Université de Montréal

Une correction à l'échelle et progressive des données Hi-C révèlent des principes fondamentaux de l'organisation tridimensionnelle et fonctionnelle du génome

par

Ilunga Benjamin Matala

Département de Biochimie et médecine moléculaire
Faculté de Médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de
Maître ès sciences (M.Sc.)
en Bio-informatique

19 décembre 2016

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé

Une correction à l'échelle et progressive des données Hi-C révèlent des principes fondamentaux de l'organisation tridimensionnelle et fonctionnelle du génome

présenté par

Ilunga Benjamin Matala

a été évalué par un jury composé des personnes suivantes :

Sylvie Hamel
(président-rapporteur)

Stephen W. Michnick
(directeur de recherche)

Nicolas Lartillot
(codirecteur)

Matthieu Blanchette
(membre du jury)

Mémoire accepté le

RÉSUMÉ

Au cours des dernières années, de nouvelles évidences semblent indiquer que, tout autant que sa séquence, l'organisation d'un génome dans l'espace et le temps est importante pour comprendre la fonction de celui-ci. Une des avancées fondamentales sur le sujet a été de présenter à l'échelle du génome la carte des interactions ADN-ADN. Ces interactions sont essentiellement de 2 types, soit entre chromosomes ou entre régions du même chromosome. Par la suite, la modélisation a permis de visualiser et appréhender la structure tridimensionnelle (3D) du génome à partir des données 3C, ou d'une modélisation purement théorique. Une question importante et centrale demeure, soit de résoudre les mécanismes responsables de l'organisation spatiale et fonctionnelle du génome. Notamment, une question est de savoir comment des processus nucléaires tels que la transcription affectent la structure du génome. Cependant, l'idée selon laquelle les données de types 3C capturent cette information dans la levure est remise en question par le fait que les modèles théoriques du génome récapitulent les caractéristiques marquantes soulignées par 3C. Pour répondre à cette question, nous avons conçu une approche qui, pour évaluer l'importance d'une interaction, se base sur la distribution d'interactions entre les 2 régions d'ADN mises en contacts. Nos résultats supportent l'hypothèse selon laquelle les éléments fonctionnels et propres aux données expérimentales de la structure 3D du génome se forment d'une manière spécifique à l'échelle de l'interaction et au type d'interactions. Par ailleurs, nos résultats indiquent qu'un grand nombre de facteurs de transcription induisent la proximité spatiale des gènes dont ils régulent l'expression.

Mots clés: structure spatial (3D) du génome, organisation fonctionnel du génome, régulation de la transcription, données de type 3C (Hi-C), correction de données 3C

SUMMARY

Over the last decade, accumulating empirical evidence suggest that, as much as its sequence, a genome spatiotemporal organization is essential to understand it's biological function. One of the major breakthroughs has been chromosome conformation capture (3C) experiments presenting DNA-DNA contact for whole genomes at unprecedented resolution (5-10kb). Along with genome-wide maps of DNA contacts came genome 3D modelling from experimental 3C data, and even from purely theoretical and biophysical basis. However, the mechanisms underlying the regulation of the genome spatial functional organization are still not well understood. Among other questions, how the regulation and event of nuclear processes such as transcription modulate genome structure or how genome structure affect these in turn is still not fully resolved. Moreover, computational models of S.cerevisae genome have recapitulated the hallmarks at larger scale of its 3D features. In order to contrast genome structural features arising from the event of biochemical and molecular activity, we have develop a method assessing the significance of structural features. The underlying principle is to consider for a given interaction, the two DNA regions put in contact and the distribution of existing interactions between these before assigning significance to the selected interaction. Using this method, we demonstrate that structural features resulting from potential biochemically active processes occur at precise scale on the genome. Our results also highlight that exact nature of the interaction (between vs across chromosomes) is crucial to such events. Finally, we have also found that a large portion of transcription factors have their targeted genes in spatial proximity.

Keywords: genome functional organization, genome spatial (3D) structure, transcriptional regulation

TABLE DES MATIÈRES

LISTE DES FIGURES

REMERCIEMENTS

Dr Stephen W. Michnick: : Je voudrais vous remercier de m'avoir donné l'opportunité de faire une maîtrise en bio-informatique, venant d'un bacc. en biochimie. Pour avoir financé et soutenu mon temps d'apprentissage à cette formation. Pour m'avoir appris et transmis une intuition, un sens et une appréciation de la biologie avec excellence admirable.

Dr Nicolas Lartillot: De m'avoir donné l'occasion d'être formé par lui, de pouvoir voir à quoi ressemble une passion et habileté remarquable à appliquer les mathématique à des problèmes biologiques.

Stephanie Lefebvre: Pour tout, de croire en moi, me soutenir quand rien n'a l'air de marché, quand je n'ai pas réussi à bien faire ! D'être là tous les jours, de partager les moments de la vie, qui font que j'ai le courage et l'esprit sain pour continuer les études.

Mes parents et ma famille: Je voudrais vous remericer d'avoir été les premiers à me soutenir, m'encourager. Votre soutien, et vos mains ont toujours été là pour me porter.

Md Elaine Meunier: D'être cosuperviseuse anonyme, celle qui rend possible tout ce cheminement. Je vous assure que votre aide a été clé et vitale pour moi ! Sans elle, je ne serais pas où j'en suis, littéralement !

Louis-Philippe Bergeron-Sandoval: Pour m'avoir aidé et dirigé dans le développement de mon projet, à des moments clés. Et dans tout ça, d'avoir cru et investi en moi, bien que mes propres efforts ou rendements n'étaient alors pas exceptionnels.

Chapitre 1

MOTIVATIONS SCIENTIFIQUES DU TRAVAIL

L'identité d'une cellule est déterminée par son ADN, qui contient l'information nécessaire à la fabrication des protéines et des ARNs de la cellule. Pour réguler l'activité d'une cellule, il est donc nécessaire de réguler en premier lieu les activités effectuées sur l'ADN, telles que la transcription ou les modifications d'épigénétiques. Au cours des dernières années, l'information acquise sur l'organisation spatiale du génome a amené à revisiter profondément notre compréhension de la génétique [Bickmore and van Steensel, 2013; Cavalli and Misteli, 2013]. Ce point soulève le fait que tout autant que sa séquence, l'organisation dans l'espace et le temps du génome est importante pour que celui-ci puisse correctement exercer sa fonction. Une des avancées fondamentales sur le sujet en son temps a été de présenter à l'échelle du génome la carte des contacts ADN-ADN (Hi-C ou dérivé de 3C) chez la levure Saccharomyces cerevisiae [Duan, Andronescu, Schutz, McIlwain, Kim, Lee, Shendure, Fields, Blau, and Noble, 2010]. Un domaine qui s'est développé en parallèle et qui est connexe est la modélisation de l'ADN génomique. Ce genre de modélisation a permis de visualiser et appréhender la structure tridimensionnelle (3D) du génome à partir des données 3C, ou d'analyses s'appuyant sur une modélisation purement théorique.

Néanmoins, une question importante et centrale qui demeure est de résoudre les mécanismes qui régulent l'organisation spatiale et fonctionnelle du génome. Afin de répondre à cette question, nous posons l'hypothèse que la régulation de la structure dynamique du génome est dictée par des processus biologiques nucléaires intimement reliés à l'ADN et l'ARN, et médiés par des complexes composés d'ARN et protéines. En effet, bon nombre d'évidences supportent l'idée que les protéines impliquées dans la régulation de la transcription sont aussi impliquées dans la formation de la conformation locale du génome. Cependant, l'idée selon laquelle les données de type 3C capturent cette information dans la levure est

remise en question par le fait que des modèles théoriques du génome récapitulent les caractéristiques marquantes de l'organisation spatiale de celui-ci [Duan et al., 2010]. Les principes sur lesquelles se basent ces modèles sont en majeure partie l'exclusion de volume (entre chromosomes), l'ancrage des centromères, les propriétés biophysique de l'ADN comme polymère, et l'enveloppe nucléaire [Kimura, Shimooka, Nishikawa, Miura, Sugiyama, Yamada, and Ohyama, 2013; Tjong, Gong, Chen, and Alber, 2012; Wong, Marie-Nelly, Herbert, Carrivain, Blanc, Koszul, Fabre, and Zimmer, 2012]. On retrouve au nombre de ces caractéristiques l'émergence de territoires chromosomiques, le positionnement en périphérie des télomères, la conformation dite « Rabl », la déplétion d'interactions entre les bras de chromosome de part et d'autre des centromères, la proximité spatiale des bras chromosomiques de taille semblable, et celle de loci partageant une fonction tel que les gènes d'ARNt et les origines de réplications. Ce qui est à retirer de ces observations est que les contraintes physiques à large échelle, ainsi que les propriétés de la biophysique de l'ADN et de la chromatine sont responsables de la majeure partie du relief de contacts ADN-ADN. Dans le cas éventuel où des interactions résultant de réactions biochimiques (actives) auraient lieu, il serait donc nécessaire de tenir compte des facteurs précédemment mentionnés. C'est dans cette perspective que nous avons conçu une approche par région où pour une interaction donnée, la distribution d'interaction entre les 2 régions d'ADN mise en contacts est utilisée pour évaluer l'importance de cette interaction. Voici donc les résultats obtenus lors de cet effort.

Chapitre 2

ORGANISATION NUCLÉAIRE DE LA LEVURE

2.1. Structure et processus nucléaire de la cellule

Depuis la découverte de la structure de l'ADN par rayon X (cristallographie) il y a presque 60 ans, beaucoup de choses sur le génome et l'organisation nucléaire ont été découvertes. Il y a ce qui a trait au génome lui-même et à sa structure, ce qui a trait à d'autres structures importantes retrouvées dans le noyau, et ce qui a trait aux différents évènements, processus nucléaires et les facteurs clés dans la coordination de ces évènements et processus. On compte parmi les structures importantes trouvées dans le noyau d'une cellule : le nucléole, les régions centromériques (autour des centromère) la membrane nucléaire et les complexes protéiques qui y sont associés en périphérie du noyau. Parmi les processus nucléaires intimement liés au génome sont comptées la transcription, la réplication, la réparation de l'ADN. Cela dit, il y a aussi des processus non liés (directement) au génome comme le traitement et l'épissage des ARN, ainsi qu'une partie de la synthèse des ribosomes, de laquelle d'ailleurs découle la structure du nucléole, qui est donc une organelle nucléaire sans membrane par extension.

2.2. Le génome de *Sacharomyces cerevisiæ*

En ce qui a trait au génome, celui de la levure est composé de 16 chromosomes. La levure peut être retrouvée sous forme haploïde (16 chromosomes), ou diploïde (2 x 16 chromosomes). La taille du génome haploïde est d'environ 12 millions de base (Mb). Comme pour la plupart des eucaryotes, l'ADN des chromosomes est associé à des protéines appelées histones, autour desquels il est enroulé. L'ensemble composé d'ADN, d'histones, d'autres protéines associés et même d'ARN est appelé chromatine. Les régions de l'ADN où celui-ci est fortement enroulé

autour des histones (régions fermées), et à l'aide d'autres protéines, forment l'hétérochromatine. Les régions plutôt ouvertes ou peu enroulées de l'ADN forment l'euchromatine. Une des distinctions particulières entre l'hétéro- et l'euchromatine est le fait que l'euchromatine de par son état ouvert facilite l'accès à l'ADN pour d'autres protéines pour accomplir diverses tâches telles que la réplication de l'ADN, sa réparation, sa transcription, sa modification, etc. Bien que la règle ne soit pas absolue, une partie importante de l'hétérochromatine se retrouve en périphérie du noyau, et plusieurs des gènes qui s'y retrouvent sont réprimés. À l'inverse, l'euchromatine tend plutôt à se retrouver vers l'intérieur du noyau. Une exception particulière est celle des pores nucléaires, canaux d'échanges entre le cytoplasme et le nucléoplasme (intérieur du noyau). En plus de la simple différence entre hétéro- et euchromatine, il y a aussi les modifications épigénétiques qui caractérisent l'état de l'ADN. Ces modifications se retrouvent soit sur l'ADN lui-même,soit sur les protéines qui lui sont associées. Plusieurs de ces modifications consistent par exemple en des méthylations ou acétylations des histones sur des positions bien spécifiques, ou encore en une méthylation de l'ADN. En général, les modifications peuvent être associées à l'hétérochromatine et d'autres à l'euchromatine, bien que plusieurs exceptions existent.

2.3. L'organisation nucléaire flexible et corps nucléaire

Il existe une autre catégorie de corps nucléaires qui peuvent être caractérisés comme organelles (nucléaire) sans membrane. Le nucléole, par exemple, est formé autour des séquences sur le génome portant les ARN ribosomaux (ADNr). Il s'agit en fait d'un environnement très compact composé d'ARNs, de protéines, et de complexes ribo protéiques, qui exclut les autres éléments ou corps nucléaires, et la chromatine du reste du génome. En fait, il y a dans le noyau noyau d'une cellule plusieurs types de corps nucléaires dépendant de l'ARN tel les « pbodies », les « cajal bodies »[Caudron-Herger and Rippe, 2012], dans lesquelles se déroulent des processus reliés à de l'ARN codant et non codant[Chinen and Tani, 2012]. Tout récemment, des travaux ont apporté des évidences du rôle de l'ARN dans la formation de certains corps nucléaires et même dans l'état de la chromatine, modifiant sa compaction conditionnellement à la présence de certains de ces ARN [Han, Kato, Xie, Wu, Mirzaei, Pei, Chen, Xie, Allen, Xiao, and McKnight, 2012; Kato, Han, Xie, Shi, Du, Wu, Mirzaei, Goldsmith, Longgood, Pei, Grishin, Frantz, Schneider, Chen, Li, Sawaya, Eisenberg, Tycko, and McKnight, 2012; Schubert, Pusch, Diermeier, Benes, Kremmer, Imhof, and Langst, 2012]. Il n'est pas rare que ce type d'organelles jouent un rôle dans la régulation et la coordination moléculaire des processus nucléaires [Nott, Craggs, and Baldwin, 2016]. Ces travaux,

avec plusieurs autres, soulignent aussi le rôle des protéines dans la formation des structures nucléaires et le remodelage de la chromatine.

2.4. Régulation de la transcription

Il a été observé que certains gènes transcriptionnellement actifs se retrouvaient à proximité de ces pores nucléaires. Une hypothèse pour expliquer cette observation serait qu'une telle localization faciliterait le transport du produit de leur transcription vers le cytoplasme. L'environnement autour de ces pores serait donc associé à l'euchromatine. Notamment, il a été observé que certains gènes actifs colocalisent avec les pores nucléaires. Une certaine séquence («zip code») a été identifiée comme nécessaire et suffisante pour une colocalisation interchromosomique des gènes en question [Brickner, Ahmed, Meldi, Thompson, Light, Young, Hickman, Chu, Fabre, and Brickner, 2012]. Qui plus est, le facteur de transcription Put3 a été identifié comme liant spécifiquement cette séquence.

2.5. La structure tridimensionnelle du génome de *S. cerevisæ*

Dans la dernière décennie, deux travaux publiés ont apporté une contribution très grande à la compréhension de l'organisation nucléaire : soit ceux de l'équipe du Dr Job Dekker [Lieberman-Aiden, van Berkum, Williams, Imakaev, Ragoczy, Telling, Amit, Lajoie, Sabo, Dorschner, Sandstrom, Bernstein, Bender, Groudine, Gnirke, Stamatoyannopoulos, Mirny, Lander, and Dekker, 2009] , et ceux de l'équipe du Dr William S. Noble [Duan et al., 2010]. Ces deux équipes ont présenté pour la première fois une image à l'échelle du génome de la plage d'interaction d'un locus avec le reste des chromosomes, à une échelle de 1 mégabase (Mb) sur le génome humain [Lieberman-Aiden et al., 2009], et 10 kilobase (Kb) pour le génome haploïde de la levure [Duan et al., 2010]. Dans les deux cas, les données présentées sont basées sur une technique de capture de la conformation des chromosomes (3C, ou « Chromosome Conformation Capture », aussi appelé Hi-C) avec certaines variantes [Schmitt, Hu, and Ren, 2016]. Le principe est de fixer les cellules, et donc leur noyau, avec du formaldéhyde. Cette étape de fixation lie ensemble, par des liaisons covalentes, l'ADN et les protéines à proximité (réagencement ou « cross-linking »). Par la suite, la chromatine ainsi figée est digérée à l'aide d'enzymes de restriction pour obtenir des fragments d'ADN plus petits. Au final, ces fragments d'ADN, retenus ensemble par l'intermédiaire des protéines auxquelles ils sont liés, vont être ligués ensemble, puis séquencés après dissolution des protéines. Au final, les données se présentent sous la forme d'une

matrice d'interactions intrachromosomiques et interchromosomiques des divers loci à travers le génome. D'autres travaux du même genre ont été menés par la suite sur d'autres organismes[Denker and de Laat, 2016] .

Suite à ces travaux, il est donc devenu possible d'explorer pour un gène donné l'ensemble de loci ou gènes avec lequel celui peut interagir pour exercer une quelconque fonction. Ces données ont aussi permis de soutenir plusieurs observations effectuées dans le cadre d'autres types d'expériences, dont la microscopie. Tout d'abord, ces données suggèrent la présence de territoires chromosomiques. En effet, la probabilité d'interaction entre deux loci situés à l'intérieur du même chromosome est en moyenne nettement plus grande que celle entres deux loci situés sur des chromosomes différents, et ce, même à de très grandes distance intrachromosomiques [Duan et al., 2010; Lieberman-Aiden et al., 2009]. Par ailleurs, l'interaction entre deux chromosomes semble plus probable pour certaines paires. Notamment, il semble que les chromosomes de taille semblable interagissent davantage entre eux que ne le font des chromosomes de tailles très différentes. Un des faits inattendus est que la fréquence d'interaction entre deux fragments d'ADN le long d'un même bras chromosomique diminue graduellement, alors que cette même différence semble abruptement réduite lorsqu'on passe d'un côté du centromère à l'autre. Le regroupement en grappes (« clustering ») de certains sites fonctionnels tels que les séquences codant les ARNt ou les sites d'origines de réplication a aussi été observé.

Il a été noté qu'il existe un lien étroit entre la fréquence d'interaction entre les fragments d'ADN et leur distance dans l'espace [Duan et al., 2010]. Plus précisément, il s'agit d'une relation inverse telle que plus la fréquence d'interaction est élevée, plus la distance entre les fragments est courte. L'observation de cette correlation a donc permis de proposer un premier modèle tridimensionnel du génome de la levure, basé sur les données de type 3C [Duan et al., 2010]. En fait, la structure tridimensionnelle est obtenue par une modélisation du génome dans l'espace et une résolution des contraintes de distances entre les paires de fragments d'ADN, telles qu'inférées par la matrice des fréquences d'interactions. Ce modèle tridimensionnel (3D) reflète plusieurs caractéristiques déjà observées. Notamment, la conformation « Rabl »déjà observée expérimentalement [Schober, Kalck, Vega-Palas, Van Houwe, Sage, Unser, Gartenberg, and Gasser, 2008] est aussi mise en évidence dans la structure 3D de la levure. Toutefois, il est important de souligner le fait que les données obtenues lors de l'expérience de capture de conformation des chromosomes sont en fait obtenues à partir d'une population de cellules, alors que l'image 3D est unique, ce qui peut entraîner certaines limitations.

2.6. Un modèle *in silico* biophysique de l'organisation tridimensionnelle du génome de S. cerevisae

Alors que plusieurs hypothèses ont été proposées quant aux causes et mécanismes qui induisent et régulent l'organisation spatiale du génome, une équipe a permis d'élucider deux phénomènes importants à considérer : l'ancrage physique des chromosomes et l'exclusion de volume [Tjong et al., 2012]. En effet, une modélisation du génome 3D de levure comme polymère d'ADN dans laquelle on impose comme contraintes d'ancrer les chromosomes au centre de microtuble (« Spindle Pole Body »ou SPB) par leur centromère, en restreignant les télomères à la périphérie du noyau et en considérant l'exclusion de volume causé par le nucléole suffit à prédire une partie importante des caractéristiques de l'organisation du génome de la levure. Aucun facteur spécifique liant l'ADN n'a été ajouté. Plus spécifiquement, la démarche d'optimisation de la structure 3D selon les contraintes ci-dessus, combinées à d'autres contraintes imposées sur les propriétés biophysiques de l'ADN, a permis de générer un très large échantillon de structures alternatives pour l'ensemble du génome. Et c'est de cet ensemble que les fréquences d'interaction ont été obtenues pour l'analyse des données. Parmi les points relevés dans ce modèle, il semblerait que l'ancrage des chromosomes par leur centromère au SPB crée une "pression", ou exclusion de volume, de sorte que les chromosomes de plus grandes tailles tendent à occuper l'espace nucléaire plus distant du SPB, tandis que les chromosomes plus courts occupent davantage un espace plus rapproché du SPB. De plus, le jeu de données des fréquences d'interaction inférées par l'ensemble des structures issues du modèle reprend aussi le fait que certaines paires de chromosomes de taille semblable tendent davantage à interagir entre elles. Il a aussi été observé que la fréquence d'interaction intrachromosomique est plus élevée sur un même bras chromosomique, que de part et d'autre d'un centromère. Ceci serait probablement dû à la pression et à l'encombrement causé par la présence de plusieurs chromosomes contraints à occuper un espace restreint à proximité du SPB, à cause de leur ancrage. Encore plus étonnant est le fait que le regroupement des loci associés au ARNt et aux origines de réplication est aussi observé. Ainsi donc, des contraintes biophysiques simples suffisent à expliquer la structure tridimensionnelle globale de la chromatine, telle qu'inférée via les expériences de capture de conformation des chromosomes (3C).

Wong et al. [2012] ont aussi proposé un modèle simulé du génome de la levure. Ce modèle, basé encore une fois sur une modélisation de la fibre chromatine pour l'ensemble du génome, a permis de relever plusieurs points de l'organisation du génome, comme celle de l'équipe de Tjong et al. [2012]. Cependant, la particularité

de ce modèle a été qu'au lieu de prédéfinir le volume occupé par le nucléole, celui-ci émerge naturellement comme résultat du fait que la fibre de chromatine de l'ADN ribosomal (ADNr) situé sur le chromosome XII est modélisée comme plus volumineuse ; ce qui représenterait l'ensemble de protéines et ARN entourant cette région du génome. Ce faisant, il semble que l'exclusion de volume place naturellement le nucléole à l'opposé du SPB, tout en permettant de souligner le point de l'organisation du génome de manière semblable au modèle de Tjong *et* al. Pour cette raison, le modèle a été qualifié de prédictif, puisqu'il permet de prédire le positionnement du nucléole, même lorsque le volume de celui-ci est modulé, pour simuler une baisse de la transcription par exemple. Ceci est très intéressant, vu l'importance du nucléole dans l'organisation de l'espace nucléaire. Cela dit, dans les deux modèles théoriques simulant le génome et son organisation, ce qui est bien modélisé est plutôt la structure à moyenne ou grande échelle de la chromatine.

Chapitre 3

MÉTHODOLOGIE D'ANALYSE DE LA PROXIMITÉ SPATIALE

3.1. Introduction

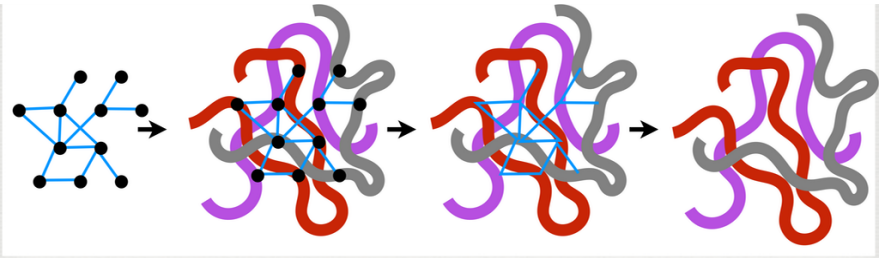

Figure 3.1. Représentation conceptuelle du but de l'analyse de données issue de méthodes de types 3C. Les données d'interactions sont identifiées aux régions chromosomiques auxquelles elles appartiennent, pour pouvoir, au final, reconstruire leur position relative dans la structure du génome.

L'hypothèse qui est faite lors de l'étude de la conformation des chromosomes est que l'organisation spatiale du génome comporte des caractéristiques structurales qui sont fonctionnelles. Ces caractéristiques peuvent en principe être dégagées par l'analyse des données issues d'une méthode dérivée du 3C (ou Hi-C) (voir figure 3.1).

Au cours des dernières années, plusieurs approches statistiques ont été développées afin de permettre d'analyser la quantité massive de données générée par ce type d'expérience.

La plupart de ces méthodes ont en fait ramené le problème à la question de la co-localisation, dans le noyau, de gènes (ou plus généralement, d'éléments codés dans l'ADN génomique) impliqués dans une même fonction, ou soumis à une même régulation. Alors que le terme *co-localisation* est couramment utilisé, celui-ci se confond à une proximité importante de sites dans l'espace, pouvant possiblement être dans un même complexe moléculaire. Cependant, dans le cas de

données 3C, la proximité peut servir à souligner une distance spatial simplement réduite, mais bien loin d'être dans un même complexe moléculaire. Nous croyons qu'une expression plus appropriée (minimisant l'ambiguité dans la signification) est donc celle de *proximité spatiale*. La question peut être reformulée de manière plus générale comme suit : est-ce que les gènes liés par une fonction biologique f donnée sont en proximité spatiale. Pour imaginer un cas simple, la fonction f sera par exemple le fait d'être ciblé par un même facteur de transcription.

Le but de ce chapitre est de présenter et de discuter quelques unes de ces méthodes, en particulier celles qui sont directement en lien avec les développements effectués durant cette maîtrise. Il sera également l'occasion d'introduire quelques notions et formalismes qui seront utiles pour la suite.

3.2. Formalisation du problème

3.2.1. Les données de type 3C : obtention, correction, normalisation

Typiquement, dans une expérience de type 3C, le génome est d'abord divisé en fragment. Ceux-ci résultent en fait de la digestion du génome par une enzyme de restriction. La taille des fragments va de quelques centaines de pairs de bases à quelques kb, avec une résolution de 10 kb au final. Dans un deuxième temps, les fréquences d'interaction (ou de cross-linking) entre paires de fragments sont mesurées expérimentalement. Cette fréquence est positivement corrélée avec la proximité spatiale des fragments en question. [1] par 3C.

À cause de biais technique de la détection d'interaction, certaines correction des données sont parfois nécessaire avant de procéder à leur analyse [Imakaev, Fudenberg, McCord, Naumova, Goloborodko, Lajoie, Dekker, and Mirny, 2012; Yaffe and Tanay, 2011].

3.2.2. Le graphe d'interactions

On peut formaliser le problème général en termes de graphe d'interactions. Chaque fragment du génome est alors représenté comme un sommet du graphe. L'ensemble des fragments (des sommets) est noté V. On notera $N = |V|$ le nombre total de fragments. Les sommets correspondant à des fragments en interaction sont alors reliés par une arête. L'ensemble des arêtes est noté E, et l'on note $G = (V, E)$ le graphe d'interactions.

1. Notons, qu'il est fort probable qu'une partie non négligeable des interactions ne soient pas détectées, et ce de manière non uniforme et dépendante de propriétés biophysiques des macromolécules. En effet, les interactions détectées dépendent des complexes que celles-ci forment à un moment, un endroit, ou une concentration donnée par exemple.

Puisqu'une interaction en deux sommers i et j est symétrique, le graphe est non-orienté. Formellement, E est donc un ensemble de paires $\{i, j\}$ d'éléments de V. Pour la suite, on définit, pour tout $(i, j) \in V^2$, $e_{ij} = 1$ si $\{i, j\} \in E$ et $e_{ij} = 0$ sinon. Pour des raisons pratiques, on définit $e_{ii} = 0$ pour tout $i \in V$.

Par ailleurs, dans certains cas, plutôt que de considérer que les fragments sont ou ne sont pas en interactions, on considère plus précisément la fréquence d'interaction normalisée m_{ij}, associée à la paire de fragments (i, j). Cette valeur peut être vue comme une pondération associée à l'arête (i, j) du graphe G, qui est donc dans ce cas un graphe pondéré.

On considère alors un ensemble de gènes liés par une certaine fonction. On note A le sous ensemble de V, représentant les fragments d'ADN associés à ces gènes, et $N_A = |A|$ le nombre de ces fragments. Dans la suite, pour simplifier, on parlera des gènes qui sont dans A. La structure d'interactions entre les gènes de A est alors donnée par le *sous-graphe de G induit par A*, que l'on notera $G_A = (A, E_A)$. Ici. E_A représente l'ensemble des paires $\{i, j\} \in E$ telles que $i \in A$ et $j \in A$.

La question fondamentale est de savoir si les gènes de A co-localisent. Ceci revient à se poser la question de savoir si le graphe G_A est significativement plus dense en interactions qu'attendu sous l'hypothèse nulle. Toute la question sera alors de spécifier correctement, d'une part, la mesure de la densité en interactions du sous-graphe G_A, et d'autre part, l'hypothèse nulle contre laquelle effectuer le test statistique.

3.3. Utilisation de la loi Hypergéométrique

L'approche présentée par Dai and Dai [2012] est basée sur l'utilisation de la loi hypergéométrique.

Le but est de tester si deux fragments (deux éléments) de l'ensemble A ont plus de chance d'être en interaction que deux éléments quelconques de V. Étant donné une paire $\{i, j\}$ de fragments d'ADN, on considère alors les deux propriétés suivantes :

— propriété 1 : les deux éléments de cette paire sont-ils tous deux dans A ?

— propriété 2 : cette paire est-elle en interaction (est-elle une arête de G) ?

Le test hypergéométrique cherche à tester si ces deux propriétés sont (positivement) corrélées. Par ailleurs, ce test est formulé directement au niveau des paires de fragments d'ADN. L'hypothèse nulle du test est donc que les deux propriétés sont indépendantes et identiquement distribuées sur l'ensemble des paires de fragments.

Pour calculer la valeur-p associée à ce test, on définit les nombres entiers suivant :

: $K = |E|$, le nombre total d'arêtes effectivement présentes dans le graphe G,

: $M = N(N-1)/2$, le nombre total de paires de sommets de G,

: $k =\mid E_A \mid$ le nombre total d'arêtes effectivement présentes dans le sous-graphe G_A induit par A,

: $m = N_A(N_A-1)/2$, le nombre total de paires de sommets du sous-graphe G_A.

Le test hypergéométrique revient à tirer aléatoirement (et sans remise) un ensemble de K parmi les M paires de sommets de V, puis à calculer le nombre X de paires $\{i, j\}$ ainsi tirées qui sont telles que $i \in A$ et $j \in A$. La valeur p est alors définie comme la probabilité que X soit au moins égal à k Pour obternir cette valeur p, la probabilité de tirer exactement x arêtes est d'abord calculée. Celle-ci correspond à

$$p(X = x) = \frac{\binom{m}{x}\binom{M-m}{K-x}}{\binom{M}{K}} \tag{3.3.1}$$

La probabilité calculée de tirer au moins k arêtes dans K est donc

$$1 - \sum_{x=0}^{k-1} p(X = x) \tag{3.3.2}$$

3.3.1. Critique du test hypergéométrique

Le test basé sur la loi hypergéométrique a été critiqué dans un article très important de Witten and Noble [2012]. Dans cet article est également introduit un test alternatif (voir ci-dessous). Le point fondamental de la critique de Witten et Noble est que le test hypergéométrique présente les failles suivantes : l'hypothèse nulle du test suppose que l'occurrence des deux propriétés mentionnées ci-dessus sont indépendantes entre elles, et identiquement distribuée (i.i.d.) à travers toutes les paires possibles. Hors, dans le cas présent, cette hypothèse est fondamentalement violée pour les deux propriétés.

Pour le voir, prenons trois éléments de V, a, b et c, et considérons les trois paires $\{a, b\}$, $\{b, c\}$ et $\{a, c\}$. Concernant la propriété 2, on voit que, si nous avons une arête entre les sommets a et b, indiquant que les deux fragments d'ADN correspondant sont très proches dans l'espace, et qu'il existe aussi une arête entre b et c, alors il est fort probable d'avoir également une arête entre a et c. (puisque a et c seront au plus à une distance $[d(a, b)+d(b, c)]$, si on imagine des distances dans un espace 3D conventionnel). Donc, si $\{a, b\}$ et $\{b, c\}$ sont dans E, $\{a, c\}$ a de fortes

chances d'être aussi dans E. Concernant la propriété 1, on voit immédiatement que, si elle est vérifiée pour $\{a, b\}$ et pour $\{b, c\}$, alors cela implique que a, b et c sont tous trois dans A, et donc, la propriété est forcément vérifiée aussi pour $\{a, c\}$.

L'effet observé est que les ensembles d'arêtes d'intérêt ont des sommets plus en grappe artificiellement que ceux de d'ensembles de même taille auxquels ils sont comparés. La distribution des p-valeurs de groupes de sommets tirés aléatoirement est alors non-uniforme, et ressemble à l'exemple supérieur droit de la figure 3.2.

3.4. Approche non paramétrique de ré-échantillonage, Witten and Noble [2012]

Afin de proposer une approche plus appropriée que celle utlisant la loi hypergéométrique, Witten and Noble [2012] ont proposé une méthode alternative, qualifiée de méthode non paramétrique de rééchantillonage. Cette méthode peut être justifiée par la remarque suivante. Le problème du test basé sur la loi hypergéométrique est qu'il revient en fait à randomiser les arêtes du graphe G, en gardant fixe le sous-ensemble de sommets A. Ce test revient donc à poser la question : le sous graphe de G induit par A contient-il significativement plus d'arêtes que les sous-graphe induits par A sur tous les graphes G' de même taille (de même nombre d'arêtes) que G ? Or, on voit ici que la question est posée dans le mauvais sens. En réalité, ce que l'on veut savoir, c'est si le sous-ensemble de gènes A est plus fortement interconnecté que les sous-ensembles de gènes quelconques et de même taille. Autrement dit, on veut savoir si le sous graphe de G induit par A contient significativement plus d'arêtes que les sous-graphes de G induits par des ensembles aléatoires de sommets de même taille que A. Il ne faut donc pas randomiser le graphe G mais l'ensemble A.

Dans cette direction, la méthode de Witten et Noble propose un calcul de la valeur-p du test par échantillonnage de Monte Carlo. Cette approche sera également utilisée dans le travail présenté dans le chapitre suivant. La démarche générale consiste à tirer des sous-ensemble de V de même taille que A (et contrôlés de diverses manières) puis à recalculer la densité d'interaction des sous-graphes induits par chacun de ces sous-ensembles de sommets. La valeur p est alors simplement la fraction des simulations pour lesquelles la densité d'interaction obtenue sur la configuration simulée est supérieure à celle observée sur la configuration réelle induite par A.

Dans le test de Witten et Noble, seules les interactions interchromosomiques sont considérées. Étant donné le sous-ensemble A de V, on note n_i, $i = 1..I$, le

nombre d'éléments de A situés sur le chromosome i. On a donc que $\sum_i n_i = N_A$, le nombre total d'éléments de A. Par ailleurs, on définit

- x, le nombre d'arêtes dans G_A correspondant à des interactions interchromosomiques.
- s, le nombre total de paires d'éléments de A présents sur deux chromosomes distincts (le nombre total d'interactions interchromosomiques possibles entre éléments de A).

La statistique utilisée par le test est définie par le rapport x/s. Pour estimer la valeur-p, la procédure Monte Carlo suivante est alors répétée en boucle B fois, où B est un nombre relativement grand (ex :1000). Nous avons donc pour $b = 1 \dots B$:

- pour $i = 1 \dots I$, tirer aléatoirement et uniformément n_i fragments du chromosome i sans remise, et former ainsi un sous-ensemble W_b de même taille et de même composition chromosomique que A.
- calculer x_b , le nombre d'arêtes interchromosomique dans le sous-graphe de G induit par W_b
- Calculer s_b , le nombre de paires interchromosomiques dans le sous-ensemble W_b.

L'estimée Monte-Carlo de la valeur-p du test est alors égale à :

$$\frac{1}{B}\sum_{b=1}^{B} I\left(x_b/s_b \geq x/s\right) \tag{3.4.1}$$

Les auteurs ont montré les problèmes du test hypergéométrique, et le bien-fondé de leur approche, en appliquant les deux tests sur 1000 sous-ensembles de départ défini aléatoirement. Dans le cas du test de rééchantillonnage, la p-valeur est bien uniformément distribuée entre 0 et 1, contrairement au cas du test hypergéométrique..

3.4.1. Résultat et conclusion de l'emploi de Witten and Noble [2012]

L'approche de Witten and Noble [2012] ne se prête qu'à l'interprétation des interactions interchromosomiquees, qui sont en fait nettement moins abondantes (10 %) que les interactions intrachromosomiquees (90 %). Aussi, en utilisant cette approche, aucune évidence de la proximité spatiale des cibles de facteurs de transcription n'a été détectée.

3.5. Approche non-paramétrique considérant la structure du génome, Paulsen et al. [2013]

3.5.1. Normalization et randomization préservant la distance génomique entre loci

L'équipe du Dr Eiving Hovig a apporté une contribution majeur dans l'approche qu'ils ont introduite[Paulsen et al., 2013]. Leur approche permet de pouvoir correctement traiter les données d'une manière avertie des propriétés du génome. En effet, le traitement des donnée prends en comptes des propriétés récemment proposées et établies du génome à partir de données 3C[Imakaev et al., 2012], élucidant certains facteurs dictant la conformation de la chromatine. L'approche prends aussi en compte l'ordre séquentiel des fragments d'ADN, ce qui est essentiel tel que nous l'avons défini dans la description de la pertinance d'une information sur l'organisation spatiale du génome. Cela dit, leur approche a été développé sur des données Hi-C (variante du 3C) sur le génome humain.

Dans cette approche, les données 3C sont utilisées de manière quantitative : on part des données de fréquences d'interactions, et l'on note m_{ij} la fréquence d'interaction entre les fragments i et j. On note également :

- : c_i le chromosome auquel appartient le fragment i
- : x_i, la position chromosomique du fragment i, définie par la paire de base marquant le début de la séquence d'ADN du fragment
- : $\delta_{ij} = |x_j - x_i|$ la distance linéaire entre les fragments i et j ($\delta_{ij} = +\infty$ pour des fragments appartenant à deux chromosomes différents)

On peut alors calculer la fréquence d'interaction moyenne entre paires de fragments situés sur un même chromosome, et à une distance k l'un de l'autre, qui sera notée $\hat{E}[m \mid \delta = k]$, ainsi que la déviation standard associée : $\hat{sd}[m \mid \delta = k]$. Ceci nous permet alors de définir m^*_{ij}, la valeur normalisé associe à la paire de fragments i, j, comme suit :

$$m^*_{ij} = \frac{e_{ij} - E[m|\delta(e_{ij})]}{std(m|\delta(e_{ij}))} \tag{3.5.1}$$

Étant donné un ensemble de gènes A, la statistique test est défini comme étant la valeur moyenne sur l'ensemble des pairs de A de la fréquence normalisée m^* :

$$t = \frac{1}{|E_A|} \sum_{e_{ij} \subseteq E_A} e^*_{ij} \tag{3.5.2}$$

Une fois cette statistique calculée sur les données réelles, la p-valeur du test est ensuite estimée en ayant recours à un test de permutation. Afin de tenir compte

de la configuration des gènes, notamment le long des chromosomes. les auteurs proposent de randomiser le sous-ensemble (A) d'intérêt de manière à préserver l'ensemble des distances consécutives entre gènes d'un même chromosome. Ceci est accompli en permutant la séquence des distances entre fragments successifs le long du chromosome. De cette façon, la distribution des distances consécutives est strictement conservée entre l'ensemble original A et les répliques de Monte Carlo. Ceci a pour résultat la méthode de randomization de Monte Carlo nommée Conserved Consécutive Distances (CCD). La procédure est répétée B fois, et la p-valeur est alors définie par :

$$p = \frac{\left[\sum_{b=1}^{B} I(t_b \geq t_{obs})\right] + 1}{B + 1} \tag{3.5.3}$$

3.5.2. Randomization par compartiment génomique

Paulsen et al. [2013] ont aussi fait une addition importante à leur méthode dans le but de prendre en considération la structure du génome. Une première observation qu'ils ont considéré est le fait que certaines régions des chromosomes, telles que les extrémités ou les régions centromériques, ont tendance à interagir davantage avec elles-même qu'avec le reste du génome. Les auteurs proposent donc de diviser les bras de chaque chromosome en 6 groupes, et d'effectuer la randomization CCD à l'intérieur de ces groupes. Une deuxième observation considérée a été celle selon laquelle les régions de chromatine ouverte tendent à avoir davantage d'interactions entre elles,et pareillement pour les régions de chromatine fermée. Paulsen et al. [2013] ont donc choisi de délimiter le génome, humain dans ce cas-ci, selon que l'état de la chromatine est ouverte ou fermée, et d'ensuite effectuer la randomization CCD à l'intérieur des ces limites. Cette dernière définition de limite est faite de manière alternative, et non additive, à la segmentation en six partie des chromosomes.

3.5.3. résultat et conclusion de l'emploi de l'approche de Paulsen et al. [2013]

La méthode présentée par Paulsen et al. [2013] a effectivement pris en considération les éléments dictant une majeure partie des interactions ADN-ADN, soit la compartimentalisation en chromatine ouverte ou fermée, et la position relative des gènes sur le chromosome, ce qui est en accord avec les résultats de Imakaev et al. [2012]. Leur méthode, sous l'hypothèse nulle, produit bien une distribution de p-valeurs uniforme (Figure 3.2), même avec des ensembles de fragments aléatoires préférentiellement choisis pour être en grappe (distance linéaire). De plus, leur approche a permis de confirmer qu'effectivement, en randomisant par région,

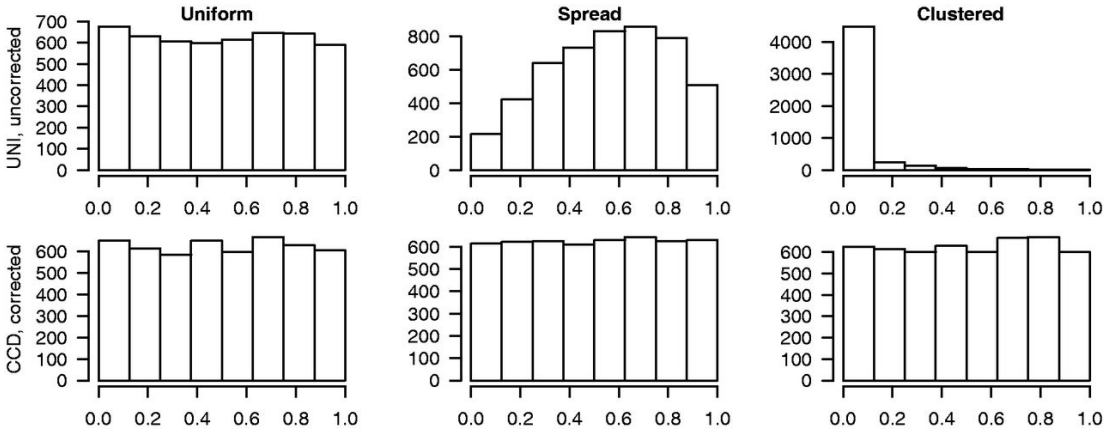

Figure 3.2. Distribution des pValeurs selon la méthode de randomisation. (figure tirée de Paulsen et al. [2013]. Chaque graphique montre l'histogramme de 5000 P-valeurs trouvées en effectuant des tests d'hypothèses sur les structures 3D simulées (basé sur une procédure de marche aléatoire, comme expliqué dans la section « Matériels et Méthodes »de Paulsen et al. [2013]) où l'hypothèse nulle est vraie. Les tests sont effectués sur les interactions intra et inter -chromosomiques simultanément. La rangée supérieure affiche l'approche la moins complexe, en utilisant la stratégie de ré-échantillonnage Monte Carlo UNI (uniforme) et la statistique de test non corrigée. La rangée du bas montre les résultats de la Monte Carlo stratégie de ré-échantillonnage CCD et la statistique de test corrigée de l'équation 3.5.1.Les trois colonnes représentent les trois configurations différentes dans la formation aléatoire des ensembles de fragments, répartie uniformément (colonne de gauche), éparpillée (colonne du milieu) et en grappe (colonne de droite). (note : Ceci est une traduction automatisée/corrigée du texte originale)

les télomères et centromères sont plus proches entre eux, par rapport au reste des autres régions. Le résultat est observé par le fait que la randomisation par compartiment réduit sensiblement la p-valeur de ces deux groupes fonctionnels par rapport à une randomisation globale.

3.6. Méthodes alternatives

3.6.1. Approche de randomisation stricte du graphe d'interactions, Kruse, Sewitz, and Babu [2013]

Une méthode qui trouve particulièrement ses racines dans la théorie des graphes, mais qui reste discutable a été introduite par Kruse et al. [2013]. Brièvement, la méthode éxécute la randomisation en considérant les données purement comme un graphe. Par la suite, pour tenter de préserver les propriétés initiales, des concepts tels que la transitivité définie d'une manière spécifique interviennent. Cela dit, il n'y a aucune notion de position relative au chromosome ou d'échelle qui intervient. Ceci est pourtant critique pour la question présente, et nécessaire.

3.6.2. Etude précédente de la colocalisation des cibles de facteurs de transcription, Ben-Elazar, Yakhini, and Yanai [2013]

Bien que Ben-Elazar et al. [2013] ont présenté des résultats sur la question de l'intitulé, il y a des points critiques qui ne sont pas suffisamment considérés. Notamment, la randomisation dans leur cas pour obtenir une distribution contrôle est faite sans même préserver le nombre de gènes par chromosome, et sans aucune considération du type de région (position relative au chromosome).

3.7. CONCLUSION DES MÉTHODES D'ANALYSE DE DONNÉES 3C

La méthode que nous trouvons la plus performante et que nous prendrons comme référence est donc celle de Paulsen et al. [2013]. Cela dit, le problème majeur est que, dans le cas de la levure, tous les éléments démontrés dans ce cas-ci sont aussi récapitulés dans les modèles 3D. C'est dire que la statistique mesurée donne une information à grande échelle, tout comme les modèles théoriques du génome. Cependant, afin de pouvoir investiguer les différences entre l'organisation du génome qui résulte de la biophysique du génome, de contraintes physiques à grande échelle, a ceux qui résultent potentiellement d'une active biochimique nucléaire, il est nécessaire de pouvoir apprécier les choses à une échelle réduite et plus précise. Cette importance critique de l'échelle devient de plus en plus reconnue dans le domaine [Schmitt et al., 2016]. Le cas échéant, il deviendra donc nécessaire d'être dans l'ordre de grandeur de la résolution des données 3C, afin d'en exploiter le potentiel autant que possible. Ceci est un point important, car sans cela, l'utilisation de technique 3C pourrait être qualifiée d'obsolète, si l'information qui en est tirée est récapitulée *in silico*.

Chapitre 4

PROGRESSIVE SCALING OF HI-C CORRECTION REVEALS PRINCIPLES OF TRANSCRIPTION-DEPENDENT AND FUNCTIONAL GENOME 3D ORGANIZATION

4.1. ABSTRACT

Back ground :

Early in this decade, notable work has allowed presenting whole genome structural information at unprecedented resolution. This was achieved for various species, including human, yeast, mouse, Drosophila, and others model organism. First done on an experimental basis, there was effort to obtain this insight *in silico*, with a purely computational model [Kimura et al., 2013; Tjong et al., 2012; Wong et al., 2012]. In yeast, one of the most frequently and extensively studied model organism, theoretical models have recapitulated many of the features of the experimentally derived genome structure organization. These models are based on large-scale physical constraints and biophysical properties of DNA and chromatin. However, the extent to which biological activity dictates yeast genome structural features has not been clearly delimited .

Results

We have developed a method which allows contrasting better the significance of DNA contacts potentially resulting from active nuclear processes. The method is a progressive scaling of interaction data correction and is region based. The approach focuses on locality and minimizes the effect of large scale physical constraints, biophysical properties of DNA and chromatin on the 3C derived DNA contact map (Hi-C). Finally, the spatial proximity significance profiles (5kb to 70kb inquired range) of TFs functional target presents evidences that genomic

spatial clustering events occur at different scales, and in a contact type dependent manner (intra-chromosomal vs. inter-chromosomal contact type). Also, the provided pieces of evidence allow drawing a model where TFs target genes distribution drives genome structure, modulating the structures arising from large-scale physical constraints and biophysical properties of DNA. Altogether, this work has shed new light on the basis of functional genome 3D features and its interplay with nuclear processes.

4.2. Introduction

What defines a cell's identity is its DNA, which contains all the required information to produce both RNA and proteins. Thus, the regulation of cellular activity first and necessarily begins at regulating DNA transcription. Currently, awareness of genome spatial conformation is deeply redefining our understanding of genetics [Bickmore and van Steensel, 2013; Cavalli and Misteli, 2013]. What it underlines is the fact that beyond genome sequence, which has been central to understand biology, the spatial organization of the genome is also essential to understand genome function. In this context, pioneering and outstanding work done by Duan et al. [2010] presented the first genome-wide map of DNA contact in yeast *Saccharomyces cerevisiæ* at a 10 kb resolution. An aspect that has followed genome-wide DNA interaction data is the three-dimensional (3D) modeling of the genome. Such model allows picturing the genome structure captured by DNA-DNA contacts. Resolving the underlying mechanism regulating genome spatial organization has become a fundamental question. Just as decoding the DNA sequence of organism has provided invaluable information for understanding fundamental basis in biology, decoding genome structure and the basis of its regulation will most likely provide essential information to a cell's biology. This will, in turn, shed new light on the genetic basis of diseases and yield valuable information on how to address these [Bickmore and van Steensel, 2013; Cavalli and Misteli, 2013].

First, we favor the view that dynamic genome structure regulation is driven by nuclear processes and mediated by the proteins and RNA. In fact, increasing evidence support the view that proteins engaged in the regulation of transcription are also implicated in building genome architecture Nott et al. [2016]. However, the assumption that 3C data captures this information in yeast could be challenged by the fact that theoretical models reproduced the hallmarks of yeast genome structures reported by Duan et al. [2010] (see Figure 4.1). These models were based on principles such as volume exclusion, chromosome tethering, and proper modeling of DNA as a polymer [Kimura et al., 2013; Tjong et al., 2012; Wong

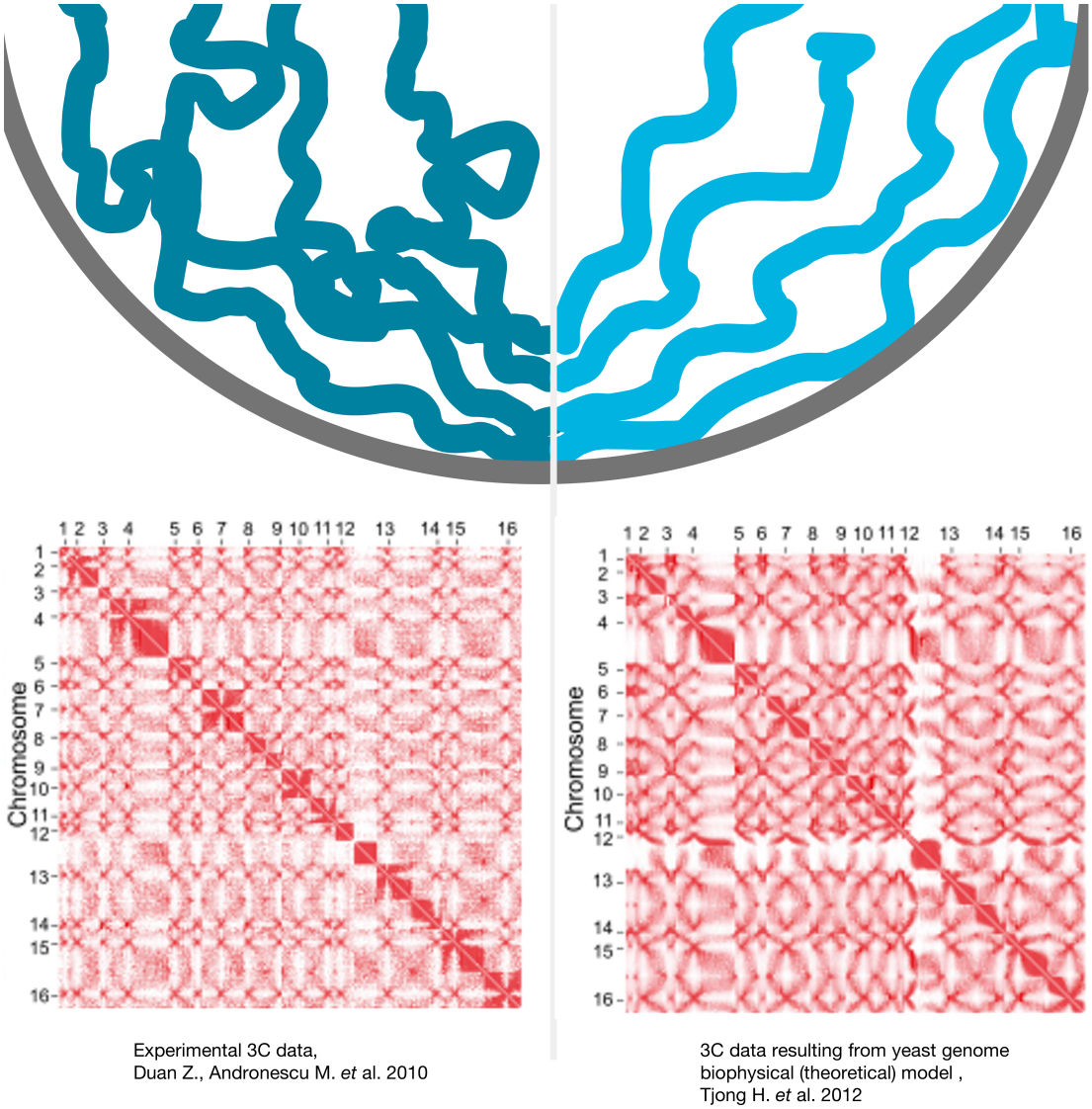

Figure 4.1. Comparison of experimental and theoretical yeast 3C data and hypothesis. The DNA contact map represented share many features [Tjong et al., 2012]. The possible structures beeing captured by the experimental data and theoretical model share many features. Some of the driving force of genome spatial organization, recapitulated in the theoretical model, are chromosome tethering to spindlepole bodies, chromosome volume exclusion, nuclear enveloppe volume constraint and DNA polymer biophysical properties. While large scale features are shared, at higher resolution (lower scale), experimental data might yield information to structural features resulting from nuclear processes regulation (e.g. transcriptional regulation), potentially introducing more variability in the genome structure compared to the theoretical model. So far, there aren't clear indication as of which additional functional features experimental data yield.

et al., 2012]. The hallmarks included evidence for the chromosomal territories, telomere positioning, Rabl like conformation, interaction depletion between chromosome arms (compared to within arms), spatial proximity of functionally associated genomic loci including early and late replication origins and tRNA genes.

The lesson drew from these observations is that large scale physical constraints and physical properties of DNA and chromatin account for much of the DNA contacts landscape. Should biochemically inferred interactions occur, it would, therefore, be necessary to account for these biophysical factors. One way to do this is to compute a significance for an interaction in a way that considers these factors. To do so, we have developed the idea of a region-based approach, using the interaction distribution between defined regions implicated in an interaction and normalizing a given interaction based on that distribution. Thus, the contact distribution between the two regions targeted would reflect the resulting conformation due to large scale physical constraints. Interactions which are resulting from active protein or RNA complexes regulation would, in theory, be better exposed, and be stand out as significant. Given that spatiotemporal coordination of molecules is inherent to cells activity regulation, we have developed an algorithm to assess the spatial proximity of a given set of loci using 3C data. Essentially what is calculated is the density of detected interaction within a set of targeted loci. This measure is then used to assess the significance of colocalization (Figure 4.2). The density is calculated from the DNA-DNA interaction data(after normalization). The fact that the algorithm results are calculated in simple and straight forward way makes the analysis very transparent and easy to interpret, as mentioned later (Figures 4.2 - 4.4). It also allows the algorithm assessing colocalization to be fast, and be computed with a large number of parameters to run numerous calculation as mentioned in following sections.

Given the role of transcription factors (TFs) in regulating transcription and in bridging DNA to protein complexes, we have first sought to assess the colocalization of transcription factors (TFs) target genes. On this question, work by Witten and Noble [2012] has established the importance when randomizing the interaction network (or a set of loci of interest) to preserve the number of loci per chromosome when doing so. Furthermore, Paulsen et al. [2013] has shown that it is important to preserve also the relative linear distance between loci, as there is an inversely proportional relation between linear distance and the probability of interaction.The essentially results from the biophysical properties of DNA as a polymer. The two principles mentioned afore relate to intra-chromosomal (within the same chromosome) interaction, but no principle address inter-chromosomal interactions (across different chromosomes). One of the differences between intra-chomosomal and inter-chromosomal interactions is that even at great distance, DNA contacts for intra-chromosomal interactions are much more abundant than

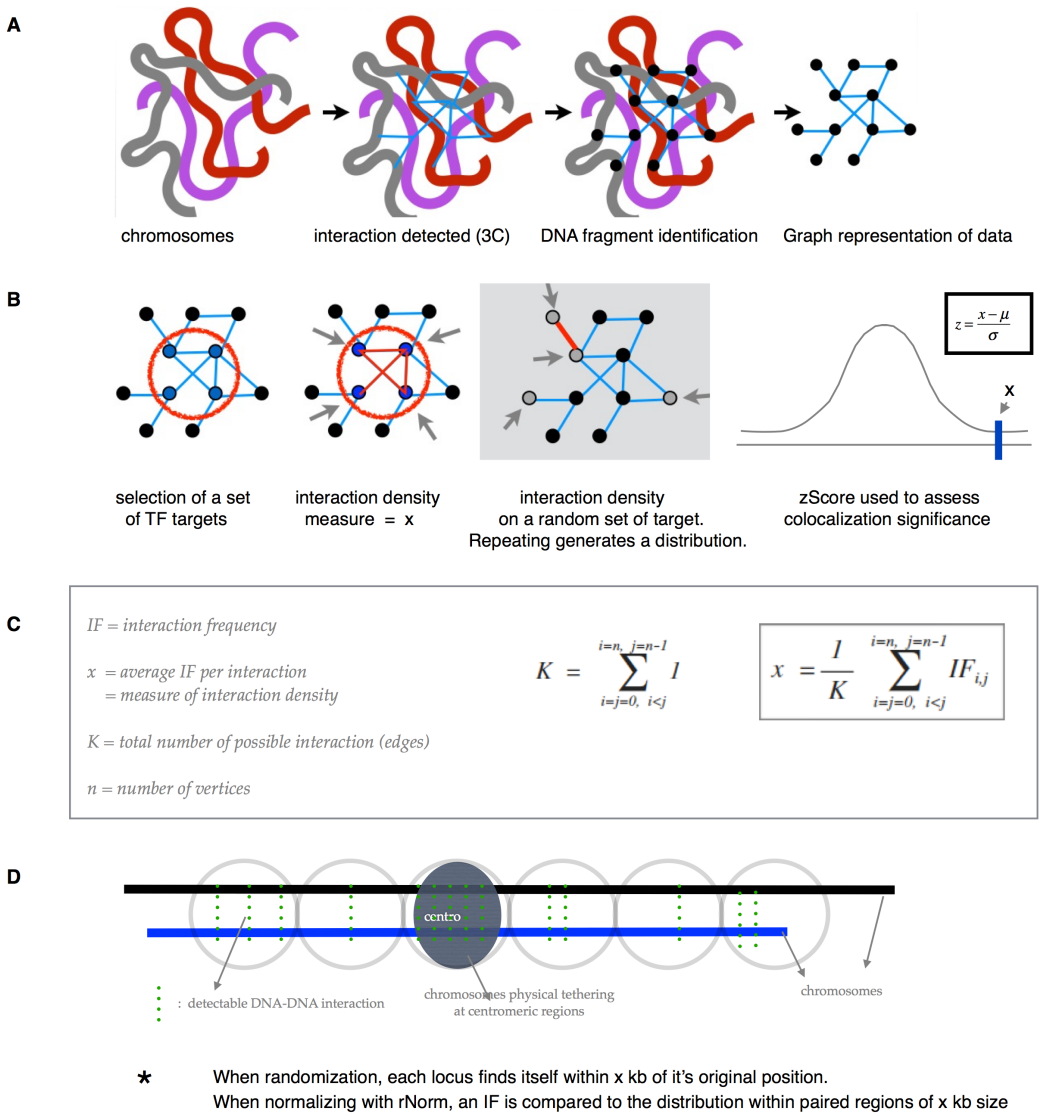

Figure 4.2. Representation of the approach used to assess colocalization significance (A) From chromosomes organization, to graph representation of DNA contacts captured (B) Schema of colocalization assessment (C) Calculation details. (D) Region based randomization and normalization representation

inter-chromosomal interactions. This is in agreement with the observation of chromosomal territories by microscopy [Hübner and Spector, 2010]. It, therefore, becomes important to determine to which type of interaction the colocalization significance of gene can be attributed. It is also important to remember that the current yeast 3C data are an average over a cell population. This means colocalization which is meaningful might be harder to capture if the colocalization the event occurs in a small fraction of the population. To the best of our knowledge,

there are no single-cell 3C experiments reported yet for yeast, as it is the case for the human genome [Nagano, Lubling, Stevens, Schoenfelder, Yaffe, Dean, Laue, Tanay, and Fraser, 2013]. In accordance with Paulsen et al. [2013], our work is based on calculating the average value of an interaction within the set of loci as a measure from which to estimate spatial proximity significance.

Work by Yaffe and Tanay [2011] highlighted the importance of correcting genome-wide chromosome conformation capture data, studying human genome. Imakaev et al. [2012] and Hu, Deng, Selvaraj, Qin, Ren, and Liu [2012] presented general methods for correcting genome-wide 3C data. More recently, Ay, Bailey, and Noble [2014] reported a method (Fit-Hi-C) "that assigns statistical confidence estimates to mid-range intra-chromosomal contacts by jointly modeling the random polymer looping effect and previously observed technical biases in Hi-C datasets" .Work of shared interest to ours, but using a different approach has been reported by Ben-Elazar et al. [2013] . The conclusions of this work indicate that indeed, a certain fraction of yeast transcription factor (64/174) show significant of spatial proximity of their functional targets.

Given the considerations addressed above, particularly to preserve a comparable distance between loci when selecting a random set, we have used a region based randomization and normalization (rNorm). Under this setting, features that arises from large physical constraints are in theory less prevalent. While other correction procedures are useful for correcting methodological and technical biases, the interest in a region based normalization is the ability to contrast better contact patterns that are resulting from active genome 3D functional reorganization (by proteins and RNA), and those resulting from large-scale physical constraints and DNA biophysical properties. A critical difference between the method presented by Paulsen et al. [2013] and rNorm is that with rNorm, interaction frequencies are normalized using the regions of interest put in contact each time. The method presented by Paulsen et al. [2013] normalizes interaction frequencies using the set of interactions having equivalent linear distance between genomic element put in contact. This, however, does not even out the interaction specific landscapes between the regions put in contact. Thus, region based normalization works in a manner that is inherently aware of genome structural properties.

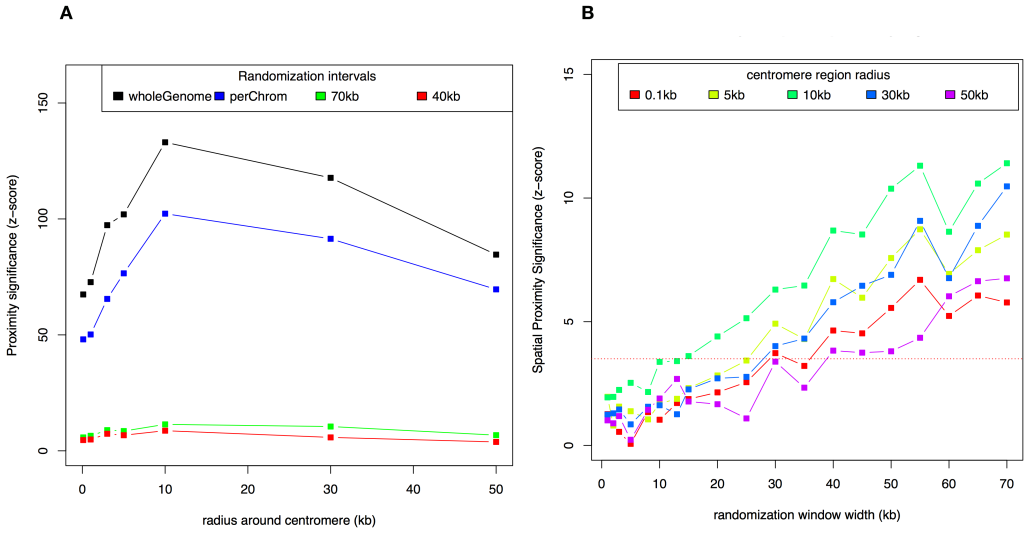

Figure 4.3. Randomization Window Width Effect on centromere colocalization significance. The radius corresponds to the distance encompassing the centromere on either side of a the centromeric sequence. The 3C dataset used is the high-confidence (1 % false discovery rate) filtered inter-chromosomal contact subset (see Materials and Methods). The region based randomization procedure was performed with a 1000 permutations. (A) Spatial proximity (colocalization) significance of centromeres under 4 randomization window width (wholeGenome, perChrom, 70kb, 40kb). (B) Randomization Window Width Effect on centromere colocalization signal, ranging from 1 to 70 kb.

4.3. Results

4.3.1. Positive controls of the region based randomization scheme

In order to assess the accuracy of the approach, we have first tested if DNA regions expected to colocalize gave a significant colocalization signal. The dataset used is the high confidence filtered (1 % false discovery rate) from Duan et al. [2010]. Centromeres, telomeres, and tRNA were chosen as control sets, as did Duan et al. [2010]. For each of these controls, significant spatial proximity was detected (Figures 4.3, 4.4). The spatial proximity significance was measured using genome-wide, per chromosome, and region based randomization approaches (i.e. random shuffling per region). Because centromeric sequences are very short sequences, we have performed the spatial proximity test by defining as centromeric regions at increasing linear distance (radius) from the centromere sequence. Intra-chromosomal interactions are not of primary interest because centromeres are located on different chromosome each. The resulting colocalization signal of

centromeres is extremely significant (z-score > 100) when permuting the fragments genome-wide, and still high when permuting DNA fragment per chromosome (Figure 4.4). However, when permuting fragment position per region (70kb or 35kb, 4.4), the spatial proximity signal drops drastically from close to 120 of to less than 10. This result strongly suggests and gives a sense that indeed, in the most extreme cases such as centromeres, the approach per region is capable of removing signal which can be attributed to large-scale physical constraints. Another observation is that the spatial proximity signal seems to peek at 10kb of linear radius when defining centromeric regions (Figure 4.3). This measure has been kept as reference for defining centromeric regions as a control set. Using the complete 3C dataset (intra + inter -chromosomal contacts) (Figure 4.4) reduces the signal nearly by half compared to using the inter-chromosomal contact subset. The effect observed is that intra-chromosomal contacts behave very differently from inter-chromosomal contacts and that using the complete 3C dataset (intra+inter) may mask or reduce the measured characteristic. As represented in the figure, intra-chromosomal contacts alone do not give any signal at smaller kb radius (10kb or less) for defining centromeric region, because the high-confidence 3C dataset filters out interactions between loci that are less than 20kb apart.

This is not the case with the unfiltered raw data (Figure 4.4), under which centromeres do give a spatial proximity signal when using intra-chromosomal contact subset, due to the presence of sub-20kb intra-chromosomal centromeric interactions. There is, therefore, a use for interactions present in the raw dataset.

Likewise, we have assessed the spatial proximity significance of telomeres (defined in [Duan et al., 2010]), and have detected significant spatial proximity (Figure 4.4). The significance was observed under genome-wide, per chromosome, and per region permutation of fragments. We did notice however that when using the intra-chromosomal contact subset, per region permutation gave a stronger signal compared to genome-wide or per chromosome permutation. This suggests an enrichment of interactions within telomeric regions which becomes apparent when contrasting locale interaction landscape but do not stand out when including the rest of the genome. Moreover, using both intra and inter -chromosomal contacts together did not reduce as much the spatial proximity signal when the permutation is done per region, compared to the inter-chromosomal contact subset alone. This is most likely due to the fact that intra-chromosomal interactions display a trend of colocalization which agrees with that of inter-chromosomal interactions.

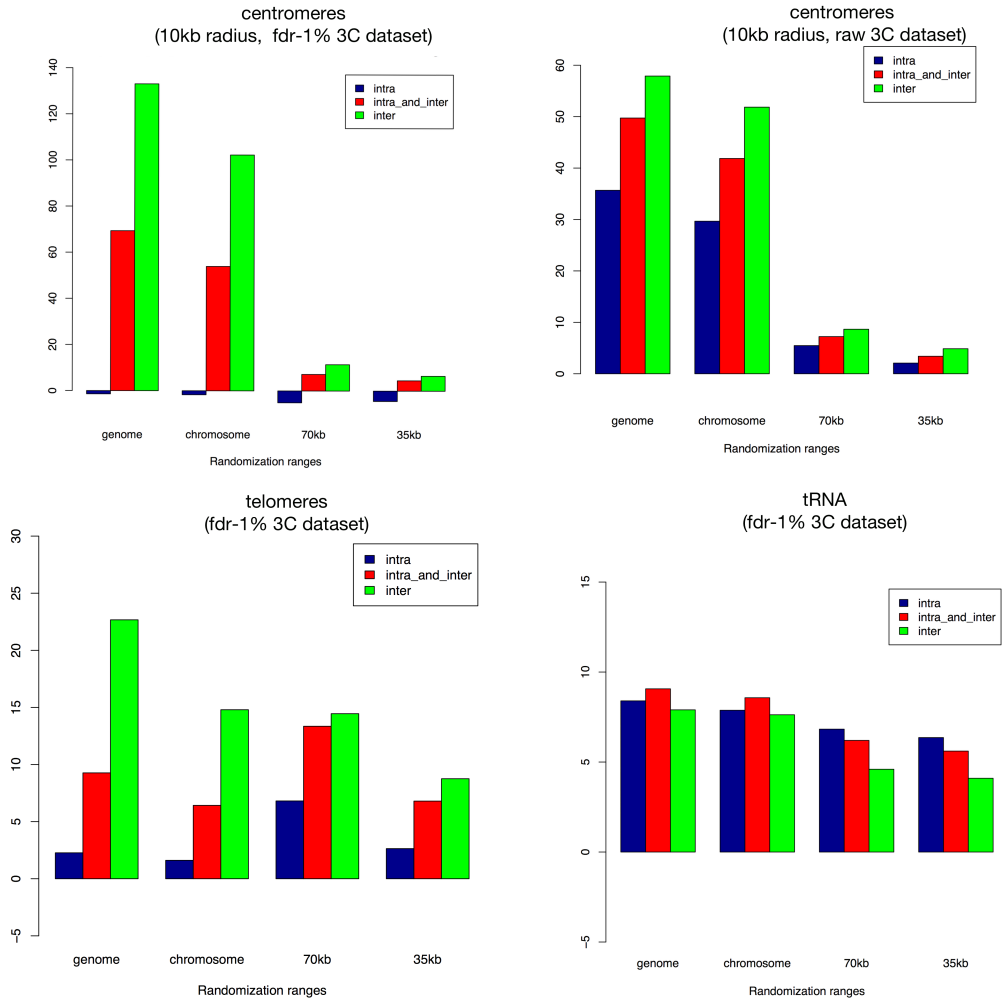

Figure 4.4. Effect of region based randomization on spatial proximity significance of positive control for colocalization. The 3C dataset used are either the high-confidence filtered (false discovery rate (FDR)-1 %) or the unfiltered (raw) dataset. In each case, spatial proximity significance of a positive control set is assessed using genome, per chromosome, randomization of loci position, and per region (70kb, 35kb region) randomization. The goal is to verify that region based randomization accounts for large-scale physical constraints such as the tethering of telomere in evaluating spatial proximity significance. (in the case of centromere, lower signal in region based is better).

For the tRNAs control set, the spatial proximity significance was also detected.The significance was lesser than centromeres' or telomeres', but was displaying less extreme variation across the 4 randomization settings. This behavior is different from that of the two set of positive controls previously shown. This suggests that at various scales, genome-wide, per chromosome, per region, the genome structures facilitates tRNAs spatial proximity. Also, for tRNAs, the intra-chromosomal dependent spatial proximity signal is as strong or greater than the inter-chromosomal. Still, under region based randomization, using both intra

and inter -chromosomal contacts does not yield a stronger signal than intra-chromosomal contacts simply. This demonstrates again how spatial proximity signal can be better captured when using a specific contact type of the 3C data.

In order to have a better sense of the effect of assessing locally genomic features using region based randomization, we have assessed the spatial proximity of centromere using inter-chromosomal contact subset with randomization window width ranging from 1kb to 70 kb, at 5 kb interval step or less (Figure 4.3B). What we observe, as expected intuitively, is that the spatial proximity significance decreases as the randomization scale is decreased. From these results, 20kb was set a reference scale for region based randomization for the following reason : 1. Given that the resolution of the [Duan et al., 2010] 3C dataset is said to be 10kb, then 20kb of randomization window seems to be extremely stringent and using less than twice the resolution before hand would seem excessive 2. At this width, only the 10kb defined centromeric regions (linear radius) give a significant signal, the significance cut-off being set at z-score = 3.5.

4.3.2. Assesment of the spatial proximity of yeast transcription factors functional targets

We have next assessed spatial proximity of transcription factors (TFs) functional target genes using this procedure. TFs were counted as significant if under any of intra-chromosomal, inter-chromosomal or both (intra + inter -chromosomal) datasets, their functional target gave a spatial proximity signal beyond cut-off. Using 5 % false error rate, and Bonferroni correction for multiple testing, the cut-off was set at a z-score value of 3.5. About a third (60) of the 174 TFs showed signal beyond the significance cut-off. Using the same strategy of permutation at 20kb window width, random sets we generated to see that indeed these TFs counts did not exceed error rate of 5 % at this cut-off. Indeed, the count of false positive TFs was 0. (Using an empirical false discovery rate of 1 % [z-score = 3.0] rather than the stringent criteria of Bonferroni correction actually gives 85 (49 %) TFs showing significant signal).

When inspecting in details the number of TFs count per contact type, we see that using inter-chromosomal contacts or the complete dataset (intra+inter -chromosomal contact, labeled as “both” in Figure 4.5) yields higher TFs count than intra-chromosomal contacts alone. We have also counted the number of TFs that required using a subset of the 3C contacts (intra-chromosomal only (unique), or inter-chromosomal only), but could not be captured using the complete 3C dataset (intra- and inter-chromosomal contacts). The number of inter-chromosomal unique TFs was 11, and none for intra-chromosomal unique. This suggests , that

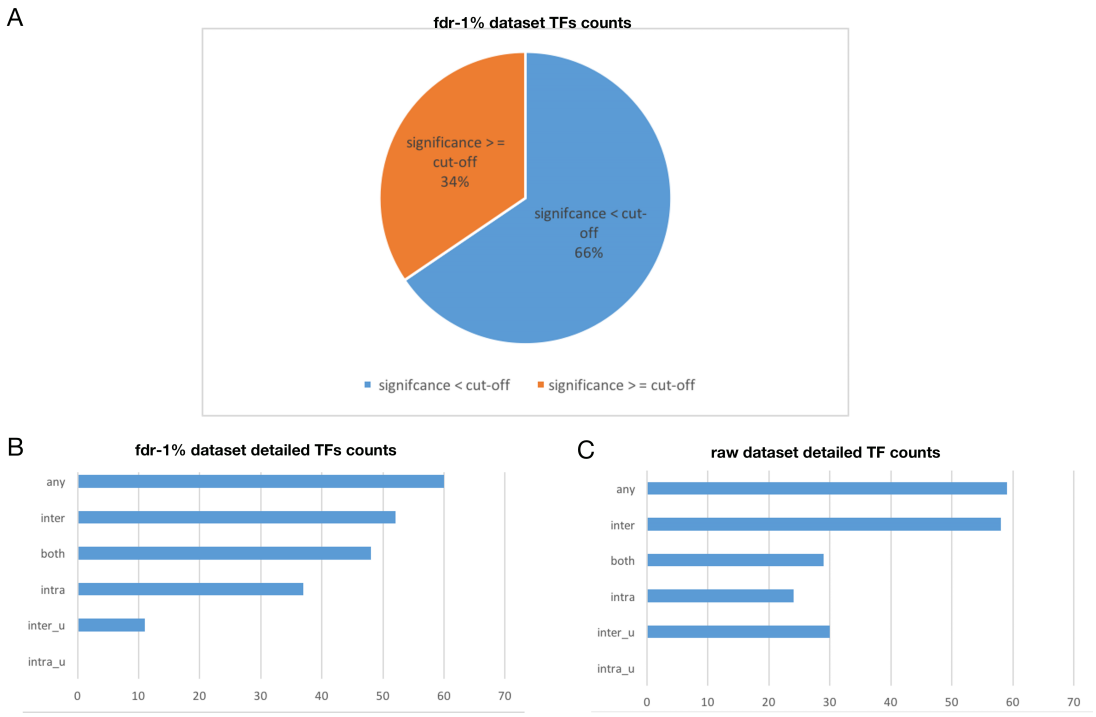

Figure 4.5. Counts of transcription factors (TFs) having functional targets in spatial proximity. There is a total of 174 TFs. The randomization window width used is 20kb, according to the region based randomization procedure, with a 1000 permutation.The 3C dataset used are either the high-confidence filtered (false discovery rate (FDR)-1 %) or the unfiltered (raw) dataset. The significance cut-off for spatial proximity is set at a z-score of 3.5 (5 % error rate, with Bonferroni correction for multiple testing). (A) Counts of TFs who's functional targets spatial proximity significance exceeds the cut-off. These will be referred to simply as counts of significant TFs in this figure. (B,C). Detailed counts of significant TFs. There are 6 counts categories as follows : 1.both, which corresponds to counts of TFs deemed significant when using the complete 3C dataset intra-chromosomal + inter-chromosomal contacts. 2. intra, which corresponds to counts of TFs deemed significant when using the intra-chromosomal subset of the 3C data. 3. inter, which corresponds to counts of TFs deemed significant when using the intra-chromosomal subset of the 3C data. 4. any,which corresponds to counts of TFs deemed significant under either of both, intra, or inter 3C data setting. 5. intra_u(intra_unique), which corresponds to counts of TFs deemed significant only when using the intra-chromosomal subset of the 3C data, but not the complete dataset (both). 6. inter_u(inter-unique), which corresponds to counts of TFs deemed significant only when using the inter-chromosomal subset of the 3C data, but not the complete dataset (both).

when using the complete 3C interaction dataset, inter-chromosomal dependent signal is more likely to be dimmed or drowned, whereas intra-chromosomal dependent remains as evident or is enhanced. This observation can also be made on the controls used in Figure 4.4.

After these observations, we sought to see if spatial proximity of TFs functional targets would reveal equivalent insight (Figure 4.5C) when applying region based randomization with the unfiltered (raw) dataset , which is considered excessively

noisy, rather than the high confidence set of contacts (FDR-1 %) from Duan et al. [2010]. We have observed a similar number of TFs (59), to be significant, and 58 to already be captured using inter-chromosomal contacts alone. However, this result would not have been possible to obtain using the complete raw dataset. Half (30 TFs) of these counts are visible when considering inter-chromosomal contacts only ("inter-chromosomal unique", in Figure 4.5) . These results obtained with the raw dataset are surprising, considering that previous analysis of TFs targets colocalization in yeast required filtering for highly significant contacts Ben-Elazar et al. [2013]. Nevertheless, using intra-chromosomal contacts subset, the TFs count is of 24, and 29 using the complete raw dataset. In this case, intra-chromosomal data subset and the complete data present a 35 % decrease of counts compared to high confidence dataset. This indicates that with the raw data, the noise does affect or eclipse intra-chromosomal based patterns. Another important observation of these counts is that inter-chromosomal contacts detected, even with the unfiltered raw dataset, are more reliable (stable). This is exemplified by the fact that raw inter-chromosomal TFs counts are of 58, vs 52 for high confidence filtered dataset.The increase is most probably due to the fact that the filtering in the high-confidence removes more useful contact information than noise.

We have observed a strong relation between the spatial proximity significance of TFs and the number of fragments targeted by a TF (i.e. fragments overlapping the functional target genes). Beyond a certain value ($\geq$150), almost all TFs give a significant signal (Figure 4.6). We have verified that this behavior is not reproduced on negative control sets generated using the 20kb window width permutation. The behavior is indeed very specific to TFs target set. This factor has not been exposed by previous studies. The effect is even more clearly seen when correlating spatial proximity signal to the number of existing contacts within a fragment set (TF functional target set) (Figure 4.6). The figure indicates that the probability of yielding a signal that is significant beyond a certain number of contacts ($\sim$ 2500) among the functional target is nearly 100 %.

4.3.3. Global trends of TFs Counts, and impact of rNorm 3C correction

To better understand the importance of locality for spatial proximity significance, we have extended the analysis by comprehensively assessing the TFs functional target 3D proximity significance at randomization scale ranging from 1kb to 70kb (Figure 4.7). The first observation is that the number of TFs selected increased inversely to scale of randomization (window width) when using the high confidence complete dataset (FDR-1 %) . This behavior is opposite to what

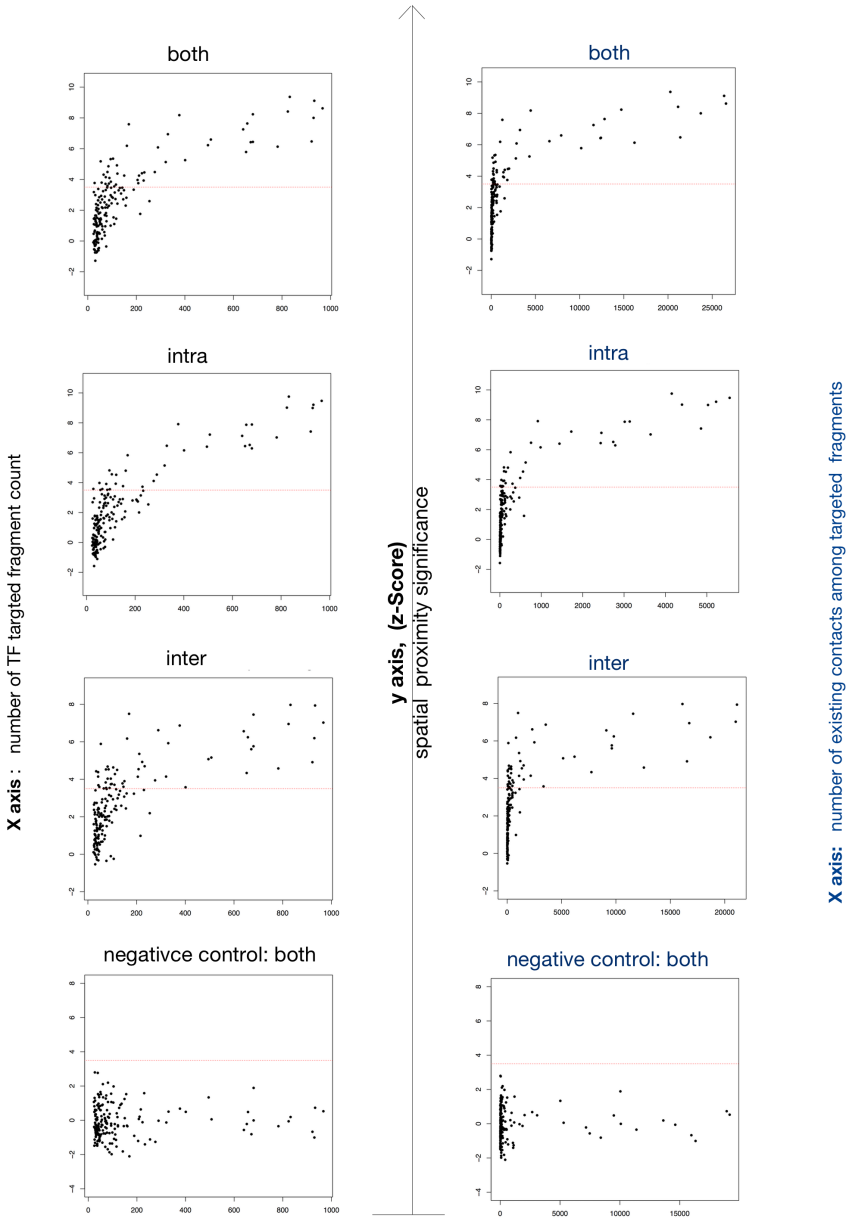

Figure 4.6. Effect of Transcription Factor number of functional target genes on colocalization significance]. The effect is represendted by correlating the spatial proximity significance to either the number of TF targeted 3C restriction enzyme digestion fragments or the number of existing 3C interactions among the TF target genes. There is a total of 174 TFs.

would be expected given that centromere colocalization significance decreased when randomization scale was decreased. Thus, this result suggests that the TFs target fragments selected are indeed far functionally relevant. The same behavior is observed and even more pronounced when considering TFs count obtained using the intra-chromosomal contact subset, most likely imprinting it's pattern unto the

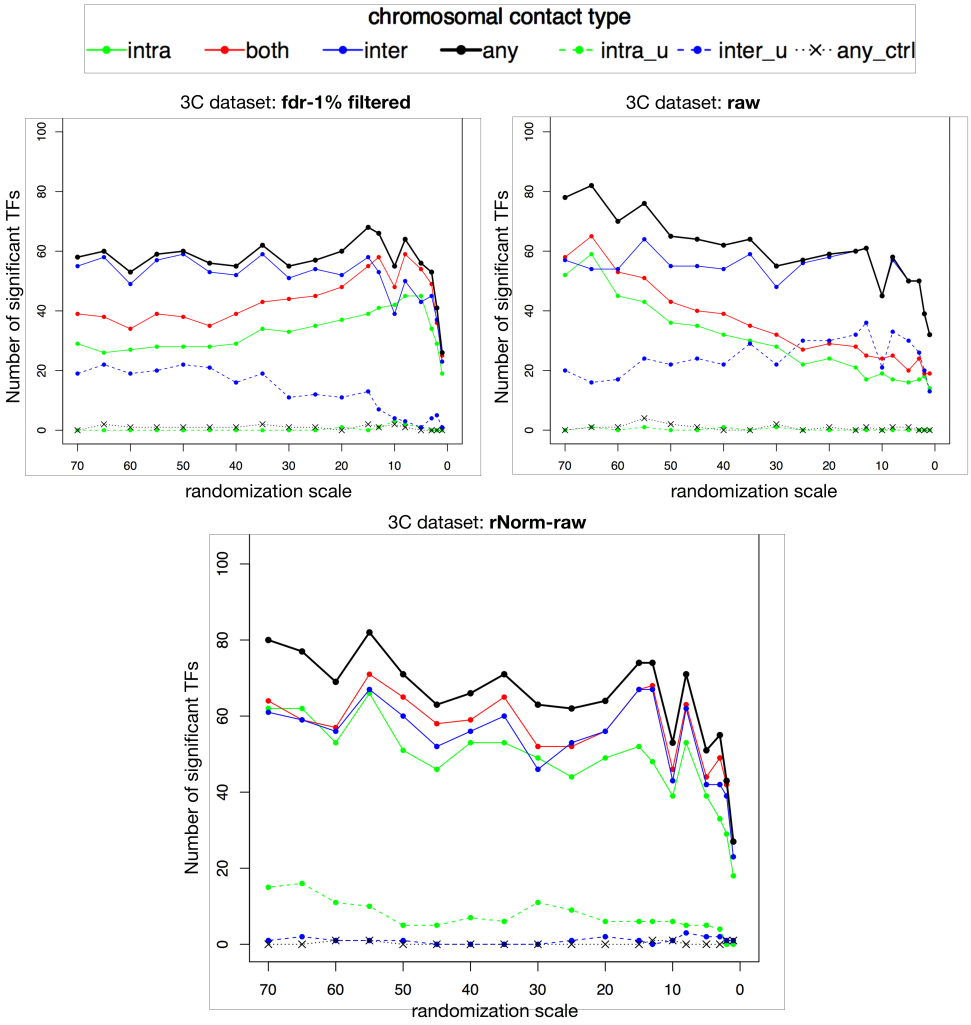

Figure 4.7. Effect of randomization scale on significant TFs counts. TFs are said to be significant if their functional target genes give a spatial proximity signal beyond the cut-off. An error rate of 5 %, with Bonferroni correction for multiple testing gives a z-score significance cut-off of 3.5. Counts were done comparing 3 different 3C dataset : 1.high-confidence filtered (false discovery rate of 1 %) 2. raw 3C dataset 3. rNorm-raw dataset, being the raw dataset normalized using region based normalization described in this article. The normalization of the rNorm dataset was done at each scale of randomization before TFs spatial proximity significance is assessed, according to the method principle. There is a total of 174 TFs. There are 6 counts categories as follows : 1.both, which corresponds to counts of TFs deemed significant when using the complete 3C dataset (i.e. intra+inter -chromosomal contacts). 2. intra, which corresponds to counts of TFs deemed significant when using the intra-chromosomal subset of the 3C data. 3. inter, which corresponds to counts of TFs deemed significant when using the intra-chromosomal subset of the 3C data. 4. any, which corresponds to counts of TFs deemed significant under either of both, intra, or inter 3C dataset. 5. intra_u(intra_unique), which corresponds to counts of TFs deemed significant only when using the intra-chromosomal subset of the 3C data, but not the complete dataset. 6. inter_u(inter-unique), which corresponds to counts of TFs deemed significant only when using the inter-chromosomal subset of the 3C data, but not the complete dataset.

complete dataset trend. A plausible explanation of these trends is that a number of genomic reorganization events, driven by TFs target genes colocalization, occur more frequently at the lower scale (25kb and under).Moreover, cis-regulatory events, under the current settings, are an important part in driving genome functional organization. When focusing on inter-chromosomal contact subset, TFs counts appear not to be correlated with the scale of the randomization, until the lower scale (10kb or less, due to data resolution limitation potentially). Though the behavior is different from that of intra-chromosomal, it still deserves attention, because the intuitively expected behavior would be a decrease.

It is interesting to bring note that the resolution of Duan et al. [2010] 3C data is of 10kb, and that the optimal detection of TFs who's functional target are in spatial proximity is between 15kb and 5kb. However, when looking at the details of the method used to evaluate contact probability, Duan et al. [2010] used a 5kb window around DNA fragment of an interaction to calculate that probability. This is more likely the reason why the detection of significance is optimal in the 15 to 5 kb range.

To gather additional insight on the matter, we have also performed the same analysis of the importance of locality on TFs spatial proximity significance using the unfiltered raw data (Figure 4.7). The number of TFs who's functional targets give a significant signal of spatial proximity when using the complete dataset decreases as the randomization scale is reduced also. This behavior is also reflected in the number of TFs selected when using intra-chromosomal contacts subset. Given the behavior observed using high confidence 3C dataset, a possible explanation is that linear distance between loci dictates the interaction frequencies in large majority. If so, the colocalization effect is decreased as randomization window is reduced, and the linear effect accounted for (i.e. minimized). However, inter-chromosomal contact TFs count remained fairly high and did not decrease gradually when the randomization width was reduced. Inter-chromosomal contacts seem to remain significant at various scales. This suggests, as mentioned before, that inter-chromosomal contacts are more stable and less sensitive to variations on the scale of randomization, prior to any data correction. This is most probably due to the weak probability of any inter-chromosomal contact to happen and be captured, even in the unfiltered raw dataset.

4.3.4. Progressively applying rNorm correction and randomization at increasing scale resolves better spatial proximity signal

After these observations, we have hypothesized that the best way to accurately assess the presence of significant 3D features at various scales would be to assess

each contact significance at a given scale. In order to do so, each interaction frequency from the unfiltered raw data was normalized using the normal distribution by considering the existing interactions between the two regions put in contact (See Materials and Methods, and Figure 4.2). We will refer to this approach and dataset as rNorm (region based normalization). Regions where delimited by a window width defined by the scale of interest (x-kb). Thus, for each randomization scale, contacts were first normalized to equal scale. The TFs functional target spatial proximity has been performed at scale ranging from 1kb to 70 kb inclusively, with 5 kb steps at most. The result presents a drastic improvement over the capability to detect structural functional features using the complete 3C dataset or intra-chromosomal contacts only (Figure 4.7). The number of TFs selected using intra-chromosomal contacts nearly doubled at scales further from the 15 to 5kb optimal range compared to the FDR-1 % filtered 3C data. Moreover, the TF counts obtained when using intra-chromosomal contacts remain relatively high as the randomization window is decreased (3 - 10kb). Likewise, the complete dataset is more able to capture signal over the range of randomization width covered, being consistently equal or higher than inter-chromosomal or intra-chromosomal subset derived TF counts. Another observation is that the inter-chromosomal unique (as defined earlier) counts using region based normalization is nearly null (0 to 2). This suggests that inter-chromosomal dependent signal is now better captured when using the complete dataset (intra+inter), and that intra-chromosomal contact noise dims less the inter-chromosomal signal in the complete dataset.

4.3.5. Individual Spatial proximity signal profiles of yeast transcription factors : rNorm insight on the scale of TFs related structural features

In order to better understand some of the characteristics of the underlying spatial organization features detected, we have analyzed for each TF the spatial proximity signal variation across the 1 to 70kb window of scale. This gave individual spatial proximity significance profile (Figure 4.8). When examining these profiles, the first striking observation is the impact of the number of targeted DNA fragment on the spatial proximity significance, as presented in Figure 4.6. What was seen at a 20kb randomization scale is even more strongly established when considering the wider randomization scale range. When the number of target DNA fragment mapping the target genes of a TF is in the lower range (20 to 50), the proportion of TF displaying spatial proximity in that range is lower. A second range can be defined, 50 to 150 fragments, where signal significance tends

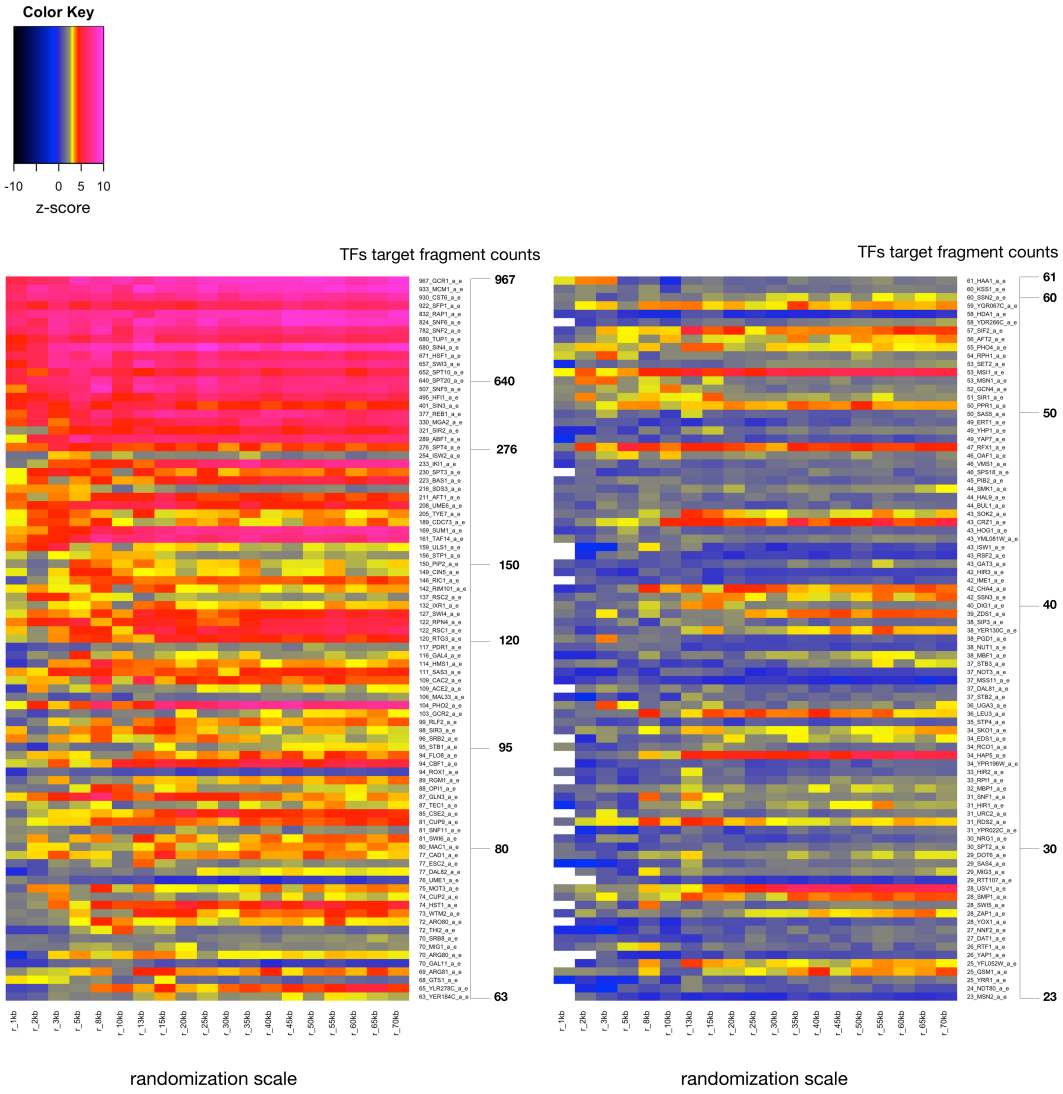

Figure 4.8. Spatial proximity significance profile of yeast transcription factors (TFs). There is a total of 174 TFs. The color scale represents spatial proximity significance z-score. The significance cut-off is a z-score of 3.5, corresponding to an error rate of 5 %, with Bonferroni correction for multiple testing. The dataset used is the corrected raw data, using region based normalization for correction (rNorm) on the complete dataset (intra+inter chromosomal interactions).The normalization of the rNorm dataset was done at each scale of randomization before TFs spatial proximity significance is assessed, according to the method principle. Each line represents a TF's spatial proximity significance over the covered randomization window width range, i.e. randomization scale. TFs profiles are sorted in descending order of the number of target 3C DNA fragments ;those having the highest number of target fragments at the top left. Thus, they are also sorted in descending order of the number of functional target genes of a TF.

to be more visible, as well the proportion of TFs displaying significant spatial proximity of their functional targets. An important observation when analyzing the profiles using rNorm dataset is that the spatial proximity signal when using the complete dataset tends to be imprinted by the strongest signal pattern from

either intra-chromosomal or inter-chromosomal contacts. (Figures 4.9, 4.10). The last range ($>$ 150 TFs target fragments) tends to have all TFs display spatial proximity significance and do so over the whole range of scales covered. A possible explanation is that numerous colocalization events to which the targeted genes participate exist at different scales. As such, when the number of targeted genes increase, the number of such detected events increases. The numerous and overlapping signals at multiple scales then give a consistent, still very significant signal over large scale range (negative controls give no significant signal, Figure 4.9). What follows this model is that there can be cases displaying spatial proximity signal at distinct randomization scale. These would be better resolved when the number of TF target fragments is lower. There are TFs such as YFL052W that displays spatial proximity signal peek at 13kb and then 55kb range, which is in both cases inter-chromosomal specific (Figure 4.9). The transcription factor SOK2, on the other hand, has an intra-chromosomal contact specific peak around 55kb, and inter-chromosomal contact specific peak at 13 kb of randomization scale. Another case is that of LEU3, which has over 4 peaks where intra-chromosomal and inter-chromosomal signal peaks are coinciding in scale. It is interesting to know that LEU3 transcription factor acts as an activator and repressor depending on the conditions, in branched amino acid biosynthesis and ammonia assimilation. A possible scenario explaining LEU3 profile is that genes correlated with activation would be associated and colocalize at certain scale and the genes co-repressed by LEU3 would give a colocalization signal at another scale. In the same order of idea, the two distinct biological processes mentioned above could also be at the origin of having multiple peaks, resulting in distinct regulatory events. MOT3, like LEU3, is also a transcription factor acting either as repressor or activator on multiple biological processes and shows similar patterns to LEU3, further supporting the hypothesis of distinct biological processes regulation leading to specific peaks on spatial proximity profile.

It now becomes imperative to investigate and determine the exact identity of the sequences at the origin of a signal peak at a given scale. This could be done by looking at a subset of TF target that is functionally related. These could possibly form some sort of cluster leading to a proximity signal, and would verify the hypothesis on the origin of the scale specific signal peak so. Identifying the clusters on the genome structure could lead to defining nuclear structural assemblies that are key or on the basis of genome functional organization, and perhaps other nuclear bodies spatial regulation also. This idea can be given further thought when considering that MOT3 transcription factor can form [MOT3+]prion, allowing to extend protein complexes, and could thus participate in forming nuclear bodies or

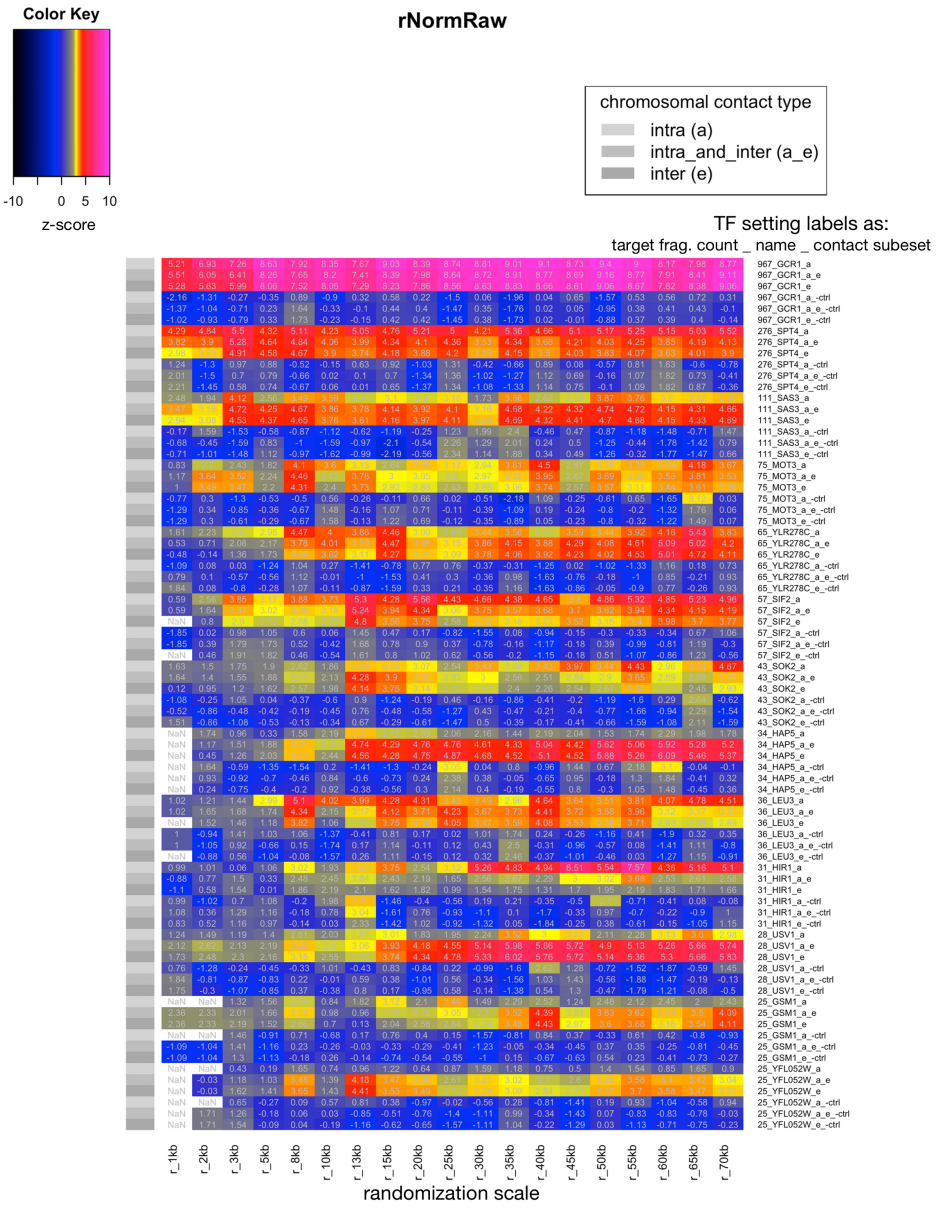

Figure 4.9. Comparative spatial proximity profiles of 13 selected transcription factors (TFs). The TFs are selected over the range of the 174 TFs fragments count. The fragment counts are the result of mapping a TF functional target genes to the corresponding 3C DNA fragments. The color scale represents spatial proximity significance z-score. The significance cut-off is a z-score of 3.5, corresponding to an error rate of 5 %, with Bonferroni correction for multiple testing. The dataset used is the corrected raw data, using region based normalization for correction (rNorm).The normalization was done at each scale of randomization prior to calculating TFs spatial proximity significance, according to the method principle. For each of the 13 (TFs), the spatial proximity significance profile was represented at randomization scale of 1kb to 70kb.The profiles were computed with either the complete dataset) , the intra-chromosomal contacts dataset or the inter-chromosomal contact dataset. Negative controls ("-ctrl" lines) are added under the 3 same data settings. Negative controls for a TF at a given scale correspond to a shuffling of the target fragments where each fragment randomized position is at most at distance equal to the scale from its original position as TF target fragment.

regulatory complexes. Once again, the profiles interpretation is in agreement with a model consisting of multiple clusters of target genes of a TF, rather than a single event for the target gene set. This said it is also possible when a spatial proximity signals spans over a large scale range to be originating from a single cluster of events. This is particularly important to remember when looking at inter-chromosomal dependent spatial proximity significance, since inter-chromosomal contact, from our previous results (Figure 4.6), tend to be less sensitive to the scale of randomization and normalization. The method of interpretation described above becomes useful and rich when multiple peaks can be identified at different scales and contact types. Considering the observations and proposed model, it is interesting to notice that there are events for which even a 1kb perturbation in the fragment position is enough to exhibit significance (given that 1kb randomization scale also displays consistent significance for at least 19 TFs, also seen in Figure 4.10).

4.3.6. Comparing the performance of rNorm method to others for detecting functional structural features of the genome

When comparing the TFs profiles with different 3C data correction (raw, FDR-1 % and rNorm) using a global and comprehensive representation (Figure 4.10), the trends seen in Figures 4.7, 4.8 and 4.9 are better grasped. When looking at the profiles obtained using raw data, there is a clear trend to better capture spatial proximity signal at 40kb to 50kb of scale and beyond. The inter-chromosomal contact based profiles of raw 3C data signal profiles spans significance from at least 5kb, up to the maximum scale assessed (70kb) . The global profiles obtained when using the complete raw dataset (labeled :"both") tend to follow more the intra-chromosomal patterns, and to a lesser extent, bear the inter-chromosomal patterns. This exemplifies the fact that when using complete raw dataset, the intra-chromosomal contact effect is predominant, shadowing the inter-chromosomal patterns. Also, in Figure 4.10, since the intra-chromosomal spatial proximity significance is better captured on the scale range from 40kb - 50 kb and beyond , it could give a relative empirical estimation of where the linear distance effect (polymer looping effect perhaps) becomes less predominant. At lower scales, the interaction frequencies being greatly influenced by raw intra-chromosomal contacts, there is little significant signal captured. This also is a result of the region-based randomization approach accounting for the linear distance between loci.

The 174 TFs profiles obtained using the high-confidence 3C dataset (FDR-1 %) have a somewhat opposite global trend to those obtained using the raw

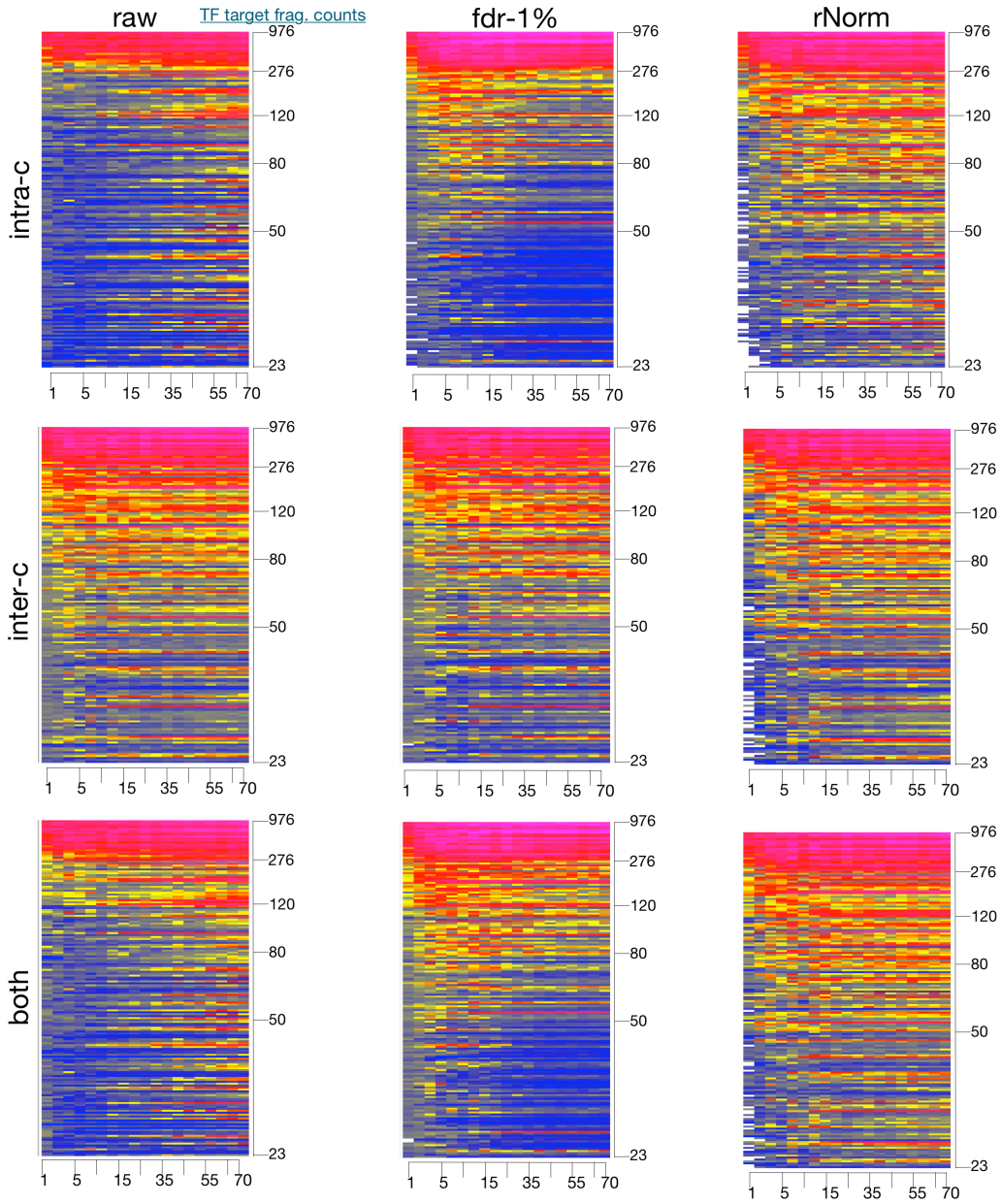

Figure 4.10. Comparison of 3C correction approaches using the global yeast TFs spatial proximity profile compilation. 3 different 3C datasets ; 1.high-confidence filtered [false discovery rate (FDR) of 1 %)], 2. raw 3. rNorm-raw, being the raw dataset normalized using region-based normalization described in this article. The color scale represents spatial proximity significance z-score. The significance cut-off is a z-score of 3.5, corresponding to an error rate of 5 %, with Bonferroni correction for multiple testing. The normalization of the rNorm dataset was done at each scale of randomization prior to assessing TFs spatial proximity significance is, according to the method principle. There is a total of 174 TFs on each color matrix. Each row represents a TF's spatial proximity significance over the covered randomization scale (1- 0 kb). TFs profiles are sorted in descending order the number of target 3C DNA fragments (top at 967, bottom at 23), and therefore essentially sorted in descending order of the number of functional target genes of a TFs.

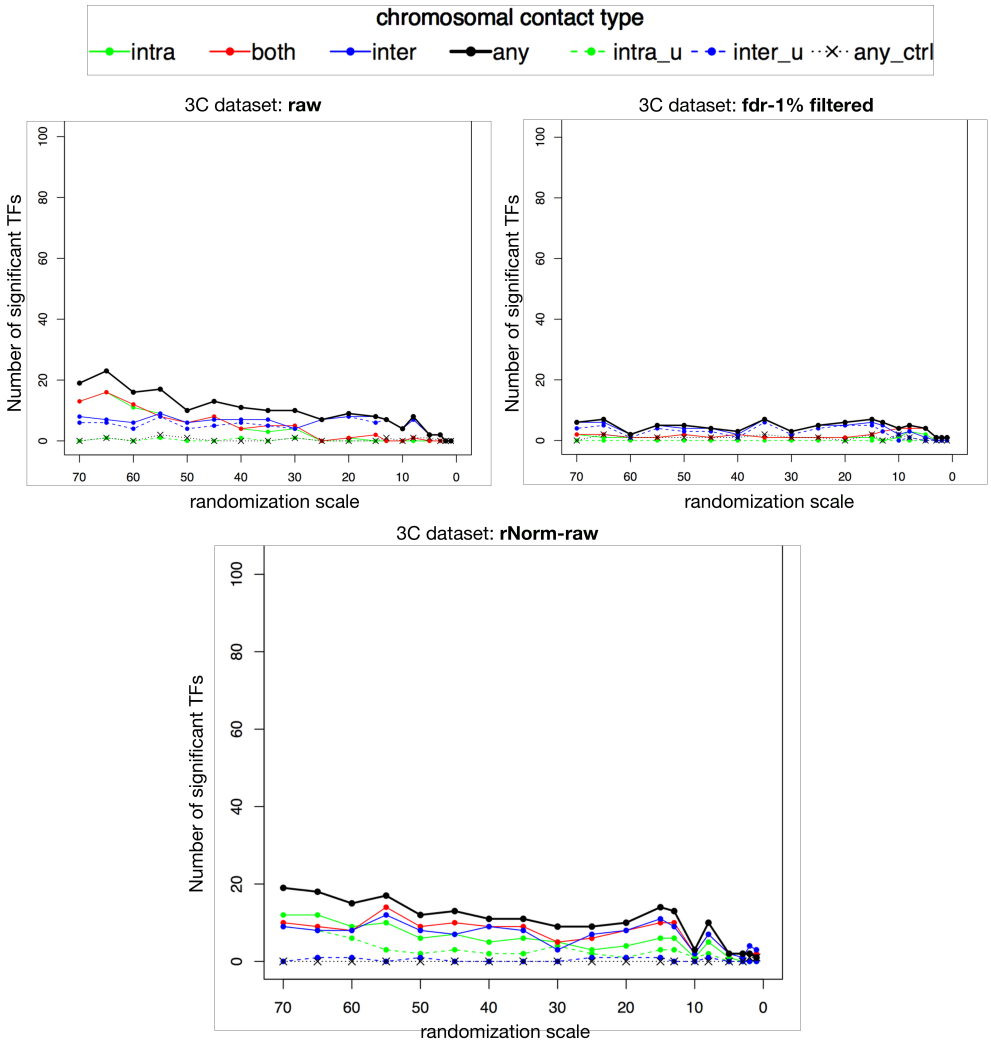

Figure 4.11. Effect of randomization scale on significant TFs counts of TFs having few target genes. The figure is the equivalent of Figure 4.7, but for TFs having 50 or less 3C DNA target fragments. TFs are said to be significant if their functional target genes give a spatial proximity signal beyond the cut-off. An error rate of 5 %, with Bonferroni correction for multiple testing, gives a z-score significance cut-off of 3.5. Counts were done comparing 3 different 3C dataset : 1.high-confidence filtered (false discovery rate of 1 %), 2. raw 3. rNorm-raw, being the raw dataset normalized using region based normalization described in this article. The normalization of the rNorm dataset was done at each scale of randomization before TFs spatial proximity significance is assessed, according to the method principle. There is a total of 174 TFs. There is 6 counts category as follows : 1.both, which corresponds to counts of TFs deemed significant when using the complete 3C dataset contacts. 2. intra, which corresponds to counts of TFs deemed significant when using the intra-chromosomal subset of the 3C data. 3. inter, which corresponds to counts of TFs deemed significant when using the intra-chromosomal subset of the 3C data. 4. any, which corresponds to counts of TFs deemed significant under either of both, intra, or inter 3C data setting. 5. intra_u(intra_unique), which corresponds to counts of TFs deemed significant only when using the intra-chromosomal subset of the 3C data, but not the complete dataset (both). 6. inter_u(inter-unique), which corresponds to counts of TFs deemed significant only when using the inter-chromosomal subset of the 3C data, but not the complete dataset (both).

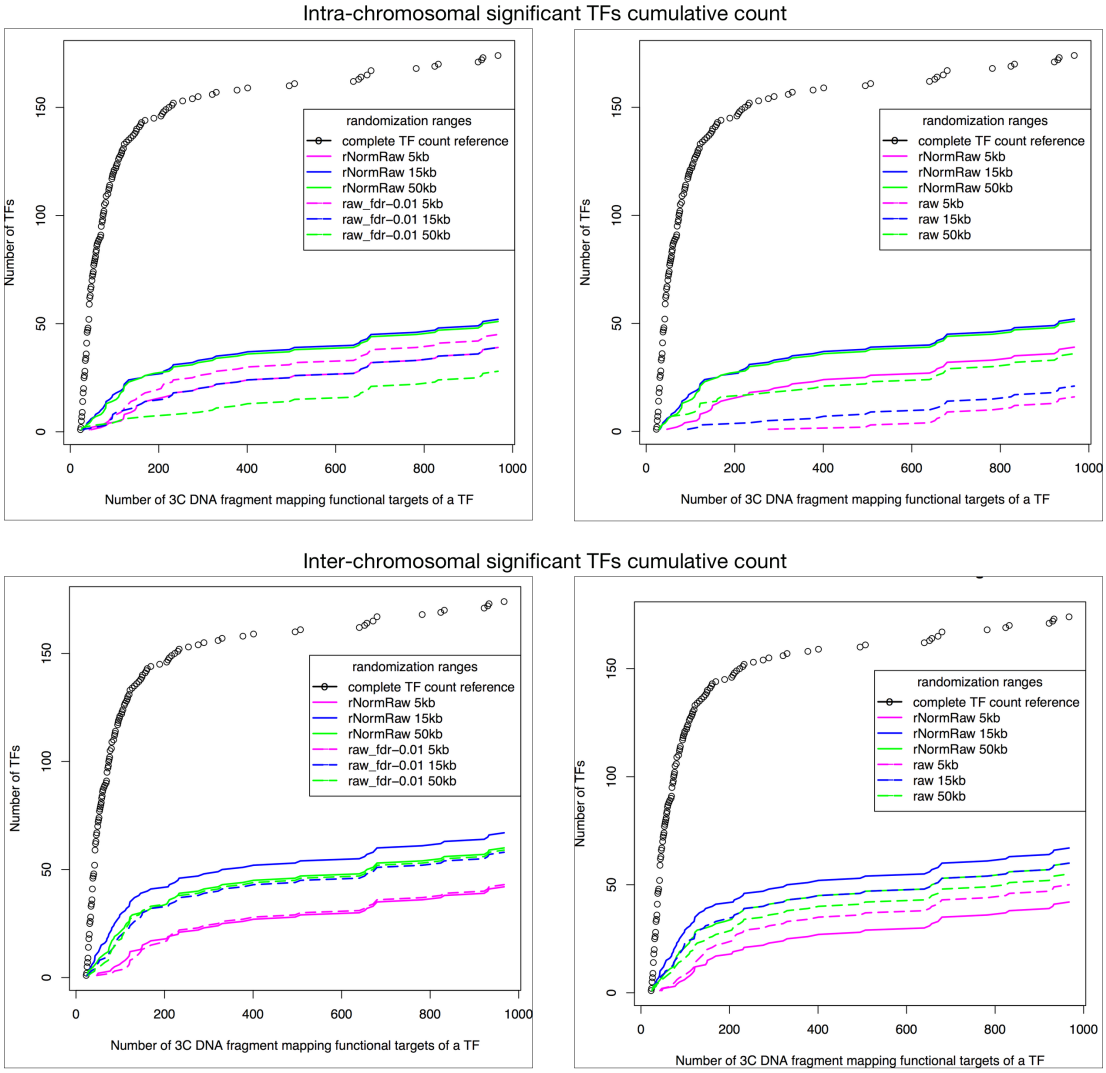

Figure 4.12. Cumulative significant TFs count. The cumulative number of significant TFs is the number of significant TFs having at least "x" number of target 3C DNA fragment. TFs are said to be significant if their functional target genes give a spatial proximity signal beyond the cut-off. An error rate of 5 %, with Bonferroni correction for multiple testing, gives a z-score significance cut-off of 3.5. Counts were done comparing 3 different 3C dataset : 1.high-confidence filtered (false discovery rate of 1 %), 2. raw 3. rNorm-raw, being the raw dataset normalized using region based normalization described in this article. The normalization of the rNorm dataset was done at each scale of randomization before TFs spatial proximity significance is assessed, according to the method principle.

data set. In this case, the scale range of intra-chromosomal contact based spatial proximity significance is shifted to the left, approximatively below 35kb. However, there is a none negligible number of TFs who's signal does span beyond 35 kb. A noticeable contrast between the raw dataset and high-confidence filtered dataset is that in the lower end (TF with a number of target fragments < 50) there is little

to no spatial proximity significance. This is most probably due to the fact high-confidence filtered data have, by the very principle, only a small fraction of the total DNA-DNA contacts captured during an experiment. Thus, when the number of contacts is limited, the statistical power is also limited. possible contacts follows the number of fragments, and in this case, the number of existing contacts. To better assess the significance of TFs intra-chromosomal signal in that interval of 50 or fewer target fragments, significant counts for that interval were plotted in Figure 4.11. The depletion is then obvious, intra-chromosomal dependant count being nearly null. As for the inter-chromosomal dependent generated profiles of the high-confidence 3C dataset, the spatial proximity significance does not have any apparent shift in scale. However, there is some depletion of spatial proximity signal.

This is best appreciated in Figure 4.11, when considering the inter-chromosomal dependant significant TFs counts, which are inferior to both the raw dataset and rNorm dataset. As for the TFs profiles obtained from the rNorm dataset, first, the intra-chromosomal based profiles present little to no shift on the range of randomization scale assessed. This point is critical and demonstrate the importance of a scale dependent normalization. A quantitative comparison with the other dataset shows that the number of significant TFs counts at lower(15kb) and higher(50kb) randomization scale is superior to both the raw dataset and high-confidence dataset (Figure 4.12). As for the inter-chromosomal rNorm profiles, just as for the two previous datasets, the profiles do not present any apparent depletion or shift, except at 5kb and below (Figure 4.10). This decrease at 5kb scale and below is due mostly to the fact that the use of a normal distribution to assess contact significance in the region-based normalization is limited at such scales. The limitation is the number of fragments on a 5kb interval (given the experiment resolution). Thus the number of existing contacts from which to infer a distribution is also limited. Another statistical approach could be used to replace the normal distribution in order to overcome this limitation, such as the one used by Duan et al. [2010] for instance. This can be done while preserving the principles and properties of region-based approach. Because of the effect of the number of targets fragments observed in Figure 4.6, we also wanted to see how the number of significan TFs varies with the number of their targets using cumulative counts. When comparing cumulative TFs counts for the inter-chromosomal contact dataset (Figure 4.12), the rNorm dataset has similar counts to the high-confidence filtered dataset at 5kb and 50kb, and slightly higher at 15kb. The rNorm counts are also similar to the raw dataset (Figure 4.12), except at 5kb, where significant TFs count are slightly inferior, due to the limitation of the method mentioned

above at this scale. This approach of analysis (in Figure 4.12) is however very limited, due to the fact that it is highly dependent on the significance cut-off. If the cut-off is set to be restrictive (as it is the case with Bonferroni correction), it will capture less dynamic changes of counts.

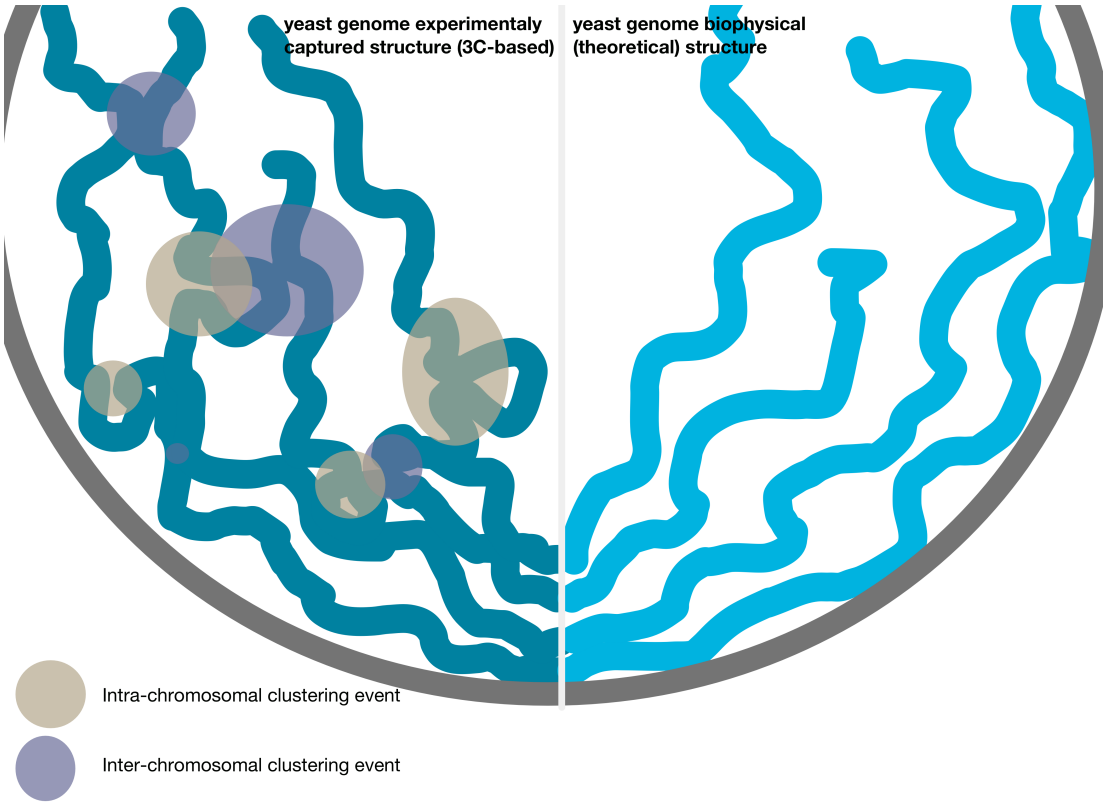

Figure 4.13. Representation of genome functional 3D organization. The yeast genome (experimental) structure results from transcription spatial regulation (e.g. transcription factories), at various scales, DNA contact type (intra- and inter-chromosomal contacts). The distribution of transcription factor targets play an important role on the resulting 3D structure dynamics. Nuclear physical constraints still emprint overall large scale features of the genome as in a theoretical model, data resulting from these structures (experimental and theoretical) thus sharing many features, which without progressive scaling of hi-c correction, would be predominant.

Reports of Single cell Hi-C experiment proposed that significant interactions are at smaller resolution mostly. We have observed these over the whole scale range assessed. One thing to remember is that 3C data is from a population average. It captures structural features that are highly significant in a small fraction of the population, just as conformations that are weaker yet significant in a greater fraction of the population. Since single cell Hi-C looks at a very limited number of cells, they are more likely to capture conformations that give a strong signal in a majority of the cells. It is likely that these tend to be at lower scale. However, given the limited number of cells, it is statistically hard or near impossible

to capture conformations that are weaker , but meaningful in a fraction of the population. An extreme example of such large scale but rare conformation at the population level would be mitosis, which happens on a smaller fraction of the cell and harder to capture at a single-cell level (without synchronizing for the event), but clearly present at a population level. Thus, a signal at larger scales might be more challenging to study at single cell level, unless a statistically sufficient number of cells is analyzed.

4.4. Discussion

We have presented in this article strong arguments of the importance of a region-based strategy, and thus locality, instead of global randomization or normalization approaches, as is often used. We have developed a method, rNorm (region based normalization), which assesses the significance of DNA-DNA contact and normalizes them per region. This method can be used for both intra-chromosomal and inter-chromosomal contact correction. On intra-chromosomal contacts, the method is particularly useful, given that intra-chromosomal interaction frequencies tend to reflect linear distance effect. An advantage of this method over some the state of the art approach (e.g. ICE, [Imakaev et al., 2012]) is that it can be applied to genome-wide dataset as well as datasets aiming a genomic region of interest. This is important because many 3C-based studies are interested on a subset of genomic regions or sequences and their relation to the rest of the genome, or on intra-chromosomal contact type only for instance. We have shown how this approach yields results when applied on unfiltered raw 3C dataset (giving rNorm dataset) which are superior to both filtering for significant contact or using unfiltered dataset. To our knowledge, previous analysis of functional genomic organization using 3C-based methodology required filtering for high confidence set of contacts, which is not necessary in our case.

We presented a methodology for inquiring genomic structural features by representing their spatial proximity significance over a range of scales. This methodology is the first, to our knowledge, allowing to have a critical insight of the relative effects of 3C data correction method and their parameters. Specifically, we have shown some of the possible advantages and drawbacks of using high confidence contacts filtering, which often used by default.

Moreover, we have further added evidence in favor of a functional genome 3D organization. We have shown pieces of evidence of functional clustering of genomic loci by studying the spatial proximity of TF target genes. Our results indicate that a significant fraction of TFs shows significant signal of spatial proximity of their functional targets. Our studied has more than doubled this proportion from

previous reports. Our analysis presents a much more detailed insight into the properties and behavior of the colocalization signal of TFs target genes. Factors such as the size of the target genes set, the type of contact from which the signal is drawn and even the scale at which such event could occur have been addressed in a quantitive way for the first time in this context. When applying the view drawn from evidence assembled in this work, the concept transcription factories can be naturaly overlayed, yet with the insight on scale and contact type.

It is noteworthy to have a scale aware significance for each interaction. This allows being able to trace and pinpoint exactly the epicenter of functional reorganization. It is, therefore, possible to determine the regions or even the potential regulatory sequences at the origin of a genome functional reorganization. Combined with the information of the binding sites of proteins such as TFs, cofactors or even chromatin remodelers, it becomes possible to start mapping the players of the regulation of genome functional structure. This asset in itself is unique and invaluable to understand functional genome features since it allows probing the sea of 3C contacts for meaningful information. Based on the information and evidence gathered, we have proposed a model of genome spatial organization where TFs targets are parts of different clusters, at differing in scale and contact type. We suggest that this architecture of organization could also support the organization of other nuclear complexes and matter.The study makes it possible to envision integrating DNA-DNA contact to protein-DNA and protein-protein interactions, in a useful way as to achieve an integrated model of molecular nuclear complexes and interactions.

Recently this year, a comprehensive review by Schmitt et al. [2016] of chromosome conformation capture derivatives and methods to analyze chromosome architecture suggested what are the three major limitations 3C and methods of analysis are facing. The first was the limited insight on single-cell to single-cell variability in the data captured, the second the was resolution required to resolve to understand gene regulation mechanism, and the third the need for multiway DNA-DNA contact capture. The edge of our method is that is not only scale aware but also ready to exploit significantly the resolution when a greater resolution is achieved.

4.5. Materials and Method : rNorm presentation

See [Duan et al., 2010] for data processing, from experiment, to sequencing, and processing required to obtain the raw dataset and high-confidence filtered (1 % false discovery rate) dataset.

4.5.1. Metric of spatial proximity

A core part of the analysis presented here depends on computing the spatial proximity significance among a number of genomic positions. This measure is calculated based on 3C data. Each position or locus is therefore mapped to a DNA region, referred to as fragment, according to the digestion of the genome by a restriction enzyme during the 3C-based experiment. The set of fragments and interaction frequencies between these fragments compose a network (or graph), with fragments as vertices and frequencies as edges. We then have a weighted undirected graph $G = (V, E)$. Let V be the set of vertices, and let E be the set of edges of G. We use $\{i, j\} \in E$ to refer to the edge between vertices i and j, and m_{ij} for the associated weight. For convenience, we set $m_{ij} = 0$ when $\{i, j\} \notin E$. Also, we set $m_{ii} = 0$ for all $i \in V$. The weights are first the interaction frequencies between fragments, and when mentionned, the normalized frequencies as presented further in the text. Given a set of genomic regions of interest, i.e. functionally related loci such as centromeric regions, telomeres, tRNA, genes targeted by a transcription factor, or any other functional relation that can define a set of loci, we define the subset of vertices A intercepted by these functionally related regions. We note $G_A = (A, E_A)$ the subgraph of G induced by A. The set of edges E_A is thus the set of all pairs $\{i, j\} \in E$ such that i and j are both in A. The set of all subgraphs of G induced by subsets of vertices of same size as A will be noted $W[A]$. The number of such subgraphes (i.e. the size of the set $W[A]$) is therefore equal to $\mid W[A] \mid = \binom{|V|}{|A|}$.

The problem of spatial proximity can be recast more generally and more formally as follows. First, we define a measure of the mean density of interaction $d(A)$ among a given subset of fragments A included in V. Second, for a particular subset of fragments A, we ask whether the mean interaction density $d(A)$ is higher on average than the mean interaction density obtained on random subsets of V of the same size.

The interaction density is calculated as follows :

$$d_{obs} = d(A) = \frac{2}{|A|(|A| - 1)} \sum_{i,j \in A} m_{ij}. \tag{4.5.1}$$

Here, $\frac{|A|(|A|-1)}{2}$ is simply the number of possible edges between the vertices in subset A.

Due to complex correlations among fragments, the null distribution for this test cannot be analytically computed. Here, the null distribution was simulated by Monte Carlo. Specifically, the subset of fragments was randomized with $B = 1000$ times. Each randomization b consists of a permutation of the indices in V, i.e. a one-to-one mapping π_b from V to V, which is further constrained so as to preserve a certain number of properties of subsets of fragments (see below). This mapping maps A to a subset $\pi_b(A)$ of same size, and the statistic $d_b = d(\pi_b(A))$, is computed on this subset. The series of values d_b, for $b = 1..B$ represents our simulated sample from the null distribution.

Based on this simulated null distribution, we looked at two classical measures of significance :

(1) The p-value : here, calculated as the frequency over Monte Carlo replicates of draws p in 1..P for which $d_b > d_{obs}$.

(2) The z-score, obtained by calculating the empirical mean μ and empirical standard deviation σ of the list of values d_b, $b = 1..B$ and then defining :

$$z = \frac{d_{obs} - \mu}{\sigma} \tag{4.5.2}$$

The p-value is most often used for assessing significance. On the other hand, Monte Carlo lacks power for estimating low p-values. As an alternative, the z-score may allow rapid discrimination between strongly and moderately rejected instances.

4.5.2. Controlled randomization schemes

Formally, it is possible that the fragments contained in A are enriched in some property that, in itself, already implies a higher interaction density than on random subsets of fragments of same size. Particular cases of this are when the fragments of A are in linear proximity along the chromosomes. Because the weight of the edge m_{ij} is highly correlated to the genomic distance between it's vertices, δ_{ij} , the permutation b must preserve distance among fragments as much as possible, i.e. $\delta_{\pi_b(i)\pi_b(j)} \simeq \delta_{i,j}$ for all i and j in V. Furthermore, it is important that not only the set of consecutive distances be similar, but also the set of distances from all pairs, because the latter can change significantly even when preserving consecutive distances.

Accordingly, permutation was implemented by constraining the randomization, so as to seek to minimize the deviation from these properties. In practice,

we found that a convenient way to enforced the properties mentioned above, is to have the mapping π such that the genomic position of fragment $\pi_b(i)$ is close to fragment i. Defining a tolerance scale α, we set :

$$\forall b, \quad \delta_{i,\pi_b(i)} \leq \alpha \tag{4.5.3}$$

$$\forall b, \quad \left[\delta_{i,\pi_b(i)} + \delta_{\pi_b(i),\pi_b(i)}\right] \leq 2\alpha,$$

$$\forall b, \quad \left[\mid \delta_{i,j} - \delta_{\pi_b(i),\pi_b(j)} \mid\right] \leq 2\alpha \tag{4.5.4}$$

Above, equation 4.5.4 rises from the fact that if the genomic distances in A for $\delta_{i,j} = a$, and v_i is permuted such that $\delta_{i,\pi_b(i)} = \alpha$, and $\delta_{j,\pi_b(j)} = \alpha$ likewise, then $\delta_{\pi_b(i),\pi_b(j)} = a \pm 2\alpha$, if the remapping occurs in opposite directions. The advantage of such an approach is that it allows to use a comparable and common basis, which preserve properties of a given region and contrast to other regions in contact. Conceptually, if $\delta_{i,\pi_b(i)} \leq \alpha$, and α is relatively small, any difference of properties between the two genomic positions, such as biophysical properties, the physical anchoring of genomic region, or even epigenetic properties possibly, will also be small or negligeable. In the method presented, this is the strategy used to better expose features, by spatial proximity significance, that are more likely to arise from active biochemical process, and less the result of properties which can be modeled computationally. Thus, beyong the distribution of genomic distances between fragment ($\delta_{\pi_b(i)\pi_b(j)}$), other propeties are potentially kept comparable. We can generalize and formalize this notion of differential properties by a function λ , the difference of properties between the genomic position of v_i and v_j, parametrized by $\lambda_{\pi_b(i)\pi_b(j)}$, similarly to the genomic distance δ. We can therefore express that if $\delta_{i,\pi_b(i)}$, is small, then $\lambda_{i,\pi_b(i)}$ will be reduced, such that $\lambda_{i,j} \approx \lambda_{\pi_b(i),\pi_b(j)}$.That is the aim when controling for the scale of permutation,α.

The first controls are centromeric regions. They are chosen because they present the strongest signal and frequencies, and known to be anchored to the nucleus membrane. The fact that the signal of spatial proximity of centromeric regions decreases drastically (more than an order of magnitude of z-score, see results), indicates that the same corrective behavior can be expected for regions put in spatial proximity due physical constraints within the nuclear space.

In our case, the region size, controlled by the window width parameter, corresponds to 2α. Given the dataset, and scale of potential functional features, it is set at values going from 1kb, to 70kb. Thus, spatial proximity significance will be depend on the scale targeted. For simplicity to the reader, we use the terms *randomization window width* for 2α, referring to the implementation. To reason

about the permutation and normalization approach in the article, the term *randomization scale* is used, but formaly corresponds to the α factor.

4.5.3. Region based normalization

A critical difference between the method presented by Paulsen et al. [2013] and rNorm is that with rNorm, interaction frequencies are normalized using the regions of interest put in contact each time. This reflects the principle stating that the significance of a single interaction depends on the distribution of interactions between the two regions put in contact, as presented in the article. Again, the method presented by Paulsen et al. [2013] normalizes interaction frequencies using the set of interactions having equivalent linear distance between genomic element put in contact. This, as mentionned in the article, does not account for the interaction specific landscapes between the regions put in contact. Thus, region based normalization works in a manner that is inherently aware of genome structural properties.

(While done with window width parameter of 1 to 70 kb, here's example with 2α = 20kb) :

: DNA fragment are the result of the division (digestion) of the yeast genome by HindIII enzyme during Chromosome Conformation Capture experiment.

: The genome is divided into 20 000 base pairs intervals (20 kb), e.g. fixed 20kb windows.)

: Normalization for a given interaction m_{ij} between 2 DNA fragments (v_i,v_j) if done by getting a z-score $Z(m_{ij})$:

: r_i = region to which fragment v_i belongs,

: r_j = region to which fragment v_j belongs,

:

$$U(m_{ij}) = \{m_{ab}, a = 1..n, b = 1..n \mid r_a = r_i, \;\; r_b = r_j\} \tag{4.5.5}$$

$$Z(m_{ij}) = \frac{m_{ij} - \hat{\mathrm{E}}\left[U(m_{ij})\right]}{std(U(m_{ij}))} \tag{4.5.6}$$

4.5.4. Algorithm basis of region based randomization

Altogether, the randomization procedure to achieve this, in our implementation, can be described in pseudo-code, as follows :

Algorithm 1 Permutation procedure

with $n =\mid V \mid$, for $G = (V, E)$
for $i = 0$ **to** 2 **do**
switch identity of v_i and v_{i+c}, such that $i \leq i + c < n - 1$ {note that $c \sim unif(0, n - 2 - i)$}
end for

When proceeding with genome-wide randomization, the n fragments are defined as the set of mappable fragments. The result are equivalent when n corresponds to the complete set of fragments, wether mappable or not.

When randomizing fragment sets per chromosome, e.g.random shuffling within each chromosome, the permutation occurs independently on for each chromosome, with n being defined as the the number of mappable DNA fragments on a given chromosome.

When randomizing locus, and thus fragments, per region, the strategy used is the same. The width of a region is first defined (ex : 20kb). Then each chromosome is subdivided into 20kb regions. The permutation then occurs within regions, with n corresponding to the number of fragments strictly contained within that region boundaries, and overlapping the lower boundary (but not the upper).

Note that a bipartite version of this test can also be defined, given two groups A and B : for instance, assuming that two transcription factors a and b physically interact each (protein-protein interaction), we could then ask if the interaction frequencies among pairs of fragments, one of which is bound by transcription factor a, and the other by transcription factor b is higher than expected. This test was implemented in the context of this work, but was not considered any further.

4.5.5. Bonferroni correction for multiple testing

174 transcription factors where included in spatial proximity significance profiles. In order to address the issue of multiple testing, Bonferroni correction was used. Let us remind that is probably too conservative and hinders the statistical power to assess significance. Nonetheless, this approach to multiple testing was used to demonstrate that even with strict criteria, the results are *clearly* seen. An error rate of 5 %, with Bonferroni correction for multiple testing, gives a z-score significance cut-off of 3.5.

Chapitre 5

CONCLUSION

Dans le manuscrit présenté, nous avons établi rigoureusement l'importance d'une approche locale, et par région plus précisément, pour correctement analyser l'organisation du génome à partir de données de type 3C. Nous avons posé les bases, développées, et mis au point une méthode nommée «rNorm» («region based normalization»), qui attribue une importance à chaque interaction 3C. Cette méthode peut être utilisée pour l'analyse de jeux donnés intrachromosomal tout comme interchromosomal, sur un sous-ensemble d'interactions ou la totalité d'un génome. Ce point relevé précédemment est essentiel, puisque lors d'analyse fonctionnelle de la conformation du génome, particulièrement lors d'approches différentielles sur des conditions biologiques qui diffèrent, il peut être nécessaire et suffisant de se pencher sur un ensemble de gènes (ou régions) d'intérêts. Les analyses basées sur l'approche «rNorm» sont robustes ne propagent pas une erreur à l'échelle du génome même une hétérogénéité de reliefs de contacts.

Nous avons aussi mis sur pied une méthodologie qui permet de sonder des caractéristiques d'éléments structuraux du génome en représentant la significativité de la proximité spatiale sur différentes échelles. Au mieux de notre connaissance, cette méthodologie est la première qui permet d'avoir un avis aussi critique sur les effets relatifs d'une méthode de correction des données de type 3C et différents paramètres sur celle-ci. Plus précisément, nous avons démontré les avantages et inconvénients possibles d'utiliser un jeu de donnée filtré par l'importance de significativité des contacts, ce qui est en général fait par défaut.

Qui plus est, nous avons ajouté diverses évidences qui suggèrent une organisation bel et bien fonctionnelle la structure 3D du génome. Nous avons présenté des éléments indiquant le regroupement (clustering) de loci liés par une fonction biologique en étudiant la proximité spatiale des cibles de facteurs de transcription (TFs). Nos résultats indiquent qu'une fraction importante des TFs ont leurs cibles en proximité spatiale. Notre étude a plus que doublé ce nombre par rapport

aux études précédentes. Notre analyse présente une compréhension beaucoup plus détaillée des caractéristiques et propriétés de la colocalisation des cibles de TFs. Des facteurs tels que la taille de l'ensemble de cibles, le type de contact causant le signal, et l'échelle à laquelle la colocalisation se produit ont été évalués de manières quantitatives pour la première fois dans ce contexte. En se basant sur l'ensemble d'évidences présentées, nous avons proposé un modèle de l'organisation spatiale du génome où les cibles de facteurs de transcriptions font partie de différentes grappes (clusters), à différentes échelles, et dépendant de différents types de contact chromosomaux. Nous suggérons que cette base d'organisation génomique pourrait aussi supporter l'organisation d'autres complexes nucléaires et matières nucléaires.

Bibliographie

Ferhat Ay, Timothy L Bailey, and William Stafford Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome research*, 24(6) :999–1011, June 2014. doi : 10.1101/gr.160374.113. URL http://genome.cshlp.org/content/24/6/999.short.

Shay Ben-Elazar, Zohar Yakhini, and Itai Yanai. Spatial localization of co-regulated genes exceeds genomic gene clustering in the Saccharomyces cerevisiae genome. *Nucleic acids research*, 41 (4) :2191–2201, February 2013. doi : 10.1093/nar/gks1360. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23303780&retmode=ref&cmd=prlinks.

Wendy A Bickmore and Bas van Steensel. Genome architecture : domain organization of interphase chromosomes. *Cell*, 152(6) :1270–1284, March 2013. doi : 10.1016/j.cell.2013.02.001. URL http://www.ncbi.nlm.nih.gov/pubmed/23498936.

Donna Garvey Brickner, Sara Ahmed, Lauren Meldi, Abbey Thompson, Will Light, Matthew Young, Taylor L Hickman, Feixia Chu, Emmanuelle Fabre, and Jason H Brickner. Transcription factor binding to a DNA zip code controls interchromosomal clustering at the nuclear periphery. *Developmental Cell*, 22(6) :1234–1246, June 2012. doi : 10.1016/j.devcel.2012.03.012. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22579222&retmode=ref&cmd=prlinks.

Maïwen Caudron-Herger and Karsten Rippe. Nuclear architecture by RNA. *Current Opinion in Genetics & Development*, 22(2) :179–187, April 2012. doi : 10.1016/j.gde.2011.12.005. URL http://linkinghub.elsevier.com/retrieve/pii/S0959437X11001870.

Giacomo Cavalli and Tom Misteli. Functional implications of genome topology. *Nature structural & molecular biology*, 20(3) :290–299, March 2013. doi : 10.1038/nsmb.2474. URL http://www.nature.com/doifinder/10.1038/nsmb.2474.

Madoka Chinen and Tokio Tani. Diverse functions of nuclear non-coding RNAs in eukaryotic gene expression. *Frontiers in bioscience (Landmark edition)*, 17 : 1402–1417, 2012. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22201811&retmode=ref&cmd=prlinks.

Zhiming Dai and Xianhua Dai. Nuclear colocalization of transcription factor target genes strengthens coregulation in yeast. *Nucleic acids research*, 40(1) :27–36, January 2012. doi : 10.1093/nar/gkr689. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=21880591&retmode=ref&cmd=prlinks.

Annette Denker and Wouter de Laat. The second decade of 3C technologies : detailed insights into nuclear organization. *Genes & development*, 30(12) :1357–1382, June 2016. doi : 10.1101/gad.281964.116. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=27340173&retmode=ref&cmd=prlinks.

Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296) : 363–367, May 2010. doi : 10.1038/nature08973. URL http://www.nature.com/doifinder/10.1038/nature08973.

Tina W Han, Masato Kato, Shanhai Xie, Leeju C Wu, Hamid Mirzaei, Jimin Pei, Min Chen, Yang Xie, Jeffrey Allen, Guanghua Xiao, and Steven L McKnight. Cell-free Formation of RNA Granules : Bound RNAs Identify Features and Components of Cellular Assemblies. *Cell*, 149(4) :768–779, May 2012. doi : 10.1016/j.cell.2012.04.016. URL http://linkinghub.elsevier.com/retrieve/pii/S0092867412005132.

Ming Hu, Ke Deng, Siddarth Selvaraj, Zhaohui Qin, Bing Ren, and Jun S Liu. HiCNorm : removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23) :3131–3133, December 2012. doi : 10.1093/bioinformatics/bts570. URL http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/bts570.

Michael R Hübner and David L Spector. Chromatin dynamics. *Annual Review of Biophysics*, 39 :471–489, 2010. doi : 10.1146/annurev.biophys.093008.131348. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=20462379&retmode=ref&cmd=prlinks.

Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nature methods*, 9(10) :999–1003, October 2012. doi : 10.

1038/nmeth.2148. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22941365&retmode=ref&cmd=prlinks.

Masato Kato, Tina W Han, Shanhai Xie, Kevin Shi, Xinlin Du, Leeju C Wu, Hamid Mirzaei, Elizabeth J Goldsmith, Jamie Longgood, Jimin Pei, Nick V Grishin, Douglas E Frantz, Jay W Schneider, She Chen, Lin Li, Michael R Sawaya, David Eisenberg, Robert Tycko, and Steven L McKnight. Cell-free Formation of RNA Granules :Low Complexity Sequence Domains Form Dynamic Fibers within Hydrogels. *Cell*, 149(4) :753–767, May 2012. doi : 10.1016/j.cell.2012.04.017. URL http://dx.doi.org/10.1016/j.cell.2012.04.017.

Hajime Kimura, Yasutoshi Shimooka, Jun-ichi Nishikawa, Osamu Miura, Shigeru Sugiyama, Shuji Yamada, and Takashi Ohyama. The genome folding mechanism in yeast. *Journal of biochemistry*, 154(2) :137–147, August 2013. doi : 10.1093/jb/mvt033. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23620598&retmode=ref&cmd=prlinks.

Kai Kruse, Sven Sewitz, and M Madan Babu. A complex network framework for unbiased statistical analyses of DNA-DNA contact maps. *Nucleic acids research*, 41(2) :701–710, January 2013. doi : 10.1093/nar/gks1096. URL http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gks1096.

Erez Lieberman-Aiden, Nynke L van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, Bryan R Lajoie, Peter J Sabo, Michael O Dorschner, Richard Sandstrom, Bradley Bernstein, M A Bender, Mark Groudine, Andreas Gnirke, John Stamatoyannopoulos, Leonid A Mirny, Eric S Lander, and Job Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950) :289–293, October 2009. doi : 10.1126/science.1181369. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=19815776&retmode=ref&cmd=prlinks.

Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469) :59–64, October 2013. doi : 10.1038/nature12593. URL http://www.nature.com/doifinder/10.1038/nature12593.

Timothy J Nott, Timothy D Craggs, and Andrew J Baldwin. Membraneless organelles can melt nucleic acid duplexes and act as biomolecular filters. *Nature chemistry*, 8(6) :569–575, June 2016. doi : 10.1038/nchem.2519. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=27219701&retmode=ref&cmd=prlinks.

Jonas Paulsen, Tonje G Lien, Geir Kjetil Sandve, Lars Holden, Ornulf Borgan, Ingrid K Glad, and Eivind Hovig. Handling realistic assumptions in hypothesis testing of 3D co-localization of genomic elements. *Nucleic acids research*, 41(10) :5164–5174, May 2013. doi : 10.1093/nar/gkt227. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=23571755&retmode=ref&cmd=prlinks.

Anthony D Schmitt, Ming Hu, and Bing Ren. Genome-wide mapping and analysis of chromosome architecture. *Nature Publishing Group*, pages 1–13, September 2016. doi : 10.1038/nrm.2016.104. URL http://dx.doi.org/10.1038/nrm.2016.104.

Heiko Schober, Véronique Kalck, Miguel A Vega-Palas, Griet Van Houwe, Daniel Sage, Michael Unser, Marc R Gartenberg, and Susan M Gasser. Controlled exchange of chromosomal arms reveals principles driving telomere interactions in yeast. *Genome research*, 18(2) :261–271, February 2008. doi : 10.1101/gr.6687808. URL http://www.genome.org/cgi/doi/10.1101/gr.6687808.

T Schubert, M C Pusch, S Diermeier, V Benes, E Kremmer, A Imhof, and G Langst. Df31 Protein and snoRNAs Maintain Accessible Higher-Order Structures of Chromatin. 48(3) :434–444, November 2012. doi : 10.1016/j.molcel.2012.08.021. URL http://www.ncbi.nlm.nih.gov/pubmed/23022379.

Harianto Tjong, Ke Gong, Lin Chen, and Frank Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome research*, 22(7) :1295–1305, July 2012. doi : 10.1101/gr.129437.111. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22619363&retmode=ref&cmd=prlinks.

Daniela M Witten and William Stafford Noble. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic acids research*, 40(9) :3849–3855, May 2012. doi : 10.1093/nar/gks012. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22266657&retmode=ref&cmd=prlinks.

Hua Wong, Hervé Marie-Nelly, Sébastien Herbert, Pascal Carrivain, Hervé Blanc, Romain Koszul, Emmanuelle Fabre, and Christophe Zimmer. A predictive computational model of the dynamic 3D interphase yeast nucleus. *Current biology : CB*, 22(20) :1881–1890, October 2012. doi : 10.1016/j.cub.2012.07.069. URL http://eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=pubmed&id=22940469&retmode=ref&cmd=prlinks.

Eitan Yaffe and Amos Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature genetics*, 43(11) :1059–1065, November 2011. doi : 10.1038/ng.947. URL

http://www.nature.com/doifinder/10.1038/ng.947.

Annexe A

CONTRIBUTION À L'ARTICLE PRÉSENTÉ

Ilunga Benjamin Matala :Conception de la méthode d'analyse principale « rNorm »et de l'approche de randomisation par région. Rédaction entière du manuscrit.

Stephen W. Michnick : Conception du projet initiale d'étude des propriété de la structure du génome en relations aux protéines et facteurs de transcriptions.

Nicolas Lartillot : Co-conception du projet initiale d'étude des propriété de la structure du génome en relations aux protéines et facteurs de transcriptions. Révisions de la rédaction.