Inferring the 3D architecture of the genome

Nelle Varoquaux

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY

MINES ParisTech

École doctorale n° 432: Sciences des métiers de l'ingénieur

Doctorat ParisTech

THÈSE

pour obtenir le grade de docteur délivré par

l'École nationale supérieure des mines de Paris

Spécialité doctorale "Bio-informatique"

présentée et soutenue publiquement par

Nelle Varoquaux

le 3 décembre 2015

Inferring the 3D structure of the genome.
Inférence de la structure tri-dimensionelle du génome.

Directeur de thèse : **Jean-Philippe Vert**

Jury

M. Emmanuel Barillot,	Directeur de Recherche	Président
M. William S. Noble,	Professor	Examinateur
M. Marc Marti-Renom,	Research Professor	Rapporteur
M. Julien Mozziconacci,	Maître de conférences	Examinateur
M. Stéphane Robin,	Directeur de Recherche	Rapporteur
M. Jean-Philippe Vert,	Directeur de Recherche	Examinateur

MINES ParisTech
Centre de Bio-Informatique (CBIO)
35 rue Saint-Honoré, 77300 Fontainebleau, France

THESE

Abstract

The structure of DNA, chromosomes and genome organization is a topic that has fascinated the field of biology for many years. Most research focused on the one-dimensional structure of the genome, studying the linear organizations of genes and genomes and their link with gene expression and regulation, splicing, DNA methylation, ... Yet, spatial and temporal three-dimensional (3D) genome architecture is also thought to play an important role in many genomic functions.

Chromosome conformation capture (3C) based methods, coupled with next generation sequencing (NGS), allow the measurement, in a single experiment, of genome wide physical interactions between pairs of loci, thus enabling to unravel the secrets behind 3D organization of genomes. These new technologies have paved the way towards a systematic and genome wide analysis of how DNA folds into the nucleus and opened new avenues to understanding many biological processes, such as gene regulation, DNA replication and repair, somatic copy number alterations and epigenetic changes. Yet, 3C technologies, as any new biotechnology, now poses important computational and theoretical challenges for which mathematically well grounded methods need to be developped.

In this thesis, we attempt to address some of the challenges faced while analysing such data.

The first chapter is dedicated to developping a robust and accurate method to infer a 3D model of the genome from Hi-C data. Previous methods often formulated the inference as an optimization problem akin to *multidimensional scaling* (MDS) based on an *ad hoc* conversion of contact counts into euclidean *wish distances*. Chromosomes are modeled with a beads-on-a-string model, and the methods attempt to place the beads in a 3D euclidean space to fullfill a number of, often non convex, constraints and such that the pairwise distances between beads are as close as possible to the corresponding *wish distances*. These approaches rely on dubious hypotheses to convert contact counts into *wish distances*, challenging the accuracy of the final 3D model. Another limitation is the MDS formulation which is only intuitively motivated, and not grounded on a clear statistical model. To alleviate these problems, our method models contact counts as a Poisson distribution where the intensity is a decreasing function of the spatial distance between elements interacting. We then formulate the 3D structure inference as a maximum likelihood problem. We demonstrate that our method infers robust and stable models across resolutions and datasets.

The second chapter focuses on the genome architecture of the *P. falciparum*, a small parasite responsible for the deadliest and most virulent form of human malaria. This

project was biologically driven and aimed at understanding whether and how the 3D structure of the genome related to gene expression and regulation at different time points in the complex life cycle of the parasite. In collaboration with the Le Roch lab and the Noble lab, we built 3D models of the genome at three time points which resulted in a complex genome architecture indicative of a strong association between the spatial genome and gene expression.

The last chapter tackles a very different question, also based on 3C-based data. Initially developped to probe the 3D architecture of the chromosomes, Hi-C and related techniques have recently been re-purposed for diverse applications: *de novo* genome assembly, deconvolution of metagenomic samples and genome annotations. We describe in this chapter a novel method, Centurion, that jointly infers the locations of all centromeres in a single yeast genome from Hi-C data, using the centromeres' tendency to strongly colocalize in the nucleus. Indeed, centromeres are essential for proper chromosome segregation, yet, despite extensive research, centromere locations are unknown for many yeast species. We demonstrate the robustness of our approach on datasets with low and high coverage on well annotated organisms. We then predict centromere coordinates for 6 yeast species that currently lack those annotations.

During the course of my PhD, I have collaborated on several other projects, for which my contributions were minor and thus which I will not describe in the main part of this manuscript. The corresponding papers can be found in appendix. The first project consists in the development of a complete pipeline to preprocess Hi-C data from reads to normalized contact counts. I have worked on a fast and memory efficient python implementation of the normalization. Despite its simplicity, it is to our knowledge the fastest implementation existing so far. The second paper is a review of the epigenetics of the *P. falciparum* following our first paper on the 3D structure of this parasite. The last project extends the Hi-C protocol to detect interactions between triplets and quadruplets of loci in addition to the usual pairwise interactions. My contribution to this last paper is the development of a method to infer the 3D structure of polyploid method which we applied to the KBM7 nearly haploid human cell line.

Résumé

La structure de l'ADN, des chromosomes et l'organisation du génome sont des sujets fascinants du monde de la biologie. La plupart de la recherche s'est concentrée sur la structure unidimensionnelle du génome, étudiant comment les gènes et les chromosomes sont organisés, et le lien entre l'organisation unidimensionnelle et la régulation des gènes, l'épissage, la méthylation, ... Cependant, le génome est avant tout organisé dans un espace euclidien tridimensionnel, et cette structure 3D, bien que moins étudiée, joue, elle aussi, un rôle important dans la fonction génomique de la cellule.

La capture de la conformation des chromosomes (3C) et les méthodes qui en sont dérivées, associées au le séquençage à haut débit (NGS) mesurent désormais en une seule expérience des interactions physiques entre paire de loci sur tout le génome, permettant ainsi aux chercheurs de découvrir les secrets de l'organisation des génomes. Ces nouvelles technologies ouvrent la voie à des études systématiques et globales sur le repliement de l'ADN dans le noyau ainsi qu'à une meilleure étude et compréhension de beaucoup de processus biologiques, comme la régulation des gènes, la replication et la réparation de l'ADN, les altérations du nombre de copies somatiques ainsi que les changements épigénétiques. Cependant, ces nouvelles méthodes 3C, comme toute nouvelle technologie, sont accompagnées de nombreux défis computationnelles et théoriques.

Dans cette thèse, nous cherchons à relever un certain nombre de ces défis.

Le premier chapitre est dédié au développement d'une méthode robuste et précise pour inférer un modèle tridimensionnel à partir de données Hi-C. Les méthodes développées précédemment formulent souvent ce problème d'inférence comme un problème d'optimisation basé sur le *positionnement multidimensionnel* (en anglais *multidimensional scaling*) (MDS), reposant sur une dérivation *ad hoc* des fréquences d'interaction en distances euclidiennes. Les chromosomes sont modélisés comme des colliers de perles, lesquels doivent être placés dans un espace euclidien de dimension 3 de telle sorte à non seulement respecter un certain nombre de contraintes (souvent non convexes) mais aussi de manière à positionner les perles de façon à ce que les distances entre elles soient les plus proches des distances dérivées des fréquences d'interaction. Ces approches reposent sur des hypothèses contestables pour transformer fréquences d'interaction en distances euclidiennes, soulevant ainsi un doute sur la validité du modèle final obtenu. Une autre limitation de ces méthodes est la formulation du problème d'inférence sous forme MDS, justifiée non pas par un modèle statistique, mais uniquement par l'intuition. Pour pallier ces problèmes, notre méthode modélise les fréquences d'interaction comme une distribution de Poisson dont l'intensité est une fonction de la distance euclidienne entre paires de loci :

nous formulons ainsi l'inférence de la structure 3D comme un problème de maximum de vraisemblance. Nous montrons que notre méthode infère des modèles plus robustes et plus stables selon les données et les résolutions de celles-ci.

Le deuxième chapitre est consacré à l'étude de l'architecture du *P. falciparum*, un petit parasite responsable de la forme la plus virulente et mortelle de la malaria. Ce projet, dont l'objectif était avant tout de répondre à une question biologique, cherchait à comprendre comment l'architecture 3D du génome du *P. falciparum* est liée à l'expression et la régulation des gènes à différent moments du cycle cellulaire du parasite. En collaboration avec les équipes de Karine Le Roch et de William Noble, spécialisées respectivement dans l'étude du *P. falciparum*, et dans le développement de méthode computationnelle pour étudier, entre autre, la structure 3D du génome, nous avons construit des modèles de l'organisation du génome à trois moments du cycle cellulaire du parasite. Ceux-ci révèlent que le génome est replié dans le noyau dans une structure complexe, où de nombreux nombreux éléments génomiques colocalisent: centromères, télomères, ADN ribosomal, famille de gènes, ... Cette architecture indique une forte association entre l'organisation spatiale du génome et l'expression des gènes.

Le dernier chapitre répond à une question très différente, mais aussi lié à l'étude des données 3C. Celles-ci, initialement développées pour étudier la structure tridimensionnelle du génome, ont été récemment utilisées pour des applications très diverses: l'assemblage de génomes *de novo*, la déconvolution d'échantillons métagénomiques et l'annotation de génomes. Nous décrivons dans ce chapitre une nouvelle méthode, Centurion, qui infère conjointement la position de tous les centromères d'un organisme, en utilisant la propriété qu'ont les centromères à colocaliser dans le noyau. Cette méthode est donc une alternative aux méthodes de détection de centromères classiques, qui, malgré des années de recherche et un enjeu économique certain, n'ont pu identifier la position des centromères dans un certain nombre d'espèces de levure. Nous démontrons dans ce projet la robustesse et la précision de notre approche sur des jeux de données à haute comme à basse couverture. Nous prédisons par ailleurs la position des centromères dans 6 espèces qui n'avaient pour l'instant aucune annotation.

J'ai par ailleurs au cours de ma thèse travaillé sur un certain nombre de projets pour lesquels ma contribution a été mineure et que je ne décrirai pas dans ce manuscript, mais dont les papiers peuvent être trouvés en appendice. Le premier projet consiste au développement d'un nouvel outil permettant le pre processing des données Hi-C afin de construire et de normaliser les cartes de fréquences d'interaction à partir des données brutes de séquençage. Ma contribution a été l'implémentation en python d'une version optimisée à la fois en mémoire et en temps de calcul de la normalisation. Cette implémentation, bien que très simple et non parallélisée, est à notre connaissance la

plus performante existant à l'heure actuelle. La deuxième publication est une revue de l'épigénétique du *P. falciparum* suite à notre premier publication sur le sujet. Le troisième papier étend la méthode Hi-C afin de détecter, en plus des paires d'interactions, des interactions entre trois et quatre éléments. Ma contribution à ce dernier projet a été le développement d'une méthode permettant l'inférence de la structure 3D de génomes polyploïdes.

Acknowledgements

First I would like to express my deep gratitude to Jean-Philippe Vert for supervising my work and sharing his expertise during these three years, for having welcomed me in his research team and giving me the opportunity to work in two prestigious and stimulating institutes: Institut Curie and Mines ParisTech. I would also like to thank William Noble for suggesting the subject and the collaborations from which my PhD relied on, and mentoring me during the past three years.

I would like to thank Stéphane Robin and Marc Marti-Renom for accepting to review my thesis, and sharing interesting comments and discussions with me. I am also grateful to Emmanuel Barillot, William Noble and Julien Mozziconacci for accepting to be part of the jury.

Many people have contributed, directly or indirectly to the work presented in this thesis, and I would like to thank them here: Nicolas Servant, for whom I am grateful for his availablity to answer my questions, for sharing his expertise on Hi-C; Karine Le Roch and the amazing people from her team, Sebastiaan Le Bol and Evelien Bunnik, for providing the collaboration, support and ideas behind our project on *P. falciparum*, once again William Noble and members of his team, Ferhat Ay, Kate Cook and Wenxiu Hu, for their expertise on the analysis of the 3D structure of the genome and our fruitful collaborations; and all my other collaborators: Édith Heard, Éric Viara, Chong-Jian Chen, Job Dekker, Bryan Lajoie, Jacques Prudhomme, Maitreya Dunham, Ivan Liachko, Jay Shendure, Josh Burton. I'd like to thank all the members and former members of the CBIO: Alice Schoenauer-Sebag for sharing her experience (and frustration) on using the High Performance Ressources at our disposal, but also her passion for the opéra; Toby Hocking and Anne-Claire Haury, without whom the CBIO was just not quite the same; Elsa Bernard, Erwan Scornet, Véronique Stoven for all the stories shared around a coffee, Thomas Walter, Yunlong Jiao, Chloé Azencott, Matahi Moarii, Pierre Chiche, Xiwei Zhang, Victor Bellon, Judith Abécassis, Svetlana Gribkhova, Nino Shervashidze, Andrea Cavagnino, Émile Richard, Édouard Pauwels, Kevin Vervier, Olivier Collier, and Azadeh Khaleghi.

I'd like to thank my parents supporting with me during those three years, my brother Gaël for convincing me to start my studies again to deepen my knowledge of machine learning.

Supplementary Notes

Supplementary Figures

Supplementary Tables

Chapter 1

Introduction and related work

Résumé

L'architecture spatiale et temporelle du génome joue un rôle important dans beaucoup de fonctions génomiques, mais est cependant à l'heure actuelle peu comprise. Le développement récent du protocol Hi-C, qui permet en une seule expérience de mesurer les fréquences d'interactions entre paire de loci sur tout le génome, ouvre la porte à une étude plus systématique de la structure tridimensionnelle du génome. Dans ce chapitre, nous introduisons les concepts sous-jacents à la capture de la conformation des chromosomes, la structure de l'ADN et aux méthodes d'inférence de l'architecture 3D du génome.

Abstract

The spatial and temporal genome architecture is thought to play an important role in many genomic functions, but is yet poorly understood. Recently, the development of the Hi-C protocol, which allows in a single experiment to assess genome wide physical interactions between pairs of loci, has paved the way for a systematic analysis of the 3D structure of DNA. We aim in this chapter at providing some background on chromosome conformation capture, the structure of DNA and the field of 3D architecture inference.

§ 1 Peeking under the hood of genome architecture

Methods to investigate the 3D structure of the genome fall broadly into two categories: bio imaging techniques and biochemical protocols. In the first category, light microscopy allows single cell visualization of specific loci and enables live cell imaging, sometimes at very high resolution [Cremer and Cremer, 2010]. Yet, these techniques limit studies to a very small number of loci. On the other hand, biochemical protocols, such as chromosome conformation capture (3C) and its derivatives, enable to measure physical interaction between DNA fragments [Dekker et al., 2002], but performing single cell experiments is troublesome, and tracking live cell impossible. To understand how DNA fold into a nucleus, one has to juggle both technologies. In this thesis, we are mostly interested in analysing 3C-based datasets.

§ 1.1 3C, 4C, 5C and Hi-C data

In recent years, the technique of chromosome conformation capture (3C) [Dekker et al., 2002], which identifies physical contacts between different genomic loci and yields information about their relative spatial distance in the nucleus, has paved the way for the systematic analysis of the 3D structure of DNA. 3C techniques and its derivatives are based on 5 experimental steps [Lieberman-Aiden et al., 2009, Kalhor et al., 2011].

- **Cross-linking** : results in the cross-linking of DNA segments to proteins and to cross-linking of proteins with each other (Figure 1-A).
- **Restriction digest** A restriction enzyme is added in excess to the cross-linked DNA (Figure 1-B). The restriction enzyme will cut the DNA at specific nucleotide sequences, separating the non-cross-linked DNA from the cross-linked chromatin. Recognition sequences in DNA differ from each restriction enzyme, producing different lengths and sequences of strands. The selection of the restriction enzyme depends on the type of studies targeted in the experiment.
- **Intramolecular Ligation** The third step is an intramolecular ligation step. DNA fragments are joined together (Figure 1-C). There are two major types of ligation junctions: the first is the ligation of two neighboring DNA fragments, and the second is the junction that is formed when ligating one end of the fragment to the other end of the same fragment.
- **Reverse Cross-links** The fourth step consists of reversing the first step: the reversal of cross-links (Figure 1-D).

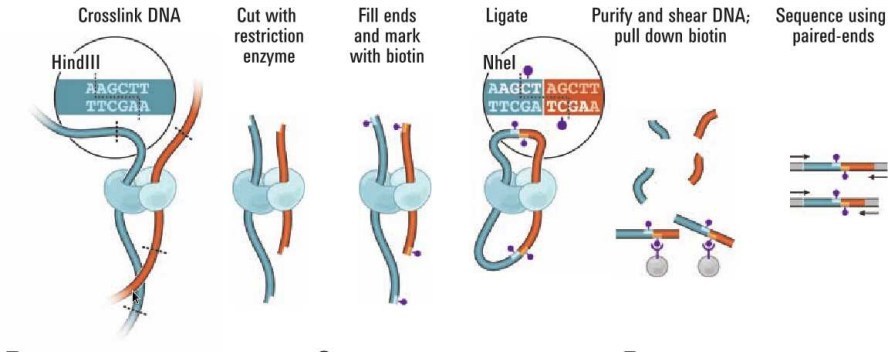


FIGURE 1: **Hi-C Protocol.** The procedure relies on cross linking, restriction enzymes digestions, intra molecular ligation, deproteinization and deep sequencing. Reads are then aligned to the reference genome, and binned at $10kb$, $40kb$ or $100kb$ depending on coverage.

- **Quantitation** Polymerase chain reaction (PCR) is used to amplify the DNA copies and to assess the frequencies of the fragments of interest, which are then sequenced (Figure 1-E).

After paired-end sequencing, each pair of reads can be associated to one [Lieberman-Aiden et al., 2009] or several [Ay et al., 2015b] DNA interactions. We can then create a symmetric matrix of integers, for which rows and columns corresponds to a specific genomic window and entries correspond to the number of times locus i and j were observed to contact on another. We denote by C the interaction frequency matrix, and c_{ij} the interaction frequency between locus i and locus j.

These protocols are complex, and yield highly biased interaction frequencies [Imakaev et al., 2012, Cournac et al., 2012, Yaffe and Tanay, 2011]. Imakaev et al. [2012] proposes a simple iterative method, called ICE, to normalize the data. In short, the authors assume that the bias of each entry c_{ij} of the matrix can be written as the product of two biases β_i and β_j corresponding to biases induced by loci. Hence, we can write $c_{ij} = \beta_i \beta_j p_{ij}$, where p_{ij} is the probability of locus i interacting with locus j. Thus, $\sum_i p_{ij} = 1$. This is a non convex optimization problem that can be solved exactly by an iterative process. To avoid degeneracies, we filter out the top 2% sparse loci from our entry matrix before applying ICE (this value needs to be adapted to each dataset). To give an intuition, this method projects each vector of interactions onto the ℓ_1 unit ball. In practice, it yields an expected interaction frequency count: kp_{ij}, where k is the average interaction frequency other all pairs of loci.

Though still quite recent, chromosome conformation capture and its genome wide derivatives are now widely used to discover how DNA folds in a bunch of different organisms [Duan et al., 2010, Sexton et al., 2012, Tanizawa et al., 2010, Ay et al., 2014b]. The challenge is now to increase the Hi-C resolution, using very large data sets with deeper

sequencing [Rao et al., 2014, Jin et al., 2013]. As any genome-wide sequencing data, Hi-C usually requires several millions or billions of paired-end sequencing reads, depending on genome size and on the desired resolution. Managing these data thus requires optimized bioinformatics workflows able to extract the contact frequencies in reasonable computational time and with reasonable storage requirements. The overall strategy to analyze Hi-C data is converging among recent studies and summarized in Lajoie et al. [2015]. Our collaborators and we have built HiC-Pro (see Appendix C, an easy-to-use and complete pipeline to process Hi-C data from raw sequencing reads to the normalized contact maps. Once these processing steps are done, one can finally proceed to the study of genome organization and DNA folding from Hi-C data in an attempt to unfold the mysteries of genome architecture.

§ 2 The study of chromosome organization

The study of chromosome organization based on contact count maps broadly falls into two categories: model-based studies and data-driven studies. The former methods consider the polymer nature of DNA to leverage the theoretical and computational work done in statistical physics of polymers to build with as few assumptions as possible many chromosome conformations. Those chromosome conformations are then used to compare against experimental data, such as Hi-C contact count matrices, in order to iteratively improve the models. These models offer mechanistical insights into the folding of DNA. The latter approaches use the experimental data to infer 3D models, by typically minizing a cost function ensuring the models are as consistent as possible with the data. These data driven models and analysis are the primary focus of this thesis.

Though we here review some of the methods used to study and build models, this is a very incomplete view of a blooming field. Rosa and Zimmer [2014] provide a more thorough (but again incomplete) overview of computational models of genome architectures.

§ 2.1 DNA as a polymer

Polymer physics divide homopolymers (polymers with identical monomers) into three main types, which are then extended to build more complex models: (1) the *random coil*, (2) the *swollen coil*, (3) the *equilibrium polymer*. These polymers are characterized by relationships such as the one between the size of a polymer subchain $L(s)$ as a function of its lengths s, between the size of the polymer $L(N)$ and the total length of this polymer N, or between the contact probability between monomers $P(s)$ and the linear distance

between monomers s. DNA being a polymer, each pair of nucleic acid forms a monomer, and the distance s is the genomic distance between two loci.

The *random coil* corresponds to an unconstrained polymer, best described by a random walk. A *random coil* of length N has an expected size of $N^{1/2}$, and so has any of its subchain: $L(s) \sim s^{1/2}$. The contact probability between two monomers is $P(s) \sim s^{-3/2}$. These relationships lead to a low density polymer, where contact between monomers is sparse. The modeling of the *random coil* does not exclude the volume occupied by monomers: when taking in account that monomers can not occupy the same chain, one obtains a new polymer model known as the *swollen coil*, best described as a self avoiding random walk. This type of polymer occupies a larger space: $L(N) \sim N^{\frac{3}{5}}$.

If the polymer is constrained in a small volume, the polymer folds into an *equilibrium globule* state. This polymer behaves as a random walk, until it bounces of the boundary of the constrained space, and starts another random walk inside the confined volume. The expected size of this polymer is $N^{1/3}$. The size of a subchain of a polymer follows the relationship: $L(s) = s^{1/2}$ for $s < N^{2/3}$ and constant elsewise: it is the same as a random coil until it plateaus. The probability of contact between two monomers is $P(s) = s^{-3/2}$ for $s < N^{2/3}$ and constant elsewise: once again, it is the same relationship as the random coil, until it becomes constant. Interestingly, this polymer is uniformely distributed in the constrained space, and the density of the polymer is independent of the total length N and the volume V.

Another interesting polymer behaviour is the *fractal globule*: when the chain is sufficiently long and the constrained volume sufficiently small, the polymer forms knotted crumples of increasing sizes. The polymer is then constrained by the available volume and the volume it itself occupies, which creates topological constraints forcing the polymer to collapse into crumples. First proposed by Grosberg et al. [1988], and further analysed by Mirny [2011], the polymer presents interesting properties: the size of any subchain follows the same law as the equilibrium globule, but without the plateau: $L(s) \sim s^{1/3}$, and the probability of contact between two monomers is inversely proportional to the linear distance that separates them: $P(s) \sim s^{-1}$.

Now that we have briefly summarized the different theoritical behaviour of polymers, let us have a closer look at the relationships we observe in practice, using DNA contact counts maps obtained through Hi-C. From figure 3, we can observe that organisms fall into two categories: the first group, composed of small genomes such as *S. cerevisae*, *P. falciparum*, behaves as an *equilibrium globule* coil, while the second group, composed of large genomes such as mammifer genomes and *A. thaliana D. drosophilae*, exhibit properties of *fractal globules*.

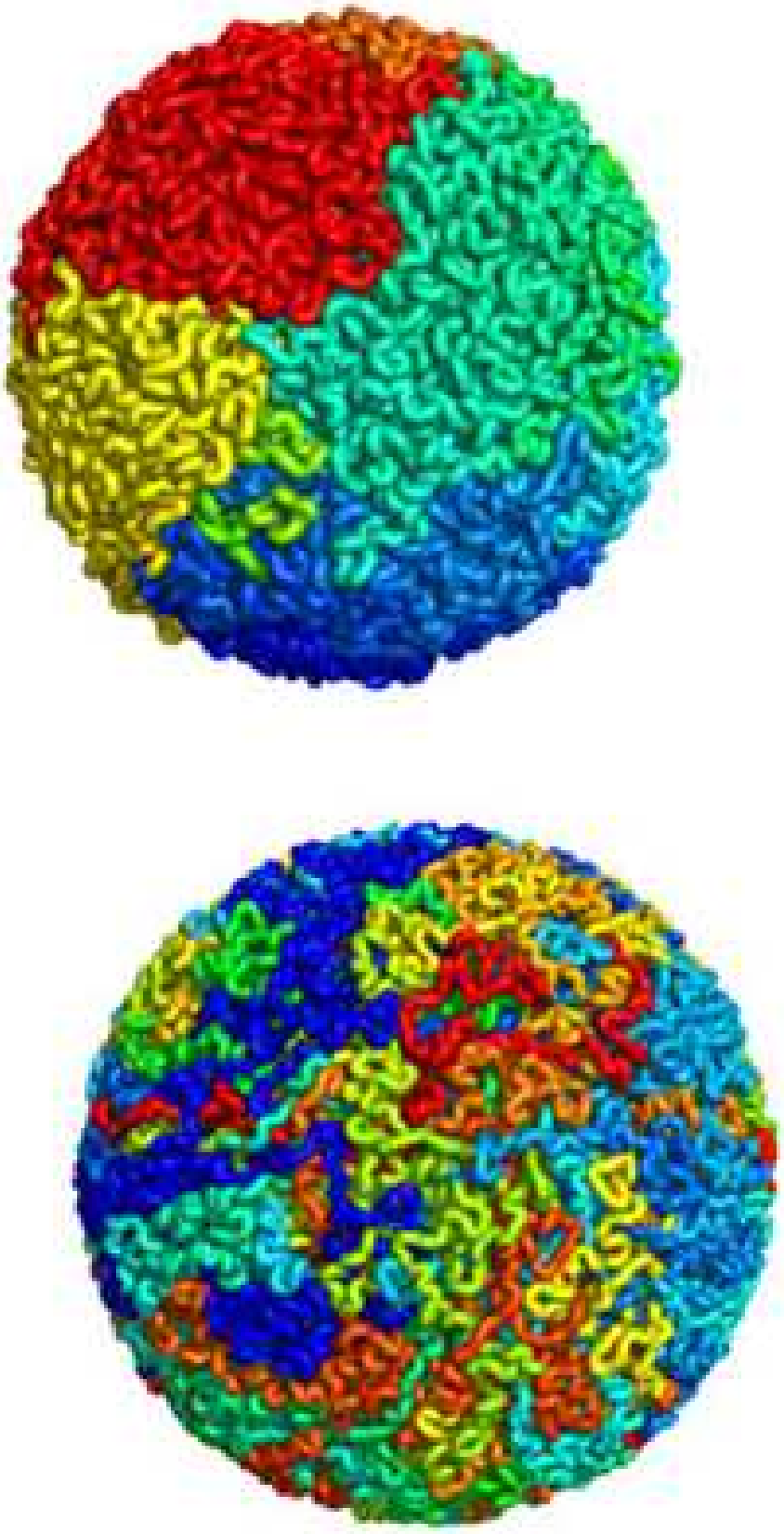


FIGURE 2: **Fractal globule versus the equilibrium globule**
This image from Mirny [2011] illustrates the difference between the *fractal globule* or crumpled globule and the *equilibrium globule*. In the first row, the fractal globule's subchain occupes a distinct territory in the nucleus, while the second row illustrates the equilibrium globule's property to occupy a wide space in the nucleus.

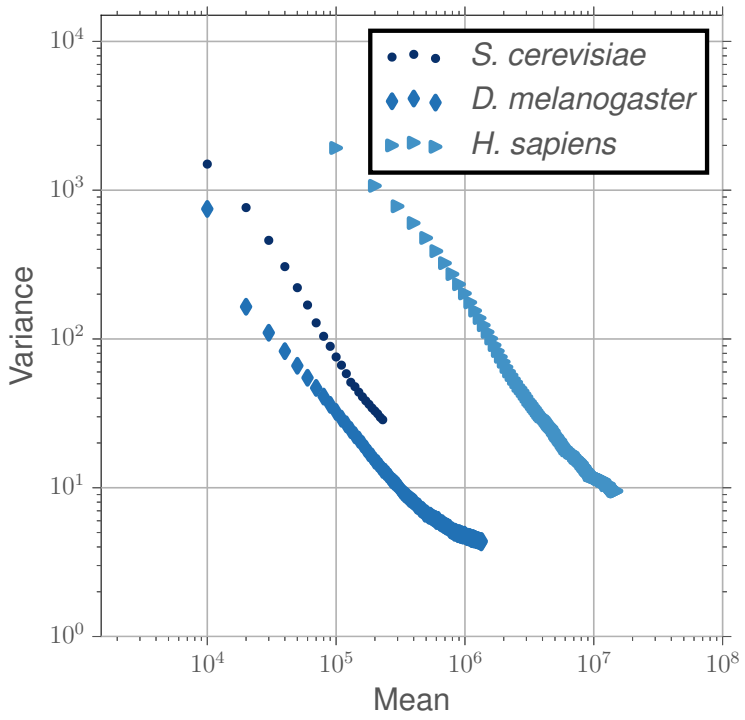


FIGURE 3: **Relationship between contact counts and genomic distances** Average contact counts as a function of genomic distance for *S. cerevisiae* [Duan et al., 2010], *D. melanogaster* [Sexton et al., 2012] and chr 1 of the KBM7 human cell line [Rao et al., 2014]. *S. cerevisiae*'s genome behaves as a *equilibrium globule*, while Sexton et al. [2012]'s *D. melanogaster* and Rao et al. [2014]'s KBM7 datasets display relationships of the *fractal crumpled globule*. Notice that *S. cerevisiae*'s average contact counts decreases more quickly with the genomic distance than *D. melanogaster*'s and KBM7's.

§ 2.2 The inference of DNA three-dimensional models

Several techniques have been developed to infer three-dimensional models of the genome from interaction counts data. They fall into three categories: the first finds an average structure by optimizing an objective function as [Tanizawa et al., 2010, Duan et al., 2010, Ben-Elazar et al., 2013]. The second samples local minima from a optimization problem leading to the study of the population of local minima [Bau et al., 2011]. The last samples the posterior distribution [Rousseau et al., 2011].

Tanizawa et al. [2010] model the 3D genome of the fission yeast (3 chromosomes) by a string of 622 beads, each bead x_i being the center of a 20kb section. The first step was to infer physical distances δ_{ij} from frequency interactions. They studied eighteen pairs of genes using FISH measurements, and fitted the Hi-C data on the distances with a non linear regression curve. The second step was to compute the coordinates of the beads, such that the distances between the beads match the inferred physical distances to the best, with additional biological motivated constraints.

Duan et al. [2010] convert the interaction frequencies into distances by examining the relationship between interaction frequencies and genomic distances. Then, a multidimensional scaling (MDS) is used to place each bead so that the wish distances are respected as well as possible.

Tanizawa et al. [2010] and Duan et al. [2010] optimize a problem of the form:

$$\begin{aligned} \underset{x_1,\ldots,x_n}{\text{minimize}} \quad & \sum_{i<j\leq n} \left(\|x_i - x_j\|_2 - \delta_{ij}\right)^2 \\ \text{subject to} \quad & \text{biological motivated non convex constraints.} \end{aligned}$$

Tanizawa et al. [2010] published one solution, but did not mention the non convexity of the problem. Hence, we assume they seeked the best local minimum Duan et al. [2010] ran the optimization process 30 times, and, observing the obtained solutions, found that they did not differ much. No formal study was done to compare the solutions.

Lesne et al. [2014] propose a method based on the classical MDS algorithm *ShRec3D*: (1) construct a graph whose vertices are the loci assessed in the Hi-C experiment, and the weights of vertices inversely proportional to the contact counts; (2) compute a matrix of shortest path between pairs of loci which we denote by the "distance" matrix; (3) apply a classical MDS on this distance matrix. This yields a fast algorithm for inferring a consensus algorithm.

Ben-Elazar et al. [2013] formulate a non metric multidimensional scaling optimization problem. They first filter the interaction count matrix so that remains only the most significant interactions. They then interpolate the missing values to obtain a smooth, symmetric, positive definite matrix. Finally, they apply a non-metric multidimensional scaling on this psd matrix.

Bau et al. [2011] use IMP (Integrative Modeling Platform), also used in nuclear magnetic resonance (NMR) microscopy to construct a 3D model of the α-globin module. Chromosomes are represented by beads, each beads linked by restraining oscillators. IMP seeks a solution at the equilibrium of those beads. Three types of restraints are used: the first corresponds to harmonic oscillators, with strengths inversely proportional to the 5C score, computed from the interaction frequencies. The second ensures that two beads cannot be too close to each other. The third ensures that two consecutive beads cannot be separated too much. The last two springs have strength only when the constraints are not fulfilled. The optimization of this problem yields different configuration with similar IMP scores. A population of 50000 structures was computed. The

10000 structures with the smaller objective function were then chosen as the population of local minima to be studied.

Rousseau et al. [2011] and Hu et al. [2013] both describe a formal probabilistic model of interaction frequencies and their relationship with physical distances. They then use a Markov Chain Monte Carlo sampling procedure to produce an ensemble of 3D structures consistant with the contact count data.

Tjong et al. [2012] construct a very simple model of the budding yeast *S. cerevisiae* by modeling chromosomes as a flexible fiber, and using additional biologically motivated constraints, such as the positioning of centromeres and telomeres, they formulate an optimization problem. Generating 200000 feasible structures, they show that Hi-C data can be fully explained by this very simple model.

Wong et al. [2013] model the budding yeast chromosomes a semi-flexible fiber constrained in a nucleus, and applies 4 sequences specific forces on this fiber to obtain certain properties: (1) centromeres are attached to a single point of the nucleus by a segment; (2) telomeres are subjected to an outward force, that pushes them towards the nuclear membrane; (3) rDNA is thicken; (4) apply a random brownian movement. This model recovers many of the known hallmarks of the 3D architecture of *S. cerevisiae*.

Nagano et al. [2013] and Paulsen et al. [2015] both propose methods to infer 3D structures from single-cell Hi-C data. The first is a constraint based modelisation: the structure is modeled as a flexible fiber as in Tjong et al. [2012], but beads are constrained to be in contact when an interaction is observed in the single-cell contact map. Paulsen et al. [2015] formulate a *manifold based optimization*, where a low rank psd matrix (and thus a distance matrix) is optimized to be as close as possible to the sparse contact count matrix. Applying classical MDS on this low rank psd matrix then yields a 3D model of the genome.

§ 3 Long range interactions

Thought mostly and initially used to study DNA folding, contact counts maps have recently been re-purposed for diverse applications: *de novo* genome assembly [Burton et al., 2013, Kaplan and Dekker, 2013], deconvolution of metagenomic samples [Burton et al., 2014, Beitel et al., 2014], and genome annotation [Marie-Nelly et al., 2014b, Varoquaux et al., 2015].

Publication	*Name*	*Consensus*	*Population*	*MDS-based*	*Statistical model*	*Availability*
Duan et al. [2010]		x		x		x
Tanizawa et al. [2010]		x		x		
Ay et al. [2014b]		x		x		
Ben-Elazar et al. [2013]		x		x		x
Varoquaux et al. [2014]	Pastis	x			x	x
Bau et al. [2011]			x			
Umbarger et al. [2011]		x				
Zhang et al. [2013]	chromSDE	x		x		x
Rousseau et al. [2011]			x		x	x
Hu et al. [2013]	Bach	x	x		x	x
Kalhor et al. [2011]			x			
Wong et al. [2012]			x			
Lesne et al. [2014]	ShRec3D	x		x		x
Trieu and Cheng [2014]		x				
Nagano et al. [2013]		x				
Paulsen et al. [2015]		x		x		x

TABLE 1: **A comparison of 3D inference methods**
In this table, we summarize properties of published methods to infer the 3D structure of the genome: (1) is it a consensus or a population based inference? (2) Is it an MDS based method? (3) or relies on a statistical modeling; (4) is the software available or not (to the best of our knowledge).

§ 3.1 *De novo* genome assembly, haplotype resolution and metagenomic sample deconvolution

De novo genome assembly is the task of assembling many short DNA reads into a whole genome. These short DNA reads can be assembled into short contigs but the process of joining short contigs into larger scaffolds is often made difficult due to the presence of repetitive sequences. Despite improvements in sequencing technology and thus the sequencing of longer reads, filling the gaps caused by these repetitive sequences in complex genome remains difficult. Burton et al. [2013] and Kaplan and Dekker [2013] propose to use the massive amount of DNA sequences produced by Hi-C to first assemble short contigs, and rely on contact counts informations between contigs to attempt to place them one relatively to the other. Indeed, the more two contigs interact, the closer in terms of genomic distances they should be (with the proper normalization in contig lengths, GC-content, mappability and so on). The contact count matrix of the ordered contigs (the "normal" contact count map) is simply a permutation of rows and columns of the contact count matrix of the unordered set of contigs. Assembling the genome consists of reordering rows and columns to obtain a suitable Hi-C contact count matrix, smooth and with a strong diagonal. Burton et al. [2013] proposes to first cluster contigs into groups that belong to the same chromosomes, then create a graph where each node is a contig, and vertex represents interactions of weight contact counts. They then apply a minimum spanning tree algorithm to identify a path in the

graph corresponding to adjacent contigs. Kaplan and Dekker [2013] and Marie-Nelly et al. [2014a] both formulate the task as finding a permutation of rows and columns to maximize a likelihood. As finding a permutation matrix is a NP-hard problem, they use heuristics to simplify the optimization problem.

Mammifer genomes contain two copies of each chromosomes, which themselves hold specific single nucleotide polymorphisme (SNP). It is sometimes interesting to identify whether SNPs or mutations are held on the same copy of the chromosome. The task of identifying which SNP belongs to which chromosome is known as resolving the haplotype. Selvaraj et al. [2013] proposes a very similar idea as before, which is that reads containing SNPs on the same homologous chromosomes interact more than reads with SNPs on the other homologous chromosome. It is thus possible to use contact count information between reads in order to determine to which homologous chromosomes SNPs belong and resolve the haplotype.

Microbiomes contain an ensemble of very small organisms in different abundances. Traditional techniques to identify which organisms are in a metagenomic sample rely on deep sequencing to produce millions of short DNA reads. These reads are then either aligned on a reference genome or used as input of a supervised learning algorithm to identify which organisms are in the sample, and in which abundance. Both of these methods need a priori knowledge of the community's composition. Recently, Burton et al. [2014] and Marbouty et al. [2014] use shotgun sequencing and Hi-C contact counts to determine which contigs belong to which organisms. Once again, the idea is very similar as before: contigs from the same organism interact amongst each other the most. Thus, a clustering algorithm that takes as input a similarity matrix (the contact counts) groups contigs from the same organism together. Once the organism's contigs are identified, one can use techniques as described before to scaffold contigs together, and therefore in a single experiment do *de novo* sequencing of many organisms, and find the composition and abundance of a metagenomic samples.

All these applications are based on a very simple and elegant idea that contact counts map hold information on the contiguity of chromosomes which can be leverage for various tasks.

§ 3.2 Genome annotations and centromeres identification

The last unusual application of Hi-C data is the annotation of genomes, and more specifically the detection of highly co-localized elements that are elsewise difficult to annotate. In particular, centromeres have proven difficult to precisely identify in many species, including highly studied and used yeasts species. Centromeres, essential for

proper chromosome segregation, have the property of clustering in 3D, and thus have very specific signal patterns in Hi-C data. Marie-Nelly et al. [2014b] proposed to annotate centromeric regions by detecting these very specific patterns in the contact maps of yeast species.

§ 4 Contributions of the thesis

This thesis brings several contributions to the fields of analyzing Hi-C data. We now review them, following the organization of the manuscript (and not in chronological order of publication).

- Chapter 2 presents a novel, stable and robust statistical method for inferring a consensus 3D model of the genome. We model contact counts as a Poisson distribution where the 3D structure is a latent variable, and formulate the inference problem as maximizing the likelihood. We show both on generated and real Hi-C data that our method is more accurate, stable and robust than previous methods. *

- Chapter 3 studies the 3D architecture of the parasite *P. falciparum* during its erythrocytic cycle and its links with gene expression. We assayed the genome architecture of the parasite at three time points to obtain high resolution contact maps which we used to construct consensus 3D models. The resulting models showed that *P. falciparum*'s genome is folded in a complex architecture, that cannot be explained by a simple volume exclusion model due to the strong co-clustering of many genomic elements in the nucleus. We observe a strong link between chromatin structure and gene expression, in particular reduced expression of genes located in spatial proximity to the repressive subtelomeric center and colocalization of distinct groups of parasite-specific genes with coordinated expression profiles. Overall, our results show that the 3D structure of the parasite is strongly correlated with gene expression during the erythrocytic cycle. This work has been done in collaboration with the Noble lab and the Le Roch lab. My contribution were (1) the 3D modeling of the genome using MDS-like approaches, (2) the comparison to the volume exclusion modeling and (3) finding the link between gene expression profiles and the 3D models using kernel CCA. Supplementaries are detailed in Appendix A.

- Chapter 4 presents work done while I was visiting the Noble lab, in collaboration with the Dunham lab and the Shendure lab. We proposed a novel method to jointly

*Our method *Pastis* is available as a free and opensource software at `http://cbio.ensmp.fr/pastis`

identify yeasts centromeres from Hi-C data, using the property centromeres have to strongly colocalize in the nucleus and thus to create a specific pattern in the contact count maps. Our method, *Centurion*, outperforms a previous published method by performing a joint optimization and using a better strategy to initialize the optimization. We show that *Centurion* is very accurate and stable both on high and low coverage datasets. [†] Supplementaries are detailed in Appendix B.

- Appendix C details a complete pipeline to preprocess Hi-C data from reads to normalized contact counts. My contribution to this pipeline is a fast and memory efficient python implementation of the normalization. Despite its simplicity, it is to our knowledge the fastest implementation existing so far.

- Appendix D presents an extension of the Hi-C protocol to detect interactions between triplets and quadruplets of loci in addition to the usual pairwise interactions. My contribution to this last paper is the development of a method to infer the 3D structure of polyploid method which we applied to the KBM7 nearly haploid human cell line.

- Appendix E reviews the multiple dimensions of epigenetic gene regulation of the *P. falciparum*.

[†]Our method *Centurion* is available as a free and opensource software at `http://cbio.ensmp.fr/centurion`

Chapter 2

A statistical approach for inferring the 3D structure of the genome

This chapter has been published in a slightly modified form in [Varoquaux et al., 2014] and presented at ISMB 2014, as joint work with Ferhat Ay, William S. Noble and Jean-Philippe Vert.

Résumé

De récents développements dans les protocoles biologiques permettent désormais de mesurer les fréquences d'interaction entre toutes les paires de loci d'un génome, et ce en une seule experience. Le défi suivant est donc d'inférer des modèles tridimensionnels du repliement des chromosomes dans le noyau de la cellule. Les inférences proposées jusqu'à présent reposent souvent sur des méthodes dites de *positionnement multidimensionnelle* (en anglais *multidimensional scaling*) (MDS), lesquelles optimisent une structure telle que les distances relatives entre éléments soient les plus proches de celles dérivées directement des fréquences d'interaction. Ces approches optimisent une fonction objective heuristique, et reposent sur des hypothèses contestables sur la biophysique de l'ADN pour trouver la fonction de

transfert entre fréquences d'interaction et distances euclidiennes, pouvant ainsi conduire à l'inférence de modèles incorrects de la structure de l'ADN.

Dans ce travail, nous proposons une nouvelle méthode pour inférer une structure 3D consensus du génome à partir de données Hi-C. Notre méthode repose sur un modèle statistique des fréquences d'interaction. Nous modélisons les fréquences d'interaction comme une distribution de Poisson dont l'intensité dépend de la distance physique entre paires de loci. Notre méthode peut automatiquement ajuster les paramètres de la fonction de transfert entre fréquences d'interaction et distances physiques, et infère un modèle expliquant au mieux les données.

Nous comparons deux variantes de notre méthode (avec ou sans optimisation des paramètres de la fonction de transfert) à quatre algorithmes basés sur MDS sur des données simulées : deux MDS dit "métriques", dont les fonctions objectives diffèrent, une version non métrique de cet algorithme, et ChromSDE qui propose une version récente et convexe du MDS. Nous démontrons que notre modèle de Poisson reconstruit des modèles plus précis que toutes les méthodes MDS, en particulier lorsque la couverture des données est faible ou que les données Hi-C sont à haute résolution, soulignant ainsi l'importance du choix de la fonction objective à optimiser. Sur des données Hi-C publiques de cellules embryoniques de souris, nous démontrons par ailleurs que les méthodes Poisson infèrent des structures plus stables, plus reproductibles et plus robustes à la fois lorsque le protocole biologique diffère, mais aussi à différentes résolutions.

Une implémentation Python de notre méthode est disponible à http://cbio.ensmp.fr/pastis.

Abstract

Motivation: Recent technological advances allow the measurement, in a single Hi-C experiment, of the frequencies of physical contacts among pairs of genomic loci at a genome-wide scale. The next challenge is to infer, from the resulting DNA-DNA contact maps, accurate three dimensional models of how chromosomes fold and fit into the nucleus. Many existing inference methods rely upon *multidimensional scaling* (MDS), in which the pairwise distances of the inferred model are optimized

to resemble pairwise distances derived directly from the contact counts. These approaches, however, often optimize a heuristic objective function and require strong assumptions about the biophysics of DNA to transform interaction frequencies to spatial distance, and thereby may lead to incorrect structure reconstruction.

Methods: We propose a novel approach to infer a consensus three-dimensional structure of a genome from Hi-C data. The method incorporates a statistical model of the contact counts, assuming that the counts between two loci follow a Poisson distribution whose intensity decreases with the physical distances between the loci. The method can automatically adjust the transfer function relating the spatial distance to the Poisson intensity and infer a genome structure that best explains the observed data.

Results: We compare two variants of our Poisson method, with or without optimization of the transfer function, to four different MDS-based algorithms—two metric MDS methods using different stress functions, a nonmetric version of MDS, and ChromSDE, a recently described, advanced MDS method—on a wide range of simulated datasets. We demonstrate that the Poisson models reconstruct better structures than all MDS-based methods, particularly at low coverage and high resolution, and we highlight the importance of optimizing the transfer function. On publicly available Hi-C data from mouse embryonic stem cells, we show that the Poisson methods lead to more reproducible structures than MDS-based methods when we use data generated using different restriction enzymes, and when we reconstruct structures at different resolutions.

Availability: A Python implementation of the proposed method is available at http://cbio.ensmp.fr/pastis.

§ 1 Introduction

Spatial and temporal three-dimensional (3D) genome architecture is thought to play an important role in many genomic functions, but is still poorly understood [van Steensel and Dekker, 2010]. In recent years, the technique of chromosome conformation capture (3C) [Dekker et al., 2002], which identifies physical contacts between different genomic loci and yields information about their relative spatial distance in the nucleus, has

paved the way for the systematic analysis of the 3D structure of DNA. Coupled with high-throughput sequencing, genome-wide conformation capture assays, broadly referred to as *Hi-C* [Lieberman-Aiden et al., 2009], have emerged as promising techniques to investigate the global structure of DNA at various resolutions. Hi-C has opened new avenues to understanding many biological processes including gene regulation, DNA replication, somatic copy number alterations and epigenetic changes [Shen et al., 2012, Ryba et al., 2010, De and Michor, 2011, Dixon et al., 2012].

A typical Hi-C experiment yields a DNA *contact map*, that is, a matrix indicating the frequency of interactions between all pairs of loci at a given resolution. A fundamental question is then to reconstruct the 3D structure of the genome from this contact map. Two general approaches have been proposed for that purpose: (i) *consensus methods* that aim at inferring a unique mean structure representative of the data and (ii) *ensemble methods* that yield a population of structures.

Consensus approaches [Duan et al., 2010, Tanizawa et al., 2010, Bau et al., 2011] model each chromosome by a chain of beads, convert the contact map frequencies into pairwise distances (which we refer as *wish distances*) using various biophysical models of DNA, and infer a 3D conformation that best matches the pairwise distances by solving a multidimensional scaling (MDS) problem [Kruskal and M., 1977]. Converting interaction counts to physical wish distances requires, however, strong assumptions which are not always met in practice. For example, this mapping may change from one organism to another [Fudenberg and Mirny, 2012], from one resolution to another [Zhang et al., 2013], from one genomic distance range to another [Ay et al., 2014a], or from one time point to another during the cell cycle [Le et al., 2013, Ay et al., 2014b].

To alleviate this problem, Zhang et al. [2013] proposed ChromSDE, a method that jointly optimizes the 3D structure and a parameter of the function that maps contact frequencies to spatial distances, in addition to modifying the objective function of MDS. Ben-Elazar et al. [2013] proposed an approach akin to *nonmetric MDS* [Kruskal, 1964], where the 3D structure and the wish distances are alternatingly optimized in an attempt to preserve coherence between the ranking of pairwise distances and the ranking of pairwise contact frequencies.

As for the ensemble methods, Rousseau et al. [2011] and Hu et al. [2013] describe two formal probabilistic models of contact frequencies and their relationship with physical distances. They then use a Markov chain Monte Carlo (MCMC) sampling procedure to produce an ensemble of 3D structures consistent with the observed contact counts. Kalhor et al. [2011] propose an optimization framework that generates a population of structures by enforcing each contact to define an active constraint in only a fraction of the inferred structures, thereby mimicking the heterogeneity of contacts coming from

each cell in the Hi-C sample. Applying a similar method to budding yeast, Tjong et al. [2012] demonstrate that a large population of structures inferred using known physical constraints of yeast genome architecture can recapitulate, to a large extent, the consensus contact map observed from Hi-C experiments.

Both consensus and ensemble models have benefits and limitations. Ensemble approaches are biologically more accurate, because Hi-C data is derived from a population of cells, each with potentially a unique 3D architecture. An inferred population of 3D structures may therefore better reflect the diversity of structures than a single consensus structure. In concordance with such ensemble methods, a recent development in Hi-C technology, assaying chromatin conformation at a single cell level, demonstrates that chromatin structure varies highly from cell to cell by modeling the single-copy X chromosomes of a male mouse cell line [Nagano et al., 2013].

However, an ensemble approach raises the question of interpretability: one often has to fall back to interpreting a mean signal from the population structure [Kalhor et al., 2011] or to selecting a few structures, representative in some way of the diversity of the population [Rousseau et al., 2011]. Consensus methods, in contrast, provide a single structure more amenable to visual inspection and analysis. This structure can be seen as a useful *model* to recapitulate the rich information captured in Hi-C data and to allow easy integration with other sources of data, such as RNA-seq, which are usually also population based. In addition, despite the stochasticity of cell-to-cell variations, certain hallmarks of genome organization observed by consensus methods, such as chromosome territories or topological domain organization, are conserved across different cells [Nagano et al., 2013, Hu et al., 2013]. Computationally, ensemble methods are more demanding than consensus methods since they need to sample from a very large dimensional space of possible structures with complicated likelihood landscapes. Optimization-based consensus methods are usually faster to converge to a local optimum, but may miss the global optimum corresponding to the best structure when the objective function is non-convex.

In this work, we focus on the consensus approach, and we propose a new method to infer a 3D structure from Hi-C data. We propose to replace the arbitrary loss function minimized by existing MDS-based approaches by a better-motivated likelihood function derived from a statistical model, similar to the one use by a previous ensemble method [Hu et al., 2013]. Specifically, our proposed method models the interaction frequency between two loci by a Poisson model (PM), the intensity of which decreases with the increasing spatial distance between the pair of loci. Similar to the problem of inferring the wish distances from interaction frequencies faced by MDS-based approaches, our model faces the difficulty of transforming spatial distances into intensities of the Poisson distribution. To solve this problem, we propose two variant methods. The first method

(PM1) uses a default transfer function motivated by a biophysical model, whereas the second method (PM2) uses a parametric family of transfer functions, the parameters of which are automatically optimized together with the 3D structure to best explain the observed data.

We compare both PM variants to four MDS-based methods, including metric MDS with two stress functions, nonmetric MDS and ChromSDE. We demonstrate on simulated data that the new models reconstruct more accurate 3D structures than all MDS-based methods, especially at low coverage and high resolution. We also assess the negative effect of using an incorrect transfer function, and we show that PM2 is able to overcome this difficulty. On real data, we show that, compared to MDS-based methods, PM1 and PM2 generate more similar models when applied to replicate experiments performed with different restriction enzymes or when applied to the same data at varying resolutions. The results suggest that the Poisson model methods we describe here provide promising alternatives to current methods for consensus DNA structure inference.

§ 2 Approach

We model chromosomes as series of beads in 3D, each bead representing a genomic window of a given length, and we denote by $\mathbf{X} = (x_1, \ldots, x_n) \in \mathbb{R}^{3\times n}$ the coordinate matrix of the structure, where n denotes the total number of beads in the genome (for example, $n = 1216$ at 10kb resolution for the yeast genome) and $x_i \in \mathbb{R}^3$ represents the 3D coordinate of the i-th bead. The Hi-C data can be summarized as an n-by-n matrix $\mathbf{c}$ in which each row and column corresponds to a genomic locus, and each matrix entry c_{ij} is a number, called the *contact frequency* or *contact count*, indicating the number of times locus i and j were observed to contact one another. The matrix is by construction square and symmetric.

§ 2.1 Data normalization

The raw contact count matrix suffers from many biases, some technical (from the sequencing and mapping) and others biological (inherent to the physical properties of chromatin) [Yaffe and Tanay, 2011, Imakaev et al., 2012]. Therefore, before inferring the 3D structure of the genome, we normalize each raw contact matrix using iterative correction and eigenvalue decomposition (ICE) [Imakaev et al., 2012], a method based on the assumption that all loci should interact equally. Due to mappability issues, some beads have zero contact counts. We remove these beads from the optimization and only try to infer the positions of beads with nonzero contact counts.

§ 2.2 MDS-based methods

§ 2.2.1 Metric MDS

Metric MDS is a classical method to infer coordinates of points given their approximate pairwise Euclidean distances [Kruskal and M., 1977]. To use MDS in the context of DNA structure inference from Hi-C data, we need to assign each pair of beads (i, j) a physical wish distance δ_{ij}—i.e., the distance that we aim to capture with our 3D model—derived from the bead pair's contact count c_{ij}. Performing this assignment requires us to decide how contact counts are transformed into physical distances. In Section § 2.4 we discuss a commonly used transformation of the form $\delta_{ij} = \gamma c_{ij}^{-3}$ if $c_{ij} > 0$ motivated by polymer physics. Metric MDS then places all the beads in 3D space such that the Euclidean distance $d_{ij}(\mathbf{X}) = \|x_i - x_j\|$ between the beads i and j is as close as possible to the wish distance δ_{ij}. Denoting by $\mathcal{D}$ the subset of indices whose distances we wish to constrain (typically, the set of pairs (i, j) with non-zero contact counts $c_{ij} > 0$), metric MDS attempts to minimize the following objective function, usually called the *raw stress*:

$$\underset{\mathbf{X}}{\text{minimize}} \qquad \sum_{(i,j)\in\mathcal{D}} \big(d_{ij}(\mathbf{X}) - \delta_{ij}\big)^2 . \tag{2.1}$$

In two previous studies that use metric MDS, Duan et al. [2010] and Tanizawa et al. [2010] infer the 3D structure of DNA from Hi-C data by solving Equation 2.1, limiting $\mathcal{D}$ to pairs of indices with statistically significant contact counts (FDR 0.01%). Both methods use additional constraints such as minimum and maximum distances between adjacent beads, minimum pairwise distances between arbitrary beads to avoid clashes, and organism-specific constraints that concern the positioning of centromeres, telomeres and ribosomal RNA coding regions. In the experiments we present here, we simply solve Equation 2.1 without any constraints but including all pairs of beads with positive counts in $\mathcal{D}$, and we call the resulting method MDS1. In general, we have observed that adding constraints related to minimal and maximal distances between beads is unnecessary, because the structures found by MDS1 typically fulfill all of these constraints (data not shown).

A drawback of the raw stress of Equation 2.1 in our context is that the quadratic form is dominated by large values, corresponding to pairs of loci with large wish distances (i.e., small contact counts). Because these counts are less reliable than large contact counts, we propose a variant of MDS1, which we call MDS2, where we weight the contribution of a pair (i, j) in the stress by a factor inversely proportional to the square wish distance

between the corresponding beads:

$$\underset{\mathbf{X}}{\text{minimize}} \quad \sum_{(i,j)\in\mathcal{D}} \delta_{ij}^{-2}\big(d_{ij}(\mathbf{X}) - \delta_{ij}\big)^2 . \tag{2.2}$$

While other weighting schemes could be proposed to decrease the influence of pairs with large wish distances, we found this formulation to be quite robust in practice. Notice that MDS2 can be thought of as a quadratic approximation of the raw stress (minimized by MDS1) applied to log-transformed distances, because in the setting $d_{ij}(\mathbf{X}) \approx \delta_{ij}$ it holds that:

$$\begin{aligned}\sum_{(i,j)\in\mathcal{D}} (\log d_{ij}(\mathbf{X}) - \log \delta_{ij})^2 &= \sum_{(i,j)\in\mathcal{D}} \log\left(\frac{d_{ij}(\mathbf{X})}{\delta_{ij}}\right)^2 \\ &\approx \sum_{(i,j)\in\mathcal{D}} \left(\frac{d_{ij}(\mathbf{X})}{\delta_{ij}} - 1\right)^2 .\end{aligned}$$

Both MDS1 and MDS2 implicitly ignore non-interacting pairs of beads (i.e., pairs with zero contact counts).

In addition to MDS1 and MDS2, we include in our benchmark ChromSDE [Zhang et al., 2013], a recently proposed method which also attempts to minimize a weighted stress function penalized by an additional term to push non-interacting pairs far from each other. In addition, ChromSDE optimizes the exponent of the transfer function that maps from contact counts to wish distances. However, it does not infer the relative positions of chromosomes. Accordingly, we compare only the reconstruction of each individual chromosome produced by each method. Note that, because intra-chromosomal counts are more reliable than inter-chromosomal counts, ChromSDE should not be penalized compared to the other methods by only considering intra-chromosomal counts.

§ 2.2.2 Nonmetric MDS (NMDS)

The derivation of the transfer function from contact counts to 3D wish distances, needed by metric MDS-based methods, relies on strong assumptions about the physics of DNA (Section § 2.4). NMDS [Shepard, 1962, Kruskal, 1964] offers an alternative way to proceed, which was proposed in the context of DNA structure inference from Hi-C data by Ben-Elazar et al. [2013]. Instead of inferring physical distances from the contact matrices, NMDS relies on the sole hypothesis that if two loci i and j are observed to be in contact more often than loci k and ℓ, then i and j should be closer in 3D space than k and ℓ. Using this hypothesis, NMDS attempts to solve the following problem:

Problem 1. Given a set of similarities c_{ij} (e.g., the contact frequency between i and j), find $\mathbf{X} \in R^{3\times n}$ such that:

$$c_{ij} \geq c_{k\ell} \Leftrightarrow \|x_i - x_j\|_2 \leq \|x_k - x_\ell\|_2 \,. \tag{2.3}$$

Equation 2.3 is known as the nonmetric constraint, or the ordinal constraint. This problem was first introduced by Shepard [1962] and formalized as an optimization problem by Kruskal [1964]. It can be solved by minimizing the cost function:

$$\underset{\mathbf{X},\Theta}{\text{minimize}} \sum_{i,j} \frac{\left(\|x_i - x_j\|_2 - \Theta(c_{ij})\right)^2}{\Theta(c_{ij})^2}, \tag{2.4}$$

with respect to the embedding $\mathbf{X}$ and the function Θ, where Θ is a decreasing function. Algorithms to solve this optimization problem involve iterating over two steps: (1) fixing Θ and minimizing the objective function with respect to $\mathbf{X}$ (hence falling back to solve MDS2), and (2) fitting Θ to the new configuration $\mathbf{X}$ subject to the ordinal constraints. This second step of the algorithm can be performed using an isotonic regression method, such as the pool adjacent violator algorithm [Best et al., 1999].

A trivial solution of this problem is to set Θ equal to 0. In this case all points will collapse on the origin. To avoid this collapse, we add additional constraints on $\mathbf{X}$ or on Θ, such as $\sum_{i,j} \|x_i - x_j\|_2 = K$ for some constant value of K.

§ 2.3 Poisson model

Instead of metric or non metric MDS-based methods, which attempt to minimize a stress function that measures a discrepancy between the wish distances and the 3D distances of the structure, we propose to cast the problem of structure inference as a maximum likelihood problem. For that purpose, we need to define a probabilistic model of contact counts parametrized by the 3D structure that we want to infer from contact count observations.

For that purpose, we take a model similar to the one used in the BACH algorithm [Hu et al., 2013] and model the contact frequencies $(c_{ij})_{(i,j)\in\mathcal{D}}$ as independent Poisson random variables, where the Poisson parameter of c_{ij} is a decreasing function of $d_{ij}(\mathbf{X})$ of the form $\beta d_{ij}(\mathbf{X})^\alpha$, for some parameters $\beta > 0$ and $\alpha < 0$. We can then express the likelihood as

$$\ell(\mathbf{X}, \alpha, \beta) = \prod_{i,j} \frac{(\beta d_{ij}^\alpha)^{c_{ij}}}{c_{ij}!} \exp(-\beta d_{ij}^\alpha) \,.$$

By maximizing the log likelihood, a new optimization problem naturally emerges from this formulation:

$$\max_{\alpha,\beta,\mathbf{X}} \quad \mathcal{L}(\mathbf{X},\alpha,\beta) = \sum_{i<j\leq n} c_{ij}\alpha \log d_{ij} + c_{ij}\log\beta - \beta d_{ij}^{\alpha} \tag{2.5}$$

With this new formulation, we can either provide the parameter α, using prior knowledge, and only optimize the structure and β (which depends on the dataset), or we can use a nonmetric approach, by inferring α. We refer to the former as PM1 and to the latter as PM2.

PM2 is solved using a coordinate-descent algorithm: first choose randomly an $\mathbf{X}$ configuration, then iterate between maximizing $\mathcal{L}$ with respect to α and β and, fixing α and β and maximizing $\mathcal{L}$ with respect to $\mathbf{X}$. In this work, we try to initialize $\mathbf{X}$ with a good approximation of the solution by first evaluating the parameters α and β using some prior knowledge and initialize $\mathbf{X}$ with the inferred structure from the MDS.

All optimization problems (MDS1, MDS2, NMDS, PM1 and PM2) were solved using IPOPT, an interior point filter algorithm [Wächter and Biegler, 2006] and the isotonic regression implementation from the Python toolbox Scikit-Learn for NMDS [Pedregosa et al., 2011].

§ 2.4 Default contact-to-distance transfer function

A prerequisite for both the MDS and the PM1 model (and for good initialization of the NMDS and PM2 methods) is a function that converts from contact counts to wish distances. Extensive previous studies of the behaviour of polymers in general and DNA in particular have yielded proposed relationships between, on the one hand, the genomic distance s and contact counts c and, on the other hand, genomic distance s and physical distances d for several classes of polymers [Grosberg et al., 1988, Lieberman-Aiden et al., 2009, Fudenberg and Mirny, 2012]. For a fractal globule polymer, representative of mammalian DNA, the contact count is inversely proportional to the genomic distance ($c \sim s^{-1}$), whereas the volume scales linearly with the subchain length ($d^3 \sim s$), from which we deduce a relationship between d and c of the form $d \sim c^{-1/3}$. For an equilibrium globule, representative of a smaller genome such as *S. cerevisae*, the relationships differ: $c \sim s^{-3/2}$ and $d \sim s^{1/2}$ up to a maximum distance, corresponding to the size of the nucleus in which the DNA is confined. Conveniently, coupling those two relationships for either type of polymer yields the same mapping between contact counts and physical distances:

$$d \sim c^{-1/3}. \tag{2.6}$$

Thus, by default we convert contact counts c_{ij} into 3D wish distances δ_{ij} using the following relationship:

$$\delta_{ij} = \gamma c_{ij}^{-1/3}, \tag{2.7}$$

where γ defines the scale of the structure. It is important to note that this relationship holds true for only a subset of the full genomic distance range and that this range varies for different genomes. In practice, we will not infer γ for the MDS and NMDS problem: the structures can easily be rescaled after convergence to match biological knowledge of the organism studied.

§ 2.5 Data

In order to test various 3D architecture inference methods, we conducted experiments on both simulated datasets and publicly available genome-wide Hi-C datasets.

For the simulation, we generated 170 data sets using the yeast genome architecture proposed by Duan et al. [2010]. Because the repetitive rDNA on yeast chromosome XII cannot be observed in practice, we discard all contacts involving these loci, and we do not infer the position of the corresponding rDNA. We generate these 170 datasets using the following model:

$$c_{ij} = P(\beta d_{ij}^{\alpha}), \tag{2.8}$$

where $\alpha = -3$ (corresponding to the theoretical exponent discussed in Section § 2.4) and β varies between 0.01 and 0.7 (0.01, 0.01, 0.02, 0.03, 0.04, 0.05, 0.06, 0.07, 0.08, 0.09, 0.1, 0.15, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7) with 10 different random generator seeds, thus obtaining 10 different datasets per parameter. The β parameter controls the number of contact counts in the datasets. A low β will yield a dataset with few counts; hence, the corresponding wish distance matrix will be less likely to be close to the true distance matrix. To estimate how noisy the generated data is, we compute the following measure of signal-to-noise ratio (SNR):

$$SNR = \frac{\sum c_{ij}}{\sqrt{\sum (\beta d_{ij}^{\alpha} - c_{ij})^2}}. \tag{2.9}$$

The numerator (the signal) corresponds to the number of counts, and the denominator (the noise) corresponds to the sum of deviation between each count and its expected value. We use this first ensemble of simulated datasets to assess the robustness to noise of the different methods. Note that in actual data, the SNR gets smaller when we sequence fewer reads or when we infer a structure at a higher resolution.

We simulated another ensemble of datasets to compare nonmetric and metric methods when the parameters provided to the different algorithms are not the correct ones. We generate 20 datasets according to Equation 2.8, with α between -4 and -2 (-4, -3.5, -3, -2.5, -2) and β between 0.4 and 0.7 (0.4, 0.5, 0.6, 0.7).

We also applied our methods to publicly available Hi-C data from mouse embryonic stem cells (mESC) [Dixon et al., 2012]. We started with the data at 20 kb resolution and considered only chromosomes 1 to 19, with both available restriction enzymes (HindIII and NcoI). We then subsampled the data at resolutions of 100 kb, 200 kb, 500 kb and 1 Mb. Note that the methods studied here infer a single copy per chromosomes, thus yielding a consensus model for both homologous chromosomes.

§ 2.6 Structure similarity measures

In order to assess the ability of a method to reconstruct a known structure from simulated data, or the stability of the reconstructed structure with respect to change in resolution or library preparation, we need quantitative measures of similarity between 3D structure. We use two such measures: the root mean square deviation (RMSD) and the distance error, which we now explain.

The RMSD is a standard way to compare two sets of structures described by their coordinates $\mathbf{X}, \mathbf{X}' \in R^{3\times n}$, widely used for example to compare protein 3D structures. It is given by:

$$RMSD = \min_{\mathbf{X}^*} \sqrt{\sum_{i=1}^{n} (\mathbf{X}_i - \mathbf{X}_i^*)^2}\,,$$

where the structure $\mathbf{X}^*$ is obtained by translating, rotating and rescaling $\mathbf{X}'$ ($\mathbf{X}^* = s\mathbf{R}\mathbf{X}' - \mathbf{t}$, where $\mathbf{R} \in R^{3\times 3}$ is a rotation matrix, $\mathbf{t} \in R^3$ is a translation vector, and s is a scaling factor). Because ChromSDE does not infer the relative position of chromosomes, the RMSD values we report below are sums of RMSDs computed independently on each chromosome.

We also directly compare the 3D distance matrices corresponding to the two structures with the distance error:

$$\text{distanceError} = \sqrt{\sum_{i,j=0}^{n} (d_{ij}(\mathbf{X}) - d_{i,j}(\mathbf{X}'))^2}\,.$$

The main difference between the optimization formulated by ChromSDE and those of the other methods is the penalty assigned to non-interacting beads. Due to this penalty, ChromSDE should recover better long distances than other MDS-based methods. This

property is not well captured by the RMSD measure, therefore, we also compute how well the distance matrix is recovered with the distance error, which assigns most of the weight to long distances. We expect that methods based on MDS, which optimize an objective function based on the distance matrix, should perform better on this measure than others.

§ 3 Results

To assess the relative strength of our new Poisson model-based methods, PM1 and PM2, we compare them to a panel of four MDS-based methods: MDS1, MDS2, NMDS and ChromSDE on simulated and real data.

§ 3.1 Simulated Hi-C data

We first tested the six methods on data simulated as explained in Section § 2.5.

§ 3.1.1 Performance as a function of SNR

We ran all six methods—MDS1, MDS2, NMDS, PM1, PM2 and ChromSDE—on the 170 simulated datasets with varying SNR levels. Our goal here is to assess how well the different methods manage to reconstruct a known 3D structure from simulated data at different SNR levels. Remember that SNR estimates how far the empirical counts differ from their expectations; in real Hi-C data, SNR typically decreases when we have fewer reads in total, or when we want to increase the resolution of the structure. In this first series of experiments, we provide the correct count-to-distance or distance-to-count transfer functions to the methods that need them (MDS1, MDS2, PM1). In this setting, for infinite SNR, all methods should consistently estimate the correct structure.

Figure 1 shows the performance of the different methods in terms of RMSD (top) and distance error (middle) as a function of the β parameter, which controls the SNR (bottom). As expected, all methods perform well when the SNR is high, but exhibit marked differences in performance for finite SNR. In the low SNR setting (SNR < 2), both PM1 and PM2 significantly outperform all MDS-based methods, in both RMSD and distance error. Interestingly, we observe no significant difference between PM1 and PM2, which shows that there is no price to pay in terms of inferred structure if we don't specify the exponent of the distance-to-count transfer function. In this setting, PM2 is able to estimate the structure accurately enough to produce a structure of the same quality as PM1. Among MDS-based methods, we see that NMDS generally outperforms

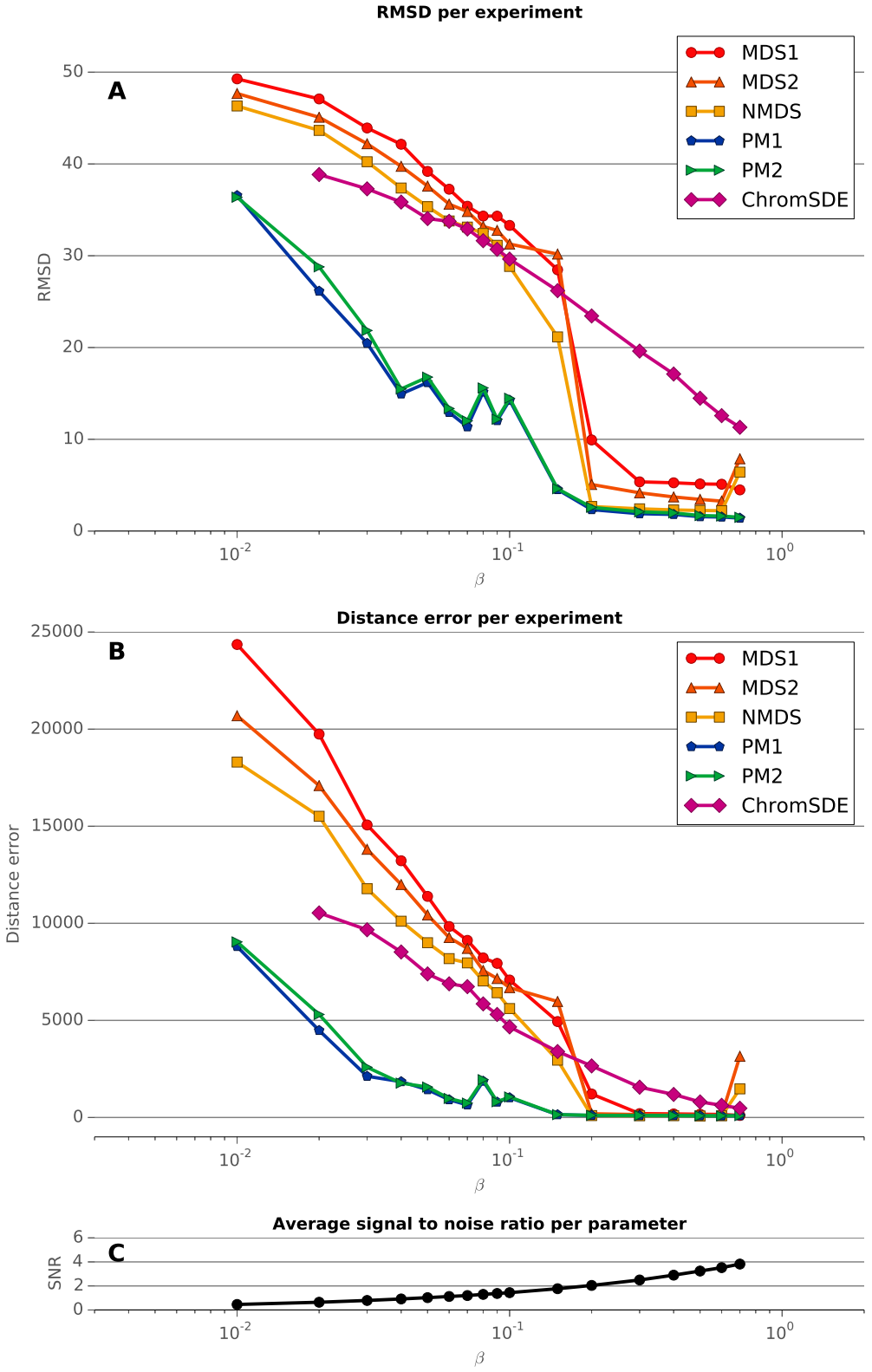


FIGURE 1: **Performance evaluation on simulated data, varying the parameter β. A** RMSD of each experiment for varying values of the parameter β. ChromSDE failed to yield consistent results for 14 experiments (It reported the wrong number of beads in the results file.), and the PM2 algorithm failed to converge at the desired precision for one experiment (It exceeded the maximum number of iterations.). **B** Distance error of each experiment for varying values of β. **C** Average SNR for each β. Higher SNR corresponds to better quality data.

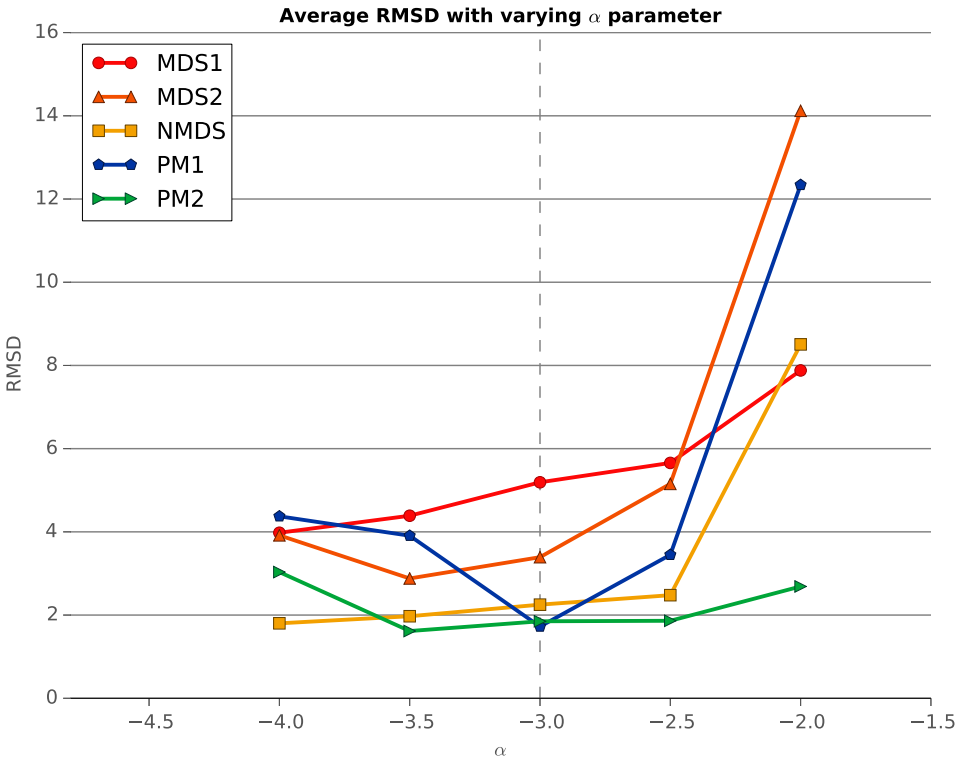


FIGURE 2: **Performance evaluation for simulated data, varying the parameter α.** The figure plots the average RMSD of the inferred structures for a range of α values. As α increases, the SNR of the dataset also increases.

MDS2, which itself outperforms MDS1. This observation highlights that in the non-asymptotic, low SNR setting, the choice of stress function influences the performance of MDS. ChromSDE performs better than other MDS-based methods on datasets with a low SNR, corresponding to datasets with low coverage and, consequently, many non-interacting pairs of beads. This may be due to the way ChromSDE explicitly handles such pairs. On the other hand, in a more favorable setting (SNR $>$ 2), ChromSDE does not perform as well as other MDS-based method; we hypothesize that when the coverage is high enough, taking into account non-interacting pairs of beads does not add any additional information. Since ChromSDE is not better than other MDS-based methods, and requires much longer to run, we do not report its performance on the next experiments and instead focus on the differences between the other MDS-based methods and the PM methods.

§ 3.1.2 Metric versus nonmetric methods: robustness to incorrect parameter estimation

Three of the methods tested, which we collectively refer to as *metric* methods, require as input a count-to-distance or distance-to-count transfer function: MDS1, MDS2 and PM1. In reality, however, the DNA may not follow the ideal physical laws underlying the default transfer function discussed in Section § 2.4, and the structures inferred from these methods may diverge from the correct one because of miss-specification of the transfer function.

To assess this phenomenon, and evaluate the robustness of the different methods (including NMDS and PM2, which automatically infer a transfer function), we now study the performance of the methods on datasets generated with varying α parameters. We therefore run the MDS1, MDS2, NMDS, PM1 and PM2 methods on the second ensemble of simulated datasets. We provide the default transfer function to all metric methods, thus inducing a miss-specification for all simulated datasets with $\alpha \neq -3$.

Figure 2 shows the RMSD of each method, averaged over the datasets with different β, as a function of α. The performance curve of PM1, which is the best method when the data are simulated with the correct parameter $\alpha = -3$, exhibits a characteristic U-shape centered around $\alpha = -3$. This curve confirms that PM1 performs better when given the true parameter and performs worse as α moves away from -3. On the other hand, the performance curves of the two other metric methods, MDS1 and MDS2, do not exactly follow this trend: MDS1 and NMDS perform increasingly better when α decreases, and MDS2 achieves the best performance when $\alpha = -3.5$. This phenomenon occurs because in our simulation, when α decreases, the SNR for a given β increases, counterbalancing the negative effect of the transfer function miss-specification. Thus, for MDS-based methods, it is apparently more important to have more data than to have a correct α parameter. Finally, we see that, as expected, the non-metric approaches, NMDS and PM2, are more robust to transfer function misspecification than the metric approaches, because they automatically estimate it. When the parameter is wrong, PM2 outperforms the other methods for low SNR, whereas for high SNR, NMDS performs better.

§ 3.2 Real Hi-C data

We now test the different methods on real Hi-C data. Since in this case the true consensus structure is unknown, we investigate the behaviors of the different methods in terms of their ability to infer consistent structures from different datasets and across resolutions.

Resolution	Corr	MDS1		MDS2		NMDS		PM1		PM2	
		RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr	RMSD	Corr
1 Mb	0.981	13.13	0.945	5.54	0.964	5.80	0.965	7.28	0.931	**4.92**	**0.976**
500 kb	0.959	10.00	0.942	5.68	0.959	5.67	0.959	7.14	0.913	**4.66**	**0.968**
200 kb	0.845	5.64	0.940	3.74	0.945	3.73	0.946	4.01	0.891	**3.42**	**0.958**
100 kb	0.605	5.07	0.736	2.53	0.676	2.52	0.666	**2.51**	0.664	2.76	**0.771**

TABLE 1: **Stability across enzyme replicates.** For each resolution, the table lists the Spearman correlation the two enzyme replicate datasets, and, for each inference method, the average RMSD and Spearman correlation between pairs of structures inferred from the two datasets. Boldface values correspond to the best RMSD or correlation values among all five methods. In general, higher resolution leads to a lower correlation between pairs of inferred structures.

§ 3.2.1 Stability to enzyme replicates

The Hi-C assay depends upon a restriction enzyme to cleave the DNA after cross-linking, and the same sequence library can be analyzed multiple times using different enzymes. Although the resulting restriction fragments will differ, we expect *a priori* that the overall genome architecture should be the same from such replicate experiments. We therefore evaluate each genome architecture inference method with respect to the similarity of the structures inferred from two replicate Hi-C experiments that differ only in the choice of restriction enzyme. Specifically, we apply each method to two enzyme replicates, HindIII and NcoI, carried out in mouse ES cells [Dixon et al., 2012] for chromosomes 1–19.

To measure the stability of the methods, we compute (1) the Spearman correlation between the two pairwise Euclidean distance matrices of the pairs of predicted structures and (2) the RMSD between the rescaled predicted structures. Note that, before computing our two error measures, we filter out from the pair of structures any beads for which the inference hasn't been done on either dataset, i.e., beads that have zero contact counts in either data set.

To give a sense of how similar the two replicate datasets are, we also compute the Spearman correlation directly on the data, rather than on the inferred structures. As expected (Table 1), the higher the resolution is, the lower the correlation between the pairs of datasets is and the more different the inferred structures are. Across different enzyme replicates, the PM2 method yielded significantly higher correlation than all of the other methods ($p < 0.05$, signed-rank test adjusted for multiple tests with a Bonferroni correction).

§ 3.2.2 Stability to resolution

Zhang et al. [2012] show that the mapping from contact counts to physical distance

	MDS1	MDS2	NMDS	PM1	PM2
RMSD	14.86	12.92	12.98	13.03	**11.48**
Correlation	0.781	0.754	0.738	0.737	**0.807**

TABLE 2: **Stability across resolution.** The table lists the average RMSD and Spearman correlation between pairs of structures of different resolutions. In bold are the lowest average RMSD and highest average Spearman correlation. These values were computed on mouse ESC HindIII libraries Dixon et al. [2012])

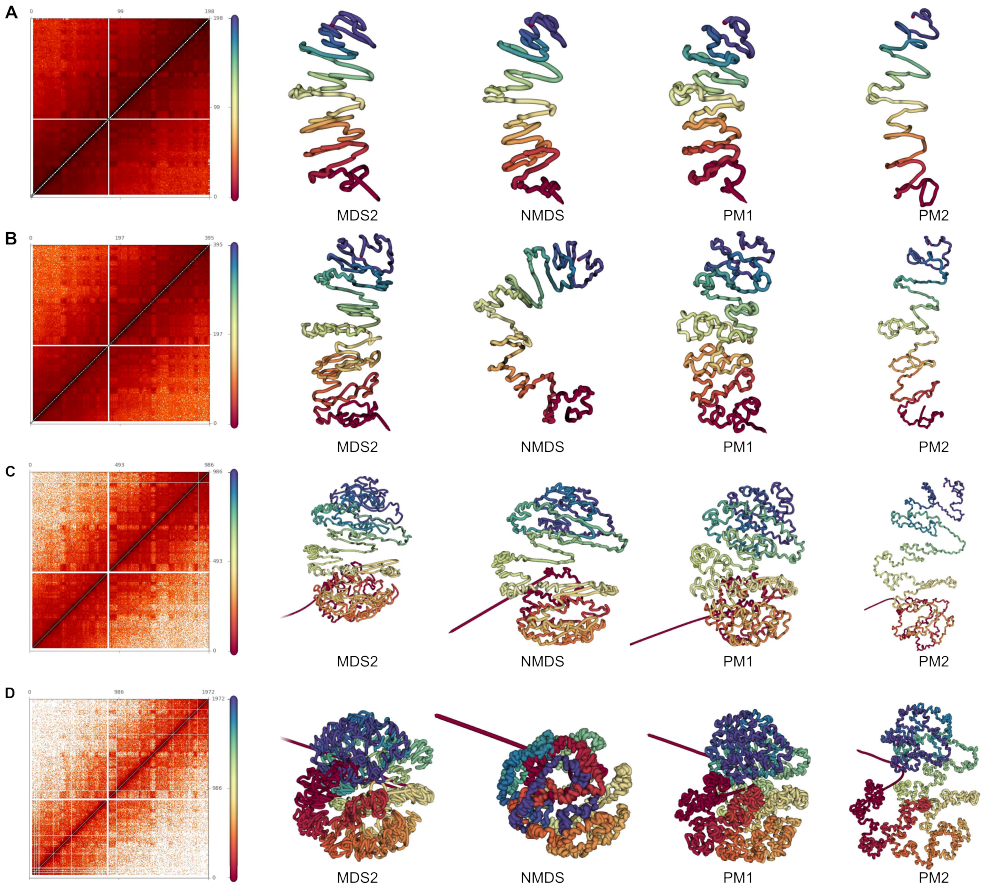


FIGURE 3: **Predicted structures for chromosome 1 at different resolution** Contact counts matrices and predicted structures for the MDS2, NMDS, PM1 and PM2 methods at 1 Mb (**A**), 500 kb (**B**), 200 kb (**C**), 100 kb (**D**)

differs from one resolution to another, underscoring the importance of good parameter estimation. To study the stability of the structure inference methods to changes in resolution, we compute the RMSD between pairs of structures inferred at different resolutions. Let $(\mathbf{X}, \mathbf{Y}) \in (R^{3\times n}, R^{3\times m})$ be a pair of predicted structures such that $n < m$ (i.e., $\mathbf{X}$ is a structure at a lower resolution than $\mathbf{Y}$). We compute a downsampled structure $\mathbf{Y}^* \in X^{3\times n}$ at the same resolution as $\mathbf{X}$ by averaging the coordinates of beads. We then compute the RMSD between this new structure $\mathbf{Y}^*$ and $\mathbf{X}$, as well as a corresponding Spearman correlation of the distance matrices.

Results are shown in Figure 3 and Table 2. PM2 is significantly ($p < 0.05$) more stable to resolution changes, both in terms of RMSD and correlation of distances.

§ 4 Discussion and conclusion

In this work, we present a novel method for inferring a consensus genomic 3D structure from Hi-C data. The method maximizes a likelihood derived from a statistical model of the relationship between the contact counts and physical distances, and includes an automatic tuning of the parameters defining the link between a 3D distance and the Poisson parameter of the corresponding contact count. We showed in simulations that the new method outperforms a panel of MDS-based approaches, including ChromSDE, which optimize an often ad-hoc stress function. The improvement is particularly important at low SNR, corresponding to more difficult problems where we want to increase the resolution of the model with a fixed total number of reads; this is typically the situation where one expects a correct maximum likelihood estimator to outperform more *ad hoc* estimators. We also showed that misspecification in the count-to-distance transfer function can harm the performance of metric methods, while our model can adapt to unknown distributions within a parametric family. Finally, we also demonstrated, on real Hi-C data, the robustness of our methods to resolution change and enzyme duplicated datasets.

Our probabilistic model of reads is similar to the model proposed by Hu et al. [2013]; however, instead of generating a family of structures by MCMC we use the model for direct maximum likelihood estimation of a consensus structure. Although the consensus structure might not be a definitive structure *in vivo*, it provides us with a rich model for further analysis, conserving hallmarks of genome organization such as the water lily form of the budding yeast [Duan et al., 2010] or topological domains [Kalhor et al., 2011].

The Poisson model underlying our approach remains very basic and could be subject to many improvements. For example, physical constraints, such as the size of the nucleus, could be incorporated into the model. Better models for zero entries may be possible, because those can either come either from non-interacting loci or from measurement errors due to, e.g., mappability problems. Overall, expressing the structure inference problem as a maximum likelihood problem offers a principled way to improve the method by improving the probabilistic model of measured dat.

Chapter 3

Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression

This chapter has been published in a slightly modified form in [Ay et al., 2014b], as joint work with Evelien Bunnik, Ferhat Ay, Sebastiaan Bol, Jacques Prudhomme, Jean-Philippe Vert, Bill Noble and Karine Le Roch.

Résumé

Dans ce projet, nous nous intéressons à l'architecture du parasite *Plasmodium falciparum*, responsable de la forme la plus virulente et mortelle du paludisme chez l'homme. Le développement de ce parasite est contrôlé par des changements précis et coordonnés dans l'expression de ses gènes au cours son cycle cellulaire. Les mécanismes régulant ces changements sont à l'heure actuelle peu connus. Nous

nous intéressons dans ce travail au lien entre l'architecture spatiale du génome et la régulation des gènes. Nous étudions la conformation du *P. falciparum* à trois moments de son cycle cellulaire asexué érythrocyte (cycle cellulaire du parasite lors de sa présence dans les cellules sanguines humaines). Grâce au protocole de capture de la conformation des chromosomes associé au séquençage haut débit, nous obtenons des cartes haute résolution des fréquences de contact entre paires de loci, à partir desquelles nous construisons des structures consensus tridimensionnelles pour chaque étape du développement cellulaire du parasite. Nous observons dans ces modèles une forte colocalisation des centromères, des télomères, de l'ADN ribosomal, ainsi que les gènes "virulence". Ces contraintes conduisent à une architecture complexe du génome, qui ne peut être simplement expliquée par un modèle de volume d'exclusion comme celui de la levure. Par ailleurs, les cartes de contacts exhibent des domaines particuliers à la position des clusters internes de gènes "virulence", suggérant l'importance du rôle de ces gènes dans l'architecture du génome. Lors de l'état trophozoite, à mi chemin dans le cycle erythrocytique alors que le génome est très fortement transcrit, celui-ci adopte une conformation plus ouverte, et les chromosomes interagissent plus entre eux. Nous observons de plus que les gènes à proximité des centres répressifs sous-télomériques sont sous-exprimés. Par ailleurs, la colocalisation de groupes de gènes spécifiques au parasite, tels que des gènes impliqués dans l'invasion des cellules sanguines humaines, ont des profils d'expression proches. Toutes ces observations suggèrent une très forte association entre l'organisation spatiale du génome de *P. falciparum* et l'expression de ses gènes. Une meilleure compréhension des processus biologiques impliqués dans la dynamique de la conformation du génome pourrait contribuer à la découverte de nouvelles stratégies pour combattre le paludisme.

Abstract

The development of the human malaria parasite *Plasmodium falciparum* is controlled by coordinated changes in gene expression throughout its complex life cycle, but the corresponding regulatory mechanisms are incompletely understood. To study the relationship between genome architecture and gene regulation in *Plasmodium*, we assayed the genome architecture of *P. falciparum* at three time

points during its erythrocytic (asexual) cycle. Using chromosome conformation capture coupled with next-generation sequencing technology (Hi-C), we obtained high-resolution chromosomal contact maps, which we then used to construct a consensus three-dimensional genome structure for each time point. We observed strong clustering of centromeres, telomeres, ribosomal DNA and virulence genes, resulting in a complex architecture that cannot be explained by a simple volume exclusion model. Internal virulence gene clusters exhibit domain-like structures in contact maps, suggesting that they play an important role in the genome architecture. Midway during the erythrocytic cycle, at the highly transcriptionally active trophozoite stage, the genome adopts a more open chromatin structure with increased chromosomal intermingling. In addition, we observed reduced expression of genes located in spatial proximity to the repressive subtelomeric center, and colocalization of distinct groups of parasite-specific genes with coordinated expression profiles. Overall, our results are indicative of a strong association between the *P. falciparum* spatial genome organization and gene expression. Understanding the molecular processes involved in genome conformation dynamics could contribute to the discovery of novel antimalarial strategies.

§ 1 Introduction

Malaria remains a major contributor to the global burden of disease, with an estimated 219 million infected individuals and 660,000 deaths annually [World Health Organization, 2012]. One of the main limiting factors for the development of novel therapies is our poor understanding of mechanisms regulating the parasite's complex life cycle, which involves several distinct parasitic stages in the human and mosquito hosts. Regulation of these developmental stages is thought to be controlled by coordinated changes in gene expression. In addition, virulence associated with the human malaria parasite, *Plasmodium falciparum*, is known to be directly linked to the parasite's ability to tightly control the expression of genes involved in antigenic variations on the surface of infected red blood cells. Some progress has been made in elucidating mechanisms controlling the expression of these virulence genes [Duraisingh et al., 2005, Freitas-Junior et al., 2005]. Furthermore, a limited number of putative sequence-specific transcription factors has

been identified in the parasite genome [Balaji et al., 2005, Coulson et al., 2004], including 27 ApiAP2 plant-like TFs, and drastic changes in chromatin structure related to transcriptional activity have been observed throughout the parasite erythrocytic cycle [Ponts et al., 2010]. However, general and specific mechanisms controlling the expression of the 6,372 parasite genes remain poorly understood.

In higher eukaryotes, several analyses have emphasized the role of genome architecture in regulating transcription. Compartmentalization of the nucleus, chromatin loops and long-range interactions contribute to a complex regulatory network [Homouz and Kudlicki, 2013, Kalhor et al., 2011, Lieberman-Aiden et al., 2009, Dixon et al., 2012]. In *P. falciparum*, little is known about the effect of genome organization on gene expression. Recent data indicate that genes involved in control of parasite virulence (*var* genes) are associated with repressive centers at the nuclear periphery [Duraisingh et al., 2005, Dzikowski et al., 2007, Lopez-Rubio et al., 2009] and that ribosomal DNA gene clusters are also colocalized [Mancio-Silva et al., 2010, Lemieux et al., 2013]. However, a global picture of the nuclear architecture throughout the parasite erythrocytic cycle progression and its role in transcriptional regulation is not yet available.

Chromosome conformation capture coupled with next generation sequencing (Hi-C) measures the population average frequency of contacts between pairs of DNA fragments in 3D space and can be used to model the spatial architecture of the genome [Lieberman-Aiden et al., 2009, Duan et al., 2010, Kalhor et al., 2011]. Here, we performed a variant of the Hi-C protocol, tethered conformation capture [Kalhor et al., 2011], to model at 10 kb resolution the spatial organization of the *P. falciparum* genome throughout its erythrocytic cycle. Our results indicate that the *P. falciparum* genome is highly structured, with strong colocalization of centromeres, telomeres, active rDNA genes and virulence gene clusters. These virulence genes exhibit distinctive contact patterns and may therefore contribute to establishing the three-dimensional structure of the *P. falciparum* genome. We identified discrete chromosomal territories during the early and late stages of the parasite erythrocytic cycle, which are partially lost in the highly transcriptionally active trophozoite stage. Global chromosome movements during the erythrocytic cycle are coherent with levels of transcriptional activity during the different stages, and the three-dimensional genome architecture shows strong correlation with gene expression levels. Collectively, our results suggest that the *P. falciparum* genome organization and gene expression are strongly interconnected.

§ 2 Results

§ 2.1 Assaying genome architecture of *P. falciparum* at three stages using Hi-C

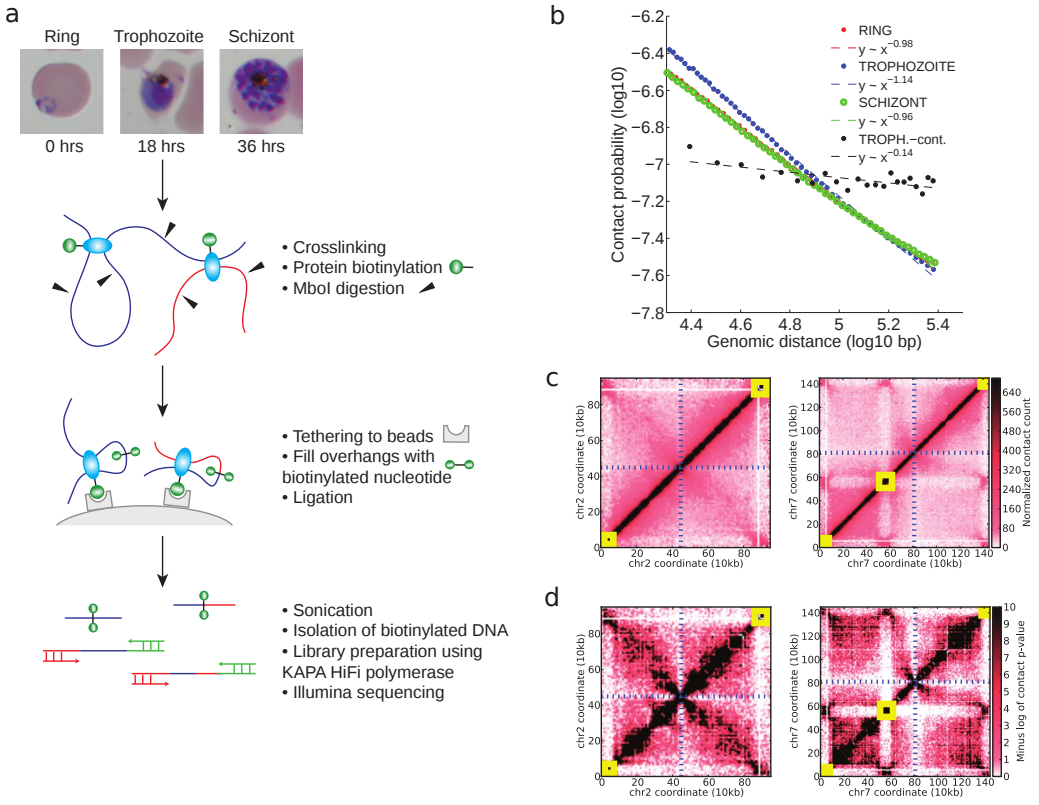


FIGURE 1: **Tethered conformation capture of the *Plasmodium falciparum* genome. a,** Experimental protocol. **b,** Contact probability as a function of genomic distance, with log-linear fits for the three erythrocytic stages, as well as an experimental control. **c,** Normalized contact count matrices at 10 kb resolution for chromosome 2 and chromosome 7 in the schizont stage. **d,** Contact p-values (negative log10 scale) for chromosome 2 and chromosome 7 in the schizont stage. In **(c)** and **(d)**, yellow boxes denote clusters of VRSM genes, and blue dashed lines indicate the centromere location.

To study the genome architecture of *P. falciparum*, we harvested parasites at three stages of the infected red blood cell cycle: after invasion of red blood cells at the ring stage (0h), during high transcriptional activity at the trophozoite stage (18h) and near the end of the cycle at the schizont stage (36h), just before the newly formed parasites are released into the bloodstream. Next, we applied the Hi-C protocol [Kalhor et al., 2011] with modifications to accommodate the extremely AT-rich genome of the malaria parasite (Fig. 1a, Appendix Note 1, Appendix File 1). As a control, we prepared a sample for which chromatin contacts were not preserved by crosslinking of DNA and proteins.

We evaluated the quality of the resulting data for each sample. First, we confirmed that the contact probability between two intrachromosomal loci exhibits a log-linear decay with increasing genomic distance (Fig. 1b, Appendix Fig. 1). Second, we obtained lower numbers of interchromosomal contacts from crosslinked samples relative to both random expectation and our control sample (Appendix Table 1). Third, we observed that the percentage of long-range contacts (either interchromosomal or intrachromosomal >20 kb) was significantly higher than control and comparable to the numbers observed in yeast [Duan et al., 2010] (Appendix Table 1). Together, these results indicated that we successfully assayed the *P. falciparum* genome architecture with a high signal-to-noise ratio. We then coalesced the mapped read pairs into a raw contact count matrix at 10 kb resolution, and we corrected for potential technical and experimental biases [Imakaev et al., 2012] (Fig. 1c, Appendix Fig. 2) . The resulting normalized contact maps were used to identify a subset of high-confidence contacts for each stage (Methods, Appendix Note 2, Appendix File 2) [Ay et al., 2014a]. We identified pairs of genes that show evidence of stage-specific contacts (Methods) and then applied gene set enrichment analysis to the set of genes that participate in such contacts. This analysis identified significant enrichment of VRSM genes for the ring and trophozoite stages (Appendix Table 2). This observation suggests that the proximity between some VRSM clusters changes from the ring to trophozoite stages, even though both stages show overall colocalization of VRSM clusters. A similar enrichment analysis conducted using contacts that are specific to two out of three stages resulted in no significant enrichment due to the small number of genes involved in such contacts.

Normalized contact count and confidence score matrices exhibit a canonical "X" shape, indicative of a folded chromosome architecture anchored at the centromere, as previously observed in yeast [Duan et al., 2010, Tanizawa et al., 2010] and the bacterium *C. crescentus* [Umbarger et al., 2011] (Fig. 1c-d, Appendix Fig. 3). However, chromosomes that harbor non-subtelomeric clusters of genes involved in antigenic variation and immune evasion (Appendix File 3; VRSM genes: *var*, *rifin*, *stevor* and *Pfmc-2tm*)—chromosomes 4, 6, 7, 8 and 12—exhibit additional folding structure (Fig. 1c-d, Appendix Fig. 3).

§ 2.2 Three-dimensional modeling recapitulates known organizational principles of *Plasmodium* genome

To better characterize the genome architecture, we generated for each stage 100 consensus 3D structures, each of which summarizes the population average (Fig. 2a, Methods), using multidimensional scaling (MDS) with two primary constraints [Duan et al., 2010]: (i) the DNA must lie within a sphere with a specified diameter [Bannister et al., 2005, Weiner et al., 2011] and (ii) adjacent 10kb loci must not be separated by more than 91

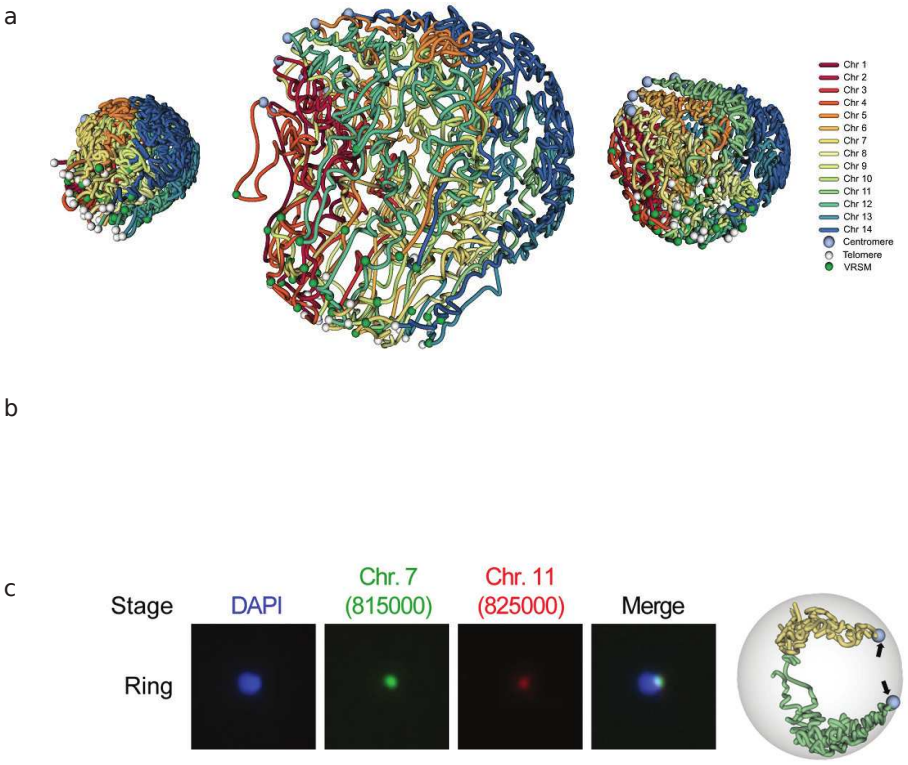


FIGURE 2: **3D modeling and validation with DNA FISH. a,** 3D structures of all three stages. The nuclear radii used to model ring, trophozoite and schizont stages were 350, 850, and 425 nm, respectively. Centromeres and telomeres are indicated with light blue and white spheres, respectively. Midpoints of VRSM gene clusters are shown with green spheres. **b,** Validation of colocalization between a pair of interchromosomal loci with VRSM genes (chr7: 550,000 - 560,000 that harbors internal VRSM genes and chr8: 40,000 - 50,000 that harbors subtelomeric VRSM genes) by DNA FISH (left) and by the three-dimensional model for the corresponding stage (right). The location of the loci in the 3D model is indicated with light blue spheres and pointed by black arrows. **c,** Validation same as in **(b)** for a pair of interchromosomal loci that harbor no VRSM genes (chr7: 810,000 - 820,000 and chr11: 820,000 - 830,000).

nm [Bystricky et al., 2004]. *P. falciparum* undergoes an atypical form of cell division, resulting in schizont stage parasites with multiple independent nuclei, each containing 1n chromosomes. Note that our model assumes that a single copy of each chromosome is present in each structure, thus averaging the signal from these multiple nuclei per cell.

We performed a series of experiments to assess the robustness of our 3D inference procedure. Our results showed only slight changes in the inferred 3D models when we varied the parameter used in conversion of contact counts to expected distances (Appendix Table 3). This was also true when we removed from the inference the two types of spatial constraints related to nuclear volume and to distances between adjacent beads (Appendix Table 4). Finally, our experiments on the impact of the initialization step (Methods) showed that structures inferred from different initial configurations are highly

similar (Appendix Fig. 4), do not fall into discrete clusters (Appendix Fig. 5) and all such structures exhibit common organizational hallmarks (Appendix Fig. 6). Because of the stability of our inference procedure, hereafter we generally present and discuss the results for only one representative structure per stage.

Although the modeling procedure contains no explicit constraints on telomere or centromere locations, we observe strong colocalization of both sets of loci across all three stages (Fig. 2a, Appendix Fig. 7, Appendix Table 5), with centromeres and telomeres localizing in distal regions of the nucleus. To understand further the colocalization patterns of centromeres and telomeres in each stage, we divided each chromosome into three compartments (left–mid–right or telomeric–centromeric–telomeric) using eigenvalue decomposition (Methods) and then performed hierarchical clustering on the matrix of pairwise distances between compartments (Appendix Fig. 8). At each stage, we observed clusters that are comprised primarily of either centromeric or telomeric compartments. In particular, during the trophozoite stage, all the centromeric compartments fall into two main clusters suggesting strong colocalization of all centromeres for this stage (Appendix Fig. 8d). Such strong colocalization has previously been observed by immunofluoresence microscopy at the trophozoite and schizont stages but not at the early ring stage [Hoeijmakers et al., 2012a]. However, when the size of the nucleus is used as a marker of the parasite asexual cycle stage [Bannister et al., 2005, Weiner et al., 2011], the cells that are presented as trophozoites in this previous study [Hoeijmakers et al., 2012a] are more similar to our ring stage parasites, indicating that centromere clustering also occurs early in the erythrocytic cycle. Furthermore, if the centromeres are stochastically distributed between a small number of foci within a population, then an assay that measures average signal, such as Hi-C, will indeed demonstrate an aggregate clustering for the centromeres and not complete dispersion as suggested by a recent study [Lemieux et al., 2013]. These results suggest that *P. falciparum* nuclei are highly structured around centromeres and telomeres, consistent with known organizational principles gathered through multiple independent microscopy experiments [Duraisingh et al., 2005, Dzikowski et al., 2007, Lopez-Rubio et al., 2009, Hoeijmakers et al., 2012a].

§ 2.3 Virulence gene clusters on different chromosomes colocalize in 3D

In addition to centromeres and telomeres, we observed for all VRSM gene clusters, both internal and subtelomeric, a significant colocalization with one another (Fig. 2a, Appendix Table 5). The significant colocalization for VRSM clusters as well as for centromeres and telomeres were all reproducible when we used contact counts instead of 3D distances to perform colocalization tests similar to Appendix Table 5 (data not

shown). Given colocalization of the telomeres, colocalization of subtelomeric clusters is not surprising. However, the proximity of internal VRSM clusters with one another and with subtelomeric clusters is unexpected under the random polymer looping model and, to the best of our knowledge, observed experimentally for the first time. To further validate these results inferred from our 3D models, we performed DNA fluorescence *in situ* hybridization (FISH) (Methods, Appendix Note 3) on an interchromosomal pair of strongly interacting (at 10 kb resolution) VRSM clusters: the internal cluster from chromosome 7 and a subtelomeric cluster from chromosome 8. We observed strong colocalization by FISH (>90% of cells, Fig. 2b, Appendix Fig. 9a, Appendix Table 6), providing independent support for the clustering of VRSM genes. Although previous FISH results indicated that *var* genes form 2 to 5 clusters in 3D per cell [Freitas-Junior et al., 2000, Lopez-Rubio et al., 2009], others recently showed single foci for the VRSM gene-associated repressive histone mark H3K9me3 and heterochromatin protein 1 (PfHP1) [Dahan-Pasternak et al., 2013], as well as for H3K36me3 that marks both active and silenced *var* genes [Ukaegbu et al., 2014]. Because our experimental strategy (Hi-C) captures a population average, we are unable to distinguish between multiple VRSM gene clusters in 3D if the genes are randomly distributed among clusters from cell to cell. Using FISH experiments, we also observed strong colocalization (>90% of cells, Fig. 2c, Appendix Fig. 9b) for a pair of interchromosomal loci located outside VRSM clusters with consistent strong interactions at all three stages, while colocalization was not observed for a pair of non-interacting interchromosomal loci (<10% of cells, Appendix Fig. 9c). These results demonstrate that our population average Hi-C data agrees with a majority of single cell FISH images.

§ 2.4 Highly transcribed rDNA units colocalize in 3D during the ring stage

Similar to VRSM genes, the rDNA genes are strictly regulated during the parasite life cycle. In *P. falciparum*, these genes are dispersed on different chromosomes in five rDNA units containing the 18S, 5.8S and 28S genes and one repeat unit consisting of three copies of the 5S gene. A previous FISH study suggested that all rDNA units localize at a single nucleolus but also claimed that the two units on chromosomes 5 and 7 that are actively transcribed during the ring stage (A-type units) are dispersed in the ring stage [Mancio-Silva et al., 2010]. However, a more recent Hi-C study of ring stage parasites demonstrated strong clustering of these two A-type units in multiple strains [Lemieux et al., 2013]. Analysis of our Hi-C data confirmed overall enrichment of contacts between chromosomes 5 and 7 in all three stages and showed a particular peak of enrichment centered at the rDNA unit on chromosome 5 among all interchromosomal

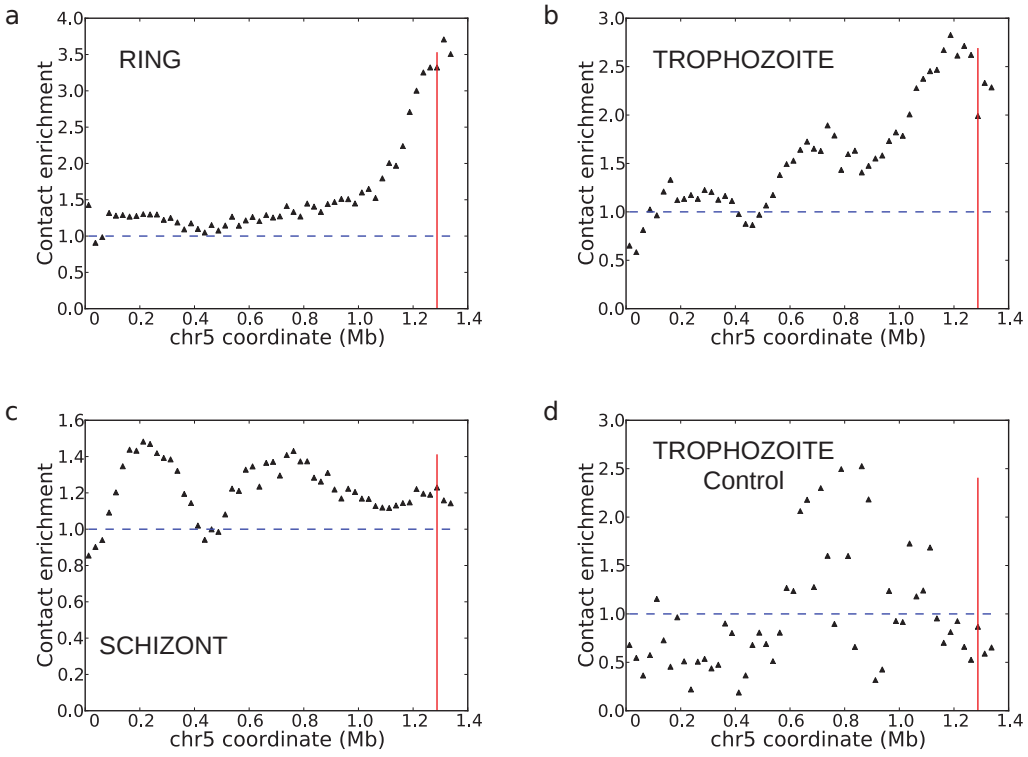


FIGURE 3: **Colocalization of highly transcribed rDNA units.** Virtual 4C plots generated at 25 kb resolution using as a bait the A-type rDNA unit on chromosome 7 from crosslinked Hi-C libraries of **(a)** ring, **(b)** trophozoite, **(c)** schizont stages and **(d)** from the trophozoite control library. Vertical red line indicates the midpoint of the A-type rDNA unit on chromosome 5. Normalized contact counts from 50 kb up- and downstream of the 25 kb bin containing the rDNA unit are used, omitting the rDNA-containing window itself to exclude repetitive DNA. For each window w on chromosome 5, the contact enrichment is calculated by dividing the contact count between the bait and w to the average interchromosomal contact count for the bait locus.

contact partners of the rDNA unit on chromosome 7 in the ring stage (3.32x, Fig. 3a). We observed less striking enrichment of contacts that are not specific to or centered on the rDNA units for the other two stages (trophozoites (1.99x), schizonts (1.23x), Fig. 3b-c) during which the two rDNA units are not transcribed [Mancio-Silva et al., 2010]. Reanalysis of the Lemieux *et al.* data using our processing pipeline also showed this enrichment consistently in three different NF54-derived strains in the ring stage (6.06x, 4.47x and 4.61x, respectively, Appendix Fig. 10a-c). Control libraries from both studies do not exhibit this enrichment (Fig. 3d, Appendix Fig. 10d). Our 3D models for the ring stage place these two A-type rDNA units near the nuclear periphery. Together with the strong colocalization between A-type rDNA, these results suggest the existence of perinuclear transcriptionally active compartments. Such compartments may play a role in separating out the single active var gene per cell from compact chromatin around (sub)telomeric regions marked by the repressive H3K9me3 modification [Lopez-Rubio

et al., 2009]. We did not observe an overall colocalization between all rDNA units in the ring stage, including the three 18S, 5.8S, 28S units and one 5S unit that are not expressed during asexual erythrocytic cycle (Appendix Table 5). This observation suggests that genomic location may influence rDNA expression by the preferential colocalization of the expressed rDNA units, away from the non-expressed units.

§ 2.5 Transcriptionally active trophozoite stage exhibits an open chromatin structure

Assaying three different time points, we observed significant changes in chromatin structure throughout the erythrocytic cycle. To visualize high-level changes, we generated animations showing the movement of chromosomes as the parasite progresses through its cell cycle (Appendix Files 4-18). We then characterized global chromatin changes by analyzing the relationship between contact frequency and genomic distance (Fig. 1b, Appendix Fig. 1). The gradient of the log-linear fit is very close to -1 in both the ring and schizont stages (-0.98 and -0.96, respectively) indicative of a fractal globule genome architecture that is usually found in higher eukaryotes [Lieberman-Aiden et al., 2009]. Intriguingly, the intermediate and most active transcriptional stage yields a log-linear fit value with gradient -1.14, a value between the fractal (-1) and the equilibrium globule (-1.5) model suggested in yeast [Fudenberg and Mirny, 2012] and indicative of more chromosomal intermingling. Indeed, a value of -1.17 has been demonstrated to correspond to a state of "unentangled rings" similar to the fractal globule state, in which the rings may correspond to long chromosomal regions looped on or anchored to a nuclear scaffold [Vettorel et al., 2009]. It is important to note that the value of the gradient is determined solely by Hi-C contact counts and, therefore, the above mentioned difference is independent of our 3D modeling and the change in the nuclear radius from one stage to another. Furthermore, the difference in the gradient value for trophozoites compared to the two other stages is consistent for each chromosome, suggesting that all chromosomes change their folding behavior during the trophozoite stage (Appendix Table 7).

In order to further investigate whether trophozoites show a more open chromatin structure than the two other stages, we systematically compared our data across all three stages. First, we computed and compared intra and interchromosomal contact probabilities for each stage (Appendix Fig. 11). We observed that intrachromosomal contacts, even at very large distances, are more prevalent than interchromosomal contacts for all three stages, suggesting the existence and preservation of chromosome territories throughout the erythrocytic cycle. However, the enrichment in intrachromosomal contacts was the lowest for trophozoite stage for distances above 300 kb, suggesting a relative loss of territories in this stage compared to the other two. Second, we quantified

how preserved the chromosomal territories are at each stage by estimating the degree of chromosome intermingling in our 3D models. We randomly sampled small spheres in the nucleus and asked, for each chromosome i, what percentage of the spheres that contain any locus from chromosome i also contain a locus from another chromosome j. Our results using different sphere sizes, and controlling for the varying nuclear diameter, consistently exhibited the highest amount of intermingling for the trophozoite stage and the highest territory preservation for the schizont stage (Appendix Fig. 12).

To understand the architectural dynamics responsible for the systematic changes in chromatin compaction, we computed the relative movements among chromosome compartments during the erythrocytic cycle. Despite the increase in nuclear volume, many interchromosomal compartment pairs came closer together in the transition from the ring to trophozoite stage (Appendix Fig. 13a, red color). Subsequently, most interchromosomal compartments moved away from each other in the transition to the schizont stage (Appendix Fig. 13b, blue color), resulting in more compact chromatin that favors formation of chromosome territories. These results are consistent with a previously proposed model, in which the *P. falciparum* nucleus exhibits a more open chromatin configuration at the trophozoite stage, enabling interchromosomal contacts and high levels of transcriptional activity [Ponts et al., 2010].

§ 2.6 *Plasmodium* genome architecture cannot be explained by volume exclusion

We next assessed whether the primary architectural features in *P. falciparum* arise from a population of constrained but otherwise random configurations of chromatin following a simple volume exclusion (VE) model, as recently shown for *Saccharomyces cerevisiae* [Tjong et al., 2012]. We therefore repeated the Tjong *et al.* simulations using the same set of constraints and successfully recovered the strong correlation between the simulated map and the experimentally observed yeast contact map (raw correlation of 0.91; normalized correlation of 0.57; Fig. 4a, Methods, Appendix Note 4, Appendix Fig. 14). In contrast, our simulations for the ring, trophozoite and schizont stages of *P. falciparum* yielded markedly lower correlations (normalized correlation of 0.34, 0.39 and 0.49, respectively) and strikingly different contact maps compared to the experimentally observed maps (Fig. 4b). One significant reason for the observed discrepancy between yeast and *P. falciparum* is the lack of structure around clusters of VSRM genes in the simulated data (Fig. 4b). Accordingly, we conclude that the simple volume exclusion model, which so convincingly explains the yeast genome architecture, is insufficient to explain the observed architecture of *P. falciparum* genome, highlighting the need for a genome-wide assay such as Hi-C to obtain accurate structural models.

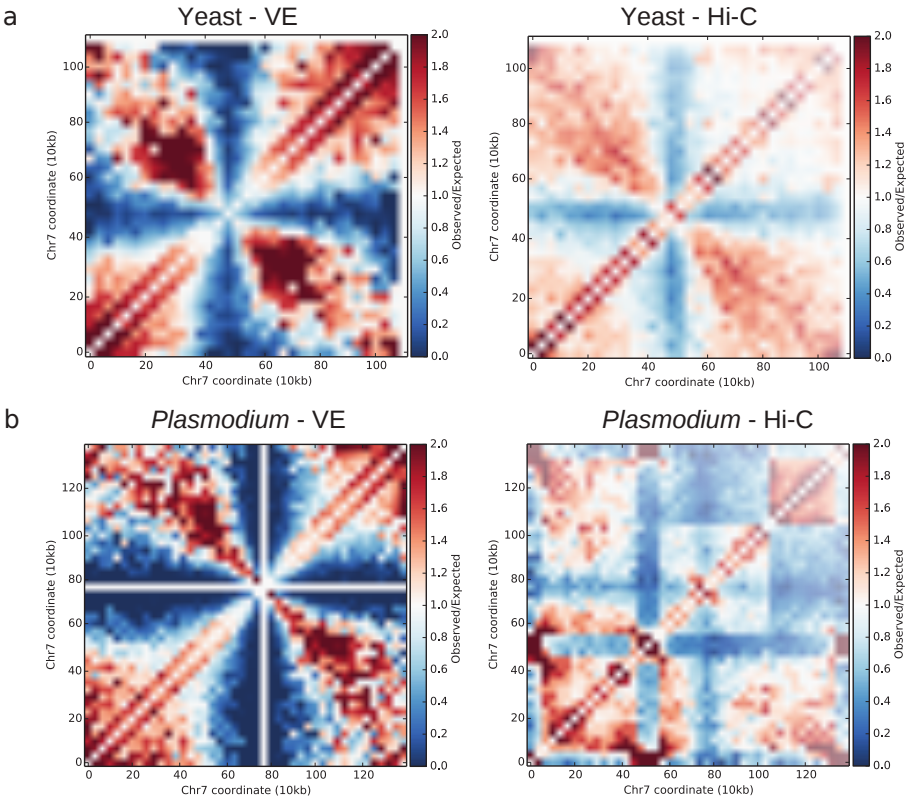


FIGURE 4: **Volume exclusion modeling.** Observed/expected contact frequency matrices illustrate, for each locus, either the depletion (blue) or enrichment (red) of interaction frequencies compared to what would be expected given their genomic distances. **a,** Observed/expected contact frequency matrices derived from *S. cerevisiae* chr 7 from volume exclusion modeling (left) and Hi-C data (right). **b,** Observed/expected matrices from volume exclusion modeling (left) and Hi-C data (right) for *P. falciparum* chr 7 during the trophozoite stage.

§ 2.7 VRSM gene clusters form domain-like structures

Our results from the volume exclusion modeling and from visual inspection of the contact maps suggest that the internal VRSM gene clusters are associated with distinctive structural features. All eight of the internal VRSM clusters induce a striking cross-like shape, both in the contact count and 3D distance matrices (Fig. 5a-b, Appendix Fig. 3). Quantification of this phenomenon revealed a consistent contact pattern across all eight internal VRSM clusters (Appendix Fig. 15), suggesting that VRSM gene clusters adopt a compact, domain-like structure. Although these domain-like structures resemble topologically associated domains (TADs) described in mammals [Dixon et al., 2012, Nora et al., 2012], the VSRM domains are much smaller (10–50 kb) compared to TADs (0.1–1 Mb). Furthermore, because VRSM genes have no orthologs in human and mouse, mechanisms regulating these domain-like structures likely differ from the one in mammalian genomes. Further understanding of how these VRSM domains are formed

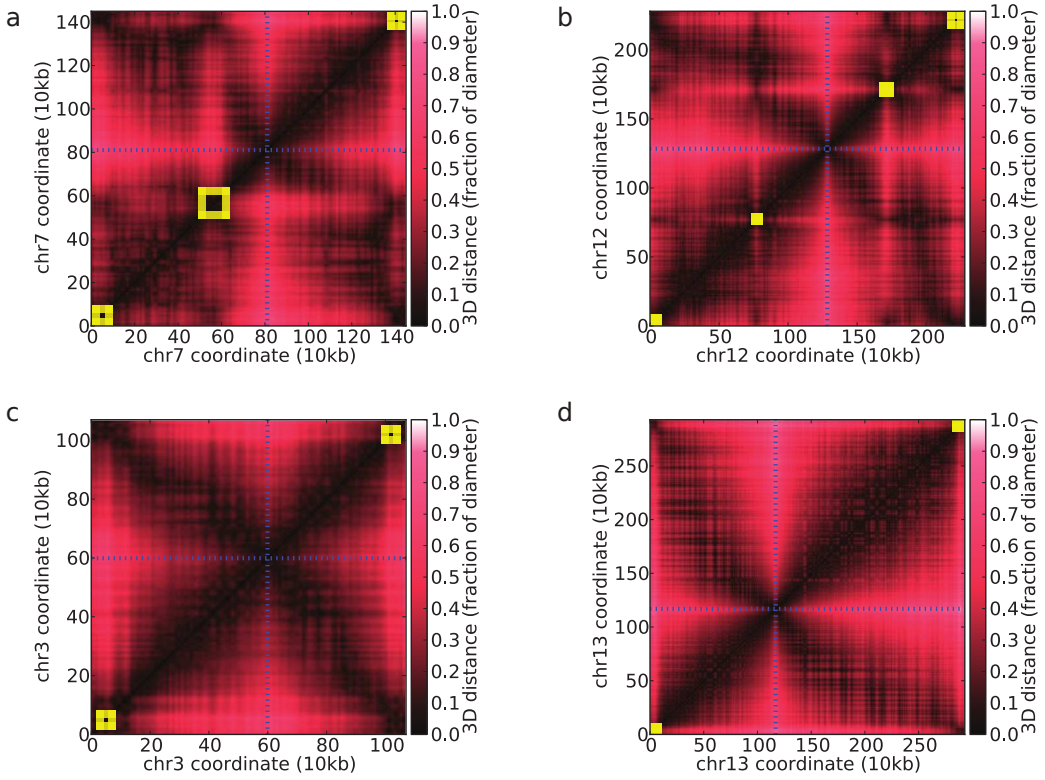


FIGURE 5: **Role of internal VRSM gene clusters in shaping genome architecture. a-d,** Heatmaps of scaled pairwise Euclidean distances derived from the 3D model at 10 kb resolution for **(a, b)** two chromosomes that harbor internal VRSM gene clusters and **(c, d)** two chromosomes that do not. Yellow boxes indicate locations of VRSM clusters.

in *Plasmodium* would shed light on genome architecture associated regulation of VRSM gene expression.

Another interesting pattern involving internal VRSM clusters emerged from further inspection of chromosome compartments. Five of the eight internal VRSM clusters (two on chromosome 4, one on chromosome 7 and both clusters on chromosome 12) occur at compartment boundaries (third and fourth rows of Appendix Fig. 3). This striking overlap suggests that VRSM genes may contribute to or rely upon the boundaries of chromosomal compartments. Taken together with the domain-like structures around these VRSM clusters, these results confirm that genome architecture is likely to be involved in the strict regulation of virulence genes during the erythrocytic cycle.

§ 2.8 Expression is highly concordant with 3D localization for *Plasmodium* genes

Next, we investigated the relationship between the three-dimensional genome structure and gene expression using four published expression data sets [Le Roch et al., 2003, Lopez-Barragan et al., 2011, Otto et al., 2010, Bunnik et al., 2013]. First, we observed that, for each of the three stages, interchromosomal pairs of genes that strongly interact (contact counts within the top 20%) as well as gene pairs that are in close proximity (<20% of the nuclear diameter) showed more correlated expression profiles than genes that are far apart (Fig. 6a,b), as previously observed in yeast [Homouz and Kudlicki, 2013]. To assess whether these observed trends are confounded by similarly expressed VRSM genes that strongly interact with each other and are placed together near telomeres by our 3D model, we repeated the above analyses by excluding all VRSM genes (Appendix Fig. 16). Even though the observed trends are weakened by exclusion of VRSM genes, the decrease in 3D distance and increase in contact count with increasing expression correlation remained significant (Appendix Fig. 16). It is also important to note that, for these analyses, we excluded intrachromosomal gene pairs to only focus on the relationship between 3D proximity and gene expression by eliminating the confounding effect caused by genes that lie nearby on a chromosome and show similar expression profiles. Second, we analyzed gene expression in relation to the repressive subtelomeric clusters [Duraisingh et al., 2005, Dzikowski et al., 2007, Lopez-Rubio et al., 2009] and other nuclear landmarks. The subset of genes that lie within 20% of the nuclear diameter to the centroid of the telomeres showed significantly lower expression levels than more distal genes (Fig. 6c). The repressive effect of the subtelomeric clusters is apparent in all three stages and is strongest at the trophozoite stage, in which subtelomeric VRSM clusters are known to be tightly repressed [Chen et al., 1998]. If we remove the VRSM genes from the analysis, the repressive effect is still significant at the trophozoite stage, which is known to be the most active transcriptional stage of the erythrocytic cycle (Appendix Fig. 17a,b). Similar analysis showed higher expression levels for genes located near the nuclear center, as well as for genes close to the centroid of the centromeres (Appendix Fig. 17c,d). Furthermore, we observed significant and consistent colocalization across all three stages for 11 of the 15 expression clusters identified in Le Roch et al. [2003] (Appendix Table 5). Strikingly, the trophozoite stage showed significant colocalization for clusters associated with genes that are repressed during this stage (clusters 1, 3, 4, and 13-15) as well as genes that exhibit high levels of expression (clusters 6, 9, 10, and 12), confirming the strong relationship between 3D location and gene expression.

To further explore the relationship between gene expression and 3D structure, we employed an unsupervised learning method known as *kernel canonical correlation analysis*

(kCCA) [Bach and Jordan, 2002]. This methodology identifies a set of orthogonal gene expression profiles that exhibit coherence with respect to the 3D structure (Methods). For all stages, the projection of gene expression patterns onto the first extracted profile exhibits a striking transcriptional gradient across the 3D structure, from the telomere cluster to the opposite side of the nucleus (Fig. 6c, Appendix Fig. 18a,c,e). The coherence with 3D structure drops significantly in the second component of the kCCA (Appendix Fig. 18b,d,f), suggesting that gene expression is strongly influenced by distance to the subtelomeric repressive center. To further interpret the kCCA results we employed gene set enrichment analysis [Subramanian et al., 2005] on the ranked lists of projections onto the first kCCA component. The results showed, for all three stages, significant enrichment (q-value < 0.01) of gene sets related to antigenic variation and translation (i.e. ribosome proteins) on the telomeric and non-telomeric side, respectively, of the extracted kCCA expression profile (Appendix Tables 8, 9, 10). Similar to the colocalization test results for expression clusters of Le Roch et al. [2003], clusters of genes that are repressed (clusters 4, 13, and 14) and expressed (clusters 6 and 9-12) in the trophozoite stage showed consistent enrichment in the strongest kCCA profile (Appendix Table 11). In addition, genes exclusively expressed in sporozoites (cluster 1) and gametocytes (clusters 3) were also strongly enriched, indicating that the repression of these genes during the asexual erythrocytic cell cycle may be related to their localization within the nucleus. Finally, for GO terms related to parasite invasion (rhoptry, myosin complex, motor activity; q-value < 0.1) and for the cluster of invasion genes (cluster 15), we observed an enrichment relative to the second kCCA component, suggesting that expression of invasion genes may also be regulated by the 3D genome structure (Appendix Tables 11, 12).

§ 3 Discussion

This study presents the first analysis of genome architecture during the cell cycle of a eukaryotic pathogen. Overall, our data demonstrate that the genome of *P. falciparum* exhibits a higher degree of organization than the similarly sized budding yeast genome. Although localization of chromosomes within the *P. falciparum* nucleus is partially dictated by size constraints, the simple volume exclusion model observed in yeast is insufficient to explain the 3D architecture of the *P. falciparum* genome. In particular, a striking spatial complexity is added by clusters of virulence genes, which function as critical structural elements that shape the genome architecture. Furthermore, our model correlates well with expression levels of parasite-specific gene sets and shows strong clustering of repressed genes and highly transcribed rDNA units, indicative of a non-random genomic organization that contributes to gene regulation during the asexual erythrocytic

cycle. Considering the strong association between nuclear architecture and gene expression as well as the observed domain-like structures, *Plasmodium* species may be excellent model organisms to study the impact of genome structure on gene regulation. The lower complexity of genome organization in organisms with similarly sized genomes, such as yeast, may indeed be less informative for such investigations.

Assaying multiple time points during the parasite's erythrocytic cycle revealed intriguing changes in genome structure between the different developmental stages. Our results show that the genome adopts a more open conformation during the trophozoite stage consistent with high transcriptional activity in this stage of the erythrocytic cycle, followed by compaction of chromosomes into discrete chromosome territories before re-invasion of a new host cell. A similar pattern was observed previously for nucleosome occupancy, with strong histone depletion at the trophozoite stage and nucleosome replacement at the schizont stage [Ponts et al., 2010]. Based on these observations, we hypothesize that the spatial genome organization of *P. falciparum*, coupled with its dynamic chromatin structure, acts as an important alternative mechanism of transcriptional regulation, possibly compensating for the lack of a diverse collection of specific transcription factors [Balaji et al., 2005, Coulson et al., 2004] and the low capacity of the parasite to regulate gene expression in response to metabolic stress [Ganesan et al., 2008, Le Roch et al., 2008]. These changes in genome architecture could mainly be indicative of differences between the various developmental stages of the parasite, but could also be related to cell cycle progression itself. Given the importance of nuclear architecture for regulation of gene expression, disruption of its genome organization is likely to interfere with parasite development through the erythrocytic cycle and could therefore be lethal to the parasite. Compounds targeting proteins involved in establishing and maintaining the three-dimensional genome structure in *P. falciparum* may thus have potent antimalarial activity.

A recently published Hi-C study suggested that chromosomal territories are absent in the ring stage parasites, especially for larger chromosomes [Lemieux et al., 2013]. In contrast, our data provides multiple lines of evidence for the existence of chromosome territories throughout the erythrocytic cell cycle. In particular, we observed that intrachromosomal contacts, even at very large distances, are more prevalent than interchromosomal contacts. This observation is supported by our own Hi-C data in three stages as well as by our reanalysis of the Lemieux *et al.* data (Appendix Fig. 11b-e). The difference between the two analyses can be traced to our improved method for discretizing the genomic distance axis, which avoids bins with few observations and, hence, high variance (Appendix Fig. 11a versus b). Even though further experiments may be necessary to reconcile these differences, our results strongly suggest that *P. falciparum* chromosomes occupy distinct territories, similar to other eukaryotic genomes.

Clustering of virulence gene families into a distinct nuclear compartment is likely to play an important role in the formation of repressive heterochromatin that controls the silencing of these genes. Heterochromatin around virulence genes is characterized by histone modifications H3K36me3 [Jiang et al., 2013] and H3K9me3 [Duraisingh et al., 2005, Lopez-Rubio et al., 2009], both of which were shown to be essential for maintaining *var* gene repression. The formation of heterochromatin is directed by the interaction of PfSIP2 with specific DNA motifs in promoters of virulence genes and in subtelomeric domains [Flueck et al., 2010], but additional factors are likely to contribute to this process. The question remains, however, how the formation of this repressive center is regulated and whether the colocalization of virulence gene clusters is a cause or a consequence of their transcriptional silencing. One experiment that would shed light on this issue would be to relocate a *var* gene to a different location in the genome and to monitor how the introduction of this novel *var* gene locus influences genome structure, although technical challenges that come with manipulation of the *P. falciparum* genome may prevent such procedures. Virulence genes are expressed on the surface of red blood cells and are therefore important antigens for the humoral immune system. A better understanding of virulence gene silencing will provide us with more opportunities to interfere with this process, which would ultimately benefit vaccine development.

In this study, we modeled the *P. falciparum* genome architecture based on the average signal from a population of parasites. However, it can be expected that considerable variability in genome conformation exists from cell to cell, as recently demonstrated in mouse [Nagano et al., 2013]. While challenging, it would be interesting to perform Hi-C analysis on individual parasites to reveal the extent of inter-cellular variation in *P. falciparum* genome architecture. This experiment would also allow a more detailed analysis of the clustering of *var* genes in one or multiple repressive centers, as well as the differential localization of the single active *var* gene.

In conclusion, this study demonstrates the unique role of genome organization in transcriptional regulation in the human malaria parasite. In other eukaryotes such as human and mouse, genome organization has been shown to participate in gene regulation through formation of specific chromatin loops that bring enhancers and enhancer-like elements in proximity to their target promoters. However, a global reorganization of the entire genome correlated with changes in transcriptional capacity, as described here for *P. falciparum*, has not been observed for any of the genomes studied so far. Therefore, our data proposes a novel mechanism of gene regulation for *P. falciparum* that can operate without relying on specific transcription factors or enhancer elements. Similar to other eukaryotes, gene expression in *P. falciparum* is likely to be regulated by multiple layers of control at both transcriptional and translational levels. However, the necessity

to transcriptionally repress distinct groups of parasite-specific genes may have driven *P. falciparum* to adopt this exceptional genome organization.

§ 4 Methods

§ 4.1 Experimental protocols

§ 4.1.1 *P. falciparum* strain and culture conditions

P. falciparum strain 3D7 was maintained in human O+ erythrocytes in 5% haematocrit according to a previously described protocol [Trager and Jensen, 1976]. Cultures were synchronized twice at ring stage with 5% D-sorbitol treatments performed eight hours apart [Lambros and Vanderberg, 1979]. Parasites were harvested 48 hours after the first sorbitol treatment (0h; ring stage), and then 18 hours (early trophozoite stage) and 36 hours (late schizont stage) thereafter. The developmental stage of the parasites was verified by microscopy using Giemsa-stained blood smears prior to harvesting.

§ 4.1.2 Cross-linking

Aspirated *P. falciparum* cultures were pooled into 50 ml centrifuge tubes and filled up to 35 ml with phosphate buffered saline (PBS) warmed to 37°C. Cultures were treated with 3 ml 16% formaldehyde (1.25% final concentration) and incubated for 25 min at 37°C while rocking. Formaldehyde was quenched with 5.2 ml 1.25 M glycine (final concentration 150 mM) for 15 min at 37°C while rocking, followed by 15 min at 4°C while rocking. PBS was used instead of formaldehyde and glycine for the not cross-linked control. Cultures were spun at 660 × g for 20 min at 4°C. Not cross-linked control parasites were treated with 5 volumes 0.15% saponin in water and incubated 10 min at 4°C while rocking. PBS was used instead of saponin for the cross-linked parasites. Parasites were spun at 660 × g for 15 min at 4°C. Pellets were washed multiple times until clean and stored at -80°C.

§ 4.1.3 Tethered conformation capture procedure

We applied an adapted Hi-C method referred to as tethered conformation capture (TCC) [Kalhor et al., 2011] to map the intra and interchromosomal contacts in *Plasmodium falciparum*. For a detailed description of the overall protocol see Appendix Note 1.

§ 4.1.4 DNA-FISH

For each 10 kb locus of interest, we determined the location for which on average the highest number of contact counts were observed and designed DNA probes targeting the 2 kb region surrounding this location. Probes were prepared using Fluorescein-High Prime and Biotin-High Prime kits (Roche) according to manufacturer's instructions. Template DNA was prepared by PCR (5 min at 95°C, 35 cycles of 30 sec at 98°C followed by 150 sec at 62°C, and 5 min at 62°C) using the KAPA HiFi DNA Polymerase HotStart ReadyMix. Sequences of primers used for probe generation are shown in Appendix Table 6. For a detailed description of the DNA-FISH protocol see Appendix Note 3. The percentage of colocalization was determined by visual inspection of >100 cells per condition.

§ 4.2 Computational methods

§ 4.2.1 Mapping and filtering of sequence data

We first trimmed each end of the paired-end reads from all samples to 40 bp. We used FastQC [Andrews, 2010] reports of aggregate read qualities for each sample to determine the amount of trimming required from each end of the read to keep the highest quality 40-bp region.

To filter out reads from human DNA, we mapped the trimmed paired-end reads to the human genome (UCSC hg19) using the short read alignment mode of BWA (v0.5.9) [Li and Durbin, 2010] with default parameter settings. Each end of the paired reads was mapped individually. We post-processed the alignment results to extract reads that mapped with an edit distance of at most 3. We then eliminated all pairs for which at least one of the ends mapped to the human genome without any filtering on the mapping quality or uniqueness. This loose mapping criteria is used to assure that any read pair that is likely to come from human blood contamination in the parasite samples is filtered out from our further analysis of *Plasmodium* genome architecture.

We mapped the remaining paired-end reads to the *Plasmodium falciparum 3D7* reference genome (PlasmoDB v9.0). We post-processed the alignment results further to extract the reads that mapped (i) uniquely to one location in the reference genome, (ii) with an alignment quality score of at least 30 (which corresponds to a 1 in 1000 chance that the mapping is incorrect), and (iii) with an edit distance of at most 2. We extracted the paired-end reads with both ends mapping to the *Plasmodium* genome. We then identified potential PCR duplicates, i.e., pairs of read-pairs with identical genomic coordinates, and retained only one copy of each. We also filtered out reads that map to intrachromosomal

loci that are ≤1 kb apart. We refer to the remaining reads as *informative reads*. We computed chromosomal contact maps using only these informative reads. Appendix File 1 summarizes the results of applying this pipeline to our sequencing libraries.

§ 4.2.2 Calculating noise level and percentage of long range contacts

We calculated two measures that provide estimates of the noise level and efficiency of the assay. The first is the interchromosomal contact probability (ICP) index [Kalhor et al., 2011]:

$$ICP = \frac{\sum \text{interchr contact counts}}{\sum \text{intrachr contact counts (>1 kb)}}$$

In the denominator, the intrachromosomal contact counts exclude contacts between pairs of loci ≤1 kb apart. Smaller ICP values indicate a better signal-to-noise ratio, assuming that the real data (signal) will be enriched for intrachromosomal contacts, whereas noise will be dominated by interchromosomal contacts. The second number is the percent of long-range contacts (PLRC) extracted from the initial set of paired-end reads that remain after filtering the reads that mapped to human genome:

$$PLRC = \frac{\sum \text{interchr contact counts} + \sum \text{intrachr contact counts (>20 kb)}}{\text{Number of raw reads after human DNA filtering}}$$

The bigger this percentage is, the more information the dataset provides about non-adjacent chromatin contacts for the amount of sequencing in hand.

§ 4.2.3 Aggregating data relative to 10 kb windows

Digesting the DNA with a frequently cutting restriction enzyme yields a very large number of possible pairs of restriction fragments (i.e., locus pairs). In our case, digesting the *Plasmodium* genome with MboI, which cuts at the 4 bp recognition site "GATC", yielded 28,784 fragments (mean length 810 bp) corresponding to 33,114,193 intrachromosomal and 336,629,028 interchromosomal locus pairs. For 3D modeling, we partitioned the *Plasmodium* genome into a collection of non-overlapping 10 kb windows, and we assigned each restriction fragment to the 10 kb window that covers the majority of the bases in the fragment. This operation reduced the number of possible fragments from 28,784 to 2,337 and the number of possible locus pairs from 3.7×10^8 to 2,715,615 (228,539 intrachromosomal and 2,487,076 interchromosomal).

§ 4.2.4 Normalizing raw contact maps

For each possible pair of 10 kb loci, we refer to the total number of informative read pairs that link the two loci as the *contact count*, and we refer to the two-dimensional matrix containing these contact counts as the *raw contact map*. We normalized the raw contact maps in two steps. First, we ranked loci by their percentage of intrachromosomal contacts with zero counts, and we filtered out the top 2% of this list. This removes all loci for which the signal to noise ratio is too low (typically, regions of low mappability). Second, we applied an iterative correction and eigenvector decomposition (ICE) method [Imakaev et al., 2012] that attempts to eliminate systematic biases in Hi-C data. The method estimates a bias vector with one entry per locus. The tensor product of the bias vector with itself generates a bias matrix B that can be used to convert the raw contact map into a normalized contact map.

§ 4.2.5 Estimating power-law fits to intrachromosomal contact probabilities

It has been observed in the literature that for a pair of intrachromosomal loci, the relationship between genomic distance and the expected contact count can be estimated by a log-linear model [Lieberman-Aiden et al., 2009, Fudenberg and Mirny, 2012]. This log-linear model is captured by a power-law fit of the form $P(s) \sim s^{\alpha}$ where s denotes the genomic distance, $P(s)$ denotes the expected contact probability at distance s and α is the gradient of the log-linear fit. For each stage, we first calculated $P(s)$ by segregating all intrachromosomal locus pairs into $b = 50$ equal-occupancy bins. This procedure involves enumerating all possible intrachromosomal locus pairs (including pairs that have a contact count of zero), sorting the pairs in increasing order according to their genomic distances, and then segregating the resulting list into b quantiles. For each bin i, we computed the average number of contact counts per locus pair $\hat{c}_i$, and the average contact distance $\hat{s}_i$ over all locus pairs in the bin. Then, for each bin i, $P(\hat{s}_i) = \frac{\hat{c}_i}{N}$ where N is the sum of all observed intrachromosomal contact counts. We then found the best linear fit to $\log P(s)$ versus $\log s$ in a given genomic distance range. Note that the control library "TROPH.-cont." was not subjected to normalization.

§ 4.2.6 Assigning statistical significance to normalized contact maps

To obtain a set of high confidence contacts for each stage, we subjected the contact maps at 10 kb resolution to a statistical confidence estimation procedure [Ay et al., 2014a]. We first accounted for the effect of genomic distance on the intrachromosomal contact

probability by fitting a smoothing spline to capture this effect. We then accounted for biases using the normalization procedure described above. Finally, we calculated p-values for intra and interchromosomal contacts and corrected them jointly for multiple hypothesis testing to compute q-values, which are used to filter contacts at a desired false discovery rate. For a detailed description of the statistical significance estimation procedure see Appendix Note 2.

§ 4.2.7 Identifying stage-specific contacts

We determined the contacts that are specific to only one stage or to two out of three stages as follows. First, we sorted the lists of contacts at 10 kb resolution according to increasing p-values computed as described above for each stage. Then, we extracted contacts that are ranked in top 1,000 in each stage and checked to see whether they appear among top 10,000 contacts for the other two stages. We labeled these contacts as stage-specific because they are among the strongest contacts for one stage but not among moderately-strong contacts for the other two stages. Similarly, we labeled contacts that are in top 1,000 in two out of three stages but not in top 10,000 for the third stage. To perform gene set enrichment analysis (GSEA), we extracted the lists of genes that are involved in stage-specific contacts (only ring, only trophozoite or only schizont) as well as contacts common to two stages (common to ring and trophozoite, common to ring and schizont or common to trophozoite and schizont).

§ 4.2.8 Inferring the 3D structures

Our method for inferring the 3D structures is based on the method of Duan et al. [2010]. Each chromosome is modeled as a series of beads on a string, spaced approximately 10 kb apart. We associated with each pair of beads x_i and x_j a physical *wish distance* δ_{ij}—i.e., the distance that we aim to capture with our 3D model—derived from the bead pair's contact count c_{ij}. We then placed all the beads in 3D space such that the distance d_{ij} between the beads i and j is as close as possible to the wish distance δ_{ij}.

Wish distances: To obtain the wish distances, we note that two proximal intrachromosomal loci are likely to come into contact due to random looping of the DNA, and that this "polymer packing" contact likelihood can be expressed as a function of the genomic distance s between the loci. We then assumed that two loci with observed contact count c_{ij} will have the same physical distance δ^{ij} as two intrachromosomal loci with expected contact count c_{ij} by polymer packing. The relationship between the expected contact frequencies and the genomic distances s suggests that *P. falciparum*'s

DNA behaves like a fractal globule polymer [Lieberman-Aiden et al., 2009] (Appendix Fig. 1). Any crumpled polymer exhibits a well-defined relationship between its genomic length s and the physical distance d [Grosberg et al., 1988]:

$$d \sim s^{1/3} \tag{3.1}$$

Therefore, using the relationship between genomic distances s and contact frequencies c, obtained by the fitting of the linear model, and the relationship between physical distances d and genomic distances s (Equation 3.1), we inferred a mapping between contact frequencies c and physical distances d up to a factor. We arbitrarily set the distance of the two beads with the smallest non-zero contact count $c_{\min}$ to be at a certain percentage β of the nucleus diameter. Note that $c_{\min}$ is not necessarily equal 1 since the contact counts are normalized. The β parameter hence sets the scaling of the physical distances. We then obtain:

$$\delta_{ij} = \frac{\beta 2r}{c_{\min}^{\alpha/3}} c_{ij}^{\alpha/3} \tag{3.2}$$

where r is the nucleus radius, and α the coefficient obtained in the linear model fitting (range: 30–500 kb, $\alpha = -0.963$ for rings, $\alpha = -1.124$ for trophozoites, $\alpha = -1.013$ for schizonts). We set all distances larger than the nucleus diameter to this value.

Optimization: Given the resulting physical wish distances, we defined the following optimization problem to find a structure $\mathbf{X} \in R^{3\times n}$, where n is the number of beads:

$$\begin{aligned} \underset{\mathbf{X}}{\text{minimize}} \quad & \sum_{\delta_{ij}\in\mathcal{D}} \tfrac{1}{\delta_{ij}^2}\left(d_{ij} - \delta_{ij}\right)^2 \\ \text{subject to} \quad & x_i^T x_i \le r_{\max}^2, && i = 1:n \\ & d_{i,i+1} \le b^{\max}, && i = \{1:n \mid \text{chr}_i = \text{chr}_{i+1}\} \end{aligned}$$

where d_{ij} is the Euclidean distance between beads x_i and x_j, $\mathcal{D} = \{\delta_{ij} | \delta_{ij} \neq 0\}$ is the set of non-zero wish distances, and $b^{\max}$ is defined below.

The constraints are as follows:

1. *All loci must lie within a spherical nucleus centered on the origin.* Electron microscopy experiments show that the nucleus roughly resembles a sphere, with the radius depending on the stage of the organism. In this work, we use a nuclear radius of $r = 350$ nm for the ring stage, $r = 850$ nm for the trophozoite stage and $r = 425$ nm for the schizont stage [Bannister et al., 2005, Weiner et al., 2011].

2. *Two adjacent loci must not to be too far apart.* 1000 bp of chromatin occupies a distance between 6.6–9.1 nm [Berger et al., 2008]. Because we use 10 kb resolution, we set $b^{\max} = 91$ nm.

Initialization: We create a population of 100 independently optimized structures by initializing $\mathbf{X}$ randomly from a standard normal distribution.

Measuring similarities between structures: To compare pairs of structures (X, Y) we used the standard RMSD measure:

$$\text{RMSD} = \min_{X^*} \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i^* - y_i)^2} \tag{3.3}$$

where X^* is obtained by translating and rotating X. To compare structures of different scale (e.g., different β values), we seek, in addition of the translation and rotation factor, the scaling factor that minimizes the RMSD between structures.

Another similarity measure we use to compare two structures is the average difference of their pairwise distance matrices (at 10 kb resolution), which we denote by *distance difference*:

$$\text{distance difference} = \frac{1}{n(n-1)/2}\sum_{i>j}|d_{ij}^X - d_{ij}^Y| \tag{3.4}$$

where d^X and d^Y are the Euclidean distance matrices of the structures X and Y.

Clustering the population of structures: In order to see whether the structures fall into discrete groups, we computed the RMSD between pairs of structures and performed hierarchical clustering on the resulting 100×100 distance matrix for each stage (Appendix Fig. 5).

Choosing the parameter β: As noted above, the parameter β controls the scaling of the inferred 3D structure. A small value of β will yield a structure with a very dense center, and a large value of β will push all beads against the nuclear envelope. The literature suggests that chromatin should abut the nuclear envelope [Weiner et al., 2011]. Assuming the chromatin should also occupy the center of the nucleus, we ran the entire optimization multiple times, and we selected a value of β that yields a chromatin density as close as possible to a uniform distribution.

This procedure required that we estimate the density of chromatin at a distance ℓ from the center of the nucleus. To do so, we first created an intermediate function

$$f(\ell) = \sum_{i=1}^{N} g\left(\ell - \sqrt{x_i^2 + y_i^2 + z_i^2}\right),$$

where $g(\cdot)$ is a Gaussian ($\mu = 0$, $\sigma = 10$ nm). The standard deviation σ of the Gaussian corresponds to the uncertainty of the position of each bead. The estimated density $D(\ell)$ was then computed as a generalized histogram, using discretized distance bins ℓ_i. To ensure that the volume was constant for each bin, the bin spacings were defined as $\ell_i = i^{1/3}\ell_1$, where we chose $\ell_1 = \frac{r}{3}$. We then normalized the histogram to sum to one.

Let D_i be the density of bin i and let n_{bins} be the number of bins. To select β, we defined the scoring function

$$\text{score} = \sqrt{\sum_{i=1}^{n_{\text{bins}}} \left(D_i - \frac{1}{n_{\text{bins}}}\right)^2}, \tag{3.5}$$

which corresponds to the mean squared error between the estimated density and the expected density. The resulting density scores are shown in Appendix Table 13, with the minimal value for each stage in boldface.

§ 4.2.9 Eigenvalue decomposition and chromatin compartments

To identify chromatin compartments, for each stage, we carried out eigenvalue decomposition on the matrix of Euclidean distances between locus pairs. For each chromosome we used the intrachromosomal 3D distance matrix at a resolution of 10 kb, where each 10 kb locus is represented by the 3D coordinate of its midpoint. We then calculated the Spearman correlation between each pair of rows of the 3D distance matrix and applied eigenvalue decomposition (using the *eig* function in MATLAB) to this correlation matrix. The sign of the first eigenvector defined a compartment assignment for each 10 kb locus at each stage. We also aggregated all three stages and calculated a set of aggregate compartments (Appendix Fig. 3, fourth row of figures on each page) which divided each chromosome into three main compartments (i.e., telomeric-centromeric-telomeric or left(L)-mid(M)-right(R)).

§ 4.2.10 Kernel canonical correlation analysis

We used an approach based on kernel canonical correlation analysis (kCCA) [Bach and Jordan, 2002, Vert and Kanehisa, 2003b,a] to extract gene expression profiles that simultaneously capture the variance of the gene expression data and exhibit coherence with respect to the 3D structure.

Let $\mathcal{G}$ be the set of n genes. Each gene $g \in \mathcal{G}$ is characterized by its log expression profile $e(g) = (e_1(g), \ldots, e_p(g)) \in \mathbb{R}^p$ at p timepoints and by its position $x(g) \in \mathbb{R}^3$ in 3D space. We assume that the set of gene expression profiles is mean centered and unit variance scaled, i.e., $\sum_{g\in\mathcal{G}} e_i(g) = 0$ and $\frac{1}{|\mathcal{G}|}\sum_{g\in\mathcal{G}} e_i(g)^2 = 1$ for $i = 1, \ldots, p$.

Let $v \in \mathbb{R}^p$ be a direction in the expression profile space. To assess whether v is representative of the observed expression profiles, we computed the percentage of variance explained among the gene expression profiles once they are projected onto v, defined by

$$V(v) = \frac{\sum_{g\in\mathcal{G}} \left(v^T e(g)\right)^2}{\|v\|^2} \,. \tag{3.6}$$

The larger $V(v)$ is, the more v explains the differences between gene expression profiles, and the more likely v is to correspond to some biological event which influences the expression of many genes. $V(v)$ is, for example, maximized by principal component analysis.

Instead of just asking the profile v to capture variance among gene expression, we simultaneously asked it to exhibit coherence with respect to the 3D structure. For that purpose, we defined for every $f \in \mathbb{R}^n$ a function $S(f)$ that quantifies how smoothly f varies in 3D. f can be thought of as a vector of scores, one score being assigned to each gene. Because we know the 3D coordinates of each gene we can imagine f as a set of scores in 3D. Following a standard approach in kernel methods [Schölkopf and Smola, 2002], we quantified the smoothness of f with the function

$$S(f) = \frac{f^\top K_{3D}^{-1} f}{||f||^2} \,, \tag{3.7}$$

where K_{3D} is the $n \times n$ matrix whose (i, j) entry is the Gaussian kernel between genes i and j, namely, $\exp\left(-||x(i) - x(j)||^2/2\sigma^2\right)$. The smaller $S(f)$ is, the more smoothly f is distributed in 3D.

We then combined the ideas of capturing variance (Equation 3.6) and being smooth in 3D (Equation 3.7) by designing a joint objective function over v and f to ensure that (i) v captures a lot of variance, (ii) f is smooth in 3D, and (iii) f is maximally correlated with the vector $\left(v^\top e(g)\right)_{g\in\mathcal{G}}$. In words, we aimed to ensure that genes which are positively

correlated with v (and those which are negatively correlated) tend to be co-localized in 3D. We designed the function by following the approach of Bach and Jordan [2002], who show that v and f can be found by solving a kCCA problem equivalent to the following generalized eigenvalue problem:

$$\begin{pmatrix} 0 & K_E K_{3D} \\ K_{3D} K_E & 0 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \rho \begin{pmatrix} (K_E + \delta I)^2 & 0 \\ 0 & (K_{3D} + \delta I)^2 \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix},$$

where K_{3D} is the $n \times n$ matrix whose (i, j)-th entry is $e(i)^\top e(j)$, and δ is a small regularization parameter. Once we found the generalized eigenvectors $(\alpha, \beta)^\top$, ranked by decreasing eigenvalue ρ, we recovered a pair (v, f) by $v = \sum_{g \in \mathcal{G}} \alpha_g e(g)$ and $f = K_{3D}\beta$.

We computed the profiles for several values of σ (0.01, 0.02, 0.05, 0.1) and δ (0.01, 0.02, 0.04, 0.06) and obtained highly correlated results (correlation > 0.99 for all pairs of profiles). Therefore, we chose $\sigma = 0.01$ and $\delta = 0.02$ for the rest of the analysis.

§ 4.2.11 Gene set enrichment analysis

To detect set of genes highly or poorly correlated kCCA profiles, we apply gene set enrichment analysis (GSEA) [Subramanian et al., 2005]. Unlike a traditional GO term enrichment analysis, this method takes as input a ranked list of genes rather than a set of genes; hence, GSEA takes full advantage of the results of the kCCA. The procedure detects sets of genes enriched at the top or at the bottom of the ranked list of genes. We applied GSEA to the ranked list of projections of expression profiles on the first and second extracted profile. Corresponding p-values were computed using $4,000$ permutations. We also used GSEA in our comparison gene sets that are involved in contacts that are specific to either one stage or two out of three stages.

§ 4.2.12 Volume exclusion model

Following the methodology of Tjong et al. [2012], we constructed a population of three-dimensional structures by modeling chromosomes as random configurations subject to the following constraints:

1. Each chromosome is modeled as a series of N beads spaced 3.2 kb apart, with consecutive beads restrained to be 30 nm apart.

2. Overlaps between beads are prevented by imposing a volume exclusion constraint for all pairs of beads.

3. All chromosomes lie within a spherical nucleus of a specified radius.

4. All centromeres are colocalized in a small sphere of radius 50 nm abutting the nuclear envelope.

5. All telomeres are located within 50 nm of the nuclear envelope.

We formulated an optimization problem that includes, in addition to the constraints, a penalty term that accounts for chromatin stiffness by placing an angular restraint between three consecutive beads:

$$\frac{1}{2}k_{\text{angle}}\sum_{i=1}^{N-2}\left(1-\frac{x_{i+1}-x_i}{\|x_{i+1}-x_i\|}\cdot\frac{x_{i+2}-x_{i+1}}{\|x_{i+2}-x_{i+1}\|}\right)^2, \tag{3.8}$$

where $x_i \in \mathbb{R}^3$ is the coordinate vector of bead i. We used the Integrated Modeling Platform (IMP) [Bau et al., 2011] to generate 5,000 budding yeast structures with a nuclear radius of 1000 nm, and 5,000 *Plasmodium* structures for each of the three stages with nuclear radii of 350 nm, 850 nm and 425 nm, respectively.

Following Tjong et al. [2012], we used the population of structures to generate a volume exclusion (VE) contact frequency matrix C, considering that two beads are in contact when they are $\leq$45 nm apart. The contact frequency matrix was then aggregated to a resolution of 32 kb and normalized following the ICE procedure as described above, resulting in a contact frequency matrix c_{ij}^{VE} for $i, j = 1, \ldots, N$ according to the VE model.

In order to compare the VE contact matrix to experimental Hi-C data, we similarly computed the Hi-C contact count matrix at a resolution of 3.2 kb, aggregated it at 32 kb, and normalized the same way as the VE contact frequency matrix to get a Hi-C contact matrix c_{ij}^{HIC} for $i, j = 1, \ldots, N$.

We then compared both matrices by computing the row-based Pearson correlation [Tjong et al., 2012] defined as the average Pearson correlation between their rows.

$$\frac{1}{N}\sum_{i=1}^{N}\frac{N\sum_{j\neq i}^{N}c_{ij}^{\text{HIC}}c_{ij}^{\text{VE}}-\sum_{j\neq i}^{N}c_{ij}^{\text{HIC}}\sum_{j\neq i}^{N}c_{ij}^{\text{VE}}}{\sqrt{N\sum_{j\neq i}^{N}(c_{ij}^{\text{HIC}})^2-(\sum_{j\neq i}^{N}c_{ij}^{\text{HIC}})^2}\sqrt{N\sum_{j\neq i}^{N}(c_{ij}^{\text{VE}})^2-(\sum_{j\neq i}^{N}c_{ij}^{\text{VE}})^2}}. \tag{3.9}$$

Furthermore, we also computed a *normalized* row-based Pearson correlation between the matrices by replacing the counts c_{ij}^{VE} and c_{ij}^{HIC} in (Equation 3.9) by their ratio to an expected count c_{ij}^{E} that we would expect if there was no structural information in the matrix, besides the obvious decrease of contacts between loci at increasing genomic distance. To estimate the expected frequencies c_{ij}^{E} used to define the ratios, we fit an isotonic regression to the mapping between genomic distance and the average contact

frequency at this genomic distance. The isotonic regression allows us to fit a non-increasing mapping between genomic distance and contact frequency, thus correcting the effect of enrichment of contact frequencies at chromosome ends. This mapping allowed us to define e_{ij}^{E} as the expected count corresponding to the genomic distance between loci i and j in the case of intrachromosomal contacts, and to the genome-wide average of inter chromosomal counts in case of interchromosomal contacts.

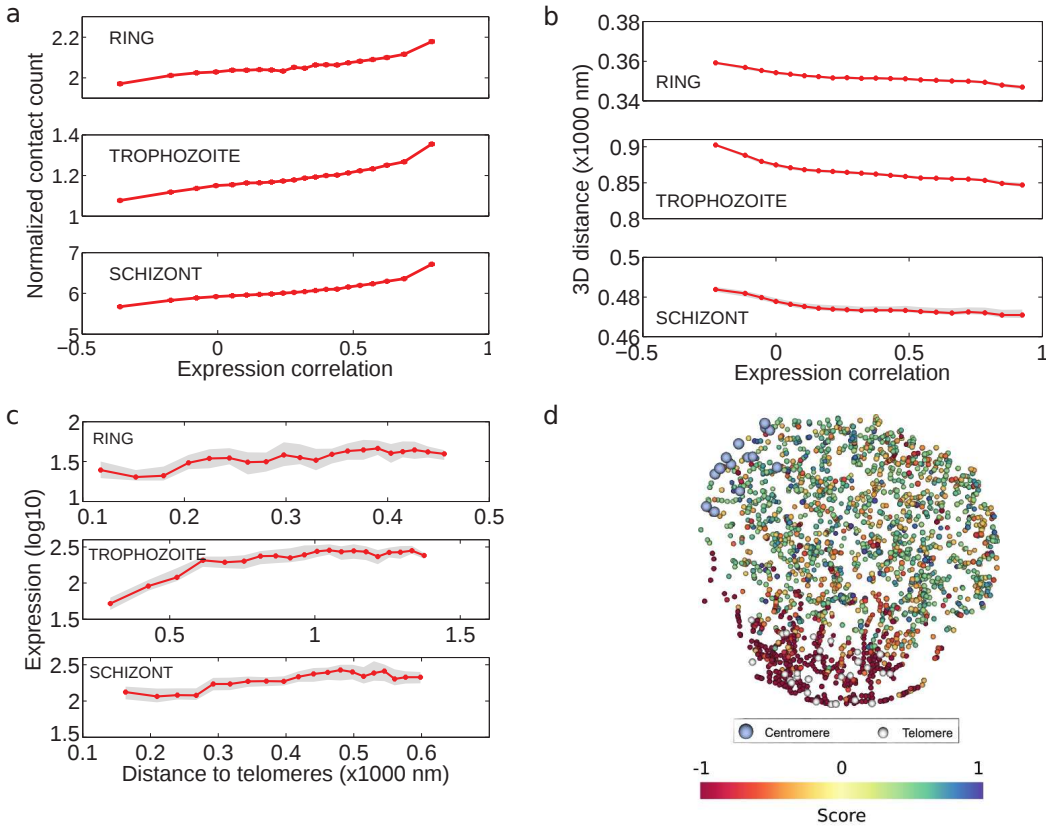


FIGURE 6: **Relationship between 3D architecture and gene expression. a,**Correlation between expression profiles of pairs of interchromosomal genes as a function of number of contacts linking the two genes. To generate this plot all interchromosomal gene pairs are first sorted in increasing order of their expression correlation and then binned into 20 equal width quantiles (5th, 10th, ..., 100th). For each bin, the average expression correlation between gene pairs (x-axis) and the average normalized contact count linking the genes in each pair together with its standard error (y-axis) are computed and plotted. Interchromosomal gene pairs that have contact counts within the top 20% for each stage have more highly correlated expression profiles than the remaining gene pairs [Wilcoxon rank-sum test, p-values 2.48e-206 (ring), 0 (trophozoite), and 0 (schizont)]. **b,** Correlation between expression profiles of pairs of interchromosomal genes as a function of 3D distance between the genes. This plot is generated similar to **a** but with using 3D distances instead of contact counts (y-axis). In order to summarize results from multiple 3D structures per each stage, we plot the median value among 100 structures with a red line and shaded the region corresponding to the interval between 5th and 95th percentile with gray. Interchromosomal gene pairs closer than 20% of the nuclear diameter have more highly correlated expression profiles than genes that are far apart [Wilcoxon rank-sum test, p-values 7.17e-221 (ring), 0 (trophozoite), and 1.57e-88 (schizont)]. **c,** Gene expression as a function of distance to telomeres. To generate this plot all genes are first sorted by increasing distance to the centroid of telomeres (x-axis) and then binned similar to **a** into 20 equal width quantiles. The average log expression value [Bunnik et al., 2013] together with its standard error (y-axis) is plotted for genes in each bin. In order to summarize results from multiple 3D structures per each stage, we plot the median value among 100 structures with a red line and shaded the region corresponding to the interval between 5th and 95th percentile with gray. Genes that lie within 20% of the nuclear diameter to the centroid of the telomeres showed significantly lower expression levels [Wilcoxon rank-sum test, p-values 1.54e-12 (ring), 1.69e-32 (trophozoite), 3.37e-20 (schizont)]. **d,** First kCCA expression profile component score, corresponding to the projection of the gene expression profile onto the extracted kCCA profile for the trophozoite stage.

Chapter 4

Accurate identification of centromere locations in yeast genomes using Hi-C

This chapter has been published in a slightly modified form in [Varoquaux et al., 2015], as joint work with Ivan Liachko, Josh Burton, Ferhat Ay, Jay Shendure, Maitreya Dunham, Jean-Philippe Vert and Bill Noble.

Résumé

Les centromères sont des éléments génomiques permettant la ségrégation correcte des chromosomes lors de la division cellulaire. Malgré leur importance dans le développement de la cellule et l'important effort de recherche qui leur est dédié, la position des centromères chez la levure est souvent, même de nos jours, difficile à inférer et est inconnue chez la plupart des espèces. Récemment, le protocole de capture de conformation des chromosome Hi-C, initialement développé pour étudier la structure 3D du génome, a été reciblé pour diverses applications: séquençage *de novo* de génome, déconvolution d'échantillon métagénomique, et inférence de la position des centromères chez la levure. Nous décrivons ici une méthode, nommée Centurion, qui permet l'inférence conjointe de la position de tous les centromères à la fois d'un organisme à partir de données Hi-C en exploitant la propriété qu'ont les

centromères de certains organismes à colocaliser dans le noyau. Nous démontrons dans un premier temps la précision de notre algorithme, en identifiant les centromères dans des données Hi-C à haute couverture chez la levure de boulanger *S. cerevisiae* et le parasite responsable de la malaria *P. falciparum*. Nous utilisons ensuite Centurion pour prédire la position des centromères dans 14 autres espèces de levure d'un échantillon métagénomique. Parmi tous les organismes que nous étudions, Centurion prédit 89% de centromères à moins de 5 kb de leur position. Nous démontrons par ailleurs la robustesse de notre approche sur des jeux de données à faible couverture. Finalement, nous inférons la position des centromeres dans 6 espèces qui n'ont pour l'instant aucune annotation. Ces résultats montrent que Centurion peut être utilisé pour l'identification de centromères pour différentes espèces de levures, ainsi que pour d'autres organismes.

Abstract

Centromeres are essential for proper chromosome segregation. Despite extensive research, centromere locations in yeast genomes remain difficult to infer, and in most species they are still unknown. Recently, the chromatin conformation capture assay, Hi-C, has been re-purposed for diverse applications, including *de novo* genome assembly, deconvolution of metagenomic samples, and inference of centromere locations. We describe a method, Centurion, that jointly infers the locations of all centromeres in a single genome from Hi-C data by exploiting the centromeres' tendency to cluster in 3D space. We first demonstrate the accuracy of Centurion in identifying known centromere locations from high coverage Hi-C data of budding yeast and a human malaria parasite. We then use Centurion to infer centromere locations in 14 yeast species. Across all microbes that we consider, Centurion predicts 89% of centromeres within 5 kb of their known locations. We also demonstrate the robustness of the approach in datasets with low sequencing depth. Finally, we predict centromere coordinates for six yeast species that currently lack centromere annotations. These results show that Centurion can be used for centromere identification for diverse species of yeast and possibly other microorganisms.

§ 1 Introduction

Centromeres are chromosomal regions whose function enables faithful chromosome segregation via formation of the kinetochore [Bloom, 2014]. These elements are also key regulators of genome stability [Feng et al., 2009] and replication timing [Koren et al., 2010, Pohl et al., 2012]. In animal and plant genomes, centromeres are large heterochromatic zones, but many yeast species have *point centromeres*, which are sequence elements as small as 125 bp [Cottarel et al., 1989]. The relative simplicity of yeast centromeres has allowed their functional dissection, and the abundance of sequenced yeast species has shed light on the evolution of centromeric elements across hundreds of millions of years of evolution [Gordon et al., 2011].

The Hemiascomycetes yeasts comprise a highly important taxon of model organisms in genetics and genomics [Dujon, 2010, Hittinger, 2013], and some are crucial in biotechnology applications such as recombinant protein expression [Böer et al., 2007]. Most yeast plasmid expression systems are dependent on locating and identifying yeast centromeres because they confer the property of stable segregation to episomal plasmids [Murray and Szostak, 1983]. However, efforts to annotate yeast centromeres are hindered by the extraordinary diversity among species [Malik and Henikoff, 2009]. Mapping centromeres in diverse species has been attempted, usually through phylogenetic tools [Gordon et al., 2011, The Génolevures Consortium et al., 2009] or chromatin immunoprecipitation [Lefrancois et al., 2009]. However, both approaches have drawbacks, the former due to the divergence of underlying functional motifs and the latter due to nonspecific signal. A method of mapping centromeres that does not rely on evolutionary predictions or rare protein-DNA interactions would therefore be useful for identifying centromeres in novel species. These new centromere sequences could then be used, for example, to build new plasmid-based strain engineering tools in species important for research and biotechnology.

Chromosome conformation capture tools such as Hi-C and related protocols use proximity ligation and massively parallel sequencing to probe the three-dimensional architecture of chromosomes within the genome [Lieberman-Aiden et al., 2009, Kalhor et al., 2011, Duan et al., 2010]. Hi-C and related techniques create a *contact map*, consisting of a matrix of genome-wide interaction counts between pairs of loci. Contact maps have recently been shown to contain long-range contiguity information: Hi-C has been used in the scaffolding of *de novo* genome assemblies [Burton et al., 2013, Kaplan and Dekker,

2013], molecular haplotyping [Selvaraj et al., 2013], and metagenomic deconvolution [Burton et al., 2014, Beitel et al., 2014]. These methods have also paved the way for a more systematic analysis of genome architecture, including long-range gene regulation and chromatin architecture [Nora et al., 2012, Dixon et al., 2012, Mizuguchi et al., 2014]. These advances raise the possibility that contact maps might be used to determine the location of subchromosomal genomic structures such as centromeres and nucleoli.

A recent study attempted to map centromere locations using Hi-C contact probability maps [Marie-Nelly et al., 2014b]. This approach exploits the strong architectural features of yeast genomes to determine centromere positions and rDNA clusters in *Saccharomyces cerevisiae*, *Naumovozyma castellii*, *Nuraishia capsulata*, and *Debaryomyces hansenii*. In yeasts, centromeres are tethered by microtubules to the spindle pole body, leading to centromere clustering [Mizuguchi et al., 2014]. Similar clustering is also present in other organisms, such as the parasite *Plasmodium falciparum* and the plant *Arabidopsis thaliana* [Ay et al., 2014b, Feng et al., 2014]. The clustering of elements creates a distinct peak of interactions between chromosomes in the *trans* Hi-C matrix, and an X-shape in the *cis*-elements of the inter-chromosomal contact counts pearson correlation matrix. Marie-Nelly et al. [2014b] exploit this X-shape structure in *trans* contact counts correlation matrices to first detect a 40 kb window containing each centromere. In a subsequent step, they carve out 40 kb-by-40 kb windows of contact counts for each pair of centromeres and refine the prediction by fitting a Gaussian on the sum of *trans* elements of these windows, a procedure similar to those used for single molecule localization or high resolution microscopy [Ober et al., 2004]. However, this method has several limitations. First, the procedure relies on the correct pre-localization of candidate centromeres. This step fails when other sequences also colocalize (for instance, rDNA sequences). Second, the last step of the procedure collapses the data of several *trans* interaction windows into a 1D profile and calls the different centromeres independently from each other, thus potentially losing some valuable information.

Here we propose a novel method, Centurion, that jointly calls all centromeres in a genome-wide Hi-C contact map. The key idea is that a joint optimization can effectively exploit the clustering of centromeres in 3D. We first compare our method to the one described by Marie-Nelly *et al.* on four publicly available high-resolution Hi-C contact maps (*S. cerevisiae* [Duan et al., 2012] and three stages of *P. falciparum* [Ay et al., 2014b]). This comparison demonstrates that Centurion infers centromere positions more accurately than the previously published method. We then apply our method to Hi-C data from 14 diverse yeast species [Burton et al., 2014], yielding high-resolution centromere location predictions for each chromosome in each species. For the eight species that already have centromere annotations available, our predictions match very closely

with the existing calls. For species with as-yet uncharacterized centromeres, our predictions will serve as the basis for targeted experimental validation and could be used to create new plasmid tools in these yeasts. Our results suggest that Centurion has great potential to identify the centromere locations of many yeasts for which standard techniques have failed to date. Furthermore, we demonstrate that Centurion works well even with very limited sequencing depth Hi-C libraries generated from pooled samples, making it a practical as well as powerful tool to use on single microorganisms and metagenomic mixtures. Centurion is freely available as open source software at http://cbio.ensmp.fr/centurion.

§ 2 Method

§ 2.1 Single organism Hi-C data

We use Hi-C data gathered in two previous studies: an asynchronous budding yeast (*S. cerevisiae*) sample [Duan et al., 2010] and three different stages of the human malaria parasite *P. falciparum* [Ay et al., 2014b]. For the budding yeast Hi-C data we download and use the files HindIII + MspI (intra and inter) from `http://noble.gs.washington.edu/proj/yeast-architecture/sup.html`. For the three stages of *P. falciparum* we download and use the Hi-C raw contact counts at 10 kb resolution from GEO archive (Accession codes: GSM1215592, GSM1215593, GSM1215594).

§ 2.2 Metagenomic Hi-C data

For Hi-C data from metagenomic samples we use the two synthetic mixtures (M-Y, M-3D) generated in [Burton et al., 2014]. We also perform additional sequencing of the M-3D sample using two restriction enzymes that cut more frequently than the 6-bp cutters HindIII and NcoI used in the original publication. We perform these additional Hi-C experiments exactly as described in [Burton et al., 2014] with the exception that we use Sau3AI (a 4-bp cutter that recognizes "GATC") and AflIII (a 6-bp cutter that recognizes "ACRYGT") to fragment the DNA. We then combine the reads from these two libraries (Sau3AI and AflIII) to produce Hi-C contact maps.

We process the Hi-C libraries from these metagenomic samples in a similar fashion to the Hi-C data from the above mentioned single organism samples, with the exception of two differences. First, we map the reads to a meta-reference genome that concatenates the reference genomes of all the organisms in the corresponding sample. This mapping strategy discards contacts which cannot be uniquely assigned to a single organism,

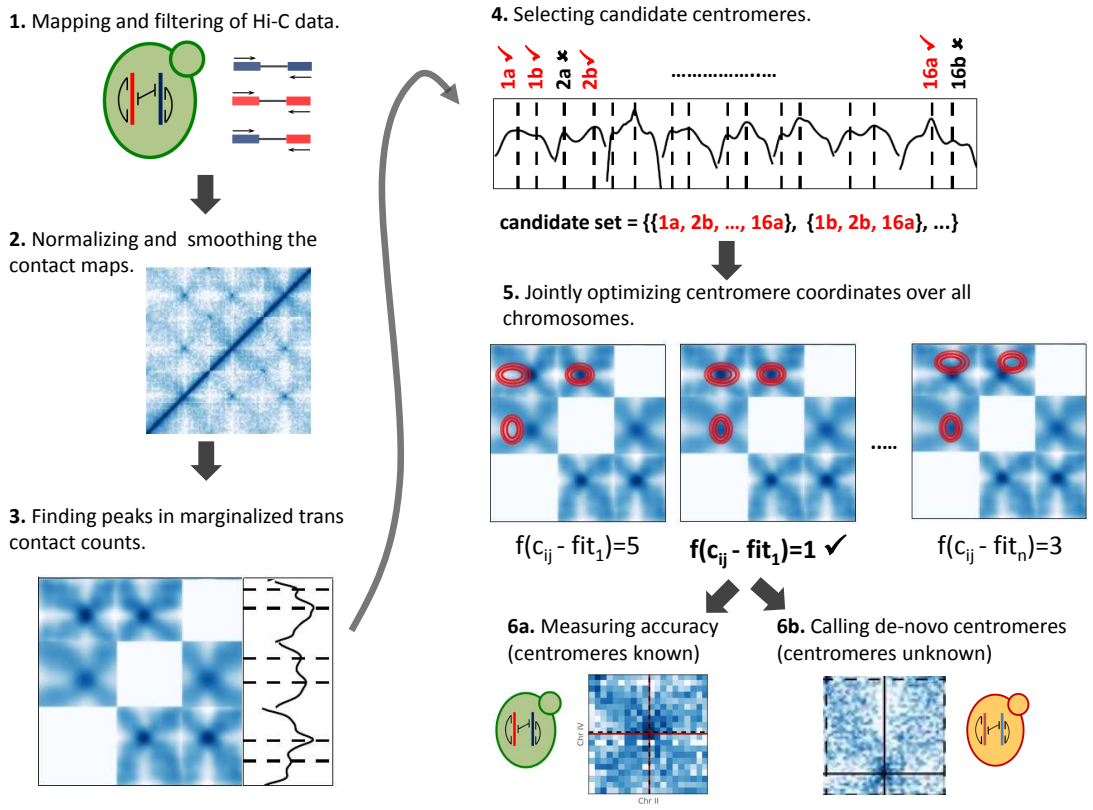


FIGURE 1: **Outline of Centurion's computational workflow 1.** Paired-end Hi-C reads are mapped and filtered to produce genome-wide contact maps (see Methods). **2.** Contact maps are normalized to correct for technical and experimental biases [Imakaev et al., 2012]. **3.** Peaks in marginalized *trans* contact counts are identified as candidate centromere locations. **4.** If necessary, a heuristic reduces the number of centromere candidates that will be used to initialize the joint optimization. **5.** A joint optimization procedure finds the best set of centromere coordinates, one per chromosome, minimizing the squared distance between the 2D Gaussian fits and the observed *trans* contact counts. **6.** For organisms with known centromere locations, the accuracy of predicted centromere locations is evaluated; otherwise, the method provides *de novo* centromere calls.

thereby reducing contamination between contact maps. Second, because of the longer read lengths for the metagenomic libraries compared to single organisms (80–101 bp versus 20–50 bp), we post-process the non-mapped reads that contain a cleavage site for the restriction enzyme used for the library generation, as previously described [Ay et al., 2015b]. This post-processing increases the number of informative contacts extracted from the metagenomic Hi-C libraries by 5-15% depending on the read length and the cleavage site frequency. The resulting set of informative contacts are processed further at appropriate resolution, as described below.

§ 2.3 Assembling the *K. wickerhamii* genome

Two input genome assemblies are used for creating the new *K. wickerhamii* reference genome. The first is the publicly available *K. wickerhamii* reference genome originally sequenced by Baker *et al.* Baker et al. [2011], and the second is the *K. wickerhamii* associated cluster from Burton *et al.* Burton et al. [2014]. These assemblies are merged with CISA [Lin and Liao, 2013] and then merged using the mate-pair library from [Burton et al., 2014] using the "scaffold" command from IDBA [Peng et al., 2012]. Hi-C reads are then aligned to this assembly, and the seven scaffolds containing the 7 *K. wickerhamii* centromeres are identified. Lastly, this assembly is run through Lachesis [Burton et al., 2013], with a restriction that the seven centromere-containing scaffolds could not be merged.

§ 2.4 Data normalization

Hi-C contact counts are subjected to many biases (GC-content, mappability, etc) [Yaffe and Tanay, 2011]. To correct for technical biases, we apply to the raw contact counts an iterative correction and eigenvector decomposition (ICE) method proposed by Imakaev et al. [2012], based on the assumption that all loci should interact equally. We then rescale the resulting matrix such that the average normalized contact count is equal to the average raw contact counts.

§ 2.5 Centromere calling

We segment the full genome into N windows of similar length ($N = 611$ for *S. cerevisae* at 20 kb) and summarize the Hi-C data by the contact count matrix $C \in \mathbb{R}^{N \times N}$, where C_{ij} is the normalized number of physical interactions captured between loci in windows i and j. For each window $i \in [1, N]$ we denote by $B(i) \in [1, L]$ the chromosome to which window i belongs, L being the total number of chromosomes ($L = 16$ for *S. cerevisae*). We also denote by x_i the genomic coordinate of the center of the i-th window. Our objective is to infer the genomic coordinates $p = (p_1, \ldots, p_L)$ of the centromeres of the L chromosomes. More precisely, centromeres usually consist of a sequence with a length ranging from several hundred base pairs for point centromeres to several thousand base pairs for regional centromeres. In this work, we infer the mean position of these sequences.

Our main assumption is that, because centromeres colocalize in the nucleus, we expect loci near centromeres in different chromosomes to be enriched in Hi-C contacts. To capture this enrichment, we model the contact counts between windows i and j of

different chromosomes k and l by a 2-D Gaussian function centered on the corresponding centromeres p_k and p_l:

$$a \exp\left(-\frac{(x_i - p_k)^2 + (x_j - p_l)^2)}{2\sigma^2}\right) + b\,,$$

with parameters a, b and $\sigma \geq 0$. Then, denoting by $\mathcal{D}$ the set of pairs of windows (i, j) from different chromosomes with non-zero counts, we jointly estimate the parameters (a, b, σ) and the positions of the L centromeres by a least-squares fit of the Hi-C count data, namely, by minimising in $a, b, \sigma \geq 0$ and $p = (p_1, \ldots, p_L)$ the following objective function:

$$\sum_{(i,j)\in\mathcal{D}} \left[C_{ij} - a \exp\left(-\frac{(x_i - p_{B(i)})^2 + (x_j - p_{B(j)})^2)}{2\sigma^2}\right) - b\right]^2 . \tag{4.1}$$

Note that in this optimization, the position of each centromere is constrained to be on its corresponding chromosome. Note also that for each non-zero entry of the contact count matrix, we only fit the Gaussian centered on the corresponding pair of loci. Thus, when the centromeres are close to a chromosome boundary, we only fit a truncated Gaussian.

§ 2.6 Initializing the optimization problem

Because the optimization problem (4.1) is non convex, the local minimum found by the algorithm depends on the initialization of the parameters, in particular of the centromeres' positions. We therefore need a heuristic to initialize centromere positions. Because centromeres tend to interact in *trans* with other centromeres, a simple heuristic is to choose the position on each chromosome at the center of the window with the largest total number of *trans* contact counts. However, we found that this heuristic was often not sufficient, because other loci besides centromeres, such as telomeres or rDNA clusters, can exhibit large numbers of *trans* interactions. We therefore implemented another heuristic to generate other good initializations and to explore more local minima. In short, on each chromosome we detect a few local maxima (typically, two per chromosome) of a smoothed *trans* contact counts curve. We then initialize the optimization by combining each choice of centromere location among the candidates on each chromosome. If time constraints do not allow us to test all such initializations (with 2 choices on 14 chromosomes, this corresponds to $2^{14} = 16384$ different initializations), then we can further reduce the exploration of local minima by starting from the best candidate on each chromosome (i.e., with the largest number of *trans* contact counts), optimizing the objective function from this initialization, and then moving to other "nearby" local

minima of the objective function by changing centromere initialization to another candidate one centromere at a a time, until no nearby local minimum is better than the one we have converged to.

A Python implementation of the proposed method is available at http://cbio.ensmp.fr/centurion.

§ 2.7 Measuring the performance

To measure the performance of the centromere position prediction on datasets for which we have the ground truth, we compute the distance in base pairs between the prediction *pred* and the segment (b, e) as follows:

$$\max\left((b - pred)_+, (pred - e)_+\right)$$

where $(u)_+$ is u if $u \geq 0$, 0 otherwise.

§ 3 Results

§ 3.1 Validating the method on *S. cerevisiae* and *P. falciparum*

To evaluate the accuracy of our centromere prediction method, we first applied it to two organisms with known centromere coordinates and available Hi-C data. The first one is the widely studied budding yeast *S. cerevisiae*. The genome of *S. cerevisiae* has 16 chromosomes and thus 16 centromeres, all of which colocalize near the spindle pole body [Jin et al., 2013]. All 32 telomeres of *S. cerevisiae* tether to the nuclear envelope. The repetitive ribosomal DNA of *S. cerevisiae* occurs on chromosome XII and is bundled into the nucleolus at the opposite side of the nucleus from the spindle pole body [Venema and Tollervey, 1999]. These organizational principles constrain the chromosomes to fold into a distinct configuration, known as the *Rabl configuration*, which resembles a water lily shape [Zimmer and Fabre, 2011]. The contacts between centromeres in *S. cerevisiae* chromosomes are known to result in a strong enrichment of centromere-to-centromere Hi-C links [Duan et al., 2012]. We sought to evaluate Centurion's ability to pinpoint the exact centromere locations directly from a Hi-C contact map [Gotta et al., 1996].

Using 40 kb-resolution Hi-C contact maps from Duan *et al.* Duan et al. [2012] (Figure 2A and 2B), Centurion predicts centromere coordinates with an average deviation of 11 kb from the known coordinates. Notably, Centurion's Gaussian fitting procedure allows the centromere calls to achieve finer resolution than is provided by the input contact

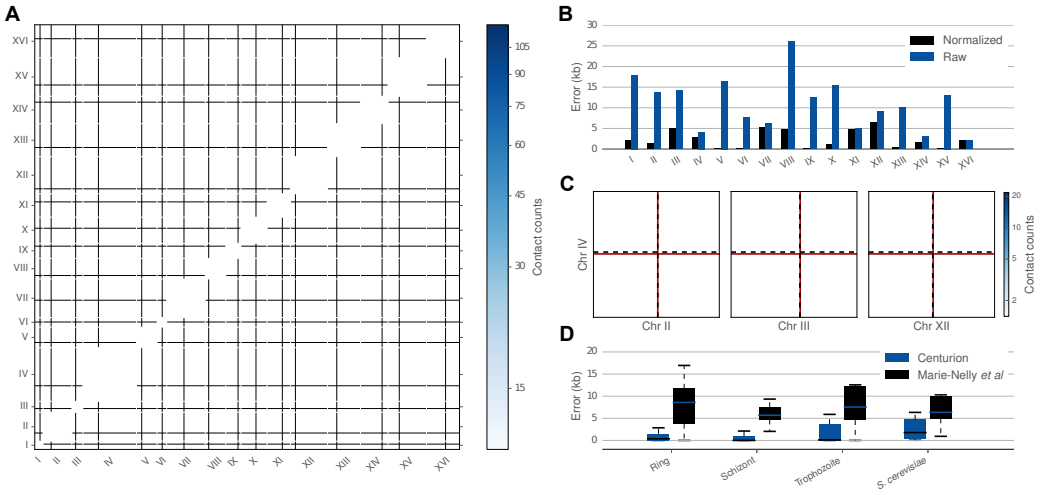


FIGURE 2: **Calling centromeres on *P. falciparum* and *S. cerevisiae* A**. Heatmap of the normalized *trans* contact counts for *S. cerevisiae* Hi-C data at 40 kb overlaid with Centurion's centromeres calls (black lines). The contact counts were smoothed with a Gaussian filter (σ = 40 kb) for visualization purposes. White lines indicate chromosome boundaries. **B.** Per chromosome errors of Centurion's centromere calls for *S. cerevisiae* using normalized (black) and raw (blue) Hi-C contact maps at 40 kb resolution. **C.** Heatmap of *trans* contact counts for *P. falciparum* trophozoite data at 40 kb overlaid with Centurion's centromere calls (dashed black line) and ground truth (red line) for chr 2, 3, 4 and 12. **D.** Average errors of centromere calls for Centurion (black) and Marie-Nelly et al. [2014b] method for *S. cerevisiae* data from Duan et al. [2012] and the three stages of *P. falciparum* when both methods are initialized with the ground truth centromere coordinates.

maps. Using 20 kb resolution contact maps, the average deviation drops to 9 kb. Furthermore, we observed that normalizing the contact maps [Imakaev et al., 2012] yields substantially improved results, reducing the average deviation to 2.5 kb for both the 20 kb and 40 kb resolution. We investigated the differences in the prediction accuracy of our method among the 16 different chromosomes. While our predictions were within 1 kb of the known centromere coordinates for the chromosomes V, VI, IX, XIII and XV (respectively, 59 bp, 235 bp, 111 bp, 289 bp and 163 bp away), they were more than 5 kb away for chromosomes III, VII and XII (respectively, 5011 bp, 5327 bp and 6457 bp away). While the cause of this fluctuation of accuracy is not yet known, chromosomes III and XII house the only major blocks of heterochromatin in this genome other than telomeres (the silent mating loci and rDNA, respectively), suggesting that linked heterochromatinized loci may interfere with accurate centromere prediction.

We then applied our method to a second species, the malaria parasite *P. falciparum*, which is responsible for the most virulent form of malaria [World Health Organization, 2012]. We recently used Hi-C to provide a global picture of the genome architecture of *P. falciparum* at three stages (ring, schizont and trophozoite) throughout its erythrocytic life cycle in human blood [Ay et al., 2014b]. Centromere coordinates for *P. falciparum*

were only identified systematically relatively recently [Hoeijmakers et al., 2012a]. We applied Centurion to the contact maps of each of these three stages at 10 kb, 20 kb and 40 kb resolutions (Appendix Fig 3). As with *S. cerevisiae*, we observe some variation in the accuracies of our predictions for each chromosome. However, overall, the accuracy is very high. At 10 kb resolution, for example, Centurion's centromere predictions fall within the known centromere location for all 14 chromosomes during the schizont stage, 13 out of 14 for the ring stage and for 11 out of 14 chromosomes in the trophozoite stage. Overall, across the three different stages Centurion correctly localizes 90%, 64%, and 45% of centromeres at 10 kb, 20 kb and 40 kb resolution, respectively. For the incorrectly called centromeres, the average distance from Centurion's prediction and the edge of the centromere is 495 bp, 1308 bp, and 2319 bp, respectively.

We next sought to understand the sources of error in our predictions. Looking closely at the contact counts matrices in the neighborhood of centromeres for which the prediction is not accurate, we observed that loci in proximity to centromeres seem to exhibit unusually sparse interaction counts. For example, Figure 2C shows that in the trophozoite stage, the centromere of chr 1 is close to a chromosome boundary and the chr 4 centromere is close to a locus with few interacting bins. The latter case leads to bias from the normalization procedure because the few nonzero entries in this sparse region are over-corrected. We also investigated whether the accuracy of our prediction varies by life cycle stage and matrix resolution (Appendix Fig 1). Many chromosomes are given consistently poor centromere calls across all life cycle stages and at all resolutions, corroborating the observations above that the predictions tend to be influenced by biases intrinsic to the genome around those centromeres, such as mappability or GC content.

We next compared the accuracy of our predictions to that of a previously published method [Marie-Nelly et al., 2014b]. Marie-Nelly *et al.* method often works well for identifying centromeres using Hi-C libraries with very high sequencing depth; however, when Hi-C sequencing depth is limited or when loci other than centromeres strongly cluster, the first step of the procedure, called "pre-localization," sometimes fails to identify the correct fixed size window in which the centromeres reside. We hypothesized that the joint centromere calling by Centurion, which leverages data from all chromosomes at once, might alleviate this instability. To test this hypothesis, we applied the Marie-Nelly *et al.* method to the same four datasets (one *S. cerevisiae* and three *P. falciparum*) described above. As shown in Appendix Figure 4, in each of these four datasets Centurion identifies centromeres with better accuracy than the Marie-Nelly *et al.* method. For instance, the colocalization of rDNA clusters and virulence genes in *P. falciparum* drastically changes the pattern of the correlation matrix used by Marie-Nelly

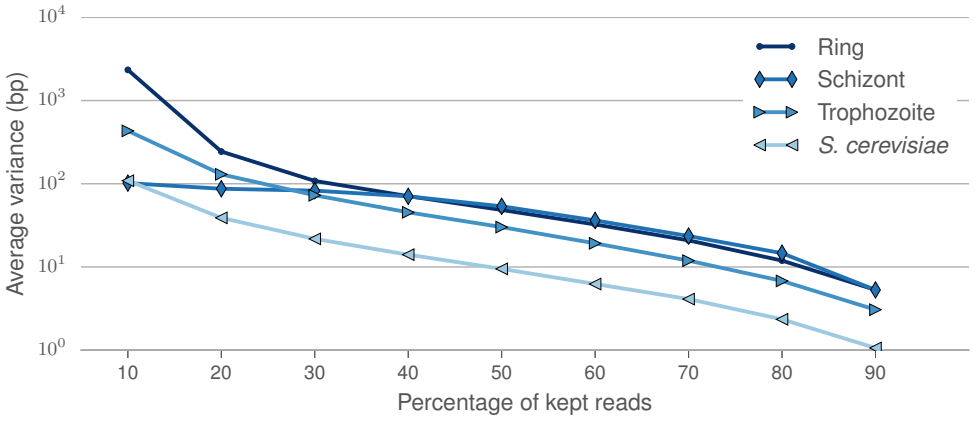


FIGURE 3: **Impact of Hi-C library sequencing depth on the stability of the centromere calls** Average variance of the results of Centurion on 500 generated datasets obtained by downsampling the raw contact counts to the desired coverage.

et al. to pre-localize their centromere calls, thus confounding their prediction (Appendix Fig. 5).

We also asked whether the improvement of Centurion over Marie-Nelly *et al.* method is due to the initialization step, or due to different objective functions used by each method. We initialized both optimization problems with the ground truth and computed the resulting error. Our results (Figure 2D) showed that Centurion's error is still between 4- and 10- fold lower, thus demonstrating the benefit of jointly calling centromeres.

§ 3.2 Resolution, sequencing depth and prediction accuracy

To assess the stability of our predictions, we simulated 500 bootstrapped data sets of *S. cerevisiae* and of each stage of *P. falciparum* with an expected total number of reads equal to the contact counts matrices. These bootstrapped samples were obtained by drawing a contact count for each pair of loci i and j from a Poisson distribution of intensity c_{ij}. We then ran the optimization process on the bootstrapped data sets, starting with initial values randomly placed within 40 kb of the centromere calls from our optimization in Appendix Tables 1, 2, 3 and 4. Our results show that the optimization is very stable (average variance of 25 bp for ring, 6 bp for schizont and 12 bp for trophozoite), suggesting that the stochastic sampling of the sequencing procedure does not significantly affect centromere predictions.

We then sought to investigate the extent to which the matrix resolution and sequencing depth affect the accuracy of Centurion's predictions. As already seen in Appendix

Figures 2 and 3, different species give different results: for *S. cerevisiae*, increasing the matrix resolution to 10 kb results in lowered accuracy of centromere calls, while in *P. falciparum* the call quality improves slightly. We speculated that our ability to call centromeres in a given species at a given resolution may depend on the choice of restriction enzyme, the sequencing depth, and the resolution of the contact map.

We next evaluated the effect of depth of sequence coverage on the quality of our centromere predictions. We generated 500 low-coverage datasets by randomly downsampling the raw contact counts. We then ran the optimization process on these downsampled datasets, initializing with perturbed calls as before. We observe that the low coverage centromere calls remain highly stable and accurate. As illustrated in Figure 3, results across all data sets only begin to degrade when downsampling to less than 10% of the total number of reads, which corresponds to less than one count per bin on average. Centurion is thus applicable to call centromeres at low cost or for low-abundance species in metagenomic samples.

§ 3.3 Centromere calls on a metagenomic dataset

We next sought to call centromeres in several species simultaneously by combining Centurion with metagenomic Hi-C libraries. We previously [Burton et al., 2014] generated two Hi-C datasets from synthetic mixtures: one containing 16 yeast strains (including four strains of *S. cerevisiae*), and one containing a mixture of 8 yeasts and 10 prokaryotic species. The two samples contain a total of 19 yeast species, some of which are much better characterized than others: centromere positions are already known for eight species (*K. lactis*, *L. kluyveri*, *L. thermotolerans*, *S. cerevisiae*, *S. kudriavzevii*, *S. mikatae*, *S. pombe*, *S. rouxii*) and partially for one more (*S. bayanus*) [Scannell et al., 2011, The Génolevures Consortium et al., 2009, McDowall et al., 2014, Dujon et al., 2004].

We aligned the reads from the metagenomic Hi-C datasets to these yeast species' reference genomes (see Appendix **??**). The quality of the individual species datasets differ greatly because the organisms vary in abundance in the metagenomic samples, and because many sequences are shared nearly identically between organisms, making the number of uniquely mappable reads for each organism range between 109 k for one of the *S. cerevisiae* strains to 26 M for the bacteria *V. fischeri*. Consequently, the sparsity of the matrices is variable (Appendix Tables 6 and 7). Furthermore, some contact counts matrices include at least one interaction count for more than 99% of all possible locus pairs, whereas other matrices are below 5%. Similarly, in the 40 kb matrices, the average number of interchromosomal contact counts per bin varies from less than 0.004 to more than 200. In particular, the matrices for the four *S. cerevisiae* strains are very

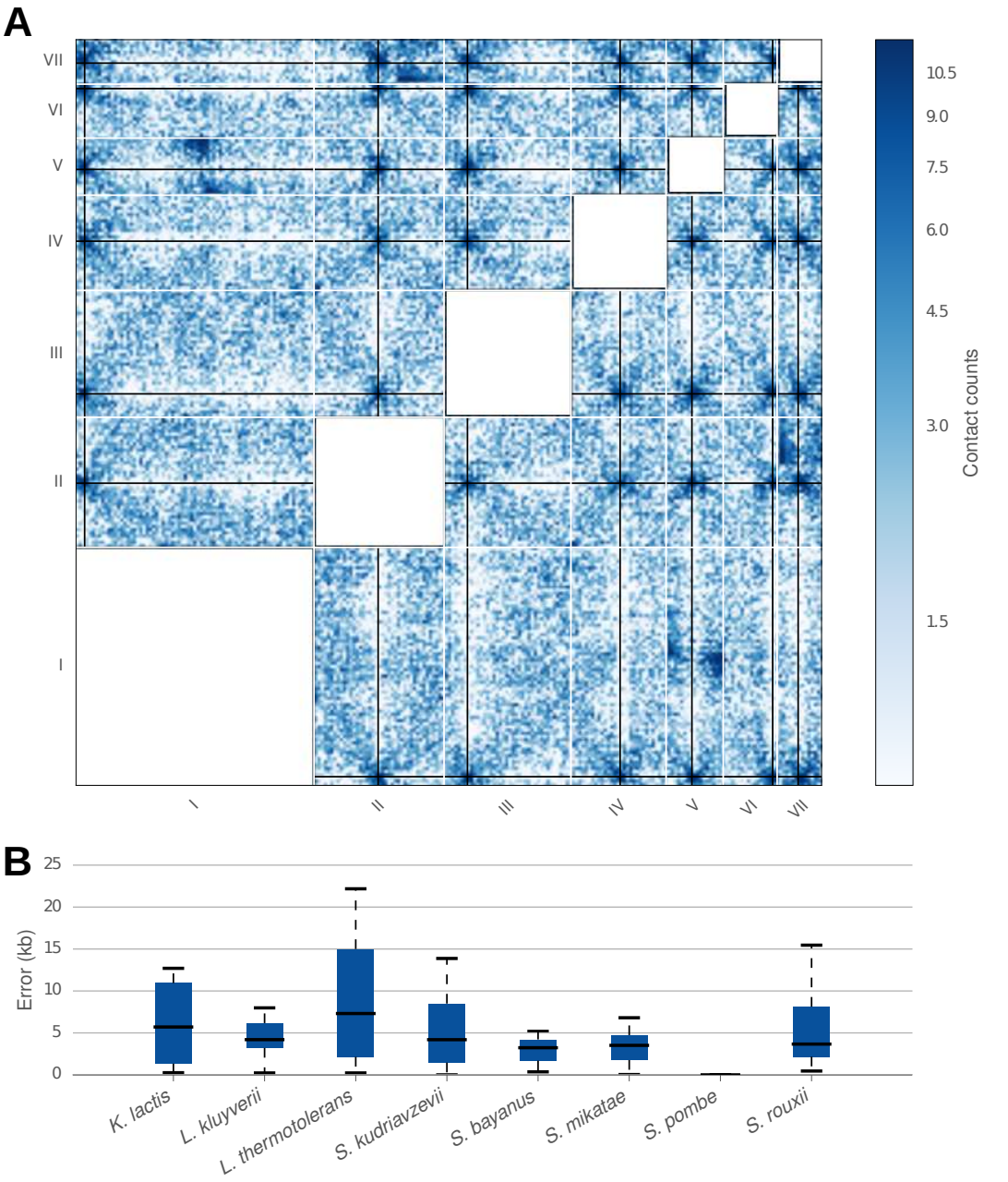


FIGURE 4: **Centromere calling on a metagenomic sample A.** Heatmap of the *trans* contact counts for *K. wickerhamii* overlaid with *de novo* centromere calls (black lines). The contact counts were smoothed with a Gaussian filter ($\sigma = 40$ *kb*) for visualization purposes. White lines indicate chromosome boundaries. **B.** Box plots indicating the error (in kb) for each chromosome in Centurion's centromere calls for eight yeasts with known centromere coordinates from the combined metagenomic Hi-C samples M-3D and M-Y of [Burton et al., 2014] on the 20 kb contact count matrices.

sparse: the reference genomes of the four strains are very similar to one another; thus, we are not able to map reads uniquely. We therefore discarded those strains from our analysis, as well as organisms with incomplete reference genomes. We applied Centurion to the remaining 14 yeasts (*E. gossypii*, *K. lactis*, *K. wickerhamii*, *L. kluyveri*, *L. waltii*, *S. bayanus*, *S. kudriavzevii*, *S. mikatae*, *S. paradoxus*, *S. stipitis*, *P. pastoris*, *L. thermotolerans*, *S. pombe*, *S. rouxii*) on both 20 kb and 40 kb contact maps.

Across these 14 species Centurion performs well, both on high-coverage datasets (*K. lactis*, *L. kluyveri*, *S. bayanus*) and low-coverage datasets (*S. mikatae*), at 20 kb and 40 kb, finding centromeres at an average deviation from the ground truth of 10 kbp (Figure 4B and Appendix Figure 6). Given this success with yeasts with known centromere positions, we next made *de novo* centromere calls for the other 6 yeast species present in the metagenomic samples. These regions, visualized in Appendix Figures 7, 8, 9, 10, 11, 12, 13, 14, are strong candidates for experimental validation by other approaches. One feature that is shared by centromeres across all studied fungi is that they reside in regions of early replication timing [Koren et al., 2010, Pohl et al., 2012]. Thus if our centromere calls lie in regions of advanced replication timing in a species for which replication timing has been profiled but centromeres have not yet been identified, this data could be used to assess the validity of our predictions. Accordingly, we overlaid the positions of our centromere calls in *P. pastoris*, where replication has been recently profiled [Liachko et al., 2014]. In all four chromosomes, *P. pastoris* centromere predictions lay in regions of early replication timing (Appendix Fig. 21), lending support to our predictions.

§ 3.4 The effect of the choice of restriction enzyme

In addition to the resolution of our contact matrices, the underlying resolution of the Hi-C data itself may limit the accuracy of our predictions. Hi-C reads can only occur near the recognition site of the restriction enzyme used in the Hi-C assay; indeed, the best resolution we can hope to achieve is a matrix in which each corresponds to one restriction enzyme fragment. Some restriction enzymes cut much more frequently than others. Thus, we speculated that a Hi-C experiment using enzymes that cut more frequently might yield more accurate results than an experiment using less frequently cutting enzymes.

To address this question, we compare the accuracy of centromere calling from two Hi-C libraries created from a single metagenomic sample using different combinations of restriction enzymes. The first library was created using the two 6 bp-cutters, HindIII and NcoI. The second library uses Sau3AI, which has a 4 bp recognition site, and AflIII, which has a 6 bp recognition site with two degenerate sites, making it effectively a 5 bp

cutter. Digestion with HindIII/NcoI yields a total of 8324 restriction fragments, whereas digestion with Sau3AI/AflIII yields 42359 restriction fragments. We corrected for the difference in Hi-C sequencing depth between Sau3AI/AflIII and the NcoI/HindIII libraries by generating downsampled datasets with an equal number of reads from each sequencing library. We then normalized the datasets and applied Centurion. The sample includes three species for which we possess the ground truth centromere locations, only one of which (*L. thermotolerans*) had enough reads in both the NcolI/HindIII (63000 reads) and the pooled Sau3AI/AflIII (55000 reads) datasets to correctly call the centromeres. The error on the downsampled Sau3AI/AflIII datasets (8 kbp) was on average half as large as the error on the the NoclI/HindIII datasets (16 kbp). Thus, we conclude that using a restriction enzyme with more frequent cutting sites enables more precise centromere calls at fine scales.

§ 4 Discussion

While centromeres are a fundamental element in the biology of genomes, their identification in diverse species has proven difficult due to sequence divergence and limitations of available tools. In this work, we have developed a novel method, Centurion, that uses centromere colocalization and the pattern it creates in Hi-C contact maps to jointly call centromeres for all chromosomes of an organism. We first established the feasibility of this approach by demonstrating that Centurion accurately calls regional centromeres on the parasite *P. falciparum* and the yeast *S. pombe* as well as point centromeres on several other yeasts with known centromere coordinates. For the species with high depth Hi-C sequencing, Centurion often identified centromeres within 1 kb of the actual coordinates (41 times out of 58 for three stages of *P. falciparum* and *S. cerevisiae* data). We then used Centurion to infer centromeres of multiple yeast species (8 with known, 6 with unknown centromere coordinates) from two metagenomic Hi-C samples. Our results showed that Centurion still accurately identifies centromere coordinates from samples with only limited sequencing depth. Thus, Centurion can be used to accurately and efficiently identify centromere locations in yeast species.

The task of centromere identification from Hi-C data has been attempted recently by others [Marie-Nelly et al., 2014b]. Centurion offers a few key differences compared to the previous approach. The first difference is in the pre-localization of candidate centromeres. Marie-Nelly *et al.*'s method uses only the *cis* Pearson correlation information independently per chromosome to identify the initial candidates. However, the pattern created by centromeres in the Pearson correlation matrix can be very similar to the

patterns generated by other genomic elements such as rDNA coding regions or by specific gene clusters (e.g., virulence genes in *P. falciparum*). Because Marie-Nelly *et al.*'s method restricts the further search for the best centromere coordinate to only the candidates from the pre-localization step, an inaccurate candidate (e.g., an rDNA region instead of a centromere) will prevent the method from finding the correct centromere location. Centurion, on the other hand, utilizes *trans* contact information jointly across all chromosomes for its pre-localization step. Furthermore, Centurion allows multiple candidates per chromosome during the second step of the optimization, thereby leaving room for correcting potential errors in the pre-localization step. The second difference between the two methods is in how they use the submatrices that correspond to *trans* contact maps flanking the pairs of candidate centromeres from the pre-localization step. For an organism with N chromosomes, Marie-Nelly *et al.*'s method carves out the N-1 *trans* submatrices for each chromosome, sums these N-1 matrices and then collapses the sum into a 1D vector of row/column sums. Then, independently for each chromosome, the method fits a Gaussian to this 1D vector, and the resulting peak corresponds to the predicted centromere location. In this procedure, both the summation of N-1 matrices and the collapsing of the resulting matrix into a 1D vector of sums result in loss of important information embedded in 2D maps. Furthermore, performing the Gaussian fit separately for each chromosome does not fully take into account the joint co-localization of the other N-1 centromeres. To address these issues, Centurion infers a 2D Gaussian fit that best explains the observed *trans* contact counts, jointly optimizing these 2D fits for all pairs of centromeres. Both of these improvements in the pre-localization and the optimization steps allow Centurion to perform better specifically for the cases with limited sequencing depth. Our approach could be improved in several respects. First, better modeling of zero contact counts may improve inference for organisms with many repeated sequences in the peri-centromeric regions, or data sets with low sequencing depth. Second, one could model contact counts as a Gaussian distribution centered on the pairs of centromere locations. Maximising the log likelihood of such a model might yield improved performance. Last, as described here, our method requires reference genomes for the metagenomic samples. It would be possible to first build reference genomes directly from the Hi-C data, using methods like Lachesis [Burton et al., 2013] or Graal [Marie-Nelly et al., 2014a], and then infer centromeres locations using the inferred references. However, the inherent structure of Hi-C contact counts for organisms with colocalizing centromeres will likely present a challenge for these methods because pericentromeric sequences on different chromosomes are likely to appear to be adjacent to one another.

Finally, our new centromere predictions have practical applications. Autonomously replicating plasmids and artificial chromosomes are useful tools for research and strain engineering [Böer et al., 2007]. Identification of centromeres in new species will facilitate building such constructs over an expanded species range. *P. pastoris*, for example, is a common industrial chassis [Cregg et al., 2009], but existing plasmid tools in the species have elevated loss rates [Liachko and Dunham, 2014] that could be stabilized by addition of a centromere. Many of our centromere calls were accurate to < 1 kb, making experimental validation possible.

§ 5 Funding

This work was supported by the European Research Council [SMAC-ERC-280032 to J-P.V., N.V.]; the European Commission [HEALTH-F5-2012-305626 to J-P.V., N.V.]; the French National Research Agency [ANR-11-BINF-0001 to J-P.V., N.V.]; the National Institute of Health/National Human Genome Research Institute [HG006283 to J.S., T32HG000035 to J.N.B.]; National Institute of Health/National Institute of General Medical Sciences [P41 GM103533 to I.L., M.J.D., W.S.N.; R01AI106775 to F. A., W.S.N.]; National Science Foundation [1243710 to I.L., M.J.D.]. M.J.D. is a Rita Allen Foundation Scholar and a Senior Fellow in the Genetic Networks program at the Canadian Institute for Advanced Research.

§ 6 Acknowledgements

We thank Celia Payen for providing the yeast centromere annotations, Stéfan van der Walt for advice on peak detection algorithms, Fabrice Varoquaux for help on understanding the specificity of *A. thaliana* genome and Chloé Azencott for helpful comments on the manuscript.

Chapter 5

Discussion

In this thesis, I have presented contributions to the analysis of Hi-C data, in particular *3D structure inference* methods. To summarize:

Biological contributions - I studied, in collaboration with the Le Roch lab and the Noble lab, the three-dimensional structure of the human malaria parasite *P. falciparum*, which led to a better understanding of links between the genome architecture and gene expression and regulation.

Methodological contributions - I focused on several methodological projects, first in the domain of *3D structure inference*, with a statistical method to infer a consensus model of the genome architecture, second in the use of Hi-C for *genome annotation*, with an approach to detect centromeric regions for organisms whose genome fold in a Rabl configuration (with centromeres colocalizing).

Software contributions - In addition to the methodological contributions and the biological contributions, I have also focused on the implementations of several methods studied or developped during the course of this thesis. I believe that high quality implementations are critical for the analysis of the huge quantity of data available in biology, while challenging. I have not only contributed to *scikit-learn*, a machine learning toolkit written in Python (with the inclusion of the isotonic regression, the metric and non-metric MDS, ...), but also released three packages, specific to analysis of Hi-C data: *iced*, *pastis* and *centurion*, which are all free and open-source softwares.

Research perspectives

The field of genome 3D structure is a young yet fast moving field. When I first started to work on 3D structure inference methods, only a handfull of papers using Hi-C were published. Nowadays, more than a paper per week on this subject is published. Yet, as

the field is still young, well grounded methods and publicly available softwares are still lacking. I here describe some research perspectives.

Quality control and normalization of Hi-C data - Between the first publication on Hi-C [Lieberman-Aiden et al., 2009] and recent work such as Rao et al. [2014], Jin et al. [2013], the resolution of the Hi-C contact maps have increased from 1 Mb to 5 kb or even 1 kb. Not only has the number of reads greatly increased, but also the protocol improved, assessing contact counts in a more robust fashion. Yet, there is still no satisfying quality control protocol or quality measures to identify what is a "good" dataset. In particular, the choice of the resolution of the contact maps is still not justified by any well-grounded method and is left to the judgement of the researcher. Developing quality control measurement seems a natural first step to help scientists set up and compare reliable protocols to assess the 3D structure of the genome.

Inference of the 3D structure of polyploid structures - So far, most methods either only dealt with haploid 3D structures or ignored the diploidy or polyploidy of genomes when building 3D models. To our knowledge, only two methods incorporated the polyploidy of genomes: Kalhor et al. [2011] and Ay et al. [2015b]. Neither have been validated on simulated data, and one might wonder, considering the complexity of solving such deconvolution problems, how accurate and reliable these are. Now that single-allele Hi-C datasets have become available [Deng et al., 2015], this challenge can be more thoroughly investigated.

Inference of a population of structures using single-cell data - Nagano et al. [2013] published a protocol to assess physical interactions in single cells, laying the foundation for studying the variability of structures amongst a population of cell. So far, only two methods exploit this type of data for 3D structure reconstruction: Nagano et al. [2013] proposes a constraint-based approach to infer structures, while Paulsen et al. [2015] proposes to infer low-rank psd matrices as close as possible to sparse contact count maps, and apply manifold learning techniques to find a euclidean embedding of the data. Neither methods attempt to leverage several single cell datasets or population contact maps to alleviate the sparsity of single-cell contact maps. Performing a joint optimization may improve the accuracy and robustness of inferring structures from single-cell data.

***De novo* sequencing using Hi-C data** - Several methods have been proposed to re-target Hi-C for *de novo* scaffolding [Burton et al., 2013, Kaplan and Dekker, 2013, Marie-Nelly et al., 2014a], but none leverage the recent work on convex relaxation for permutation problems. Yet Fogel et al. [2013] proposes to solve exactly the challenge faced in *de novo* sequencing using Hi-C data: finding a permutation matrix to reorder

rows and columns such that strongly interacting elements are close one another. Instead of relying on heuristics, one may attempt to use these recent convex relaxation approaches.

This list of research perspectives is of course a very incomplete list of possible extensions to this thesis, and I believe that in a short period of time, many more challenges will arise in the field of Hi-C analysis.

Appendix A

Supplementaries - Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression

Files and datasets that are too large to include in this supplement are made available through http://noble.gs.washington.edu/proj/plasmo3d.

SuppFile1-mapping-and-filtering.xlsx This file summarizes the results of applying our mapping and filtering pipeline to the sequences from each Hi-C library generated in this work.

SuppFile2-contacts-at-0.1-FDR.xlsx This file lists, for each stage (ring, trophozoite, schizont), the set of contacts at 10 kb that were assigned a q-value < 0.1 (Methods). Rows are sorted from lowest to highest q-value and are colored using two other q-value thresholds (0.05 and 0.01).

SuppFile3-Var,Rif,Stevor,MC(VRSM)-clusters.xlsx This file contains the chromosomal coordinates of all *var*, *rifin*, *stevor*, and *Pfmc-2tm* (VRSM) genes, as well as the boundaries of subtelomeric and internal VRSM gene clusters.

SuppFile4-dynamic-model-all-chromosomes.mov This movie shows the dynamic changes in the architecture of all chromosomes during the *Plasmodium* erythrocytic cycle inferred by a linear interpolation of bead positions from one stage to the next by aligning the structures of adjacent stages. The movie starts and ends at the ring stage (ring–trophozoite–schizont–ring). Each chromosome is represented by a different color, and purple regions mark VRSM gene clusters. Telomeres are indicated by white spheres.

SuppFile{5–18}-dynamic-model-chr{1–14}.mov These movies are the same as the previous movie, but each focuses on a single chromosome.

Supplementary Tables

TABLE 1: **Quality measures for Hi-C data.**

P. falciparum libraries are presented in this work and *S. cerevisiae* libraries from Duan et al. Duan et al. [2010] are listed here for comparison. Rows marked with bold are control libraries that were generated without the cross-linking step of the Hi-C protocol. Interchromosomal contact probability (*ICP* Kalhor et al. [2011]) and percent of long-range contacts (*PLRC*) values are computed as described in Methods.

Organism	*Library*	*ICP*	*PLRC*
P. falciparum	Ring	1.13	9.04%
	Trophozoite	0.66	7.64%
	Schizont	0.74	22.04%
	Trophozoite (not cross-linked)	**7.82**	**3.05%**
S. cerevisiae Duan et al. [2010]	HindIII-MspI	1.92	8.99%
	HindIII-MseI	2.31	12.08%
	EcoRI-MspI	1.71	3.99%
	EcoRI-MseI	1.86	4.19%
	HindIII-MspI (not cross-linked)	**4.26**	**3.39%**

TABLE 2: **GSEA results for genes involved in stage-specific contacts.**

For each stage, GSEA is applied to the set of genes that participate in contacts that are specific to that stage (Methods). For the *Type* column CC denotes "Cellular Component", MF denotes "Molecular Function" and BP denotes "Biological Process". Enrichments with q-value < 0.1 are shown.

Stage	*GO term*	*Description*	*Type*	*q-value*
Ring	GO:0020033	antigenic variation	BP	0.099
Trophozoite	GO:0020002	host cell plasma membrane	CC	0.004
	GO:0020030	infected host cell surface knob	CC	0.008
	GO:0016021	integral to membrane	CC	0.015
	GO:0004872	receptor activity	MF	0.007
	GO:0050839	cell adhesion molecule binding	MF	0.020
	GO:0020033	antigenic variation	BP	0.010
	GO:0009405	pathogenesis	BP	0.010
	GO:0020013	modulation by symbiont of host erythrocyte aggregation	BP	0.012
	GO:0020035	cytoadherence to microvasculature	BP	0.016
	GO:0016337	cell-cell adhesion	BP	0.022

TABLE 3: **Assessing sensitivity of the 3D inference to different parameter settings.**

RMSD and distance difference values in nanometers (nm) between structures inferred from an unconstrained MDS with five different β values ranging from 0.4 to 0.6.

Stage	*RMSD* Mean (Standard deviation)	*Distance difference* Mean (Standard deviation)
Ring	10.39 (4.24)	5.75 (2.68)
Trophozoite	17.76 (6.57)	10.62 (4.65)
Schizont	12.90 (5.71)	8.10 (4.08)

TABLE 4: **Assessing sensitivity of the 3D inference to spatial constraints.**

RMSD and distance difference values in nanometers (nm) between a structure inferred using constrained MDS and a structure from the corresponding unconstrained MDS.

Stage	*RMSD*	*Distance difference*
Ring	8.05	0.01
Trophozoite	61.99	0.83
Schizont	7.86	0.01

TABLE 5: **Colocalization test for 21 gene/locus sets.**

We applied a previously described statistical test Witten and Noble [2012] to assess whether the loci in each set colocalize more than expected by chance (only interchromosomal pairs are considered). This test involves calculation of a colocalization statistic, which requires labeling of each locus pair as "close" or "far". We used varying distance thresholds (10%, 20% and 40% of the nuclear diameter) to deem a locus pair "close" and labeled all remaining pairs in the set as "far". We generated 3000 random locus sets to compute a p-value for each test. We corrected the p-values for multiple hypothesis testing using the Benjamini-Hochberg procedure Benjamini and Hochberg [1995] to compute the associated q-value. Grey color indicates a q-value < 0.05. Centromere coordinates were extracted from Hoeijmakers et al. Hoeijmakers et al. [2012a]. Telomeres were defined as 20 kb regions at each end of each chromosome. The sets of internal and subtelomeric VRSM genes were tested all together as well as separately. The rDNA set consists of five units of 18S-5.8S-28S rDNA genes and one tandem of three 5S rDNA genes Mancio-Silva et al. [2010]. Clusters 1–15 correspond to expression clusters described in Le Roch et al. Le Roch et al. [2003].

Gene set	*Ring*			*Trophozoite*			*Schizont*		
	10%	**20%**	**40%**	**10%**	**20%**	**40%**	**10%**	**20%**	**40%**
Centromeres	0.003	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Telomeres	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VRSM (all)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
VRSM (internal)	0.388	0.215	0.070	0.152	0.023	0.079	0.246	0.077	0.025
VRSM (sub-telomeric)	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
rDNA genes	1.000	0.277	0.463	0.124	0.037	0.060	1.000	1.000	0.442
Cluster 1	0.103	0.011	0.115	0.018	0.133	0.133	0.135	0.007	0.013
Cluster 2	0.449	0.069	0.014	0.219	0.054	0.015	0.947	0.809	0.117
Cluster 3	0.215	0.014	0.075	0.437	0.002	0.001	0.758	0.809	0.033
Cluster 4	0.106	0.000	0.000	0.002	0.000	0.000	0.701	0.089	0.000
Cluster 5	0.449	0.701	0.291	0.779	0.045	0.002	0.508	0.528	0.682
Cluster 6	0.849	0.228	0.000	0.029	0.000	0.000	0.396	0.441	0.026
Cluster 7	0.291	0.000	0.000	0.596	0.001	0.074	0.704	0.523	0.011
Cluster 8	0.047	0.054	0.016	0.015	0.006	0.033	0.208	0.117	0.592
Cluster 9	0.508	0.063	0.014	0.040	0.001	0.002	0.468	0.809	0.117
Cluster 10	0.048	0.043	0.007	0.000	0.000	0.000	0.355	0.208	0.007
Cluster 11	0.198	0.019	0.063	0.411	0.000	0.000	0.601	0.446	0.013
Cluster 12	0.028	0.010	0.000	0.006	0.000	0.000	0.523	0.007	0.000
Cluster 13	0.091	0.021	0.075	0.033	0.003	0.000	0.711	0.028	0.001
Cluster 14	0.155	0.082	0.046	0.688	0.000	0.000	0.751	0.039	0.000
Cluster 15	0.046	0.014	0.103	0.016	0.058	0.002	0.758	0.809	0.011

TABLE 6: **Sequences of primers used for the generation of FISH probes.**

Chr.	*Annotation*	*Locus (kb)*	*Forward primer*	*Reverse primer*
7	VRSM	550 - 560	5′-GATGGTAGAAGATAATAGGG -3′	5′-GACAAGTATAAGAACCAACC-3′
8	VRSM	40 - 50	5′-CGAAAGATAGTAGTGATGGT-3′	5′-CACTTATGCATTTCCATCCA-3′
7	Non-VRSM	810 - 820	5′-GCTTCCTTAATTGGACATTC-3′	5′-GAATTCGTTGGAGATTCTGT-3′
11	Non-VRSM	820 - 830	5′-CACTGAACAAGTAGTGTAATCA-3′	5′-GTTTCATCTTCAGAAGTAAGAG-3′
2	Non-VRSM	440 - 450	5′-GTTCCTACAGGTTTAGATCT-3′	5′-CATGAGGACATATTCACTTG-3′
4	Non-VRSM	1,160 - 1,170	5′-AAGTACAGGTGTAGGTAAAG-3′	5′-CGTAGCTTTAACCTGTTGTA-3′

TABLE 7: **Gradient values of the log-linear fits that best capture the scaling of contact probability with genomic distance for each chromosome.**

Gradient (α) values for each chromosome at each stage calculated by fitting a power-law curve of the form $P(s) \sim s^{\alpha}$ to the intrachromosomal contact probability $P(s)$ as a function of genomic distance s. The reported α values are computed using raw contact maps at a single restriction enzyme fragment resolution for a genomic distance range of 20–250 kb.

Chromosome	*Ring*	*Trophozoite*	*Schizont*
1	-1.02	-1.18	-1.04
2	-0.99	-1.22	-1.01
3	-0.99	-1.20	-0.98
4	-0.97	-1.13	-0.99
5	-0.97	-1.14	-0.96
6	-1.01	-1.19	-1.00
7	-1.02	-1.27	-1.01
8	-1.00	-1.19	-0.98
9	-0.99	-1.11	-0.94
10	-0.97	-1.14	-0.96
11	-0.97	-1.11	-0.94
12	-0.99	-1.14	-0.97
13	-0.97	-1.07	-0.93
14	-0.98	-1.09	-0.93

TABLE 8: **GSEA results for the ring stage on the first component of the kCCA.**

GSEA is applied to the ranked list of genes per projection on the kCCA component. For the *Enrichment* column, *t* denotes enrichment near the telomeres, and *n-t* denotes enrichment in non-telomeric regions. For the *Type* column CC denotes "Cellular Component", MF denotes "Molecular Function" and BP denotes "Biological Process". Enrichments with q-value < 0.1 are shown.

GO term	*Description*	*Type*	*Enrichment*	*q-value*
GO:0020002	host cell plasma membrane	CC	t	0.000
GO:0020030	infected host cell surface knob	CC	t	0.000
GO:0020036	Maurer's cleft	CC	t	0.000
GO:0005840	ribosome	CC	n-t	0.000
GO:0005622	intracellular	CC	n-t	0.000
GO:0022627	cytosolic small ribosomal subunit	CC	n-t	0.000
GO:0022625	cytosolic large ribosomal subunit	CC	n-t	0.000
GO:0015935	small ribosomal subunit	CC	n-t	0.000
GO:0005783	endoplasmic reticulum	CC	n-t	0.000
GO:0015934	large ribosomal subunit	CC	n-t	0.000
GO:0005789	endoplasmic reticulum membrane	CC	n-t	0.000
GO:0005839	proteasome core complex	CC	n-t	0.006
GO:0005794	Golgi apparatus	CC	n-t	0.008
GO:0005829	cytosol	CC	n-t	0.008
GO:0005739	mitochondrion	CC	n-t	0.018
GO:0016021	integral to membrane	CC	t	0.034
GO:0005694	chromosome	CC	n-t	0.064
GO:0004872	receptor activity	MF	t	0.000
GO:0050839	cell adhesion molecule binding	MF	t	0.000
GO:0046789	host cell surface receptor binding	MF	t	0.000
GO:0003735	structural constituent of ribosome	MF	n-t	0.000
GO:0004175	endopeptidase activity	MF	n-t	0.010
GO:0003677	DNA binding	MF	n-t	0.011
GO:0005215	transporter activity	MF	n-t	0.034
GO:0004298	threonine-type endopeptidase activity	MF	n-t	0.035
GO:0003676	nucleic acid binding	MF	n-t	0.094
GO:0016881	acid-amino acid ligase activity	MF	n-t	0.094
GO:0020033	antigenic variation	BP	t	0.000
GO:0009405	pathogenesis	BP	t	0.000
GO:0020035	cytoadherence to microvasculature	BP	t	0.000
GO:0020013	modulation by symbiont of host erythrocyte aggregation	BP	t	0.000
GO:0016337	cell-cell adhesion	BP	t	0.000
GO:0006412	translation	BP	n-t	0.000
GO:0006886	intracellular protein transport	BP	n-t	0.000
GO:0006511	ubiquitin-dependent protein catabolic process	BP	n-t	0.001
GO:0045454	cell redox homeostasis	BP	n-t	0.019
GO:0007264	small GTPase mediated signal transduction	BP	n-t	0.025
GO:0006281	DNA repair	BP	n-t	0.025
GO:0006260	DNA replication	BP	n-t	0.025
GO:0016192	vesicle-mediated transport	BP	n-t	0.027
GO:0006414	translational elongation	BP	n-t	0.029
GO:0015031	protein transport	BP	n-t	0.029

TABLE 9: **GSEA results for the trophozoite stage on the first component of the kCCA.**

GSEA is applied to the ranked list of genes per projection on the kCCA component. For the *Enrichment* column, *t* denotes enrichment near the telomeres, and *n-t* denotes enrichment in non-telomeric regions. For the *Type* column CC denotes "Cellular Component", MF denotes "Molecular Function" and BP denotes "Biological Process". Enrichments with q-value < 0.1 are shown.

GO term	*Description*	*Type*	*Enrichment*	*q-value*
GO:0005840	ribosome	CC	n-t	0.000
GO:0005622	intracellular	CC	n-t	0.000
GO:0022627	cytosolic small ribosomal subunit	CC	n-t	0.000
GO:0005789	endoplasmic reticulum membrane	CC	n-t	0.000
GO:0005783	endoplasmic reticulum	CC	n-t	0.000
GO:0022625	cytosolic large ribosomal subunit	CC	n-t	0.000
GO:0015935	small ribosomal subunit	CC	n-t	0.000
GO:0020002	host cell plasma membrane	CC	t	0.000
GO:0020030	infected host cell surface knob	CC	t	0.000
GO:0020036	Maurer's cleft	CC	t	0.000
GO:0016021	integral to membrane	CC	t	0.000
GO:0015934	large ribosomal subunit	CC	n-t	0.001
GO:0005794	Golgi apparatus	CC	n-t	0.005
GO:0005739	mitochondrion	CC	n-t	0.028
GO:0005839	proteasome core complex	CC	n-t	0.084
GO:0005829	cytosol	CC	n-t	0.087
GO:0003735	structural constituent of ribosome	MF	n-t	0.000
GO:0004872	receptor activity	MF	t	0.000
GO:0050839	cell adhesion molecule binding	MF	t	0.000
GO:0046789	host cell surface receptor binding	MF	t	0.000
GO:0005215	transporter activity	MF	n-t	0.003
GO:0003677	DNA binding	MF	n-t	0.059
GO:0005509	calcium ion binding	MF	n-t	0.088
GO:0020033	antigenic variation	BP	t	0.000
GO:0009405	pathogenesis	BP	t	0.000
GO:0016337	cell-cell adhesion	BP	t	0.000
GO:0020035	cytoadherence to microvasculature	BP	t	0.000
GO:0020013	modulation by symbiont of host erythrocyte aggregation	BP	t	0.000
GO:0006412	translation	BP	n-t	0.000
GO:0045454	cell redox homeostasis	BP	n-t	0.011
GO:0006886	intracellular protein transport	BP	n-t	0.012
GO:0006511	ubiquitin-dependent protein catabolic process	BP	n-t	0.099

TABLE 10: **GSEA results for the schizont stage on the first component of the kCCA.**

GSEA is applied to the ranked list of genes per projection on the kCCA component. For the *Enrichment* column, t denotes enrichment near the telomeres, and n-t denotes enrichment in non-telomeric regions. For the *Type* column CC denotes "Cellular Component", MF denotes "Molecular Function" and BP denotes "Biological Process". Enrichments with q-value < 0.1 are shown.

GO term	*Description*	*Type*	*Enrichment*	*q-value*
GO:0020030	infected host cell surface knob	CC	t	0.000
GO:0005840	ribosome	CC	n-t	0.000
GO:0005622	intracellular	CC	n-t	0.000
GO:0022627	cytosolic small ribosomal subunit	CC	n-t	0.000
GO:0022625	cytosolic large ribosomal subunit	CC	n-t	0.000
GO:0005783	endoplasmic reticulum	CC	n-t	0.000
GO:0015935	small ribosomal subunit	CC	n-t	0.000
GO:0020036	Maurer's cleft	CC	t	0.000
GO:0020002	host cell plasma membrane	CC	t	0.000
GO:0015934	large ribosomal subunit	CC	n-t	0.001
GO:0005789	endoplasmic reticulum membrane	CC	n-t	0.001
GO:0005839	proteasome core complex	CC	n-t	0.002
GO:0016021	integral to membrane	CC	t	0.003
GO:0005794	Golgi apparatus	CC	n-t	0.008
GO:0005739	mitochondrion	CC	n-t	0.026
GO:0004872	receptor activity	MF	t	0.000
GO:0050839	cell adhesion molecule binding	MF	t	0.000
GO:0046789	host cell surface receptor binding	MF	t	0.000
GO:0003735	structural constituent of ribosome	MF	n-t	0.000
GO:0004175	endopeptidase activity	MF	n-t	0.008
GO:0005215	transporter activity	MF	n-t	0.011
GO:0003677	DNA binding	MF	n-t	0.013
GO:0004298	threonine-type endopeptidase activity	MF	n-t	0.015
GO:0020033	antigenic variation	BP	t	0.000
GO:0009405	pathogenesis	BP	t	0.000
GO:0020013	modulation by symbiont of host erythrocyte aggregation	BP	t	0.000
GO:0020035	cytoadherence to microvasculature	BP	t	0.000
GO:0016337	cell-cell adhesion	BP	t	0.000
GO:0006412	translation	BP	n-t	0.000
GO:0006511	ubiquitin-dependent protein catabolic process	BP	n-t	0.002
GO:0006886	intracellular protein transport	BP	n-t	0.003
GO:0045454	cell redox homeostasis	BP	n-t	0.016
GO:0007264	small GTPase mediated signal transduction	BP	n-t	0.017
GO:0006260	DNA replication	BP	n-t	0.028
GO:0015031	protein transport	BP	n-t	0.043
GO:0006281	DNA repair	BP	n-t	0.091
GO:0016192	vesicle-mediated transport	BP	n-t	0.091

TABLE 11: **kCCA enrichment of 15 expression clusters.**

GSEA is applied to the ranked list of genes per projection on the first and second kCCA components (*Comp*), relative to 15 expression clusters defined by Le Roch et al. Le Roch et al. [2003] (*Cluster*). For the enrichment column (*Enr.*), *t* refers to an enrichment in telomeric regions, *n-t* to an enrichment in non-telomeric regions, *c* to enrichment in the centromeric regions, and *n-c* to enrichment in non-centromeric regions. Enrichments with q-values < 0.05 are shaded grey.

Comp	*Cluster*	*Ring*		*Trophozoite*		*Schizont*	
		q-value	*Enr.*	*q-value*	*Enr.*	*q-value*	*Enr.*
1	1	0.000	n-t	0.000	n-t	0.000	n-t
	2	0.358	n-t	0.730	n-t	0.667	n-t
	3	0.000	t	0.000	t	0.000	t
	4	0.000	n-t	0.000	n-t	0.000	n-t
	5	0.482	t	0.216	t	0.997	t
	6	0.000	t	0.003	t	0.003	t
	7	0.036	t	0.818	n-t	1.000	n-t
	8	0.759	n-t	0.897	n-t	0.819	n-t
	9	0.000	t	0.000	t	0.000	t
	10	0.000	t	0.000	t	0.000	t
	11	0.011	t	0.000	t	0.000	t
	12	0.000	t	0.000	t	0.000	t
	13	0.000	t	0.000	t	0.000	t
	14	0.000	t	0.000	t	0.000	t
	15	0.230	n-t	0.703	n-t	0.924	n-t
2	1	0.000	n-c	0.000	n-c	0.000	n-c
	2	0.006	n-c	0.002	n-c	0.005	n-c
	3	0.000	c	0.000	c	0.000	c
	4	0.000	c	0.000	c	0.000	c
	5	0.686	c	1.000	c	0.980	n-c
	6	0.838	c	1.000	c	0.999	c
	7	0.000	c	0.000	c	0.000	c
	8	0.956	n-c	1.000	c	1.000	n-c
	9	0.004	c	0.000	c	0.000	c
	10	0.071	n-c	0.996	c	1.000	n-c
	11	0.000	n-c	0.000	n-c	0.000	n-c
	12	0.000	n-c	0.000	n-c	0.000	n-c
	13	0.000	n-c	0.000	n-c	0.002	n-c
	14	0.000	n-c	0.000	n-c	0.000	n-c
	15	0.000	n-c	0.000	n-c	0.000	n-c

TABLE 12: **GSEA results for the second component of the kCCA.**

GSEA is applied to the ranked list of genes per projection on the kCCA component. For the *Enrichment* column, *c* denotes enrichment near the centromeres, *n-c* denotes enrichment in non centromeric regions. For the *Type* column CC denotes "Cellular Component", MF denotes "Molecular Function" and BP denotes "Biological Process". Enrichments with q-value < 0.1 are shown.

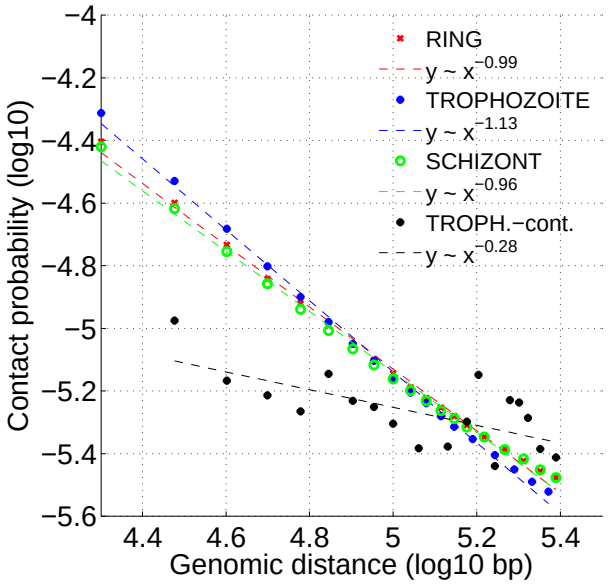
Stage	*GO term*	*Description*	*Type*	*Enrichment*	*q-value*
Ring	GO:0020008	rhoptry	CC	n-c	0.014
	GO:0016459	myosin complex	CC	n-c	0.020
	GO:0005839	proteasome core complex	CC	n-c	0.031
	GO:0008234	cysteine-type peptidase activity	MF	n-c	0.001
	GO:0004713	protein tyrosine kinase activity	MF	n-c	0.007
	GO:0003779	actin binding	MF	n-c	0.009
	GO:0004175	endopeptidase activity	MF	n-c	0.015
	GO:0004298	threonine-type endopeptidase activity	MF	n-c	0.018
	GO:0003774	motor activity	MF	n-c	0.035
	GO:0005516	calmodulin binding	MF	n-c	0.039
	GO:0016255	attachment of GPI anchor to protein	BP	n-c	0.066
Trophozoite	GO:0005839	proteasome core complex	CC	n-c	0.033
	GO:0020008	rhoptry	CC	n-c	0.058
	GO:0016459	myosin complex	CC	n-c	0.084
	GO:0008234	cysteine-type peptidase activity	MF	n-c	0.000
	GO:0004175	endopeptidase activity	MF	n-c	0.028
	GO:0004713	protein tyrosine kinase activity	MF	n-c	0.029
	GO:0004298	threonine-type endopeptidase activity	MF	n-c	0.029
	GO:0003779	actin binding	MF	n-c	0.048
	GO:0003774	motor activity	MF	n-c	0.090
	GO:0005516	calmodulin binding	MF	n-c	0.093
	GO:0016740	transferase activity	MF	n-c	0.099
	GO:0016255	attachment of GPI anchor to protein	BP	n-c	0.036
Schizont	GO:0005839	proteasome core complex	CC	n-c	0.026
	GO:0020008	rhoptry	CC	n-c	0.044
	GO:0016459	myosin complex	CC	n-c	0.083
	GO:0008234	cysteine-type peptidase activity	MF	n-c	0.000
	GO:0004175	endopeptidase activity	MF	n-c	0.013
	GO:0004298	threonine-type endopeptidase activity	MF	n-c	0.018
	GO:0004713	protein tyrosine kinase activity	MF	n-c	0.019
	GO:0003779	actin binding	MF	n-c	0.053

TABLE 13: **Density score for varying values of β parameter at different stages.**

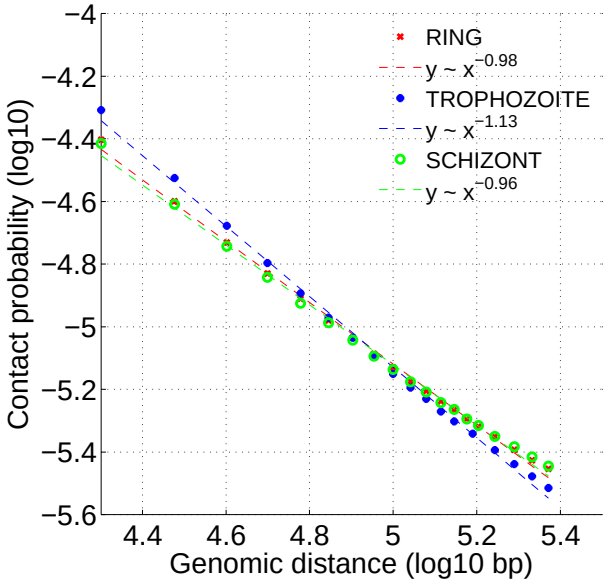
For each stage the β value that yields the minimal density score (shown in boldface) is used for three-dimensional modeling.

Stage	$\beta = 0.4$	$\beta = 0.45$	$\beta = 0.5$	$\beta = 0.55$	$\beta = 0.6$
Ring	0.109	0.077	**0.057**	0.063	0.110
Trophozoite	0.127	0.087	0.048	**0.044**	0.540
Schizont	0.051	**0.048**	0.128	0.313	0.591

Supplementary Figures

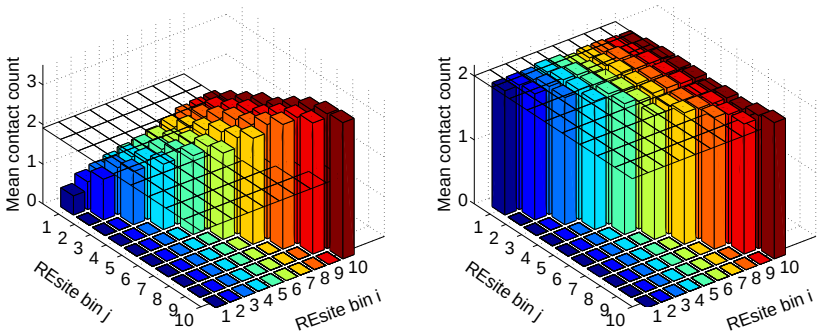


(A)



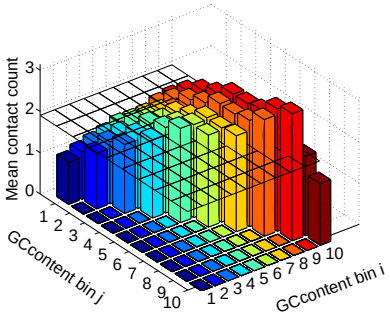
(B)

FIGURE 1: **Power-law fits to 10 kb aggregated data.**
A power law of the form $P(s) \sim s^{\alpha}$ is fit to the intrachromosomal contact probability $P(s)$ as a function of genomic distance s for each stage (Methods). These log-linear fits are visualized by dashed lines and the corresponding gradient (α) values are reported in the legend for (a) raw and (b) normalized Hi-C contact maps at 10 kb resolution.

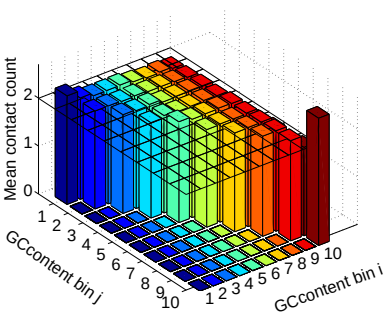


(A) Before normalization - No. of RE cut sites

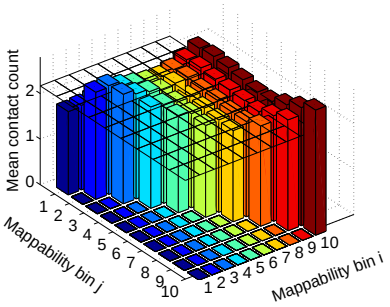
(B) After normalization - No. of RE cut sites



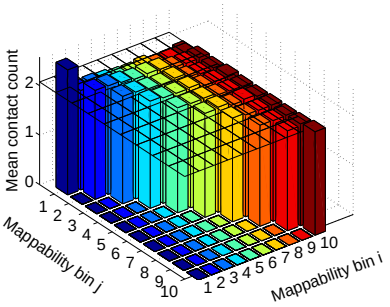
(C) Before normalization - Mappability



(D) After normalization - Mappability



(E) Before normalization - GC content

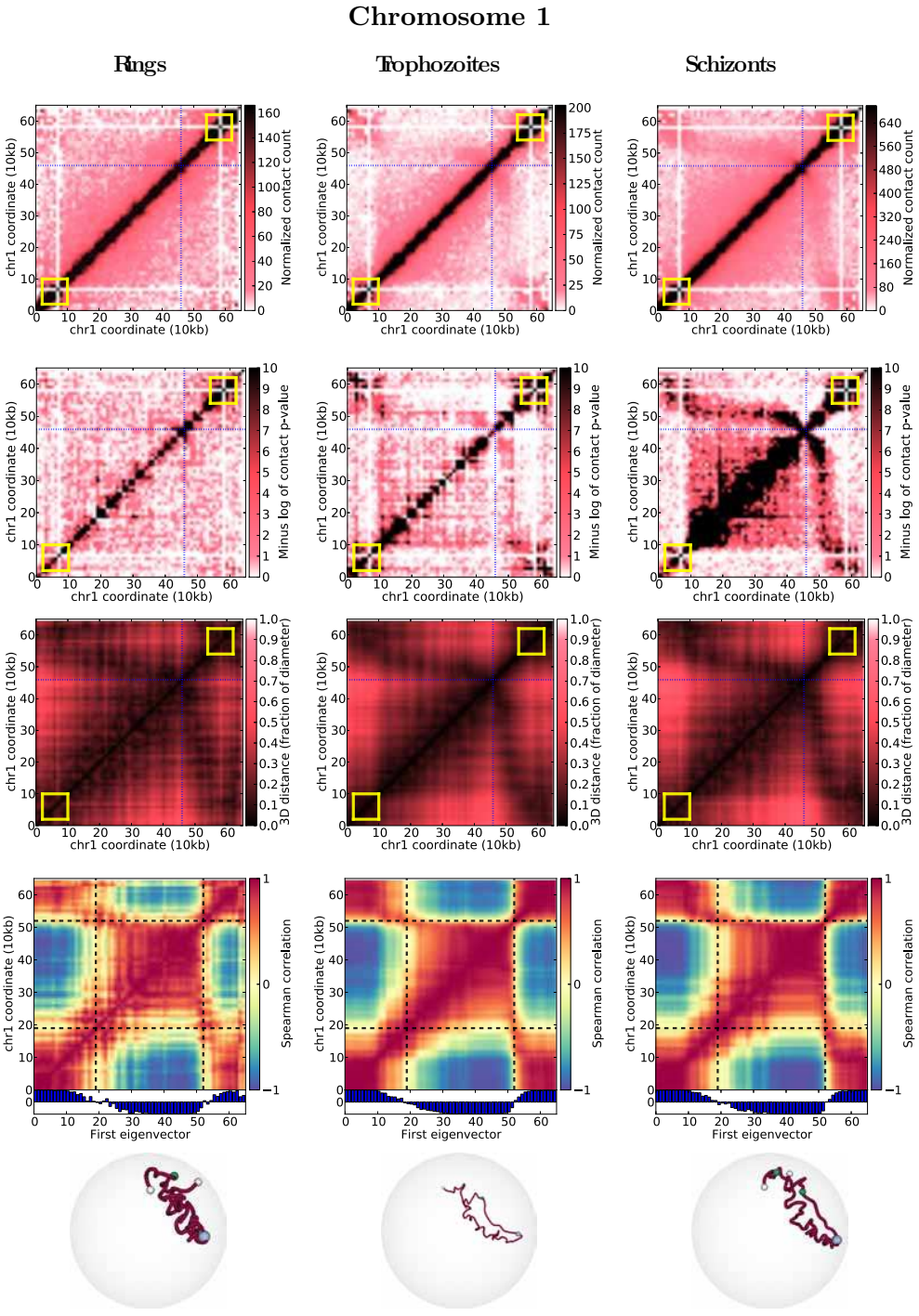


(F) After normalization - GC content

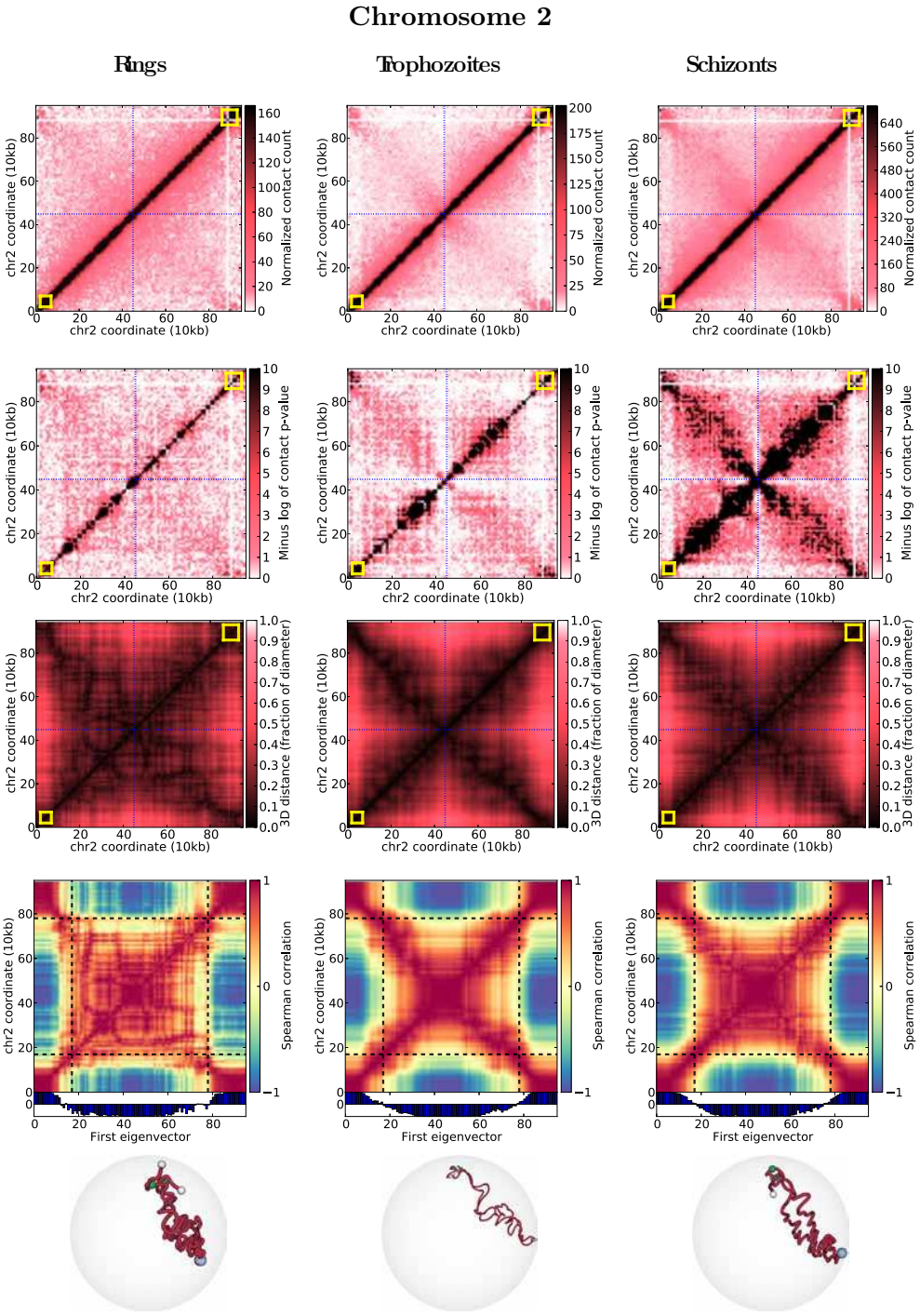
FIGURE 2: **Biases in raw and corrected contact maps for ring stage.** For each non-overlapping 10 kb window in the genome we compute a genomic feature such as the number of restriction enzyme (RE) cut sites, the fraction of uniquely mappable bases and GC content. For each feature, we group all windows into 10 equal sized bins based on the feature value. Each possible locus pair belongs to one specific bin pair (2D bin) which is indexed by the two horizontal axes. For each 2D bin we compute the mean contact count using all locus pairs that fall into that bin. The black, horizontal grid plane corresponds to the overall mean. For perfectly unbiased data all vertical bars will be of equal height and equal to the overall mean. (a, c, e) and (b, d, f) plots show biases for each indicated feature before and after normalization, respectively. Plots for trophozoite and schizont stages are similar (data not shown).

FIGURE 3: **Chromosome visualizations.**

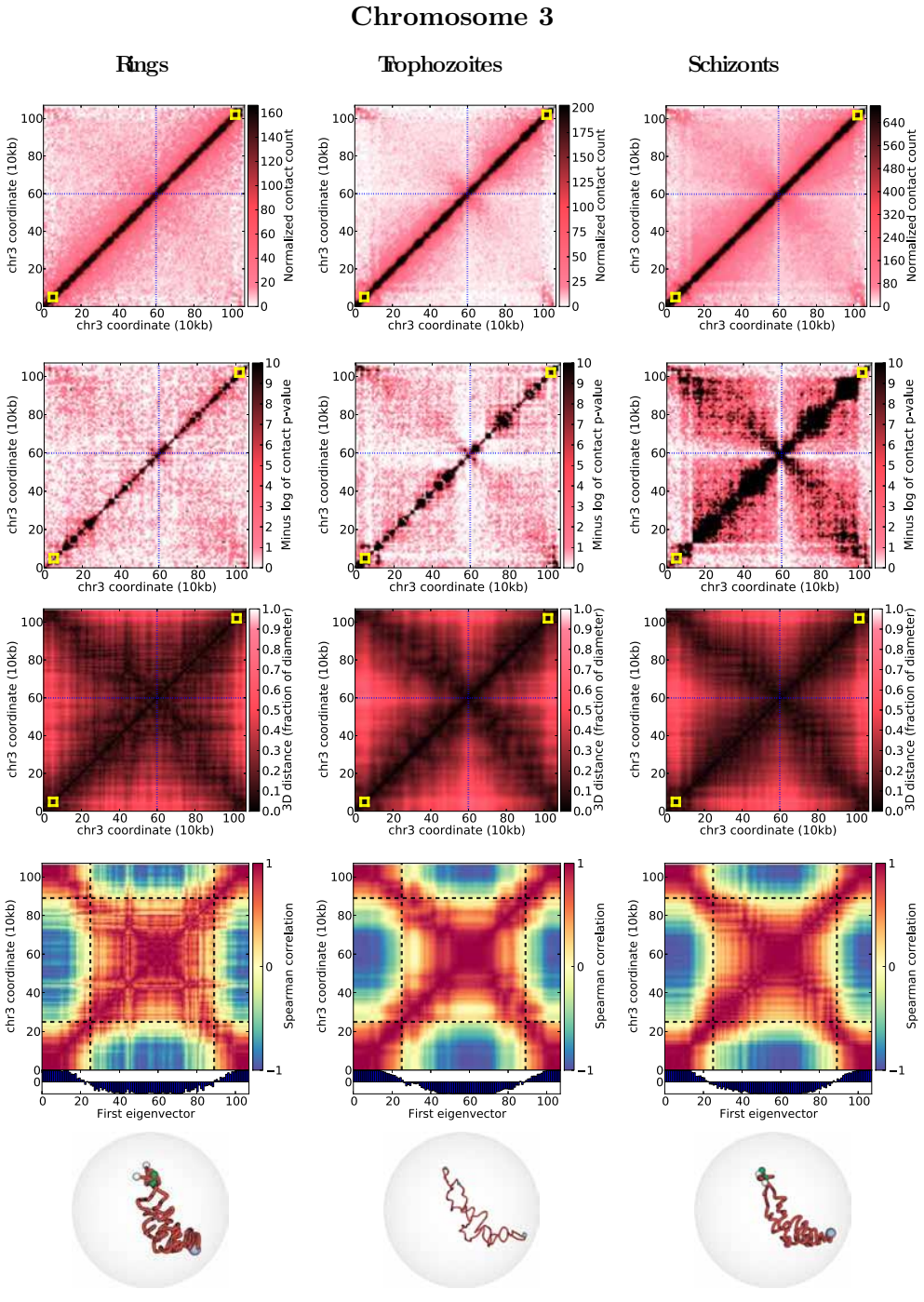
In the following fourteen pages, each page of figures corresponds to one chromosome, with the three time points (ring, trophozoite, schizont) arranged in three columns. Within each column, the top panels show the 10kb resolution contact count matrix after normalization using ICE Imakaev et al. [2012], the p-values assigned to contacts, and the pairwise Euclidean distances derived from the 3D model. Within each matrix, clusters of VSRM genes are indicated with yellow boxes, and centromere locations are indicated with blue dotted lines. The fourth panel in each column illustrates the eigenvalue analysis, with compartment boundaries aggregated over the three stages (Methods) indicated by black dotted lines. The bottom panel shows the chromosome's inferred configuration in 3D with light blue spheres indicating centomeres, white spheres indicating telomeres and green spheres indicating midpoints of VRSM gene clusters.



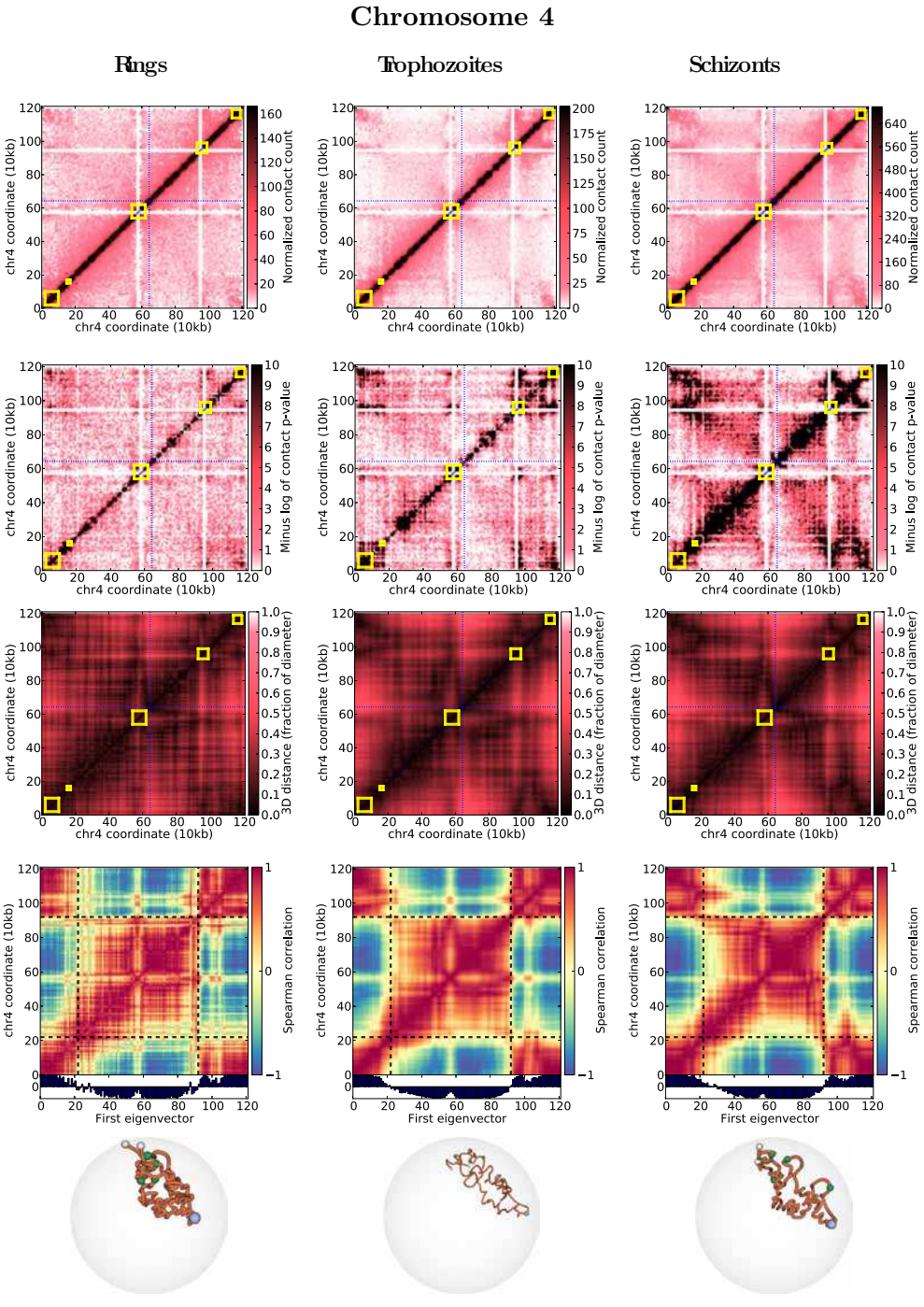
(A) Chromosome 1



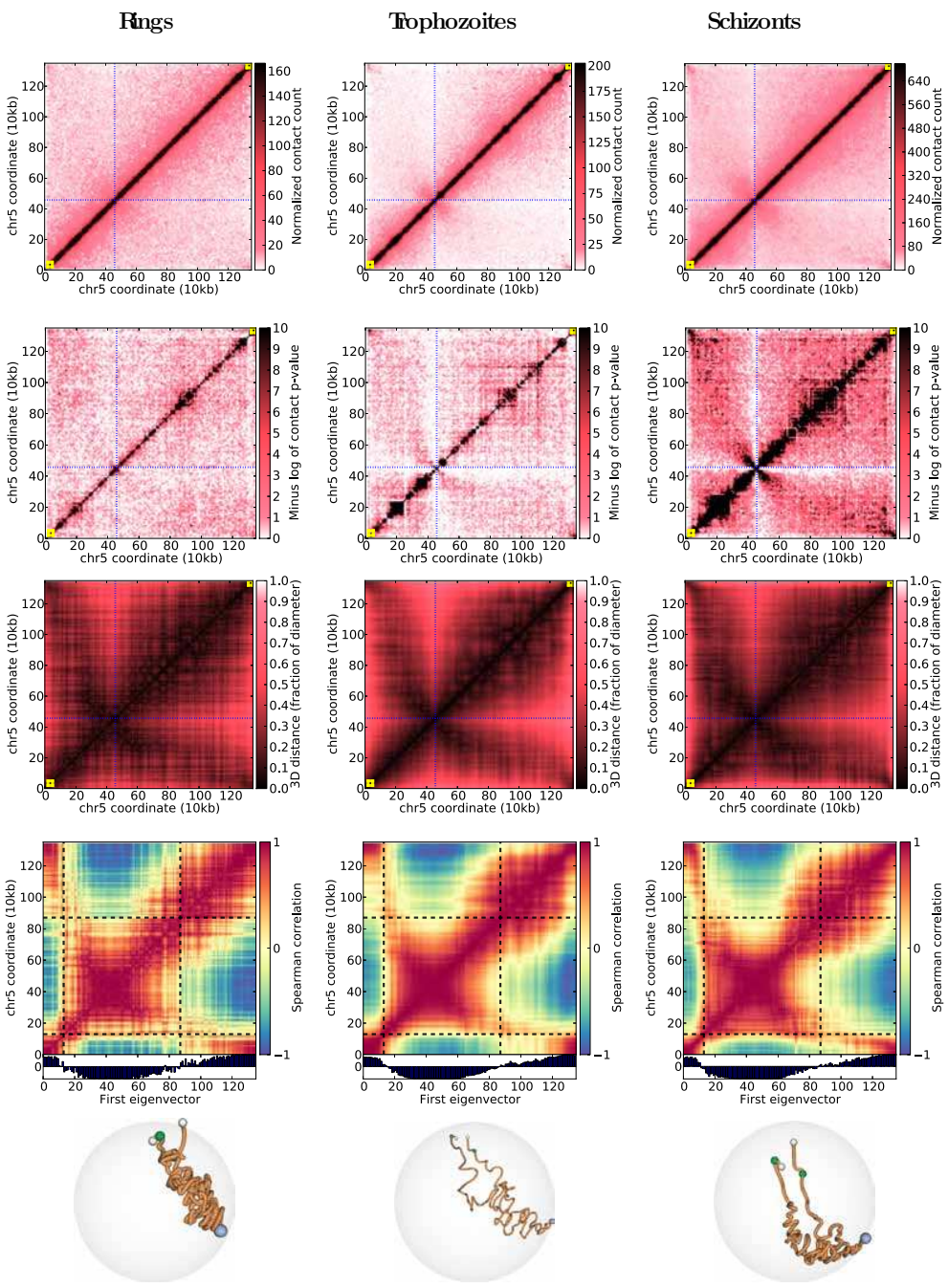
(B) Chromosome 2



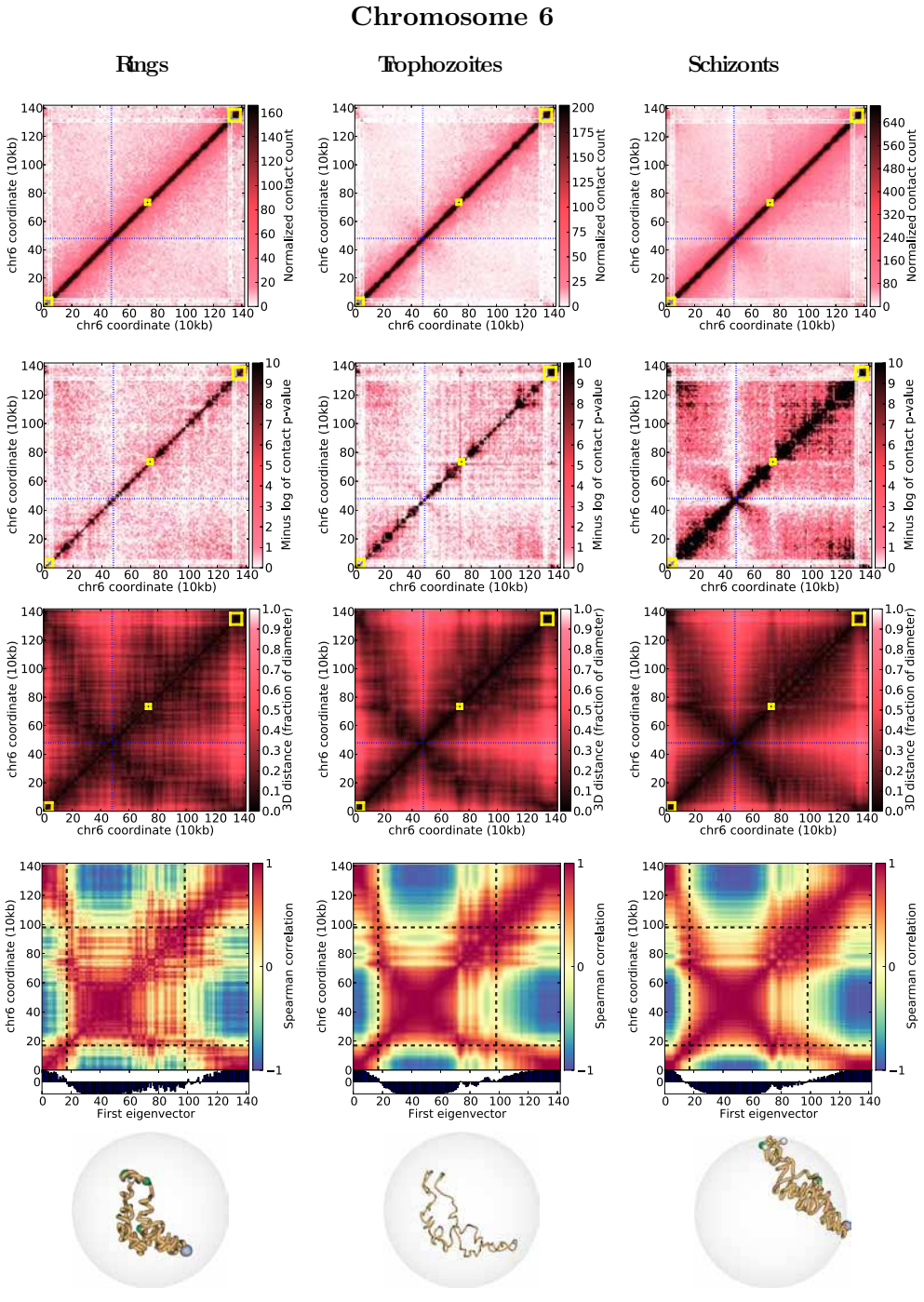
(c) Chromosome 3



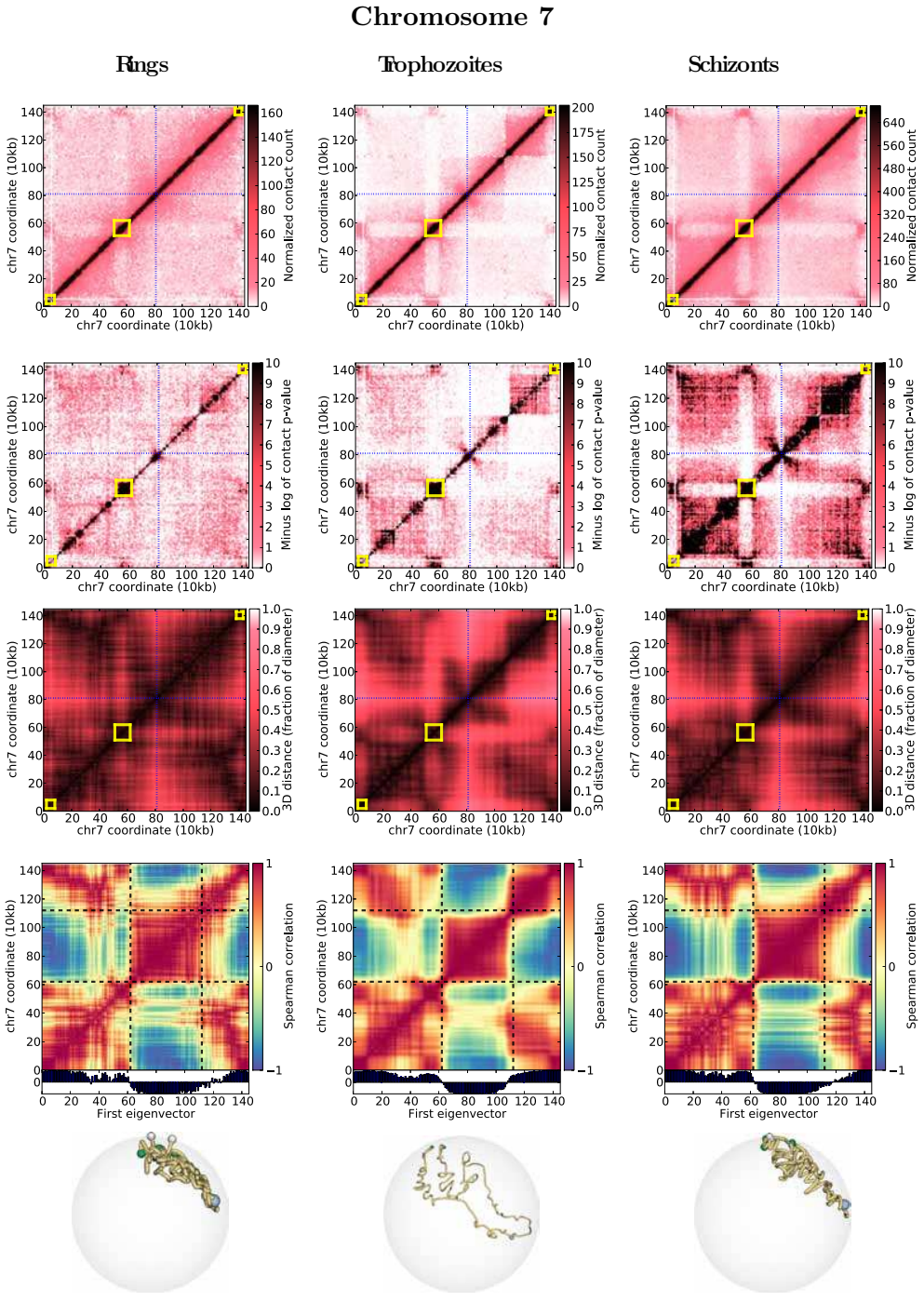
(D) Chromosome 4



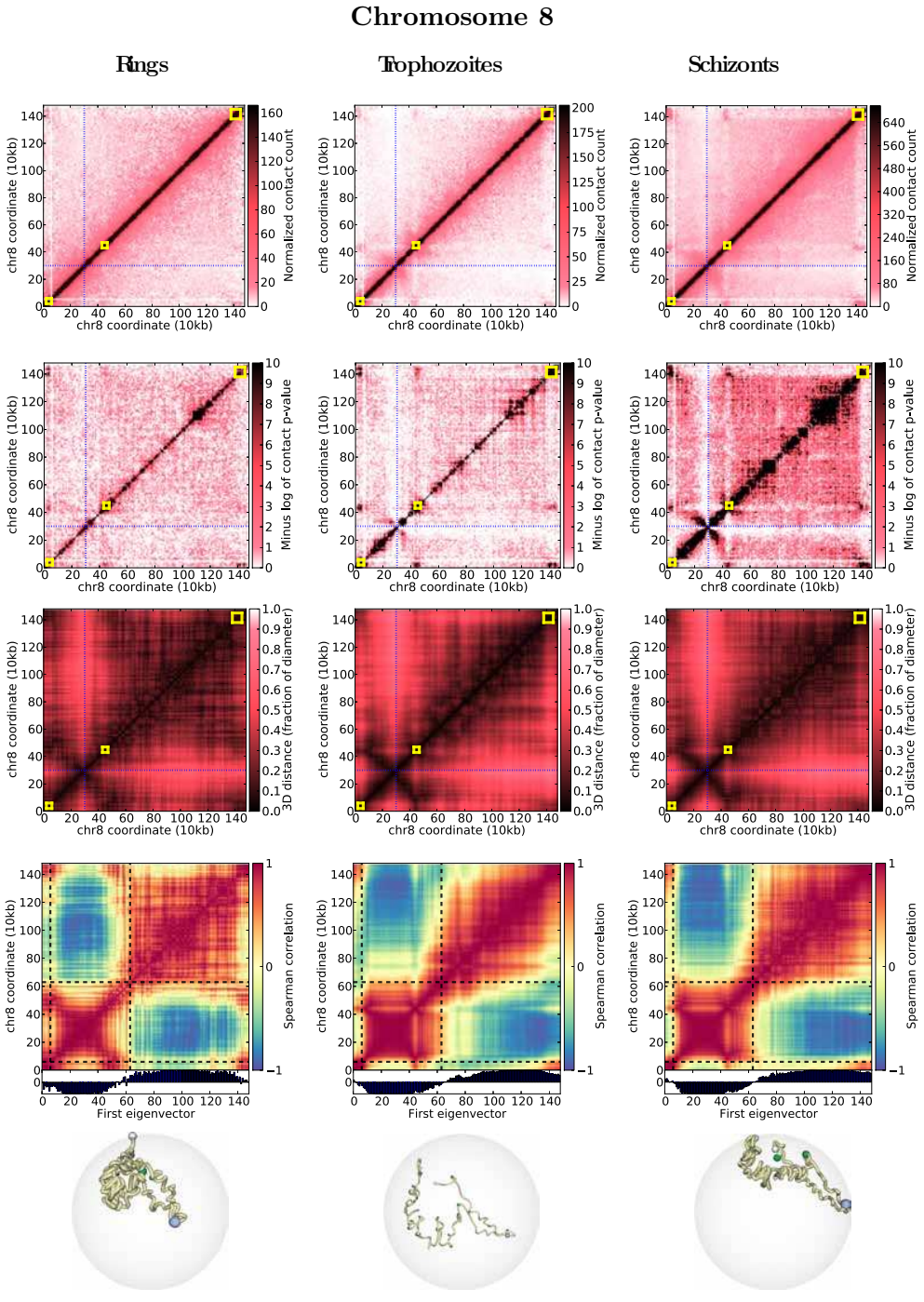
(E) Chromosome 5



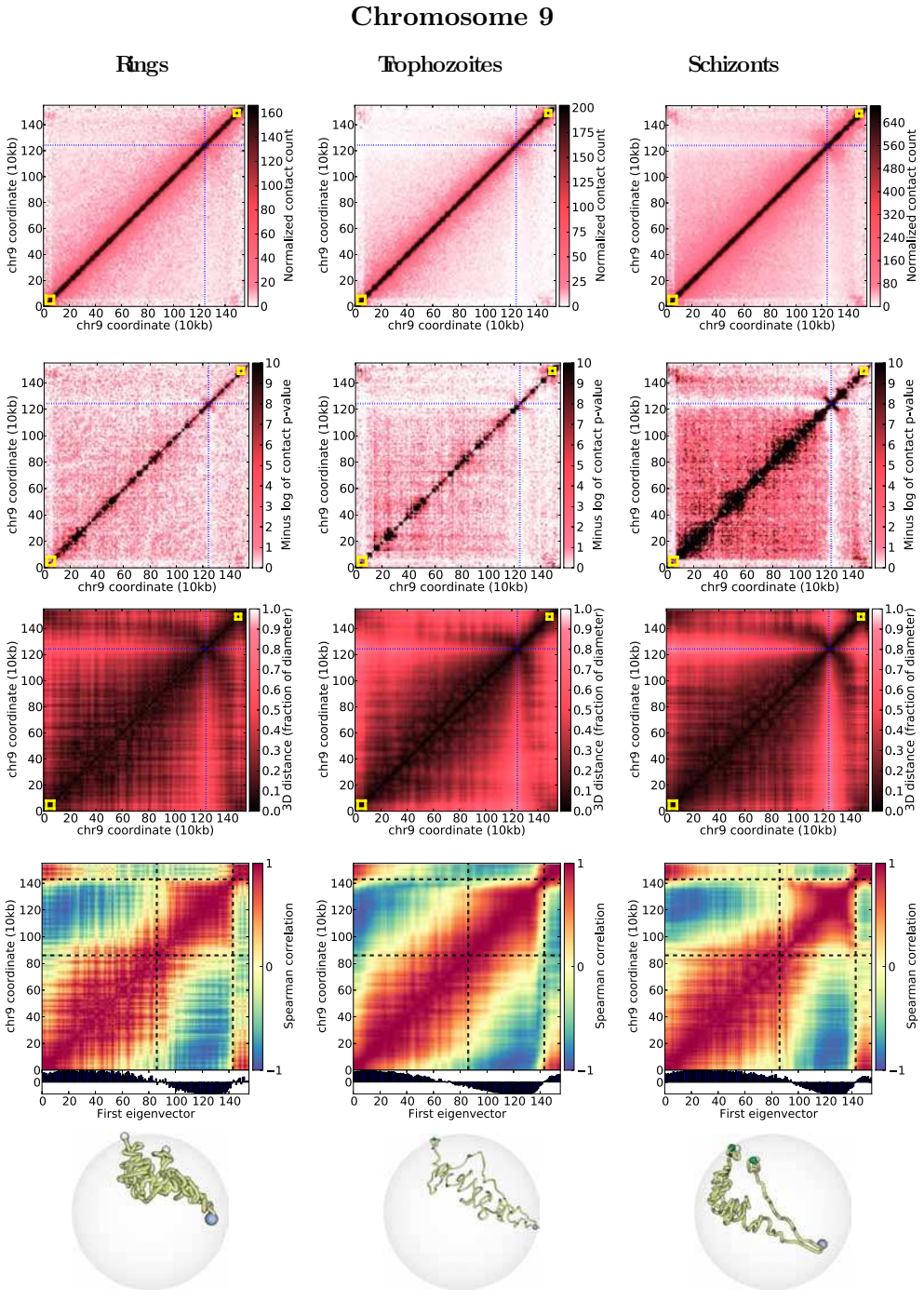
(F) Chromosome 6



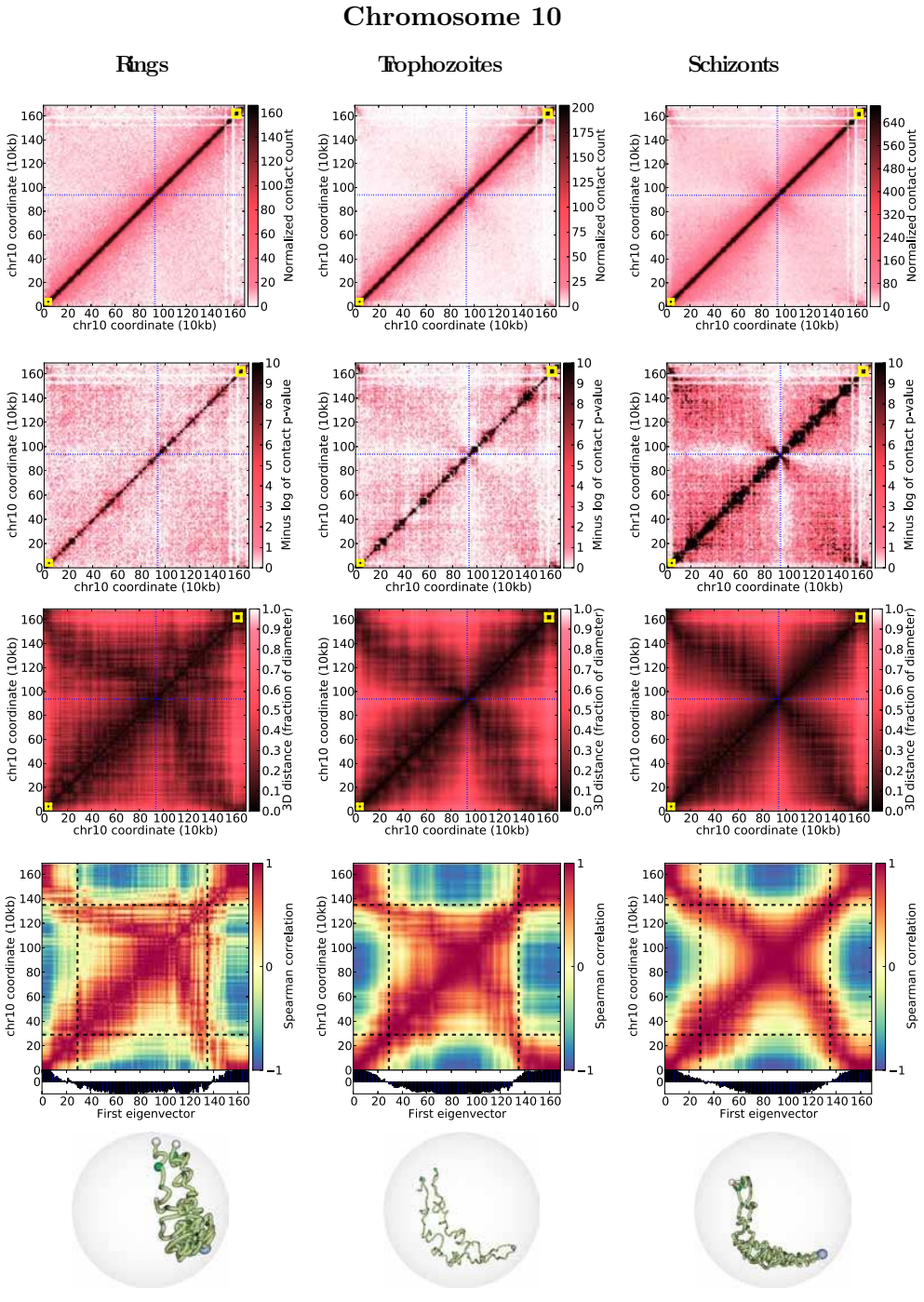
(G) Chromosome 7



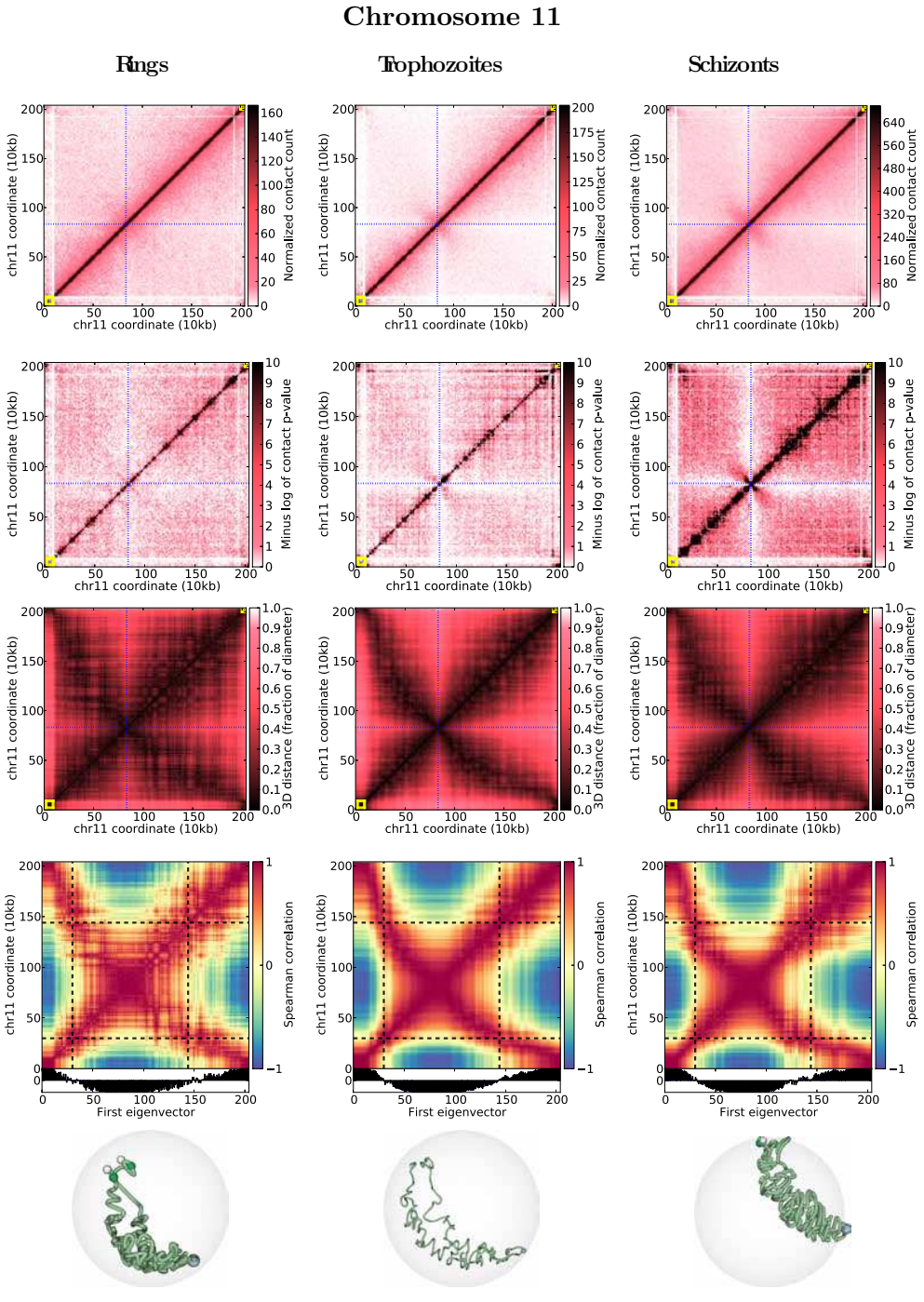
(H) Chromosome 8



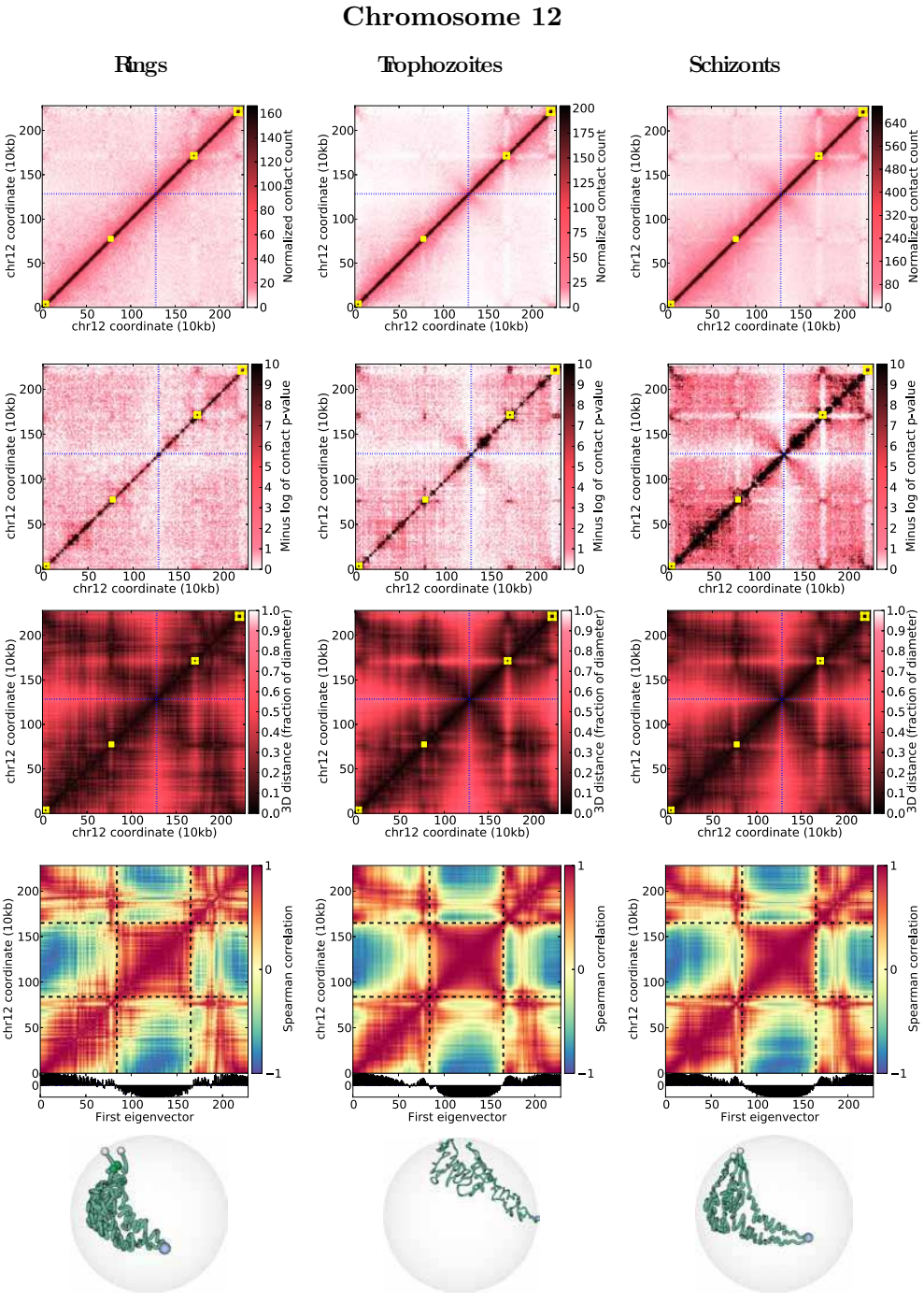
(1) Chromosome 9



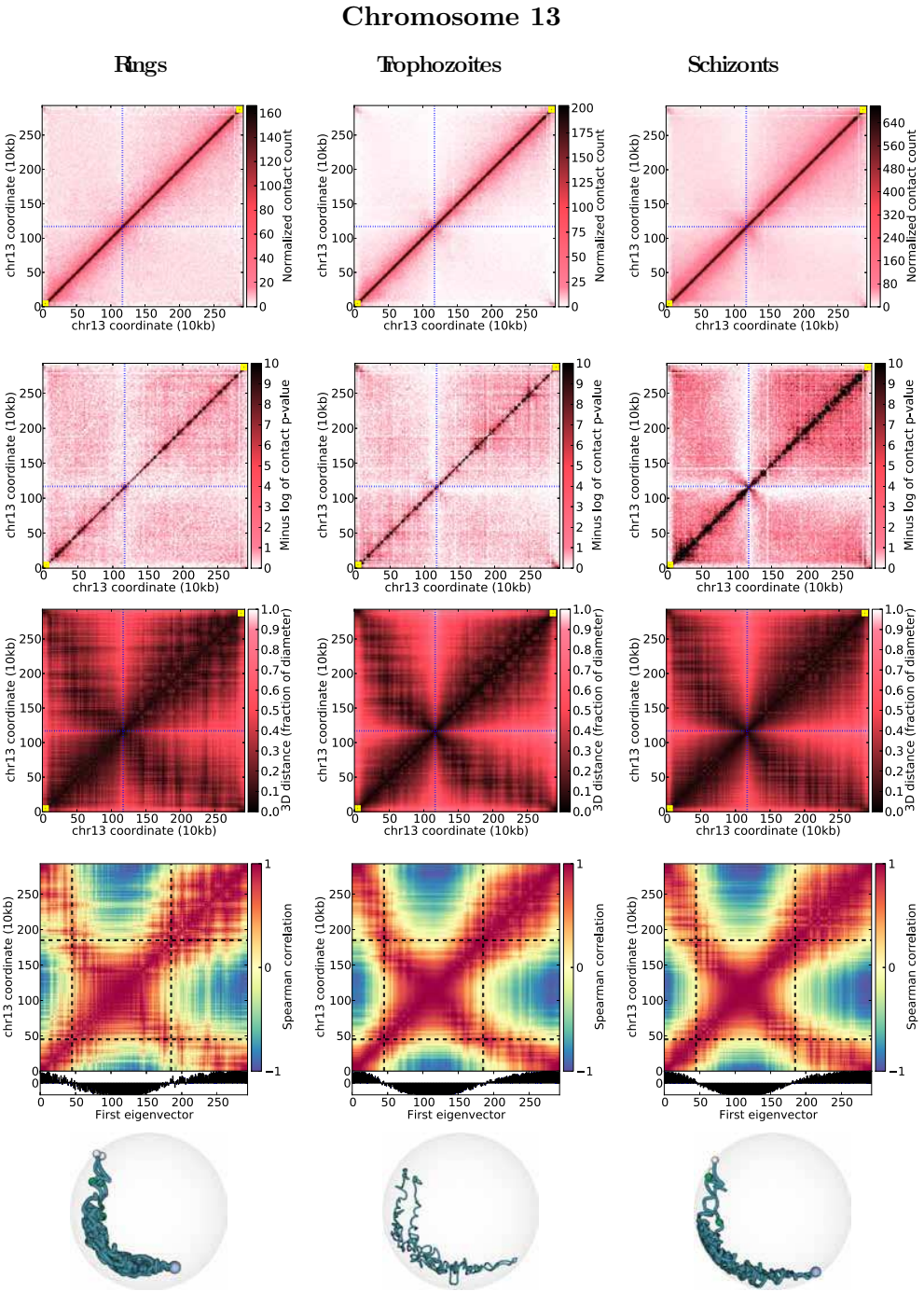
(J) Chromosome 10



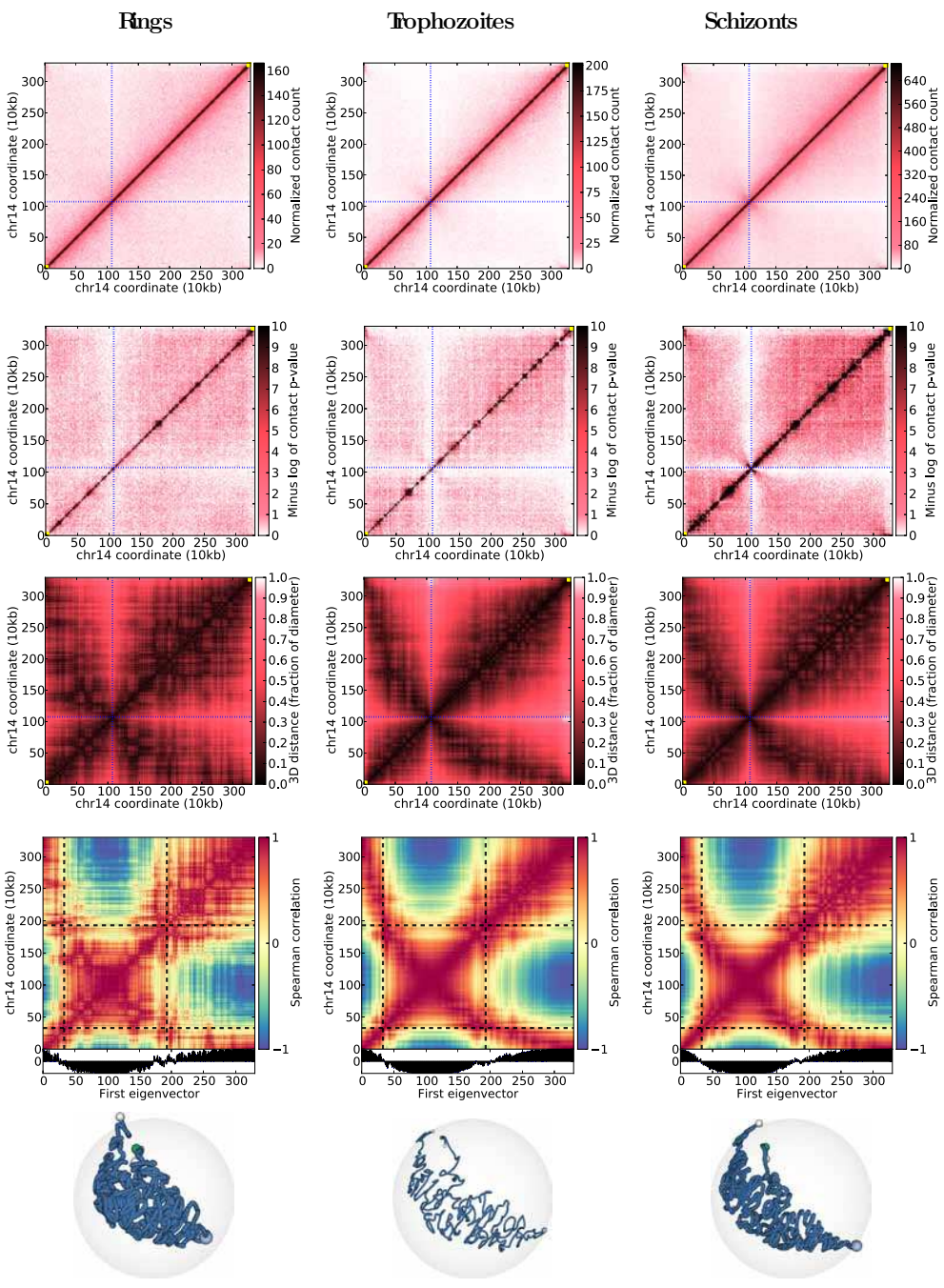
(K) Chromosome 11



(L) Chromosome 12



(M) Chromosome 13



(N) Chromosome 14

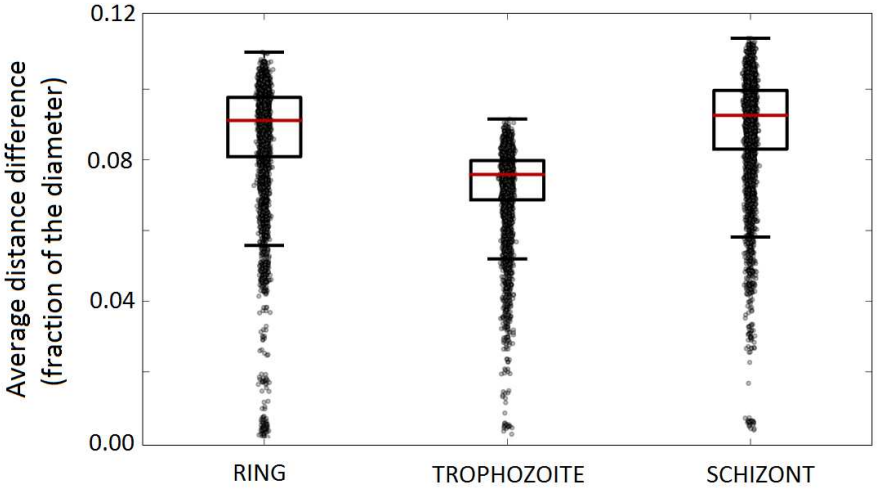
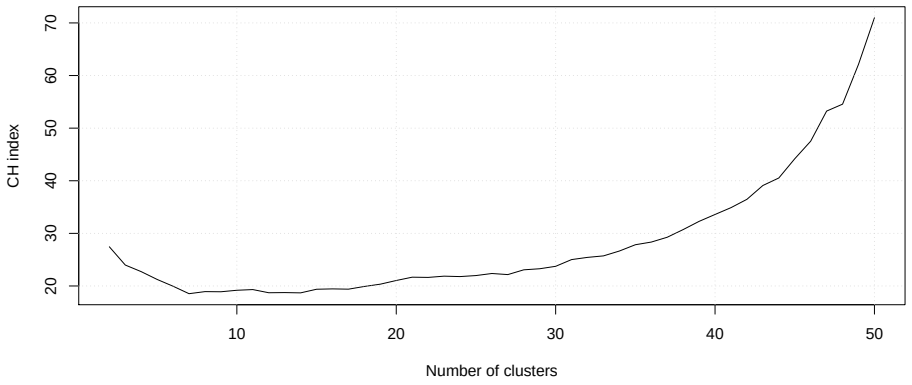
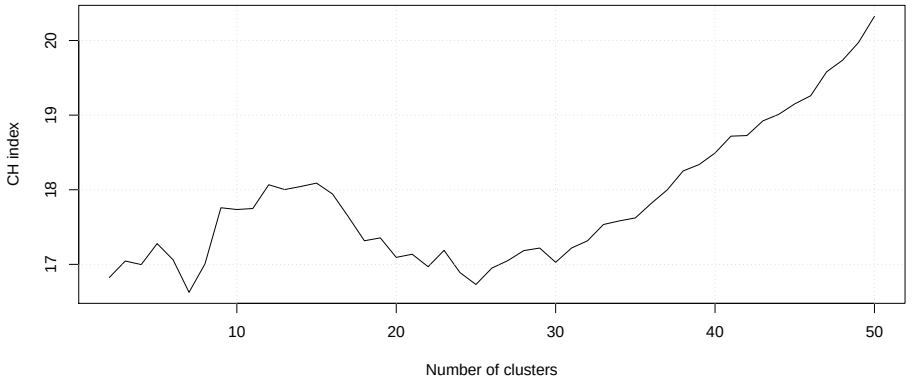


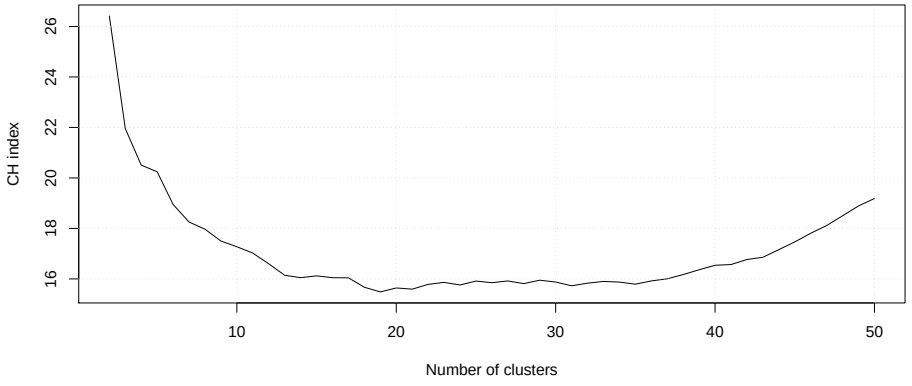
FIGURE 4: **Similarity between 3D models inferred from 100 different initializations.**
We computed the average distance differences (Methods) for each pair of structures (i.e., $\binom{100}{2}$) that are inferred from different initializations and summarized these difference using a box plot for each stage. Each box extends from the lower to upper quartile values with a red line at the median. These results show that the 3D distance between a pair of loci varies, on the average, less than 10% of the nuclear diameter from one structure to another.



(A) Ring



(B) Trophozoite



(C) Schizont

FIGURE 5: **Clustering of the 100 structures using pairwise RMSD values.** To assess whether the 100 structures generated from different random initializations fall into discrete clusters we performed hierarchical clustering on the pairwise RMSD matrix of each stage. We computed and plotted the Calinski-Harabasz (CH) index Calinski and Harabasz [1974] for each clustering while varying the number of clusters from 2 to 50. None of the stages exhibited a clear peak of the CH index, suggesting that the set of structures do not fall into discrete clusters.

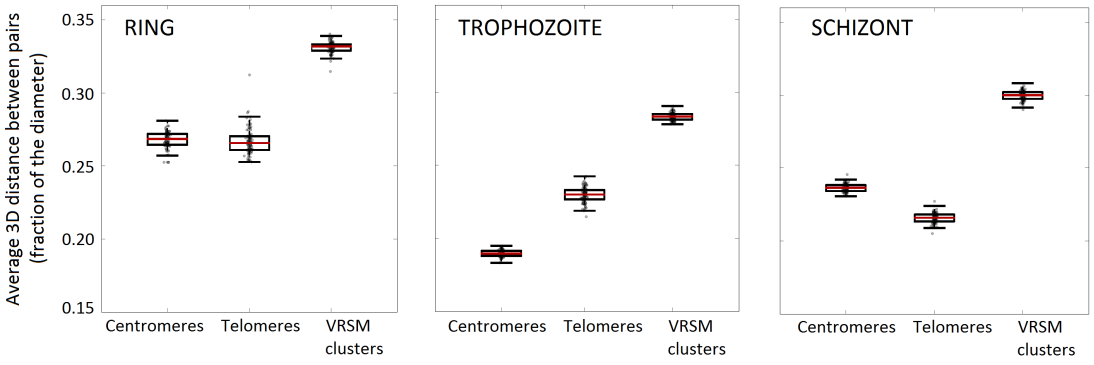


FIGURE 6: **Conservation of centromere, telomere and VRSM gene colocalizations across 100 different initializations.**
We computed the average 3D distance between pairs of centromeres ($\binom{14}{2}$ pairs), telomeres ($\binom{28}{2}$ pairs) and VRSM clusters (8 internal, 28 subtelomeric clusters and a total of $\binom{36}{2}$ pairs) for each of the 100 structures inferred from different initializations and summarized these average distances using a box plot for each stage. Each box extends from the lower to upper quartile values with a red line at the median. These results suggest that the major organizational hallmarks concerning colocalization of centromere, telomere and VRSM gene regions are common to all structures gathered from different initializations.

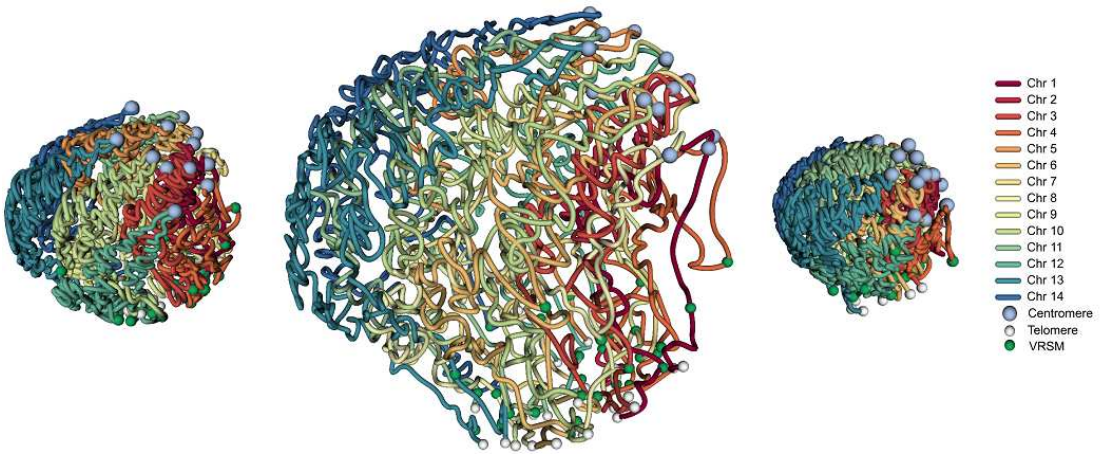
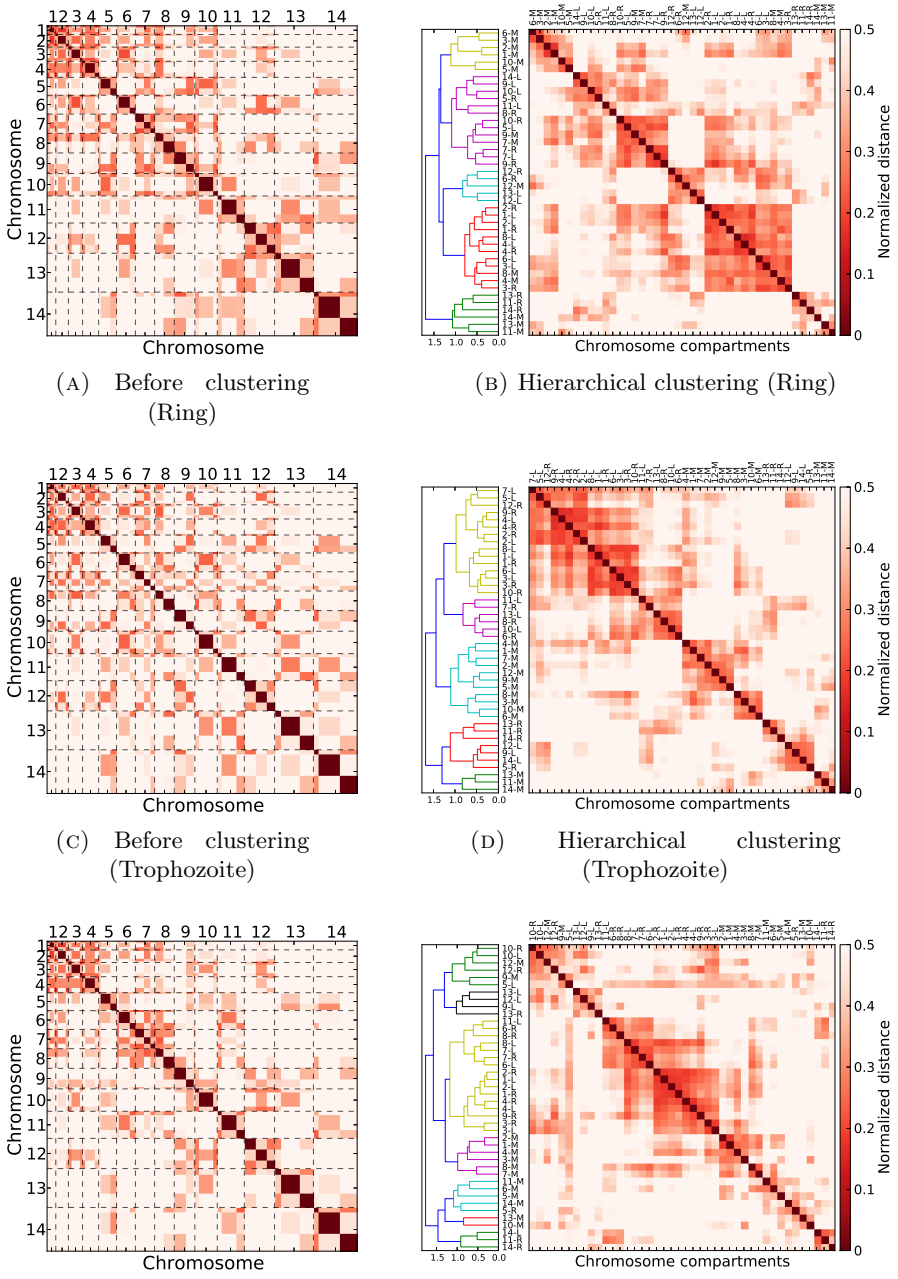


FIGURE 7: **3D structures of all three stages (centromere clustering).** This figure is identical to Main Figure 2a except the view is rotated to visualize the centromere clustering for each stage. Centromeres and telomeres are indicated with light blue and white spheres, respectively. Midpoints of VRSM gene clusters are shown with green spheres.



(A) Before clustering (Ring)

(B) Hierarchical clustering (Ring)

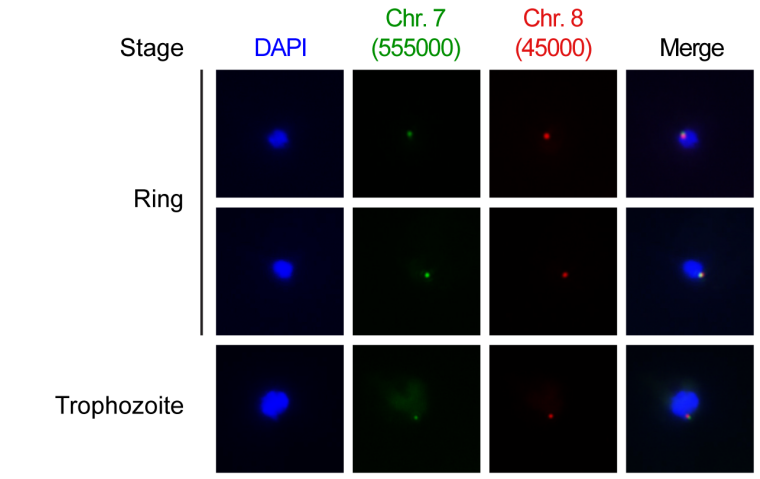
(C) Before clustering (Trophozoite)

(D) Hierarchical clustering (Trophozoite)

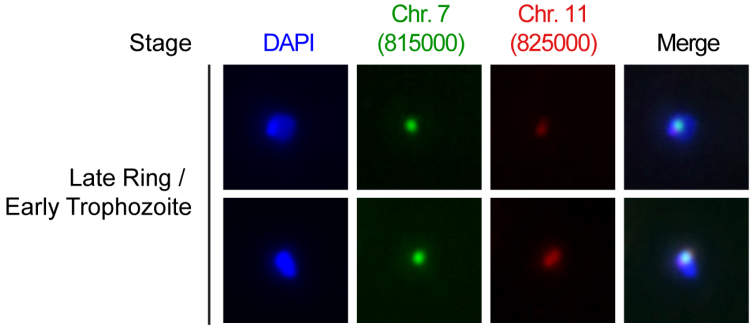
(E) Before clustering (Schizont)

(F) Hierarchical clustering (Schizont)

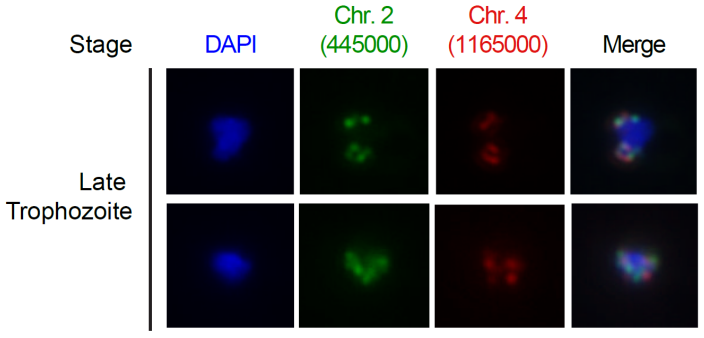
FIGURE 8: **Hierarchical clustering of compartment distance matrices.** Pairwise compartment distance matrices (42×42, three compartments on each chromosome) that are identified by eigenvalue decomposition (Methods) for (a) ring, (c) trophozoite and (e) schizont stages. Distances are averaged over all pairs of loci between the two compartments and normalized using nuclear diameter to result in a fraction between 0 and 1. In the figure, the actual length of each compartment and each chromosome are preserved. Each compartment is colored separately, with dashed lines segregating adjacent chromosomes. Hierarchical clustering of pairwise compartment distance matrices for (b) ring, (d) trophozoite and (f) schizont stages. Clustering was performed using the average linkage score. Each compartment is represented by a fixed length, and L, M, R denote left, mid, right compartments, respectively. For all panels the color bars extend from 0 to 0.5 (i.e., distance equals nuclear radius).



(A)

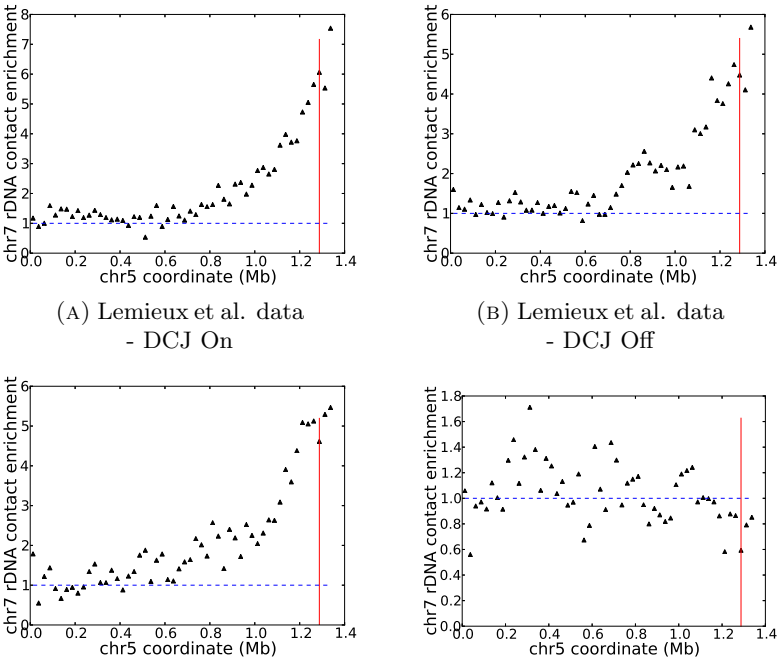


(B)



(C)

FIGURE 9: **Validation of 3D models with DNA FISH.**
Additional FISH images for (a) a pair of interchromosomal loci with VRSM genes (chr7:550,000-560,000 containing internal VRSM genes and chr8:40,000-50,000 containing subtelomeric VRSM genes) (b) a pair of interchromosomal loci that harbor no VRSM genes (chr7:810,000-820,000 and chr11:820,000-830,000). (c) FISH images showing lack of colocalization as a negative control for a pair of interchromosomal loci that harbor no VRSM genes and have no contacts in trophozoite stage (chr2:440,000-450,000 and chr4:1,160,000-1,170,000).



(A) Lemieux et al. data - DCJ On

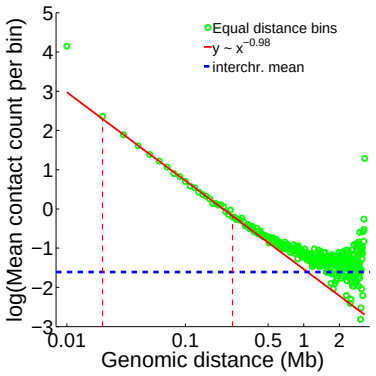
(B) Lemieux et al. data - DCJ Off

(C) Lemieux et al. data - B15C2

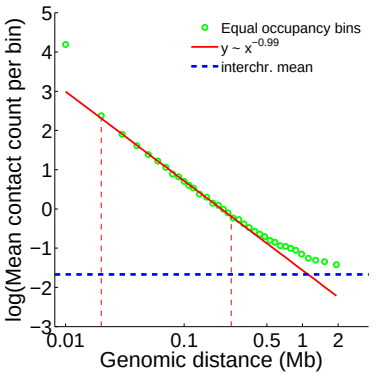
(D) Lemieux et al. data - NF54 control

FIGURE 10: **Clustering of highly transcribed rDNA units in Lemieux et al. data.**

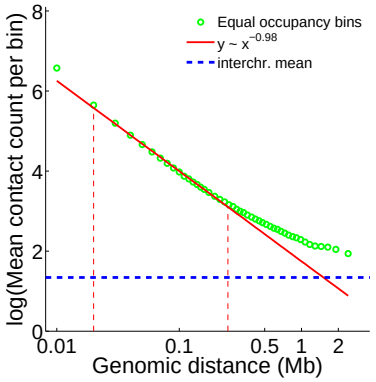
Hi-C libraries generated with MboI restriction enzyme from Lemieux et al. Lemieux et al. [2013] were mapped to the *P. falciparum* genome and further processed using the pipeline we processed our data with to generate and normalize contact maps at 25 kb resolution. The normalized contact maps were used for virtual 4C plots using as a bait the A-type rDNA unit on chromosome 7. As suggested in Lemieux et al., contact counts from 50 kb up- and downstream of the 25 kb bin containing rDNA unit were used, and the rDNA-containing window itself was removed from the analysis. For each window w on chromosome 5, the contact enrichment was calculated by dividing the contact count between the bait and w to the average interchromosomal contact count for the bait locus.



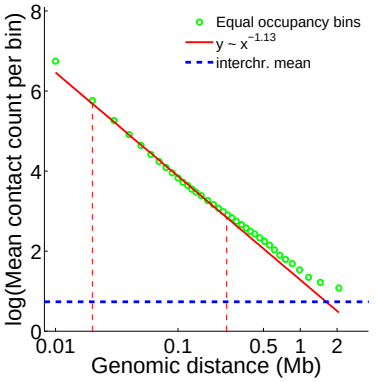
(a) Lemieux et al. data - B15C2



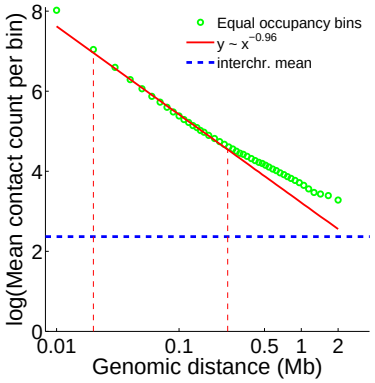
(b) Lemieux et al. data - B15C2



(c) Our data - Ring



(d) Our data - Trophozoite



(e) Our data - Schizont

FIGURE 11: **Comparison of inter and intrachromosomal contact prevalence.** The relationship between contact count and genomic distance is estimated using bins with equal genomic distances (e.g., 10 kb, 20 kb) in Lemieux et al. Lemieux et al. [2013]. Due to the diminishing number of possible locus pairs with increasing genomic distance (e.g., only one locus pair for the bin with the largest genomic distance) this estimation leads to many high variance bins for large genomic distances. This issue can be addressed by using variable-width bins that contain equal numbers of contacts (see Methods). Plotted are the log (base e) of mean contact count per bin when using (a) equal distance binning, (b) equal occupancy binning for B15C2 library of Lemieux et al. Lemieux et al. [2013] and (c-e) equal occupancy binning for ring, trophozoite and schizont stage data from this work. Dashed vertical red lines denote the range used to compute the log-linear fit.

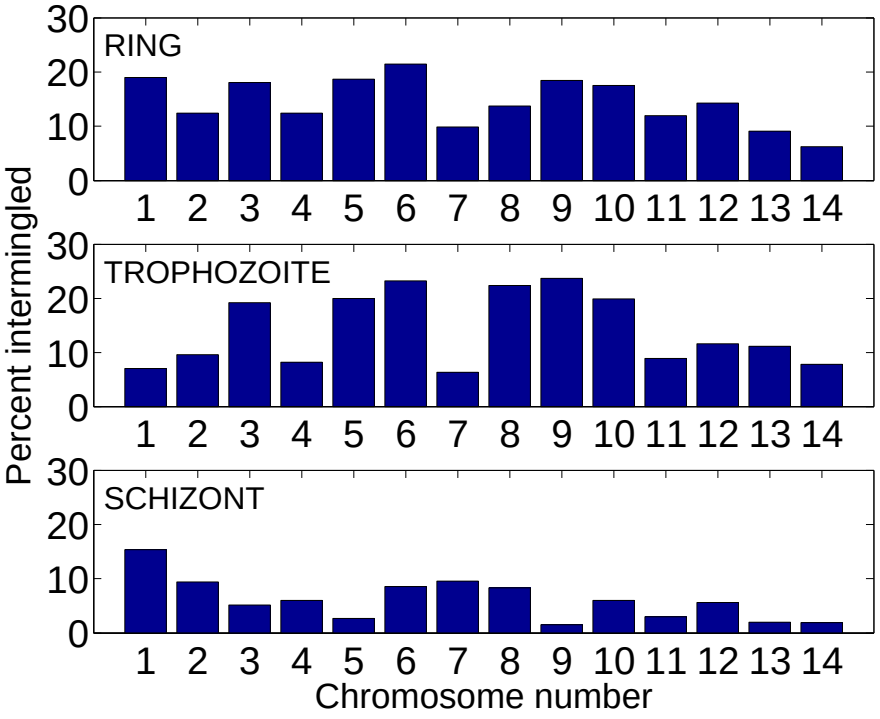
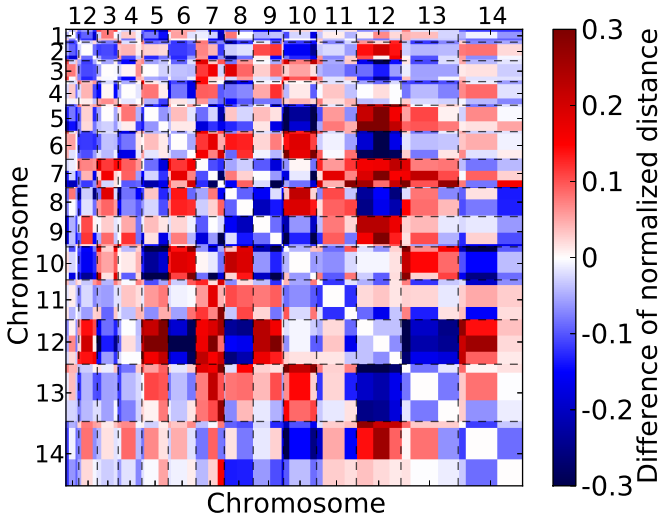
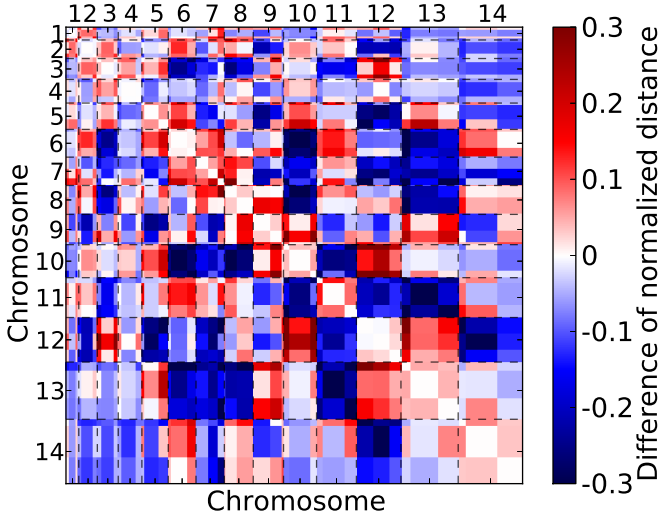


FIGURE 12: **Changes in chromosome territories during the erythrocytic cycle.** The extent to which a chromosome intermingles with other chromosomes is characterized by the percentage of nuclear volume that is jointly occupied by the chromosome of interest and at least one other chromosome, relative to the entire volume occupied by the chromosome. To compute the percentages on the y-axis, the nuclear volume was sampled using 1,000,000 randomly generated small spheres with radius 5% of the actual nuclear radius. For each chromosome i, two numbers were calculated: the number of spheres that contain a locus from chromosome i (n_i) and the number of such spheres that contain no locus from another chromosome (e_i). The percent intermingled (y-axis) for chromosome i is computed as $100 \times \frac{n_i - e_i}{n_i}$. Because the exact percentages are highly dependent on the selection of the random sphere size, the procedure was repeated using spheres with radii 2%, 10% and 20% of the nuclear volume. For each setting, the trophozoite stage exhibited the highest amount of intermingling, whereas the schizont stage showed the lowest. Also, the larger chromosomes (i.e., chromosomes with higher numbers) consistently showed lower intermingling compared to smaller chromosomes at each stage and for each threshold.



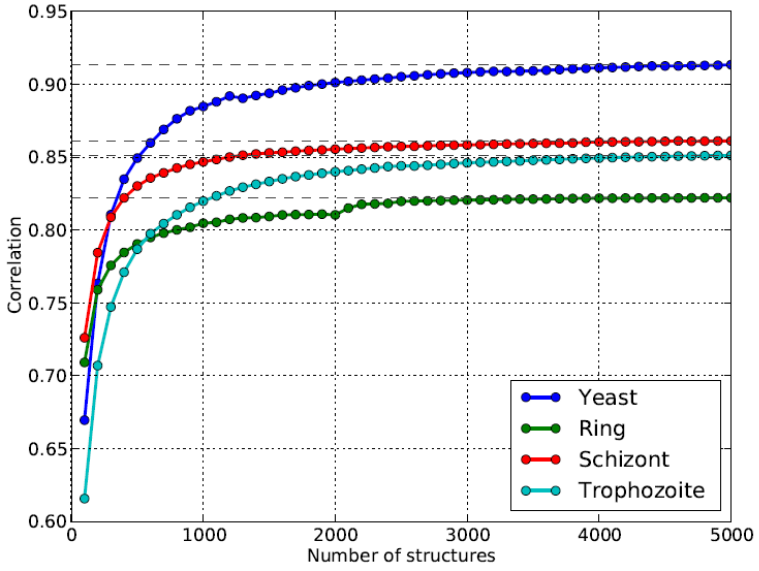
(A) From ring to trophozoite



(B) From trophozoite to schizont

FIGURE 13: **Movement of chromosome compartments with respect to each other.**

Each compartment movement matrix is generated by subtracting the pairwise compartment distance matrix (Supplementary Fig. 8) of one stage from the matrix of the preceding stage. Plotted are the movements (a) from ring to trophozoite (i.e., trophozoite minus ring), (b) from trophozoite to schizont (i.e., schizont minus trophozoite). Red color indicates that a pair of compartments are closer in the later stage compared to the earlier, and blue color indicates vice versa.

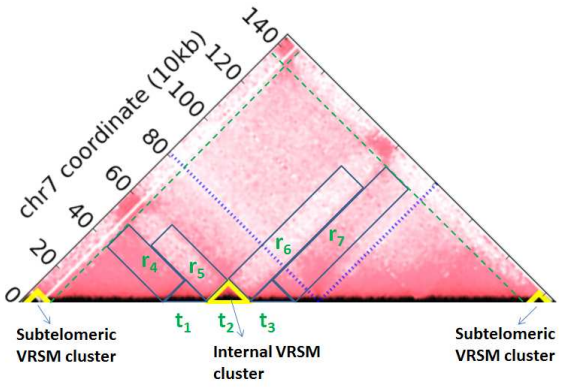


(A)

Contact map 1	*Contact map 2*	*Row-based corr.*	*Normalized row-based corr.*
Yeast *(Hi-C)* [Duan et al., 2010]	Yeast *(VE)*	0.915	0.573
	Yeast *(expected)*	0.922	0.115
Ring *(Hi-C)*	Ring *(VE)*	0.843	0.340
	Ring *(expected)*	0.928	0.072
Trophozoite *(Hi-C)*	Trophozoite *(VE)*	0.848	0.392
	Trophozoite *(expected)*	0.908	0.063
Schizont *(Hi-C)*	Schizont *(VE)*	0.864	0.487
	Schizont *(expected)*	0.923	0.081

(B)

FIGURE 14: **Volume exclusion modeling and correlation calculation.** (a) Row-based Pearson correlation between the observed Hi-C contact map and the average contact map from volume exclusion modeling as a function of the number of simulated structures. (b) Row-based Pearson correlation and normalized row-based Pearson correlation between the two contact maps listed in each row for various Hi-C libraries. *VE* refers to contact maps obtained from 5000 structures generated by volume exclusion and *expected* refers to matrices with expected contact counts generated from observed *Hi-C* matrices as described in Methods.

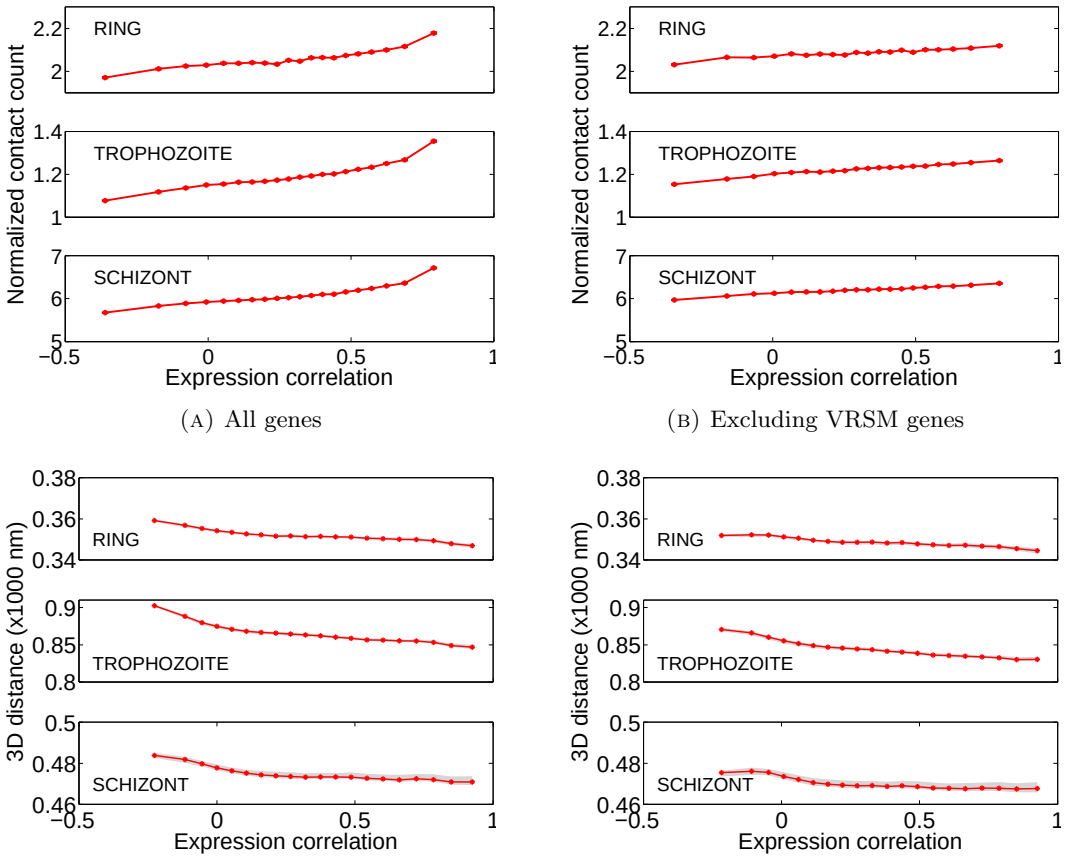


(A)

Internal VRSM	*Stage*	t_2 vs. t_1	t_2 vs. t_3	r_5 vs. r_4	r_6 vs. r_7
	R	0	0.7	0	-0.09
chr4(1)	T	0.06	0.2	-0.11	-0.25
	S	0.1	0.16	0.13	-0.14
	R	0.13	0.12	-0.38	-0.27
chr4(2)	T	0.15	0.15	-0.42	-0.33
	S	0.14	0.13	-0.37	-0.3
	R	0.03	0.08	-0.37	-0.34
chr4(3)	T	0.03	0.06	-0.55	-0.41
	S	0.03	0.01	-0.48	-0.42
	R	0.12	0.1	0.02	-0.06
chr6(1)	T	0.19	0.2	-0.09	-0.16
	S	0.16	0.18	-0.08	-0.14
	R	0.11	0.2	-0.19	-0.18
chr7(1)	T	0.19	0.28	-0.36	-0.3
	S	0.08	0.19	-0.27	-0.27
	R	0.15	0.08	-0.05	-0.09
chr8(1)	T	0.17	0.09	-0.17	-0.13
	S	0.14	0.09	-0.11	-0.11
	R	0.07	0.05	-0.02	-0.01
chr12(1)	T	0.05	0.14	-0.04	-0.02
	S	0.09	0.12	-0.07	-0.08
	R	0.09	0.09	-0.09	-0.11
chr12(2)	T	0.17	0.1	-0.27	-0.24
	S	0.09	0.07	-0.19	-0.23

(B)

FIGURE 15 *(preceding page)*: **Quantification of domain-like behavior of VRSM gene clusters.**(a) Each internal VRSM gene cluster is characterized by a set of strong intra-cluster contacts (t_2) and two sets of contacts with adjacent regions (r_5 and r_6) that are weak. For comparison, we also consider flanking, non-VSRM regions of the same size as the original VRSM cluster, including their "intra-cluster" contacts (t_1 and t_3) which should be similar to t_2 for a contact map without domain-like structures around VRSM clusters and contacts with adjacent regions (r_4 and r_7) which are comparable to (r_5 and r_6). As seen in this example, a domain-like structure for a VRSM cluster leads to stronger contacts (+ sign) within t_2 compared to both t_1 and t_3, and weaker contacts (- sign) within r_4 and r_7 compared to r_5 and r_6. (b) The table reports, for each internal VRSM gene cluster and each stage, the average normalized difference between the intra-cluster contacts within the cluster compared to its two flanking control regions, and similarly for the contacts with adjacent regions. The metric we use for comparing two contact sub-matrices X, Y of dimension $N \times M$ is $\frac{1}{NM} \sum_{i=1}^{N} \sum_{j=1}^{M} \frac{x_{ij}-y_{ij}}{\frac{1}{2}(x_{ij}+y_{ij})}$ where x_{ij} and y_{ij} are the ijth entries of X and Y, respectively. Values that have signs inconsistent with the expected pattern (i.e., +, +, -, -) are indicated with a grey background. Every internal VRSM cluster exhibits the expected sign pattern in at least one stage.



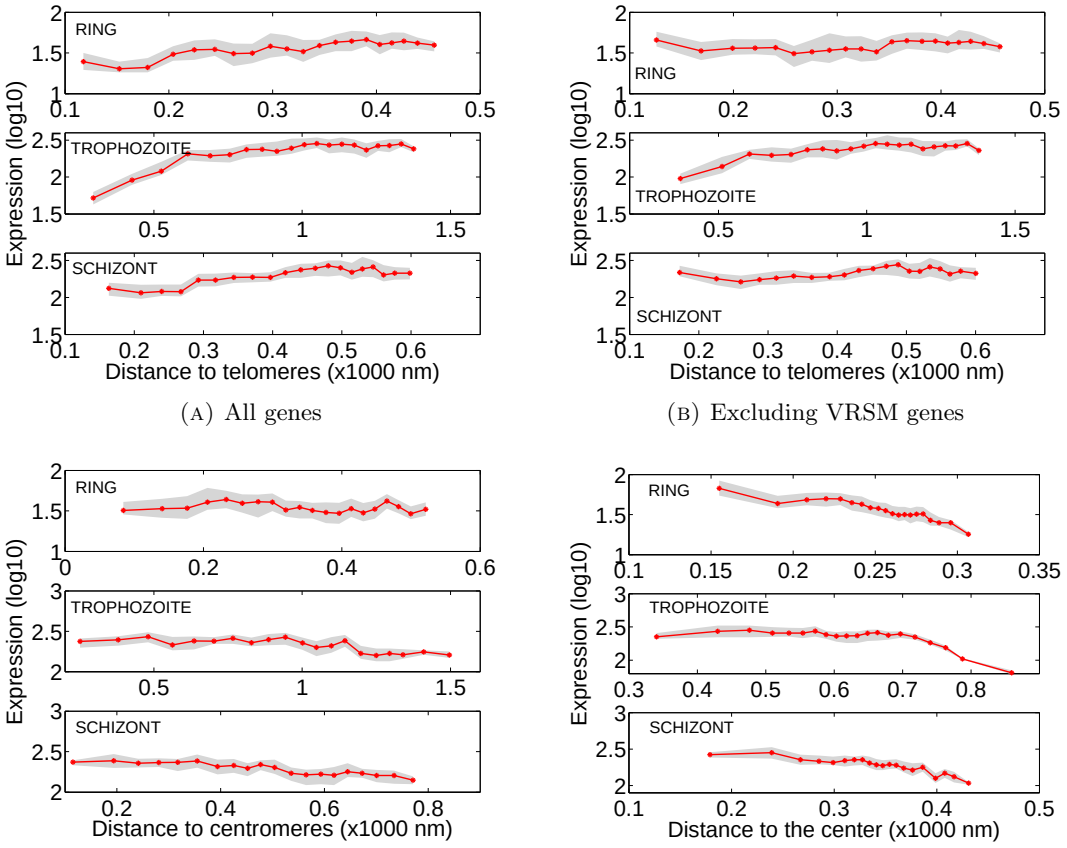
(A) All genes

(B) Excluding VRSM genes

(C) All genes

(D) Excluding VRSM genes

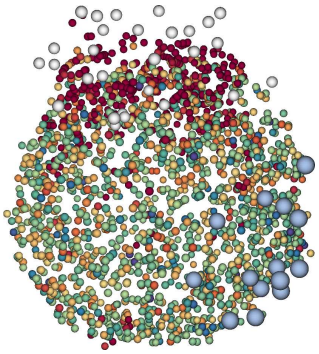
FIGURE 16: **Revisiting the relationship between 3D architecture and gene expression by excluding VRSM genes.**
(a) is identical to Main Figure 6a and (b) is generated identical to (a) except all gene pairs involving at least one VRSM gene are omitted from the analysis. Re-evaluation of our hypothesis that interchromosomal gene pairs that have contact counts within the top 20% for each stage have more highly correlated expression profiles than the remaining gene pairs still yielded significant p-values for each stage [Wilcoxon rank-sum test, p-values 1.07e-70 (ring), 0 (trophozoite), and 1.68e-302 (schizont)]. (c) is identical to Main Figure 6b and (d) is generated identical to (c) except all gene pairs involving at least one VRSM gene are omitted from the analysis. Re-evaluation of our hypothesis that interchromosomal gene pairs closer than 20% of the nuclear diameter have more highly correlated expression profiles than other genes still yielded significant p-values for each stage [Wilcoxon rank-sum test, p-values 3.27e-48 (ring), 1.32e-157 (trophozoite), and 2.16e-5 (schizont)].



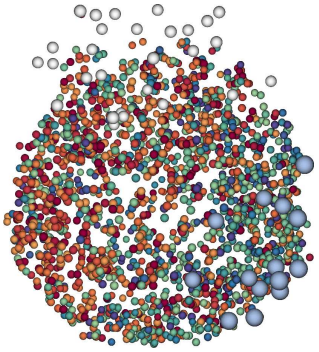
(A) All genes (B) Excluding VRSM genes

(C) All genes (D) All genes

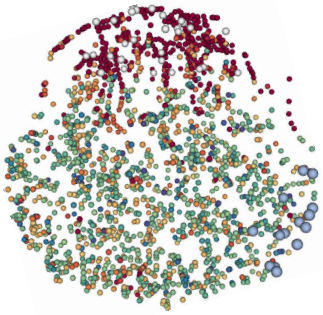
FIGURE 17: **The relationship between distance to the telomeres, nuclear center and centromeres versus the gene expression.**
(a) is identical to Main Figure 6c and (b) is generated identical to (a) except all VRSM genes are omitted from the analysis. Re-evaluation of our hypothesis that genes which lie within a distance of 20% of the nuclear diameter to the centroid of the telomeres exhibit lower expression levels yielded a significant p-value for trophozoite stage but not for ring and schizont stages at a significance threshold of 0.01 [Wilcoxon rank-sum test, p-values 0.21 (ring), 1.5e-3 (trophozoite), and 0.035 (schizont)]. (c) and (d) are generated identical to (a) expect the distance of genes are measured to (c) the centroid of the centromeres and (d) the nuclear center. For each figure, genes are first sorted in increasing order according to their distances to the landmark of interest and then binned into 20 equal width quantiles (5th, 10th, ..., 100th). For each bin, the average distance to the landmark (x-axis) and the average log expression value [Bunnik et al., 2013] together with its standard error (y-axis) are computed and plotted.



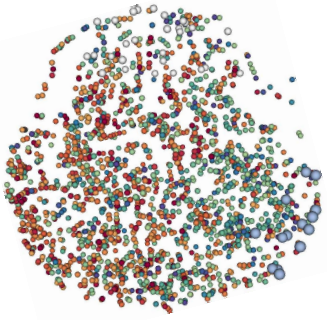
(a) First component (Ring)



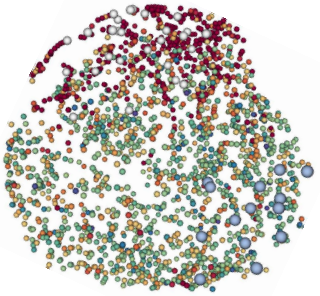
(b) Second component (Ring)



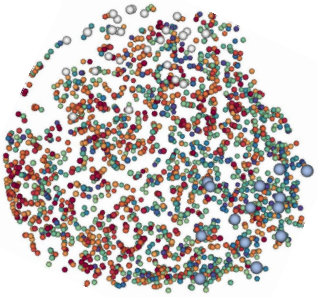
(c) First component (Trophozoite)



(d) Second component (Trophozoite)



(e) First component (Schizont)



(f) Second component (Schizont)

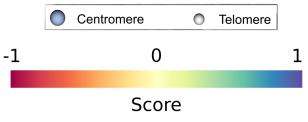


FIGURE 18: **kCCA expression profiles component score.**
Each panel shows the projection of the gene expression profile onto one of the two extracted kCCA profiles for a specified erythrocytic stage, with the score of the projection encoded on the color scale. For the first kCCA component, the projections consistently exhibit a striking gradient from the telomeric region across the nucleus, while for the second component, which is less coherent with the 3D structure, the projection gradient extends from the centromeres across the nucleus.

Supplementary Notes

Supplementary Note 1: Tethered conformation capture procedure

Day 1 Parasite pellets were thawed on ice in 550 μl Hi-C lysis buffer (25 mM Tris-HCl at pH 8.0, 10 mM NaCl, 2 mM AEBSF, Roche Complete Mini EDTA-free protease inhibitor cocktail [Roche, Basel, Switzerland], 0.25% Igepal CA-630) per 140 mg. Parasite membranes were disrupted by passing the lysate through a 26.5 gauge needle 15 times using a syringe. Samples were spun at 2,500 $\times$ g for 5 min at room temperature (RT). Pellets were washed twice with 1 ml ice-cold wash buffer (50 mM Tris-HCl at pH 8.0, 50 mM NaCl, 1 mM EDTA) and resuspended in the same buffer to a final volume of 250 μl. Samples were mixed with 95 μl 2% SDS to a final concentration of 0.5% and incubated at 55°C for 15 min. Suspensions were cooled down to RT before they were mixed with 105 μl 25 mM EZ-link Iodoacetyl-PEG2-Biotin (IPB) (Thermo Fisher Scientific, Waltham, MA, USA) to biotinylate proteins. After incubating for 1 h at RT while rotating, the SDS was neutralized by adding 1.3 ml 1$\times$ NEBuffer 2 (New England Biolabs [NEB], Ipswich, MA, USA). Samples were mixed with 225 μl 10% Triton X-100 to a final concentration of 1% and incubated for 10 min on ice, followed by 10 min at 37°C. Five μl 1 M DTT, 100 μl 10$\times$ NEBuffer 2, 415 μl water and 35 μl MboI restriction enzyme (NEB) (25 units/μl) was added to digest the DNA overnight at 37°C in a total volume of 2,530 μl.

Day 2 After digestion, samples were loaded into a Slide-A-Lyzer Dialysis Cassette G2 (Thermo Fisher Scientific) and dialyzed for 4 h at RT against 1 L of dialysis buffer (10 mM Tris-HCl at pH 8.0, 1 mM EDTA) to eliminate excess IPB remaining from the biotinylation step. Dialysis buffer was renewed after 3 h. Four hundred μl MyOne Streptavidin T1 beads (Life Technologies, Carlsbad, CA, USA) were washed 3 times with PBS + 0.01% Tween-20 (PBST) and beads were resuspended in 2 ml PBST. Dialyzed samples were divided into 5 equal aliquots of 500 μl in 1.7 ml prelubricated microcentrifuge tubes (Corning, Corning, NY, USA). Four hundred μl beads were added to each tube and samples were incubated for 30 min at RT while rotating. To prevent interference of unbound streptavidin on the beads with later steps (adding biotinylated dCTP) 5 μl neutralized IPB was added to each tube. IPB was neutralized by adding an equimolar amount of 2-mercaptoethanol. Samples were incubated for an additional 15 min at RT while rotating. Not biotinylated chromatin and not cross-linked DNA was removed by washing the magnetic T1 beads once with 600 μl PBST and once with 600 μl wash buffer (10 mM Tris-HCl at pH 8.0, 50 mM NaCl, 0.4% Triton X-100). Beads were resuspended in 100 μl of the same wash buffer. MboI generated 5$'$ overhangs were

filled in by adding 63 μl water, 1 μl 1 M MgCl, 10 μl 10× NEBuffer 2, 0.7 μl 10 mM dATP, 0.7 μl 10 mM dTTP, 0.7 μl 10 mM 2'-Deoxyguanosine-5'-O-(1-thiotriphosphate), sodium salt, Sp-isomer (Axxora, San Diego, CA, USA), 15 μl 0.4 mM Biotin-14-dCTP (Life Technologies), 4 μl 10% Triton X-100 and 5 μl 5U/μl DNA Polymerase I, Large (Klenow) Fragment (NEB). Samples were incubated for 40 min at RT while rotating. Reaction was stopped by adding 5 μl 0.5 M EDTA to the suspension. After 2 min of incubation at RT while rotating, beads were washed twice with 600 μl buffer (50 mM Tris-HCl at pH 7.4, 0.4% Triton X-100, 0.1 mM EDTA) and resuspended in 500 μl of the same buffer. Each sample was transferred into a 15 ml centrifuge tube. For blunt-end ligation under dilute conditions 500 μl sample was mixed with 4 ml water, 250 μl 10× Ligase Buffer (NEB), 100 μl 1 M Tris-HCl at pH 7.4, 90 μl 20% Triton X-100, 50 μl 100× BSA and 2 μl 2,000 U/μl T4 DNA Ligase (NEB), and incubated overnight at 16°C.

Day 3 The ligation reaction was stopped by adding 200 μl 0.5 M EDTA to each of the five 15 ml tubes. The magnetic T1 beads were collected on the wall of the tube using a magnet and the solution was aspirated out of the tube. The beads were resuspended in 400 μl extraction buffer (50 mM Tris-HCl at pH 8.0, 0.2% SDS, 1 mM EDTA, 500 mM NaCl) and the mix was transferred into a new microcentrifuge tube. Samples were treated with 5 μl RNase A (20 mg/ml) (Life Technologies) for 45 min at 37°C and with 20 μl Proteinase K (20 mg/ml) (NEB) overnight at 45°C.

Day 4 An additional 5 μl Proteinase K was added and samples were incubated for another 2 h at 45°C. Beads were collected on the wall of the tube and DNA was extracted from the supernatant twice with an equal volume of phenol:chloroform:isoamyl alcohol (25:24:1) and once with an equal volume of chloroform. The aqueous phase was mixed with sodium chloride and glycogen to a final concentration of 200 mM and 25 μg/ml, respectively. DNA was precipitated by adding 900 μl ice-cold 200 proof pure ethanol and incubation at -20°C overnight or at -80°C for $>$ 1 h. Precipitated DNA was pelleted by centrifugation at 16,100 × g for 30 min at 4°C. Pellets were washed with ice-cold 80% ethanol, spun down at 16,100 × g for 15 min at 4°C and resuspended in 20 μl 10 mM Tris-HCl at pH 8.0.

Day 5 Two to five μg purified DNA was treated with Exonuclease III (NEB) (60 units per μg DNA) in 120 μl 1× NEBuffer 1 for one h at 37°C. The reaction was ended by adding 2.7 μl 0.5 M EDTA and 2.7 μl 5 M NaCl, and subsequent incubation at 70°C for 20 min. DNA was transferred into TPX microtubes (Diagenode, Denville, NJ, USA) and sonicated using a Bioruptor UCD-200 (Diagenode) at high intensity for 30 min using

30 sec on, 30 sec off cycles. Agencourt AMPure XP beads (Beckman Coulter, Brea, CA, USA) were used to purify DNA, which was eluted in 50 μl water.

Day 6 All amounts mentioned for subsequent end-repair and adding of A-overhangs are per μg of DNA used as input at the start of Day 5. DNA ends were repaired by treating the DNA with 1 U of DNA Polymerase I, Large (Klenow) Fragment (NEB), 3 U of T4 DNA Polymerase (NEB), 10 U of T4 Polynucleotide Kinase (NEB) in 100 μl 1$\times$ T4 DNA Ligase Buffer (NEB) with 0.4 mM of dNTPs for 30 min at 20°C. Importantly, T4 DNA Polymerase and not T4 DNA Ligase should be used for end-repair (Reza Kalhor, personal communication). This was apparently written incorrectly in the original TCC protocol Kalhor et al. [2011]. DNA was purified using magnetic beads and eluted in 40 μl water. A-overhangs were added by treating the DNA with 3 U of Klenow Fragment ($3' \rightarrow 5'$ exo–) (NEB) in 50 μl 1$\times$ NEBuffer 2 with 0.2 mM dATP for 30 min at 37°C. The reaction was ended by adding 1 μl of 0.5 M EDTA. Ten μl of MyOne Streptavidin C1 magnetic beads (Invitrogen) were washed twice with 500 μl 1$\times$ Bind & Wash (B&W) buffer (5 mM Tris-HCl at pH 7.4, 0.5 mM EDTA, 1 M NaCl) and resuspended in 50 μl 2$\times$ B&W buffer. The DNA sample and the C1 beads were mixed and incubated at RT for 30 min. The beads were washed once with 500 μl 1$\times$ B&W buffer with 0.1% Triton, once with 500 μl 10 mM Tris-HCl at pH 8.0 and were resuspended in 10 μl water.

The Encore NGS Multiplex System (Nugen, San Carlos, CA, USA) was used for adapter ligation and library preparation of the cross-linked and non-cross-linked trophozoite samples. Amplification conditions were 45 sec at 98°C, 5 cycles of 15 sec at 98°C, 30 sec at 55°C and 30 sec at 62°C, followed by 10 cycles of 15 sec at 98°C, 30 sec at 63°C and 30 sec at 72°C, and a final elongation of 5 min at 72°C. NEBNext Multiplex Oligos for Illumina (NEB) and NEBNext Library Prep Reagents Set (NEB) were used for adapter ligation and library preparation of the ring and schizont samples. Amplification conditions were 45 sec at 98°C, 8 cycles of 15 sec at 98°C, 30 sec at 55°C and 30 sec at 62°C, followed by 3 cycles of 15 sec at 98°C, 30 sec at 63°C and 30 sec at 72°C, and a final elongation of 5 min at 72°C. KAPA HiFi DNA Polymerase HotStart ReadyMix (Kapa Biosystems, Woburn, MA, USA) was used for all PCRs. DNA in the supernatant was purified with Agencourt AMPure XP beads. Library quantification was performed using a 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA). Libraries were subsequenctly sequenced on a HiSeq 2000 system (Illumina, San Diego, CA, USA) at the Institute for Integrative Genome Biology (University of California, Riverside, USA), generating 50 bp paired-end sequence reads.

Supplementary Note 2: Assigning statistical significance to normalized contact maps

We can describe our confidence estimation procedure as follows. Let N_{inter}, N_{intra} denote the total number of observed informative paired-end reads between inter and intrachromosomal locus pairs and M_{inter}, M_{intra} denote the number of such inter and intrachromosomal locus pairs, respectively. If we assume that an observed paired-end read is equally likely to come from any locus pair, then the null probability that the read comes from a specific locus pair is $p_{inter} = \frac{1}{M_{inter}}$ and $p_{intra} = \frac{1}{M_{intra}}$ for intra-chromosomal and interchromosomal pairs, respectively. We use a previously described iterative procedure [Imakaev et al., 2012] to estimate locus-specific biases and adjust the interchromosomal probability accordingly: $\bar{p}_{ij} = p_{inter} * B_i * B_j$, where B_i and B_j are the estimate bias terms.

For intrachromosomal locus pairs the assumption that each read is equally likely to come from any locus pair fails due to the significant effect of genomic distance on the contact probability. To account for this effect, we used a method that estimates the prior contact probability between two loci given their genomic distance by fitting a smooth spline and refining the underlying null distribution of contact probabilities [Ay et al., 2014a]. For intrachromosomal locus pair (ℓ_i, ℓ_j) with genomic distance d, this spline is used to estimate the contact probability $p_{intra}(d)$. Similar to the interchromosomal pairs, this probability is corrected for biases of each locus ℓ_i and ℓ_j resulting in $\bar{p}_{ij} = p_{intra}(d) * B_i * B_j$.

Once the corrected null probabilities $\bar{p}_{ij}$ are computed for each possible inter and intra-chromosomal locus pair, we computed the significance of observing k_{ij} informative reads between (ℓ_i, ℓ_j) among either $N = N_{inter}$ or $N = N_{intra}$ total reads, depending on the contact type. Dropping the subscripts from $\bar{p}_{ij}$ and k_{ij}, we calculated the significance as the p-value from the binomial distribution:

$$p(K \geq k) = \sum_{i=k}^{N} \Pr(K = i) \tag{A.1}$$

where

$$\Pr(K = k) = \binom{N}{k} \bar{p}^k \left(1 - \bar{p}\right)^{N-k}.$$

Finally, we corrected the combined collection of p-values for multiple testing by estimating, for a given p-value threshold, the proportion of false positive contacts with p-values below the threshold. This proportion is known as the *false discovery rate* (FDR), which can be estimated using standard methods [Benjamini and Hochberg, 1995].

Supplementary Note 3: DNA-FISH

DNA-FISH experiments were performed according to a recently published protocol [Contreras-Dominguez et al., 2010] with minor modifications. P. falciparum-infected erythrocytes were pelleted by centrifuging at 800 × g for 5 min at 4°C, with minimal braking (brake = 1). To lyse erythrocyte membranes, double sorbitol-synchronized ring and trophozoite stage parasites were treated with 5 volumes of 0.015% cold saponin in cold PBS on ice for 20 or 10 min, respectively. Parasites were spun down at 4,200 × g for 10 min at 4°C, with minimal braking, and washed up to 7 times (2,000 × g, 10 min, brake = 5) with cold PBS. Parasites were then resuspended 4% formaldehyde (in PBS at room temperature) and fixed on ice for 15 min. After this fixation, parasites were washed 2 times in cold PBS (4,200 × g, 1 min, maximum brake) and resuspended in cold PBS.

A monolayer of parasites was deposited within a 9 × 9 mm frame-seal slide chamber (Bio-Rad, Hercules, CA, USA) that was prepared on a standard microscopy slide, and slides were air-dried for 30 min at RT. The fixed, air-dried parasites were washed with PBS for 5 min at RT, treated with 0.1% Triton X-100 in PBS for 5 min at RT and washed twice with PBS for 5 min at RT. Hybridization solution (50% formamide, 10% dextran sulfate, 2 × SSPE, 250 μg/ml single-stranded DNA from salmon testes) containing the denatured (5 min at 95°C) probes was applied and slide chambers were covered with a coverslip. Slides were denatured at 80°C for 30 min followed by hybridization at 37°C overnight. After removal of the coverslip and the hybridization solution, slides were washed in 2 × SSC/50% formamide for 30 min at 37°C, followed by 1 × SSC for 10 min at 50°C, 2 × SSC for 10 min at 50°C and 4 × SSC for 10 min at 50°C. Parasites were equilibrated in M solution (100 mM maleic acid, 150 mM NaCl, 1% bovine serum albumin) set at neutral pH, for 5 min at RT in a humid chamber, protected from light. M solution was removed and replaced with M solution containing Avidin, NeutrAvidin, Rhodamine Red-X Conjugate (Life Technologies) (1:1,000) for detection of the biotin probes. Slides were incubated for 30 min at RT, in a humid chamber, protected from light, and subsequently washed 3 times in TNT solution (100 mM Tris-HCl at pH 7.5, 150 mM NaCl, 0.5% Tween 20) for 10 min at RT with agitation. Cells were stained with DAPI (0.5 μg/ml in TNT solution) for 2 - 3 seconds. Slides were then air-dried (protected from light) and mounted using gelvatol with 2.5% Dabco anti-fade (Sigma-Aldrich, St. Louis, MO, USA). Images were acquired using an Olympus BX40 epifluorescent microscope (Olympus, Center Valley, PA, USA).

Supplementary Note 4: Volume exclusion modeling

Tjong et al. [2012] show the budding yeast's dominant architectural features can be entirely explained by a simple volume exclusion model, modeling chromatin as a random flexible polymer with few biologically motivated architectural constraints. Following their methodology, we computed a population of 5000 structures for the budding yeast using the same sets of constraints, and we successfully recovered high correlation between the contact maps generated from the population of structures and the observed Hi-C matrix (Supplementary Fig. 14(a)).

Even though the row-based correlation has been used as a measure of consistency between two contact maps [Tjong et al., 2012, Imakaev et al., 2012], we hypothesized that this measure may be dominated by the strong diagonal trend of contact maps and, hence, may not capture non-random similarity between two contact matrices. To test this hypothesis, we generated an *expected contact matrix* by setting each interchromosomal contact count to the expected contact count for its genomic distance, as defined in Methods. We obtained an even higher correlation between the observed Hi-C matrix and this structureless expectation matrix (Supplementary Fig. 14(b)).

To account for this problem, we developed a new scoring measure, the *normalized* row-based Pearson correlation, which replaces each count value with its ratio to an expected count in the correlation computation (Methods). Supplementary Fig. 14(b) demonstrates that the normalized row-based Pearson correlation is more effective for comparing contact maps: indeed, the correlations between structureless matrices (marked as *expected*) and observed Hi-C matrices are close to zero, while the correlations between the simulated (*VE*) and observed Hi-C contact matrices are conserved.

Appendix B

Supplementary information for Varoquaux et al. [2015]

FIGURE 1: **Error on centromere calls for *P. falciparum* on raw and normalized contact counts (40 kb)**

A. ring stage **B.** trophozoite stage **C.** schizont stage

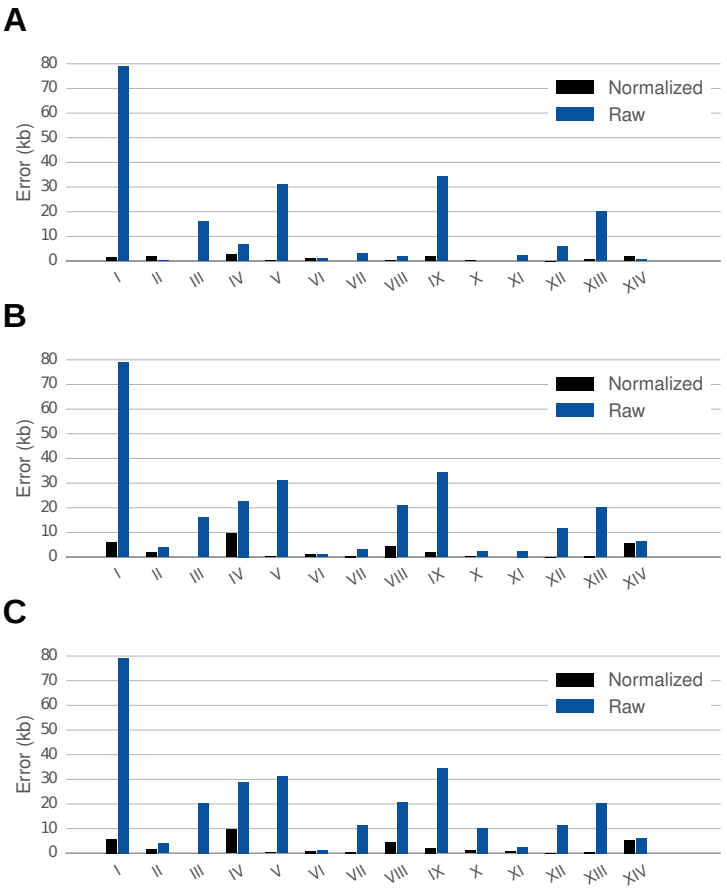


FIGURE 2: **Error on centromere calls for *S. cerevisiae* at different resolutions (10 kb, 20 kb, 40 kb)**

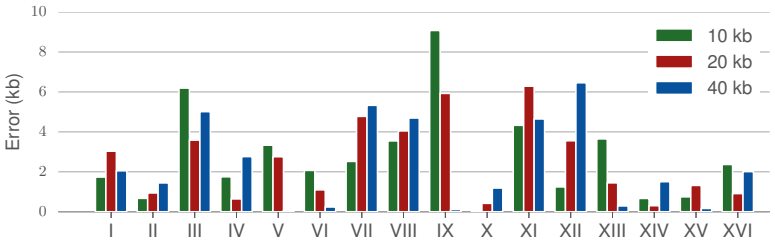


FIGURE 3: **Error on centromere calls for *P. falciparum* at different resolutions (10 kb, 20 kb, 40 kb)**

A. ring stage **B.** trophozoite stage **C.** schizont stage

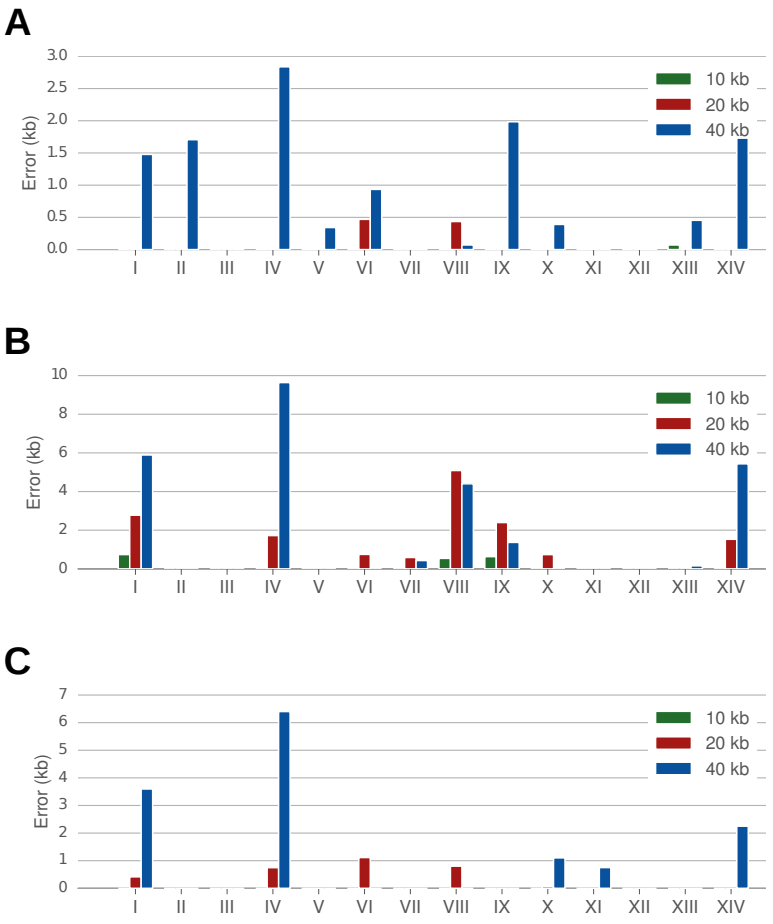


TABLE 1: **Centromere calls for *S. cerevisiae*, ground truth and errors**

Chromosome	Ground truth	10 kb		20 kb		40 kb	
		Call	Error	Call	Error	Call	Error
I	151 584 - 151 584	153 319	1735	154 614	3030	153 633	2049
II	238 325 - 238 325	238 994	669	237 386	939	236 883	1442
III	114 499 - 114 499	108 309	6190	110 914	3585	109 488	5011
IV	449 819 - 449 819	451 567	1748	450 459	640	452 579	2760
V	152 103 - 152 103	155 434	3331	149 350	2753	152 162	59
VI	148 622 - 148 622	150 691	2069	149 718	1096	148 387	235
VII	497 042 - 497 042	499 561	2519	501 816	4774	502 369	5327
VIII	105 698 - 105 698	102 152	3546	101 652	4046	101 007	4691
IX	355 742 - 355 742	364 818	9076	361 667	5925	355 631	111
X	436 418 - 436 418	436 467	49	435 999	419	437 603	1185
XI	439 889 - 439 889	444 216	4327	446 174	6285	444 533	4644
XII	150 946 - 150 946	149 704	1242	147 393	3553	144 489	6457
XIII	268 149 - 268 149	264 502	3647	266 704	1445	267 860	289
XIV	628 877 - 628 877	629 542	665	629 178	301	627 374	1503
XV	326 703 - 326 703	327 448	745	328 019	1316	326 866	163
XVI	556 070 - 556 070	553 705	2365	555 162	908	554 062	2008

TABLE 2: **Centromere calls for *P. falciparum* (ring stage), ground truth and errors**

Chromosome	Ground truth	10 kb		20 kb		40 kb	
		Call	Error	Call	Error	Call	Error
I	456 871 - 461 511	458 108	0	457 710	0	455 394	1477
II	446 771 - 450 941	448 688	0	448 953	0	452 647	1706
III	597 014 - 601 275	599 187	0	597 486	0	599 779	0
IV	641 019 - 645 339	644 178	0	644 931	0	648 176	2837
V	454 543 - 458 793	456 329	0	455 929	0	454 201	342
VI	477 756 - 482 016	480 602	0	477 287	469	482 950	934
VII	808 365 - 812 875	811 744	0	810 236	0	812 304	0
VIII	297 895 - 302 515	299 983	0	297 460	435	297 824	71
IX	1 241 081 - 1 245 451	1 242 788	0	1 242 570	0	1 247 435	1984
X	935 682 - 937 823	937 162	0	936 247	0	938 213	390
XI	830 782 - 835 432	832 728	0	832 858	0	834 051	0
XII	1 281 521 - 1 285 941	1 284 567	0	1 285 214	0	1 285 943	2
XIII	1 167 070 - 1 171 720	1 166 999	71	1 168 375	0	1 172 174	454
XIV	1 070 909 - 1 075 369	1 072 595	0	1 072 131	0	1 069 179	1730

TABLE 3: **Centromere calls for *P. falciparum* (trophozoite stage), ground truth and errors**

Chromosome	Ground truth	10 kb		20 kb		40 kb	
		Call	Error	Call	Error	Call	Error
I	456 871 - 461 511	456 134	737	454 096	2775	450 980	5891
II	446 771 - 450 941	448 623	0	447 562	0	448 953	0
III	597 014 - 601 275	598 035	0	597 348	0	597 426	0
IV	641 019 - 645 339	645 248	0	647 059	1720	654 977	9638
V	454 543 - 458 793	455 899	0	455 305	0	457 291	0
VI	477 756 - 482 016	480 552	0	477 010	746	480 417	0
VII	808 365 - 812 875	810 348	0	807 779	586	807 937	428
VIII	297 895 - 302 515	297 355	540	292 808	5087	293 495	4400
IX	1 241 081 - 1 245 451	1 240 449	632	1 238 687	2394	1 239 714	1367
X	935 682 - 937 823	936 765	0	938 559	736	937 531	0
XI	830 782 - 835 432	832 425	0	833 938	0	833 994	0
XII	1 281 521 - 1 285 941	1 284 634	0	1 284 403	0	1 282 674	0
XIII	1 167 070 - 1 171 720	1 168 647	0	1 168 225	0	1 166 916	154
XIV	1 070 909 - 1 075 369	1 071 170	0	1 069 381	1528	1 065 476	5433

TABLE 4: **Centromere calls for *P. falciparum* (schizont stage), ground truth and errors**

Chromosome	Ground truth	10 kb		20 kb		40 kb	
		Call	Error	Call	Error	Call	Error
I	456 871 - 461 511	458 269	0	456 457	414	453 275	3596
II	446 771 - 450 941	448 913	0	448 374	0	450 640	0
III	597 014 - 601 275	599 091	0	597 694	0	598 616	0
IV	641 019 - 645 339	644 640	0	646 085	746	651 741	6402
V	454 543 - 458 793	455 379	0	454 880	0	455 537	0
VI	477 756 - 482 016	479 786	0	476 643	1113	479 443	0
VII	808 365 - 812 875	810 510	0	809 034	0	810 833	0
VIII	297 895 - 302 515	299 463	0	297 094	801	299 394	0
IX	1 241 081 - 1 245 451	1 242 617	0	1 242 663	0	1 244 982	0
X	935 682 - 937 823	936 637	0	935 616	66	934 581	1101
XI	830 782 - 835 432	832 307	0	831 740	0	830 033	749
XII	1 281 521 - 1 285 941	1 284 123	0	1 284 309	0	1 283 900	0
XIII	1 167 070 - 1 171 720	1 168 687	0	1 169 725	0	1 171 143	0
XIV	1 070 909 - 1 075 369	1 072 384	0	1 071 648	0	1 068 660	2249

TABLE 5: **Centromere calls for *A. thaliana*, annotation units and errors**

Chromosome	**Ground truth**	**40 kb**	
		Call	**Error**
1	15 086 046 - 15 087 045	15 047 165	39 380
2	3 607 930 - 3 608 929	3 841 087	232 657
3	14 132 042 - 14 208 952	14 177 317	6820
4	3 956 022 - 3 957 021	3 754 384	202 137
5	11 725 025 - 11 726 024	12 055 189	329 665

FIGURE 4: **Centurion vs Marie-Nelly et al. [2014b]'s method**

Centurion and Marie-Nelly et al. [2014b]'s whole pipeline centromere calls error on 40 kb contact counts matrices. Marie-Nelly et al. [2014b]'s method fails to prelocalize properly centromeres.

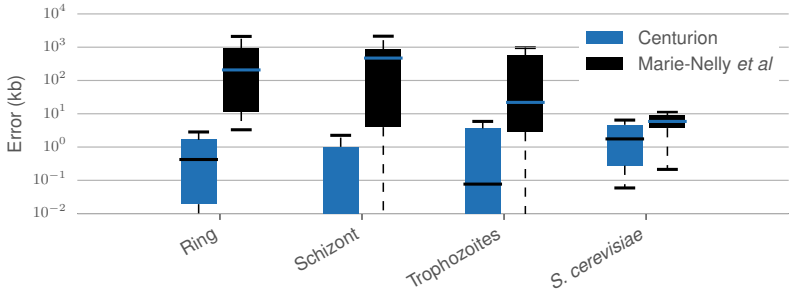


FIGURE 5: **Pearson correlation matrix of *P. falciparum*'s chr XII.**

Dashed black line indicates the centromere. Because var genes strongly colocalize, the typical X-shape found in *S. cerevisiae*'s Pearson correlation matrices completely disappears, consequently causing Marie-Nelly et al. [2014b]'s first step to fail to prelocalize centromeres.

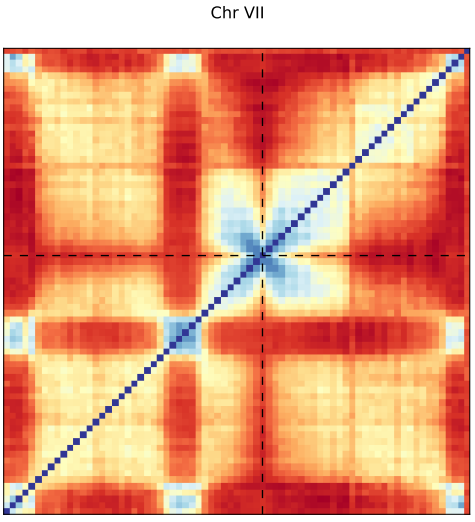


TABLE 6: **M-3D multi-sample statistics for each organism's contact counts matrices (20 kb)**

For each contact count matrix, we compute several statistics: (1) the average number of contact counts off-diagonal, (2) the percentage of non-zero element off-diagonal, (3) the average number of *trans* contact counts, (4) the percentage of non-zero *trans* contact counts

Organism	**Number of chrom**	**Average contact counts per bin**	**sparsity**	**Average *trans* contact counts per bin**	***trans* sparsity**
Acinetobacter sp. ADP1	1	91.11	99.44	-	-
Vibrio fischeri ES114	2	87.14	99.53	57.23	100.00
Methanococcus maripaludis	2	82.29	96.47	0.00	0.00
Burkholderia thailandensis E264	2	37.52	99.70	27.92	100.00
Escherichia coli str. K-12 substr. DH10B	1	35.79	90.39	-	-
Flavobacterium johnsoniae UW101	1	30.88	99.55	-	-
Rhodopseudomonas palustris CGA009	1	30.58	99.61	-	-
Bacillus subtilis subsp. subtilis str. 168	1	13.53	99.21	-	-
Schizosaccharomyces pombe	3	0.91	32.15	0.35	24.38
Pichia pastoris GS115	4	0.72	32.47	0.39	28.06
Zygosaccharomyces rouxii strain CBS732	7	0.52	24.97	0.28	20.80
Kluyveromyces thermotolerans strain CBS6340	8	0.48	23.24	0.27	20.25
Saccharomyces cerevisiae S288c	16	0.28	16.01	0.18	14.32
Pseudomonas fluorescens Pf0-1	1	0.02	1.35	-	-

TABLE 7: **M-Y multi-sample statistics for each organism's contact counts matrices (20 kb)**

For each contact count matrix, we compute several statistics: (1) the average number of contact counts off-diagonal, (2) the percentage of non-zero element off-diagonal, (3) the average number of *trans* contact counts, (4) the percentage of non-zero *trans* contact counts

Organism	**Number of chrom**	**Average contact counts per bin**	**sparsity**	**Average *trans* contact counts per bin**	***trans* sparsity**
Kluyveromyces lactis	6	3.20	74.45	1.85	72.64
Lachancea kluyveri	8	2.83	67.70	1.49	65.44
Lachancea waltii	8	2.38	60.39	1.22	57.89
Kluyveromyces wickerhamii	7	1.43	45.38	0.66	41.49
Scheffersomyces stipitis	8	1.38	45.08	0.72	42.01
Saccharomyces mikatae	16	1.35	48.58	0.82	46.65
Saccharomyces bayanus	16	0.94	35.49	0.51	33.19
Saccharomyces paradoxus	16	0.69	26.57	0.35	24.46
Pichia pastoris GS115	4	0.37	25.48	0.28	23.71
Eremothecium gossypii	7	0.13	7.06	0.06	4.82
Saccharomyces kudriavzevii	16	0.11	5.98	0.05	4.73
Saccharomyces cerevisiae SK1	16	0.02	0.92	0.01	0.65
Saccharomyces cerevisiae S288c	16	0.00	0.11	0.00	0.07

FIGURE 6: **Errors on metagenomic sample**

Box plots indicating the error (in kb) for each chromosome in Centurion's centromere calls for eight yeasts with known centromere coordinates from the combined metagenomic Hi-C samples M-3D and M-Y on the 40 kb contact count matrices.

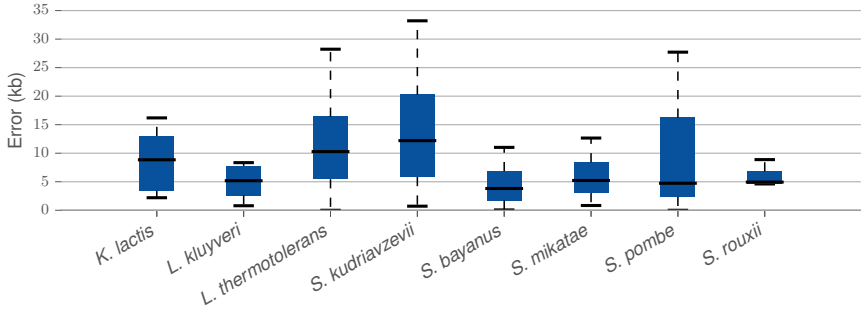


FIGURE 7: **Centromere calls for *K. lactis***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

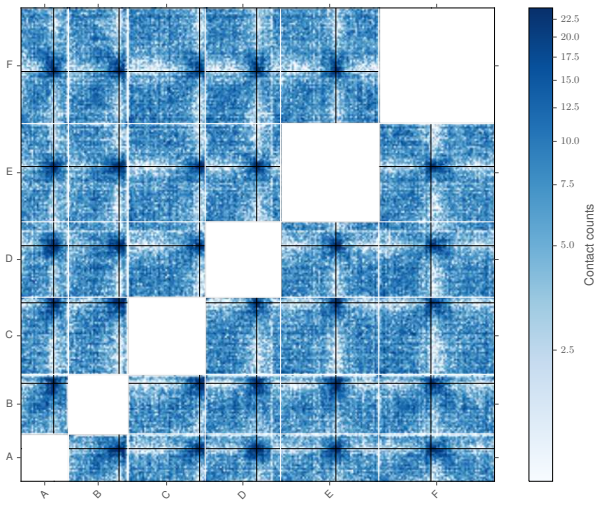


TABLE 8: ***K. lactis* centromere calls, ground truth and errors**

Chromosome	Ground truth	20 kb		40 kb	
		Call	Error	Call	Error
A	760 404 - 760 598	747 703	12 701	744 213	16 191
B	1 168 861 - 1 169 058	1 156 659	12 202	1 155 652	13 209
C	1 638 151 - 1 638 347	1 633 885	4266	1 632 850	5301
D	1 187 303 - 1 187 500	1 180 157	7146	1 174 906	12 397
E	1 263 806 - 1 264 001	1 264 257	256	1 260 994	2812
F	1 187 015 - 1 187 211	1 186 655	360	1 189 411	2200

FIGURE 8: **Centromere calls for *L. kluyveri***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

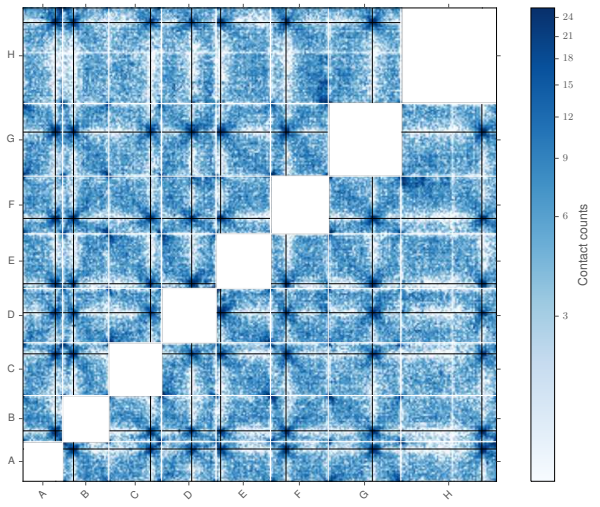


TABLE 9: ***L. kluyveri* centromere calls, ground truth and errors**

Chromosome	Ground truth	20 kb		40 kb	
		Call	Error	Call	Error
A	777 082 - 777 277	784 872	7595	782 908	5631
B	272 171 - 272 366	268 270	3901	263 855	8316
C	1 009 526 - 1 009 330	1 008 047	1479	1 011 248	1918
D	737 092 - 737 289	729 108	7984	728 743	8349
E	108 420 - 108 235	113 926	5691	111 117	2882
F	383 306 - 383 110	378 812	4494	375 717	7589
G	1 064 569 - 1 064 371	1 068 157	3786	1 069 100	4729
H	1 963 796 - 1 963 599	1 963 570	226	1 964 393	794

FIGURE 9: **Centromere calls for *S. bayanus***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

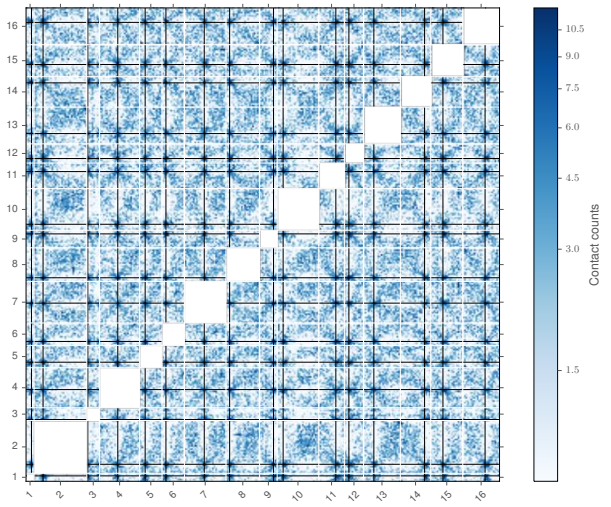


TABLE 10: ***S. bayanus*** **centromere calls, partial ground truth and errors**

Chromosome	**Ground truth**	**20 kb**		**40 kb**	
		Call	**Error**	**Call**	**Error**
1	128 493 - 128 610	133 793	5183	135 493	6883
2	-	227 362	-	225 392	-
3	24 732 - 24 851	24 374	358	24 672	60
4	447 057 - 447 177	449 935	2758	453 979	6802
5	127 728 - 127 885	131 258	3373	138 605	10 720
6	-	107 819	-	97 831	-
7	-	490 402	-	496 199	-
8	102 036 - 102 155	101 405	631	91 014	11 022
9	342 506 - 342 624	345 821	3197	348 290	5666
10	-	151 519	-	148 837	-
11	424 482 - 424 587	424 027	455	426 525	1938
12	-	113 924	-	113 109	-
13	258 015 - 258 136	253 649	4366	257 327	688
14	609 003 - 609 122	604 920	4083	605 657	3346
15	301 003 - 301 123	305 724	4601	304 941	3818
16	560 121 - 560 240	556 888	3233	558 583	1538

FIGURE 10: **Centromere calls for *S. mikatae***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

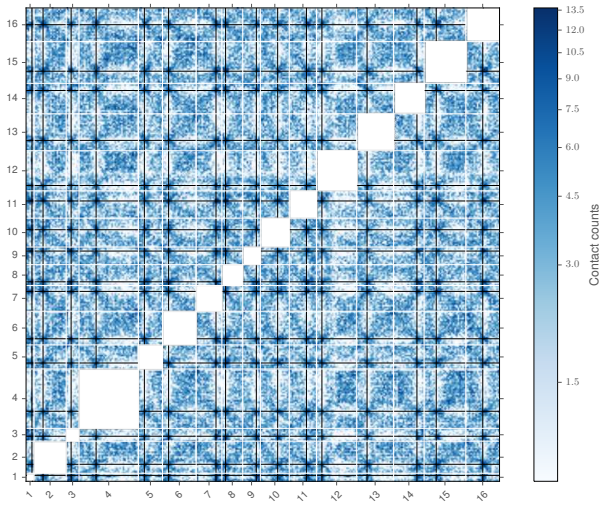


TABLE 11: ***S. mikatae* centromere calls, ground truth and errors**

Chromosome	Ground truth	20 kb		40 kb	
		Call	Error	Call	Error
1	134 639 - 134 759	139 227	4468	147 416	12 657
2	225 947 - 226 062	226 196	134	223 853	2094
3	112 549 - 112 682	123 596	10 914	123 147	10 465
4	428 996 - 429 115	425 527	3469	425 807	3189
5	155 935 - 156 053	152 599	3336	148 911	7024
6	155 876 - 155 995	156 306	311	152 732	3144
7	488 820 - 488 935	491 034	2099	493 846	4911
8	84 409 - 84 527	90 055	5528	92 846	8319
9	331 647 - 331 767	335 329	3562	339 195	7428
10	433 451 - 433 569	429 847	3604	430 766	2685
11	426 605 - 426 749	430 382	3633	428 157	1408
12	137 938 - 138 058	134 774	3164	133 389	4549
13	259 194 - 259 326	259 361	35	258 359	835
14	587 008 - 587 124	580 201	6807	581 480	5528
15	293 868 - 293 999	294 747	748	302 652	8653
16	432 872 - 432 990	419 838	13 034	422 199	10 673

FIGURE 11: **Centromere calls for *S. kudriavzevii***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

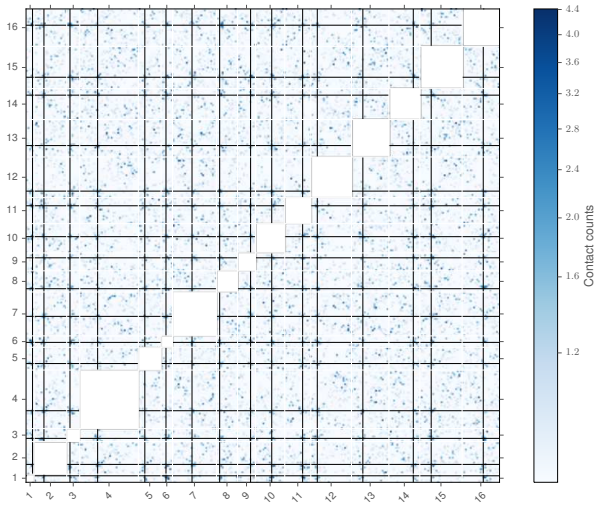


TABLE 12: ***S. kudriavzevii* centromere calls, ground truth and errors**

Chromosome	Ground truth	20 kb		40 kb	
		Call	Error	Call	Error
1	126 503 - 126 621	139 821	13 200	159 827	33 206
2	218 375 - 218 494	280 091	61 597	239 855	21 361
3	93 380 - 93 501	79 993	13 387	80 093	13 287
4	441 296 - 441 418	440 008	1288	439 893	1403
5	148 755 - 148 877	220 186	71 309	159 993	11 116
6	144 259 - 144 379	140 074	4185	120 144	24 115
7	499 997 - 500 118	500 033	0	480 114	19 883
8	87 050 - 87 170	80 118	6932	80 113	6937
9	326 489 - 326 613	320 133	6356	320 091	6398
10	403 891 - 404 009	399 992	3899	400 093	3798
11	421 054 - 421 176	420 045	1009	439 781	18 605
12	142 068 - 142 189	139 973	2095	159 948	17 759
13	253 924 - 254 043	240 052	13 872	279 914	25 871
14	595 631 - 595 753	599 954	4201	599 937	4184
15	286 560 - 286 681	279 996	6564	280 098	6462
16	520 694 - 520 812	519 870	824	519 977	717

FIGURE 12: **Centromere calls for *L. thermotolerans***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

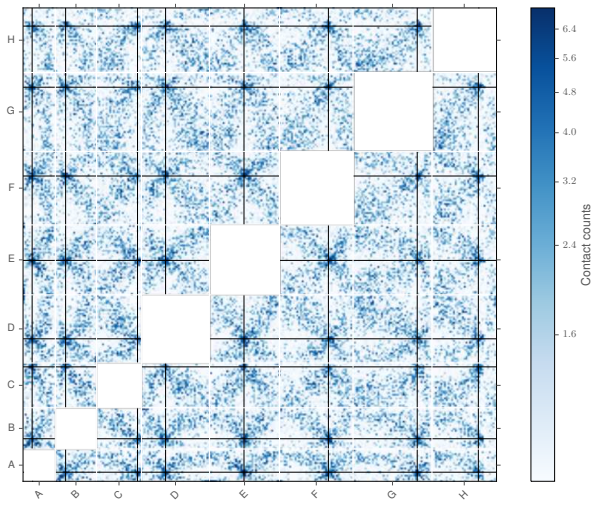


TABLE 13: ***L. thermotolerans*** **centromere calls, ground truth and errors**

Chromosome	**Ground truth**	**20 kb**		**40 kb**	
		Call	**Error**	**Call**	**Error**
A	186 515 - 186 379	201 225	14 846	202 706	16 327
B	238 312 - 238 187	235 631	2681	229 282	9030
C	912 837 - 912 964	920 560	7596	912 875	0
D	555 337 - 555 463	553 006	2331	538 813	16 524
E	761 047 - 760 921	767 727	6806	767 725	6804
F	1 078 717 - 1 078 842	1 094 222	15 380	1 090 376	11 534
G	1 432 769 - 1 432 902	1 432 357	412	1 431 366	1403
H	1 062 917 - 1 063 043	1 040 425	22 492	1 034 688	28 229

FIGURE 13: **Centromere calls for *S. pombe***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

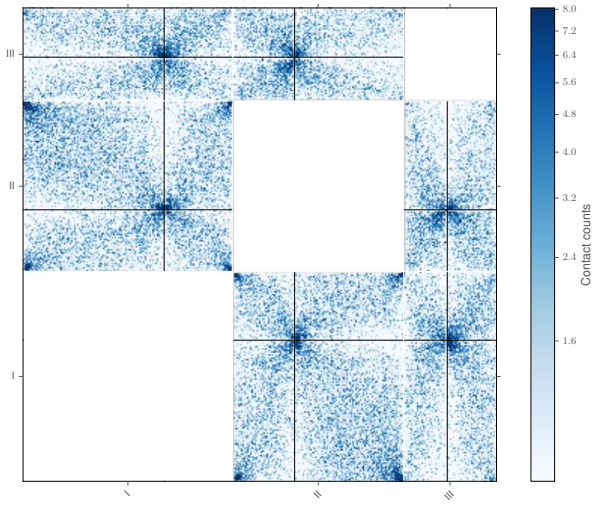


TABLE 14: ***S. pombe*** **centromere calls, ground truth and errors**

Chromosome	**Ground truth**	**20 kb**		**40 kb**	
		Call	**Error**	**Call**	**Error**
I	3 753 687 - 3 789 421	3 764 436	0	3 767 270	0
II	1 602 264 - 1 644 747	1 619 912	0	1 649 483	4736
III	1 070 904 - 1 137 003	1 121 329	0	1 164 716	27 713

FIGURE 14: **Centromere calls for *Z. rouxii***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

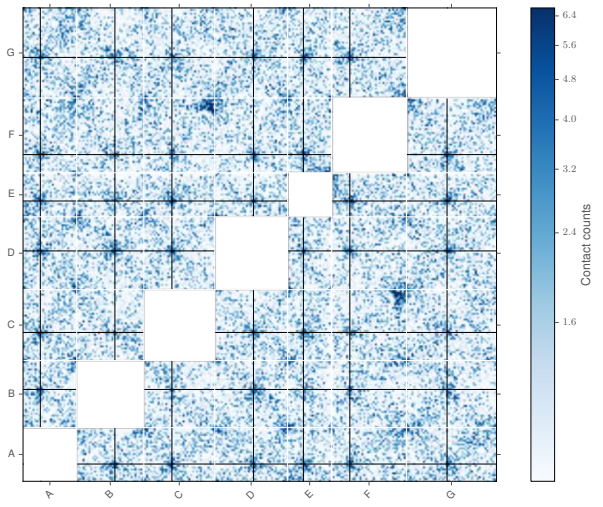


TABLE 15: ***Z. rouxii*** **centromere calls, ground truth and errors**

Chromosome	**Ground truth**	**20 kb**		**40 kb**	
		Call	**Error**	**Call**	**Error**
A	369 077 - 369 243	353 671	15 406	360 198	8879
B	788 730 - 788 896	782 871	5859	796 526	7630
C	581 961 - 581 795	582 298	503	586 407	4612
D	807 719 - 807 885	804 544	3175	808 613	728
E	335 012 - 334 844	333 919	1093	330 354	4658
F	372 701 - 372 867	376 542	3675	378 835	5968
G	852 551 - 852 385	841 169	11 382	847 614	4937

FIGURE 15: **Centromere calls for *P. pastoris***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

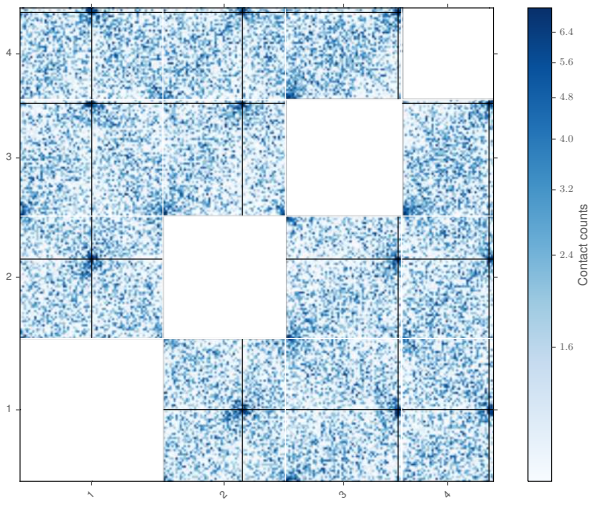


TABLE 16: ***P. pastoris* de novo centromere calls**

Chromosome	20 kb call	40 kb call
1	1 408 908	1 404 605
2	1 556 231	1 556 450
3	2 226 823	2 209 846
4	1 719 280	1 712 207

FIGURE 16: **Centromere calls for *E. gossypii***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

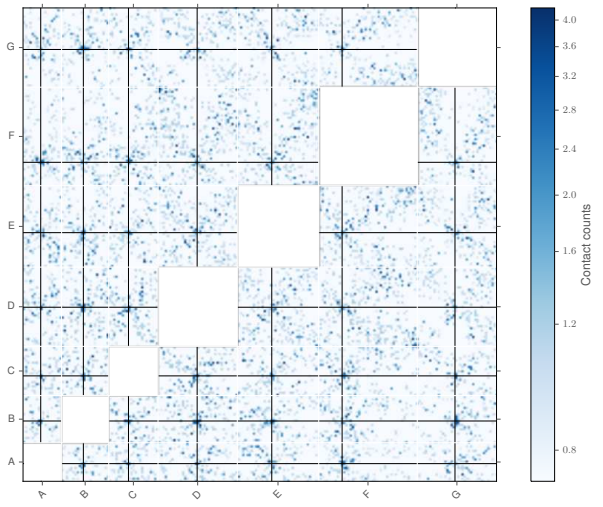


TABLE 17: ***E. gossypii* de novo centromere calls**

Chromosome	20 kb call	40 kb call
A	338 620	329 920
B	399 593	406 065
C	357 805	368 603
D	717 357	730 541
E	601 683	643 434
F	491 379	436 967
G	718 695	704 238

FIGURE 17: **Centromere calls for *K. wickerhamii***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

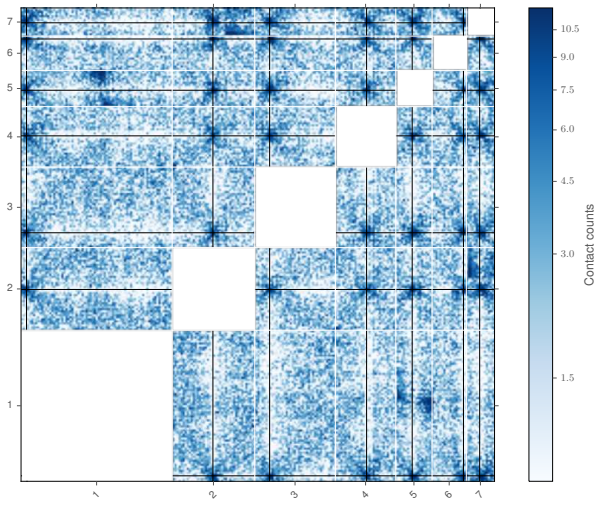


TABLE 18: ***K. wickerhamii* de novo centromere calls**

Chromosome	20 kb call	40 kb call
1	107 436	108 558
2	807 861	809 232
3	290 904	295 270
4	618 467	620 001
5	323 875	325 408
6	622 741	623 503
7	266 146	264 963

FIGURE 18: **Centromere calls for *L. waltii***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

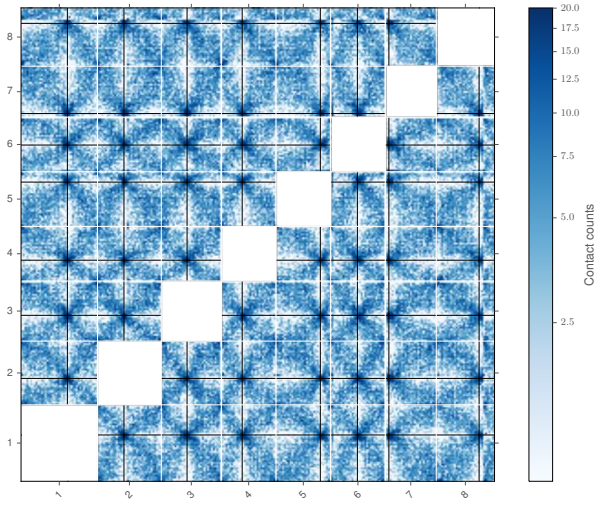


TABLE 19: ***L. waltii* de novo centromere calls**

Chromosome	20 kb call	40 kb call
1	1 028 089	1 017 632
2	587 954	580 429
3	565 659	566 391
4	454 852	457 549
5	971 551	973 444
6	589 260	587 604
7	80 178	79 371
8	935 869	941 058

FIGURE 19: **Centromere calls for *S. paradoxus***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

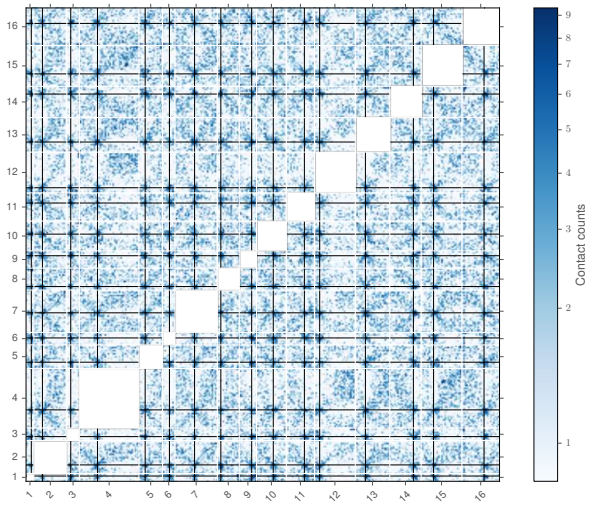


TABLE 20: ***S. paradoxus* de novo centromere calls**

Chromosome	20 kb call	40 kb call
1	138 267	134 145
2	222 275	217 965
3	100 865	100 219
4	458 330	462 705
5	154 493	159 191
6	178 545	178 292
7	494 545	499 430
8	87 986	89 589
9	317 389	316 309
10	414 694	427 637
11	455 761	461 330
12	128 761	131 623
13	256 588	258 273
14	601 000	602 387
15	316 026	317 082
16	564 137	564 849

FIGURE 20: **Centromere calls for *S. stipitis***

Smoothed *trans* contact counts (with $\sigma = 40$ kb) overlaid with Centurion's centromere calls (black line). White lines represent chromosome boundaries.

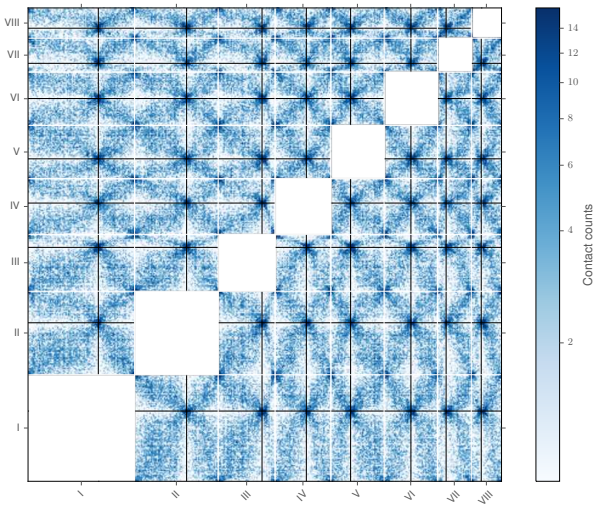
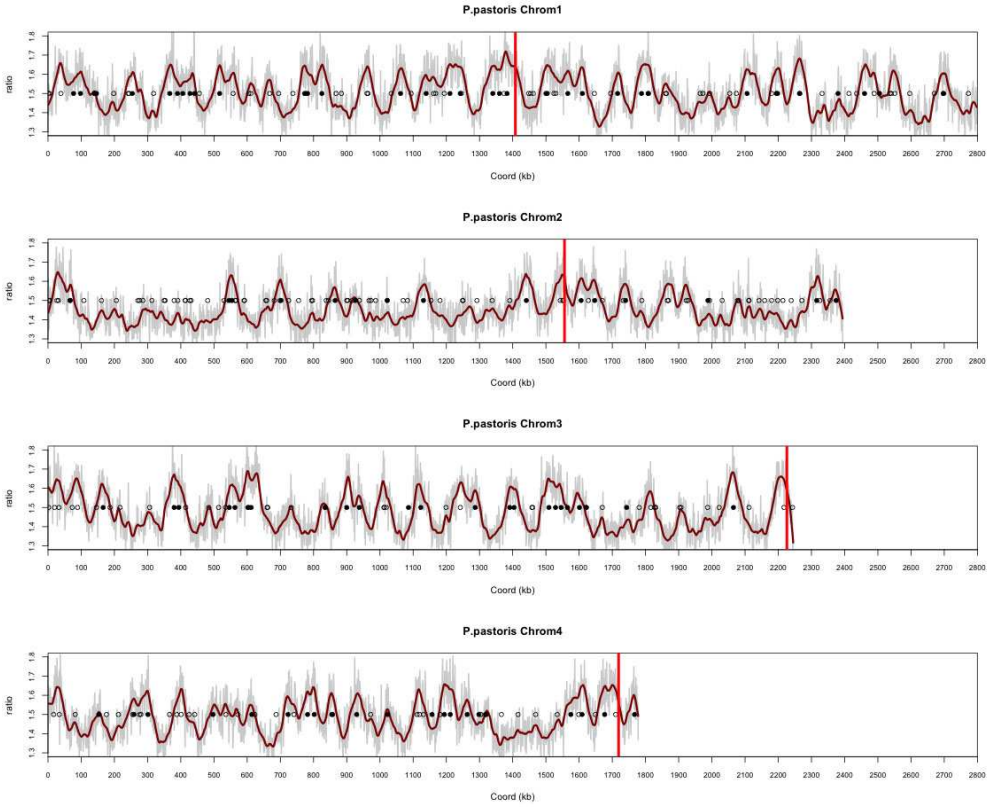


TABLE 21: ***S. stipitis* de novo centromere calls**

Chromosome	20 kb call	40 kb call
I	2 309 980	2 320 851
II	1 708 033	1 717 536
III	1 451 523	1 448 891
IV	1 039 527	1 032 779
V	655 745	654 571
VI	893 268	886 606
VII	279 875	289 537
VIII	325 140	330 100

FIGURE 21: **Replication timing profile across the *P. pastoris* genome**

Adapted from Figure 4 and Supplementary Figure 6 in Liachko et al. [2014]. The curve represents the smoothed copy number ratio of genomic DNA in cells undergoing S phase versus cells in G1 phase. Peaks correspond to positions of early replication (replication origins) and valleys represent late replicating regions (replication termini). Circles represent potential replication initiation sites. The positions of centromeres predicted by Centurion are indicated as red lines.



Supplementary Notes

Initializing the optimization

??

The optimization problem being non convex, the local minimum found by the algorithms depends on the starting point. We therefore implemented heuristics to initialize the optimization with several sets of centromere positions. Our implementation of Centurion also allows the user to specify the starting point (*ie* the rough centromere location).

For each chromosome, the centromeric regions are expected to be enriched in *trans* contact counts. We thus seek a few local maxima in the marginalized *trans* contact count profile $p(i) = \sum_{i,j|\mathcal{B}(i)\neq\mathcal{B}(j)} c_{ij}$ for each chromosome. In order to select only k

candidates per chromosome, we smooth the contact counts profile p with a Gaussian filter of parameter σ, setting σ such that there are k peaks in the profile. We consequently obtain a set of k centromere candidates per chromosome, and thus can initialize the optimization with all possible combination of these candidates.

To reduce computation time, we implemented a set of heuristics to decrease the number of candidates. First, note that the higher the contact count enrichment peak is, the more likely a candidate is to be the in the centromeric region. Second, remember that we attempt to jointly optimize centromeres location: we optimize L variables at once, L being the number of chromosomes, and each variable corresponding to a chromosome position. To reduce the number of candidates per chromosome, we first compute a baseline, by performing the optimization using as starting point the set of most likely candidate for each of our L chromosomes (the candidate with the highest peak for each of the chromosomes). Then, for each candidate p_i of the l-th chromosome, we perform the optimization once, using as starting point the set of most likely candidates, replacing the l-th one by p_i. If the objective function value is higher than our baseline (thus, using p_i as a candidate for chromosome l did not improved the fit), we remove the candidate from our list. We thus reduced the number of candidates in a small number of steps and can proceed with initializing the optimization with the all possible combination of this reduced set of candidates. Our implementation allows the user to specify whether or not to perform this filtering step.

Appendix C

HiC-Pro: An optimized and flexible pipeline for Hi-C data processing

This chapter has been publishel in a slightly modified form in [Servant et al., 2015] as joint work with Nicolas Servant, Bryan R. Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Job Dekker, Edith Heard and Emmanuel Barillot.

Abstract

HiC-Pro is an optimized and flexible pipeline for processing Hi-C data from raw reads to normalized contact maps. HiC-Pro maps reads, detects valid ligation products, performs quality controls and generates intra and inter-chromosomal contact maps. It includes a fast implementation of the iterative correction method and is based on a memory-efficient data format for Hi-C contact maps. In addition, HiC-Pro can use phased genotype data to build allele-specific contact maps. We applied HiC-Pro on different Hi-C dataset demonstrating its ability to easily process large data in a reasonable time. Source code and documentation are available at http://github.com/nservant/HiC-Pro.Introduction

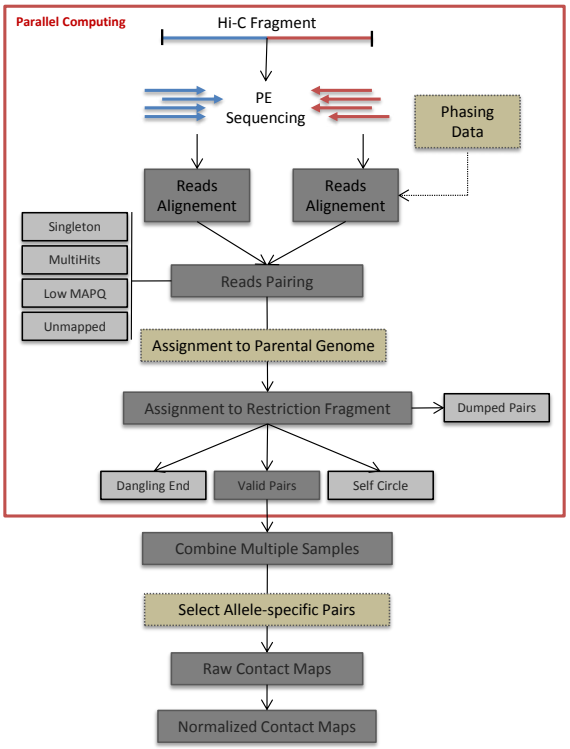


FIGURE 1: HiC-Pro workflow. Reads are first aligned on the reference genome. Only uniquely aligned reads are kept and assigned to a restriction fragment. Interactions are then classified and invalid pairs are discarded. If phased genotyping data and N-masked genome are provided, HiC- Pro will align the reads and assign them to a parental genome. These first steps can be performed in parallel for each read chunk. Data from multiple chunks are then merged and binned to generate a single genome-wide interaction map. For allele-specific analysis, only pairs with at least one allele specific read are used to build the contact maps. The normalization is finally applied to remove Hi-C systematic bias on the genome-wide contact map.

§ 1 Introduction

High-throughput chromosome conformation capture methods are now widely used to map chromatin interactions within regions of interest and across the genome. The use of Hi-C has notably changed our vision of genome organization and its impact on chromatin and genes regulation [de Wit and de Laat, 2012, Barutcu et al., 2015]. The Hi-C technique involves sequencing pairs of interacting DNA fragments, where each mate is associated with one interacting locus. Briefly, cells are crossed-linked, DNA is fragmented using a restriction enzyme or a nuclease, and interacting fragments are ligated together. After paired-end sequencing, each pair of reads can be associated to one DNA interaction [Lieberman-Aiden et al., 2009].

In recent years, the Hi-C technique has demonstrated that the genome is partitioned into domains of different scale and compaction level. The first Hi-C application has

Mapping	Detection of valid interactions	Binning	Correction of systematic noise	Parallel implementation	Allele-specific Analysis
HOMER		x	x		
HICUP	x	x			x
HiCorrector			x	x	
Hiclib	x	x	x	x	
HiC-Pro	x	x	x	x	x

TABLE 1: Comparing solutions for Hi-C data processing. HOMER offers several programs to analysis Hi-C data from aligned reads. HICUP proposes a complete pipeline until the detection of valid interaction products. It can be used together with the SNPsplit software to extract allele specific mapped reads. The hiclib python library can be applied for all analysis steps but requires good programming skills and cannot be used in a single command-line manner. None of these softwares offers to easily process very large data in a parallel mode. The HiCorrector software [Li et al., 2015] provides a parallel implementation of the iterative correction algorithm for dense matrix. Note that HOMER and hiclib also offer additional functions for downstream analysis. In the case of HiC-Pro, the downstream analysis is supported by the HiTC BioConductor package [Servant et al., 2012].

described that the genome is partitioned into distinct compartments of open and close chromatin [Lieberman-Aiden et al., 2009]. Higher throughput and resolution have then suggested the presence of megabase-long and evolutionary conserved smaller domains. These topologically associating domains are characterized by a high frequency of intra-domain chromatin interactions but infrequent inter-domain chromatin interactions [Nora et al., 2012, Dixon et al., 2012]. More recently, very large data sets with deeper sequencing have been used to increase the Hi-C resolution in order to detect loops across the entire genome [Rao et al., 2014, Jin et al., 2013].

As any genome-wide sequencing data, Hi-C usually requires several millions to billions of paired- end sequencing reads, depending on genome size and on the desired resolution. Managing these data thus requires optimized bioinformatics workflows able to extract the contact frequencies in reasonable computational time and with reasonable resources and storage requirements. The overall strategy to analyze Hi-C data is converging among recent studies [Lajoie et al., 2015], but there remains a lack of stable, flexible and efficient bioinformatics workflows to process such data. Solutions such as the HOMER [Heinz et al., 2010] or HICUP programs are already available. HOMER offers several functions to analysis Hi-C data from aligned reads. HICUP proposes a complete pipeline until the detection of valid interaction products. Using HICUP together with the SNPsplit program allows extracting allele-specific interaction products whereas HOMER does not allow extracting allele- specific information. None of these softwares offers a means of correcting contact maps from systematic bias or of processing very large data in a parallel mode. The hiclib package is currently the most commonly used solution for Hi-C data processing [Imakaev et al., 2012]. However, hiclib is a python library that

requires programming skills such as knowledge of Python and advanced linux command-line, and cannot be used in a single command-line manner. In addition, parallelization is not straightforward and it has limitations for the analysis and normalization of very high-resolution data (Table 1).

Here, we present HiC-Pro, an easy-to-use and complete pipeline to process Hi-C data from raw sequencing reads to the normalized contact maps. When phased genotypes are available, HiC-Pro is able to distinguish allele specific interactions and to build both maternal and paternal contact maps. It is optimized and offers a parallel mode for very high-resolution data as well as a fastimplementation of the iterative correction method [Imakaev et al., 2012].

§ 2 Methods

§ 2.1 HiC-Pro Workflow

HiC-Pro is organized into four distinct modules following the main steps of Hi-C data analysis; i) read alignment, ii) detection and filtering of valid interaction products, iii) binning and iv) contact maps normalization (Figure 1).

Mapping. Read pairs are first independently aligned on the reference genome to avoid any constraint on the proximity between the two reads. Most read pairs are expected to be uniquely aligned on the reference genome. A few percent however, are likely to be chimeric reads, meaning that at least one read spans the ligation junction and therefore both interacting loci. As an alternative to the iterative mapping strategy proposed by Imakaev et al. [2012], we propose a two-step approach to rescue and align those reads (Figure 2A). Reads are first aligned on the reference genome using the bowtie2 end-to-end algorithm [Langmead and Salzberg, 2012]. At this point, unmapped reads are mainly composed of chimeric fragments spanning the ligation junction. In a second step, the ligation site of these reads is identified using an exact matching procedure and only their 5' fraction is aligned back on the genome. Both mapping steps are then merged in a single alignment file. Low mapping quality reads, multiple hits and singletons can be discarded.

Detection of valid interactions. Each aligned read can be assigned to one restriction fragment according to the reference genome and the selected restriction enzyme. Both reads are expected to map near a restriction site, and with a distance within the range of molecule size distribution after shearing. Fragments with a size outside the expected range can be discarded if specified but are usually the results of random breaks or star activity of the enzyme, and can therefore be included in downstream analysis

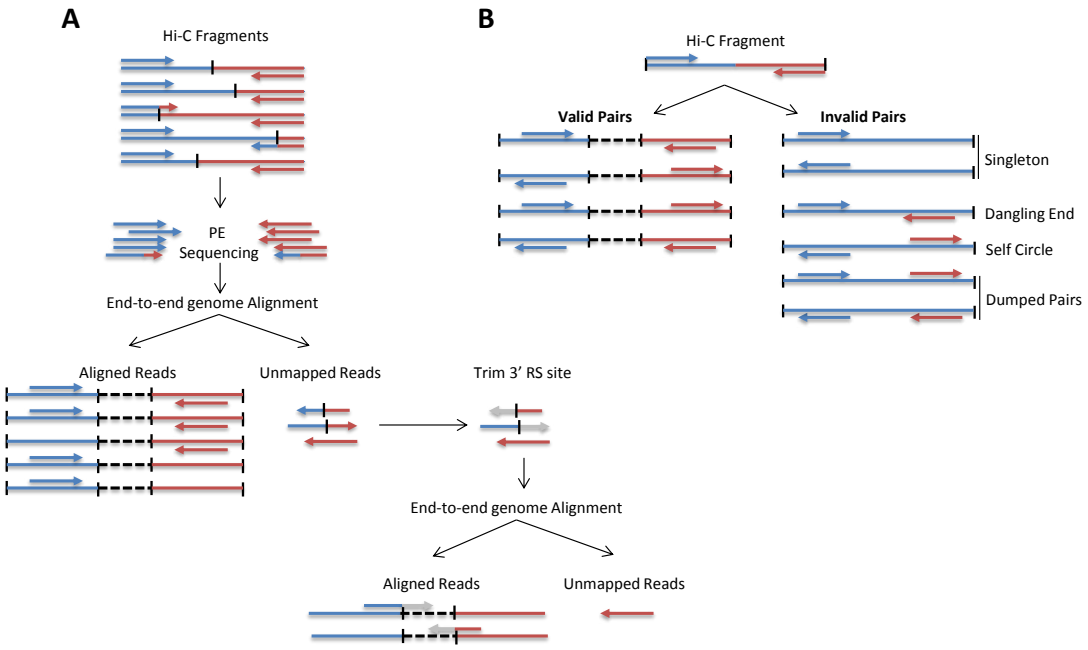


FIGURE 2: **Read pair alignment and filtering. A.** Read pairs are first independently aligned to the reference genome using an end-to-end algorithm. Then, reads spanning the ligation junction which were not aligned on the first step are trimmed at the ligation site and their 5' extremity is realigned on the genome. All aligned reads after these two steps are used for further analysis. **B.** Following the Hi-C protocol, digested fragments are ligated together to generate Hi-C products. A valid Hi-C product is expected to involve two different restriction fragments. Read pairs aligned on the same restriction fragment are classified as dangling end or self-circle products, and are not used to generate the contact maps.

[Imakaev et al., 2012]. Read pairs from invalid ligation products such as dangling end and self-circle ligation are discarded (Figure 2B). Only valid pairs involving two different restriction fragments are used to build the contact maps. Duplicated valid pairs due to PCR artefacts can also be filtered out. Each read is finally tagged in a BAM file according to its mapping and fragment properties (Figure S1).

Binning. In order to generate the contact maps, the genome is divided into bins of equal size, and the number of contacts observed between each pair of bins is reported. A single genome wide interaction map containing both raw intra and inter-chromosomal maps is generated for a set of resolutions defined by the user in the configuration file.

Normalization. In theory, the raw contact counts are expected to be proportional to the true contact frequency between two loci. However, as for any sequencing experiment, it is known that Hi-C data contain different biases mainly due to GC content,

mappability and effective fragment length [Yaffe and Tanay, 2011, Hu et al., 2012]. An appropriate normalization method is therefore mandatory to correct these biases. Over the last few years, several methods have been proposed either using an explicit-factor model for bias correction [Hu et al., 2012] or implicit matrix balancing algorithm [Imakaev et al., 2012, Cournac et al., 2012]. Among the matrix balancing algorithm, the iterative correction of biases based on the Sinkhorn-knopp algorithm has been widely used by recent studies due to its conceptual simplicity, parameter-free nature and ability to correct for unknown biases, although its assumption of the equal visibility across all loci may require further exploration. In theory, a genome-wide interaction matrix is of size $O(N^2)$, where N is the number of genomic bins. Therefore, applying a balancing algorithm on such matrix can be difficult in practice, as it requires a significant amount of memory and computational time.

The degree of sparsity of the Hi-C data is dependent on the bin size and on the sequencing depth of coverage. Even for extremely large sequencing coverage, the interaction frequency between intra-chromosomal loci is expected to decrease as the genomic distance between them increases. High resolution data are therefore usually associated with a high level of sparsity. Exploiting matrix sparsity in the implementation can improve the performance of the balancing algorithm for high resolution. HiC-Pro proposes a fast sparse based implementation of the iterative correction method [Imakaev et al., 2012] allowing to normalize genome-wide high resolution contact matrices in a short time and with reasonable memory requirement.

§ 2.2 Quality Controls

To assess the quality of a Hi-C experiment, HiC-Pro performs a variety of quality controls at different steps of the pipeline (Figure 3). The alignment statistics is the first available quality metric. According to the reference genome, a high-quality Hi-C experiment is usually associated with a high mapping rate. The number of reads aligned in the second mapping step is also an interesting control as it reflects the proportion of reads spanning the ligation junction. An abnormal level of chimeric reads can reflect a ligation issue during the library preparation. Once the reads are aligned on the genome, the fraction of singleton or multiple hits is usually expected to be low. The ligation efficiency can also be assessed using the filtering of valid and invalid pairs. As the ligation is a random process, 25% of each valid ligation class defined by distinct read pairs orientation, is expected. In the same way, a high level of dangling-end or self-circle read pairs is associated with a bad quality experiment, and reveals a problem during the digestion, fill-in or ligation steps.

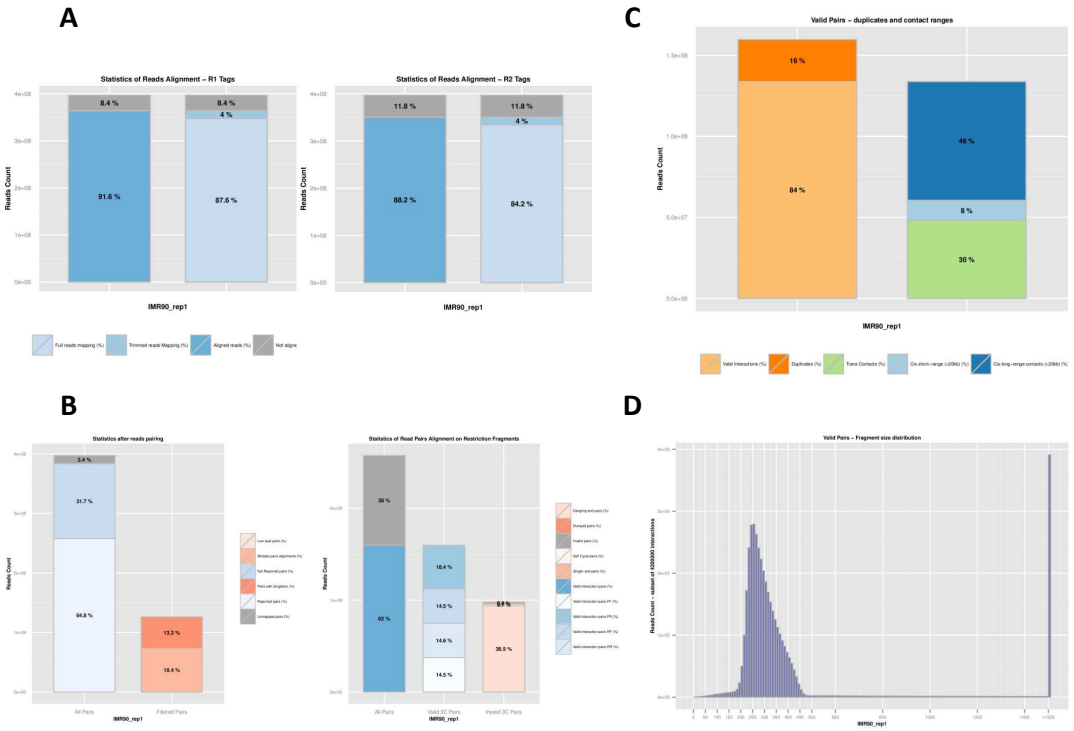


FIGURE 3: **HiC-Pro Quality Controls.** Quality controls reported by HiC-Pro (IMR90, Dixon et al. [2012] data). **A.** Read pairs statistics after alignment. Singleton and multiple hits are usually removed at this step. **B.** Read pairs are assigned to a restriction fragment. Invalid pairs such as dangling-end and self-circle are good indicators of the library quality and are tracked but discarded for subsequent further analysis. **C.** Fraction of duplicated reads, as well as short range versus long range interactions. **D.** Distribution of insert size calculated on a subset of valid pairs.

Additional quality controls such as fragment size distribution can be extracted from the list of valid interaction products. A high level of duplication indicates a poor molecular complexity and a potential PCR bias. Finally, an important metric is to look at the fraction of intra and inter- chromosomal interactions, as well as long range versus short range intra-chromosomal interactions. As two genomic loci close on the linear genome are more likely to randomly interact, a strong diagonal is expected on the raw contact maps. A low quality experiment will result in a low fraction of intra-chromosomal interactions depending on the organism and the biological context. A high quality Hi-C experiment on Human genome is typically characterized by at least 40% of intra-chromosomal interactions [Lajoie et al., 2015]. In the same way, a high quality experiment is usually characterized by a significant fraction (¿40%) of long range intra-chromosomal valid pairs [Rao et al., 2014].

	Dense format (MB)	Sparse Symmetric format (MB)
IMR90_CCL186 1Mbp	27	49
IMR90_CL186 500kbp	82	181
IMR90_CCL186 150kbp	822	911
IMR90_CCL186 40kbp	12 000	1 900
IMR90_CL186 20kbp	45 000	2 600
IMR90_CL186 5kbp	720 000	4 200

TABLE 2: **Comparison of contact maps format.** Disk space for IMR90_CCL186 genome-wide contact map generated either using the classical dense format or the sparse symmetric format at different resolution.

§ 2.3 Speed and scalability

Generating genome-wide contact maps at 40 to 1 kb resolution requires a sequencing depth from hundred of millions to multi-billions paired-end reads according to the organism [Dixon et al., 2012, Rao et al., 2014]. However, the main processing steps from read mapping to fragment reconstruction can be optimized using parallel computation of read chunks, significantly reducing the time taken in the Hi-C data processing. Next, all valid interactions are merged to remove the duplicates and to generate the final contact maps. The user can easily run the complete analysis workflow with a single command-line either on a single laptop, or on a computational cluster. Analysis parameters are all defined in a single configuration file. In addition, HiC-Pro is modular and sequential allowing the user to focus on a sub-part of the processing without running the complete workflow. In this way, HiC-Pro can also be used in complement to other methods, for instance by running the workflow from already aligned files, or by simply normalizing published raw contact maps.

The main steps of the pipeline are implemented in Python and C++ programming languages and are based on efficient data structures, such as compressed sparse row matrices for contact count data. Using an adequate data structure allows to speed up data processing but also to circumvent memory limitations. In this way, HiC-Pro allows a genome wide iterative correction to be run at very high resolution and in a short time. Our normalization implementation exploits numpy's dense array format and fast operations, scipy's sparse matrices representation and Cython to combine C and Python to reach performances of C executables with the ease of use and maintainability of the Python language.

§ 2.4 Contact maps storage

The genome-wide contact maps are generated for the resolutions defined by the user. A contact map is defined as a matrix of contact counts and a description of the associated

genomic bins and is usually stored as a matrix, divided into bins of equal size. The bin size represents the resolution at which the data will be analyzed. For instance, a Human 20 Kb genome-wide map is represented by a square matrix of 150000 rows and columns which can be difficult to manage in practice. To address this issue, we propose a standard contact maps format based on two main observations. Contact maps at high resolution are i) usually sparse and ii) are expected to be symmetric. Storing the non null contacts from half of the matrix is therefore enough to summarize all the contact frequencies. Using this format leads to a 10 to 150-fold reduction of the disk space compared to dense format (Table 2).

§ 2.5 Allele specific analysis

HiC-Pro is able to incorporate phased haplotype information in the Hi-C data processing in order to generate allele specific contact maps (Figure 1). In this context, the sequencing reads are first aligned on a reference genome for which all polymorphic sites were first N-masked. This masking strategy avoids systematic bias toward the reference allele, compared to standard procedure where reads are mapped on an unmasked genome. Once aligned HiC-Pro browses all reads spanning a polymorphic site, locates the nucleotide at the appropriate position, and assigns the read either to the maternal or paternal allele. Reads without SNPs information as well as reads with conflicting allele assignment or unexpected allele at polymorphic sites are flagged as unassigned. A BAM file with an allele specific tag for each read is generated and can be used for further analysis. Then, we classified as allele specific, all pairs for which both reads are assigned to the same parental allele or for which one read is assigned to one parental allele and the other is unassigned. These allele specific read pairs are then used to generate a genome-wide contact maps for each parental genome. Finally, the two allele specific genome-wide contact maps are independently normalized using the iterative correction algorithm.

§ 3 Results

§ 3.1 HiC-Pro results and performances

We processed Hi-C data from two public datasets; IMR90 human cell lines from [Dixon et al., 2012] (IMR90) and from [Rao et al., 2014] (IMR90_CCL186). The latter is currently one of the biggest dataset available, used to generate up to 5kb contact maps. For each dataset, we ran HiC-Pro and generated normalized contact maps at 20 kb, 40 kb,

Dataset	IMR90	IMR90	IMR90	IMR90_CCL186
# Reads	397200000	397200000	397200000	1535222082
Pipeline	hiclib	HiC-Pro	HiC-Pro parallel	HiC-Pro parallel
# Input files	10	10	84	160
# Jobs	1	1	42	80
# CPU per Job	8	8	4	4
Max mem	10	7	7	24
Wall time	28:24	14:32	02:15	11:49
- Mapping	22:03	10:31	00:21	05:56
- Filtering	00:30	03:10	00:05	00:36
- Merge	—	00:20	00:18	00:50
- Contacts maps	01:45	00:15	00:15	00:42
- ICE	04:06	01:16	01:16	03:49

TABLE 3: **HiC-Pro performances and comparison with hiclib.** HiC-Pro was run on IMR90 Hi-C dataset from Dixon et al. and Rao et al. in order to generate contact maps at resolution 20kb, 40kb, 150kb, 500kb and 1Mb. Contact maps at 5kb were also generated for the IMR90_CCL186 dataset. CPU time for each step of the pipeline is reported and compared to the hiclib python library. The reported results include I/O time of writing contact maps in text format.

150 kb, 500 kb and 1 Mb resolution. Normalized contact maps at 5 kb were only generated for the IMR90_CCL186 dataset. The datasets were either used in their original form or split into chunks containing 10 or 20 million of read pairs. Using HiC-Pro, the processing of the Dixon's dataset (397.2 million read pairs split in 84 read chunks) was completed in 2 hours using 168 CPUs (Table 3). Each chunk was mapped on the Human genome requiring 4 CPUs (2 for each mate) and 7Go of RAM. Processing the 84 chunks in parallel allows to extract the list of valid interactions in less than 30 minutes. All chunks are then merged to generate and normalize the genome-wide contact map.

In order to compare our results with the hiclib library, we ran HiC-Pro on the same dataset, and without initial reads split, using 8 CPUs. In addition for ease of use, HiC-Pro performed the complete analysis in less than 15 hours compared to 28 hours for the hiclib pipeline. The main difference in speed is explained by our two-steps mapping strategy compared to the iterative mapping strategy of hiclib which aligned the 35pb reads in 4 steps. The optimization of the binning process and the implementation of the normalization algorithm lead to a three-fold decrease in timeto generate and normalize the genome-wide contact.

The IMR90 sample from the Rao's dataset (1.5 billion read pairs split in 160 read chunks) was processed in parallel using 320 CPUs to generate up to 5kb contact maps in 12 hours, demonstrating the ability of HiC-Pro to analyze very large data in a reasonable time. The merged list of valid interactions was generated in less than 7.5 hours. The normalization of the genome- wide contact map at 1Mb, 500Kb, 150Kb, 40Kb, 20Kb and 5Kb was

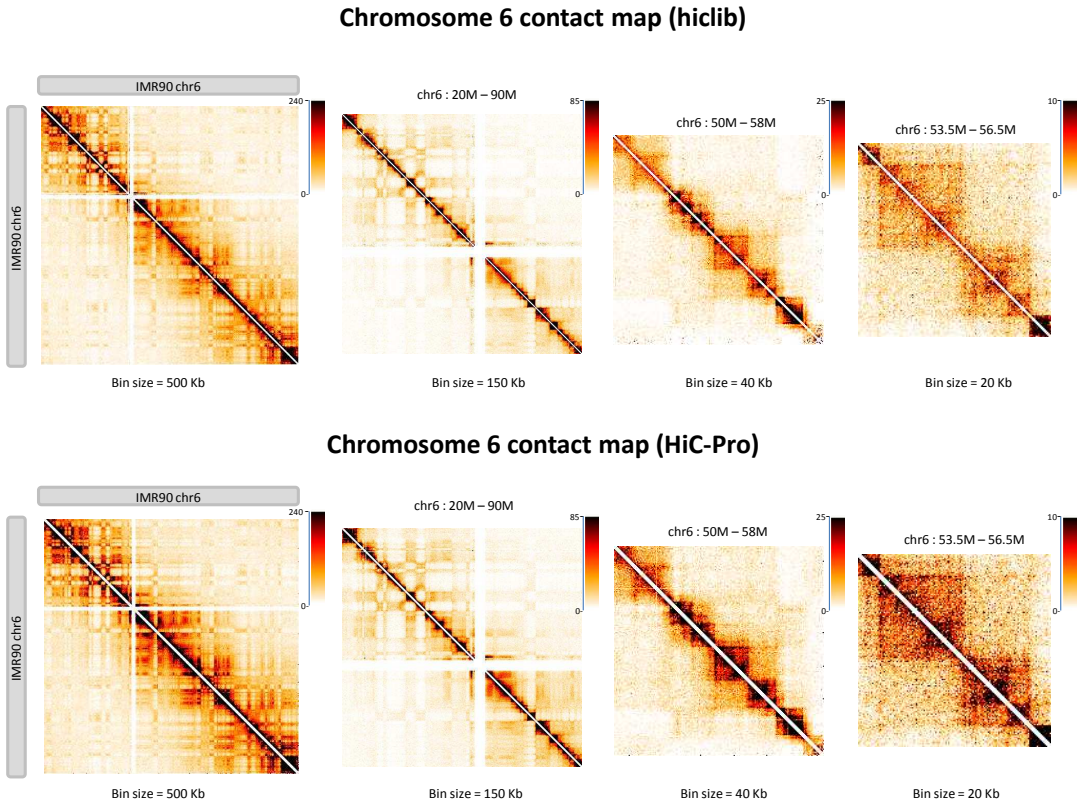


FIGURE 4: **Comparison of HiC-Pro and hiclib contact maps.** Chromosome 6 contact maps generated by hiclib (top) and HiC-Pro (bottom) at different resolutions. The chromatin interaction data generated by the two pipelines are highly similar.

performed in less than 4 hours. Details about the results and the implementation of the different solutions are available in supplementary materials.

Finally, we compared the Hi-C processing results of hiclib and HiC-Pro on the IMR90 dataset. Although the different processing and filtering steps of the two pipelines are not exactly the same, we observed a good concordance in the results (Table S1). Using default parameters, HiC-Pro is less stringent than hiclib and used more valid interactions to build the contact maps. The two sets of normalized contact maps generated at different resolutions are highly similar as illustrated on Figure 4. We further explored the similarity between the maps generated by two pipelines by computing the Spearman correlation of the normalized intra-chromosomal maps. The average correlation coefficient across all chromosomes at different resolutions was 0.83 [0.65 – 0.95]. Finally, since the inter-chromosomal data are usually very sparse, we summarized the inter- chromosomal signal using the two one-dimensional coverage vectors of rows and columns [Yaffe and Tanay, 2011, Hu et al., 2012]. The average Spearman correlation coefficient of all coverage vectors between hiclib and HiC- Pro inter-chromosomal contact maps was 0.75 [0.46 – 0.98] (Figure S2).

	HiC-Pro - Iced (dense – 1 CPU)	HiC-Pro - Iced (sparse – 1 CPU)	HiCorrector - MES (dense – 1 CPU)	HiCorrector – MEP (dense – 8 CPU)
IMR90 1Mbp	00:00:12	00:00:25	00:00:25	00:00:06
IMR90 500kbp	00:00:40	00:01:30	00:02:15	00:00:22
IMR90 150kbp	-	00:04:28	00:13:21	00:03:10
IMR90 40kbp	-	00:07:19	02:35:34	00:35:43
IMR90 20kbp	-	00:08:36	12:57:17	02:34:05

TABLE 4: **Performances of iterative correction on IMR90 data.** HiC-Pro is based on a fast implementation of the iterative correction algorithm. We therefore compare our method with the HiCorrector software [Li et al., 2015] for Hi-C data normalization (hours:minutes:seconds). All algorithms were terminated after 20 iterations (see supplementary material for details).

§ 3.2 Implementation of the iterative correction algorithm

We propose an implementation of the iterative correction procedure which emphasizes ease of use, performance, memory-efficiency and maintainability. We obtain a higher or similar performance on a single core when compared to the original ICE implementation from the hiclib library (Table 3) and from the HiCorrector package [Li et al., 2015]. The HiCorrector package proposes a parallel version of the iterative correction for dense matrices. We therefore compared the performance of HiCorrector with the HiC-Pro normalization at different Hi-C resolution (Table 4). All algorithms were terminated after 20 iterations for the purpose of performance comparison, as each iteration requires nearly the same running time.

Choosing dense- or sparse-matrix based implementation is dependent on the Hi-C data resolution and on the depth of coverage. Although our implementation can be run either in sparse or in dense mode, the available data published at resolution of 5-40Kb are currently characterized by a high degree of sparsity. At each level of Hi-C contact map resolution, we compared our dense or sparse implementation with the parallel and/or sequential version of HiCorrector. Our results demonstrate that using a compressed sparse row matrices structure is more efficient on high resolution contact maps (¡40kb) than using parallel computing on dense matrices. As expected for low resolutioncontact maps (1Mb, 500Kb), using dense matrix implementation is more efficient in time although the gain in practice remains negligible.

The code for the normalization is available as a standalone package (https://github.com/hiclib/iced) and included into HiC-Pro. Our implementation based on sparse row matrices is able to normalize a 20kb genome map in less than 30 minutes and 5 Go of RAM (Table 4). Genome-wide normalization at 5kb can be achieved in less than 2.5 hours with 24 Go of RAM. Thus, our implementation substantially speeds up and facilitates the normalization of Hi-C data prior to downstream analysis.

A

Total number of read pairs	826 414 879
Total number of valid pairs	503 536 186 (100%)
Number of pairs assigned to G1	28 391 258 (5.64%)
Number of pairs assigned to G2	28 308 925 (5.62%)
Number of trans G1/G2 pairs	603 213 (0.12)
Number of unassigned reads	446 171 241 (88.60%)
Number of conflicting reads	61 549 (0.01%)

B

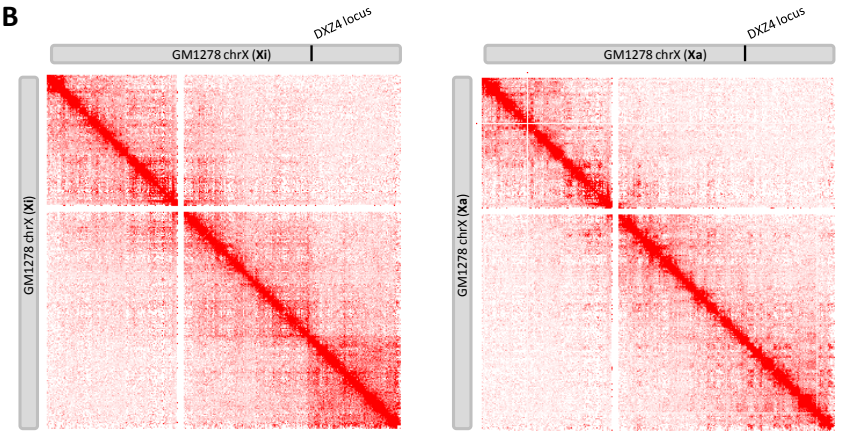


FIGURE 5: **Allele specific analysis. A.** Allele specific analysis of GM12878 cell line. Phasing data were gathered from the Illumina Platinum Genomes Project. In total, 2,210,222 high quality SNPs from GM12878 data were used to distinguish both alleles. Around 6% of the read pairs were assigned to each parental allele and used to build the allele-specific contact maps. **B.** Intra- chromosomal contact maps of inactive and active X chromosome of GM12878 at 500 Kb resolution. The inactive copy of chromosome X is partitioned into two mega-domains which are not seen in the active X chromosome. The boundary between the two mega-domains lies near the DXZ4 micro-satellite.

§ 3.3 Allele specific contact maps

We used HiC-Pro to generate the allele specific contact maps of human GM12878 cell line. Differences in paternal and maternal X chromosome organization were recently described, with the presence of mega-domains on the inactive X chromosome, which are not seen in the active X chromosome [Rao et al., 2014, Minajigi et al., 2015]. Here, we used HiC-Pro to generate the active and inactive chromosome X contacts maps of GM12878 cell line using the Hi-C dataset published by [Selvaraj et al., 2013]. Phasing data were gathered from the Illumina Platinum Genomes Project. Only good quality heterozygous SNPs were selected. The final list contained 2,210,222 SNPs. We then masked the Human genome hg19 by replacing the SNP position by an 'N' using the BEDTools utilities [Quinlan and Hall, 2010] and generated the new bowtie2 indexes. In practice, the allele specific analysis can be easily performed by simply specifying to HiC-Pro the list of SNPs and the N-masked indexes for reads alignment through the

configuration file. Among the initial 826 millions of read pairs, 61% were classified as valid interactions by HiC-Pro.

Around 6% of valid interactions were then assigned to either the paternal or maternal genome and used to construct the haploid maps. As expected, the inactive X chromosome map is partitioned into two mega-domains (Figure 5). The boundary between the two mega-domains lies near the DXZ4 micro-satellite.

§ 4 Conclusion

As the Hi-C technique become mature, it is now important to develop bioinformatics solutions which can be shared and used for any project. HiC-Pro is a flexible and efficient pipeline for Hi-C data processing. It is freely available as a collaborative project at https://github.com/nservant/HiC-Pro. It is optimized to address the challenges of processing high-resolution data and proposes an efficient format for contact maps sharing.

In addition for ease of use, HiC-Pro performs quality controls, and can process Hi-C data from the raw sequencing reads to the normalized and ready-to-use genome-wide contact maps. The intra and inter-chromosomal contact maps generated by HiC-Pro are highly similar with the onesgenerated by the hiclib package. In addition, when phased genotyping data are available, HiC-Pro allows to easily generate allele-specific maps for homologous chromosomes. Finally, HiC-Pro includes an optimized version of the iterative correction algorithm which substantially speeds up and facilitates the normalization of Hi-C data. The code is also available as a standalone package (https://github.com/hiclib/iced).

A complete online manual is available at http://nservant.github.io/HiC-Pro. The raw and normalized contact maps are compatible with the HiTC Bioconductor package [Servant et al., 2012], and can therefore be loaded in the R environment for visualization and further analysis.

Supplementary information

The following additional software and libraries are required:

- Bowtie2 mapper [Langmead and Salzberg, 2012] (http://bowtie-bio.sourceforge.net/bowtie2)
- R and the BioConductor packages RColorBrewer, ggplot2, grid.

- Samtools (≥0.1.19, http://samtools.sourceforge.net/)
- Python (≥2.7) with the pysam, bx.python, numpy and scipy libraries
- The g++ compiler

Note that the Bowtie2 ≥2.2.2 is strongly recommended for allele-specific analysis, as since this version, reads alignment on N-masked genome has been highly improved.

Most of the installation steps are fully automatic using a simple command line. The Bowtie2 and Samtools software are automatically downloaded and installed if not detected on the system.

The HiC-Pro pipeline can be installed on a Linux/UNIX-like operating system.

Public dataset used

We applied the HiC-Pro pipeline on three public dataset available on GEO. The IMR90 Hi-C contact maps were first published by Dixon et al. [2012] at a resolution of 20Kb and 40Kb. The five run of IMR90 replicate 1 (GSM862724) were used and merged, for a total number of 397.2 million read pairs. We refer to this sample in the manuscript as IMR90. More recently, Rao et al. [2014] generate genome-wide contact maps at a resolution of 1-5kb (GSE63525) for nine different cell lines. For the purpose of this paper, we applied HiC-Pro on the IMR90 cell line (GSM1551599, GSM1551600, GSM1551601, GSM1551602, GSM1551603, GSM1551604, GSM1551605). The combined samples represent a sequencing depth of 1.5 billion reads. We refer to this sample in the manuscript as IMR90_CCL186.

The allele specific analysis was performed using the human GM12878 Hi-C data published by Selvaraj et al. [2013] (GSE48592). Phasing data were gathered from the Illumina Platinum Project v7 (http://www.illumina.com/platinumgenomes/).

Results and implementation

All pipelines and software were run on the high-performance computing resource of the Institut Curie. Each node has a total of 32 or 48 processors (Intel Xeon 2.2 GHz) and 128 GB memory.

The HiC-Pro version 2.6.0 was used and the hiclib library was downloaded from http://mirnylab.bitbucke
In order to compare the performance between both solutions, we run the pipeline described in the hiclib's repository (hiclib/examples/pipeline2014/), on a single node with

8 CPUs. All default parameters were used. Following the hiclib's help pages, the binnedData and highResBinnedData classes were respectively used for low (≥100kb) and high resolution data (≤100kb) as illustrated in the testHighResHiC.py script. The HiC-Pro pipeline was run either in normal or parallel mode. HiC-Pro and hiclib were compared until the generation of genome-wide normalized contact maps at a resolution of 1Mb, 500Kb, 150Kb, 40Kb and 20Kb. Both pipelines were run with default parameters. The running time includes the export of contact maps in text format.

In order to compare the results generated by both pipelines, we calculated the Spearman correlation coefficient between HiC-Pro and hiclib intra and inter-chromosomal maps at different resolutions. By default hiclib is removing the matrix diagonal before doing the normalization. We therefore apply the same filter on the HiC-Pro contact maps. The Spearman correlation coefficients were calculated between all intra-chromosomal maps. Since the inter-chromosomal contact maps are sparse, instead of measuring the correlation directly between the two maps, we computed the Spearman correlation of the one-dimensional coverage vectors of inter- chromosomal maps as proposed by Yaffe and Tanay [2011], and Hu et al. [2012]. The results are available in Figure 7.

The HiCorrector package (version 1.1) was downloaded and compiled using openmpi-1.4.5. We compared the performance of the iterative correction algorithm included in HiC-Pro with HiCorrector on the Dixon et al. IMR90 dataset. We first split the dense matrix files using the split_data_parallel tool and the following command line; "mpirun -np 8 split_data_paralllel DENSE_MATRIX_FILE NB_ROWS ./ 8 1024 job_id" where DENSE_MATRIX_FILE is the path to the dense matrix and NB_ROWS the number of matrix rows. The genome wide contact maps were therefore split into 7 sub-matrices for 1M, 500Kb, 150Kb resolutions, 28 sub-matrices for the 40Kb resolution and 91 for the 20 Kb resolution.

The iterative correction was then applied using the ICE-MES and ICE-MEP methods on the genome-wide contact map. All algorithms were terminated after 20 iterations.

We ran the ICE_MEP method using the following parameters; "mpirun -np 8 ic_mep – useSplitInputFiles –numRows=NB_RAWS –maxIteration=20 –numTask=8 –memSizePerTask=1024 –jobID=job_id". The ICE_MES method was run using the following parameters; ""ic_mes DENSE_MATRIX_FILE 5000 3115 20 0 0". The HiC-Pro normalization (1 CPU) was run using the ice script and the following parameters; "- -max_iter 20 –eps 1e-15 –filtering_perc 0". The "–dense" option was added for the dense matrices. All input and output files were stored in the local scratch folder to limit the I/O time due to NFS.

Total read pairs	397 194 480	397 194 480
Uniquely aligned read pairs	231 047 307 (58.17%)	257 502 619 (64.83%)
Self-Circle	1 569 902 (0.68%)	1 793 553 (0.69%)
Dangling-end	79 701 493 (34.49%)	94 024 488 (36.51%)
Valid interactions	141 686 863 (61.32%)	159 737 835 (62.03%)
Filtered valid interactions	107 977 460 (46.73%)	133 761 282 (51.9%)
Intra-chromosomal contacts	66 619 145 (61.69%)	85 694 952 (64.06%)
Inter-chromosomal contacts	41 358 315 (38.30%)	48 066 330 (35.93%)

TABLE 5: **Comparison of hiclib and HiC-Pro processing steps.**
Both pipelines are generating concordant results across the processing steps. The fraction of uniquely aligned read pairs is calculated on the total number of initial reads. Self-circle and dangling-end fractions are calculated on the total number of aligned read pairs. Intra and inter-chromosomal contacts are calculated as a fraction of filtered valid interactions.

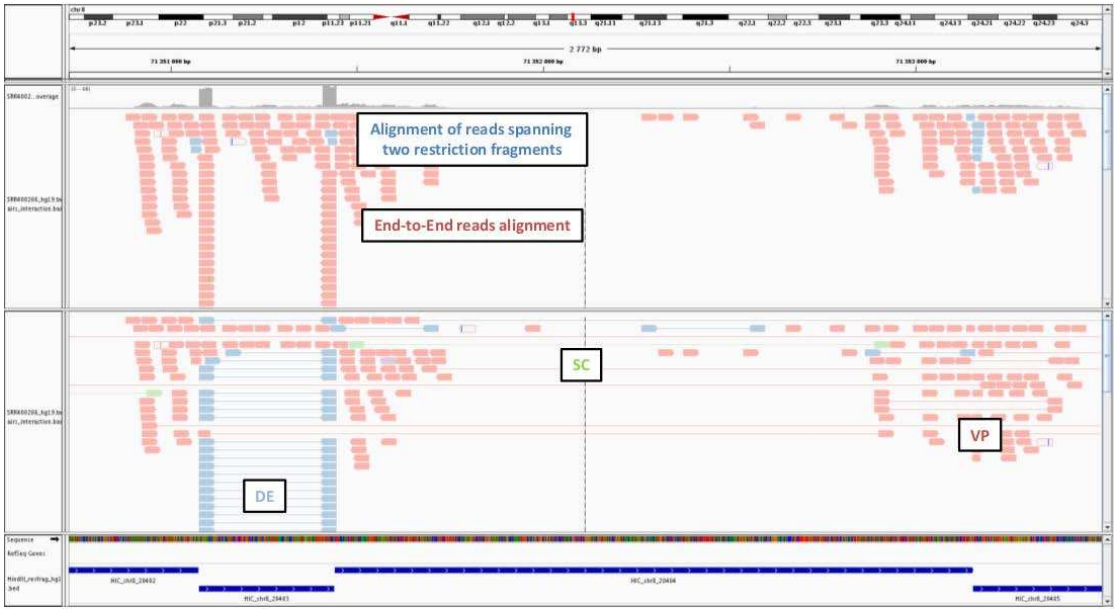


FIGURE 6: **IGV screenshot of BAM file after mapping and fragment reconstruction.**
Top panel. The reads are colored according to the alignment procedure. Blue reads were trimmed before mapping, and flanked the restriction fragment borders. Bottom panel. Read pairs are colored according to their classification. Valid interactions are in red, dangling end in blue and self-circle ligation in green.

§ 4.1 Supplementary table

Supplementary figures

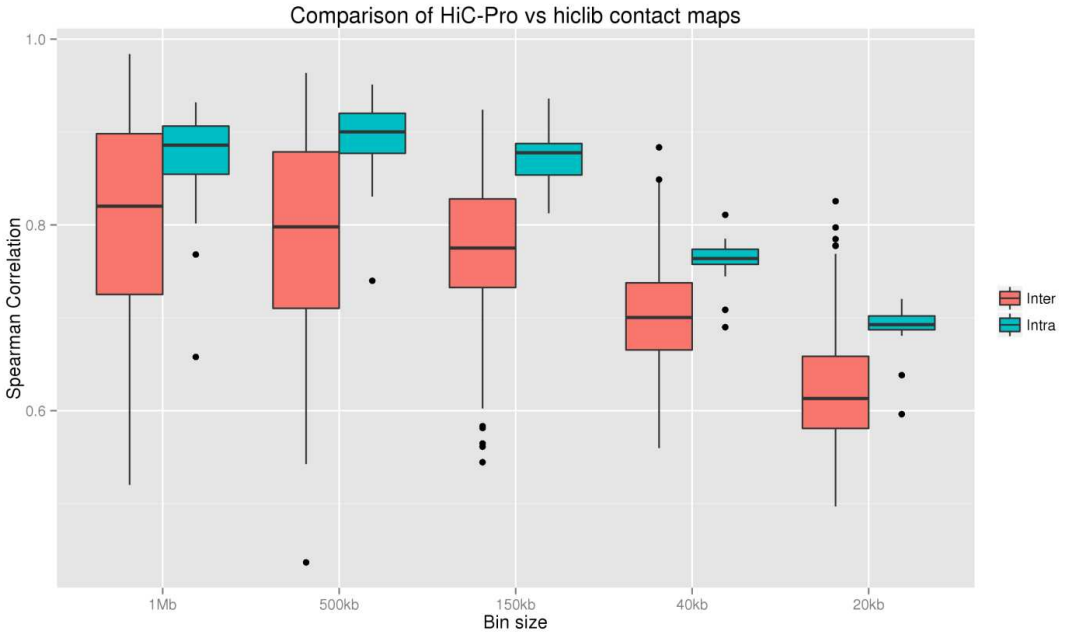


FIGURE 7: **Correlation of intra and inter-chromosomal contact maps generated by hiclib and HiC-Pro.**
Boxplots of the Spearman correlation coefficients of intra and inter-chromosomal maps generated at different resolutions by both pipelines.

Appendix D

Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C

This chapter has been published in a slightly modified form in Ay et al. [2015b], as joint work with Ferhat Ay, Thanh H. Vu, Michael J. Zeitz, Jan E. Carette, Jean-Philippe Vert, Andrew R. Hoffman and William S. Noble.

Abstract

Background: Several recently developed experimental methods, each an extension of the chromatin conformation capture (3C) assay, have enabled the genome-wide profiling of chromatin contacts between pairs of genomic loci in 3D. Especially in complex eukaryotes, data generated by these methods, coupled with other genome-wide datasets, demonstrated that non-random chromatin folding correlates strongly with cellular processes such as gene expression and DNA replication.

Results: We describe a genome architecture assay, tethered multiple 3C (TM3C), that maps genome-wide chromatin contacts via a simple protocol of restriction enzyme digestion and religation of fragments upon agarose gel beads followed by paired-end sequencing. In addition to identifying contacts between pairs of loci,

TM3C enables identification of contacts among more than two loci simultaneously. We use TM3C to assay the genome architectures of two human cell lines: KBM7, a near-haploid chronic leukemia cell line, and NHEK, a normal diploid human epidermal keratinocyte cell line. We confirm that the contact frequency maps produced by TM3C exhibit features characteristic of existing genome architecture datasets, including the expected scaling of contact probabilities with genomic distance, megabase scale chromosomal compartments and sub-megabase scale topological domains. We also confirm that TM3C captures several known cell type-specific contacts, ploidy shifts and translocations, such as Philadelphia chromosome formation (Ph+) in KBM7. We confirm a subset of the triple contacts involving the *IGF2-H19* imprinting control region (ICR) using PCR analysis for KBM7 cells. Our genome-wide analysis of pairwise and triple contacts demonstrates their preference for linking open chromatin regions to each other and for linking regions with higher numbers of DNase hypersensitive sites (DHSs) to each other. For near-haploid KBM7 cells, we infer whole genome 3D models that exhibit clustering of small chromosomes with each other and large chromosomes with each other, consistent with previous studies of the genome architectures of other human cell lines.

Conclusion: TM3C is a simple protocol for ascertaining genome architecture and can be used to identify simultaneous contacts among three or four loci. Application of TM3C to a near-haploid human cell line revealed large-scale features of chromosomal organization and multi-way chromatin contacts that preferentially link regions of open chromatin.

Keywords: genome architecture, chromatin conformation capture, multi-locus chromatin contacts, near-haploid human cells, leukemia, three-dimensional modeling.

§ 1 Background

A variety of microscopic imaging techniques have long been used to study chromatin architecture and nuclear organization [Langer-Safer et al., 1982, Manders et al., 2003,

Cremer et al., 2008]. Recent advances triggered by the invention of chromatin conformation capture (3C) enable ascertainment of genome architecture on a genome-wide scale for virtually any genome, including human [Lieberman-Aiden et al., 2009, Dixon et al., 2012, Jin et al., 2013], mouse [Zhang et al., 2012, Dixon et al., 2012], budding yeast [Duan et al., 2010], bacteria [Umbarger et al., 2011], fruit fly [Sexton et al., 2012] and a malarial parasite [Ay et al., 2014b]. These studies have revealed that the three-dimensional form of the genome *in vivo* is highly related to genome function through processes such as gene expression and replication timing. Therefore, understanding how chromosomes fold and fit within nuclei and how this folding relates to function and fitness is crucial in gathering a thorough picture of epigenetic control of gene regulation for eukaryotic organisms.

Hi-C was the first molecular assay to measure genome architecture on a genome-wide scale [Lieberman-Aiden et al., 2009], and the assay continues to be widely used [Jin et al., 2013, Naumova et al., 2013, Ay et al., 2014b]. Hi-C involves seven steps: (1) crosslinking cells with formaldehyde, (2) digesting the DNA with a six-cutter restriction enzyme, (3) filling overhangs with biotinylated residues, (4) ligating the fragments, (5) creating a sequence library using streptavidin pull-down, (6) high-throughput paired-end sequencing, and (7) mapping paired ends independently to the genome to infer contacts. A subsequently described assay by Duan et al. [2010] is more complex, involving a pair of restriction enzymes (REs) applied in three separate steps (RE1, RE2, circularization, then RE1 again), as well as the introduction of EcoP151 restriction sites to produce paired tags of 25–27bp. More recently, the tethered conformation capture (TCC) assay enhances the signal-to-noise ratio by carrying out a Hi-C-like protocol using DNA that is tethered to a solid substrate [Kalhor et al., 2011].

One limitation of current genome architecture assays is their inability to identify simultaneous interactions among multiple loci. Chromosomes are composed of complex higher order chromatin structures that bring many distal loci into close proximity. In particular, evidence suggests that eukaryotic transcription occurs in factories containing many genes [Cook, 1999]. Recently, multiple-gene interaction complexes associated with promoters were found to contain an average of nearly nine genes [Li et al., 2012]. However, currently available experimental data cannot ascertain to what extent these multiple gene interactions occur simultaneously or are confined to different sub-populations of nuclei. This distinction is analogous to the distinction between "party hubs" and "date hubs" in protein-protein interaction networks, in which a hub protein interacts either simultaneously or in a serial fashion with a series of partner proteins [Han et al., 2004]. In the context of genome architecture assays, distinguishing between "party loci" and "date loci" will be a crucial first step in elucidating the role of combinatorial regulation of gene expression.

A molecular colony technique recently developed by Gavrilov et al. [2014] investigated multicomponent interactions among remote enhancers and active β-globin genes in mouse erythroid cells. This assay, however, is PCR-based and requires a primer design step, which prevents it from providing a genome-wide picture of potential multicomponent contacts. An earlier genome-wide assay by Sexton et al., which is adapted from the traditional Hi-C protocol and is similar to the assay we present here, acknowledged the existence of multi-locus contacts that can be identified from paired-end reads in their data [Sexton et al., 2012]. However, due to a number of differences in that protocol compared to TM3C (e.g., size selection for larger fragments, shorter read lengths and no in-gel ligation step), identifying a substantial number of multi-locus contacts was not possible when we apply our two-phase mapping pipeline to the Sexton et al. data (<0.0004% triples and no quadruples). Therefore, genome-wide methods that distinguish between simultaneous contacts among multiple loci and pairwise contacts that happen in different sub-populations of cells are still necessary.

To address this issue, we developed the tethered multiple chromosome conformation capture assay (TM3C), which involves a simple protocol of restriction enzyme digestion and religation of fragments within agarose gel beads (tethering step) followed by high throughput paired-end sequencing (Figure 1, steps 1–4). We apply TM3C to two human cell lines and confirm that the DNA–DNA contact matrices produced by TM3C exhibit features characteristic of existing genome architecture datasets, including the expected scaling of contact probabilities with genomic distance, enrichment of intrachromosomal contacts, megabase scale chromosomal compartments and sub-megabase scale topological domains. We confirm that TM3C in KBM7 cells captures several known cell type-specific contacts, ploidy shifts and translocations, such as Ph+ formation. In addition, we demonstrate that TM3C enables genome-wide identification of contacts among more than two loci simultaneously. We identify multi-locus contacts involving three (triple) or four (quadruple) loci by a two-phase mapping strategy that separately maps chimeric subsequences within a single read (Figure 1, steps 5–8). This mapping strategy potentially allows us to identify co-regulation or combinatorial regulation events, while also greatly increasing the number of distinct pairwise contacts (doubles) identified. We also validate a subset of the triple contacts involving the *IGF2-H19* imprinting control region (ICR) using PCR for KBM7 cells. We demonstrate that pairwise and triple contacts prefer to link open chromatin regions to each other and regions with higher numbers of DHSs to each other. Finally, we use the contact maps gathered from TM3C to infer a local 3D structure of the *IGF2-H19* region at 40 kb resolution and a whole genome 3D model at 1 Mb resolution for the near-haploid KBM7 genome. Our 3D models place *H19* and *IGF2* genes far away from each other, consistent with their opposite transcriptional status, and place gene-rich small chromosomes (chrs. 16, 17, 19–22) and

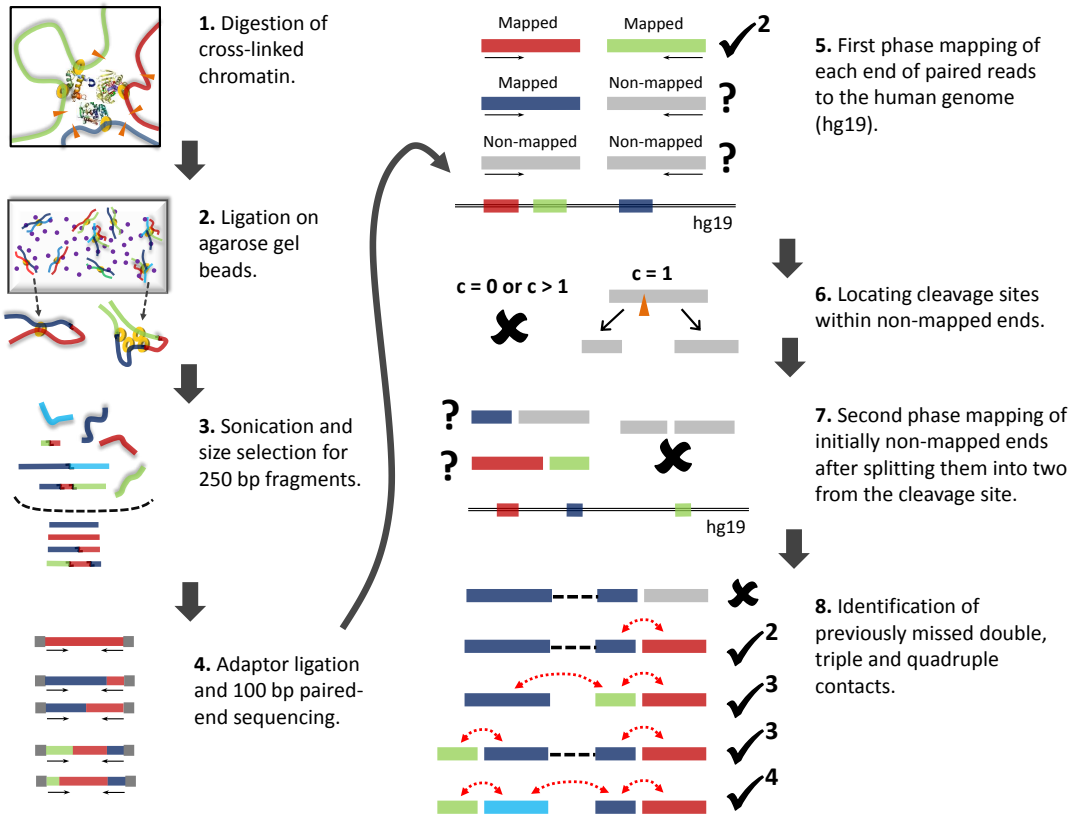


FIGURE 1: **Overview of TM3C experimental protocol and mapping of paired-end reads to human genome. 1.** Cells are treated with formaldehyde, covalently crosslinking proteins to one another and to the DNA. The DNA is then digested with either a single 4-cutter enzyme (DpnII) or a cocktail of enzymes (AluI, DpnII, MspI, and NlaIII). **2.** Melted low-melting agarose solution is added to the digested nuclei to tether the DNA to agarose beads. Thin strings of the hot nuclei plus agarose solution is then transferred to an ice-cold ligation cocktail overnight. **3.** After reversal of formaldehyde crosslinks and purification via gel extraction, the TM3C molecules are sonicated and size-selected for 250 bp fragments. **4.** Size-selected fragments are paired-end sequenced (100 bp per end) after addition of sequencing adaptors. **5.** Each end of paired-end reads are mapped to human reference genome. If both ends are mapped then the pair is considered a *double* and retained because it is informative for genome architecture. **6.** Read ends that do not map to the reference genome are identified and segregated according to the number of cleavage sites they contain for the restriction enzyme(s) used for digestion. **7.** Reads with exactly one cleavage site are considered for the second phase of mapping. These reads are split into two from the cleavage site and each of these two pieces are mapped back to the reference genome. **8.** Read pairs with either one or both ends not mapped in the first mapping phase are reconsidered after second phase. Depending on how many pieces stemming from the original reads are mapped in the second phase, such pairs lead to either no informative contacts, *doubles*, *triples* or *quadruples*.

large chromosomes (chrs. 1–5) near each other, confirming previous observations of gene-density-correlated arrangements of higher-order chromatin in human cells [Bolzer et al., 2005].

§ 2 Results

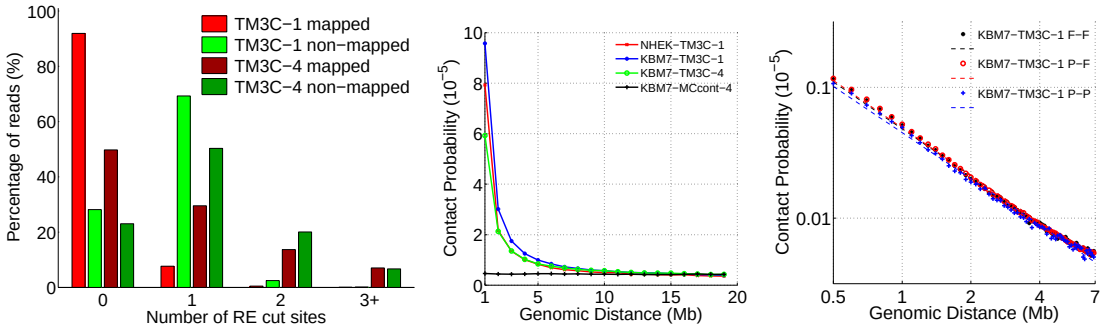
§ 2.1 Tethered multiple chromatin conformation capture (TM3C)

To identify simultaneous chromatin contacts among two or more loci, we digest crosslinked chromatin with one or more 4-cutter restriction enzymes (REs) (Step 1 of Figure 1). When using multiple REs, we select a set of enzymes such that sticky or blunt ends left by one enzyme are incompatible with the ends left by any other, thereby preventing ligation between fragments generated by different enzymes. We then encapsulate and ligate the digested DNA within agarose beads (Step 2 of Figure 1), which replaces the tethering step of Kalhor et al. [2011]. We then size-select DNA fragments of around 250 bp and subject the selected fragments to high throughput paired-end sequencing (Steps 3, 4 of Figure 1). Our assay differs from the original Hi-C assay in three primary ways: (i) TM3C can use multiple REs simultaneously, (ii) TM3C does not include a step where sticky ends of restriction fragments are biotinylated, and (iii) TM3C carries out the ligation step within agarose gel beads. Digestion using multiple REs greatly increases the resolution that can be achieved via these genome-wide 3C-based techniques (Supplementary Fig. 7). However, comparison of two libraries, one generated with four 4-cutters and the other with only one, suggests that the noise-to-signal ratio is much higher for the multiple 4-cutters case. Our second modification, elimination of the biotinylation step, greatly reduces the complexity of the overall protocol and has already been applied successfully by Sexton et al. [2012]. This simplification, however, comes with the drawback of sequencing many uninformative, unligated sonication products both for the TM3C and the Sexton et al. protocols. Because detection of such uninformative read pairs is computationally trivial, this simplification, fortunately, does not contribute an additional noise factor. The third modification we implement, in-gel ligation, is similar to but simpler than the tethering achieved using protein biotinylation in the tethered conformation capture (TCC) assay [Kalhor et al., 2011]. Our initial experimental data which omitted the in-gel ligation demonstrated that without this step the resulting signal-to-noise ratio for the case of four 4-cutters is very low (95% of the contacts are interchromosomal). Addition of in-gel ligation step improved the percentage of intrachromosomal contacts from 5% to 20% and 48% for the four 4-cutter (KBM7-TM3C-4) and one 4-cutter (KBM7-TM3C-1) libraries, respectively. Therefore, we only

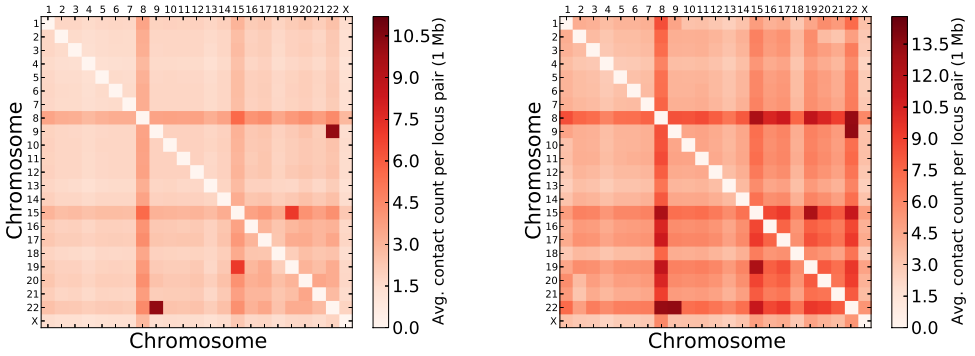
present the results from the libraries generated using the in-gel ligation and focus mainly on the results from our one 4-cutter library for both KBM7 and NHEK cell lines.

We use TM3C to investigate the chromatin architecture of the near haploid cell line KBM7 (25, XY, +8, Ph+) extracted from a heterogeneous chronic leukemia cell line [Kotecki et al., 1999], and NHEK, a normal diploid human keratinocyte primary cell line (Lonza Walkersville Inc.). We construct libraries using only one four-base cutter restriction enzyme (TM3C-1) for both KBM7 and NHEK. We also create two libraries from KBM7 cells using four different four-base cutters, one from crosslinked cells (KBM7-TM3C-4) and one from non-crosslinked cells (KBM7-MCcont-4) as a control (Table 1). In what follows, we report results from application of TM3C to these two human cell lines mainly focusing on KBM7.

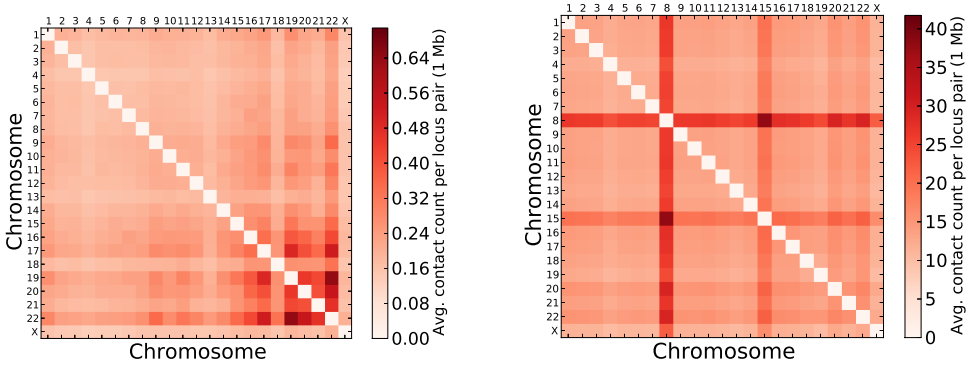
§ 2.2 TM3C reveals multi-locus chromatin contacts



(A) KBM7 (B) (C) KBM7



(D) KBM7-TM3C-1 (E) KBM7-TM3C-4



(F) NHEK-TM3C-1 (G) KBM7-MCcont-4

FIGURE 2: **Consistency of TM3C data with known organizational principles and KBM7 karyotype.** **(a)** Number of RE cut sites within reads that are fully mapped and nonmapped in the first phase mapping for KBM7 libraries. **(b)** Scaling of contact probability with genomic distance for three crosslinked libraries and one non-crosslinked control library. **(c)** Scaling of contact probability in log–log scale for three different sets of contacts identified in KBM7-TM3C-1 library. Pairwise chromosome contact matrices for **(d)** KBM7-TM3C-1, **(e)** KBM7-TM3C-4, **(f)** NHEK-TM3C-1 and **(g)** KBM7-MCcont-4 libraries. For these plots contact counts are averaged over all pairs of mappable 1 Mb windows between the two chromosomes.

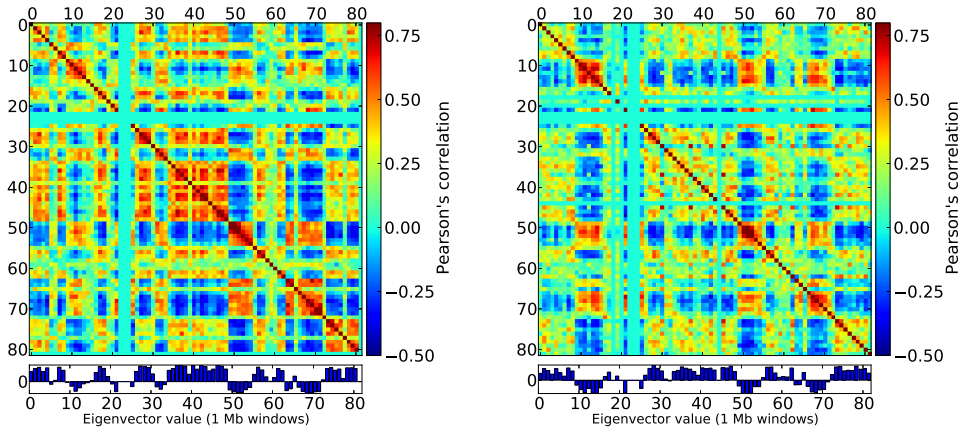
In addition to providing higher resolution, the use of frequently cutting REs (4-cutters) or multiple REs together allows identification of simultaneous contacts among more than two loci, even with reads as short as 100 bp. The original Hi-C method only retains read pairs in which both reads map completely to the reference genome. Here we refer to this type of contacts as type **F-F** (fully mapped/fully mapped, Step 5 of Figure 1). Unlike current Hi-C mapping pipelines, after identifying **F-F** pairs, we further process the unmapped paired-end reads to see whether we can still rescue some informative chromatin contacts from them. Our motivation to pursue these reads stems from the striking difference between the number of restriction sites within fully-mapped versus non-mapped reads (Figure 2a). In both the TM3C-1 and TM3C-4 libraries, greater than 70% of the non-mapped reads contain at least one RE cut site, whereas 90% of the mapped reads contain no cut sites for the TM3C-1 library (two sample Kolmogorov-Smirnov test p-values for both TM3C-1 and TM3C-4 are approximately equal to 0). This difference suggests that read ends that fail to map as a whole can still be informative of chromatin contacts because they potentially contain real ligation events leading to chimeric reads. In order to extract this contact information, we further process the read ends containing one restriction site, thereby identifying contacts between a partially mapped read and a fully mapped read (**P-F**) or between two partially mapped reads (**P-P**, Steps 6–8 of Figure 1, Methods). This two-phase mapping strategy not only identifies a greater number of pairwise contacts (doubles) but also allows us to identify contacts involving three or four loci from only one paired-end read. Step 8 of Figure 1 summarizes the different cases arising from the second mapping phase for a read pair that did not qualify as F-F in the first phase. Overall, after excluding intrachromosomal contacts with genomic distance <20 kb, we identify more than 210K triples from our KBM7-TM3C-1 library together with 10.1M and 857K additional pairwise contacts from P-F and P-P type read pairs, respectively (Table 2, Additional file 2). We also investigate the mapping orientations (signs) of ligated fragments that create different contact types (Table 3). The distribution of reads among all possible sign combinations is expected to have a bias for reads that are sonication products (undigested or religated) and to be uniform for de novo chromatin contacts due to ligation events. Table 3 shows this is the case for both the contacts that are identified by traditional Hi-C pipelines (F-F) as well as for the contacts we identify here that produce triples. Since we size select for fragments that are approximately 250 bp, the genomic distance threshold of 1 kb eliminates all sonication products, resulting in uniform distribution for the remaining contacts from TM3C.

§ 2.3 Two-phase mapping rescues contacts informative of genome architecture

Following identification of all three types of contacts (F-F, P-F, and P-P), we evaluate the quality of the resulting contact sets for each library in four ways. First, we confirm that the contact probability between two intrachromosomal loci exhibits a sharp decay with increasing genomic distance for crosslinked libraries but not for the control library when all contact types are pooled (Figure 2b). Second, we observe that this scaling relationship is consistent for different contact types (Figure 2c), and the scaling is log-linear for the genomic distance range of 0.5–7 Mb, consistent with observations from Hi-C data [Lieberman-Aiden et al., 2009]. Third, we confirm visually and quantitatively that the interchromosomal contact maps we obtain from each contact type are consistent with each other (Supplementary Fig. 8, pairwise matrix correlations are 0.997, 0.964 and 0.954 for (F-F, P-F), (F-F, P-P) and (P-F, P-P), respectively) and that the contact maps are consistent with known organizational hallmarks of human genome architecture, such as the increased number of contacts between small chromosomes (16–22 except 18) (Figure 2d–f, Supplementary Fig. 8). Fourth, we confirm that our contact profiles capture known karyotypic abnormalities of KBM7 cells, such as diploidy of chromosome 8 (+8), partial diploidy of chromosome 15, and t(9;22)(q34;q11)) translocation between chromosomes 9 and 22 that leads to Philadelphia chromosome formation [Kotecki et al., 1999, Bürckstümmer et al., 2013] (Figure 2d, e, Supplementary Fig. 9). Normal diploid human keratinocyte (NHEK) cells exhibit no karyotypic abnormalities except higher average contact counts between chromosomes 17, 19 and 22 (Figure 2f). For the non-crosslinked KBM7 control library, only the changes related to copy number (e.g., diploidy) are apparent from the heatmap (Figure 2g). Translocations are not visible in the control because digestion of non-crosslinked chromatin does not preserve genomic distances. Together, these results indicate that TM3C successfully assays genome architecture of human cells and suggests that contacts recovered by our two-phase mapping strategy, which are traditionally discarded from Hi-C analysis, are consistent with traditionally retained contacts. Therefore, for all remaining analyses with pairwise contacts we combine all three types (F-F, P-F, P-P) into an aggregated contact map for each library.

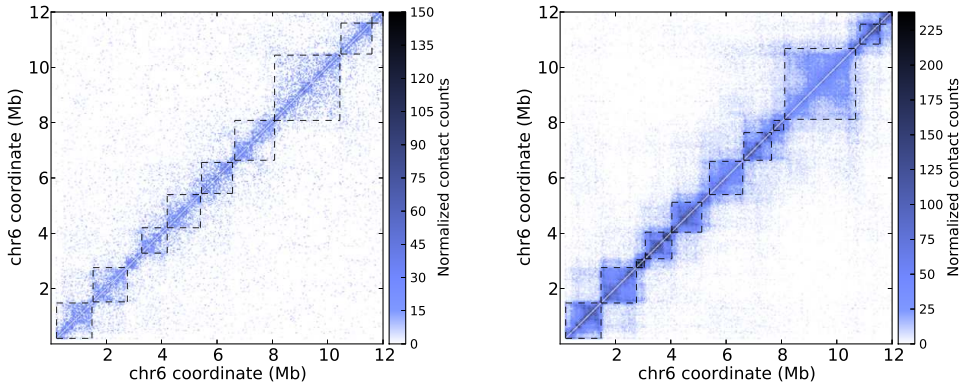
§ 2.4 TM3C data confirms chromatin compartments and topological domains

In addition to evaluating whether results from the TM3C data sets are consistent with polymer models of chromatin folding and karyotypic properties of assayed cell lines, we assess whether TM3C contact maps exhibit the expected compartment-scale and



(A) KBM7

(B) GM06990 [Lieberman-Aiden et al., 2009]



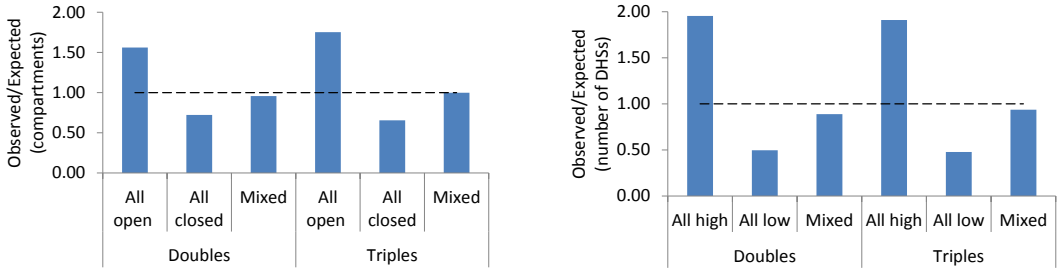
(C) KBM7

(D) IMR90 [Dixon et al., 2012]

FIGURE 3: **Figure 3 - Comparison of TM3C data with existing genome architecture datasets** Eigenvalue decomposition to identify open/closed chromatin compartments of chromosome 17 **(a)** from the KBM7 cell line assayed by TM3C and **(b)** from GM06990 cell line assayed by Hi-C [Lieberman-Aiden et al., 2009]. Topological domain calls and contact count heatmaps of a 6 Mb region of chromosome 6 **(c)** for the KBM7 cell line assayed by TM3C and **(d)** for the IMR90 cell line assayed by Hi-C [Dixon et al., 2012].

domain-scale organization. For this purpose we perform eigenvalue decomposition on our contact maps and compare our compartment calls to those of previous Hi-C data sets on other human cell lines [Lieberman-Aiden et al., 2009, Dixon et al., 2012]. The resulting compartment calls exhibit a nearly perfect overlap for chromosome 17 between KBM7 and GM06990 (Figures 3a–b) and a high level of genome-wide conservation (82%) between these two cell lines. Conservation between pairs of contact maps from the five previously published contact maps ranged between 70–82%.

Similarly, we perform topological domain decomposition at 40 kb resolution on KBM7 contact maps and compare our calls to those of two human cell lines published by Dixon et al. [2012] (Methods). Figures 3c–d demonstrate the significant overlap of



(A) (B)

FIGURE 4: **Figure 4 - Genome-wide characterization of triple contacts (a)** Observed over expected percentages of double and triple contacts that link 1 Mb regions with the same (either open or closed) or different (mixed) compartment labels for the KBM7-TM3C-1 library (Methods). Both double and triple contacts prefer to link open compartments to each other with triples showing slightly more enrichment for this trend. **(b)** Similar percentages as in **(a)** but when 1 Mb windows are segregated according to the number of DHSs they contain (Methods). Contacts linking regions with higher numbers of DHSs than the median number are enriched within the doubles and the triples of the KBM7-TM3C-1 library. Due to lack of DNase data for KBM7 cells, we use data from six other human cell lines for this analysis. Since the results are very similar among different cell lines, here we only plot the results for K562 which is also a leukemia cell line.

topological domain calls from KBM7 and IMR90 contact maps on a 12 Mb region of chromosome 6. Overall, 73% of IMR90 and 72.8% of ESC domain boundaries overlap with the boundaries that we identify for the KBM7 cell line (Fisher's exact test p-values compared to random overlap are $< 10^{-100}$ for each case).

Together, the compartment-scale and domain-scale similarities between our data and previous Hi-C data suggests that TM3C, a simpler protocol, provides similar results to Hi-C and that KBM7, which has a distinct karyotype, preserves the large scale organizational features of other human cell lines.

§ 2.5 Genome-wide characterization of triple contacts

After identifying chromatin compartments at 1 Mb resolution and topological domains at 40 kb resolution for the KBM7 cell line, we evaluate whether the triple contacts identified by TM3C preferentially link regions with the same compartment labels and regions within the boundaries of a topological domain. Figure 4a shows that triple contacts, similar to doubles, are enriched among regions of open chromatin (observed 14.6% compared to expected 8.33%, Methods). Out of all intrachromosomal triples (triples that link three loci on the same chromosome), we see that 16.5% are within the same topological domain. Note that we exclude from this percentage all short range intrachromosomal triples (<20 kb) as well as all those that link at least two loci within the same 40 kb

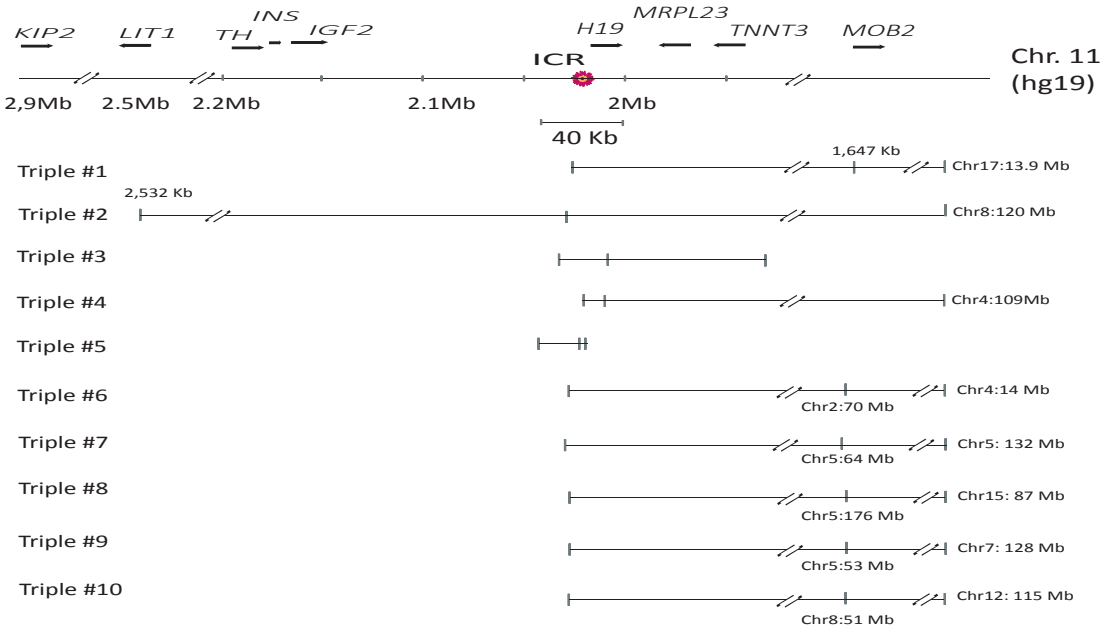
window which would otherwise inflate the reported percentage. We assess the significance of this observed percentage of intradomain triples by generating a null model with 100 shuffled topological domain decompositions for each chromosome (Methods). The median and the mean percentages are both ∼14.1% with a standard deviation of 0.16% for the null model suggesting a statistically significant enrichment of intradomain triples for the observed domain decomposition compared to shuffled configurations (p-value= 0, z-score= 14.67).

Next we carry out an analysis similar to the compartment label analysis described above using the numbers of DNase hypersensitive sites within each 1 Mb window (Methods). Figure 4b shows that, consistent with and slightly surpassing the enrichment for open chromatin compartments, triple contacts as well as doubles are enriched among regions with higher numbers of DHSs (for triples observed 23.7% compared to expected 12.4%, Methods).

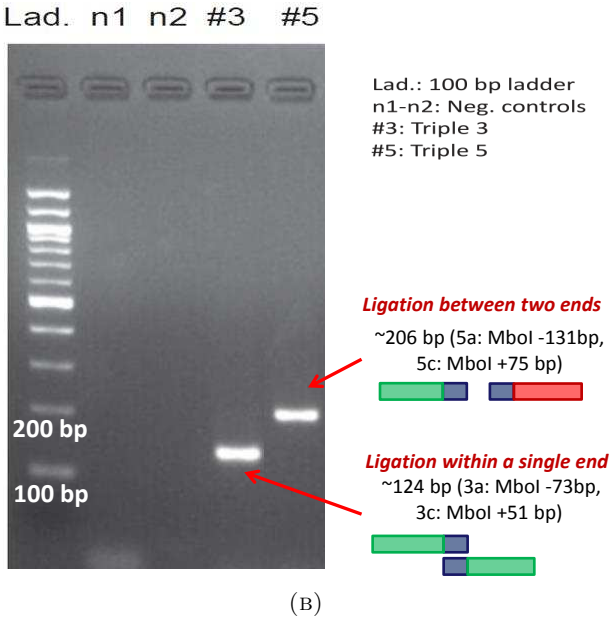
§ 2.6 Verification of triples involving *IGF2-H19* locus

We next investigate whether the multi-locus contacts identified by the TM3C assay correspond to possible combinatorial regulatory interactions in KBM7 cells. Specifically, we focus on triples (contacts involving three loci) involving the *IGF2-H19* locus, which is a classic example of imprinting that leads to allele-specific gene expression and regulation in both mouse and human [Bartolomei et al., 1991, Ling et al., 2006, Vu et al., 2010, Murrell et al., 2004]. Our previous work in human cells has shown that a region that is located just upstream of the *H19* promoter which is differentially methylated between maternal and paternal copies is involved in formation of allele-specific long-range chromatin loops [Vu et al., 2010]. Methylation status of this imprinting control region (ICR) determines whether *IGF2* is transcribed (paternal allele) or not (maternal allele). Because KBM7 cells are haploid for chromosome 11, we expect our TM3C data to be consistent with only one mode of operation of this ICR. Analyzing the triples inferred from KBM7-TM3C-1 data involving the ICR region (±20 kb), we observe contacts that link this ICR region to distal loci on the same chromosome as well as to a *trans* loci on other chromosomes (Figure 5a).

In order to verify these contacts, we design three primers per each triple and perform PCR experiments (Supplementary Table 1). We test whether pairs of forward/reverse primers give rise to PCR products with expected sizes to confirm contacts identified from our two-phase mapping (Figure 5b). For triple 3, we use primers 3a and 3c designed for two loci that are 80 kb away and are linked by a contact found from a ligation occurring within one end of a paired-end read. For triple 5, we use primers 5a and 5c that link two loci that are 24 kb apart on chromosome 11 and are found (one of them



(A)



(B)

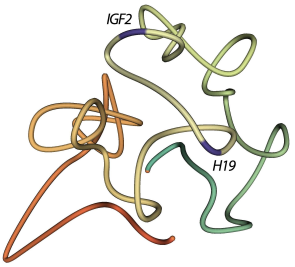
FIGURE 5: **Figure 5 - Validation of triples using PCR (a)** Ten triples extracted from the KBM7-TM3C-1 library that have at least one of their three ends in the 40 kb region surrounding the imprinting control region (ICR) of *IGF2* and *H19* genes. These triples involve short- and long-range contacts within chromosome 11 which are all indicated by tick marks with coordinates in kilobases (kb) displayed only for long-range contacts. Interchromosomal contacts with other chromosomes are indicated by the chromosome identifier followed by the coordinate in megabases (Mb). Orientation of the displayed locus is in the direction of *IGF2* and *H19* transcription. **(b)** PCR verification of pairwise contacts from triples 3 and 5. One pair of forward/reverse primers is used for each gel (Supplementary Table 1).

only partially) in two separate ends of a paired-end read. For both of these cases we observe PCR products near the expected size from our primer design (Supplementary Table 1). Validation of contacts found by our two-phase mapping either within a single end of a read or from two different ends supports the idea that chimeric reads contain information about genuine chromatin contacts.

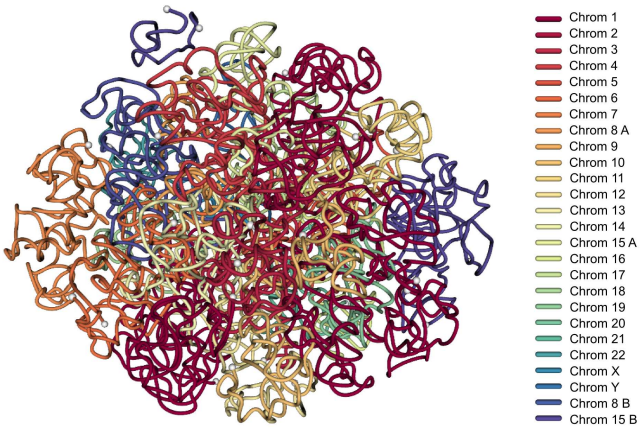
Next, we perform PCR on all the triples shown in Figure 5a using all three primers simultaneously. Out of 10 triples tested, 6 of them (triples 1–6) resulted in either one or more PCR products that have the expected size(s), confirming these contacts (Supplementary Figs. 10, 11, Supplementary Table 1). Detailed analysis of the distal loci that are contact partners of ICR (either interchromosomal or interchromosomal with distance >40 kb to ICR) in these six triples reveal that most of these loci (6 out of 8, Supplementary Figs. 12–14) lie in regions consisting mainly of unmethylated CpGs in K562 cells and mainly methylated CpGs in at least one other cell line assayed by ENCODE [ENCODE Project Consortium, 2012]. These contacts suggest existence of complex chromatin loops that bring together the differentially methylated ICR in 3D with loci that show cell type-specific methylation and specifically unmethylation in K562 cells. These results together with our preliminary methylation analysis of the ICR suggest a 3D organization which silences *IGF2* by restricting enhancer access to its promoter similar to *Igf2* silencing of the maternal copy of mouse chromosome 7 [Qiu et al., 2008]. In order to test our hypothesis that the single copy of chromosome 11 in KBM7 corresponds to the maternal allele, we check the expression status of *H19* and *IGF2* genes from a recently published data set [Bürckstümmer et al., 2013]. Supplementary Fig. 15 shows that *H19* is expressed but *IGF2* is not as we expected. This expression data confirms the prediction from TM3C data for the parent-of-origin of chromosome 11 in KBM7 cells. Furthermore, our 3D model of the 2 Mb region centered on the ICR (Figure 6a) demonstrate that the two genes, expressed *H19* and non-expressed *IGF2*, are placed in distal chromatin domains, consistent with the proposed gene regulation model in the maternal copy of the homologous region in mouse [Murrell et al., 2004]. However, the depth of our data is not sufficient to do a finer scale 3D modeling that can distinguish between allele specific loops established by several differentially methylated regions and CTCF binding sites.

§ 2.7 Three-dimensional modeling of KBM7 genome recapitulates known organizational principles of human cells

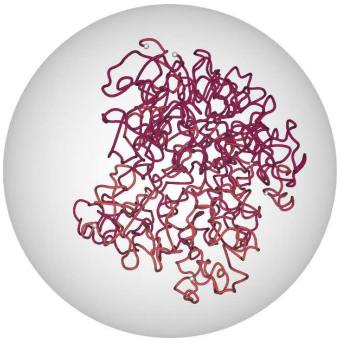
Finally, to visualize the genome architecture of near-haploid KBM7 cells, we generated a set of 3D structures using an optimization framework that alternates between inferring the 3D configuration of beads that best summarize TM3C contacts [Varoquaux et al.,



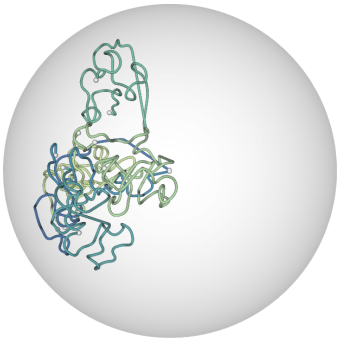
(A) Chromosome 11 1Mb–3Mb



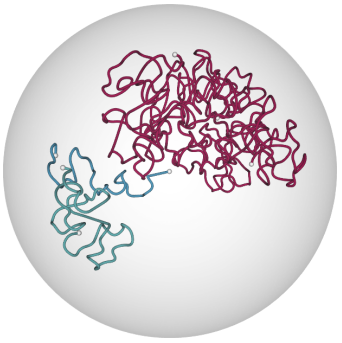
(B) All chromosomes



(C) Chromosomes 1, 2, 3, 4 and 5



(D) Chromosomes 16, 17, 19, 20, 21, and 22



(E) Chromosomes 1, 2, 20 and 21

FIGURE 6: **Three-dimensional modeling of KBM7 genome architecture (a)** Three-dimensional structure of the 2 Mb region of chromosome 11 (chr11:1,000,000-3,000,000) which is centered around *IGF2-H19* imprinting control region. This structure is inferred from normalized contact counts of KBM7-TM3C-1 data at 40 kb resolution using the Poisson model from Varoquaux et al. [2014]. **(b)** Three-dimensional structure of the KBM7 genome, which is haploid for all chromosomes other than diploid chromosome 8 (8A, 8B) and partially diploid chromosome 15 (15A, 15B) (see Methods for details of the 3D inference). Different colors represent different chromosomes, and white balls represent chromosome ends. Same 3D structure as **(b)** when confined to **(c)** only a subset of long chromosomes, **(d)** only a subset of small chromosomes, **(e)** two small and two large chromosomes.

2014] and re-estimating the distribution of contact counts between diploid chromosomes. Since this optimization is non-convex, we ran the optimization 1000 times and selected the 100 structures with the highest log likelihoods (Methods). Figure 6b–e plots the structure with highest likelihood inferred at 1 Mb resolution. Visual observation of Figure 6b suggests that individual chromosomes preserve their territories in 3D (see also Additional File 3). In order to better visualize which chromosomes are closer to each other, we plot subsets of different chromosomes in Figures 6c–e. Consistent with previous models [Lieberman-Aiden et al., 2009] and our contact count heatmaps, we observe strong colocalization among the small gene-rich chromosomes (16, 17, 19, 20, 21 and 22). However, chromosome 18, which is small but gene-poor, does not colocalize with gene-rich small chromosomes in 3D (Figure 6c, Supplementary Fig. 16). We also observe colocalization of large chromosomes with each other, but not as strongly as small chromosomes (Figure 6d). Visualization of two large and two small chromosomes clearly demonstrates that the two sets of chromosomes are far from each other in our 3D models (Figure 6e).

§ 3 Discussion

Catalyzed by the availability of genome-wide chromatin architecture data generated using chromatin conformation capture assays, the field of regulatory genomics has recently witnessed increased interest in the functional role of higher order DNA structure. Organizational principles of eukaryotic nuclei that are uncovered by these genome wide assays range from large scale patterns such as open/closed chromatin compartments [Lieberman-Aiden et al., 2009] and topological domains [Dixon et al., 2012] to more local patterns such as silencing or activating of individual genes by altering the 3D proximity of enhancers to gene promoters [Ferraiuolo et al., 2010, Li et al., 2012]. However, one important question that remains to be answered is how the simultaneous proximity of more than two loci in the nucleus impacts gene regulation. Current conformation capture assays cannot address this question because they only characterize pairwise contacts that involve exactly two loci.

Here we demonstrated how to discover simultaneous multi-locus contacts using a straightforward conformation capture assay. We aimed at distinguishing between proximity of multiple loci measured from different nuclei in the form of pairwise contacts and simultaneous proximity between these loci within a single nucleus. We showed that our TM3C assay, which can employ more than one restriction enzyme at a time to increase chromatin digestion, results in chimeras even within a single end of a short paired-end

read. Accordingly, we developed a two-phase mapping pipeline that uses cleavage information to extract from these chimeras informative contacts that involve two, three or four loci. An additional advantage of TM3C is that it is significantly simpler and yet provides increased resolution for the resulting contact maps compared to current Hi-C assays.

It is important to note, however, that there are two drawbacks to our assay compared to traditional Hi-C or TCC assays. The first drawback is a tradeoff between resolution and the noise level of the data. Frequent digestion of chromatin with multiple 4-cutters increases the resolution but also the noise level of the data, as measured by the ratio between inter and intrachromosomal reads (Additional file 2). The second drawback is a tradeoff between the simplicity of the assay and the proportion of informative reads from the paired-end sequencing. In the TM3C assay we omit the steps of RE overhang biotinylation and streptavidin pull-down which are present in both the Hi-C and TCC assays. This omission results in a higher percentage of sonication products (non-informative read pairs) in the sequencing libraries of TM3C (Additional file 2) which we discard after read mapping.

Despite these drawbacks, we believe that TM3C is an effective assay in profiling genome architecture—evident by the consistency of our results with characteristic features of genome organization—with the added benefit of revealing multi-locus contacts. In order to demonstrate the utility of TM3C, we applied it to two human cell lines. We specifically chose one of these cell lines as the near-haploid KBM7 which has been used in settings where having multiple copies of a chromosome is problematic, such as loss-of-function genetic screens [Carette et al., 2009, Bürckstümmer et al., 2013]. We first established that TM3C contact maps are consistent with karyotypic features of KBM7 and that KBM7 cells share common large scale organization with other mammalian cell lines previously assayed by Hi-C. Focusing on a well-studied locus (*IGF2-H19*) that has been shown to be involved in parent of origin specific long-range chromatin loops, we showed that TM3C identifies multi-locus contacts (triples), more than half of which were validated using PCR. Confirmed triples involved intrachromosomal loops bringing together regions that are more than megabases away in genomic distance as well as regions from different chromosomes. Together with results from previous FISH experiments that reveal *IGF2* is located outside of its chromosome territory in the majority of nuclei [Mahy et al., 2002], our findings suggest that complex regulation of *IGF2* and *H19* may involve interactions with multiple distal regions simultaneously.

Another important aspect of our work is the modeling of 3D organization of a human cell line without averaging data from multiple copies of a chromosome or resolving the haplotype. To date 3D modeling efforts on the human genome have been limited to

haploid chromosomes such as the X chromosome in male cells [Nagano et al., 2013], one chromosome or one portion of a chromosome at a time [Nagano et al., 2013, Bau et al., 2011] or have assumed artificially that only one copy of each chromosome exists per cell [Zhang et al., 2013]. In this work the near-haploid karyotype of KBM7 allowed us to overcome these limitations to infer whole-genome 3D models. By extending an algorithm that we developed previously for haploid genomes [Varoquaux et al., 2014] to handle the diploid portions of KBM7 cells, we generated 3D models for this leukemia cell line. Due to the lack of independent data available on KBM7 cells, we were unable to verify our 3D models further or correlate them with features such as histone modifications and transcription binding. However, our models are consistent at the large scale with previous observations that suggest chromosomes with similar sizes tend to be closer to each other in 3D. It is also important to note that, similar to many previous approaches, our 3D models are consensus structures that summarize the genome architecture of a cell population. Capturing the heterogeneity of genome architecture across cells may be possible in the future, especially in conjunction with single-cell techniques [Nagano et al., 2013].

Overall, we showed that TM3C provides a framework to identify multi-locus contacts genome-wide in conjunction with commonly used next generation sequencing platforms that produce short paired-end reads (e.g., 100 bp Illumina). We believe that with broader use of longer reads (e.g., Pacific Biosciences) TM3C will be able to profile a larger number of multi-locus contacts with higher signal-to-noise ratio. Such profiling is important in understanding better the combinatorial regulation of gene expression and complex chromatin loops that involve more than two loci simultaneously.

§ 4 Conclusion

TM3C is a simple protocol for ascertaining genome architecture and can be used to identify simultaneous contacts among three or four loci. Application of TM3C to a near-haploid human cell line revealed large-scale features of chromosomal organization and multi-way chromatin contacts that preferentially link regions of open chromatin.

§ 5 Materials and methods

§ 5.1 TM3C library generation

Approximately six million NHEK and ten million KBM7 cells were fixed in 1.5% formaldehyde at room temperature for 10 minutes. The fixed cells were washed with TN buffer

(10 mM Tris, 40 mM NaCl, pH 7.5) and collected by centrifugation at 600 g for 3 minutes. To increase digestion efficiency, fixed cells (6 or 10 million / 122 ul) were treated with SDS (add 3.8 ul of 10% SDS to a final of 0.30% SDS) at 64°C for 10 minutes and then at 37°C overnight (15 hours). The SDS concentration was reduced gradually to 0.10% by adding five times of 50 ul (1 x DpnII digestion buffer or NEB buffer 4 for multiple enzymes) with mixing. Triton X-100 (38 ul of 20% Triton X) was added to 1.8% concentration and the sample was incubated at 37°C for 1 hour. Sample volume was adjusted to 600 ul by adding 1 X restriction buffer, ATP (0.2 mM final) and BSA (100 ug/ml final). Digestion with appropriate restriction enzymes (300 units each) was carried out on a rotate shaker at 37°C for 15 hours. We used high concentration NEB enzymes to keep the final volume of the enzyme mixture less than 60 ul (1/10 reaction volume).

The digested samples were deactivated at 65°C for 15 minutes and then centrifuged at 15,000 g for 5 min. We recovered ∼95% of cellular DNA in the pellet fraction. The pellet fraction was re-suspended with T4 ligation buffer (15 ul 10 x buffer, 65 ul total) heated at 65°C and mixed with 100 ul of melted 2.5% low-melting agarose. We used 200 ul pipette to deliver the hot agarose sample to ice-cold ligation buffer (800 ul of 1 x ligation buffer containing T4 ligase (4000 units, NEB) in a steady fashion within ∼5 seconds, on melted ice. Strings of gel bead appeared instantly at 0°C. We sealed the tube with parafilm and perform ligation at RT (23°C.) overnight on top of a shaker (∼300 rpm), then transfer the tube to a iced water bath.

The sample pellet was recovered by centrifugation at 20,000 g for 2 minutes, then 10 ul of 1% SDS (0.05% final) was added and heated at 80°C for 1 hour. Cross-links were reversed by treatment with Proteinase K (200 ug/ml) at 65°C and 300 rpm overnight (12 hours). Melted TM3C-agarose sample was incubated with RNase A (10 ug / 210 ul) at 55°C for 15 minutes and then purified by QIAquick gel extraction protocol (QUIAGEN Inc., CA). Purified TM3C DNA was quantified using both a NanoDrop spectrophotometer (Thermo Scientific) and a Qubit 2.0 Fluorometer. The Qubit quantification represents the more accurate DNA concentration.

§ 5.2 First phase mapping of sequence data

We mapped the paired-end reads to the human reference genome (hg19) using the short read alignment mode of BWA (v0.5.9) with default parameter settings. Each end of the paired reads was mapped individually. We post-processed the alignment results to extract the reads that satisfy the following three criteria: (i) mapped uniquely to one location in the reference genome, (ii) mapped with an alignment quality score of at least 30, (iii) mapped with an edit distance of at most 3. Reads that satisfy these criteria are

named *fully-mapped* (**F**), and the rest of the mapped reads that did not satisfy these criteria are discarded from further analysis. We identified pairs of fully-mapped reads that share a common identifier to generate the set of contacts that we denote as **F-F** (fully-mapped - fully-mapped). The reads that did not map to any location in this phase of mapping are named *non-mapped* and are analyzed further.

§ 5.3 Second phase mapping of non-mapped reads

Re-mapping the reads that are deemed *non-mapped* in the initial mapping is necessary to avoid discarding a significant number of informative reads for an assay such as TM3C that uses a frequently cutting restriction enzyme (or enzymes) for digestion. Due to the high frequency of cleavage sites in the genome, TM3C is highly likely to capture ligations between DNA fragments from two different loci in a single end of a read. We call each such read *chimeric* because the sequences do not come from a continuous piece of DNA but instead from two loci that are in proximity in the three-dimensional space. Therefore, for these chimeric ends, after splitting into smaller fragments from the cleavage sites of the restriction enzymes used in the digestion step, we applied a second phase of mapping.

Within each non-mapped read, we first counted the number of cleavage sites, taking into account all the restriction enzymes that are used in the digestion step for that specific library. We discarded reads that contain more than two cleavage sites. We also discarded reads that contain no cleavage sites because such reads surely are not chimeric. We split the remaining reads that contain only one cleavage site into two smaller fragments, preserving the entire cleavage site on both adjacent fragments. We mapped the two resulting fragments to the genome using BWA with default parameter settings. The 3-point filtering criteria mentioned in the previous section are applied to the aligned reads, but allowing an edit distance of at most 1 to make sure we only extract the unique and high quality mappings. The reads that are extracted from this phase of mapping are named *partially-mapped* (**P**) because they did not map as a whole, but their constituent fragments were successfully mapped to different loci. The two classes of mapped reads (fully-mapped (**F**) and partially-mapped (**P**)) yield three possible types of contacts, namely **F-F**, **F-P** and **P-P**. The first set (F-F) is extracted after the initial mapping in which each paired-end read can contribute at most one interaction between two loci. The second set (P-F) consists of paired-end reads with one end fully mapped and the other end having either one or two smaller fragments that mapped to the genome. If the latter contains only one mapped fragment, then the only interaction is between this fragment and the fully-mapped end. However, if the end has two mapped fragments, then this paired-end read produces three contacts: one between the two

mapped fragments on the partially-mapped end and two others that have one side from a fragment from the partially-mapped end and the other side from the fully-mapped end. In addition, the same paired-end read produces one triple (i.e., interaction among three loci) of type P-F. For the contacts of the third type (P-P), each paired-end can produce either one, three or six pairwise contacts, depending on whether one or two fragments from each end are successfully mapped. If only one fragment from one end and two from the other is mapped, then, similar to the case of P-F, three pairwise contacts and one triple is produced. If both ends have two mapped fragments, then six pairwise contacts, four triples (of type P-P) and one quadruple (i.e., contact among four loci) are produced.

§ 5.4 Normalization of contact maps

For each possible pair of 1 Mb loci, we refer to the total number of read pairs that link the two loci as the *contact count*, and we refer to the two-dimensional matrix containing these contact counts as the *raw contact map*. To normalize the 3113×3113 raw contact maps, we extended the iterative correction procedure, ICE [Imakaev et al., 2012], for a nearly haploid genome. First, we corrected for the bias caused by the partial diploidy of the genome. For that, we constructed a "deduplicated" contact counts matrix, where contact counts associated with diploid loci are divided into two equal parts, each of which is associated with one of the homologous chromosomes. Contact counts between two different copies of diploid chromosomes/regions are set to 0. The deduplicated matrix is akin to an artificially created allele-specific contact counts matrix, where homologous chromosomes interact in identical ways and do not interact with each other. As a preprocessing step, we ranked loci by their percentage of intrachromosomal contacts with zero counts and filter out the top 10% of this list. This filtering removes all loci for which the signal to noise ratio is too low (typically, regions of low mappability). Last, we applied ICE, a method that attempts to eliminate systematic biases in Hi-C data. ICE assumes that the bias for each entry can be decomposed as the product of the biases associated with each locus, and estimates a bias vector β under the equal visibility hypothesis: the coverage of counts should be uniform. The tensor product $\beta \otimes \beta$ generates a bias matrix that can be used to convert the raw contact map into a normalized contact map. To generate a contact count matrix of the original size, we summed all counts from homologous chromosomes associated with the same loci. This procedure yields a (3113×3113) contact counts matrix for which diploid loci interact twice as much as haploid loci.

§ 5.5 Eigenvalue decomposition

We carried out eigenvalue decomposition on the normalized contact maps of KBM7 and NHEK TM3C datasets as described in Lieberman-Aiden et al. [2009]. For each chromosome we used the intrachromosomal contact matrices at 1 Mb resolution. We calculated the Pearson correlation between each pair of rows of the contact matrix and apply eigenvalue decomposition (using the eig function in MATLAB) to the correlation matrix. The sign of either the first or the second eigenvector defines chromosome compartments for each chromosome. Similar to Lieberman-Aiden et al. [2009], we used the second eigenvector in cases where the first eigenvector values are either all positive or all negative. To map signs of eigenvectors to open/closed compartment labels we used GC content as a marker. For each chromosome the sign with higher GC content is selected as open chromatin. We then compared the percentage of 1 Mb bins that are assigned the same compartment label by TM3C data versus previously published Hi-C data in four human cell lines (H1-hESC, IMR90 [Dixon et al., 2012]; K562, GM06990 [Lieberman-Aiden et al., 2009]).

§ 5.6 Topological domain analysis

We identified topological domains using a previously described hidden Markov model-based software tool [Dixon et al., 2012]. To facilitate direct comparison with the previously published topological domains in human cell lines, we carried out the domain calling for these published datasets using the human GRCh36/hg19 assembly. We applied the topological domain calling on normalized contact maps of our TM3C data at 40 kb resolution. To measure the consistency between the topological domains inferred from TM3C and those from published Hi-C data, we calculated the overlap of domain boundaries obtained between these two assays. We deemed two boundaries, one from each assay, as overlapping if they overlap by at least 1 bp or are adjacent to each other, as described in Dixon et al. [2012].

§ 5.7 Contacts among regions with the same compartment label

We used compartment labels assigned by the eigenvalue decomposition as described above and computed the number of read pairs that define double and triple contacts between two or among three regions all with the same compartment label (all open or all closed) or at least two with opposite labels (mixed). We used only interchromosomal doubles and interchromosomal triples (linking three different chromosomes) for this analysis and eliminated regions that have less than 50% uniquely mappable bases. We

then computed the number of all possible pairs and triples of 1 Mb windows and segregated this number into three groups (all open, all closed, mixed) giving us the expected percentages of contacts that should fall into each group. With exactly equal numbers of open and closed compartments for each chromosome, these percentages would be 25%, 25%, 50% for pairs of compartments and 12.5%, 12.5%, 75% for triples of compartments for the groups of all open, all closed and mixed, respectively. We then reported the ratio between the percentage of observed double and triple contacts to expected percentages within each of these three groups. A ratio >1 represents an enrichment for the observed contacts for that compartment label group.

§ 5.8 Contacts among regions with similar numbers of DHSs

We performed an analysis similar to the compartment label analysis described above using joint (UW–Duke) DNase hypersensitivity peak calls for the six Tier 1 cell lines (GM12878, H1-hESC, HeLa-S3, HepG2, HUVEC, K562) downloaded from http://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/openchrom/jan2011/fdrPeaks. Since there is no DNase data for KBM7 we reported results for only K562 which is also a leukemia cell line. We computed for each 1 Mb window with mappability of at least 50% the number of DHS peaks that overlap with this window. We sorted all these windows by decreasing number of DHSs and labeled the top 50% as "high" and bottom 50% as "low" DNase sensitivity. We then calculated and reported the expected over observed percentage of doubles and triples as described for compartment labels.

§ 5.9 Contacts within the same topological domain

After carrying out the topological domain calling using our KBM7-TM3C-1 data, we computed the percentage of intrachromosomal doubles and triples that link loci within the same topological domain. To estimate the significance of the observed percentages, we randomly shuffled topological domains by preserving the distribution of the domain lengths for each chromosome arm as described in Ay et al. [2014a]. We reported the mean and the standard deviation for the percentage of within domain doubles and triples across 100 randomized shufflings.

§ 5.10 Inference of the 3D structure

We modeled each chromosome as a series of beads on a string, spaced approximately 1 Mb apart. We denote by $\mathbf{X} = (x_1, \ldots, x_n) \in \mathbb{R}^{3\times n}$ the coordinate matrix of the

structure, where n denotes the total number of beads in the genome including the newly introduced chromosomes 8B and 15B ($n = 3289$ for the KBM7 genome), and $x_i \in \mathbb{R}^3$ represents the 3D coordinates of the i-th bead. Contacts from TM3C data can be summarized as an $m \times m$ matrix $\mathbf{c}$, where each entry c_{kl} corresponds to the observed contact count between loci k and l. Because contact information does not distinguish between homologous chromosomes, m only includes one copy of each chromosome and $m < n$. For loci in diploid regions, the contact counts are the sum of contact counts due to each copy of the region. If we denote by $\Phi : [1, n] \to [1, m]$ the mapping that associates a bead i to a locus $\Phi(i)$ of the contact count matrix, this means that the contact count c_{kl} between loci k and l is the sum of counts due to interactions between beads in $\Phi^{-1}(k)$ and $\Phi^{-1}(l)$. For any two beads i and j mapping respectively to loci $k = \Phi(i)$ and $l = \Phi(j)$, let us denote by $0 \leq \mu_{ij} \leq 1$ the proportion of counts in c_{kl} due to interactions between beads i and j. Since all contact counts must be accounted for by interactions between beads, we must have for any loci k and l:

$$\sum_{i \in \Phi^{-1}(k), j \in \Phi^{-1}(l)} \mu_{ij} = 1 \,.$$

We propose to jointly infer the structure $\mathbf{X}$ and the distributions of contact counts μ_{ij}'s by maximizing the likelihood of the observed contact counts. For that purpose, we modeled the contact frequencies $(\mu_{ij} c_{\Phi(i)\Phi(j)})_{(i,j)\in\mathcal{D}}$ ($\mathcal{D}$ is the set of non-zero contact counts) as independent Poisson random variables, where the Poisson parameter of $\mu_{ij} c_{\Phi(i)\Phi(j)}$ is a decreasing function of the Euclidean distance $d_{ij}(\mathbf{X})$ between beads i and j. Our and others' previous work suggested that the relationship between $\mu_{ij} c_{\Phi(i)\Phi(j)}$ and d_{ij} is approximately of the form $d_{ij}(\mathbf{X})^{\alpha}$, with $\alpha = -3$ [Lieberman-Aiden et al., 2009, Fudenberg and Mirny, 2012, Varoquaux et al., 2014]. We can then express the likelihood of the model as:

$$\ell(\mathbf{X}, \mu) = \prod_{i,j} \frac{(d_{ij}^{\alpha})^{\mu_{ij} c_{\Phi(i)\Phi(j)}}}{(\mu_{ij} c_{\Phi(i)\Phi(j)})!} \exp(-d_{ij}^{\alpha}) \,. \tag{D.1}$$

To infer the position of each bead, we maximized the log likelihood of the model which is:

$$\mathcal{L}(\mathbf{X}, \mu) = \sum_{i,j} \mu_{ij} c_{\Phi(i)\Phi(j)} \alpha \log(d_{ij}) - d_{ij}^{\alpha} - \log(\mu_{ij} c_{\Phi(i)\Phi(j)}!) \,. \tag{D.2}$$

In practice, we solved the following relaxation since $\mu_{ij} c_{\Phi(i)\Phi(j)}$ may not have integer values

$$\mathcal{L}(\mathbf{X}, \mu) = \sum_{i,j} \mu_{ij} c_{\Phi(i)\Phi(j)} \alpha \log(d_{ij}) - d_{ij}^{\alpha} - \log(\Gamma(\mu_{ij} c_{\Phi(i)\Phi(j)} + 1)) \,, \tag{D.3}$$

with the following constraints:

- $d_{ij} \leq d_{max}$. To find a suitable d_{max}, we first computed the expected distances $c_{i,i+1}^{-1/3}$ for adjacent beads of haploid chromosomes. We set d_{max} to the 97% quantile, thus excluding outliers values arising in the normalization procedure.

- $0.3 \leq \mu_{ij} \leq 0.7$, where i and j corresponds to loci from the same copy of a diploid chromosomes.

To optimize this non-convex function, we iterated between two steps: (1) infer the 3D structure $\mathbf{X}$; (2) re-estimate the distribution of contact counts μ_{ij} between diploid chromosomes. The first step is solved using an interior point method, as described in Varoquaux et al. [2014]. For the second step, the optimization problem can be performed with respect to each pair of loci k and l independently. Thus we perform a grid search on $\{\mu_{ij} | \Phi(i) = k, \Phi(j) = l\}$, with a step size of 0.01.

We ran the optimization 1000 times varying the initialization of the distribution of the contact counts, and another 1000 times varying the initial structure $\mathbf{X}$. We then selected the top 100 structures with the highest log likelihoods.

§ 6 List of abbreviations used

3C: chromatin conformation capture, TM3C: tethered multiple 3C, TCC: tethered 3C, ICR: imprinting control region, PCR: polymerase chain reaction, RE: restriction enzyme, Ph+: Philadelphia chromosome positive, CpG: Cytosine—phosphate—Guanine, DHS: DNase hypersensitive site, TSS: transcription start site.

§ 7 Tables

Table 1 - Summary of datasets generated in this paper.

Cell Type	**Tethering**	**Restriction Enzymes (REs)**				**Identifier**
		AluI AG|CT	**MboI/DpnII** |GATC	**MspI** C|CGG	**NlaIII** CATG|	
NHEK	Yes		✓			NHEK-TM3C-1
KBM7	Yes		✓			KBM7-TM3C-1
KBM7	Yes	✓	✓	✓	✓	KBM7-TM3C-4
KBM7 (gDNA)	No	✓	✓	✓	✓	KBM7-MCcont-4

Table 2 - Summary of informative pairwise and multi-locus contacts for each KBM7 library.

Library	Total Reads	Doubles (pairwise)	Triples	Quadruples
KBM7-TM3C-1	95,000,000	14,830,477 (15.61%) inter: 8,036,033 intra: 6,794,444	211,249 (0.22%) inter: 92,959 intra: 28,930 mixed: 89,360	1,676 (0.002%) inter: 672 intra: 38 mixed: 966
KBM7-TM3C-4	72,800,218	13,858,985 (19.04%) inter: 11,544,137 intra: 2,314,848	816,625 (1.12%) inter: 594,052 intra: 22,787 mixed: 199,786	25,158 (0.034%) inter: 15,889 intra: 85 mixed: 9,184

Table 3 - Summary of intrachromosomal read orientations for different contact types (KBM7-TM3C-1).

Contact Type	Genomic Dist.	Read Orientations (end1/end2)			
		+/+	+/-	-/+	-/-
Doubles (F-F)	All	1.8%	48.2%	48.2%	1.8%
	> 1 kb	24.9%	25.1%	25.1%	24.9%
		+/++,-/--	+/+-,-/-+	+/-+,-/+-	+/--,-/++
Triples (F-P)	All	0.1%	49.7%	0.2%	50%
	> 1 kb	24.5%	25.8%	25.3%	24.4%
		++/+,--/-	++/-,--/+	+-/+,-+/-	+-/-,-+/+
Triples (P-F)	All	0.2%	49.9%	49.7%	0.2%
	> 1 kb	25.6%	24.1%	25.4%	25.0%

§ 8 Supplementary Figures

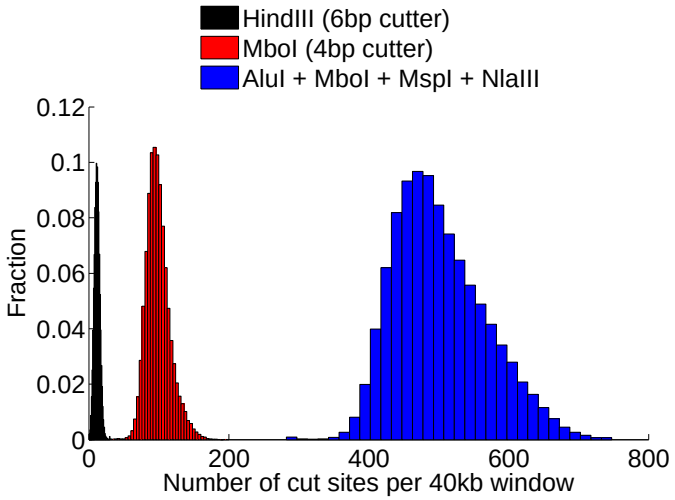
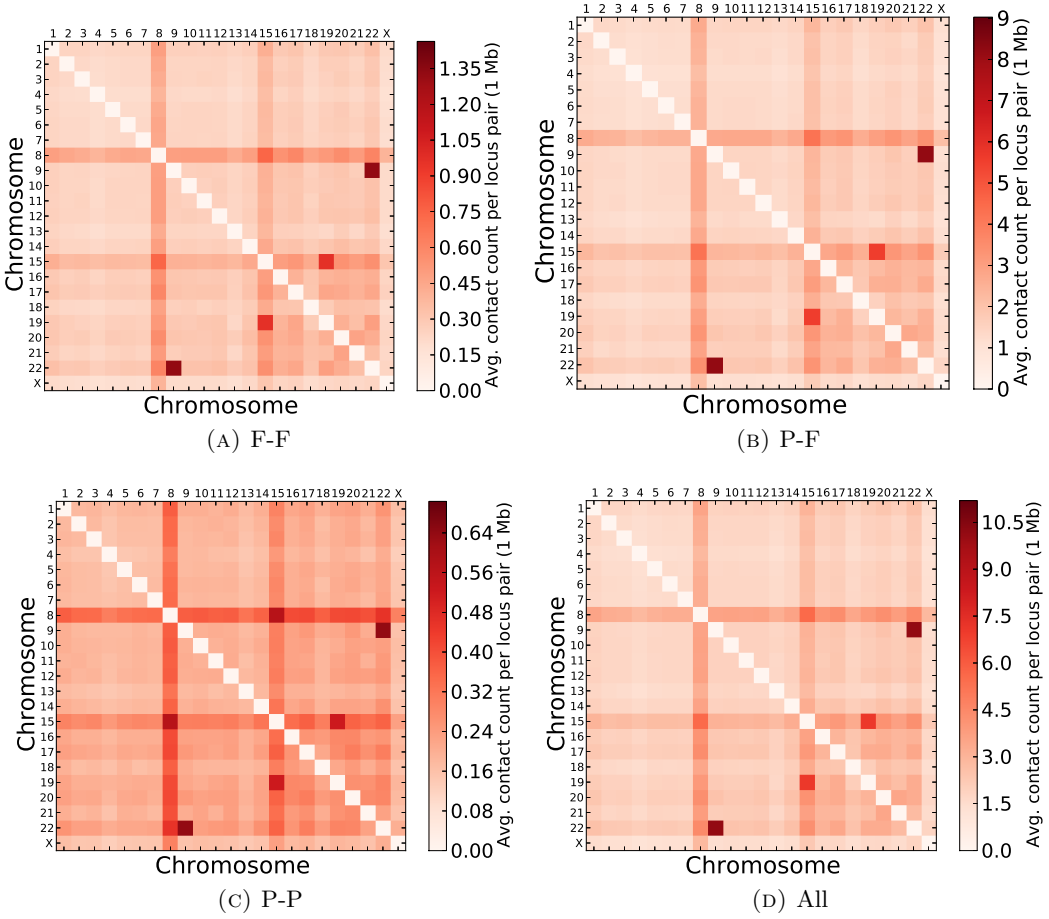


FIGURE 7: **Number of restriction enzyme cut sites across the human genome.** Histograms of the number of cut sites within 40 kb windows across the human genome for each digestion system. HindIII, the most commonly used RE for creating Hi-C libraries, recognizes a 6 bp cut site, whereas AluI, MboI, MspI and NlaIII all cleave from 4 bp cut sites. For TM3C-1 libraries only MboI was used. For TM3C-4 libraries all four 4 bp cutters were used together to digest crosslinked chromatin. The theoretical resolution that can be achieved by each digestion system is inversely proportional to the RE cut site frequency. The mean number of cut sites per 40 kb suggests that TM3C-1 (99.5) and TM3C-4 (501.7) can achieve around 9 and 43 times higher resolution compared to using HindIII (11.7).



(A) F-F (B) P-F

(C) P-P (D) All

FIGURE 8: **Chromosome contact maps of different contacts types for KBM7-TM3C-1.**

Pairwise raw contact counts are averaged over all pairs of mappable 1 Mb windows between the two chromosomes. Contacts that are of type (a) fully mapped/fully mapped (both ends mapped completely), (b) partially mapped/fully mapped (one end mapped completely and one end mapped after second phase of mapping) and (c) partially mapped/partially mapped (both ends mapped after second phase of mapping) are plotted separately. For (d) we aggregated all these three types of contacts.

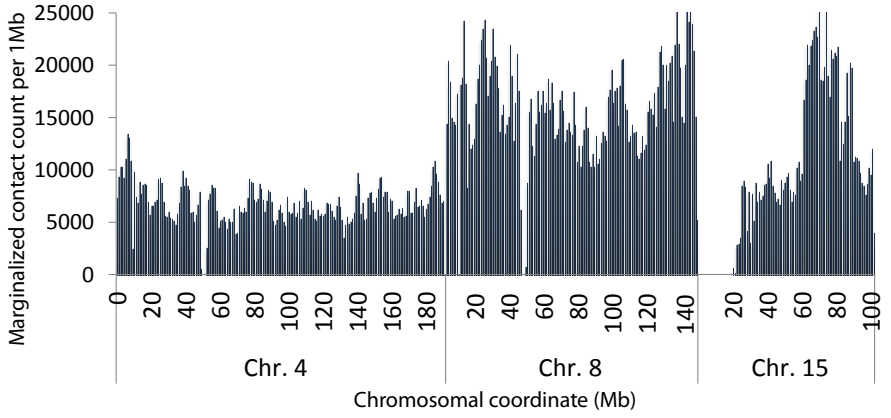


FIGURE 9: **Ploidy track for select chromosomes from KBM7 TM3C data.** Plot of total contact count from each 1 Mb region to all other regions (both intra and interchromosomal) in the genome for a haploid (chr. 4), a diploid (chr. 8) and a partially diploid (chr. 15) chromosome. The diploid region of chromosome 15 is approximately 29 Mb (61–90 Mb).

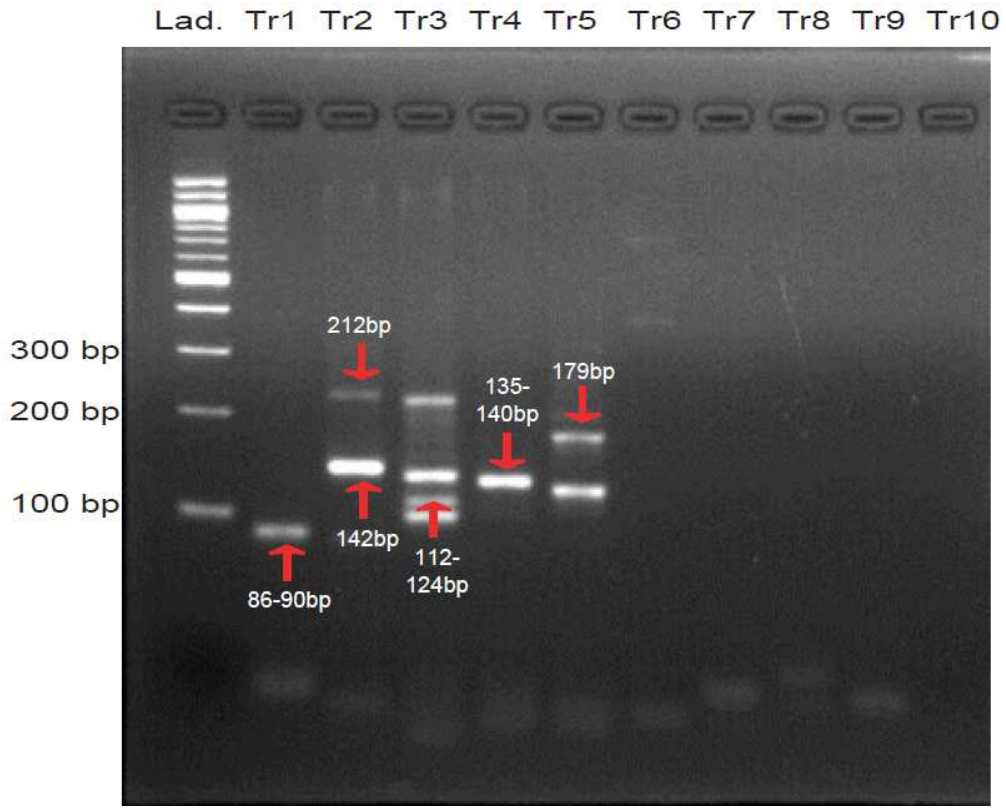
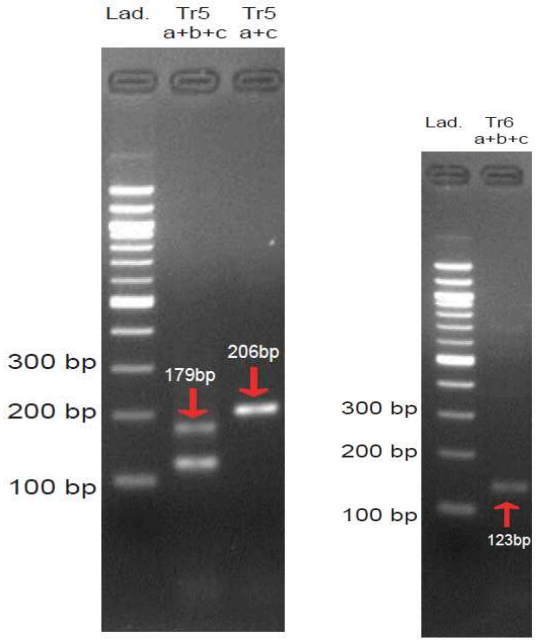


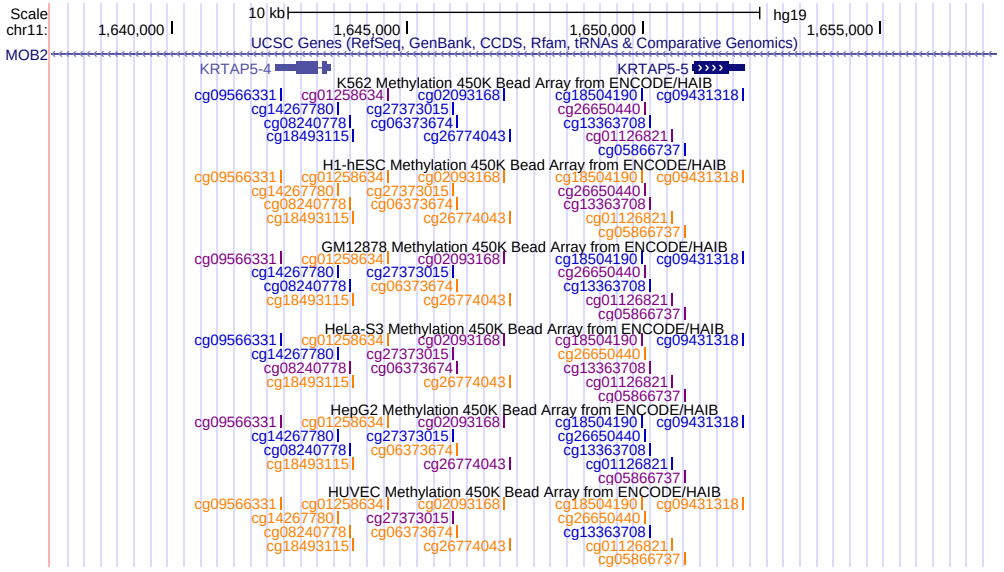
FIGURE 10: **PCR verification of triples 1–10 listed in Main Figure 5.**
Lanes are: 100 bp ladder, triples 1 to 10. For each experiment all three primers (e.g., 1a+1b+1c) designed for that triple are used simultaneously. Expected product sizes for each triple are listed in Supplementary Table 1 and corresponding PCR products with approximate sizes are indicated by red arrows. Triples 1–5 showed PCR products in this gel, and triple 6 showed one product in another gel (Supplementary Fig. 11).



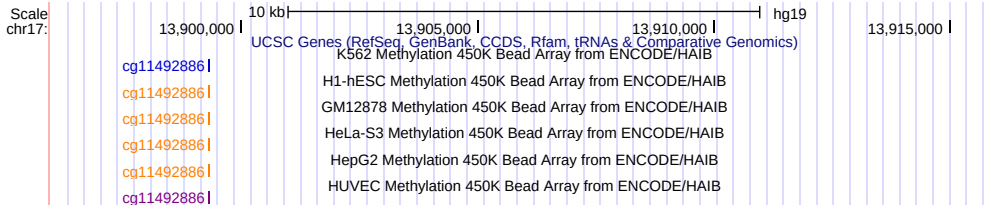
(A) Triple 5 only

(B) Triple 6 only

FIGURE 11: **Additional PCR experiments for triples 5 and 6.** (a) Triple 5 PCR gel. Lanes are: 100 bp ladder, 5a+5b+5c primers and only 5a+5c primers. (b) Triple 6 PCR gel. Lanes are: 100 bp ladder and 6a+6b+6c primers. Expected product sizes for each case are listed in Supplementary Table 1 and corresponding PCR products with approximate sizes are indicated by red arrows.



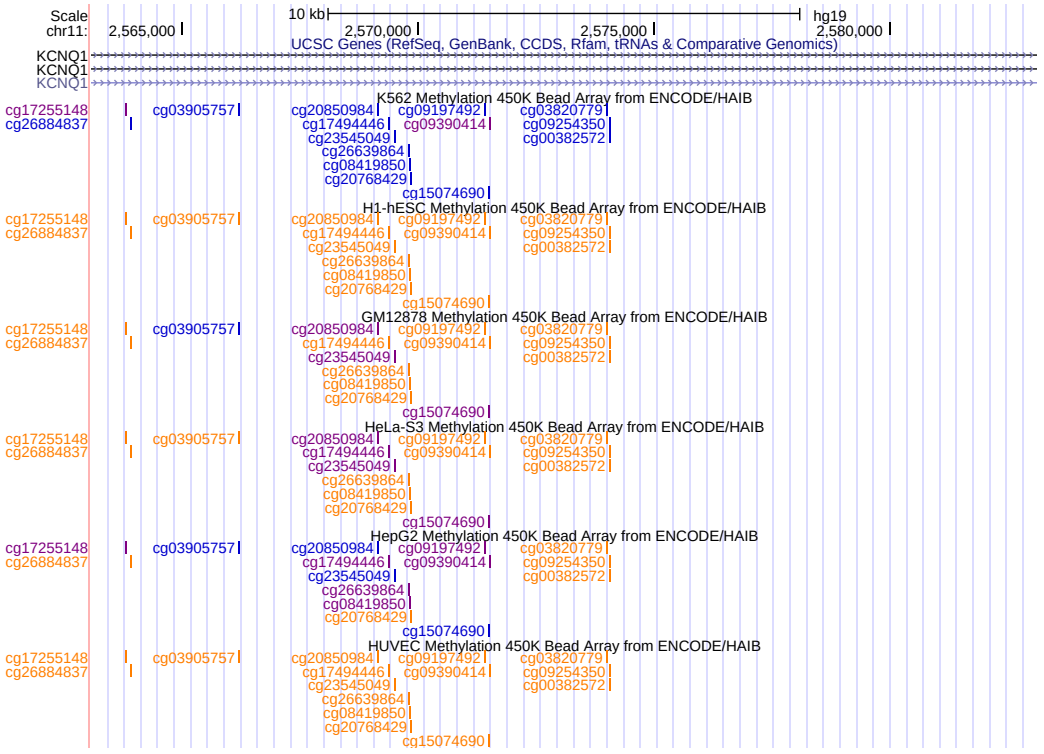
(A) Triple 1–end 1



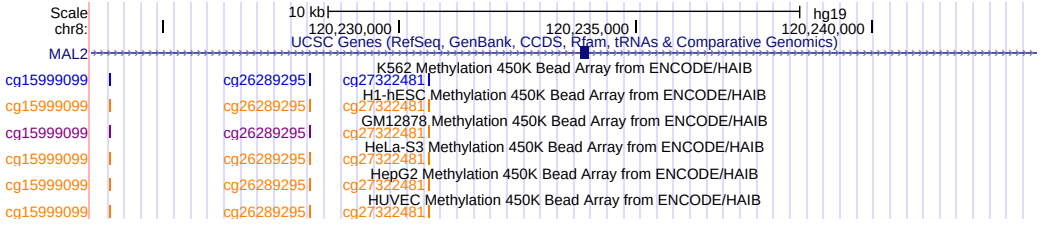
(B) Triple 1–end 3

FIGURE 12: **Methylation status of the distal contact partners of *IGF2-H19* ICR for triple 1.**

UCSC genome browser snapshots of the tracks that display the methylation status of CpG dinucleotides in six ENCODE cell types for the two loci that are distal contact partners of ICR in triple 1. Methylation scores are color coded with orange for "methylated", purple for "partially methylated" and blue for "unmethylated". The figure displays a 20 kb region centered on (a) triple 1–end 1 located on chromosome 11, and (b) triple 1–end 3 located on chromosome 17. End 2 is located within 40 kb of the ICR.



(A) Triple 2–end 2

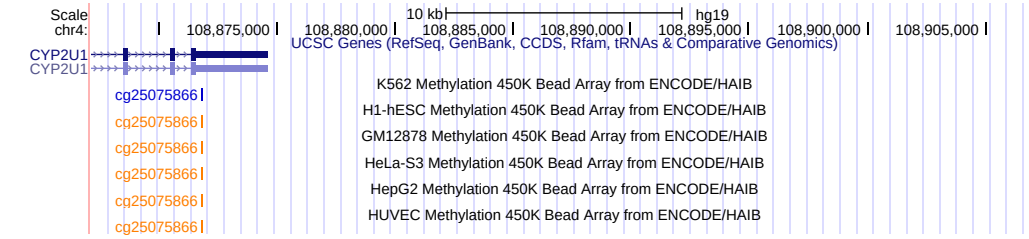


(B) Triple 2–end 3

FIGURE 13: **Methylation status of the distal contact partners of *IGF2-H19* ICR for triple 2.**

Similar snapshots as Supplementary Fig. 12 above for 20 kb region centered on (a) triple 2–end 2 located on chromosome 11, and (b) triple 2–end 3 located on chromosome 8. End 1 is located within 40 kb of the ICR.

(A) Triple 3–end 1



(B) Triple 4–end 3

FIGURE 14: **Methylation status of the distal contact partners of *IGF2-H19* ICR for triples 3 and 4.**

Similar snapshots as Supplementary Fig. 12 above for (a) 20 kb region centered on triple 3–end 1 located on chromosome 11, and (b) 40 kb region centered on triple 4–end 3 located on chromosome 4. Ends 2 and 3 for triple 3, and ends 1 and 2 for triple 4 are located within 40 kb of the ICR.

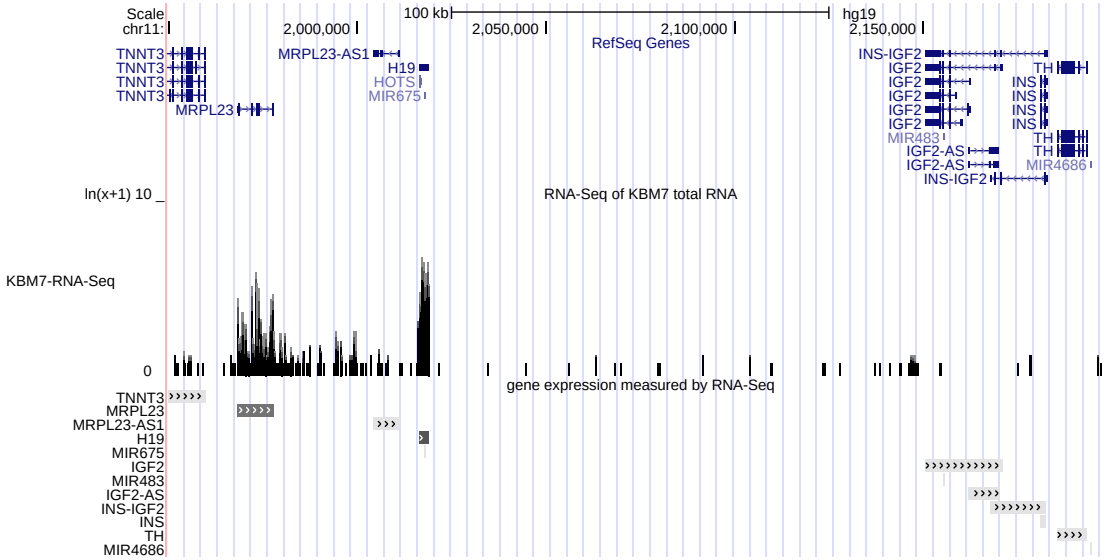
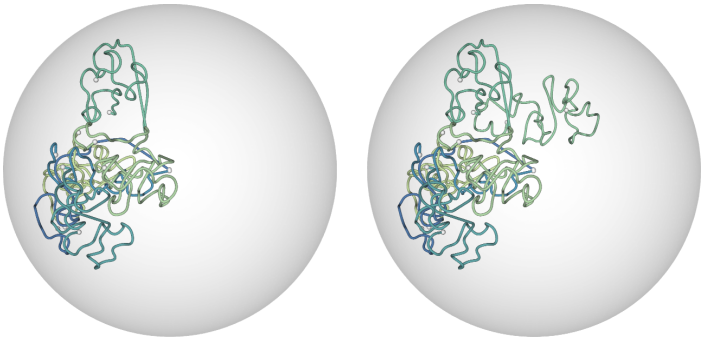


FIGURE 15: **Gene expression measured by RNA-seq for the *IGF2-H19* locus.** Snapshot of 200 kb region taken from the KBM7 genome browser (Bürckstümmer et al.) that includes the *IGF2* and *H19* genes. RNA-seq measurements show that *H19* is expressed, whereas *IGF2* is not. This mode of *IGF2-H19* expression is consistent with the maternal expression pattern of human chromosome 11.



(A) Chromosomes 16, 17, 19, 20, 21, and 22

(B) Chromosomes 16–22

FIGURE 16: **Gene-poor chromosome 18 does not colocalize strongly with other small chromosomes that are gene-rich.**

§ 9 Supplementary Tables

TABLE 1: **Sequences of primers used for PCR verification.**

These primers are designed to test 10 triple contacts listed in Main Figure 5 that involve *IGF2-H19* (ICR). Individual paired-end reads that produced each triplet are analyzed to construct forward/reverse primer pairs that are upstream/downstream of the MboI junction site that resulted in the corresponding chimera. All primer sequences are reported in 5′ to 3′ orientation even though reverse complements are used for reverse primers. The "Expected sizes" column lists the PCR products that are expected to be amplified when each primer triple is used.

Label	Chr.	Dist. to junction	Primer sequence	Strand	Expected sizes (bp)
1a	11	+44	5′-GTGACTGTGAACATTTTAACATGCATGTTTAACGC-3′	forward	
1b	11	-42	5′-TGCTGCACCCACATTAGCAGATTATCTCA-3′	reverse	86, 90
1c	17	-46	5′-AGGGTGATTTTCTTACTGTTTGTAAATAGTGCC-3′	reverse	
2a	8	+121	5′-AGGTGGTAGTCAGAGAATCAGTAAAG-3′	forward	
2b	11	+51	5′-TATAAGCCAAGGAGAGAGGCCTTGGAG-3′	forward	142, 212
2c	11	-91	5′-ACCCTTTCTCTTTTCCCCATTGGTGGTG-3′	reverse	
3a	11	-73	5′-TGTGAGCTGGTGCCAAGGACAGAGGCATCA-3′	reverse	
3b	11	-39	5′-CTCTTCCTTTTGGGGTGAAGACTGTCACCTTCTG-3′	reverse	112, 124
3c	11	+51	5′-ATTCAGAGGCATGACAGTCTCAAGTTCTTGGGA-3′	forward	
4a	11	-61	5′-TTGGCCACGGGCTCTGGAGGCCAGTGCCT-3′	reverse	
4b	4	+74	5′-GTAGGGAGGGAGAAACGGTAATGCTGGTCA-3′	forward	135, 140
4c	11	+79	5′-GGAGGCTCAGGTGAGCCCAGGTCTCCCTCTC-3′	forward	
5a	11	-131	5′-CACCATCCTCCCTCCTGAGAGCTCATTCACTCC-3′	reverse	
5b	11	+48	5′-GCAGCAGTGGCGCTCCCAGCTCTTTAGCA-3′	forward	179, 206
5c	11	+75	5′-TCGTAGGAGACTTTCACGGAGTGCCTGGTCTCC-3′	forward	
6a	2	+61	5′-GCTTATTCTCCATCGGTTTCTAAAGTTGTTCAT-3′	forward	
6b	11	-62	5′-ATTTCATCTCTGACCCAACCAATCAGCACTCCCTA-3′	reverse	96, 123
6c	4	+35	5′-ATTGTTTCCCAGTTCTGGAGTCCAGAAGTCCAA-3′	forward	
7a	11	+73	5′-TGCTTGCTCCTCCGGATGTCCCCTGTGTTTT-3′	forward	
7b	5	+88	5′-CCCAAAGTCATTGATATGGTTTGGCTGCATGTC-3′	forward	130, 145
7c	5	-57	5′-TGCTGATGAATATCTTGGCATCTAGGGGTCAAA-3′	reverse	
8a	5	+47	5′-CAGAAGTTAGGAGAGTCTTGAGTGTGCCTGTTT-3′	forward	
8b	11	+74	5′-TGTGGGCAAATTCACCTCTCCACGTGCCAACTA-3′	forward	154, 171
8c	15	-97	5′-AAAAATATGTTTCCCAGAAACTAGAGACTGGAG-3′	reverse	
9a	7	+61	5′-TGCCCATAGAAACAATTTACTCCAAGGGTCAAT-3′	forward	
9b	11	+49	5′-ACACCCGAGCCATCGAACATCCTAACCCCATCA-3′	forward	98, 110
9c	5	-37	5′-AGCACATGCTAATGCTATCATGAAGTCATACAC-3′	reverse	
10a	8	+55	5′-CCTCTTGTATTTGCTTTTTCCTCTTATCTCTCT-3′	forward	
10b	11	+49	5′-ACACCCGAGCCATCGAACATCCTAACCCCATCA-3′	forward	113, 119
10c	12	-64	5′-TCCCTCTCTCTTTTGTTTTTTGTACTTTATTTG-3′	reverse	

§ 10 Description of additional data files

Additional file 2 — Summary of the two-phase mapping results (XLSX).

This file contains separate worksheet describing in detail the numbers of reads processed at each step of our two-phase mapping.

Additional file 3 — Rotating view of our KBM7 3D model (MP4).

This file contains a movie of the KBM7 3D structure that resulted in the highest log likelihood inferred by our algorithm. Each chromosome is colored as indicated in Figure 6b.

Appendix E

Multiple dimensions of epigenetic gene regulation in the malaria parasite *Plasmodium falciparum*

This chapter has been published in a slightly modified form in [Ay et al., 2015a] as a joint work with Ferhat Ay, Evelien Bunnik, Jean-Philippe Vert, Karine Le Roch and William S. Noble and

Abstract

Plasmodium falciparum is the most deadly human malaria parasite, responsible for an estimated 207 million cases of disease and 627,000 deaths in 2012. Recent studies reveal that the parasite actively regulates a large fraction of its genes throughout its replicative cycle inside human red blood cells and that epigenetics plays an important role in this precise gene regulation. Here we discuss recent advances in our understanding of three aspects of epigenetic regulation in *P. falciparum*: changes in histone modifications, nucleosome occupancy and the three-dimensional genome structure. We compare these three aspects of the *P. falciparum* epigenome to those of other eukaryotes, showing that large-scale compartmentalization is particularly important in determining histone decomposition and gene regulation in *P. falciparum*. We conclude by presenting a gene regulation model for *P. falciparum* which

combines the described epigenetic factors and by discussing the implications of this model for the future of malaria research.

Keywords: malaria, nucleosome occupancy, histone modifications, three-dimensional genome organization, epigenetics, gene regulation, virulence genes.

Abbreviations: PfEMP1, Plasmodium falciparum Erythrocyte Membrane Protein 1; var, family of genes that encode PfEMP1 proteins; ApiAP2, a family of transcription factors in Plasmodium; mRNA, messenger RNA; FISH, fluorescent in situ hybridization; 3C, chromatin conformation capture; 4C, circularized chromatin conformation capture; Hi-C, chromatin conformation capture coupled to next-generation sequencing; ChIA-PET, Chromatin Interaction Analysis by Paired-End Tag Sequencing; PTM, post-translational modification; TSS, transcription start site; ChIP, chromatin immunoprecipitation; TAD, topologically associated domain; H4K20me3, histone H4 lysine 20 trimethylation; H3K9ac, histone H3 lysine 9 acetylation; H3KNme3, histone H3 lysine N trimethylation; H2A, histone H2A; H2A.Z, H2B.Z, variants of histone H2A and H2B; SHH, sonic hedgehog gene; Hox, a group of homeobox genes; OR, olfactory receptors; hpi, hours post invasion.

§ 1 Introduction

The complex life cycle of *Plasmodium falciparum* includes multiple stages in both the human host and the mosquito vector (reviewed in Greenwood et al. [2008]) (Fig 1). Human infection starts with the bite of an infected female Anopheles mosquito, resulting in the transfer of sporozoites that quickly migrate to the liver. Inside liver cells (hepatocytes), these sporozoites multiply extensively over a period of approximately two weeks and are then released into the bloodstream in the form of thousands of merozoites (Fig. 1 - liver stage). During the next stage of its life cycle, the parasite replicates in red blood cells (erythrocytes) by means of an unusual process of cell division called schizogony. While the parasite progresses through three distinct developmental stages (ring, trophozoite and schizont), it undergoes multiple rounds of nuclear replication followed by division of the multinucleated parasite into 16 to 32 daughter merozoites (Fig. 1 - asexual cycle). Upon bursting out of the host cell, these merozoites are released into the bloodstream and will invade new erythrocytes. During the asexual cycle, the parasite can commit to

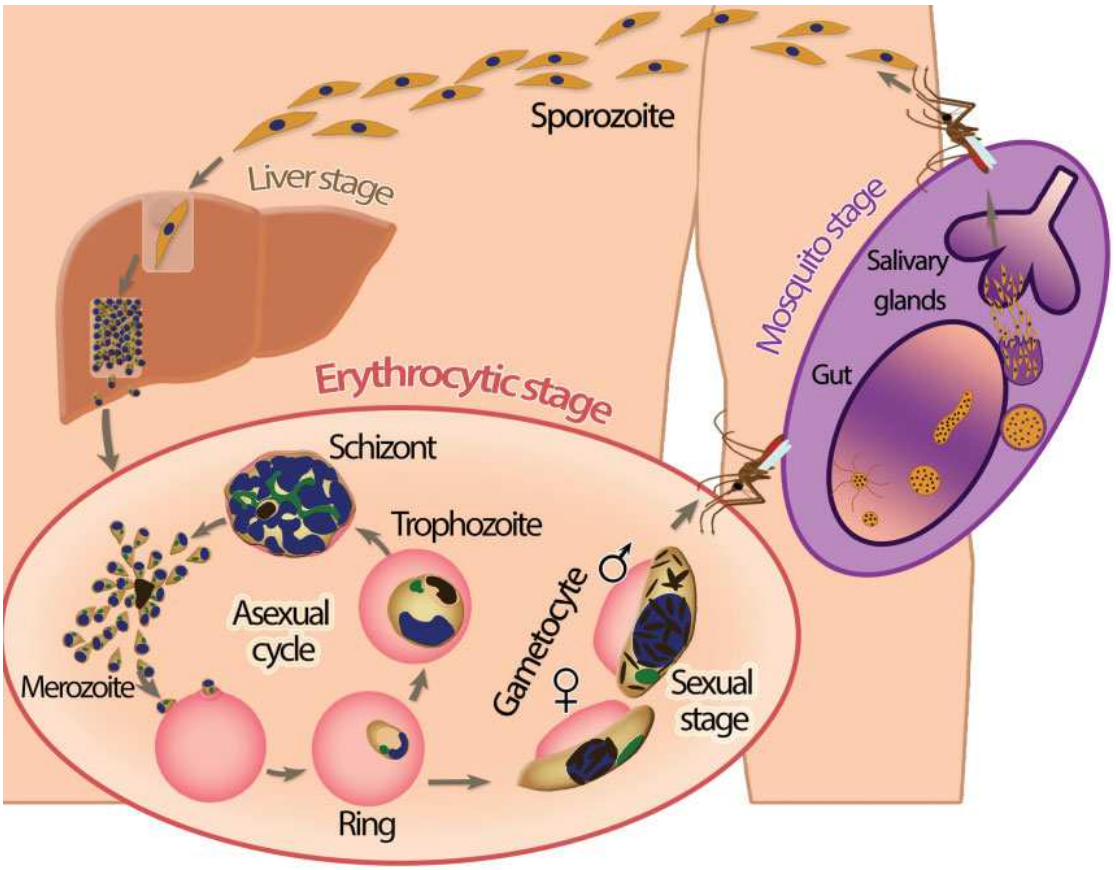


FIGURE 1: Overview of the *P. falciparum*

sexual development (reviewed in Baker [2010]), resulting in differentiation into a male or female gametocyte (Fig. 1 - sexual stage). The uptake of mature gametocytes by a feeding mosquito followed by the further development of the parasite in the mosquito midgut completes the *P. falciparum* life cycle (Fig. 1 - mosquito stage).

The asexual replication cycle is responsible for symptomatic disease and for the complications that are associated with severe malaria, such as anemia due to rupturing of red blood cells. In addition, severe disease can result from cytoadherence, the attachment of *P. falciparum*-infected erythrocytes to the smallest blood vessels, preventing clearance by the spleen and causing organ dysfunction. This cytoadherence is mediated by a family of parasite virulence proteins that are expressed on the erythrocyte surface, *Plasmodium falciparum* Erythrocyte Membrane Protein 1 (PfEMP1) [Baruch et al., 1995, Smith et al., 1995, Su et al., 1995]. Each *P. falciparum* parasite has approximately 60 different PfEMP1 variants encoded by var genes, only one of which is expressed at any time. Switching var gene expression enables the parasite to escape from host immune responses [Bull et al., 1998, Roberts et al., 1992]. This process of antigenic variation is one example of the excellent adaptation of the parasite to survive in the human host.

The development of *P. falciparum* through the different stages of its life cycle is thought

to be driven by coordinated changes in gene expression. Over the last decade, it has become clear that the parasite relies on an unusual combination of regulatory mechanisms for gene expression, and that these mechanisms are largely dependent on epigenetic processes (reviewed in [Cui and Miao, 2010, Duffy et al., 2012, Hoeijmakers et al., 2012b, Horrocks et al., 2009, Deitsch et al., 2007, Voss et al., 2014]). In higher eukaryotes, gene expression is often mediated by transcription factors that bind to cell- or tissue-specific promoters and give rise to the expression of a subset of genes specific to that cell type or tissue [Dunham et al., 2012]. However, despite extensive computational searches, relatively few transcription factors have been identified in *P. falciparum* [Balaji et al., 2005, Coulson et al., 2004], only a handful of which are known to be specific to a certain stage [Campbell et al., 2010]. A notable example is PfAP2-G, a member of the ApiAP2 transcription factor family, that drives expression of gametocyte-specific genes and is crucial for the development of gametocytes [Kafsack et al., 2014, Sinha et al., 2014]. On the other hand, a relatively large number of genes are predicted to encode proteins involved in chromatin structure, mRNA decay and translation rates [Coulson et al., 2004], suggesting that alternative mechanisms of gene regulation, at the epigenetic as well as post-translational levels, may be more important for gene regulation in *P. falciparum.* Here we focus on three important aspects of epigenetic gene regulation in *P. falciparum*, all of which are related to how DNA is packed in the nucleus (see Chung et al. [2009], Le Roch et al. [2011], Suvorova and White [2014], Kramer [2014], Bunnik et al. [2013] for articles discussing post-transcriptional regulation and see Ponts et al. [2013] for a discussion on DNA methylation, which is not well-characterized in *P. falciparum*). Similar to other eukaryotes, *P. falciparum* packages its DNA in the form of a condensed DNA-protein complex called chromatin. The basic packaging unit is a nucleosome, a stretch of approximately 147 bp of DNA wrapped around a core of eight histone proteins. Several layers of higher-order compaction of these strings of nucleosomes together create a highly structured nucleus. The organization of chromatin at both local and global levels is known to be involved in transcriptional regulation [Jenuwein and Allis, 2001, Zentner and Henikoff, 2013, Nora et al., 2013, Belmont, 2014]. Local chromatin structure encompasses two main regulatory processes: the post-translational modification (PTM) of histone proteins that form nucleosomes, and nucleosome occupancy, which comprises the location, frequency, binding strength and protein composition (i.e., variant versus canonical histones) of nucleosomes on DNA.

At the global level, the organization of chromatin has been studied extensively, initially using gene-by-gene approaches such as immunofluorescent microscopy and fluorescent in situ hybridization (FISH) and, more recently, with chromatin conformation capture (3C)-based next-generation sequencing assays. 3C-based assays have enabled genome-wide profiling of chromatin contacts for various organisms including human, mouse, fruit

fly, budding yeast and *P. falciparum* [Ay et al., 2014b, Dixon et al., 2012, Duan et al., 2010, Lemieux et al., 2013, Lieberman-Aiden et al., 2009, Sexton et al., 2012]. These profiles have yielded significant insights into the relation between chromatin organization and transcription, revealing for example the compartmentalization of the genome into regions of transcriptionally active euchromatin and transcriptionally silent heterochromatin. Furthermore, for the haploid *P. falciparum* genome, the 3D models inferred from these contact profiles allowed tracking changes in nuclear organization throughout different stages of the parasite life cycle [Ay et al., 2014b].

In the following sections, we provide an overview of our current understanding of chromatin organization and its role in transcriptional regulation in *P. falciparum*. We first describe various characteristics of local chromatin structure and subsequently focus on three-dimensional genome architecture. Finally, we combine these local and global views of chromatin to provide a model that explains our current understanding of the overall nuclear organization in *P. falciparum* and the role of the epigenome in regulating gene expression.

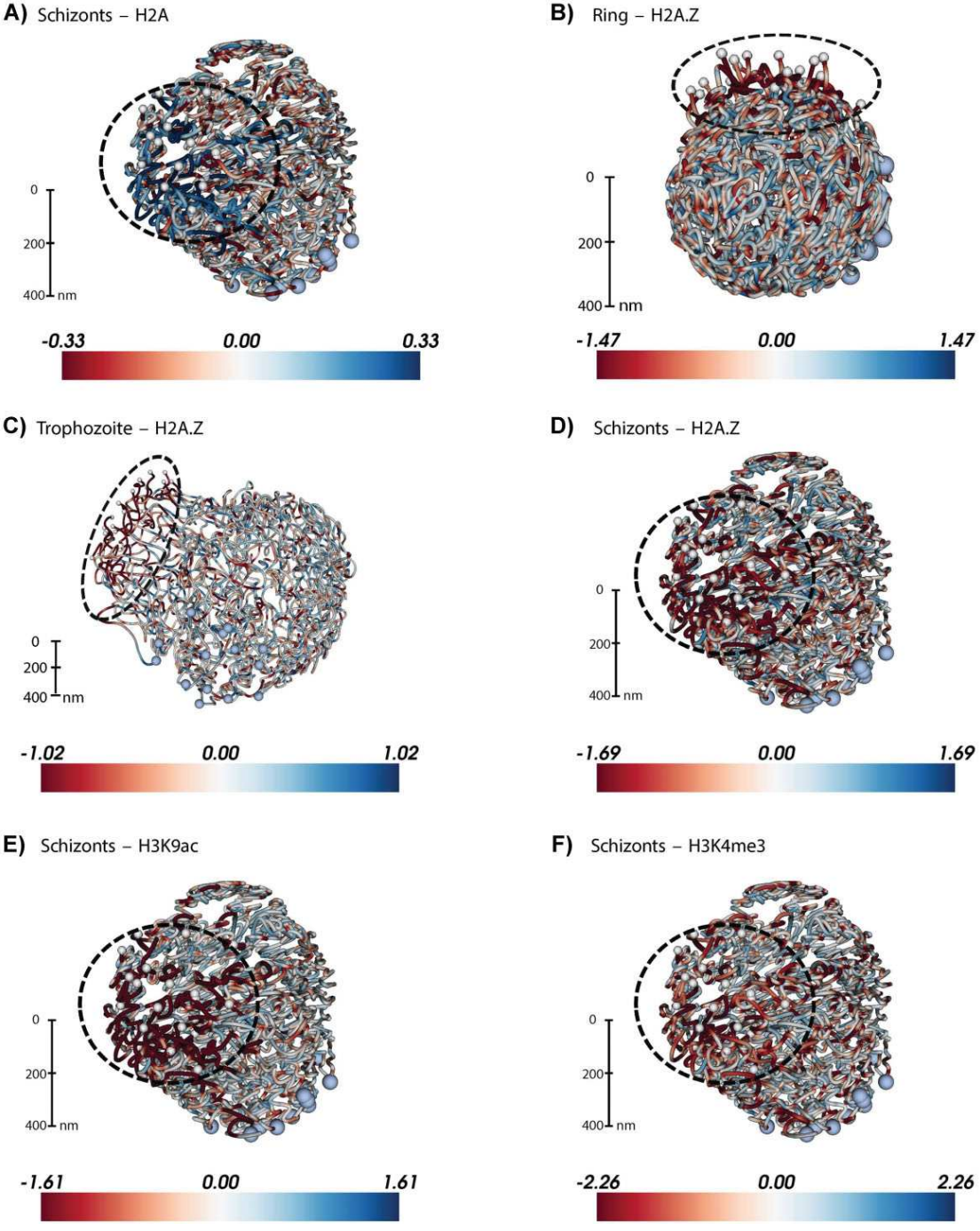
§ 2 Histone modification landscape of the *P. falciparum* genome favors euchromatin

§ 2.1 Post-translational modification of histone proteins

Histone proteins consist of a globular core structure and an N-terminal tail that protrudes from this core domain. Many amino acid residues in the core domain and in particular in the N-terminal tail can be chemically modified, with various effects on chromatin organization (Fig. 2). In general, the addition of an acetyl group neutralizes the positive charge of histone proteins and thereby disrupts the stability of the DNA-histone interaction. This destabilization results in a more open chromatin structure and promotes a transcriptionally permissive state. On the other hand, methylations are uncharged and do not directly interfere with the interaction between histones and DNA. Rather, methylations mostly function by recruiting other effector molecules to the locus, resulting in further modifications of the chromatin.

§ 2.2 Activating histone marks are abundant and broadly distributed

In *P. falciparum*, mass spectrometry experiments have identified at least 50 different histone post-translational modifications (PTMs), including methylation, acetylation, phosphorylation, ubiquitylation, and sumoylation [Lasonder et al., 2012, Miao et al.,


A) Schizonts – H2A
0
200
400 nm
-0.33
0.00
0.33
B) Ring – H2A.Z
0
200
400 nm
-1.47
0.00
1.47
C) Trophozoite – H2A.Z
0
200
400 nm
-1.02
0.00
1.02
D) Schizonts – H2A.Z
0
200
400 nm
-1.69
0.00
1.69
E) Schizonts – H3K9ac
0
200
400 nm
-1.61
0.00
1.61
F) Schizonts – H3K4me3
0
200
400 nm
-2.26
0.00
2.26

2006, Treeck et al., 2011, Trelle et al., 2009]. Subsequent chromatin immunoprecipitation (ChIP) studies have given us insight into the genome-wide distribution of these histone marks in the asexual cycle (Table 1). In contrast to multicellular eukaryotes, a large proportion of the genome in *P. falciparum* is constitutively acetylated [Miao et al., 2006, Lopez-Rubio et al., 2009]. An abundance of activating marks has also been observed for other unicellular organisms, such as *Saccharomyces cerevisiae* and *Tetrahymena thermophila* [Garcia et al., 2007]. Inhibition of histone acetyltransferase and deacetylase activity influences the expression levels of the majority of genes and interferes with parasite growth [Cui et al., 2008, 2007, Chaal et al., 2010], indicative of the importance of acetylation for regulating transcription levels. Activating marks H3K9ac and H3K4me3 are mainly located in intergenic regions [Bartfai et al., 2010, Jiang et al., 2013, Salcedo-Amaya et al., 2009]. Highly transcribed genes carry more H3K9ac marks in their promoter [Bartfai et al., 2010], and this marking extends into the 5' coding region [Salcedo-Amaya et al., 2009].

§ 2.3 Repressive histone marks are scarce and localized to specific regions

Typical repressive marks, in particular H3K9me3, are almost exclusively found in repressive clusters containing genes belonging to the virulence families, such as var, rifin, stevor, and pfmc-2tm [Lopez-Rubio et al., 2009, Jiang et al., 2013, Chookajorn et al., 2007, Lopez-Rubio et al., 2007]. Interestingly, H3K9me3 is also present at several additional loci, including the gene encoding the gametocyte-specific transcription factor

FIGURE 2 *(preceding page)*: Large-scale depletion of the transcriptionally permissive histone variant H2A.Z and activating histone marks in the telomeric cluster visualized on the 3D *P. falciparum* genome. ChIP-seq data from Bartfai et al. Bartfai et al. [2010] for four histone variants or marks were downloaded from GEO (accession number: GSE23787) and mapped to the *P. falciparum* genome (PlasmoDB v9.0) using the short read alignment mode of BWA (v0.5.9) [Li and Durbin, 2010] with default parameter settings. Reads were post-processed, and only the reads that map uniquely with a quality score above 30 and with at most two mismatches were retained for further analysis. Retained reads were subjected to PCR duplicate elimination and then were aggregated for each non-overlapping 5 kb bin across the *P. falciparum* genome. The number of reads for each 5 kb bin was normalized using the overall sequencing depth of the corresponding ChIP-seq library. Plotted are the log2 ratios of sequence-depth normalized number of reads from the ChIP-seq library versus the corresponding input library (red: depletion, blue: enrichment) for **A:** H2A at 40 hours post invasion (hpi), **B:** H2A.Z at 10 hpi, **C:** H2A.Z at 30 hpi, **D:** H2A.Z at 40 hpi, **E:** H3K9ac at 40 hpi, and **F:** H3K4me3 at 40 hpi. 3D models for the ring, trophozoite and schizont stages were generated in Ay et al. [2014b] and were colored with ChIP-seq enrichment/depletion from 10, 20, and 40 hpi, respectively. Light blue and white spheres indicate centromeres and telomeres, respectively. The black dashed circle denotes the telomeric cluster for each stage. See Supporting information or http://noble.gs.washington.edu/proj/plasmo-epigenetics for the rotating 3D figure of each available ChIP-seq library.

Histone PTM/variant	*Other eukaryotes*	*P. falciparum*
H3K4me3	Promoters of active genes [Bernstein et al., 2005, Kim et al., 2005, Wang et al., 2008, Barski et al., 2007]	Widely distributed in intergenic regions [Bartfai et al., 2010, Salcedo-Amaya et al., 2009]
H3K9ac	Promoters of active genes [Wang et al., 2008, Nishida et al., 2006]	Widely distributed in intergenic regions [Bartfai et al., 2010, Salcedo-Amaya et al., 2009]
H3K4me3	Promoters of active genes [Bernstein et al., 2005, Kim et al., 2005, Wang et al., 2008, Barski et al., 2007]	Widely distributed in intergenic regions [Bartfai et al., 2010, Salcedo-Amaya et al., 2009]
H3K9ac	Promoters of active genes [Wang et al., 2008, Nishida et al., 2006]	Widely distributed in intergenic regions [Bartfai et al., 2010, Salcedo-Amaya et al., 2009]
H3K9me3	Silent genes [Wang et al., 2008, Barski et al., 2007]	Repressed var genes [Lopez-Rubio et al., 2009, Chookajorn et al., 2007, Lopez-Rubio et al., 2007]
H3K27me3	Promoters of silent/poised genes [Wang et al., 2008, Barski et al., 2007, Mikkelsen et al., 2007], absent in yeast [Lachner et al., 2004]	Not detected [Trelle et al., 2009]
H3K36me3	Enriched in pericentromeric heterochromatin [Chantalat et al., 2011]; Transcribed regions of active genes [Wang et al., 2008, Barski et al., 2007]	TSS of repressed var genes [Jiang et al., 2013]; 3' end coding region active genes [Jiang et al., 2013]
H4K20me3	Silencing of telomeres, transposons and long terminal repeats [Barski et al., 2007, Lachner et al., 2004]; inactive promoters [Wang et al., 2008]	Repressed var genes [Jiang et al., 2013] and broad distribution across additional loci [Lopez-Rubio et al., 2009]
H2A.Z	Enriched in nucleosomes bordering active promoter (reviewed in [Zlatanova and Thakar, 2008, Talbert and Henikoff, 2010])	Widely distributed in intergenic regions [Hoeijmakers et al., 2013, Petter et al., 2013]
H2B.Z	Lineage-specific variants with specialized functions, for example enriched at TSS in *Trypanosoma brucei* [Siegel et al., 2009]	Widely distributed in intergenic regions [Hoeijmakers et al., 2013, Petter et al., 2013]

TABLE 1: Overview of most-studied histone modifications and variants in *P. falciparum* and comparison of their genome-wide distribution or function in other eukaryotes.

PfAP2-G [Lopez-Rubio et al., 2009] that is tightly repressed during the asexual cycle. Transcription start sites of silent var genes are also enriched for H3K36me3 [Jiang et al., 2013], while this modification is found at equal levels inside coding regions of active and repressed var genes [Jiang et al., 2013, Ukaegbu et al., 2014]. H3K36me3 is present at lower levels in the rest of the genome and is enriched at the 3' end of coding regions of active *P. falciparum* genes, in agreement with its role in transcriptional elongation in other eukaryotes. The repressive mark H4K20me3 is also mainly present in var gene clusters, although its enrichment is not as strong as for H3K9me3 and H3K36me3 [Jiang et al., 2013]. On the other hand, the single active var gene, out of $\tilde{6}0$ family members, is enriched in active histone marks, such as H3K9ac, H3K4me3, and H4 acetylations [Jiang et al., 2013, Lopez-Rubio et al., 2007]. Finally, the repressive mark H3K27me3 has not been detected in the parasite [Trelle et al., 2009], similar to yeast. The *P. falciparum* genome organization thus seems unusual in that a large fraction of its chromatin is continuously in a transcriptionally permissive state, while the formation of heterochromatin seems to be limited to virulence and specific sexual genes.

§ 3 Histone variants and nucleosome occupancy are associated with gene expression

§ 3.1 Plasmodium exhibits a distinctive nucleosome landscape around coding regions relative to other eukaryotes

Nucleosome occupancy plays an important role in regulating gene expression by allowing or restricting access of the transcription machinery to the DNA. Nucleosomes are not placed uniformly along the genome, but show a distinct distribution around coding regions [Brogaard et al., 2012, Buenrostro et al., 2014, Jansen and Verstrepen, 2011, Lee et al., 2007, Mavrich et al., 2008]. In yeast and higher eukaryotes, the promoter is characterized by a nucleosome-depleted region, bordered on either side by strongly positioned -1 and +1 nucleosomes, respectively, both of which are enriched for the variant histone H2A.Z [Raisner et al., 2005, Guillemette et al., 2005, Tolstorukov et al., 2009]. The +1 nucleosome is located at a fixed distance relative to the transcription start site (TSS), although this distance varies between organisms [Lee et al., 2007]. The +2, +3 and subsequent nucleosomes form an array of nucleosomes with increasingly more fuzzy positioning towards the 3' end of the gene. Finally, the transcription stop site is again demarcated by a strongly positioned nucleosome, followed by another nucleosome-depleted region.

Nucleosome organization in *P. falciparum* is similar to other eukaryotes in several respects. First, the promoter region is depleted of nucleosomes [Bunnik et al., 2014, Ponts et al., 2010, Westenberger et al., 2009], the level of which correlates with transcriptional activity. Second, highly expressed genes have a more open chromatin organization at their core promoter than silent genes [Bunnik et al., 2014, Ponts et al., 2010]. However, the *P. falciparum* nucleosome landscape also exhibits a number of unusual features. Notably, the TSS is not marked by a strongly positioned +1 nucleosome; instead, the strongest nucleosomes are the first and last nucleosomes within the coding region [Bunnik et al., 2014, Ponts et al., 2010]. Furthermore, telomeric repeats and subtelomeric regions that contain the virulence gene families (var, rifin, etc) have higher nucleosome occupancy levels than the bulk of the genome [Bunnik et al., 2014, Ponts et al., 2010, Segal et al., 2006]. Intergenic regions, on the other hand, contain lower nucleosome levels than coding regions [Bunnik et al., 2014, Ponts et al., 2010, Segal et al., 2006, Ponts et al., 2011], which is likely to be related to their extremely high AT-content (90-95%). AT-rich DNA is inherently inflexible, hampering the winding of DNA around the histone core [Tillo and Hughes, 2009, Segal and Widom, 2009]. Finally, intergenic regions in *P. falciparum* are exclusively occupied by nucleosomes composed of histone variants H2A.Z and H2B.Z [Hoeijmakers et al., 2013, Petter et al., 2013], which are thought to have adopted a specialized function in *P. falciparum* to allow nucleosome assembly in these highly AT-rich regions. These histone variants are thus not restricted to promoter flanking nucleosomes but have a much broader distribution.

§ 3.2 Nucleosome dynamics change in concordance with transcriptional activity during the asexual cycle

Another unconventional feature of nucleosome organization in *P. falciparum* is that nucleosome levels vary considerably during the asexual replication cycle, in parallel with changes in transcriptional activity [Bunnik et al., 2014, Ponts et al., 2010]. At the transcriptionally most active trophozoite stage, histone levels decrease by approximately two-fold [Bunnik et al., 2014, Ponts et al., 2010]. This nucleosome depletion occurs in a genome-wide fashion and is not restricted to genes that are expressed in the trophozoite stage. As the asexual cycle progresses into the schizont stage, nucleosomes are re-assembled, resulting in condensation of DNA as the parasites prepare for egress and re-invasion of a new red blood cell. Given the correlation between nucleosome density in promoter regions and gene expression levels, the dynamic nucleosome landscape in *P. falciparum* may have evolved to compensate for a paucity of specific transcription factors. Interestingly, *Trypanosoma brucei*, a parasite causing sleeping sickness in humans, has also developed an unusual nucleosome landscape, where certain combinations

of canonical and variant histones mark the transcription initiation and termination sites in its genome [Siegel et al., 2009]. Reminiscent of the lack of transcription factors in *P. falciparum*, transcription factors have remained elusive in *T. brucei*, indicating that these parasites may have followed parallel evolutionary pathways towards the use of the nucleosome landscape as a mechanism to regulate gene expression.

§ 4 Three-dimensional conformation of the *P. falciparum* genome

§ 4.1 Principles of nuclear organization in *P. falciparum*

It has been long known that the eukaryotic nucleus is a highly structured entity. In addition to three-dimensional conformation of the chromatin-packaged DNA, key structural landmarks include the nuclear envelope, nuclear pores and nucleoli. For decades, various microscopic imaging techniques have been the "go-to" tools for understanding nuclear organization and chromatin architecture in many different organisms [Cremer et al., 2006, Misteli, 2007, Takizawa et al., 2008]. In *P. falciparum*, FISH applications have been instrumental in demonstrating important characteristics of genome organization in the parasite. In particular, silent var genes were shown to colocalize with each other near the nuclear periphery, while the single active var gene is located elsewhere

[Lopez-Rubio et al., 2009, Freitas-Junior et al., 2000, Ralph et al., 2005]. Together with the other epigenetic mechanisms outlined above — histone modifications, histone variants and nucleosome occupancy — the non-random organization of DNA into repressive centers is believed to play a crucial role in the one-at-a-time expression of 60 genes in the var family. Another intriguing discovery from FISH experiments was that the ribosomal DNA loci that are distributed in a seemingly random fashion on different *P. falciparum* chromosomes show non-random colocalization in 3D [Mancio-Silva et al., 2010]. A more recent study employed several ultrastructural microscopy techniques to study the distribution of nuclear pore complexes and chromatin throughout the *P. falciparum* asexual cycle [Weiner et al., 2011], demonstrating a striking increase in pore density during the transcriptionally active trophozoite stage, as well as chromatin decomposition near the nuclear envelope. These changes parallel previously observed changes in transcriptional activity and nucleosome occupancy that have been discussed above [Ponts et al., 2010].

§ 4.2 Profiling of eukaryotic genome architecture using next-generation sequencing applications

Within the last decade, the field of genome architecture has been revolutionized by breakthroughs in combining next generation sequencing with molecular assays that measure proximities of DNA regions to certain nuclear landmarks (e.g., lamina, nucleolus) or to other regions in cis or trans (e.g., 4C, Hi-C, ChIA-PET) [Duan et al., 2010, Lieberman-Aiden et al., 2009, Fullwood et al., 2009, Guelen et al., 2008, van Koningsbruggen et al., 2010, Vogel et al., 2007, Zhao et al., 2006] (see [Steensel and Dekker, 2010] for review). Applications of these techniques to multiple genomes including human and mouse have revealed the organizational hallmarks of genome architecture. These include localization of gene-rich regions near the nuclear center and heterochromatin near the nuclear lamina [Guelen et al., 2008], colocalization of ribosomal DNA loci near nucleoli [van Koningsbruggen et al., 2010], and megabase-scale open/closed chromatin compartments [Lieberman-Aiden et al., 2009]. In addition, genomes of higher eukaryotes are partitioned into megabase-sized topologically associated domains (TADs) that are enriched for interactions within but not across domains and are separated from each other by insulator proteins [Dixon et al., 2012, Nora et al., 2012, Sofueva et al., 2013] (see [Nora et al., 2013] for review). Finally, these studies have provided us with examples of cell type-specific chromatin loops bringing distal regulatory elements in close 3D proximity. Long-range chromatin loops that play regulatory roles in gene expression include Hox cluster silencing [Ferraiuolo et al., 2010, Rousseau et al., 2014], control of SHH gene by an enhancer that is located 1 Mb away in human [Li et al., 2012] and a validated set of cell type-specific enhancers in mouse [Shen et al., 2012].

§ 4.3 Profiling of *P. falciparum* genome architecture during the asexual cycle

As is the case for many other next generation sequencing-based assays, application of these genome architecture assays has been challenging for the AT-rich genome of *P. falciparum*. However, within the last year, two groups have published their results using Hi-C, one profiling the genome architecture of different *P. falciparum* strains [Lemieux et al., 2013] and the other modeling the 3D structure of *P. falciparum*-3D7 at three key stages during its asexual replication cycle within human red blood cells [Ay et al., 2014b]. These studies revealed key characteristics of *P. falciparum* genome structure (Table 2), including colocalization of centromeres, colocalization of telomeres near the nuclear periphery, colocalization of both internal and subtelomeric virulence gene clusters near the telomeres, colocalization of rDNA loci that are active in ring stage parasites

Feature	*Ring*	*Trophozoite*	*Schizont*
Nuclear size	Small (∼700 nm diameter) [Weiner et al., 2011, Bannister et al., 2005]	Large (∼700 nm diameter) [Weiner et al., 2011, Bannister et al., 2005]	Small (∼850 nm diameter) [Weiner et al., 2011, Bannister et al., 2005]
Nuclear pores	Few (3-7), clustered together [Weiner et al., 2011]	Many (12-58), uniformly distributed [Weiner et al., 2011]	Few per daughter nucleus (2-6), clustered together [Weiner et al., 2011]
Nucleosome occupancy	High [Bunnik et al., 2014, Ponts et al., 2010]	Low [Bunnik et al., 2014, Ponts et al., 2010]	High [Bunnik et al., 2014, Ponts et al., 2010]
Chromatin compaction	Compact [Ay et al., 2014b, Bunnik et al., 2014, Ponts et al., 2010, Weiner et al., 2011]	Open [Ay et al., 2014b, Bunnik et al., 2014, Ponts et al., 2010, Weiner et al., 2011]	Compact [Ay et al., 2014b, Bunnik et al., 2014, Ponts et al., 2010, Weiner et al., 2011]
Chromosome territories	Conflicting reports (absent [Lemieux et al., 2013] vs present [Ay et al., 2014b])	Partially lost [Ay et al., 2014b]	Present [Ay et al., 2014b]
Centromere locations	Conflicting reports (colocalized [Ay et al., 2014b] vs dispersed [Lemieux et al., 2013, Hoeijmakers et al., 2012a])	Colocalized [Ay et al., 2014b, Hoeijmakers et al., 2012a]	Colocalized [Ay et al., 2014b, Hoeijmakers et al., 2012a]
Telomere locations	Colocalized near periphery [Ay et al., 2014b, Freitas-Junior et al., 2000]	Colocalized near periphery [Ay et al., 2014b, Freitas-Junior et al., 2000]	Colocalized near periphery [Ay et al., 2014b, Freitas-Junior et al., 2000]
Virulence gene locations	Colocalized [Lemieux et al., 2013] near periphery [Ay et al., 2014b, Lopez-Rubio et al., 2009, Freitas-Junior et al., 2000]	Colocalized near periphery [Ay et al., 2014b, Lopez-Rubio et al., 2009, Freitas-Junior et al., 2000]	Colocalized near periphery [Ay et al., 2014b, Lopez-Rubio et al., 2009, Freitas-Junior et al., 2000]
rDNA gene locations	Conflicting reports (all loci clustered [Mancio-Silva et al., 2010] vs strong clustering of only active loci [Ay et al., 2014b, Lemieux et al., 2013])	Conflicting reports (dispersed [Mancio-Silva et al., 2010] vs weak clustering of only active loci [Ay et al., 2014b])	Conflicting reports (dispersed [Mancio-Silva et al., 2010] vs weak clustering of only active loci [Ay et al., 2014b])

TABLE 2: Summary of organizational features of *P. falciparum* nucleus and genome at three distinct stages during asexual parasite replication in human red blood cells (asexual cycle).

and maintenance of chromosomes territories (see Ay et al. [2014b] for details). Furthermore, Hi-C profiles from Ay et al. [2014b] exhibit different polymer behavior in the most transcriptionally active trophozoite stage compared to the other two stages, suggesting a link between overall chromatin compaction and transcriptional activity. The degree of telomere colocalization and the repressive effect of the telomeric compartment is also most pronounced in this trophozoite stage, suggesting a strict compartmentalization to segregate genes that need to be repressed from the rest. Finally, both the Hi-C contact maps and the 3D models inferred from them suggest a tight correlation between the 3D location of a gene and its expression. Gene pairs located nearby in 3D have significantly higher expression correlation compared to other pairs, even after discarding intra-chromosomal pairs that would be biased by their genomic distance in 1D [Ay et al., 2014b]. Overall, these observations suggest that *P. falciparum* chromatin is highly structured at the large scale and that this structure provides a potential epigenetic mechanism to regulate gene expression.

The folded chromosome structure seen in *P. falciparum* is similar to what has been observed in budding and fission yeast [Duan et al., 2010, Tanizawa et al., 2010]. However, chromosome looping to achieve localization of var genes in repressive perinuclear compartments results in a more complex three-dimensional organization of the *P. falciparum* genome compared to yeast, even though these organisms have similarly sized genomes [Ay et al., 2014b]. Interestingly, the clonal var gene expression and clustering of all remaining var genes in repressive heterochromatin is strikingly similar to the epigenetic signature of the 1,400 olfactory receptor genes in the mouse, all except one of which are located in heterochromatic foci enriched for H3K9me3 and H4K20me3, resulting in monogenic and monoallelic expression [Magklara et al., 2011, Lyons et al., 2013]. In comparison to higher eukaryotes, such as human, mouse and fly, the *P. falciparum* genome organization is relatively simple and does not display TADs. The nuclear architecture in *P. falciparum* thus exploits features from both unicellular and multicellular organisms.

§ 5 A combined model of epigenetic gene regulation in *P. falciparum*

§ 5.1 Nuclear organization and gene regulation

The epigenetic makeup of the *P. falciparum* genome, as outlined above, points towards a binary nuclear organization, with the majority of the genome present in the form of euchromatin, while a limited number of genes are organized into strongly repressed

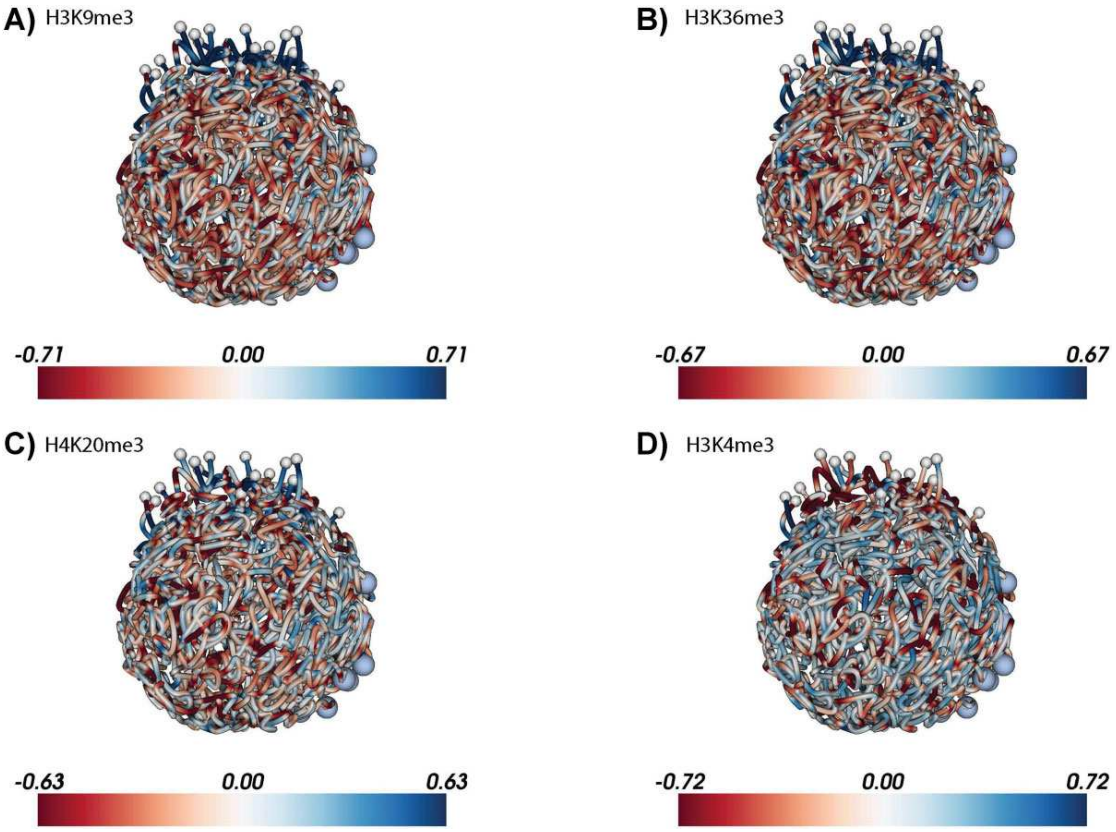


FIGURE 3: **Visualization of ChIP-seq data from Jiang et al. [46] on the 3D *P. falciparum* genome at the ring stage**. ChIP-seq data from Jiang et al. for 5 histone marks were downloaded from SRA (accession number: SRP022761) and processed as described in the caption of Figure 2. Due to lack of input libraries from this publication, the input libraries from Bartfai et al. at different time points were pooled into one aggregated input library which is then used for normalization of each Jiang et al. ChIP-seq library. Similar to Figure 2, log2 ratios of ChIP-seq versus input were plotted for **A:** H3K9me3, **B:** H3K36me3, **C:** H4K20me3, and **D:** H3K4me3 at 18 hpi. The 3D model for the ring stage from [Ay et al., 2014b] was used to visualize enrichment/depletion of each histone mark. See http://noble.gs.washington.edu/proj/plasmo-epigenetics for the rotating 3D figure of each available ChIP-seq library.

heterochromatin. This heterochromatin is localized at the nuclear periphery and is characterized by high nucleosome density (Fig. 2A), the presence of repressive histone marks H3K9me3, H3K36me3 and H4K20me3 (Fig. 3A-C), and the absence of the transcription-associated histone variant H2A.Z (Fig. 2B-D) and histone marks H3K9ac (Fig. 2E) and H3K4me3 (Fig. 2F and 3D). It was recently demonstrated that heterochromatin protein 1 (HP1) and *P. falciparum* histone deacetylase 2 (PfHda2) are both essential for maintaining heterochromatic regions [Brancucci et al., 2014, Coleman et al., 2014]. Depletion of either HP1 or PfHda2 resulted in an arrest of parasite development at the trophozoite stage and a loss of var gene repression. In addition, an increase in the number of parasites differentiating into gametocytes was observed, indicating that the gametocyte transcription factor locus pfap2-g is also under strict epigenetic control. The remaining euchromatic fraction of the genome has several notable features,

including perinuclear compartments containing the active var gene or active rDNA genes (Fig. 4A). In addition, clustering of silent genes that are specific to other stages of the parasite's life cycle [Ay et al., 2014b], suggests the presence of small heterochromatic islands, as observed at the trophozoite stage by advanced transmission and scanning electron microscopy [Weiner et al., 2011].

§ 5.2 Remodeling of the nuclear organization during the asexual cycle

Microarray and RNA-seq studies have shown that 70-80% of all genes are expressed in the asexual replication cycle, in particular during the trophozoite stage [Bunnik et al., 2013, Le Roch et al., 2003, Otto et al., 2010]. During the 48-hour cycle, the nucleus and chromatin are dramatically remodeled to facilitate this high transcriptional activity (Fig. 4B and Table 2). First, the nucleus expands in size [Weiner et al., 2011], which can also be readily observed in microscopy images of Giemsa stained parasites [Ay et al., 2014b]. Second, the number of nuclear pores increases drastically, from 3-7 clustered pores in the ring stage to 12-58 pores that are uniformly distributed around the nucleus in the trophozoite stage [Weiner et al., 2011]. Third, in line with the increased nuclear volume, the chromatin opens up [Ay et al., 2014b, Weiner et al., 2011], accompanied by removal of nucleosomes [Bunnik et al., 2014, Ponts et al., 2010] and increased intermingling of chromosomes [Ay et al., 2014b]. Despite these large-scale nuclear dynamics, the centromeres, telomeres and repressed var genes remain clustered. The correlation of nucleosome density of gene promoters with transcriptional activity of individual genes suggests that local chromatin organization may play an important role in regulating the level of gene expression [Bunnik et al., 2014]. The transitioning of the parasite from the trophozoite stage to the schizont stage is characterized by a reversion of nuclear changes, including reassembly of nucleosomes and re-establishment of chromosomal territories, which results in recompaction of the genome. Finally, during DNA replication, the nucleus divides into multiple small daughter nuclei, each with a small number of the nuclear pores that were present in the original nucleus [Weiner et al., 2011].

§ 6 Outstanding questions

§ 6.1 Clustering of repressive heterochromatin

Whether heterochromatin containing silent var genes is organized into a single large repressive center or is divided over a small number of perinuclear foci remains a topic of debate. FISH images visualizing the location of telomere-associated repeat elements

or var gene promoters typically show 2-6 foci distributed around the nucleus [Lopez-Rubio et al., 2009, Freitas-Junior et al., 2000, Ralph et al., 2005, Voss et al., 2006]. On the other hand, single foci were observed by immunofluorescence microscopy for H3K9me3, H3K36me3, and heterochromatin protein 1 [Ukaegbu et al., 2014, Dahan-Pasternak et al., 2013], all of which are strongly associated with the repressed var genes. In addition, the Hi-C-derived three-dimensional models of the *P. falciparum* genome showed strong clustering of centromeres and telomeres [Ay et al., 2014b] (Fig. 4A), a chromosome configuration that has been observed in other organisms [Duan et al., 2010, Tanizawa et al., 2010, Umbarger et al., 2011]. These models suggested the organization of subtelomeric var genes into a single cluster at the nuclear perimeter. Such organization, even though seemingly contradicting the FISH data, may be due to aggregation across a large population of cells for Hi-C experiments. If each var gene cluster is randomly located in one of multiple repressive clusters in each cell, then the aggregate signal would suggest colocalization of all var genes. However, it may conceivably be beneficial to locate all repressed genes in close proximity of each other to regulate the expression of a single var gene and the tight repression of all remaining family members. Additional experiments will be necessary to unravel the precise mechanisms by which var gene expression is controlled, by further dissecting the effect of gene localization, nuclear architecture, and gene-to-gene communication on this process. In particular, Hi-C experiments on single cells would likely provide significant insight into the localization of active and repressed var genes, as well as the extent of cell-to-cell variability.

§ 6.2 Mediators of epigenetic control and nuclear remodeling

Drastic remodeling of the nucleus and chromatin are likely to be driving forces behind the wave of transcriptional activity during the trophozoite stage. Components involved in these dynamic processes may thus be promising targets for antimalarial drugs. Future research should therefore focus on understanding the molecular mechanisms involved in chromatin and nuclear remodeling. For example, very little is known about proteins and enzymes that regulate the formation of heterochromatin and the global nuclear architecture, with the exception of the role of HP1 in maintaining repressive perinuclear chromatin containing the var genes and the pfap2-g locus. A multitude of such proteins has been identified in other organisms, most notably RNA polymerase III-associated factor (TFIIIC), cohesin and CCCTC binding factor (CTCF) (reviewed in [Gomez-Diaz and Corces, 2014]), and are likely to have homologues in *P. falciparum*. Other potential drug targets include key components involved in expansion of the nuclear membrane and chromatin remodeling enzymes that regulate the global nucleosome eviction and reassembly during the trophozoite and schizont stages. Analysis of chromatin-associated

proteins by proteomics-based approaches will likely identify many candidates that may be involved in these processes. In addition, the application of novel genetic engineering tools in *P. falciparum*, such as the CRISPR/Cas9 system [Ghorbal et al., 2014, Zhang et al., 2014, Wagner et al., 2014], may enable us to study the effect of gene deletion or translocation on genome structure to better understand the determinants of nuclear architecture.

§ 6.3 Epigenetic control in other parasite stages

The epigenetic regulation model we present here is based on profiles taken during the asexual replication cycle. During this phase of the parasite's life cycle, the genome seems to be largely shaped by the strict one-at-a-time expression of the var genes. The absence of var gene expression in all other parasite stages may have a large impact on chromatin organization. In addition, while some genes may be constitutively expressed during the parasite's life cycle, others may be silenced or activated in these alternative and highly variable stages, ranging from the male and female gametocyte, via the diploid zygote in the mosquito midgut, to the haploid sporozoite. Therefore, we expect generating genome-wide profiles of histone modifications, nucleosome landscape and three-dimensional architecture during these other parasite stages to be of great interest to further explore the epigenetic regulatory mechanisms in *P. falciparum*.

Furthermore, we know very little about the role of epigenetic control in transcriptional regulation in other Plasmodium species. *P. vivax*, for example, has a much lower AT-content (on average 57%), which is likely to influence the binding kinetics and preferences of nucleosomes. In addition, *P. vivax* expresses a large proportion of its gene family encoding for variant surface proteins (vir) during the blood stage [Bozdech et al., 2003, Fernandez-Becerra et al., 2005]. The absence of clonal expression as seen for the var family in *P. falciparum* may relieve the requirements for strictly repressive heterochromatin in *P. vivax*. Determining the nucleosome landscape, the location of histone modifications and the three-dimensional structure of the *P. vivax* genome will therefore also be extremely informative for our understanding of epigenetic gene regulation.

§ 7 Conclusions and prospects

An increasing amount of data highlights the importance of epigenetic mechanisms in regulating gene expression in *P. falciparum* and other eukaryotes, including human and mouse [Ay et al., 2014b, Dixon et al., 2012, Duan et al., 2010, Lemieux et al., 2013, Lieberman-Aiden et al., 2009, Sexton et al., 2012]. Here we have discussed multiple

layers of epigenetic control, including histone modifications, nucleosome occupancy, histone variants and genome architecture, which are involved in the precise gene regulation during the asexual replication cycle of the malaria parasite, *P. falciparum*. We summarized the current understanding of the interplay among these different layers and how these layers shape the overall nuclear organization and connect to overall transcriptional activity and to the one-at-a-time expression of var genes.

Better characterization of epigenetic regulation in *P. falciparum* will stimulate interest in several exciting directions in malaria research. Further studies into the establishment and maintenance of strong repressive compartments in the nucleus may reveal the underlying regulatory mechanisms and lead to the identification of proteins involved in this process. Disrupting the function of proteins responsible for maintaining heterochromatin, such as HP1 [Brancucci et al., 2014], could be an effective strategy to block parasite replication during the asexual cycle. Another important event in the malaria life cycle is gametocytogenesis, which was recently shown to be driven by the transcription factor PfAP2-G [Kafsack et al., 2014, Sinha et al., 2014]. It would be interesting to fully characterize the epigenetic factors, such as genome architecture, that help PfAP2-G target and regulate gametocyte-specific genes. In addition to layers of epigenetic regulation we focused on here, post-transcriptional and translational controls are likely to be involved in the timing of protein expression [Suvorova and White, 2014, Kramer, 2014, Bunnik et al., 2013, Le Roch et al., 2004]. Increased insight into these regulatory processes would significantly advance our understanding of parasite biology and could mark a major breakthrough in our fight against malaria.

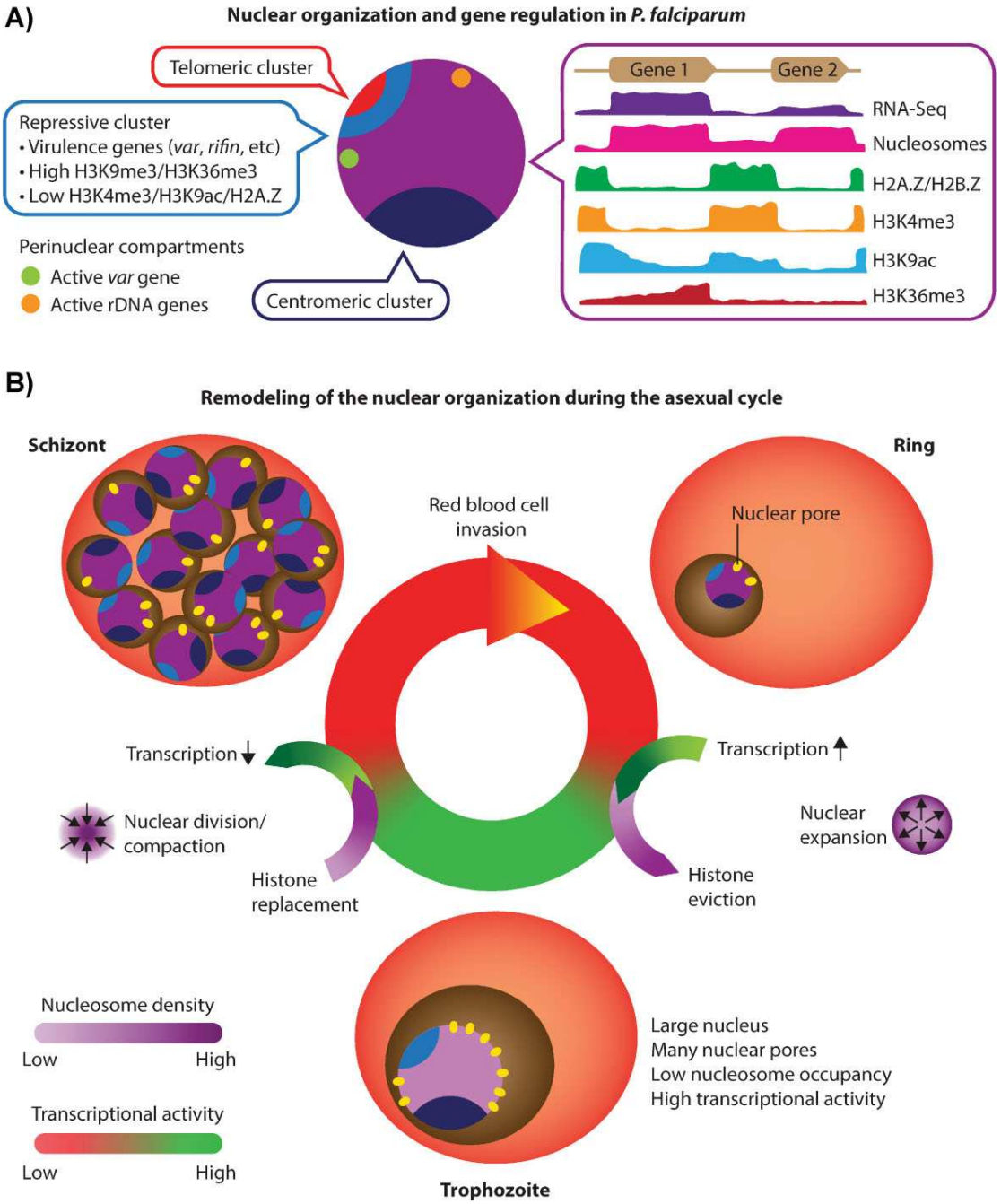

A)
Nuclear organization and gene regulation in P. falciparum
Telomeric cluster
Repressive cluster
• Virulence genes (var, rifin, etc)
• High H3K9me3/H3K36me3
• Low H3K4me3/H3K9ac/H2A.Z
Perinuclear compartments
Active var gene
Active rDNA genes
Centromeric cluster
Gene 1
Gene 2
RNA-Seq
Nucleosomes
H2A.Z/H2B.Z
H3K4me3
H3K9ac
H3K36me3
B)
Remodeling of the nuclear organization during the asexual cycle
Schizont
Ring
Red blood cell invasion
Nuclear pore
Transcription
Transcription
Nuclear division/ compaction
Nuclear expansion
Histone replacement
Histone eviction
Nucleosome density
Low
High
Transcriptional activity
Low
High
Large nucleus
Many nuclear pores
Low nucleosome occupancy
High transcriptional activity
Trophozoite

FIGURE 4 *(preceding page)*: **Model for *P. falciparum* epigenetic gene regulation. A:** Nuclear organization and gene regulation in *P. falciparum*. Centromeric (dark blue) and telomeric (red) clusters are localized at the nuclear periphery. Subtelomeric virulence genes (blue) are anchored to the nuclear perimeter and cluster with internally located var genes in repressive center(s), characterized by repressive histone marks H3K9me3 and H3K36me3. The single active var gene (green) is located in a perinuclear compartment away from the repressive center(s). In addition, active rDNA genes (orange) also cluster at the nuclear periphery. The remaining genome (purple) is largely present in an open, euchromatic state with a number of notable features. (i) Nucleosome levels are high in genic and lower in intergenic regions, while gene expression correlates with nucleosome density at the transcription start site. (ii) Intergenic regions are bound by nucleosomes containing histone variants H2A.Z and H2B.Z. (iii) Intergenic regions contain H3K4me3, the level of which does not influence transcriptional activity. (iv) H3K9ac is mainly found in intergenic regions and extends into 5' ends of coding regions, with highly expressed genes showing higher levels of H3K9ac. (v) Active genes are marked with H3K36me3 towards their 3' end. **B:** Remodeling of the nuclear organization during the asexual cycle. Extensive remodeling of the nucleus takes place as the parasite progresses through the ring, trophozoite and schizont stages. In the transition from the relatively inert ring stage to the transcriptionally active trophozoite stage, the size of the nucleus and the number of nuclear pores increase, accompanied by a decrease in genome-wide nucleosome levels, resulting in an open chromatin structure that allows high transcription rates. In the schizont stage, the nucleus divides and recompacts, histones are re-assembled and transcription is shut-down, to facilitate egress of the parasites' daughter cells and re-invasion of new red blood cells.

Bibliography

S. Andrews. FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc, 2010.

F. Ay, T. L. Bailey, and W. S. Noble. Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts. *Genome Research*, 24:999–1011, 2014a.

F. Ay, E. M. Bunnik, N. Varoquaux, S. M. Bol, J. Prudhomme, J.-P. Vert, W. S. Noble, and K. G. Le Roch. Three-dimensional modeling of the *P. falciparum* genome during the erythrocytic cycle reveals a strong connection between genome architecture and gene expression. *Genome Research*, 24:974–988, 2014b.

F. Ay, E. M. Bunnik, N. Varoquaux, J.-P. Vert, W. S. Noble, and K. G. Le Roch. Multiple dimensions of epigenetic gene regulation in the malaria parasite *Plasmodium falciparum*. *Bioessays*, 37(2):182–194, 2015a.

F. Ay, T. H. Vu, M. J. Zeitz, N. Varoquaux, J. E. Carette, J.-P. Vert, A. R. Hoffman, and W. S. Noble. Identifying multi-locus chromatin contacts in human cells using tethered multiple 3C. *BMC Genomics*, 16(121), 2015b.

F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.

C. R. Baker, B. B. Tuch, and A. D. Johnson. Extensive DNA-binding specificity divergence of a conserved transcription regulator. *Proc. Natl. Acad. Sci. U.S.A.*, 108(18): 7493–7498, May 2011.

D. A. Baker. Malaria gametocytogenesis. *Mol. Biochem. Parasitol.*, 172(2):57–65, Aug 2010.

S. Balaji, M. M. Babu, L. M. Iyer, and L. Aravind. Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Research*, 33(13):3994–4006, 2005.

L. H. Bannister, G. Margos, and J. M. Hopkins. Making a home for *Plasmodium* post-genomics: ultrastructural organization of the blood stages. In *Molecular Approaches to Malaria*, pages 24–49. ASM Press, 2005.

A. Barski, S. Cuddapah, K. Cui, T. Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129:823–837, 2007.

R. Bartfai, W. A. Hoeijmakers, A. M. Salcedo-Amaya, A. H. Smits E. Janssen-Megens, A. Kaan, M. Treeck, T. W. Gilberger, K. J. Francoijs, and H. G. Stunnenberg. H2A.Z demarcates intergenic regions of the *Plasmodium falciparum* epigenome that are dynamically marked by H3K9ac and H3K4me3. *PLoS Pathogens*, 6(12):e1001223, 2010.

M.S. Bartolomei, S. Zemel, and S. M. Tilghman. Parental imprinting of the mouse H19 gene. *Nature*, 351(6322):153–155, 1991.

D. I. Baruch, B. L. Pasloske, H. B. Singh, X. Bi, X. C. Ma, M. Feldman, T. F. Taraschi, and R. J. Howard. Cloning the *P. falciparum* gene encoding PfEMP1, a malarial variant antigen and adherence receptor on the surface of parasitized human erythrocytes. *Cell*, 82(1):77–87, Jul 1995.

A. R. Barutcu, A. J. Fritz, S. K. Zaidi, A. J. vanWijnen, J. B. Lian, J. L. Stein, J. A. Nickerson, A. N. Imbalzano, and G. S. Stein. C-ing the Genome: A Compendium of Chromosome Conformation Capture Methods to Study Higher-Order Chromatin Organization. *J. Cell. Physiol.*, Jun 2015.

D. Bau, A. Sanyal, B. R. Lajoie, E. Capriotti, M. Byron, J. B. Lawrence, J. Dekker, and M. A. Marti-Renom. The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. *Nat Struct Mol Biol*, 18(1):107–114, 2011.

C. W. Beitel, L. Froenicke, J. M. Lang, I. F. Korf, R. W. Michelmore, J. A. Eisen, and A. E. Darling. Strain- and plasmid-level deconvolution of a synthetic metagenome by sequencing proximity ligation products. *PeerJ*, 2:e415, 2014.

A. S. Belmont. Large-scale chromatin organization: the good, the surprising, and the still perplexing. *Curr Opin Cell Biol*, 26C:69–78, 2014.

S. Ben-Elazar, Z. Yakhini, and I. Yanai. Spatial localization of co-regulated genes exceeds genomic gene clustering in the saccharomyces cerevisiae genome. *Nucleic Acids Res*, 41(4):2191–2201, Feb 2013.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B*, 57:289–300, 1995.

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

B. E. Bernstein, M. Kamal, K. Lindblad-Toh, S. Bekiranov, D. K. Bailey, D. J. Huebert, S. McMahon, E. K. Karlsson, E. J. Kulbokas, T. R. Gingeras, S. L. Schreiber, and E. S. Lander. Genomic maps and comparative analysis of histone modifications in human and mouse. *Cell*, 120(2):169–181, 2005.

M. J. Best, N. Chakravarti, and V. A. Ubhaya. Minimizing separable convex functions subject to simple chain constraints. *SIAM J. on Optimization*, 10(3):658–672, July 1999. ISSN 1052-6234. doi: 10.1137/S1052623497314970. URL http://dx.doi.org/10.1137/S1052623497314970.

K. S. Bloom. Centromeric heterochromatin: the primordial segregation machine. *Annu. Rev. Genet.*, 48:457–484, Nov 2014.

E. Böer, G. Steinborn, G. Kunze, and G. Gellissen. Yeast expression platforms. *Appl. Microbiol. Biotechnol.*, 77(3):513–523, Dec 2007.

A. Bolzer, G. Kreth, I. Solovei, D. Koehler, K. Saracoglu, C. Fauth, S. Müller, R. Eils, C. Cremer, M. R. Speicher, and T. Cremer. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS Biology*, 3 (5):e157, 2005.

Z. Bozdech, M. Llinas, B. L. Pulliam, E. D. Wong, J. Zhu, and J. L. DeRisi. The transcriptome of the intraerythrocytic developmental cycle of *Plasmodium falciparum*. *PLoS Biology*, 1(1):e5, 2003.

N. M. Brancucci, N. L. Bertschi, L. Zhu, I. Niederwieser, W. H. Chin, R. Wampfler, C. Freymond, M. Rottmann, I. Felger, Z. Bozdech, and T. S. Voss. Heterochromatin protein 1 secures survival and transmission of malaria parasites. *Cell Host Microbe*, 16(2):165–176, Aug 2014.

K. Brogaard, L. Xi, J. P. Wang, and J. Widom. A map of nucleosome positions in yeast at base-pair resolution. *Nature*, 486(7404):496–501, Jun 2012.

J. D. Buenrostro, C. L. Araya, L. M. Chircus, C. J. Layton, H. Y. Chang, M. P. Snyder, and W. J. Greenleaf. Quantitative analysis of RNA-protein interactions on a massively parallel array reveals biophysical and evolutionary landscapes. *Nat. Biotechnol.*, 32 (6):562–568, Jun 2014.

P. C. Bull, B. S. Lowe, M. Kortok, C. S. Molyneux, C. I. Newbold, and K. Marsh. Parasite antigens on the infected red cell surface are targets for naturally acquired immunity to malaria. *Nat. Med.*, 4(3):358–360, Mar 1998.

E. M. Bunnik, D. W. Chung, M. Hamilton, N. Ponts, A. Saraf, J. Prudhomme, L. Florens, and K. G. Le Roch. Polysome profiling reveals translational control of gene expression in the human malaria parasite *Plasmodium falciparum*. *Genome Biology*, 14(11):R128, 2013.

E. M. Bunnik, A. Polishko, J. Prudhomme, N. Ponts, S. S. Gill, S. Lonardi, and K. G. Le Roch. DNA-encoded nucleosome occupancy is associated with transcription levels in the human malaria parasite *Plasmodium falciparum*. *BMC Genomics*, 15:347, 2014.

T. Bürckstümmer, C. Banning, P. Hainzl, R. Schobesberger, C. Kerzendorfer, F. M. Pauler, D. Chen, N. Them, F. Schischlik, M. Rebsamen, M. Smida, F. Fece de la Cruz, A. Lapao, M. Liszt, B. Eizinger, P. M. Guenzl, V. A. Blomen, T. Konopka, B. Gapp, K. Parapatics, B. Maier, J. Stöckl, W. Fischl, S. Salic, M. R. Taba Casari, S. Knapp, K. L. Bennett, C. Bock, J. Colinge, R. Kralovics, G. Ammerer, G. Casari, T. R. Brummelkamp, G. Superti-Furga, and S. M. Nijman. A reversible gene trap collection empowers haploid genetics in human cells. *Nature Methods*, 10(10):965–971, 2013.

J. N. Burton, A. Adey, R. P. Patwardhan, R. Qiu, J. O. Kitzman, and J. Shendure. Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nat Biotechnol*, 31(12):1119–1125, 2013.

J. N. Burton, I. Liachko, M. J. Dunham, and J. Shendure. Species-level deconvolution of metagenome assemblies with Hi-C-based contact probability maps. *G3 (Bethesda)*, 4(7):1339–1346, 2014.

K. Bystricky, P. Heun, L. Gehlen, J. Langowski, and S. M. Gasser. Long-range compaction and flexibility of interphase chromatin in budding yeast analyzed by high-resolution imaging techniques. *Proceedings of the National Academy of Sciences of the United States of America*, 101(47):16495–16500, 2004.

R.B. Calinski and J. Harabasz. A dendrite method for cluster analysis. *Comm. in Statistics*, 3:1–27, 1974.

T. L. Campbell, E. K. de Silva, K. L. Olszewski, O. Elemento, and M. Llinas. Identification and genome-wide prediction of DNA binding specificities for the ApiAP2 family of regulators from the malaria parasite. *PLoS Pathogens*, 6(10):e1001165, 2010.

Jan E. Carette, Carla P. Guimaraes, Malini Varadarajan, Annie S. Park, Irene Wuethrich, Alzbeta Godarova, Maciej Kotecki, Brent H. Cochran, Eric Spooner, Hidde L. Ploegh, and Thijn R. Brummelkamp. Haploid genetic screens in human cells identify host factors used by pathogens. *Science*, 326(5957):1231–1235, 2009.

B. K. Chaal, A. P. Gupta, B. D. Wastuwidyaningtyas, Y. H. Luah, and Z. Bozdech. Histone deacetylases play a major role in the transcriptional regulation of the *Plasmodium falciparum* life cycle. *PLoS Pathog.*, 6(1):e1000737, Jan 2010.

S. Chantalat, A. Depaux, P. Hery, S. Barral, J. Y. Thuret, S. Dimitrov, and M. Gerard. Histone H3 trimethylation at lysine 36 is associated with constitutive and facultative heterochromatin. *Genome Res.*, 21(9):1426–1437, Sep 2011.

Q. Chen, V. Fernandez, A. Sundstrom, M. Schlichtherle, S. Datta, P. Hagblom, and M. Wahlgren. Developmental selection of *var* gene expression in *Plasmodium falciparum*. *Nature*, 394(6691):392–395, 1998.

T. Chookajorn, R. Dzikowski, M. Frank, F. Li, A. Z. Jiwani, D. L. Hartl, and K. W. Deitsch. Epigenetic memory at malaria virulence genes. *Proc. Natl. Acad. Sci. U.S.A.*, 104(3):899–902, Jan 2007.

D. W. Chung, N. Ponts, S. Cervantes, and K. G. Le Roch. Post-translational modifications in *Plasmodium*: more than you think! *Mol. Biochem. Parasitol.*, 168(2): 123–134, Dec 2009.

B. I. Coleman, K. M. Skillman, R. H. Jiang, L. M. Childs, L. M. Altenhofen, M. Ganter, Y. Leung, I. Goldowitz, B. F. Kafsack, M. Marti, M. Llinas, C. O. Buckee, and M. T. Duraisingh. A *Plasmodium falciparum* histone deacetylase regulates antigenic variation and gametocyte conversion. *Cell Host Microbe*, 16(2):177–186, Aug 2014.

M. Contreras-Dominguez, C. B. Moraes, T. Dorval, A. Genovesio, F. M. Dossin, and L. H. Freitas-Junior. A modified fluorescence *in situ* hybridization protocol for *Plasmodium falciparum* greatly improves nuclear architecture conservation. *Molecular and Biochemical Parasitology*, 173(1):48–52, 2010.

P. R. Cook. The organization of replication and transcription. *Science*, 284:1790–1795, 1999.

G. Cottarel, J. H. Shero, P. Hieter, and J. H. Hegemann. A 125-base-pair CEN6 DNA fragment is sufficient for complete meiotic and mitotic centromere functions in *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, 9(8):3342–3349, Aug 1989.

R. M. Coulson, N. Hall, and C. A. Ouzounis. Comparative genomics of transcriptional control in the human malaria parasite *Plasmodium falciparum*. *Genome Research*, 14 (8):1548–1554, 2004.

A. Cournac, H. Marie-Nelly, M. Marbouty, R. Koszul, and J. Mozziconacci. Normalization of a chromosomal contact map. *BMC Genomics*, 13:436, 2012.

J. M. Cregg, I. Tolstorukov, A. Kusari, J. Sunga, K. Madden, and T. Chappell. Expression in the yeast *Pichia pastoris*. *Meth. Enzymol.*, 463:169–189, 2009.

M. Cremer, F. Grasser, C. Lanctot, S. Muller, M. Neusser, R. Zinner, I. Solovei, and T. Cremer. Multicolor 3D fluorescence in situ hybridization for imaging interphase chromosomes. *Methods Mol Biol*, 463:205–239, 2008.

T. Cremer and M. Cremer. Chromosome territories. *Cold Spring Harb Perspect Biol*, 2 (3):a003889, 2010.

T. Cremer, M. Cremer, S. Dietzel, S. Muller, I. Solovei, and S. Fakan. Chromosome territories-a functional nuclear landscape. *Curr. Opin. Cell Biol.*, 18:307–316, 2006.

L. Cui and J. Miao. Chromatin-mediated epigenetic regulation in the malaria parasite *Plasmodium falciparum*. *Eukaryotic Cell*, 9(8):1138–1149, Aug 2010.

L. Cui, J. Miao, and L. Cui. Cytotoxic effect of curcumin on malaria parasite *Plasmodium falciparum*: inhibition of histone acetylation and generation of reactive oxygen species. *Antimicrob. Agents Chemother.*, 51(2):488–494, Feb 2007.

L. Cui, J. Miao, T. Furuya, Q. Fan, X. Li, P. K. Rathod, X. Z. Su, and L. Cui. Histone acetyltransferase inhibitor anacardic acid causes changes in global gene expression during in vitro *Plasmodium falciparum* development. *Eukaryotic Cell*, 7(7):1200–1210, Jul 2008.

N. Dahan-Pasternak, A. Nasereddin, N. Kolevzon, M. Pe'er, W. Wong, V. Shinder, L. Turnbull, C. B. Whitchurch, M. Elbaum, and W. Timgilberger. Pfsec13 is an unusual chromatin-associated nucleoporin of *Plasmodium falciparum* that is essential for parasite proliferation in human erythrocytes. *J Cell Sci*, 126(Pt 14):3055–3069, 2013.

S. De and F. Michor. DNA replication timing and long-range DNA interactions predict mutational landscapes of cancer genomes. *Nat Biotechnol*, 29(12):1103–1108, 2011.

E. de Wit and W. de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes Dev*, 26(1):11–24, 2012.

K. Deitsch, M. Duraisingh, R. Dzikowski, A. Gunasekera, S. Khan, K. Le Roch, M. Llinas, G. Mair, V. McGovern, D. Roos, J. Shock, J. Sims, R. Wiegand, and E. Winzeler. Mechanisms of gene regulation in *Plasmodium*. *Am. J. Trop. Med. Hyg.*, 77(2):201–208, Aug 2007.

J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, 2002.

X. Deng, W. Ma, V. Ramani, A. Hill, F. Yang, F. Ay, J. B. Berletch, C. A. Blau, J. Shendure, Z. Duan, W. S. Noble, and C. M. Disteche. Bipartite structure of the inactive mouse X chromosome. *Genome Biol.*, 16:152, 2015.

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

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

Z. Duan, M. Andronescu, K. Schutz, C. Lee, J. Shendure, S. Fields, W. S. Noble, and C. Anthony Blau. A genome-wide 3C-method for characterizing the three-dimensional architectures of genomes. *Methods*, 58(3):277–288, 2012.

M. F. Duffy, S. A. Selvarajah, G. A. Josling, and M. Petter. The role of chromatin in *Plasmodium* gene expression. *Cell. Microbiol.*, 14(6):819–828, Jun 2012.

B. Dujon. Yeast evolutionary genomics. *Nat. Rev. Genet.*, 11(7):512–524, Jul 2010.

B. Dujon, D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuveglise, E. Talla, N. Goffard, L. Frangeul, M. Aigle, V. Anthouard, A. Babour, V. Barbe, S. Barnay, S. Blanchin, J. M. Beckerich, E. Beyne, C. Bleykasten, A. Boisrame, J. Boyer, L. Cattolico, F. Confanioleri, A. De Daruvar, L. Despons, E. Fabre, C. Fairhead, H. Ferry-Dumazet, A. Groppi, F. Hantraye, C. Hennequin, N. Jauniaux, P. Joyet, R. Kachouri, A. Kerrest, R. Koszul, M. Lemaire, I. Lesur, L. Ma, H. Muller, J. M. Nicaud, M. Nikolski, S. Oztas, O. Ozier-Kalogeropoulos, S. Pellenz, S. Potier, G. F. Richard, M. L. Straub, A. Suleau, D. Swennen, F. Tekaia, M. Wesolowski-Louvel, E. Westhof, B. Wirth, M. Zeniou-Meyer, I. Zivanovic, M. Bolotin-Fukuhara, A. Thierry, C. Bouchier, B. Caudron, C. Scarpelli, C. Gaillardin, J. Weissenbach, P. Wincker, and J. L. Souciet. Genome evolution in yeasts. *Nature*, 430(6995):35–44, Jul 2004.

I. Dunham, A. Kundaje, S. F. Aldred, P. J. Collins, C. A. Davis, F. Doyle, C. B. Epstein, S. Frietze, J. Harrow, R. Kaul, J. Khatun, B. R. Lajoie, S. G. Landt, B. K. Lee, F. Pauli, K. R. Rosenbloom, P. Sabo, A. Safi, A. Sanyal, N. Shoresh, J. M. Simon, L. Song, N. D. Trinklein, R. C. Altshuler, E. Birney, J. B. Brown, C. Cheng, S. Djebali, X. Dong, I. Dunham, J. Ernst, T. S. Furey, M. Gerstein, B. Giardine, M. Greven, R. C. Hardison, R. S. Harris, J. Herrero, M. M. Hoffman, S. Iyer, M. Kellis, J. Khatun, P. Kheradpour, A. Kundaje, T. Lassmann, Q. Li, X. Lin, G. K. Marinov, A. Merkel, A. Mortazavi, S. C. Parker, T. E. Reddy, J. Rozowsky, F. Schlesinger, R. E. Thurman, J. Wang, L. D. Ward, T. W. Whitfield, S. P. Wilder, W. Wu, H. S. Xi, K. Y. Yip, J. Zhuang, M. J. Pazin, R. F. Lowdon, L. A. Dillon, L. B. Adams, C. J. Kelly, J. Zhang, J. R. Wexler, E. D. Green, P. J. Good, E. A. Feingold, B. E. Bernstein, E. Birney, G. E. Crawford, J. Dekker, L. Elnitski, P. J. Farnham, M. Gerstein, M. C. Giddings, T. R. Gingeras, E. D. Green, R. Guigo, R. C. Hardison, T. J. Hubbard, M. Kellis, W. Kent, J. D. Lieb, E. H. Margulies, R. M. Myers, M. Snyder, J. A. Stamatoyannopoulos, S. A. Tenenbaum, Z. Weng, K. P. White, B. Wold, J. Khatun, Y. Yu, J. Wrobel, B. A. Risk, H. P. Gunawardena, H. C. Kuiper, C. W. Maier, L. Xie, X. Chen, M. C. Giddings, B. E. Bernstein, C. B. Epstein, N. Shoresh, J. Ernst, P. Kheradpour, T. S. Mikkelsen, S. Gillespie, A. Goren, O. Ram, X. Zhang, L. Wang, R. Issner, M. J. Coyne, T. Durham, M. Ku, T. Truong, L. D. Ward, R. C. Altshuler, M. L. Eaton, M. Kellis, S. Djebali, C. A. Davis, A. Merkel, A. Dobin, T. Lassmann, A. Mortazavi, A. Tanzer, J. Lagarde, W. Lin, F. Schlesinger, C. Xue, G. K. Marinov, J. Khatun, B. A. Williams, C. Zaleski, J. Rozowsky, M. Roder, F. Kokocinski, R. F. Abdelhamid, T. Alioto, I. Antoshechkin, M. T. Baer, P. Batut, I. Bell, K. Bell, S. Chakrabortty, X. Chen, J. Chrast, J. Curado, T. Derrien, J. Drenkow, E. Dumais, J. Dumais, R. Duttagupta, M. Fastuca, K. Fejes-Toth, P. Ferreira, S. Foissac, M. J. Fullwood, H. Gao, D. Gonzalez, A. Gordon, H. P. Gunawardena, C. Howald, S. Jha, R. Johnson, P. Kapranov, B. King, C. Kingswood, G. Li, O. J. Luo, E. Park, J. B. Preall, K. Presaud, P. Ribeca, B. A. Risk, D. Robyr, X. Ruan, M. Sammeth, K. S. Sandhu, L. Schaeffer, L. H. See, A. Shahab, J. Skancke, A. M. Suzuki, H. Takahashi, H. Tilgner, D. Trout, N. Walters, H. Wang, J. Wrobel, Y. Yu, Y. Hayashizaki, J. Harrow, M. Gerstein, T. J. Hubbard, A. Reymond, S. E. Antonarakis, G. J. Hannon, M. C. Giddings, Y. Ruan, B. Wold, P. Carninci, R. Guigo, T. R. Gingeras, K. R. Rosenbloom, C. A. Sloan, K. Learned, V. S. Malladi, M. C. Wong, G. P. Barber, M. S. Cline, T. R. Dreszer, S. G. Heitner, D. Karolchik, W. Kent, V. M. Kirkup, L. R. Meyer, J. C. Long, M. Maddren, B. J. Raney, T. S. Furey, L. Song, L. L. Grasfeder, P. G. Giresi, B. K. Lee, A. Battenhouse, N. C. Sheffield, J. M. Simon, K. A. Showers, A. Safi, D. London, A. A. Bhinge, C. Shestak, M. R. Schaner, S. K. Kim, Z. Z. Zhang, P. A. Mieczkowski, J. O. Mieczkowska, Z. Liu, R. M. McDaniell, Y. Ni, N. U. Rashid, M. J. Kim, S. Adar, Z. Zhang, T. Wang, D. Winter, D. Keefe, E. Birney, V. R. Iyer, J. D.

Lieb, G. E. Crawford, G. Li, K. S. Sandhu, M. Zheng, P. Wang, O. J. Luo, A. Shahab, M. J. Fullwood, X. Ruan, Y. Ruan, R. M. Myers, F. Pauli, B. A. Williams, J. Gertz, G. K. Marinov, T. E. Reddy, J. Vielmetter, E. Partridge, D. Trout, K. E. Varley, C. Gasper, A. Bansal, S. Pepke, P. Jain, H. Amrhein, K. M. Bowling, M. Anaya, M. K. Cross, B. King, M. A. Muratet, I. Antoshechkin, K. M. Newberry, K. McCue, A. S. Nesmith, K. I. Fisher-Aylor, B. Pusey, G. DeSalvo, S. L. Parker, S. Balasubramanian, N. S. Davis, S. K. Meadows, T. Eggleston, C. Gunter, J. Newberry, S. E. Levy, D. M. Absher, A. Mortazavi, W. H. Wong, B. Wold, M. J. Blow, A. Visel, L. A. Pennachio, L. Elnitski, E. H. Margulies, S. C. Parker, H. M. Petrykowska, A. Abyzov, B. Aken, D. Barrell, G. Barson, A. Berry, A. Bignell, V. Boychenko, G. Bussotti, J. Chrast, C. Davidson, T. Derrien, G. Despacio-Reyes, M. Diekhans, I. Ezkurdia, A. Frankish, J. Gilbert, J. M. Gonzalez, E. Griffiths, R. Harte, D. A. Hendrix, C. Howald, T. Hunt, I. Jungreis, M. Kay, E. Khurana, F. Kokocinski, J. Leng, M. F. Lin, J. Loveland, Z. Lu, D. Manthravadi, M. Mariotti, J. Mudge, G. Mukherjee, C. Notredame, B. Pei, J. M. Rodriguez, G. Saunders, A. Sboner, S. Searle, C. Sisu, C. Snow, C. Steward, A. Tanzer, E. Tapanari, M. L. Tress, M. J. van Baren, N. Walters, S. Washietl, L. Wilming, A. Zadissa, Z. Zhang, M. Brent, D. Haussler, M. Kellis, A. Valencia, M. Gerstein, A. Reymond, R. Guigo, J. Harrow, T. J. Hubbard, S. G. Landt, S. Frietze, A. Abyzov, N. Addleman, R. P. Alexander, R. K. Auerbach, S. Balasubramanian, K. Bettinger, N. Bhardwaj, A. P. Boyle, A. R. Cao, P. Cayting, A. Charos, Y. Cheng, C. Cheng, C. Eastman, G. Euskirchen, J. D. Fleming, F. Grubert, L. Habegger, M. Hariharan, A. Harmanci, S. Iyengar, V. X. Jin, K. J. Karczewski, M. Kasowski, P. Lacroute, H. Lam, N. Lamarre-Vincent, J. Leng, J. Lian, M. Lindahl-Allen, R. Min, B. Miotto, H. Monahan, Z. Moqtaderi, X. J. Mu, H. O'Geen, Z. Ouyang, D. Patacsil, B. Pei, D. Raha, L. Ramirez, B. Reed, J. Rozowsky, A. Sboner, M. Shi, C. Sisu, T. Slifer, H. Witt, L. Wu, X. Xu, K. K. Yan, X. Yang, K. Y. Yip, Z. Zhang, K. Struhl, S. M. Weissman, M. Gerstein, P. J. Farnham, M. Snyder, S. A. Tenenbaum, L. O. Penalva, F. Doyle, S. Karmakar, S. G. Landt, R. R. Bhanvadia, A. Choudhury, M. Domanus, L. Ma, J. Moran, D. Patacsil, T. Slifer, A. Victorsen, X. Yang, M. Snyder, T. Auer, L. Centanin, M. Eichenlaub, F. Gruhl, S. Heermann, B. Hoeckendorf, D. Inoue, T. Kellner, S. Kirchmaier, C. Mueller, R. Reinhardt, L. Schertel, S. Schneider, R. Sinn, B. Wittbrodt, J. Wittbrodt, Z. Weng, T. W. Whitfield, J. Wang, P. J. Collins, S. F. Aldred, N. D. Trinklein, E. C. Partridge, R. M. Myers, J. Dekker, G. Jain, B. R. Lajoie, A. Sanyal, G. Balasundaram, D. L. Bates, R. Byron, T. K. Canfield, M. J. Diegel, D. Dunn, A. K. Ebersol, T. Frum, K. Garg, E. Gist, R. Hansen, L. Boatman, E. Haugen, R. Humbert, G. Jain, A. K. Johnson, E. M. Johnson, T. V. Kutyavin, B. R. Lajoie, K. Lee, D. Lotakis, M. T. Maurano, S. J. Neph, F. V. Neri, E. D. Nguyen, H. Qu, A. P. Reynolds, V. Roach, E. Rynes, P. Sabo, M. E. Sanchez, R. S. Sandstrom, A. Sanyal, A. O. Shafer, A. B. Stergachis, S. Thomas, R. E. Thurman,

B. Vernot, J. Vierstra, S. Vong, H. Wang, M. A. Weaver, Y. Yan, M. Zhang, J. M. Akey, M. Bender, M. O. Dorschner, M. Groudine, M. J. MacCoss, P. Navas, G. Stamatoyannopoulos, R. Kaul, J. Dekker, J. A. Stamatoyannopoulos, I. Dunham, K. Beal, A. Brazma, P. Flicek, J. Herrero, N. Johnson, D. Keefe, M. Lukk, N. M. Luscombe, D. Sobral, J. M. Vaquerizas, S. P. Wilder, S. Batzoglou, A. Sidow, N. Hussami, S. Kyriazopoulou-Panagiotopoulou, M. W. Libbrecht, M. A. Schaub, A. Kundaje, R. C. Hardison, W. Miller, B. Giardine, R. S. Harris, W. Wu, P. J. Bickel, B. Banfai, N. P. Boley, J. B. Brown, H. Huang, Q. Li, J. J. Li, W. S. Noble, J. A. Bilmes, O. J. Buske, M. M. Hoffman, A. D. Sahu, P. V. Kharchenko, P. J. Park, D. Baker, J. Taylor, Z. Weng, S. Iyer, X. Dong, M. Greven, X. Lin, J. Wang, H. S. Xi, J. Zhuang, M. Gerstein, R. P. Alexander, S. Balasubramanian, C. Cheng, A. Harmanci, L. Lochovsky, R. Min, X. J. Mu, J. Rozowsky, K. K. Yan, K. Y. Yip, and E. Birney. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74, Sep 2012.

M. T. Duraisingh, T. S. Voss, A. J. Marty, M. F. Duffy, R. T. Good, J. K. Thompson, L. H. Freitas-Junior, A. Scherf, B. S. Crabb, and A. F. Cowman. Heterochromatin silencing and locus repositioning linked to regulation of virulence genes in *Plasmodium falciparum*. *Cell*, 121(1):13–24, 2005.

R. Dzikowski, F. Li, B. Amulic, A. Eisberg, M. Frank, S. Patel, T. E. Wellems, and K. W. Deitsch. Mechanisms underlying mutually exclusive expression of virulence genes by malaria parasites. *EMBO Reports*, 8(10):959–965, 2007.

ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489:57–74, 2012.

S. Feng, S. J. Cokus, V. Schubert, J. Zhai, M. Pellegrini, and S. E. Jacobsen. Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in *Arabidopsis*. *Mol. Cell*, 55(5):694–707, Sep 2014.

W. Feng, J. Bachant, D. Collingwood, M. K. Raghuraman, and B. J. Brewer. Centromere replication timing determines different forms of genomic instability in *Saccharomyces cerevisiae* checkpoint mutants during replication stress. *Genetics*, 183(4):1249–1260, Dec 2009.

C. Fernandez-Becerra, O. Pein, T. R. de Oliveira, M. M. Yamamoto, A. C. Cassola, C. Rocha, I. S. Soares, C. A. de Braganca Pereira, and H. A. del Portillo. Variant proteins of *Plasmodium vivax* are not clonally expressed in natural infections. *Mol. Microbiol.*, 58(3):648–658, Nov 2005.

M. A. Ferraiuolo, M. Rousseau, C. Miyamoto, S. Shenker, X. Q. Wang, M. Nadler, M. Blanchette, and J. Dostie. The three-dimensional architecture of *Hox* cluster silencing. *Nucleic Acids Res*, 21:7472–7484, 2010.

C. Flueck, R. Bartfai, I. Niederwieser, K. Witmer, B. T. Alako, S. Moes, Z. Bozdech, P. Jenoe, H. G. Stunnenberg, and T. S. Voss. A major role for the *Plasmodium falciparum* ApiAP2 protein PfSIP2 in chromosome end biology. *PLoS Pathogens*, 6 (2):e1000784, 2010.

Fajwel Fogel, Rodolphe Jenatton, Francis Bach, and Alexandre D'Aspremont. Convex relaxations for permutation problems. In C.J.C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1016–1024. Curran Associates, Inc., 2013. URL http://papers.nips.cc/paper/4986-convex-relaxations-for-permutation-problems.pdf.

L. H. Freitas-Junior, E. Bottius, L. A. Pirrit, K. W. Deitsch, and C. Scheidig. Frequent ectopic recombination of virulence factor genes in telomeric chromosome clusters of *P. falciparum*. *Nature*, 407(6807):1018–1022, 2000.

L. H. Freitas-Junior, R. Hernandez-Rivas, S. A. Ralph, D. Montiel-Condado, O. K. Ruvalcaba-Salazar, A. P. Rojas-Meza, L. Mâncio-Silva, R. J. Leal-Silvestre, A. M. Gontijo, S. Shorte, and A. Scherf. Telomeric heterochromatin propagation and histone acetylation control mutually exclusive expression of antigenic variation genes in malaria parasites. *Cell*, 121(1):25–36, 2005.

G. Fudenberg and L. A. Mirny. Higher-order chromatin structure: bridging physics and biology. *Curr Opin Genet Dev.*, 22(2):115–124, 2012.

M. J. Fullwood, M. H. Liu, Y. F. Pan, J., H. Xu, Y. B. Mohamed, Y. L. Orlov, S. Velkov, A. Ho, P. H. Mei, E. G. Chew, P. Y. Huang, W. J. Welboren, Y. Han, H.S. Ooi, P. N. Ariyaratne, V. B. Vega, Y. Luo, P. Y. Tan, P. Y. Choy, K. D. Wansa, B. Zhao, K. S. Lim, S. C. Leow, J. S. Yow, R. Joseph, H. Li, K.V. Desai, J. S. Thomsen, Y. K. Lee, R. K. Karuturi, T. Herve, G. Bourque, H.G. Stunnenberg, X. Ruan, V. Cacheux-Rataboul, W. K. Sung, E. T. Liu, C. L. Wei, E. Cheung, and Y. Ruan. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, 462(7269):58–64, 2009.

K. Ganesan, N. Ponmee, L. Jiang, J. W. Fowble, J. White, S. Kamchonwongpaisan, Y. Yuthavong, P. Wilairat, and P. K. Rathod. A genetically hard-wired metabolic transcriptome in *Plasmodium falciparum* fails to mount protective responses to lethal antifolates. *PLoS Pathogens*, 4:e10000214, 2008.

B. A. Garcia, S. B. Hake, R. L. Diaz, M. Kauer, S. A. Morris, J. Recht, J. Shabanowitz, N. Mishra, B. D. Strahl, C. D. Allis, and D. F. Hunt. Organismal differences in post-translational modifications in histones H3 and H4. *J. Biol. Chem.*, 282(10):7641–7655, Mar 2007.

A. A. Gavrilov, H. V. Chetverina, E. S. Chermnykh, S. V. Razin, and A. B. Chetverin. Quantitative analysis of genomic element interactions by molecular colony technique. *Nucleic Acids Research*, 42(5):e36, 2014.

M. Ghorbal, M. Gorman, C. R. Macpherson, R. M. Martins, A. Scherf, and J. J. Lopez-Rubio. Genome editing in the human malaria parasite *Plasmodium falciparum* using the CRISPR-Cas9 system. *Nat. Biotechnol.*, 32(8):819–821, Aug 2014.

E. Gomez-Diaz and V. G. Corces. Architectural proteins: regulators of 3D genome organization in cell fate. *Trends Cell Biol.*, 24(11):703–711, Nov 2014.

J. L. Gordon, K. P. Byrne, and K. H. Wolfe. Mechanisms of chromosome number evolution in yeast. *PLoS Genet.*, 7(7):e1002190, Jul 2011.

M. Gotta, T. Laroche, A. Formenton, L. Maillet, H. Scherthan, and S. M. Gasser. The clustering of telomeres and colocalization with Rap1, Sir3, and Sir4 proteins in wild-type *Saccharomyces cerevisiae*. *J. Cell Biol.*, 134(6):1349–1363, Sep 1996.

B. M. Greenwood, D. A. Fidock, D. E. Kyle, S. H. Kappe, P. L. Alonso, F. H. Collins, and P. E. Duffy. Malaria: progress, perils, and prospects for eradication. *J. Clin. Invest.*, 118(4):1266–1276, Apr 2008.

A. Y. Grosberg, S. K. Nechaev, and E. I. Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de Physique*, 49(12):2095–2100, 1988.

L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel. Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. *Nature*, 453 (7197):948–951, 2008.

B. Guillemette, A. R. Bataille, N. Gevry, M. Adam, M. Blanchette, F. Robert, and L. Gaudreau. Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning. *PLoS Biol.*, 3(12):e384, Dec 2005.

J. D. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and M. Vidal. Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430: 88–93, 2004.

S. Heinz, C. Benner, N. Spann, E. Bertolino, Y. C. Lin, P. Laslo, J. X. Cheng, C. Murre, H. Singh, and C. K. Glass. Simple combinations of lineage-determining transcription factors prime¡ i¿ cis¡/i¿-regulatory elements required for macrophage and b cell identities. *Molecular cell*, 38(4):576–589, 2010.

C. T. Hittinger. *Saccharomyces* diversity and evolution: a budding model genus. *Trends Genet.*, 29(5):309–317, May 2013.

W. A. Hoeijmakers, C. Flueck, K. J. Francoijs, A. H. Smits, J. Wetzel, J. C. Volz, A. F. Cowman, T. Voss, H. G. Stunnenberg, and R. Bártfai. *Plasmodium falciparum* centromeres display a unique epigenetic makeup and cluster prior to and during schizogony. *Cell Microbiology*, 14(9):1391–1401, 2012a.

W. A. Hoeijmakers, H. G. Stunnenberg, and R. Bartfai. Placing the *Plasmodium falciparum* epigenome on the map. *Trends Parasitol.*, 28(11):486–495, Nov 2012b.

W. A. Hoeijmakers, A. M. Salcedo-Amaya, A. H. Smits, K. J. Francoijs, M. Treeck, T. W. Gilberger, H. G. Stunnenberg, and R. Bartfai. H2A.Z/H2B.Z double-variant nucleosomes inhabit the AT-rich promoter regions of the *Plasmodium falciparum* genome. *Mol. Microbiol.*, 87(5):1061–1073, Mar 2013.

D. Homouz and A.S. Kudlicki. The 3D organization of the yeast genome correlates with co-expression and reflects functional relations between genes. *PLoS ONE*, 8(1):e54699, 2013.

P. Horrocks, E. Wong, K. Russell, and R. D. Emes. Control of gene expression in *Plasmodium falciparum* - ten years on. *Mol. Biochem. Parasitol.*, 164(1):9–25, Mar 2009.

M. Hu, K. Deng, S. Selvaraj, Z. Qin, B. Ren, and J. S. Liu. HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*, 28(23):3131–3133, 2012.

M. Hu, K. Deng, Z. Qin, J. Dixon, S. Selvaraj, J. Fang, B. Ren, and J. S. Liu. Bayesian inference of spatial organizations of chromosomes. *PLoS Comput Biol*, 9(1):e1002893, 2013.

M. Imakaev, G. Fudenberg, R. P. McCord, N. Naumova, A. Goloborodko, B. R. Lajoie, J. Dekker, and L. A. Mirny. Iterative correction of Hi-C data reveals hallmarks of chromosome organization. *Nat Methods*, 9:999–1003, 2012.

A. Jansen and K. J. Verstrepen. Nucleosome positioning in *Saccharomyces cerevisiae*. *Microbiol. Mol. Biol. Rev.*, 75(2):301–320, Jun 2011.

T. Jenuwein and C. D. Allis. Translating the histone code. *Science*, 293(5532):1074–1080, Aug 2001.

L. Jiang, J. Mu, Q. Zhang, T. Ni, P. Srinivasan, K. Rayavara, W. Yang, L. Turner, T. Lavstsen, and T. G. Theander. PfSETvs methylation of histone H3K36 represses virulence genes in *Plasmodium falciparum*. *Nature*, 499(7457):223–227, 2013.

F. Jin, Y. Li, J. R. Dixon, S. Selvaraj, Z. Ye, A. Y. Lee, C. A. Yen, A. D. S., C. A. Espinoza, and B. Ren. A high-resolution map of the three-dimensional chromatin interactome in human cells. *Nature*, 503(7475):290–294, 2013.

B. F. Kafsack, N. Rovira-Graells, T. G. Clark, C. Bancells, V. M. Crowley, S. G. Campino, A. E. Williams, L. G. Drought, D. P. Kwiatkowski, D. A. Baker, A. Cortes, and M. Llinas. A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature*, 507(7491):248–252, Mar 2014.

R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen. Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. *Nat Biotechnol*, 30(1):90–98, 2011.

N. Kaplan and J. Dekker. High-throughput genome scaffolding from in vivo DNA interaction frequency. *Nat Biotechnol*, 31(12):1143–1147, 2013.

T. H. Kim, L. O. Barrera, M. Zheng, C. Qu, M. A. Singer, T. A. Richmond, Y. Wu, R. D. Green, and B. Ren. A high-resolution map of active promoters in the human genome. *Nature*, 436(7052):876–880, Aug 2005.

A. Koren, H. J. Tsai, I. Tirosh, L. S. Burrack, N. Barkai, and J. Berman. Epigenetically-inherited centromere and neocentromere DNA replicates earliest in S-phase. *PLoS Genet.*, 6(8):e1001068, Aug 2010.

M. Kotecki, P. S. Reddy, and B. H. Cochran. Isolation and characterization of a near-haploid human cell line. *Exp Cell Res*, 252(2):273–280, 1999.

S. Kramer. RNA in development: how ribonucleoprotein granules regulate the life cycles of pathogenic protozoa. *Wiley Interdiscip Rev RNA*, 5(2):263–284, 2014.

J. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29:1–27, 1964.

J. B. Kruskal and Wish. M. *Multidimensional Scaling*. Sage Publications, Beverly Hills, CA, 1977.

M. Lachner, R. Sengupta, G. Schotta, and T. Jenuwein. Trilogies of histone lysine methylation as epigenetic landmarks of the eukaryotic genome. *Cold Spring Harb. Symp. Quant. Biol.*, 69:209–218, 2004.

B.R. Lajoie, J. Dekker, and N. Kaplan. The Hitchhiker's guide to Hi-C analysis: practical guidelines. *Methods*, 72:65–75, 2015.

C. Lambros and J. P. Vanderberg. Synchronization of *Plasmodium falciparum*, erythrocytic stages in culture. *The Journal of Parasitology*, 65(3):418–420, 1979.

P. R. Langer-Safer, M. Levine, and D. C. Ward. Immunological method for mapping genes on *Drosophila* polytene chromosomes. *Proceedings of the National Academy of Sciences of the United States of America*, 79(14):4381–4385, 1982.

B. Langmead and S. L. Salzberg. Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, 9(4):357–359, Apr 2012.

E. Lasonder, M. Treeck, M. Alam, and A. B. Tobin. Insights into the *Plasmodium falciparum* schizont phospho-proteome. *Microbes Infect.*, 14(10):811–819, Aug 2012.

T. B. K. Le, M. V. Imakaev, L. A. Mirny, and M. T. Laub. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science*, 342(6159):731–734, 2013.

K. G. Le Roch, Y. Zhou, P. L. Blair, M. Grainger, J. K. Moch, J. D. Haynes, P. de la Vega, A. A. Holder, S. Batalov, D. J. Carucci, and E. A. Winzeler. Discovery of gene function by expression profiling of the malaria parasite life cycle. *Science*, 301(5639): 1503–1508, 2003.

K. G. Le Roch, J. R. Johnson, L. Florens, Y. Zhou, A. Santrosyan, M. Grainger, S. F. Yan, K. C. Williamson, A. A. Holder, and D. J. Carucci. Global analysis of transcript and protein levels across the *Plasmodium falciparum* life cycle. *Genome Research*, 14: 2308–2318, 2004.

K. G. Le Roch, J. R. Johnson, H. Ahiboh, D. W. Chung, J. Prudhomme, D. Plouffe, K. Henson, Y. Zhou, W. Witola, J. R. Yates, III, C. B. Mamoun, E. A. Winzeler, and H. Vial. A systematic approach to understand the mechanism of action of the bisthiazolium compound T4 on the human malaria parasite, *Plasmodium falciparum*. *BMC Genomics*, 9:513, 2008.

K. G. Le Roch, D. W. Chung, and N. Ponts. Genomics and integrated systems biology in *Plasmodium falciparum*: a path to malaria control and eradication. *Parasite Immunol*, 34(2–3):50–60, 2011.

W. Lee, D. Tillo, N. Bray, R. H. Morse, R. W. Davis, T. R. Hughes, and C. Nislow. A high-resolution atlas of nucleosome occupancy in yeast. *Nat. Genet.*, 39(10):1235–1244, Oct 2007.

P. Lefrancois, G. M. Euskirchen, R. K. Auerbach, J. Rozowsky, T. Gibson, C. M. Yellman, M. Gerstein, and M. Snyder. Efficient yeast ChIP-Seq using multiplex short-read DNA sequencing. *BMC Genomics*, 10:37, 2009.

J. E. Lemieux, S. A. Kyes, T. D. Otto, A. I. Feller, R. T. Eastman, R. A. Pinches, M. Berriman, X. Z. Su, and C. I. Newbold. Genome-wide profiling of chromosome interactions in *Plasmodium falciparum* characterizes nuclear architecture and reconfigurations associated with antigenic variation. *Mol Microbiol*, 90(3):519—-537, 2013.

Annick Lesne, Julien Riposo, Paul Roger, Axel Cournac, and Julien Mozziconacci. 3D genome reconstruction from chromosomal contacts. *Nat. Methods*, 11(11):1141–1143, 2014.

G. Li, X. Ruan, R. K. Auerbach, K. S. Sandhu, M. Zheng, P. Wang, H. Mei Poh, Y. Goh, J. Lim, J. Zhang, H. S. Sim, S. Q. Peh, F. H. Mulawadi, C. T. Ong, Y. L. O., S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W. Sung, M. Snyder, and Y. Ruan. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, 148(1):84–98, Jan 2012.

H. Li and R. Durbin. Fast and accurate long-read alignment with Burrows–Wheeler transform. *Bioinformatics*, 26(5):589–595, 2010.

W. Li, K. Gong, Q. Li, F. Alber, and X.J. Zhou. Hi-Corrector: a fast, scalable and memory-efficient package for normalizing large-scale Hi-C data. *Bioinformatics*, 31 (6):960–962, 2015.

I. Liachko and M. J. Dunham. An autonomously replicating sequence for use in a wide range of budding yeasts. *FEMS Yeast Res.*, 14(2):364–367, Mar 2014.

I. Liachko, R. A. Youngblood, K. Tsui, K. L. Bubb, C. Queitsch, M. K. Raghuraman, C. Nislow, B. J. Brewer, and M. J. Dunham. GC-rich DNA elements enable replication origin activity in the methylotrophic yeast *Pichia pastoris*. *PLoS Genet.*, 10(3): e1004169, Mar 2014.

E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny,

E. S. Lander, and J. Dekker. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, 326(5950):289–293, 2009.

S. H. Lin and Y. C. Liao. CISA: contig integrator for sequence assembly of bacterial genomes. *PLoS ONE*, 8(3):e60843, 2013.

J. Q. Ling, T. Li, J. F. Hu, T. H. Vu, H. L. Chen, X. W. Qiu, A. M. Cherry, and A. R. Hoffman. CTCF mediates interchromosomal colocalization between *Igf2/H19* and *Wsb1/Nf1*. *Science*, 312(5771):269–272, 2006.

M. J. Lopez-Barragan, J. Lemieux, M. Quinones, K. C. Williamson, A. Molina-Cruz, K. Cui, C. Barillas-Mury, K. Zhao, and X.-Z. Su. Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC Genomics*, 12:587, 2011.

J. J. Lopez-Rubio, A. M. Gontijo, M. C. Nunes, N. Issar, R. Hernandez Rivas, and A. Scherf. 5' flanking region of var genes nucleate histone modification patterns linked to phenotypic inheritance of virulence traits in malaria parasites. *Mol. Microbiol.*, 66 (6):1296–1305, Dec 2007.

J.-J. Lopez-Rubio, L. Mancio-Silva, and A. Scherf. Genome-wide analysis of heterochromatin associates clonally variant gene regulation with perinuclear repressive centers in malaria parasites. *Cell*, 5:179–190, 2009.

D. B. Lyons, W. E. Allen, T. Goh, L. Tsai, G. Barnea, and S. Lomvardas. An epigenetic trap stabilizes singular olfactory receptor expression. *Cell*, 154(2):325–336, Jul 2013.

A. Magklara, A. Yen, B. M. Colquitt, E. J. Clowney, W. Allen, E. Markenscoff-Papadimitriou, Z. A. Evans, P. Kheradpour, G. Mountoufaris, C. Carey, G. Barnea, M. Kellis, and S. Lomvardas. An epigenetic signature for monoallelic olfactory receptor expression. *Cell*, 145(4):555–570, May 2011.

N. L. Mahy, P. E. Perry, and W. A. Bickmore. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *Journal of Cell Biology*, 159:753–763, 2002.

Harmit S. Malik and Steven Henikoff. Major evolutionary transitions in centromere complexity. *Cell*, 138(6):1067 – 1082, 2009. ISSN 0092-8674.

L. Mancio-Silva, Q. Zhang, C. Scheidig-Benatar, and A. Scherf. Clustering of dispersed ribosomal DNA and its role in gene regulation and chromosome-end associations in malaria parasites. *Proceedings of the National Academy of Sciences of the United States of America*, 107(34):15117–15122, 2010.

E. M. M. Manders, A. E. Visser, A. Koppen, W. C. de Leeuw, R. van Liere, G. J. Brakenhof, and R. van Driel. Four-dimensional imaging of chromatin dynamics during the assembly of the interphase nucleus. *Chromosome Research*, 11:537–547, 2003.

Martial Marbouty, Axel Cournac, Jean-François Flot, Hervé Marie-Nelly, Julien Mozziconacci, and Romain Koszul. Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms. *eLife*, 3, 2014.

H. Marie-Nelly, M. Marbouty, A. Cournac, J.-F. Flot, G. Liti, D. P. Parodi, S. Syan, N. Guillén, A. Margeot, C. Zimmer, and R. Koszul. High-quality genome (re)assembly using chromosomal contact data. *Nature Communications*, 5:5695+, December 2014a. ISSN 2041-1723. doi: 10.1038/ncomms6695.

H. Marie-Nelly, M. Marbouty, A. Cournac, G. Liti, G. Fischer, C. Zimmer, and R. Koszul. Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinformatics*, 30(15):2105–2113, 2014b.

T. N. Mavrich, C. Jiang, I. P. Ioshikhes, X. Li, B. J. Venters, S. J. Zanton, L. P. Tomsho, J. Qi, R. L. Glaser, S. C. Schuster, D. S. Gilmour, I. Albert, and B. F. Pugh. Nucleosome organization in the *Drosophila* genome. *Nature*, 453(7193):358–362, May 2008.

M. D. McDowall, M. A. Harris, A. Lock, K. Rutherford, D. M. Staines, J. Bahler, P. J. Kersey, S. G. Oliver, and V. Wood. PomBase 2015: updates to the fission yeast database. *Nucleic Acids Res.*, Oct 2014.

J. Miao, Q. Fan, L. Cui, J. Li, J. Li, and L. Cui. The malaria parasite *Plasmodium falciparum* histones: organization, expression, and acetylation. *Gene*, 369:53–65, 2006.

T. S. Mikkelsen, K. Manching, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T. K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander, and B. Bernstein. Genome-wide maps of chromatin state in pluripotent and lineage-commited cells. *Nature*, 448:553–560, 2007.

A. Minajigi, J. E. Froberg, C. Wei, H. Sunwoo, B. Kesner, D. Colognori, D. Lessing, B. Payer, M. Boukhali, W. Haas, and J. T. Lee. A comprehensive Xist interactome reveals cohesin repulsion and an RNA-directed chromosome conformation. *Science*, 349(6245), Jul 2015.

L. A. Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome Research*, 19(1):37–51, 2011.

T. Misteli. Beyond the sequence: Cellular organization of genome function. *Cell*, 128 (4):787–800, 2007.

T. Mizuguchi, G. Fudenberg, S. Mehta, J.-M. Belton, N. Taneja, H. D. Folco, P. FitzGerald, J. Dekker, L. Mirny, J. Barrowman, and S. I. S. Grewal. Cohesin-dependent globules and heterochromatin shape 3d genome architecture in *S. pombe*. *Nature*, 516 (7531):432–435, 2014.

A. W. Murray and J. W. Szostak. Pedigree analysis of plasmid segregation in yeast. *Cell*, 34(3):961–970, Oct 1983.

A. Murrell, S. Heeson, and W. Reik. Interaction between differentially methylated regions partitions the imprinted genes *Igf2* and *H19* into parent-specific chromatin loops. *Nature Genetics*, 36(8):889–893, 2004.

T. Nagano, Y. Lubling, T. J. Stevens, S. Schoenfelder, E. Yaffe, W. Dean, E. D. Laue, A. Tanay, and P. Fraser. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

N. Naumova, M. Imakaev, G. Fudenberg, Y. Zhan, B. R. Lajoie, L. A. Mirny, and J. Dekker. Organization of the mitotic chromosome. *Science*, 342(6161):948–953, 2013.

H. Nishida, T. Suzuki, S. Kondo, H. Miura, Y. Fujimura, and Y. Hayashizaki. Histone H3 acetylated at lysine 9 in promoter is associated with low nucleosome density in the vicinity of transcription start site in human cell. *Chromosome Res.*, 14(2):203–211, 2006.

E. P. Nora, B. R. Lajoie, E. G. Schulz, L. Giorgetti, I. Okamoto, N. Servant, T. Piolot, N. L. van Berkum, J. Meisig, J. Sedat, J. Gribnau, E. Barillot, N. Bluthgen, J. Dekker, and E. Heard. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398):381–385, 2012.

Elphège P Nora, Job Dekker, and Edith Heard. Segmental folding of chromosomes: a basis for structural and regulatory chromosomal neighborhoods? *Bioessays*, 35(9): 818–828, 2013.

R. J. Ober, S. Ram, and E. S. Ward. Localization accuracy in single-molecule microscopy. *Biophys. J.*, 86(2):1185–1200, Feb 2004.

T. D. Otto, D. Wilinski, S. Assefa, T. M. Keane TM, L. R. Sarry, U. Böhme, J. Lemieux, B. Barrell, A. Pain, M. Berriman, C. Newbold, and M. Llinás. New insights into the blood-stage transcriptome of *Plasmodium falciparum* using RNA-seq. *Molecular Microbiology*, 76(1):12–24, 2010.

Jonas Paulsen, Odin Gramstad, and Philippe Collas. Manifold based optimization for single-cell 3d genome reconstruction. *PLoS Comput Biol*, 11(8):e1004396, 08 2015. doi: 10.1371/journal.pcbi.1004396. URL http://dx.doi.org/10.1371%2Fjournal.pcbi.1004396.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

Y. Peng, H. C. Leung, S. M. Yiu, and F. Y. Chin. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, 28(11):1420–1428, Jun 2012.

M. Petter, S. A. Selvarajah, C. C. Lee, W. H. Chin, A. P. Gupta, Z. Bozdech, G. V. Brown, and M. F. Duffy. H2A.Z and H2B.Z double-variant nucleosomes define intergenic regions and dynamically occupy var gene promoters in the malaria parasite *Plasmodium falciparum*. *Mol. Microbiol.*, 87(6):1167–1182, Mar 2013.

T. J. Pohl, B. J. Brewer, and M. K. Raghuraman. Functional centromeres determine the activation time of pericentric origins of DNA replication in *Saccharomyces cerevisiae*. *PLoS Genet.*, 8(5):e1002677, 2012.

N. Ponts, E. Y. Harris, J. Prudhomme, I. Wick, C. Eckhardt-Ludka, G. R. Hicks, G. Hardiman, S. Lonardi, and K. G. Le Roch. Nucleosome landscape and control of transcription in the human malaria parasite. *Genome Research*, 20(2):228–238, 2010.

N. Ponts, E. Y. Harris, S. Lonardi, and K. G. Le Roch. Nucleosome occupancy at transcription start sites in the human malaria parasite: A hard-wired evolution of virulence? *Infection, Genetics and Evolution*, 11(4):716–724, 2011. doi: 10.1016/j.meegid.2010.08.002.

N. Ponts, L. Fu, E. Y. Harris, J. Zhang, D. W. Chung, M. C. Cervantes, J. Prudhomme, V. Atanasova-Penichon, E. Zehraoui, E. M. Bunnik, E. M. Rodrigues, S. Lonardi, G. R. Hicks, Y. Wang, and K. G. Le Roch. Genome-wide mapping of DNA methylation in the human malaria parasite *Plasmodium falciparum*. *Cell Host Microbe*, 14(6):696–706, Dec 2013.

X. Qiu, T. H. Vu, Q. Lu, J. Q. Ling, T. Li, A. Hou, S. K. Wang, H. L. Chen, J. F. Hu, and A. R. Hoffman. A complex deoxyribonucleic acid looping configuration associated with the silencing of the maternal *Igf2* allele. *Mol Endocrinol*, 22(6):1476–1488, 2008.

A. R. Quinlan and I. M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

R. M. Raisner, P. D. Hartley, M. D. Meneghini, M. Z. Bao, C. L. Liu, S. L. Schreiber, and O. J. Rando H. D. Madhani. Histone variant h2a.z marks the 5' ends of both active and inactive genes in euchromatin. *Cell*, 123(2):233–248, 2005.

S. A. Ralph, C. Scheidig-Benatar, and A. Scherf. Antigenic variation in *Plasmodium falciparum* is associated with movement of var loci between subnuclear locations. *Proceedings of the National Academy of Sciences of the United States of America*, 102 (15):5414–5419, 2005.

Suhas SP Rao, Miriam H Huntley, Neva C Durand, Elena K Stamenova, Ivan D Bochkov, James T Robinson, Adrian L Sanborn, Ido Machol, Arina D Omer, Eric S Lander, and Erez Lieberman Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 59(7):1665–1680, 2014.

D. J. Roberts, A. G. Craig, A. R. Berendt, R. Pinches, G. Nash, K. Marsh, and C. I. Newbold. Rapid switching to multiple antigenic and adhesive phenotypes in malaria. *Nature*, 357(6380):689–692, Jun 1992.

A. Rosa and C. Zimmer. Computational models of large-scale genome architecture. *Int. Rev. Cell Mol. Biol.*, 307:275–349, 2014.

M. Rousseau, J. Fraser, M. Ferraiuolo, J. Dostie, and M. Blanchette. Three-dimensional modeling of chromatin structure from interaction frequency data using Markov chain Monte Carlo sampling. *BMC Bioinformatics*, 12(1):414, October 2011. ISSN 1471-2105.

M. Rousseau, J. L. Crutchley, H. Miura, M. Suderman, M. Blanchette, and J. Dostie. Hox in motion: tracking HoxA cluster conformation during differentiation. *Nucleic Acids Res.*, 42(3):1524–1540, Feb 2014.

T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome Res*, 20(6):761–770, 2010.

A. M. Salcedo-Amaya, M. A. van Driel, B. T. Alako, M. B. Trelle, A. M. van den Elzen, A. M. Cohen, E. M. Janssen-Megens, M. van de Vegte-Bolmer, R. R. Selzer, A. L. Iniguez, R. D. Green, R. W. Sauerwein, O. N. Jensen, and H. G. Stunnenberg. Dynamic histone H3 epigenome marking during the intraerythrocytic cycle of *Plasmodium falciparum*. *Proc. Natl. Acad. Sci. U.S.A.*, 106(24):9655–9660, Jun 2009.

D. R. Scannell, O. A. Zill, A. Rokas, C. Payen, M. J. Dunham, M. B. Eisen, J. Rine, M. Johnston, and C. T. Hittinger. The Awesome Power of Yeast Evolutionary Genetics: New Genome Sequences and Strain Resources for the *Saccharomyces sensu stricto* Genus. *G3 (Bethesda)*, 1(1):11–25, Jun 2011.

B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, MA, 2002.

E. Segal and J. Widom. Poly(dA:dT) tracts: major determinants of nucleosome organization. *Curr. Opin. Struct. Biol.*, 19(1):65–71, Feb 2009.

E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thøaström, Y. Field, I. K. Moore, J. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 44(17): 772–778, 2006.

S. Selvaraj, R. Dixon J, V. Bansal, and B. Ren. Whole-genome haplotype reconstruction using proximity-ligation and shotgun sequencing. *Nat Biotechnol*, 31(12):1111–1118, 2013.

N. Servant, B. R. Lajoie, E. P. Nora, L. Giorgetti, C. J. Chen, E. Heard, J. Dekker, and E. Barillot. Hitc: exploration of high-throughput 'C' experiments. *Bioinformatics*, 28 (21):2843–2844, 2012.

Nicolas Servant, Nelle Varoquaux, Bryan Lajoie, Eric Viara, Chong-Jian Chen, Jean-Philippe Vert, Edith Heard, Job Dekker, and Emmanuel Barillot. HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biology*, 16(1):259, 2015. ISSN 1474-760X. doi: 10.1186/s13059-015-0831-x. URL http://genomebiology.com/2015/16/1/259.

T. Sexton, E. Yaffe, E. Kenigsberg, F. Bantignies, B. Leblanc, M. Hoichman, H. Parrinello, A. Tanay, and G. Cavalli. Three-dimensional folding and functional organization principles of the *Drosophila* genome. *Cell*, 148(3):458–472, 2012.

Y. Shen, F. Yue, D. F. McCleary, Z. Ye, L. Edsall, S. Kuan, U. Wagner, J. Dixon, L. Lee, V. V. Lobanenkov, and B. Ren. A map of the *cis*-regulatory sequences in the mouse genome. *Nature*, 488:116–120, 2012.

R. Shepard. The analysis of proximities: Multidimensional scaling with an unknown distance function. i. *Psychometrika*, 27:125–140, 1962. ISSN 0033-3123. URL http://dx.doi.org/10.1007/BF02289630. 10.1007/BF02289630.

T. N. Siegel, D. R. Hekstra, L. E. Kemp, L. M. Figueiredo, J. E. Lowell, D. Fenyo, X. Wang, S. Dewell, and G. A. Cross. Four histone variants mark the boundaries of polycistronic transcription units in *Trypanosoma brucei*. *Genes Dev.*, 23(9):1063–1076, May 2009.

A. Sinha, K. R. Hughes, K. K. Modrzynska, T. D. Otto, C. Pfander, N. J. Dickens, A. A. Religa, E. Bushell, A. L. Graham, R. Cameron, B. F. Kafsack, A. E. Williams, M. Llinas, M. Berriman, O. Billker, and A. P. Waters. A cascade of DNA-binding proteins for sexual commitment and development in *Plasmodium*. *Nature*, 507(7491): 253–257, Mar 2014.

J. D. Smith, C. E. Chitnis, A. G. Craig, D. J. Roberts, D. E. Hudson-Taylor, D. S. Peterson, R. Pinches, C. I. Newbold, and L. H. Miller. Switches in expression of *Plasmodium falciparum* var genes correlate with changes in antigenic and cytoadherent phenotypes of infected erythrocytes. *Cell*, 82(1):101–110, Jul 1995.

S. Sofueva, E. Yaffe, W. C. Chan, D. Georgopoulou, M. Vietri Rudan, H. Mira-Bontenbal, S. M. Pollard, G. P. Schroth, A. Tanay, and S. Hadjur. Cohesin-mediated interactions organize chromosomal domain architecture. *EMBO J.*, 32(24):3119–3129, Dec 2013.

B. V. Steensel and J. Dekker. Genomics tools for unraveling chromosome architecture. *Nat Biotechnol*, 28:1089–1095, 2010.

X. Z. Su, V. M. Heatwole, S. P. Wertheimer, F. Guinet, J. A. Herrfeldt, D. S. Peterson, J. A. Ravetch, and T. E. Wellems. The large diverse gene family var encodes proteins involved in cytoadherence and antigenic variation of em Plasmodium falciparum-infected erythrocytes. *Cell*, 82(1):89–100, Jul 1995.

A. Subramanian, P Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 102(43):15545–15550, 2005.

E. S. Suvorova and M. W. White. Transcript maturation in apicomplexan parasites. *Curr. Opin. Microbiol.*, 20:82–87, Aug 2014.

T. Takizawa, K. J. Meaburn, and T. Misteli. The meaning of gene positioning. *Cell*, 135(1):9–13, Oct 2008.

P. B. Talbert and S. Henikoff. Histone variants–ancient wrap artists of the epigenome. *Nat. Rev. Mol. Cell Biol.*, 11(4):264–275, Apr 2010.

H. Tanizawa, O. Iwasaki, A. tanaka, J. R. Capizzi, P. Wickramasignhe, M. Lee, Z. Fu, and K. Noma. Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation. *Nucleic Acids Res*, 38(22):8164–8177, 2010.

The Génolevures Consortium, J. L. Souciet, B. Dujon, C. Gaillardin, M. Johnston, P. V. Baret, P. Cliften, D. J. Sherman, J. Weissenbach, E. Westhof, P. Wincker, C. Jubin, J. Poulain, V. Barbe, B. Segurens, F. Artiguenave, V. Anthouard, B. Vacherie, M. E. Val, R. S. Fulton, P. Minx, R. Wilson, P. Durrens, G. Jean, C. Marck, T. Martin, M. Nikolski, T. Rolland, M. L. Seret, S. Casaregola, L. Despons, C. Fairhead, G. Fischer, I. Lafontaine, V. Leh, M. Lemaire, J. de Montigny, C. Neuveglise, A. Thierry, I. Blanc-Lenfle, C. Bleykasten, J. Diffels, E. Fritsch, L. Frangeul, A. Goeffon, N. Jauniaux, R. Kachouri-Lafond, C. Payen, S. Potier, L. Pribylova, C. Ozanne, G. F. Richard, C. Sacerdot, M. L. Straub, and E. Talla. Comparative genomics of protoploid *Saccharomycetaceae*. *Genome Res.*, 19(10):1696–1709, Oct 2009.

D. Tillo and T. R. Hughes. G+C content dominates intrinsic nucleosome occupancy. *BMC Bioinformatics*, 10:442, 2009.

H. Tjong, K. Gong, L. Chen, and F. Alber. Physical tethering and volume exclusion determine higher-order genome organization in budding yeast. *Genome Res*, 22(7): 1295–1305, 2012.

M. Y. Tolstorukov, P. V. Kharchenko, J. A. Goldman, R. E. Kingston, and P. J. Park. Comparative analysis of H2A.Z nucleosome organization in the human and yeast genomes. *Genome Res.*, 19(6):967–977, Jun 2009.

William Trager and James B Jensen. Human malaria parasites in continuous culture. *Science*, 193(4254):673–675, 1976.

M. Treeck, J. L. Sanders, J. E. Elias, and J. C. Boothroyd. The phosphoproteomes of *Plasmodium falciparum* and *Toxoplasma gondii* reveal unusual adaptations within and beyond the parasites' boundaries. *Cell Host Microbe*, 10(4):410–419, Oct 2011.

M. B. Trelle, A. M. Salcedo-Amaya, A. M. Cohen, H. G. Stunnenberg, and O. N. Jensen. Global histone analysis by mass spectrometry reveals a high content of acetylated lysine residues in the malaria parasite *Plasmodium falciparum*. *J. Proteome Res.*, 8 (7):3439–3450, Jul 2009.

Tuan Trieu and Jianlin Cheng. Large-scale reconstruction of 3D structures of human chromosomes from chromosomal contact data. *Nucleic Acids Res.*, 42(7):e52, 2014.

U.E. Ukaegbu, S. P. Kishore, D. L. Kwiatkowski, C. Pandarinath, N. Dahan-Pasternak, R. Dzikowski, and K. W. Deitsch. Recruitment of PfSET2 by RNA Polymerase II to variant antigen encoding loci contributes to antigenic variation in *P. falciparum*. *PLoS Pathog*, 10(1):e1003854, 2014.

M. A. Umbarger, E. Toro, M. A. Wright, G. J. Porreca, D. Bau, S. Hong, M. J. Fero, L. J. Zhu, M. A. Marti-Renom, H. H. McAdams, L. Shapiro, J. Dekker, and G. M. Church. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Molecular Cell*, 44:252–264, 2011.

S. van Koningsbruggen, M. Gierlinski, P. Schofield, D. Martin, G. Barton, Y. Ariyurek, J. T. den Dunnen, and A. I. Lamond. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Mol Biol Cell*, 21(21):3735–3748, 2010.

B. van Steensel and J. Dekker. Genomics tools for the unraveling of chromosome architecture. *Nature Biotechnology*, 28(10):1089–1095, 2010.

N. Varoquaux, F. Ay, W. S. Noble, and J.-P. Vert. A statistical approach for inferring the 3D structure of the genome. *Bioinformatics*, 30(12):i26–i33, 2014.

N. Varoquaux, I. Liachko, F. Ay, J. N. Burton, J. Shendure, M. Dunham, J-P. Vert, and W.S. Noble. Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Research*, 43(11):5331–5339, 2015.

J. Venema and D. Tollervey. Ribosome synthesis in *Saccharomyces cerevisiae*. *Annu. Rev. Genet.*, 33:261–311, 1999.

J.-P. Vert and M. Kanehisa. Extracting active pathways from gene expression data. *Bioinformatics*, 19(Suppl 2):ii238–244, 2003a.

J.-P. Vert and M. Kanehisa. Graph-driven features extraction from microarray data using diffusion kernels and kernel CCA. In *Advances in Neural Information Processing Systems 15*, pages 1425–1432, Cambridge, MA, 2003b. MIT Press.

T. Vettorel, A. Y. Grosberg, and K. Kremer. Statistics of polymer rings in the melt: a numerical simulation study. *Phys. Biol.*, 6(2):025013, 2009.

M. J. Vogel, D. Peric-Hupkes, and B. van Steensel. Detection of in vivo protein-DNA interactions using DamID in mammalian cells. *Nat Protoc*, 2(6):1467–1478, 2007.

T. S. Voss, J. Healer, A. J. Marty, M. F. Duffy, J. K. Thompson, J. G. Beeson, J. C. Reeder, B. S. Crabb, and A. F. Cowman. A var gene promoter controls allelic exclusion of virulence genes in *Plasmodium falciparum* malaria. *Nature*, 439(7079):1004–1008, Feb 2006.

T. S. Voss, Z. Bozdech, and R. Bartfai. Epigenetic memory takes center stage in the survival strategy of malaria parasites. *Curr. Opin. Microbiol.*, 20:88–95, Aug 2014.

T.H. Vu, A. H. Nguyen, and A. R. Hoffman. Loss of IGF2 imprinting is associated with abrogation of long-range intrachromosomal interactions in human cancer cells. *Hum Mol Genet*, 19(5), 2010.

A. Wächter and L. T. Biegler. On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. *Math Program*, 106(1):25–57, May 2006. ISSN 0025-5610. doi: 10.1007/s10107-004-0559-y. URL http://dx.doi.org/10.1007/s10107-004-0559-y.

J. C. Wagner, R. J. Platt, S. J. Goldfless, F. Zhang, and J. C. Niles. Efficient CRISPR-Cas9-mediated genome editing in *Plasmodium falciparum*. *Nat. Methods*, 11(9):915–918, Sep 2014.

Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T. Roh, W. Peng, M. Q. Zhang, and K. Zhao. Combinatorial patterns of histone acetylations and methylations in the human genome. *Nature Genetics*, 40(7):897–903, 2008.

A. Weiner, N. Dahan-Pasternak, E. Shimoni, V. Shinder, P. von Huth, M. Elbaum, and R. Dzikowski. 3D nuclear architecture reveals coupled cell cycle dynamics of chromatin and nuclear pores in the malaria parasite *Plasmodium falciparum*. *Cell Microbiology*, 13(7):967–977, 2011.

S. J. Westenberger, L. Cui, N. Dharia, E. Winzeler, and L. Cui. Genome-wide nucleosome mapping of *Plasmodium falciparum* reveals histone-rich coding and histone-poor intergenic regions and chromatin remodeling of core and subtelomeric genes. *BMC Genomics*, 10:610, 2009.

D. M. Witten and W. S. Noble. On the assessment of statistical significance of three-dimensional colocalization of sets of genomic elements. *Nucleic Acids Research*, 40(9): 3849–3855, 2012.

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. *Curr. Biol.*, 22(20):1881–1890, 2012.

Hua Wong, Jean-Michel Arbona, and Christophe Zimmer. How to build a yeast nucleus. *Nucleus*, 4(5):361–366, 2013.

World Health Organization. World malaria report, 2012.

E. Yaffe and A. Tanay. Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nat Genet*, 43:1059–1065, 2011.

G. E. Zentner and S. Henikoff. Regulation of nucleosome dynamics by histone modifications. *Nat. Struct. Mol. Biol.*, 20(3):259–266, Mar 2013.

C. Zhang, B. Xiao, Y. Jiang, Y. Zhao, Z. Li, H. Gao, Y. Ling, J. Wei, S. Li, M. Lu, X. Z. Su, H. Cui, and J. Yuan. Efficient editing of malaria parasite genome using the CRISPR/Cas9 system. *MBio*, 5(4):e01414–01414, 2014.

Y. Zhang, R. P. McCord, Y. Ho, B. R. Lajoie, D. G. Hildebrand, A. C. Simon, M. S. Becker, F. W. Alt, and J. Dekker. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. *Cell*, 148:1–14, 2012.

Z. Zhang, G. Li, K.-C. Toh, and W.-K. Sung. Inference of spatial organizations of chromosomes using semi-definite embedding approach and Hi-C data. In *Proceedings of the 17th International Conference on Research in Computational Molecular Biology*, volume 7821 of *Lecture Notes in Computer Science*, pages 317–332, Berlin, Heidelberg, 2013. Springer-Verlag.

Z. Zhao, G. Tavoosidana, M. Sjolinder, A. Gondor, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet*, 38(11): 1341–1347, 2006.

C. Zimmer and E. Fabre. Principles of chromosomal organization: lessons from yeast. *Journal of Cell Biology*, 192(5):723–733, 2011.

J. Zlatanova and A. Thakar. H2A.Z: view from the top. *Structure*, 16(2):166–179, Feb 2008.

Inférence de l'architecture 3D du génome

RÉSUMÉ : La structure de l'ADN, des chromosomes et l'organisation du génome sont des sujets fascinants du monde de la biologie. La plupart de la recherche s'est concentrée sur la structure unidimensionnelle du génome, l'organisation linéaire et la régulation des gènes. Cependant, le génome est avant tout organisé dans un espace euclidien tridimensionnel, et cette structure 3D, bien que moins étudiée, joue elle aussi un rôle important dans la fonction génomique de la cellule.

La capture de la conformation des chromosomes mesure en une seule expérience des interactions physiques entre paires de loci sur tout le génome, ouvrant la voie à des études systématiques et globales sur le repliement de l'ADN dans le noyau. Cependant, ces nouvelles technologies sont accompagnées de nombreux défis computationnelles et théoriques.

Dans cette thèse, je cherche à relever un certain nombre de ces défis, en particulier concernant l'inférence de modèle 3D de l'architecture du génome.

Mots clés : Inférence, structure tri-dimensionnelle, Hi-C, génome

Inferring the 3D structure of the genome

ABSTRACT : The structure of DNA, chromosomes and genome organization is a topic that has fascinated the field of biology for many years. Most research focused on the one-dimensional structure of the genome, studying the linear organizations of genes and genomes. Yet, the three-dimensional genome architecture is also thought to play an important role in many genomic functions.

Chromosome conformation capture based methods allow the measurement of physical interactions between pairs of loci, paving the way towards a systematic and genome wide analysis of how DNA folds into the nucleus. Yet, these technologies now poses important computational and theoretical challenges for which mathematically well grounded methods need to be developped.

In this thesis, we attempt to address some of the challenges faced while analysing such data, in particular on the topic of building 3D models of genome architecture from this data.

Keywords : Inference, three-dimensional structure, Hi-C, genome

MINES ParisTech

ParisTech
INSTITUT DES SCIENCES ET TECHNOLOGIES
PARIS INSTITUTE OF TECHNOLOGY