Modélisation des systèmes biologiques, analyse de données multiomiques et visualisation à large échelle

Thibault Poinsignon

université PARIS-SACLAY

Modélisation des systèmes biologiques, analyse de données multiomiques et visualisation à large échelle

Modeling of biological systems, multi-omics data analysis and large-scale visualization

Thèse de doctorat de l'université Paris-Saclay

École doctorale n°577, Structure et Dynamique des Systèmes Vivants (SDSV)
Spécialité de doctorat : Sciences de la vie et de la santé
Graduate School : Life Sciences and Health
Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche **I2BC (Université Paris-Saclay, CEA, CNRS)** et **Institut Jacques Monod (CNRS, Université Paris Cité)**, sous la direction de **Gaëlle LELANDAIS**, professeure des Universités, la co-direction de **Pierre POULAIN**, maître de conférences et le co-encadrement de **Mélina GALLOPIN**, maître de conférences.

Thèse soutenue à Paris-Saclay, le 27 septembre 2023, par

Thibault POINSIGNON

Composition du Jury

Membres du jury avec voix délibérative

Sébastien BLOYER Professeur à l'Université Paris Saclay	Président
Nadia PONTS Chargée de recherche, HDR à l'INRA, Bordeaux (Institut National de la Recherche Agronomique)	Rapportrice & Examinatrice
Matthieu MONTES Professeur au CNAM, Paris (Conservatoire National des Arts et Métiers)	Rapporteur & Examinateur
Stéphanie BURY MONÉ Professeure à l'Université Paris Saclay	Examinatrice

THESE DE DOCTORAT
NNT : 2023UPASL078

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Modélisation des systèmes biologiques, analyse de données multiomiques et visualisation à large échelle

Mots clés : Données Hi-C, modélisation 3D, intégration multiomique, visualisation des données, organisation du génome.

Résumé : L'ampleur de la complexité des systèmes biologiques fait de leur représentation en modèles simplifiés une étape cruciale de leur étude. Dans le cas du système cellulaire, l'évolution rapide des méthodes de mesure a créé un changement d'échelle dans la quantité de données disponibles. La génération de données omiques est maintenant courante et les biologistes sont confrontés au défi de traiter toujours plus de données. Alors que la construction de modèles à grande échelle devient possible, la visualisation de ces modèles reste un défi crucial pour la structuration des connaissances. Ainsi, cette thèse étudie l'enjeu de la visualisation à grande échelle des données omiques. Dans cette optique, nous avons exploré l'utilisation des modèles 3D des génomes générés à partir des données Hi-C comme support à la visualisation et à l'intégration de données omiques. Pour cela nous avons conçu un workflow allant des données brutes Hi-C aux modèles 3D entièrement annotés et nous avons réanalysé des ensembles de données omiques publiques disponibles pour trois espèces modèles de champignons : *S. cerevisiae*, *S. pombe*, *N. crassa*.

Title: Modeling of biological systems, multi-omics data analysis and large-scale visualization

Keywords: Hi-C data, 3D modeling, integration of omics data, data visualization, genome organization.

Abstract: The sheer complexity of biological systems makes their representation in simplified models a crucial step in their study. In the case of the cellular system, the rapid evolution of measurement methods has created a change of scale in the amount of data available for modeling. Omics data generation is now widespread, and biologists are faced with the challenge of processing an ever-increasing amount of data. While large-scale modeling is becoming possible, visualization of these models remains a crucial challenge for structuring knowledge. Thus, this thesis investigates the challenge of large-scale visualization of omics data. We explored the use of 3D models of genomes generated from Hi-C data as a support for the visualization and integration of omics data. To this end, we assembled a workflow from raw Hi-C data to fully annotated 3D models and reanalyzed publicly available omics datasets for three fungal model species: *S. cerevisiae*, *S. pombe*, *N. crassa*.

Ce projet a bénéficié du financement MINOMICS ANR-19-CE45-0017.

REMERCIEMENTS

Je voudrais tout d'abord remercier grandement ma directrice, mon co-directeur et ma co-encadrante de thèse Mme Gaëlle Lelandais, professeure à l'Université Paris Saclay, Mr Pierre Poulain, maître de conférences à l'Université Paris cité, et Mme Mélina Gallopin, maître de conférences à l'Université Paris Saclay, pour leur encadrement durant ces trois années. Leurs expertises, leurs points de vue complémentaires, leur soutien constant et leur enthousiasme pour mon sujet ont été essentiels à la réalisation de cette thèse.

J'adresse tous mes remerciements à Mr Sébastien Bloyer, professeur à l'Université Paris Saclay, Mme Nadia Ponts, chargée de recherche à l'INRA, Mr Matthieu Montès, professeur au CNAM, Mme Séphanie Bury Moné, professeure à l'Université Paris Saclay, pour leur participation à mon jury de thèse.

Je tiens aussi à remercier à nouveau Mme Séphanie Bury Moné et Mme Nadia Ponts pour leurs retours précieux lors du développement du workflow 3D Genome Builder.

Je remercie Mme Nelle Varoquaux, chercheuse au CNRS, pour son expertise et son aide précieuse sur l'utilisation de l'outil de modélisation Pastis.

Je voudrais également remercier les membres de l'équipe EDC à l'I2BC : Mme Fabienne Malagnac, professeure à l'Université Paris Saclay, Mr Pierre Grognet, maître de conférences à l'Université Paris Saclay, Mme Mengyuan Li, doctorante, Mme Eléonore Pillot-Lucas et Mr Damien Rémy pour les discussions quotidiennes et l'environnement de travail stimulant et agréable.

Enfin, je tiens à remercier Mr Jean Michel Camadro, directeur de recherche au cnrs, et toute son équipe à l'IJM pour leur accueil toujours chaleureux.

Merci enfin à tous ceux qui ont contribué de près ou de loin à la réalisation de cette thèse,

Thibault Poinsignon

1 INTRODUCTION

1.1 Modéliser l'immense complexité des systèmes biologiques

1.1.1 Qu'est-ce qu'un système biologique ? Une diversité d'échelles et de natures vertigineuse

1.1.1.1 La notion d'échelle dans le vivant, imbrication sans fin de systèmes complexes

Un système biologique peut être défini comme un ensemble d'acteurs interagissant entre eux, à différentes échelles, pour remplir une fonction. Cette définition très large met en valeur la difficulté de délimiter la biologie en tant que discipline, du fait de l'immense complexité de son sujet d'étude. Par complexité on entend qu'il est presque impossible de bien concevoir le nombre d'acteurs impliqués, le nombre d'interactions entre ces acteurs et le nombre d'échelles d'étude possible.

A l'échelle la plus large, l' « International Union for Conservation of Nature » divise la biosphère en 110 écosystèmes (Keith et al., 2022). La compréhension de ces systèmes a évolué rapidement : leurs acteurs ont été catégorisés et leurs interactions détaillées. Le « Catalogue of Life » (Bánki et al., 2023) liste 2,3 millions d'espèces identifiées sur l'arbre phylogénétique du vivant. Toutefois, ces espèces ne représentent qu'une fraction de la biosphère : la majorité des acteurs reste inconnue (Tricou, 2022). Décrire la complexité des écosystèmes est donc loin d'être une tâche accomplie et face à une telle complexité, les méthodes de mesure limitent la finesse de la description.

La technologie du LIDAR est une méthode de télédétection par laser si efficace qu'elle permet de traquer des pyramides sous la jungle (Prümers et al., 2022) ou bien de scanner la France pour en créer une immense carte 3D formée de voxels avec des résultats allant du m^3 au cm^3 [1]. Cette méthode de mesure permet aussi de reconstituer numériquement une forêt à l'arbre prêt (Zhong et al., 2022). Pourtant, l'écosystème "forêt tempérée" (Keith et al., 2022) est bien plus qu'une population d'arbres. En effet, une étude des forêts de l'est de la Hongrie a pu échantillonner 1 125 espèces différentes d'oiseaux, d'insectes, d'arachnides, de champignons et de plantes (Tinya et al., 2021) dans une zone de 312 km^2. Les nombreux individus de chaque espèce entretiennent des relations de prédation, parasitisme et symbiose évoluant au cours des saisons. Ainsi, même une méthode aussi résolutive et à large échelle que le LIDAR ne donne accès qu'à une fraction de la complexité d'un écosystème.

On pourrait supposer que réduire la taille du système biologique étudié réduirait la complexité. Mais un zoom sur un des acteurs vivants d'un écosystème (sans même s'arrêter à l'étude des populations) révèle des organismes formés de systèmes

[1] https://www.ign.fr/institut/lidar-hd-vers-une-nouvelle-cartographie-3d-du-territoire

interagissant en permanence pour maintenir un équilibre dynamique stable. L'étude des systèmes biologiques rejoint ici la médecine, qui illustre la complexité de cette échelle à travers la description précise de l'organisme humain. Un exemple parlant est le système olfactif humain : il est formé d'environ 6 à 10 millions de neurones sensoriels (Firestein, 2001) exprimant chacun un des ~1 100 gènes humains codant pour des récepteurs olfactifs. Chaque odeur est donc associée à une combinaison de récepteurs spécifiques activés. Une infime variation de structure d'une molécule odorante peut changer l'odeur perçue (Malnic et al., 1999). La complexité s'accentue encore à l'échelle cellulaire, système à la limite du visible où les acteurs sont moléculaires. Les cellules sont les unités de base de la vie, le socle de toute la complexité évoquée jusqu'ici (Theobald, 2010). Ce zoom rapide jusqu'à l'échelle cellulaire qui sera étudiée ici, illustre que la complexité des systèmes biologiques ne s'arrête pas à la cellule et son noyau. Les problématiques de modélisation et de visualisation à large échelle développées dans cette thèse le sont donc dans une fraction de l'image complète, un pixel d'une image bien plus vaste.

1.1.1.2 La cellule : une entité dynamique et organisée au cœur de la complexité biologique

La cellule est l'unité fondamentale de tous les organismes vivants. C'est un système biologique hautement organisé, capable de maintenir un équilibre dynamique interne et de répondre aux stimuli environnementaux. Ces usines moléculaires invisibles à l'œil nu réalisent des milliers de réactions chimiques par seconde, maintenant ainsi leur équilibre dynamique et leur capacité de réplication. L'immense diversité des acteurs moléculaires est divisée en quatre grandes familles (Hasin et al., 2017). Le génome est l'ensemble des molécules d'ADN de la cellule, qui contiennent toutes les informations nécessaires à son fonctionnement. Le transcriptome est l'ensemble des molécules d'ARN présentes dans la cellule qui sont produites à partir du génome. Le protéome est l'ensemble des protéines de la cellule, synthétisées à partir de l'ARN et responsables de la majorité des fonctions cellulaires. Enfin, le métabolome est constitué des métabolites, petites molécules qui sont les substrats des réactions enzymatiques.

Chacune de ces grandes familles contient en réalité plusieurs milliers d'acteurs moléculaires de nature et de fonction différentes. À elles seules, les protéines peuvent jouer le rôle d'enzymes, de transporteurs, de récepteurs et de composants structurels. À chacune de ces fonctions sont associées des milliers de structures, chacune présente en grand nombre. Par exemple, à un instant donné, environ 42 millions de copies des 5 307 protéines (identifiées avec évidences expérimentales sur UniProt) sont présentes dans une cellule de la levure *S. cerevisiae* (Ho et al., 2018), (voir **Figure 1**).

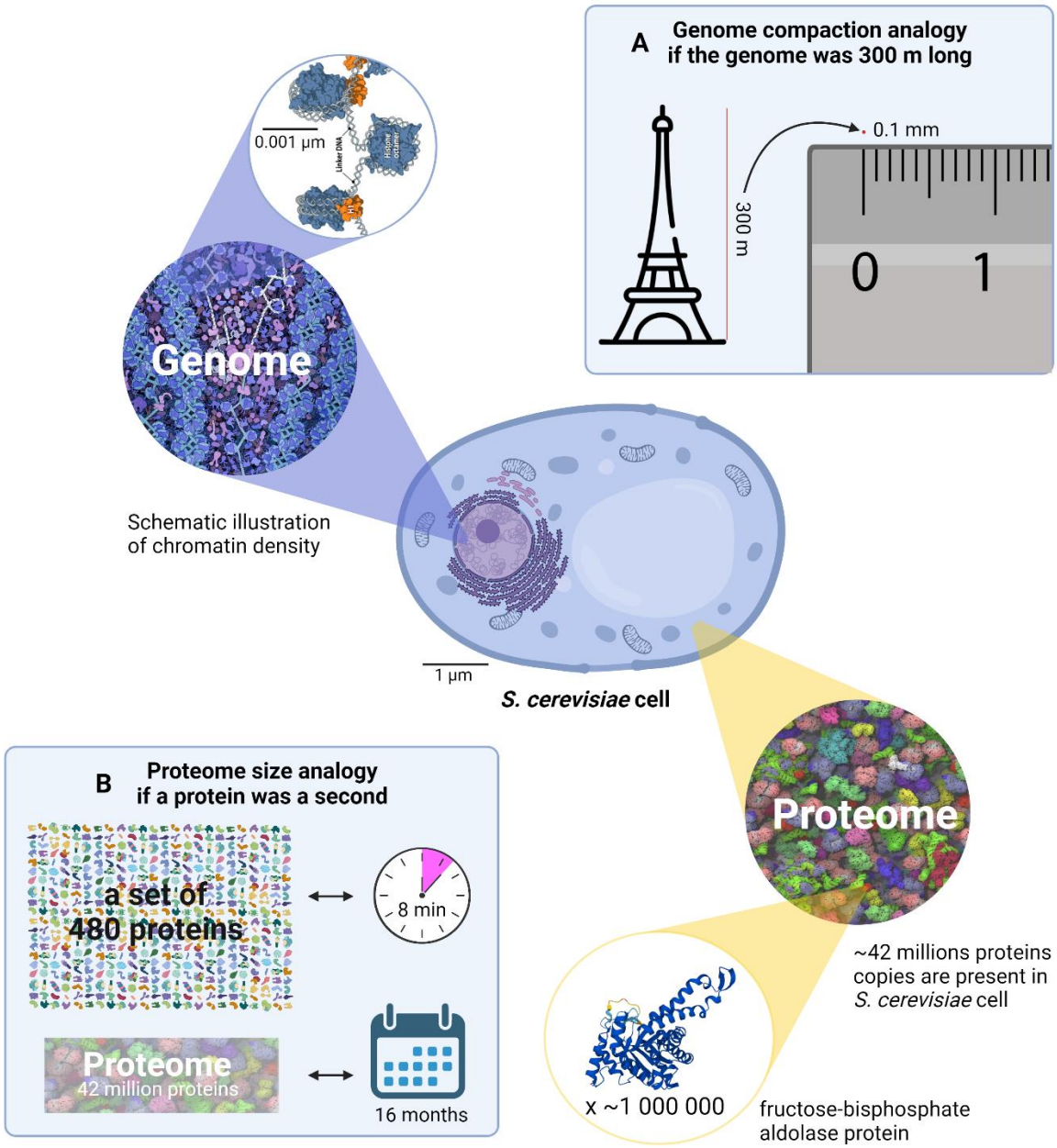


Figure 1 : L'exemple de *S. cerevisiae* pour illustrer la densité et la complexité cellulaire. La quantité et la diversité des entités cellulaires soulignent cette complexité. La densité de la chromatine est illustrée dans l'encadré **(A)** : si, toutes proportions gardées, le génome de *S. cerevisiae* mesurait 300m de long (la hauteur de la tour Eiffel), il serait stocké dans l'équivalent d'un grain de sable, une sphère de diamètre 0.1 mm. Cette compaction est rendue possible par le diamètre de la molécule d'ADN (2 nm) et la structure de la chromatine, visible dans le zoom x1 000 par rapport à l'échelle de la cellule (illustration de la chromatine par Goodsell). La diversité du protéome est illustrée dans l'encadré **(B)** : s'il est possible de visualiser 480 protéines à la fois, il est difficile de visualiser les 42 millions de copies de protéines qui constituent le protéome de *S. cerevisiae*. Si l'on était capable de visualiser une protéine par seconde, on pourrait visualiser les 480 protéines en 8 min seulement. Pour visualiser le protéome en entier, il faudrait attendre 16 mois. Le protéome est également extrêmement dense (illustration du protéome par (McGuffee & Elcock, 2010)), avec certaines protéines comme la fructose-bisphosphate aldolase présentes en million d'exemplaires par cellule.

Le génome est un autre bon exemple de la complexité cellulaire, ne serait-ce que par sa topologie (voir **Figure 1**). Un polymère d'environ 4 mètre (pour *S. cerevisiae* (Dickerson et al., 1982)) est compacté dans le volume micrométrique du noyau (2 µm pour *S. cerevisiae* (Thelen et al., 2021)) chez les eucaryotes, et il contient dans sa séquence presque (à l'exception de l'ADN mitochondriale) toute l'information génétique d'une espèce donnée. La chromatine (formée de l'ADN et des protéines histones) dans le noyau est donc hautement compactée suivant des motifs imbriqués : des compartiments, eux-mêmes formés de domaines qui rassemblent plusieurs boucles (voir (Misteli, 2020) pour une revue). Comme détaillé par Misteli et al., ces motifs sont de plus de puissants modulateurs de l'activité du génome, ce qui illustre que la complexité des fonctions est inscrite dans la complexité des structures. Son aspect dynamique est tout autant vertigineux, comme l'illustre bien la combinatoire de l'épissage d'un gène (voir (Marasco & Kornblihtt, 2023; Shi, 2017) pour des revues). Un génome contient des dizaines de milliers de gènes dont les exons peuvent être alternativement inclus ou exclus dans différentes combinaisons, sachant qu'un gène eucaryote contient en moyenne 9 exons. L'exon skipping (évitage d'exon) augmente encore cette modularité et à l'épissage s'ajoute les modifications post traductionnelles : glycosylation, phosphorylation, acétylation, etc.

Ces exemples ne représentent qu'une fraction de la complexité cellulaire. On peut citer par exemple l'importance des marques épigénétiques, qui sera illustrée ensuite ici, ou le rôle de la séparation de phases liquides dans l'organisation du nucléole et du noyau (Farr et al., 2021; Feric et al., 2016; Strom & Brangwynne, 2019).

1.1.2 La nécessité des espèces modèles dans la compréhension des systèmes biologiques complexes

Appréhender les systèmes biologiques complexes à travers l'étude de modèles simplifiés est donc une nécessité. La focalisation sur des systèmes donnés permet de réduire la dimension, de créer un point de départ à l'exploration. Comme dans de nombreux domaines de la biologie, l'adoption d'espèces modèles en biologie cellulaire possède ainsi une forte dimension historique et est liée aux limites des méthodes de mesure disponibles (voir (Müller & Grossniklaus, 2010) pour une revue).

Au 19e siècle, en parallèle du travail de catalogage sans a priori des espèces réalisé par Darwin (qui aboutira à la théorie de l'évolution, modèle du mécanisme de l'évolution), Mendel choisit d'étudier en particulier le pois (*Pisum sativum*). Il déduisit de ses observations des lois de l'héritabilité, créant ainsi les bases de la génétique moderne (Gayon, 2016). En partant d'une espèce facilement cultivable avec des phénotypes simples et visibles (couleurs et aspects des pois, voir **Figure 2**), Mendel pu élaborer un modèle passé initialement inaperçu mais qui sera ensuite redécouvert et développé grâce aux progrès de la microscopie (Gayon, 2016). Au début du 20e siècle, Morgan travaille sur la drosophile (*Drosophila melanogaster*) et son équipe est la première à associer un gène spécifique à un chromosome particulier (voir **Figure 2**), posant ainsi

les bases de la génétique des populations (Huxley, 1924). Les drosophiles peuvent facilement être étudiées sur plusieurs générations (centaines d'individus en 15 jours), ont un dimorphisme sexuel reconnaissable et des chromosomes visibles au microscope optique. Dans les années 1920 et 1930, le maïs (*Zea mays*) est devenu un organisme modèle important pour l'étude de la génétique des plantes grâce aux travaux pionniers de McClintock. Cette chercheuse a découvert les éléments transposables dans le maïs (McClintock, 1956), mettant en évidence la plasticité du génome et des mécanismes de régulation génétique des années avant que le processus moléculaire ne soit élucidé dans les années 60-70 chez les bactéries, phages et levures grâce aux techniques modernes de biologie moléculaire (McClintock, 1961). Ses conclusions sur les transposons viennent encore une fois d'observation de traits visibles des graines de maïs (voir **Figure 2**).

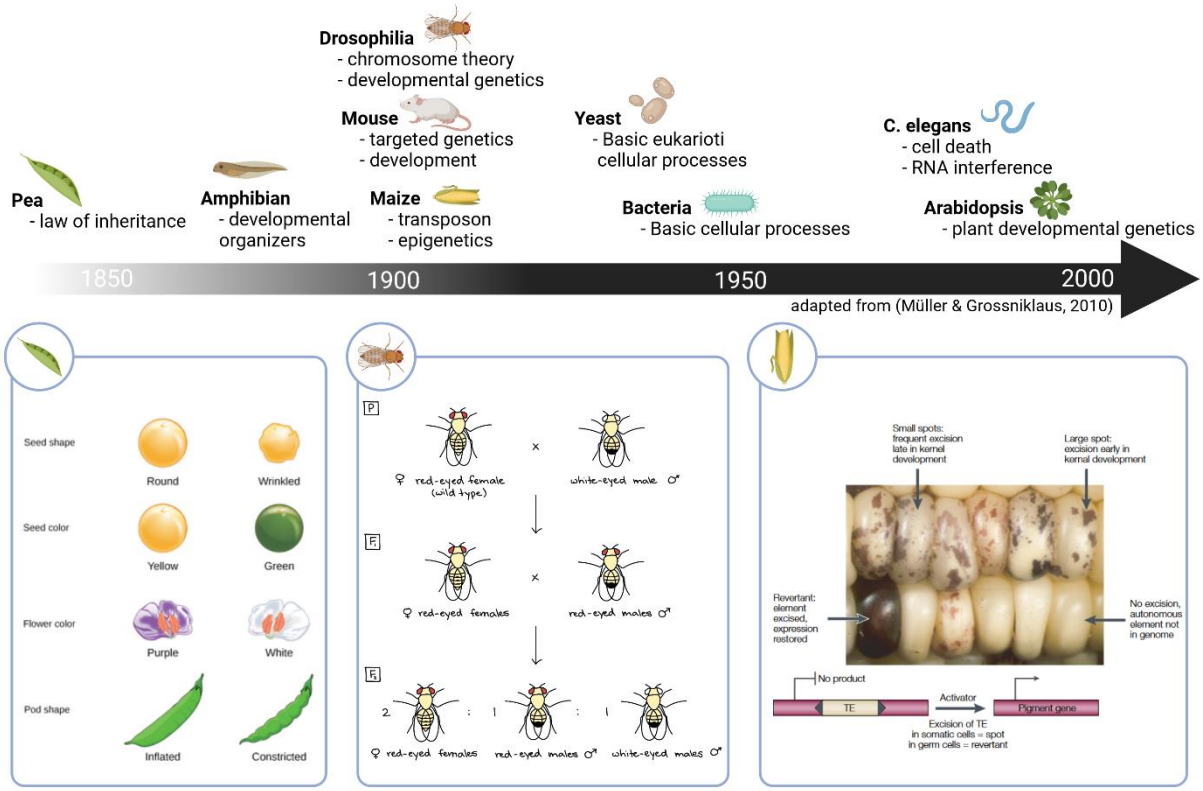


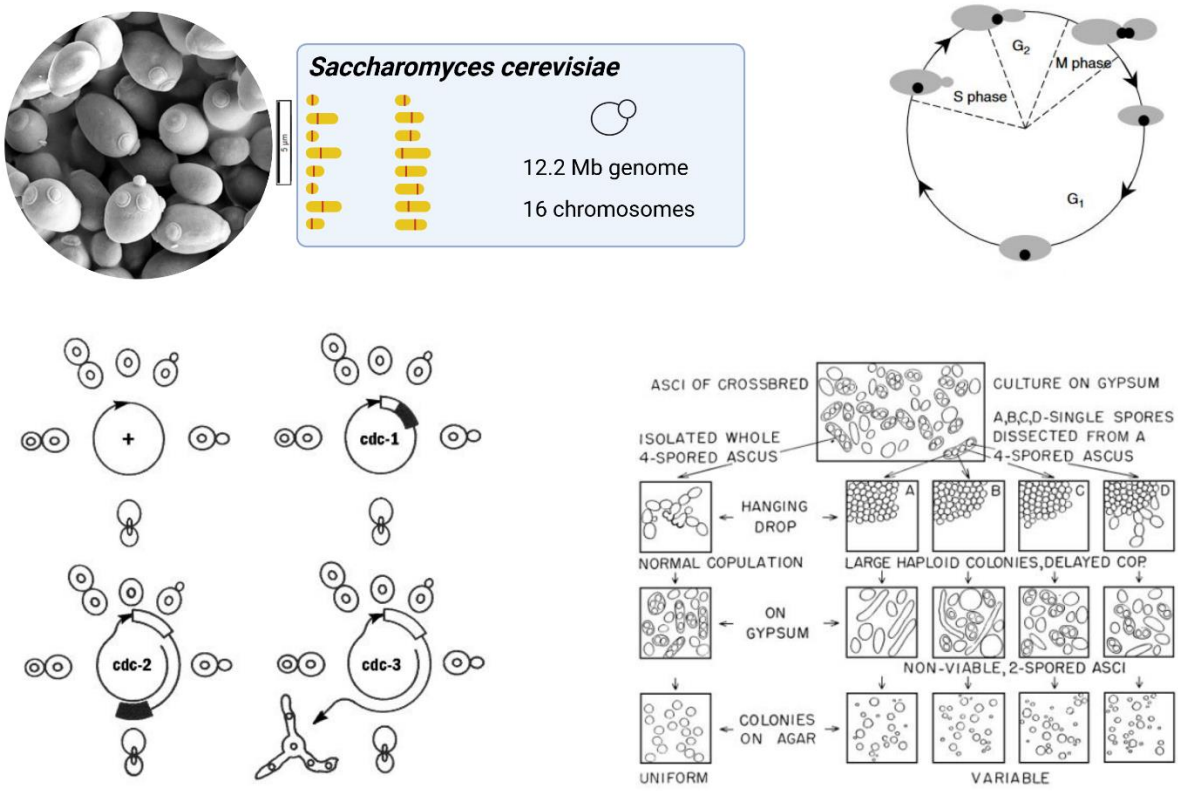
Figure 2 : Exemples historiques d'espèces modèles de la biologie cellulaire. Chaque exemple de la frise chronologique est associé aux domaines pour lesquels il a été déterminant. L'encadré de gauche illustre les phénotypes utilisés par Mendel pour élaborer son modèle de l'hérédité. L'encadré central schématise l'hérédité de la mutation « white eyes » étudié par Morgan dans son modèle de l'hérédité génétique (Morgan, 1910). L'encadré de droite (issu de (Feschotte et al., 2002) illustre le phénotype des grains de maïs étudié par McClintock dans son modèle de l'interaction entre gènes et transposons (McClintock, 1956).

Les trois exemples de la **Figure 2** illustrent comment un choix judicieux de modèle permet de conceptualiser des processus dont les bases moléculaires dans le système cellulaire sont pourtant encore largement invisibles à l'époque. C'est le principe même de la modélisation : créer une représentation de la réalité (dans la limite des connaissances disponible) qui facilite la génération de nouvelles hypothèses. On peut noter dès à présent l'importance de la question de la visualisation, sur laquelle nous reviendrons plus loin.

Les modèles unicellulaires ont dans la même logique joué un rôle prédominant dans la compréhension du système cellulaire. La facilité de culture et la relative simplicité du fonctionnement unicellulaire de la bactérie *Escherichia coli* (*E. coli)* et de la levure *Saccharomyces cerevisiae* en ont fait des espèces modèles incontournables (Müller & Grossniklaus, 2010). Au milieu du 20e siècle, *S. cerevisiae* et deux autres champignons : la levure *Schizosaccharomyces pombe* et le champignon filamenteux *Neurospora crassa* sont devenus des espèces modèles clés pour l'étude de la biologie cellulaire et moléculaire. Ce travail de thèse repose sur l'étude de ces trois modèles du fait de leur intérêt scientifique, de l'expertise de l'environnement de recherche ainsi que de l'intérêt et du nombre des jeux de données publiques disponibles.

1.1.3 Modéliser le système cellulaire, l'exemple de trois champignons

1.1.3.1 La levure S. cerevisiae, espèce modèle de référence pour les eucaryotes



1970 Hartwell diagram on cdc mutants

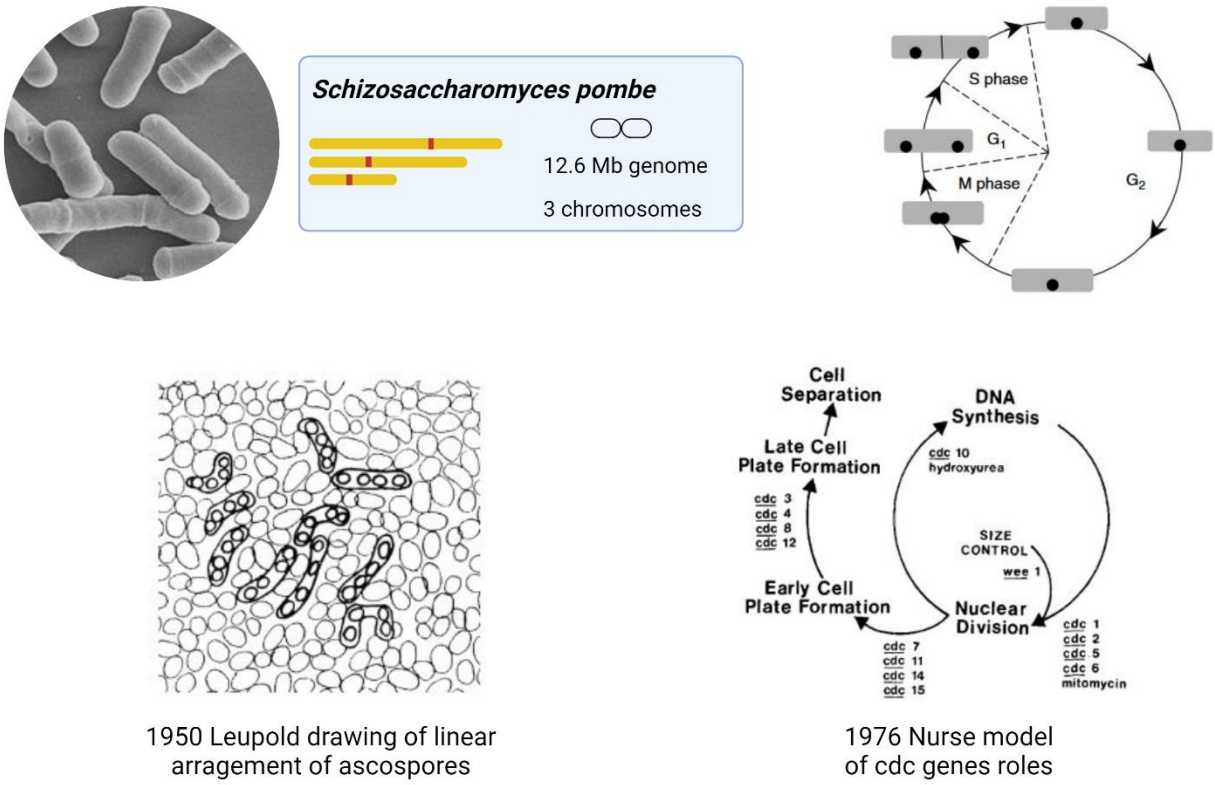
1943 Lindegren model illustration

Figure 3 : *Saccharomyces cerevisiae*. Photo de microscopie électronique à balayage (encadré rond, Wikipédia). Le génome de cette levure mesure 12.2 Mb et comporte 16 chromosomes. Son cycle cellulaire est illustré en haut à gauche. *S. cerevisiae* est une espèce modèle historique phare, étudiée notamment pour la compréhension de mécanismes du cycle cellulaire comme le contrôle génétique par les gènes cdc (schéma à gauche, (Hartwell et al., 1970)). L'hétérothallisme, i. e. l'existence de type sexuel chez les champignons a également été mis en évidence chez *S. cerevisiae* (schéma à droite, (Barnett, 2007)).

La levure *Saccharomyces cerevisiae*, largement connue pour ses propriétés fermentaires dans la production de pain, de bières et de vins, a également acquis une renommée en tant que modèle biologique pour l'étude de la cellule eucaryote dès le début du 20ème siècle (voir **Figure 3**) (Müller & Grossniklaus, 2010) (voir (Barnett, 2007) pour une revue). Cette espèce de levure présente de nombreuses caractéristiques cellulaires et génétiques communes avec les organismes multicellulaires, tout en offrant les avantages expérimentaux d'une culture facile et d'un cycle de vie rapide. *S. cerevisiae* est capable de se développer par bourgeonnement ou par reproduction sexuée entre deux types sexuels alpha et bêta. Son génome haploïde est dupliqué (Wolfe, 2015) et comprend 12,2 Mb répartis en 16 chromosomes (voir **Figure 3**). Premier génome eucaryote à être séquencé en 1996 (Goffeau et al., 1996), de nombreux mécanismes cellulaires fondamentaux ont pu être mis en évidence chez ce modèle (Engel et al.,

2014; Müller & Grossniklaus, 2010). Pour donner deux exemples, l'étude du cycle cellulaire chez cette levure a permis notamment d'en décrypter les mécanismes de régulation (Hartwell et al., 1970; Hartwell & Weinert, 1989) et *S. cerevisiae* a également joué un rôle crucial dans la compréhension des mécanismes de réponse au stress cellulaire (Barnett, 2007). Les similitudes de la structure tri-dimensionnelle (3D) des génomes des levures et des métazoaires (Torres et al., 2023) ont également érigé *S. cerevisiae* comme modèle d'étude du génome en 3D (voir (Noma, 2017; Torres et al., 2023) pour revues). Les 16 chromosomes sont organisés dans une conformation Rabl, où les centromères et les télomères sont réunis en clusters localisés à des points opposés de l'enveloppe nucléaire. La cohésine et la condensine ont été identifiées comme des acteurs clefs du fonctionnement de la chromatine de *S. cerevisiae*. Ces complexes protéiques en anneaux structurent les chromatines sœurs, maintiennent les boucles et domaines de la chromatine et organisent l'évolution de sa compaction au cours du cycle cellulaire (Hoencamp & Rowland, 2023; Oldenkamp & Rowland, 2022).

1.1.3.2 La levure S. pombe, un autre modèle unicellulaire complémentaire



1950 Leupold drawing of linear arragement of ascospores

1976 Nurse model of cdc genes roles

Figure 4 : *Schizosaccharomyces pombe*. Photo de microscopie électronique à balayage (encadré rond, Wikipédia). Le génome de cette levure mesure 12.6 Mb et comporte 3 chromosomes. Son cycle cellulaire est illustré en haut à gauche. *S. pombe* est la seconde espèce modèle de levure emblématique, étudiée également pour la compréhension de mécanismes du cycle cellulaire comme le contrôle génétique par les gènes cdc (schéma à droite, (Nurse et al., 1976)). Le dessin de gauche illustre l'arrangement linéaire des ascospores de *S. pombe*, utilisé par Leupold pour étudier la polyploïdie.

Schizosaccharomyces pombe est le sixième organisme dont le génome ait été séquencé (Yanagida, 2005). Il est constitué de trois chromosomes haploïdes de 12,6 Mb au total (voir **Figure 4**). Les cellules de *S. pombe* sont de forme cylindrique (~3,5 µm de diamètre) et se divisent symétriquement en mitose. Elles peuvent également se développer par reproduction sexuée entre types sexuels P et M en condition de stress. Cette levure a joué un rôle complémentaire de celui de *S. cerevisiae* comme espèce modèle phare en biologie cellulaire (Hoffman et al., 2015; Müller & Grossniklaus, 2010; Yanagida, 2005). En effet, *S. pombe* a divergé évolutivement de *S. cerevisiae* il y a environ 350 millions d'années. Depuis *S. cerevisiae* a perdu de nombreux gènes (338) qui ont été conservés chez *S. pombe* et les mammifères (Hoffman et al., 2015) et a subi une duplication de son génome. Ces deux levures sont donc des modèles complémentaires de la cellule eucaryote, à la fois des organismes unicellulaires simples à cultiver et à étudier, tout en étant évolutivement éloignés. Les mécanismes cellulaires fondamentaux qui ont été mis en évidence chez *S. cerevisiae* ont également été étudié en parallèle chez *S. pombe*, comme par exemple le contrôle du cycle cellulaire (voir **Figures 3 et 4**) (Nurse et al., 1976). Pour citer un autre exemple, l'étude de *S. pombe* a été déterminante pour l'élaboration du « code des histones », c'est-à-dire la description des marquages épigénétiques comme d'un système de régulation de la chromatine (Jenuwein & Allis, 2001). À noter que Jenuwein et al. précisaient ici que *S. cerevisiae* n'était pas un modèle adapté étant donné l'absence de certains mécanismes comme l'inhibition de l'expression génique par la protéine HP1 ; ce qui illustre de nouveau la complémentarité de ces espèces modèles. La structure du génome de *S. pombe* est similaire à celle de *S. cerevisiae* décrite précédemment (Noma, 2017; Torres et al., 2023). Les trois chromosomes sont organisés en une conformation Rabl et la dynamique de la chromatine est structurée par la cohésine et la condensine (Kakui et al., 2020; Kim, 2021) au long du cycle cellulaire (Tanizawa et al., 2017).

1.1.3.3 Le champignon filamenteux N. crassa, espèce modèle multicellulaire

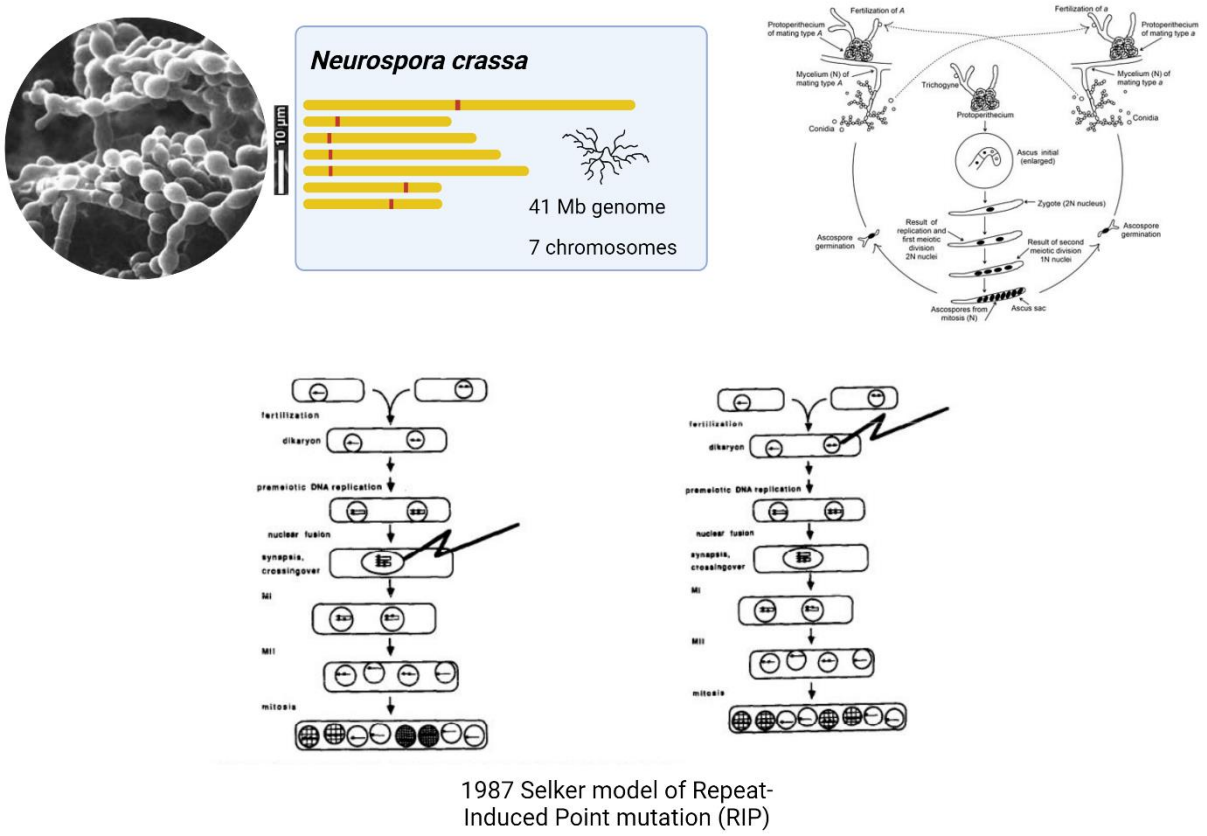


Figure 5 : *Neurospora crassa*. Photo de microscopie électronique à balayage (encadré rond, The ASM Microbe Library). Le génome de *N. Crassa* est de taille 41 Mb et comporte 7 chromosomes. Son cycle cellulaire est décrit en haut à droite. Cette espèce modèle a notamment permis la découverte du RIP (Repeat-Induced Point Mutation).

Neurospora crassa est un champignon filamenteux. Sa taille moyenne varie de 100 à 500 mm de longueur, avec des cellules individuelles atteignant une longueur d'environ 10 µm (voir **Figure 5**). *N. crassa* possède un cycle de vie majoritairement haploïde et un taux de division cellulaire rapide. Ces caractéristiques en ont fait un organisme multicellulaire modèle, dont l'étude est à l'origine de l'hypothèse de Beadle and Tatum « un gène-une protéine », proposant que chaque gène code pour une protéine spécifique (Gayon, 2016; Müller & Grossniklaus, 2010). Le génome de *N. crassa*, séquencé en 2003 (Galagan et al., 2003), est plus grand que celui des deux levures présentées précédemment : 41 Mb divisé en 7 chromosomes (voir **Figure 5**). Il est en revanche lui aussi structuré dans le noyau selon une conformation Rabl, mise en évidence par microscopie à fluorescence (Klocko et al., 2016; Torres et al., 2023). *N. crassa* présente des caractéristiques épigénétiques dont sont dépourvues *S. cerevisiae* et *S. pombe* comme la méthylation de l'ADN et le système de méthylation H3K27. Cela en a fait un modèle de l'étude épigénétique qui a notamment permis l'identification du RIP («repeat-induced point mutation »), le premier système de défense du génome dépendant de l'homologie découvert chez les eucaryotes (voir **Figure 5**) (Aramayo & Selker, 2013).

1.2 Le changement d'échelle provoqué par l'évolution des techniques de mesure a changé le paradigme de la modélisation du système cellulaire

1.2.1 L'émergence de la "big data" en biologie : nouveau paradigme et conséquences

1.2.1.1 Les données omiques, changement de l'échelle d'étude

Comme illustré précédemment, les méthodes de mesure disponibles limitent la démarche de modélisation : les informations non détectées expérimentalement sont absentes du modèle final, dont la précision ne dépasse donc pas celle de la méthode de mesure utilisée. Mais l'évolution rapide des techniques expérimentales de biologie cellulaire depuis les années 90 a permis une accélération de l'accumulation de connaissance et a ainsi créé un défi interdisciplinaire entre biologie et informatique (voir « *Working with Omics Data: An Interdisciplinary Challenge at the Crossroads of Biology and Computer Science*" en annexe). La terminologie des omiques a commencé à être utilisée régulièrement dans les années 2000, chaque domaine omique correspondant à l'étude d'une des familles de molécules citées plus haut : génomique, transcriptomique, protéomique et métabolomique (voir (Arivaradarajan & Misra, 2018) pour une revue). Bien que l'évolution des techniques d'analyse en spectrométrie de masse ait eu un effet considérable en protéomique et métabolomique, nous nous concentrerons ici sur les applications du séquençage haut débit d'acides nucléiques pour illustrer le changement de paradigme causé par les approches omiques.

La génomique a débuté dès 1977 avec l'application de la méthode de séquençage sur gel développée par Sanger, pour séquencer pour la première fois le génome entier d'un virus : le phage phiX. Treize ans plus tard, en 1990, le « human genome project » débute avec pour but de séquencer les 3 milliards de bases du génome humain, en utilisant le séquençage capillaire (Karger & Guttman, 2009). Plus de dix ans et près de 3 milliards de dollars plus tard, cette tâche titanesque a été accomplie (International Human Genome Sequencing Consortium, 2004). La technologie des puces à ADN reste également emblématique du développement de la génomique (Schena et al., 1995). Dans les années 2000, elles représentaient la clé de voûte d'une discipline alors appelée "post-génomique" (Souciet, 2011). Derrière cette terminologie, l'idée était qu'une fois les génomes entièrement séquencés, de nouvelles études pourraient être menées pour comprendre leur fonctionnement. Les puces à ADN sont alors apparues comme un outil prometteur pour suivre l'expression des gènes, notamment chez les levures (Eisen et al., 1998; Gasch et al., 2000; Rustici et al., 2004; Spellman et al., 1998) . À partir de 2007, de nouvelles méthodes appelées "séquençage de nouvelle génération" (NGS) ont permis de réduire considérablement le coût, les difficultés techniques et la durée du processus de séquençage (Metzker, 2010). La **Figure 6** illustre le changement de paradigme entre la méthode Sanger et la méthode Illumina.

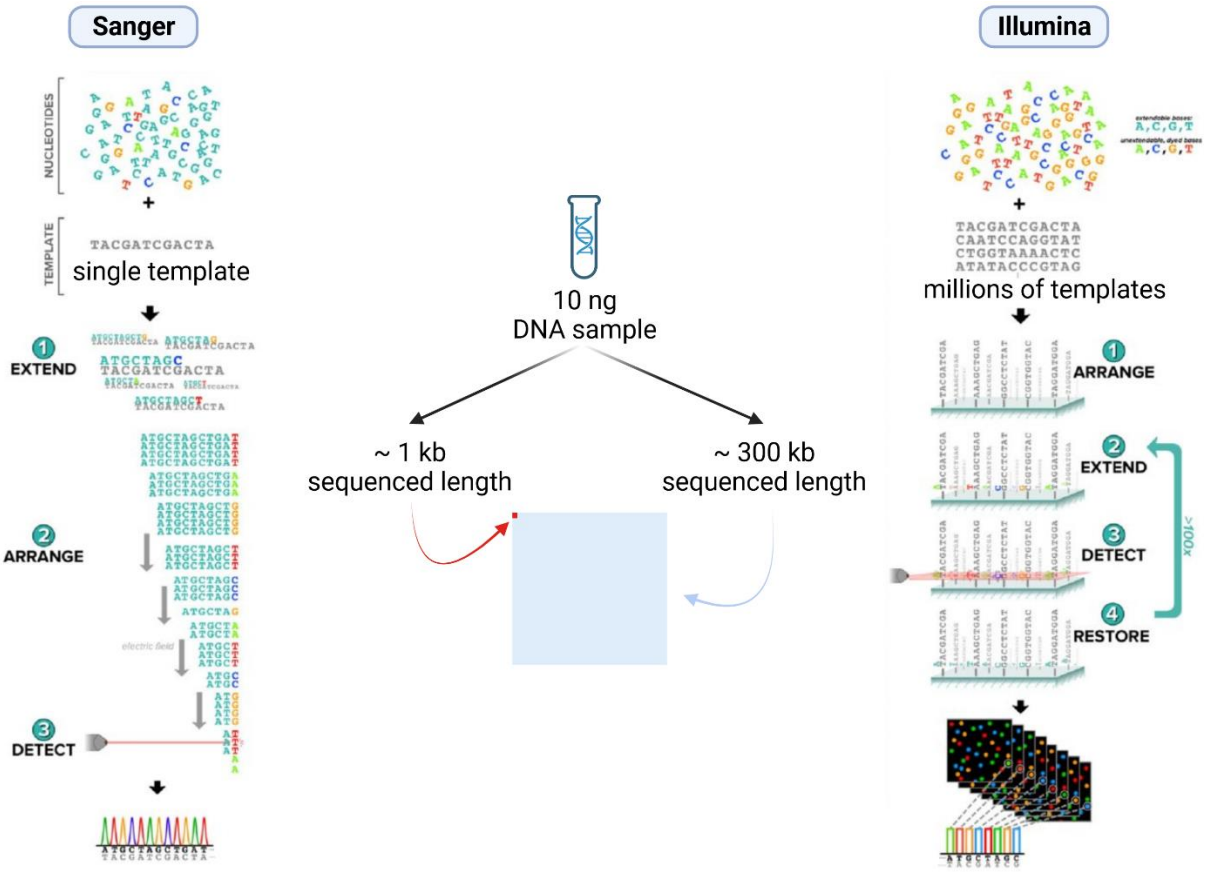


Figure 6 : Illustration du changement d'échelle des méthodes de séquençage de nouvelle génération. Le processus de séquençage Sanger est schématisé à gauche et celui d'Illumina est illustré à droite (adaptés de (Muzzey et al., 2015)). L'atout majeur apporté par le séquençage Illumina est l'introduction de la parallélisation du séquençage qui permet d'augmenter considérablement la taille des séquences identifiées. Cette idée est illustrée au centre de la figure par des carrés à l'échelle des tailles de séquence obtenues par Sanger ou Illumina à partir de 10 ng d'ADN (valeurs issues d'illumina.com).

Le séquençage haut débit a ainsi rendu accessible les séquences de nombreux génomes, ouvrant la voie à la génomique comparée (Souciet, 2011). En transcriptomique, les NGS ont permis d'évaluer les transcrits sans apriori comme avec les sondes des puces à ADN, mais à l'échelle du génome entier grâce aux RNAseq (*RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis*, 2019). Le séquençage des génomes entier a également ouvert la voie à des méthodes d'étude de leurs structures 3D (voir (Jerkovic´ & Cavalli, 2021) pour une revue). La méthode de « chromatin conformation capture » ou 3C a permis dès les année 2000 de détecter la fréquence d'interaction entre deux loci génomiques donnés (Dekker et al., 2002). Mais c'est la technique du Hi-C («High-throughput Chromatin Conformation Capture ») qui est rapidement devenue la méthode d'étude prédominante de la structure 3D des génomes (Lieberman-Aiden et al., 2009). Suivant le même principe que la 3C, elle tire profit du séquençage haut débit pour mesurer les fréquences d'interaction non plus entre deux loci mais à l'échelle du génome entier. Expérimentalement, les régions de la chromatine proches dans l'espace sont fixées entre elles par le formaldéhyde puis

ligaturées et marquées à la biotine après digestion enzymatique. Ces séquences chimériques sont ensuite isolées par différentes étapes (Jerkovic´ & Cavalli, 2021) et c'est leur séquençage « pair end » haut débit qui permet de générer une carte de fréquence de contact à l'échelle du génome. Chaque séquence chimérique détectée contenant deux régions données de la chromatine devient une indication de « contact ». Cette méthode a notamment permis de décrire les caractéristiques de la structure 3D des génomes de *S. cerevisiae*, *S. pombe* et *N. crassa* décrites plus haut (Torres et al., 2023).

Ces applications du séquençage haut débit illustrent le changement d'échelle provoqué par l'évolution récente des techniques de mesures expérimentales de biologie cellulaire. Pourtant la génomique comparée, les analyses RNA-seq et les analyses Hi-C ne sont qu'une fraction des possibilités ouvertes par les données omiques toujours plus nombreuses. Le stockage et le partage de ces données brutes omiques sont donc des enjeux importants pour leur utilisation.

1.2.1.2 Comment stocker et partager ce déluge de données ?

L'ensemble des bases de données biologiques publiques hébergeant des données omiques est très vaste et en constante évolution. Entre 1991 et 2016, 1 727 bases de données uniques de biologie moléculaire ont été créées, avec en moyenne 104 bases de données créées par an. Le projet ELIXIR a été lancé en 2013 avec pour but d'unifier tous les centres européens et les ressources bio-informatiques en une infrastructure unique et coordonnée (Harrow et al., 2021). Ce projet produit notamment les ELIXIR Core Data Resources (créées en 2017), un ensemble de bases de données européennes sélectionnées, répondant à des exigences définies. Les bases de données du National Center for Biotechnology Information (NCBI) aux États-Unis sont également des références. Étant donné la nature "brute" des ensembles de données omiques, ils sont stockés dans des dépôts de données d'archives : données brutes provenant d'articles scientifiques, partagées sur des bases de données facilement accessibles à des fins de reproductibilité. À l'exception de Sequence Read Archive (SRA), les bases de données citées ici sont mixtes : Elles hébergent des données de séquençage brutes et des données secondaires résultants de leurs analyses. Pour les données génomiques, la base de données NCBI Genome (Benson et al., 2010) et la base de données européenne EMBL-EBI (membre d'ELIXIR) Ensembl (Howe et al., 2021) sont des références. Elles organisent les séquences génomiques avec des annotations et proposent des outils de comparaison de séquences et d'exploration visuelle. Les données transcriptomiques peuvent être déposées dans plusieurs bases de données, comme Gene Expression Omnibus (GEO) (Barrett et al., 2012) qui est structurée en échantillons formant des ensembles de données. La Sequence Read Archive (SRA) (Leinonen et al., 2011) accepte les données brutes de séquençage. PRIDE (Perez-Riverol et al., 2019) est une base de données de référence pour les données protéomiques issues d'expérience de spectrométrie de masse.

Ces quelques exemples de bases de données sont des références généralistes incontournables, mais il existe de nombreuses bases de données plus spécialisées. 89 nouvelles bases de données sont listées dans le numéro 2021 de « NAR database issue » et une douzaine d'entre elles sont spécifiques aux omiques (Rigden & Fernández, 2021). Des efforts constants sont déployés pour relier les différentes références et citations d'acteurs biologiques (gènes, protéines, métabolites) à travers la diversité des bases de données. Chacune d'entre elles représente des pétaoctets d'informations biologiques (43 000 téraoctets de données de séquençage rien que pour SRA), et l'échelle du réseau qu'elles forment par recoupement est difficile à conceptualiser. L'approche omique a considérablement augmenté la part mesurable de la complexité des systèmes biologiques. Cette complexité se retrouve donc naturellement dans le nombre, la taille et la diversité des résultats générés.

1.2.2 La recherche menée par l'hypothèse, par les données ou par le modèle

La recherche moderne est menée selon la méthode scientifique : l'observation permet l'élaboration d'une hypothèse, qui est ensuite testée par l'expérience. L'analyse des résultats expérimentaux permet d'accepter, de rejeter ou d'affiner l'hypothèse. Chaque nouvelle hypothèse validée permet de préciser un peu plus le modèle du système étudié. Les études des modèles historiques citées précédemment en sont de bons exemples. Les expériences sur les mutants de *N. crassa* ont permis à Beadle et Tatum de proposer le modèle « un gène-une protéine ». Ce modèle a par la suite été progressivement complexifié pour prendre en compte notamment les résultats sur l'épissage alternatif et les ARN non codants.

Ce paradigme de recherche a été remis en question par le développement des technologies haut débit et le déluge de données –omiques qu'elles ont provoqué. Le changement d'échelle a rendu possible l'exploration sans apriori des données pour trouver des motifs, des associations et des relations qui n'avaient pas été envisagés. C'est le principe de recherche par les données (« data driven »). Par exemple un réseau de régulation contenant uniquement des facteurs de transcription et des gènes a permis à Alon et al. de créer le concept de 'motifs' dans un réseau de régulation biologique (Alon, 2007), cité depuis plus de 2000 fois en métabolomique et biologie synthétique depuis.

Bien que pouvant apparaître comme contradictoire avec une logique de recherche par l'hypothèse, ce nouveau paradigme de recherche ne concerne finalement que l'étape « expérience » de la méthode scientifique : l'augmentation de la quantité d'information générée n'implique pas de ne plus tester d'hypothèses. Au contraire, plus il existe d'informations (intégrées et cohérentes) sur un système biologique, plus il est possible de poser des questions complexes et d'y répondre de manière précise et intégrée. C'est la démarche holistique de la biologie des systèmes et des analyses multiomiques (Hasin et al., 2017; Subramanian et al., 2020; Veenstra, 2021). Mais la complexité des données omiques détaillée plus haut implique un enjeu de modélisation à l'étape d'analyse des

résultats (Smalheiser, 2002; Veenstra, 2021). Comment créer une représentation de la réalité qui facilite la génération de nouvelles hypothèses à partir de données à larges échelles et de natures variées ?

1.2.3 Le réseau comme modèle à large échelle ?

1.2.3.1 Un système biologique est un immense réseau d'interactions

Les réseaux sont un outil puissant pour modéliser un grand nombre d'acteurs densément interconnectés comme le sont les acteurs moléculaires du fonctionnement cellulaire (voir (Barabási & Oltvai, 2004) pour une revue). Les données omiques donnent accès à leurs interactions à large échelle, il est donc par exemple possible de construire des réseaux d'interactions protéines-protéines (Alberghina et al., 2012; Ekman et al., 2006), des voies métaboliques ou bien des réseaux de régulation transcriptionnelle (Monteiro et al., 2020). Les données Hi-C présentées précédemment peuvent également être représentées comme un réseau de contacts (Ye et al., 2020). Dans tous ces exemples les acteurs moléculaires sont modélisés par des nœuds, reliés entre eux par des arêtes représentant les interactions étudiées. Cette modélisation permet d'identifier et de quantifier les caractéristiques du réseau (densité, taille, connectivité, robustesse, etc.) et de caractériser les sous-réseaux, les clusters et les motifs présents (Barabási & Oltvai, 2004). Des outils bio informatiques spécifiquement adaptés à la représentation et surtout à l'exploration des réseaux biologiques ont été développés pour tirer profit des possibilités qu'offre la théorie des réseaux. Certains sont spécifiques à des types de réseaux particuliers comme Escher (King et al., 2015) ou bien MicrobiomeAnalyst (Lu et al., 2023) ; d'autres sont modulaires comme la référence Cytoscape (Franz et al., 2015) ou encore NetworkAnalyst (Zhou et al., 2019). L'utilisation fréquente de ces outils par la communauté (Les articles de MétaboAnalyst (Pang et al., 2021) et NetworkAnalyst (Zhou et al., 2019) sont cités plus de 1000 fois chacun d'après GoogleScholar) illustre l'efficacité de la modélisation en réseaux pour l'analyse des données omiques. L'efficacité des concepts abstraits de nœud et d'arête permet de structurer la complexité des données omiques, la rendant ainsi analysable à large échelle.

1.2.3.2 L'abstraction en réseau du système cellulaire limite la visualisation

L'abstraction des acteurs moléculaires dans un réseau permet de faciliter l'analyse, mais la lisibilité du modèle reste une limite. Le grand nombre d'acteurs étudiés (et donc de nœuds et d'arêtes) provoque rapidement l'effet « hairball » (Zhou & Xia, 2018) (voir **Figure 7A**) qui rend difficile l'exploration sans a priori du modèle. Sans une liste de gènes ou bien un motif cible, un réseau de régulations transcriptionnelles peut être difficile à lire dans son ensemble. Les outils présentés précédemment proposent de nombreuses solutions : Cytoscape offre par exemple de nombreuses options d'affichage, de filtrage et d'annotation qui permettent de clarifier les réseaux générés. OmicsNet propose même une visualisation dans un espace 3D, séparant les différents acteurs omiques du réseau en différents plans (Zhou & Xia, 2018) (voir **Figure 7B**).

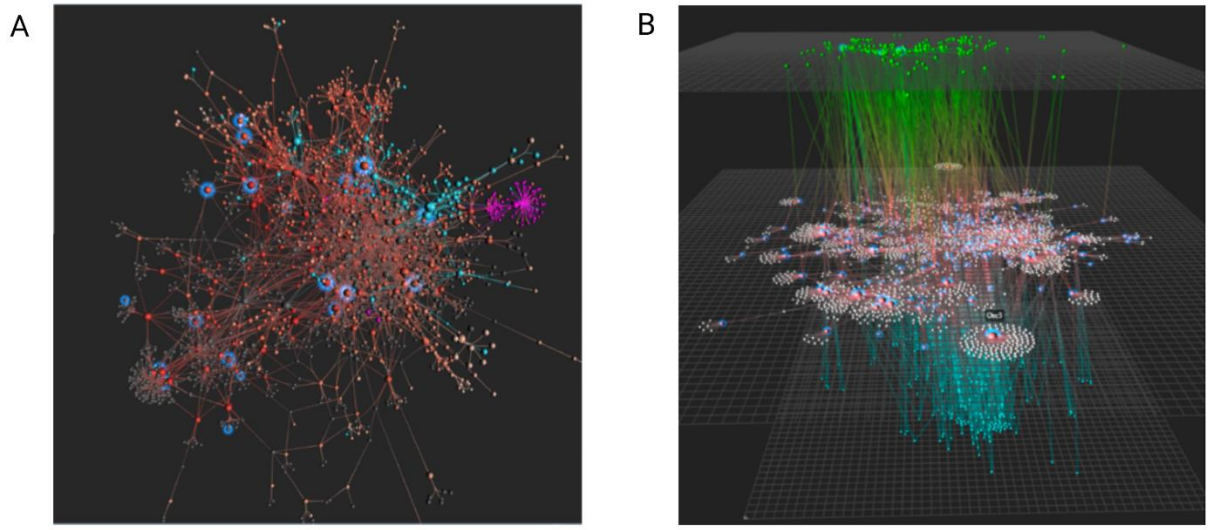


Figure 7 : L'échelle des réseaux biologiques limite leur visualisation en réseau. Extrait de la figure 2 de (Zhou & Xia, 2018). (**A)** L'effet « hairball » est visible sur ce réseau d'interaction protéines-protéines composé de ~ 2 000 nœuds et ~ 4 000 arêtes. Pour y remédier, l'outil OmicsNet développé par les auteurs propose des solutions de mise en valeur par coloration et surbrillance. **(B)** Un autre mode du logiciel permet de structurer un réseau en plans 2D correspondant aux différents types d'acteurs de ce réseau : gènes, facteurs de transcription et miRNAs. Ce choix de visualisation illustre l'intérêt de compenser le haut niveau d'abstraction de la représentation en nœuds et arêtes en associant de l'information à la position des nœuds.

Ces solutions mettent en valeur que la limite de lisibilité des réseaux vient de la perte d'information qu'implique nécessairement une modélisation abstraite, en nœud et en arête. Ainsi, en replaçant dans des plans différents les nœuds d'un réseau multiomique, Zhou et al. ajoute un sens biologique à la position des nœuds du réseau. Malgré tout, les possibilités de visualisation offertes par un réseaux reste limitées par son haut niveau d'abstraction comparé à la complexité de la réalité cellulaire. La question de la lisibilité de la visualisation peut paraitre secondaire mais est en réalité un enjeu complexe ; non seulement pour les réseaux mais aussi pour les autres méthodes de modélisation des données omiques existantes (Gehlenborg et al., 2010).

1.3 L'importance de la visualisation dans la modélisation large échelle de systèmes complexes

1.3.1 L'importance de la visualisation dans la méthode scientifique et la modélisation

Nous avons souligné plus tôt comment l'observation expérimentale directe a joué un rôle dans le choix des espèces modèles historiques comme la drosophile et le maïs (voir **Figure 2**). Mais au-delà de l'observation expérimentale, la visualisation est également fondamentale à l'étape de l'analyse des données. Intégrer les informations extraites des données brutes dans un modèle graphique permet d'en déduire des connaissances (O'Donoghue, 2021). En 1858 déjà, la statisticienne et infirmière

Nightingale créait un « Diagramme des causes de mortalité dans l'armée d'Orient »[2] pour visualiser l'ampleur des décès survenus dans l'armée anglaise à la suite de maladies évitables (en bleu), par rapport aux décès dus à des blessures ou à d'autres causes (en rouge et en noir) (voir **Figure 8**).

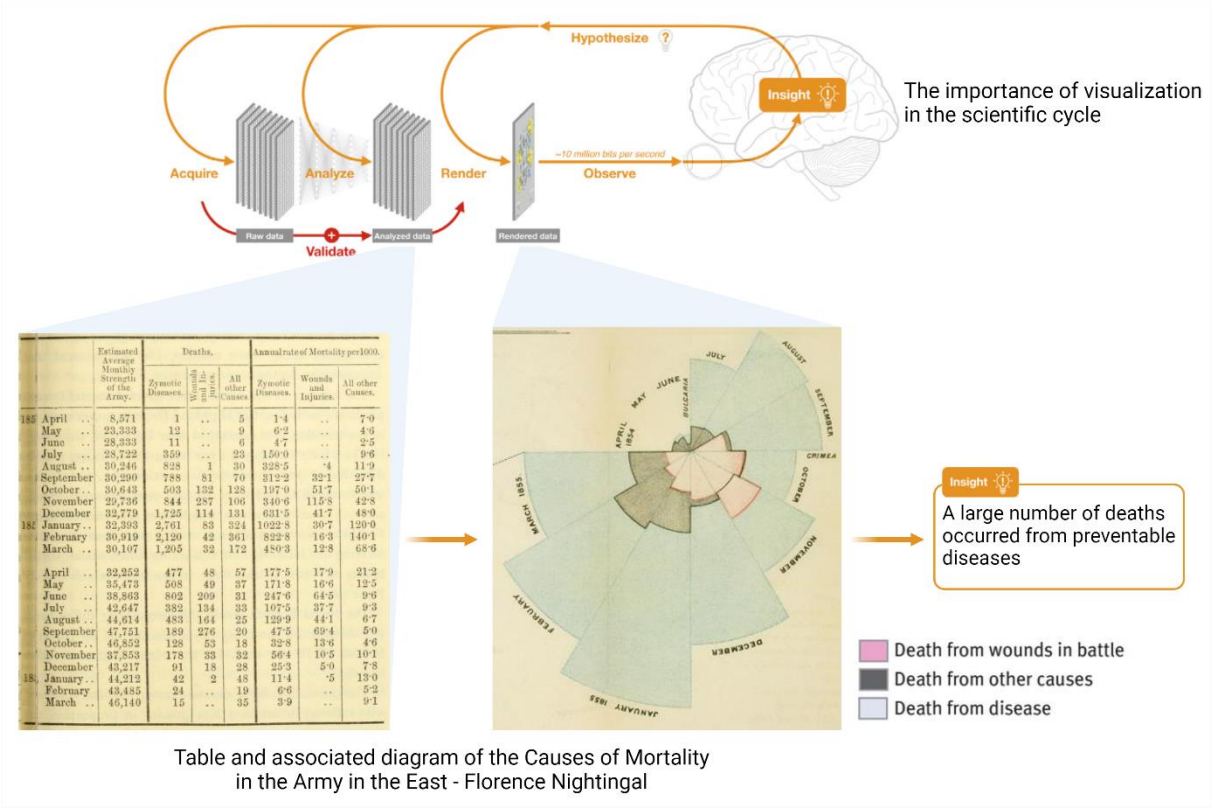
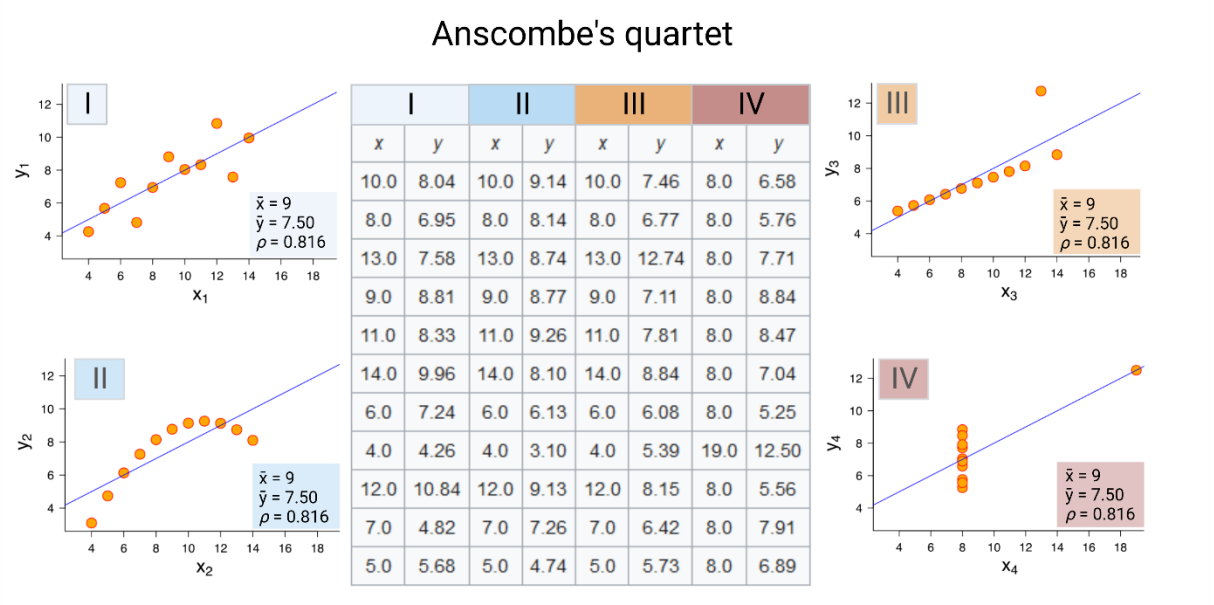


Figure 8 : Illustration de l'importance de la visualisation dans la démarche scientifique. Le cycle de la démarche scientifique, figure 1 de (O'Donoghue, 2021). L'étape de visualisation permet l'analyse, elle est donc à la base de l'élaboration de nouvelles hypothèses. Nightingale fut une pionnière de l'utilisation de la visualisation pour l'interprétation de données dans son « Notes on Matters Affecting the Health, Efficiency and Hospital Administration of the British Army » en 1858. Le diagramme (centre) des causes de mortalité qu'elle a produit rend compte d'une réalité beaucoup moins visible à la lecture des données brutes (tableau à gauche) : les maladies tuent plus les soldats anglais que les combats. Cet exemple illustre la rapidité à laquelle l'œil humain détecte un motif dans une visualisation face à un motif dans un tableau de données brutes.

Les diagrammes n'ont depuis cessé d'évoluer et de se diversifier, tirant profit des excellentes capacités humaines de reconnaissance de patterns visuels (O'Donoghue, 2021). Le Quartet d'Anscome présenté **Figure 9** illustre simplement cette idée que des représentations visuelles sont plus efficaces pour discriminer des données brutes (Matejka & Fitzmaurice, 2017).

[2] https://www.rct.uk/collection/1075240/notes-on-matters-affecting-the-health-efficiency-and-hospital-administration-of



I		II		III		IV	
x	y	x	y	x	y	x	y
10.0	8.04	10.0	9.14	10.0	7.46	8.0	6.58
8.0	6.95	8.0	8.14	8.0	6.77	8.0	5.76
13.0	7.58	13.0	8.74	13.0	12.74	8.0	7.71
9.0	8.81	9.0	8.77	9.0	7.11	8.0	8.84
11.0	8.33	11.0	9.26	11.0	7.81	8.0	8.47
14.0	9.96	14.0	8.10	14.0	8.84	8.0	7.04
6.0	7.24	6.0	6.13	6.0	6.08	8.0	5.25
4.0	4.26	4.0	3.10	4.0	5.39	19.0	12.50
12.0	10.84	12.0	9.13	12.0	8.15	8.0	5.56
7.0	4.82	7.0	7.26	7.0	6.42	8.0	7.91
5.0	5.68	5.0	4.74	5.0	5.73	8.0	6.89

Figure 9 : Le Quartet d'Anscome. Cet ensemble de données est formé de quatre jeux de données distincts (numéroté de I à IV, chacun composé de 11 paires de coordonnées (x, y)) qui produisent les mêmes statistiques (moyennes $\bar{x}$, $\bar{y}$ et corrélation ρ) tout en produisant des graphiques visuellement très différents.

1.3.2 En biologie cellulaire, l'omniprésence des dessins scientifiques et le défi de la modélisation large échelle

La visualisation est particulièrement importante dans l'étude du système cellulaire du fait de son échelle. La cellule est organisée à l'échelle mésoscopique, à la frontière de ce que les techniques d'imagerie comme la microscopie cryo-électronique permettent de voir (Goodsell et al., 2020). Mais les techniques omiques haut débit décrites précédemment et de nombreuses autres techniques comme la cristallographie au rayon X permettent de caractériser les acteurs moléculaires qui le composent (voir **Figure 10A**) (Goodsell et al., 2020). Bien qu'invisibles, la nature de ces molécules rend leur représentation en formes géométriques intuitive et efficace. L'utilisation de « schémas de synthèse » pour résumer les conclusions d'un article illustre cette idée (voir **Figure 10B**). Ils sont systématiques dans les revues (Lajoie et al., 2015; Misteli, 2020; *RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis*, 2019; Torres et al., 2023) et leur importance a été soulignée dans le numéro spécial 2010 de Nature Methods sur la visualisation des données biologiques (Wong, 2011).

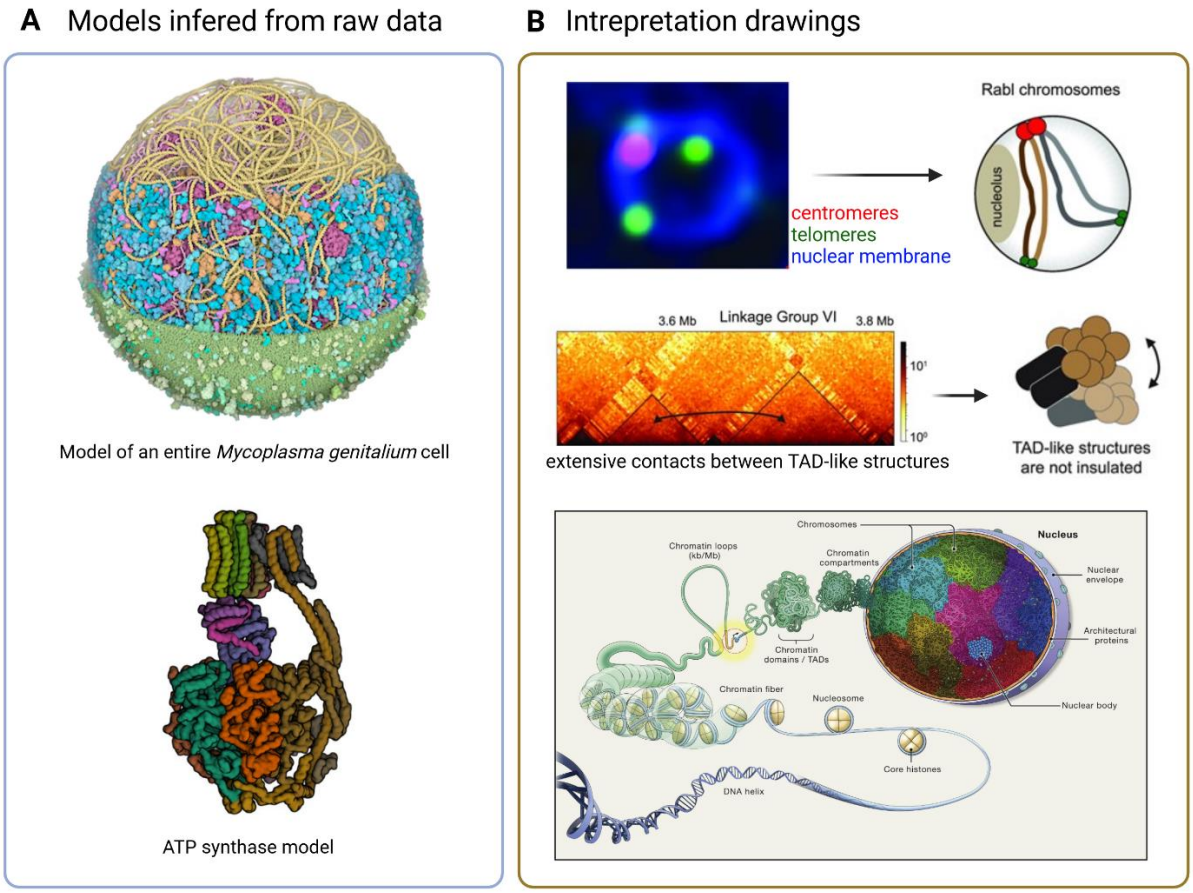


Figure 10 : L'importance de la visualisation en biologie cellulaire. (A) Modèles inférés à partir de données expérimentales. Le modèle de *Mycoplasma genitalium* a été approximé à une sphère d'un rayon de ~145 nm. Chaque protéine est représentée par une structure 3D provenant soit de modélisation par homologie, soit de données expérimentales, soit d'homologues provenant de la Protein Data Bank. Le modèle de structure de l'ATP synthase mitochondriale bovine a été inféré à partir de données de microscopie cryo-électronique (identifiant Protein Data Bank 5ARE). **(B)** Exemples de schémas de synthèse de la structure de la chromatine de *N. crassa* (Torres et al., 2023) et de l'organisation générale des génomes eucaryotes (Misteli et al., 2020). Microscopie à fluorescence de cellules de *N. crassa* exprimant CenH3::iRFP (rouge) pour éclairer les centromères, TZA1::GFP (vert) pour mettre en évidence les télomères, et ISH1::BFP (bleu) pour délimiter la membrane nucléaire (Klocko et al. 2016). Le schéma de droite résume les caractéristiques de la configuration Rabl mises en évidence par le marquage : clustering respectif des centromères et des télomères à la membrane nucléaire. De même, les données HiC mettent en valeur des contacts entre compartiments de la chromatine (flèche noire) et le modèle de synthèse propose un repliement sur eux-mêmes de ces deux compartiments.

Comme illustré plus haut pour les réseaux, la complexité des données omiques impose donc un défi de taille : comment adapter l'étape cruciale de visualisation des résultats à la nouvelle échelle des données expérimentales ? (O'Donoghue, 2021) (voir (O'Donoghue et al., 2018) pour une revue). O'Donoghue et al. soulignent la nécessité de développer de nouvelles stratégies de visualisation. Le travail de Goodsell est un exemple emblématique d'une approche originale de visualisation à large échelle

(Goodsell, 2009). Ses illustrations sont présentées comme « Molecule of the Month » sur le site de la « Protein Data Bank » depuis plus de vingt ans, mais il a plus récemment participé à la modélisation structurale d'une cellule entière de *Mycoplasma genitalium* (Maritan et al., 2022) (voir **Figure 10A**). Les petites molécules, les ions et l'eau ne sont pas représentés sur l'illustration, ils remplissent les espaces entre les macromolécules. La section supérieure du modèle met en évidence les ribosomes (magenta), les filaments d'ADN (jaune) et d'ARNm (rose) ; la section centrale montre le nucléoïde bactérien dans le contexte de macromolécules solubles (protéines liant l'ADN en orange, protéines cytoplasmiques en nuances de bleu, ARNt en rose vif) ; la section inférieure montre la membrane cellulaire (gris/vert) avec les protéines membranaires associées (nuances de vert). Cet exemple montre comment la création de visualisation à l'échelle mésoscopique est au carrefour de multiples disciplines : biologie, physique, informatique et art. Une telle visualisation offre un autre regard sur l'organisation interne d'une cellule. En effet, la complexité discutée jusqu'ici prend un sens physique : la diversité et le nombre d'acteurs moléculaires se traduit en un système extrêmement dense. Les auteurs décrivent que « l'aspect le plus long et le plus hétérogène de ce processus est la première étape - la collecte et la sélection des données pour soutenir la modélisation » (Maritan et al., 2022). De telles visualisations sont donc rendues possibles par un immense travail d'analyse et d'assemblage de connaissances. Est-il possible d'aller vers une lisibilité similaire dans le cas de l'analyse et l'intégration directe de données brutes omiques ?

1.4 L'opportunité de la modélisation 3D des génomes à partir de données Hi-C

1.4.1 La méthode Hi-C, « High-throughput Chromatin Conformation Capture »

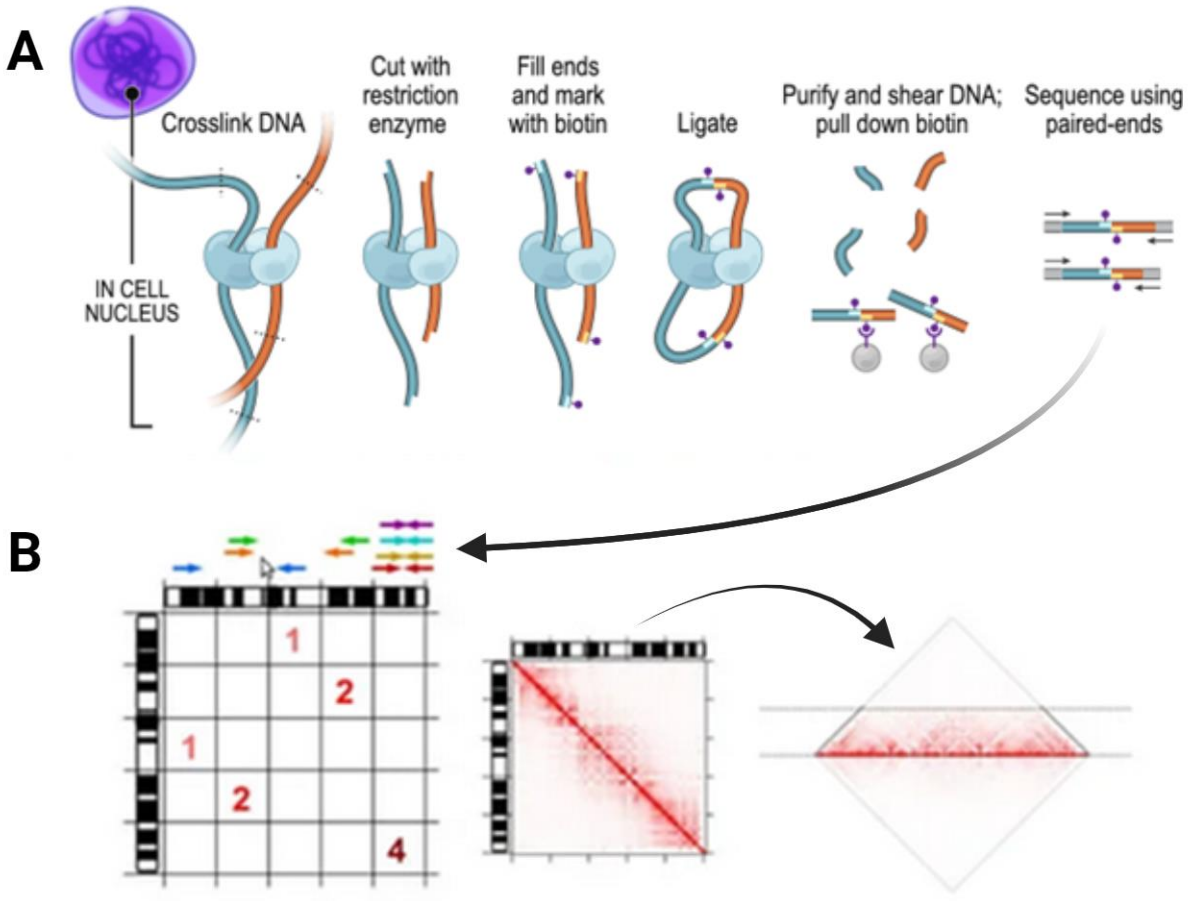


Figure 11 : Les étapes principales de la méthode Hi-C. Schéma issu de (Rao et al, 2015). **(A)** Résumé schématique des principales étapes expérimentales qui permettent d'obtenir l'information de contact entre chaque paire de régions de la chromatine. **(B)** Schéma de la création d'une carte de contact. Chaque pixel correspond au nombre de paires reads alignés aux extrémités de la séquence chimérique formée par les deux régions en abscisse et en ordonnée. La carte de contacts en heatmap est obtenue après normalisation et transformation logarithmique de la matrice de contact.

Comme présenté précédemment et illustré dans la **Figure 11**, la première étape de la méthode générique d'Hi-C est la fixation des régions de la chromatine proches dans l'espace du noyau. Ces fixations sont réalisées grâce au formaldéhyde, qui est le plus petit des aldéhydes (CH_2O). Il est électrophile et va donc réagir avec les protéines et l'ADN, formant ainsi des complexes. Sa petite taille lui permet de lier deux groupes distaux d'environ 2Å en réagissant avec les groupements aminés de la chaîne principale ou latérale des protéines et/ou des quatre bases de l'ADN. La chromatine est ensuite

découpée avec une ou plusieurs enzymes de restriction. Les extrémités de ces fragments sont marquées à la biotine pour pouvoir être sélectionnées ensuite. L'étape suivante de ligation des régions fixées est cruciale, puisque c'est le séquençage des séquences chimériques ainsi formées qui donne indirectement accès à l'information de proximité. Les reads « pair-end » du séquençage sont alignés sur le génome de référence et décomptés pour chaque paire de régions (voir **Figure 11**).

Un point important est que les fragments du génome découpés par l'enzyme de restriction sont très petits comparés au génome entier. En utilisant par exemple sur le génome humain une enzyme de restriction coupant 6 pb, il y a près de 10^6 fragments de restriction, ce qui conduit à un espace d'interaction de l'ordre de 10^{12} interactions possibles par paire (Lajoie et al., 2015). La plupart du temps, le nombre de reads séquencés n'est pas suffisant pour couvrir tous ces fragments. Il est donc nécessaire de créer des groupes ("bins"), réunissant le signal de plusieurs fragments de restriction. Plus une expérience d'Hi-C repose sur une enzyme de restriction coupant souvent (petits fragments) et une grande profondeur de séquençage (un grand nombre de reads), moins il est nécessaire de fusionner le signal de plusieurs fragments ensemble. C'est-à-dire qu'il est possible de garder une résolution plus fine. Le rapport bruit/signal a également un impact sur le résultat et cette problématique de la résolution (i. e. la taille choisie pour les bins, les groupes de fragments) est plus marquée sur les contacts longue distance que les contacts proches, comme illustré en **Figure 12**. Avec des bins de taille 10 Kb, le signal des contacts longues distances visible à 50 Kb (flèche blanche) est perdu. En revanche, le signal des contacts courte distances est conservé (ils sont plus fréquents puisque plus des régions sont proches, plus des contacts sont détectés) et profite de la meilleure résolution.

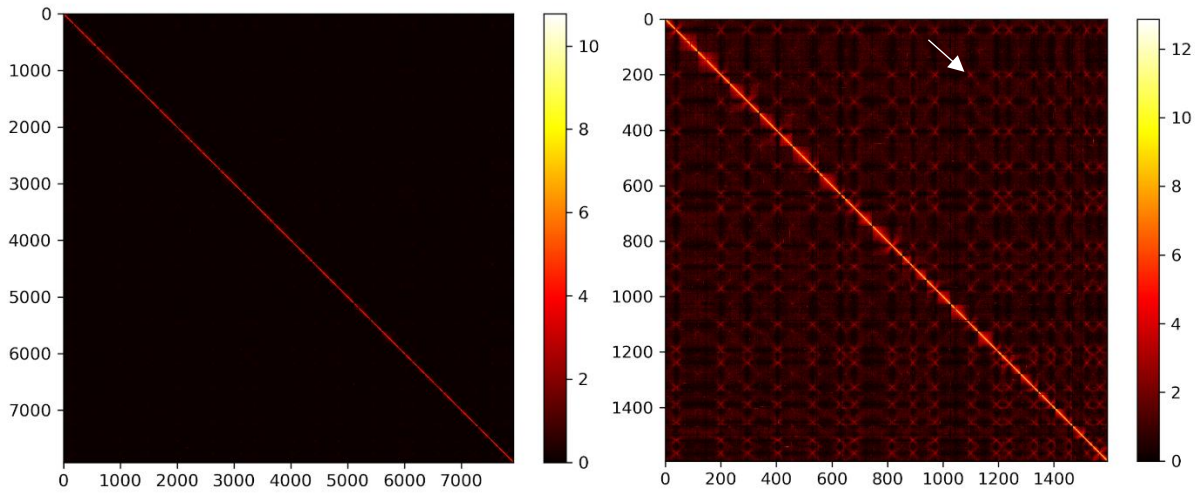


Figure 12 : L'impact du choix de la résolution sur le signal Hi-C. Cartes de contacts à 10 Kb (gauche) et à 50 Kb (droite) de résolution du génome du champignon *P. striiformis*. Cartes de contacts créées avec 3DGB à partir des données de (Xia et al., 2022). A 10 Kb, la carte de contact est de taille 8 000 par 8 000 bins. A 50 Kb, la carte de contact est de taille 1 500 par 1 500 bins. Ici, la profondeur de séquence ne permet pas une lisibilité à 10 Kb des contacts entre les

centromères (exemple à la flèche blanche), ce qui illustre que le choix de la résolution dépend de la profondeur de séquençage. Obtenir ou non des cartes de contacts lisibles à une résolution fine est dépendant du séquençage en amont.

Une carte de contact Hi-C est très dense puisqu'elle résume l'information contenue dans le séquençage de millions de reads en une matrice de taille 800 par 800 (par exemple, voir table annexe 1). En plus de sa densité, L'information visualisée dans les cartes de contacts est une représentation abstraite de la structure 3D du génome. L'accès à cette information nécessite une mesure indirecte : le séquençage et le comptages des reads. L'Hi-C utilise ces comptages de contacts comme un proxy de l'information qui nous intéresse, i.e. la distances entre les différentes portions du génome. Pour visualiser ces distances, on utilise les cartes de contact et un gradient de couleurs pour représenter l'intensité des comptages.

1.4.2 Modéliser la structure d'un génome : comment passer de fréquences de contact à des positions dans l'espace ?

1.4.2.1 Méthodes probabilistes ou basées sur les distances, plusieurs types de modélisations des distances dans l'espace

A cet égard, la modélisation 3D des cartes de contact Hi-C présentées précédemment est une stratégie alternative intéressante (O'Donoghue et al., 2018). Elle repose sur le calcul de coordonnées 3D (x, y, z) pour toutes les régions génomiques représentées dans une carte de contact, de sorte que leurs distances euclidiennes par paire restent cohérentes avec leurs fréquences de contact dans la carte originale. Le modèle obtenu est un ensemble de sphères dans l'espace, chacune représentant une région génomique donnée. La modélisation 3D de l'organisation spatiale des chromosomes n'est pas triviale, mais il existe de nombreux logiciels basés sur différentes stratégies (voir Oluwadare et al., 2019 pour une revue). Nous présentons ici les deux types de stratégies possibles pour déduire les coordonnées 3D à partir des cartes de contacts.

Dans toute cette section, on note c_{ij} le nombre de reads observés entre le locus i et le locus j. Pour chaque locus i, on note $X_i = x_i, y_i, z_i$ ses coordonnées dans l'espace. On note X la matrice qui contient toutes les coordonnées de tous les points. On note $d_{ij} = || X_i - X_j ||$ où d est la distance euclidienne entre le locus i et le locus j.

1.4.2.2 Les méthodes basées sur des techniques de positionnement multidimensionnel ou MDS (multidimensional scaling)

Les méthodes dites « basées sur les distances » convertissent les fréquences de contacts en distances et résolvent ensuite un problème d'optimisation pour en déduire des coordonnées 3D (J. Li et al., 2018; Rieber & Mahony, 2017). Ces méthodes nécessitent souvent d'intégrer au modèle des contraintes physiques liées à des caractéristiques structurelles connues des génomes comme le regroupement des centromères et des télomères dans des clusters opposés à l'intérieur du noyau ou le

regroupement de l'ADNr à l'extérieur de la structure globale pour former le nucléole.

Plus précisément, on définit une fonction Θ qui permet d'associer chaque fréquence de contact c_{ij} à une distance, par exemple $\Theta(c_{ij}) = \beta\, c_{ij}^{\alpha}$ où α et β sont des paramètres fixés qui dépendent de propriétés physiques des polymères. On cherche alors les coordonnées X des locus dans l'espace qui minimise la quantité suivante mesurant l'écart entre les distances entre les points dans le modèle 3D et les distances entre les points obtenues à partir des comptages de la matrice de fréquence de contact Hi-C et de la fonction Θ : $\sum_{(i,j)} \frac{(d_{ij} - \Theta(c_{ij}))^2}{\Theta(c_{ij})}$

Ces approches dépendent de nombreux paramètres (ici, α et β) et le choix de la quantité à minimiser est arbitraire. Les modèles 3D de la littérature décrits ensuite pour *N. crassa*, *S. cerevisiae* et *S. pombe* (voir **Figure 15**) ont été obtenus en utilisant ce type de méthodes. Ces approches ne prennent pas en compte que les comptages observés dans les matrices de fréquences de contact sont très aléatoires du fait des caractéristiques intrinsèques de la technologie de séquençage et de la stochasticité du phénomène étudié. Les stratégies probabilistes modélisent les nombres de contacts à l'aide de variables aléatoires (Varoquaux et al., 2023).

1.4.2.3 Inférence des coordonnées 3D à l'aide d'un modèle probabiliste

La méthode Pastis-NB (implémentée dans le logiciel Pastis (Varoquaux et al., 2023)) met en œuvre une stratégie probabiliste basée sur des variables aléatoires binomiales négatives. Ce type de variables aléatoires est un choix de modélisation adapté aux données de comptage générées par les technologies de séquençage, comme l'illustre la large adoption des distributions binomiales négatives pour modéliser les données RNA-seq (Love et al., 2014).

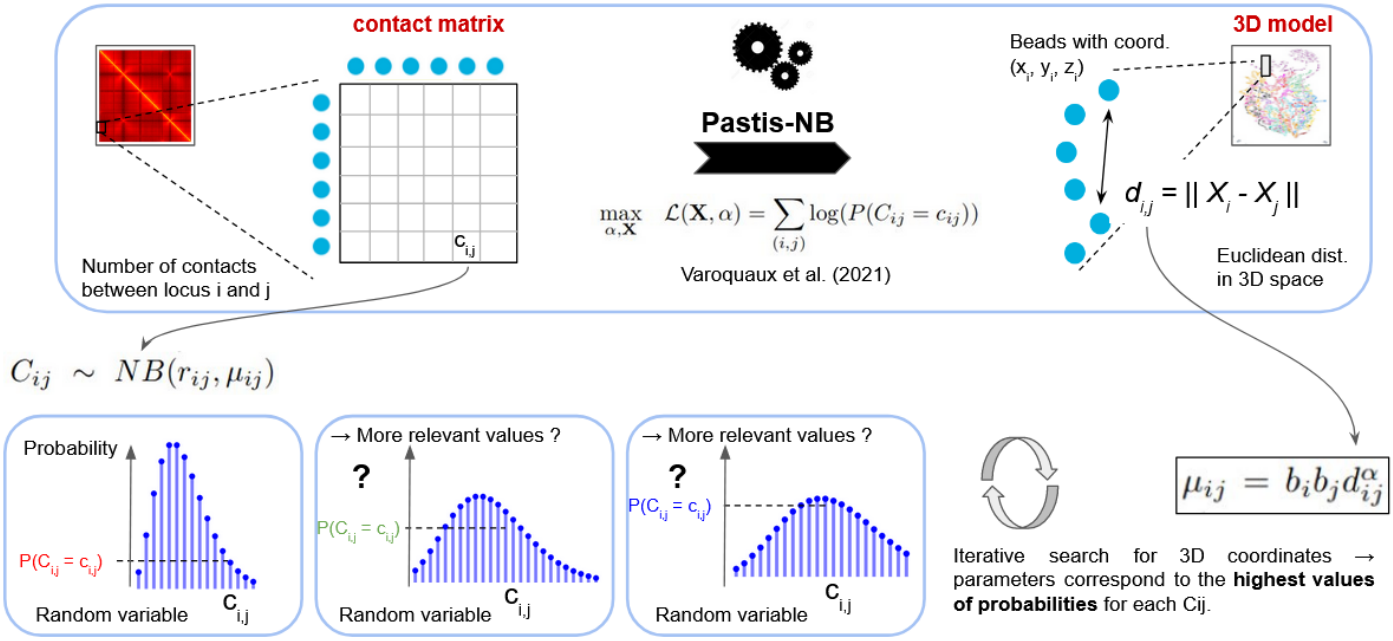


Figure 13 : Représentation schématique du principe de l'inférence des coordonnées 3D du génome à partir des comptages observés des données Hi-C. La quantité c_{ij} représente le nombre de reads séquencés entre le locus i et le locus j du génome. On considère que ce comptage est la réalisation d'une variable aléatoire C_{ij} qui suit une loi binomiale négative, dont le paramètre de moyenne dépend de X, les coordonnées 3D de nos locus dans l'espace par l'intermédiaire des quantités d_{ij} les distances euclidiennes entre les locus i et j, pour tous les locus. Cette moyenne dépend également d'un paramètre α et de paramètres de normalisations b_i et b_j corrigeant les biais techniques de séquençage. Le paramètre r_{ij} modélise la dispersion des données. Les coordonnées 3D sont des paramètres inconnus. Pour inférer ces paramètres à partir des données, la méthode Pastis recherche le maximum de vraisemblance du modèle à l'aide d'un algorithme itératif.

Comme illustré en **Figure 13**, la méthode Pastis-NB modélise chaque nombre de contacts observés c_{ij} entre deux régions i et j comme une réalisation d'une variable aléatoire négative binomiale de dispersion r_{ij} et de moyenne μ_{ij} inconnue. Le lien avec l'espace 3D est fait grâce à l'expression de la moyenne comme d'une fonction de la distance Euclidienne entre les régions i et j, donc indirectement de leurs coordonnées 3D. Cette hypothèse modélise l'idée intuitive que plus deux régions sont proches l'une de l'autre, plus leur nombre de contacts est élevé. Pastis-NB cherche itérativement les paramètres qui maximiseront la probabilité d'observer c_{ij}. À l'échelle de la matrice entière, la méthode Pastis-NB modélise donc l'inférence d'une structure 3D consensus comme un problème de maximum de vraisemblance (Varoquaux et al., 2021). Cette approche tient ainsi compte de la sur-dispersion, fonctionne bien dans des contextes de faible couverture et ne nécessite aucune contrainte physique initiale.

1.4.3 Modéliser dans l'espace le génome, hub central du système cellulaire, une opportunité pour l'intégration visuelle

1.4.3.1 Exemples d'intégrations multiomiques grâce à la structure 3D des génomes

Les modèles 3D de la chromatine sont des modèles statistiques dont la résolution reste faible, de l'ordre de 5 à 50 kb. Chaque sphère est une représentation abstraite d'une région génomique. L'échelle du modèle n'est donc pas du tout comparable à l'échelle atomique des modèles protéiques. Malgré cela, la modélisation 3D des génomes représente une riche opportunité pour la visualisation et l'intégration des données omiques à large échelle. Quelques études utilisent en effet des modèles 3D de génomes comme support à l'intégration de données multiomiques (Ay et al., 2014; Wang et al., 2022). Un modèle du génome de la souris *Mus musculus* a par exemple permis d'illustrer la localisation des gènes codant pour des récepteurs olfactifs proches de l'enveloppe nucléaire lorsqu'ils ne sont pas exprimés (Tan et al., 2019). Dans un autre exemple d'approche « cellule unique » chez la souris, Stevens et al. ont illustré l'organisation des compartiments de régulation transcriptionnelle (Stevens et al., 2017).

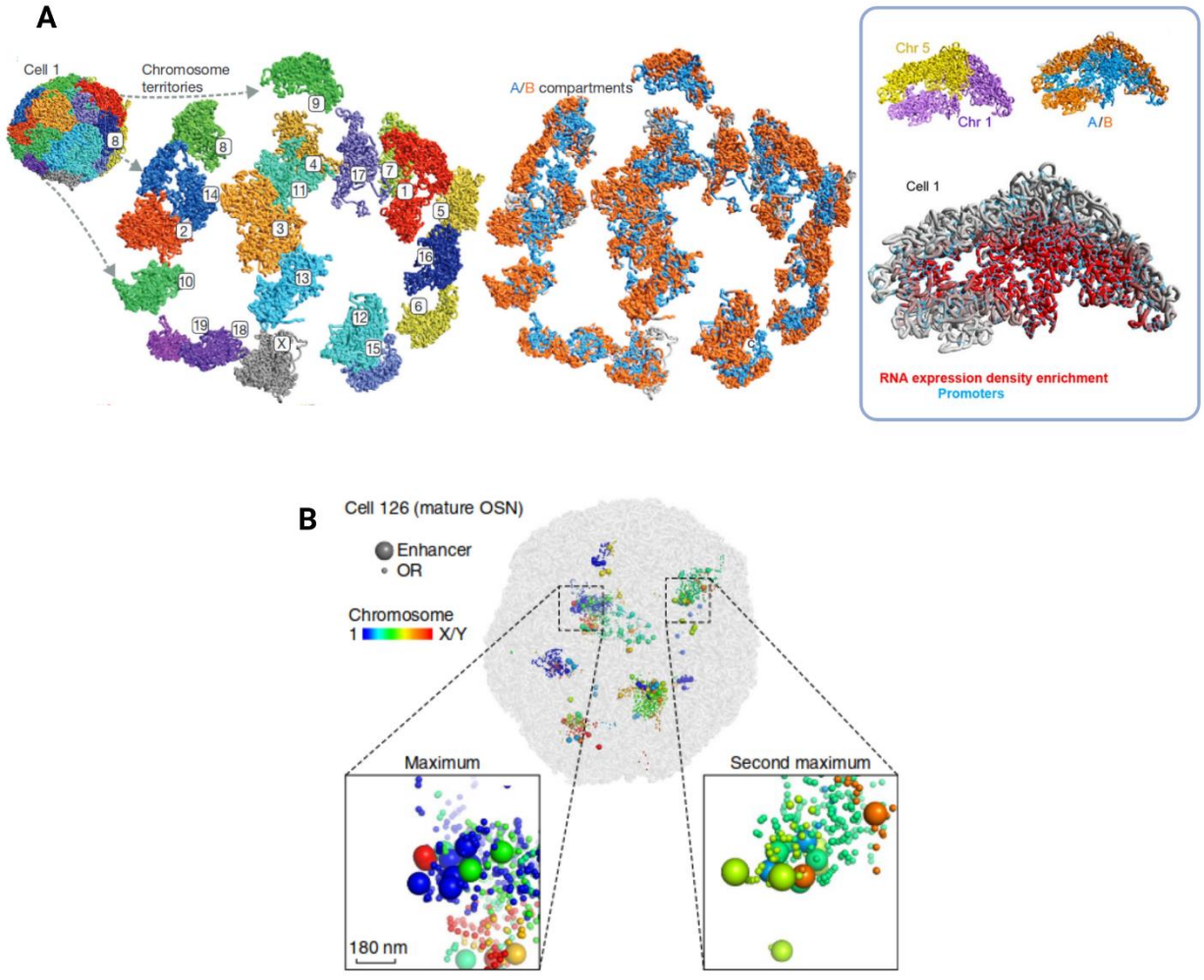


Figure 14 : deux exemples d'intégrations omiques sur la structure 3D du génome de *Mus musculus*. (A) Représentations issues de (Stevens et al., 2017), structure 3D (générée avec

NucDynamics, développé pour les données Hi-C unicellulaires) d'un génome de cellule souche embryonnaire haploïde de *M. musculus* avec un zoom dispersé des territoires chromosomiques (à gauche), et la distribution spatiale des compartiments A (bleu) et B (rouge) (à droite). L'encadré met en valeur le lien entre l'organisation des compartiments structuraux A et B et celle de la densité de transcription des gènes. **(B)** Représentations issues de (Tan et al., 2019), modèle 3D du génome d'un neurone olfactif mature représentatif chez *M. musculus*, construit avec hickit (développé pour les données Hi-C diploïdes unicellulaires). Les gènes des récepteurs olfactifs et leurs activateurs sont représentés sous forme de sphères avec des rayons de particules différents pour plus de clarté visuelle. L'échelle de couleur correspond au chromosome d'appartenance. Les deux agrégats mis en valeur illustrent la colocalisation des gènes et de leurs régulateurs dans l'espace du noyau.

Les figures de ces articles montrent que l'intérêt des modèles 3D pour la visualisation et l'intégration des données omiques à large échelle repose sur leur nature intuitive. Ces modèles statistiques ne sont pas des images de la cellule mais l'information contenue dans les données brutes est directement représentée pour ce qu'elle est : une distance dans l'espace. Cette abstraction limitée facilite la lecture du modèle et donc celle des autres données omiques intégrées à la visualisation.

1.4.3.2 La modélisation 3D des génomes et de son utilisation pour l'intégration multiomique chez S. cerevisiae, S. pombe et N. crassa

Au-delà des quelques exemples cités précédemment, la grande majorité des articles présentant des données Hi-C n'utilisent pas cette méthode de visualisation 3D en complément de l'analyse de carte de contact en heatmap. Ce constat est paradoxal étant donné qu'un large panel d'outils existe pour créer les modèles 3D (Oluwadare et al., 2019) et les visualiser : Genome3D (Asbury et al., 2010), GMOL (Nowotny et al., 2016), GenomeFlow (Trieu, Oluwadare, Wopata, et al., 2019), HiC-3DViewer (Djekidel et al., 2017), Csynth (Todd et al., 2021) et WashU Epigenome Browser (D. Li et al., 2022). Les génomes de *S. cerevisiae* et *S. pombe* ont fait partie des premiers génomes modélisés en 3D (Duan et al., 2010; Tanizawa et al., 2010). Le génome de *N. crassa* a également été modélisé, pour chaque chromosome indépendamment (Galazka et al., 2016).

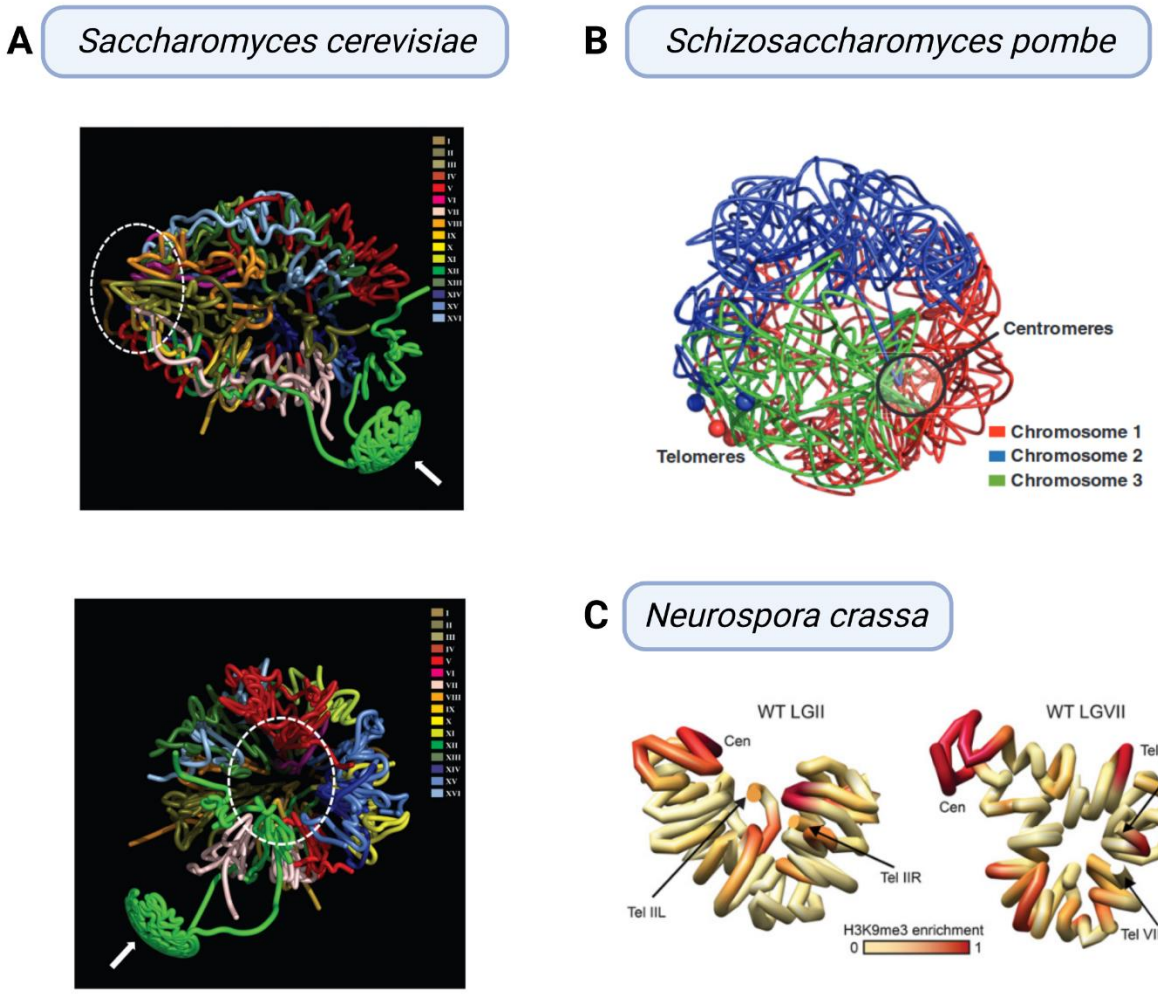


Figure 15 : Les modèles 3D existants pour les génomes de *S. cerevisiae* et *S. pombe* et *N. crassa*. (A) Modèle 3D du génome de *S. cerevisiae* présenté dans (Duan et al., 2010). **(B)** La même année, Tanizawa et al. présentent un modèle 3D du génome entier de *S. pombe*. Une caractéristique importante des deux modèles de génome entier est leur construction guidée par des contraintes de position. Les télomères et les centromères doivent former deux clusters périphériques, tous les points du modèle doivent être dans une sphère donnée et le nucléole doit former également un cluster périphérique. **(C)** Deux exemples des modèles de chromosomes de *N. crassa* issus de (Galazka et al., 2016), inférés indépendamment les uns des autres, excluant ainsi les informations de contacts interchromosomiques.

Pourtant, les études Hi-C plus récentes ne présentent pas de nouveau modèles 3D pour illustrer leurs résultats. (Rodriguez et al. 2022 ; Tanizawa et al. 2017 ; Costantino et al. 2020). Au regard du défi de la visualisation à l'ère des données omiques, nous avons donc exploré la modélisation 3D des génomes entiers chez ces organismes modèles. Les nouvelles méthodes de modélisation probabiliste et les derniers résultats de la littérature permettent de mieux comprendre l'organisation spatiale des chromosomes fongiques et d'améliorer l'intégration visuelle des données omiques.

2 RESULTATS

2.1 Intégrer un réseau de régulations transcriptionnelles sur le modèle 3D du génome de *S. cerevisiae*

Les résultats et figures de cette première partie sont publiés dans « Additional insights into the organization of transcriptional regulatory modules based on a 3D model of the Saccharomyces cerevisiae genome », BMC research Note 2022, voir en annexe.

2.1.1 Objectifs et sources des données publiques utilisées

Cette première approche a pour objectif de rechercher des informations supplémentaires sur l'organisation des modules de régulation transcriptionnelle (MRT) à partir du modèle 3D du génome de *S. cerevisiae* à l'interphase présenté en introduction (Duan et al., 2010). Les MRT ont été explorées sous un nouvel angle, qui intègre les informations fonctionnelles et spatiales actuellement disponibles, et répond à la question suivante : les gènes cibles associés à un facteur de transcription commun (donc appartenant au même MRT) sont-ils disséminés au hasard dans le noyau, ou sont-ils colocalisés ?

Cette problématique s'inscrit dans le prolongement de deux analyses précédentes des modules de régulation transcriptionnelle chez *S. cerevisiae*, présentées dans la littérature. La première est celle de (Monteiro et al., 2020). Les auteurs ont évalué les caractéristiques de régulation du réseau transcriptionnel actuel de *S. cerevisiae*, en tirant parti de la dernière version de leur base de données YEASTRACT, qui comprend près de 200 000 interactions, dont 220 facteurs de transcription et 6 886 gènes cibles. La seconde est celle de (Sun et al., 2019). Les auteurs ont utilisé le modèle 3D du génome de *S. cerevisiae* pour étudier l'organisation spatiale du réseau de régulations transcriptionnelles de *S. cerevisiae*.

L'étude de Monteiro et al. repose sur un travail colossal de collecte, de nettoyage et d'organisation des régulations transcriptionnelles identifiées dans plus d'un millier de publications. Les auteurs ont notamment fourni des informations sur le niveau de confiance de chaque régulation, ce qui a permis d'obtenir des données de très grande qualité. Ils ont observé des propriétés topologiques intéressantes du réseau transcriptionnel global de *S. cerevisiae* et ont discuté de la complexité des processus de régulation de la transcription qui contrôlent l'expression des gènes.

Nous explorons ici le rôle potentiel de l'organisation 3D du génome dans le fonctionnement de ce réseau. Sun et al. ont eu l'idée originale de placer les régulations transcriptionnelles dans le contexte du modèle de génome 3D disponible chez *S. cerevisiae*. Ils concluent que "le réseau de régulation transcriptionnelle de *S. cerevisiae* présente une structure optimisée dans l'espace pour s'adapter aux exigences

fonctionnelles" (Sun et al., 2019). Pour tester notre approche nous avons donc choisi d'explorer cette structure optimisée en utilisant les régulations transcriptionnelles vérifiées de Monteiro et al. et en explorant individuellement les MRT.

2.1.2 Faire le lien entre le modèle 3D et les gènes

Les données supplémentaires de Monteiro et al. fournissent toutes les régulations transcriptionnelles de YEASTRACT, annotées en fonction de "binding evidence" (preuves de liaison), "expression evidence" (preuves d'expression) ou "both" (les deux). Seules les associations de régulation qui s'appuient uniquement sur des "preuves de liaison" ont été conservées ici. Elles représentent 176 facteurs de transcription avec 6 475 gènes cibles, connectés par 45 209 associations (23% de l'ensemble des données d'associations). Les coordonnées tridimensionnelles de 9 185 « features » du génome de *S. cerevisiae* (dont 6 572 ORF) ont ensuite été interpolées sur la structure 3D du génome et utilisées pour calculer les distances spatiales euclidiennes entre toutes les paires de « features » du génome (42 177 520 distances).

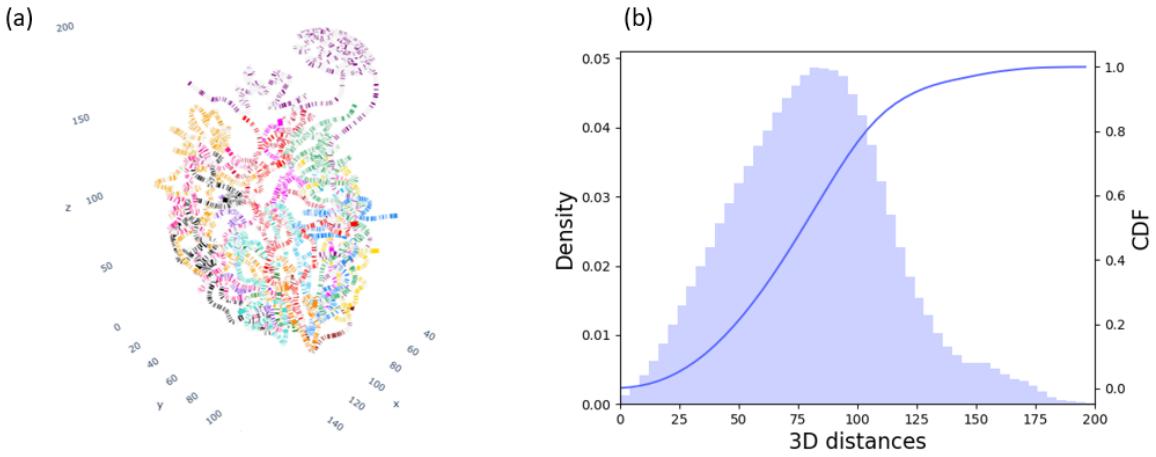


Figure 16 : Le génome de *S. cerevisiae* modélisé dans 3D-Scere. a) Capture d'écran du modèle 3D du génome disponible dans l'outil 3D-Scere. b) Histogramme de densité (bleu clair) de toutes les distances euclidiennes entre toutes les « features » chromosomiques dans le modèle 3D et fonction de distribution cumulative (CDF, bleu foncé) de l'histogramme de densité.

Pour chaque MRT défini par les 176 facteurs de transcription différents, les distances par paire entre les gènes cibles ont été sélectionnées. La distribution des distances obtenues avec toutes les « features » du génome de *S. cerevisiae* et avec le sous-ensemble de gènes appartenant à une MRT particulier ont été superposées et utilisées pour quantifier un biais potentiel de colocalisation (distances plus faibles) entre les gènes cibles dans les MRT. Un test de Kolmogorov Smirnov (KS) avec une correction de Bonferroni a été effectué pour quantifier la déviation par rapport à la distribution de tous les gènes.

En conséquence, plusieurs facteurs de transcription pour lesquels les gènes cibles présentaient des localisations atypiques dans le noyau ont été observés. Ces facteurs

de transcription sont énumérés dans le tableau (voir BMC note en annexe), et les distributions de distance des quatre TRM présentant la statistique KS la plus élevée (c'est-à-dire l'écart le plus important par rapport à la distribution de toutes les cibles) sont illustrées.

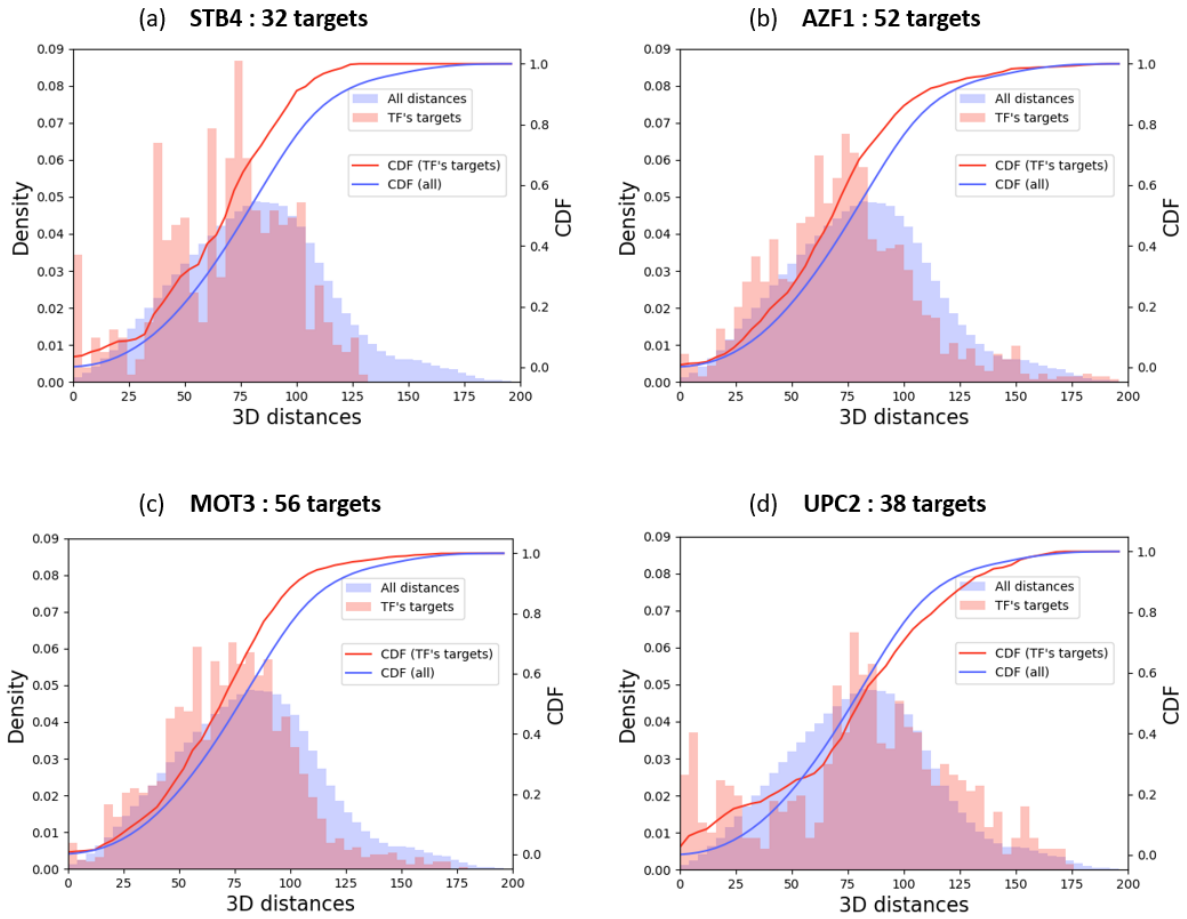


Figure 17 : Exemples de modules de régulation transcriptionnelle dans lesquels les cibles sont préférentiellement colocalisées dans le noyau. Les histogrammes des distances (roses) pour les cibles de quatre facteurs de transcription (STB4, AZF1, MOT3 et UPC2) sont illustrés avec l'histogramme de toutes les distances (bleu clair) tel que présenté dans la Figure 16. Ces facteurs de transcription ont été sélectionnés parce que (i) ils ont un nombre de gènes cibles > 30, (ii) ils présentent les valeurs les plus élevées de la statistique de Kolmogorov Smirnov, avec (iii) des valeurs p ajustées significatives associées (< 0,05).

Un outil open-source a été développé en python pour la visualisation et l'exploration interactives. Le code source est disponible[3] et l'outil est librement utilisable en ligne à l'adresse https://3d-scere.ijm.fr/ . Il permet de visualiser n'importe quelle liste de gènes dans le contexte du modèle 3D du génome de *S. cerevisiae*. D'autres informations peuvent être facilement ajoutées, comme des annotations fonctionnelles (GO annotations) ou des mesures d'expression génique.

[3] https://github.com/data-fun/3d-scere

2.1.3 Les limites du prototype et leurs implications

La première limite de cette approche concerne la pertinence biologique du modèle 3D du génome *de S. cerevisiae* qui a été utilisé. Ce modèle structurel ne représente qu'une vue statique moyenne du positionnement relatif des 16 chromosomes dans le noyau à l'interphase. Il a été obtenu à partir de données d'expériences 3C, qui ont dû être traitées à l'aide de procédures numériques complexes, afin de trouver une solution optimale. Parce que "optimal" ne garantit pas "réel", toutes les observations qui émergent de ce modèle doivent être validées. À cet égard, de nouvelles données générées par les techniques Hi-C plus récentes, à différents stades du cycle cellulaire de *S. cerevisiae* pour saisir la dynamique de l'organisation de son génome, pourraient être d'un grand intérêt.

La définition même des MRT pose également une limite à cette intégration multiomique. Nous avons défini un MRT comme un ensemble de gènes dont l'expression est modulée par un facteur de transcription commun. Mais un gène cible peut appartenir à plusieurs MRT et peut également nécessiter, pour être régulé transcriptionnellement, l'association de plusieurs facteurs de transcription. De tels gènes pourraient être étudiés spécifiquement pour des colocalisations particulières sur le modèle 3D du génome de *S. cerevisiae*.

2.2 Développement d'un workflow d'analyse allant des données brutes aux modèles 3D

2.2.1 Une approche plus modulaire, partant des données brutes

Afin de simplifier la création de modèles 3D nous avons développé 3D Genome Builder (3DGB), un workflow bio-informatique qui automatise la génération de modèles 3D pour visualiser et explorer l'organisation spatiale des chromosomes, sur la base des résultats expérimentaux Hi-C. La **Figure 18** présente une vue d'ensemble du workflow.

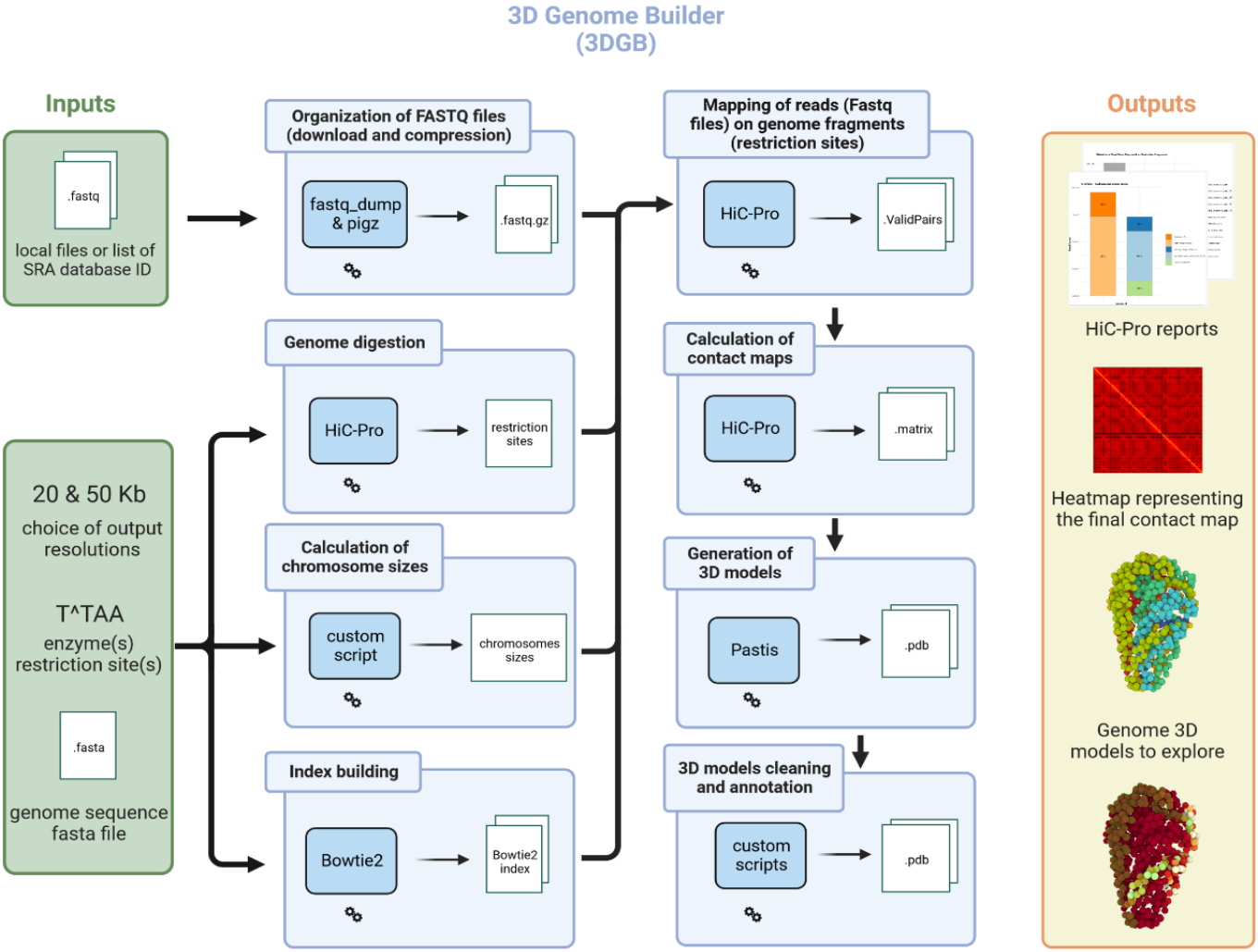


Figure 18 : Vue d'ensemble du workflow 3D Genome Builder (3DGB). Les entrées et les sorites sont respectivement représentés par des boîtes vertes ou oranges (panneaux de gauche et de droite). Le workflow est détaillé sous la forme d'une séquence de boîtes bleues (panneau du milieu). Chaque boîte bleue représente une tâche d'intérêt particulier. Le workflow 3DGB est orchestré par Snakemake (voir méthodes). Toutes les tâches sont automatiquement enchaînées les unes après les autres jusqu'à ce que les résultats ciblés soient produits. Les résultats finaux comprennent des rapports de contrôle de qualité générés par HiC-Pro, des cartes de contact et des fichiers PDB des modèles 3D. Les fichiers PDB sont enrichis d'annotations supplémentaires (voir le texte principal pour plus de détails) et peuvent être visualisés avec n'importe quel logiciel de visualisation PDB (mol* dans notre cas).

À partir des fichiers FASTQ bruts Hi-C, 3DGB exécute automatiquement les étapes bio-informatiques critiques nécessaires pour i) calculer les fréquences de contact Hi-C, ii) déduire les modèles 3D associés de l'organisation de la chromatine et iii) annoter et contrôler la qualité des modèles 3D. Ces modèles 3D sont stockés dans des fichiers PDB standard, de sorte qu'ils peuvent être étudiés avec des outils de visualisation complémentaires. Les modèles 3D peuvent éventuellement être enrichis de données omiques quantitatives supplémentaires, telles que des signaux ChIP-seq ou RNA-seq. 3DGB a été conçu pour être le plus ergonomique possible. Par conséquent, il ne

nécessite que quatre inputs à spécifier par l'utilisateur (**Figure 18**, cases vertes) : Les identifiants des fichiers FASTQ, un fichier FASTA contenant la séquence du génome de référence, la liste des sites de restriction pour la ou les enzymes utilisées pendant les expériences Hi-C et les résolutions ciblées pour l'analyse des données Hi-C. Ce dernier paramètre a un impact sur les modèles 3D finaux, c'est-à-dire que plus la valeur (spécifiée en pb) est petite, plus le modèle 3D est détaillé et contient de sphères.

Les huit principales étapes nécessaires au traitement des données Hi-C sont représentées par des cases bleues dans la **Figure 18** et s'appuient sur deux logiciels. La première étape utilise HiC-Pro (Servant et al., 2015), une référence dans le traitement des données Hi-C, citée plus de 800 fois. Il traite les fichiers FASTQ bruts, effectue un contrôle qualité et génère des comptes de contacts normalisés et les figures associées (présentant des statistiques importantes pour évaluer la qualité de l'étape de « mapping » des reads et justifier un éventuel filtrage). Ensuite, Pastis (Varoquaux et al., 2021) calcule itérativement des modèles 3D de l'organisation des chromosomes, par le biais d'une modélisation binomiale négative originale du nombre de contacts (Pastis-NB, voir introduction). Il est intéressant de noter que Pastis produit un modèle consensuel qui, par son unicité, simplifie grandement les analyses et les interprétations en aval. Autour de ces deux composants principaux (HiC-Pro et Pastis), 3DGB centralise la configuration, effectue l'analyse Hi-C, génère des cartes de contact, construit un modèle 3D et ajoute un traitement supplémentaire du modèle 3D en sortie sous forme de fichiers PDB. 3DGB est open source, disponible sur GitHub[4] et archivé dans Software Heritage[5].

2.2.2 Un outil pour faciliter le partage de l'approche choisie

Les principaux résultats de 3DGB sont des modèles 3D de l'organisation spatiale du génome. Ces modèles 3D sont composés de sphères dont les coordonnées 3D (x, y, z) ont été déduites des informations de contact (voir introduction). Une sphère représente plusieurs milliers de paires de bases correspondant à la résolution choisie lors de l'analyse des données Hi-C. Pour faciliter l'étude de ces modèles, nous avons enrichi les structures produites par Pastis en fichiers PDB par une procédure en quatre étapes. Tout d'abord, 3DGB formate et enrichit le modèle 3D produit par Pastis pour la visualisation et l'intégration des données en annotant chaque sphère avec le numéro de chromosome. Cela permet de distinguer les chromosomes lors de la visualisation de structures complètes (voir la **Figure 18** pour illustration). Deuxièmement, il reconstruit automatiquement les sphères pour lesquelles aucune coordonnée n'a pu être calculée par Pastis, en interpolant les coordonnées manquantes à partir des coordonnées existantes (voir Méthodes). Il est à noter que nous n'extrapolons pas les coordonnées manquantes, ce qui signifie que les sphères dont les coordonnées

[4] https://github.com/data-fun/3d-genome-builder

[5] swh:1:dir:26b6504724952e6d0d7db34c394e052217523754

manquantes sont situées aux extrémités du modèle sont supprimées. Troisièmement, les sphères aberrantes, c'est-à-dire les sphères placées en dehors du modèle global, sont filtrées et supprimées sur la base d'une valeur de distance seuil qui peut être spécifiée par l'utilisateur. Quatrièmement, les valeurs quantitatives, telles que les données ChIP-seq, peuvent être utilisées pour colorer le modèle, ce qui permet une intégration visuelle des données omiques sur la structure 3D du génome. Toutes ces fonctionnalités supplémentaires ont été mises en œuvre dans des scripts Python (Guido Van Rossum & Fred L. Drake, 2009) intégrés dans le workflow de Snakemake (voir Méthodes).

Dans l'ensemble, ces améliorations permettent aux analystes de données Hi-C de visualiser et d'explorer des modèles 3D plus facilement. Nous avons fourni la structure du modèle 3D au format PDB, un format de fichier traditionnellement utilisé pour stocker les coordonnées des structures moléculaires. La plupart des logiciels utilisés en biologie structurale pour visualiser et manipuler les structures peuvent traiter le format PDB. Nous avons utilisé de préférence Mol* (Sehnal et al., 2021), un outil de visualisation, avec une interface web conviviale permettant la visualisation et la personnalisation des modèles 3D en seulement quelques clics de souris. Nous avons également utilisé HiC3D-Viewer (Djekidel et al., 2017), un outil de visualisation spécialement développé pour les structures chromatiniennes. A noter que HiC3D-viewer nécessite la conversion du format de fichier PDB en format G3D. 3DGB fournit également des modèles 3D au format de fichier G3D, ce qui permet aux utilisateurs de choisir le logiciel de visualisation qu'ils préfèrent. À titre d'illustration, les images produites par Mol* et HiC3D-viewer sont présentées dans la **Figure 20**.

2.2.3 Détection de contigs

Dans le cadre de nos analyses des données Hi-C chez *N. crassa*, nous avons été confrontés à une situation inattendue en ce qui concerne la séquence génomique de référence de cet organisme. En effet, l'assemblage du génome de *N. crassa*, initialement publié en 2003 (Galagan et al., 2003), est composé de 7 chromosomes et de 13 super contigs, tous disponibles dans GenBank sous l'accession "assembly nc12". En 2016, Galazka et al. (Galazka et al., 2016) ont identifié une inversion du contig nommé "12.304", correspondant à une grande région du chromosome 6. En 2022, Rodriguez et al. (Rodriguez et al., 2022) ont encore amélioré l'assemblage original "nc12" en se basant sur l'analyse des données Hi-C. L'assemblage mis à jour qui intègre les améliorations de Galazka et al. et de Rodriguez et al. est disponible dans la base de données GEO sous l'accession "assembly nc14". Lorsque nous avons commencé à utiliser 3DGB pour créer un modèle 3D du chromosome de *N. crassa* de type sauvage, nous avons utilisé comme génome de référence l'assemblage "nc12", disponible dans la base de données GenBank. De manière surprenante, nous avons observé dans notre modèle 3D déduit, une incohérence dans l'ordre des régions génomiques, par rapport à la succession de nos sphères (**Figure 19**).

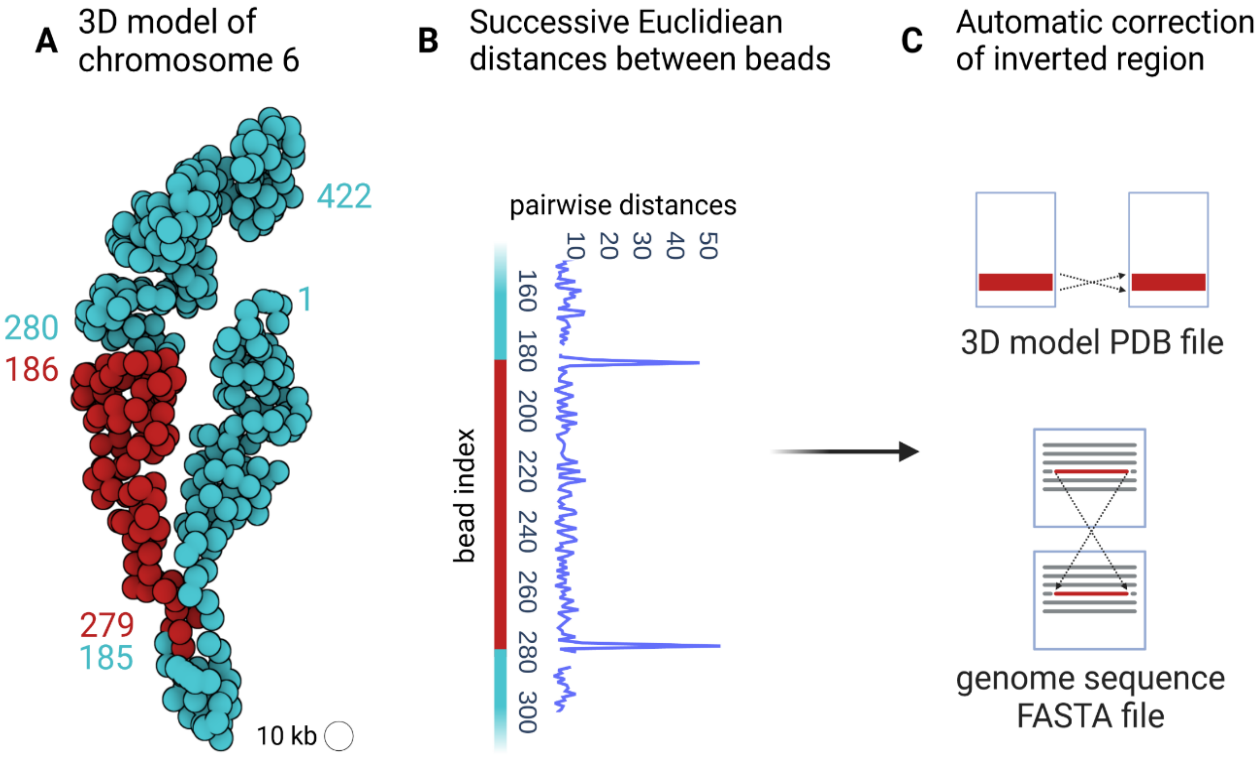


Figure 19 : Incohérence entre l'organisation spatiale du chromosome 6 chez *N. crassa* et l'assemblage du génome de référence " nc12 ". (A) Modèle 3D du chromosome 6 obtenu avec 3DGB en utilisant les données Hi-C de (Galazka et al. 2016) et l'assemblage "nc12" de la séquence génomique de *N. crassa* comme référence. Chaque sphère représente une région chromosomique de 10 kb. Les numéros étiquettent les sphères en fonction de la région génomique à laquelle elles sont associées. Par conséquent, le numéro 1 de la sphère correspond à l'intervalle]0, 10] kb de la séquence génomique linéaire, la sphère numéro 2 correspond à l'intervalle]10, 20] kb, la sphère numéro 3 correspond à l'intervalle]20, 30] kb, etc. Les transitions inattendues entre les paires de sphères 185 - 279 et 186 - 280 sont représentées par des couleurs différentes (bleu et rouge). **(B)** Distances euclidiennes entre les sphères avec des numéros successifs. Des valeurs élevées inattendues sont observées entre les paires de billes 185 - 186 et 279 - 280. Elles correspondent à la structure présentée en (A). **(C)** Sur la base de l'incohérence révélée dans les distances euclidiennes, il est possible dans 3DGB de corriger automatiquement la séquence génomique utilisée comme référence.

Notre modèle 3D a ainsi directement mis en évidence le contig inversé 12.304 sur le chromosome 6, trouvé à l'origine par Galazka et al. A 10 kb de résolution, les 95 sphères représentant cette région ont été placées dans l'espace 3D en fonction des seules informations de contacts Hi-C, indépendamment de l'assemblage du génome de référence. Ceci explique pourquoi nous avons pu mettre en évidence un décalage entre l'enchaînement des sphères dans les modèles 3D et la numérotation de ces sphères selon la séquence du génome (**Figure 19A**). A noter qu'en utilisant l'assemblage nc14 du génome comme référence, le modèle 3D du chromosome 6 que nous avons obtenu était cohérent à la fois spatialement (ordre des sphères) et séquentiellement (ordre des régions génomiques). A partir de cette expérience, nous avons développé une

procédure automatisée (intégrée à 3GDB en option) pour détecter et corriger les contigs inversés en comparant l'enchaînement des sphères dans le modèle 3D (basé sur la distance entre les sphères adjacentes, **Figure 19B**) et la numérotation de la séquence génomique écrite dans le fichier FASTA. Cette méthode produit également une version corrigée du modèle 3D et de l'assemblage du génome (**Figure 19C**).

2.2.4 Evaluation de la cohérence biologique des modèles 3D

Pour tester les performances et la pertinence biologique de 3DGB, nous avons choisi trois espèces fongiques emblématiques présentées en introduction : *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* et *Neurospora crassa*. Nous avons créé des modèles 3D de l'organisation de leur génome et confronté ces modèles aux connaissances actuelles de la littérature. Les données Hi-C de *S. cerevisiae* proviennent d'une étude récente de Costantino et al. (Costantino et al., 2020), les données Hi-C de *S. pombe* proviennent de Tanizawa et al. (Tanizawa et al., 2017) et les données Hi-C de *N. crassa* proviennent de Galazka et al. (Galazka et al., 2016). Des informations détaillées concernant la source des données sont données dans le tableau 1 en annexe. Nous avons ainsi obtenu de nouveaux modèles des génomes de *S. cerevisiae, S. pombe* et *N. crassa*. Trois d'entre eux, qui correspondent à des situations de type sauvage, sont présentés dans la **Figure 20**.

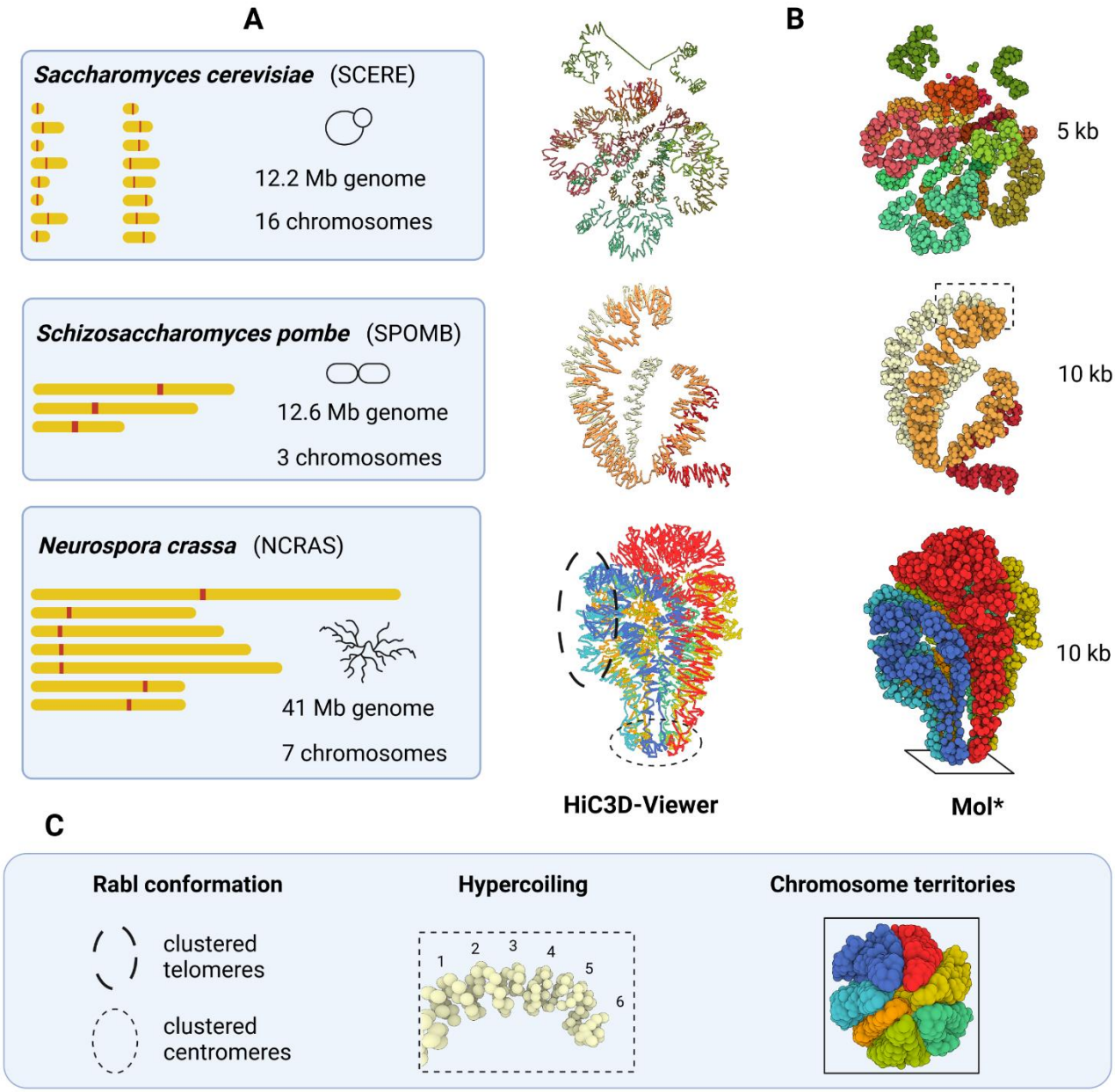


Figure 20 : Modélisation 3D des génomes sauvages des levures modèles *S. cerevisiae* et *S. pombe* et du champignon filamenteux *N. crassa*. (A) Représentations 2D des chromosomes dans chaque espèce. Ils sont affichés à la même échelle, ce qui montre la diversité des génomes utilisés dans cette étude. **(B)** Modèles 3D obtenus avec le workflow 3DGB. La source des données Hi-C utilisées pour les créer est présentée dans le tableau 1 en annexe. Ils sont visualisés avec HiC3D-Viewer (à gauche) et Mol* (à droite). A noter que HiC3D-Viewer relie les billes artificiellement, alors qu'avec Mol*, les billes restent individualisées (elles représentent chacune une région chromatinienne de 5 ou 10 kb, selon l'espèce, voir le tableau 1 en annexe). **(C)** Les principales caractéristiques de l'organisation du génome sont mises en évidence. La configuration Rabl, l'hyper-enroulement des chromosomes et leur distribution en territoires distincts sont bien observés sur les modèles 3D obtenus avec 3DGB et présentés en (B) (voir cercles et rectangles).

Les principales caractéristiques de l'organisation spatiale des génomes des trois

espèces ont été observées, à savoir la conformation de Rabl, l'hyper-enroulement des fibres de chromatine et les territoires chromosomiques. L'observation de ces caractéristiques connues a constitué une étape importante dans la validation de la pertinence du flux de travail de 3DGB, d'autant plus si l'on considère que 3DGB utilise une stratégie basée sur les probabilités pour déduire des modèles 3D (Pastis-NB), qui est libre de contraintes initiales, telles que la position des centromères.

2.3 Modèle 3D du génome de *N. crassa* et illustration de la modification du profil épigénétique chez le mutant hpo

2.3.1 Quantification de la stabilité des modèles 3D au bruit aléatoire avec des utilisations multiples de 3DGB

L'avantage d'utiliser un workflow comme 3DGB est la possibilité d'automatiser les tâches. Nous avons évalué la stabilité de l'organisation 3D des chromosomes par rapport aux inexactitudes potentielles dans les fréquences de contact mesurées par Hi-C. Notre stratégie a consisté à altérer les valeurs originales des fréquences de contact (issues des données Hi-C) en ajoutant un "shot noise" aux fréquences de contact, puis à exécuter 3DGB pour créer un modèle 3D associé. Si l'intensité du bruit de fond était faible, le modèle de sortie devait être proche du modèle original (obtenu à partir des données originales). Le score RMSD résultant, calculé en comparant la structure 3D déduite de la carte de contact de *N. crassa* de référence et la structure 3D déduite de la carte de *N. crassa* modifiée, devait également être faible. Au total, nous avons évalué 23 niveaux d'intensité du bruit (voir Méthodes) et généré automatiquement un total de 1 150 modèles. Nos résultats sont résumés dans la **Figure 21**.

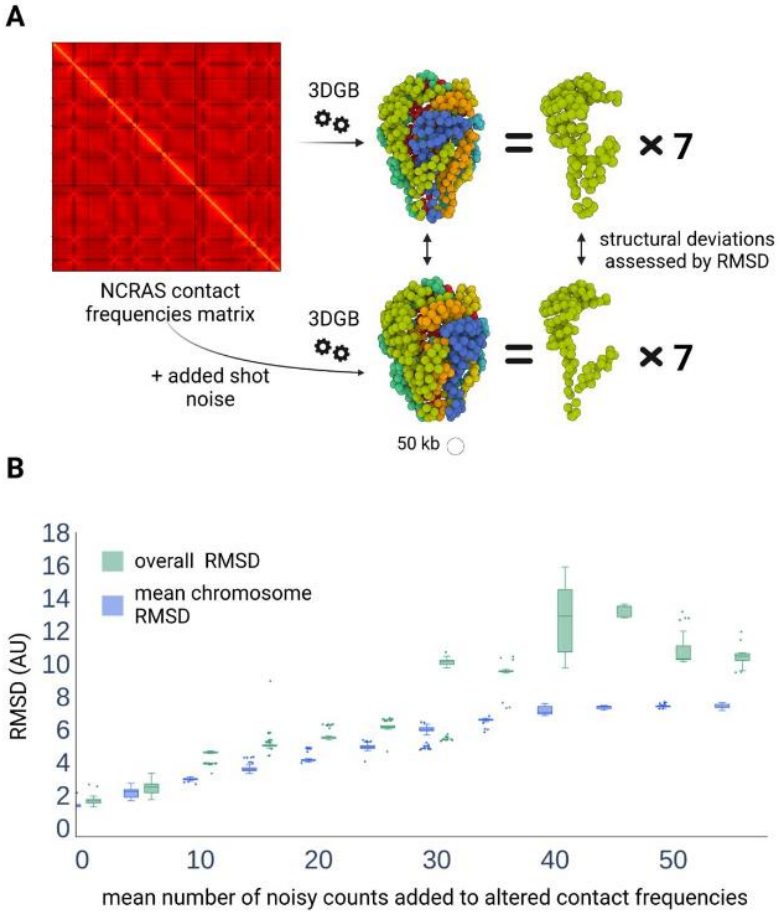


Figure 21 : Quantification de la stabilité des modèles 3D au bruit aléatoire ajouté aux fréquences de contact originales. (A) Chaque fréquence de contacts par paire est modifiée de manière aléatoire en ajoutant une valeur de comptage tirée d'une distribution de Poisson avec un paramètre moyen spécifié (voir Méthodes). Les modèles 3D sont créés avec 3DGB (à une résolution de 50 kb) et comparés en calculant les scores RMSD. **(B)** Valeurs des scores RMSD (unité arbitraire) obtenues pour différentes valeurs de bruit, c'est-à-dire différentes valeurs du paramètre moyen utilisé dans la distribution de Poisson. Pour une valeur de bruit donnée, 50 modèles sont générés, le score RMSD est soit calculé sur la structure globale composée de tous les chromosomes (boîtes vertes), soit moyenné sur les scores RMSD obtenus pour les 7 chromosomes individuels (boîtes bleues).

Comme attendu, nous avons observé une augmentation des scores RMSD lorsque l'intensité du bruit de fond augmente. Ce phénomène est plus frappant pour la structure globale (**Figure 21B**, boxplot verts) que pour les chromosomes individuels (**Figure 21B**, boxplot bleus). Cette observation souligne la plus grande sensibilité au « shot noise » de l'organisation des territoires chromosomiques par rapport à l'organisation interne des chromosomes.

2.3.2 Illustration de la modification du profil épigénétique chez le mutant hpo

Dans le génome de *N. crassa*, les régions hétérochromatiques sont une composante

majeure de la conformation des chromosomes (Galazka et al., 2016). L'hétérochromatine constitutive et l'hétérochromatine facultative sont respectivement des régions génomiques qui contiennent peu de gènes avec peu de transcription (hétérochromatine constitutive) et des régions génomiques qui contiennent des gènes avec une inhibition de l'expression génique régulée (hétérochromatine facultative). Au niveau moléculaire, l'hétérochromatine constitutive et l'hétérochromatine facultative peuvent être distinguées par la présence de marques d'histone H3K9me3 ou H3K27me2/3 (Galazka et al., 2016). Il a été observé que la réduction de H3K9me3 dans l'hétérochromatine constitutive provoque la redistribution de H3K27me2/3. En particulier, dans un contexte génétique où la protéine 1 de l'hétérochromatine (Hp1, qui reconnaît H3K9me3) est perdue, H3K27me3 est appauvri dans l'hétérochromatine facultative et H3K27me2 est gagné dans l'hétérochromatine constitutive (Jamieson et al., 2016). Notre objectif était d'évaluer la visualisation de ce phénomène à l'échelle du génome complet. Pour ce faire, nous avons commencé par créer des modèles 3D avec 3DGB, à partir des données Hi-C générées à partir de souches de type sauvage (WT) et déficientes en HP1 (hpo). Ces modèles ont ensuite été utilisés pour l'intégration visuelle des données ChIP-seq, en se référant à la localisation génomique des histones avec des modifications post-traductionnelles H3K9me3 et H3K27me2/3 (voir Méthodes). Les résultats sont présentés **Figure 22A**.

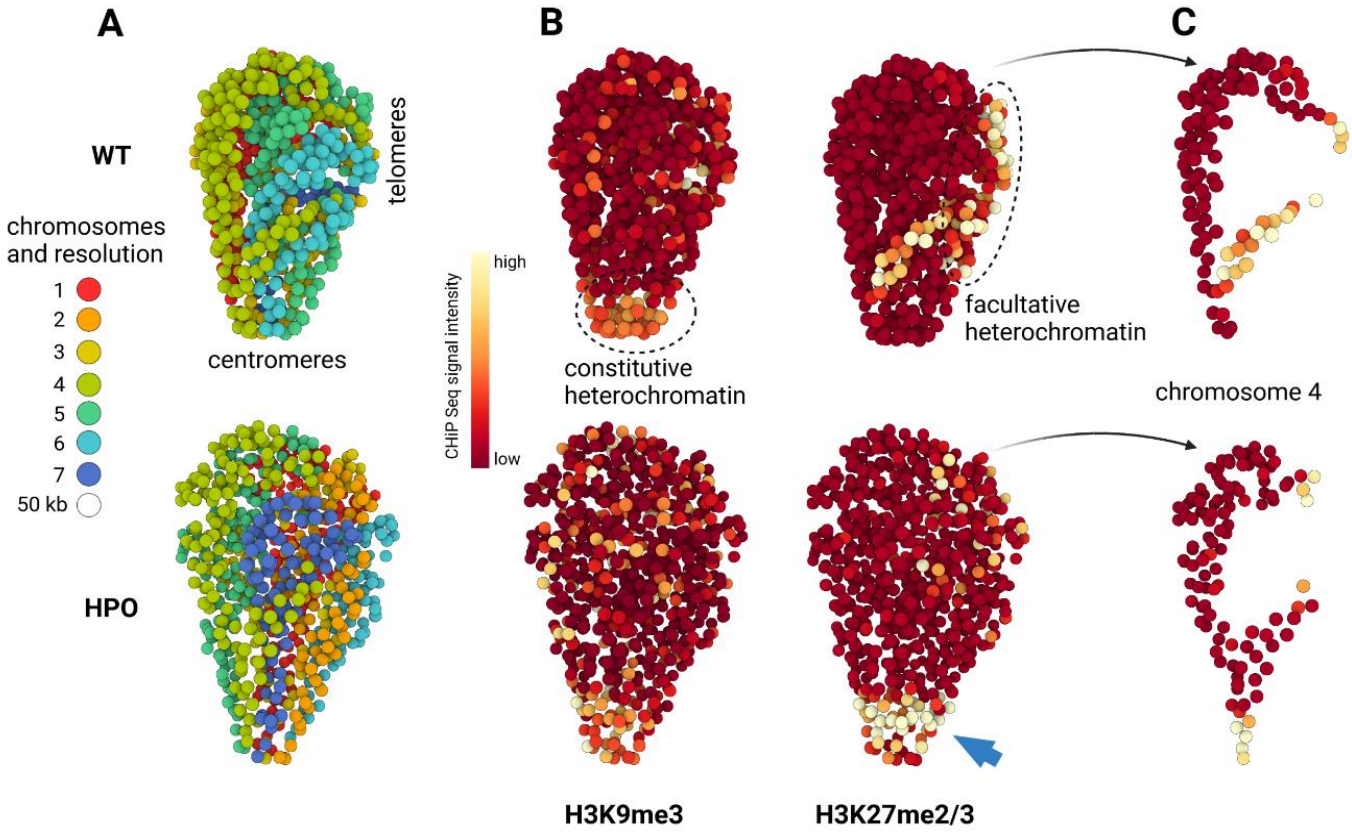


Figure 22 : Intégration visuelle des données ChIP-seq à l'aide de modèles 3D du génome de *N. crassa*. (A) Modèles 3D obtenus avec 3DGB à partir des données Hi-C de (Galazka et al.

2016), respectivement dans les souches de type sauvage (WT) et déficientes en HP1 (hpo) (voir le tableau 1 en annexe et la section Méthodes pour les détails techniques). Chaque sphère représente une région chromosomique de 50 kb et est colorée en fonction du chromosome auquel elle appartient (de 1 à 7). La colocalisation des centromères et des télomères est observée dans les deux modèles (centromères en bas et télomères à droite). **(B)** Mêmes modèles qu'en (A), en utilisant l'intensité du signal ChIP-seq comme code couleur pour les billes (voir Méthodes). Les données ChIP-seq proviennent de (Basenko et al. 2015) et se réfèrent à la localisation génomique des histones avec des modifications post-traductionnelles H3K9me3 et H3K27me2/3, respectivement. Dans le modèle WT, l'intensité du signal ChIP-seq pour H3K9me3 est particulièrement élevée (couleur jaune) autour des centromères (par rapport au reste du génome, couleur rouge foncé), tandis que l'intensité du signal ChIP-seq pour H3K27me3 est particulièrement élevée au niveau des télomères. Ces observations sont cohérentes avec les fonctions connues de l'hétérochromatine constitutive et facultative. Dans la souche hpo, des changements importants dans les intensités des signaux ChIP-seq sont observés avec notamment une relocalisation massive du signal de haute intensité de H3K27me3 vers les centromères (flèche bleue). **(C)** Isolation du chromosome 4, mettant mieux en évidence les changements d'intensité des signaux ChIP-seq H3K27me2/3 entre les souches WT et hpo.

Comme attendu d'après la littérature, nous avons observé une légère différence entre les deux organisations modélisées, essentiellement l'intensité de la compaction de la chromatine et la position relative des chromosomes 6 et 7 (**Figure 22A**). Pour prolonger cette observation, nous avons calculé pour chaque région de 50 kb représentée par une sphère dans les modèles 3D, l'intensité des signaux ChIP-seq issus des expériences ciblant les modifications post-traductionnelles des histones H3K9me3 et H3K27me2/3 (voir Méthodes). Les résultats sont présentés à la **Figure 22B** (génome entier) et à la **Figure 22C** (chromosome 4). Chez le WT, nous avons pu observer comme prévu une accumulation de marques histones H3K9me3 sur les centromères, correspondant à l'hétérochromatine constitutive, et une accumulation de marques histones H3K27me2/3 dans les régions sub-télomériques, correspondant à l'hétérochromatine facultative. Chez le mutant hpo, les signaux ChIP-seq sont apparus très désorganisés, en particulier pour les marques H3K27me2/3 qui est massivement relocalisés vers les centromères (flèche bleue, **Figure 22B**). Cette observation était attendue d'après la littérature, mais pour la première fois elle peut être vue à large échelle d'une manière simple et intégrée.

Cet exemple illustre l'intérêt des modèles 3D dans le contexte de l'intégration des données multiomiques. Si les données Hi-C nous permettent de mieux comprendre l'organisation des chromosomes, d'autres données omiques (comme ici ChIP-seq) nous éclairent sur les mécanismes de fonctionnement du génome. Il est important de garder à l'esprit que les images présentées sont issues d'une très grande quantité de données, à savoir plusieurs millions de mesures (fréquences de contacts, probabilités d'interactions, valeurs de qualité, etc.) Ces modèles 3D peuvent confirmer ou infirmer des hypothèses sur le fonctionnement global des génomes, et éventuellement conduire à la formulation de nouvelles hypothèses.

2.4 Autre intérêt de la modélisation 3D des génomes : la visualisation de la dynamique moléculaire de la chromatine chez *S. pombe*

Le but de cette application est d'évalué la possibilité de voir, avec des modèles 3D, des changements dans l'organisation des chromosomes au cours du cycle cellulaire. En effet, les chromosomes sont soumis à des contraintes structurelles importantes et subissent des réarrangements aux différentes étapes du cycle cellulaire (G2, M, G1 et S). Nous avons choisi d'explorer cette question en utilisant les données de Tanizawa et al. (Tanizawa et al., 2017). Les auteurs ont appliqué un protocole Hi-C in situ pour suivre l'organisation du génome de la levure *S. pombe* tout au long du cycle cellulaire. Ils ont observé que pendant la mitose, les chromosomes sont structurés en grands (300 kb à 1 Mb) et petits (30 - 40 kb) domaines, qui sont respectivement structurés par les complexes protéiques de la condensine et de la cohésine (Tanizawa et al., 2017). En se basant exclusivement sur leur interprétation des cartes de contact Hi-C, ils ont montré que si l'organisation mitotique en grands domaines se dissout progressivement au cours du cycle cellulaire, les petits domaines restent relativement stables. Ils ont également émis l'hypothèse que la conformation de Rabl était stable en interphase mais perturbée pendant la mitose. Avec cet ensemble de données, nous avons évalué notre capacité à observer ces caractéristiques structurelles dans les modèles 3D obtenus avec 3DGB, à différents stades du cycle cellulaire de *S. pombe*. Nous avons collecté les données Hi-C de l'article original (voir tableau 1 en annexe) et produit des modèles 3D (voir Méthodes) montrant l'organisation des chromosomes à différents moments du cycle de vie de *S. pombe*, après une synchronisation initiale des cellules en phase G2. Nos résultats sont présentés dans la **Figure 23**.

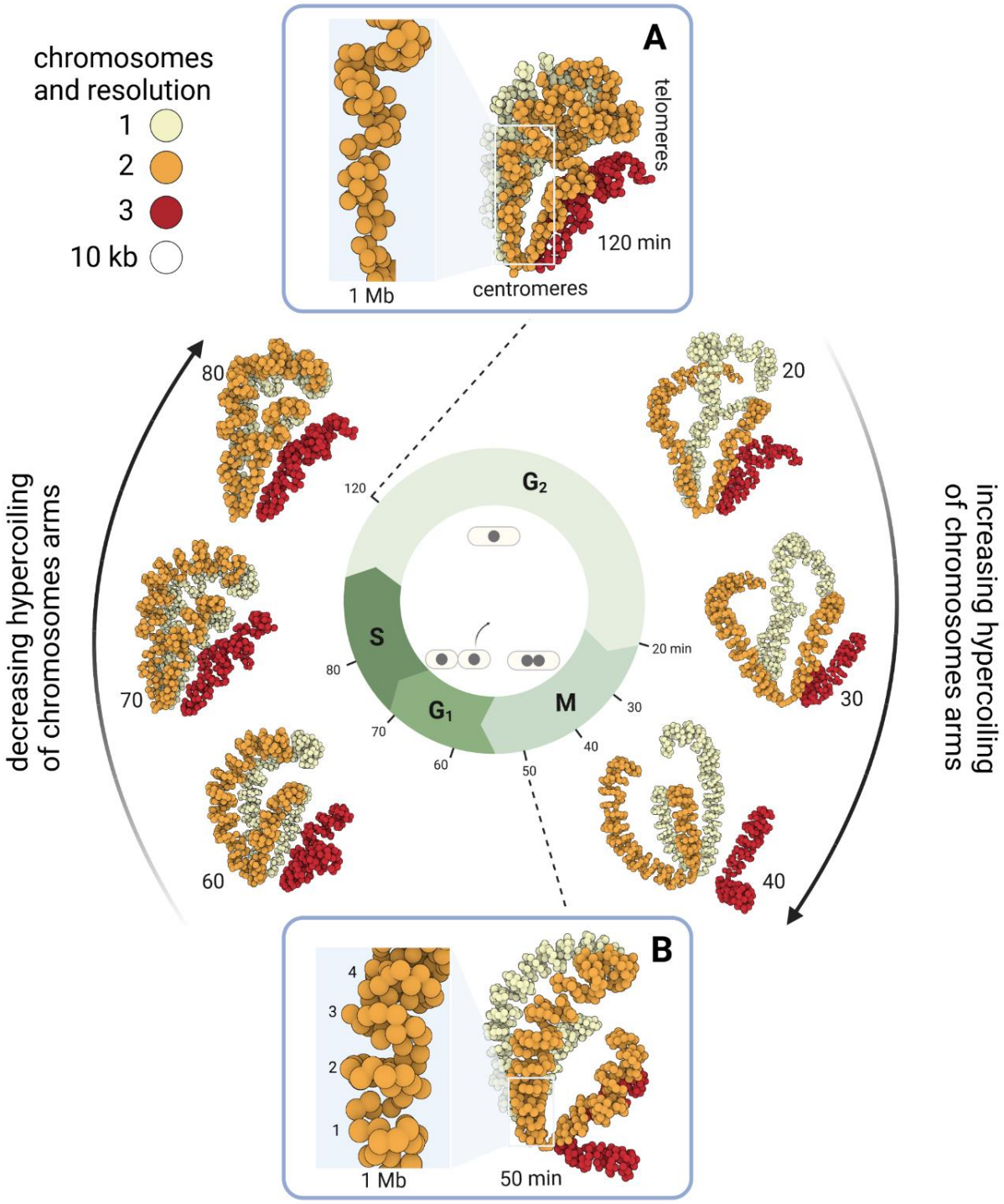


Figure 23 : Dynamique de l'organisation spatiale des chromosomes de *S. pombe* au cours du cycle cellulaire. Modèles 3D obtenus avec 3DGB à partir des données Hi-C de (Tanizawa et al. 2017). Chaque sphère représente une région chromosomique de 10 kb et est colorée en fonction du chromosome auquel elle appartient (de 1 à 3). Huit modèles sont présentés ici. Ils correspondent à différentes organisations des chromosomes de *S. pombe* au cours du cycle cellulaire de la levure. Après la synchronisation des cellules en phase G2, les points de temps suivants ont été analysés : 20, 30, 40, 50, 60, 70, 80 et 120 minutes. Leur correspondance avec les phases du cycle cellulaire est indiquée au milieu. Deux points de temps sont donc en phase

G2 (20 et 120 minutes), trois points de temps sont en phase M (30, 40 et 50 minutes), deux points de temps sont en phase G1 (60 et 70 minutes), et enfin un point de temps est en phase S. Les régions du centromère et du télomère sont annotées. Les centromères de tous les chromosomes sont constamment regroupés, quel que soit le moment du cycle cellulaire. Ceci est cohérent avec les observations précédentes (voir le texte principal). L'enroulement des bras chromosomiques, tel que décrit dans la littérature (Mizuguchi et al. 2014 ; Tanizawa et al. 2017), est visible sur les modèles 3D. Des zooms sont fournis dans les cases **(A)** et **(B)**, qui correspondent à des points temporels (respectivement 120 et 50 minutes) pour lesquels l'intensité de l'enroulement est minimale et maximale (voir le texte principal pour plus d'explications).

Comme prévu, la conformation de Rabl est visible, en particulier pendant l'interphase (modèles G1, S et G2, entre 60 et 120 minutes). Cependant, elle devient quelque peu désorganisée à la transition entre G2 et Mitose (**Figure 23**, modèle 20 min), même si le cluster de centromères reste stable pendant tout le cycle cellulaire. Les télomères des chromosomes 1 (couleur jaune) et 2 (couleur orange) restent également associés, tandis que le chromosome 3 (couleur rouge), dans lequel se trouvent les répétitions de l'ADN ribosomique (non visible dans le modèle), occupe une position externe cohérente avec la compartimentation de l'ADNr dans le nucléole. Nous avons également observé que le niveau de compaction de la chromatine varie au cours du cycle cellulaire, avec un minimum observé en G2 (**Figure 23A**) et un maximum observé en fin de mitose (**Figure 23B**). En effet, le modèle de 120 minutes montre des bras chromosomiques non enroulés, avec un pliage lâche de type globule (**Figure 23A**) alors que les bras chromosomiques du modèle de 50 minutes sont structurés en bobines régulières de ~250 kb (**Figure 23B**). Cette organisation plus compacte (voir le zoom sur la région de 1 Mb de la **Figure 23A-B**) et plus structurée (enroulement régulier) apparaît progressivement sur les modèles de 20, 30 et 40 minutes (**Figure 23**, flèche de droite). La structure mitotique hautement organisée s'estompe progressivement au cours des phases G1 et S pour revenir à la conformation Rabl (**Figure 23**, flèche gauche). Enfin, il est important de garder à l'esprit que les données Hi-C de chaque modèle ont été obtenues indépendamment, ce qui renforce la pertinence biologique des similitudes observées dans la position relative des chromosomes et le niveau de compaction.

En résumé, les modèles 3D obtenus avec 3DGB illustrent de manière cohérente les oscillations dans l'enroulement des bras chromosomiques au cours du cycle cellulaire chez *S. pombe*. Cela renforce les conclusions de l'étude originale de Tanizawa et al.

2.5 La structure de la chromatine chez *S. cerevisiae*, intégration multiomique et changement de perspective

Dix ans après le premier modèle 3D du génome de *S. cerevisiae*, (Costantino et al., 2020) ont produit de nouvelles données Hi-C et ChIP-seq, afin de mieux comprendre la relation entre les motifs identifiés des domaines chromatiniens (informations dérivées de l'analyse des données Hi-C) et les régions de résidence de la cohésine (informations dérivées de l'analyse des données ChIP-seq). Ils ont utilisé des cellules arrêtées en mitose et ont étudié à la fois la souche de type sauvage et les mutants mutés pour la cohésine ou ses régulateurs. Ils ont utilisé une amélioration récente de la technique Hi-C, appelée Micro-C XL (Hsieh et al., 2020), pour décrire finement la formation des boucles chromatiniennes. Cette méthode a l'avantage d'améliorer grandement la détection des interactions à courte portée (à l'échelle des nucléosomes), tout en permettant de détecter les interactions chromatiniennes du génome entier. Ceci est particulièrement intéressant pour une espèce comme *S. cerevisiae*, dans laquelle les chromosomes sont très petits par rapport aux autres espèces (les chromosomes de *S. cerevisiae* varient de 0,23 à 1,53 Mb seulement). En conséquence, ils ont montré que dans les cellules mitotiques de *S. cerevisiae*, les régions d'ancrage de la cohésine à haute résidence (appelées "CARs" et détectées par ChIP-seq) correspondent aux limites des boucles de chromatine à l'échelle du génome (appelées "domaines CAR" et détectées par Micro-C XL). Il s'agit d'une observation importante qui i) soutient l'idée que le génome de la levure est organisé en structures chromatiniennes récurrentes définies et délimitées par la cohésine et ii) que cette organisation spatiale des chromosomes peut avoir un impact sur les fonctions du génome, comme c'est le cas chez les mammifères. Cependant, dans leur article original, les auteurs ne présentent que des cartes de contact Hi-C pour étayer leurs interprétations. Pour aller plus loin dans cette visualisation et l'enrichir, nous avons collecté leurs données Micro-C XL et les avons réanalysées avec 3DGB, pour créer un modèle 3D actualisé de l'organisation spatiale des chromosomes de *S. cerevisiae* (voir Méthodes). Compte tenu de la très grande précision des données, nous nous attendions à pouvoir observer à la fois la configuration générale (territoires chromosomiques et conformation de Rabl) et les domaines chromatiniens fins (enroulement), en zoomant et en dézoomant à l'aide d'un logiciel de visualisation. Nos résultats sont présentés dans la **Figure 24**.

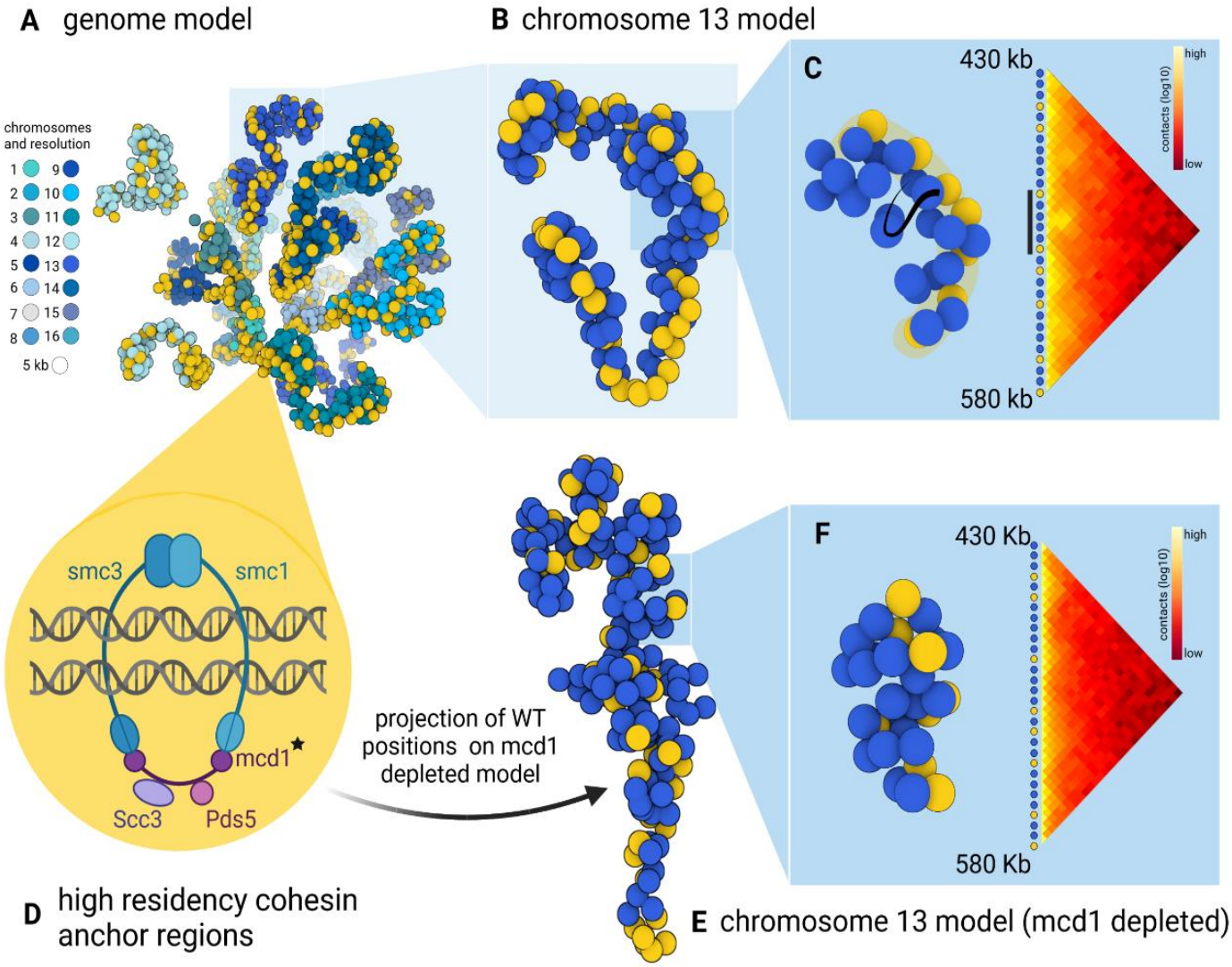


Figure 24 : Illustration du squelette 3D formé par la cohésine dans la chromatine de *S. cerevisiae*. (A) Modèle 3D de la chromatine de *S. cerevisiae*, obtenu avec 3DGB, en utilisant les données Micro-C XL de (Costantino et al. 2020). Chaque sphère représente une région chromosomique de 5 kb. La couleur jaune indique les régions génomiques avec une valeur élevée de l'intensité du signal ChIP-seq de la sous-unité de cohésine Mcd1p, dans les cellules mitotiquement arrêtées (voir Méthodes). **(B)** Chromosome 13 isolé de la structure globale, montrant l'alignement spatial des sites de liaison à l'ADN de la cohésine. La cohésine structure un « squelette » chromosomique sur lequel des boucles de chromatine peuvent se former. **(C)** Juxtaposition du modèle 3D et de la carte de contact associée, pour les billes situées entre 430 et 580 kb. Sur la structure 3D, le squelette de la cohésine est surligné en jaune clair. La visualisation de cet alignement spatial de la cohésine est complémentaire de l'information contenue dans la carte des contacts, c'est-à-dire que la fréquence des contacts est plus élevée à l'intérieur des régions génomiques délimitées par les billes jaunes. **(D)** Représentation schématique du complexe cohésine dans *S. cerevisiae*. La protéine Mcd1p, absente dans les cellules utilisées pour créer le modèle présenté en **(E)** et **(F)**, est indiquée par une étoile noire. Les régions génomiques qui sont attachées par la cohésine sont appelées CARs, c'est-à-dire régions d'ancrage de la cohésine à haute résidence, et sont visualisées en jaune sur les modèles

3D. (E) Chromosome 13 isolé de la structure globale, obtenue avec le mutant mcd1 (voir le texte principal). Les billes jaunes correspondent au squelette de cohésine telle qu'il a été identifié dans le WT (B). (F) Juxtaposition du modèle 3D et de la carte des contacts associée, pour les billes situées entre 430 et 580 kb. Il s'agit de la même région génomique exposée en (C).

Les données ChIP-seq ont également été réanalysées (voir Méthodes) pour localiser les CAR sur le nouveau modèle 3D. Ils sont représentés en jaune dans la **Figure 24**. Comme prévu dans les cartes de contact Hi-C présentées dans l'article original de Costantino et al., nous avons observé une distribution continue des CARs le long des chromosomes de *S. cerevisiae* (**Figure 24A**). Pour compléter cette observation, nous avons zoomé sur les chromosomes individuellement. Un exemple du chromosome 13 est présenté à la **Figure 24B**. Nous avons pu observer que les CARs étaient alignés dans l'espace et ce indépendamment de la longueur des boucles de chromatine dont ils représentent les frontières (**Figure 24C**). Cette observation est vraie pour tous les chromosomes. La modélisation 3D des contacts Hi-C est donc ici d'un intérêt significatif, révélant l'existence d'un squelette de cohésine, assurant la stabilité structurelle de l'organisation spatiale des chromosomes de *S. cerevisiae*. De plus, cette observation nouvelle est cohérente avec le modèle d'extrusion de boucle de chromatine par la cohésine proposé dans l'étude originale.

3 DISCUSSION

3.1 Résumé de la démarche et des principaux résultats de la thèse

Nous avons exposé en introduction comment l'étude du système cellulaire impose de travailler avec des modèles, particulièrement depuis que l'approche omique a considérablement augmenté la part mesurable de la complexité de ce système. La visualisation est une étape fondamentale de la méthode scientifique et permet de faciliter l'analyse de systèmes complexes. Il est donc crucial d'adapter la visualisation à la nouvelle échelle des données expérimentales. Pour certains systèmes mésoscopiques, l'accumulation de connaissances est telle que les premières modélisations graphiques d'une cellule entière sont possibles (voir modèle de *Mycoplasma genitalium* (Maritan et al., 2022)). Ces modèles graphiques proposent une vision synthétique d'une grande quantité de données, mais l'image finale n'est pas ancrée dans un contexte expérimental. Obtenir une lisibilité similaire dans le cas de l'analyse et l'intégration directe de données brutes est un enjeu complexe.

Nous avons exploré dans cette thèse l'utilisation des structures 3D de l'organisation des chromosomes inférées à partir des données Hi-C comme support à l'intégration d'autres données omiques. Nous avons dans un premier temps mené une intégration multiomique en utilisant un modèle existant du génome de *S. cerevisiae*. Nous avons ensuite assemblé le workflow 3DGB allant des données brutes Hi-C aux modèles 3D entièrement annotés et nous avons réanalysé des ensembles de données omiques publiques disponibles pour trois espèces modèles. Outre les propriétés connues de l'organisation spatiale de leurs chromosomes (conformation de Rabl, hyper-enroulement et territoires chromosomiques), nos résultats ont mis en évidence i) chez *Saccharomyces cerevisiae*, l'organisation linéaire des régions d'ancrage de la cohésine, qui sont alignées tout au long des bobines formées par le surenroulement des chromosomes, ii) chez *Schizosaccharomyces pombe*, les oscillations de l'enroulement des bras chromosomiques tout au long du cycle cellulaire et iii) chez *Neurospora crassa*, la relocalisation massive des marques épigénétiques dans un mutant d'un régulateur de l'hétérochromatine. La modélisation 3D des chromosomes offre de nouvelles possibilités d'intégration visuelle des données omiques. Cette perspective holistique favorise l'intuition et pose les bases de nouveaux concepts. Grâce à 3DGB, nous avons fourni des fichiers PDB enrichis pour une visualisation avancée avec le logiciel MolStar. Le workflow 3DGB est à la disposition d'autres scientifiques qui souhaiteraient explorer cette possibilité d'intégration multiomique.

3.2 L'intégration visuelle de données omiques en utilisant les modèles 3D des génomes

La pertinence des modèles 3D de l'organisation spatiale des chromosomes est une question importante. Au-delà de la production d'une "belle image", quelle est la valeur ajoutée de ces modèles pour les analystes de données Hi-C ? Peut-on vraiment avoir une idée de la structure du génome ? Il est en effet important de garder à l'esprit que ces modèles statistiques ne sont pas des scans 3D de l'intérieur des noyaux cellulaires. Ils sont inférés à partir de données expérimentales dont la qualité peut varier et avoir un impact significatif sur le modèle déduit. En définitive, malgré l'échelle des mesures omiques, ces modèles représentent des mesures de fréquences de contact observées à l'échelle de populations cellulaires (synchronisées ou non) et sont donc des représentations simplifiées de la réalité. Malgré cela, les trois exemples de génomes fongiques explorés ont montré que la visualisation de modèles 3D est profitable pour plusieurs raisons. Tout d'abord, elle donne une représentation globale de l'organisation des chromosomes. Alors que les cartes de contact sont très utiles pour identifier les structures locales de la chromatine (< 1 kb), les modèles 3D permettent un zoom arrière complet, à l'échelle du chromosome entier, donnant une visualisation globale du signal derrière l'ensemble des fréquences de contact.

Deuxièmement, les modèles 3D peuvent à la fois servir de visualisation finale des résultats d'une étude Hi-C, mais aussi de point de départ pour une intégration plus poussée des données. Les résultats que nous avons présentés pour les génomes de *S. pombe* sont un exemple de résultat de visualisation finale. Nos modèles révèlent dans l'espace 3D et à l'échelle du génome la dynamique d'enroulement (**Figure 24**) décrite dans l'étude initiale mais qui est difficile à conceptualiser à partir des seules cartes de contact.

Les résultats que nous avons présentés pour *N. crassa* et *S. cerevisiae* sont des exemples d'intégration de données omiques (**Figures 22 et 24**). Cette fois, les modèles 3D sont des points de départ pour une exploration supplémentaire. La coloration de la représentation 3D des modèles, basée sur des mesures quantitatives issues d'expériences "omiques", bien qu'il s'agisse d'une idée simple, est particulièrement efficace pour étudier des phénomènes globaux, tels que la relocalisation massive des marques d'histones dans un mutant, comme cela a été montré ici pour *N. crassa* ou bien l'organisation linéaire dans l'espace des régions de fixation de la cohésine chez *S. cerevisiae*.

3.3 Les limites de la modélisation des génomes viennent de la méthode de mesure de la Hi-C

Malgré sa cohérence et son efficacité d'illustration, la modélisation de l'organisation tridimensionnelle du génome basée sur les contacts présente certaines limites. Ces modèles restent une représentation 3D d'un réseau de contacts basé sur des données Hi-C, avec des régions chromosomiques modélisées comme des sphères indépendantes, chacune représentant de 5 000 à 50 000 paires de bases, en fonction du modèle (voir **Figures 22, 23, 24**). Cette résolution est en elle-même une simplification conséquente.

La profondeur de séquençage des données Hi-C et la résolution choisie pour leur analyse ont un impact direct sur la précision du modèle final : pour un nombre donné de "reads", plus la résolution est faible, moins les contacts inter chromosomiques sont détectés. Nous avons observé que lorsque la profondeur de séquençage est faible, les chromosomes dans les modèles 3D sont plus éloignés les uns des autres, par exemple dans les modèles mitotiques de *S. cerevisiae* et *S. pombe* (**Figure 23**). Dans une telle situation, il est difficile de faire la distinction entre le compactage biologique de la chromatine et les limitations dues à la technique expérimentale Hi-C, si l'on ne considère que le volume final occupé par un modèle. Cependant, des caractéristiques structurelles telles que le niveau d'enroulement sont encore visibles (**figures 23 et 24**) et peuvent donner une idée du niveau de compaction biologique de la chromatine, en compensant les limitations techniques (qui se traduisent par des chromosomes plus espacés).

Comme indiqué précédemment, une autre limite inhérente aux modèles 3D est qu'il s'agit d'images "statiques", correspondant à des "moyennes de population", d'un système biologique (l'organisation de la chromatine) connu pour être très dynamique et variable d'une cellule à l'autre. Ces artefacts existeront tant que les données Hi-C seront des mesures obtenues à partir de populations de cellules. Néanmoins, il est important de noter que les conformations de Rabl et les territoires chromosomiques ont été observés dans toutes les espèces étudiées ici, même avec le modèle de *N. crassa*, qui est une représentation moyenne dérivée d'expériences Hi-C réalisées sur une population de cellules asynchrones. Pour les modèles de *S. pombe*, comme les cellules étaient initialement synchronisées, les jeux de données Hi-C produits à différents stades du cycle cellulaire ont révélé plusieurs autres organisations intéressantes du génome dans les modèles 3D déduits (**Figure 23**). Nous avons ainsi réussi à rendre compte de la dynamique de la chromatine. En ce qui concerne la limite de la "moyenne de la population", les stratégies Hi-C de cellules uniques se développent (Stevens et al. 2017), ouvrant la perspective de créer des modèles génomiques 3D de cellules uniques.

Une dernière limite des modèles 3D actuels concerne les régions manquantes des génomes fongiques. À titre d'illustrations, les modèles présentés ici sont construits à

partir de séquences génomiques dans lesquelles les répétitions de l'ADNr ont été supprimées, empêchant ainsi la représentation correcte des caractéristiques structurelles du nucléole (expliquant l'espace vide que l'on peut observer dans le chromosome 12 des modèles de *S. cerevisiae*, **Figures 20, 24**).

Une autre simplification importante découle de la représentation conjointe des chromatides sœurs et de la complexité du travail avec des génomes diploïdes (dans les analyses de données Hi-C classiques, les mesures de contact ne peuvent pas être distinguées entre les séquences identiques). Des stratégies émergent pour résoudre les haplotypes (Oomen et al. 2020 ; Mitter et al. 2020) et les logiciels utilisés dans 3DGB dispose d'options adaptées à cet effet : à l'étape de la cartographie des reads, HiC-Pro peut construire des cartes de contact spécifiques à un allèle si l'information SNP est fournie. Au stade de la modélisation 3D, la méthode de Poisson de Pastis a été étendue à la résolution des haplotypes. Même si cela nous permet d'imaginer la création de modèles plus complets dans le futur, il est important de garder à l'esprit que les génomes fongiques examinés ici, i) restent haploïdes pendant le cycle cellulaire et ii) ont des chromatides sœurs étroitement maintenues ensemble par cohésine. La représentation reste donc informative.

Dans l'ensemble, les limites soulignent la dépendance de la modélisation 3D à l'égard de la qualité des données brutes Hi-C expérimentales, en plus de la méthode de modélisation. Comme exposé en introduction, les méthodes de mesure restent la limite de la capacité de modélisation, même si elles ont énormément évolué dernièrement. Les améliorations techniques introduites à un rythme rapide réduisent progressivement ces limites.

3.4 Les nombreuses possibilités d'évolution de l'intégration visuelle multiomiques

Il est important de noter que, même si le plafond technique n'est pas encore atteint, nous avons déjà réussi à mettre en évidence des caractéristiques importantes de génomes fongiques d'une manière nouvelle en réanalysant des ensembles de données publiques. Pour *N. crassa*, les six modèles 3D produits par 3DGB ont condensé les informations de cinq articles de recherche en une riche illustration à grande échelle (**Figure 22**). Ils offrent de nouvelles possibilités d'intégration visuelle des données omiques : alors que l'analyse Hi-C implique souvent une logique de "zoom avant", en se concentrant sur des régions précises d'une carte de contact, la modélisation 3D complète cette logique avec une vision "zoom arrière".

Nous avons exploré l'intégration à large échelle de données épigénétiques ChIP-seq dans le contexte 3D de la chromatine, mais toute combinaison appropriée d'ensembles de données omiques peut être réalisée en utilisant le modèle de génome comme support visuel. Par exemple, nous étudions actuellement l'interpolation des gènes sur la structure 3D du génome. 3DGB pourrait être adapté pour fournir un modèle au

niveau des gènes (chaque gène étant modélisé par une sphère), afin de permettre la cartographie des données omiques dépendantes des gènes (RNA-seq) sur la structure chromosomique. La position des gènes constitue un moyen pratique de tracer des données omiques sur l'organisation 3D de la chromatine. Il pourrait être également intéressant d'intégrer des informations sur l'occupation des protéines ou les réseaux d'interaction des facteurs de transcription. Considérer les domaines de la chromatine comme des boucles d'une bobine plutôt que comme des entités distinctes (**Figure 24**) pourrait ouvrir de nouvelles perspectives sur la régulation des gènes entre les domaines.

4 METHODES

4.1 Assemblage de 3D-genome-builder (3DGB) et gestion du workflow

4.1.1 Détails techniques

3DGB a été assemblé avec le système de gestion de workflow open-source Snakemake (Mölder et al., 2021), qui automatise les différentes étapes d'une analyse de données dans un langage Python lisible par l'homme. Les environnements logiciels ont été déployés et isolés dans un environnement Conda (pour Pastis et tous les scripts Python personnalisés) et un conteneur Singularity (pour HiC-Pro). 3DGB nécessite les inputs suivants : Les fichiers Hi-C FASTQ (fournis sous forme d'ID SRA ou de fichiers FASTQ locaux), une séquence génomique de référence fournie dans un fichier FASTA (sans l'ADN mitochondrial), un ou plusieurs motifs de sites de restriction enzymatiques, et enfin, les valeurs de la résolution de sortie pour dessiner la carte des contacts et générer des modèles 3D.

4.1.2 Principales étapes de l'analyse

Les premières étapes du workflow formatent les informations nécessaires à l'étape de mapping des reads : (i) les fichiers FASTQ sont téléchargés et compressés s'ils ne sont pas déjà fournis par l'utilisateur ; (ii) une liste de fragments dérivés du fichier FASTA du génome et du motif du site de restriction de l'enzyme est générée par HiC-Pro (Servant et al., 2015) version 3.1.0 ; (iii) la taille de chaque chromosome du fichier FASTA du génome est extraite avec un script personnalisé et (iv) le génome de référence est indexé avec Bowtie2 (Langmead et Salzberg 2012) version 2.4.4. Ensuite, HiC-Pro génère un ensemble de "paires de reads valides", qui sont les reads pertinents utilisés pour générer et normaliser les fréquences (ou comptages) de contacts (voir introduction). Pour une résolution donnée, ces valeurs sont ensuite utilisées pour calculer des cartes de contact (heatmaps) et pour servir d'entrée à Pastis (Varoquaux et al., 2021) (version du 21 juillet 2021) pour construire un modèle 3D consensuel. Dans 3DGB, la méthode Pastis-NB est utilisée. Il calcule itérativement des modèles 3D de l'organisation des chromosomes, par modélisation binomiale négative du nombre de contacts. Pour une meilleure reproductibilité et portabilité, HiC-Pro est utilisé dans une image Singularity fournie par les auteurs du logiciel. Pastis est installé dans un environnement conda. Dans 3DGB, une étape finale pour affiner le modèle 3D fourni par Pastis est disponible. Elle consiste à prédire les coordonnées de sphères supplémentaires, pour lesquelles les calculs initiaux manquaient dans le modèle 3D de Pastis. Les coordonnées manquantes se trouvent généralement dans les régions centromériques et télomériques. Les prédictions sont effectuées par interpolation, en utilisant des « monotonic cubic splines », comme implémenté par la méthode pchip dans la bibliothèque Python SciPy (Virtanen et al., 2020). Notons que les sphères dont les coordonnées manquantes sont localisées aux extrémités des chromosomes sont

écartées et que les sphères dont les coordonnées sont aberrantes sont filtrées en fonction d'un seuil appliqué à la valeur de la distance euclidienne calculée entre les sphères voisines.

4.1.3 Sorties 3DGB

Les modèles 3D sont stockés dans des fichiers PDB. À partir du modèle brut produit par Pastis, 3DGB annote les chromosomes par des numéros, sur la base de la séquence de référence spécifiée dans les inputs de 3DGB. Des exemples de modèles, entièrement annotés avec 3DGB, sont fournis sous forme de fichiers PDB dans le GitHub. Dans les fichiers PDB, l'annotation du chromosome est présente dans l'identifiant du résidu (1, 2, 3...), dans l'identifiant de la chaîne (A, B, C...) et dans le nom du résidu (C01, C02, C03...). Cette adaptation des champs de résidus et de chaînes standard dans les fichiers PDB est particulièrement utile, car elle permet de visualiser immédiatement les chromosomes à l'aide d'un logiciel de visualisation. Enfin, des valeurs quantitatives peuvent être associées aux sphères du modèle 3D (par exemple l'intensité du signal ChIP-seq), en utilisant le champ du facteur B dans les fichiers PDB. La plupart des outils de visualisation de fichiers PDB peuvent afficher les facteurs B en couleur sur les structures. C'est le cas de Mol*, utilisé pour créer les images présentées dans cet article.

4.2 Analyses des jeux de données expérimentaux

4.2.1 Accès aux données brutes de la base de données SRA

Tous les fichiers FASTQ bruts Hi-C-seq, Micro-C XL-seq et ChIP-seq ont été téléchargés à partir de la base de données SRA, voir le tableau 1 en annexe pour tous les identifiants SRA. Les génomes de référence de *S. cerevisiae* et *S. pombe* ont été obtenus à partir de la base de données NCBI Genome, respectivement version R64 (S288C) et ASM294v2.19. La version nc14 du génome de *N. crassa* a été téléchargée à partir des données supplémentaires de (Rodriguez et al., 2022) sur NCBI GEO, GEO dataset GSE173593. Des fichiers FASTA contenant uniquement des séquences chromosomiques ont été générés en tant qu'inputs de 3DGB.

4.2.2 L'application de 3DGB pour la création de modèles

3DGB a été appliqué avec des paramètres par défaut (voir le tableau 1 en annexe pour tous les détails des FASTQ) et les fichiers de configuration peuvent être consultés sur Zenodo et Github. Pour *S. cerevisiae*, deux modèles ont été générés à une résolution de 5 kb à partir de quatre fichiers FASTQ (total de 31 028 357 paires de reads valides) et les informations ChiP-Seq de deux fichiers FASTQ ont été intégrées. Pour *S. pombe*, huit modèles à une résolution de 10 kb ont été construits à partir de dix-neuf fichiers FASTQ (total de 136 321 555 paires valides). Pour N. crassa, trois modèles ont été créés à des résolutions de 50 kb et 10 kb, à partir de quatorze fichiers FASTQ (total de 115 588 951 paires valides) et les informations ChiP-seq de neuf fichiers FASTQ ont été intégrées. Les résolutions choisies pour les modèles ont été définies en fonction de la

technique utilisée (Hi-C ou Micro-C XL), du nombre de reads et de leur qualité.

4.2.3 Évaluation de la stabilité du modèle 3D au bruit aléatoire

La matrice de fréquence de contact du mutant hpo de *N. crassa* a été utilisée comme "référence" pour générer des "matrices de fréquence de contact bruitées". Le mutant hpo a été choisi parce qu'il possède le plus petit nombre de paires de lectures valides (2 922 526) et que cet échantillon est le plus susceptible d'être sensible au bruit. Soit x_{ij} le nombre de comptages dans la matrice de fréquence de contact de cet échantillon, avec i et j les indices des sphères allant de 1 à n = 825, avec n le nombre total de sphères pour cet échantillon. Pour créer une matrice de fréquence de contact "altérée" ou "bruitée" de *N. crassa*, nous avons fixé $y_{ij} = x_{ij} + e_{ij}$ le nombre de comptages dans la matrice de fréquence de contact bruitée, où e_{ij} est une valeur échantillonnée à partir d'une distribution de Poisson avec le paramètre λ, indépendamment pour chaque paire de sphères (i, j). On peut noter que les matrices de fréquence des contrats sont symétriques et que $e_{ij} = e_{ji}$ pour chaque paire de sphère. La distribution de Poisson a été utilisée car elle modélise le bruit de fond, c'est-à-dire la variabilité des nombres de reads qui est due au processus d'échantillonnage des reads plutôt qu'à des variations biologiques (Anders & Huber, 2010). Le niveau de bruit (c'est-à-dire le paramètre moyen λ de la distribution de Poisson utilisée pour générer le « shot noise ») varie entre les valeurs suivantes : 0,1, 5, 10, 15, 20 et 110. Pour chaque valeur de λ, 50 répétitions ont été générées, c'est-à-dire 50 matrices de fréquence de contact bruitées. Pour évaluer la stabilité des structures 3D, les RMSD ont été calculées entre la structure 3D déduite de la matrice de fréquence de contact de *N. crassa* de référence (la matrice de contact brute avec des comptes x_{ij}) et la structure 3D déduite des cartes de *N. crassa* altérées (la matrice de fréquence de contact bruitée avec des comptes y_{ij}). Cette procédure conduit à 50 valeurs de RMSD pour chaque valeur de λ. La **Figure 21** représente les valeurs RMSD pour λ allant de 0,1 à 50. Au-delà de la valeur λ = 50, les valeurs RMSD atteignent un seuil et n'augmentent plus avec la valeur de λ. Les RMSD entre chacun des sept chromosomes de *N. crassa* ont également été calculés, pour chaque valeur de λ et chacune des 50 répétitions.

4.2.4 Intégration visuelle des données omiques

Le même processus d'analyse a été utilisé pour l'intégration des données ChIP-seq de *N. crassa* et *S. cerevisiae*. Pour réaliser l'alignement, Bowtie2 a été utilisé avec les paramètres par défaut voir (tableau 1 en annexe). Le programme samtools (Danecek et al., 2021) a été utilisé pour trier et indexer les fichiers SAM en fichiers .bam, et le programme bamCoverage a été utilisé pour générer des fichiers .bedgraph de 5 kb ou 50 kb (pour *S. cerevisiae* et *N. crassa* respectivement) avec normalisation RPKM. Chaque bin dans le .bedgraph correspond à une sphère du modèle 3D. Il est à noter que la résolution du .bedgraph et du modèle 3D est la même, ce qui permet l'intégration des données dans le fichier PDB. Pour *S. cerevisiae*, le seuil a été fixé à 80 comptages après normalisation, convertissant le signal continu de ChIP-seq en un signal binaire mettant

en évidence les CAR à forte résidence (Costantino et al. 2020). Pour *N. crassa*, les données ChIP-seq ont été conservées en continu pour mettre en évidence la distribution des marques épigénétiques. Les valeurs inférieures à 20 comptages sont éliminées pour la visualisation.

4.2.5 Accès aux données

Le workflow 3DGB est disponible sur GitHub : https://github.com/data-fun/3d-genome-builder . Les fichiers .PDB des treize modèles présentés dans cette étude sont disponibles sur Zenodo 10.5281/zenodo.7740302, ainsi que des GIF animés pour quatre d'entre eux. Les fichiers .YML (fichiers de configuration) pour la génération de ces treize modèles sont disponibles sur GitHub et sur Zenodo.

5 REFERENCES

Alberghina, L., Mavelli, G., Drovandi, G., Palumbo, P., Pessina, S., Tripodi, F., Coccetti, P., & Vanoni, M. (2012). Cell growth and cell cycle in Saccharomyces cerevisiae: Basic regulatory design and protein–protein interaction network. *Biotechnology Advances*, *30*(1), 52–72. https://doi.org/10.1016/j.biotechadv.2011.07.010

Alon, U. (2007). Network motifs: Theory and experimental approaches. *Nature Reviews Genetics*, *8*(6), 450–461. https://doi.org/10.1038/nrg2102

Anders, S., & Huber, W. (2010). *Differential expression analysis for sequence count data*.

Aramayo, R., & Selker, E. U. (2013). Neurospora crassa, a Model System for Epigenetics Research. *Cold Spring Harbor Perspectives in Biology*, *5*(10), a017921–a017921. https://doi.org/10.1101/cshperspect.a017921

Arivaradarajan, P., & Misra, G. (Eds.). (2018). *Omics Approaches, Technologies And Applications: Integrative Approaches For Understanding OMICS Data*. Springer Singapore. https://doi.org/10.1007/978-981-13-2925-8

Asbury, T. M., Mitman, M., Tang, J., & Zheng, W. J. (2010). Genome3D: A viewer-model framework for integrating and visualizing multi-scale epigenomic information within a three-dimensional genome. *BMC Bioinformatics*, *11*(1), 444. https://doi.org/10.1186/1471-2105-11-444

Ay, F., Bunnik, E. M., Varoquaux, N., Bol, S. M., Prudhomme, J., Vert, J.-P., Noble, W. S., & Le Roch, K. G. (2014). Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between

genome architecture and gene expression. *Genome Research*, *24*(6), 974–988. https://doi.org/10.1101/gr.169417.113

Bánki, O., Roskov, Y., Döring, M., Ower, G., Vandepitte, L., Hobern, D., Remsen, D., Schalk, P., DeWalt, R. E., Keping, M., Miller, J., Orrell, T., Aalbu, R., Abbott, J., Adlard, R., Adriaenssens, E. M., Aedo, C., Aescht, E., Akkari, N., … Legume Phylogeny Working Group (LPWG). (2023). *Catalogue of Life Checklist* (Version 2023-05-15). Catalogue of Life. https://doi.org/10.48580/dfs6

Barabási, A.-L., & Oltvai, Z. N. (2004). Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, *5*(2), 101–113. https://doi.org/10.1038/nrg1272

Barnett, J. A. (2007). A history of research on yeasts 10: Foundations of yeast genetics1. *Yeast*, *24*(10), 799–845. https://doi.org/10.1002/yea.1513

Barrett, T., Wilhite, S. E., Ledoux, P., Evangelista, C., Kim, I. F., Tomashevsky, M., Marshall, K. A., Phillippy, K. H., Sherman, P. M., Holko, M., Yefanov, A., Lee, H., Zhang, N., Robertson, C. L., Serova, N., Davis, S., & Soboleva, A. (2012). NCBI GEO: Archive for functional genomics data sets—update. *Nucleic Acids Research*, *41*(D1), D991–D995. https://doi.org/10.1093/nar/gks1193

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2010). GenBank. *Nucleic Acids Research*, *38*(suppl_1), D46–D51. https://doi.org/10.1093/nar/gkp1024

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

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*(2), giab008. https://doi.org/10.1093/gigascience/giab008

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

Dickerson, R. E., Drew, H. R., Conner, B. N., Wing, R. M., Fratini, A. V., & Kopka, M. L. (1982). The Anatomy of A-, B-, and Z-DNA. *Science*, *216*(4545), 475–485. https://doi.org/10.1126/science.7071593

Djekidel, M. N., Wang, M., Zhang, M. Q., & Gao, J. (2017). HiC-3DViewer: A new tool to visualize Hi-C data in 3D space. *Quantitative Biology*, *5*(2), 183–190. https://doi.org/10.1007/s40484-017-0091-8

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), 363–367. https://doi.org/10.1038/nature08973

Eisen, M. B., Spellman, P. T., Brown, P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences*, *95*(25), 14863–14868. https://doi.org/10.1073/pnas.95.25.14863

Ekman, D., Light, S., Björklund, Å. K., & Elofsson, A. (2006). What properties characterize

the hub proteins of the protein-protein interaction network of Saccharomyces cerevisiae? *Genome Biology*, *7*(6), R45. https://doi.org/10.1186/gb-2006-7-6-r45

Engel, S. R., Dietrich, F. S., Fisk, D. G., Binkley, G., Balakrishnan, R., Costanzo, M. C., Dwight, S. S., Hitz, B. C., Karra, K., Nash, R. S., Weng, S., Wong, E. D., Lloyd, P., Skrzypek, M. S., Miyasato, S. R., Simison, M., & Cherry, J. M. (2014). The Reference Genome Sequence of *Saccharomyces cerevisiae*: Then and Now. *G3 Genes|Genomes|Genetics*, *4*(3), 389–398. https://doi.org/10.1534/g3.113.008995

Farr, S. E., Woods, E. J., Joseph, J. A., Garaizar, A., & Collepardo-Guevara, R. (2021). Nucleosome plasticity is a critical element of chromatin liquid–liquid phase separation and multivalent nucleosome interactions. *Nature Communications*, *12*(1), 2883. https://doi.org/10.1038/s41467-021-23090-3

Feric, M., Vaidya, N., Harmon, T. S., Mitrea, D. M., Zhu, L., Richardson, T. M., Kriwacki, R. W., Pappu, R. V., & Brangwynne, C. P. (2016). Coexisting Liquid Phases Underlie Nucleolar Subcompartments. *Cell*, *165*(7), 1686–1697. https://doi.org/10.1016/j.cell.2016.04.047

Feschotte, C., Jiang, N., & Wessler, S. R. (2002). Plant transposable elements: Where genetics meets genomics. *Nature Reviews Genetics*, *3*(5), 329–341. https://doi.org/10.1038/nrg793

Firestein, S. (2001). How the olfactory system makes sense of scents. *Nature*, *413*(6852), 211–218. https://doi.org/10.1038/35093026

Franz, M., Lopes, C. T., Huck, G., Dong, Y., Sumer, O., & Bader, G. D. (2015). Cytoscape.js: A graph theory library for visualisation and analysis. *Bioinformatics*, btv557.

https://doi.org/10.1093/bioinformatics/btv557

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C. B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., … Birren, B. (2003). The genome sequence of the filamentous fungus Neurospora crassa. *Nature*, *422*(6934), 859–868. https://doi.org/10.1038/nature01554

Galazka, J. M., Klocko, A. D., Uesaka, M., Honda, S., Selker, E. U., & Freitag, M. (2016). Neurospora chromosomes are organized by blocks of importin alpha-dependent heterochromatin that are largely independent of H3K9me3. *Genome Research*, *26*(8), 1069–1080. https://doi.org/10.1101/gr.203182.115

Gasch, A. P., Spellman, P. T., Kao, C. M., Carmel-Harel, O., Eisen, M. B., Storz, G., Botstein, D., & Brown, P. O. (2000). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes□D. *Molecular Biology of the Cell*, *11*.

Gayon, J. (2016). From Mendel to epigenetics: History of genetics. *Comptes Rendus Biologies*, *339*(7–8), 225–230. https://doi.org/10.1016/j.crvi.2016.05.009

Gehlenborg, N., O'Donoghue, S. I., Baliga, N. S., Goesmann, A., Hibbs, M. A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D., & Gavin, A.-C. (2010). Visualization of omics data for systems biology. *Nature Methods*, *7*(S3), S56–S68. https://doi.org/10.1038/nmeth.1436

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*

(New York, N.Y.), *274*(5287), 546, 563–567. https://doi.org/10.1126/science.274.5287.546

Goodsell, D. S. (2009). *The machinery of life* (2nd ed., corrected). Copernicus Books.

Goodsell, D. S., Olson, A. J., & Forli, S. (2020). Art and Science of the Cellular Mesoscale. *Trends in Biochemical Sciences*, *45*(6), 472–483. https://doi.org/10.1016/j.tibs.2020.02.010

Guido Van Rossum & Fred L. Drake. (2009). *Python 3 Reference Manual*. CreateSpace.

Harrow, J., Drysdale, R., Smith, A., Repo, S., Lanfear, J., & Blomberg, N. (2021). ELIXIR: Providing a sustainable infrastructure for life science data at European scale. *Bioinformatics*, *37*(16), 2506–2511. https://doi.org/10.1093/bioinformatics/btab481

Hartwell, L. H., Culotti, J., & Reid, B. (1970). Genetic Control of the Cell-Division Cycle in Yeast, I. Detection of Mutants. *Proceedings of the National Academy of Sciences*, *66*(2), 352–359. https://doi.org/10.1073/pnas.66.2.352

Hartwell, L. H., & Weinert, T. A. (1989). Checkpoints: Controls That Ensure the Order of Cell Cycle Events. *Science*, *246*(4930), 629–634. https://doi.org/10.1126/science.2683079

Hasin, Y., Seldin, M., & Lusis, A. (2017). Multi-omics approaches to disease. *Genome Biology*, *18*(1), 83. https://doi.org/10.1186/s13059-017-1215-1

Ho, B., Baryshnikova, A., & Brown, G. W. (2018). Unification of Protein Abundance Datasets Yields a Quantitative Saccharomyces cerevisiae Proteome. *Cell Systems*, *6*(2), 192-205.e3. https://doi.org/10.1016/j.cels.2017.12.004

Hoencamp, C., & Rowland, B. D. (2023). Genome control by SMC complexes. *Nature Reviews Molecular Cell Biology*, 1–18. https://doi.org/10.1038/s41580-023-00609-8

Hoffman, C. S., Wood, V., & Fantes, P. A. (2015). An Ancient Yeast for Young Geneticists: A Primer on the *Schizosaccharomyces pombe* Model System. *Genetics*, *201*(2), 403–423. https://doi.org/10.1534/genetics.115.181503

Howe, K. L., Achuthan, P., Allen, J., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Azov, A. G., Bennett, R., Bhai, J., Billis, K., Boddu, S., Charkhchi, M., Cummins, C., Da Rin Fioretto, L., Davidson, C., Dodiya, K., El Houdaigui, B., Fatima, R., … Flicek, P. (2021). Ensembl 2021. *Nucleic Acids Research*, *49*(D1), D884–D891. https://doi.org/10.1093/nar/gkaa942

Hsieh, T.-H. S., Cattoglio, C., Slobodyanyuk, E., Hansen, A. S., Rando, O. J., Tjian, R., & Darzacq, X. (2020). Resolving the 3D Landscape of Transcription-Linked Mammalian Chromatin Folding. *Molecular Cell*, *78*(3), 539-553.e8. https://doi.org/10.1016/j.molcel.2020.03.002

Huxley, J. S. (1924). The Mechanism of Mendelian Heredity. *Nature*, *113*(2841), Article 2841. https://doi.org/10.1038/113518a0

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. https://doi.org/10.1038/nature03001

Jamieson, K., Wiles, E. T., McNaught, K. J., Sidoli, S., Leggett, N., Shao, Y., Garcia, B. A., & Selker, E. U. (2016). Loss of HP1 causes depletion of H3K27me3 from facultative

heterochromatin and gain of H3K27me2 at constitutive heterochromatin. *Genome Research*, *26*(1), 97–107. https://doi.org/10.1101/gr.194555.115

Jenuwein, T., & Allis, C. D. (2001). Translating the Histone Code. *Science*, *293*(5532), 1074–1080. https://doi.org/10.1126/science.1063127

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

Kakui, Y., Barrington, C., Barry, D. J., Gerguri, T., Fu, X., Bates, P. A., Khatri, B. S., & Uhlmann, F. (2020). Fission yeast condensin contributes to interphase chromatin organization and prevents transcription-coupled DNA damage. *Genome Biology*, *21*(1), 272. https://doi.org/10.1186/s13059-020-02183-0

Karger, B. L., & Guttman, A. (2009). DNA sequencing by CE. *ELECTROPHORESIS*, *30*(S1), S196–S202. https://doi.org/10.1002/elps.200900218

Keith, D. A., Ferrer-Paris, J. R., Nicholson, E., Bishop, M. J., Polidoro, B. A., Ramirez-Llodra, E., Tozer, M. G., Nel, J. L., Mac Nally, R., Gregr, E. J., Watermeyer, K. E., Essl, F., Faber-Langendoen, D., Franklin, J., Lehmann, C. E. R., Etter, A., Roux, D. J., Stark, J. S., Rowland, J. A., … Kingsford, R. T. (2022). A function-based typology for Earth's ecosystems. *Nature*, *610*(7932), 513–518. https://doi.org/10.1038/s41586-022-05318-4

Kim, K.-D. (2021). Potential roles of condensin in genome organization and beyond in fission yeast. *Journal of Microbiology*, *59*(5), 449–459. https://doi.org/10.1007/s12275-021-1039-2

King, Z. A., Dräger, A., Ebrahim, A., Sonnenschein, N., Lewis, N. E., & Palsson, B. O. (2015). Escher: A Web Application for Building, Sharing, and Embedding Data-Rich Visualizations of Biological Pathways. *PLOS Computational Biology*, *11*(8), e1004321. https://doi.org/10.1371/journal.pcbi.1004321

Klocko, A. D., Ormsby, T., Galazka, J. M., Leggett, N. A., Uesaka, M., Honda, S., Freitag, M., & Selker, E. U. (2016). Normal chromosome conformation depends on subtelomeric facultative heterochromatin in *Neurospora crassa*. *Proceedings of the National Academy of Sciences*, *113*(52), 15048–15053. https://doi.org/10.1073/pnas.1615546113

Lajoie, B. R., Dekker, J., & Kaplan, N. (2015). The Hitchhiker's guide to Hi-C analysis: Practical guidelines. *Methods*, *72*, 65–75. https://doi.org/10.1016/j.ymeth.2014.10.031

Leinonen, R., Sugawara, H., Shumway, M., & on behalf of the International Nucleotide Sequence Database Collaboration. (2011). The Sequence Read Archive. *Nucleic Acids Research*, *39*(Database), D19–D21. https://doi.org/10.1093/nar/gkq1019

Li, D., Harrison, J. K., Purushotham, D., & Wang, T. (2022). Exploring genomic data coupled with 3D chromatin structures using the WashU Epigenome Browser. *Nature Methods*, *19*(8), 909–910. https://doi.org/10.1038/s41592-022-01550-y

Li, J., Zhang, W., & Li, X. (2018). 3D Genome Reconstruction with ShRec3D+ and Hi-C Data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *15*(2), 460–468. https://doi.org/10.1109/TCBB.2016.2535372

Lieberman-Aiden, E., Van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling,

A., Amit, I., Lajoie, B. R., Sabo, P. J., Dorschner, M. O., 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

Love, M. I., Huber, W., & Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, *15*(12), 550. https://doi.org/10.1186/s13059-014-0550-8

Lu, Y., Zhou, G., Ewald, J., Pang, Z., Shiri, T., & Xia, J. (2023). MicrobiomeAnalyst 2.0: Comprehensive statistical, functional and integrative analysis of microbiome data. *Nucleic Acids Research*, gkad407. https://doi.org/10.1093/nar/gkad407

Malnic, B., Hirono, J., Sato, T., & Buck, L. B. (1999). *Combinatorial Receptor Codes for Odors*.

Marasco, L. E., & Kornblihtt, A. R. (2023). The physiology of alternative splicing. *Nature Reviews Molecular Cell Biology*, *24*(4), Article 4. https://doi.org/10.1038/s41580-022-00545-z

Maritan, M., Autin, L., Karr, J., Covert, M. W., Olson, A. J., & Goodsell, D. S. (2022). Building Structural Models of a Whole Mycoplasma Cell. *Journal of Molecular Biology*, *434*(2), 167351. https://doi.org/10.1016/j.jmb.2021.167351

Matejka, J., & Fitzmaurice, G. (2017). Same Stats, Different Graphs: Generating Datasets with Varied Appearance and Identical Statistics through Simulated Annealing. *Proceedings of the 2017 CHI Conference on Human Factors in Computing*

Systems, 1290–1294. https://doi.org/10.1145/3025453.3025912

McClintock, B. (1956). Controlling Elements and the Gene. *Cold Spring Harbor Symposia on Quantitative Biology*, *21*(0), 197–216. https://doi.org/10.1101/SQB.1956.021.01.017

McClintock, B. (1961). Some Parallels Between Gene Control Systems in Maize and in Bacteria. *The American Naturalist*, *95*(884), 265–277. https://doi.org/10.1086/282188

McGuffee, S. R., & Elcock, A. H. (2010). Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm. *PLoS Computational Biology*, *6*(3), e1000694. https://doi.org/10.1371/journal.pcbi.1000694

Metzker, M. L. (2010). Sequencing technologies—The next generation. *Nature Reviews Genetics*, *11*(1), 31–46. https://doi.org/10.1038/nrg2626

Misteli, T. (2020). The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell*, *183*(1), 28–45. https://doi.org/10.1016/j.cell.2020.09.014

Mölder, F., Jablonski, K. P., Letcher, B., Hall, M. B., Tomkins-Tinch, C. H., Sochat, V., Forster, J., Lee, S., Twardziok, S. O., Kanitz, A., Wilm, A., Holtgrewe, M., Rahmann, S., Nahnsen, S., & Köster, J. (2021). Sustainable data analysis with Snakemake. *F1000Research*, *10*, 33. https://doi.org/10.12688/f1000research.29032.1

Monteiro, P. T., Pedreira, T., Galocha, M., Teixeira, M. C., & Chaouiya, C. (2020). Assessing regulatory features of the current transcriptional network of Saccharomyces cerevisiae. *Scientific Reports*, *10*(1), 17744. https://doi.org/10.1038/s41598-020-74043-7

Müller, B., & Grossniklaus, U. (2010). Model organisms—A historical perspective. *Journal of Proteomics*, *73*(11), 2054–2063. https://doi.org/10.1016/j.jprot.2010.08.002

Muzzey, D., Evans, E. A., & Lieber, C. (2015). Understanding the Basics of NGS: From Mechanism to Variant Calling. *Current Genetic Medicine Reports*, *3*(4), 158–165. https://doi.org/10.1007/s40142-015-0076-8

Noma, K. (2017). The Yeast Genomes in Three Dimensions: Mechanisms and Functions. *Annual Review of Genetics*, *51*(1), 23–44. https://doi.org/10.1146/annurev-genet-120116-023438

Nowotny, J., Ahmed, S., Xu, L., Oluwadare, O., Chen, H., Hensley, N., Trieu, T., Cao, R., & Cheng, J. (2015). Iterative reconstruction of three-dimensional models of human chromosomes from chromosomal contact data. *BMC Bioinformatics*, *16*(1), 338. https://doi.org/10.1186/s12859-015-0772-0

Nowotny, J., Wells, A., Oluwadare, O., Xu, L., Cao, R., Trieu, T., He, C., & Cheng, J. (2016). GMOL: An Interactive Tool for 3D Genome Structure Visualization. *Scientific Reports*, *6*(1), 20802. https://doi.org/10.1038/srep20802

Nurse, P., Thuriaux, P., & Nasmyth, K. (1976). Genetic control of the cell division cycle in the fission yeast Schizosaccharomyces pombe. *Molecular and General Genetics MGG*, *146*(2), 167–178. https://doi.org/10.1007/BF00268085

O'Donoghue, S. I. (2021). Grand Challenges in Bioinformatics Data Visualization. *Frontiers in Bioinformatics*, *1*, 669186. https://doi.org/10.3389/fbinf.2021.669186

O'Donoghue, S. I., Baldi, B. F., Clark, S. J., Darling, A. E., Hogan, J. M., Kaur, S., Maier-

Hein, L., McCarthy, D. J., Moore, W. J., Stenau, E., Swedlow, J. R., Vuong, J., & Procter, J. B. (2018). *Visualization of Biomedical Data*.

Oldenkamp, R., & Rowland, B. D. (2022). A walk through the SMC cycle: From catching DNAs to shaping the genome. *Molecular Cell*, *82*(9), 1616–1630. https://doi.org/10.1016/j.molcel.2022.04.006

Oluwadare, O., Highsmith, M., & Cheng, J. (2019). An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biological Procedures Online*, *21*(1), 7. https://doi.org/10.1186/s12575-019-0094-0

Pang, Z., Chong, J., Zhou, G., de Lima Morais, D. A., Chang, L., Barrette, M., Gauthier, C., Jacques, P.-É., Li, S., & Xia, J. (2021). MetaboAnalyst 5.0: Narrowing the gap between raw spectra and functional insights. *Nucleic Acids Research*, *49*(W1), W388–W396. https://doi.org/10.1093/nar/gkab382

Perez-Riverol, Y., Csordas, A., Bai, J., Bernal-Llinares, M., Hewapathirana, S., Kundu, D. J., Inuganti, A., Griss, J., Mayer, G., Eisenacher, M., Pérez, E., Uszkoreit, J., Pfeuffer, J., Sachsenberg, T., Yılmaz, Ş., Tiwary, S., Cox, J., Audain, E., Walzer, M., … Vizcaíno, J. A. (2019). The PRIDE database and related tools and resources in 2019: Improving support for quantification data. *Nucleic Acids Research*, *47*(D1), D442–D450. https://doi.org/10.1093/nar/gky1106

Prümers, H., Betancourt, C. J., Iriarte, J., Robinson, M., & Schaich, M. (2022). Lidar reveals pre-Hispanic low-density urbanism in the Bolivian Amazon. *Nature*, *606*(7913), 325–328. https://doi.org/10.1038/s41586-022-04780-4

Rieber, L., & Mahony, S. (2017). miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics*, *33*(14), i261–i266. https://doi.org/10.1093/bioinformatics/btx271

Rigden, D. J., & Fernández, X. M. (2021). The 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection. *Nucleic Acids Research*, *49*(D1), D1–D9. https://doi.org/10.1093/nar/gkaa1216

RNA Sequencing Data: Hitchhiker's Guide to Expression Analysis. (2019). 37.

Rodriguez, S., Ward, A., Reckard, A. T., Shtanko, Y., Hull-Crew, C., & Klocko, A. D. (2022). The genome organization of *Neurospora crassa* at high resolution uncovers principles of fungal chromosome topology. *G3 Genes|Genomes|Genetics*, *12*(5), jkac053. https://doi.org/10.1093/g3journal/jkac053

Rustici, G., Mata, J., Kivinen, K., Lió, P., Penkett, C. J., Burns, G., Hayles, J., Brazma, A., Nurse, P., & Bähler, J. (2004). Periodic gene expression program of the fission yeast cell cycle. *Nature Genetics*, *36*(8), 809–817. https://doi.org/10.1038/ng1377

Schena, M., Shalon, D., Davis, R. W., & Brown, P. O. (1995). Quantitative Monitoring of Gene Expression Patterns with a Complementary DNA Microarray. *Science*, *270*(5235), 467–470. https://doi.org/10.1126/science.270.5235.467

Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S. K., Koča, J., & Rose, A. S. (2021). Mol* Viewer: Modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research*, *49*(W1), W431–W437. https://doi.org/10.1093/nar/gkab314

Servant, N., Varoquaux, N., Lajoie, B. R., Viara, E., Chen, C.-J., Vert, J.-P., Heard, E., Dekker,

J., & Barillot, E. (2015). HiC-Pro: An optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, *16*(1), 259. https://doi.org/10.1186/s13059-015-0831-x

Shi, Y. (2017). Mechanistic insights into precursor messenger RNA splicing by the spliceosome. *Nature Reviews Molecular Cell Biology*, *18*(11), 655–670. https://doi.org/10.1038/nrm.2017.86

Smalheiser, N. R. (2002). Informatics and hypothesis-driven research. *EMBO Reports*, *3*(8), 702–702. https://doi.org/10.1093/embo-reports/kvf164

Souciet, J.-L. (2011). Ten years of the Génolevures Consortium: A brief history. *Comptes Rendus Biologies*, *334*(8–9), 580–584. https://doi.org/10.1016/j.crvi.2011.05.005

Spellman, P. T., Sherlock, G., Zhang, M. Q., Anders, K., Eisen, M. B., Brown, P. O., & Futcher, B. (1998). Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization□D. *Molecular Biology of the Cell*, *9*, 25.

Stevens, T. J., Lando, D., Basu, S., Atkinson, L. P., Cao, Y., Lee, S. F., Leeb, M., Wohlfahrt, K. J., Boucher, W., O'Shaughnessy-Kirwan, A., Cramard, J., Faure, A. J., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M. G. S., Lehner, B., Di Croce, L., … Laue, E. D. (2017). 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, *544*(7648), 59–64. https://doi.org/10.1038/nature21429

Strom, A. R., & Brangwynne, C. P. (2019). The liquid nucleome – phase transitions in the nucleus at a glance. *Journal of Cell Science*, *132*(22), jcs235093. https://doi.org/10.1242/jcs.235093

Subramanian, I., Verma, S., Kumar, S., Jere, A., & Anamika, K. (2020). Multi-omics Data Integration, Interpretation, and Its Application. *Bioinformatics and Biology Insights*, *14*, 117793221989905. https://doi.org/10.1177/1177932219899051

Sun, D., Tian, L., & Ma, B. (2019). Spatial organization of the transcriptional regulatory network of Saccharomyces cerevisiae. *FEBS Letters*, *593*(8), 876–884. https://doi.org/10.1002/1873-3468.13371

Tan, L., Xing, D., Daley, N., & Xie, X. S. (2019). Three-dimensional genome structures of single sensory neurons in mouse visual and olfactory systems. *Nature Structural & Molecular Biology*, *26*(4), 297–307. https://doi.org/10.1038/s41594-019-0205-2

Tanizawa, H., Iwasaki, O., Tanaka, A., Capizzi, J. R., Wickramasinghe, P., Lee, M., Fu, Z., & Noma, K. (2010). Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research*, *38*(22), 8164–8177. https://doi.org/10.1093/nar/gkq955

Tanizawa, H., Kim, K.-D., Iwasaki, O., & Noma, K. (2017). *Architectural alterations of the fission yeast genome during the cell cycle*. 31.

Thelen, N., Defourny, J., Lafontaine, D. L. J., & Thiry, M. (2021). Visualization of Chromatin in the Yeast Nucleus and Nucleolus Using Hyperosmotic Shock. *International Journal of Molecular Sciences*, *22*(3), 1132. https://doi.org/10.3390/ijms22031132

Theobald, D. L. (2010). A formal test of the theory of universal common ancestry.

Nature, *465*(7295), Article 7295. https://doi.org/10.1038/nature09014

Tinya, F., Kovács, B., Bidló, A., Dima, B., Király, I., Kutszegi, G., Lakatos, F., Mag, Z., Márialigeti, S., Nascimbene, J., Samu, F., Siller, I., Szél, G., & Ódor, P. (2021). Environmental drivers of forest biodiversity in temperate mixed forests – A multi-taxon approach. *Science of The Total Environment*, *795*, 148720. https://doi.org/10.1016/j.scitotenv.2021.148720

Todd, S., Todd, P., McGowan, S. J., Hughes, J. R., Kakui, Y., Leymarie, F. F., Latham, W., & Taylor, S. (2021). CSynth: An interactive modelling and visualization tool for 3D chromatin structure. *Bioinformatics*, *37*(7), 951–955. https://doi.org/10.1093/bioinformatics/btaa757

Torres, D. E., Reckard, A. T., Klocko, A. D., & Seidl, M. F. (2023). Nuclear genome organization in fungi: From gene folding to Rabl chromosomes. *FEMS Microbiology Reviews*, *47*(3), fuad021. https://doi.org/10.1093/femsre/fuad021

Tricou, T. (2022). *Détecter et exploiter les flux de gènes dans une biodiversité majoritairement inconnue*.

Trieu, T., Oluwadare, O., & Cheng, J. (2019). Hierarchical Reconstruction of High-Resolution 3D Models of Large Chromosomes. *Scientific Reports*, *9*(1), 4971. https://doi.org/10.1038/s41598-019-41369-w

Trieu, T., Oluwadare, O., Wopata, J., & Cheng, J. (2019). GenomeFlow: A comprehensive graphical tool for modeling and analyzing 3D genome structure. *Bioinformatics*, *35*(8), 1416–1418. https://doi.org/10.1093/bioinformatics/bty802

Varoquaux, N., Noble, W. S., & Vert, J.-P. (2021). *Inference of genome 3D architecture by*

modeling overdispersion of Hi-C data [Preprint]. Bioinformatics. https://doi.org/10.1101/2021.02.04.429864

Veenstra, T. D. (2021). Omics in Systems Biology: Current Progress and Future Outlook. *PROTEOMICS*, *21*(3–4), 2000235. https://doi.org/10.1002/pmic.202000235

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Vázquez-Baeza, Y. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wang, H., Yang, J., Zhang, Y., Qian, J., & Wang, J. (2022). Reconstruct high-resolution 3D genome structures for diverse cell-types using FLAMINGO. *Nature Communications*, *13*(1), 2645. https://doi.org/10.1038/s41467-022-30270-2

Wolfe, K. H. (2015). Origin of the Yeast Whole-Genome Duplication. *PLOS Biology*, *13*(8), e1002221. https://doi.org/10.1371/journal.pbio.1002221

Wong, B. (2011). Points of view: The overview figure. *Nature Methods*, *8*(5), 365–365. https://doi.org/10.1038/nmeth0511-365

Yanagida, M. (2005). The model unicellular eukaryote, Schizosaccharomyces pombe. *Current Biology*, *15*(16), R613–R614. https://doi.org/10.1016/j.cub.2005.08.015

Ye, C., Paccanaro, A., Gerstein, M., & Yan, K.-K. (2020). The corrected gene proximity map for analyzing the 3D genome organization using Hi-C data. *BMC Bioinformatics*, *21*(1), 222. https://doi.org/10.1186/s12859-020-03545-y

Zhong, H., Lin, W., Liu, H., Ma, N., Liu, K., Cao, R., Wang, T., & Ren, Z. (2022). Identification of tree species based on the fusion of UAV hyperspectral image and LiDAR data in a coniferous and broad-leaved mixed forest in Northeast China. *Frontiers in Plant Science*, *13*, 964769. https://doi.org/10.3389/fpls.2022.964769

Zhou, G., Soufan, O., Ewald, J., Hancock, R. E. W., Basu, N., & Xia, J. (2019). NetworkAnalyst 3.0: A visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Research*, *47*(W1), W234–W241. https://doi.org/10.1093/nar/gkz240

Zhou, G., & Xia, J. (2018). OmicsNet: A web-based tool for creation and visual analysis of biological networks in 3D space. *Nucleic Acids Research*, *46*(W1), W514–W522. https://doi.org/10.1093/nar/gky510

Tableau 1

A

Model number	Model name	Genome size (in bp)	Number of chromosomes	Hi-C data sources	SRA ID	strain type	restriction enzymes	number of raw reads	% of reported mapped reads pairs	% of HiC valid read pairs	% of non duplicated merged valid read pairs	final number of merged valid read pairs	resolution (bp)	number of beads in the 3D model (Pastis output)	number of beads in the 3D model (final output)
1	*Neurospora crassa*	41 035 892	7	Rodriguez et al. 2022	SRR14362684	WT	DpnII & MseI	23 462 245	68	46	81	107 636 304	10 000	4 107	3976
					SRR14362685			46 740 610	60	48	83				
					SRR16761088			456 036 737	73	46	71				
					SRR16761089			56 574 891	62	52	81				
					SRR16761090			42 099 719	79	40	73				
					SRR16761091			137 502 307	73	50	77				
					SRR16761092			25 338 577	74	38	75				
2				Galazka et al. 2016	SRR2105869	WT	HindIII	27 468 135	60	22	61	5 030 121	50 000	825	800
					SRR2105870			25 384 746	61	23	72				
					SRR2105871			19 058 145	47	3	56				
					SRR2105872			71 911 026	60	18	65				
3					SRR2105876	HPO mutant		25 204 445	63	22	39	2 922 526			808
					SRR2105877			25 386 582	63	20	50				
					SRR2105878			50 591 027	63	21	44				
4					SRR5149251	WT 20min		29 480 945	27	81	98	15 511 107			1 223
					SRR5149252			41 525 872	28	81	98				
					SRR5149253			25 957 692	30	73	98	14 377 600			1 211
5					SRR5149254	WT 30min		40 305 026	31	73	98				
6					SRR5149255	WT 40min		24 695 113	30	76	98	14 394 280			1 204
					SRR5149256			38 982 137	31	76	98				
7					SRR5149257	WT 50min		33 202 062	29	71	98	15 497 423			1 200
					SRR5149258			42 977 834	30	71	98				
					SRR5149259			29 344 104	29	70	98				
8	*Schizosaccharomyces pombe*	12 571 820	3	Tanizawa et al. 2017	SRR5149260	WT 60min	MboI	42 681 824	29	70	98	19 035 344	10 000	1 258	1 213
					SRR5942526			25 500 551	39	49	97				
9					SRR5149261	WT 70min		30 771 374	30	64	98	13 841 928			1 224
					SRR5149263			43 485 076	30	64	98				
					SRR5149264			28 676 303	36	71	98	21 936 238			1201
10					SRR5149265	WT 80min		41 156 922	36	71	98				
					SRR5942527			25 451 561	38	48	97				
					SRR5149266			28 600 421	31	67	98	21 727 635			1 189
11					SRR5149267	WT 120min		42 440 615	31	68	98				
					SRR5942528			34 930 401	39	54	97				
12	*Saccharomyces cerevisiae*	12 157 105	16	Constantino et al. 2020	SRR11893084	WT	MNase	30 734 607	29	89	91	16 847 912	5 000	2 423	2 332
					SRR11893085			38 231 647	32	90	90				
13					SRR11893086	mcd1 mutant		35 781 165	25	87	91	14 180 445			2 323
					SRR11893087			33 785 215	26	88	91				
									% of the I column	% of the J column	% of the K column				

B

Model name	ChIP Seq data sources	SRA ID	phenotype	epigenetic mark / anchored protein	number of raw reads	alignment (%)
Neurospora crassa	Jamieson et al. 2016	SRR2026383	WT	WT H3K27me2/3	506 019	84
	Basenko et al. 2015	SRR2036141	WT		16 253 730	86
	Basenko et al. 2015	SRR2036142	WT		14 549 060	86
	Jamieson et al. 2016	SRR2026380	HPO	HPO H3K27me2/3	9 944 186	92
	Basenko et al. 2015	SRR2036174	HPO		18 723 702	23
	Jamieson et al. 2016	SRR2026386	HPO	HPO H3K9me3	3 565 406	88
	Basenko et al. 2015	SRR2036173	HPO		129 366 356	87
	Jamieson et al. 2016	SRR2026390	WT	WT H3K9me3	2 888 817	94
	Basenko et al. 2015	SRR2036168	WT		6 423 302	91
Saccharomyces cerevisiae	Costantino et al. 2020	SRR11872088	WT	WT Mcd1	7 002 691	94
		SRR11872089			11 721 218	94

BMC Research Notes

RESEARCH NOTE

Open Access

Additional insights into the organization of transcriptional regulatory modules based on a 3D model of the *Saccharomyces cerevisiae* genome

Thibault Poinsignon[1], Mélina Gallopin[1], Jean-Michel Camadro[2], Pierre Poulain[2*] and Gaëlle Lelandais[2*]

Abstract

Objectives: Transcriptional regulatory modules are usually modelled via a network, in which nodes correspond to genes and edges correspond to regulatory associations between them. In the model yeast *Saccharomyces cerevisiae*, the topological properties of such a network are well-described (distribution of degrees, hierarchical levels, organization in network motifs, etc.). To go further on this, our aim was to search for additional information resulting from the new combination of classical representations of transcriptional regulatory networks with more realistic models of the spatial organization of *S. cerevisiae* genome in the nucleus.

Results: Taking advantage of independent studies with high-quality datasets, i.e. lists of target genes for specific transcription factors and chromosome positions in a three dimensional space representing the nucleus, particular spatial co-localizations of genes that shared common regulatory mechanisms were searched. All transcriptional modules of *S. cerevisiae*, as described in the latest release of the YEASTRACT database were analyzed and significant biases toward co-localization for a few sets of target genes were observed. To help other researchers to reproduce such analysis with any list of genes of their interest, an interactive web tool called 3D-Scere (https://3d-scere.ijm.fr/) is provided.

Keywords: Transcriptional regulations, Chromosome conformation capture, Yeast, 3D-Scere

Introduction

Normal cell functioning requires appropriate gene expression, which depends on multiple regulatory layers (see [1] for review). In this context, transcriptional regulatory modules (TRMs) were extensively studied (for instance [2–5]). By definition, a TRM is a set of genes for which transcriptional activity is modulated by a specific transcription factor (TF) [6]. In the model yeast *Saccharomyces cerevisiae*, TRMs are well described [2–5] and public databases like YEASTRACT [7] or SGD [8], provide lists of target genes for any TF. All together TRMs were explored to better understand their individual organizations, but also their collective relationships [4, 5, 9, 10]. In most studies, questions were addressed via a representation of TRMs as networks. In these networks, TF and target genes are the nodes, which are connected by directed edges (from TF to related targets). Topological properties of such networks were analysed to reveal the design principles underlying transcriptional regulations. It allowed the discovery of important regulatory motifs, surprisingly consistent across very different species [10, 11].

In addition to this information, spatial organization of the 16 chromosomes of *S. cerevisiae* was reported in the literature [1]. Experimental techniques derived from

*Correspondence: pierre.poulain@u-paris.fr; gaelle.lelandais@universite-paris-saclay.fr
[2] Institut Jacques Monod, CNRS, Université de Paris, 75006 Paris, France
Full list of author information is available at the end of the article

chromosome conformation capture (3C) were used to obtain a tridimensional (3D) model [12]. This model is based on the idea that interphase chromosomes are not positioned randomly within the nucleus. In particular, chromosomes should adopt a "Rabl configuration", in which centromeres are clustered together at one pole of the nucleus, whereas arms are extended in several directions until telomeres, which are abutted to the nuclear envelope. Moreover, chromosome 12, which carries the rDNA repeats in *S. cerevisiae*, is expected to extend outward to join the nucleolus, i.e. the site of ribosome biogenesis (Additional file 1). This 3D model is relevant with the existence of a repressive chromatin structure, i.e. silent chromatin, which is known in yeasts for a long time (see [13] for a review) and affects mating-type loci, telomeres or rDNA repeats. More recently, this 3D model was used to study potential connections between interchromosomal DNA contacts and gene co-expressions [14]. Significant correlations were found, thus supporting the idea that a non-random nature of the genome organization helps to coordinate transcriptional processes in groups of genes, like those found in TRMs.

In this work, our aim was to search for additional insights into the organization of TRMs based on the 3D model of the *S. cerevisiae* genome at interphase. The TRMs were explored from a new perspective, which integrates functional and spatial information presently available, and addressed the following question: are target genes associated to a common TF (TRM) randomly disseminated within the nucleus, or are they preferentially co-localized? In the literature, this question was only partially answered, focusing essentially on spatial distances between genes coding for TFs and associated targets [15]. Our analysis represents an additional step in this context, reporting all distances between genes that belong to any TRM, as described in the latest release of the YEASTRACT database. Statistical parameters are provided, to quantify the intensity of potential bias observed in distributions of pairwise Euclidean distances calculated between lists of genes. A web tool called 3D-Scere (https://3d-scere.ijm.fr/) was also developed. With this tool, any researcher can retrieve information for all pairs of genes that belong to a list of his/her interest.

Main text

This study extends on two previous analyses of transcriptional regulatory modules, recently presented in the literature. The first one is that of Monteiro et al. [2], published in 2020. The authors assessed the regulatory features of the current transcriptional network of *S. cerevisiae*, taking advantage of the latest release of their YEASTRACT database, which comprised almost 200,000 interactions including 220 TFs and 6886 target genes. The second one is that of Sun et al. [15], published in 2019. The authors used the 3D model of *S. cerevisiae* genome proposed by Duan et al. [12], to study the spatial organization of the regulatory network of *S. cerevisiae*. We propose that the perspectives and the data from the two studies can be elegantly combined to increase the scope of the results presented in each of them. Indeed, both studies have strengths, but also limitations. On one hand, the study of Monteiro et al. is based on a colossal work to collect, clean, and organize the transcriptional regulations identified in more than a thousand publications in peer-reviewed international journals. Notably, the authors also provided confidence level information for each regulation, thus delivering very high-quality data. They observed interesting topological properties of the global *S. cerevisiae* transcriptional network and discussed the complexity of the transcription regulatory processes that control gene expression. In that respect, searching for a potential role of genome organization in the functioning of this network, represents a natural perspective. On the other hand, Sun et al. had the original idea to place the transcriptional regulations between genes in the context of the 3D genome model available in *S. cerevisiae*. They concluded that "the transcriptional regulatory network of *S. cerevisiae* presents an optimized structure in space to adapt to functional requirements". Undoubtedly very promising, we think that this conclusion (i) suffers from the use of transcriptional regulations, which were only partially verified and (ii) lacks individual analyses of TRMs.

The work presented in this article was performed in three steps. First, the TRMs were extracted from the study of Monteiro et al. [2]. The supplementary data provided all the YEASTRACT transcriptional regulations, annotated according to "binding evidence", "expression evidence" or "both". We decided to focus on regulatory associations which relied on "binding evidence" only. They represent 176 TFs with 6475 target genes, connected with 45,209 associations (23% of the full regulatory associations dataset). Second, the 3D model from the study of Duan et al. [12] was recovered. In the related supplementary data, the 3D coordinates for 26,538 "points" were found. Each point can be seen as a precise location in space, defined by 3D-coordinates (x, y, z). All together the points define all chromosomes of *S. cerevisiae* genome (Fig. 1a). Each chromosome was arranged into pairs of successive points, which thus delimit chromosomal regions in space. Note here that the obtained regions were of variable sizes because the points in the initial 3D model were not equidistant. We for instance observed that in situations where chromosomes are folded or change direction in space, more points were present to model the same length of DNA

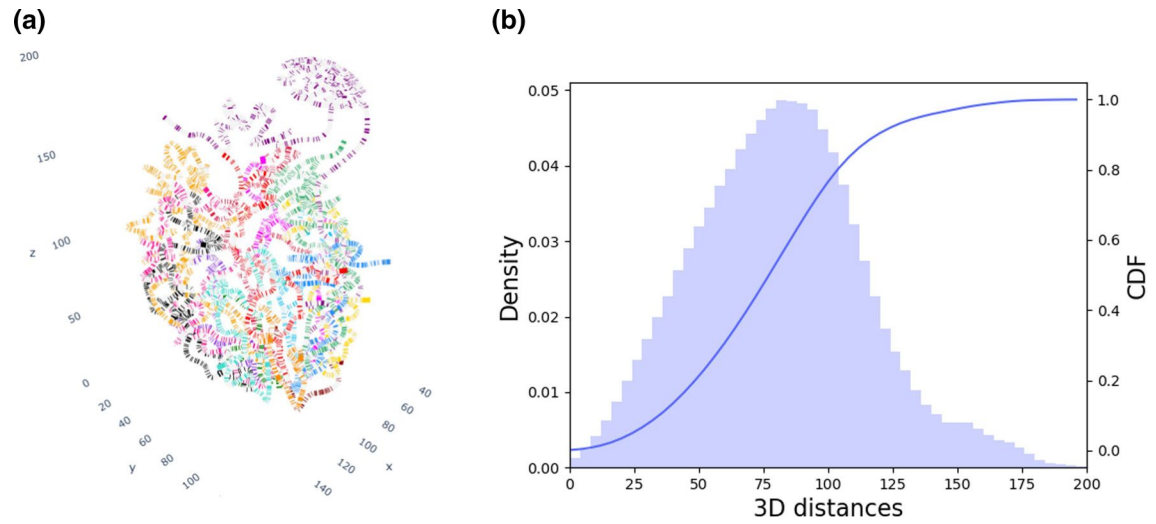


Fig. 1 Spatial organization of the 16 chromosomes of *S. cerevisiae*. **a** Screenshot of the 3D genome model available in the 3D-Scere tool. **b** Density histogram (light blue) of all Euclidean distances between all chromosomal features in the 3D model (see "Main text"), and Cumulative Distribution Function (CDF, dark blue) of the density histogram

base pairs. Tridimensional coordinates for 9185 *S. cerevisiae* genome features (including 6572 ORFs) were next derived (Additional file 2) and used for calculations of spatial Euclidean distances between all pairs of genome features (this represents 42,177,520 distances) (Fig. 1b). All distance calculations are available as Additional file 3. For each TRM defined by the 176 different TFs, pairwise distances between target genes were selected. Distance distributions obtained with all features of the *S. cerevisiae* genome and with the subset of genes that belong to a particular TRM were finally superimposed and used to quantify a potential bias for co-localization (smaller distances) between target genes in TRMs. All results are available as Additional file 4. A Kolmogorov Smirnov (KS) test with a Bonferonni correction to quantify the deviation from the distribution of all genes was performed. As a result, several TFs for which the target genes exhibited atypical locations within the nucleus were observed. These TFs are listed in Table 1, and the distance distributions of the four TRMs with the highest KS statistic (i.e. highest deviation from the distribution of all targets) are shown in Fig. 2. An interesting situation, regarding the Upc2 transcriptional module, is detailed in Additional file 6.

Finally, an open-source tool was developed, for interactive visualization and exploration. Source code is available on GitHub https://github.com/data-fun/3d-scere and the tool is freely usable online at https://3d-scere.ijm.fr/. It allows the visualization of any list of genes in the context of the 3D model of *S. cerevisiae* genome (Additional file 5 for screenshots). Further information can easily be added, like functional annotations (GO terms) or gene expression measurements. Qualitative or quantitative functional properties are highlighted in the large-scale 3D context of the genome with only a few mouse clicks.

Limitations

We see in this work three main limitations. The first one concerns the biological relevance of the 3D model of the *S. cerevisiae* genome that was used. Created more than 10 years ago [12], this structural model represents only a static (and averaged) view of the relative positioning of the 16 chromosomes in the nucleus at interphase. It was obtained from 3C experiment data, which had to be processed with complex numerical procedures, to find an optimal solution. Because "optimal" does not guarantee "real", all observations that emerge from this model must be further validated. In that respect, new data generated with the latest and most powerful Hi-C techniques, at different stages of the *S. cerevisiae* cell cycle to capture the dynamics of its genome organization could be of great interest. The second limitation concerns the lack of landmarks for the localization of genes, within the nucleus. Are they located near the nuclear envelope and possibly near pores allowing, for instance, the rapid export of transcripts to the cytoplasm? Such information is presently missing from our analyses. One solution could be to calculate additional distances with referential points on chromosomes such as centromeres, telomeres, or the outside emblematic region of rDNA repeats. Finally, the third limitation, in our point of view, relies on the

Table 1 Statistical parameters derived from the study of the organization of transcriptional regulatory modules based on a 3D model of the *Saccharomyces cerevisiae* genome

TF	Description (SGD database)	# targets	KS value	p-val	p-val (adjusted)
STB4	Putative transcription factor; contains a Zn(II)2Cys6 zinc finger domain characteristic of DNA-binding proteins; computational analysis suggests a role in regulation of expression of genes encoding transporters; binds Sin3p in a two-hybrid assay	32	0.166	2e−12	3e−10
AZF1	Zinc-finger transcription factor; involved in diauxic shift; in the presence of glucose, activates transcription of genes involved in growth and carbon metabolism; in nonfermentable carbon sources, activates transcription of genes involved in maintenance of cell wall integrity; relocalizes to the cytosol in response to hypoxia	52	0.153	1e−27	1e−25
MOT3	Transcriptional repressor, activator; role in cellular adjustment to osmotic stress including modulation of mating efficiency; involved in repression of subset of hypoxic genes by Rox1p, repression of several DAN/TIR genes during aerobic growth, ergosterol biosynthetic genes in response to hyperosmotic stress; contributes to recruitment of Tup1p-Cyc8p general repressor to promoters; relocalizes to cytosol under hypoxia; forms [MOT3+] prion under anaerobic conditions	56	0.144	2e−28	5e−26
UPC2	Sterol regulatory element binding protein; induces sterol biosynthetic genes, upon sterol depletion; acts as a sterol sensor, binding ergosterol in sterol rich conditions; relocates from intracellular membranes to perinuclear foci upon sterol depletion; redundant activator of filamentation with ECM22, up-regulating the expression of filamentous growth genes; contains a Zn[2]-Cys[6] binuclear cluster; UPC2 has a paralog, ECM22, that arose from the whole genome duplication	38	0.106	2e−07	4e−05
PHO2	Homeobox transcription factor; regulatory targets include genes involved in phosphate metabolism; binds cooperatively with Pho4p to the PHO5 promoter; phosphorylation of Pho2p facilitates interaction with Pho4p; relocalizes to the cytosol in response to hypoxia	134	0.100	6e−78	1e−75
DAL80	Negative regulator of genes in multiple nitrogen degradation pathways; expression is regulated by nitrogen levels and by Gln3p; member of the GATA-binding family, forms homodimers and heterodimers with Gzf3p; DAL80 has a paralog, GZF3, that arose from the whole genome duplication	57	0.097	1e−13	2e−11
YAP3	Basic leucine zipper (bZIP) transcription factor	39	0.095	2e−06	0.0004
PLM2	Putative transcription factor, contains Forkhead Associated domain; found associated with chromatin; target of SBF transcription factor; induced in response to DNA damaging agents and deletion of telomerase; PLM2 has a paralog, TOS4, that arose from the whole genome duplication	182	0.093	5e−125	9e−123
RSF2	Zinc-finger protein; involved in transcriptional control of both nuclear and mitochondrial genes, many of which specify products required for glycerol-based growth, respiration, and other functions; RSF2 has a paralog, TDA9, that arose from the whole genome duplication; relocalizes from nucleus to cytoplasm upon DNA replication stress	35	0.092	7e−05	0.012
RPH1	JmjC domain-containing histone demethylase; targets tri- and dimethylated H3K36; associates with actively transcribed regions and promotes elongation; repressor of autophagy-related genes in nutrient-replete conditions; damage-responsive repressor of PHR1; phosphorylated by the Rad53p-dependent DNA damage checkpoint pathway and by a Rim1p-mediated event during starvation; target of stress-induced hormesis; RPH1 has a paralog, GIS1, that arose from the whole genome duplication	91	0.090	7e−30	1e−27

The ten TFs with the highest values of KS statistics are shown here. Results for all other TFs are available as Additional file 4

definition of TRMs by themself. We defined a TRM as a set of genes for which the expression is modulated by a common TF. In this work, we reasoned by individual TRM. But a target gene can belong to several TRMs and also can require, to be transcriptionally regulated, the association between several TFs. Such genes could be studied specifically for particular co-localizations on the 3D model of the *S. cerevisiae* genome. Our strategy thus opens interesting research perspectives in the context of the study of gene lists that belong to transcriptional modules, but it can be of interest for any list of genes. The spatial proximity could be studied, between strongly (or weakly) expressed genes, or between genes which encode proteins involved in common metabolic pathways or which associate within complexes, etc. In this context, the online tool (https://3d-scere.ijm.fr/) will be of interest to the community, allowing any researcher to query any list of genes for which he/she has a particular interest in.

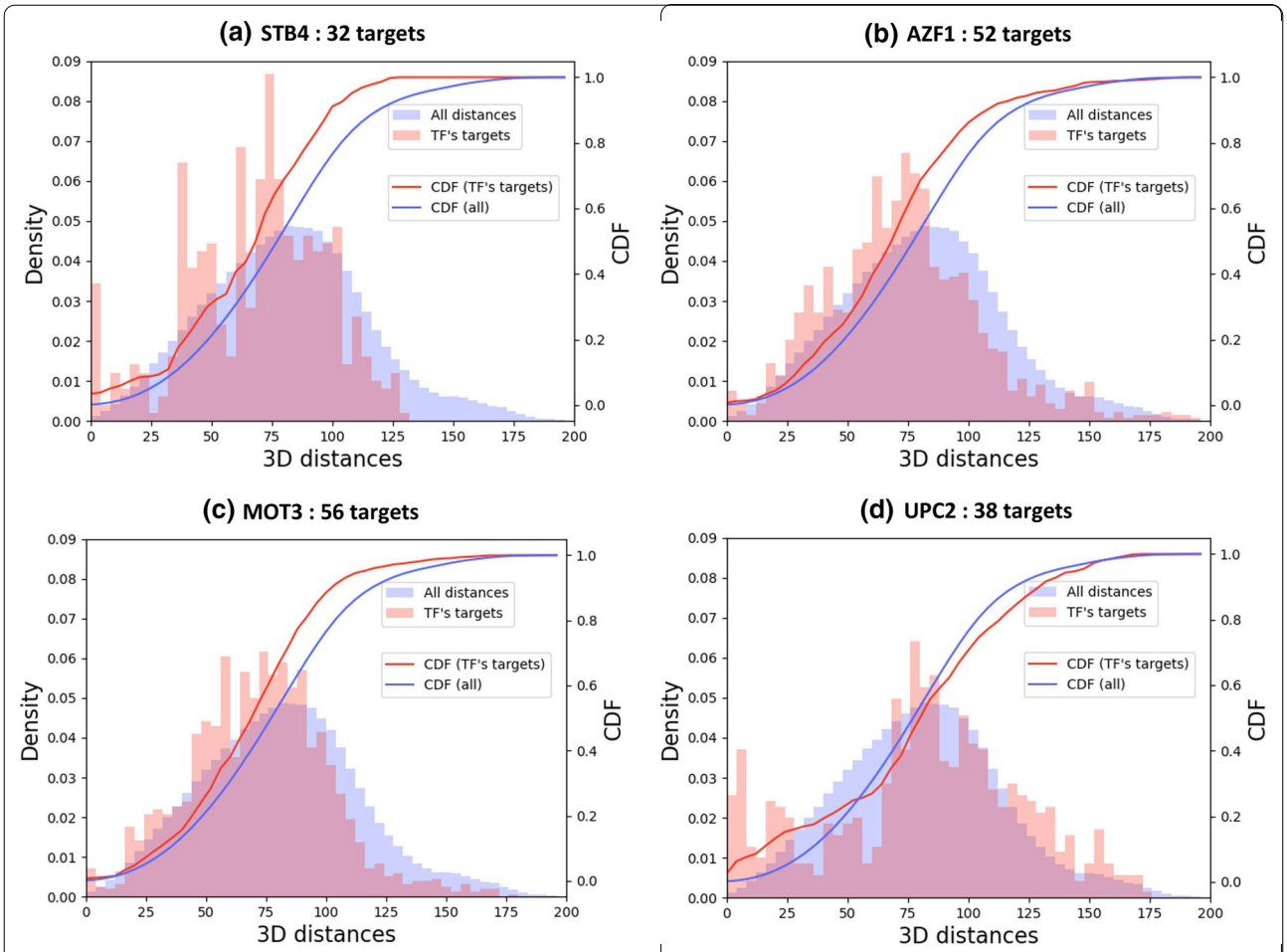


Fig. 2 Examples of transcriptional regulatory modules in which targets are preferentially co-localized within the nucleus. Distance histograms (pink) for the targets of four TFs (STB4, AZF1, MOT3 and UPC2) are shown and compared to the distance histogram of all distances (light blue) as presented in Fig. 1. These TFs were selected because (i) they have a number of target genes > 30, (ii) they exhibit the highest values of Kolmogorov Smirnov statistics, with (iii) associated significant adjusted p-values (< 0.05)

Abbreviations

TRM: Transcriptional regulatory module; TF: Transcription factor; 3C: Chromosome conformation capture; ORF: Open reading frame; GO: Gene ontology; Scere: *Saccharomyces cerevisiae*; SGD: Saccharomyces genome database.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13104-022-05940-5.

Additional file 1. Pictures of the 3D model of the S. cerevisiae genome, as it is available in the 3D-Scere tool.

Additional file 2. Text file describing the method that was applied to associate the 9185 *S. cerevisiae* chromosomal features with the spatial coordinates of the 3D model.

Additional file 3. ZIP file with distance calculations between all pairwise chromosomal features for which spatial coordinates were associated to the 3D model (link from Zenodo repository: https://zenodo.org/record/5841177/files/3D_distances.parquet.gzip?download=1).

Additional file 4. ZIP file with graphical representations associated to each transcriptional module: (link from Zenodo repository: https://zenodo.org/record/5841177/files/supplementary-data-file-S4.zip?download=1).

Additional file 5. General overview of the 3d-Scere tool. A public web acces is available at https://3d-scere.ijm.fr. Three different uses are proposed to users: (1) «GO term projection», (2) Quantitative variable projection and (3) 3D distances histogram and network. Each access starts with the upload of a list of genes of interest for the user. Note that the list of Upc2 targets can be loaded as a «demo data». From the list of genes, users can manipulate either qualitative information (Access 1) or quantitative information (Access 2) and obtain graphics showing location on the chromosome or location on the 3D model of S. cerevisiae genome. Distribution of pairwise distances between genes is obtain with the Access 3.

Additional file 6. New insights into the transcriptional module related to Upc2 transcription factor.

Authors' contributions

TP and GL designed the analysis; TP and PP developed the 3d-Scere webtool; MG and JMC gave feedback regarding the applied statistical procedure and

the relevance of TRMs, respectively; TP and GL wrote the first draft of the manuscript. All authors read and approved the final manuscript.

Funding and Acknowledgements
This work was funded by the Agence Nationale pour la Recherche (MINOMICS and SLIM projects, Grant Number ANR-19-CE45-0017 and ANR-18-CE44-0014). We thank Fabienne Malagnac and Pierre Grognet for helpful discussions.

Availability of data and materials
Source code written to generate figures is available on GitHub: https://github.com/data-fun/3d-scere-scripts. Source code written to develop the web tool (3d-Scere) is also available on GitHub: https://github.com/data-fun/3d-scere. Results are archived in the Zenodo repository: https://zenodo.org/record/5841177.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, 91198 Gif-sur-Yvette, France. [2]Institut Jacques Monod, CNRS, Université de Paris, 75006 Paris, France.

Received: 11 November 2021 Accepted: 31 January 2022
Published online: 19 February 2022

References

1. Noma K. The yeast genomes in three dimensions: mechanisms and functions. Annu Rev Genet. 2017;51(1):23–44.
2. Monteiro PT, Pedreira T, Galocha M, Teixeira MC, Chaouiya C. Assessing regulatory features of the current transcriptional network of *Saccharomyces cerevisiae*. Sci Rep. 2020;10(1):17744.
3. Pilpel Y, Sudarsanam P, Church GM. Identifying regulatory networks by combinatorial analysis of promoter elements. Nat Genet. 2001;29(2):153–9.
4. Lee TI. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. Science. 2002;298(5594):799–804.
5. Guelzim N, Bottani S, Bourgine P, Képès F. Topological and causal structure of the yeast transcriptional regulatory network. Nat Genet. 2002;31(1):60–3.
6. Goudot C, Etchebest C, Devaux F, Lelandais G. The reconstruction of condition-specific transcriptional modules provides new insights in the evolution of yeast AP-1 proteins. PLoS ONE. 2011;6(6): e20924.
7. Teixeira MC, Monteiro PT, Palma M, Costa C, Godinho CP, Pais P, et al. YEASTRACT: an upgraded database for the analysis of transcription regulatory networks in *Saccharomyces cerevisiae*. Nucleic Acids Res. 2018;46(D1):D348–53.
8. Ng PC, Wong ED, MacPherson KA, Aleksander S, Argasinska J, Dunn B, et al. Transcriptome visualization and data availability at the *Saccharomyces* genome database. Nucleic Acids Res. 2020;48(D1):D743–8.
9. Balaji S, Babu MM, Iyer LM, Luscombe NM, Aravind L. Comprehensive analysis of combinatorial regulation using the transcriptional regulatory network of yeast. J Mol Biol. 2006;360(1):213–27.
10. Ouma WZ, Pogacar K, Grotewold E. Topological and statistical analyses of gene regulatory networks reveal unifying yet quantitatively different emergent properties. PLOS Comput Biol. 2018;14(4): e1006098.
11. Alon U. Network motifs: theory and experimental approaches. Nat Rev Genet. 2007;8(6):450–61.
12. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, et al. A three-dimensional model of the yeast genome. Nature. 2010;465(7296):363–7.
13. Huang Y. Transcriptional silencing in *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Nucleic Acids Res. 2002;30(7):1465–82.
14. Homouz D, Kudlicki AS. The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes. PLoS ONE. 2013;8(1): e54699.
15. Sun D, Tian L, Ma B. Spatial organization of the transcriptional regulatory network of *Saccharomyces cerevisiae*. FEBS Lett. 2019;593(8):876–84.

Publisher's Note

Chapter 10

Working with Omics Data: An Interdisciplinary Challenge at the Crossroads of Biology and Computer Science

Thibault Poinsignon, Pierre Poulain, Mélina Gallopin, and Gaëlle Lelandais

Abstract

Nowadays, generating omics data is a common activity for laboratories in biology. Experimental protocols to prepare biological samples are well described, and technical platforms to generate omics data from these samples are available in most research institutes. Furthermore, manufacturers constantly propose technical improvements, simultaneously decreasing the cost of experiments and increasing the amount of omics data obtained in a single experiment. In this context, biologists are facing the challenge of dealing with large omics datasets, also called "big data" or "data deluge." Working with omics data raises issues usually handled by computer scientists, and thus cooperation between biologists and computer scientists has become essential to efficiently study cellular mechanisms in their entirety, as omics data promise. In this chapter, we define omics data, explain how they are produced, and, finally, present some of their applications in fundamental and medical research.

Key words Genomics, Transcriptomics, Proteomics, Metabolomics, Big data, Computer science, Bioinformatics

1 Introduction

There are different types of omics data, each revealing an aspect of cell complexity. To illustrate this complexity, we propose in Fig. 1 an analogy between the functions of a cell and that of a factory. The different omics data types are replaced there, in their specific context. Cells are the building blocks of living organisms. They can be pictured as microscopic, automated factories, made up of thousands of biological molecules (or molecular components) that work together to perform specific functions. Basically, there are four main types of molecular components: DNA, RNA, proteins, and metabolites. The whole population of one type of cellular component is named with the suffix -ome, i.e., gen*ome* (DNA), transcript*ome* (RNA), prote*ome* (proteins), and metabol*ome* (metabolites) (*see* Fig. 1). The scientific fields, which aim at studying those respective populations, are named with the suffix -omics,

Olivier Colliot (ed.), *Machine Learning for Brain Disorders*, Neuromethods, vol. 197, https://doi.org/10.1007/978-1-0716-3195-9_10,

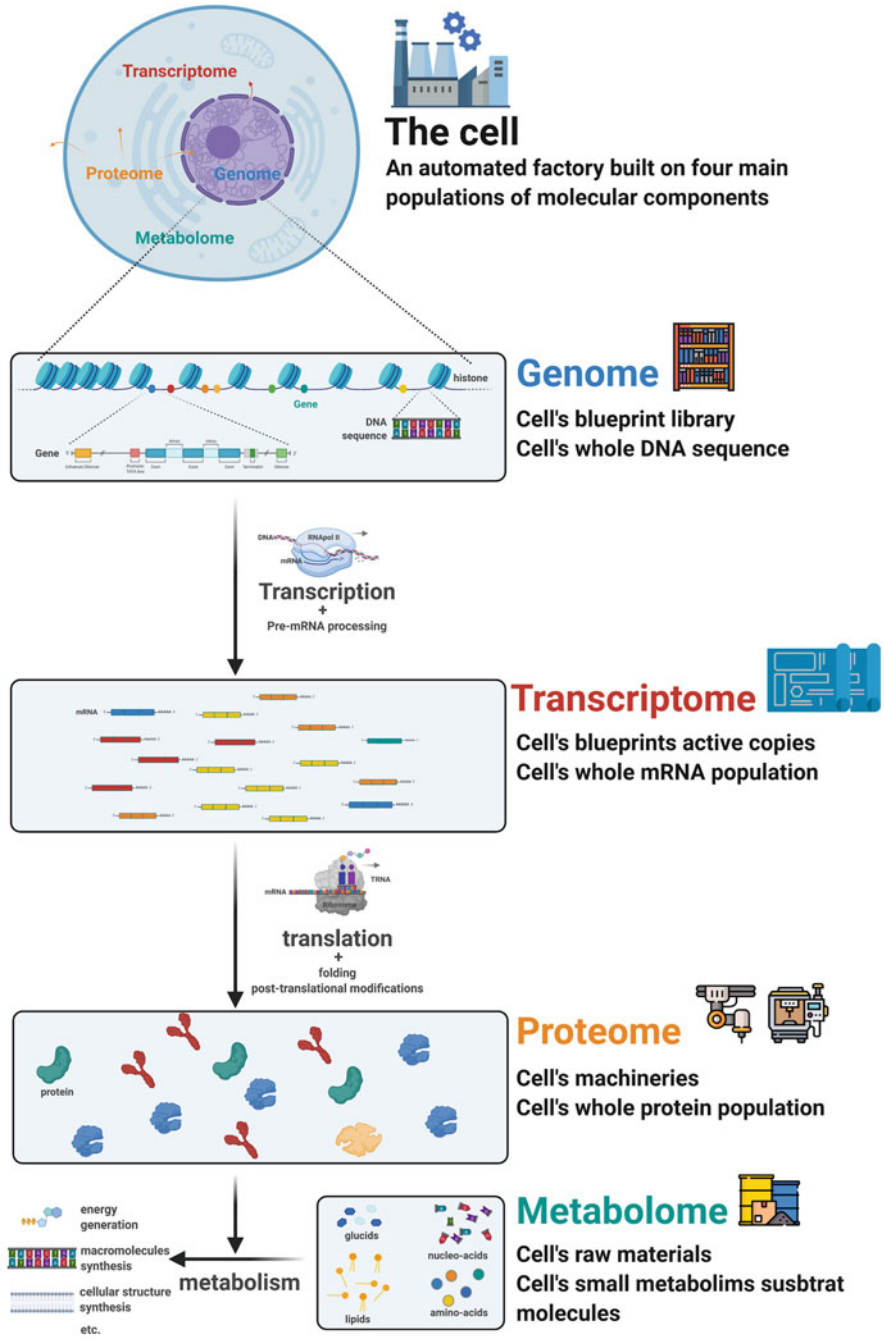


Fig. 1 The four main -omes and an analogy of their functions. The genome designates all cell's DNA molecules. The transcriptome, the proteome, and the metabolome refer, respectively, to the cell's whole set of RNA, proteins, or metabolites at a given time

i.e., gen*omics*, transcript*omics*, prote*omics*, and metabol*omics*. The common point between the different types of omics data is that they all arise from high-throughput experimental strategies that allow the simultaneous observation of all individual components that constitute either the genome, the transcriptome, the proteome, or the metabolome [1].

The genome is made of DNA molecules, which are the carrier of genetic information. It can be imagined as the blueprint library of the cell (*see* Fig. 1). From a chemical point of view, DNA molecules are polymers (or sequences) of simpler chemical units called nucleotides. There are four main types of nucleotides: adenine (A), thymine (T), cytosine (C), and guanine (G). DNA molecules are organized into chromosomes, which are compacted in the cell nucleus. The genome is directly connected to the transcriptome and the proteome (*see* next sections). The information to synthesize RNA molecules (transcriptome) and proteins (proteome) is encoded in specific regions of the DNA sequence called genes (*see* Fig. 1). Genes are made of successive nucleotides (clustered into codons), which correspond to amino acids, i.e., the molecules that constitute the proteins. The correspondence between nucleotides, codons, and amino acids is known as the genetic code. To summarize, a genomics dataset thus contains the sequences of DNA molecules present in a cell (or a population of cells) and can be seen as a copy of the cell's blueprint library (its genome) written as a long sequence of A, T, C, and G.

The transcriptome is made of RNA molecules. Multiple types exist, and they can be roughly classified into messenger RNA (mRNA), ribosomal RNA (rRNA), transfer RNA (tRNA), and non-coding RNA (ncRNA). Transcriptomics datasets mainly focus on mRNAs, which are the intermediate messengers between the genome and the proteome (*see* previous paragraph). The transcriptome is thus intimately connected to the genome and the proteome (*see* Fig. 1). Notably, the RNA polymerase is required to generate mRNA, reading the genome during transcription. In eukaryotes, mRNAs exit the nucleus to be used as templates by ribosomes (a macromolecular complex made of rRNA and proteins), to synthesize proteins by assembling amino acids (following the genetic code) during translation. Compared to the genome, the transcriptome is much more dynamic. The cell population of mRNA molecule varies according to cell requirement in proteins, and a transcriptomics dataset lists all sequences of mRNA present at a given time. They can be seen as snapshots of which parts of the genome are currently transcribed and in which proportion. Following up on the genome analogy presented in Fig. 1, mRNAs can be seen as active copies of the cell's blueprints that are more or less actively used.

The proteome is made of proteins, i.e., macromolecules made with one or several polymers of amino acids. Proteins are extraordinarily diverse in their three-dimensional (3D) conformations and associated functions. To illustrate this diversity, some proteins constitute the backbone of the cell structure, others detect or transmit external or internal chemical signals, and a large portion of them (enzymes) catalyze chemical reactions of the metabolism (the whole set of chemical reactions sustaining the cell). Proteins are also responsible for the regulation and expression (transcription and translation) of the genetic information (*see* previous paragraph). Protein functions are closely linked to their 3D spatial conformation, and all processes of the cells are based on protein activities (*see* Fig. 1). The proteome is as dynamic as the transcriptome because the set of proteins present at a given time in a cell varies accordingly to the current state and function of this cell. Proteomics datasets give a snapshot of which proteins are present at a given moment in the life of the cell. Genomics, transcriptomics, and proteomics resume the classical central dogma of biology, as first stated by Francis Crick in 1957. Even if it has been further detailed since, with, for instance, a better understanding of epigenomics, it still effectively summarizes the principal flow of information between the main molecular components of the cell: DNA is transcribed into RNA which is translated into proteins.

To end this description of omics data types, we believe it is important to mention the metabolome (*see* Fig. 1). The metabolome is made of metabolites, small molecules that are protein substrates in chemical reactions. Nucleotides and amino acids, cited before, are metabolites, as well as other molecules like lipids (forming bilayer membranes that compartmentalize the cell) or ATP (a molecule used as intracellular energy transfer). To extend, again, the analogy, metabolites can be seen as the raw materials used by the automated microscopic factory (*see* Fig. 1). Metabolomics datasets peek into the population of metabolites in a cell at a given time. Again, it is important to specify that if each cited "omics" field gives an assessment of its associated "ome" population, it is quite a "blurred" one. Everything is intertwined in a cell. Moreover, most omics studies give only an average observation on a population of cells. Multi-omics and single-cell techniques are trying to overcome these limitations.

In this chapter, we detail the different types of files used for omics data and present examples of databases where they are stored. We introduce different methods for generating omics data and finally provide some applications of omics data in fundamental research, cancer research, and pandemic response.

2 What Are Omics Data?

2.1 Results from High-Throughput Studies Written in Multiple Binary and Text Files

To describe the files used to store omics information, it is necessary to consider genomics and transcriptomics on one side and proteomics and metabolomics on the other side. Indeed, these files are generated by different experimental techniques, which are, respectively, sequencing (for genomics and transcriptomics) and mass spectrometry (for proteomics and metabolomics) (*see* Fig. 2). For each group, two types of files must be distinguished: the ones that are directly obtained after the applications of experimental protocols, i.e., the raw omics data files, and the ones that are generated by downstream informatic analyses, i.e., the processed omics data files (*see* Fig. 2). Experimental protocols and the informatic treatments applied to raw data files will be detailed in the next section.

Genomics and transcriptomics raw data files are essentially nucleotide sequence files. In that respect, the FASTA and the FASTQ text formats are commonly used. FASTA was created by Lipman and Pearson in 1985 as an input for their software [2] and became a de facto standard, without any clear statement acknowledging it [3]. This probably explains the absence of a common file extension (e.g., .fasta, .fna, .faa) even if FASTA is a unified file type. FASTA files contain one or several sequences. A sequence begins with a description line starting with the character ">". NCBI databases (*see* next sections) have unified rules to write this line.[1]

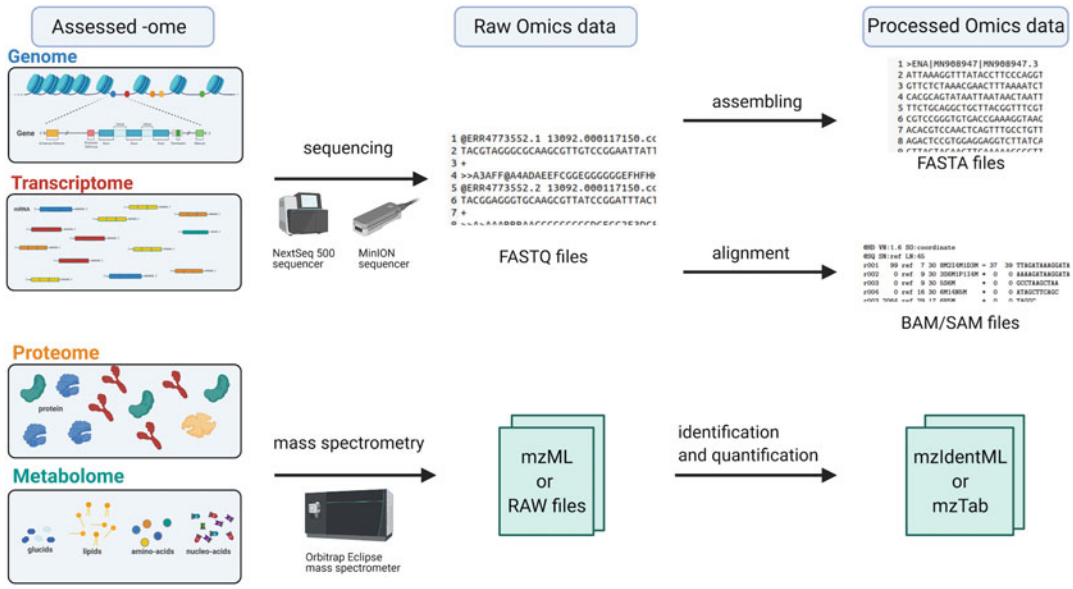


Fig. 2 Omics data are assessments of -ome populations. Raw omics data are generated through sequencing (for DNA and cDNA) or mass spectrometry (for proteins and metabolites)

[1] https://www.ncbi.nlm.nih.gov/genbank/fastaformat/

Subsequent lines contain the sequence itself split into multiple blocks of 60 to 80 characters (one per line). With nucleic acid sequences, the sequence lines are a series of A/T/C/G/U characters, representing the nucleic acids: adenine, thymine, cytosine, guanine, and uracil (the latter replacing thymine in RNA). FASTQ is the file format for the raw data generated by the sequencer in genomics and transcriptomics (*see* Fig. 2). The first two lines are similar as with FASTA file: identification line starts with "@" instead of ">" and the second line contains the nucleic sequence, but a quality score is associated with each position of the sequence (i.e., each letter in the sequence line). This score is called "Phred score," and it codes the probability of error in the identification of this nucleotide [3]. It goes from 0 to 62 and is coded in ASCII symbols. This allows to code any score using a single symbol, keeping the same length as the sequence line. FASTA and FASTQ files can be opened with any text editor software. FASTQ files are mainly lists of short sequences called "reads" (between 50 and 200 nucleic acids), which need to be processed (aligned or assembled) to be further analyzed. Alignment data files are one type of processed data. Indeed, reads in FASTQ files can be aligned to a reference genome sequence to allow further analyses (*see* below for pipeline description and example of applications). The text file format used in this case is the SAM[2] (sequence alignment and mapping) format [4, 5]. It can be further compacted into its binary equivalent, which are BAM or CRAM formats [6].

The file formats for proteomics and metabolomics data are not as homogeneous as for genomics and transcriptomics. At least 17 types of formats exist for mass spectrometry files (*see* below) [7]. Each machine manufacturer created its own, adapted to proprietary software to read and analyze it, thus multiplying formats. In an effort to facilitate data exchange and to avoid data loss (in case of no more readable old file formats), HUPO [8] and PSI[3] created the open-source mzML[4] format (XML text file with specific tag syntax) in 2011 [9]. In the main databases that host mass spectrometry result files, most of the files are in the RAW format, developed by Thermo Fisher Scientific. These binary files contain retention time, intensity, and mass-to-charge ratios (*see* later sections). Software like Peaks, Mascot, MaxQuant, or Progenesis [10, 11] use these files to identify proteins present in the sample and to quantify them. Results from these analyses are shared through two other text file formats: mzIdentML[5] and mzTab.[6]

[2] Sequence Alignment/Map Format Specification

[3] HUPO Proteomics Standards Initiative

[4] mzML 1.1.0 Specification | HUPO Proteomics Standards Initiative|mzML 1.1.0 Specification | HUPO Proteomics Standards Initiative

[5] mzIdentML | HUPO Proteomics Standards Initiative|mzIdentML | HUPO Proteomics Standards Initiative

[6] mzTab Specifications | HUPO Proteomics Standards Initiative|mzTab Specifications | HUPO Proteomics Standards Initiative

Note that many other file formats exist. One of the most critical for omics data analyses concerns the annotations of features on a DNA, RNA, or protein sequence. They are shared through the General Feature Format (GFF[7]) that is a text file with nine tabulated separated fields: sequence, source of the annotation, feature, start of the feature on the sequence, end of the feature, score, strand, phase, and attributes.

2.2 Results from High-Throughput Studies Shared Through Multiple Public Databases

The set of public biological databases hosting omics data is large and constantly evolving. Omics terminology started being regularly used in the 2000s. Between 1991 and 2016 (25 years), more than 1500 "molecular biology" databases were presented in publications, with a proliferation rate of more than 100 new databases each year [12]. These numbers are only the visible part of existing databases. How many have been created without being published? Around 500 of those databases are roughly co-occurrent with the apparition of the World Wide Web, the very Internet application allowing the creation of online databases. The availability of molecular biology databases decreased by only 3.8% per year from 2001 to 2016 [12]. This shows a sustained motivation from the community to create and maintain public platforms to share data. But it also highlights that this motivation comes more from a shared need for easy access to data rather than a supervised effort to coordinate approaches and unify sources. Such efforts indeed exist, for example, the ELIXIR project started in 2013 as an effort to unify all European centers and core bioinformatics resources into a single, coordinated infrastructure [13]. This notably produces the ELIXIR Core Data Resources (created in 2017), a set of selected European databases, meeting defined requirements, and the website "bio.tools," i.e., a comprehensive registry of available software programs and bioinformatics tools. The US National Center for Biotechnology Information (NCBI[8]) databases are also main references.

Given the "raw" nature of omics dataset, they are stored in archive data repositories: raw data from scientific articles, shared on databases easily accessible for reproducibility. Except for the Sequence Read Archive (SRA), the databases cited here are mixed ones: they host raw archive data and knowledge extracted from them. For genomics dataset, NCBI database Genome [14] and EMBL-EBI (member of ELIXIR) database Ensembl [15] are references. They organize genome sequences together with annotations and include sequence comparison and visual exploration tools. Transcriptomics data can be deposited into several databases, like Gene Expression Omnibus (GEO) [16] initially dedicated to microarray datasets, which is structured into samples forming

[7] GFF/GTF File Format
[8] NCBI

datasets. Tools are available to query and download gene expression profiles. The Sequence Read Archive (SRA) [17] accepts raw sequencing data. PRIDE [18] is a reference database for mass spectrometry-based proteomics data. Raw files containing spectra are available with associated identification and quantification information. For metabolomics data, MetaboLights [19] is an archive data repository and a knowledge database. It lists metabolite structures, functions, and locations alongside reference raw spectra. Those databases are generalist references, and many more specialized databases exist: 89 new databases are reported in the 2021 NAR database issue, and a dozen of them are omics specific [20]. For example, AtMAD is a repository for large-scale measurements of associations between omics in *Arabidopsis thaliana*, and Aging Atlas gathers aging-related multi-omics data [21, 22]. Finally, noteworthy is the existence of general-purpose open repositories like Zenodo,[9] which allow researchers to deposit articles, research datasets, source codes, and any other research-related digital information. Researchers thus receive credit by making their work more easily findable and reusable and hence support the application of the FAIR (findable, accessible, interoperable, reusable) data principles.[10]

Consistent efforts are made to cross-reference biological components (genes, proteins, metabolites) through the diversity of databases. Each database represents terabytes and petabytes of biological information (43,000 terabytes of sequence data just for SRA[11]), and the scale of the network they form through cross-reference is hard to conceptualize. This is the "big data" in biology and even more are generated every day.

3 How to Generate Omics Data?

Genomics started in 1977 with the application of the gel-based sequencing method developed by Sanger, to sequence for the first time the whole genome of a virus: the phage phiX. Only 13 years later, in 1990, the Human Genome Project began, aiming at sequencing three billion bases of the human genome, using capillary sequencing [23]. More than 10 years and almost three billion dollars later, this titanic task was accomplished [24]. When we think of omics analyses, microarray technology remains emblematic [25]. In the 2000s, the microarray represented the keystone of a discipline then called "post-genomics" [26]. Behind this terminology, the idea was that once the genomes are entirely sequenced,

[9] https://zenodo.org/

[10] https://www.go-fair.org/fair-principles/

[11] NCBI Insights: The wait is over... NIH's Public Sequence Read Archive is now open access on the cloud

new studies could be performed to understand their functioning. Microarrays thus emerged as a promising tool to monitor gene expression. They allow the quantification of the abundances of transcripts, which are associated with several thousands of different genes, simultaneously. Briefly, microarrays are slides, made of glass, on which probes have been attached. These probes are small DNA molecules, which have the particularity of being specific to one (and only one) gene. The experiment then consists of extracting mRNA molecules from a population of cells and transcribing them into complementary DNA (cDNA), labeled with a fluorescent molecule. These cDNAs are then hybridized on the glass slide and end up attached to the probes which are specific to them. They create a local fluorescent signal there. The higher the amount of mRNA, the more fluorescent signal is measured at each probe location position. Microarrays have been used to successfully study many biological processes, some fundamental such as the cell cycle [27] and others directly related to health issues such as human cancer [28]. It thus paved the road to new applications for sequencing technologies (*see* below).

3.1 High-Throughput Sequencing Technologies

From 2007, new methods called next-generation sequencing (NGS) [29] helped to considerably reduce cost, technical difficulties, and duration of the process.

Illumina is the currently predominant NGS method (*see* Fig. 3). After extraction, the DNA molecules are sequenced by synthesis (SBS) on a flow cell. Thanks to sequence adaptors, each DNA molecule is amplified by bridge amplification as a cluster of copies on the flow cell. The reading of the flow cell is based on optical detection: each time a DNApol adds a new nucleotide, a flash of light is detected. NGS advantage, compared to older Sanger techniques, is to allow massive parallel sequencing of large numbers of short sequences (between 50 and 250 nucleotides) called "reads." The limit of this technique is the size of the fragments, but Illumina technology has very high fidelity (very low error rate).

MinION of Oxford Nanopore is another well-established NGS technology [30]. It is based on electronic detection through a nanopore (*see* Fig. 3). When there is an electric potential around a membrane (measurable as a voltage between the two sides), the passage of a macromolecule through a nanopore (a modified biological protein canal) triggers small changes in this electric potential. The changes are distinctive in function of the current nucleotide in the nanopore. So, the succession of electronic potential variation can be associated as the nucleotide sequence. This is the fundamental concept behind MinION technology, and the main advantage is the length of the sequenced molecules. Without the technical necessity of flow cells, the sequence passing through the nanopore can be very long (order of magnitude of a thousand instead of a hundred base pairs) [31]. But given that the physical

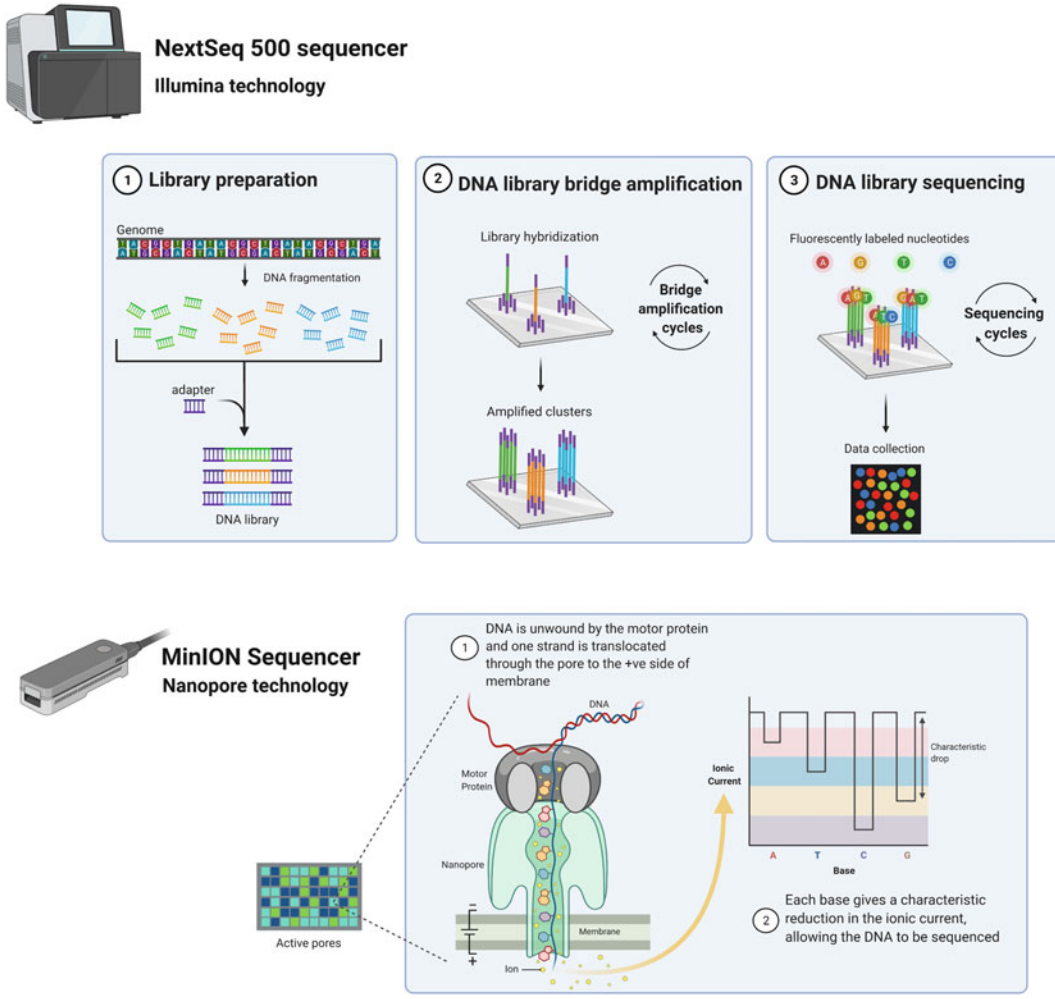


Fig. 3 Illumina and MinION sequencing technologies. Illumina is a sequencing by synthesis technology that allows massive parallel sequencing of small DNA molecules. MinION is a nanopore-based technology that allows the sequencing of longer DNA molecules

signal detected is small variations of an electric potential, the sequencing is less reliable (higher error rate). Depending on the fidelity of the sequencing or the size of the sequence needed, SBS and nanopore-based techniques are complementary.

The sequencing machine output is a group of FASTQ files (*see* previous section). For genomics data, fragments must be assembled to obtain a single sequence of the genome. For transcriptomics data, fragments can be aligned on a reference genome to observe which genes are transcribed at a given time (transcriptome de novo assembly is also possible but still very challenging). Therefore, to extract information from the FASTQ files produced by the sequencer, two main processing steps are needed. The numerous small sequences (reads) stored in the file must be aligned to a

reference genome (mapping), and then the count of reads aligned to a gene sequence gives an estimation of its level of transcription (quantification). Dozens of bioinformatics tools have been developed over the years for mapping (STAR [31], TopHat [32], HISAT2, Salmon [33]) and quantification (featureCounts [34], Cufflinks [35]). Benchmarking studies highlight similar performance for most of them [36–38]. Interestingly, TopHat2 exhibits an alignment recall on simulated malaria data that varies from under 3% using defaults to over 70% using optimized parameters [39]. This underlines the impact of parameter optimization on result quality. Quantification tools generate a text file summarizing the level of transcription of each gene in each condition into a matrix of counts.

3.2 Mass Spectrometry Technologies

Since the first use of a mass spectrometer for protein sequencing in 1966 by Biemann,[12] the improvement of mass spectrometer is closely linked to proteomics and metabolomics development [40]. Metabolites and proteins cannot be read as templates like DNA or RNA, and so they neither can be amplified nor sequenced by synthesis. To access their sequence, the main tool is the mass spectrometer. In the classical bottom-up approach, proteins are digested into small peptides, which pass through a chromatography column. They are then sequentially sprayed as ions into the spectrometer. Migration through the spectrometer allows separation of the peptides according to their mass-to-charge ratio. For each fraction exiting the column, an abundance is calculated. In a data-dependent acquisition (DDA), a few peptides with an intensity superior to a given threshold are isolated one at the time. They are fragmented, and additional spectra (mass-to-charge ratio and intensity) are generated for each fragmented ion. In a data-independent acquisition (DIA), a spectrum is generated for all fractions coming out of the chromatography column. Obtained spectra are a combination of spectra corresponding to each peptide present in each original fraction. Comparison with a peptide spectrum library generated in silico is therefore required to allow the deconvolution of those complex spectra. All this information (abundances in fractions, mass-to-charge ratios, intensities) is stored into .raw files, which can only be read by dedicated software (*see* Subheading 2.1).

3.3 Single-Cell Strategies

Most omics experiments are bulked, and they are an average measure done on a population of cells, which is more or less homogeneous. Single-cell omics allow a more precise measurement, highlighting the plasticity of the cell system. Single-cell techniques started with manual separation of a single cell under a microscope in 2009 [41] and quickly evolved toward techniques allowing the

[12] HUPO—Proteomics Timeline

parallel sequencing of thousands of cells [42]. Plate-based techniques use flow cytometry to separate isolated cells into the different wells of a plate, allowing processing of hundreds of cells. The introduction of nanometric droplets to separate isolated cells allowed the parallel processing of thousands of cells thanks to individual barcoding [43, 44]. Cells isolated from tissues are mixed with microparticles in a buffer that forms droplets in oil. Most droplets are empty, but some contain both a microparticle and a cell. After cell lysis, oligonucleotide primers on the microparticles allow the capture of the cell mRNA (by oligo-dT and polyA tail complementarity). Primers on the same microparticle are barcoded, thus creating a cell tag on each sequence. Amplification and sequencing can be bulked without losing the cell of origin for each transcript. Several bioinformatics tools are specialized for single-cell transcriptomics data [45]. For example, Cell Ranger and Loupe Browser are, respectively, four pipelines (mapping, quantification, and downstream analysis) and a visualization tool developed by 10× Genomics [44]. Single-cell transcriptomics data are challenging for bioinformatics analysis because of their high level of technical noise and the multifactorial variability between cells [45]. Transcriptomics is the more advanced single-cell omics, but single-cell genomics is also used in SNP and copy number variation screening (*see* Subheading 4.2).

Proteomics and metabolomics data are still challenging to obtain at a single cell level: one cell yields only 250–300 pg [46] of proteins when MS in-depth measurement still necessitates population scale yield. But thanks to innovations in sample preparation and experimental design, single-cell proteomics assessments scaled up from a few hundred to more than a thousand identified proteins in just 4 years [47].

4 Which Applications for Omics Data?

4.1 In Fundamental Research

Describing biological systems implies to identify, quantify, and functionally connect their individual molecular components. Given the diversity of cellular components and their multiple interlocking functions, the large scale of omics data empowers the characterization of biological systems. As stated before, each type of "omics" is an assessment of a specific subpopulation of molecular components. Mining omics data thus allows bulk identification of the nature (sequence and structure), location, function, and abundance of molecular components in those subpopulations.

Genomics data are making the genome sequences of thousands of species accessible. The first direct application of these resources is the annotation of genomic features onto those genomic sequences: protein-coding genes, tRNA and rRNA genes, pseudogenes, transposons, single-nucleotide polymorphisms, repeated regions, telomeres, centromeres... Genomic features are numerous, and DNA sequences alone can be enough to recognize

patterns specific to some of them. For example, specific tools exist to detect protein-coding genes, like Augustus[13] [48]. The annotation can be based only on sequence patterns or also on comparison with another sequence. Comparative genomics, i.e., the comparison of genome sequences, allows the transfer of knowledge for homolog genes (evolutionarily related genes) between species. Bioinformatics tools exist to infer evolutionary relationships between genes based on their sequence similarity [49]. Understanding the evolution of the genome helps to understand the dynamics behind phenotypic convergence, population evolutions, speciation events, and natural selection processes. For example, the study of 17 marine mammals' genomes offered insight into the macroevolutionary transition of marine mammal lineages from land to water [50].

Transcriptomics data give insight on the levels of gene transcription. The resulting count matrix (*see* previous section) is mainly used to carry out differential expression analysis (DEA) of genes between conditions. Conditions differ by the variation of a single factor: a mutation, a different medium, or a stimulus. Basic DEA is a multi-step workflow [51] that allows the detection of statistically significant variations in expression across conditions. The final goal is to deduce insight on the gene's functions from the observed variations. Transcriptomics data are also used to increase the quality of genome annotation. The presence of hypothetical genes can be verified by their transcription, the exact structure of known genes can be refined (size of UTRs and exons; *see* Fig. 1), and previously undetected genes can be observed [52].

Proteomics data allows the identification and quantification of proteome. Proteome does not totally correlate with transcriptome. RNA can be spliced (assembly of the mRNA from exons, not always the same and in the same order), and proteins undergo several post-translational modifications (minor changes in the chemical structure of the protein) and re-localization [53]. Cellular pathways and phenotypes thus cannot be fully understood only through transcriptomics assessments. Proteomics completes the information given by genomics and transcriptomics. It describes the third -ome of the central dogma of biology (*see* Fig. 1).

Multi-omics analysis, taking advantage of several omics insights in the same experimental approach, comes with several challenges. Generating several types of omics data comes with a significant investment in time, skilled manpower, and money [1]. Even if generated in the same experimental approach, omics data are heterogeneous by nature, thus complexifying their integration. If challenging, multi-omics datasets are also a step toward the systemic description of biological systems [54].

[13] Augustus/ABOUT.md at master

4.2 In Medical Research

An early application of genomics in medical research is the genome-wide association studies (GWAS). By comparing genome sequences from a large population of individuals (both healthy and sick), GWAS highlight SNPs (single-nucleotide polymorphisms) that are significantly more frequent in individuals with the disease. Correlation does not mean causality, but GWAS can give a first clue of the metabolic pathways or cellular components involved in the disease [55]. This strategy has proven to be efficient in the case of "common complex diseases." Unlike Mendelian diseases (which are rarer), the heritability (genetic origin) of these diseases depends on hundreds of SNPs with small effect sizes, which GWAS studies help identify [56]. Alzheimer's disease and cancers are examples of "common complex diseases" whose genetic underpinnings have been explored through GWAS [55, 57].

Most cancers emerge from the successive alteration of cell functioning (by accumulation of mutations), leading to abnormal growth causing tumors and metastasis. Multi-omics studies can highlight the underlying molecular mechanisms of cancer development, better explain resistance to treatment, and help classify cancer types. Screening cohorts of patients helps assess alleles associated with the development of certain types of cancer. The different subtypes for breast cancer are a well-documented example [58].

Single-cell genomics is the only way of characterizing rare cellular types such as cancer stem cells [59]. Single-cell omics data are also used to follow the rapid evolution of cancer cell population inside tumors. Understanding and describing cancer cell population dynamics is crucial: the characteristic accelerated rate of mutation can be the cause of treatment resistance. Omics data specific to cancer cell lines are shared on specific databases driven and maintained by global consortium such as the Cancer Genome Atlas Program[14] (over 2.5 petabytes of genomics, epigenomics, transcriptomics, and proteomics data) or the International Cancer Genome Consortium [60].

Omics data proved to be a priceless resource in pandemic response. The virus severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) causing the COVID-19 disease quickly spread around the world, causing more than six million deaths (as of March 2022) and a global health crisis. Its RNA sequence was obtained in January 2020 and allowed the development of detection kits and later RNA-based vaccines. Since the beginning of the pandemic, the genomic evolution of the virus is followed almost in real time, as new variants (with mutations affecting mostly the spike protein of the virus envelope) are sequenced. Variant profiling allows the World Health Organization to closely monitor variants of concern. The precise characterization of the virus structure

[14] https://www.cancer.gov/tcga

opens the research of therapeutic targets. Multi-omics studies helped specify the COVID-19 biomarkers, pathophysiology, and risk factors [61].

Getting omics data in brain tissue studies is promising but challenging because of brain specificity. Indeed, except in a few specific diseases where in vivo resections are performed (brain tumors, surgically treated epilepsy, etc.), human brain samples are collected postmortem, when the less stable molecule populations are already significantly altered. For example, studies of the brain transcriptome are deeply impacted. On the other hand, some omics studies target peripheral fluids (e.g., plasma, cerebrospinal fluid, etc.) with the aim to find biomarkers, but the relationships between observations in peripheral fluids and pathophysiological mechanisms in the brain are far from clear. Moreover, the brain is organized as a network of intricate substructures, constituted of several cell types (glial cells and different neuron types) with distinct function and thus different omics landscape [62]. Nonetheless, multi-omics exploratory studies are describing complex diseases in a systematic paradigm, highlighting diversity of cellular dysregulations linked to complex pathologies like Alzheimer's disease [57].

5 Conclusion

Genomics, transcriptomics, proteomics, and metabolomics are arguably the most developed and used omics, but they are not the only ones. Other omics describe other sides of the functioning of the cell, which require intricate relationships between omics levels. For example, epigenomics describes the transitory chemical modifications of DNA, and lipidomics looks at the lipidic subpopulation of metabolites (*see* Fig. 1). Omics diversity mirrors the complexity of cell systems. With the constant improvement of measurement techniques, possibilities to assess ever larger subsystems of the cells are increasing. Omics dataset generation is paired with the development of software, essential tools to generate, read, and analyze them. By design, computer science is therefore omnipresent in modern "big data" biology. The need for more gold standard analysis pipelines and file formats grows with the scale and complexity of produced datasets.

Acknowledgments

This work was funded by the Agence Nationale pour la Recherche (MINOMICS, grant number ANR-19-CE45-0017).
The authors are grateful to Sarah Cohen-Boulakia for reviewing this chapter.
Figures were made on BioRender, using icons from Flaticon.

References

1. Hasin Y, Seldin M, Lusis A (2017) Multi-omics approaches to disease. Genome Biol 18:83
2. Lipman DJ, Pearson WR (1985) Rapid and sensitive protein similarity searches. Science 227:1435–1441
3. Cock PJA, Fields CJ, Goto N et al (2010) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771
4. Li H, Handsaker B, Wysoker A et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079
5. Danecek P, Bonfield JK, Liddle J et al (2021) Twelve years of SAMtools and BCFtools. GigaScience 10:giab008
6. Hsi-Yang Fritz M, Leinonen R, Cochrane G et al (2011) Efficient storage of high throughput DNA sequencing data using reference-based compression. Genome Res 21:734–740
7. Deutsch EW (2012) File formats commonly used in mass spectrometry proteomics. Mol Cell Proteomics 11:1612–1621
8. Deutsch EW, Lane L, Overall CM et al (2019) Human proteome project mass spectrometry data interpretation guidelines 3.0. J Proteome Res 18:4108–4116
9. Martens L, Chambers M, Sturm M et al (2011) mzML—a community standard for mass spectrometry data. Mol Cell Proteomics 10(R110): 000133
10. Ma B, Zhang K, Hendrie C et al (2003) PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. Rapid Commun Mass Spectrom 17:2337–2342
11. Välikangas T, Suomi T, Elo LL (2017) A comprehensive evaluation of popular proteomics software workflows for label-free proteome quantification and imputation. Brief Bioinform
12. Imker HJ (2018) 25 years of molecular biology databases: a study of proliferation, impact, and maintenance. Front Res Metr Anal 3:18
13. Harrow J, Drysdale R, Smith A et al (2021) ELIXIR: providing a sustainable infrastructure for life science data at European scale. Bioinformatics 37:2506–2511
14. Benson DA, Karsch-Mizrachi I, Lipman DJ et al (2010) GenBank. Nucleic Acids Res 38: D46–D51
15. Howe KL, Achuthan P, Allen J et al (2021) Ensembl 2021. Nucleic Acids Res 49:D884–D891
16. Barrett T, Wilhite SE, Ledoux P et al (2012) NCBI GEO: archive for functional genomics data sets—update. Nucleic Acids Res 41: D991–D995
17. Leinonen R, Sugawara H, Shumway M et al (2011) The sequence read archive. Nucleic Acids Res 39:D19–D21
18. Perez-Riverol Y, Csordas A, Bai J et al (2019) The PRIDE database and related tools and resources in 2019: improving support for quantification data. Nucleic Acids Res 47: D442–D450
19. Haug K, Cochrane K, Nainala VC et al (2019) MetaboLights: a resource evolving in response to the needs of its scientific community. Nucleic Acids Res gkz1019
20. Rigden DJ, Fernández XM (2021) The 2021 *Nucleic Acids Research* database issue and the online molecular biology database collection. Nucleic Acids Res 49:D1–D9
21. Lan Y, Sun R, Ouyang J et al (2021) AtMAD: *Arabidopsis thaliana* multi-omics association database. Nucleic Acids Res 49:D1445–D1451
22. Aging Atlas Consortium, Liu G-H, Bao Y et al (2021) Aging Atlas: a multi-omics database for aging biology. Nucleic Acids Res 49:D825–D830
23. Karger BL, Guttman A (2009) DNA sequencing by CE. Electrophoresis 30:S196–S202
24. International Human Genome Sequencing Consortium (2004) Finishing the euchromatic sequence of the human genome. Nature 431: 931–945
25. Schena M, Shalon D, Davis RW et al (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470
26. Gershon D (1997) Bioinformatics in a post-genomics age. Nature 389:417–418
27. Spellman PT, Sherlock G, Zhang MQ et al (1998) Comprehensive identification of cell cycle–regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9:25
28. DeRisi J, Penland L, Brown PO, Bittner ML, Meltzer PS, Ray M, Chen Y, Su YA, Trent JM (1996) Use of a cDNA microarray to analyse gene expression patterns in human cancer. Nat Genet 14(4):457–460
29. Metzker ML (2010) Sequencing technologies — the next generation. Nat Rev Genet 11:31–46

30. Deamer D, Akeson M, Branton D (2016) Three decades of nanopore sequencing. Nat Biotechnol 34:518–524
31. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21
32. Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36
33. Patro R, Duggal G, Love MI et al (2017) Salmon provides fast and bias-aware quantification of transcript expression. Nat Methods 14: 417–419
34. Liao Y, Smyth GK, Shi W (2014) featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930
35. Trapnell C, Roberts A, Goff L et al (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc 7:562–578
36. Teng M, Love MI, Davis CA et al (2016) A benchmark for RNA-seq quantification pipelines. Genome Biol 17:74
37. The RGASP Consortium, Engström PG, Steijger T et al (2013) Systematic evaluation of spliced alignment programs for RNA-seq data. Nat Methods 10:1185–1191
38. Schaarschmidt S, Fischer A, Zuther E et al (2020) Evaluation of seven different RNA-Seq alignment tools based on experimental data from the model plant Arabidopsis thaliana. Int J Mol Sci 21:1720
39. Baruzzo G, Hayer KE, Kim EJ et al (2017) Simulation-based comprehensive benchmarking of RNA-seq aligners. Nat Methods 14: 135–139
40. Biemann K, Tsunakawa S, Sonnenbichler J et al (1966) Structure of an odd nucleoside from serine-specific transfer ribonucleic acid. Angew Chem Int Ed Engl 5:590–591
41. Tang F, Barbacioru C, Wang Y et al (2009) mRNA-Seq whole-transcriptome analysis of a single cell. Nat Methods 6:377–382
42. Svensson V, Vento-Tormo R, Teichmann SA (2018) Exponential scaling of single-cell RNA-seq in the past decade. Nat Protoc 13: 599–604
43. Macosko EZ, Basu A, Satija R et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 161:1202–1214
44. Zheng GXY, Terry JM, Belgrader P et al (2017) Massively parallel digital transcriptional profiling of single cells. Nat Commun 8:14049
45. Stein CM, Weiskirchen R, Damm F et al (2021) Single-cell omics: overview, analysis, and application in biomedical science. J Cell Biochem 122:1571–1578
46. Jehan Z (2019) Single-cell omics: an overview. In: Single-cell omics. Elsevier, pp 3–19
47. Kelly RT (2020) Single-cell proteomics: progress and prospects. Mol Cell Proteomics 19: 1739–1748
48. Stanke M, Morgenstern B (2005) AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res 33:W465–W467
49. Quest for Orthologs consortium, Altenhoff AM, Boeckmann B et al (2016) Standardized benchmarking in the quest for orthologs. Nat Methods 13:425–430
50. Yuan Y, Zhang Y, Zhang P et al (2021) Comparative genomics provides insights into the aquatic adaptations of mammals. Proc Natl Acad Sci 118:e2106080118
51. Van den Berge K, Hembach KM, Soneson C, Tiberi S, Clement L, Love M, Patro R, Robinson MD (2019) RNA sequencing data: Hitchhiker's guide to expression analysis. Annu Rev Biomed Data Sci 2:139–173
52. Chen G, Shi T, Shi L (2017) Characterizing and annotating the genome using RNA-seq data. Sci China Life Sci 60:116–125
53. Arivaradarajan P, Misra G (eds) (2018) Omics approaches, technologies and applications: integrative approaches for understanding OMICS data. Springer Singapore, Singapore
54. Veenstra TD (2021) Omics in systems biology: current progress and future outlook. Proteomics 21:2000235
55. Tam V, Patel N, Turcotte M et al (2019) Benefits and limitations of genome-wide association studies. Nat Rev Genet 20:467–484
56. Uitterlinden A (2016) An introduction to genome-wide association studies: GWAS for dummies. Semin Reprod Med 34:196–204
57. Hampel H, Nisticò R, Seyfried NT et al (2021) Omics sciences for systems biology in Alzheimer's disease: state-of-the-art of the evidence. Ageing Res Rev 69:101346
58. Kohler BA, Sherman RL, Howlader N et al (2015) Annual report to the nation on the status of cancer, 1975–2011, featuring

incidence of breast cancer subtypes by race/ethnicity, poverty, and state. JNCI J Natl Cancer Inst 107

59. Liu J, Adhav R, Xu X (2017) Current progresses of single cell DNA sequencing in breast cancer research. Int J Biol Sci 13:949–960
60. Zhang J, Bajari R, Andric D et al (2019) The international cancer genome consortium data portal. Nat Biotechnol 37:367–369
61. Overmyer KA, Shishkova E, Miller IJ et al (2021) Large-scale multi-omic analysis of COVID-19 severity. Cell Syst 12:23–40.e7
62. Naumova OY, Lee M, Rychkov SY et al (2013) Gene expression in the human brain: the current state of the study of specificity and spatio-temporal dynamics. Child Dev 84:76–88

Visual integration of omics data to improve 3D models of fungal chromosomes

Thibault Poinsignon[1,2], Mélina Gallopin[1], Pierre Grognet[1], Fabienne Malagnac[1], Gaëlle Lelandais[1]* and Pierre Poulain[2]*

*Equal contribution, corresponding authors: gaelle.lelandais@universite-paris-saclay.fr ; pierre.poulain@u-paris.fr.

[1]Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Université Paris-Saclay, 91198 Gif-sur-Yvette, France.

[2]Université Paris Cité, CNRS, Institut Jacques Monod, F-75013 Paris, France.

Running title: Seeing the 3D organization of fungal chromosomes

Abstract

The functions of eukaryotic chromosomes and their spatial architecture in the nucleus are reciprocally dependent. Hi-C experiments are routinely used to study chromosome 3D organization by probing chromatin interactions. Standard representation of the data has relied on contact maps that show the frequency of interactions between parts of the genome. In parallel, it has become easier to build 3D models of the entire genome based on the same Hi-C data, and thus benefit from the methodology and visualization tools developed for structural biology. 3D modeling of entire genomes leverages the understanding of their spatial organization. However, this opportunity for original and insightful modeling is under exploited. In this paper, we show how seeing the spatial organization of chromosomes can bring new perspectives to Hi-C data analysis. We assembled state-of-the-art tools into a workflow that goes from Hi-C raw data to fully annotated 3D models and we re-analysed public Hi-C datasets available for three fungal species. Besides the well-described properties of the spatial organization of their chromosomes (Rabl conformation, hypercoiling and chromosome territories), our 3D models highlighted *i)* in *Saccharomyces cerevisiae*, the backbones of the cohesin anchor regions, which were aligned all along the chromosomes, *ii)* in *Schizosaccharomyces pombe*, the oscillations of the coiling of chromosome arms throughout the cell cycle and *iii)* in *Neurospora crassa*, the massive relocalization of histone marks in mutants of heterochromatin regulators. 3D modeling of the chromosomes brings new opportunities for visual integration. This holistic perspective supports intuition and lays the foundation for building new concepts.

Introduction

What if it were possible to see all the details of chromosomes inside the nucleus of a cell? In eukaryotic cells, the nucleus is a dynamic organelle which is highly organized and characterized by extensive compartmentalization of structural components in its three-dimensional space (see (Razin et al. 2014; Dundr and Misteli 2001; Cremer and Cremer 2001; Arifulin et al. 2018) for reviews). In such a crowded environment, the arrangement of chromosomes is constrained and requires the formation of multiple chromatin domains to limit gene positions to preferred locations within the nuclear space (Misteli 2020). The spatial organization of chromosomes is of great interest, helping molecular biologists represent the objects they work with, and understand their interactions. Even if immense progress has been made in cell imaging, biological molecules (e.g. DNA, RNA, proteins) are too small to be individualized with optical microscopes (Wong and Eleftheriades 2013; Jensen 2013) and consequently, the interior of the cell (and the interior of a nucleus even more) remains largely invisible to the human eye. Alternative solutions are based on molecular-scale techniques like X-ray crystallography (Smyth 2000), NMR microscopy (Reckel et al. 2005) or electron microscopy (Radulović et al. 2022). By analyzing atom arrangements in molecules, these techniques produce informative views of macromolecular complexity (Nogales and Scheres 2015; Baumeister 2022), but they require complex technical skills and expensive equipment. Furthermore, it is important to keep in mind that in the end, these images are still artificial representations of reality. In other words, they are “models”.

The use of models is widespread in biology. From model organisms to model systems, their interest is to understand a phenomenon in a simplified context, in order to, later, generalize to more complex situations. In cell biology for instance, structural models are used to represent cell components (membranes, nucleus, cytoplasm, etc.), to understand their organization, and to describe their constituent molecules (Im et al. 2016). The work of David Goodsell provides an emblematic example (Goodsell 2009). His drawings representing cellular compartments and their molecular actors are so striking because of the unexpected density of molecules and the

complexity of their organization. Goodsell's illustrations have been featured as “Molecule of the Month" on the Protein Data Bank website for over twenty years and the scientific journal Nature chose one of his paintings to make the cover of a special issue on COVID-19 (August 20, 2020 issue). Models make it possible to represent and summarize, in an intuitive but still scientifically rigorous way, the massive knowledge of cell molecular structures. Creating them thus represents a stimulating challenge, at the crossroads of multiple disciplines (biology, physics, computer science, art) (O’Donoghue 2021).

Modeling chromosomes is challenging because they belong to the mesoscale, *i. e.* a length-scale that is larger than discrete molecular complexes yet still remains intracellular (Sear et al. 2015). As a consequence, they are both too small to be observed with precision under optical microscopes, and too large to be fully modeled at the atomic scale. In eukaryotes, chromosomes are long DNA molecules, tightly packed to fit within the nucleus, a space only a fraction of their length. To this end, DNA is wrapped around histone proteins to form nucleosomes, stacked to form chromatin fibers, themselves arranged into higher-order chromatin architecture (Misteli 2020). In this study, we are specifically interested in seeing the spatial organization of fungal chromosomes. Our laboratory has long-standing expertise in functional genomics projects in yeasts (Denecker et al. 2020; Poinsignon et al. 2022; Sénécaut et al. 2022) or filamentous fungi (Grognet et al. 2019; Carlier et al. 2021; Lelandais et al. 2022), and spatial genome organization in fungi has already been investigated, for model species like *Saccharomyces cerevisiae* (Duan et al. 2010; Tokuda et al. 2012), *Schizosaccharomyces pombe* (Grand et al. 2014; Tanizawa et al. 2010; Gallardo et al. 2019; Noma 2017) and *Neurospora crassa* (Galazka et al. 2016; Rodriguez et al. 2022). Notably, the nuclear architecture of *N. crassa*, a multicellular fungus that grows as a mycelium with a network of hyphae (Galagan et al. 2003), has structural homology (thanks to the existence of heterochromatin and euchromatin) with the human genome (Rodriguez et al. 2022). This makes the *N. crassa* genome a cost-efficient model to study chromosome conformation. *S. cerevisiae* and *S. pombe* are distantly related yeast species (evolutionary distance of at least 400 Mya) that represent very different models of unicellular

eukaryotes. Their genomes, of similar size (~12 Mb), are organized into different sets of chromosomes (16 chromosomes for *S. cerevisiae* and 3 chromosomes for *S. pombe*). These yeast genomes are more than three times shorter than the genome of *N. crassa* (genome size is 41 Mb, with 7 chromosomes ranging from 4 to 10 Mb). Altogether, *N. crassa*, *S. cerevisiae* and *S. pombe* represent an interesting diversity of genomic situations for which much knowledge of nuclear organization is available in the literature (to view representative data from several publications, see **Supplementary Figure S1**).

The most emblematic feature highlighted in these articles is the “Rabl conformation”. This is a particular positioning of chromosomes in which their centromeres and telomeres are respectively clustered at distinct peripheral locations, inside the nuclear envelope. The Rabl conformation has been observed in *N. crassa*, *S. cerevisiae* and *S. pombe* (**Supplementary Figure S1**). It has been found in plants (Rodriguez-Granados et al. 2016), mice (Stevens et al. 2017; Zhang et al. 2020) and flies (Bauer et al. 2012), underpinning the idea that it represents a conserved constrained genome structure to limit topological entanglement of chromosomes (Pouokam et al. 2019). The clustering of centromeres and telomeres at distinct peripheral locations inside the nuclear envelope induces chromosomes to be organized in a “clothespin-like” structure, meaning that they are folded in half at the centromere, their arms against each other, their ends slightly bent. In *S. pombe*, the clustering of centromeres and telomeres is reinforced by protein anchors to the nuclear envelope and heterochromatin epigenetic marks (**Supplementary Figure S1**). At the molecular level, 3D genome organization in yeasts is determined by cohesin and condensin mediated structures (36,37). In yeasts, cohesin seems to form 40~50 kb globules (Mizuguchi et al. 2014; Tanizawa et al. 2017), and condensin 300~500 kb large domains (Noma 2017; Tanizawa et al. 2017; Kim et al. 2016). This highlights how 3D genome structures are organized at successive, nested scales of condensation patterns. Another important feature relates to the definition of “chromosome territories", where chromosomes are folded within their own space in the nucleus (Fritz et al. 2019). Those territories are another well described example of the functional importance of chromatin structure and are observed in many different organisms. Well

known in yeasts, chromosome territories are still debated in *N. crassa* (Rodriguez et al. 2022). Nonetheless correlations between 3D structuration and heterochromatin/euchromatin epigenetic marks were revealed by Hi-C and ChIP-seq studies (Rodriguez et al. 2022; Pouokam et al. 2019; Fritz et al. 2019). Centromeres and telomeres are heterochromatic (as in yeasts) and euchromatic regions are condensed into 20~40 kb globules (consistent with yeast cohesin globules).

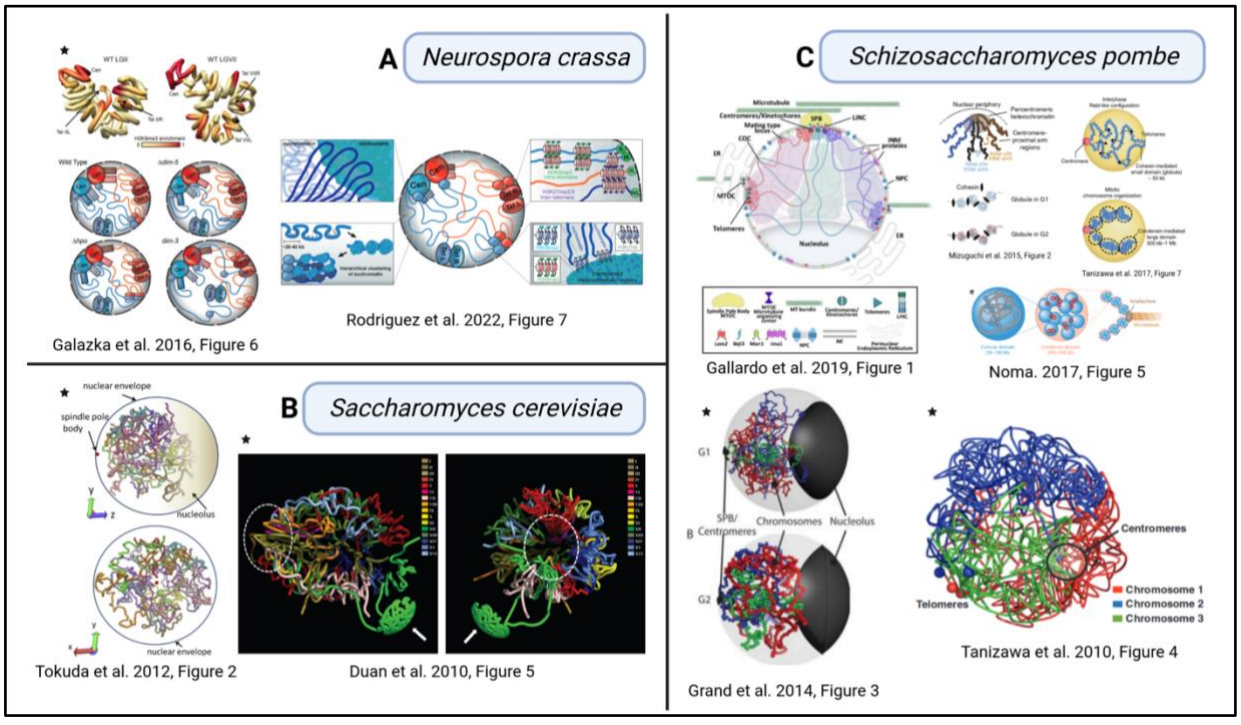
This knowledge of the spatial organization of *N. crassa*, *S. cerevisiae* and *S. pombe* chromosomes was mainly built on Hi-C data analyses. The initial 3C technique evolved rapidly with NGS technologies, and recent technical improvements allow single cell analysis, with better spatial resolution (Hsieh et al. 2016) or haploid investigations (Oomen et al. 2020; Mitter et al. 2020) (see (Kempfer and Pombo 2020; Jerkovic´ and Cavalli 2021) for reviews). Various types of Hi-C data are thus available in public databases, differing in accuracy and resolution, for numerous species (Hoencamp et al. 2022), far beyond fungal models. Their processing steps are now well described (see (Jerkovic´ and Cavalli 2021) for review), and bioinformatics tools like the Juicer suite (Durand et al. 2016) or the HiC-Pro pipeline (Servant et al. 2015) are available and widely used by the community. Generally, the analysis of Hi-C data ends with the representation of a "contact map" (symmetrical heatmap). A contact map shows the frequency of contacts observed between different portions of a genome, whose sizes depend on a predefined resolution (generally around 5 to 10 kb). Contact maps present the advantage of summarizing in a single image, all pairwise distances between genomic regions, both at short and long-range. Contact maps thus reveal chromosome clusters, their organization into domains and the formation of DNA loops. They are valuable because of their precision (resolution can go down to a few kb) and the information density they convey. Still, their biological interpretation is not trivial and, in the end, contact maps remain abstract representations of the spatial organization of chromosomes.

3D modeling of Hi-C contact maps is an interesting strategy to boost interpretation of Hi-C experimental results (Yardımcı and Noble 2017; Oluwadare et al. 2019). It consists in calculating 3D coordinates (x, y, z) for all genomic intervals shown in a contact map, so that their pairwise Euclidean distances remain consistent with their contact frequencies in the original map. 3D

modeling of the spatial organization of chromosomes is not trivial, but several software packages based on different strategies exist. For instance, “distance-based methods” convert the number of contacts stored in the contact maps into distances and then resolve an optimization problem to fit 3D coordinates to those distances (Rieber and Mahony 2017; Li et al. 2018). Such methods often require fulfilling physical constraints related to known structural features of the genomes (for instance centromere and telomere clustering in opposite locations inside the nucleus or rDNA clustering outside the overall structure to form the nucleolus). This is how the 3D models already proposed for *N. crassa*, *S. cerevisiae* and *S. pombe* were obtained (**Supplementary Figure S1**, black stars). Alternatively, “probability-based strategies" model contact numbers using random variables. This offers the advantage of taking into account the nature of Hi-C measurements (an average over a cell population) (Varoquaux et al. 2014). As an illustration, the Pastis-NB method (implemented in the Pastis software (Varoquaux et al. 2021)) implements a probability-based strategy based on negative binomial random variables. This distribution properly models count data generated by sequencing technologies, as illustrated by the wide adoption of negative binomial distributions to model RNA-seq data (Love et al. 2014). The model therefore accounts for over-dispersion, performs well in low coverage settings and doesn't require any initial physical constraints. Once a 3D model has been generated, a large panel of tools exists to visualize the 3D coordinates of genome models: Genome3D, GMOL, GenomeFlow, HiC-3DViewer, Csynth, WashU Epigenome Browser (Asbury et al. 2010; Nowotny et al. 2016; Trieu et al. 2019; Djekidel et al. 2017; Todd et al. 2021; Li et al. 2022). Initially limited to the context of viewing biomolecules (nanoscale), visualization software was thus extended to representation of the 3D organization of chromosomes at the mesoscale.

Surprisingly, recent Hi-C studies on *S. cerevisiae*, *S. pombe* and *N. crassa* do not feature 3D models (Rodriguez et al. 2022; Tanizawa et al. 2017; Costantino et al. 2020). In a context where *i*) an increasing number of Hi-C studies are generated, *ii)* methods for 3D modeling of Hi-C contacts exist and *iii)* software for 3D visualization is routinely available in structural biology, one may wonder why contact maps are not associated with 3D models more often. In the present

study, we explored the potential of using 3D models as well as Hi-C contact maps to better understand the spatial organization of fungal chromosomes. For this purpose, we created a workflow called 3DGB to simplify the creation of 3D models from Hi-C data. Open source and freely available on GitHub, 3DGB generates contact maps, builds 3D models and adds further processing of the 3D model output as PDB files, suitable for advanced visualization with molecular viewer software. Using 3DGB, we created several models of the *S. cerevisiae*, *S. pombe* and *N. crassa* genomes, starting from Hi-C data available in public databases. Different strains were analyzed (wild type and mutants for heterochromatin organization or structural proteins), at different stages of the life cycle. Our models showcase known characteristics of the chromosomal organization of these genomes. But they also reveal, thanks to the visual integration of omics data, important properties of regulatory proteins with critical functions for the maintenance of spatial organization. We thus demonstrate the interest of 3D modeling of Hi-C contacts for studies of genome organization.



Supplementary Figure S1: Spatial organization of fungal chromosomes, as described in the literature. *3D models of Neurospora crassa (A), Saccharomyces cerevisiae (B) and Schizosaccharomyces pombe (C) genomes, from several previously published articles. Notably, these models are either drawings,*

meaning that they are interpretations of the data made by the authors, or real 3D objects (see black stars), meaning that they arise from the application of dedicated algorithms for the calculation of spatial coordinates, based on Hi-C contact measurements. In both situations, these models summarize current knowledge of the overall organization of genomes of these three species of fungi (see the main text for descriptions of interesting properties).

Results

Part 1. Simplifying 3D model creation for Hi-C data analysts with a robust and reproducible workflow

General overview of 3D Genome Builder

To simplify the creation of 3D models for researchers interested in Hi-C data analysis, we created 3D Genome Builder (3DGB), a bioinformatics workflow that streamlines the generation of 3D models to visualize and explore the spatial organization of chromosomes, based on Hi-C experimental results. A general overview of the workflow is presented in **Figure 1**. Starting from Hi-C raw FASTQ files, 3DGB automatically performs the critical bioinformatics steps required to *i)* compute Hi-C contact frequencies, *ii)* infer associated 3D models of the chromatin organization and *iii*) annotate and control the quality of the 3D models. These 3D models are stored in standard PDB files, so they can be further investigated with complementary visualization tools (see below). 3D models can optionally be enriched with supplementary quantitative omics data, such as ChIP-seq or RNA-seq signals (see next section). 3DGB has been designed to remain as simple as possible. Therefore, it requires only four basic inputs to be specified by the user (**Figure 1**, green boxes): FASTQ file identifiers, a FASTA file with the sequence of the reference genome, the list of restriction sites for the enzymes used during the Hi-C experiments, and the targeted resolutions for the Hi-C data analysis. This final parameter has an impact on the final 3D models, *i.e.* the smaller the value (specified in bp), the more detailed the 3D model.

The eight main steps required for Hi-C data processing are represented in blue boxes in **Figure 1** and rely on two state-of-the-art software packages. The first step utilized HiC-Pro (Servant et al. 2015), a reference in Hi-C data processing, cited more than 800 times. It processes raw FASTQ files, performs quality control and generates normalized contact counts and associated figures (presenting important statistics to evaluate the quality of read mapping and justify potential read filtering). Then, Pastis (Varoquaux et al. 2021) iteratively computes 3D models of the organization of chromosomes, through an original negative binomial contact count modelization

(referred to as Pastis-NB). Interestingly, it outputs a consensus model, which, by its uniqueness, greatly simplifies downstream analyses and interpretations. Around those two main components (HiC-Pro and Pastis), 3DGB centralizes the configuration, performs contact calculations, generates contact heatmaps, builds 3D models and adds further processing of the 3D model output as PDB files. 3DGB is open source, available in GitHub (https://github.com/data-fun/3d-genome-builder) and archived in Software Heritage (swh:1:dir:26b6504724952e6d0d7db34c394e052217523754).

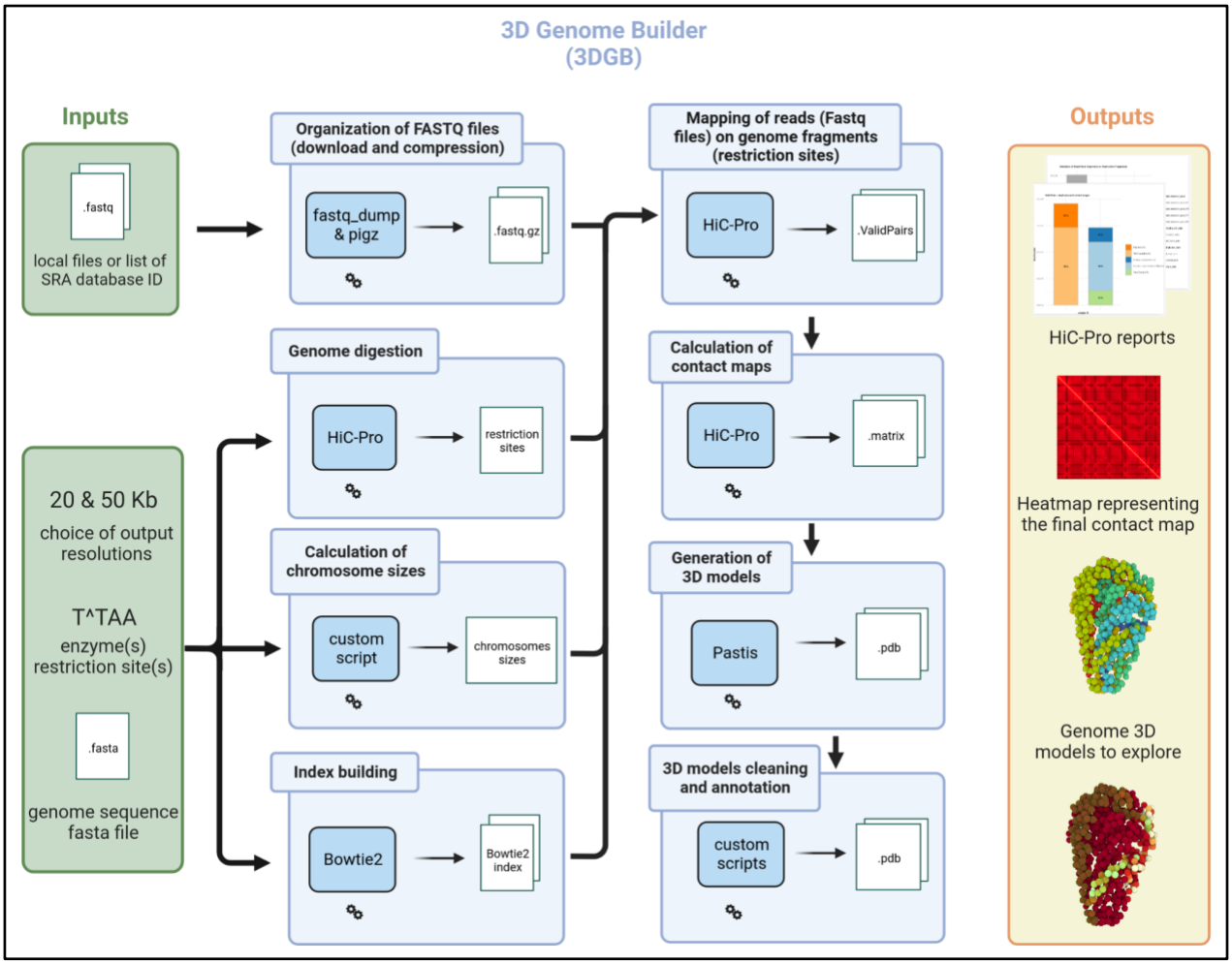
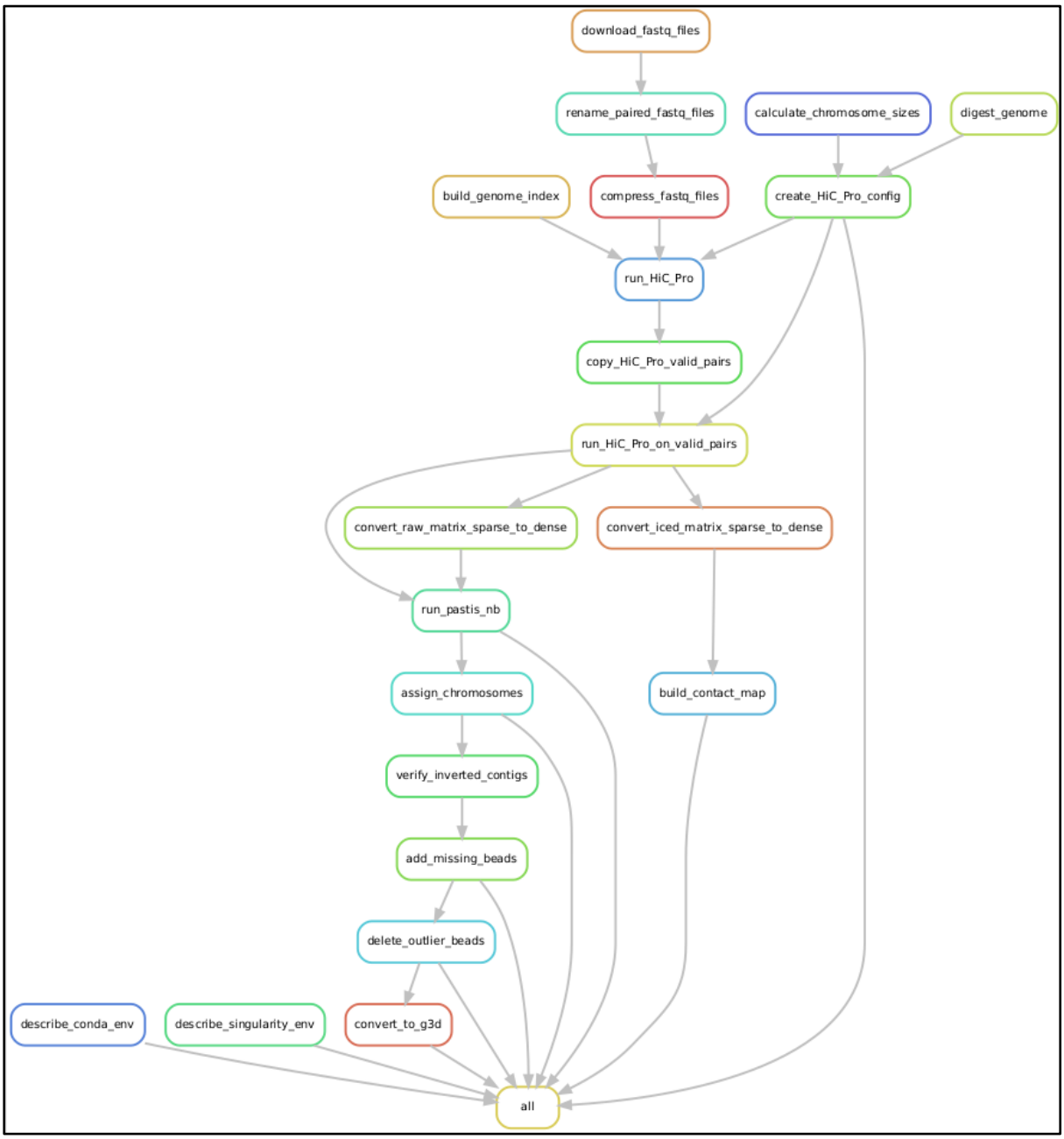


Figure 1: General overview of the 3D Genome Builder (3DGB) workflow. *Inputs and outputs are respectively represented with green and orange boxes (left and right panels). The workflow is detailed as a sequence of blue boxes (middle panel). Each blue box represents a task of particular interest. The 3DGB workflow is orchestrated with Snakemake (see* ***Supplementary Data S2*** *and* ***Methods****). All tasks are*

automatically chained one after the other until target outputs are produced. Final results comprise quality control reports generated by HiC-Pro, contact heatmaps and PDB files with the 3D models. PDB files are enriched with additional annotations (see main text for more details) and can be visualized with any PDB viewer software (mol in our case).*



Supplementary Figure S2: Directed acyclic graph of the 3DGB workflow as exported from Snakemake*. Each box refers to a Snakemake rule with defined inputs and outputs. Rules can be run in parallel by Snakemake. More information can be found in https://github.com/data-fun/3d-genome-builder.*

Enriched output PDB files for advanced visualization with molecular viewer software

The main 3DGB outputs are 3D models of genome organization. These 3D models are composed of beads for which 3D coordinates (x, y, z) were inferred from contact information (see **Methods**). One bead represents several thousand base pairs corresponding to the chosen resolution during the Hi-C data analysis (see previous section). To facilitate the study of these models, we enriched the structures produced by Pastis as PDB files with a four-step procedure. First, 3DGB formats and enriches the 3D model output by Pastis for visualization and data integration by annotating each bead with the chromosome number. This is useful to distinguish chromosomes when viewing complete structures (see **Figure 2** for illustrations). Second, it automatically reconstructs beads for which no coordinates could be calculated by Pastis, by interpolating missing coordinates from the existing ones (see **Methods**). Note that we do not extrapolate missing coordinates, meaning that beads with missing coordinates located at the extremities of the model are discarded. Third, outlier beads, *i.e.* beads placed outside the overall model, are filtered out and deleted based on a threshold value which can be specified by the user. Four, quantitative values, such as ChIP-seq data, can be used to color the model, allowing visual integration of omics data on the 3D structure of the genome (see next section for detailed examples of such integration of omics data). All these additional functionalities were implemented in Python (Guido Van Rossum and Fred L. Drake 2009) scripts integrated within the Snakemake workflow (see **Methods)**.

Altogether, these enhancements provide convenient 3D models to be viewed and explored by Hi-C data analysts. For chromatin 3D model file formats, we provided the 3D model structure in the PDB format, a traditional file format widely used to store coordinates of molecular structures. Most software used in structural biology to view and manipulate structures can handle the PDB format. In this paper, we preferentially used Mol* (Sehnal et al. 2021), a ubiquitous viewer for large scale molecular structures, with a user-friendly web interface allowing visualization and customization

of 3D models with only a few mouse clicks. We also used HiC3D-Viewer (Djekidel et al. 2017), a viewer especially developed for chromatin structures. Note that HiC3D-viewer requires conversion of the PDB file format into the G3D format. 3DGB also provides 3D models in the G3D file format, allowing users to choose the viewing software they prefer. As illustrations, images produced by Mol* and HiC3D-viewer are presented in **Figure 2**, for three different 3DGB models inferred in yeasts and a filamentous fungus (see next section).

Part 2. Seeing the spatial organization of fungal chromosomes with 3D Genome Builder (3DGB)

Creation of wild type models and visual consistency with the literature

To test the performances and the relevance of 3DGB, we chose three emblematic fungal species: *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe* and *Neurospora crassa*. We wanted to create 3D models of their genome organization and confront the models with current knowledge found in the literature. The *S. cerevisiae* Hi-C data are from a recent study from Constantino et al. (Costantino et al. 2020), the *S. pombe* Hi-C data are from Tanizawa et al. (Tanizawa et al. 2017) and the *N. crassa* Hi-C data are from Galaska et al. (Galazka et al. 2016). Detailed information regarding the source of the data are given in **Table 1**. As a result, we obtained new models of *S. cerevisiae*, *S. pombe* and *N. crassa* genomes. Three of them, which correspond to wild type situations, are shown in **Figure 2**. As expected, the key features of the spatial organization of the three species' genomes were observed, *i.e.* the Rabl conformation, the hypercoiling of chromatin fibers and the chromosome territories. Finding these well-described characteristics was an important step in validating the relevance of the 3DGB workflow, especially considering that 3DGB uses a probabilistic-based strategy to infer 3D models (Pastis-NB), which is free of initial constraints, such as the relative position of chromosomes.

Model	Fungal species (genome version)	Strain	Hi-C data sources	SRA IDs	Final number of valid read pairs	Number of beads in the final model	Figure (model resolution in bp)
1	*N. crassa* (nc14)	WT	Rodriguez et al. 2022	SRR14362684 SRR14362685 SRR16761088 SRR16761089 SRR16761090 SRR16761091 SRR16761092	107,636,304	3,976	2 (10,000)
2		WT	Galazka et al. 2016	SRR2105869 SRR2105870 SRR2105871 SRR2105872	5,030,121	800	5 (50,000)
3		hpo		SRR2105876 SRR2105877 SRR2105878	2,922,526	808	
4	*S. pombe* (ASM294v 2.19)	WT	Tanizawa et al. 2017	SRR5149251 SRR5149252	15,511,107	1,223	4 (10,000)
5				SRR5149253 SRR5149254	14,377,600	1,211	
6				SRR5149255 SRR5149256	14,394,280	1,204	
7				SRR5149257 SRR5149258	15,497,423	1,200	
8				SRR5149259 SRR5149260 SRR5942526	19,035,344	1,213	

9				SRR5149261	13,841,928	1,224	
				SRR5149263			
10				SRR5149264	21,936,238	1,201	
				SRR5149265			
				SRR5942527			
11				SRR5149266	21,727,635	1,189	
				SRR5149267			
				SRR5942528			
12	*S. cerevisiae (R64, S288C)*	WT	Constantino et al. 2020	SRR11893084	16,847,912	2,332	2 and 3 (5,000)
				SRR11893085			
13		Mcd1 depleted		SRR11893086	14,180,445	2,323	
				SRR11893087			

Table 1: Main characteristics of raw data used in this study to create 3D models of fungal chromosomes. *The sources of Hi-C data (original articles and SRA identifiers) are given. All datasets were obtained on wild type strains, with the exception of models #3 and #13 which correspond to the N. crassa hpo mutant (used in* ***Figure 5****) and S. cerevisiae mcd1 depleted (used in* ***Figure 3****). The number of valid pairs of reads gives an estimation of the overall quality of the data and indicates the density of the contact frequencies measured in the experiment. The number of beads in the 3D model is determined by the chosen resolution during Hi-C data analysis and the length of the reference genome.*

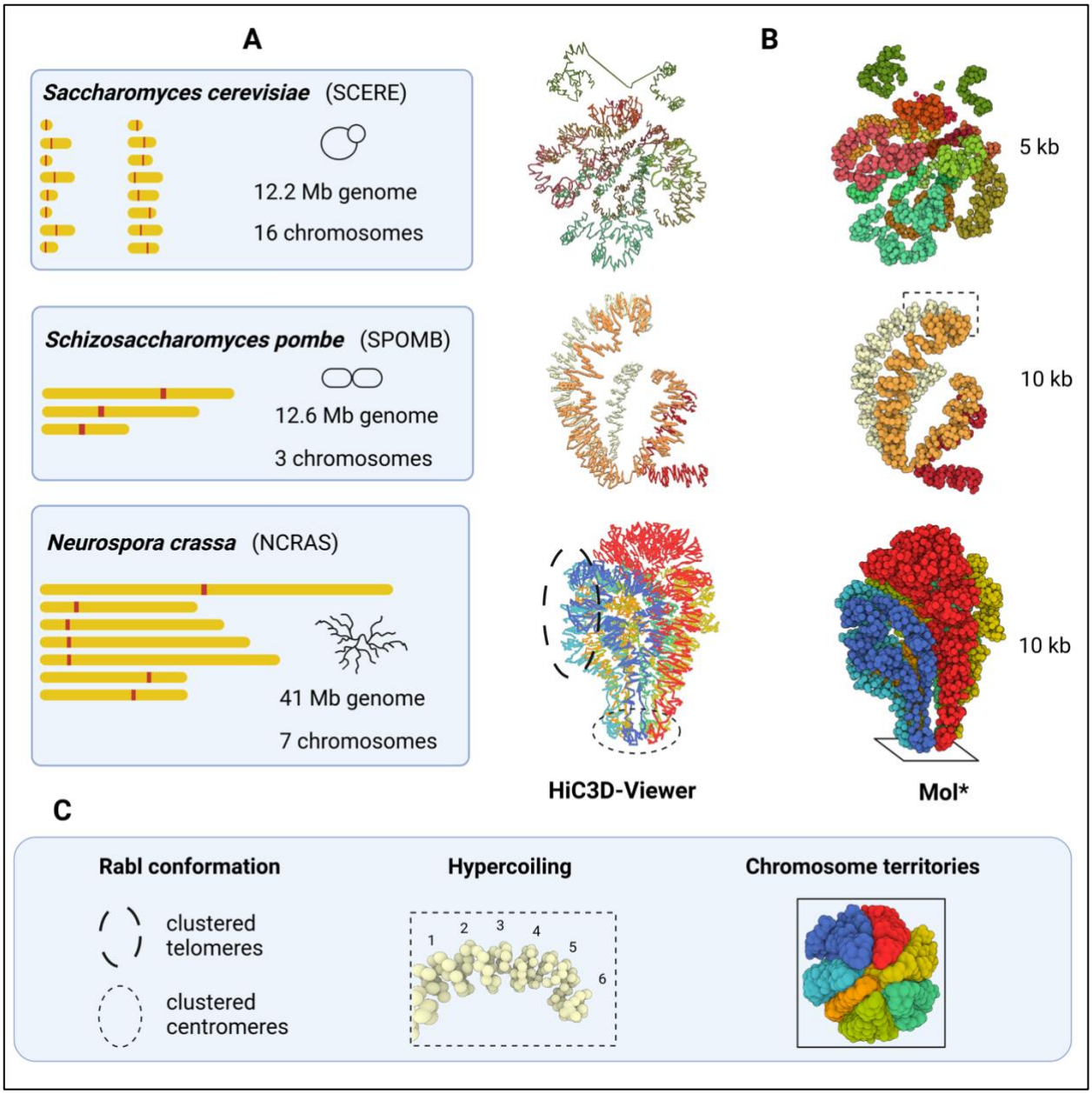


Figure 2: 3D modeling of fungal chromosomes in wild type situations, in the model yeasts S. cerevisiae and S. pombe and the filamentous fungus N. crassa. ***(A)*** *2D representations of chromosomes in each species. They are displayed with the same scale, showing the diversity of genomes used in this study to create 3D models.* ***(B)*** *3D models obtained with the 3DGB workflow. The source of the Hi-C data used to create them are presented in* ***Table 1****. They are visualized with HiC3D-Viewer (left) and Mol* (right). Note that HiC3D-Viewer links beads artificially, whereas with Mol*, beads remain individualized (they each represent a 5 or 10 kb chromatin region, depending on the species, see* ***Table 1****).* ***(C)*** *Key features of genome organization are highlighted. The Rabl configuration, the hypercoiling of*

chromosomes and their distribution into separate territories are well observed on the 3D models obtained with 3DGB and presented in (B) (see circles and rectangles).

Added value of 3DGB models for visualization: examples ranging from a detailed view to a more general view of chromatin organization

Example #1: Spatial alignment of cohesin binding sites along chromosomes in *S. cerevisiae*

Ten years after the first 3D model of the *S. cerevisiae* genome (see (Duan et al. 2010) and **Supplementary Figure S1**), Costantino et al. (Costantino et al. 2020) produced new Hi-C and ChIP-seq data, in order to better understand the relationship between the identified patterns of chromatin domains (information derived from Hi-C data analysis) and the cohesin residency regions (information derived from ChIP-seq data analysis). They used cells arrested in mitosis, and studied both the wild type strain and mutants altered for cohesin or its regulators. Notably, to finely describe chromatin loop formation, they used a recent improvement of the Hi-C technique, called Micro-C XL (Hsieh et al. 2016). This method has the advantage of greatly improving the detection of short-range interactions (at the scale of nucleosomes), while still allowing the detection of whole genome chromatin interactions. This is of particular interest for a species like *S. cerevisiae*, in which chromosomes are very small compared to the other species (*S. cerevisiae* chromosomes range from 0.23 to 1.53 Mb only, see **Figure 2A**). As a result, they showed that in *S. cerevisiae* mitotic cells, high residency cohesin anchor regions (named “CARs” and detected with ChIP-seq) correspond to the genome-wide boundaries of chromatin loops (named “CAR domains” and detected with Micro-C XL). This was an important observation which *i*) supports the idea that the yeast genome is organized into recurrent defined chromatin structures delimited by cohesin and *ii*) that this spatial organization of chromosomes can impact genome functions, as in mammals. Still, in their original article, the authors only present Hi-C contact heatmaps to support their interpretations.

To go further in this and enrich visualization, we collected their Micro-C XL data and reanalyzed them with 3DGB, to create an updated 3D model of the spatial organization of the *S. cerevisiae* chromosomes (see **Methods**). Considering the very high precision of the data, we expected to

observe both the general configuration (chromosome territories and Rabl conformation) and the fine chromatin domains (coiling), by zooming in and out with visualization software. Our results are presented in **Figure 3**. Note that the overall model (**Figure 3A**) is the same as the one shown in **Figure 2B**, but with different (color-blind friendly) colors used to identify the 16 chromosomes of the *S. cerevisiae* genome. ChIP-seq data were also reanalyzed (see **Methods**) to locate the CARs on the new 3D model. They are shown in yellow in **Figure 3**. As expected from Hi-C contact maps presented in the original paper of Costantino et al. (Costantino et al. 2020), we observed a continuous distribution of CARs along the *S. cerevisiae* chromosomes (**Figure 3A**). To complete this observation, we zoomed in on the chromosomes individually. An example of chromosome 13 is shown **Figure 3B**. Notably, we could observe that CARs, which are still represented by yellow beads on the 3D models (**Figure 3D**), were aligned in space, and this, independently of the length of the chromatin loops whose boundaries they represent (**Figure 3C**). This observation was true for all chromosomes (another example can be found in **Supplementary Figure S3**). 3D modeling of Hi-C contacts is therefore of significant interest here, revealing the existence of a cohesin skeleton, ensuring structural stability of the spatial organization of the *S. cerevisiae* chromosomes.

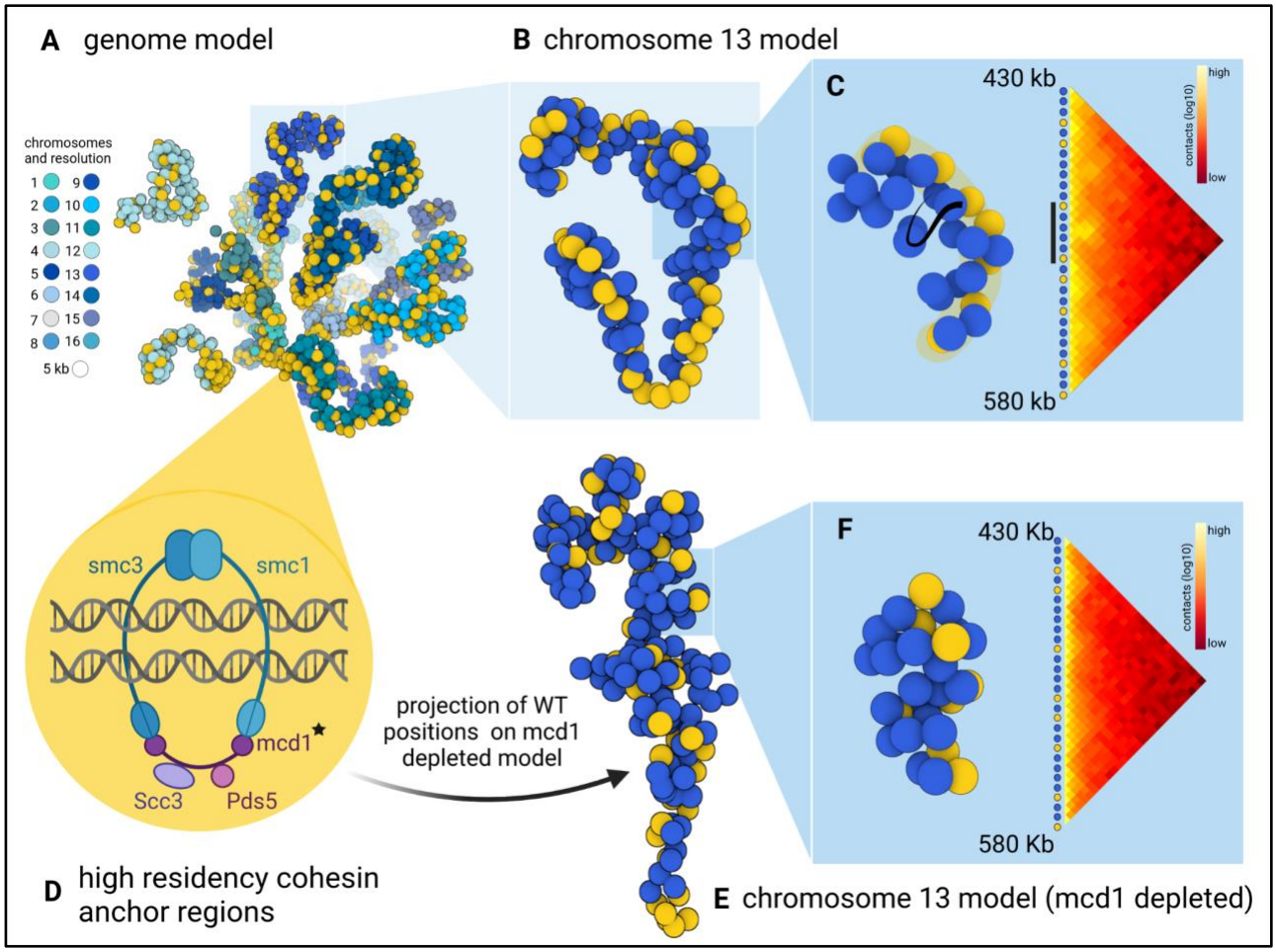
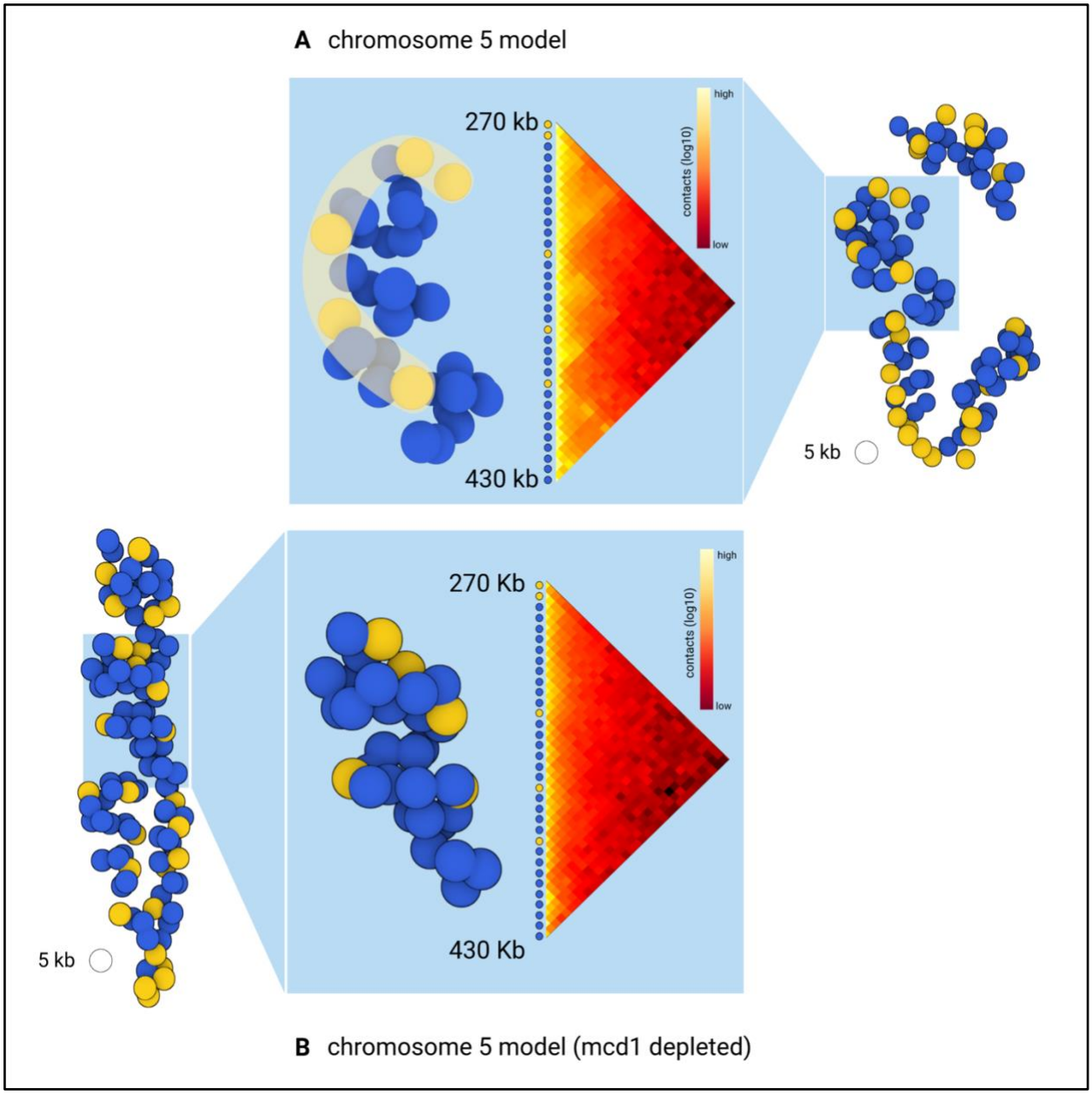


Figure 3: Revealing the 3D skeleton formed by cohesin in S. cerevisiae chromatin. (A) *3D model of S. cerevisiae chromatin, obtained with 3DGB, using the Micro-C XL data from Costantino et al.* (Costantino et al. 2020)*. Each bead represents a 5 kb chromosomal region. Yellow color indicates genomic regions with a high value of ChIP-seq signal intensity of the Mcd1p cohesin subunit, in mitotically arrested cells.* ***(B)*** *Isolation of chromosome 13 from the overall structure, showing the spatial alignment of the DNA binding sites of cohesin. Cohesin structures a chromosomal skeleton on which chromatin loops can form.* ***(C)*** *Juxtaposition of the 3D model and the associated contact map, for the beads which are located between 430 and 580 kb. On the 3D structure, the cohesin backbone is highlighted in yellow. Viewing this spatial alignment of cohesin is complementary to the information contained in the contact map, i.e. the frequency of contacts is very high inside genomic regions delimited by yellow beads.* ***(D)*** *Schematic representation of the cohesin complex in S. cerevisiae. The protein Mcd1p, depleted in cells used to create the model shown in (E) and (F), is indicated with a black star. Genomic regions, which are attached by cohesin are referred to as CARs, i.e. high residency cohesin anchor regions, and are visualized in yellow on the 3D models.* ***(E)***

Isolation of chromosome 13 from the overall structure, obtained with the mcd1 mutant (see the main text). The yellow beads correspond to the cohesive backbone as it was originally identified in the wild type (B). ***(F)*** *Juxtaposition of the 3D model and the associated contact map, for the beads which are located between 430 and 580 kb. This is the same genomic region exposed in (C).*



Supplementary Figure S3: 3D skeleton formed by cohesin in S. cerevisiae chromatin, example of chromosome 5. (A) *Isolation of chromosome 5 from the overall structure obtained with the wild type strain and* ***(B)*** *isolation of chromosome 5 from the overall structure obtained with the strain depleted for the Mcd1 cohesin subunit. Yellow beads correspond to the backbone of CARs (see the main text), as defined in the wild type situation.*

To verify the relevance of this observation, we also reanalyzed the Hi-C data obtained with an *S. cerevisiae* strain in which the gene encoding the subunit Mcd1 of the cohesin complex was depleted (**Figure 3D,** black star). Results are presented in **Figure 3E**. This time, the 3D models showed more irregular coiling of chromosomes and globule-like structures. Again, this observation is relevant to the original article of Costantino et al. in which the authors found that in the absence of cohesin, the mitotic chromatin is not entirely disorganized, but rather structured into globular domains, dependent on histone modifications (Costantino et al. 2020). An important question then was what happens, in this context, to the cohesin skeleton previously observed in the wild type strain. We therefore mapped the genomic positions of CARs, as they were defined based on the ChIP-seq data in the wild type strain, on the 3D "mutant" model (**Figure 3E**, yellow beads) and observed a complete destabilization of the cohesin backbone initially detected in wild type.

In summary, we illustrate here the interest of placing known 2D patterns of chromatin loops in a 3D context. Our observations reinforce the idea that a cohesin backbone stabilizes the spatial organization of chromosomes in *S. cerevisiae*. 3DGB brings a new perspective to the visualization compared to Hi-C contact maps: the flat loops are arranged in a coil supported by a backbone of aligned CAR regions.

Example #2: Oscillations in coiling of chromosome arms during the cell cycle in *S. pombe*

In the previous section, we were able to observe details of *S. cerevisiae* chromosome organization by zooming into the 3D model we built with 3DGB. This model was associated with cells arrested in mitosis and therefore represents a static view of the chromosomal organization at a very particular point in the yeast cell cycle. The goal of our second example was to evaluate the possibility of seeing, with 3D models, changes in the organization of chromosomes during the cell cycle. Indeed, chromosomes are subject to major structural constraints and undergo rearrangements at the different stages of the cell cycle (G2, M, G1 and S). We chose to explore this question using the data of Tanizawa et al., published in 2017 (Tanizawa et al. 2017). The

authors applied an *in situ* Hi-C protocol to follow the organization of the fission yeast *S. pombe* genome throughout the cell cycle. The authors observed that during mitosis, chromosomes are structured into large (300 kb to 1 Mb) and small (30 - 40 kb) domains, which are respectively structured by condensin and cohesin protein complexes (Tanizawa et al. 2017). Based exclusively on their interpretation of Hi-C contact maps, they showed that if the mitotic organization into large domains gradually dissolves across the cell cycle, small domains remain relatively stable. They also hypothesized that the Rabl conformation was stable in interphase but disrupted during mitosis. With this dataset, we assessed our ability to observe these structural features in the 3D models obtained with 3DGB, at different stages of the *S. pombe* cell cycle. We collected Hi-C data from the original article (**Table 1**) and produced 3D models (see **Methods**), showing the organization of chromosomes at different time points of the S. pombe life cycle, after an initial synchronization of cells in phase G2. Our results are presented in **Figure 4**.

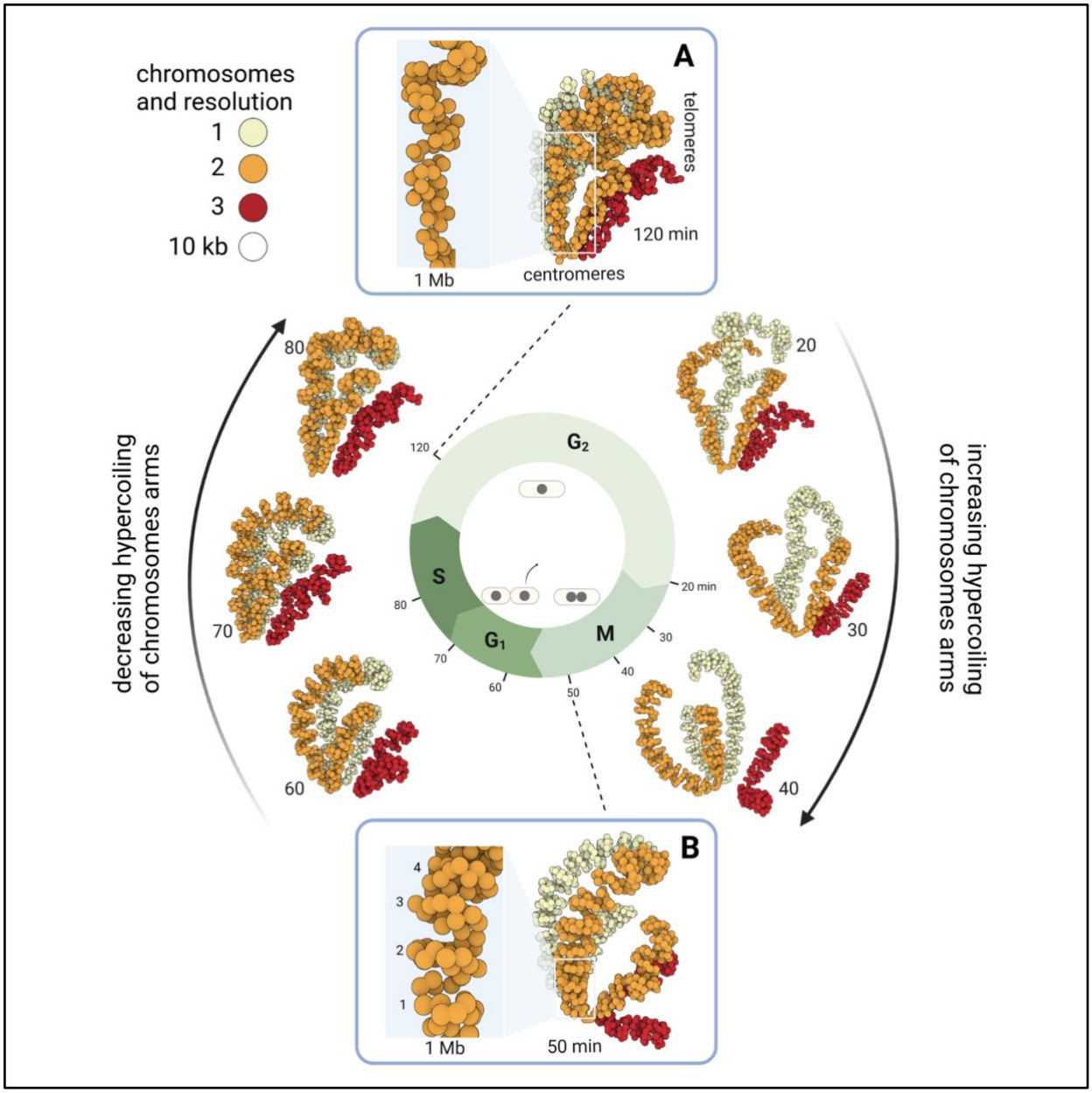


Figure 4: Dynamics of the spatial organization of* S. pombe *chromosomes during the cell cycle. *3D models obtained with 3DGB using Hi-C data from Tanizawa et al.* (Tanizawa et al. 2017)*. Each bead represents a 10 kb chromosomal region and is colored according to the chromosome to which they belong (from 1 to 3). Eight models are shown here. They correspond to different organization of* S. pombe *chromosomes during the fission yeast cell cycle. After the synchronization of cells in phase G2, the following time points were analyzed: 20, 30, 40, 50, 60, 70, 80 and 120 minutes. Their correspondence with the cell cycle phases is shown in the middle. Two time points are thus in the G2 phase (20 and 120 minutes), three time points are in the M phase (30, 40 and 50 minutes), two time points are in the G1 phase (60 and 70 minutes), and finally one time point is in the S phase. Centromere and telomere regions are*

annotated. Centromeres of all chromosomes are constantly clustered, whatever the cell cycle time point. This is consistent with previous observations (see the main text). Coiling of chromosome arms, as described in the literature (Mizuguchi et al. 2014; Tanizawa et al. 2017)*, is visible on the 3D models. Zooms are provided in (A) and (B) boxes, which correspond to time points (respectively 120 and 50 minutes) for which the intensity of coiling is minimal and maximal (see the main text for further explanation).*

As expected, the Rabl conformation is visible, particularly during interphase (G1, S and G2 models, between 60 and 120 minutes). However, it becomes somewhat disorganized at the transition between G2 and Mitosis (**Figure 4**, 20 min model), even though the cluster of centromeres remains stable during the entire cell cycle. Telomeres of chromosomes 1 (yellow color) and 2 (orange color) also remain associated, whereas chromosome 3 (red color), in which ribosomal DNA repeats are found, occupied an external position consistent with compartmentalization of rDNA in the nucleolus. We also observed that the level of chromatin compaction varies over the cell cycle, with an observed minimum in G2 (**Figure 4A**) and an observed maximum in late mitosis (**Figure 4B**). Indeed, the 120 minute model shows uncoiled chromosome arms, with loose globule-like folding (**Figure 4A**) whereas the 50 minute model chromosome arms are structured into regular coils of ~250 kb (**Figure 4B**). This more compact (see the zoom-in 1 Mb region **Figure 4A-B**) and structured (regular coiling) organization can be seen gradually appearing on the 20, 30 and 40 minute models (**Figure 4**, right arrow). Finally, it is important to keep in mind that the Hi-C data of each model were obtained independently, reinforcing the biological relevance of the observed similarities in relative position of chromosomes and level of compaction. Therefore, the highly organized mitotic structure gradually fades though the G1 and S phases to return to the Rabl conformation (**Figure 4**, left arrow).

In summary, the 3D models obtained with 3DGB consistently illustrate the oscillations in coiling of chromosome arms during the cell cycle in *S. pombe*. This strengthens the original study by Tanizawa et al. and opens new perspectives for further analysis, for example omics data integration.

Example #3: Massive relocalisation of histone marks in *N. crassa* heterochromatin regulator mutants

In the *N. crassa* genome, heterochromatic regions are a major component of the chromosome conformation (Galazka et al. 2016). Constitutive and facultative heterochromatin are respectively genomic regions that contain few genes with little transcription (constitutive heterochromatin) and genomic regions that contain genes with regulated gene repression (facultative heterochromatin). At the molecular level, constitutive and facultative heterochromatin can be distinguished by the presence of H3K9me3 or H3K27me2/3 histone marks (Galazka et al. 2016). Interactions between constitutive and facultative heterochromatin are complex and an important subject of discussion in the literature. In that respect, an emblematic observation is that the reduction of H3K9me3 in constitutive heterochromatin, causes the redistribution of H3K27me2/3. In particular, in a genetic context in which the heterochromatin protein 1 (Hp1, which recognize H3K9me3) is lost, H3K27me3 is depleted from facultative heterochromatin and H3K27me2 is gained at constitutive heterochromatin (Jamieson et al. 2016). Our objective was to assess the ability to see this phenomenon at the scale of the complete genome. To do so, we started by creating 3D models with 3DGB, from the wild type (WT) and HP1-deficient (*hpo*) strains. These models were then used for visual integration of ChIP-seq data, referring to the genomic location of histones with H3K9me3 and H3K27me2/3 post translational modifications (see **Methods**). Results are shown **Figure 5A**.

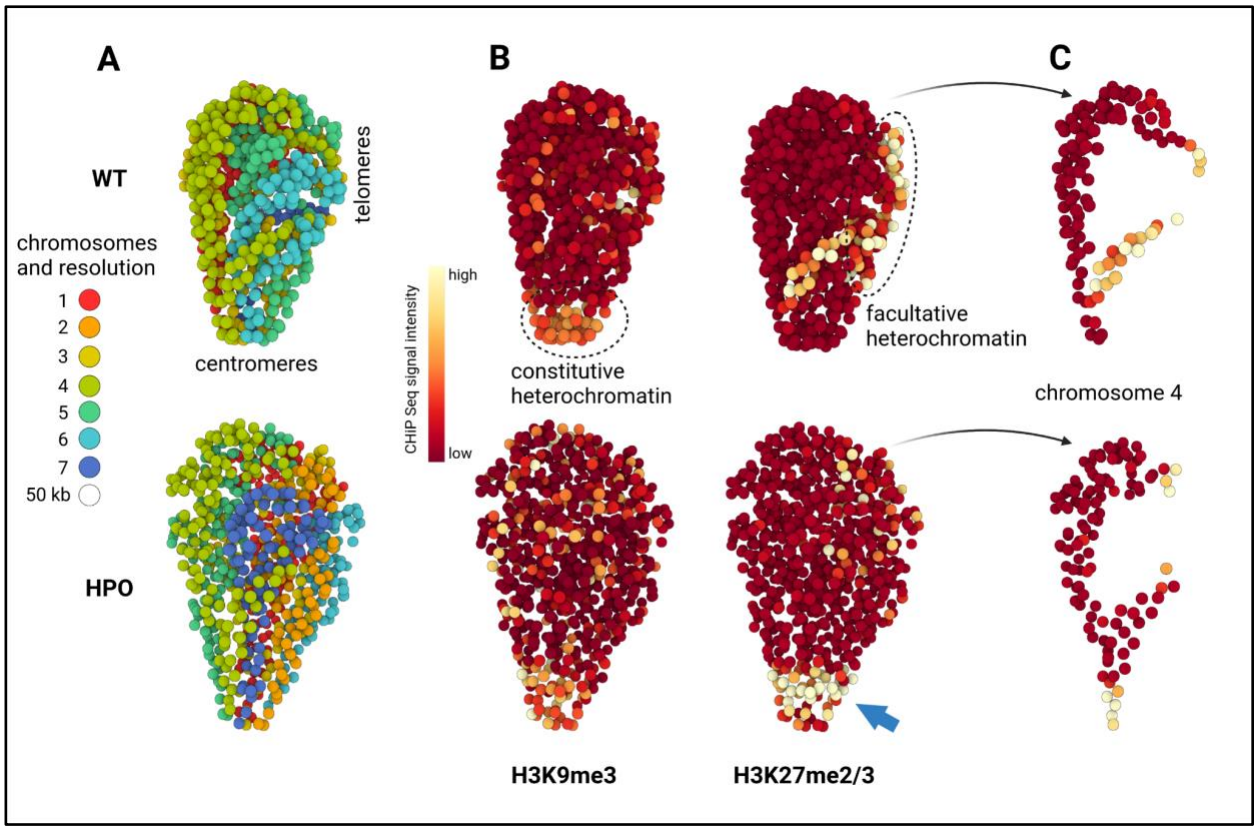


Figure 5: Visual integration of ChiP-seq data using 3D models of the* N. crassa *genome. (A) *3D models obtained with 3DGB using Hi-C data from Galaska et al.* (Galazka et al. 2016)*, respectively in the wild type (WT) and HP1-deficient (hpo) strains (see* ***Table 1*** *and* ***Methods*** *section for technical details). Each bead represents a 50 kb chromosomal region and is colored according to the chromosome to which it belongs (from 1 to 7). The colocalization of centromeres and telomeres is observed in both models (centromeres on the bottom and telomeres on the right).* ***(B)*** *Same models as shown in (A), using the ChiP-seq signal intensity as a color code for beads (see* ***Methods****). ChIP-seq data are from Basenko et al.* (Basenko et al. 2015) *and refer to the genomic location of histones with H3K9me3 and H3K27me2/3 post translational modifications, respectively. In the WT model, the intensity of the ChIP-seq signal for H3K9me3 is particularly high (yellow color) around centromeres (compared to the rest of the genome, dark red color), whereas the intensity of the ChIP-seq signal for H3K27me3 is particularly high at telomeres. These observations are relevant to known functions of constitutive and facultative heterochromatin (see the main text). In the hpo strain, important changes in the intensities of ChIP-seq signals are observed with, notably, a massive relocalization of H3K27me3 high intensity signal to the centromeres (blue arrow).* ***(C)*** *Isolation*

of chromosome 4, better highlighting the changes in H3K27me2/3 ChIP-seq signal intensity between WT and hpo strains.

As expected from the literature, we observed a slight difference between the two model organizations, essentially the intensity of chromatin compaction and the relative position of chromosomes 6 and 7 (**Figure 5A**). To extend this observation, we calculated for each 50 kb region represented by a bead in the 3D models, the intensity of ChIP-seq signals arising from experiments targeting H3K9me3 and H3K27me2/3 post-translational histone modifications (see **Methods**). Results are shown **Figure 5B** (whole genome) and **Figure 5C** (chromosome 4). In WT, we could observe as expected an accumulation of H3K9me3 histones in centromeres, relevant to constitutive heterochromatin, and an accumulation of H3K27me2/3 histones in sub-telomeric regions, relevant to facultative heterochromatin. In *hpo*, ChIP-seq signals appeared to be quite disorganized, especially for H3K27me2/3 marks, and massively relocalized to centromeres (blue arrow, **Figure 5B**). This observation was expected according to the literature, but for the first time it can be seen clearly, in a simple and integrated way.

This example illustrates the interest of 3D models in the context of multi-omics data integration. If Hi-C data allows us to better understand the organization of chromosomes, other omics data (such as ChIP-seq here) bring insights on the mechanisms of genome function. It is important to keep in mind that the images presented abstract a very large quantity of data, *i.e.* several million measurements (frequencies of contacts, probability of interactions, quality values, etc.). These 3D models can confirm or invalidate hypotheses on the overall functioning of genomes, and possibly lead to the formulation of new hypotheses.

Additional benefits of using 3DGB

Detection of inconsistencies between the spatial organization of chromosomes derived from Hi-C experiments and the genomic sequence used as reference

In the context of our analyses of Hi-C data in *N. crassa*, we faced an unexpected situation with respect to the reference genome sequence for this organism. Indeed, the *N. crassa* genome assembly, originally published in 2003 (Galagan et al. 2003), is composed of 7 chromosomes and 13 supercontigs, all available in GenBank under the accession “assembly nc12”. In 2016, Galazka et al. (Galazka et al. 2016) identified an inversion of the contig named “12.304”, corresponding to a large region in chromosome 6. In 2022, Rodriguez et al. (Rodriguez et al. 2022) further improved the original assembly “nc12” based on Hi-C data analysis. The updated assembly that integrates the improvements of Galazka et al. and Rodriguez et al. is available in the GEO database under the accession “assembly nc14”. When we started using 3DGB to create a 3D model of the wild-type *N. crassa* chromosome (**Figure 2**), we used as reference genome the “nc12” assembly, available from the GenBank database. Surprisingly, we observed in our inferred 3D model, an inconsistency in the order of the genomic regions, with respect to the succession of our beads (**Figure 6**). Our 3D model thus directly highlighted the inverted contig 12.304 on the chromosome 6, originally found by Galazka et al. At 10 kb resolution, the 95 beads used to represent this region were placed in the 3D space according to the information of Hi-C contacts only, independently of the reference genome assembly. This explains why we were able to highlight a mismatch between the chaining of beads in the 3D models and the numbering of these beads according to the genome sequence (**Figure 6A**). Note that when using the nc14 assembly of the genome as reference, the obtained 3D model of chromosome 6 we obtained was coherent both spatially (order of the beads) and sequentially (order of the genomics regions). From this experience, we developed an automated procedure (integrated to 3GDB as an option) to detect and correct inverted contigs by comparing the chaining of beads in the 3D model (based on the distance between adjacent beads, **Figure 6B**) and the numbering from the genomic

sequence written in the FASTA file. This method also produces a corrected version of the 3D model and of the genome assembly (**Figure 6C**).

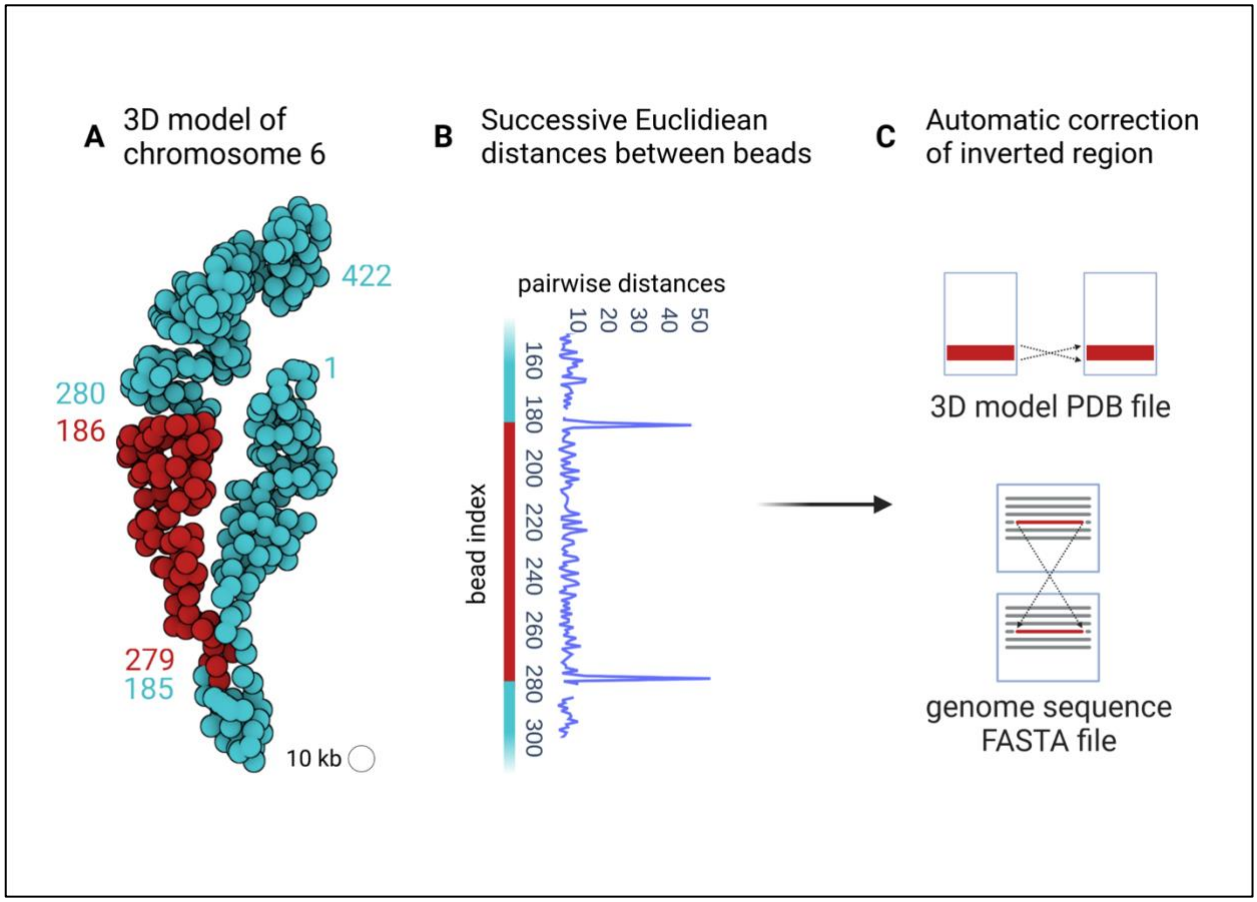
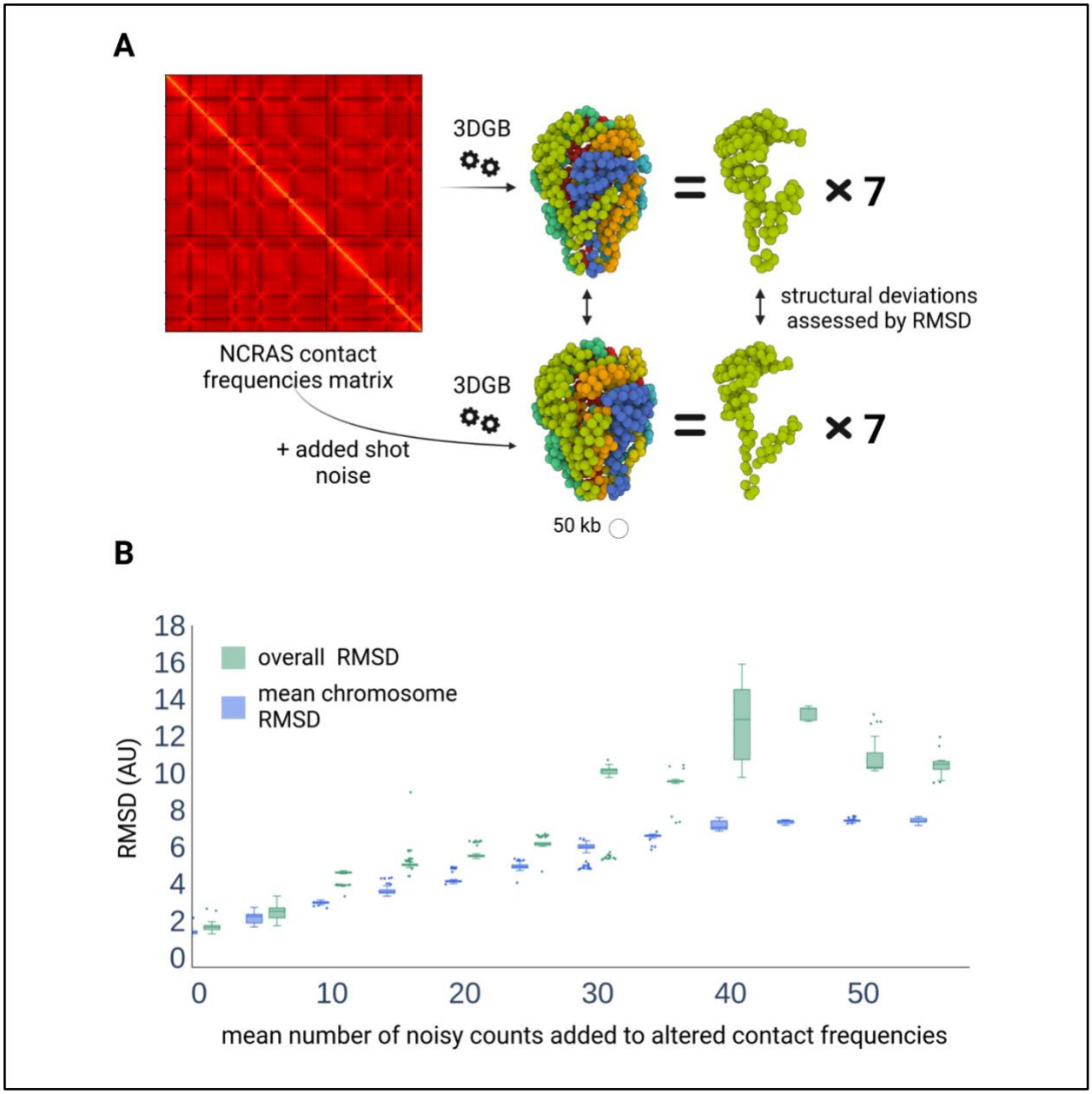


Figure 6: Inconsistency between the spatial organization of chromosome 6 in N. crassa and the reference genome assembly "nc12". ***(A)*** *3D model of chromosome 6 obtained with 3DGB using Hi-C data from Galaska et al. and the "nc12" assembly of the* N. crassa *genomic sequence as reference. Each bead represents a 10 kb chromosomal region. The numbers label the beads, according to the genomic region (DNA sequence) with which they are associated. Therefore, the bead number 1 corresponds to the interval]0, 10] kb of the linear genomic sequence, the bead number 2 corresponds to the interval]10, 20] kb, the bead number 3 corresponds to the interval]20, 30] kb, etc. Unexpected transitions between the pairs of beads 185 - 279 and 186 - 280 are shown with different colors (blue and red).* ***(B)*** *Euclidean distances between 3D beads with successive numbers as defined with genomic intervals on the DNA sequence. Unexpected high values are observed between bead pairs 185 - 186 and 279 - 280. They are in line with the structure shown in (A).* ***(C)*** *Based on the inconsistency revealed in the Euclidean distances, it is possible in 3DGB to automatically correct the genomic sequence used as a reference.*

Quantification of the stability of 3D models to random noise with multiple uses of 3DGB

The advantage of using a workflow like 3DGB is the possibility of automating tasks. In this part, we assess the stability of the 3D organization of the chromosomes with respect to potential inaccuracies in the contact frequencies measured with Hi-C (see **Methods**). Our strategy was to alter the original values of contact frequencies (issued from the Hi-C data) by adding a “shot noise” on the contact frequencies, and then run 3DGB to create an associated 3D model. If the intensity of the shot noise was low, the output model was expected to be close to the original model (obtained from the original data). The resulting RMSD score, calculated by comparing the 3D structure inferred from the reference *N. crassa* contact map and the 3D structure inferred from the altered *N. crassa* map, was also expected to be low. Altogether, we assessed 23 levels of intensity of the shot noise (see **Methods**) and generated, automatically, a total of 1150 models. Our results are summarized in **Supplementary Figure S5**. As expected, we observed an increase in the RMSD scores, when the intensity of the shot noise increases. This was more striking for the global structure (**Supplementary Figure 5B**, green boxplots) than for individual chromosomes (**Supplementary Figure 5B**, blue boxplots). This observation underlines the greater sensitivity of the organization of the chromosomal territories compared to the internal organization of the individual chromosomes.



Supplementary Figure S5: Quantification of the stability of 3D models to random noise added to original contact frequencies. (A) *Each pairwise frequency of contacts is randomly modified by adding a count value drawn from a Poisson distribution with a specified mean parameter (see* ***Methods****). 3D models are created with 3DGB (at 50 kb resolution) and compared by calculating RMSD scores.* ***(B)*** *Values of RMSD scores (arbitrary unit) obtained for different values of noise, i.e. different values for the mean parameter used in the Poisson distribution. For a given noise value, 50 models are generated, the RMSD score is either calculated on the overall structure composed of all chromosomes (green boxes) or averaged on the RMSD scores obtained for the 7 individual chromosomes (blue boxes).*

Discussion

The objective of this work was to explore the interest of creating 3D visualizations of chromosome organization to complement the classic contact maps producing during Hi-C data analysis. We found that, although methods to create 3D models exist, they are not often exploited, and this is particularly true of fungal genome studies. The emblematic 3D representation of the complete genome of *S. cerevisiae* had been obtained more than 10 years ago, but only partial structures were generated for *S. pombe* and *N. crassa* (see **Supplementary Figure S1**). Our first intention was therefore to create, from existing bioinformatics tools and Hi-C datasets, complete models for these three species. Unlike what has been done for *S. cerevisiae* by Duan et al. (Duan et al. 2010), we wanted to build models without specifying initial physical constraints regarding chromosome positioning. The Pastis-NB method was particularly suitable for this task. To automate the generation of 3D models, we built 3DGB, a high-throughput Hi-C data processing workflow (**Figure 1**). With this workflow, we provided enriched PDB files for advanced visualization with molecular viewer software (**Figure 2**) and we created several hundred models in reasonable computing times (**Supplementary Figure S5**). The 3DGB workflow is available to other scientists who would like to add 3D models to their analyses of Hi-C data.

The relevance of 3D models of the spatial organization of chromosomes is an important question of our study. Beyond the production of a "beautiful image", what is the added value of these models for Hi-C data analysts? It is indeed important to keep in mind that these models are not "pictures" of the interior of cell nuclei. They are based on experimental data, whose quality may vary and significantly impact the inferred model. Ultimately, they represent measurements of contact frequencies observed at the scale of cell populations and therefore, they are simplified representations of reality. This work on fungal genomes taught us that the visualization of 3D models is profitable for several reasons. First, it gives an overall representation of the organization of the chromosomes (**Figure 2**). While contact maps are very useful for identifying local chromatin structures (at the scale of kilo base pairs), 3D models allow a complete zoom out, at the scale of

the entire chromosome, giving a global representation of all contact frequencies. Second, 3D models can both serve as the final visualization of the results of a Hi-C study, but also as a starting point for further data integration. The results we presented for the *S. cerevisiae* and *S. pombe* genomes are examples of final visualization results. On the one hand, our models revealed important features of the organization of chromosomes, respectively a backbone of cohesin (**Figure 3**) and hypercoiling dynamics (**Figure 4**), which are difficult to fully understand from the contact heatmaps alone. On the other hand, the results we presented for *N. crassa* are examples of original data integration (**Figure 5**). This time, 3D models are starting points for additional exploration. The coloring of the 3D representation of the models, based on quantitative measurements from "omics" experiments, albeit a simple idea, is particularly effective for studying global phenomena, such as the massive relocalization of histone marks in a mutant, as shown here for *N. crassa*.

Still, it is worth considering that contact-based modeling of the genome 3D organization has some limits. They remain a 3D representation of a Hi-C based contact network, with chromatin regions modeled as independent beads, each representing several kilo base pairs. The sequencing depth of the Hi-C data and the chosen resolution for their analysis have a direct impact on the accuracy of the final model: for a given number of reads, the lower the resolution, the fewer the inter chromosomal contacts are detected. We observed that when the sequencing depth is low, the chromosomes in the 3D models are more distant from each other, for example in the mitotic *S. cerevisiae* and *S. pombe* models (**Figure 2**). In such a situation, it is difficult to discriminate between the biological compaction of the chromatin and the limitations due to the Hi-C experimental technique, when only looking at the final volume occupied by a model. However, structural features like level of coiling are still visible (**Figures 3** and **4**) and can provide insight at the level of the biological compaction of chromatin, compensating for technical limitations (resulting in chromosomes which are more spaced). As stated previously, another inherent limitation of 3D models is that they are “static”, “population averaged” pictures of a biological system (chromatin organization) known to be highly dynamic and variable from cell to cell. These

artifacts will exist as long as the Hi-C data are snapshots obtained from cell populations. Still, it is important to note that Rabl conformations and chromosome territories were observed in all species (**Figure 2**), even with the *N. crassa* model, which is an average representation derived from Hi-C experiments performed on an asynchronous cell population. Therefore, with synchronous cells, the situation can be expected to be even better. This is what we observed with *S. pombe* models. Because the cells were initially synchronized, the Hi-C datasets produced at different stages of the cell cycle revealed several other interesting genome organizations in the inferred 3D models (**Figure 4**) and we thus managed to render the dynamics of chromatin. As for the "population average" limit, single cell Hi-C strategies are developing (Stevens et al. 2017), opening the promising perspective of creating 3D genome models of single cells. A final limit of current 3D models is the missing regions of the fungal genomes. As illustrations, the models presented here are built using genomic sequences in which the rDNA repeats were deleted, thus preventing the correct representation of structural features in the nucleolus (explaining the hole we can observe on chromosome 12 in *S. cerevisiae* models, **Figures 2** and **3**). An additional important simplification arises from the joint representation of sister chromatids and the complexity of working with diploid genomes (in classical Hi-C data analyses, the contact measurements cannot be distinguished between identical sequences). Strategies are emerging to solve haplotypes (Oomen et al. 2020; Mitter et al. 2020) and the software used in 3DGB has tuned options for this: at the mapping step, HiC-Pro can build allele-specific contact maps if SNP information is provided. At the 3D modeling stage, the Poisson-based method of Pastis has been extended to haplotype resolution. Even if this allows us to imagine creation of more complete models in the future, it is important to keep in mind that the fungal genomes examined here, *i)* remain haploid during the cell cycle and *ii)* have sister chromatids tightly maintained together by cohesion. We therefore believe that representation of fungal chromosomes as one string of beads is informative.

Overall, the limits underline the dependence of 3D modeling on the quality of experimental Hi-C raw data and generation methods. We are confident however, that the technical improvements

that are being introduced at a rapid pace will progressively minimize these limitations. Importantly, even though the technical ceiling is not yet reached, we have already managed to highlight important fungal genome characteristics in a novel way by reanalyzing public datasets. For *N. crassa*, the 6 3D models produced by 3DGB condensed the information of 5 research articles and about 20 raw FASTQ data files into one rich large-scale illustration (**Figure 5**). They bring new opportunities for visual integration of omics data: while Hi-C analysis often implies a "zoom-in" logic, focusing on precise regions of a contact map, 3D modeling completes and enhances this logic with a "zoomed-out" vision.

In conclusion, we have presented a holistic approach that is favorable to intuition and new hypotheses. We explored large-scale integration of epigenetic ChIP-seq data into the 3D context of chromatin, but any suitable combination of omics datasets can be made using the genome model as a visual support. For example, we are currently exploring the interpolation of genes on the 3D genome structure. 3DGB could be adapted to provide a model at the gene (one bead each) level, to allow the mapping of gene-dependent omics data, such as RNA-seq, on the chromosome structure. Gene positions need not be accurate, but rather provide a convenient way to plot omics data on top of the 3D organization of the chromatin. Going even further, it could be of interest to integrate information from protein occupancy or transcription factor interaction networks. Seeing chromatin domains as loops of a coil instead of distinct entities could open new insights on inter-domain gene regulation.

Methods

3DGB implementation and workflow management

Technical details

3DGB was orchestrated with the open-source workflow management system Snakemake (Mölder et al. 2021), which automates the different steps of a data analysis in a human-readable, Python-based language. 3DGB is structured in multiple individual rules depicted in **Supplementary Figure S1**. Software environments were deployed and isolated in a Conda environment (for Pastis and all custom Python scripts) and a Singularity container (for HiC-Pro).

3DGB inputs

3DGB requires the following inputs: Hi-C FASTQ files (provided as SRA IDs or as local FASTQ files), a reference genome sequence provided in a FASTA file (without mitochondrial DNA), one or several enzyme restriction site motifs, and finally, values for the output resolution to draw the contact map and generate 3D models.

Main steps in the analysis

The first steps of the workflow format the necessary information for the read mapping step: *(i)* FASTQ files are downloaded and compressed if not already provided by the user ; *(ii)* a list of fragments derived from the genome FASTA file and the enzyme restriction site motif is generated by HiC-Pro (Servant et al. 2015) version 3.1.0 ; *(iii)* the size of each chromosome from the genome FASTA file is extracted with a custom script and *(iv)* the reference genome is indexed with Bowtie2 version 2.4.4. Next, HiC-Pro generates a set of ‘valid pairs’, which are relevant reads used to generate and normalize contact frequencies (or counts). For a given resolution, these values are then used to compute contact maps (heatmaps) and to serve as inputs for Pastis (Varoquaux et al. 2014, 2021) (version as of July, 21st, 2021) to build a consensus 3D model. In 3DGB, the Pastis-NB method is used. It iteratively computes 3D models of the organization of

chromosomes, through negative binomial contact count modelization. For better reproducibility and portability, HiC-Pro is used in a Singularity image provided by the authors of the software. Pastis is installed in a conda environment. In 3DGB, a final step to refine the 3D model provided by Pastis is available. It consists in predicting the coordinates for additional beads, for which initial calculations were missing in the 3D model from Pastis. Missing coordinates usually occur in centromeric and telomeric regions. Predictions are performed by interpolation, using monotonic cubic splines, as implemented by the pchip method in the Scipy Python library (Virtanen et al. 2020). Note that beads with missing coordinates localized at the extremities of chromosomes are discarded and beads with aberrant coordinates are filtered out based on a threshold applied to the Euclidean distance value calculated between neighboring beads.

3DGB outputs

3D models are stored in PDB files. From the raw model produced by Pastis, 3DGB annotates chromosomes by numbers, based on the reference sequence specified in the 3DGB inputs. Examples of models, fully annotated with 3DGB, are provided as PDB files in Supplementary Data (see source code and data availability). In PDB files, the chromosome annotation is present in the residue id (1, 2, 3…), in the chain id (A, B, C…) and in the residue name (C01, C02, C03…). This adaptation of the standard residue and chain fields in PDB files is particularly useful, allowing immediate visualization of chromosomes with visualization software. Eventually, quantitative values can be associated with the beads of the 3D model (for instance ChIP-Seq signal intensity), using the B factor field in PDB files. Most PDB viewers can display B factors on top of the structures. This is the case for Mol*, used to create images presented in this article.

Analyses of experimental datasets in fungal species

Access to raw data from SRA database

All the Hi-C seq, Micro-C XL seq and ChIP seq raw FASTQ files were downloaded from the SRA database, see **Table 1** and **Supplementary Table 1** (second tab for ChIP seq) for all SRA ids.

The reference genomes of *S. cerevisiae* and *S. pombe* were obtained from the NCBI Genome database, respectively version R64 (S288C) and ASM294v2.19. The nc14 version of the *N. crassa* genome was downloaded from the supplementary data of Rodriguez et al. on NCBI GEO database, GEO dataset GSE173593. FASTA files with chromosome sequences only were generated as 3DGB input.

Application of 3DGB to create 3D models

3DGB was applied with default parameters (FASTQ files, see **Table 1** for all details) and configuration files can be accessed in Zenodo and Github. For *S. cerevisiae*, 2 models were generated at 5 kb resolution from 4 FASTQ files (total of 31 028 357 valid pairs) and ChiP-Seq information from 2 FASTQ files were integrated. For *S. pombe*, 8 models at 10 kb resolution were built, from 19 FASTQ files (total of 136 321 555 valid pairs). For *N. crassa*, 3 models were created at 50 kb and 10 kb resolutions, from 14 FASTQ files (total of 115 588 951 valid pairs) and ChiP-Seq information from 9 FASTQ files were integrated. The chosen resolutions of the models were defined according to the technique used (Hi-C or Micro-C XL), the number of reads and their quality.

Evaluation of 3D model stability to random noise

The *hpo* mutant *N. crassa* contact frequency matrix was used as a "reference" to generate “noisy contact frequency matrices”. The *hpo* mutant was selected because it has the lowest number of valid read pairs (2,922,526) and this sample is the most likely to be sensitive to noise. Let x_{ij} be the number of counts in the contact frequency matrix of this sample, with i and j the indices of the beads ranging from 1 to $n = 825$, with n the total number of beads for this sample. To create an "altered" or “noisy” *N. crassa* contact frequency matrix, it was set $y_{ij} = x_{ij} + e_{ij}$ the number of counts in the noisy contact frequency matrix, where e_{ij} is a value sampled from a Poisson distribution with parameter λ, independently for each pair of beads (i, j). Note that the contract frequency matrices are symmetric and $e_{ij} = e_{ji}$ for each pair of beads. The Poisson distribution

was used, because it models shot noise, *i.e.* the variability in read counts that is due to the read sampling process rather than to biological variations (Anders and Huber 2010). The level of noise (*i.e.* the mean parameter λ of the Poisson distribution used to generate shot noise) varies among the following values 0.1, 5, 10, 15, 20 up to 110. For each value of λ , 50 replicates were generated, *i.e.* 50 noisy contact frequency matrices. To assess the stability of 3D structures, the RMSD were computed between the 3D structure inferred from the reference *N. crassa* contact frequency matrix (the raw contact matrix with counts x_{ij}) and the 3D structure inferred from the altered *N. crassa* maps (the noisy contact frequency matrix with counts y_{ij}). This procedure leads to 50 values of RMSD for each value of λ. The **Supplementary Figure S5** represents the RMSD values for λ ranging from 0.1 to 50. Above the value $\lambda = 50$, the RMSD values reach a threshold and do not increase with the value of λ. The RMSD between each one of the seven chromosomes of *N. crassa* were also computed, for each value of λ and each of the 50 replicates.

Visual integration of omics data

The same pipeline was used for integration of ChIP-Seq data from *N. crassa* and *S. cerevisiae*. To perform alignment, Bowtie2 was used with default parameters (**Supplementary Table 1, tab 2)**. The program samtools was used to sort and index the SAM files into bam files, and the program bamCoverage was used to generate 5 kb or 50 kb .bedgraph files with RPKM normalization (for *S. cerevisiae* and *N. crassa* respectively). Each bin in the bedgraph corresponds to one bead of the 3D model. Note that the resolution of the bedgraph and of the 3D model are the same, allowing the integration of the data into the PDB file. For *S. cerevisiae*, the threshold was set to 80 counts after normalization, converting the continuous ChIP-seq signal into a binary signal highlighting the high residency CARs (Costantino et al. 2020). For *N. crassa*, the ChIP-seq data was kept continuous to highlight the epigenetic mark distribution. The values below 20 counts are discarded for visualization.

Data access

The 3DGB workflow is available on GitHub: https://github.com/data-fun/3d-genome-builder. The .PDB files of the 13 models presented in this study are available on Zenodo 10.5281/zenodo.7740302, as well as animated GIF for 4 of them. The .YML (config files) for generating those 13 models are available on GitHub and on Zenodo.

Competing interest statement

The authors declare that they have no competing interests.

Acknowledgments

We thank Nelle Varoquaux for useful discussions. This work was funded by the Agence Nationale pour la Recherche (MINOMICS project, Grant Number ANR-19-CE45-0017). Figures were made on BioRender.

References

Anders S, Huber W. 2010. Differential expression analysis for sequence count data.

Arifulin EA, Musinova YR, Vassetzky YS, Sheval EV. 2018. Mobility of Nuclear Components and Genome Functioning. *Biochemistry Moscow* **83**: 690–700.

Asbury TM, Mitman M, Tang J, Zheng WJ. 2010. Genome3D: A viewer-model framework for integrating and visualizing multi-scale epigenomic information within a three-dimensional genome. *BMC Bioinformatics* **11**: 444.

Basenko EY, Sasaki T, Ji L, Prybol CJ, Burckhardt RM, Schmitz RJ, Lewis ZA. 2015. Genome-wide redistribution of H3K27me3 is linked to genotoxic stress and defective growth. *Proc Natl Acad Sci USA* **112**: E6339–E6348.

Bauer CR, Hartl TA, Bosco G. 2012. Condensin II Promotes the Formation of Chromosome Territories by Inducing Axial Compaction of Polyploid Interphase Chromosomes ed. R.S. Hawley. *PLoS Genet* **8**: e1002873.

Baumeister W. 2022. Cryo-electron tomography: The power of seeing the whole picture. *Biochemical and Biophysical Research Communications* **633**: 26–28.

Carlier F, Li M, Maroc L, Debuchy R, Souaid C, Noordermeer D, Grognet P, Malagnac F. 2021. Loss of EZH2-like or SU(VAR)3–9-like proteins causes simultaneous perturbations in H3K27 and H3K9 tri-methylation and associated developmental defects in the fungus Podospora anserina. *Epigenetics & Chromatin* **14**: 22.

Costantino L, Hsieh T-HS, Lamothe R, Darzacq X, Koshland D. 2020. Cohesin residency determines chromatin loop patterns. *eLife* **9**: e59889.

Cremer T, Cremer C. 2001. Chromosome territories, nuclear architecture and gene regulation in mammalian cells. *Nat Rev Genet* **2**: 292–301.

Denecker T, Zhou Li Y, Fairhead C, Budin K, Camadro J-M, Bolotin-Fukuhara M, Angoulvant A, Lelandais G. 2020. Functional networks of co-expressed genes to explore iron homeostasis processes in the pathogenic yeast Candida glabrata. *NAR Genomics and Bioinformatics* **2**: lqaa027.

Djekidel MN, Wang M, Zhang MQ, Gao J. 2017. HiC-3DViewer: a new tool to visualize Hi-C data in 3D space. *Quant Biol* **5**: 183–190.

Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS. 2010. A three-dimensional model of the yeast genome. *Nature* **465**: 363–367.

Dundr M, Misteli T. 2001. Functional architecture in the cell nucleus.

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Systems* **3**: 95–98.

Fritz AJ, Sehgal N, Pliss A, Xu J, Berezney R. 2019. Chromosome territories and the global regulation of the genome. *Genes Chromosomes Cancer* **58**: 407–426.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma L-J, Smirnov S, Purcell S, et al. 2003. The genome sequence of the filamentous fungus Neurospora crassa. *Nature* **422**: 859–868.

Galazka JM, Klocko AD, Uesaka M, Honda S, Selker EU, Freitag M. 2016. Neurospora chromosomes are organized by blocks of importin alpha-dependent heterochromatin that are largely independent of H3K9me3. *Genome Res* **26**: 1069–1080.

Gallardo P, Barrales RR, Daga RR, Salas-Pino S. 2019. Nuclear Mechanics in the Fission Yeast. *Cells* **8**: 1285.

Goodsell DS. 2009. *The machinery of life*. 2nd ed., corrected. Copernicus Books, New York.

Grand RS, Pichugina T, Gehlen LR, Jones MB, Tsai P, Allison JR, Martienssen R, O'Sullivan JM. 2014. Chromosome conformation maps in fission yeast reveal cell cycle dependent sub nuclear structure. *Nucleic Acids Research* **42**: 12585–12599.

Grognet P, Timpano H, Carlier F, Aït-Benkhali J, Berteaux-Lecellier V, Debuchy R, Bidard F, Malagnac F. 2019. A RID-like putative cytosine methyltransferase homologue controls sexual development in the fungus Podospora anserina ed. E. Gladyshev. *PLoS Genet* **15**: e1008086.

Guido Van Rossum, Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace.

Hoencamp C, Dudchenko O, Elbatsh AMO, Brahmachari S, Raaijmakers JA. 2022. 3D genomics across the tree of life identifies condensin II as a determinant of architecture type.

Hsieh T-HS, Fudenberg G, Goloborodko A, Rando OJ. 2016. Micro-C XL: assaying chromosome conformation from the nucleosome to the entire genome. *Nat Methods* **13**: 1009–1011.

Im W, Liang J, Olson A, Zhou H-X, Vajda S, Vakser IA. 2016. Challenges in structural approaches to cell modeling. *Journal of Molecular Biology* **428**: 2943–2964.

Jamieson K, Wiles ET, McNaught KJ, Sidoli S, Leggett N, Shao Y, Garcia BA, Selker EU. 2016. Loss of HP1 causes depletion of H3K27me3 from facultative heterochromatin and gain of H3K27me2 at constitutive heterochromatin. *Genome Res* **26**: 97–107.

Jensen EC. 2013. Overview of Live-Cell Imaging: Requirements and Methods Used. *Anat Rec* **296**: 1–8.

Jerkovic´ I, Cavalli G. 2021. Understanding 3D genome organization by multidisciplinary methods. *Nat Rev Mol Cell Biol*. http://www.nature.com/articles/s41580-021-00362-w (Accessed June 18, 2021).

Kempfer R, Pombo A. 2020. Methods for mapping 3D chromosome architecture. *Nat Rev Genet* **21**: 207–226.

Kim K-D, Tanizawa H, Iwasaki O, Noma K. 2016. Transcription factors mediate condensin recruitment and global chromosomal organization in fission yeast. *Nat Genet* **48**: 1242–1252.

Lelandais G, Remy D, Malagnac F, Grognet P. 2022. New insights into genome annotation in Podospora anserina through re-exploiting multiple RNA-seq data. *BMC Genomics* **23**: 859.

Li D, Harrison JK, Purushotham D, Wang T. 2022. Exploring genomic data coupled with 3D chromatin structures using the WashU Epigenome Browser. *Nat Methods* **19**: 909–910.

Li J, Zhang W, Li X. 2018. 3D Genome Reconstruction with ShRec3D+ and Hi-C Data. *IEEE/ACM Trans Comput Biol and Bioinf* **15**: 460–468.

Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Misteli T. 2020. The Self-Organizing Genome: Principles of Genome Architecture and Function. *Cell* **183**: 28–45.

Mitter M, Gasser C, Takacs Z, Langer CCH, Tang W, Jessberger G, Beales CT, Neuner E, Ameres SL, Peters J-M, et al. 2020. Conformation of sister chromatids in the replicated human genome. *Nature* **586**: 139–144.

Mizuguchi T, Fudenberg G, Mehta S, Belton J-M, Taneja N, Folco HD, FitzGerald P, Dekker J, Mirny L, Barrowman J, et al. 2014. Cohesin-dependent globules and heterochromatin shape 3D genome architecture in S. pombe. *Nature* **516**: 432–435.

Mölder F, Jablonski KP, Letcher B, Hall MB, Tomkins-Tinch CH, Sochat V, Forster J, Lee S, Twardziok SO, Kanitz A, et al. 2021. Sustainable data analysis with Snakemake. *F1000Res* **10**: 33.

Nogales E, Scheres SHW. 2015. Cryo-EM: A Unique Tool for the Visualization of Macromolecular Complexity. *Molecular Cell* **58**: 677–689.

Noma K. 2017. The Yeast Genomes in Three Dimensions: Mechanisms and Functions. *Annu Rev Genet* **51**: 23–44.

Nowotny J, Wells A, Oluwadare O, Xu L, Cao R, Trieu T, He C, Cheng J. 2016. GMOL: An Interactive Tool for 3D Genome Structure Visualization. *Sci Rep* **6**: 20802.

O'Donoghue SI. 2021. Grand Challenges in Bioinformatics Data Visualization. *Front Bioinform* **1**: 669186.

Oluwadare O, Highsmith M, Cheng J. 2019. An Overview of Methods for Reconstructing 3-D Chromosome and Genome Structures from Hi-C Data. *Biol Proced Online* **21**: 7.

Oomen ME, Hedger AK, Watts JK, Dekker J. 2020. Detecting chromatin interactions between and along sister chromatids with SisterC. *Nat Methods* **17**: 1002–1009.

Poinsignon T, Gallopin M, Camadro J-M, Poulain P, Lelandais G. 2022. Additional insights into the organization of transcriptional regulatory modules based on a 3D model of the Saccharomyces cerevisiae genome. *BMC Res Notes* **15**: 67.

Pouokam M, Cruz B, Burgess S, Segal MR, Vazquez M, Arsuaga J. 2019. The Rabl configuration limits topological entanglement of chromosomes in budding yeast. *Sci Rep* **9**: 6795.

Radulović S, Sunkara S, Rachel R, Leitinger G. 2022. Three-dimensional SEM, TEM, and STEM for analysis of large-scale biological systems. *Histochem Cell Biol* **158**: 203–211.

Razin SV, Borunova VV, Iarovaia OV, Vassetzky YS. 2014. Nuclear matrix and structural and functional compartmentalization of the eucaryotic cell nucleus. *Biochemistry Moscow* **79**: 608–618.

Reckel S, Löhr F, Dötsch V. 2005. In-Cell NMR Spectroscopy. *ChemBioChem* **6**: 1601–1606.

Rieber L, Mahony S. 2017. miniMDS: 3D structural inference from high-resolution Hi-C data. *Bioinformatics* **33**: i261–i266.

Rodriguez S, Ward A, Reckard AT, Shtanko Y, Hull-Crew C, Klocko AD. 2022. The genome organization of *Neurospora crassa* at high resolution uncovers principles of fungal chromosome topology ed. J. Dekker. *G3 Genes|Genomes|Genetics* **12**: jkac053.

Rodriguez-Granados NY, Ramirez-Prado JS, Veluchamy A, Latrasse D, Raynaud C, Crespi M, Ariel F, Benhamed M. 2016. Put your 3D glasses on: plant chromatin is on show. *EXBOTJ* **67**: 3205–3221.

Sear RP, Pagonabarraga I, Flaus A. 2015. Life at the mesoscale: the self-organised cytoplasm and nucleoplasm. *BMC Biophys* **8**: 4.

Sehnal D, Bittrich S, Deshpande M, Svobodová R, Berka K, Bazgier V, Velankar S, Burley SK, Koča J, Rose AS. 2021. Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. *Nucleic Acids Research* **49**: W431–W437.

Sénécaut N, Poulain P, Lignières L, Terrier S, Legros V, Chevreux G, Lelandais G, Camadro J-M. 2022. Quantitative Proteomics in Yeast: From bSLIM and Proteome Discoverer Outputs to Graphical Assessment of the Significance of Protein Quantification Scores. In *Yeast Functional Genomics* (ed. F. Devaux), Vol. 2477 of *Methods in Molecular Biology*, pp. 275–292, Springer US, New York, NY https://link.springer.com/10.1007/978-1-0716-2257-5_16 (Accessed March 3, 2023).

Servant N, Varoquaux N, Lajoie BR, Viara E, Chen C-J, Vert J-P, Heard E, Dekker J, Barillot E. 2015. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol* **16**: 259.

Smyth MS. 2000. x Ray crystallography. *Molecular Pathology* **53**: 8–14.

Stevens TJ, Lando D, Basu S, Atkinson LP, Cao Y, Lee SF, Leeb M, Wohlfahrt KJ, Boucher W, O'Shaughnessy-Kirwan A, et al. 2017. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature* **544**: 59–64.

Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K. 2010. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Research* **38**: 8164–8177.

Tanizawa H, Kim K-D, Iwasaki O, Noma K. 2017. Architectural alterations of the fission yeast genome during the cell cycle. 31.

Todd S, Todd P, McGowan SJ, Hughes JR, Kakui Y, Leymarie FF, Latham W, Taylor S. 2021. CSynth: an interactive modelling and visualization tool for 3D chromatin structure ed. A. Valencia. *Bioinformatics* **37**: 951–955.

Tokuda N, Terada TP, Sasai M. 2012. Dynamical Modeling of Three-Dimensional Genome Organization in Interphase Budding Yeast. *Biophysical Journal* **102**: 296–304.

Trieu T, Oluwadare O, Wopata J, Cheng J. 2019. GenomeFlow: a comprehensive graphical tool for modeling and analyzing 3D genome structure ed. B. Berger. *Bioinformatics* **35**: 1416–1418.

Varoquaux N, Ay F, Noble WS, Vert J-P. 2014. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics* **30**: i26–i33.

Varoquaux N, Noble WS, Vert J-P. 2021. *Inference of genome 3D architecture by modeling overdispersion of Hi-C data*. Bioinformatics http://biorxiv.org/lookup/doi/10.1101/2021.02.04.429864 (Accessed January 14, 2022).

Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, Burovski E, Peterson P, Weckesser W, Bright J, et al. 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**: 261–272.

Wong AMH, Eleftheriades GV. 2013. An Optical Super-Microscope for Far-field, Real-time Imaging Beyond the Diffraction Limit. *Sci Rep* **3**: 1715.

Yardımcı GG, Noble WS. 2017. Software tools for visualizing Hi-C data. *Genome Biol* **18**: 26.

Zhang C, Xu Z, Yang S, Sun G, Jia L, Zheng Z, Gu Q, Tao W, Cheng T, Li C, et al. 2020. tagHi-C Reveals 3D Chromatin Architecture Dynamics during Mouse Hematopoiesis. *Cell Reports* **32**: 108206.