creating genetic diversity: From directed evolution to synthetic biology. *Biotechnol Adv* **50**, 107762 (2021).
15. C. Payen, G. Fischer, C. Marck, C. Proux, D. J. Sherman, J.-Y. Coppée, M. Johnston, B. Dujon, C. Neuvéglise, Unusual composition of a yeast chromosome arm is associated with its delayed replication. *Genome Res* **19**, 1710–1721 (2009).
16. F. Labroussaa, A. Lebaudy, V. Baby, G. Gourgues, D. Matteau, S. Vashee, P. Sirand-Pugnet, S. Rodrigue, C. Lartigue, Impact of donor–recipient phylogenetic distance on bacterial genome transplantation. *Nucleic Acids Research* **44**, 8501–8511 (2016).
17. E. Ruiz, V. Talenton, M.-P. Dubrana, G. Guesdon, M. Lluch-Senar, F. Salin, P. Sirand-Pugnet, Y. Arfi, C. Lartigue, CReasPy-Cloning: A Method for Simultaneous Cloning and Engineering of Megabase-Sized Genomes in Yeast Using the CRISPR-Cas9 System. *ACS Synth. Biol.* **8**, 2547–2557 (2019).
18. C. Lartigue, S. Vashee, M. A. Algire, R.-Y. Chuang, G. A. Benders, L. Ma, V. N. Noskov, E. A. Denisova, D. G. Gibson, N. Assad-Garcia, N. Alperovich, D. W. Thomas, C. Merryman, C. A. Hutchison, H. O. Smith, J. C. Venter, J. I. Glass, Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science* **325**, 1693–1696 (2009).
19. S. Karlin, J. Mrázek, Compositional differences within and between eukaryotic genomes.

Proceedings of the National Academy of Sciences **94**, 10227–10232 (1997).
20. K. Brogaard, L. Xi, J.-P. Wang, J. Widom, A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496–501 (2012).
21. M. T. Ocampo-Hafalla, Y. Katou, K. Shirahige, F. Uhlmann, Displacement and re-accumulation of centromeric cohesin during transient pre-anaphase centromere splitting. *Chromosoma* **116**, 531–544 (2007).
22. K.-L. Chan, M. B. Roig, B. Hu, F. Beckouët, J. Metson, K. Nasmyth, Cohesin's DNA Exit Gate Is Distinct from Its Entrance Gate and Is Regulated by Acetylation. *Cell* **150**, 961–974 (2012).
23. L. Lazar-Stefanita, V. F. Scolari, G. Mercy, H. Muller, T. M. Guérin, A. Thierry, J. Mozziconacci, R. Koszul, Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *EMBO J* **36**, 2684–2697 (2017).
24. Z. Duan, A three-dimensional model of the yeast genome. *Nature* **465**, 363–367 (2010).
25. H. Wong, H. Marie-Nelly, S. Herbert, P. Carrivain, H. Blanc, R. Koszul, E. Fabre, C. Zimmer, A predictive computational model of the dynamic 3D interphase yeast nucleus. *Curr Biol* **22**, 1881–1890 (2012).
26. A. Y. Grosberg, S. K. Nechaev, E. I. Shakhnovich, The role of topological constraints in the kinetics of collapse of macromolecules. *J. Phys. France* **49**, 2095–2100 (1988).
27. L. A. Mirny, The fractal globule as a model of chromatin architecture in the cell. *Chromosome Res* **19**, 37–51 (2011).
28. L. Dauban, R. Montagne, A. Thierry, L. Lazar-Stefanita, N. Bastié, O. Gadal, A. Cournac, R. Koszul, F. Beckouët, Regulation of Cohesin-Mediated Chromosome Folding by Eco1 and Other Partners. *Mol Cell* **77**, 1279-1293.e4 (2020).
29. L. Costantino, T.-H. S. Hsieh, R. Lamothe, X. Darzacq, D. Koshland, Cohesin residency determines chromatin loop patterns. *Elife* **9**, e59889 (2020).
30. S. A. Schalbetter, A. Goloborodko, G. Fudenberg, J.-M. Belton, C. Miles, M. Yu, J. Dekker, L. Mirny, J. Baxter, SMC complexes differentially compact mitotic chromosomes according to genomic context. *Nat Cell Biol* **19**, 1071–1080 (2017).
31. S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, E. L. Aiden, A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
32. A. Goloborodko, M. V. Imakaev, J. F. Marko, L. Mirny, Compaction and segregation of sister chromatids via active loop extrusion. *Elife* **5**, e14864 (2016).
33. J. H. I. Haarhuis, R. H. van der Weide, V. A. Blomen, J. O. Yáñez-Cuna, M. Amendola, M. S. van Ruiten, P. H. L. Krijger, H. Teunissen, R. H. Medema, B. van Steensel, T. R. Brummelkamp, E. de Wit, B. D. Rowland, The Cohesin Release Factor WAPL Restricts Chromatin Loop Extension. *Cell* **169**, 693-707.e14 (2017).
34. E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, J. Dekker, Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
35. N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, J. Dekker, Organization of the Mitotic Chromosome. *Science* **342**, 948–953 (2013).
36. J. H. Gibcus, K. Samejima, A. Goloborodko, I. Samejima, N. Naumova, J. Nuebler, M. T. Kanemaki, L. Xie, J. R. Paulson, W. C. Earnshaw, L. A. Mirny, J. Dekker, A pathway for mitotic chromosome formation. *Science* **359**, eaao6135 (2018).
37. G. Spracklin, Diverse silent chromatin states modulate genome compartmentalization and loop extrusion barriers. *Nat Struct Mol Biol* **30**, 38–51 (2023).
38. E. P. Nora, A. Goloborodko, A.-L. Valton, J. H. Gibcus, A. Uebersohn, N. Abdennur, J. Dekker, L. A. Mirny, B. G. Bruneau, Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell* **169**, 930-944.e22 (2017).
39. M. Kabi, G. J. Filion, Heterochromatin: did H3K9 methylation evolve to tame transposons?

Genome Biology **22**, 325 (2021).
40. L. L. Breeden, T. Tsukiyama, Quiescence in Saccharomyces cerevisiae. *Annual Review of Genetics* **56**, 253–278 (2022).
41. M. Guidi, M. Ruault, M. Marbouty, I. Loïodice, A. Cournac, C. Billaudeau, A. Hocher, J. Mozziconacci, R. Koszul, A. Taddei, Spatial reorganization of telomeres in long-lived quiescent cells. *Genome Biol.* **16**, 206 (2015).
42. K. Struhl, E. Segal, Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**, 267–273 (2013).
43. E. Routhier, A. Bin Kamruddin, J. Mozziconacci, keras_dna: a wrapper for fast implementation of deep learning models in genomics. *Bioinformatics* **37**, 1593–1594 (2021).
44. R. V. Chereji, D. J. Clark, Major Determinants of Nucleosome Positioning. *Biophys J* **114**, 2279–2289 (2018).
45. Z. Gvozdenov, Z. Barcutean, K. Struhl, Functional analysis of a random-sequence chromosome reveals a high level and the molecular nature of transcriptional noise in yeast cells. *Molecular Cell* **83**, 1786-1797.e5 (2023).
46. B. R. Camellato, R. Brosh, H. J. Ashe, M. T. Maurano, J. D. Boeke, Synthetic reversed sequences reveal default genomic states. *Nature* **628**, 373–380 (2024).
47. K. Struhl, E. Segal, Determinants of nucleosome positioning. *Nat Struct Mol Biol* **20**, 267–273 (2013).
48. I. Luthra, X. Chen, C. Jensen, A. M. Rafi, A. Salaudeen, C. G. de Boer, Biochemical activity is the default DNA state in eukaryotes. *bioRxiv*, doi: https://doi.org/10.1101/2022.12.16.520785 (2022).
49. P. A. Ginno, P. L. Lott, H. C. Christensen, I. Korf, F. Chédin, R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell* **45**, 814–825 (2012).
50. J. R. Lobry, Asymmetric substitution patterns in the two DNA strands of bacteria. *Mol Biol Evol* **13**, 660–665 (1996).
51. N. Bastié, C. Chapard, A. Cournac, S. Nejmi, H. Mboumba, O. Gadal, A. Thierry, F. Beckouët, R. Koszul, Sister chromatid cohesion halts DNA loop expansion. *Mol Cell* **84**, 1139-1148.e5 (2024).
52. E. J. Banigan, "Transcription shapes 3D chromatin organization by interacting with loop extrusion" in *Proceedings of the National Academy of Sciences 120* (2023), p. 2210480120.
53. B. J. H. Dequeker, M. J. Scherr, H. B. Brandão, J. Gassler, S. Powell, I. Gaspar, I. M. Flyamer, A. Lalic, W. Tang, R. Stocsits, I. F. Davidson, J.-M. Peters, K. E. Duderstadt, L. A. Mirny, K. Tachibana, MCM complexes are barriers that restrict cohesin-mediated loop extrusion. *Nature*, 1–7 (2022).
54. C. Allen, S. Büttner, A. D. Aragon, J. A. Thomas, O. Meirelles, J. E. Jaetao, D. Benn, S. W. Ruby, M. Veenhuis, F. Madeo, M. Werner-Washburne, Isolation of quiescent and nonquiescent cells from yeast stationary-phase cultures. *Journal of Cell Biology* **174**, 89–100 (2006).
55. M. F. Laughery, T. Hunter, A. Brown, J. Hoopes, T. Ostbye, T. Shumaker, J. J. Wyrick, New vectors for simple and streamlined CRISPR-Cas9 genome editing in Saccharomyces cerevisiae. *Yeast* **32**, 711–720 (2015).
56. N. Agier, A. Fleiss, S. Delmas, G. Fischer, A Versatile Protocol to Generate Translocations in Yeast Genomes Using CRISPR/Cas9. *Methods Mol Biol* **2196**, 181–198 (2021).
57. J. Luo, X. Sun, B. P. Cormack, J. D. Boeke, Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature* **560**, 392–396 (2018).
58. K. Labun, T. G. Montague, M. Krause, Y. N. Torres Cleuren, H. Tjeldnes, E. Valen, CHOPCHOP v3: expanding the CRISPR web toolbox beyond genome editing. *Nucleic Acids Res* **47**, W171–W174 (2019).
59. J.-P. Concordet, M. Haeussler, CRISPOR: intuitive guide selection for CRISPR/Cas9 genome editing experiments and screens. *Nucleic Acids Research* **46**, W242–W245 (2018).
60. T. W. Christianson, R. S. Sikorski, M. Dante, J. H. Shero, P. Hieter, Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110**, 119–122 (1992).

61. R. Koszul, S. Caburet, B. Dujon, G. Fischer, Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments. *The EMBO Journal* **23**, 234–243 (2004).
62. D. Viterbo, A. Marchal, V. Mosbach, L. Poggi, W. Vaysse-Zinkhöfer, G.-F. Richard, A fast, sensitive and cost-effective method for nucleic acid detection using non-radioactive probes. *Biology Methods and Protocols* **3**, bpy006 (2018).
63. L. Dauban, R. Montagne, A. Thierry, L. Lazar-Stefanita, N. Bastié, O. Gadal, A. Cournac, R. Koszul, F. Beckouët, Regulation of Cohesin-Mediated Chromosome Folding by Eco1 and Other Partners. *Molecular Cell* **77**, 1279-1293.e4 (2020).
64. B. Hu, N. Petela, A. Kurze, K.-L. Chan, C. Chapard, K. Nasmyth, Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq. *Nucleic Acids Res* **43**, e132 (2015).
65. D. Challal, M. Barucco, S. Kubik, F. Feuerbach, T. Candelli, H. Geoffroy, C. Benaksas, D. Shore, D. Libri, General Regulatory Factors Control the Fidelity of Transcription by Restricting Non-coding and Ectopic Initiation. *Mol Cell* **72**, 955-969.e7 (2018).
66. S. Granneman, G. Kudla, E. Petfalski, D. Tollervey, Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proceedings of the National Academy of Sciences* **106**, 9613–9618 (2009).
67. B. Langmead, S. L. Salzberg, Fast gapped-read alignment with Bowtie 2. *Nat Meth* **9**, 357–359 (2012).
68. A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, J. Mozziconacci, Normalization of a chromosomal contact map. *BMC Genomics* **13**, 436 (2012).
69. N. Abdennur, L. A. Mirny, Cooler: scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics* **36**, 311–316 (2020).
70. F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, T. Manke, deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res* **44**, W160-165 (2016).
71. P. Danecek, J. K. Bonfield, J. Liddle, J. Marshall, V. Ohan, M. O. Pollard, A. Whitwham, T. Keane, S. A. McCarthy, R. M. Davies, H. Li, Twelve years of SAMtools and BCFtools. *Gigascience* **10**, giab008 (2021).
72. J. Serizay, R. Koszul, Epigenomics coverage data extraction and aggregation in R with tidyCoverage. *Bioinformatics* **40**, btae487 (2024).
73. J. Serizay, C. Matthey-Doret, A. Bignaud, L. Baudry, R. Koszul, Orchestrating chromosome conformation capture analysis with Bioconductor. *Nat Commun* **15**, 1072 (2024).
74. Y. Liao, G. K. Smyth, W. Shi, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
75. M. I. Love, W. Huber, S. Anders, Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**, 550 (2014).
76. J. Serizay, J. Ahringer, periodicDNA: an R/Bioconductor package to investigate k-mer periodicity in DNA. *F1000Res* **10**, 141 (2021).
77. M. Jiang, J. Anderson, J. Gillespie, M. Mayne, uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics* **9**, 192 (2008).
78. D. G. Batrakou, C. A. Müller, R. H. C. Wilson, C. A. Nieduszynski, DNA copy-number measurement of genome replication dynamics by high-throughput sequencing: the sort-seq, sync-seq and MFA-seq family. *Nat Protoc* **15**, 1255–1284 (2020).
79. C. Matthey-Doret, L. Baudry, A. Breuer, R. Montagne, N. Guiglielmoni, V. Scolari, E. Jean, A. Campeas, P. H. Chanut, E. Oriol, A. Méot, L. Politis, A. Vigouroux, P. Moreau, R. Koszul, A. Cournac, Computer vision for pattern detection in chromosome contact maps. *Nature Communications* **11**, 5795 (2020).
80. Keras 3: A new multi-backend Keras, Keras (2023); https://github.com/keras-team/keras.
81. M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V.

Vasudevan, P. Warden, M. Wicke, Y. Yu, X. Zheng, "TensorFlow: a system for large-scale machine learning" in *Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation* (USENIX Association, USA, 2016)*OSDI'16*, pp. 265–283.
82. E. Routhier, E. Pierre, G. Khodabandelou, J. Mozziconacci, Genome-wide prediction of DNA mutation effect on nucleosome positions for yeast synthetic genomics. *Genome Res* **31**, 317–326 (2021).
83. D. R. Kelley, Y. A. Reshef, M. Bileschi, D. Belanger, C. Y. McLean, J. Snoek, Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Res* **28**, 739–750 (2018).
84. A. Majdandzic, C. Rajesh, P. K. Koo, Correcting gradient-based interpretations of deep neural networks for genomics. *Genome Biol* **24**, 109 (2023).
85. T. L. Bailey, J. Johnson, C. E. Grant, W. S. Noble, The MEME Suite. *Nucleic Acids Res* **43**, W39-49 (2015).
86. M. Gotta, T. Laroche, S. M. Gasser, Analysis of nuclear organization in Saccharomyces cerevisiae. *Methods Enzymol* **304**, 663–672 (1999).
87. J. G. Henikoff, J. A. Belsky, K. Krassovsky, D. M. MacAlpine, S. Henikoff, Epigenome characterization at single base-pair resolution. *Proc Natl Acad Sci U S A* **108**, 18318–18323 (2011).
88. M. Bodmer-Glavas, K. Edler, A. Barberis, RNA polymerase II and III transcription factors can stimulate DNA replication by modifying origin chromatin structures. *Nucleic Acids Research* **29**, 4570–4580 (2001).
89. A. Koren, S. A. McCarroll, Random replication of the inactive X chromosome. *Genome Res* **24**, 64–69 (2014).
90. K. Nasmyth, Disseminating the genome: joining, resolving, and separating sister chromatids during mitosis and meiosis. *Annu. Rev. Genet.* **35**, 673–745 (2001).

1.2. Résultats supplémentaires

1.2.1 Analyse protéomique des chromosomes bactériens chez *S. cerevisiae*

Les chromosomes bactériens, en particulier le chromosome Mpneumo, sont spontanément transcrits. **On peut alors se demander si les transcrits bactériens sont traduits** ? En collaboration avec la plateforme de protéomique, nous avons réalisé des analyses par spectrométrie de masse (Bernard et al., 2022). Nous avons comparé le protéome de la souche sauvage (W303), de la souche XVIfMmyco (le chromosome Mmyco est fusionné au chromosome XVI) et de la souche XVIfMpneumo (le chromosome Mpneumo est fusionné au chromosome XVI).

Pour identifier les séquences codantes, nous avons généré les séquences protéiques, c'est-à-dire, d'acide aminés, comprenant les annotations de la levure S288c et les annotations bactériennes. D'abord, les ORFs eucaryotes putatives des séquences bactériennes sont générés puis les séquences identifiées sont traduites en séquences d'acide aminé en utilisant le code eucaryote. Le traitement des données générées après la spectrométrie de masse a été réalisé avec le logiciel Maxquant. Nous avons identifié 2983 protéines, réparties entre les trois conditions (**Figure 1. A**).

Seize protéines correspondant à des ORFs de protéines bactériennes ont été détectées. Parmi ces seize candidats, nous nous sommes concentrés sur les protéines ayant au moins quatre peptides associés (**Figure 1. B**). Dans la condition Mpneumo, une seule protéine est identifiée avec 5 peptides. Dans la condition Mmyco, deux protéines sont identifiées. Pour ces trois protéines, nous avons vérifié si un transcrit était détecté, et c'est le cas pour les trois (exemple de ftsZ en **Figure 1. C**).

Nous avons donc identifié trois protéines putatives :

- Une pour Mpneumo, qui pourrait correspondre au gène ptsH (protéine phosphocarrière HPr) ou à une ORF proche.
- Deux pour Mmyco, qui pourraient correspondre aux gènes ftsZ (GTPase impliquée dans la division cellulaire procaryote) et rplK (protéine de la sous-unité ribosomique 50S) ou à des ORFs proches.

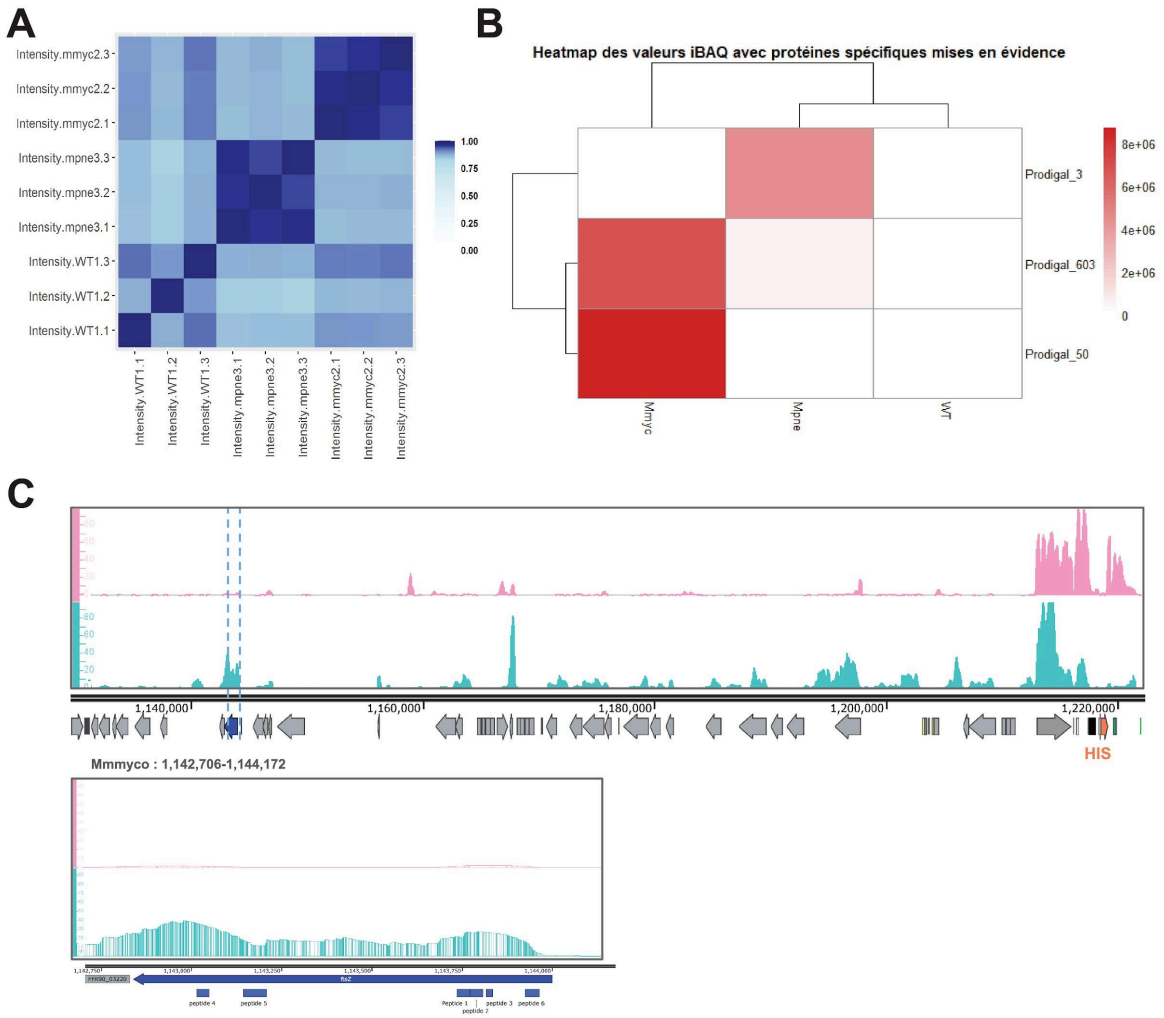


Figure 1. Trois protéines bactériennes putatives semblent être exprimées.

A) Cluster entre les différentes conditions, réalisées en triplicats. **B)**Heatmap des valeurs iBAQ des 3 protéines bactériennes putatives. iBAQ : Valeur calculée pour une protéine donnée correspondant à la somme des intensités de tous les peptides détectés divisée par le nombre de peptides théoriques observables (peptide d'une longueur comprise entre 6 et 30 acides aminés obtenu après digestion in silico de la protéine par la trypsine). Cette valeur donne une idée de l'abondance (relative) d'une protéine dans l'échantillon par rapport aux autres. **C)** Capture d'écran IGV des régions Mmmyco:1,129,581-1,222,199 et 1,142,706-1,144,172. Identification des peptides traduits du chromosome bactérien. L'identifiant "Prodigal_50" semble correspondre au gène bactérien ftsZ.

Ces résultats suggèrent que quelques protéines issues des chromosomes bactériens sont exprimées. Ces résultats restent préliminaires et nécessitent une validation expérimentale, par exemple par Western blot avec des anticorps spécifiques (par exemple en utilisant un anticorps anti ftsZ) ou l'ajout d'un tag dans les séquences codantes identifiées. Actuellement, le jeu de données nous permet d'avoir 2300 protéines identifiées, et il serait pertinent de faire une analyse différentielle des différentes protéines, levures et bactériennes, exprimées dans chaque condition.

2. Ingénierie d’une séquence d'ADN aléatoire dans un noyau eucaryote

2.1. Article 2 : Activité et structure d'une séquence aléatoire à travers différentes sources de carbone chez *Saccharomyces cerevisiae*

2.1.1 Article

La transcription est un processus omniprésent qui non seulement régule le métabolisme cellulaire, mais joue également un rôle clé dans la modulation ou la conduite de multiples niveaux de structuration des chromosomes. Les règles qui régulent la transcription des séquences eucaryotes sont de mieux en mieux comprises et les combinaisons de motifs placés à diverses positions *cis* des promoteurs laissant entrevoir des règles complexes. L'intégration d’ADN aléatoire dans un génome a été proposé comme solution pour explorer les principes de la régulation transcriptionnelle en l'absence de contraintes évolutives. Cependant, la plupart des travaux antérieurs exploitant cette stratégie se concentrent principalement sur la transcription sans tenir compte de l’organisation du génome.

Dans ce second projet, en co-autrice avec Hélène Bordelet, nous tirons parti d'une séquence aléatoire de 100 kb intégrée dans le chromosome IV de *S. cerevisiae* pour explorer, dans différentes conditions métaboliques, non seulement l’activité transcriptionnelle de cette séquence, mais aussi son organisation tridimensionnelle.

Nous montrons que l'assemblage des nucléosomes, l'activité transcriptionnelle et le repliement 3D de cette région dépendent de la source de carbone. La séquence aléatoire est non seulement plus sensible aux changements de sources de carbone mais aussi à la voie de dégradation NNS. Finalement, nous démontrons ensuite que nous maîtrisons suffisamment les règles liant les déterminants de la séquence, la transcription et l'organisation, pour introduire des changements clés qui nous permettent de la façonner à notre guise.

Hélène Bordelet m’a encadrée et formée tout au long du projet. Nous avons réfléchi ensemble à la conception du projet et conçu les constructions génétiques nécessaires au projet. J’ai pu réappliquer les méthodes développées dans le premier projet sur la séquence synthétique (MNase-seq, ATAC-seq), ainsi que les méthodes du laboratoire (ChIP-seq, Hi-C, extraction d’ARN pour le RNA-seq). Nous avons généré et caractérisé plusieurs constructions génétiques intégrées dans la région Syn100. Nous avons réalisé la majorité des analyses, l’interprétation des données et la génération des figures avec l’aide de Jacques Serizay. Enfin, nous avons rédigé ensemble le manuscrit.

Bien qu'ils soient présentés sous forme d'article, certaines expériences nécessitent d'être reproduites et approfondies pour renforcer nos conclusions. De plus, de nombreuses analyses complémentaires et plus systématiques doivent être réalisées avec les différents jeux de données.

Je présenterai d'abord le papier avec une explication des prochaines expériences et analyses prévues. Puis je présenterai des résultats supplémentaires où nous essayons de mettre en place une régulation longue-distance dans le système de la levure.

Activity and structure of a random sequence through different carbon sources in *Saccharomyces cerevisiae*

Léa Meneu[1,2], Hélène Bordelet[1,2,#], Agnès Thierry[1], Jacques Serizay[1], Fabien Girard[1], Alexandros Minakakis[3], Domenico Libri[3], Romain Koszul[1,#]

Affiliations

[1] Régulation spatiale des génomes, Institut Pasteur, CNRS UMR3525, 75015 Paris, France

[3] Institut de Génétique Moléculaire de Montpellier (IGMM), 34090 Montpellier, France

[2] These authors contributed equally.

Abstract

Transcription is an ubiquitous process that not only regulates cellular metabolism but also plays key roles in modulating or driving multiple levels of chromosome structuration. The rules that regulate transcription of eukaryotic sequences are increasingly well understood, with combinations of motifs positioned at various cis-positions from promoters hinting at intricate rules. Random DNA has been proposed as a solution for exploring the principles of transcriptional regulation in the absence of evolutionary constraints. However, most previous work exploiting this strategy were based on relatively short random DNA sequences of a few kb or on oligonucleotides, and focused primarily on transcription. Here, we take advantage of a random 100 kb sequence in *Saccharomyces cerevisiae* to explore, under different metabolic conditions, not only the spontaneous transcription patterns of this sequence but also its three-dimensional organization. We show that nucleosome assembly, transcriptional activity and 3D folding of this region depends on the carbon source. From this random sequence, we then demonstrate that we have sufficient mastery of the rules linking sequence determinants, transcription and organization, to introduce key changes that allow us to shape it as we wish.

Introduction

Random and/or exogenous DNA has been proposed as a powerful and convenient solution to explore the principles of transcriptional regulation in the absence of evolutionary constraints. Various teams have used random, non-biological, or exogenous DNA in yeast studies, including an 18 kb synthetic sequence (Gvozdenov et al., 2023), a 254 kb dataStorage Chromosome, "dChr" (Luthra et al., 2022; Zhou, 2022), complete exogenous bacterial genomes from Mycoplasma species (Chapard et al., 2023), or also human YACs (Luthra et al., 2022), and reversed human gene (Camellato et al., 2022). These sequences, which lack evolved yeast regulatory sequences, spontaneously form nucleosomes (Chapard et al., 2023; Gvozdenov et al., 2023) and show discrete RNA products and active chromatin signatures in yeast (Camellato et al., 2022; Chapard et al., 2023; Gvozdenov et al., 2023; Luthra et al., 2022). However, these random sequences are often too short for in-depth studies of transcription regulation, such as searching for motifs and regulators. Plus, the characterisation of the chromatin and transcriptional activity of these regions have only been studied under glucose conditions, despite pervasive transcription being influenced by different culture conditions (Nevers et al., 2018). Finally, most previous work exploiting this strategy focused primarily on transcription and not on 3D structures and their regulation.

Here, we take advantage of a 100kb size random, homogeneous GC composition of 52% (called Syn100) inserted in a yeast chromosome. Characterization of this region showed that the region is chromatinized and transcribed in agreement with previous works. Syn100 is folded by cohesins in G2/M and this folding is dictated by transcriptional convergence, similarly to the native *S. cerevisiae* genome. Characterization of Syn100 in different carbon sources reveals two levels of transcriptional and post-transcriptional regulation that predominantly shape the transcriptome of the Syn100 sequence and not the yeast genome. Finally, by fine-tuning and regulating transcription termination we were able to control the 3D organization of Syn100 and form a discrete and reproducible cohesin-dependent loop between two loci of choice. This work is a proof of concept for the use of an innovative synthetic genomics approach to tackle complex biological problems.

Results

Design and chromosomal integration of Syn100, a 100kb random DNA region

To characterize the chromatin composition, activity, and 3D folding of an unbiased, non-evolved large DNA molecule, we designed *in silico* a 100 kb DNA random region (GC%: 52%) referred to as Syn100. Syn100 was first assembled on a centromeric plasmid using targeted associated recombination, (see **Methods**). To increase stability of the region and control its copy number that may result from the circularity of the molecule, Syn100 was then transferred into *S. cerevisiae* chr. IV right arm using a CRISPR-Cas9 targeted approach (Agier et al., 2021) (**Fig. 1A, Methods)**. The strain carrying Syn100 did not display a growth defect (**Fig. S1A**).

Nucleosomes form spontaneously and lead to spurious transcription

We characterized the nucleosome pattern of Syn100 by performing pair-end MNase-seq on cells growing in glucose medium (**Methods**). Syn100 region displayed clear nucleosome patterns, similar to that of the neighboring and ectopic native regions (**Fig. 1B**). Nucleosome arrays are similar to yeast's with a linker length of ~14 bp and a nucleosome repeat length (NRL) of 160 bp (**Fig. 1D, E**). Nucleosome depleted regions (NDR) are less frequent in Syn100 compared to the clear chromatin pattern of yeast chromosomes (**Fig. 1C**). In Syn100, nucleosomes are non-randomly positioned, and few NDR regions are detected. These observations are consistent with a recent study exploring the chromatin of a circular 18-kb DNA segment with a 50% GC (Gvozdenov et al., 2023).

We next performed RNA-seq to measure transcription levels along Syn100, and quantified RNA levels from random-sequence and genomic DNA. Syn100 is fully transcribed with most RNAs present at levels comparable to those from genomic DNA (**Fig. S1B**), consistent with previous studies (Chapard et al., 2023; Gvozdenov et al., 2023; Luthra et al., 2022).

The yeast transcriptome has a clear transcriptional architecture, with distinct and spaced sense and antisense transcription. The stranded RNA-seq data showed that the Syn100 region is split into large (~20 kb) RNA tracts for the most part oriented unidirectionally, with a few overlapping positions (**Fig. 1B**, red triangle). To test whether the unidirectional tracks reflect directly de novo transcription or RNA molecule following post-transcriptional

regulation, we profiled nascent transcripts by CRAC-seq, a technique that quantifies RNA molecules still bound to RNA Pol II, i.e. prior to post-transcriptional regulation. CRAC-seq tracks reveal an accumulation of peaks at the level of the beginning of transcription tracks observed with the steady-state RNAseq stranded analysis (**Fig. S1D**, red triangle). This result suggests that these unidirectional tracks reflect directly transcription and are not post-transcriptional regulation of bidirectional transcription.

To further confirm this hypothesis, we explored the positioning of nucleosome-depleted regions (NDRs) along the Syn100 sequence with respect to transcription. Indeed, NDRs act as bidirectional promoters that give rise to equal levels of transcription in both directions(Xu et al., 2009). The positions of transcripts identified in both RNA-seq and CRAC-seq experiments correlate with NDRs sites, and result in regularly, highly positioned nucleosomes similar to NDRs along yeast chromosomes (**Fig. 1C, Fig. S1E**). We then explored whether the main motifs of promoter sequences could play a role in NDRs formation and transcription. In Syn100, 29 TATA boxes (TATAWAWR) and 454 TATA-like are detected. Moreover, for the initiator motif (INR, YYANWYY) 1359 for Syn100 against 2199 for yeast. The fact that there are fewer NDRs coupled to fewer TATA boxes may explain why transcription track ranges are longer and fewer (~7 tracks) along the random 100 kb Syn100 sequence than in the yeast genome where 100 kb are usually covered by ~50 genes. Altogether, these observations suggest that few NDRs in Syn100 leads to long tracks of bidirectional transcription.

Syn100 spontaneously reorganizes into three dimension during metaphase

To study how a long, non-evolved random sequence folds in three dimensions in the nuclear space, we generated Hi-C contact maps of the strain synchronized in G1 and G2/M. The positioning of the SMC cohesin, a major chromatin organizer of the yeast genome during G2/M (Costantino et al., 2020; Dauban et al., 2020), was further assessed using ChIP-seq of the Scc1 subunit (**Methods**). For all experiments, the carbon source of the growth media is glucose (**Methods**). Synchronization was verified using flow cytometer (**Fig. S2A**), as well as properties displayed by Hi-C contact maps of the native yeast genome in G2/M and G1. In G1 the contact frequency as a function of genomic distance (i.e. the contact decay curve, or p(s)) was flat, reflecting the uncompacted genome whereas in G2/M, it was characteristic of yeast chromosomes compacted into arrays of chromatin loops at this stage (**Fig. 1F; S2B,C ;S2D,E**)(Dauban et al., 2020; Lazar-Stefanita et al., 2017; Schalbetter et al., 2017).

Despite a slight reduction of contacts between 10 -100 kb compared to control regions, the p(s) of Syn100 is rather flat in G1, probably reflecting the poor structuration of yeast genome at this stage (Fig 1F). However in G2/M, Syn100 DNA shows compaction as p(s) curve is displaced with an enrichment of contact within 15 - 100 kb and as the diagonal of the contact map is thickened (**Fig. 1F-H**). On average, a 100 kb region of the native *S. cerevisiae* genome displays approximately 4 to 6 cohesin mediated chromatin loops (**Fig. S2D,E**)(Costantino et al., 2020), with enrichment on cohesin peaks overlapping with Hi-C maps loop basis in G2/M. In contrast to yeast chromosomes, no discrete loop patterns are detected along the Syn100 contact map (**Fig. 1H**). These regions nevertheless display an atypical contact signal in the form of stripe-like structures (**Fig. 1H**, white triangle) where cohesin accumulates. Cohesin accumulation at the basis of both Syn100 stripe-like structure and *S. cerevisiae* chromatin loops coincides with transcriptionally convergent regions demonstrating that cohesin deposition on Syn100 obeys the same rules as for the native yeast genome (**Fig. 1H**). One base of this stripe-like structure, as for native to loop bases corresponds to an area of narrow, reproducible transcriptional convergence (**Fig. 1H (1**)). In contrast, the other base of the structure is long and diffused coinciding with overlapping sense and antisense transcription (**Fig. 1H (2)).** This difference of structure may be explained by the diffuse nature of this large transcriptional convergence that deposits cohesin in the region with low reproducibility at the cell population scale. It is interesting to note that the p(s) inflection point of Syn100 is displaced toward longer distances compared to the control flanking regions. This is consistent with the small frequency of cohesin borders in the regions.

As the native yeast genome, Syn100 3D organization is cell cycle dependent. In G2/M Syn100 gets compacted by the cohesin complex deposited at more or less defined converging transcription sites.

Transcription pattern changes dramatically depending on carbon source

Naive regions have only been characterized under glucose conditions. How these random, unevolved regions behave under other carbon sources has not yet been studied.

We performed Mnase-seq and RNAseq in galactose and lactate (**Methods**). Compared to glucose, Syn100 nucleosome arrays have a longer linker length in galactose and lactate (**Fig. 2C**). Autocorrelations of the nucleosome track also show a faster decline in galactose and lactate than in glucose suggesting that chromatin is less periodically arranged in these

carbon sources (**Fig. 2D**). Strikingly, total RNA-seq profiles showed higher signals and levels in galactose and lactate compared to glucose (**Fig. 2A, B**). This increase affects the entire Syn100 region, unlike the rest of the genome, where only a part of RNAs are over-represented (**Fig. 2A**).

We tested if this global increase of transcription on Syn100 in galactose condition was due to the recruitment of the galactose activator Gal4. We performed a ChIP-seq against Gal4 and enrichments of Gal4 coincides with UAS motifs identified in Syn100 sequence (**Fig. SX**). Interestingly, Syn100 shows a much higher density of fortuitous UAS sites than in the native *S. cerevisiae* genome (28 against 4 per 100 kb), suggesting that evolution has limited the abundance of this motif, restricting transcription activation upon galactose. Gal4 enrichments coincide with specific increase in RNA signal in galactose condition (**Fig. 2B**) and nucleosome reorganization with the apparition of NDRs as natively observed at Gal7, Gal10 and Gal1 genes (**Fig. S3F G**). In conclusion, Gal4 promotes Syn100 transcription in galactose condition. However, Syn100 RNAs are also increased at distances from Gal4 sites in galactose, and also in lactate condition which is where Gal4 is not active anyway suggesting that another regulation is driving the sharp increase in Syn100 transcripts.

Syn100 3D organization by cohesin is regulated by carbon source

In other carbon sources, the Syn100 transcriptional pattern changes and transcription affects cohesin deposition along the chromatin and 3D DNA folding(Lengronne et al., 2004). To investigate the impact of these changes, we generated Hi-C contact maps of both strains synchronized in G1 and G2/M and Scc1 ChIP in galactose and lactate. We confirmed synchronization by flow cytometry (**Fig. S4A**) and p(s) (**Fig. S4B**). Galactose induction is confirmed by the strong border detected at the GAL genes on the Hi-C map in galactose but not in lactate (**Fig. S4C**). In G1, no major changes are detected between the 3 carbon conditions (**Fig. S4D**). In G2/M, the stripe-like structure detected in glucose is lost in both galactose and lactate conditions (**Fig. 2E,F**). This could be explained by the strong modification of transcription activity in the region disrupting the sharp transcriptional convergence site where cohesin accumulated in glucose (**Fig. 1H (1)**). Disappearance of cohesin accumulation at this site is confirmed by Scc1 ChIP in galactose and lactate (**Fig. 2E,F**). Taken together, these results suggest that Syn100 organization is influenced by carbon sources that impacts transcription and consequently cohesin accumulation.

RNA degradation could be impaired by carbon source

The massive increase in Syn100 RNAs under galactose and lactate conditions could also be associated with a slowdown in their degradation in both carbon sources. To test this hypothesis, we performed total RNA-seq in glucose in mutants of two RNA degradation pathways : RRP6 (involved in the Nrd1-Nab3-Sen1 pathway, nuclear regulation) and UPF1 (Nonsense-mediated mRNA decay, cytoplasmic regulation). First, we confirmed the mutants were functional by checking the expected increase of CUTs (**Fig. S5A**),(Malabat et al., 2015).

All Syn100 transcripts are elevated in rrp6Δ & rrp6Δ/upf1Δ but not in upf1Δmutants suggesting that 1) almost all Syn100 transcriptome is unstable, and 2) it is regulated by the NNS RNA is degradation pathway in glucose (**Fig. 3A**).

The over-representation of Syn100 RNAs in the rrp6Δ/upf1Δ mutant is similar to the increase detected in WT galactose or lactate conditions (**Fig. 2A, 3A,B**). Unlike in glucose, rrp6Δ/upf1Δ deletions don't cause a massive increase of RNAs in galactose or lactate (**Fig. S5D**). This apparent absence of additivity suggests that at least one of the two RNA degradation pathways is inhibited in galactose and lactate conditions. Consistently, RNA-seq results show that several genes encoding NNS factors (Nrd1) and exosome factors (RRP41, RRP42, RRP46, RRP40, MTR3) involved in the NNS pathway are less expressed in galactose/lactate than glucose supporting a downregulation of the NNS pathway in those carbon sources. As the NNS pathway plays a crucial role in the orientation of transcripts, its downregulation may explain the increase of overlapping transcription observed in galactose and lactate. Systematic analysis will confirm if in general transcripts regulated by NNS are also overrepresented in gal/lac. In addition, the expression of many genes involved in translation and ribosome biogenesis is repressed in galactose and lactate (**Fig. S5C**). Since the Nonsense-mediated mRNA decay pathway is coupled to translation, it's reasonable to consider that this degradation pathway could also be slowed down in galactose and lactate. In conclusion, Syn100 is pervasively and strongly transcribed but generated RNAs are rapidly degraded by the NNS pathway in glucose. In contrast, at least one RNA degradation pathway seems to be downregulated in galactose and in lactate, partly explaining the high amount of Syn100 RNAs detected. Thus carbon source seems to affect synthetic RNA persistence in the cell.

Depending on the two-carbon source, we observe an intertwining of 1) transcriptional modifications explained by the presence of UAS and 2) post-transcriptional modifications

with reduced expression of degradation-related pathways. Naive sequences are more prone to instability and an increase in transcripts that have not evolved and lack the functional elements to constrain transcripts.

Syn100 folding by cohesin can be fine tuned to create a stable and reproducible loop

The interplay between loops and cohesin are now well known and described. In budding yeast, a cohesin-enriched site forms loops another the adjacent cohesin-enriched site mostly accumulating at converging transcription sites. Syn100 is compacted by the cohesin complex in G2/M (**Fig. 1H**). If the Syn100 contact map shows no canonical cohesin loop pattern, "stripe-like" structures are present between sharp and less defined cohesin enrichment sites in glucose (**Figure 4A, 1**). We wondered if we could transform this stripe-like pattern into a stable and reproducible loop by modifying the sequence of Syn100. We took advantage of the first, well defined, cohesin-enriched site "base 1" and decided to engineer a second site 32 kb apart, "base 2" (**Figure 4A, 1**). Via a CRISPR-Cas9 approach we inserted different constructions at the base 2 to dissect the sequences involved in loop formation that were checked by Hi-C and Scc1 ChIP.

First, we tested whether removing the 10 kb-long poorly defined transcriptional converging region or replacing it by a GFP coding sequence (Δ10kb strain and Δ10kb_GFP) would fine-tune the "strip-like" pattern into a defined loop. Both constructions were not sufficient to form a loop or an accumulation of cohesin (**Fig 4A 2, 3**). Then, we inserted the GFP coding sequence flanked by two terminators, t_{GUO}, sufficient for mRNA 3'-end formation in *S. cerevisiae* and to arrest transcription (Guo et al, 1996, curran et al 2015). These two 39 bp were designed *in silico* to stop the "pervasive" transcription from Syn100 (Δ10kb_GFP + tGUO terminators). On the Hi-C contact map, a dot is observed at the intersection of the two bases: base 1, which corresponds to the base initially present, and base 2, genetically engineered (**Fig. 4A 4, S6A**). The final construct inserted, where the 10kb of the poorly defined transcriptional converging region is replaced by, GFP gene, the pGAL1 promoter, terminator ADH1 and two t_{GUO} terminators, reveals two dots (**Fig. 4A 5).** The first one, between base 1 and 2 is better defined (**Fig 4A.4, S6A**) than with the previous construct (**Fig 4A.3, S6A**). The second loop corresponds to base 2 and another cohesin-enriched site at the end of the Syn100 insertion, base 3 (**Figure 4A.4, 4B , S6A**). ChIP cohesin reveals two successive peak of cohesin in the construct (Δ10kb_GFP + tGUO terminators) whereas in the final construct we have a single well-defined peak (**S6A**). These experiments were carried

out in glucose. The pGAL1 promoter was inactive and the GFP gene was not transcribed. The terminators therefore appear to be the key elements in forming this loop, probably by generating a zone of fine transcriptional convergence that favors the discrete accumulation of cohesins.

Thus, we can precisely modify and position cohesin-dependent loops in a random region by playing with transcriptional terminator sequences.

We then wondered how this modified Syn100 sequence is organized depending on carbon sources. To test if the loop is maintained in other carbon sources, we performed Hi-C on the strain Δ10kb_pGAL1-GFP grown in lactate and galactose. In galactose, the reproducible loop between base 1 and base 2 is lost because of the loss of cohesin accumulation at the base 1 already observed in the original sequence (**Fig. 4C**). As expected, in galactose condition, the inserted sequence pGAL1-GFP is transcribed thanks to pGAL1 activation. In a strain expressing GAL4-ER-VP16 fusion protein, allowing inducible activation of GAL promoters in response to estradiol, even in the presence of glucose. In this condition, loop pattern is conserved in spite of pGAL1-GFP activation. Surprisingly, in lactate condition, the Syn100 region reveals an organization with numerous well-defined loops. In conclusion, carbon source influences 3D folding of a naive sequence possibly by impacting its transcription.

Discussion

Various teams have used random, non-biological, or exogenous DNA in yeast studies. Here, we characterized a 100 kb random sequence integrated into a yeast chromosome. Characterization of the nucleosomes and transcription shows a spontaneously transcribed region, mainly linked to the NDR and impacting the 3D cohesin-dependent organization. Remarkably, changing carbon sources reveals that the Syn100 region is sensitive to environmental growth conditions. The rrp6Δ mutant reveals that transcripts of this naive sequence are regulated by the exosome. We uncovered that the strong increase in RNA level from Syn100 detected in galactose or lactate is likely associated with a decay in NNS degradation in these carbon sources. Finally, we were able to modify and set-up a cohesive-dependent loop by using transcriptional terminator sequences. We show that a naïve sequence is more prone to transcriptional and post-transcriptional alterations driven by carbon sources than the native genome, affecting multiple layers of organization.

Transcriptional and post-transcriptional changes

We have exploited a naive sequence, with a GC content of 52%, which explains the presence of sequences usually repressed in the host genome, such as UAS sequences. These sequences are rare in the yeast genome and not in the Syn100 region, and respond particularly well under galactose conditions. We have shown that, under different carbon source conditions, there is a difference in the expression of NNS and translation pathways. These changes could explain the increase in naïve transcripts detected. This raises the question of whether, under stress conditions, yeast allows the expression of new transcripts potentially useful for yeast. Surprisingly, these sequences have many stop codons and are therefore more likely to be taken up by the NMD system. However, the UPF1 mutant had no impact on the amount of RNA detected. The naïve region seems to be more dependent on the nuclear degradation pathway.

Random sequence engineering

We have shown that we can precisely modify a sequence and create stable, reproducible loops. Carbon sources have a considerable effect on the 3D organization of this naïve sequence, and these observations should be taken into account for future synthetic genomics approaches. Interestingly, the use of synthetic promoters that enable the activation of regulatory sequences without changing the carbon source seems to be an alternative to constraint experiments in one carbon source. Genetic engineering of these random sequences would be nice tools for deciphering different biological processes and could be platforms for revealing structures from nucleosomes to higher structures.

Figures

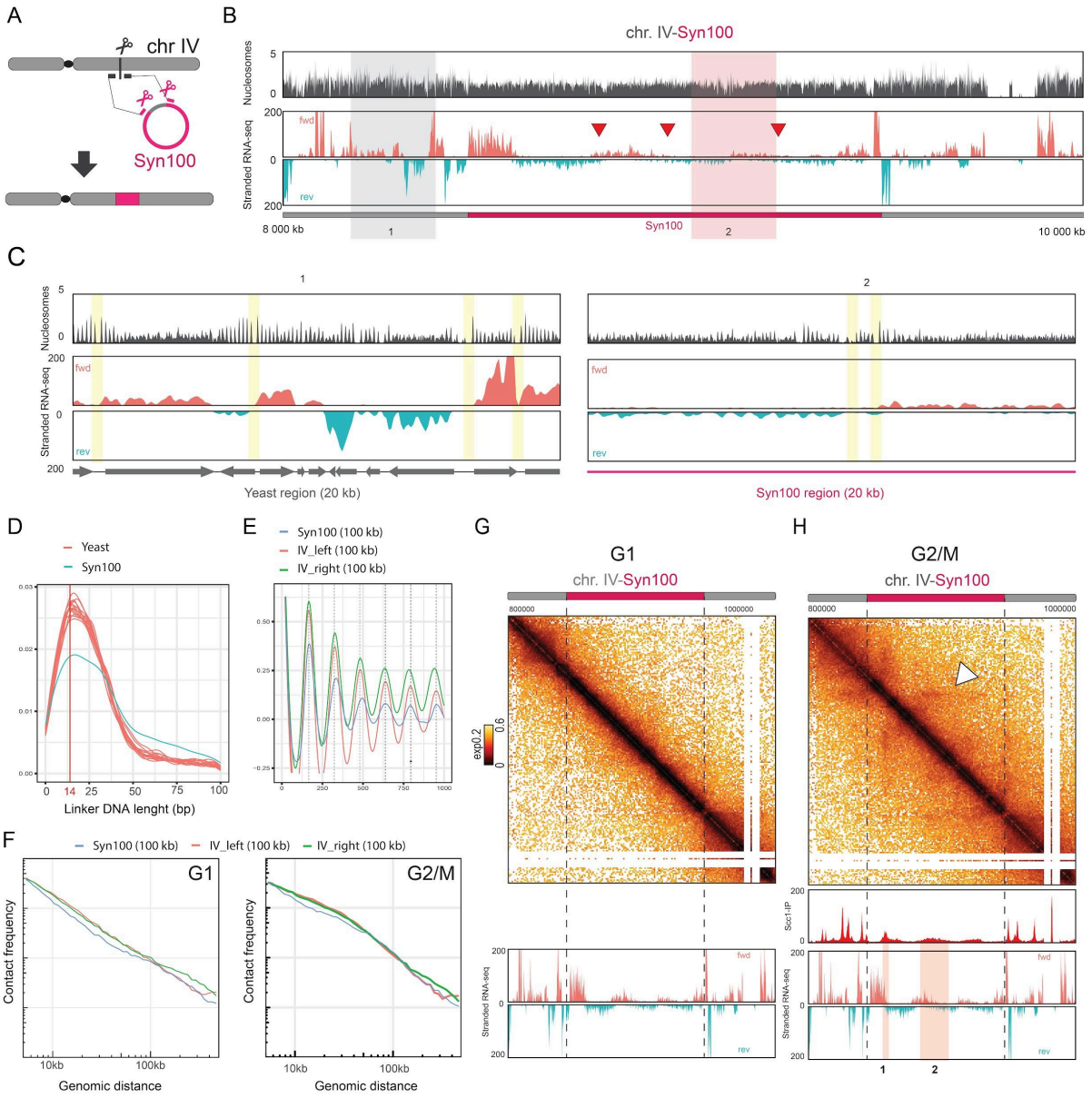

A
chr IV
Syn100
B
chr. IV-Syn100
Nucleosomes
Stranded RNA-seq
fwd
rev
8 000 kb
Syn100
10 000 kb
C
Yeast region (20 kb)
Syn100 region (20 kb)
D
Yeast
Syn100
Linker DNA lenght (bp)
E
Syn100 (100 kb)
IV_left (100 kb)
IV_right (100 kb)
F
G1
G2/M
Contact frequency
Genomic distance
G
H
chr. IV-Syn100
Scc1-IP

Fig. 1. Syn100 is spontaneously chromatinized and organized in 3D in glucose.
A. Schematic representation of a random synthetic region (referred to as Syn100, in pink) inserted at position 845091 of chromosome IV using the CRISPR-Cas9 approach.
B. Nucleosome trace (in gray) and stranded RNA-seq profiles along 200 kb of the Chr.IV-Syn100 yeast chromosome. Pink and turquoise represent forward and reverse transcription, respectively. Data are from the strain RSGY1129.
C. Nucleosome trace (in gray) and stranded RNA-seq profiles along 200 kb of the Chr.IV-Syn100 yeast chromosome. Pink and turquoise represent forward and reverse transcription, respectively. Data are from the strain RSGY1129.
D. Frequency of nucleosome linker DNA length with a red line indicating 14 bp.
E. Autocorrelation of the nucleosome track showing the nucleosome repeat length (NRL) for 100 kb of the left and right sides of Chr.IV and Syn100.
F. Contact frequency (p) as a function of genomic distance (s) plots in G1 and G2/M for two 100 kb yeast regions (IV_left, IV_right) and the Syn100 region.
G. Top: 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G1. Bottom: Stranded RNA-seq profiles of the strain RSGY1129, asynchronous.
H. Top: 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G2/M. Bottom: Scc1 ChIP-seq deposition profile of the strain RSGY1129 synchronized in G2/M, and stranded RNA-seq profiles of the strain RSGY1129, asynchronous.

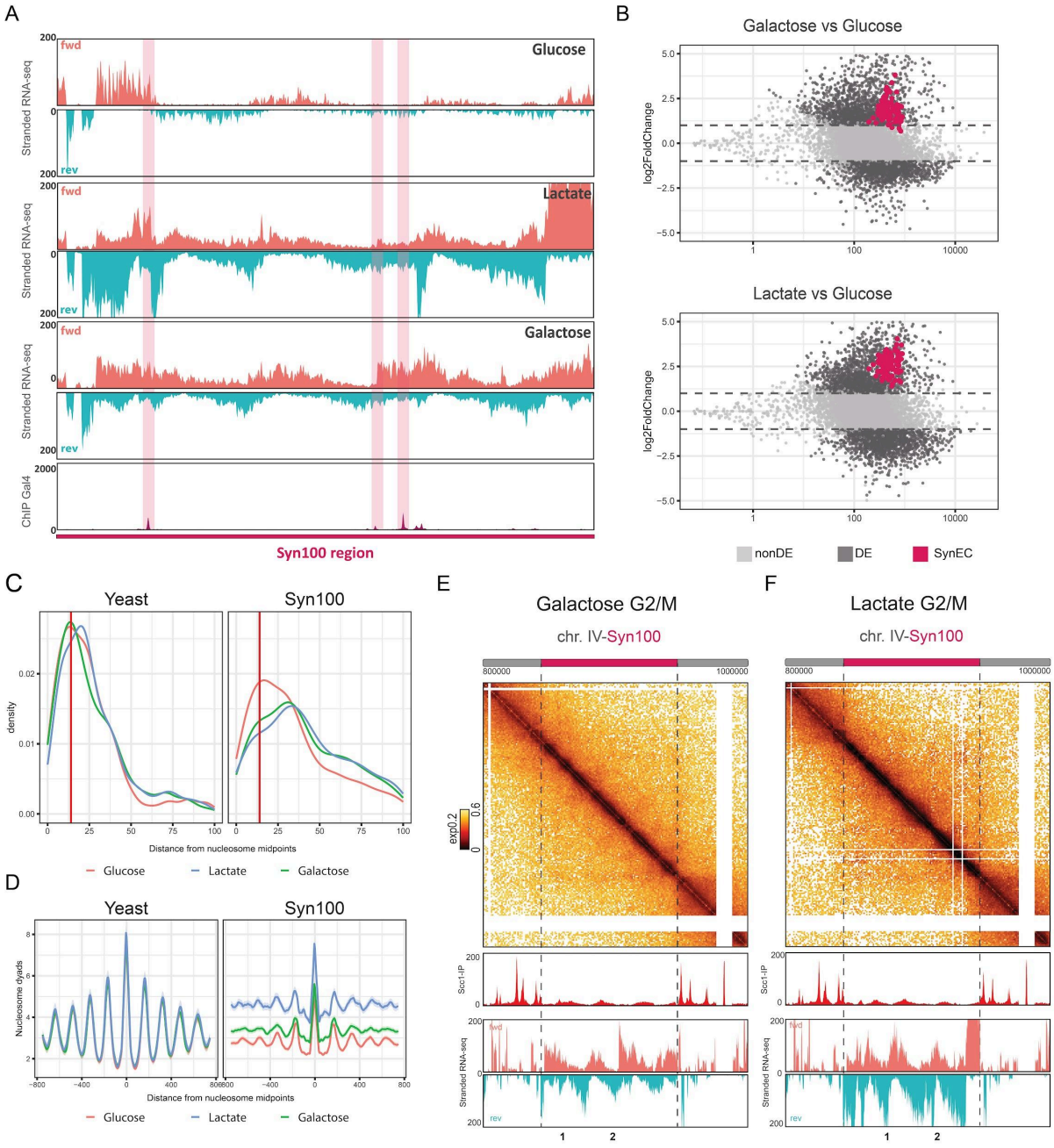

A
Glucose
Lactate
Galactose
fwd
rev
Stranded RNA-seq
ChIP Gal4
Syn100 region
B
Galactose vs Glucose
Lactate vs Glucose
log2FoldChange
nonDE
DE
SynEC
C
Yeast
Syn100
density
Distance from nucleosome midpoints
Glucose
Lactate
Galactose
D
Nucleosome dyads
E
Galactose G2/M
chr. IV-Syn100
Scc1-IP
F
Lactate G2/M

Fig. 2. : Syn100 is differentially transcribed, impacting its 3D organization under galactose and lactate conditions.
A. Stranded RNA-seq profiles along 200 kb of the Chr.IV-Syn100 yeast chromosome under glucose, lactate, and galactose conditions. Pink and turquoise represent forward and reverse transcription, respectively. Gal4 ChIP-seq deposition profile of the strain RSGY1129 in galactose is shown.
B. Differential analysis of expression between galactose/lactate and glucose conditions, showing yeast genomic regions (gray) and Syn100 region.
C. Frequency of nucleosome linker DNA length with a red line indicating 14 bp.
D.
E. Top: 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G2/M under galactose. Bottom: Scc1 ChIP-seq deposition profile of the strain RSGY1129 synchronized in G2/M, and stranded RNA-seq profiles of the strain RSGY1129, asynchronous.
F. Top: 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G2/M under lactate. Bottom: Scc1 ChIP-seq deposition profile of the strain RSGY1129 synchronized in G2/M, and stranded RNA-seq profiles of the strain RSGY1129, asynchronous.

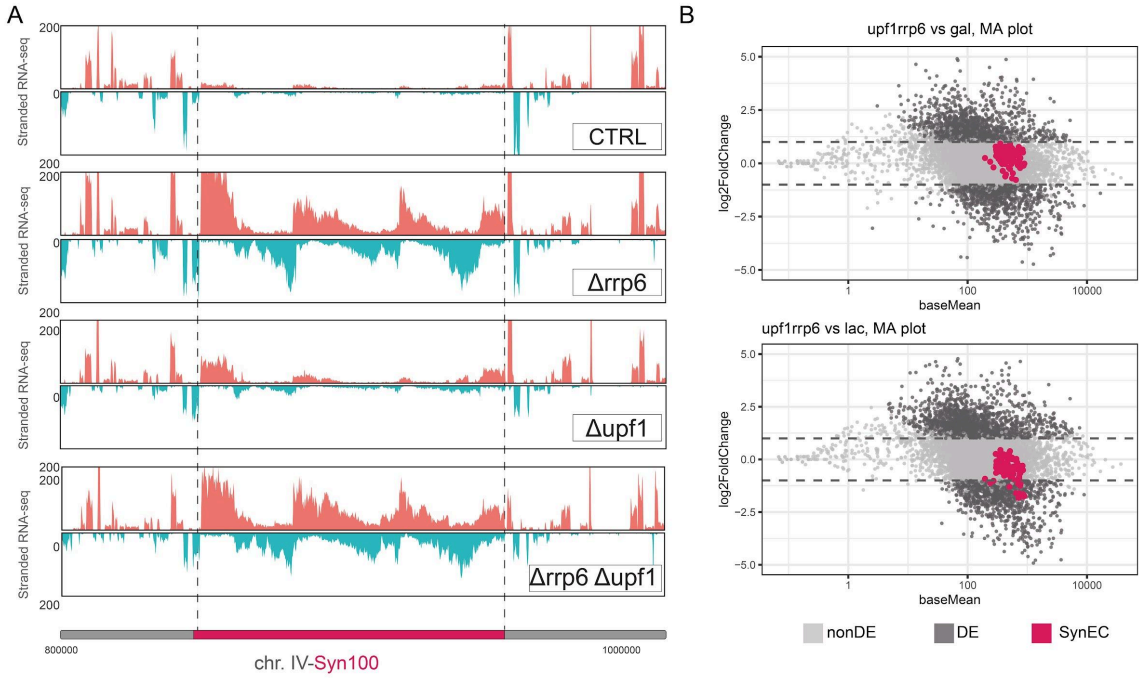

A
Stranded RNA-seq
200
0
CTRL
Δrrp6
Δupf1
Δrrp6 Δupf1
800000
1000000
chr. IV-Syn100
B
upf1rrp6 vs gal, MA plot
upf1rrp6 vs lac, MA plot
log2FoldChange
baseMean
nonDE
DE
SynEC

Figure 3. Syn100 transcripts are more sensitive to nuclear degradation pathways than the yeast genome.

A. Stranded RNA-seq profiles along 200 kb of the Chr.IV-Syn100 yeast chromosome in glucose for different strains: RSGY1129 (control), RSG_Y001353 (rrp6Δ), RSG_Y001271 (upf1Δ), and RSG_Y001299 (rrp6Δ/upf1Δ). Forward transcription is shown in pink, and reverse transcription is shown in turquoise.

B. Differential expression analysis of yeast genomic regions (gray) and the Syn100 region between the rrp6Δ/upf1Δ strain and control strain under glucose conditions.

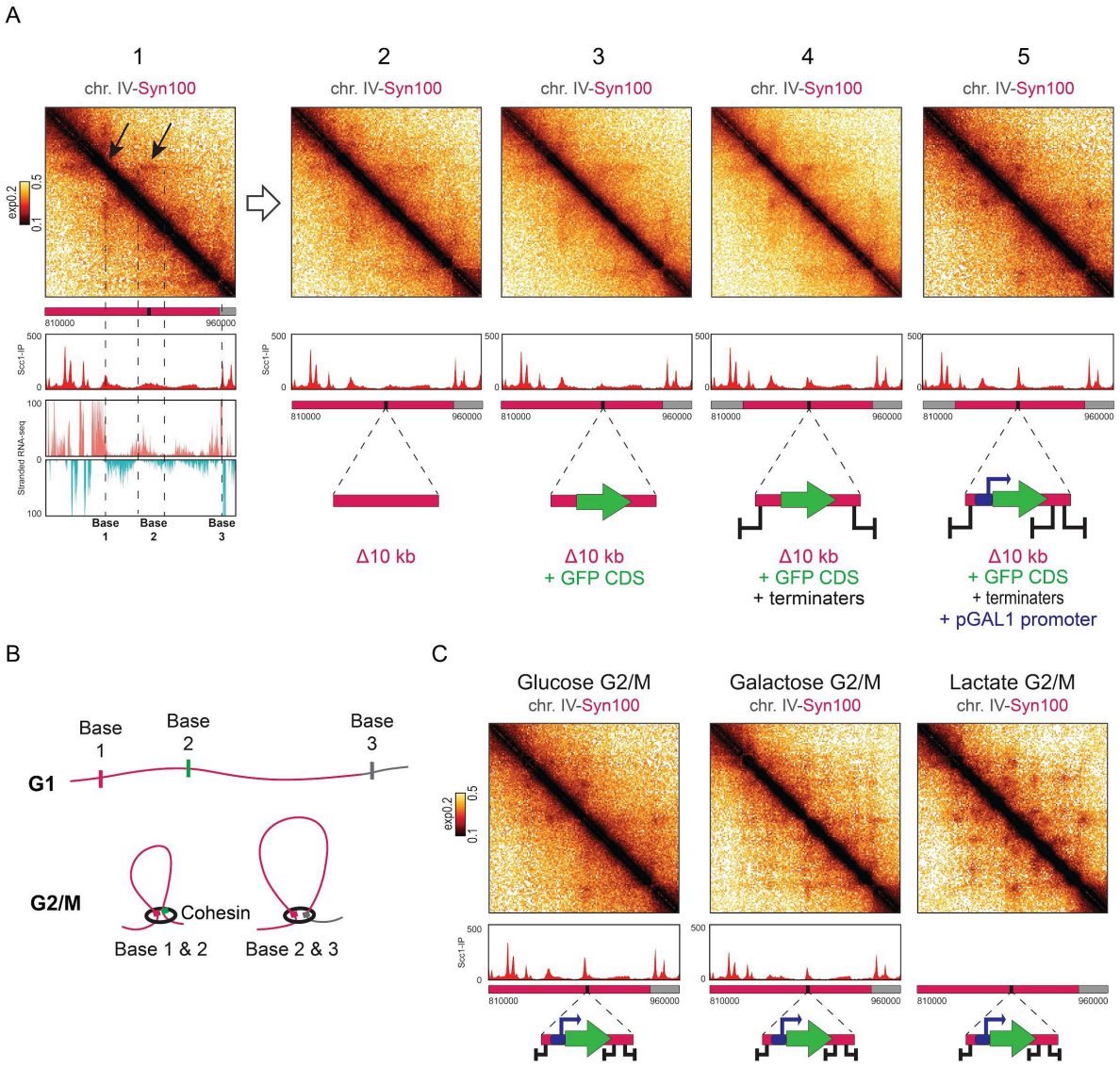

A
1
2
3
4
5
chr. IV-Syn100
exp0.2
0.5
0.1
810000
960000
Scc1-IP
500
0
Stranded RNA-seq
100
Base 1
Base 2
Base 3
Δ10 kb
Δ10 kb
+ GFP CDS
Δ10 kb
+ GFP CDS
+ terminaters
Δ10 kb
+ GFP CDS
+ terminaters
+ pGAL1 promoter
B
Base 1
Base 2
Base 3
G1
G2/M
Cohesin
Base 1 & 2
Base 2 & 3
C
Glucose G2/M
Galactose G2/M
Lactate G2/M
chr. IV-Syn100

Figure 4 : Genetic manipulation of Syn100 allows the formation of a reproducible cohesin-dependent loop.

A. Insertion of various constructs into the Syn100 region using CRISPR-Cas9. Top: 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 for different strains: RSGY1129 (control), RSG_Y001354 (Δ10kb), RSG_Y001359 (Δ10kb_GFP), RSG_Y001371 (Δ10kb_GFP_terminators), and RSG_Y001165 (Δ10kb_pGAL1-GFP-terminators) synchronized in G2/M. Bottom: Scc1 ChIP-seq deposition profile of the strain RSGY1129 synchronized in G2/M and schematic representation of each construct introduced into Syn100, showing the relative positions of engineered sites.

B. Schematic diagram illustrating the spatial organization of Syn100 and the engineered cohesin loops, highlighting the modified regions and their effect on loop formation.

C. 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 region in G2/M under galactose and lactate conditions for the strain Δ10kb_pGAL1-GFP. Bottom: Scc1 ChIP-seq deposition profile of the strain RSG_Y001165 synchronized in G2/M and schematic representation of each construct introduced into Syn100, showing the relative positions of engineered sites.

Supplementary figures

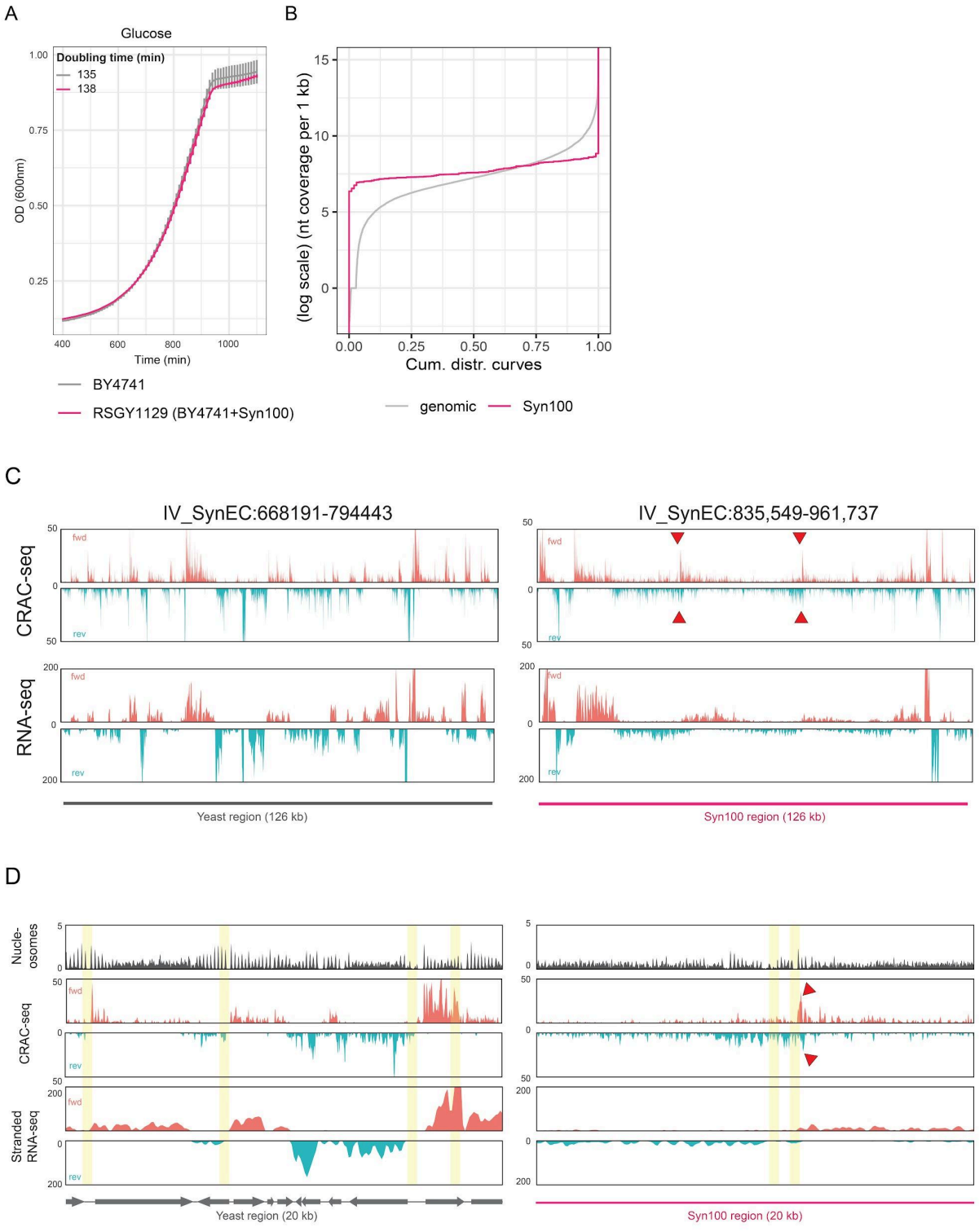


Fig. Sup. 1. Syn100 is spontaneously chromatinized and organized in 3D under glucose conditions.

A. Growth curves of the BY4741 (control) and RSGY1129 (Chr.IV-Syn100) strains in YPD medium. Two independent cultures were performed for each strain

B. Steady-state RNA levels expressed from genomic (gray) and Syn100 (pink) DNA, measured by nucleotide coverage per 1-kb window and sorted in ascending order.

C. Stranded CRAC-seq profiles along 126 kb of the IV_SynEC in glucose of the strain RSGY1129. Stranded RNA-seq profiles along 126 kb of the IV_SynEC in glucose of the strain RSGY1129. Pink and turquoise represent forward and reverse transcription, respectively.

D. Stranded CRAC-seq profiles along 20 kb of the IV_SynEC in glucose of the strain RSGY1129. Stranded RNA-seq profiles along IV_SynEC:820000-840000 and in glucose of the strain RSGY1129. Pink and turquoise represent forward and reverse transcription, respectively.

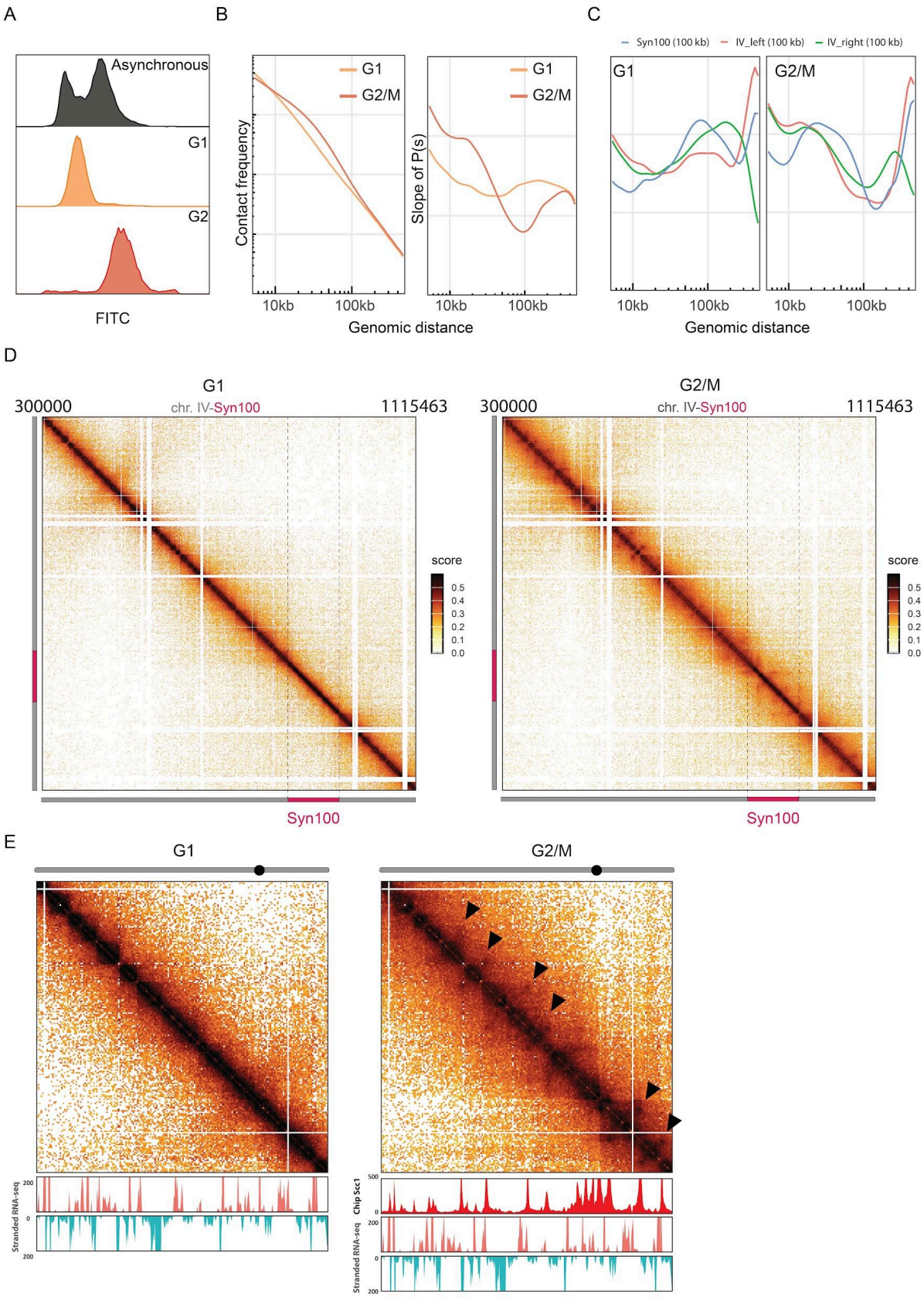

A
Asynchronous
G1
G2
FITC
B
G1
G2/M
Contact frequency
Slope of P(s)
10kb
100kb
Genomic distance
C
Syn100 (100 kb)
IV_left (100 kb)
IV_right (100 kb)
G1
G2/M
Genomic distance
D
G1
300000
chr. IV-Syn100
1115463
score
Syn100
G2/M
E
Stranded RNA-seq
Chip Scc1

Fig. Sup. 2. Synchronization and 3D organization of the strain RSGY1129 (Chr.IV-Syn100) in glucose.

A. Cell synchronization is monitored by flow cytometry in G1 and G2/M phases in glucose.

B. Contact frequency (p) as a function of genomic distance (s) plots of RSGY1129 chromosomes of G1 and G2/M, and their respective derivative curves.

C. Derivative curves of contact frequency (p) versus genomic distance (s) for two 100 kb yeast regions (IV_left, IV_right) and the Syn100 region in G1 and G2/M. The derivative curves provide a more detailed view of the compaction levels and changes in the 3D organization across different regions.

D. 700 kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G1 and G2/M, showing large-scale chromatin interactions and the overall 3D structure of the Syn100 region compared to neighboring regions.

E. Top: 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G1 and G2/M. Bottom: Scc1 ChIP-seq deposition profiles showing cohesin binding along the Syn100 region and stranded RNA-seq profiles of the strain RSGY1129, asynchronous.

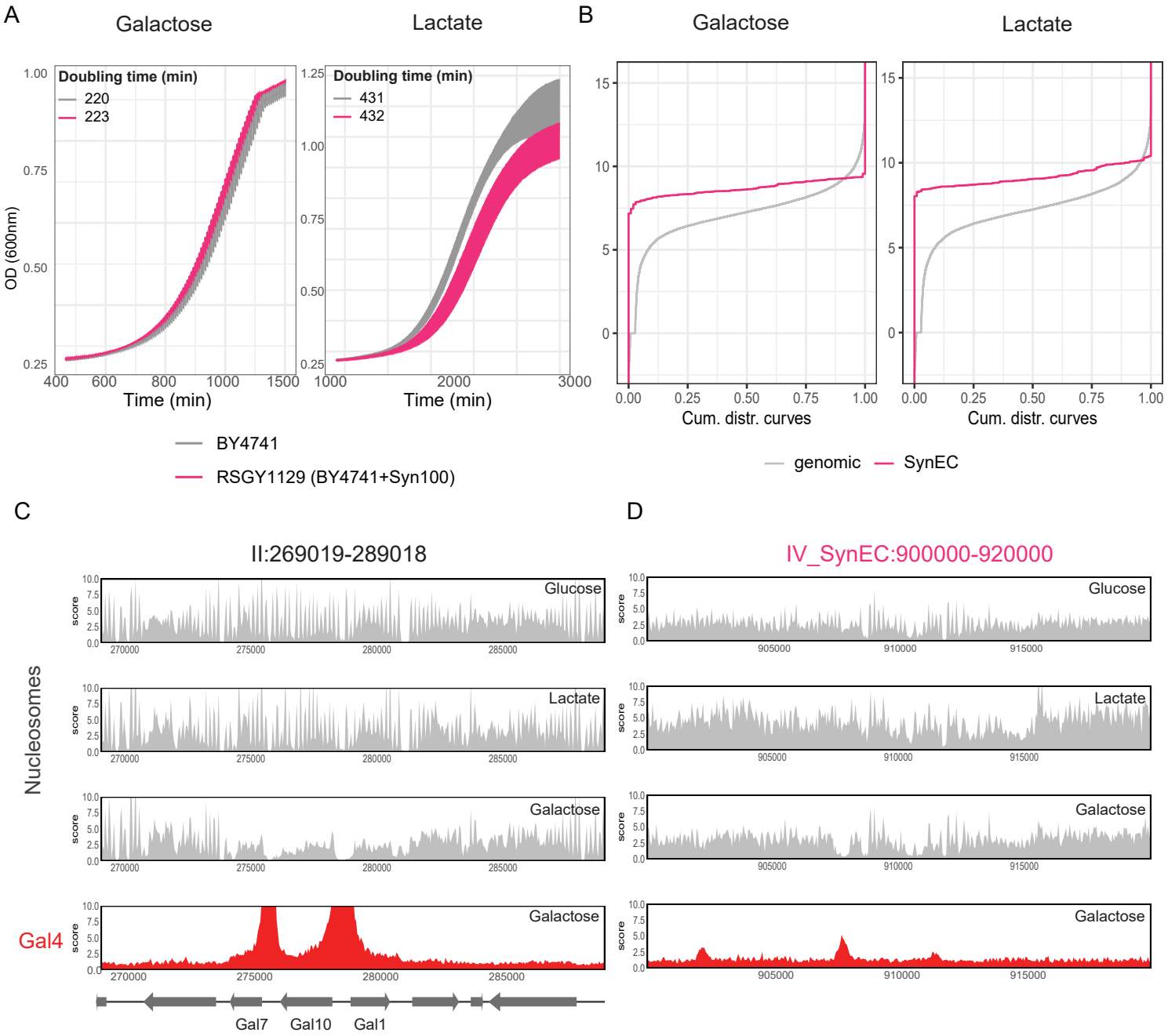

A
Galactose
Lactate
Doubling time (min)
220
223
431
432
OD (600nm)
Time (min)
BY4741
RSGY1129 (BY4741+Syn100)
B
Galactose
Lactate
Cum. distr. curves
genomic
SynEC
C
II:269019-289018
Nucleosomes
Glucose
Lactate
Galactose
Gal4
Gal7
Gal10
Gal1
D
IV_SynEC:900000-920000

Fig. Sup. 3. Nucleosome positioning and transcriptome profiles of Syn100 depend on carbon source conditions.

A. Growth curves of BY4741 (control strain) and RSGY1129 (Chr.IV-Syn100) strains in galactose and lactate media. Two independent cultures were performed for each strain.

B. Steady-state RNA levels expressed from genomic (gray) and Syn100 (pink) DNA, measured by nucleotide coverage per 1-kb window and sorted in ascending order.

C. Nucleosome trace (in gray) for a region of the yeast genome (II:269019-289018), illustrating the nucleosome arrangement of the GAL genes, under glucose, lactate and galactose conditions.

D. Nucleosome trace (in gray) along a 20 kb region of Chr.IV-Syn100 (IV_SynEC:900000-920000) under glucose, lactate and galactose conditions.

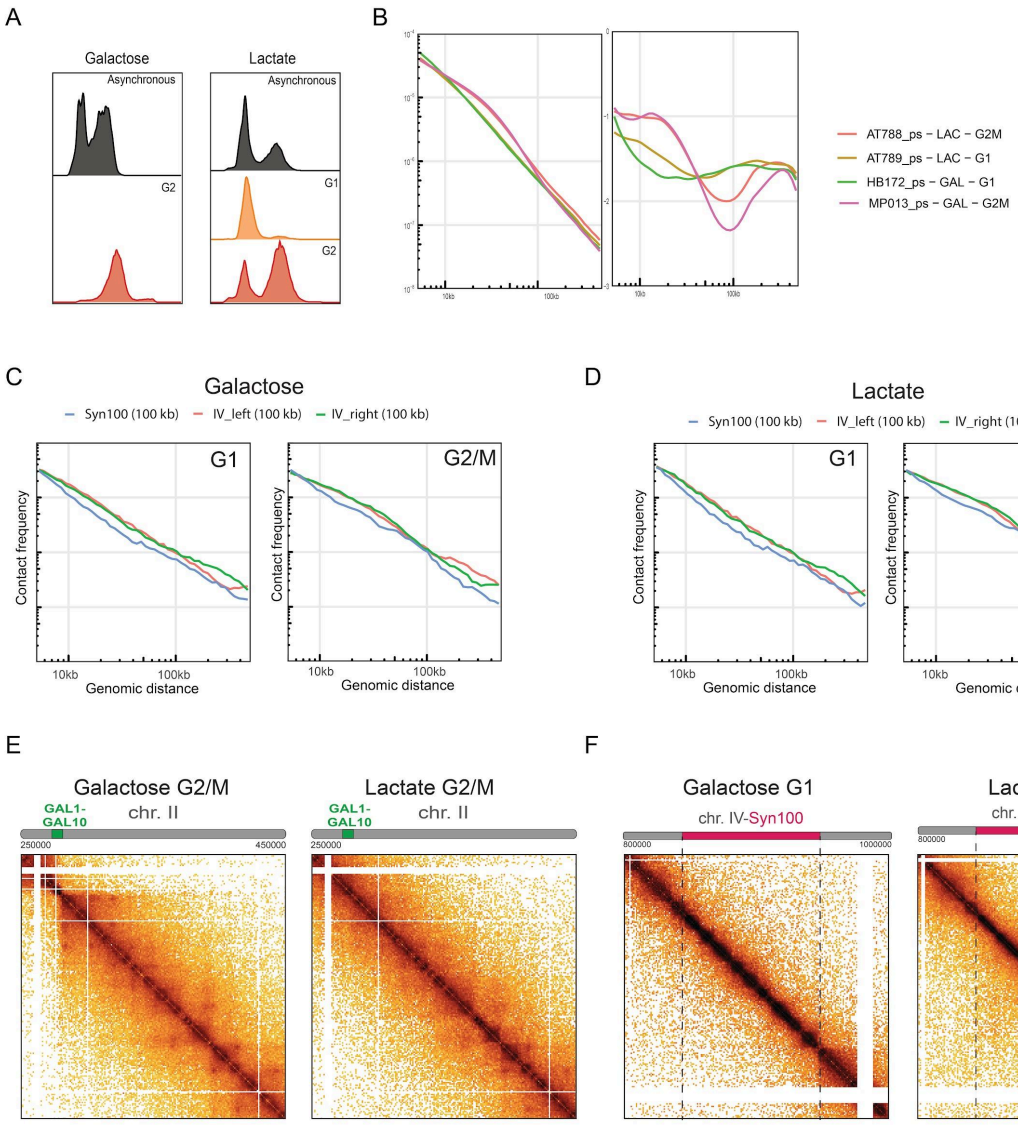

A
Galactose
Lactate
Asynchronous
G2
G1
B
AT788_ps – LAC – G2M
AT789_ps – LAC – G1
HB172_ps – GAL – G1
MP013_ps – GAL – G2M
C
Galactose
Syn100 (100 kb)
IV_left (100 kb)
IV_right (100 kb)
G1
G2/M
Contact frequency
10kb
100kb
Genomic distance
D
Lactate
E
Galactose G2/M
Lactate G2/M
GAL1-GAL10
chr. II
250000
450000
F
Galactose G1
Lactate G1
chr. IV-Syn100
800000
1000000

Fig. Sup. 4. Synchronization and 3D organization of the strain RSGY1129 (Chr.IV-Syn100) in lactate and galactose conditions.

A. Flow cytometry profiles showing the synchronization of the RSGY1129 strain in G1 and G2/M phases under lactate and galactose conditions.

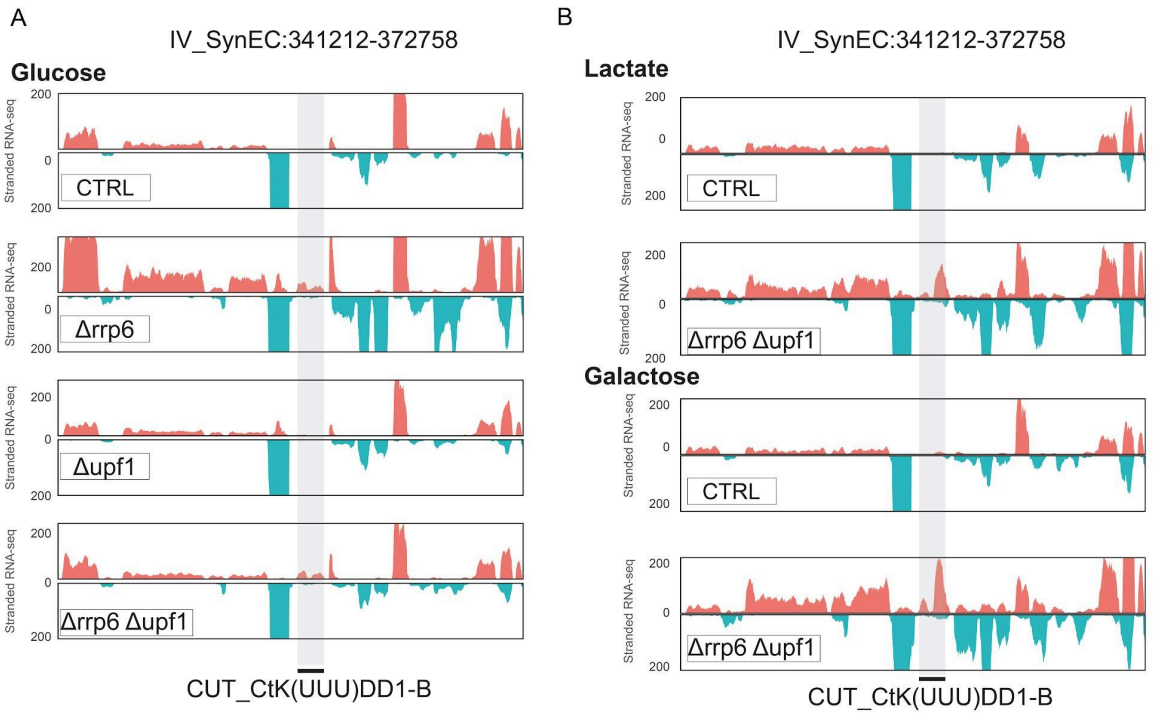
B. Contact frequency (p) as a function of genomic distance (s) plots for the chromosomes of the RSGY1129 strain in G1 and G2/M under lactate and galactose conditions, along with their respective derivative curves.

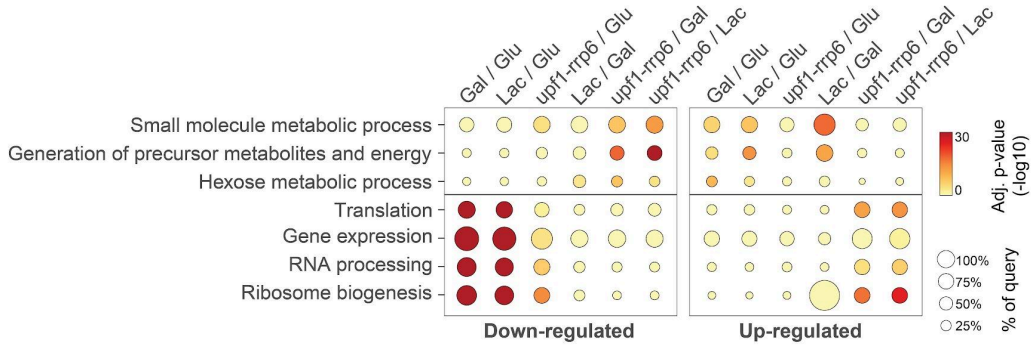
C. Contact frequency (p) as a function of genomic distance (s) for two 100 kb yeast regions (IV_left, IV_right) and the Syn100 region in G1 and G2/M under lactate conditions.

D. Contact frequency (p) as a function of genomic distance (s) for two 100 kb yeast regions (IV_left, IV_right) and the Syn100 region in G1 and G2/M under galactose conditions.

E. Top: 200 kb windows of the region of the chromosome II from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G1 and G2/M. Bottom: Scc1 ChIP-seq deposition profiles showing cohesin binding along the Syn100 region and stranded RNA-seq profiles of the strain RSGY1129, asynchronous.

F. Top: 200 kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1129 synchronized in G1 and G2/M. Bottom: Scc1 ChIP-seq deposition profiles showing cohesin binding along the Syn100 region and stranded RNA-seq profiles of the strain RSGY1129, asynchronous.


A
IV_SynEC:341212-372758
Glucose
Stranded RNA-seq
200
0
200
CTRL
Δrrp6
Δupf1
Δrrp6 Δupf1
CUT_CtK(UUU)DD1-B
B
IV_SynEC:341212-372758
Lactate
Galactose
CTRL
Δrrp6 Δupf1
CUT_CtK(UUU)DD1-B


C
Gal / Glu
Lac / Glu
upf1-rrp6 / Glu
Lac / Gal
upf1-rrp6 / Gal
upf1-rrp6 / Lac
Small molecule metabolic process
Generation of precursor metabolites and energy
Hexose metabolic process
Translation
Gene expression
RNA processing
Ribosome biogenesis
Down-regulated
Up-regulated
30
0
Adj. p-value (-log10)
100%
75%
50%
25%
% of query

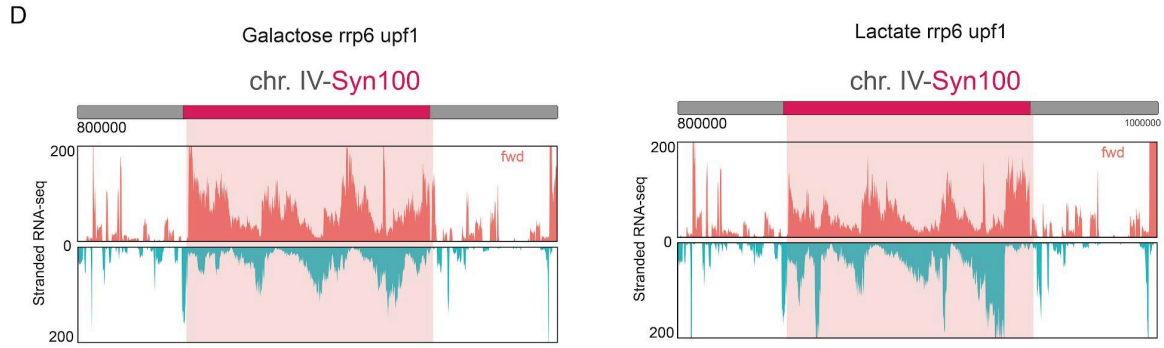
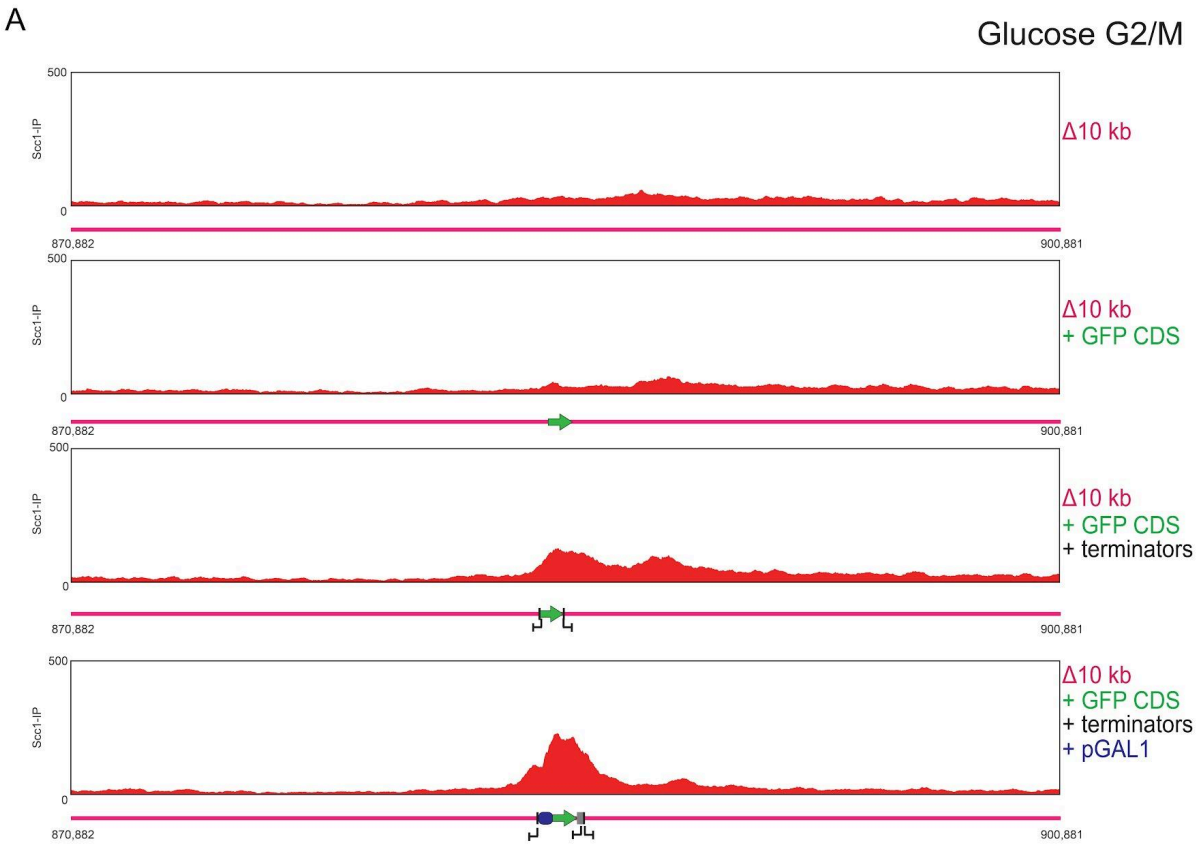
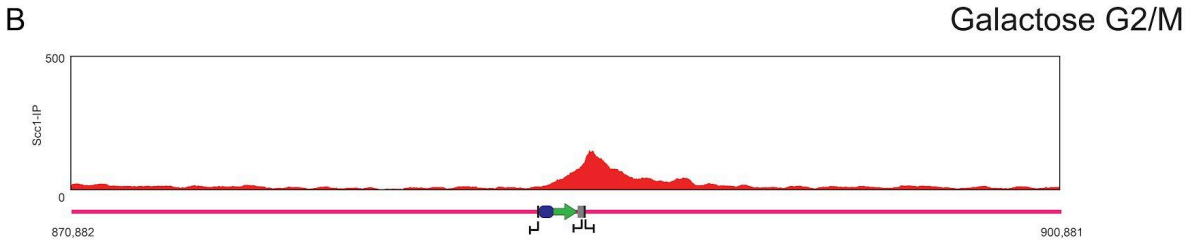

D
Galactose rrp6 upf1
chr. IV-Syn100
800000
1000000
200
0
fwd
Stranded RNA-seq
Lactate rrp6 upf1

Fig. Sup. 5. Impact of RNA degradation pathways on Syn100 transcriptional profiles.
A. Stranded RNA-seq profiles along IV_SynEC:341212-372758 in glucose for the strains: RSGY1129 (control strain), RSG_Y001353 (rrp6Δ), RSG_Y001271 (upf1Δ), and RSG_Y001299 (rrp6Δ/upf1Δ). Forward transcription is shown in pink, and reverse transcription is shown in turquoise.

B. Stranded RNA-seq profiles along IV_SynEC:341212-372758 in lactate and galactose conditions for the strains RSGY1129 (control strain) and RSG_Y001299 (rrp6Δ/upf1Δ).
C. Gene Ontology (GO) analysis of genes involved in RNA degradation pathways.
D. Stranded RNA-seq profiles along 200 kb of the Chr.IV-Syn100 yeast chromosome in glucose for the strains: RSGY1129 (control strain), RSG_Y001353 (rrp6Δ), RSG_Y001271 (upf1Δ), and RSG_Y001299 (rrp6Δ/upf1Δ).


A
Glucose G2/M
Scc1-IP
500
0
870,882
900,881
Δ10 kb
Δ10 kb
+ GFP CDS
Δ10 kb
+ GFP CDS
+ terminators
Δ10 kb
+ GFP CDS
+ terminators
+ pGAL1


B
Galactose G2/M
Scc1-IP
500
0
870,882
900,881

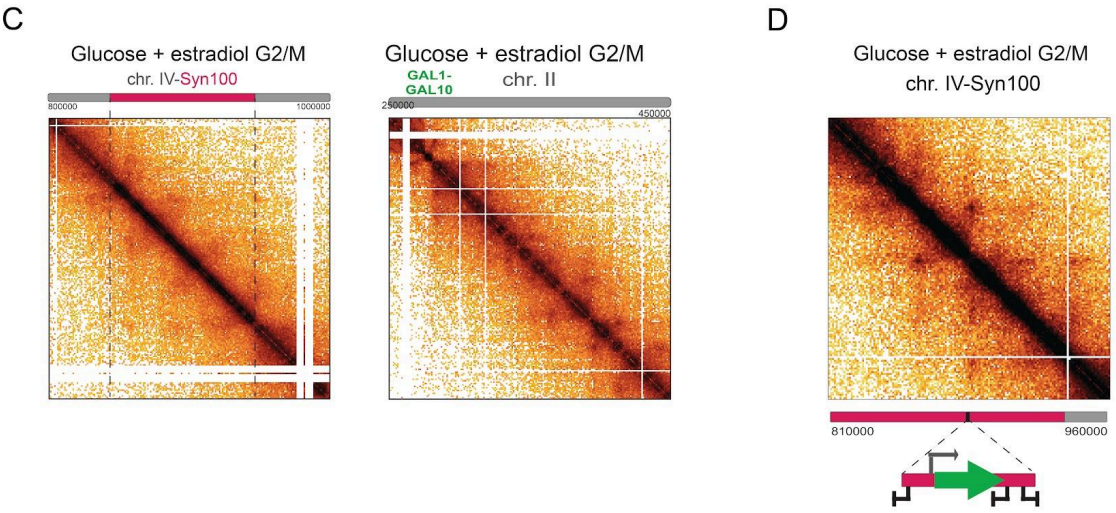

C
Glucose + estradiol G2/M
chr. IV-Syn100
800000
1000000
Glucose + estradiol G2/M
GAL1-
GAL10
chr. II
250000
450000
D
Glucose + estradiol G2/M
chr. IV-Syn100
810000
960000

Fig. Sup. 6. Genetic modifications on the 3D organization of Syn100 in glucose and the impact of carbon sources.

A. Scc1 ChIP-seq deposition profiles showing cohesin binding along the Syn100 region for different strains: RSGY1129 (control), RSG_Y001354 (Δ10kb), RSG_Y001359 (Δ10kb_GFP), RSG_Y001371 (Δ10kb_GFP_terminators), and RSG_Y001165 (Δ10kb_pGAL1-GFP-terminators) synchronized in G2/M, in glucose.

B. Scc1 ChIP-seq deposition profiles showing cohesin binding along the Syn100 region for RSG_Y001165 (Δ10kb_pGAL1-GFP-terminators) synchronized in G2/M, in galactose.

C. 200kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1288 (RSGY1129 with the cassette GAL4BDB-ER-Msn2-AD) synchronized in G2/M. Region ChrIV_Syn100:800000-1000000 and II:200000-400000.

D. 200kb windows from Hi-C contact maps of the Chr.IV-Syn100 of the strain RSGY1289(RSGY1165 with the cassette GAL4BDB-ER-Msn2-AD) synchronized in G2/M. Region ChrIV_Syn100:800000-1000000 and II:200000-400000.

Material and Methods

Strains and medium culture conditions

All yeast strains used in this study are derivatives of BY4741 and are listed in the Table “Strain list” (Table S1). The original strain carries the circular plasmid of Syn100. Yeast strains were grown overnight at 30°C in 150mL of suitable media to attain 4,2x108 cells. Cells were grown in a synthetic complete medium deprived of histidine (SC -His) (0.67% yeast nitrogen base without amino acids) (Difco), supplemented with a mix of amino-acids, 2% glucose) or in in rich medium (YPD): 1% bacto peptone (Difco), 1% bacto yeast extract (Difco), and 2% glucose. Yeast cells were synchronized in G1 by adding α-factor (Proteogenix, WY-13) in the media every 30 min during 2h30 (1 µg/mL final). To arrest cells in metaphase, cells were washed twice in fresh YPD after G1 arrest and released in rich medium (YPD) containing Nocodazole (Sigma-Aldrich, M1404-10MG) during 1h30. Cell synchronization was verified by Flow-cytometry.

CRISPR-Cas9 engineering

We used a CRISPR–Cas9 strategy to integrate the 100 kb of Syn100 into the chromosome IV of the strain BY4741. Nous nous sommes inspirés du protocole de Agier et al. The transformation utilized two plasmids: pAEF5, along with two gRNAs (gRNAs) and an additional pAEF5 modified with histidine selection. The strain was co-transformed with 500 ng of two targeting plasmids and 10 µL of DNA donor at a concentration of 100 µM. Successful integration was confirmed by colony PCR and Hi-C analysis. For further modifications, a single plasmid was used to target the specific position of the Syn100 region. The insertion of Syn100 was validated by PCR and Hi-C analysis. Details of all gRNA sequences, DNA donor sequences, and PCR primers used are provided in Table 1.

Bibliography

Agier, N., Fleiss, A., Delmas, S., & Fischer, G. (2021). A Versatile Protocol to Generate Translocations in Yeast Genomes Using CRISPR/Cas9. In W. Xiao (Ed.), *Yeast Protocols* (Vol. 2196, pp. 181–198). Springer US. https://doi.org/10.1007/978-1-0716-0868-5_14

Camellato, B., Brosh, R., Maurano, M. T., & Boeke, J. D. (2022). *Genomic analysis of a synthetic reversed sequence reveals default chromatin states in yeast and mammalian cells* (p. 2022.06.22.496726). bioRxiv. https://doi.org/10.1101/2022.06.22.496726

Chapard, C., Meneu, L., Serizay, J., Routhier, E., Ruault, M., Bignaud, A., Gourgues, G., Lartigue, C., Piazza, A., Taddei, A., Beckouët, F., Mozziconacci, J., & Koszul, R. (2023). *Exogenous chromosomes reveal how sequence composition drives chromatin assembly, activity, folding and compartmentalization* (p. 2022.12.21.520625). bioRxiv. https://doi.org/10.1101/2022.12.21.520625

Costantino, L., Hsieh, T.-H. S., Lamothe, R., Darzacq, X., & Koshland, D. (2020). Cohesin residency determines chromatin loop patterns. *eLife*, *9*, e59889. https://doi.org/10.7554/eLife.59889

Dauban, L., Montagne, R., Thierry, A., Lazar-Stefanita, L., Bastié, N., Gadal, O., Cournac, A., Koszul, R., & Beckouët, F. (2020). Regulation of Cohesin-Mediated Chromosome Folding by Eco1 and Other Partners. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2020.01.019

Gvozdenov, Z., Barcutean, Z., & Struhl, K. (2023). Functional analysis of a random-sequence chromosome reveals a high level and the molecular nature of transcriptional noise in yeast cells. *Molecular Cell*, *83*(11), 1786-1797.e5. https://doi.org/10.1016/j.molcel.2023.04.010

Lazar-Stefanita, L., Scolari, V. F., Mercy, G., Muller, H., Guérin, T. M., Thierry, A., Mozziconacci, J., & Koszul, R. (2017). Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *The EMBO Journal*, *36*(18), 2684–2697. https://doi.org/10.15252/embj.201797342

Lengronne, A., Katou, Y., Mori, S., Yokobayashi, S., Kelly, G. P., Itoh, T., Watanabe, Y., Shirahige, K., & Uhlmann, F. (2004). Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature*, *430*(6999), Article 6999. https://doi.org/10.1038/nature02742

Luthra, S., Kumar, A., Sharma, M., Arturo Garza-Reyes, J., & Kumar, V. (2022). An analysis of operational behavioural factors and circular economy practices in SMEs: An emerging economy perspective. *Journal of Business Research*, *141*, 321–336. https://doi.org/10.1016/j.jbusres.2021.12.014

Malabat, C., Feuerbach, F., Ma, L., Saveanu, C., & Jacquier, A. (2015). Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *eLife*, *4*, e06722. https://doi.org/10.7554/eLife.06722

Nevers, A., Doyen, A., Malabat, C., Néron, B., Kergrohen, T., Jacquier, A., & Badis, G. (2018). Antisense transcriptional interference mediates condition-specific gene repression in budding yeast. *Nucleic Acids Research*, *46*(12), 6009–6025. https://doi.org/10.1093/nar/gky342

Schalbetter, S. A., Goloborodko, A., Fudenberg, G., Belton, J.-M., Miles, C., Yu, M., Dekker, J., Mirny, L., & Baxter, J. (2017). SMC complexes differentially compact mitotic chromosomes according to genomic context. *Nature Cell Biology*, *19*(9), 1071–1080. https://doi.org/10.1038/ncb3594

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., & Steinmetz, L. M. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature*, *457*(7232), 1033–1037. https://doi.org/10.1038/nature07728

Zhou, J. (2022). Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nature Genetics*, 1–10. https://doi.org/10.1038/s41588-022-01065-4

2.1.2 Analyses et expériences prévues

Pour finaliser la rédaction de l'article, plusieurs analyses et expériences sont en cours pour approfondir la compréhension des régions synthétiques de la levure, notamment Syn100.

Expériences en Cours :

Un approfondissement de la caractérisation de la région Syn100 est en cours (identification des motifs, comparaison à une séquence “contrôle” de levure). Des expériences d'ATAC-seq sont également prévues pour compléter notre compréhension de l'accessibilité chromatinienne. Des expériences de CRAC-seq sont en cours dans des conditions de galactose et de lactate nous permettront de discriminer si l'augmentation du signal RNA-seq est due à une augmentation de la transcription ou à un ralentissement de la dégradation des transcrits.

Analyses en Cours :

Bien que cette analyse ne puisse pas être terminée avant le premier rendu du manuscrit, nous approfondissons les comparaisons entre les niveaux d’expressions de la région Syn100 avec les niveaux des différents types d’ARN (mRNA, CUTs, SUTs, XUTs...) pour évaluer leur expression différentielle en galactose, lactate, et glucose. Cela permettra de déterminer si ces ARN sont aussi affectés ou si l’augmentation des transcrits est spécifique à la séquence Syn100. Une analyse systématique des régions dépourvues de nucléosomes (NDR) est en cours, notamment pour rechercher les sites de démarrage de la transcription (TSS).

2.2. Résultats supplémentaires

2.2.1 Régulation longue distance dans la séquence aléatoire, Syn100

Les liens entre cohésine et transcription sont multiples. Contrairement aux métazoaires, le génome de *S. cerevisiae* est dense en gènes et pauvre en séquences non codantes. La régulation transcriptionnelle en *cis* est de courte portée (< 1 kb), avec de courtes séquences d'enhancers (UAS) situées près des promoteurs. Par ailleurs, les cohésines compactent le génome de *S. cerevisiae* durant la phase G2/M, contrairement aux métazoaires où les boucles cohésines dépendantes sont présentes en interphase. **Est-ce que l’organisation en *cis* du génome par les cohésines peut affecter la transcription chez *S. cerevisiae* ?** Cette relation causale est une question brûlante et difficile à résoudre chez les métazoaires où l’importance des cohésines dans la régulation de promoteurs par leurs « enhancers » distants de plusieurs dizaines de kb n’est pas claire.

Pour étudier cette question, nous avons modifié Syn100 pour qu'il présente des boucles discrètes médiées par la cohésine qui relient un UAS et son promoteur pGAL1 séparés par plusieurs dizaines de kb (**Figure 2. A**). Ce promoteur contrôle le gène rapporteur GFP. Nous avons synchronisé la souche en phase G2/M en glucose, et lorsque la boucle est formée entre l'UAS et le promoteur, nous avons induit l'activation de Gal4-ER en ajoutant l'oestradiol (Pincus et al., 2014) afin de tester la capacité de cette structure à promouvoir la transcription du gène rapporteur.

Deux heures après, nous avons réalisé le Hi-C, extrait les ARN pour faire de la RTqPCR ainsi que de la microscopie. Lorsque l'UAS est situé à proximité du promoteur, le gène rapporteur est transcrit, comme attendu (**Figure 2. CD**). Cependant, lorsque l'UAS est à distance en *cis* du promoteur, l'expression de GFP n'est pas détectée par RTqPCR **(Figure 2. C**) ni par microscopie (**Figure 2. D**) et ce malgré un enrichissement des contacts détectés entre l'UAS et le promoteur pGAL1. La boucle est préservée lorsque l'UAS est distant, les contacts P-E étant enrichis grâce à la cohésine (**Figure 2. B**).

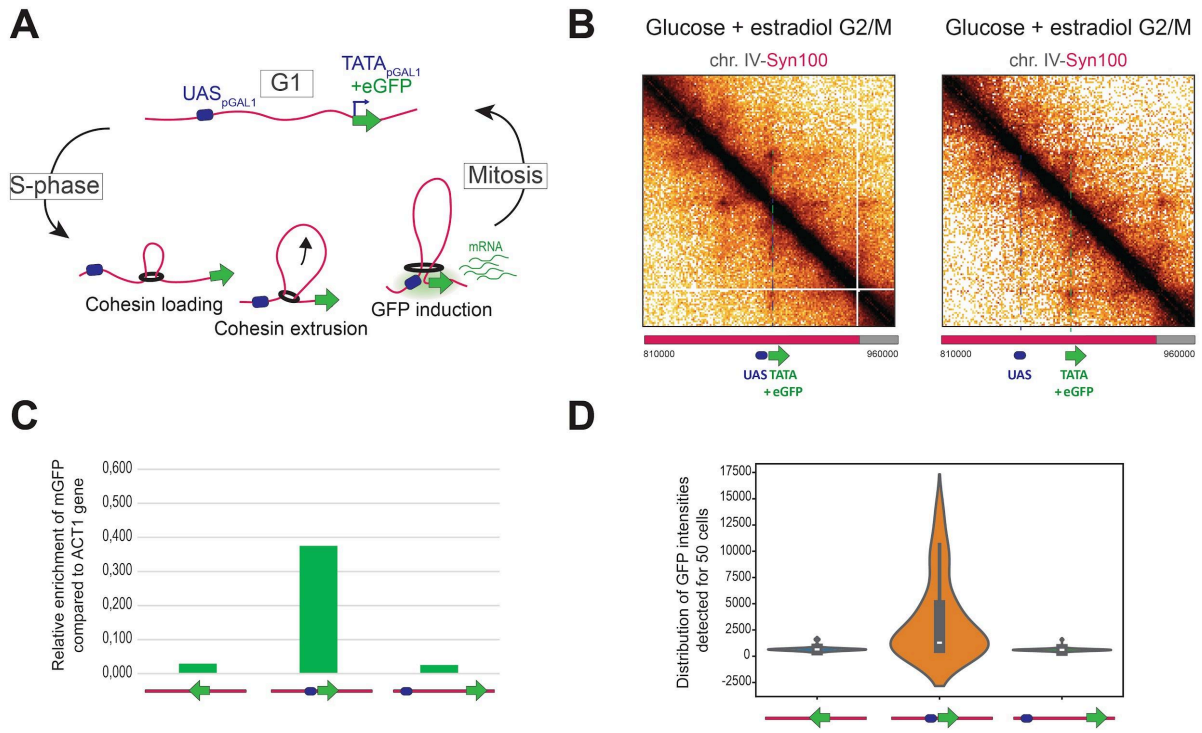


Figure 2 : Malgré le contact enrichi entre l'enhancer et le promoteur, aucune expression de GFP n'est détectée.

A) Schéma de la mise en place du système dans la région Syn100. **B)** Matrices des souches RSGY1287 et RSGY1289 synchronisées en G2/M suivies de l'ajout d'œstradiol (500 µM) pendant 2 heures. **C)** Mesure de l'expression de la GFP par RTqPCR, relative à l'expression du gène ACT1. A la suite : séquence codante de la GFP sans promoteur ; séquence codante de la GFP avec le promoteur pGAL1 adjacent ; séquence codante de la GFP avec le promoteur pGAL1 placé à 30 kb. **D)** Mesure de l'expression de la GFP par Microscopie. Les conditions sont similaires à C).

Puis, nous avons essayé de forcer les contacts entre le promoteur et l'UAS. Nous avons exploité un design pour forcer le contact P-E. Nous avons pris avantage du domaine POZ de la protéine GAGA chez la drosophile, impliquée dans la régulation de gènes notamment au cours du développement. Ce dernier est un domaine d'oligomérisation et de formation de boucle d'ADN chez la drosophile (Li et al., 2023). De plus, le recrutement d'un promoteur par GAGA permet une activation à longue distance chez la levure (Petrascheck et al., 2005). Nous avons construit des cassettes permettant l'expression de LacI-GAGApoz et inséré des sites LacO à côté de l'UAS et du promoteur (**Figure 3. A**).

Cependant, ce design ne conduit pas à un renforcement notable de l'interaction entre l'UAS et le promoteur détectée par Hi-C en G2/M (**Figure 3. C**). Ce pontage n'est pas non plus détecté en G1 par HiC (**Figure 3. C),** suggérant que, s'il se forme, il reste très transitoire. De manière cohérente, aucun signal de transcription du gène GFP n'est détecté lorsque LacI est exprimé et l'UAS est placé à distance du promoteur (**Figure 3. B**). En l'absence de production de GFP et surtout de confirmation du pontage médié par LacI-GAGApoz, nous ne sommes pas en mesure de conclure si un contact entre E-P, distant en *cis,* médié par la cohésine et stabilisé est impliqué dans la régulation transcriptionnelle.

Ces résultats préliminaires suggèrent que le simple pontage médié par la cohésine entre un promoteur et son enhancer distant n'est pas suffisant pour promouvoir la transcription, suggérant que cette régulation chez les métazoaires nécessite des facteurs supplémentaires pour établir l'interaction entre le promoteur et l'enhancer. La cohésine, par sa capacité à former des boucles dynamiques, permettrait la formation de ce contact entre E et P qui serait alors stabilisé par d'autres facteurs. Nous n'avons malheureusement pas pu confirmer cette hypothèse par la formation de pontage stable par le système LacI-GAGApoz. Ce système montre néanmoins que nous sommes désormais en mesure de concevoir une grande structure chromatinienne dans la levure bourgeonnante afin d'explorer les relations structure-fonction.

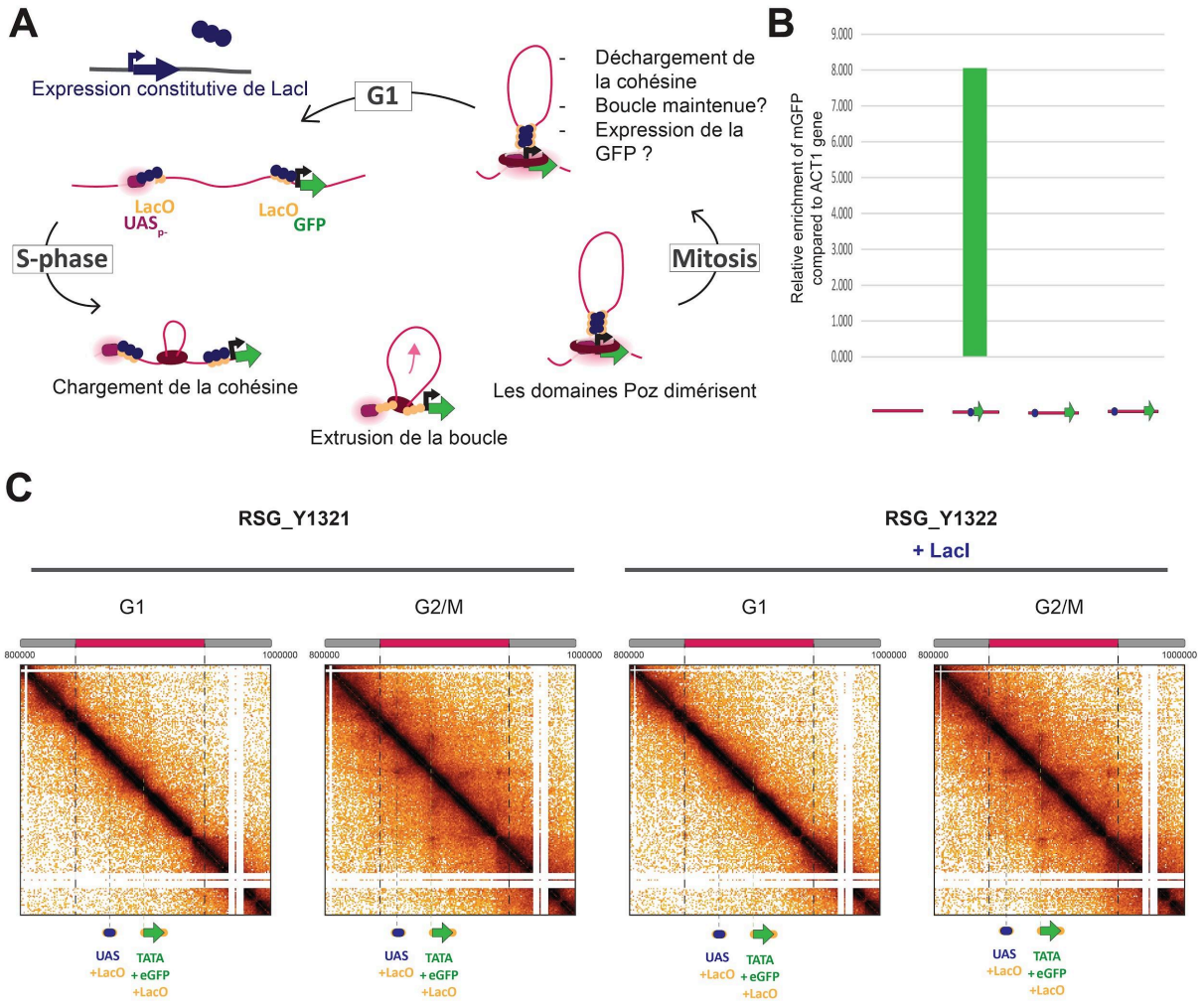


Figure 3 : Le système pour forcer le contact entre le promoteur et l'UAS ne permet pas de maintenir le contact P-E en G1.

A) Stratégie pour forcer le contact entre le P-E via l'expression constitutive de LacI-GAGA/POZ (en bleu). LacI-GAGA/POZ s'associe aux séquences LacO. Lorsque les sites sont en contact, les protéines LacI-GAGA/POZ dimérisent et maintiennent le contact. **B)** Mesure de l'expression de la GFP par RTqPCR, relative à l'expression d'ACT1. Les conditions évaluées sont : région vierge de Syn100; Syn100 avec la séquence codante de la GFP avec le promoteur pGAL1 adjacent (contrôle positif); séquence codante de la GFP avec le promoteur pGAL1 placé à 30 kb avec le système de dimérisation ; séquence codante de la GFP avec le promoteur pGAL1 placé à 30 kb sans le système de dimérisation. **C)** Fenêtres de 200 kb à partir des cartes de contact Hi-C des banques Hi-C des souches RSGY1321 et RSGY1322 synchronisées en phases G1 et G2/M avec ajout d'œstradiol (500 nM). La souche RSGY1321 n'exprime pas la protéine LacI, tandis que la souche RSGY1322 l'exprime.

Discussion

De nombreux travaux ces dernières années ont révélé plusieurs niveaux d'organisation 3D des génomes eucaryotes. Les génomes de différentes espèces sont ainsi organisés en boucles de chromatine dépendantes de complexes SMCs qui régulent et interagissent avec différents processus biologiques tels que la transcription, la réparation et la réplication. Les mécanismes moléculaires qui régulent cette organisation chromatinienne et les liens avec des fonctions de l'ADN sont en effet de mieux en mieux compris. Au cours de ma thèse, j'ai utilisé une approche un peu différente de ces études, en exploitant des séquences d'ADN "naïves", i.e. des molécules n'ayant pas subi de contraintes évolutives au sein de l'organisme hôte, au contraire de son génome natif. Et ce, pour voir si l'approche permet d'éclairer des processus biologiques sous un jour nouveau. Cette approche a déjà été utilisée au préalable, et par exemple tant chez la levure que chez l'homme, l'insertion de telles séquences naïves ont montré une activité transcriptionnelle. Cependant, ces travaux antérieurs se sont principalement focalisés sur la transcription et peu sur l'organisation spatiale du génome (Camellato et al., 2024; Gvozdenov et al., 2023; Luthra et al., 2024).

Mes travaux de thèse portent plus précisément sur l'exploitation de séquences exogènes et aléatoires où nous caractérisons leurs activités chromatiniennes : nucléosomes, transcription et finalement l'organisation spatiale du génome. Dans le premier projet, nous montrons que l'ADN génomiques de Mpneumo, 40%GC, adopte un état chromatinien de type Y (yeast) similaire à celui de la levure (38%GC), tandis que Mmyco, 24%GC, adopte un état chromatinien de type U (unorganized) qui est spatialement séparé du type Y dans le noyau. Ces deux types de chromatine pourraient être des archétypes de l'euchromatine et de l'hétérochromatine chez les métazoaires.

Dans le second projet, nous tirons parti d'une région aléatoire de 100 kb insérée dans un chromosome de levure. Cette région est chromatinisée et transcrite en accord avec des travaux antérieurs. Syn100 est replié par les cohésines en G2/M et ce repliement est dicté par la convergence transcriptionnelle, à l'instar du génome natif de *S. cerevisiae.* La caractérisation de Syn100 dans différentes sources de carbone révèle un niveau de régulation transcriptionnelle et un niveau post-transcriptionnel qui façonnent principalement le transcriptome de la séquence Syn100 et, de là, l'organisation 3D du génome de la levure. En tirant parti de notre connaissance des mécanismes régulant la formation de boucles par la cohésine via l'extrusion de boucles, nous avons modifié la séquence de telle sorte à affecter la terminaison de la transcription, afin de contrôler l'organisation 3D de Syn100 et former une boucle discrète et reproductible dépendante de la cohésine entre deux loci choisis. Enfin, nous avons essayé, pour l'instant sans succès, d'intégrer une régulation longue distance de type P-E au sein de Syn100, en exploitant la formation de la boucle de chromatine stable.

Ces deux projets illustrent le potentiel d'intégration de séquences exogènes et aléatoires dans un génome hôte, ouvrant de nouvelles perspectives en biologie génomique.

Les deux manuscrits ont déjà des discussions. Je développe ici quelques points supplémentaires à ceux des articles.

1. Pertinence des séquences exogènes dans l'organisme

1.1 Faire évoluer les séquences exogènes

L'intégration de séquences exogènes dans un génome soulève des questions immédiates quant à leur impact potentiel sur des processus essentiels liés à l'ADN, tels que la régulation des gènes, la réparation de l'ADN et l'organisation 3D. Ces séquences obéissent à de nouvelles règles sous lesquelles elles n'ont jamais évolué et qui les placent sous un jour différent. L'ensemble des résultats présentés dans ce manuscrit et les résultats d'autres équipes (Camellato et al., 2024; Chapard et al., 2023; Gvozdenov et al., 2023; Luthra et al., 2024) indiquent que les séquences naïves sont transcrites. Il est donc envisageable que des ORFs contenues dans ces transcrits soient traduites. Pour voir si c'est le cas, notamment concernant le génome de *Mycoplasma pneumoniae*, très transcrit, nous avons effectué une expérience de spectrométrie de masse. Nous avons détecté quelques peptides dans les deux souches (Résultats 1.2.1). Cela soulève plusieurs questions sur l'évolution des séquences introduites au cours du temps et sur leur maintien dans différents milieux.

Premièrement, il serait intéressant d'évaluer dans quelle mesure la levure va apprivoiser les fragments chromosomiques bactériens (Mmyco et Mpneumo) maintenus sous pression de sélection dans une culture continue en y introduisant des mutations au fil du temps, et lesquelles. Par exemple, on peut envisager que la stabilité des régions passera par une chromatinisation plus proche de celle de la levure, une répression de la transcription, etc… Nous avons fait deux expériences préliminaires en collaboration avec Jean-Baptiste Boulé (MNHN). La souche XVIfMmyco a été mise en culture dans un bioréacteur durant une centaine de générations. Pour le moment, les analyses Hi-C n'ont pas permis d'identifier de changement d'organisation ou de perte de séquences au niveau de la population, ni sur trois sous-clones. À court terme, il faudrait regarder plus finement la couverture pour identifier une perte d'une région de Mmyco mais à première vue, la couverture est identique et je n'ai pas identifié de perte de séquence. Une expérience est en cours où un mutagène, le MMS (méthanesulfonate de méthyle), est ajouté régulièrement de manière contrôlée dans la culture. Après séquençage, nous caractériserons les mutations des sous-clones isolés à partir des souches évoluées dans ces conditions de croissance, et nous comparerons les régions natives avec les régions naïves afin de déterminer les différences. Nous verrons si les chromosomes artificiels accumulent plus de

mutations que les régions natives, et si oui, s' il existe des biais dans leur nature.

Il serait par ailleurs intéressant de réduire la quantité d'un acide aminé pour forcer la pression de sélection. L'optimisation des bioréacteurs permettra de tester de plus en plus de conditions avec un contrôle précis des paramètres environnementaux sur de longues périodes (Bertaux et al., 2022).

La comparaison des deux chromosomes bactériens, avec des taux de GC différents, nous permettra d'étudier le comportement de ces deux séquences. Des analyses de protéomique et des expérimentations d'évolution dirigée pourraient révéler si ces séquences sont maintenues, éliminées ou modifiées sous pression sélective, ainsi que leur influence sur l'évolution du génome. Un résultat intéressant, par exemple, pourrait être soit l'utilisation de séquence bactérienne pour produire des protéines d'intérêt et/ou l'apparition de néogènes.

1.2 Augmenter la diversité des séquences

Dans le premier projet, nous avons mis en évidence des comportements très différents en fonction du taux de GC. ll nous a fallu appuyer ces résultats en utilisant des séquences exogènes d'autres organismes (*Plasmodium falciparum* d'origine eucaryote, *Phytoplasma vitis* d'origine procaryote) qui ont révélé une organisation similaire à celle de *Mycoides mycoides* (Papier ; Fig. Supp 7). Dans les projets futurs, il sera intéressant d'utiliser diverses séquences d'ADN d'autres organismes et/ou aléatoires. La levure reste l'outil idéal, car elle est facilement manipulable pour l'introduction et l'ingénierie de séquences, et nous disposons déjà de plusieurs librairies de souches contenant des séquences d'ADN exogènes.

La génomique synthétique, cherche donc de plus en plus à synthétiser de grandes régions de génomes. Cependant, il reste encore complexe de synthétiser de longues séquences. La première stratégie est de réutiliser des banques de YACs ou des BACs (Bacterial Artificial Chromosome) comme nous avons commencé à le faire. Nous avons, avec une étudiante en stage de M2, tenté d'intégrer un YAC fabriqué à partir d'un génome bactérien riche en GC (*Streptomyces ambofaciens*, 72% GC), mais malheureusement, l'expérience a échoué à chaque tentative jusqu'à présent. Il est donc possible que l'intégration GC riche dans le génome de la levure soit plus complexe, notamment en raison du manque de séquences réplicatives (comme les ARS, riches en AT). Noskov et al. ont ajouté des séquences réplicatives de levure dans la séquence de *Synechococcus elongatus* (55% G+C) et peuvent ainsi obtenir des YACs allant jusqu'à 454 kb (Noskov et al., 2012). Des outils sont en train d'être développés, comme MenDEL, pour trouver des BAC qui couvrent de longues régions d'intérêt et permettent de trier les résultats en fonction de plusieurs critères définis par l'utilisateur - longueur totale, nombre de BAC, BAC le plus long, etc (German et al., 2022). L'objectif ultime étant d'obtenir

un panel le plus complet de tous les taux de GC, de génomes eucaryotes et procaryotes, dans la levure et discriminer l'impact du taux de GC d'autres motifs.

Un des résultats importants de l'article "Myco" est l'utilisation de réseaux neuronaux (https://github.com/Alexwestbrook/bacterial_genome) capables de prédire la position de la polymérase, la cohésine et la MNase sur les séquences bactériennes, ensuite confirmés expérimentalement. Pour l'instant, les laboratoires ont soit utilisé des séquences d'autres organismes soit généré aléatoirement des portions d'ADN. Cependant, personne n'a synthétisé directement de longs fragments dessinés à l'aide d'un modèle prédictif concernant la composition de la chromatine. Ce projet nécessite une collaboration avec un laboratoire spécialisé en deep learning pour la génomique, comme par exemple celle que nous avons avec l'équipe de Julien Mozziconacci au Muséum d'Histoire Naturelle. Les coûts de synthèse de longs fragments d'ADN restent malheureusement élevés, mais nous avons tout de même initié ce projet au sein de l'équipe, grâce au travail de Manon Perrot, nouvelle doctorante.

Grâce à ces approches, la génération de nouveaux jeux de données permettra d'apprendre toujours davantage sur les règles et les relations de cause à effet entre la séquence et la composition, l'activité et l'organisation nucléaire de la chromatine sur des régions de grande taille. Cela nous permettra de tirer parti de ces règles pour générer des souches dont les structures chromosomiques sont modifiées, afin d'aborder des questions autrement difficiles à traiter dans un génome natif.

2. Régulation transcriptionnelle

2.1 Différence d'activité transcriptionnelle entre les séquences bactériennes et les séquences aléatoires

Des chercheurs ont intégré des brins d'ADN contenant des séquences aléatoires dans des cellules de levure (Camellato et al., 2024; Gvozdenov et al., 2023; Luthra et al., 2024) et de souris (Camellato et al., 2024) afin de déterminer l'état transcriptionnel par défaut de leur génome. Chez la levure, plusieurs séquences ont été introduites, comme chrXVII (18 kb, 50%GC), dChr (254 kb, 40%GC), les gènes HPRT1 et HPRT1R (100 kb, 41%GC), Mpneumo (0.8 Mb, 40%GC), Syn100 (100 kb, 50%GC). Pour toutes ces séquences, des produits d'ARN discrets et des signatures de chromatine active sont observés dans la levure. En revanche, la chromatine de Mmyco (1.2 Mb, 24%GC), est silencieuse.

Il est intéressant de comparer les différences transcriptionnelles entre ADN aléatoire et ADN bactérien (évolué). Les transcrits Mpneumo et Syn100 sont souvent plus longs que ceux de la levure, ce qui pourrait être dû, par exemple, à la présence d'ORF bactériennes évoluées, avec certains biais

de composition favorisant l'élongation ou limitant la terminaison, dans l'une des deux séquences. Par contre, Syn100 présente une transcription beaucoup plus bidirectionnelle avec des transcrits antisens qui chevauchent la transcription sens, une caractéristique qui n'est pas présente sur Mpneumo. Cela indique potentiellement une moindre régulation de la terminaison transcriptionnelle dans les séquences aléatoires de Syn100, comparé aux séquences de Mpneumo. Dans la séquence aléatoire, la régulation de la transcription est largement dépendante du système NNS (Nrd1-Nab3-Sen1) pour la terminaison, mais cette dépendance n'est pas observée dans les séquences bactériennes. La transcription bidirectionnelle (ou divergente) est un mécanisme conservé à travers les bactéries, les archées et les eucaryotes (Warman et al., 2021). Il existerait donc des éléments conservés dans Mpneumo qui permettraient un sens des transcrits unidirectionnel, contrairement à ce que l'on observe dans des séquences aléatoires comme Syn100. Cette différence met en lumière l'importance de certains éléments de régulation, intrinsèques à la séquence, conservés même chez les espèces dépourvues de nucléosomes, influençant l'orientation et l'expression des transcrits.

Pour poursuivre l'étude des transcrits, une approche comme le CRAC-seq sur d'autres sous-unités de la polymérase pourrait être envisagée. Bien que l'ATAC-seq fournisse des informations sur l'accessibilité de la chromatine, utiliser des techniques comme le CRAC-seq ou le NET-seq couplées avec le nanopore permettrait de mieux caractériser la longueur des transcrits et d'identifier les sites de départ de transcription. Par ailleurs, l'insertion de terminateurs ou la modification de motifs par CRISPR-Cas9 pourrait éclairer le rôle de différents éléments impliqués dans la régulation transcriptionnelle.

Les ADN exogènes (en l'occurrence bactériens) ne sont pas neutres d'un point de vue évolutif même quand ils sont introduits dans un contexte eucaryote. L'avantage de région comme Syn100, facilement éditable, en fait potentiellement une plateforme pour disséquer le comportement de transcrits issus d'une région naïve et neutre d'un point de vue évolutif. Les deux types de séquences, exogènes et aléatoires, apporteront donc des informations complémentaires. Les études utilisant ce type d'outils doivent donc être considérées en fonction de ces observations, et les conclusions tirées de l'une ou l'autre des approches doivent être généralisées avec attention.

2.2 Mettre en place une régulation longue distance chez *S. cerevisiae*

Pour réguler l'expression, les enhancers doivent se rapprocher de leur gène cible. Cependant, la relation entre la proximité tridimensionnelle et régulation de l'activité transcriptionnelle reste relativement floue (de Wit & Nora, 2023). L'objectif initial de Syn100 dans la levure était de se servir de cette région comme d'une matrice pour essayer de mettre en place une régulation longue distance, dépendante d'une boucle de cohésine. Nous avons montré que les changements de source de carbone

peuvent impacter l'organisation 3D. L'utilisation du système oestradiol nous permet de garder une boucle cohésine dépendante stable et reproductible tout en activant le promoteur pGAL1 en glucose (Pincus et al., 2014). Nous avons inséré dans Syn100 les différents éléments (UAS, core promoteur, séquence rapportrice) en variant les distances et avons confirmé le contact par Hi-C. Pourtant nous n'avons pas été capable de détecter de l'expression du gène rapporteur (Partie Résultats 2.2.1).

D'abord, nous sommes limités par la résolution (résolution habituelle en Hi-C = 1kb, UAS = 123 pdb, core promoteur 319 pdb) et il est nécessaire de se tourner vers l'approche Micro-C. Des résultats préliminaires encourageants, avec une résolution de 100 pdb, nous permettent de voir le contact entre le site 1 et le site 2 (**Figure 1**). Cette approche nous aidera à améliorer les futures constructions.

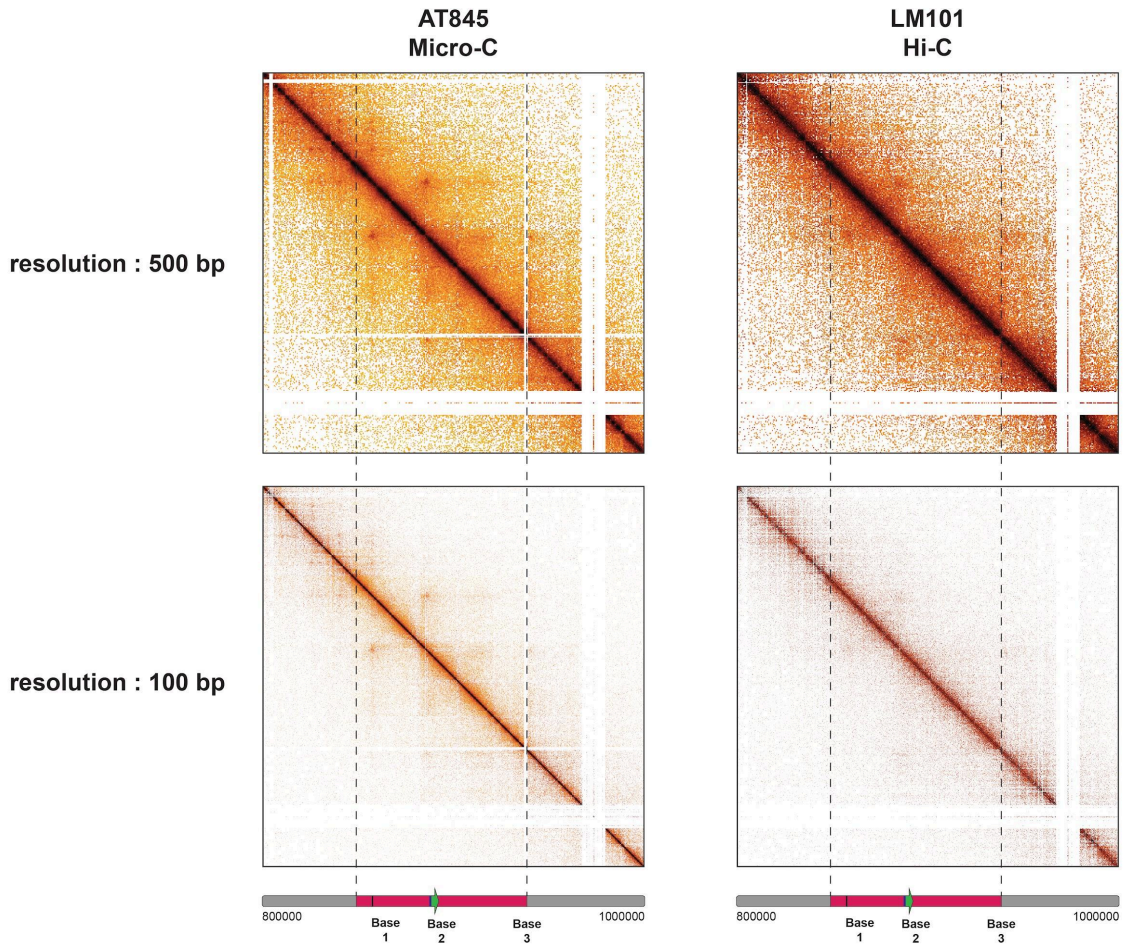


Figure 1 : Comparaison de fenêtres de 200kb provenant de la banque LM101 (souche RSGY1165, synchronisée en G2/M, Hi-C) et de la Banque AT845 (souche RSGY1165, synchronisée en G2/M, Micro-C). Résolutions : 500 pdb, 100 pdb.

Pour l'instant, les résultats de Hi-C pointent vers un contact physique entre UAS et promoteur. Alors pourquoi ne détectons nous pas le gène rapporteur ? Une première hypothèse à explorer serait le positionnement incorrect de l'UAS. Nous pourrions tester plusieurs positions de l'UAS et évaluer l'activité du rapporteur. Une seconde stratégie serait d'utiliser d'autres séquences régulatrices connues pour interagir à longue distance comme les séquences promotrices des protéines du choc thermique, les protéines HSP (Chowdhary et al., 2019; Kainth et al., 2021). Pour tester ces nombreuses hypothèses, nous mesurerons la GFP produite par plusieurs méthodes comme la microscopie, la cytométrie en flux, ou des lecteurs de plaques comme le TECAN permettant le multiplexage. Mais il serait aussi utile de remplacer la GFP par des rapporteurs mesurant la croissance d'une souche avec un marqueur auxotrophique (Dobi & Winston, 2007).

Les interactions entre les enhancers et les promoteurs sont dynamiques et peuvent nécessiter des mécanismes supplémentaires pour assurer leur stabilité. En effet, les contacts entre ces éléments régulateurs peuvent être trop transitoires, ce qui suggère la nécessité de stabiliser ces interactions pour une régulation génique efficace (Pollex et al., 2024). De plus, les séquences enhancers induisent une transcription bidirectionnelle (T.-K. Kim et al., 2010), ce qui peut influencer l'organisation du génome de manière significative. Ce phénomène montre l'importance de stabiliser adéquatement les boucles d'interaction entre les éléments régulateurs. Si la cohésine ne sert qu'à rapprocher ces éléments sans les stabiliser de manière prolongée, il pourrait être nécessaire d'envisager l'ajout d'autres composants ou même de forcer artificiellement le contact, par exemple avec des systèmes tels que LexA, que nous avons commencé à exploiter (**Résultats 2.2.1, Figure 2**). Une autre stratégie serait de stabiliser la cohésine en mutant des sous-unités de la cohésine, comme Wapl, permettant ainsi de maintenir ou d'augmenter le nombre de boucles d'interaction P-E.

Le laboratoire de Lu Bai a développé un système pour étudier les interactions fonctionnelles à longue distance chez la levure. En insérant un promoteur MET3 isolé flanqué de séquences invariables de 1 kb dans divers loci génomiques, ils ont pu identifier des positions génomiques qui induisent une expression du rapporteur différente. Avec la méthode 3C, ils ont confirmé que ces positions génomiques interagissent avec des régions parfois localisées sur d'autres chromosomes (Du et al., 2017).

3. La chromatine "AT-riche" de Mmyco

Le résultat le plus remarquable concerne l'organisation en globule du chromosome Mmyco et les interactions en *trans* quand ces régions sont placées sur le chromosome XVI et XIII. Le taux de GC apparaît comme le composant principal, ce qui est confirmé par les autres séquences riches en AT.

3.1 Tester la chromatine silencieuse

La chromatine de Myco rappelle l'hétérochromatine observée chez d'autres métazoaires, bien qu'elle ne soit pas dépendante des marques H3K9me3 ou HP1 connues chez les mammifères, ni des protéines SIR, typiquement présentes dans l'hétérochromatine constitutive de la levure. Quel est l'impact de l'U-chromatine, c'est-à-dire la chromatine de *Mycoplasma mycoides*, sur l'expression de gènes qui seraient inclus dans un tel domaine inactif ? En d'autres termes, est-ce que l'environnement 3D d'un gène actif va influencer son activité? Bien que *S. cerevisiae* ait un génome très compact, et qu'on puisse s'attendre à ce que cet effet soit limité, il est néanmoins intéressant de tester cette hypothèse. On peut ainsi imaginer intégrer des gènes sous contrôle de promoteurs de force variable pour tester leurs activités. Ce qui serait particulièrement intéressant serait l'intégration d'un gène rapporteur. Si ce dernier est réprimé dans la chromatine U de Mmyco par rapport à la chromatine Y de levure, on pourrait ensuite cribler de nombreux mutants et mesurer l'expression du rapporteur. Cependant, nous sommes limités par la complexité de la transformation dans une séquence riche en AT et il est essentiel de bien maîtriser l'édition de ce type de séquence, en augmentant les tailles d'homologies et en testant plusieurs sites de coupure par CRISPR-Cas9.

3.2 Identifier les mécanismes régulant la compartimentation chez Mmyco

Les compartiments chromosomiques ont été identifiés dans plusieurs clades, y compris chez les archées (Takemata et al. 2019, Cockram et al. 2021). Récemment, une étude sur le bombyx, *B. mori*, a révélé la formation de territoires chromosomiques (CTs) bien définis et la ségrégation spatiale des chromosomes en compartiments A actifs et B inactifs, similaire à celles observées chez d'autres eucaryotes. De manière inattendue, ils ont également découvert une nouvelle structure de compartiment qui ne présente pas d'interactions préférentielles avec les compartiments A et B. La caractérisation du nouveau compartiment révélée par l'exploitation du chromosome bactérien Mmyco représente une piste de recherche particulièrement intéressante pour comprendre les mécanismes impliqués dans cette compartimentation (Belmont, 2022; Mirny & Dekker, 2021).

Nous avons commencé à utiliser le système degron dans nos souches de levure en exploitant la dégradation des complexes SMCs. Ce système nous permet d'étudier les effets directs de la déplétion des acteurs principaux impliqués dans l'organisation du génome alors que leur délétion est létale. Nous n'avons pas encore étudié l'effet direct de la dégradation de la cohésine, la condensine ou de SMC5/6 dans ces souches. Bien que ce système soit fonctionnel dans une souche dépourvue de chromosome bactérien, nous rencontrons des difficultés à intégrer les constructions (Scc1-AID et SMC2-AID) dans la souche XVIfMmycot1. Remarquablement, le tag Scc1-AID impacte la croissance

de la souche XVIfMmyco1. Il me semble nécessaire d'utiliser des systèmes OsTIR1 encore plus stricts (Yesbolatova et al., 2020). En plus d'étudier la famille des SMCs, nous pourrions étendre l'utilisation de ce système à d'autres protéines impliquées dans l'organisation nucléaire ou la transcription.

La longueur inhabituelle de 10 pdb de la taille du linker entre les nucléosomes, mesurée dans le chromosome Mmyco (Article, Fig. 1D,E) suggère que l'Histone 1 (H1) pourrait lier les linkers d'ADN entre les nucléosomes consécutifs conduisant à la formation d'une chromatine de type U. La délétion de HHO1, l'équivalent de l'histone H1 chez les mammifères, n'impactent pas la compartimentation de Mmyco (Fig. 5A-D, S9A,B). Il pourrait être intéressant d'étudier la délétion de Hmo1, une autre protéine rappelant les caractéristiques de l'histone H1 (Panday & Grove, 2016).

Pour finir, nous aimerions aussi dissocier le taux de GC des autres caractéristiques identifiées dans les séquences de Myco. L'objectif serait d'explorer si un taux de GC similaire à celui de la levure, mais sans transcription, pourrait permettre de séparer les différents éléments responsables de la compartimentation chromosomique.

L'exploration de ces mécanismes apportera de nouvelles perspectives sur les processus biologiques associés à la formation des compartiments chez les métazoaires.

Bibliographie

Abdennur, N., & Mirny, L. A. (2020). Cooler: Scalable storage for Hi-C data and other genomically labeled arrays. *Bioinformatics (Oxford, England)*, *36*(1), 311–316. https://doi.org/10.1093/bioinformatics/btz540

Agmon, N., Temple, J., Tang, Z., Schraink, T., Baron, M., Chen, J., Mita, P., Martin, J. A., Tu, B. P., Yanai, I., Fenyö, D., & Boeke, J. D. (2020). Phylogenetic debugging of a complete human biosynthetic pathway transplanted into yeast. *Nucleic Acids Research*, *48*(1), 486–499. https://doi.org/10.1093/nar/gkz1098

Annaluru, N., Muller, H., Mitchell, L. A., Ramalingam, S., Stracquadanio, G., Richardson, S. M., Dymond, J. S., Kuang, Z., Scheifele, L. Z., Cooper, E. M., Cai, Y., Zeller, K., Agmon, N., Han, J. S., Hadjithomas, M., Tullman, J., Caravelli, K., Cirelli, K., Guo, Z., … Chandrasegaran, S. (2014). Total synthesis of a functional designer eukaryotic chromosome. *Science (New York, N.Y.)*, *344*(6179), 55–58. https://doi.org/10.1126/science.1249252

Aragón, L. (2018). The Smc5/6 Complex: New and Old Functions of the Enigmatic Long-Distance Relative. *Annual Review of Genetics*, *52*(Volume 52, 2018), 89–107. https://doi.org/10.1146/annurev-genet-120417-031353

Aragon, L., Martinez-Perez, E., & Merkenschlager, M. (2013). Condensin, cohesin and the control of chromatin states. *Current Opinion in Genetics & Development*, *23*(2), 204–211. https://doi.org/10.1016/j.gde.2012.11.004

Bai, L., & Morozov, A. V. (2010). Gene regulation by nucleosome positioning. *Trends in Genetics*, *26*(11), 476–483. https://doi.org/10.1016/j.tig.2010.08.003

Banigan, E. J., Tang, W., van den Berg, A. A., Stocsits, R. R., Wutz, G., Brandão, H. B., Busslinger, G. A., Peters, J.-M., & Mirny, L. A. (2023). Transcription shapes 3D chromatin organization by interacting with loop extrusion. *Proceedings of the National Academy of Sciences*, *120*(11), e2210480120. https://doi.org/10.1073/pnas.2210480120

Baptista, T., Grünberg, S., Minoungou, N., Koster, M. J. E., Timmers, H. T. M., Hahn, S., Devys, D., & Tora, L. (2017). SAGA Is a General Cofactor for RNA Polymerase II Transcription. *Molecular Cell*, *68*(1), 130-143.e5. https://doi.org/10.1016/j.molcel.2017.08.016

Barbieri, M., Chotalia, M., Fraser, J., Lavitas, L.-M., Dostie, J., Pombo, A., & Nicodemi, M. (2012). Complexity of chromatin folding is captured by the strings and binders switch model. *Proceedings of the National Academy of Sciences*, *109*(40), 16173–16178. https://doi.org/10.1073/pnas.1204799109

Basehoar, A. D., Zanton, S. J., & Pugh, B. F. (2004). Identification and Distinct Regulation of Yeast TATA Box-Containing Genes. *Cell*, *116*(5), 699–709. https://doi.org/10.1016/S0092-8674(04)00205-3

Bastié, N., Chapard, C., Dauban, L., Gadal, O., Beckouët, F., & Koszul, R. (2022). Smc3 acetylation, Pds5 and Scc2 control the translocase activity that establishes cohesin-dependent chromatin loops. *Nature Structural & Molecular Biology*, *29*(6), 575–585. https://doi.org/10.1038/s41594-022-00780-0

Batty, P., Langer, C. C., Takács, Z., Tang, W., Blaukopf, C., Peters, J., & Gerlich, D. W. (2023). Cohesin‐mediated DNA loop extrusion resolves sister chromatids in G2 phase. *The EMBO Journal*, *42*(16), e113475. https://doi.org/10.15252/embj.2023113475

Beckouët, F., Hu, B., Roig, M. B., Sutani, T., Komata, M., Uluocak, P., Katis, V. L., Shirahige, K., & Nasmyth, K. (2010). An Smc3 Acetylation Cycle Is Essential for Establishment of Sister Chromatid Cohesion. *Molecular Cell*, *39*(5), 689–699. https://doi.org/10.1016/j.molcel.2010.08.008

Belmont, A. S. (2022). Nuclear Compartments: An Incomplete Primer to Nuclear Compartments, Bodies, and Genome Organization Relative to Nuclear Architecture. *Cold Spring Harbor Perspectives in Biology*, *14*(7), a041268. https://doi.org/10.1101/cshperspect.a041268

Belton, J.-M., Lajoie, B. R., Audibert, S., Cantaloube, S., Lassadi, I., Goiffon, I., Baù, D.,

Marti-Renom, M. A., Bystricky, K., & Dekker, J. (2015). The Conformation of Yeast Chromosome III Is Mating Type Dependent and Controlled by the Recombination Enhancer. *Cell Reports*, *13*(9), 1855–1867. https://doi.org/10.1016/j.celrep.2015.10.063

Benabdallah, N. S., Williamson, I., Illingworth, R. S., Kane, L., Boyle, S., Sengupta, D., Grimes, G. R., Therizols, P., & Bickmore, W. A. (2019). Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation. *Molecular Cell*, *76*(3), 473-484.e7. https://doi.org/10.1016/j.molcel.2019.07.038

Berger, A. B., Cabal, G. G., Fabre, E., Duong, T., Buc, H., Nehrbass, U., Olivo-Marin, J.-C., Gadal, O., & Zimmer, C. (2008). High-resolution statistical mapping reveals gene territories in live yeast. *Nature Methods*, *5*(12), 1031–1037. https://doi.org/10.1038/nmeth.1266

Bernard, C., Locard-Paulet, M., Noël, C., Duchateau, M., Giai Gianetto, Q., Moumen, B., Rattei, T., Hechard, Y., Jensen, L. J., Matondo, M., & Samba-Louaka, A. (2022). A time-resolved multi-omics atlas of Acanthamoeba castellanii encystment. *Nature Communications*, *13*(1), 4104. https://doi.org/10.1038/s41467-022-31832-0

Bignaud, A., Cockram, C., Borde, C., Groseille, J., Allemand, E., Thierry, A., Marbouty, M., Mozziconacci, J., Espéli, O., & Koszul, R. (2024). Transcription-induced domains form the elementary constraining building blocks of bacterial chromosomes. *Nature Structural & Molecular Biology*, *31*(3), 489–497. https://doi.org/10.1038/s41594-023-01178-2

Bintu, B., Mateo, L. J., Su, J.-H., Sinnott-Armstrong, N. A., Parker, M., Kinrot, S., Yamaya, K., Boettiger, A. N., & Zhuang, X. (2018). Super-resolution chromatin tracing reveals domains and cooperative interactions in single cells. *Science*, *362*(6413), eaau1783. https://doi.org/10.1126/science.aau1783

Blackburn, E. H. (1985). Artificial chromosomes in yeast. *Trends in Genetics*, *1*, 8–12. https://doi.org/10.1016/0168-9525(85)90007-1

Boettiger, A., & Murphy, S. (2020). Advances in Chromatin Imaging at Kilobase-Scale Resolution. *Trends in Genetics*, *36*(4), 273–287. https://doi.org/10.1016/j.tig.2019.12.010

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. R., & Cremer, T. (2005). Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLOS Biology*, *3*(5), e157. https://doi.org/10.1371/journal.pbio.0030157

Botta, M., Haider, S., Leung, I. X. Y., Lio, P., & Mozziconacci, J. (2010). Intra‐ and inter‐chromosomal interactions correlate with CTCF binding genome wide. *Molecular Systems Biology*, *6*(1), 426. https://doi.org/10.1038/msb.2010.79

Brückner, D. B., Chen, H., Barinov, L., Zoller, B., & Gregor, T. (2023). Stochastic motion and transcriptional dynamics of pairs of distal DNA loci on a compacted chromosome. *Science*, *380*(6652), 1357–1362. https://doi.org/10.1126/science.adf5568

Buratowski, S. (2009). Progression through the RNA Polymerase II CTD Cycle. *Molecular Cell*, *36*(4), 541–546. https://doi.org/10.1016/j.molcel.2009.10.019

Busslinger, G. A., Stocsits, R. R., van der Lelij, P., Axelsson, E., Tedeschi, A., Galjart, N., & Peters, J.-M. (2017). Cohesin is positioned in mammalian genomes by transcription, CTCF and Wapl. *Nature*, *544*(7651), Article 7651. https://doi.org/10.1038/nature22063

Camellato, B. R., Brosh, R., Ashe, H. J., Maurano, M. T., & Boeke, J. D. (2024). Synthetic reversed sequences reveal default genomic states. *Nature*, 1–8. https://doi.org/10.1038/s41586-024-07128-2

Canzio, D., Nwakeze, C. L., Horta, A., Rajkumar, S. M., Coffey, E. L., Duffy, E. E., Duffié, R., Monahan, K., O'Keeffe, S., Simon, M. D., Lomvardas, S., & Maniatis, T. (2019). Antisense lncRNA Transcription Mediates DNA Demethylation to Drive Stochastic Protocadherin α Promoter Choice. *Cell*, *177*(3), 639-653.e15. https://doi.org/10.1016/j.cell.2019.03.008

Capelson, M., Liang, Y., Schulte, R., Mair, W., Wagner, U., & Hetzer, M. W. (2010). Chromatin-Bound Nuclear Pore Components Regulate Gene Expression in Higher Eukaryotes. *Cell*, *140*(3), 372–383. https://doi.org/10.1016/j.cell.2009.12.054

Cello, J., Paul, A. V., & Wimmer, E. (2002). Chemical Synthesis of Poliovirus cDNA: Generation of

Infectious Virus in the Absence of Natural Template. *Science*, *297*(5583), 1016–1018. https://doi.org/10.1126/science.1072266

Chan, L. Y., Kosuri, S., & Endy, D. (2005). Refactoring bacteriophage T7. *Molecular Systems Biology*, *1*(1), 2005.0018. https://doi.org/10.1038/msb4100025

Chowdhary, S., Kainth, A. S., Pincus, D., & Gross, D. S. (2019). Heat Shock Factor 1 Drives Intergenic Association of Its Target Gene Loci upon Heat Shock. *Cell Reports*, *26*(1), 18-28.e5. https://doi.org/10.1016/j.celrep.2018.12.034

Ciosk, R., Shirayama, M., Shevchenko, A., Tanaka, T., Toth, A., Shevchenko, A., & Nasmyth, K. (2000). Cohesin's Binding to Chromosomes Depends on a Separate Complex Consisting of Scc2 and Scc4 Proteins. *Molecular Cell*, *5*(2), 243–254. https://doi.org/10.1016/S1097-2765(00)80420-7

Corsi, F., Rusch, E., & Goloborodko, A. (2023). Loop extrusion rules: The next generation. *Current Opinion in Genetics & Development*, *81*, 102061. https://doi.org/10.1016/j.gde.2023.102061

Costantino, L., Hsieh, T.-H. S., Lamothe, R., Darzacq, X., & Koshland, D. (2020). Cohesin residency determines chromatin loop patterns. *eLife*, *9*, e59889. https://doi.org/10.7554/eLife.59889

Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R., & Mozziconacci, J. (2012). Normalization of a chromosomal contact map. *BMC Genomics*, *13*(1), 436. https://doi.org/10.1186/1471-2164-13-436

Cremer, T., & Cremer, C. (2001). Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nature Reviews Genetics*, *2*(4), 292–301. https://doi.org/10.1038/35066075

Cremer, T., & Cremer, M. (2010). Chromosome Territories. *Cold Spring Harbor Perspectives in Biology*, *2*(3), a003889. https://doi.org/10.1101/cshperspect.a003889

Cuartero, S., Weiss, F. D., Dharmalingam, G., Guo, Y., Ing-Simmons, E., Masella, S., Robles-Rebollo, I., Xiao, X., Wang, Y.-F., Barozzi, I., Djeghloul, D., Amano, M. T., Niskanen, H., Petretto, E., Dowell, R. D., Tachibana, K., Kaikkonen, M. U., Nasmyth, K. A., Lenhard, B., … Merkenschlager, M. (2018). Control of inducible gene expression links cohesin to hematopoietic progenitor self-renewal and differentiation. *Nature Immunology*, *19*(9), 932–941. https://doi.org/10.1038/s41590-018-0184-1

Cutter, A. R., & Hayes, J. J. (2015). A brief review of nucleosome structure. *FEBS Letters*, *589*(20PartA), 2914–2922. https://doi.org/10.1016/j.febslet.2015.05.016

Danino, T., Prindle, A., Kwong, G. A., Skalak, M., Li, H., Allen, K., Hasty, J., & Bhatia, S. N. (2015). Programmable probiotics for detection of cancer in urine. *Science Translational Medicine*, *7*(289), 289ra84. https://doi.org/10.1126/scitranslmed.aaa3519

Dauban, L., Montagne, R., Thierry, A., Lazar-Stefanita, L., Bastié, N., Gadal, O., Cournac, A., Koszul, R., & Beckouët, F. (2020). Regulation of Cohesin-Mediated Chromosome Folding by Eco1 and Other Partners. *Molecular Cell*. https://doi.org/10.1016/j.molcel.2020.01.019

David, L., Huber, W., Granovskaia, M., Toedling, J., Palm, C. J., Bofkin, L., Jones, T., Davis, R. W., & Steinmetz, L. M. (2006). A high-resolution map of transcription in the yeast genome. *Proceedings of the National Academy of Sciences*, *103*(14), 5320–5325. https://doi.org/10.1073/pnas.0601091103

Davidson, I. F., Bauer, B., Goetz, D., Tang, W., Wutz, G., & Peters, J.-M. (2019). DNA loop extrusion by human cohesin. *Science*, *366*(6471), 1338–1345. https://doi.org/10.1126/science.aaz3418

Davidson, I. F., & Peters, J.-M. (2021). Genome folding through loop extrusion by SMC complexes. *Nature Reviews Molecular Cell Biology*, *22*(7), 445–464. https://doi.org/10.1038/s41580-021-00349-7

Davies, J. O. J., Oudelaar, A. M., Higgs, D. R., & Hughes, J. R. (2017). How best to identify chromosomal interactions: A comparison of approaches. *Nature Methods*, *14*(2), 125–134. https://doi.org/10.1038/nmeth.4146

de Boer, C. G., Vaishnav, E. D., Sadeh, R., Abeyta, E. L., Friedman, N., & Regev, A. (2020). Deciphering eukaryotic gene-regulatory logic with 100 million random promoters. *Nature Biotechnology*, *38*(1), 56–65. https://doi.org/10.1038/s41587-019-0315-8

de Bruin, D., Zaman, Z., Liberatore, R. A., & Ptashne, M. (2001). Telomere looping permits gene

activation by a downstream UAS in yeast. *Nature*, *409*(6816), 109–113. https://doi.org/10.1038/35051119

de Wit, E., & Nora, E. P. (2023). New insights into genome folding by loop extrusion from inducible degron technologies. *Nature Reviews Genetics*, *24*(2), Article 2. https://doi.org/10.1038/s41576-022-00530-4

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing chromosome conformation. *Science (New York, N.Y.)*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

Despang, A., Schöpflin, R., Franke, M., Ali, S., Jerković, I., Paliou, C., Chan, W.-L., Timmermann, B., Wittler, L., Vingron, M., Mundlos, S., & Ibrahim, D. M. (2019). Functional dissection of the Sox9–Kcnj2 locus identifies nonessential and instructive roles of TAD architecture. *Nature Genetics*, *51*(8), 1263–1271. https://doi.org/10.1038/s41588-019-0466-z

Di Pierro, M., Zhang, B., Aiden, E. L., Wolynes, P. G., & Onuchic, J. N. (2016). Transferable model for chromosome architecture. *Proceedings of the National Academy of Sciences*, *113*(43), 12168–12173. https://doi.org/10.1073/pnas.1613607113

DiCarlo, J. E., Norville, J. E., Mali, P., Rios, X., Aach, J., & Church, G. M. (2013). Genome engineering in Saccharomyces cerevisiae using CRISPR-Cas systems. *Nucleic Acids Research*, *41*(7), 4336–4343. https://doi.org/10.1093/nar/gkt135

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. S., & Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), Article 7398. https://doi.org/10.1038/nature11082

Dixon, T. A., Walker, R. S. K., & Pretorius, I. S. (2023). Visioning synthetic futures for yeast research within the context of current global techno-political trends. *Yeast*, *40*(10), 443–456. https://doi.org/10.1002/yea.3897

Djebali, S., Davis, C. A., Merkel, A., Dobin, A., Lassmann, T., Mortazavi, A., Tanzer, A., Lagarde, J., Lin, W., Schlesinger, F., Xue, C., Marinov, G. K., Khatun, J., Williams, B. A., Zaleski, C., Rozowsky, J., Röder, M., Kokocinski, F., Abdelhamid, R. F., … Gingeras, T. R. (2012). Landscape of transcription in human cells. *Nature*, *489*(7414), Article 7414. https://doi.org/10.1038/nature11233

Dobi, K. C., & Winston, F. (2007). Analysis of Transcriptional Activation at a Distance in Saccharomyces cerevisiae. *Molecular and Cellular Biology*, *27*(15), 5575–5586. https://doi.org/10.1128/MCB.00459-07

Dostie, J., Richmond, T. A., Arnaout, R. A., Selzer, R. R., Lee, W. L., Honan, T. A., Rubio, E. D., Krumm, A., Lamb, J., Nusbaum, C., Green, R. D., & Dekker, J. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. *Genome Research*, *16*(10), 1299–1309. https://doi.org/10.1101/gr.5571506

Doudna, J. A., & Charpentier, E. (2014). The new frontier of genome engineering with CRISPR-Cas9. *Science*, *346*(6213). https://doi.org/10.1126/science.1258096

Downs, J. A., Kosmidou, E., Morgan, A., & Jackson, S. P. (2003). Suppression of Homologous Recombination by the Saccharomyces cerevisiae Linker Histone. *Molecular Cell*, *11*(6), 1685–1692. https://doi.org/10.1016/S1097-2765(03)00197-7

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y. J., Lee, C., Shendure, J., Fields, S., Blau, C. A., & Noble, W. S. (2010). A three-dimensional model of the yeast genome. *Nature*, *465*(7296), Article 7296. https://doi.org/10.1038/nature08973

Falk, M., Feodorova, Y., Naumova, N., Imakaev, M., Lajoie, B. R., Leonhardt, H., Joffe, B., Dekker, J., Fudenberg, G., Solovei, I., & Mirny, L. A. (2019). Heterochromatin drives compartmentalization of inverted and conventional nuclei. *Nature*, *570*(7761), Article 7761. https://doi.org/10.1038/s41586-019-1275-3

Filion, G. J., van Bemmel, J. G., Braunschweig, U., Talhout, W., Kind, J., Ward, L. D., Brugman, W., de Castro, I. J., Kerkhoven, R. M., Bussemaker, H. J., & van Steensel, B. (2010). Systematic protein location mapping reveals five principal chromatin types in Drosophila cells. *Cell*, *143*(2), 212–224. https://doi.org/10.1016/j.cell.2010.09.009

Finn, E. H., Pegoraro, G., Brandão, H. B., Valton, A.-L., Oomen, M. E., Dekker, J., Mirny, L., & Misteli, T. (2019). Extensive Heterogeneity and Intrinsic Variation in Spatial Genome Organization. *Cell*, *176*(6), 1502-1515.e10. https://doi.org/10.1016/j.cell.2019.01.020

Fleiss, A., O'Donnell, S., Fournier, T., Lu, W., Agier, N., Delmas, S., Schacherer, J., & Fischer, G. (2019). Reshuffling yeast chromosomes with CRISPR/Cas9. *PLOS Genetics*, *15*(8), e1008332. https://doi.org/10.1371/journal.pgen.1008332

Flemming, W. (1882). *Zellsubstanz, Kern und Zelltheilung*. Vogel.

Fredens, J., Wang, K., de la Torre, D., Funke, L. F. H., Robertson, W. E., Christova, Y., Chia, T., Schmied, W. H., Dunkelmann, D. L., Beránek, V., Uttamapinant, C., Llamazares, A. G., Elliott, T. S., & Chin, J. W. (2019). Total synthesis of Escherichia coli with a recoded genome. *Nature*, *569*(7757), 514–518. https://doi.org/10.1038/s41586-019-1192-5

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., & Mirny, L. A. (2016). Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, *15*(9), 2038–2049. https://doi.org/10.1016/j.celrep.2016.04.085

Fyodorov, D. V., Zhou, B.-R., Skoultchi, A. I., & Bai, Y. (2018). Emerging roles of linker histones in regulating chromatin structure and function. *Nature Reviews Molecular Cell Biology*, *19*(3), 192–206. https://doi.org/10.1038/nrm.2017.94

Gabriele, M., Brandão, H. B., Grosse-Holz, S., Jha, A., Dailey, G. M., Cattoglio, C., Hsieh, T.-H. S., Mirny, L., Zechner, C., & Hansen, A. S. (2022). Dynamics of CTCF- and cohesin-mediated chromatin looping revealed by live-cell imaging. *Science*, *376*(6592), 496–501. https://doi.org/10.1126/science.abn6583

Ganji, M., Shaltiel, I. A., Bisht, S., Kim, E., Kalichava, A., Haering, C. H., & Dekker, C. (2018). Real-time imaging of DNA loop extrusion by condensin. *Science*, *360*(6384), 102–105. https://doi.org/10.1126/science.aar7831

Garcia-Luis, J., Lazar-Stefanita, L., Gutierrez-Escribano, P., Thierry, A., Cournac, A., García, A., González, S., Sánchez, M., Jarmuz, A., Montoya, A., Dore, M., Kramer, H., Karimi, M. M., Antequera, F., Koszul, R., & Aragon, L. (2019). FACT mediates cohesin function on chromatin. *Nature Structural & Molecular Biology*, *26*(10), 970–979. https://doi.org/10.1038/s41594-019-0307-x

Gardner, M. J., Tettelin, H., Carucci, D. J., Cummings, L. M., Aravind, L., Koonin, E. V., Shallom, S., Mason, T., Yu, K., Fujii, C., Pederson, J., Shen, K., Jing, J., Aston, C., Lai, Z., Schwartz, D. C., Pertea, M., Salzberg, S., Zhou, L., … Hoffman, S. L. (1998). Chromosome 2 Sequence of the Human Malaria Parasite Plasmodium falciparum. *Science*, *282*(5391), 1126–1132. https://doi.org/10.1126/science.282.5391.1126

Gartenberg, M. R., & Smith, J. S. (2016). The Nuts and Bolts of Transcriptionally Silent Chromatin in Saccharomyces cerevisiae. *Genetics*, *203*(4), 1563–1599. https://doi.org/10.1534/genetics.112.145243

Gibcus, J. H., Samejima, K., Goloborodko, A., Samejima, I., Naumova, N., Nuebler, J., Kanemaki, M. T., Xie, L., Paulson, J. R., Earnshaw, W. C., Mirny, L. A., & Dekker, J. (2018). A pathway for mitotic chromosome formation. *Science*, *359*(6376), eaao6135. https://doi.org/10.1126/science.aao6135

Gibson, D. G., Benders, G. A., Axelrod, K. C., Zaveri, J., Algire, M. A., Moodie, M., Montague, M. G., Venter, J. C., Smith, H. O., & Hutchison, C. A. (2008). One-step assembly in yeast of 25 overlapping DNA fragments to form a complete synthetic Mycoplasma genitalium genome. *Proceedings of the National Academy of Sciences*, *105*(51), 20404–20409. https://doi.org/10.1073/pnas.0811011106

Gibson, D. G., Glass, J. I., Lartigue, C., Noskov, V. N., Chuang, R.-Y., Algire, M. A., Benders, G. A., Montague, M. G., Ma, L., Moodie, M. M., Merryman, C., Vashee, S., Krishnakumar, R., Assad-Garcia, N., Andrews-Pfannkoch, C., Denisova, E. A., Young, L., Qi, Z.-Q., Segall-Shapiro, T. H., … Venter, J. C. (2010). Creation of a Bacterial Cell Controlled by a Chemically Synthesized Genome. *Science*, *329*(5987), 52–56. https://doi.org/10.1126/science.1190719

Giorgetti, L., & Heard, E. (2016). Closing the loop: 3C versus DNA FISH. *Genome Biology*, *17*(1), 215. https://doi.org/10.1186/s13059-016-1081-2

Glynn, E. F., Megee, P. C., Yu, H.-G., Mistrot, C., Unal, E., Koshland, D. E., DeRisi, J. L., & Gerton, J. L. (2004). Genome-wide mapping of the cohesin complex in the yeast Saccharomyces cerevisiae. *PLoS Biology*, *2*(9), E259. https://doi.org/10.1371/journal.pbio.0020259

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., & Oliver, S. G. (1996). Life with 6000 Genes. *Science*, *274*(5287), 546–567. https://doi.org/10.1126/science.274.5287.546

Gotta, M., Laroche, T., Formenton, A., Maillet, L., Scherthan, H., & Gasser, S. M. (1996). The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type Saccharomyces cerevisiae. *Journal of Cell Biology*, *134*(6), 1349–1363. https://doi.org/10.1083/jcb.134.6.1349

Guacci, V., Koshland, D., & Strunnikov, A. (1997). A direct link between sister chromatid cohesion and chromosome condensation revealed through the analysis of MCD1 in S. cerevisiae. *Cell*, *91*(1), 47–57. https://doi.org/10.1016/s0092-8674(01)80008-8

Guérin, T. M., Barrington, C., Pobegalov, G., Molodtsov, M. I., & Uhlmann, F. (2023). *Cohesin chromatin loop formation by an extrinsic motor* (p. 2023.11.30.569410). bioRxiv. https://doi.org/10.1101/2023.11.30.569410

Guérin, T. M., Béneut, C., Barinova, N., López, V., Lazar-Stefanita, L., Deshayes, A., Thierry, A., Koszul, R., Dubrana, K., & Marcand, S. (2019). Condensin-Mediated Chromosome Folding and Internal Telomeres Drive Dicentric Severing by Cytokinesis. *Molecular Cell*, *75*(1), 131-144.e3. https://doi.org/10.1016/j.molcel.2019.05.021

Gvozdenov, Z., Barcutean, Z., & Struhl, K. (2023). Functional analysis of a random-sequence chromosome reveals a high level and the molecular nature of transcriptional noise in yeast cells. *Molecular Cell*, *83*(11), 1786-1797.e5. https://doi.org/10.1016/j.molcel.2023.04.010

Hampsey, M. (1998). Molecular Genetics of the RNA Polymerase II General Transcriptional Machinery. *Microbiology and Molecular Biology Reviews*, *62*(2), 465–503. https://doi.org/10.1128/mmbr.62.2.465-503.1998

Hanasaki, M., & Masumoto, H. (2019). CRISPR/Transposon gene integration (CRITGI) can manage gene expression in a retrotransposon-dependent manner. *Scientific Reports*, *9*(1), 15300. https://doi.org/10.1038/s41598-019-51891-6

Hassler, M., Shaltiel, I. A., & Haering, C. H. (2018). Towards a Unified Model of SMC Complex Function. *Current Biology: CB*, *28*(21), R1266–R1281. https://doi.org/10.1016/j.cub.2018.08.034

Hauer, M. H., & Gasser, S. M. (2017). Chromatin and nucleosome dynamics in DNA damage and repair. *Genes & Development*, *31*(22), 2204–2221. https://doi.org/10.1101/gad.307702.117

Hediger, F., & Gasser, S. M. (2006). Heterochromatin protein 1: Don't judge the book by its cover! *Current Opinion in Genetics & Development*, *16*(2), 143–150. https://doi.org/10.1016/j.gde.2006.02.013

Heyneker, H. L., Shine, J., Goodman, H. M., Boyer, H. W., Rosenberg, J., Dickerson, R. E., Narang, S. A., Itakura, K., Lin, S., & Riggs, A. D. (1976). Synthetic lacoperator DNA is functional in vivo. *Nature*, *263*(5580), 748–752. https://doi.org/10.1038/263748a0

Hildebrand, E. M., & Dekker, J. (2020). Mechanisms and Functions of Chromosome Compartmentalization. *Trends in Biochemical Sciences*, *45*(5), 385–396. https://doi.org/10.1016/j.tibs.2020.01.002

Hirano, T., Kobayashi, R., & Hirano, M. (1997). Condensins, chromosome condensation protein complexes containing XCAP-C, XCAP-E and a Xenopus homolog of the Drosophila Barren protein. *Cell*, *89*(4), 511–521. https://doi.org/10.1016/s0092-8674(00)80233-0

Hirano, T., & Mitchison, T. J. (1994). A heterodimeric coiled-coil protein required for mitotic chromosome condensation in vitro. *Cell*, *79*(3), 449–458. https://doi.org/10.1016/0092-8674(94)90254-2

Hoencamp, C., Dudchenko, O., Elbatsh, A. M. O., Brahmachari, S., Raaijmakers, J. A., van Schaik, T., Sedeño Cacciatore, Á., Contessoto, V. G., van Heesbeen, R. G. H. P., van den Broek, B., Mhaskar, A. N., Teunissen, H., St Hilaire, B. G., Weisz, D., Omer, A. D., Pham, M., Colaric, Z., Yang, Z., Rao, S. S. P., … Rowland, B. D. (2021). 3D genomics across the tree of life reveals condensin II as a determinant of architecture type. *Science*, *372*(6545), 984–989. https://doi.org/10.1126/science.abe2218

Holmquist, G. P. (1989). Evolution of chromosome bands: Molecular ecology of noncoding DNA. *Journal of Molecular Evolution*, *28*(6), 469–486. https://doi.org/10.1007/BF02602928

Hoose, A., Vellacott, R., Storch, M., Freemont, P. S., & Ryadnov, M. G. (2023). DNA synthesis technologies to close the gene writing gap. *Nature Reviews Chemistry*, *7*(3), 144–161. https://doi.org/10.1038/s41570-022-00456-9

Hsieh, T.-H. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Darzacq, X., & Tjian, R. (2022). Enhancer–promoter interactions and transcription are largely maintained upon acute loss of CTCF, cohesin, WAPL or YY1. *Nature Genetics*, *54*(12), Article 12. https://doi.org/10.1038/s41588-022-01223-8

Hsieh, T.-H. S., Fudenberg, G., Goloborodko, A., & Rando, O. J. (2016). Micro-C XL: Assaying chromosome conformation from the nucleosome to the entire genome. *Nature Methods*, *13*(12), 1009–1011. https://doi.org/10.1038/nmeth.4025

Hsieh, T.-H. S., Weiner, A., Lajoie, B., Dekker, J., Friedman, N., & Rando, O. J. (2015). Mapping Nucleosome Resolution Chromosome Folding in Yeast by Micro-C. *Cell*, *162*(1), 108–119. https://doi.org/10.1016/j.cell.2015.05.048

Hsieh, Y.-Y. P., Makrantoni, V., Robertson, D., Marston, A. L., & Murray, A. W. (2020). Evolutionary repair: Changes in multiple functional modules allow meiotic cohesin to support mitosis. *PLOS Biology*, *18*(3), e3000635. https://doi.org/10.1371/journal.pbio.3000635

Hua, S., Qiu, M., Chan, E., Zhu, L., & Luo, Y. (1997). Minimum Length of Sequence Homology Required for*in Vivo*Cloning by Homologous Recombination in Yeast. *Plasmid*, *38*(2), 91–96. https://doi.org/10.1006/plas.1997.1305

Hudson, T. J., Stein, L. D., Gerety, S. S., Ma, J., Castle, A. B., Silva, J., Slonim, D. K., Baptista, R., Kruglyak, L., Xu, S.-H., Hu, X., Colbert, A. M. E., Rosenberg, C., Reeve-Daly, M. P., Rozen, S., Hui, L., Wu, X., Vestergaard, C., Wilson, K. M., … Lander, E. S. (1995). An STS-Based Map of the Human Genome. *Science*, *270*(5244), 1945–1954. https://doi.org/10.1126/science.270.5244.1945

Hughes, R. A., & Ellington, A. D. (2017). Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *Cold Spring Harbor Perspectives in Biology*, *9*(1), a023812. https://doi.org/10.1101/cshperspect.a023812

Hutchison, C. A., Chuang, R.-Y., Noskov, V. N., Assad-Garcia, N., Deerinck, T. J., Ellisman, M. H., Gill, J., Kannan, K., Karas, B. J., Ma, L., Pelletier, J. F., Qi, Z.-Q., Richter, R. A., Strychalski, E. A., Sun, L., Suzuki, Y., Tsvetanova, B., Wise, K. S., Smith, H. O., … Venter, J. C. (2016). Design and synthesis of a minimal bacterial genome. *Science*, *351*(6280), aad6253. https://doi.org/10.1126/science.aad6253

Itakura, K., Hirose, T., Crea, R., Riggs, A. D., Heyneker, H. L., Bolivar, F., & Boyer, H. W. (1977). Expression in Escherichia coli of a Chemically Synthesized Gene for the Hormone Somatostatin. *Science*, *198*(4321), 1056–1063. https://doi.org/10.1126/science.412251

Jacquier, A. (2009). The complex eukaryotic transcriptome: Unexpected pervasive transcription and novel small RNAs. *Nature Reviews Genetics*, *10*(12), Article 12. https://doi.org/10.1038/nrg2683

Jäger, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H. E., Heindl, A., Whiffin, N., Carnicer, M. J., Broome, L., Dryden, N., Nagano, T., Schoenfelder, S., Enge, M., Yuan, Y., Taipale, J., Fraser, P., Fletcher, O., & Houlston, R. S. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, *6*(1), 6178. https://doi.org/10.1038/ncomms7178

Jeppsson, K., Pradhan, B., Sutani, T., Sakata, T., Umeda Igarashi, M., Berta, D. G., Kanno, T., Nakato,

R., Shirahige, K., Kim, E., & Björkegren, C. (2024). Loop-extruding Smc5/6 organizes transcription-induced positive DNA supercoils. *Molecular Cell*, *84*(5), 867-882.e5. https://doi.org/10.1016/j.molcel.2024.01.005

Jeppsson, K., Sakata, T., Nakato, R., Milanova, S., Shirahige, K., & Björkegren, C. (2022). Cohesin-dependent chromosome loop extrusion is limited by transcription and stalled replication forks. *Science Advances*, *8*(23), eabn7063. https://doi.org/10.1126/sciadv.abn7063

Jerkovic´, I., & Cavalli, G. (2021). Understanding 3D genome organization by multidisciplinary methods. *Nature Reviews Molecular Cell Biology*, *22*(8), 511–528. https://doi.org/10.1038/s41580-021-00362-w

Jiang, C., & Pugh, B. F. (2009). Nucleosome positioning and gene regulation: Advances through genomics. *Nature Reviews Genetics*, *10*(3), 161–172. https://doi.org/10.1038/nrg2522

Jin, Y., Eser, U., Struhl, K., & Churchman, L. S. (2017). The Ground State and Evolution of Promoter Region Directionality. *Cell*, *170*(5), 889-898.e10. https://doi.org/10.1016/j.cell.2017.07.006

Kagey, M. H., Newman, J. J., Bilodeau, S., Zhan, Y., Orlando, D. A., van Berkum, N. L., Ebmeier, C. C., Goossens, J., Rahl, P. B., Levine, S. S., Taatjes, D. J., Dekker, J., & Young, R. A. (2010). Mediator and cohesin connect gene expression and chromatin architecture. *Nature*, *467*(7314), Article 7314. https://doi.org/10.1038/nature09380

Kane, L., Williamson, I., Flyamer, I. M., Kumar, Y., Hill, R. E., Lettice, L. A., & Bickmore, W. A. (2022). Cohesin is required for long-range enhancer action at the Shh locus. *Nature Structural & Molecular Biology*, *29*(9), Article 9. https://doi.org/10.1038/s41594-022-00821-8

Kervestin, S., & Jacobson, A. (2012). NMD: A multifaceted response to premature translational termination. *Nature Reviews Molecular Cell Biology*, *13*(11), Article 11. https://doi.org/10.1038/nrm3454

Khorana, H. G., Agarwal, K. L., Büchi, H., Caruthers, M. H., Gupta, N. K., Klbppe, K., Kumar, A., Ohtsuka, E., RajBhandary, U. L., van de Sande, J. H., Sgaramella, V., Tebao, T., Weber, H., & Yamada, T. (1972). CIII. Total synthesis of the structural gene for an alanine transfer ribonucleic acid from yeast. *Journal of Molecular Biology*, *72*(2), 209–217. https://doi.org/10.1016/0022-2836(72)90146-5

Kim, Y., Shi, Z., Zhang, H., Finkelstein, I. J., & Yu, H. (2019). Human cohesin compacts DNA by loop extrusion. *Science*, *366*(6471), 1345–1349. https://doi.org/10.1126/science.aaz4475

Koster, C. C., Postma, E. D., Knibbe, E., Cleij, C., & Daran-Lapujade, P. (2022). Synthetic Genomics From a Yeast Perspective. *Frontiers in Bioengineering and Biotechnology*, *10*. https://doi.org/10.3389/fbioe.2022.869486

Kouprina, N., & Larionov, V. (2008). Selective isolation of genomic loci from complex genomes by transformation-associated recombination cloning in the yeast Saccharomyces cerevisiae. *Nature Protocols*, *3*(3), Article 3. https://doi.org/10.1038/nprot.2008.5

Krietenstein, N., Abraham, S., Venev, S. V., Abdennur, N., Gibcus, J., Hsieh, T.-H. S., Parsi, K. M., Yang, L., Maehr, R., Mirny, L. A., Dekker, J., & Rando, O. J. (2020). Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell*, *78*(3), 554-565.e7. https://doi.org/10.1016/j.molcel.2020.03.003

Kunes, S., Botstein, D., & Fox, M. S. (1985). Transformation of yeast with linearized plasmid DNA: Formation of inverted dimers and recombinant plasmid products. *Journal of Molecular Biology*, *184*(3), 375–387. https://doi.org/10.1016/0022-2836(85)90288-8

Kutyna, D. R., Onetto, C. A., Williams, T. C., Goold, H. D., Paulsen, I. T., Pretorius, I. S., Johnson, D. L., & Borneman, A. R. (2022). Construction of a synthetic Saccharomyces cerevisiae pan-genome neo-chromosome. *Nature Communications*, *13*(1), 3628. https://doi.org/10.1038/s41467-022-31305-4

Lartigue, C., Glass, J. I., Alperovich, N., Pieper, R., Parmar, P. P., Hutchison, C. A., Smith, H. O., & Venter, J. C. (2007). Genome Transplantation in Bacteria: Changing One Species to Another. *Science*, *317*(5838), 632–638. https://doi.org/10.1126/science.1144622

Lartigue, C., Vashee, S., Algire, M. A., Chuang, R.-Y., Benders, G. A., Ma, L., Noskov, V. N., Denisova, E. A., Gibson, D. G., Assad-Garcia, N., Alperovich, N., Thomas, D. W., Merryman,

C., Hutchison, C. A., Smith, H. O., Venter, J. C., & Glass, J. I. (2009). Creating bacterial strains from genomes that have been cloned and engineered in yeast. *Science (New York, N.Y.)*, *325*(5948), 1693–1696. https://doi.org/10.1126/science.1173759

Laughery, M. F., Hunter, T., Brown, A., Hoopes, J., Ostbye, T., Shumaker, T., & Wyrick, J. J. (2015). New vectors for simple and streamlined CRISPR–Cas9 genome editing in Saccharomyces cerevisiae. *Yeast*, *32*(12), 711–720. https://doi.org/10.1002/yea.3098

Lazar-Stefanita, L., Luo, J., Haase, M. A. B., Zhang, W., & Boeke, J. D. (2023). Two differentially stable rDNA loci coexist on the same chromosome and form a single nucleolus. *Proceedings of the National Academy of Sciences of the United States of America*, *120*(9), e2219126120. https://doi.org/10.1073/pnas.2219126120

Lazar-Stefanita, L., Scolari, V. F., Mercy, G., Muller, H., Guérin, T. M., Thierry, A., Mozziconacci, J., & Koszul, R. (2017). Cohesins and condensins orchestrate the 4D dynamics of yeast chromosomes during the cell cycle. *The EMBO Journal*, *36*(18), 2684–2697. https://doi.org/10.15252/embj.201797342

Le, T. B. K., Imakaev, M. V., Mirny, L. A., & Laub, M. T. (2013). High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome. *Science*, *342*(6159), 731–734. https://doi.org/10.1126/science.1242059

Lee, M., & Struhl, K. (1995). Mutations on the DNA-Binding Surface of TATA-Binding Protein Can Specifically Impair the Response to Acidic Activators In Vivo. *Molecular and Cellular Biology*, *15*(10), 5461–5469. https://doi.org/10.1128/MCB.15.10.5461

Lejeune, J., Turpin, R., & Gautier, M. (1959). [Mongolism; a chromosomal disease (trisomy)]. *Bulletin De l'Academie Nationale De Medecine*, *143*(11–12), 256–265.

Lengronne, A., Katou, Y., Mori, S., Yokobayashi, S., Kelly, G. P., Itoh, T., Watanabe, Y., Shirahige, K., & Uhlmann, F. (2004). Cohesin relocation from sites of chromosomal loading to places of convergent transcription. *Nature*, *430*(6999), Article 6999. https://doi.org/10.1038/nature02742

Li, X., Tang, X., Bing, X., Catalano, C., Li, T., Dolsten, G., Wu, C., & Levine, M. (2023). GAGA-associated factor fosters loop formation in the Drosophila genome. *Molecular Cell*, *83*(9), 1519-1526.e4. https://doi.org/10.1016/j.molcel.2023.03.011

Lieberman-Aiden, E., Berkum, N. L. van, Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., Sandstrom, R., Bernstein, B., Bender, M. A., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L. A., Lander, E. S., & Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Luo, J., Sun, X., Cormack, B. P., & Boeke, J. D. (2018). Karyotype engineering by chromosome fusion leads to reproductive isolation in yeast. *Nature*, *560*(7718), 392–396. https://doi.org/10.1038/s41586-018-0374-x

Lupiáñez, D. G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J. M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S. A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., … Mundlos, S. (2015). Disruptions of Topological Chromatin Domains Cause Pathogenic Rewiring of Gene-Enhancer Interactions. *Cell*, *161*(5), 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

Lupiáñez, D. G., Spielmann, M., & Mundlos, S. (2016). Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends in Genetics: TIG*, *32*(4), 225–237. https://doi.org/10.1016/j.tig.2016.01.003

Luppino, J. M., Park, D. S., Nguyen, S. C., Lan, Y., Xu, Z., Yunker, R., & Joyce, E. F. (2020). Cohesin promotes stochastic domain intermingling to ensure proper regulation of boundary-proximal genes. *Nature Genetics*, *52*(8), 840–848. https://doi.org/10.1038/s41588-020-0647-9

Luthra, I., Jensen, C., Chen, X. E., Salaudeen, A. L., Rafi, A. M., & de Boer, C. G. (2024). Regulatory activity is the default DNA state in eukaryotes. *Nature Structural & Molecular Biology*, *31*(3), 559–567. https://doi.org/10.1038/s41594-024-01235-4

Malabat, C., Feuerbach, F., Ma, L., Saveanu, C., & Jacquier, A. (2015). Quality control of transcription start site selection by nonsense-mediated-mRNA decay. *eLife*, *4*, e06722. https://doi.org/10.7554/eLife.06722

Malcı, K., Walls, L. E., & Rios-Solis, L. (2020). Multiplex Genome Engineering Methods for Yeast Cell Factory Development. *Frontiers in Bioengineering and Biotechnology*, *8*. https://doi.org/10.3389/fbioe.2020.589468

Maquat, L. E. (2004). Nonsense-mediated mRNA decay: Splicing, translation and mRNP dynamics. *Nature Reviews. Molecular Cell Biology*, *5*(2), 89–99. https://doi.org/10.1038/nrm1310

Marie-Nelly, H., Marbouty, M., Cournac, A., Liti, G., Fischer, G., Zimmer, C., & Koszul, R. (2014). Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinformatics (Oxford, England)*, *30*(15), 2105–2113. https://doi.org/10.1093/bioinformatics/btu162

Matthey-Doret, C., Baudry, L., Breuer, A., Montagne, R., Guiglielmoni, N., Scolari, V., Jean, E., Campeas, A., Chanut, P. H., Oriol, E., Méot, A., Politis, L., Vigouroux, A., Moreau, P., Koszul, R., & Cournac, A. (2020). Computer vision for pattern detection in chromosome contact maps. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-020-19562-7

Matthey-Doret, C., Baudry, L., Mortaza, S., Moreau, P., Koszul, R., & Cournac, A. (2022). Normalization of Chromosome Contact Maps: Matrix Balancing and Visualization. *Methods in Molecular Biology (Clifton, N.J.)*, *2301*, 1–15. https://doi.org/10.1007/978-1-0716-1390-0_1

Mavrich, T. N., Ioshikhes, I. P., Venters, B. J., Jiang, C., Tomsho, L. P., Qi, J., Schuster, S. C., Albert, I., & Pugh, B. F. (2008). A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Research*, *18*(7), 1073–1083. https://doi.org/10.1101/gr.078261.108

Mercy, G., Mozziconacci, J., Scolari, V. F., Yang, K., Zhao, G., Thierry, A., Luo, Y., Mitchell, L. A., Shen, M., Shen, Y., Walker, R., Zhang, W., Wu, Y., Xie, Z.-X., Luo, Z., Cai, Y., Dai, J., Yang, H., Yuan, Y.-J., … Koszul, R. (2017). 3D organization of synthetic and scrambled chromosomes. *Science (New York, N.Y.)*, *355*(6329). https://doi.org/10.1126/science.aaf4597

Michaelis, C., Ciosk, R., & Nasmyth, K. (1997). Cohesins: Chromosomal proteins that prevent premature separation of sister chromatids. *Cell*, *91*(1), 35–45. https://doi.org/10.1016/s0092-8674(01)80007-6

Millán-Zambrano, G., Burton, A., Bannister, A. J., & Schneider, R. (2022). Histone post-translational modifications—Cause and consequence of genome function. *Nature Reviews. Genetics*, *23*(9), 563–580. https://doi.org/10.1038/s41576-022-00468-7

Mitchell, L. A., McCulloch, L. H., Pinglay, S., Berger, H., Bosco, N., Brosh, R., Bulajić, M., Huang, E., Hogan, M. S., Martin, J. A., Mazzoni, E. O., Davoli, T., Maurano, M. T., & Boeke, J. D. (2021). De novo assembly and delivery to mouse cells of a 101 kb functional human gene. *Genetics*, *218*(1), iyab038. https://doi.org/10.1093/genetics/iyab038

Mitter, M., Gasser, C., Takacs, Z., Langer, C. C. H., Tang, W., Jessberger, G., Beales, C. T., Neuner, E., Ameres, S. L., Peters, J.-M., Goloborodko, A., Micura, R., & Gerlich, D. W. (2020). Conformation of sister chromatids in the replicated human genome. *Nature*, *586*(7827), 139–144. https://doi.org/10.1038/s41586-020-2744-4

Muller, H., Scolari, V. F., Agier, N., Piazza, A., Thierry, A., Mercy, G., Descorps-Declere, S., Lazar-Stefanita, L., Espeli, O., Llorente, B., Fischer, G., Mozziconacci, J., & Koszul, R. (2018). Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for Hi-C. *Molecular Systems Biology*, *14*(7), e8293. https://doi.org/10.15252/msb.20188293

Nagano, T., Lubling, Y., Stevens, T. J., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E. D., Tanay, A., & Fraser, P. (2013). Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, *502*(7469), 59–64. https://doi.org/10.1038/nature12593

Nasmyth, K., & Haering, C. H. (2005). The structure and function of SMC and kleisin complexes. *Annual Review of Biochemistry*, *74*, 595–648.

https://doi.org/10.1146/annurev.biochem.74.082803.133219
Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B. R., Mirny, L. A., & Dekker, J. (2013). Organization of the Mitotic Chromosome. *Science*, *342*(6161), 948–953. https://doi.org/10.1126/science.1236083
Nevers, A., Doyen, A., Malabat, C., Néron, B., Kergrohen, T., Jacquier, A., & Badis, G. (2018). Antisense transcriptional interference mediates condition-specific gene repression in budding yeast. *Nucleic Acids Research*, *46*(12), 6009–6025. https://doi.org/10.1093/nar/gky342
Nora, E. P., Goloborodko, A., Valton, A.-L., Gibcus, J. H., Uebersohn, A., Abdennur, N., Dekker, J., Mirny, L. A., & Bruneau, B. G. (2017). Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization. *Cell*, *169*(5), 930-944.e22. https://doi.org/10.1016/j.cell.2017.05.004
Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., & Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, *485*(7398), Article 7398. https://doi.org/10.1038/nature11049
Noskov, V. N., Karas, B. J., Young, L., Chuang, R.-Y., Gibson, D. G., Lin, Y.-C., Stam, J., Yonemoto, I. T., Suzuki, Y., Andrews-Pfannkoch, C., Glass, J. I., Smith, H. O., Hutchison, C. A., Venter, J. C., & Weyman, P. D. (2012). Assembly of Large, High G+C Bacterial DNA Fragments in Yeast. *ACS Synthetic Biology*, *1*(7), 267–273. https://doi.org/10.1021/sb3000194
Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N., & Mirny, L. (2018). Chromatin Organization by an Interplay of Loop Extrusion and Compartmental Segregation. *Biophysical Journal*, *114*(3), 30a. https://doi.org/10.1016/j.bpj.2017.11.211
Nusbaum, C., Slonim, D. K., Harris, K. L., Birren, B. W., Steen, R. G., Stein, L. D., Miller, J., Dietrich, W. F., Nahf, R., Wang, V., Merport, O., Castle, A. B., Husain, Z., Farino, G., Gray, D., Anderson, M. O., Devine, R., Horton, L. T., Ye, W., … Lander, E. S. (1999). A YAC-based physical map of the mouse genome. *Nature Genetics*, *22*(4), 388–393. https://doi.org/10.1038/11967
Ohno, M., Ando, T., Priest, D. G., & Taniguchi, Y. (2021). Hi-CO: 3D genome structure analysis with nucleosome resolution. *Nature Protocols*, *16*(7), 3439–3469. https://doi.org/10.1038/s41596-021-00543-z
Oomen, M. E., Hedger, A. K., Watts, J. K., & Dekker, J. (2020). Detecting chromatin interactions between and along sister chromatids with SisterC. *Nature Methods*, *17*(10), 1002–1009. https://doi.org/10.1038/s41592-020-0930-9
O'Sullivan, J. M., Tan-Wong, S. M., Morillon, A., Lee, B., Coles, J., Mellor, J., & Proudfoot, N. J. (2004). Gene loops juxtapose promoters and terminators in yeast. *Nature Genetics*, *36*(9), Article 9. https://doi.org/10.1038/ng1411
Otto, M., Skrekas, C., Gossing, M., Gustafsson, J., Siewers, V., & David, F. (2021). Expansion of the Yeast Modular Cloning Toolkit for CRISPR-Based Applications, Genomic Integrations and Combinatorial Libraries. *ACS Synthetic Biology*, *10*(12), 3461–3474. https://doi.org/10.1021/acssynbio.1c00408
Oudet, P., Gross-Bellard, M., & Chambon, P. (1975). Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell*, *4*(4), 281–300. https://doi.org/10.1016/0092-8674(75)90149-x
Paliou, C., Guckelberger, P., Schöpflin, R., Heinrich, V., Esposito, A., Chiariello, A. M., Bianco, S., Annunziatella, C., Helmuth, J., Haas, S., Jerković, I., Brieske, N., Wittler, L., Timmermann, B., Nicodemi, M., Vingron, M., Mundlos, S., & Andrey, G. (2019). Preformed chromatin topology assists transcriptional robustness of Shh during limb development. *Proceedings of the National Academy of Sciences*, *116*(25), 12390–12399. https://doi.org/10.1073/pnas.1900672116
Passarge, E. (1979). Emil Heitz and the concept of heterochromatin: Longitudinal chromosome differentiation was recognized fifty years ago. *American Journal of Human Genetics*, *31*(2), 106–115. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1685768/

Pelechano, V., & Steinmetz, L. M. (2013). Gene regulation by antisense transcription. *Nature Reviews Genetics*, *14*(12), 880–893. https://doi.org/10.1038/nrg3594

Petrascheck, M., Escher, D., Mahmoudi, T., Verrijzer, C. P., Schaffner, W., & Barberis, A. (2005). DNA looping induced by a transcriptional enhancer in vivo. *Nucleic Acids Research*, *33*(12), 3743–3750. https://doi.org/10.1093/nar/gki689

Pincus, D., Aranda-Díaz, A., Zuleta, I. A., Walter, P., & El-Samad, H. (2014). Delayed Ras/PKA signaling augments the unfolded protein response. *Proceedings of the National Academy of Sciences*, *111*(41), 14800–14805. https://doi.org/10.1073/pnas.1409588111

Postlethwait, J. H., Yan, Y.-L., Gates, M. A., Horne, S., Amores, A., Brownlie, A., Donovan, A., Egan, E. S., Force, A., Gong, Z., Goutel, C., Fritz, A., Kelsh, R., Knapik, E., Liao, E., Paw, B., Ransom, D., Singer, A., Thomson, M., … Talbot, W. S. (1998). Vertebrate genome evolution and the zebrafish gene map. *Nature Genetics*, *18*(4), 345–349. https://doi.org/10.1038/ng0498-345

Rajderkar, S., Barozzi, I., Zhu, Y., Hu, R., Zhang, Y., Li, B., Alcaina Caro, A., Fukuda-Yuzawa, Y., Kelman, G., Akeza, A., Blow, M. J., Pham, Q., Harrington, A. N., Godoy, J., Meky, E. M., von Maydell, K., Hunter, R. D., Akiyama, J. A., Novak, C. S., … Pennacchio, L. A. (2023). Topologically associating domain boundaries are required for normal genome function. *Communications Biology*, *6*(1), 1–10. https://doi.org/10.1038/s42003-023-04819-w

Rao, S. S. P., Huang, S.-C., Glenn St Hilaire, B., Engreitz, J. M., Perez, E. M., Kieffer-Kwon, K.-R., Sanborn, A. L., Johnstone, S. E., Bascom, G. D., Bochkov, I. D., Huang, X., Shamim, M. S., Shin, J., Turner, D., Ye, Z., Omer, A. D., Robinson, J. T., Schlick, T., Bernstein, B. E., … Aiden, E. L. (2017). Cohesin Loss Eliminates All Loop Domains. *Cell*, *171*(2), 305-320.e24. https://doi.org/10.1016/j.cell.2017.09.026

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., & Aiden, E. L. (2014). A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Redden, H., & Alper, H. S. (2015). The development and characterization of synthetic minimal yeast promoters. *Nature Communications*, *6*(1), 7810. https://doi.org/10.1038/ncomms8810

Rhee, H. S., & Pugh, B. F. (2012). Genome-wide structure and organization of eukaryotic pre-initiation complexes. *Nature*, *483*(7389), 295–301. https://doi.org/10.1038/nature10799

Sainsbury, S., Bernecky, C., & Cramer, P. (2015). Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, *16*(3), 129–143. https://doi.org/10.1038/nrm3952

Schalbetter, S. A., Goloborodko, A., Fudenberg, G., Belton, J.-M., Miles, C., Yu, M., Dekker, J., Mirny, L., & Baxter, J. (2017). SMC complexes differentially compact mitotic chromosomes according to genomic context. *Nature Cell Biology*, *19*(9), 1071–1080. https://doi.org/10.1038/ncb3594

Schindler, D., Walker, R. S. K., Jiang, S., Brooks, A. N., Wang, Y., Müller, C. A., Cockram, C., Luo, Y., García, A., Schraivogel, D., Mozziconacci, J., Pena, N., Assari, M., Sánchez Olmos, M. del C., Zhao, Y., Ballerini, A., Blount, B. A., Cai, J., Ogunlana, L., … Cai, Y. (2023). Design, construction, and functional characterization of a tRNA neochromosome in yeast. *Cell*, *186*(24), 5237-5253.e22. https://doi.org/10.1016/j.cell.2023.10.015

Schmidt, R., Cnops, G., Bancroft, I., & Dean, C. (1992). Construction of an Overlapping YAC Library of the Arabidopsis thaliana Genome. *Functional Plant Biology*, *19*(4), 341–351. https://doi.org/10.1071/pp9920341

Schulz, D., Schwalb, B., Kiesel, A., Baejen, C., Torkler, P., Gagneur, J., Soeding, J., & Cramer, P. (2013). Transcriptome Surveillance by Selective Termination of Noncoding RNA Synthesis. *Cell*, *155*(5), 1075–1087. https://doi.org/10.1016/j.cell.2013.10.024

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N. A., Huber, W., Haering, C. H., Mirny, L., & Spitz, F. (2017). Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, *551*(7678), Article 7678.

https://doi.org/10.1038/nature24281
Scolari, V. F., Mercy, G., Koszul, R., Lesne, A., & Mozziconacci, J. (2018). Kinetic Signature of Cooperativity in the Irreversible Collapse of a Polymer. *Physical Review Letters*, *121*(5), 057801. https://doi.org/10.1103/PhysRevLett.121.057801
Seila, A. C., Calabrese, J. M., Levine, S. S., Yeo, G. W., Rahl, P. B., Flynn, R. A., Young, R. A., & Sharp, P. A. (2008). Divergent Transcription from Active Promoters. *Science*, *322*(5909), 1849–1851. https://doi.org/10.1126/science.1162253
Sémon, M., & Duret, L. (2004). Evidence that functional transcription units cover at least half of the human genome. *Trends in Genetics*, *20*(5), 229–232. https://doi.org/10.1016/j.tig.2004.03.001
Serizay, J., Matthey-Doret, C., Bignaud, A., Baudry, L., & Koszul, R. (2024). Orchestrating chromosome conformation capture analysis with Bioconductor. *Nature Communications*, *15*(1), 1072. https://doi.org/10.1038/s41467-024-44761-x
Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., & Cavalli, G. (2012). Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. *Cell*, *148*(3), 458–472. https://doi.org/10.1016/j.cell.2012.01.010
Shao, Y., Lu, N., Wu, Z., Cai, C., Wang, S., Zhang, L.-L., Zhou, F., Xiao, S., Liu, L., Zeng, X., Zheng, H., Yang, C., Zhao, Z., Zhao, G., Zhou, J.-Q., Xue, X., & Qin, Z. (2018). Creating a functional single-chromosome yeast. *Nature*, *560*(7718), 331–335. https://doi.org/10.1038/s41586-018-0382-x
Shaw, W. M., Khalil, A. S., & Ellis, T. (2023). A Multiplex MoClo Toolkit for Extensive and Flexible Engineering of Saccharomyces cerevisiae. *ACS Synthetic Biology*, *12*(11), 3393–3405. https://doi.org/10.1021/acssynbio.3c00423
Shi, S., Liang, Y., Zhang, M. M., Ang, E. L., & Zhao, H. (2016). A highly efficient single-step, markerless strategy for multi-copy chromosomal integration of large biochemical pathways in Saccharomyces cerevisiae. *Metabolic Engineering*, *33*, 19–27. https://doi.org/10.1016/j.ymben.2015.10.011
Sikorska, N., & Sexton, T. (2020). Defining Functionally Relevant Spatial Chromatin Domains: It is a TAD Complicated. *Journal of Molecular Biology*, *432*(3), 653–664. https://doi.org/10.1016/j.jmb.2019.12.006
Smith, H. O., Hutchison, C. A., Pfannkoch, C., & Venter, J. C. (2003). Generating a synthetic genome by whole genome assembly: phiX174 bacteriophage from synthetic oligonucleotides. *Proceedings of the National Academy of Sciences of the United States of America*, *100*(26), 15440–15445. https://doi.org/10.1073/pnas.2237126100
Solovei, I., Kreysing, M., Lanctôt, C., Kösem, S., Peichl, L., Cremer, T., Guck, J., & Joffe, B. (2009). Nuclear Architecture of Rod Photoreceptor Cells Adapts to Vision in Mammalian Evolution. *Cell*, *137*(2), 356–368. https://doi.org/10.1016/j.cell.2009.01.052
Soutourina, J. (2018). Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology*, *19*(4), 262–274. https://doi.org/10.1038/nrm.2017.115
Stemmer, W. P. C., Crameri, A., Ha, K. D., Brennan, T. M., & Heyneker, H. L. (1995). Single-step assembly of a gene and entire plasmid from large numbers of oligodeoxyribonucleotides. *Gene*, *164*(1), 49–53. https://doi.org/10.1016/0378-1119(95)00511-4
Stephens, A. D., Quammen, C. W., Chang, B., Haase, J., Taylor, R. M., & Bloom, K. (2013). The spatial segregation of pericentric cohesin and condensin in the mitotic spindle. *Molecular Biology of the Cell*, *24*(24), 3909–3919. https://doi.org/10.1091/mbc.E13-06-0325
Struhl, K. (2007). Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nature Structural & Molecular Biology*, *14*(2), 103–105. https://doi.org/10.1038/nsmb0207-103
Su, J.-H., Zheng, P., Kinrot, S. S., Bintu, B., & Zhuang, X. (2020). Genome-Scale Imaging of the 3D Organization and Transcriptional Activity of Chromatin. *Cell*, *182*(6), 1641-1659.e26. https://doi.org/10.1016/j.cell.2020.07.032
Taddei, A., & Gasser, S. M. (2004). Multiple pathways for telomere tethering: Functional implications of subnuclear position for heterochromatin formation. *Biochimica et Biophysica Acta (BBA) - Gene Structure and Expression*, *1677*(1), 120–128.

https://doi.org/10.1016/j.bbaexp.2003.11.014
Taddei, A., & Gasser, S. M. (2012). Structure and Function in the Budding Yeast Nucleus. *Genetics*, *192*(1), 107–129. https://doi.org/10.1534/genetics.112.140608
Taddei, A., Houwe, G. V., Nagai, S., Erb, I., Nimwegen, E. van, & Gasser, S. M. (2009). The functional importance of telomere clustering: Global changes in gene expression result from SIR factor dispersion. *Genome Research*, *19*(4), 611–625. https://doi.org/10.1101/gr.083881.108
Taddei, A., Schober, H., & Gasser, S. M. (2010). The Budding Yeast Nucleus. *Cold Spring Harbor Perspectives in Biology*, *2*(8), a000612. https://doi.org/10.1101/cshperspect.a000612
Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., Sheffield, N. C., Stergachis, A. B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L., Canfield, T. K., Diegel, M., Dunn, D., Ebersol, A. K., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82. https://doi.org/10.1038/nature11232
Tjio, J. H., & Levan, A. (1956). The Chromosome Number of Man. *Hereditas*, *42*(1–2), 1–6. https://doi.org/10.1111/j.1601-5223.1956.tb03010.x
Uhlmann, F. (2016). SMC complexes: From DNA to chromosomes. *Nature Reviews. Molecular Cell Biology*, *17*(7), 399–412. https://doi.org/10.1038/nrm.2016.30
Uhlmann, F., Lottspeich, F., & Nasmyth, K. (1999). Sister-chromatid separation at anaphase onset is promoted by cleavage of the cohesin subunit Scc1. *Nature*, *400*(6739), 37–42. https://doi.org/10.1038/21831
Veron, M., Zou, Y., Yu, Q., Bi, X., Selmi, A., Gilson, E., & Defossez, P.-A. (2006). Histone H1 of Saccharomyces cerevisiae Inhibits Transcriptional Silencing. *Genetics*, *173*(2), 579–587. https://doi.org/10.1534/genetics.105.050195
Wang, D. (2018). GCevobase: An evolution-based database for GC content in eukaryotic genomes. *Bioinformatics*, *34*(12), 2129–2131. https://doi.org/10.1093/bioinformatics/bty068
Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*, *24*(11), 437–440. https://doi.org/10.1016/S0968-0004(99)01460-7
Wendt, K. S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., & Peters, J.-M. (2008). Cohesin mediates transcriptional insulation by CCCTC-binding factor. *Nature*, *451*(7180), 796–801. https://doi.org/10.1038/nature06634
Wit, E. de, & Laat, W. de. (2012). A decade of 3C technologies: Insights into nuclear organization. *Genes & Development*, *26*(1), 11–24. https://doi.org/10.1101/gad.179804.111
Wutz, G., Ladurner, R., St Hilaire, B. G., Stocsits, R. R., Nagasaka, K., Pignard, B., Sanborn, A., Tang, W., Várnai, C., Ivanov, M. P., Schoenfelder, S., van der Lelij, P., Huang, X., Dürnberger, G., Roitinger, E., Mechtler, K., Davidson, I. F., Fraser, P., Lieberman-Aiden, E., & Peters, J.-M. (2020). ESCO1 and CTCF enable formation of long chromatin loops by protecting cohesinSTAG1 from WAPL. *eLife*, *9*, e52091. https://doi.org/10.7554/eLife.52091
Wutz, G., Várnai, C., Nagasaka, K., Cisneros, D. A., Stocsits, R. R., Tang, W., Schoenfelder, S., Jessberger, G., Muhar, M., Hossain, M. J., Walther, N., Koch, B., Kueblbeck, M., Ellenberg, J., Zuber, J., Fraser, P., & Peters, J.-M. (2017). Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *The EMBO Journal*, *36*(24), 3573–3599. https://doi.org/10.15252/embj.201798004
Wyers, F., Rougemaille, M., Badis, G., Rousselle, J.-C., Dufour, M.-E., Boulay, J., Régnault, B., Devaux, F., Namane, A., Séraphin, B., Libri, D., & Jacquier, A. (2005). Cryptic Pol II Transcripts Are Degraded by a Nuclear Quality Control Pathway Involving a New Poly(A) Polymerase. *Cell*, *121*(5), 725–737. https://doi.org/10.1016/j.cell.2005.04.030
Xu, X., Meier, F., Blount, B. A., Pretorius, I. S., Ellis, T., Paulsen, I. T., & Williams, T. C. (2023). Trimming the genomic fat: Minimising and re-functionalising genomes using synthetic biology. *Nature Communications*, *14*(1), 1984. https://doi.org/10.1038/s41467-023-37748-7

Xu, Z., Wei, W., Gagneur, J., Perocchi, F., Clauder-Münster, S., Camblong, J., Guffanti, E., Stutz, F., Huber, W., & Steinmetz, L. M. (2009). Bidirectional promoters generate pervasive transcription in yeast. *Nature*, *457*(7232), 1033–1037. https://doi.org/10.1038/nature07728

Yang, C., Bolotin, E., Jiang, T., Sladek, F. M., & Martinez, E. (2007). Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, *389*(1), 52–65. https://doi.org/10.1016/j.gene.2006.09.029

Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., & Rando, O. J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)*, *309*(5734), 626–630. https://doi.org/10.1126/science.1112178

Zhang, Y., Wang, J., Wang, Z., Zhang, Y., Shi, S., Nielsen, J., & Liu, Z. (2019). A gRNA-tRNA array for CRISPR-Cas9 based rapid multiplexed genome editing in Saccharomyces cerevisiae. *Nature Communications*, *10*(1), 1053. https://doi.org/10.1038/s41467-019-09005-3

Zhang, Y., Zhang, X., Ba, Z., Liang, Z., Dring, E. W., Hu, H., Lou, J., Kyritsis, N., Zurita, J., Shamim, M. S., Presser Aiden, A., Lieberman Aiden, E., & Alt, F. W. (2019). The fundamental role of chromatin loop extrusion in physiological V(D)J recombination. *Nature*, *573*(7775), 600–604. https://doi.org/10.1038/s41586-019-1547-y

Zhang, Z., & Dietrich, F. S. (2005). Mapping of transcription start sites in Saccharomyces cerevisiae using 5' SAGE. *Nucleic Acids Research*, *33*(9), 2838–2851. https://doi.org/10.1093/nar/gki583

Zhao, Y., Coelho, C., Lauer, S., Majewski, M., Laurent, J. M., Brosh, R., & Boeke, J. D. (2023). CREEPY: CRISPR-mediated editing of synthetic episomes in yeast. *Nucleic Acids Research*, *51*(13), e72. https://doi.org/10.1093/nar/gkad491

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Singh Sandhu, K., Singh, U., Pant, V., Tiwari, V., Kurukuti, S., & Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nature Genetics*, *38*(11), 1341–1347. https://doi.org/10.1038/ng1891

Zhou, J. (2022). Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nature Genetics*, 1–10. https://doi.org/10.1038/s41588-022-01065-4

Zuin, J., Roth, G., Zhan, Y., Cramard, J., Redolfi, J., Piskadlo, E., Mach, P., Kryzhanovska, M., Tihanyi, G., Kohler, H., Eder, M., Leemans, C., van Steensel, B., Meister, P., Smallwood, S., & Giorgetti, L. (2022). Nonlinear control of transcription through enhancer–promoter interactions. *Nature*, *604*(7906), Article 7906. https://doi.org/10.1038/s41586-022-04570-y