A probabilistic approach for genome assembly from high-throughput chromosome conformation capture data

Hervé Marie-Nelly

THÈSE DE DOCTORAT

DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité : Mathematiques Appliquées - Bioinformatique

École doctorale : "iViv: Interdisciplinaire pour le vivant"

réalisée

à l'Institut Pasteur

présentée par

Hervé MARIE-NELLY

pour obtenir le grade de :

DOCTEUR DE L'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

A probabilistic approach for genome assembly from high-throughput chromosome conformation capture data

soutenue le 13 décembre 2013

devant le jury composé de :

M.	Hugues Roest Crollius	Rapporteur
M.	Claude Thermes	Rapporteur
M.	Frédéric Devaux	Examinateur
Mme	Dawn Thompson	Examinateur
M.	Erwan Le Pennec	Examinateur
M.	Romain Koszul	Directeur de these
M.	Christophe Zimmer	Directeur de these

Abstract

High throughput DNA sequencing technologies are fuelling an accelerating trend to assemble genomes de novo or to complete unfinished assemblies of previously sequenced genomes. Unfortunately, common DNA sequencing technology is limited to reading stretches of a few hundreds or thousands of base pairs only. Therefore, computational methods are needed to assemble entire genomes from large numbers of short DNA strands. However, standard algorithms that piece together DNA strands with overlapping sequences face important limitations due, for example, to regions of repeated sequences, thus leaving many genome assemblies incomplete (Alkan et al., 2011 [2]).

We set out to develop a new methodology for genome assembly that promises to address some of these limitations. The method is based on Hi-C, a recent biochemical technique initially developed to analyse the 3D architecture of genomes (Lieberman-Aiden et al., 2009 [78]). In Hi-C experiments, DNA is crosslinked, cut by restriction enzymes, then diluted and religated. In standard Hi-C studies, a previously assembled genome is used to identify chimeric sequences among the ligation products, and map them to pairs of chromosomal loci, thereby yielding a genome-wide matrix of contact frequencies (Cournac et al., 2012 [27]). Our method essentially reverses this approach: Hi-C data are used to test for the physical continuity of the chromatin fibre as expected from a set of DNA segments (representing either a complete or incomplete chromosomal set). Physical-interactions aberrations in the contact matrix reveal structural incongruity, and lead to the reordering of chromosomal segments with respect to the physical properties and continuity of the fibre. This procedure improves genome assembly and/or identification of structural variants in re-sequenced genomes. Our approach uses a probabilistic (Bayesian) framework that assigns probabilities to different assemblies based on the experimental Hi-C data and on laws describing the physical properties of chromosomes (Wong et al. [146]). We will explain the methodology and the developed algorithms and provide results of applications to simulated and real Hi-C data from mutant and natural structural variants of yeast and fungi (Marie-Nelly et al., in prep). We also have developed algorithm that allow us to identify functional sequences in genomes from genomewide contact matrices. Notably, we annotated the centromeric position of the Naumovozyma castellii, an intriguing RNAi-containing yeast where centromere positions could not be determined with standard techniques (Marie-Nelly et al., submitted).

Contents

Chapter 1

Introduction

Contents

Over the past few years we have developed and implemented multiple algorithms to analyze and model complex biological mechanisms that take place inside the cell nucleus. These algorithms are required as, at this moment, it is often impossible to obtain accurate and robust observations of the cell nucleus by biological experiments alone.

Besides the technical and numerical aspects that come along with modeling, it is crucial to reason from a background that takes into account the characteristics of experimental setups as well as current knowledge of the biological objects under study. To provide such a framework we will first discuss the basics of nuclear organization, the spatial and functional genome architecture, and current experimental techniques that are available to acquire three dimensional genomic information. Only afterwards we are able to introduce the four parts of our research, concerning respectively image analysis, normalization of contact normalization, genome assembly and identification of centromeres from contact data, from the right context.

1.1 Nuclear organization

The cell is a basic unit that constitutes a living organism. Most living organisms are single cells such as bacteria or yeast. Other cells acquire specialized functions and cooperate with other cells to form large multi-cellular organisms such as animals. The cell functions thus as an individual unit and as a contributing part of a larger organism. As an individual unit, the cell is capable of consuming nutrients, synthesizing several types of molecules, generating its own energy and replicating itself in order to produce succeeding generations.
Cells are basic membrane based bound units, constituted of many structures which allow cell growing, division and function. The cells internal architecture defines its membership to the eukaryotic or prokaryotic kingdom. Eukaryotic cells are composed of membrane-bound structures or organelles like mitochondria and nucleus whereas the prokaryotic cells do not have any membrane-bound organelles, resulting in a freely floating cellular material within the cell (figure 1.1).

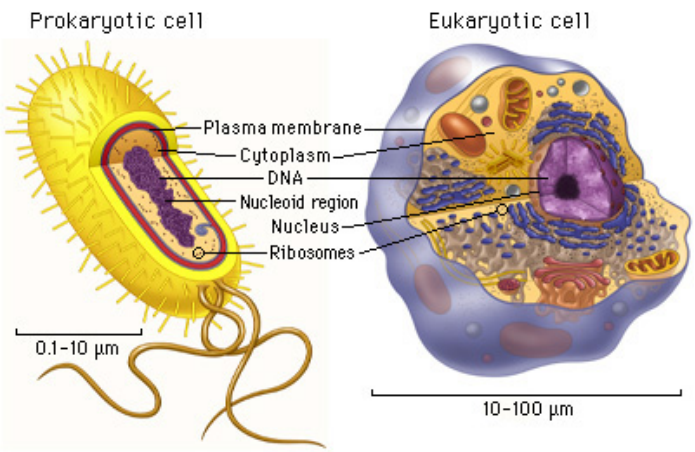


Figure 1.1: Representation of a prokaryotic and a eukaryotic cell structure.

1.1.1 Structure of the cell

All cells have similar components that work together to keep the cell alive. However, the organization of these components can differ from one cell type to another. The development of microscopes has been a key feature in providing a structural characterization of the cells.

Eukaryotic cells are large cells (from 10 to 100 μm) constituted of cytoplasmic membrane that forms a selective barrier, a cytoplasm that contains several small structures or organelles, and the nucleus that contains the genetic material necessary for cell growth and reproduction.

The cytoplasm or cytosol defines the space in which the cell structures evolve. It is delimited by the cytoplasmic membrane and contains an organized framework of fibres that constitute the cytoskeleton. This latter plays a key role in cellular function as it gives the cell its shape and facilitates the organelles movements as well as the cell movements. Besides, the cytosol contains several types of molecules involved in cellular biosynthesis.

The cytoplasmic organelles perform specific functions such as mitochondria which are responsible for energy production and cell survival. Other organelles are dedicated for example to the digestion of unwanted material in the cell (lysosome), to the direction of molecule production, processing, sorting and location (endoplasmic reticulum and the Golgi apparatus) or to the production of nutrients (chloroplasts).

The nucleus or the information center of the cell is a structure that is separated from the cytoplasm by the nuclear envelope. It contains the genetic information of the cell. For a cell, the loss of the nucleus is synonymous of a short lifespan and no cellular division. This is the case for red blood cells that eject their nucleus to accommodate maximum space for oxygen transport.

1.1.2 The nucleus

The nucleus is a doubled membrane compartment that contains the nuclear genome. It is the quintessential feature that defines eukaryotes. Compared to eukaryotes, the prokaryote (like bacteria) genome is present in a single circular DNA chromosome within the cytoplasm and is attached to the cytoplasmic membrane's surface. The phylogenetic tree of life points to the existence of three types of organisms: Bacteria, Archaea and Eucarya (Woese et al., 1990 [145]). It is commonly believed that eukaryotes precursors originate from the merge of bacterial and archaeal cells. This is supported by the independent symbiotic events that gave rise to both

mitochondria and chloroplasts (Margulis, 1970 [85]).
From a genomic point of view, the transition from prokaryotes to eukaryotes was the most radical change ever observed in cell organisation, where the appearance of the nucleus was accompanied by a huge burst of genomic material duplication. These genetic eukaryotic novelties brought along a significant sophistication of the mechanisms involved in DNA compartmentalization and segregation but also involved cellular metabolism. The apparition of the nucleus was accompanied by the assembly of endomembranes around the DNA to form the nuclear envelop by means of an evolution of nuclear pore complex that constitutes the crucial channel, bridging nucleoplasm to cytoplasm. Finally it was accompanied by the apparition of centromeres and mitotic spindle allowing for the stable inheritance of the genome along cell divisions.
The nucleus is constituted mainly of nuclear DNA that contains the necessary information for constructing the cell and directing its functions. In the following section we briefly explain the functional and structural features of the nucleus.

1.1.2.1 Functional and structural features of the nucleus

Global structures The nuclear genome is located within the cell nucleus that is separated from the cellular cytoplasm by the nuclear envelope. The nuclear envelope consists of two membranes, separated by the perinuclear space and constitute the geographical marker between the nuclear and cytoplasmic compartments. The nuclear pores are large protein complexes that cross the nuclear envelop. Through the pores not only nuclear products migrate but also cytoplasmic produced molecules and other elements that serve as communication mechanisms between the cellular environment, the cytoplasm and the nucleus. Import and export through the nuclear pores is however a selective mechanism.
The most prominent and visible structure that is found within the nucleus of all eukaryotic cells is the nucleolus. The number of nucleoli within a cell depends on the activity or differentiation state of the cell. Nucleolus is known to contain the ribosomic DNA (rDNA). But there are also other elements that characterize the nucleus. Such elements are composed of non-membrane-delineated structures that are called subnuclear bodies. These include Cajal bodies (structures bound to the nucleolus) (Gall et al., 1999), promyelocytic leukemia bodies (PML) (reviewed in Bernardi and Pandolfi, 2007 [9]), paraspeckles and speckles (Fox et al., 2002 [46]; Lamond and Spector, 2003 [75]). The different bodies indicate that the nucleoplasm is not composed of a uniform mixture but is organized into different functional subdomains.

Nuclear DNA DNA is a double helix chain consisting four bases: A, T, C and G (Adenine, Thymine, Cytosine and Guanine). The pairing of the bases A-T and C-G cause is responsible for maintaining the double helix structure. DNA is formed of millions of these aligned bases which at first sight seem to have a random distribution. The genetic information is contained in this apparent random sequence of bases.
If a DNA strand of a single cell would be completely stretched out it would measure several meters in length. In order to fit in a nucleus of only 10 μm in diameter the double DNA helix is wrapped around proteins called histones, thereby forming a complex called the chromatin. The chromatin forms several fibers in the nucleus, called chromosomes. The chromatin is organized in such a way that the information it contains can be transmitted accurately, rapidly and selectively.
A good example of the compaction and folding that the chromatin can undergo can be observed during cell division. Mitosis is the unique stage where highly condensed individual chromosomes can be visualized.

Nuclear function The primary function of the nucleus is the expression of a selected subset of genetic information that is contained in the DNA. This subset of genetic information is formed out of genes and code for proteins. Genes are defined as the transcribed regions of the genome and are constituted of a consecutive alignment of exons and introns. The genes are not directly decoded into proteins, though. The first step consists of making a transcription or copy of DNA into a messenger called RNA. Even if only the exons contain the coding sequence of the genes, during transcription the entire gene is transcribed in a pre-mRNA (including exons and introns). During transcription, the single stranded pre-mRNA is spliced, due to removing introns and joining exons. Besides, the RNA is becoming matured by the addition of several post-transcriptional modifications within the nucleus. The mature RNAs (mRNAs) are then transported from the nucleus into the cytoplasm passing through the nuclear envelope and the nuclear pores. Once in the cytoplasm, the RNAs are translated as they serve as templates for protein synthesis. This translation is triggered by ribosomes which are constituted of a ribosomal RNAs and protein complex.
In the following section we will explore the strong relation between the three dimensional conformation of DNA fiber and the regulation of gene expression.

1.2 Spatial and functional organization of the genome

The genomic DNA is organized into chromatin, which is organized into discrete functional units called chromosomes. Chromosomes consist of a single fragment of double-helix DNA wrapped around histone complexes, thereby forming nucleosomes. DNA-nucleosome units represent the second level of the chromotin organization into a 10 nm fiber which resembles a string of pearls. About 147 bp of DNA is wrapped around a heart that is composed of multiple histone dimers (two H3-H4 and two H2A-H2B). The nucleosomes are separated from one another by 10-80 bp of DNA 'linker' associated with histone H1. It has been shown *in vitro* that this 10 nm fiber is forming a helical fiber which has a diameter greater than 30 nm and contains 6 to 11 nucleosomes per helical turn (Finch and Klug, 1976 [45], Gerchman and Ramakrishnan, 1987 [52]). However, recent *in vivo* studies have questioned the existence of this helical fiber in both interphase and in metaphase (Fussner et al., 2011 [48], Eltsove et al., 2008 [40]; Maeshima and Eltsov, 2008 [83]). This goes against the idea of the existence of a higher level of genome organization in a fiber that is larger than the 10nm formed by basic alignment of chromatin forming nucleosomes.

Several studies have made it possible to accumulate evidence showing that the three-dimensional structure of chromatin is closely related to the function of the genome. The highly condensed and transcriptionally inactive heterochromatin as well as the relatively relaxed and transcriptionally active euchromatin form the two classic states of folded chromatin. These two chromatin states are marked by distinct epigenetic features (reviewed in Dillon, 2004 [33]). The folding of the chromatin corresponds to a number of principles that were demonstrated over the last decade.
The first principle is based on the non-random organization of chromosomes in the interphase nucleus into discrete domains called chromosome territories, which are to a certain level intertwined (Cremer and Cremer, 2010 [30]). The second principle governing chromatin folding corresponds to the evidence that the chromosomes are organized into large structural domains of approximately 1 Mb in length, corresponding to DNA replication units (Hiratani et al., 2008 [61]; Schwaiger et al., 2009 [124]; Ryba et al., 2010 [120]). In addition to these replication units other chromatin domains exist, such as those formed by the association of epigenetic repressive

marks like Polycomb or H3K9me (Tolhuis et al., 2006 [138], Schwartz et al., 2010[125]; Wen et al., 2009 [144]), those that are formed by the interaction of chromatin with nuclear structures such as peripheral lamina or the nucleolus (Lamina associating domains LAD and Nucleolar associating domains NAD) (Prokocimer et al., 2009 [110]; Németh et al., 2010 [98], Van Koningsbruggen et al., 2010 [140]), and topological domains characterized by frequent intra-domain interactions (TAD) (Dixon et al., 2012 [34]; Nora et al., 2012 [99]).

One of the impressive characteristics of the genome is its ability to recognize the parental origin of chromosomes. This has been observed in two well known mammalian mechanisms: genomic imprinting and X-chromosome inactivation. These mechanisms have shown to involve both interchromosomal and intrachromosomal interactions.

1.2.1 Parental origin recognition

1.2.1.1 Genomic imprinting

Genomic imprinting is an epigenetic phenomenon that is responsible for the monoallelic expression of a subset of genes (around 100 genes in mice). The imprint has been widely reported and studied in eutherian mammals and marsupials (reviewed in Reik and Jorn, 2001 [112]; Morison et al., 2005 [96]; Renfree et al., 2009 [113]). Only recently it has been reported that the mechanism is also present in flowering plants and in insects (reviewed in Kohler et al., 2012 [70]; Lloyd, 2000 [80]).

The study of imprinted genes has led to the discovery of long-range cis and trans-acting control elements which epigenetic state regulates both small clusters of genes and long non-coding RNAs.

The discovery of genomic imprinting has arisen from the observation that parthenogenetic reproduction of mammals (without fertilization) lead to an embryonic lethality, therefore indicating that the maternal genome alone cannot support a normal embryonic development (Markert, 1982 [86]). Besides, nuclear transfer experiments have shown that the paternal genome is not capable of ensuring a normal embryonic development either (Mcgrath and Solter, 1984[89]; Surani et al., 1984 [131]). In the same way, it has been shown that several human diseases are related to a disruption of the imprint within the imprinted loci (such as the Prader-Willi or the Beckwith–Wiedemann syndroms) (reviewed in Feinberg, 2007 [43]).

Further, it is important to note that genomic imprinting is responsible for a difference that has been observed centuries ago. Namely, the difference between a mule and a hinny and a liger and a tigon. Those animals that are the result of a breeding between two close species (a horse and a donkey, and lion and tiger) present different phenotypes depending on the parental sex. For example, a male lion and a female tiger will produce a liger (which grows larger than either of its parents) whereas a female lion and a male tiger will produce a tigon (which tends to be as large as its parents).

The mechanisms by which the imprinting is triggered involve epigenetic silencing through repressive DNA methylation of gene promoters, but also interchromosomal and intrachromosomal interactions. For example, the well characterized imprinted gene cluster is the one that contains the paternally expressed *Igf2* gene and the maternally expressed *H19* non coding RNA. The *H19* imprinting control region has shown to cause the silencing of the maternally inherited *Igf2* gene (Bartolomei et al., 1991 [7]). The silencing of the *Igf2* gene on the maternal allele is driven through a looping of *H19* with regions flanking the *Igf2* locus. This is what is causing the sequestration of the maternal copy of the gene into a small loop of silent chromatin (Kurukuti et al., 2006 [73]). Besides, it has been shown that the *H19* locus interacts in trans with up to four different

chromosomes (Zhao et al., 2006 [152]).

1.2.1.2 X chromosome inactivation

In mammals, a pair of chromosomes, called sex chromosomes, determines the male and female phenotype. Males have one X and one Y chromosome in their cells (XY) where females present two X chromosomes (XX). This cytological difference between the two sexes leads to an unbalance in the dosage of X-linked genes between males and females. In order to adjust this unbalance females have developed a mechanism during evolution, called X-chromosome inactivation (XCI), which role it is to compensate the dosage and equalize the X-linked gene products between the two sexes.

The studies over the past 50 years have characterized XCI mainly by using the mouse as model. XCI is a paradigm for epigenetic regulation at the level of the entire chromosome, rendering the cells functionally monosomic for the X chromosome. This monosomic expression of the X-chromosome is triggered by different epigenetic modifications like the expression of long non-coding RNAs, recruitment of repressive histone marks, repressive DNA methylation of X-linked gene promoters and specific nuclear positioning of the inactive chromosome territory close to heterochromatic compartments such as nuclear or nucleolar peripheries.

XCI is coordinated by a specific region which is located on the X-chromosome itself, the X-inactivation center (*Xic*). The X-inactivation center controls most, if not all of the steps of XCI including X-chromosome counting, the random choice of the X-chromosome to inactivate, and the initiation and maintenance of silencing almost all the 1000 genes of the X (Brown et al., 1991[15]).

X-inactivation is closely related to cell differentiation during *in vivo* development of female embryos but also during *in vitro* in female cells. The counting, the choice, and the initiation of silencing of one of the X chromosomes is completed in the preimplantation mouse embryo within the epiblast lineage. This lineage gives rise to the somatic cells or adult cells of the embryo itself (Puck et al., 1992). Once established, the pattern of XCI with random inactivation of either the paternal (X^P) or maternal X-chromosome (X^M) is stably propagated during mitosis, rendering thereby the female individuals' mosaic for XCI (figure 1.2).

Where in the epiblast, an embryonic lineage the choice of the X-chromosome for inactivation is random, in the extra-embryonic lineages of the mouse the paternal X-chromosome (X^P) is always inactivated (Takagi and Sasaki, 1975[134]). This fixed choice of paternal X-chromosome inactivation is called imprinted XCI (I-XCI) (figure 1.2).

In mice, I-XCI of the X^P is initiated early after fertilization, around the 4 cell stage. This silent state is later maintained in specifically the extra-embryonic lineages (the Trophectoderm and the Primitive endoderm) (figure 1.2).

X-chromosome inactivation controlling region XCI is controlled by the *Xic*, a 1 Mb wide region located on the X chromosome. This region has been identified by a series of experiments involving translocations and truncations of different regions of the X chromosome. The region is enriched in long non-coding RNAs (lncRNAs) (Brockdorff et al., 1991[14], Brown et al., 1991[15], Borsani et al., 1991[11]) amongst which the major actor of X inactivation: the *Xist* gene (figure 1.3).

Xist is a long ncRNA (17 Kb) that coats the chromosome from which it is expressed to form a repressive domain (Clemson et al., 1996 [21]). The silencing is triggered by the monoallelic upregulation of *Xist* during female development. Deletion from the X or inducible expression of the *Xist* gene on autosomes, have shown

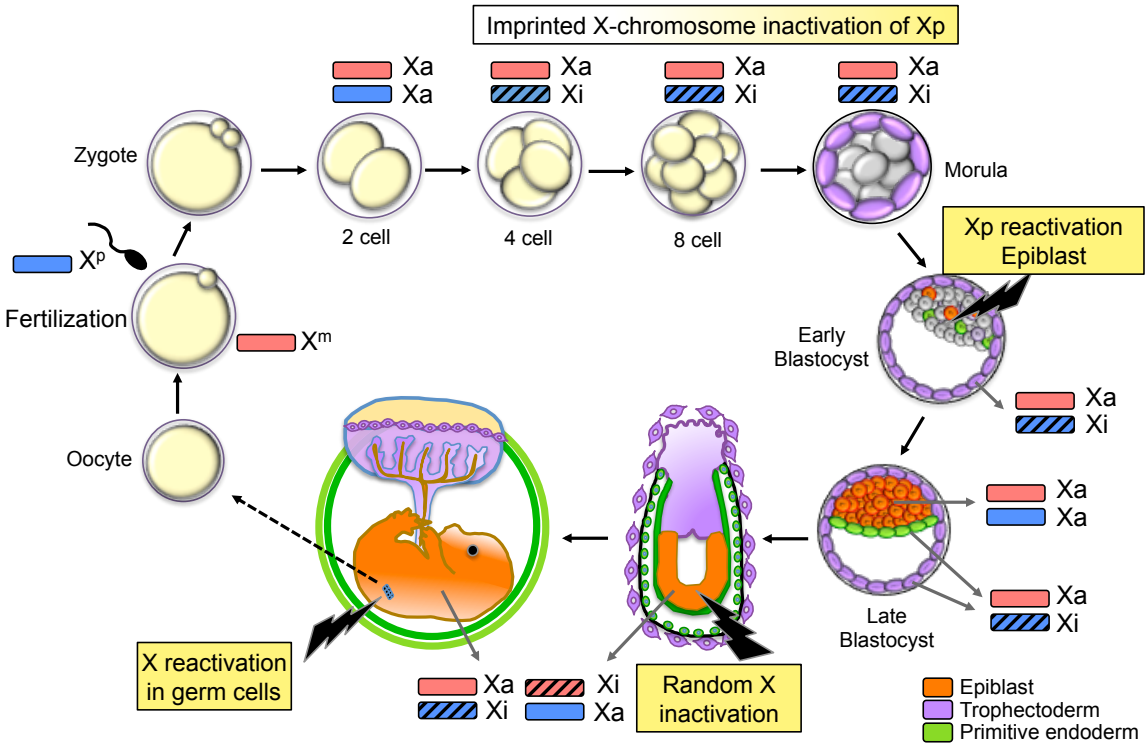


Figure 1.2: The X-chromosome inactivation and X-reactivation cycle during female mouse development.
A first wave of X-inactivation takes place around the 4 cell stage with a systematic inactivation of the paternal X-chromosome (X^P). This inactivation is called Imprinted X-chromosome inactivation (I-XCI) and is established progressively during the early preimplantation development. Where extra-embryonic lineages, including the trophectoderm (purple) and the primitive endoderm (green), maintain the imprinted X^P inactivation, the epiblast lineage, an embryonic tissue (orange), undergoes X^P-chromosome reactivation and initiates the second wave of X-inactivation which is random (X-inactivation of either X^P or X^M).

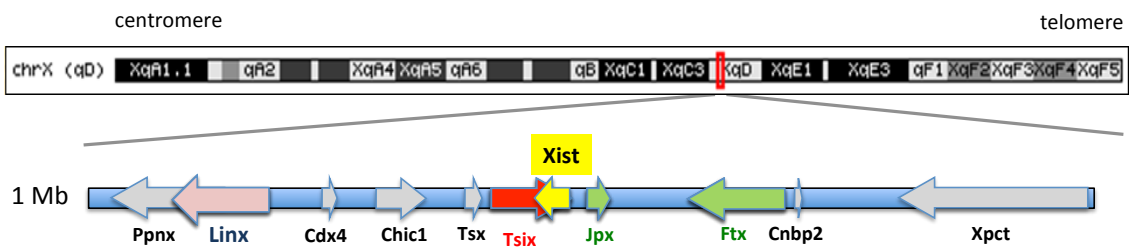


Figure 1.3: The mouse X-inactivation center (*Xic*).
Besides the *Xist* gene, the *Xic* includes several non coding RNAs (in green and red) as well as protein coding genes (in grey). The *Linx* gene (pink) is not well characterized. However, preliminary data suggest that *Linx* is a ncRNA which shares exons with the protein coding gene *Ppnx*. For all the genes the directions of the arrows indicate the transcriptional direction.

that this ncRNA is responsible for the *cis*-inactivation of the chromosome from which it is expressed (Penny et al., 1996[108], Marahrens et al., 1997[84]).

Besides, *Xist* has also been shown to be initially localized to distal sites across the chromosome by exploiting the three-dimensional conformation of the chromosome (Engreitz et al., 2013[41]) and gradually spread across in order to induce spatial reorganization of the inactive X-chromosome by creating a repressive nuclear compartment devoid of the transcription machinery and into which genes are recruited during their silencing (Chaumeil et al., 2006[18]).

Finally, the monoallelic upregulation of *Xist* in female cells involves different levels of control. This control takes

place mainly at the transcriptional level where *Xist* is found to be regulated by the binding of transcription factors but also by the expression of several long ncRNA.

1.2.2 Yeast nuclear organization

1.3 Techniques to investigate genome architecture

The genome architecture, including three-dimensional folding and the spatial organization of the chromosomes within the cell, is mainly studied by using two approaches: one involving imaging approaches and the other involving chromosome conformation capture strategies.

1.3.1 Imaging techniques

1.3.1.1 Fixed-cell imaging

Traditionally, the nuclear organization is studied by microscopy using an electron microscopy or fluorescent in situ hybridization approaches. The latter consists of a visualization at a large scale of the chromosome positions and organization but also of the visualization of chromatin domains and individual genes.

Fluorescent *in situ* hybridization For several decades, this *in situ* hybridization method has provided information concerning genome organization and the nuclear location of distinct and specific DNA sequences. The principle of this technique is to denature the chromosomal DNA within the cell and renature it in the presence of fluorescent labeled complementary DNA probes. After the renaturation or hybridization step, the fluorescent-labelled probes are detected by the use of a microscope, capturing and analyzing its images. Since the first application of this technique, several improvements have been added to enhance the probes specificity and increase the detection sensitivity and the resolution of the captured images.
The FISH technique applied on interphase or metaphase chromosome preparations has led to important discoveries. The technique has shown that chromosomes occupy distinct territories within the interphase cell nucleus. They preferably adopt a radial position within the nucleus (Cremer et al., 1982 [28], Bolzer et al., 2005 [10]; Cremer and Cremer, 2006 [29]). Therefore, large chromosomes are found more at the nuclear periphery whereas small chromosomes are more internalized. Besides, they have observed neighbor chromosomes intermingling at the periphery of their territories (Branco and Pombo, 2006 [12]).
Besides, the coupling of genome annotation with FISH experiments has showed that contiguous gene-poor and gene-rich regions are spatially separated within the nucleus (Shopland et al., 2006 [127])(figure 1.4).

The coupling between gene expression information and *in situ* hybridization observations suggests that the nuclear positioning of the genes affects their transcriptional activity. It has been shown that gene expression activation is accompanied by a looping out from the chromosomal territory (Chambeyron and Bickmore 2004 [17]; Ferrai et al. 2010b [44]) and that gene positioning compared to nuclear periphery and lamina association, pericentromeric heterochromatin or nucleolus, determines their transcriptional state (Kosak et al., 2002 [72]; Ragoczy et al., 2006 [111]; Meister et al., 2010 [90]).
Further, it has been suggested that the nuclear organization of genes is involved in their co-regulation as FISH experiments have shown that some genes present a spatial proximity regardless of their chromosomal location (Osborne et al. 2004 [103]).

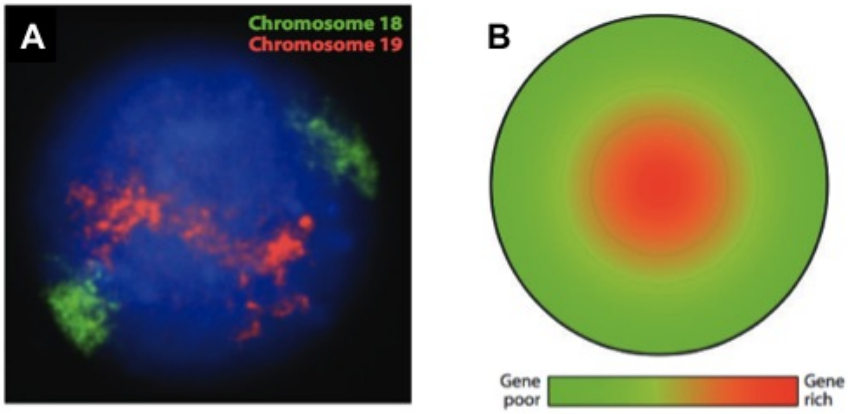


Figure 1.4: Nuclear organization of chromosomes in mammalian cells. A. Fluorescence in situ hybridization of a human cell where the gene-rich chromosome 19 is shown in red and the gene-poor chromosome 18 in green. The nuclear DNA is stained with Dapi. B. A scheme that presents the radial nuclear organization of gene-rich chromosome domains in the nucleus center where the gene-poor regions are located around the nuclear periphery (Bickmore, 2013).

This observation suggests that some genes present an ability to move within the nucleus and search for their "relatives" with limited constrains.

Electron microscopic imaging The sub-nuclear organization characterization is largely based on fluorescence and electron microscopy methods. However, fluorescence based approaches are not able to reveal the nuclear and macromolecular environment in which the tagged chromatin fibre or protein is evolving. Differential nuclear structure detection by fluorescent imaging is limited as visualization of the nuclear compartmentalization requires every time a specific labeling.

Therefore, it is largely recognized that higher-resolution techniques are required to define the ultra-structural landscape of the nucleus. Towards this goal, electron microscopy (EM) has played an important role in characterizing the nuclear ultra-structure. EM imaging has contributed to the discovery of the nucleosomal subunits (Olins and Olins, 1974 [102]; Oudet et al., 1975 [105]) and the actively transcribing ribosomal genes (Miller et al., 1969 [93]), thereby affecting greatly our conception of how DNA is organized and transcribed within the nucleus (figure 1.5).

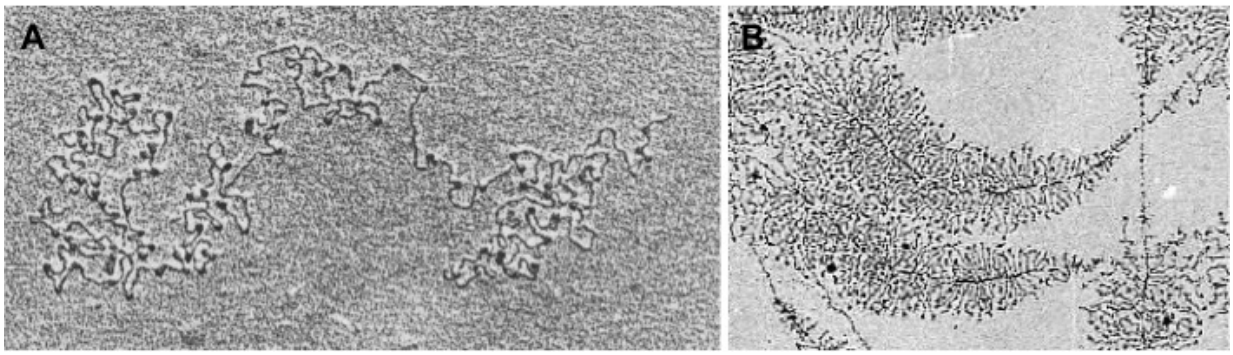


Figure 1.5: Electron microscopy imaging of chromatin. A. Nucleosome units or beads on a string imaging by electron microscopy (Oudet et al., 1975). B. Transcription of the DNA encoding ribosomal RNA (rRNA) molecules in the nucleolus (Miller laboratory).

The conventional EM involves the use of molecular stains which are responsible for the observed contrast on the images. However, this contrast is greatly influenced by the condensed/decondensed states of the chromatin within the nucleus. That means that in interphase nuclei the mainly decondensed chromatin results in low contrast images, which can be overcome by using other electron dependent imaging approaches.
The use of these methods allows access to both the nuclear surface and the interior by visualizing the nuclear envelope, the nuclear pore complexes, the spindle pole body and the imaging of the chromatin fibre at a high resolution (Kiseleva et al., 2004 [68]; O'Toole et al., 1999 [107]; Drummond et al., 2006 [37]; Arhel et al, 2007 [4]). Electron microscopy approaches are also coupled with fluorescent *in situ* hybridization, allowing for a targeted labeling and high-resolution imaging (Gerard et al., 2005 [51]).
However, the cause and effect questions concerning the mechanisms that are involved in the nuclear organization (such as peripheral positioning of gene-poor and heterochromatic regions) and their temporal parameters are not well defined with the fixed-cell approaches as the nucleus is pictured as a static structure. This apparent inertly of the genome can be overcome by using live-cell imaging.

1.3.1.2 Live-cell imaging

Genome organization has also been investigated by using a live imaging approach. Experiments designed to track particular regions within the genome reveal how a chromatin moves in response to transcriptional activation, how the chromatin organization influences the cellular state and also how the nuclear positioning influences the transcriptional efficiency.
Loci tagging that uses for example bacterial operator sequences or fluorescent dyes that allow for a spatial and temporal tracking of genomic regions has shown that the chromatin ability to move within the nucleus is limited. It has been shown that at large scales the chromatin of yeast, Drosophila or mammalian cells moves mostly by constrained diffusion. Fluorescence recovery after bleaching experiments (FRAP) has shown that the interphase chromatin is immobile over distance scales of 0.25 μm to 0.4 μm for time periods greater than one hour (Abney et al., 1997 [1]). In contrast to the chromatin's relative immobility, nuclear proteins involved in diverse nuclear processes (such as nucleosomal binding proteins, splicing factors, and rDNA processing proteins) have been shown to diffuse over the nucleus in less than a minute (Phair et al., 2000 [109]).
Further, the chromatin that is located at the nuclear or nucleolar periphery presents higher constrained motion compared to the one that is located more internally within the nucleus (Chubb et al., 2002 [20]; Marshall et al., 1997 [87]; Vazquez et al., 2001 [141]; Thakar et al., 2006 [135]). The interphase chromatin presents thus a certain diffusion but this diffusive motion is constrained such that a given chromatin fragment is free to move within a limited nuclear subregion. As this diffusion was shown to be independent from metabolic activities it has been suggested that it results from a classic passive Brownian motion rather than from an active motion (Marshall et al., 1997 [87]).
The restrictions on large-scale chromatin motion in the interphase nucleus are related to the attachment of the chromatin to nuclear substructures like the nuclear periphery, the nucleolus or the nuclear matrix. The tracking of chromosomes at different cell cycle stages has suggested that the interactions between chromosomes and internal nuclear structures modulate the range and the rate of chromatin diffusion (Vazquez et al., 2001 [141]; Chubb et al., 2002 [20]). However, genome regulating processes such as homologous chromosomes pairing, distant enhancer regulation of gene promoters, homology searching that accompanies DNA repair or recombination,

clearly implicate long distance movements of the interphase chromatin within the nucleus.
A study has shown that the induced targeting of a VP16-lac-repressor fusion protein to a repeated sequence of a lac-operator transgene induces its relocation from the nuclear periphery to a more interior position. This relocation from unfavorable to favorable transcriptional zones takes place through a series of curvilinear long-range movements interspersed with periods of constrained motion within a small radius (Chuang et al., 2006 [19]). This unidirectional and perpendicularly oriented movement along curvilinear paths presents a velocity of 0.1 to 0.9 μm/min. over a distance of 1 to 5 μm. The VP16 relocation suggests that a dynamic and active mechanism is involved in the fast and directed long-range movements of genomic loci which is directly or indirectly dependent on cytoskeletal proteins (Chuang et al., 2006 [19]) (figure 1.6).

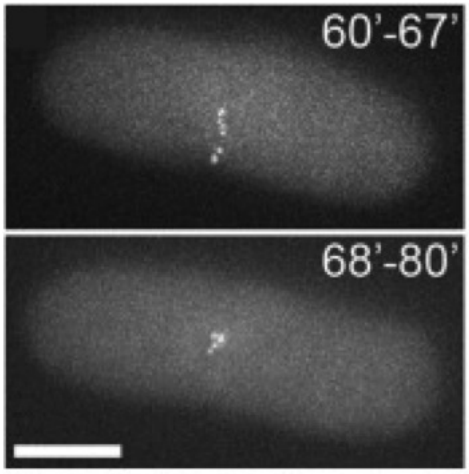


Figure 1.6: Long-distance movements along curvilinear trajectories of the VP16 tagged locus in live-cell after induction. The linear trajectory of the locus (60'-67') is followed by a long-period of localized short-range movements (68'-80')(Chuang et al., 2006 [19]).

Peripheral position of the genes within the nucleus is not always synonym of transcriptional repression. Observations of the lac/tet operator tagged galactose inducible genes in yeast have shown that upon activation those genes become confined to the nuclear envelope and that this association enhances their transcriptional activity. Therefore, this association has been shown to be dependent on several proteins such as export factors and nuclear-pore complex components (Cabal et al., 2006 [16]; Drubin et al., 2006 [36]; Taddei et al., 2006 [133]). This suggests that the proximity to the peripheral nuclear export machinery increases gene expression by facilitating RNA export or processing.
The fluorescent approaches using lac/tet operator systems allow live imaging of target genomic regions where the integration of large exogenous arrays of DNA-binding sequences is required, which lie around 10 kb into the genomic region of interest. However, this tagging is laborious, especially in mammalian cells, and not always efficient. The challenge of endogenous genomic sequence labeling can be overcome by using promising TALEs technologies.
TALEs (Transcription Activator-Like Effectors) are sequence specific DNA-binding proteins which recognize specific base pair sequences through their tandem amino acid repeats. Coupled to a fluorescent protein like GFP, TALEs mark directly their targeted sequence, for example, repeated elements in living cells (Le Cong et al., 2012 [25]; Miyanari et al., 2013 [95]).
The power of FISH methods and related microscopy approaches (such as live imaging) rely on the fact that they

allow single-cell analysis at the single-locus level. Besides, with this approach we can easily obtain information concerning the behavior of several loci in the same cell at the same time. These techniques present however some limitations such as a lack of large-scale genome information and limited throughput and resolution. Because of these restrictions, the imaging observations are limited and cannot be generalized to the whole genome.

1.3.2 Chromosome conformation capture techniques

All chromosome conformation techniques come from a study by Dekker et al. (Dekker et al., 2002 [32]) that describes the 3C method. This and other 3C-derived methods are used to establish a representation of the three-dimensional organization of the genome. To this end, the chromatin is first fixed by formaldehyde and then cut with a restriction enzyme that recognizes a particular motif within the DNA sequence. Subsequently, the sticky ends of the cross linked fragments are ligated under conditions that favor ligation between the pairs of cross linked fragments. In this way, DNA fragments are ligated that are distant according to the linear template but close to each other in the nuclear space.
At the end of the 3D conformation, the establishment of a particular locus or chromosome is determined by the measurements of ligation event numbers between linear non neighboring sites. This measurement can be made by different approaches such as PCR amplification (Polymerase chain reaction) or sequencing (figure 1.7).

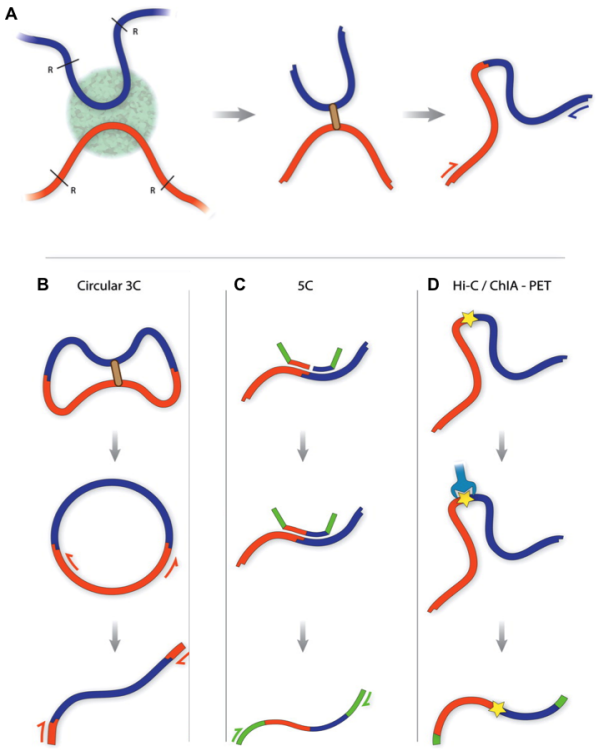


Figure 1.7: Chromosome conformation capture methods. A. Traditional 3C method. Spatially co-localized DNA sequences are cross linked and cut using a restriction enzyme. Cut DNA ends are ligated, reverse cross linked and purified. Ligation events are then quantified by PCR or sequencing. B. Circularized 3C or 4C. C. Carbon-Copy Chromosome Conformation Capture or 5C. D. HiC or genome wide chromosome conformation capture (Osborne et al, 2011 [104])

The 3C method The first 3C study on yeast has shown that the population average of 3D conformation of the chromosmome III forms a contorted ring. Besides, it has been shown in the same study that functionally

distinct AT and GC rich domains exhibit different conformations (Dekker et al., 2002 [32]).
After this study, the 3C technique was rapidly adapted to measure long-range interactions in mammalian cells showing a direct contact between distant regulatory elements and their target genes via a loop formation. 3C has been used to capture the physical association between the β-globine gene promoter and its locus control region, which is linearly located 50 kb upstream (Tolhuis et al., 2002 [139]).
3C experiments have also suggested that specific interactions can occur between loci that are located on different chromosomes. These inter-chromosomal interactions have been shown in two distinct mechanisms, one involving immune response regulation (Spilianakis et al., 2005 [129]) and the other involving homologous pairing of the X chromosomes before the onset of X-inactivation (Xu et al., 2007 [148]).
Based on 3C, several techniques have been developed to increase the throughput of interaction event measurements. These 3C related techniques can be categorized into 4 groups: the 3C, 4C (Circularized Chromosome Conformation Capture), 5C (Carbon-Copy Chromosome Conformation Capture) and HiC which consists of a whole genome conformation capture (figure 1.7).

The 4C method The 4C strategy presents a significant advantage over the 3C initial method because only one of the sites of interest needs to be known (Zhao et al., 2006 [152]; Simonis et al., 2006 [128]; Lomvardas et al., 2006 [81]; Wurtele and Chartrand, 2006 [147]). The circularized 3C consists of the study of DNA circles that contain the 'bait' sequence and an interacting partner sequence. The unknown interacting partner is finally amplified by inverse PCR or identified by sequencing. The 4C approach has been used to investigate the spatial environment of the homeotic gene *HoxB1* upon its induction. Besides, the use of 4C has showed that active and inactive chromatin domains form distinct interaction clusters. These domains correlate with the largely admitted view of a spatially segregated active and silent chromatin into euchromatin and heterochromatin.

The 5C method The 5C technique derived from the 3C allows for the parallel detection of interactions between many selected loci. This capturing of multiple interacting genomic regions concurrently gives a wider view of the genome interaction. The 5C technique developed by Dostie et al. (Dostie et al., 2006 [35]) allows us to gain insight into the nuclear organization and "interactome" at the β-globin locus. This method uses ligation-mediated amplifications, where primers that anneal at the 3C restriction sites are ligated together to generate an interaction library. The latter can be assayed by sequencing. Further, the 5C technique was used to characterize the interaction profiles at the X-inactivation center locus, a key region that controls X-chromosome inactivation (Nora et al., 2012 [99]).

The HiC method Recent improvements of the 3C method allows for the visualization of the entire genome interaction map. The Genome Conformation Capture (GCC) developed by Rodley et al. (2009 [117]) was used to capture yeast chromosome interactions with the sequencing of the whole 3C library without selection of ligation products. This means that the whole genome was sequenced, which was feasible because of the small size of the yeast genome. For larger genomes the ligation product selection was necessary.
The purification or selection of ligation junctions was undertaken by introducing a biotinylated nucleotide or an oligonucleotide tag sequence in between interacting ligated fragments.
Using the incorporated biotin-based ligation approach, the first megabase resolution, whole-genome conformation of the human nuclei was obtained (Lieberman-Aiden et al., 2009 [78]).

This study gave insight into the organization of active and inactive chromatin domains, chromosomes folding and preferred spatial arrangements of chromosome territories. Besides, this study has confirmed the fractal globular state of the chromosomes, a long-lived intermediate state of a compact polymer that was originally proposed by Grosberg et al. in 1988 ([59]).
HiC has been applied to different model systems (like *Drosophila* or mice) and improvements of sequencing technologies allows for the increase of coverage depth and resolution.
A recent improvement of the HiC approach has come from the Fraser laboratory, where they were able to generate a single-cell HiC map (Nagano et al., 2013 [97]). This improvement is crucial for the analysis of cell-to-cell chromosome structure variability. Hence, bridging the current gap between genomics and microscopic studies of the nuclear organization. However, even if the wide genome 3C method gives a great insight into the genome conformation, and even though we know that the link between spatial and functional organization of the genome is very tight, we have to take great care when analyzing this data and inferring any functional conclusions.

1.4 Our thesis work

Within the context of the theoretical framework that we have provided above, we were able to develop several algorithms to model complex biological mechanisms. These algorithms contribute to roughly four different components, to each of which we dedicate a chapter.

First, in chapter 2 we propose a way to detect and localize fluorescent probes in FISH images as well as a method to quantify the amount of nascent mRNA at transcription sites.

Chapter 3 presents our contribution to the treatment of HiC data. We discuss the way HiC data is analyzed, its workflow and its normalization procedures. In particular we analyze some of the main biases in HiC based experiments.

Chapter 4 then, embodies probably the heaviest component of our research: We discuss a computational method for complete genome assembly. Today's DNA sequencing technology is limited to reading stretches of a few hundreds or thousands of base pairs only. The algorithms we have developed enable us to assemble entire genomes from large numbers of such experimentally acquired short DNA strands.

Lastly, in chapter 5 we present how we can identify centromeres from contact data. In other words, we demonstrate how we can directly extract functional information about the sequence organization of genomes from a chromosome contact map. We apply these approaches to the complete genomic annotation of several yeast species.

The relevance of each of the methods we propose is based on the fact that it is not (yet) possible to obtain sufficiently detailed, accurate, and robust information from biological experiments alone. Our approaches are an attempt to infer as much accurate information as possible from the tools we have at our disposal. More knowledge about biological processes inside the cell nucleus would form valuable information for a vast amount of research domains, including our own existence. Hopefully, our work can get us a tiny step further in unraveling the mysteries of the cell nucleus.

Chapter 2

Image analysis

Contents

2.1 Introduction

In this chapter we will first briefly introduce some basic concepts about microscopy images in order to explain some of the algorithms we have implemented to detect and localize fluorescent probes in FISH images. The second part of this chapter is a published work introducing a methodology to quantify the amount of nascent mRNA at transciption sites. In the third part we will introduce an on going work about the 3D organization of the X chromosome.

2.1.1 Image formation model

Image formation model Because of the limited resolution of a visible light microscope, the image of a sample will not result in the exact reproduction of the reality. As a matter of fact, the image of a point source (a Dirac distribution of intensity) will not produce a point but a certain distribution of intensity. This is the impulse response of the optical device called the Point Spread Function (PSF).
The image yields by the camera can be considered as convolution of the PSF of the microscope with the image sample. In addition to the blurring caused by the PSF, the detected image contains a background signal coming from auto fluorescence, scattering and other electronic noise. This is why we will consider the simple image formation model that follows:

$$I(x, y, z) = (PSF \otimes f)(x, y, z) + b(x, y, z)$$

where, I is the final image, f the sample, and b the background which is made of random noise. This model does not take into account the presence of Poisson noise.

Gaussian Approximation of the Point Spread Function It has been demonstrated that the PSF of a wide field fluorescence microscope can be well approximated by a Gaussian function (Zhang, Zerubia and

Olivo-Marin, 2007 [151]). This function can be defined as:

$$g_{\sigma_\rho,\sigma_z}(x,y,z) = \exp\left(-\frac{x^2+y^2}{2\sigma_\rho^2} - \frac{z^2}{2\sigma_z^2}\right)$$

Where the optimal parameters , $\sigma^\star = \{\sigma_\rho^\star, \sigma_z^\star\}$ according to the least square criterion:

$$\sigma^\star = argmin_{\sigma>0}||PSF - g_{\sigma_\rho,\sigma_z}||_2^2$$

are:

$$\sigma_\rho^\star = \frac{\sqrt{2}}{k_{em}NA}\sigma_z^\star = 2\sqrt{6}\frac{n}{k_{em}NA^2}$$

where NA is the numerical aperture of the microscope and k_{em} the emission wavelength of the fluorophore.($k_{em}^{CY3} =$ 561 and $k_{em}^{FITC} = 538$).

If we assume that the background b follows a gaussian distribution$\mathcal{N}(\mu, \sigma^2)$ the image formation model becomes:

$$I = \left(Ag_{\sigma_\rho,\sigma_z} \otimes f\right) + \mathcal{N}(\mu, \sigma^2)$$

and the SNR can be defined as:

$$SNR = \left(\frac{A}{\sigma}\right)$$

Figure 2.1 displays the theoretical PSF and the real PSF in CY3 images.

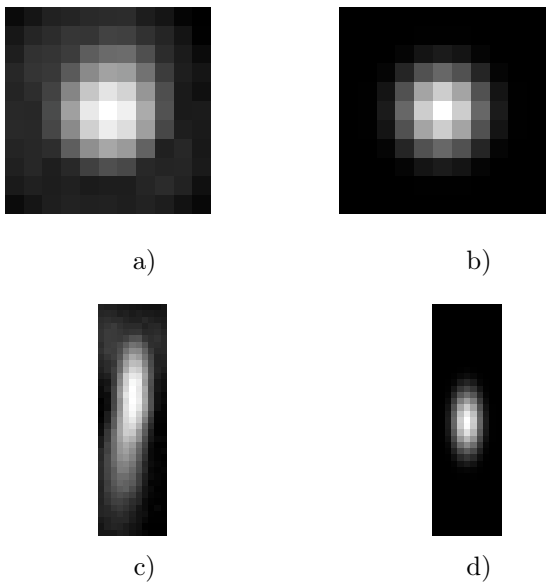


Figure 2.1: Real and theoretical PSF: The top and bottom rows display maximum intensity projection of 3D stacks along z and x respectively. a) and c) real images, b) and d) theoretical PSF

2.1.2 Predetection algortihm

We have chosen to implement, using MATLAB, the predetection technique developed by Thomman et al. (2002 [136]). Thomman et al. (2002 [136]) aim to detect and localize with great accuracy fluorescent tag in 3D. We

have analyzed the predetection scheme and tried to improve it.

2.1.2.1 Thomann et al. (2002) method

This method can be split in two parts: the Noise removal and the Single Spot Detection

Noise removal The study of living cell implies that the samples cannot be exposed to irradiation during a long time. Therefore, typical fluorescence images have a very low SNR ≈ 5 .
The cells of the experiments have been fixed so that there is no problem of short exposure. However, the SNR we have measured remains low.

Images	SNR CY3	SNR FITC
Data Set 1	20.61 db	22.11 db
Data Set 2	21.52 db	20.34 db
Data Set 3	17.99 db	20.42 db

In order to suppress local maximum coming from background noise and increase the SNR, the authors proposed to filter the raw data with a 3D Gaussian kernel which parameters correspond approximate the PSF.
We have decided to adopt the following strategy to remove the background of the images and increase the SNR.

- Convolve the image with a kernel at least twice bigger than the theoretical Gaussian kernel

$$I_{2k} = I \otimes g_{2\sigma_\rho, 2\sigma_z}$$

- Substract this image to the original one

$$I_{filt1} = I - I_{2k}$$

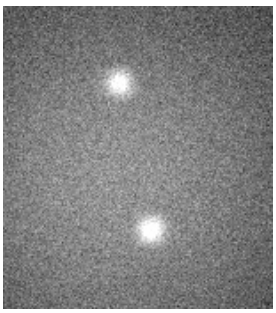
  If we assume that the signal detected I is made of the true signal and a background b, this process tends to decrease the background noise and increase the SNR. This process has been implemented on a synthetic image. The intensity profile and the pre and post filtered images are shown in figure 2.2.

- Convolve the filtered image with the theoretical Gaussian kernel.
  After the substraction process, the spots do not spread in space as much as before this operation. In order to avoid over-blurring the images are not convolved with the theoretical gaussian kernel but with a kernel which parameters are smaller.

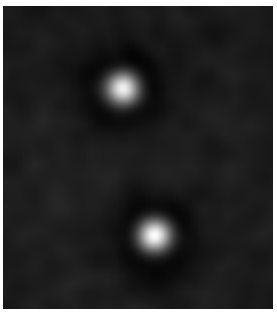
$$I_{filt2} = I_{filt1} \otimes g_{\alpha\sigma_\rho, \alpha\sigma_z}$$

  We have decided to set $\alpha = \frac{1}{2}$ for the two images. The results of this process on our data is show in figure 2.3.

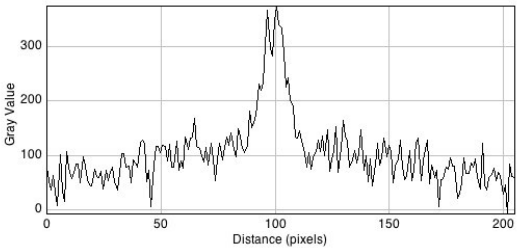
Single Spot Detection After denoising, a non maximum suppression algorithm is applied to the image. This process produces a set of spots candidates L. However, because of the low SNR of the data many recorded positions do not correspond to real RNA. Therefore, a new score has been defined to decide whether a maximum ξ is a "true" spot or not.
The authors have proposed a discriminating factor based on the curvature of the intensity distribution κ and



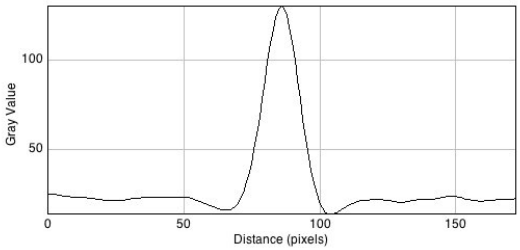
Spot in the raw data (I)



Spot after filtering(I_{filt2})

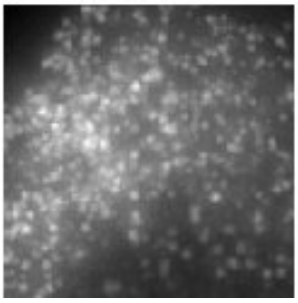


Intensity profil of the raw data

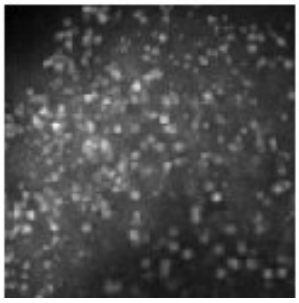


Intensity profile after filtering

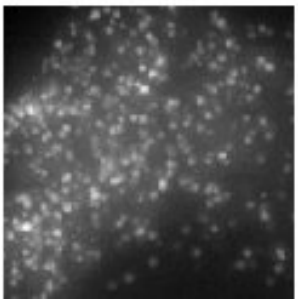
Figure 2.2: Denoising by Substraction on synthetic data



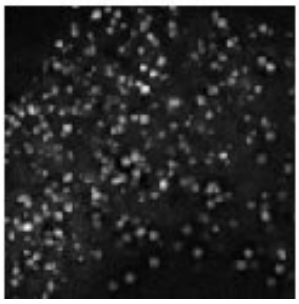
CY3 Spots SNR= 17.59db



CY3 Spots after filtering SNR=21.49db



FITC Spots SNR= 20.93db



FITC Spots after filtering SNR= 23.31db

Figure 2.3: Denoising by Substraction on real data data

the mean intensity $\tilde{I}$ in a surrounding region around the spots.

The curvature of a given pre candidate ξ is defined as follows:

$$\kappa(\xi) = det(H(\xi))$$

$$\text{where } H(\xi)) = \bigtriangledown \bigtriangledown^t I(\xi)$$

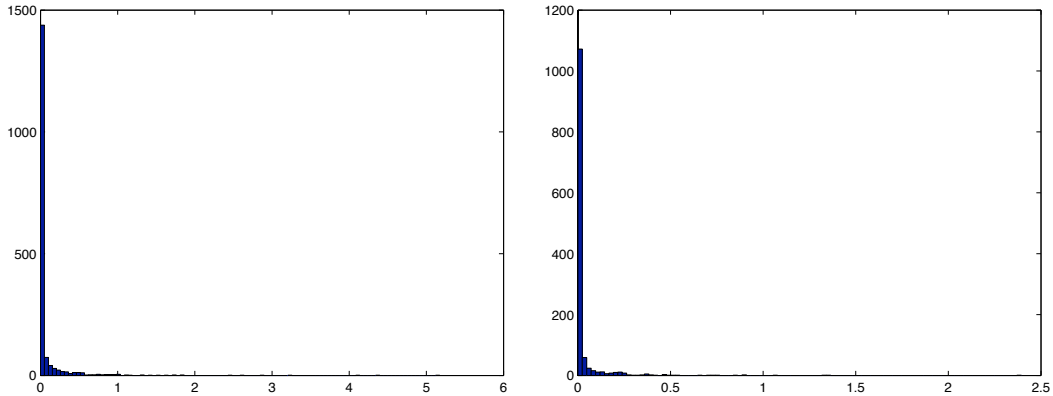
H is the Hessian matrix of the intensity $I(\xi)$.

The mean intensity is defined on a volume surrounding the spot which size is $5\sigma_\rho \times 5\sigma_\rho \times 5\sigma_z$.

Once these two values have been computed the score of a candidate is defined as :

$$s_\xi = \tilde{I}(\xi)\kappa(\xi)$$

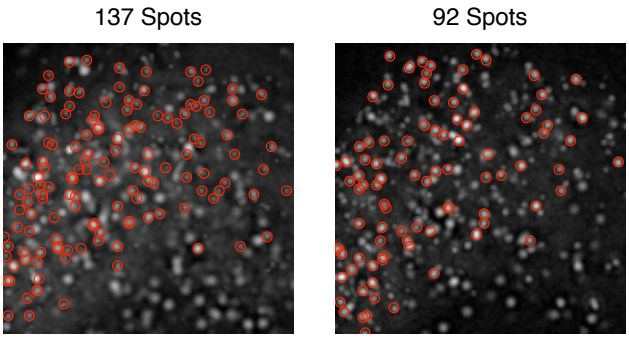
A cumulative histogram of the data $s_\xi \forall \xi \in L$ is then computed. In figure 2.4 an example of this simple statistical analysis is given. The result of the detection, after thresholding, is displayed in figure 2.5. The threshold is



CY3 image. FITC image

Figure 2.4: Histogram of the score

defined as the value where the slope of the histogram starts to flatten. Automatic thresholding has not been



CY3 image. FITC image

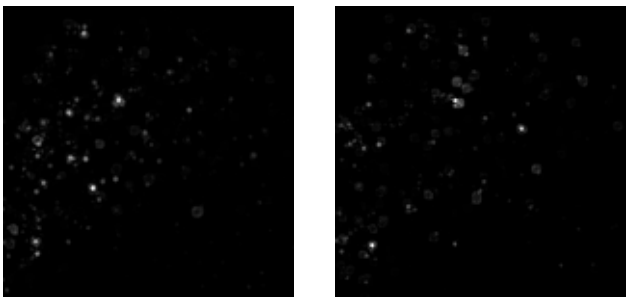
Figure 2.5: Detected spots using Thomann et al. (2002) method

implemented, because we had to face two main difficulties:

- The score does not seem to be very efficient as we can see on figure 2.6

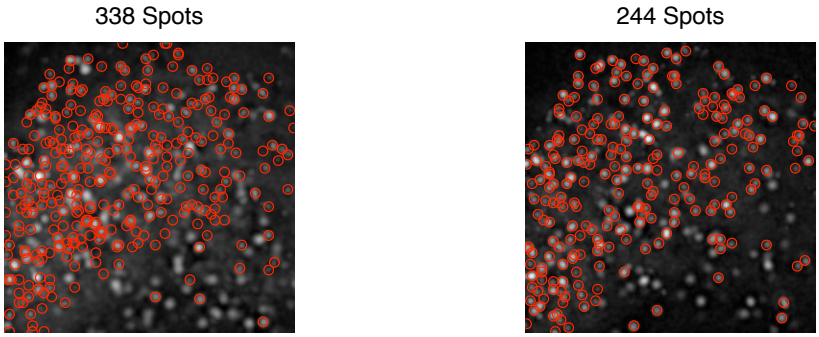
  We have displayed the maximum intensity projection of the score. A lot of spots remain in areas where the score is very low.

- Even with a low threshold many spots remain undetected: figure 2.7



CY3 image. FITC image.

Figure 2.6: Maximum intensity projection of the score



CY3 detected spots with a low threshold. FITC detected spots with a low threshold.

Figure 2.7: Detected spots with a low threshold: The spots located at the bottom right remain undetected.

2.1.2.2 Improvement of predetection based on eigenvalues

Because of the limitations of the method presented above we decided to use another score in order to improve the detection. We defined our new score by studying the properties of the Hessian matrix.

Hessian matrix properties Given a real valued function $f(x, y, z)$ the Hessian matrix is defined as follow:

$$H(f) = \begin{pmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial x \partial z} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} & \frac{\partial^2 f}{\partial y \partial z} \\ \frac{\partial^2 f}{\partial z \partial x} & \frac{\partial^2 f}{\partial z \partial y} & \frac{\partial^2 f}{\partial z^2} \end{pmatrix}$$

If all the eigenvalues $\lambda_1, \lambda_2, \lambda_3$ of H at a non-degenerate critical point ξ are negative, than the function atteins a local maximum.

A blolb-like structure will correspond to high negative negative values of the three eigenvalues [47]. Thus, we could define a boolean test on each predetected spots:

$$\text{if } \lambda_1 < 0 \text{ and } \lambda_2 < 0 \text{ and } \lambda_1 < 0 \text{ then } \xi \text{ is a "true spot"}$$

However, because of the low SNR of the images, a boolean test will not be discriminating enough to avoid the detection of noise.

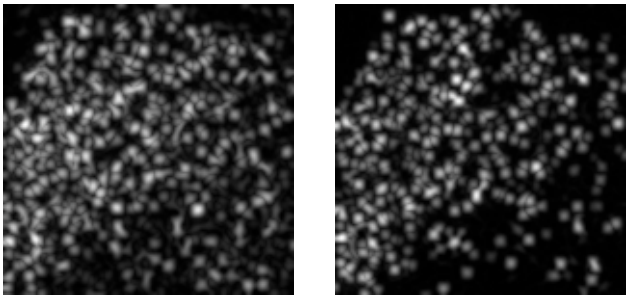
New score We defined our score as follow:

$$s_\xi = \tilde{I}(\xi)\kappa(\xi)$$

where,

$$\kappa(\xi) = -min\{\lambda_1, \lambda_2, \lambda_3\}$$

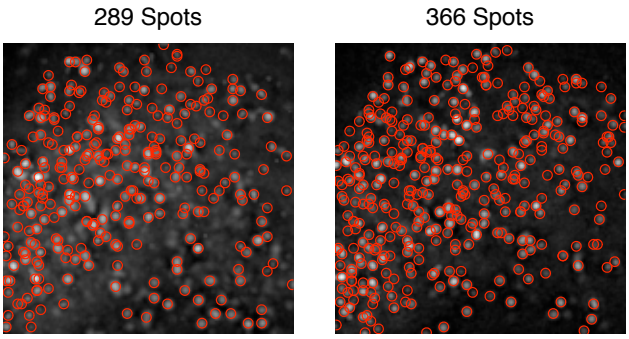
and $\tilde{I}(\xi)$ remains as defined above. The maximum intensity of this score is displayed in figure 2.8 The threshold



CY3 image. FITC image.

Figure 2.8: Maximum intensity projection of the new score

for the detection is determined manually as we did for the previous method. The detected spots are shown in figure 2.9 We decided to compute the empirical cumulative distribution functions of the score of our method



CY3 image. FITC image.

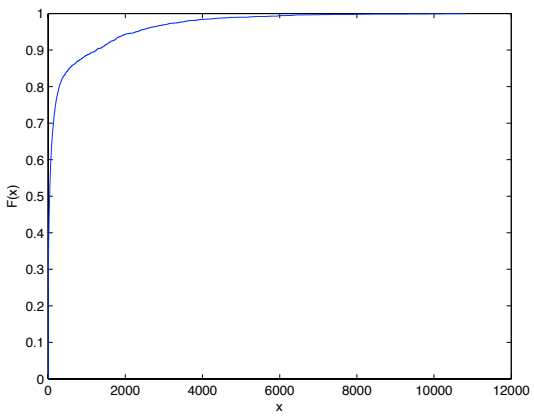
Figure 2.9: Detected spots using the new score

and of the score of the previous technique (figure 2.10. While the latter increases faster than the first one. This property is very interesting because it allows us to set our threshold in a very easy way. As a matter of fact we will need to vary the criterion of our detection to optimize the computation of the distances.

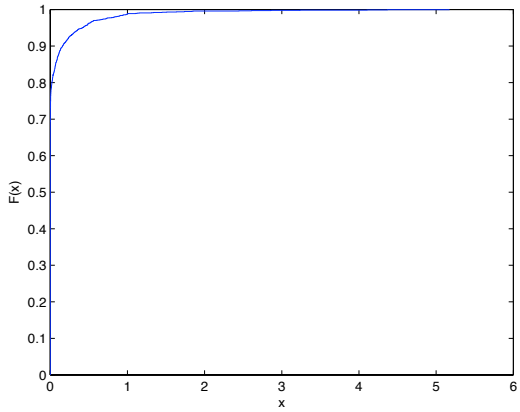
Comparison of the two methods We ran a few tests on synthetic data to compare the quality of the two methods. We generated two random sets of 100 and 300 CY3 spots (figure 2.11). We checked that the SNR of the new image was in the same rage of our data ($\approx$ 20 dB) by adding a white Gaussian noise.

We have fixed the number of detected spots and checked the number of true positives and false positives. These results are displayed in table 2.12 and table 2.13.

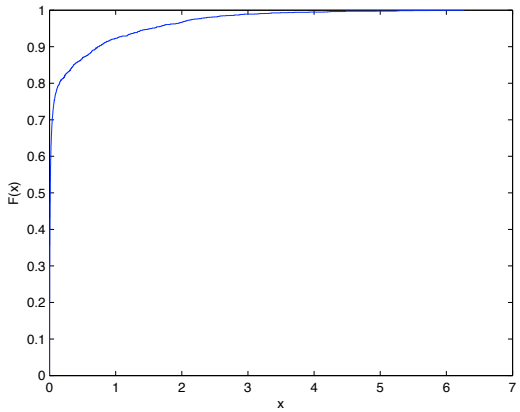
Figure 2.14 displays the detected synthetic spots.



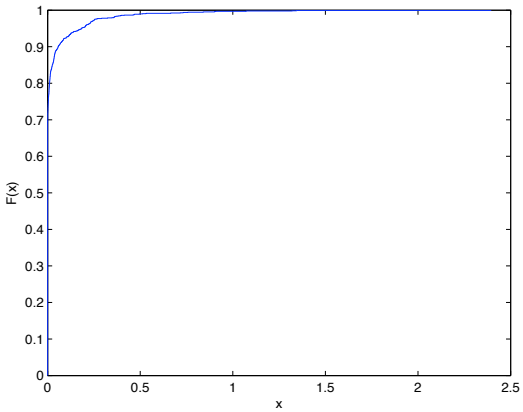
ECDF of our score on CY3 image.



ECDF of the first method's score on CY3 image.

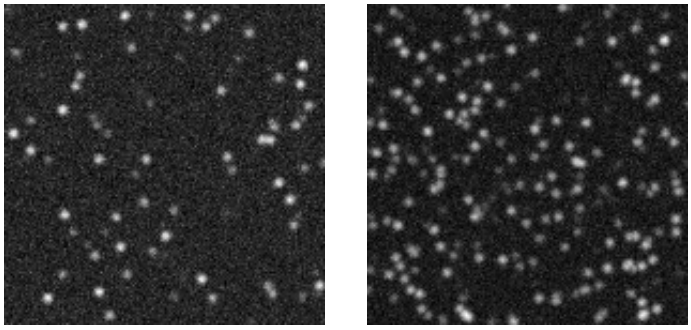


ECDF of our score on FITC image.



ECDF of the first method's score on FITC image.

Figure 2.10: Empirical cumulative functions



a) b)

Figure 2.11: Maximum Intensity projection of synthetic images:a) 100 CY3 spots, b) 300 CY3 spots

N of detected Spots	300		250		240		200	
Score	Our	Thom.	Our	Thom.	Our	Thom.	Our	Thom.
True Positive	250	232	246	227	238	226	200	200
False Positive	50	68	4	23	2	14	0	0

Figure 2.12: Detection performance of the two methods on a 300 CY3 spots synthetic image

2.1.3 Subpixel Localization

Because of physical limitations, the resolution of a fluorescence microscope does not allow to distinguish details that are smaller than $l \approx 250nm$ [136]. However, it is still possible to localize at a subpixel level the position of

N of detected Spots	100		80		70	
Score	Our	Thom.	Our	Thom.	Our	Thom
True Positive	87	78	80	77	70	70
False Positive	13	22	0	3	0	0

Figure 2.13: Detection performance of the two methods on a 100 CY3 spots synthetic image

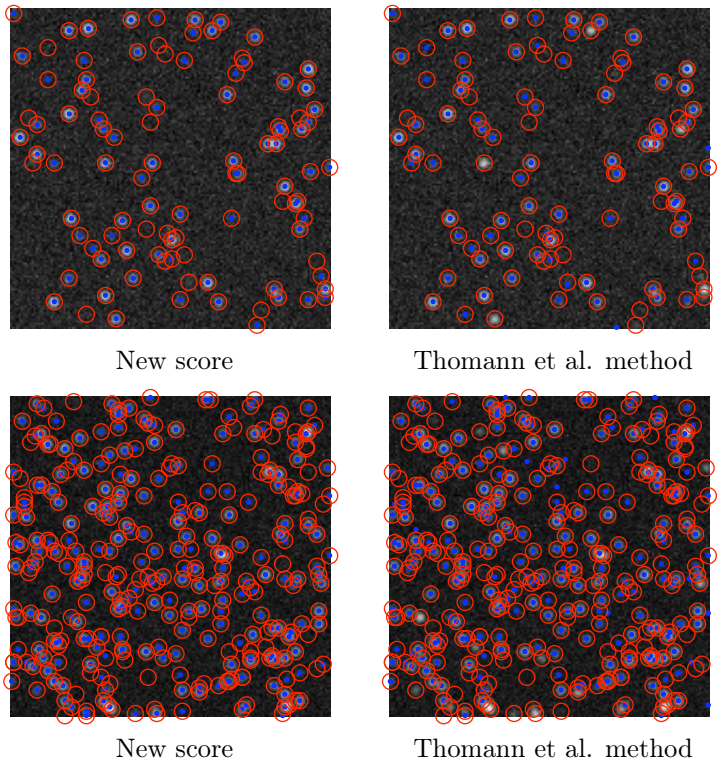


Figure 2.14: Detected synthetic spots. Top row 80 spots detected on a 100 spots image. Bottom row 240 spots detected on a 300 spots image. The true positions of the spots are displayed by the red circles while blue dots correspond to detected spots.

a given spot.

2.1.3.1 Curve Fitting

Each spot defines a sub volume Ω that we will try to fit on the theoretical gaussian approximation of the PSF g_{σ_ρ,σ_z}. Our model is defined as:

$$F_\chi(x,y,z) = A\exp\left(-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma_\rho^2} - \frac{(z-\mu_z)^2}{2\sigma_z^2}\right) + b$$

where:

$$\chi = \{\mu_x, \mu_y, \mu_z, \sigma_\rho, \sigma_z, A, b\}$$

We have computed a least-squares fitting so that we are looking for the coefficients χ that best fit the equation:

$$\min_\chi \parallel F_\chi - \Omega \parallel_2^2 = \min_\chi \sum_i \sum_j \sum_k \mid F_\chi(x_i,y_j,z_k) - \Omega(x_i,y_j,z_k) \mid^2$$

We have implemented this process in MATLAB using the function *lsqcurvefit* which compute a non linear curve

fitting in a least-squares sense.

We set the starting values as follow:

- $X_0 = \mu_{x_0}, \mu_{y_0}, \mu_{z_0} =$ center of mass of Ω
- b_0 is the background so we assume that after filtering $b_0 = 0$
- A_0 is set as the maximum intensity value in the image.
- $\sigma_{\rho_0} = \sigma_\rho^\star$
- $\sigma_{z_0} = \sigma_z^\star$

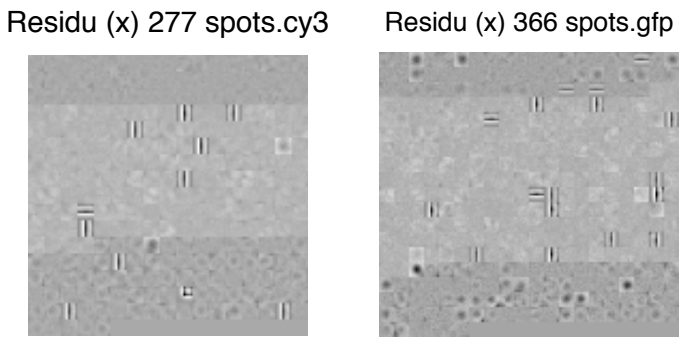
The model defined above is applied only to spots which subvolume Ω remains perfectly inside the image. For spots which are detected at the extremity of the global volume (located too high or too low) we implement a 2D curve fitting. Whereas the localization along the z axis will not be done, as we assume that our kernerl is separable, this method is still valid.

2.1.3.2 Control of the fitting

We have controlled the quality of the fitting by checking the residual of process. These results are show in figure 2.15. Moreover, the matlab function we have implemented allows us to visualize the shape of the spots (figure 2.16 andfigure 2.17). The spots are classified upon their localization in the stack. Therefore, the first ones correspond to spots detected at the top of the volume when the last ones correspond to those detected at the lowest level.

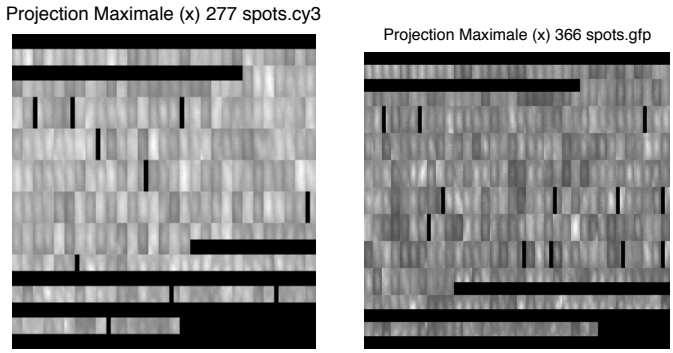
As expected, the residual of the spots located at the extremity of the stacks are worst than the ones of those located deeper in the stack. However it does not mean that the localization is not good. These spots could be discarded for the rest of the analysis.

In order to verify that our localization does not suffer any pixellique artefacts we made another test to assure that the fitting was correct. We checked that the decimal parts of the new positions were correctly spread between 0 and 1 (figure 2.18)



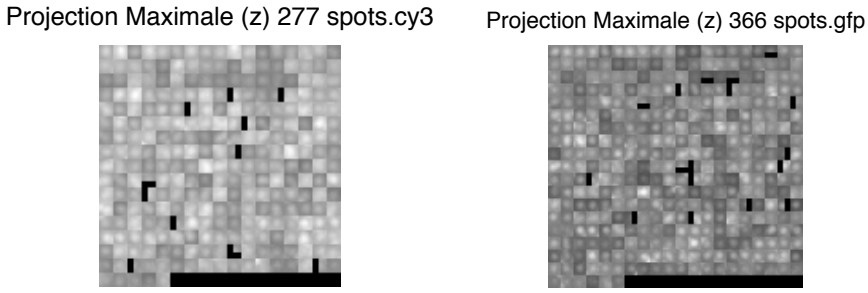
CY3 spots. FITC spots

Figure 2.15: Residual of the least-squares gaussian fitting



CY3 spots. FITC spots

Figure 2.16: Maximum intensity projection of the detected Spots along the x axis. The presence of black stripes in the images arise from the fact that some spots are not completely defined in the volume



CY3 spots. FITC spots

Figure 2.17: Maximum intensity projection of the detected Spots along the z axis.The presence of black stripes in the images arise from the fact that some spots are not completely defined in the volume

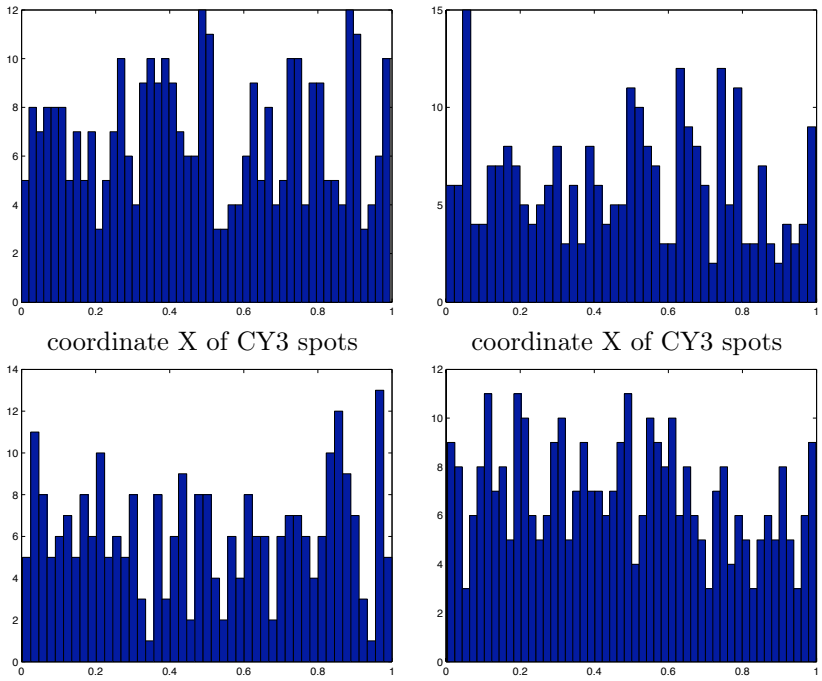


Figure 2.18: Histograms of the decimal part of the sub pixellic localization

2.2 Quantification of the transcriptional activity

CORRESPONDENCE

FISH-quant: automatic counting of transcripts in 3D FISH images

To the Editor: Transcription is inherently stochastic even in clonal cell populations[1]. Studies at the single-cell, single-molecule level enable a quantitative understanding of the underlying regulatory mechanisms[2,3]. A widely used technique is single-molecule RNA fluorescence *in situ* hybridization (FISH), in which fluorescent probes target the mRNA of interest, and individual molecules appear as bright, diffraction-limited spots (**Fig. 1a**)[3]. Recent experimental progress has made FISH easy to use[4], but a dedicated image analysis tool is currently lacking. Available methods allow counting of isolated mature mRNAs but cannot reliably quantify the dense mRNA aggregates at transcription sites in three dimensions, particularly those of highly transcribed genes. We developed FISH-quant to close this gap (**Supplementary Note 1**).

FISH-quant first detects and then localizes mature mRNA in three dimensions by fitting Gaussians to fluorescent spots[5]; each mRNA then undergoes a quality test based on the measured point-spread function (PSF) (**Fig. 1a** and **Supplementary Note 2**). This provides the three-dimensional (3D) position of mRNAs inside cells and the distribution of mRNA counts across the cell population (**Fig. 1a**). We validated this method in simulations and experimentally by dual-color FISH against mRNA of *RPB1*, which encodes the largest subunit of RNA polymerase II, and we obtained excellent agreement in the number of estimated spots and a high degree of colocalization (**Fig. 1b–d**, **Supplementary Note 3** and **Supplementary Methods**).

Counting nascent mRNAs is more challenging because their high local density prevents the resolution of individual molecules. FISH-quant offers two solutions: (i) comparison of the integrated intensity of the transcription site to that of mature mRNA and (ii) reconstruction of the transcription site signal by iterative superposition of weighted PSFs[5] (**Supplementary Note 4**). Whereas simple methods such as comparing maximum intensities[6] ignore the 3D extent of the transcription site, FISH-quant explicitly accounts for it. Also, method (ii) uses the experimentally measured PSF and restricts the intensity of the weighted PSFs to the range previously determined for mature transcripts, thereby taking into consideration aberration effects and variable labeling efficiency. Note that FISH-quant estimates an equivalent number of fully elongated transcripts; different positioning of the FISH probes on the mRNA affects the results and can be used to infer important

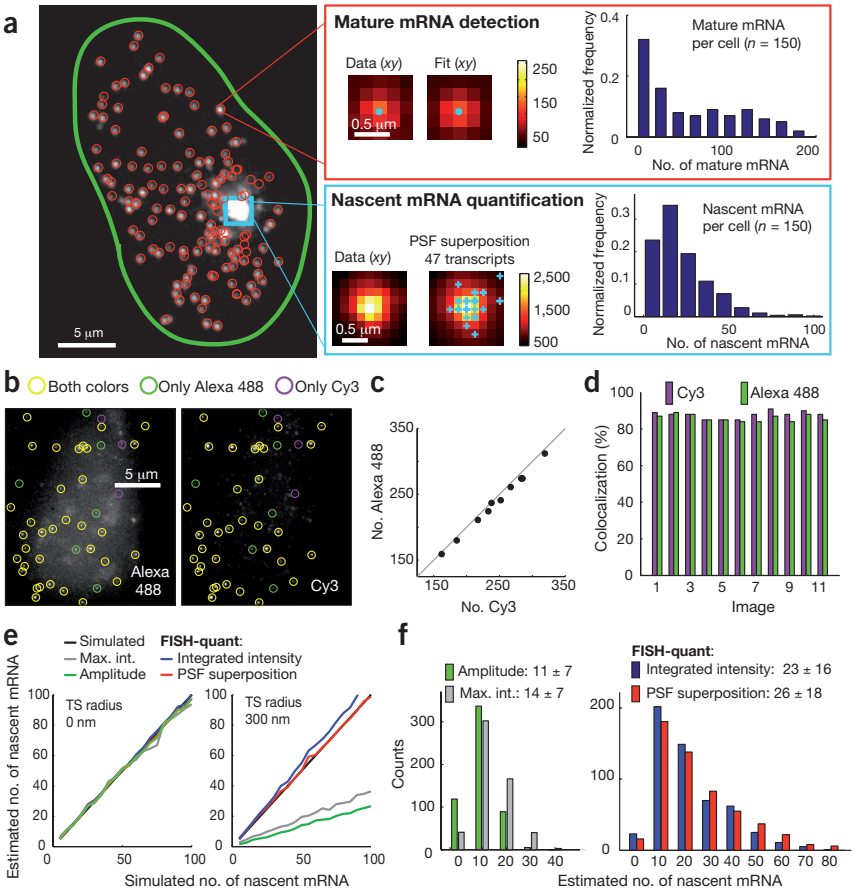


Figure 1 | Counting mature and nascent mRNA in FISH-quant. (**a**) FISH image of Hygro-MS2-x96-bGH reporter (**Supplementary Note 6**). Shown is a cell (green outline) with detected mRNA (red circles) and transcription site (TS; blue box). Red box indicates individual mRNA ("Data") fit with a 3D Gaussian ("Fit"), yielding localization (blue dot) with subpixel accuracy; blue box indicates the TS with transcripts (plus signs) quantified using PSF superposition. Histograms show distributions of nascent and mature mRNA per cell, obtained by batch processing of 150 cells. (**b**) Validation of mature mRNA detection by dual-color FISH against RPB1. Detection with 15 Alexa 488 probes (left) and 15 Cy3 probes (right). (**c**) Spot counts from two colors in 11 analyzed images. (**d**) Colocalization in each image. (**e**) Validation of TS quantification in simulations: resolution-limited TS (left) and spatially extended TS (right). Max. int., maximum intensity. (**f**) TS quantification of Hygro-MS2x96-bGH with simple methods (left) and FISH-quant (right). Values listed are mean ± s.d. of nascent mRNA counts.

properties of transcription[2,6] (**Supplementary Note 4**). To validate these methods, we first used simulations (**Supplementary Note 5**). For transcription sites smaller than the optical resolution, all methods yielded accurate estimates. However, for larger transcription sites, the simple methods led to gross underestimates, whereas FISH-quant gave accurate results (**Fig. 1e**). For elongated transcription sites, only the PSF superposition approach worked reliably. For experimental validation, we used an artificial reporter with transcription sites frequently exceeding the resolution. An RNase protection assay provided a rough, but independent, estimate of the ratio of mature versus nascent mRNA. The assay yielded ratios in the same range as the FISH-based quantifications, confirming their general validity (**Supplementary Note 6** and **Supplementary Methods**). For a more accurate assessment of simple methods and FISH-quant, we then compared the nascent transcript counts. Much as for the large simulated transcription sites, the simple methods led to underestimated counts (**Fig. 1f**). Thus, FISH-quant accurately quantified nascent mRNA even when simple approaches did not. Finally, we used FISH-quant to analyze β-actin mRNA after serum induction and measured more than twice the amount of nascent mRNA than we did with simple methods, which illustrates the importance of accurate quantification even for endogenous genes (**Supplementary Note 6** and **Supplementary Methods**). FISH-quant could also be applied to other structures with a dense accumulation of mRNA, such as processing (P)-bodies or stress granules.

FISH-quant is controlled via graphical user interfaces in Matlab and requires no computational expertise. A batch mode allows users to automatically process multiple images. FISH-quant is available at http://code.google.com/p/fish-quant/ with a detailed manual and test data.

Note: Supplementary information is available at http://dx.doi.org/10.1038/nmeth.2406.

ACKNOWLEDGMENTS
This research was supported by the Agence Nationale de la Recherche (ANR-10-BLAN-1222, ANR-09-PIRI-0024, ANR-10-INTB-1401), Sidaction, the Région Ile-de-France under C'Nano IdF (the Center of Competences in NanoSciences for the Paris region), the Fondation pour la Recherche Médicale en France (FRM) and the Institut Pasteur. We thank J. McNally for editing the manuscript.

COMPETING FINANCIAL INTERESTS
The authors declare no competing financial interests.

Florian Mueller[1,2], Adrien Senecal[2], Katjana Tantale[3], Hervé Marie-Nelly[1], Nathalie Ly[2], Olivier Collin[2], Eugenia Basyuk[3], Edouard Bertrand[3], Xavier Darzacq[2] & Christophe Zimmer[1]

[1]Institut Pasteur, Unité Imagerie et Modélisation, Centre National de la Recherche Scientifique (CNRS) Unité de Recherche Associée 2582, Paris, France. [2]Imagerie de la machinerie transcriptionnelle, Institut de Biologie de l'Ecole Normale Supérieure, CNRS Unité Mixte de Recherche 8197, Paris, France. [3]Institut de Génétique Moléculaire de Montpellier, CNRS, Unité Mixte de Recherche 5535, Montpellier, France.
e-mail: edouard.bertrand@igmm.cnrs.fr, darzacq@ens.fr or czimmer@pasteur.fr

1. Li, G.-W. & Xie, X.S. *Nature* **475**, 308–315 (2011).
2. Zenklusen, D., Larson, D.R. & Singer, R.H. *Nat. Struct. Mol. Biol.* **15**, 1263–1271 (2008).
3. Itzkovitz, S. & Van Oudenaarden, A. *Nat. Methods* **8**, S12–S19 (2011).
4. Raj, A., Van den Bogaard, P., Rifkin, S.A., Van Oudenaarden, A. & Tyagi, S. *Nat. Methods* **5**, 877–879 (2008).
5. Thomann, D., Rines, D.R., Sorger, P.K. & Danuser, G. *J. Microsc.* **208**, 49–64 (2002).
6. Boireau, S. *et al. J. Cell Biol.* **179**, 291–304 (2007).

Protein instability following transport or storage on dry ice

To the Editor: It is common practice to place protein solutions on dry ice for storage or transport, but this may lead to an unrecognized problem. A series of assay failures after short-term storage of antibody solutions on dry ice led to our observation that the pH of the thawed solutions was between 5.5 and 6.0 even though they had been formulated at pH 7.2. We hypothesized that exposure of the solutions to CO_2 caused the formation of carbonic acid, resulting in protein damage from the pH drop. Protein properties affected by pH include tertiary and quaternary structure, enzymatic rate constants, solubility, tendency to aggregate, susceptibility to chemical degradation and propensity to adsorb to surfaces[1]. We therefore examined possible interactions between dry ice and sealed frozen protein solutions.

We evaluated four types of cryogenic vials, three types of conical tubes, two types of glass vials and one type of microtube (**Supplementary Methods**). Vessels containing a buffered pH indicator solution were placed on dry ice or into a −70 °C freezer for 48 h. Upon thawing, most samples placed on dry ice experienced a substantial decrease in pH (**Supplementary Table 1**), and no container closure system consistently prevented acidification. pH changes were not observed in −70 °C freezer controls.

Sample acidification appears to result from two distinct events. First, CO_2 enters the container's headspace but is unreactive, having negligible solubility in ice. If we vented headspace before sample thawing, no acidification was observed (**Fig. 1a**). Also, placing samples into a −70 °C freezer for 96 h allowed the CO_2 to dissipate, after which no acidification was observed. The second event occurs if the sample is thawed while CO_2 is still in the headspace. Acidification was seen to originate at the liquid-gas interface and expand through the sample as it warmed (**Fig. 1b** and **Supplementary Video 1**).

We calculated pH as a function of headspace CO_2 for 1.5-ml microtubes containing Tris buffer (**Fig. 1c**). Predicted drops in pH ranged from 1.5 to 2.7 pH units depending on the starting pH and sample volume. Calculations for other buffer systems such as phosphate-buffered saline produced similar results.

Proteins generally exhibit low solubility near their isoelectric point (pI). Therefore, acidic proteins have an increased tendency to aggregate or precipitate as pH falls below physiological levels. As a model, we formulated β-lactoglobulin (pI 5.2) in citrate or phosphate buffers between pH 4.8 and 7.3. Aggregation index measurements (**Supplementary Methods**) were not substantially different at 2, 5 and 24 h post-formulation for solutions between pH 5.8 and 7.3, but aggregation index values increased with decreasing pH and increasing time below pH 5.8 (**Fig. 1d**). Stressing the samples by vortexing caused a marked increase in aggregation index values below, but not at or above pH 5.8.

We further examined the acidic protein carbonic anhydrase (pI 5.9) and the basic protein lysozyme (pI 9.3), each in 10 mM Tris, pH 7.3, or 10 mM phosphate, pH 7.3. Exposure to dry ice for 48 h before thawing resulted in substantial acidification of all solutions. Aggregation index values increased substantially for the acidic but not the basic proteins (**Supplementary Table 2**). Returning the samples to a −70 °C freezer for 96 h before thawing prevented the pH drop and increase in aggregation index value.

2.3 Tridimensional organization of the X-chromosome and the *Xic*

2.3.1 The biological context of the work

Random X-inactivation observed in adult tissue is extremely stable and irreversible. Only reprogramming of the cells by over-expression of pluripotency factors can reverse the inactive state. However, the comparison of X-chromosome inactivation stability in different cell lines derived from the three blastocyst lineages has revealed an important instability or plasticity of the inactive state, specifically in trophoblast stem cells (TS cells). In these cells characterized by an imprinted XCI, it has been shown that like the differentiated trophoblast giant cells (Corbel et al., 2013), the progenitors cells show an orchestrated cycle of reactivation and de novo inactivation of some paternal X-linked genes (Dubois et al., 2013). Thus, the reversal silencing of these genes, associated with a local loss of repressive histone marks, underlie a certain plasticity of TS cells concerning X-inactivation. Besides, this relaxed state is associated with a relocation of the genes outside the nuclear compartment, formed by the inactive X. The relaxation of XCI for some genes in particular is still unclear. Are these genes involved in a specific function in the cell and does this explain their relaxed regulation? Or conversely, does the "inefficient" X-inactivation relate to the absence of any need for precise regulation of these genes? Besides the functional aspect, we can also wonder whether the master region of the X inactivation center (*Xic*), responsible for X-inactivation maintanance, is implicated in this relaxation or not.
In order to analyze the implication of the *Xic* in the plasticity of imprinted XCI in TS cells we investigate the interplay between the three levels of genome regulation, which are: nuclear organization of this locus, gene expression, and when possible, chromatin composition within the 1Mb region at the single cell level.

2.3.2 Experimental strategies to analyze the *Xic*

In this study we use for as far as possible approaches which involve single cells and which allow us to evaluate the relationship between the three levels of genome regulation by taking into account TS stem cell population heterogeneity. Besides, these approaches allow us to tell apart the states of the two loci: the maternal active X-chromosome (Xa) versus the paternal inactive X-chromosome (Xi). In the following section, we introduce only the nuclear organization that is part of the project.

2.3.2.1 Analysis of the *Xic* tridimensional topology

For this analysis we use an imaging approach where we apply a tridimensional FISH approach to distance measurements in order to characterize and compare the topology of the *Xic* on both active and inactive X chromosomes in fixed cells. Our approach is summarized in figure 2.19. After cell fixation, a first step of probe hybridization (4 probes) in RNA-FISH is realized which allows us first to distinguish the active from the inactive X chromosome (the *Xist* cloud covers specifically the inactive X), and to determine the transcriptional status of the probe covered regions. Following RNA-FISH, hybridization in DNA-FISH of the same 4 probes is realized on the same cells in order to determine the tridimensional structure of the *Xics* chromatin fiber. This latter is obtained by measurements of the three-dimensional distances between the different parts of the *Xic* that are covered by the used probes on both active and inactive X-chromosomes. We take care that we maintain the nuclear integrity during this experimental procedure. The probes we use are around 40kb in length and regularly spaced from one another (the spacing is from 1 to 8 kb), covering the major part of the *Xic* as defined

in the literature.

For every measured distance we analyze at least 150 nuclei. After data treatment we obtain a minimum of 100 measures per distance and per locus. Besides, every combination of 4 probes gives us 6 measures between the probe pairs on both active and inactive chromosomes. These measurements are of course corrected by the chromatic aberration values inherent to the microscope opticals and are estimated before the image capture.

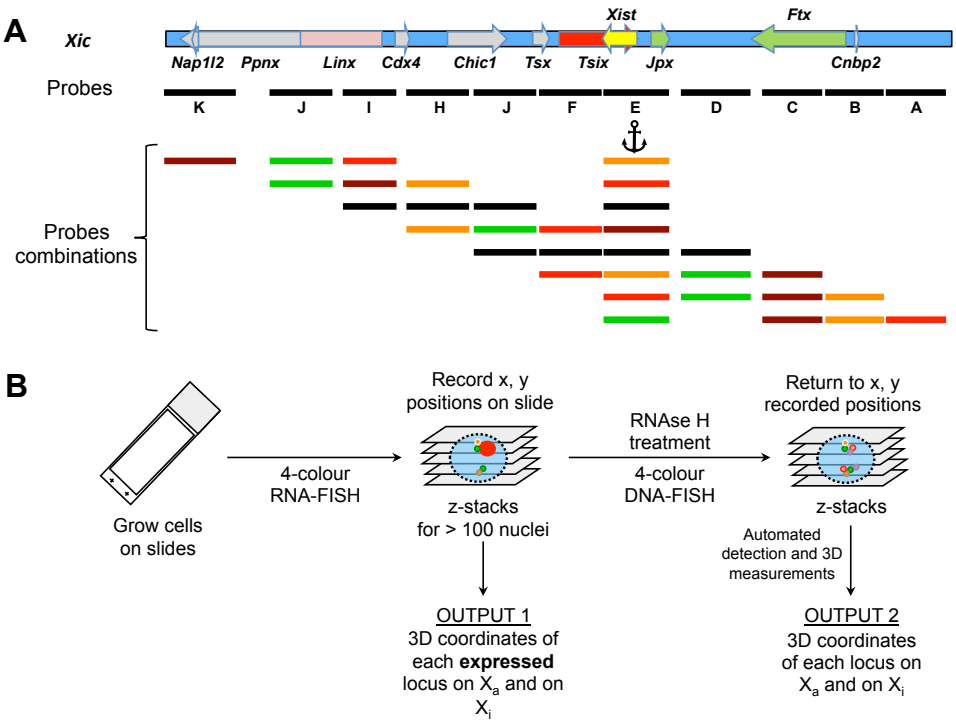


Figure 2.19: Experimental strategy used to analyze the three-dimensional topology of the X-inactivation center (DNA-FISH) and its correlation with its transcriptional status (RNA-FISH). A. Scheme representing the name and position along the *Xic* of the 11 probes hybridized in different combinations on cultured cells. B. Scheme representing the experimental strategy used for probe hybridization in RNA-FISH followed by a second hybridization in DNA-FISH. For every type of hybridization the output is specified.

This approach presents a particular specificity because it allows us to correlate on the same nuclei the tridimensional state of the chromatin fiber with its transcriptionnal activity. However, because of the experimental limitations the sequential hybridization of the probes along the *Xic* cannot be realized on the same cells. This may lead to biases due to intrinsic biological variability (different cells) and experimental manipulations (FISH experiments and image captures).

The analysis of the measured distances can be undertaken at several levels and the data can be considered from different points of view. The measured distances can first be observed by pairs considering their consecutive positioning along the *Xic* fiber or considering a fixed or anchor point. Secondly, the measured distances can be observed between three probes at the same time, allowing the identification of a loop formation. The third point of view consists of considering the probes four by four, which allows the addition of the three dimensional component.

In the following section we present some of the data we obtained regarding the first and second level of analysis.

Consecutive probe analysis: The measured distances along the *Xic* show almost no significant difference between the Xa and the Xi. The boxplot distribution shown below (figure 2.20) shows the

measured distances along the *Xic* in nm between consecutive probes. The measured distances in the 5' end of the *Xic* (including *Nap1l2*, *Ppnx*, *Linx* and *Cdx4* genes) are more important compared to the measured distances in the 3' end of this locus (from *Chic1* to *Ftx* genes). These differences suggest that the 5' end of the *Xic* is more relaxed compared to the 3' end that seems to be more condensed.

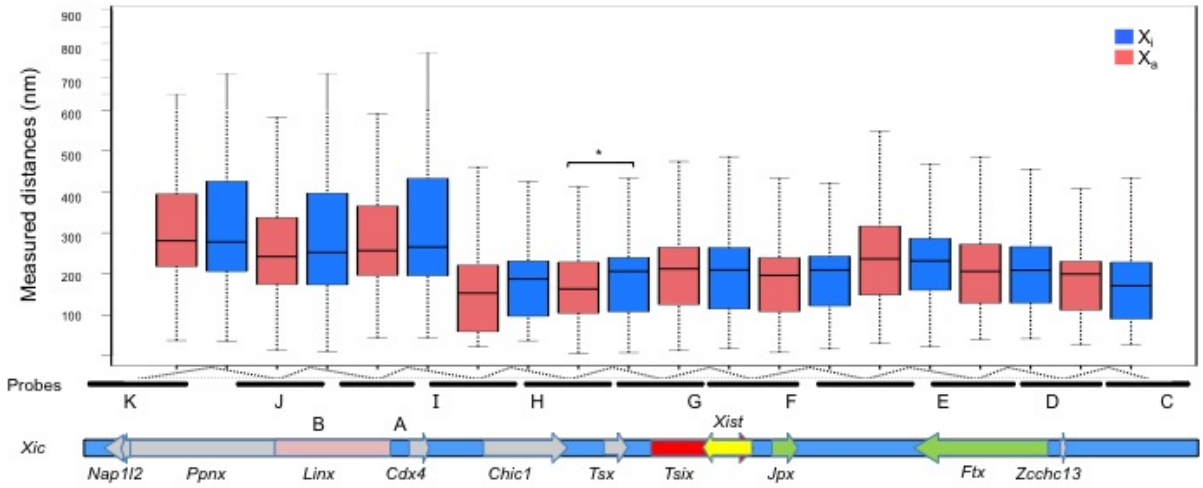


Figure 2.20: Distribution of the measured distances between consecutive probes along the *Xic* on the active X (Xa in red) and inactive X (Xi in blue). The only significant difference observed between the Xi and the Xa concerns the probe pairs G and F (*KS test P-value=0.008).

The statistical analysis of the different distributions show almost no difference in the measured distances between the active X-chromosome (Xa) and the inactive X-chromosome (Xi). The only statistical difference concerns the measured distances between the probe pair G and F, covering respectively *Chic1* / *Tsx* and *Tsix* genes (KS tets P-value=0.008). In fact it seems that the genes *Chic1* and *Tsix* are more distant on the Xi than on the Xa. Besides, these genes present high levels of active histone marks on the Xa compared to the Xi (not shown), suggesting that the two genes are transcriptionally active, especially the Xa. This specific tridimensional proximity of the two genes might be related to their positive transcriptional activity, especially on the Xa. Whether this spatial proximity is triggered by transcription or not, is however unknown.

The consecutive probe analysis: The measured distances along the *Xic* show an important heterogeneity. A reflection of a dynamic topology? The violin plots representation of the measured distances (figure 2.21) allow us to determine the existence of different subpopulations of measures within the global population. In fact the diagrams show that for certain pairs of probes the distribution is heterogeneous (several subpopulations with an undefined delimitation between one another) whereas for some other distributions several delimited subpopulations are clearly observed (three major subpopulations).

These different configurations of the measured distance distributions on both Xa and Xi suggest an important dynamic of the *Xic* fiber. In fact the hetergenous distributions suggest a discrete dynamic of the fiber covered by the analyzed probes, whereas the distinct subpopulation distributions suggest that the chromatin fiber at these loci is present in the cell population in three states.

However, we cannot exclude a highly dynamic passage from one state to another which cannot be captured by experimental approaches that use fixed cells such as the DNA-FISH.

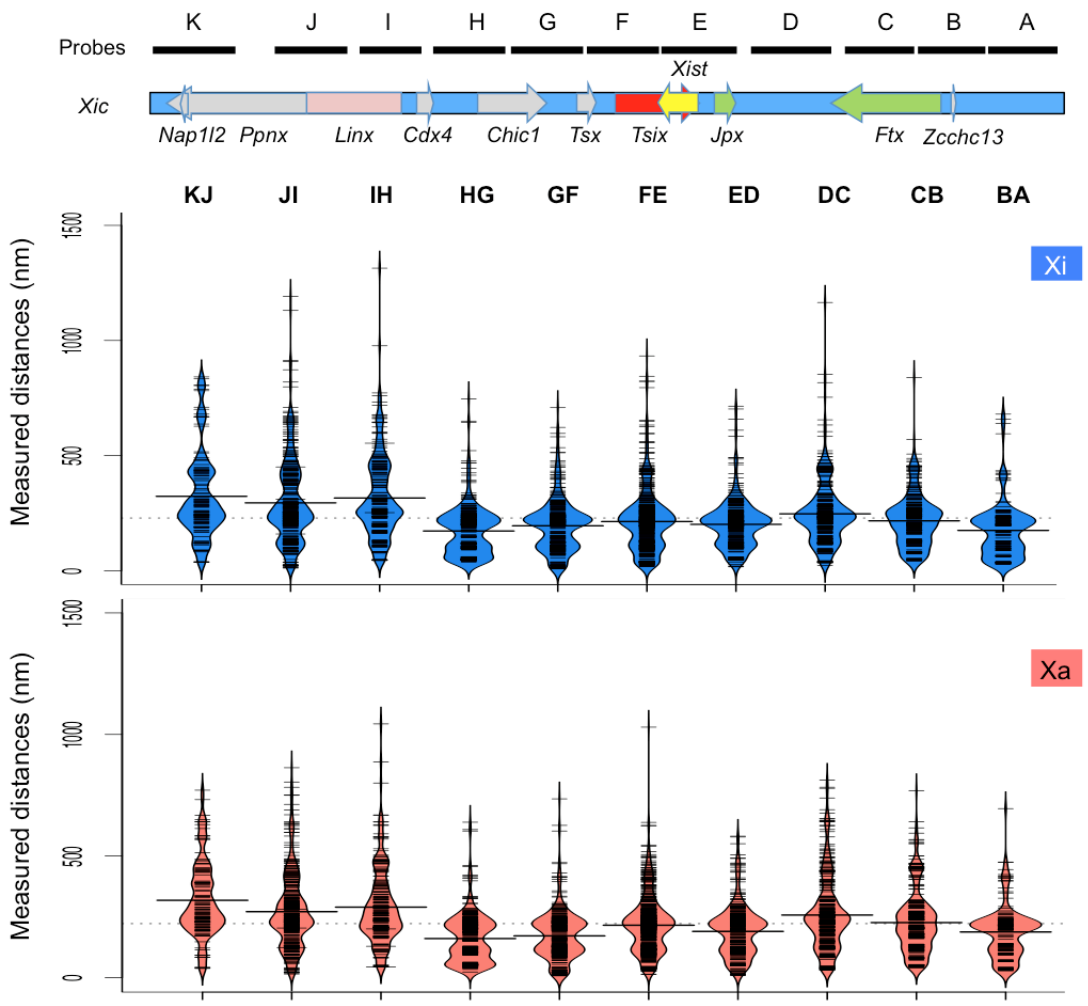


Figure 2.21: Violin plots of the measured distances between pairs of consecutive probes along the *Xic* on the active X (Xa in red) and inactive X (Xi in blue).

A method to model the dynamic topology of the *Xic* fiber Because of the experimental limitations of the DNA-FISH approach all the used probes cannot be hybridized on the same cells at the same time. In fact for every combination of 4 probes we obtain a 4x4 matrix of measured euclidean distances, for every locus (Xa and Xi) in every cell. So in order to model the whole *Xic* fiber (covered by the 11 probes) and to take into account particularly the cell population heterogeneity, we would like to use these information to give metric scale to a stochastic bayesian procedure based on HiC data.

The probes centered analysis: The *Xic* seems to be divided in two parts The analysis of the measured distances between a probe within the *Xic* and the anchor probe (covering the *Xist* gene) shows two major behaviors within this region. The measured distances between the probes located at the 3' end of the *Xic* and the probe covering *Xist* show a median value, equal or less than 200 nm, whereas the measured distances between the probes located at the 5' end and *Xist*, present for the majority a median value higher than 200 nm (figure 2.22).

This suggests that the 5' of the *Xic* is significantly more distant from the *Xist* gene compared to the 3' end of this region, independently of the genomic distance. This partitioning of the *Xic* reminds us of a recent study

that has shown that the *Xic* is partitioned into small interacting domains or TAD (Topologically Associating Domains) (Nora et al., 2012[99]).

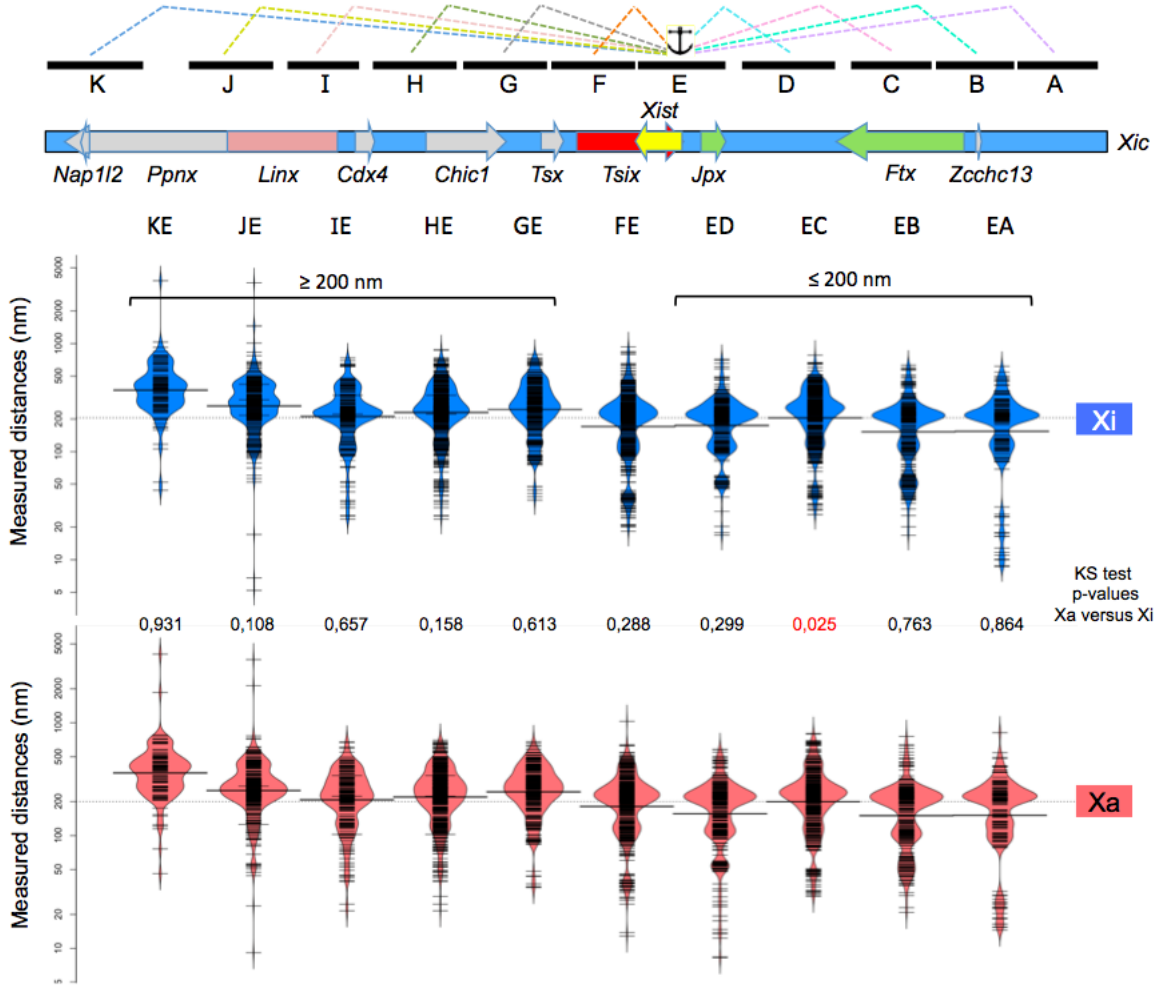


Figure 2.22: Violin plots of the measured distances between the different probes covering the *Xic* and the *Xist* gene on the active X (Xa in red) and inactive X (Xi in blue). The *Xic* can be divided into two parts. One showing median distances equal or less than 200 nm and the other showing a median distances higher than 200 nm.

Loops formation within the *Xic* The analysis of the measured distances between three probes at the same time allows us to highlight the existence of loops within the *Xic*. The analysis of the 5' end of the *Xic* show that the probes J and H are close to each other physically in the nucleus and this is responsible for the extrusion of the region covered by probe I. This extrusion suggests a loop formation, which consists of the representation in a large cloud, suggesting that this loop has various sizes in the cell population (figure 2.23). Besides, the analysis of triplet KIE shows that the I probe is close to the E probe despite their important genomic distance on both Xa and Xi. In fact it seems that the region covered by probe I loops out to get into close contact with the region covered by probe E. This suggests that the promoter region of the *Linx* gene, "interacts" with the *Xist* gene. In addition, it has been shown recently, using a 5C approach that the *Linx* gene interacts with the *Xist* antagonist gene *Tsix* in mouse ES cells. In this study it has been suggested that *Linx* might be a cis-regulatory factor of *Tsix* (Nora et al., 2012[99]).

The functional relevance of this interaction is unknown. In TS stem cells, the *Linx* gene is not expressed,

whereas the*Xist* gene is expressed only from the Xi. Therefore, this interaction between the two genes cannot be related directly to X-chromosome inactivation as it is observed on both X chromosomes.

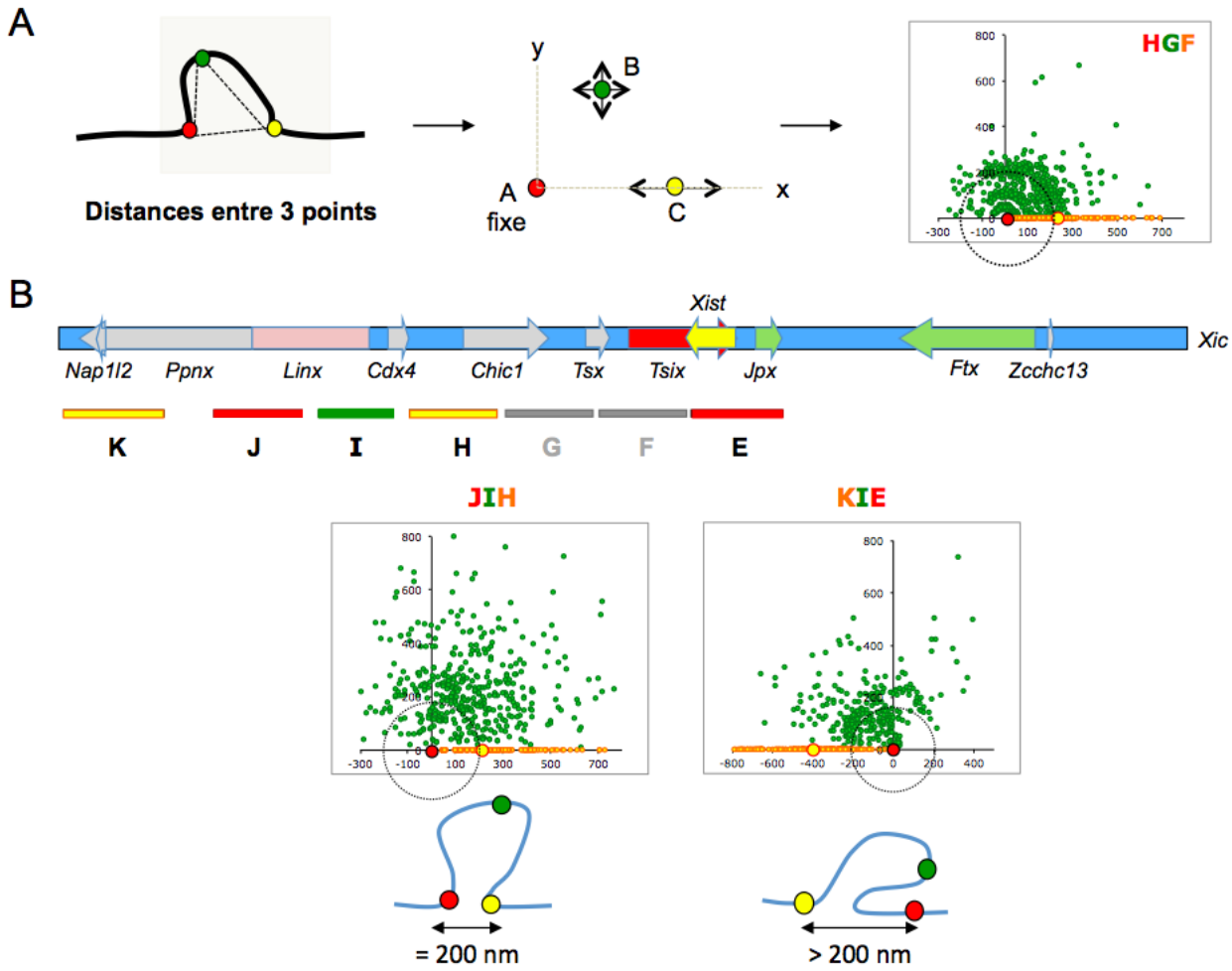


Figure 2.23: 2D-graphical representation of loops formation within the *Xic*.
A. Scheme showing the graphical 2D represention of the tridimensional organization between three probes represented in red, green and yellow. Probe A represents the reference point with its position fixed to 0. The position of probe B within the graph on the x and y axis corresponds to the measured distances between pair A-B and pair B-C. Probe C is represented only on the x axis according to the measured distances between C and A. The median distance between A and C is represented by the yellow spot.
B. The two graphs represent the behavior of a part of the 5' end of the *Xic* where the green spots represent probe I covering the promoter region of the *Linx* gene.

Chapter 3

Normalization of contact matrix

3.1 Introduction

In this chapter we will first give a brief overview of the way HiC data are analyzed before discussing some key points of the workflow. We will insist in particular on some of the aspects of the output of the available normalization procedures. The second part of this chapter consists of a published work in which we describe and analyze some of the main biases present in HiC based experiments.

3.1.1 Contact matrix generation

The first step in every data flow of HiC or 3C analysis experiment is the mapping of the raw reads on a reference genome. This operation is very critical for two reasons:

- Most of the modern aligners are not designed to process chimeric paired reads (paired reads whose mates do not belong to the same chromosome). For instance, Bowtie2 (Langmead and Salzberg, 2012 [76]) is very efficient to map paired-end reads. These sequences are mainly use to perform assembly and correspond to DNA sequences which are separated by a fixed distance d. Aligner software will use this knowledge to discard pairs and also to speed up their inner processes by decreasing the size of the mapping area. Bowtie 2 (Langmead and Salzberg, 2012 [76]) for instance, performs this operation by giving full latitude to the user in the choice of d. Surprisingly, we have noticed that even when we force the aligners not to use this prior knowledge, many reads are artificially located close to each other. This problem is very critical when structural variants of a given specie have to be analyzed.

- The second issue we would like to address concerns the repeated sequences. Not only do these areas pose a problem to every assembly algorithm available, but they also lead to artifacts and data loss in the mapping process. When a read of a given pair is mapped on a region which is present more than once in the genome, it automatically receives a penalty score that propagates to its partner. Therefore, pre and/or post processing of the data is necessary to take the repeated sequences into account.

The second critical operation that concerns the data flow is the estimation of the experimental biases.

3.1.2 Normalization techniques

At the time of writing, every published work on normalizing contact matrices belongs to one of the following categories:

- Probabilistic method. Yaffe and Tanay (2011 [149]) introduced a probabilistic framework to analyze some of the biases that affect HiC experiments. They consider three major streams of artifacts which are the local: GC content of a sequence, the length, and the mappability of a restriction fragment. While the framework of this analysis is very robust, the authors assume implicitly that no other biases may corrupt the HiC signal. This is a strong assumption which is desirable to be investigated.

- Ad-hoc methodology. Cournac et al. (2012 [27], presented in the second part of this chapter) as well as Imakaev et al. (2012, [63]) make no exhaustive assumption about the artifact streams. Instead they propose an iterative procedure that forces the rows and columns of the contact matrix to sum to one. In our view, despite the overall increase of contrast of the matrices after this process, these objects cannot be seen as pure normalized contact matrices and special care has to be given in their interpretation. We will give two graphical examples that illustrate our concerns.

Graphical signature of the iterative normalization We performed the iterative normalization procedure described in [63] and [27] on the standard test image of Lenna. The output of the process is displayed in figure 3.1:the normalization procedure seems to generate artifacts.

Figure 3.1: Iterative normalization procedure. A) The raw image. C) is the normalized image while B) is just a scaled version of A) so that B) and C) have the same global intensity. There are no details lost by the process but there are artifacts generated, as displayed in D), and the overall contrast of the image is lowered.

3.2 Published work

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

BMC Genomics

METHODOLOGY ARTICLE Open Access

Normalization of a chromosomal contact map

Axel Cournac[1], Hervé Marie-Nelly[2,3,4], Martial Marbouty[5,6], Romain Koszul[5,6*] and Julien Mozziconacci[1*]

Abstract

Background: Chromatin organization has been increasingly studied in relation with its important influence on DNA-related metabolic processes such as replication or regulation of gene expression. Since its original design ten years ago, capture of chromosome conformation (3C) has become an essential tool to investigate the overall conformation of chromosomes. It relies on the capture of long-range trans and cis interactions of chromosomal segments whose relative proportions in the final bank reflect their frequencies of interactions, hence their spatial proximity in a population of cells. The recent coupling of 3C with deep sequencing approaches now allows the generation of high resolution genome-wide chromosomal contact maps. Different protocols have been used to generate such maps in various organisms. This includes mammals, drosophila and yeast. The massive amount of raw data generated by the genomic 3C has to be carefully processed to alleviate the various biases and byproducts generated by the experiments. Our study aims at proposing a simple normalization procedure to minimize the influence of these unwanted but inevitable events on the final results.

Results: Careful analysis of the raw data generated previously for budding yeast *S. cerevisiae* led to the identification of three main biases affecting the final datasets, including a previously unknown bias resulting from the circularization of DNA molecules. We then developed a simple normalization procedure to process the data and allow the generation of a normalized, highly contrasted, chromosomal contact map for *S. cerevisiae*. The same method was then extended to the first human genome contact map. Using the normalized data, we revisited the preferential interactions originally described between subsets of discrete chromosomal features. Notably, the detection of preferential interactions between tRNA in yeast and CTCF, PolII binding sites in human can vary with the normalization procedure used.

Conclusions: We quantitatively reanalyzed the genomic 3C data obtained for *S. cerevisiae*, identified some of the biases inherent to the technique and proposed a simple normalization procedure to analyse them. Such an approach can be easily generalized for genomic 3C experiments in other organisms. More experiments and analysis will be necessary to reach optimal resolution and accuracies of the maps generated through these approaches. Working with cell population presenting highest levels of homogeneity will prove useful in this regards.

Background

Chromosomes from both eukaryotes and prokaryotes not only convey information through their linear DNA sequence but also contribute to the regulation of a number of DNA-related metabolic processes through their three dimensional arrangements [1-3]. Since an original publication by Dekker and co-workers ten years ago, chromosome conformation capture (3C) technique and its derivatives have become essential to the investigation of chromosome organization [4-6]; for a brief overview of the various techniques published so far see [7]. The general principles of these protocols remain the same and rely on formaldehyde fixation to capture long-range trans and cis chromosomal interactions in living cells. The crosslinked cells are incubated with a restriction enzyme that will cut the DNA in a number of restriction fragments (RFs). Because of the crosslink, several RFs can be covalently linked within molecular complexes. A ligation step in diluted conditions will favor ligation events between RFs trapped within the same complex. After a decrosslinking step, the resulting 3C template consists in a collection of ligation products of two specific RFs, whose relative abundance (after normalization) reflects the frequency with which these two chromatin segments were

*Correspondence: romain.koszul@pasteur.fr; mozziconacci@lptmc.jussieu.fr
[5] Institut Pasteur, Spatial regulation of genomes group, Department of Genomes and Genetics, F-75015 Paris, France
[1] LPTMC, UMR 7600, Tour 12-13/13-23, Boîte 121, 4, Place Jussieu, 75252 Paris Cedex 05, France
Full list of author information is available at the end of the article

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

crosslinked in the population. The exhaustive analysis of this collection enables the generation of chromosomal contact maps, that allows deciphering the average positioning of loci of interest with respects with each others within the nucleus. In the past few years, quantification of the abundance of ligation products has evolved from semi-quantitative PCR [4] to deep-sequencing techniques [8]. The later approach now enables genome-wide analysis of chromosome organization. A typical result of such experiment is the number of times each pair of RF is sequenced at the final step. These numbers are then arranged in a symmetric matrix representing all the possible pairs of RFs from the genome, generating a genome-wide contact map. Those matrices represent the relative frequency of physical interaction for each RF in the genome with all of the other RFs. Different experimental protocols have been used so far, and genome-wide contact maps have been obtained for Lymphoblastoid cells [8,9], mouse [10,11], *Schizosaccharomyces pombe* [12], *S. cerevisiae* [13,14], and fruit fly [15].

3C derived experiments are likely to generate biases given the complexity of the protocols, and necessitate a dedicated effort to experimentally identify and limit the generation of byproducts at each step [16]. However, it appears impossible to entirely prevent unwanted DNA molecules to be present in the final banks, and subsequently in the sequence data. Therefore, these data need to be carefully processed in order to identify these sequences, and limit the introduction of biases in the final analysis. Although not necessarily rewarding, such (re-)processing is essential not only to accurately analyze the data from a specific experiment but also to provide important feedback for the design of future experiments. For instance, GC content and RF lengths induced biases present in the Hi-C databank of the Human genome were recently identified [17]; see also [18]. Here, we have reassessed the genomic 3C data from the experimental protocol used to obtain the first comprehensive dataset in *S. cerevisiae* in a pioneering study published recently (Figure 1A; [13]). Using HindIII as 3C restriction enzyme, the interactions between 4454 sites along the 12 Mbp yeast genome were mapped and a symmetric matrix of 4454 rows per 4454 columns was generated. A number of interesting features, some of them expected, such as centromere clustering resulting from the Rabl configuration, and others less obvious, such as early replication origins clustering, were identified from this matrix [13]. Interestingly, the re-analysis of the raw data obtained through this protocol lead to the characterization of a number of events and biases unidentified before. Back-and-forth comparison between these biases and the protocol steps allowed us to identify the different sources for these events.

Having properly identified and quantified all these biases, we developed a normalization procedure which allows us to correct the data for all those biases at one time. Overall, and as expected from the original analysis, the conclusions drawn from the corrected maps do not differ significantly from the original publication. However, the corrected map gives a more contrasted view of chromosomal contacts, and present sharper features when it comes to preferential interactions between telomeres or chromosomal arms. It also ponders some of the conclusions drawn regarding clustering of specific genetics elements, which will be discussed. We then used this approach on the genomic 3C (Hi-C) human dataset obtained by Dekker and co-workers [8] and showed that proper normalisation is a prerequisite to assess relevant contacts. The methodology described here allows for an efficient and simple analysis of chromosomal contact-maps, and is potentially of great convenience to any team interested to use similar approach.

Results and discussion

Quantification of the ligation products

During the ligation step, one can envision to recover different types of products (Figure 1A, step 3). Firstly, a RF can simply be circularized on itself (step 3i), resulting in a loop. Secondly, two consecutive RF on the genome can be re-assembled together (step 3ii). This type of event will be designated as a religation event. Note that religation events are virtually indistinguishable from non-digested restriction site (RS) given the original sequence is then restored. A third type of product can be recovered at this step, especially if the digestion is partially incomplete which will always be the case: longer DNA fragments formed out of two continuous RFs can be circularized during the ligation step (step 3iii). Finally, two RFs that are not consecutive on the genome can be ligated together (step 3iv). These products are the nuggets the experiment is digging for, and will be termed here as long-range interactions. Long-range interactions can either be intra or inter-chromosomal. Although inter-chromosomal events are easily identified through mapping of the pair-end reads along the genome, intra-chromosomal events necessitate a more careful examination of the positions of the sequences. A convenient way to identify the type of an intra-chromosomal ligation product is to use the orientation of the sequences obtained from the pair-end sequencing run. Each RF exhibits two extremities. The one with the highest coordinate according to the yeast genome conventional representation is labeled "+" and the other one "-". Every ligation event therefore falls within one of these four categories: -/-, +/+, -/+ and +/- (see Additional file 1: Figure S1A). Whereas long range interactions should not happen with any preferential orientation of the fragment extremities, a circularized RF will always connect its – extremity with its + extremity (Additional file 1: Figure S1A). The distribution of interaction types

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

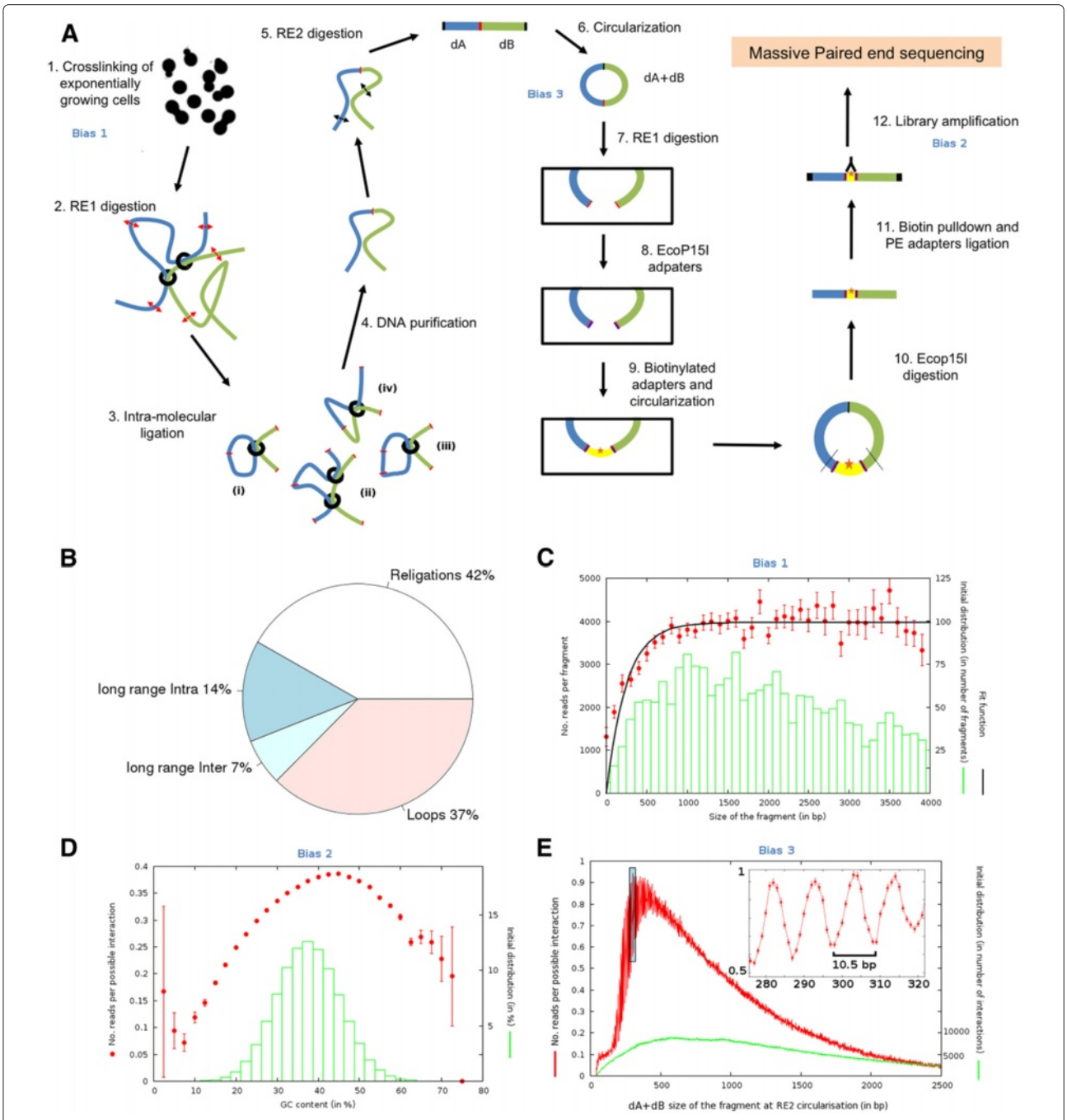


Figure 1 The different steps of the original genomic 3C experiment in yeast and their associated biases [13]. A) Experimental steps. 1: Yeast cells are fixed with formaldehyde. 2: the genome is digested using a 6 cutter restriction enzyme (RE1; red double-headed arrows). 3: extraction of protein/DNA complexes and ligation in diluted conditions that favor DNA-end interactions and religation within the same complex. During this process, some RF will simply circularize (i), while others will religate in their original orientation (ii). Religation products are also expected between non-collinear restriction fragments (iii), whereas collinear RF separated by one, or more, RF will also interact together (iv). 4: de-crosslinking and DNA purification. 5: digestion of DNA products using a frequent 4 cutter restriction enzyme (RE2; black double-headed arrows). 6: DNA is ligated in diluted conditions, favoring intra-molecular circularization of single DNA molecules. Remaining linear fragments are degraded. 7: DNA circles containing a RE1 site are re-opened using RE1. 8: short DNA sequences, containing EcoP15I recognition site and a biotinylated nucleotide are added at both ends of the linear fragments. 9: circularization of linear fragments. 10: EcoP15I digestion of the DNA segments 25 bp apart from the enzyme recognition site. 11: pull-down of the DNA fragments containing biotinylated nucleotides. 12: amplification of the DNA fragment isolated and sequencing. **B**) Pie-chart representation of the different types of events obtained at step 3: religations, long range intra, long range inter, loops (from 50 millions pair-end sequences analyzed from the HindIII-MspI condition A and B experiments). **C**) Quantification of the fragment length bias. **D**) Quantification of the GC bias. **E**) Quantification of the circularization length bias.

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

(+/+, -/-, -/+ and +/-) can be plotted for self-interacting fragments as well as for contiguous fragments (i.e. separated by only one RS), and then separated by two, and more RSs. For the later category no preferential orientations are distinguishable (Additional file 1: Figure S1B). A strong enrichment in +/- interactions is observed for pairs of collinear RFs. This enrichment is due to the presence of religation events (ii) as well as detection of sites which escaped the digestion step. The formation of type (iii) products is revealed by the fact that interactions between contiguous fragments on the genome are more often found in the -/+ configuration, which corresponds to a loop, than in a -/- or +/+ configuration. The relative number of those different products can be represented with a pie chart (Figure 1B). Loops and religation appear to be very frequent events (about 80% of the original data). Those inevitable byproducts were removed from all subsequent analysis. In addition, fragments with no restriction site for the secondary enzyme and therefore that should not be detected according to the experimental protocol were also discarded. Similarly, fragments whose extremities align ambiguously along the reference genome were removed as well (see Methods for details). In total, more than 80% of the initial raw reads were removed for subsequent analysis, which is consistent with other experiments in the field, and leaves room for a lot of improvement.

Identification of major biases in the experimental protocol

Complex protocols involving a large number of steps are likely to generate biases in the data that has to be careful sought for. What we call biases here is a variability which is larger than the expected noise and can be explained primarily by properties of the fragment itself. In the following, three major biases likely to affect the number of detected interactions between fragment pairs were identified: the length of RFs, GC content of the paired-end reads, and the length of DNA segments at the circularization of steps 6 and 9.

The distribution of the number of reads per fragment as a function of the fragment size L is presented on Figure 1C. Given the number of positions accessible to fixating agents along a RF increases with its size, one would expect the interaction probability to increase linearly with RF size. For RF under 800 bp, the number of reads per fragment increases, suggesting that indeed the probability for a cross-linking event to occur depends on the length of the fragment. However, for longer RFs, a plateau is reached, suggesting that the maximum probability for at least one cross-linking event to occur along that length is reached. In other words, the probability of longer fragments not to be cross-linked at least once is constant and very small (Methods).

Formaldehyde fixation, which is the first step of 3C based protocols, therefore introduces a length bias for sizes under 800bp. In this range, the longer a RF is, the more likely it will be cross-linked with other RF during the fixation step.

The distribution of reads per possible interaction between two RF extremities was plotted as a function of the GC content of these extremities (Figure 1D). From this figure one can see that extreme GC content extremities tend to be under represented in the final interaction reads. Therefore, the PCR reaction or/and the deep-sequencing steps can introduce additional biases, notably by favoring reads with a GC content of about 45%. The bias of GC content in short reads data from high-throughput DNA sequencing has indeed been reported (see Figure 2 in [19]). However, such biases do not appear to affect many interactions (see Figure 1D).

Quite surprisingly we also identified an original, but retrospectively not unexpected, bias in the two steps involving circularization of DNA segments (Figure 1A, step 6 and 9). It is known that the mechanical properties of DNA are such that the length of a fragment can strongly influence the efficiency of a circularization reaction. If the fragment is too small, the bending persistence of DNA is such that both ends cannot be ligated. If the fragment is two long, the entropic contribution to the free energy will also disfavor ligation. Here indeed, the distribution of the sum of the sizes (d_A+d_B) of two interacting RF A and B presents a typical circularization efficiency profile, including an optimal circularization length close to 500 bp (Figure 1E, [20]).

Intriguingly, a 10.5 bp periodicity of the circularization efficiency could be observed for the average number of circularization events for which $d_A + d_B < 500$ bp, overall (i.e for the HindIII-MspI experiment, about 15% of the interactions fall into this category). Such a periodicity is actually predicted by polymer physics and results from the natural twist of the double helix which is 10.5 bp [21]. Here, the phenomenon can be observed at an unprecedented resolution (see inset of Figure 1E) and consists in a bias that could affect any experimental procedure involving a circularization through ligation step.

Due to those various biases, some RFs will be involved in more interactions than expected, whereas others will be underrepresented in the final bank (see Additional file 1: Figure S2). Since this variability results from the experimental protocol rather than the biological reality, it is worth minimizing theses effects by either correcting or normalizing the observed frequencies of interactions [17,18]. These correspond to two different approaches: in order to correct the data, one needs to quantify the biases and then to divide each interaction frequency by its expected value, knowing the bias. On the other hand,

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

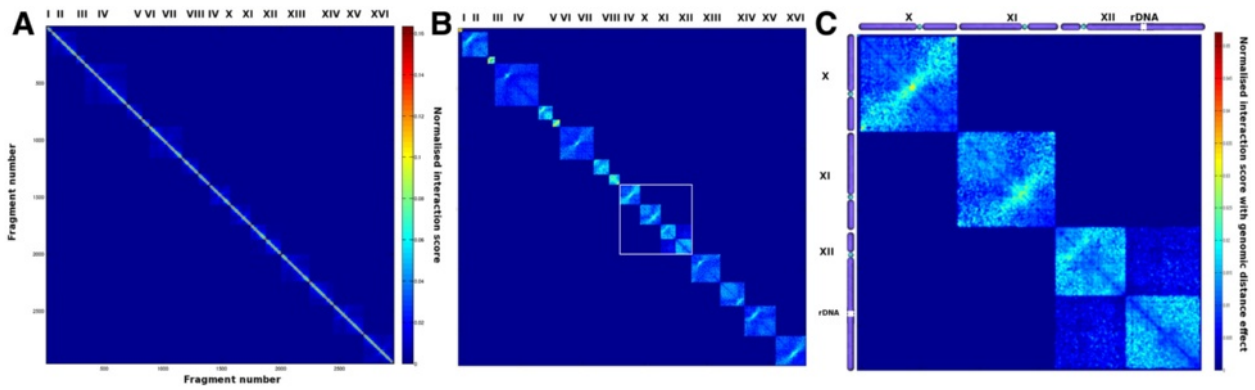


Figure 2 Normalized intra-chromosomal contact map of *S. cerevisiae*. The color scale represents the normalized interaction frequencies between fragments which is calculated with the Sequential Component Normalization. **A**) Matrices of the sixteen chromosomes from *S. cerevisiae*. The strongest interactions are at the diagonale i.e. for close fragments along the chromosome. **B**) The normalized interaction score is calculated with the SCN method and taking into account the effect of the genomic distance. **C**) Zoom on chromosomes X, XI and XII. Chromosome XII is spatially segregated in two compartments by the rDNA locus.

no prior knowledge of the bias is needed to normalize the data: the procedure consists in dividing each interaction frequency between two fragments by the product of the sums, or the norms, of the total interaction reads involving those fragments (see below).

Generation of a normalized contact map through the "Sequential Component Normalization" (SCN) methodology

The correction method developed for the human Hi-C dataset is not readily adaptable to the yeast dataset since there is an additional circularization bias to the RF length and GC content bias [17]. A important issue with the circularization bias is that it is highly non monotonous: for example, it favors circularization lengths of 261 bp, but disfavors circularization length of 266 bp and again favors circularization lengths of 271 bp and so on and so forth (see inset in Figure 1D). A similar methodology that was previously described in [17] was first applied in order to correct for this bias. However, the nature of the bias did not allow reaching a satisfying solution because of the non-monotonous specificity. In the following, instead of correcting each of the interactions frequencies individually, contact maps were normalized globally through what we called the SCN approach, which can be applied to any genomic contact map and independently from the protocol that was used to generate it. The normalization described below is based on the interactions exhibited by the entire restriction fragments, before the second digestion, in order to remain as broadly generalizable as possible to other experimental protocols. The reason why we applied normalization on the fragment instead on the extremities is that for each pair of fragment there are four possibilities to make religation event. Each of those four possibilities will exhibit a different GC content and a different dA+dB and therefore the biases described in Figure 1D and 1E, that depends on the extremities, will be smoothen out when aggregating the combinations together. This point was also discussed in the original paper [13]. The advantage of this method is that it smoothens out all the biases described above and therefore provides a cleaner view of the frequency of interaction between any pair of restriction segments in the genome.

Intra- and inter-chromosomal interactions were treated separately but using the same procedure. Firstly, normalization will give an equal weight to each fragment in the contact map. Therefore, RF with very low number of reads, corresponding to RF that could not be properly detected, are likely to introduce noise in the normalized contact map and have to be removed (see Additional file 1: Figure S3). In order to identify these fragments, we computed the distribution of reads in the contact map (see Additional file 1: Figure S2B). This distribution is roughly gaussian, with a long tail corresponding to low interaction fragments. Based on this distribution, we cut the tail of the distribution (see Methods for further information).

Once low interacting fragments are removed, we wish to normalize all rows and columns of the contact map to one so that the matrix remains symmetric. This was done through the following simple procedure. Firstly, each column vector was normalized to one, using the euclidian norm. Then each line vector of the resulting matrix was normalized to one. The whole process was repeated sequentially until the matrix become symmetric again with each row and each column normalized to one (Additional file 1: Figure S4 and Methods). Usually, two or three iterations are sufficient to insure convergence.

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

Since it involves a sequential normalization of column and line vectors of the matrix, this method was named Sequential Component Normalization (SCN). This normalization can be viewed as a sequence of extensions and shrinking of interaction vectors so that they tend to reach the sphere of radius one in the interactions space. A similar and faster approach is to divide all the matrix elements c_{ij} by the product of the norms of row i and column j : $c^*_{ij} = \frac{c_{ij}}{|c_{ik}||c_{kj}|}$. This method yields to a normalized contact map overall very similar to SCN (Additional file 1: Figures S5 and S6). However since the sum of each component is not necessarily equal using this method, it may bias further analysis such as assessing the 3D colocalization of genomic elements (see below). An alternative normalization method has been used so far by other groups [9], that use the sum of the components instead of the euclidian norm : $c^*_{ij} = \frac{c_{ij}}{\sum_k c_{ik} \sum_k c_{kj}}$. We noticed that this method yields to a contact map with lower contrast than the SCN (Additional file 1: Figure S5 and S6) and therefore recommend SCN use in further works. The normalization using the sum will give more weight to fragments wich makes fewer interactions whereas our normalization will give more weight to fragments interacting moderately with many fragments. Intra and inter-chromosomal interactions were separated in two datasets and the corresponding normalized contact matrices between RFs were plotted as a function of their position along chromosomes (Figure 2A and 3A, respectively).

S. cerevisiae contact maps after SCN

The normalized maps overall are similar to those observed before [13]. Since the probability of interaction between monomers along a polymer is decreasing with the linear distance between them, the diagonal which represents neighboring RFs presents the highest interactions score [4]. In order to increase the contrast and observe interactions between non-adjacent intra-chromosomal RF we then divided the number of interactions between fragments separated by a genomic distance D_g by the average interaction count between fragments separated by the same distance D_g (see Methods). Some features appear more contrasted with respect to the original analysis, with a typical X shape pattern centered on the centromere for each chromosome (Figure 2B). This pattern reflects the fact that the centromere does not interact much with the chromosome arms whereas both arms can interact together. In addition, interactions between RF located on both arms appear clearly more constrained when at symmetrical distances from the centromere and within its vicinity (Figure 2C). In addition, the bipartite structure of chromosome 12 due to the insulating presence of the nucleolar rDNA repeats remains clearly apparent [13]. The corrected contact maps for inter-chromosomal interactions also reveal striking features (Figure 3A). Centromere clustering is clearly apparent and results in all the centromeres interacting with each other's on the map, as in [13]. The interactions between two chromosome arms along their length are also extremely clear. The X shaped patterns at inter-centromeric interactions observed in the matrix indicate that centromeres are somehow isolated from the rest of the chromosomal arm sequence (see for instance chromosome VII and chromosome XVI on Figure 3B). This feature is even more striking when the correlation matrix is drawn similarly to [8] (Additional file 1: Figure S7). In this matrix, each element c_{ij} is the Pearson coefficient between the vectors i and j.

In addition, telomeres are also found to have enriched contact frequencies (for instance chromosome XIII and chromosome IV on Figure 3C). To investigate the role of the chromosomal arm length in the inter-chromosomal interaction frequencies, all chromosomal arms were ranked with respect to their length and the corresponding contact maps were drawn (Figure 4A). This layout conveniently reveals global interaction patterns in respect to chromosomal arm size: shorter arms tend to interact with shorter arms whereas longer arms tend to interact with longer arms (from the upper left corner to the lower right corner). On the contrary, shorter arms tend to make very few contacts with longer ones (upper right and lower left corners on Figure 4A). Zooming on the five shorter arms on the contact map reveals that the interaction frequencies between subtelomeres from shorter arms are important, sometimes even more than centromeres (e.g arms III-L and IX-R, see Figure 4B). To investigate the arm length relationship with subtelomere interactions, we computed the mean interaction frequencies between all sub-telomere pairs for both the normalized and original data. The normalized data exhibit two types of preferred subtelomeric interactions, one for short and one for long chromosome arms, whereas the orginal analysis mostly emphasized short arms interactions (see Additional file 1: Figure S8). Given that the measurements reflect a population average, it is impossible to know from this data if all the telomeres interact preferentially in a similar ways in all cells taken individually. However, similar preferred interactions have been observed in single cells using fluorescent microscopy approaches [22,23] as well as in recent modeling approaches [24]. In addition, the rDNA now appears not only as an intra-chromosomal insulator region, but also modifies the interacting properties of the two DNA segments it delimits. Whereas a gradual shift in interaction frequencies from centromere to telomere is observed for long arms, for chromosome 12 the DNA segment located between the rDNA and the telomere seems

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

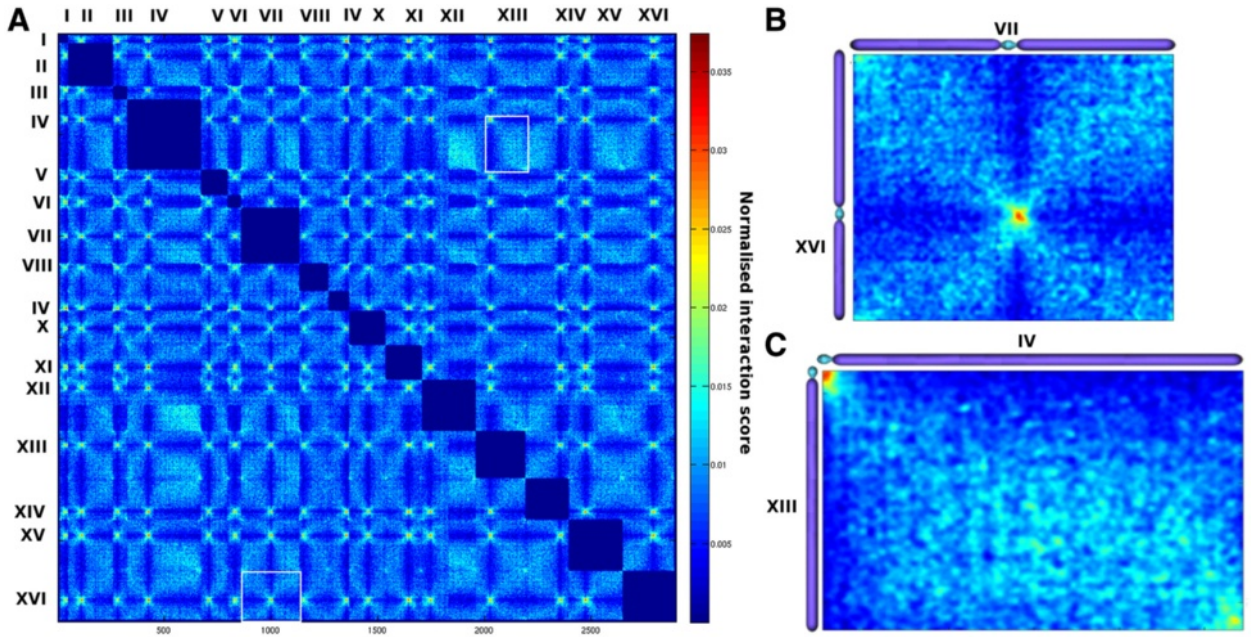


Figure 3 Normalized inter-chromosomal contact map of *S. cerevisiae*. The color scale represents the normalized interaction frequencies between fragments which is calculated with the Sequential Component Normalization. **A**) Matrix of the sixteen chromosomes from *S. cerevisiae*. **B**) Zoom on chromosomes VII and XVI. **C**) Zoom on chromosomes IV and XIII.

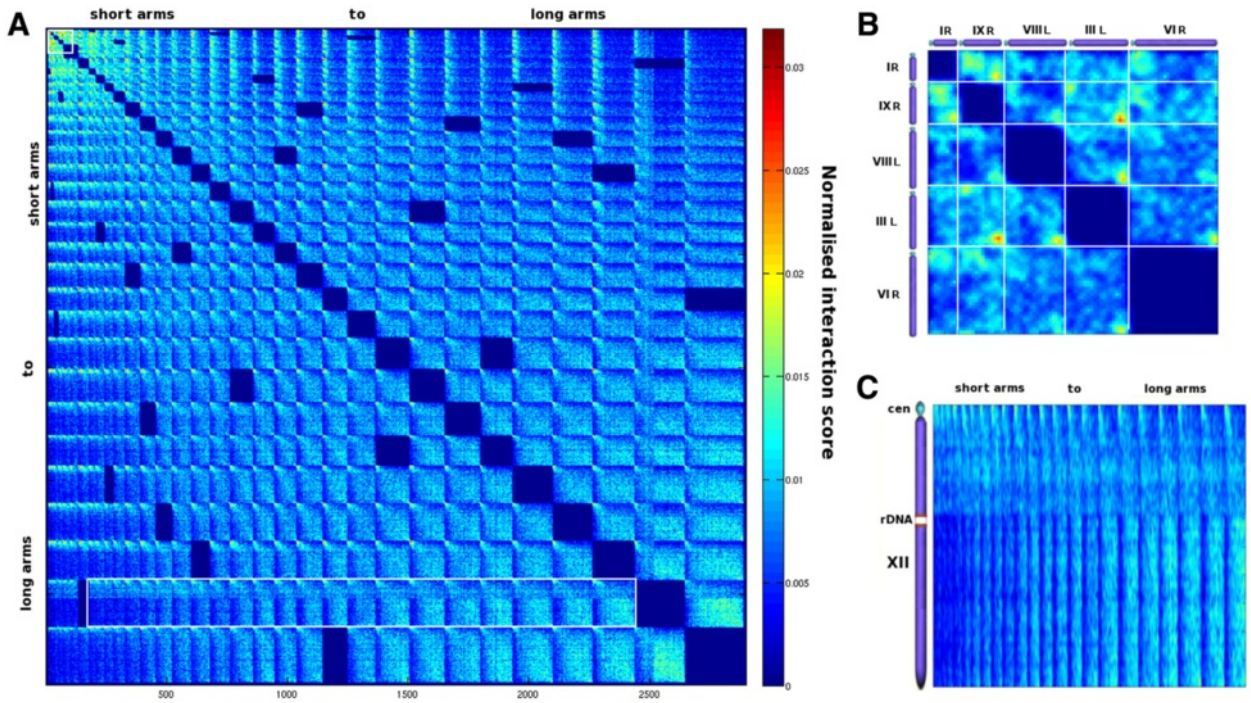


Figure 4 Normalized inter-chromosomal contact map of *S. cerevisiae*. A) Inter-chromosomal contact map of chromosomal arms ranked according to their size, from the shortest (left) to the longest (right). The white empty squares correspond to specific emphasis on the five shortest arms (**B**), and on chromosome XII (**C**).

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

less constrained that the one before the rDNA cluster (Figure 4C).

Re-assessing the 3D colocalization of genomic elements

The influence of this normalization procedure on the preferential interactions detected previously was addressed. In the original analysis, receiver operating curve (ROC) confirmed an expected enrichment of interactions for centromeres and telomeres resulting from the Rabl configuration [13,23]. More interestingly, early replication origins [25] were also shown to interact preferentially, a result experimentally supported [3]. Finally, two preferential interactions regions where identified for tRNA genes, one around the spindle pole body (SPB) and one in the vicinity of thenucleolus [13].

In this paper, we used a different method than the originally published ROC analysis. The initial ROC analysis asked the question: among the pool of strong interactions, is there an enrichment in interactions between two fragments which both carry the genomic object of interest. We ask the question: among the pool of strong interactions carrying one feature of interest, is there an enrichment for interactions with a fragment carrying the same feature (for details about the implementation, see Methods). ROC analysis on the normalized data confirmed the expected centromeres and telomeres preferential interactions (see Figure 5A). In addition, enrichment in interactions between early replication origins was also observed. However, the frequencies of interactions between restriction fragments containing tRNA genes did not exhibit significant increase when using the normalized data (Figure 5B, compare the right panel with the left panel). This was found to be true for all RFs containing tRNAs or for RFs containing only tRNAs previously found to interact preferentially with the SPB or with the nucleolus (see Figure 5B).

The previously described preferential interaction between tRNA genes was lost because it resulted from the fact that, without normalization, two fragments interacting overall more with the whole genome will interact together more frequently than other fragments. This is actually the case for tRNA fragments (see Additional file 1: Figure S9). The reason why tRNA bearing RF interact more frequently than others with all other fragments does not depend on their size, and remain open. A local improvement in cross-linking efficiency resulting from the chromatin state and/or presence of protein complexes is a possibility. Of course, we do not exclude the possibility of actual preferential interactions between tRNA as observed experimentally [26,27] and suggested by other approaches [24]. However, more experiments and higher resolution will be needed to detect those through genomic 3C approaches.

Normalization of the human genome contact map using SCN

In order to test how the SCN approach can be applied to the interaction map of a larger genome, we used the human genome-wide dataset published in 2009 by Lieberman et al. [8]. The restriction enzyme used in this dataset cuts the human genome over 830,000 times. Therefore, the number of potential interaction in the experiment is higher than 340 billion. Since the typical number of reads obtained in such experiment hardly reaches one billion [11], the resulting genome wide contact matrix is very sparsed. In order to get enough information to build a contact map, one can bin the matrix by adding the contacts over several fragments along the genome together. For intra-chromosomal interactions, a typical bin size of about ten fragments is adequate since most of the interaction detected in such an experiment are intra-chromosomal and since the number of possible intra-chromosomal interactions is much lower than the number of possible inter-chromosomal interactions. For inter-chromosomal interaction the bin size has to be increased considerably. We used a bin of one hundred fragments to build the corresponding contact map for the human genome and normalized it through the SCN method. The resulting map clearly shows preferential interactions between small chromosomes and between the long arm of long chromosomes (Additional file 1: Figure S10). Importantly, ROC curves which are used to determine the genomic elements enriched at high interaction hotspot strongly depend to whether or not the data were normalized. We performed ROC analysis on the binding sites of the CCCTC-binding factor (CTCF), a zinc finger protein that plays an important role in the organization of chromatin by mediating inter and intra-chromosomal contacts between distant loci [28,29], PolII, the centromeres and the telomeres. The results for both raw and normalized data clearly show that the preferential interactions of CTCF, PolII and centromeres are only seen on the properly normalized data (Figure 6).

Conclusions

The method described above consists in an easy and convenient way to normalize and represent genomic 3C data. It is worth recalling that before doing any normalization procedure, one has to identify the products and filter out all those that do not correspond to what is expected from the experimental protocol. It represents here more than 90% of the total reads. Depending on the protocol used, the biases in the data will vary, generating an extra number of reads that should not be used in the analysis. Among those identified in the present study, the original circularization bias is certainly of importance for any experimental protocol involving a similar step. While increasing contrast and visibility of the Rabl yeast

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

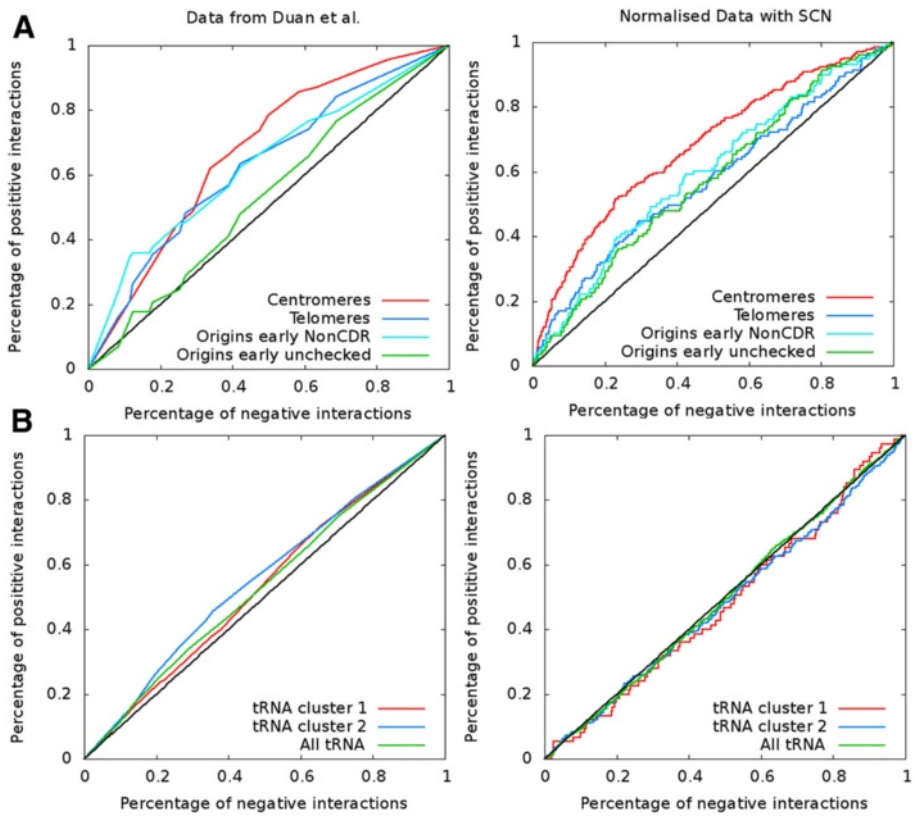


Figure 5 Receiver operating curves to assess 3D colocalization of genomic elements for the yeast contact map. Receiver operating curves (ROC) were used to assess 3D colocalization of different genomic elements. Data from Duan et al. [13] (left column) and normalized data (right column) were used. **A**) Centromeres, Telomeres, early origins of replication give positive signal with both types of data. **B**) The group of tRNA was assessed for 3D colocalization. Two clusters proposed by [13] were assessed with both data: cluster 1 of tRNA genes proposed to colocalize near rDNA and cluster 2 of tRNA genes proposed to colocalize near centromeres. The data from [13] give a positive signal contrary to the data normalized with SCN.

genome organization, the procedure described here confirms the preferential interactions of specific elements, such as early replication origins. However, it also revealed that what could appear like enrichment in interactions between other elements has to be carefully interpreted.

The SCN normalization procedure proposed here will be helpful once higher density contact maps of *S. cerevisiae* become available, and can be conveniently adapted to any other organisms. Increasing the resolution of these contact-maps will likely reveal more features, and can be addressed either through alternative protocols addressing the "invisible" zones of the genome (for instance by increasing the length of the sequenced reads or using various restriction enzymes), or through increasing the number of reads.

Methods

Alignment of the reads on the reference genome

The paired-end sequence reads from banks (SRP002120) were aligned along the yeast genome of the sequenced strain S288C (2011-02) with Bowtie2 [30]. Raw data were converted into fastq files and sent to the aligner. Only reads exhibiting non-ambiguous alignment on the genome were retained. This was done by using the preset parameter "–very-sensitive" and setting a threshold on the mapping quality. The mapping quality Q is defined as Q = -10 × log10(p) where p is the probability that the reported position is false. The higher Q, the more unique is the positioning. Reads with a score lower than 30 were discarded which means that there is one in a thousand chance that a reported position is wrong.

Statistical analysis of the different biases in the contact frequencies

In the following, we analyzed separately each different experiment conducted in [13] since different protocols can produce different results. Notably, the use of the secondary enzyme (MspI or MseI) change the potential interactions that can be observed.

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

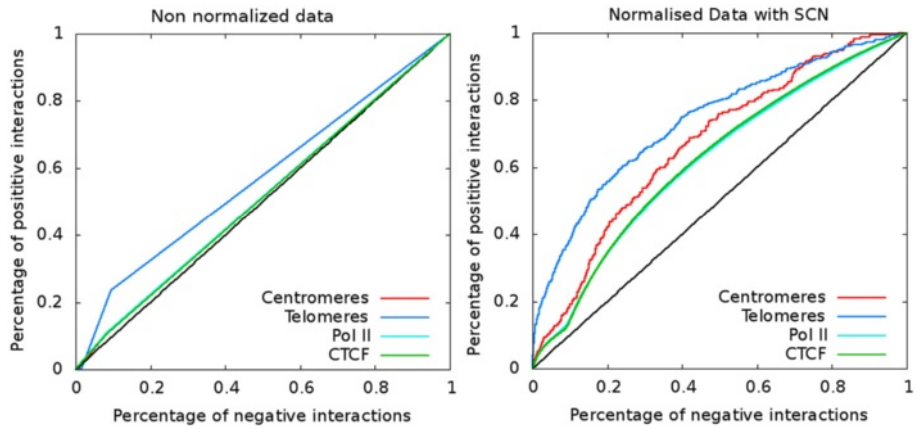


Figure 6 Receiver operating curves to assess 3D colocalization of genomic elements for the human contact map. Receiver operating curves (ROC) were used to assess 3D colocalization of different genomic elements for the human contacts map of Lieberman et al [8]. Non normalized data (left column) and normalized data (right column) were used. Only Telomeres give positive signal when using the non normalized data (curves for Centromeres, PolII are superimposed with the CTCF curve). When using the data normalized with SCN, all genomic elements tested give positive signal to the ROC test (curve for PolII is superimposed with CTCF curve).

Only the reads exhibiting a position on the genome reconcilable with the protocol design were retained (Figure 1A). Firstly, they are expected to map at a distance of about 20 bp to the nearest Hind III restriction site due to the use of the enzyme Ecop15I at the step 10 of the protocol (Figure 1A). We computed the number of read pairs as a function of the distance between the beginning of the read to the next RE1 site for each experiment. We found little difference between condition A and condition B (conditions A and B differ in the DNA concentration at the 3C step: A: 0.5 $\mu g/ml$, B: 0.3 $\mu g/ml$). Whereas reads from datasets HindIII-MspI-A and HindIII-MseI-A have maximums for distances equals to 20, 21 and 22 bp, HindIII-MspI-B, HindIII-MseI-B and HindIII-MseI-uncross-control-B exhibit maximums for distances equals to 21, 22 and 23 bp (see Additional file 1: Figure S11). We only kept reads with distance between the beginning of the read and the next RE1 site equal to 20, 21 and 22 bp for condition A and equals to 21, 22 and 23 bp for condition B. Secondly, interactions involving fragments which have no restriction site for the secondary enzyme or a secondary site with a position located less than 20 bp from the first restriction site were also discarded. Finally, interactions corresponding to self-circularization (loops) and ligation of adjacent fragments (religation events) were removed from the analysis.

Bias of fragments sizes

The influence of the size of the RF on the observed frequency of interaction was analyzed as followed. Firstly, the sizes of each fragment were binned into equally sized windows (bin size: 100 bp). For each bin, the number of possible fragments N_i was counted according to the initial distribution of fragment sizes. The number of detected reads in the experiment R_i is counted for each bin. Then, the number of reads per fragment r_i was calculated from these two numbers, with $r_i = R_i/N_i$. We fitted the data points with the following function: $f(x) = A(1 - (1 - p_c)^x)$ which is related to the probability that the fragment is crosslinked at least one time. A is a normalization constant and p_c is the probability of crosslink by base paire (we found $A \simeq 4000$ and $p_c \simeq 0.004$). The effect of the fragments size on the number of interaction reads before and after SCN is represented on Additional file 1: Figure S12 in the additional documentation.

Bias of GC content

The GC content influence was determined by binning the GC content of the mean of the two reads of each interaction (taking the sequence of the 20 bp before or after the restriction site RE1 according to the orientation of the read) into equally sized bins (bin size: 2.5%). For each bin, the number of possible interactions N_i according to the initial distribution of GC contents, and the number of detected reads in the experiment R_i were estimated. These two numbers were divided to generate the number of reads per possible interaction: $r_i = R_i / N_i$.

Bias in the circularization steps

The effect of the lengths of the DNA segment during circularization steps was analyzed by binning the size of the circularization segment into equally sized bins (bin size: 1 bp). The lengths were calculated using the coordinates of the positions of RE1 and RE2 restriction sites (MspI or MseI) on the reference genome. For each bin, the number

Cournac *et al. BMC Genomics* 2012, **13**:436
http://www.biomedcentral.com/1471-2164/13/436

of possible interactions N_i according to the initial distribution of segment lengths and the number of detected reads in the experiment R_i were estimated. These two numbers were divided to give the number of reads per possible interaction: $r_i = R_i/N_i$.

Generation of matrices

Before the normalization step, we removed an important number of restriction fragments that could not be correctly detected in the experiment. First, non-mappable fragments were discarded. They correspond to fragments whose both extremities give ambiguous mapping (i.e the 20 bp sequence of the read can be located in several loci in the genome due to the presence of repeated sequences). 104 fragments felt into this category, most of them positioned in the subtelomeric regions of the chromosomes which are indeed enriched in repeated sequences. Second, all RFs that did not present a RE2 site were discarded (i.e. a MspI site for the experiment carried out with HindIII and MspI as RE1 and RE2, respectively). Intriguingly, these fragments are still detected in the experiment but with a smaller number of reads: Additional file 1: Figure S2 A represents the distribution of the number of reads per fragment. Two groups can be distinguished: a group corresponding to fragments that do not exhibit a secondary enzyme restriction site (having a number of reads inferior to 1000) and a second group corresponding to fragments having a RE2 site. Overall, 1217 RFs were concerned, which left 3098 RFs from the original 4454 for the MspI-HindIII experiment.

In addition, several RFs still exhibited a very small number of interaction reads with respect to the average (less than a few dozens reads re. the HindIII-MspI experiment), as seen on Additional file 1: Figure S2 B were the distribution of the euclidian norms of all fragments is plotted. Fragments with a norm under 30 were discarded from the analysis. 168 fragments felt into this category when considering inter-chromosomal interactions (see Additional file 1: Figure S2 B) and, in good agreement with the biases identified above, they exhibited either low GC content at their extremities, or the length of the two ligated fragments dA + dB had disfavored circularization.

Then each column vector was normalized to one, using the euclidian norm.

Then each line vector of the resulting matrix was normalized to one. The whole process was repeated sequentially until the matrix become symmetric again with each row and each column normalized to one. Convergence is not mathematically guaranteed for any matrix. For positive matrices which we have to deal with, it is generally attained in two or three iterations. For graphic representation the matrix was blurred using a convolution matrix, with as kernel the 3x3 matrix [0.05 0.05 0.05; 0.05 0.05 0.05; 0.05 0.05 0.05]. The convolution was repeated 10 times so that the structures appear clearly.

For the intra-chromosomal interactions, an extra step was added before normalization to take into account the effect of the genomic distance. First, we average the number of reads per possible interaction for every possible genomic distance. For each bin, the number of possible interactions N_i according to the initial distribution of genomic distances was estimated as well as the number of detected reads in the experiment R_i. Then, these two numbers were divided to generate the number of reads per possible interaction: $r_i = R_i / N_i$. Then, we use polynomial functions to fit the data points (see Additional file 1: Figure S13). Finally, we divide the number of reads of the experiment for each interaction by the expected value given by the fit at the genomic distance of the interaction.

This normalization step allows us to see interactions that are stronger than what it was expected due to the genomic distance effect. The SCN can be applied subsequently.

Re-assessing the 3D colocalization of genomic elements

We used the statistical tool called Receiver Operating Curve (ROC) to look for 3D colocalization of several genomic elements. We slightly modified the initial method. We process as follows: first, we selected only the interactions containing one or two fragments containing the genomic element (centromere, telomeres, early origins of replication [25] or tRNA) instead of taking all detected interactions. We ranked the interactions of this set by p-values for the data of [13] and by the normalized interaction score for the normalized data. A interaction is labeled "positive" if both fragments contain the genomic element and negative in the other case. The ROC is generated by traversing the ranked list and plotting the percentage of positive and negative interaction above the threshold (p-value or normalized interaction score). If a genomic element tends to have strong interactions then the percentage of the positive interactions would be higher and the corresponding curve will be above the line x=y. Telomeres regions were determined as the last ten RF from each arm. Positions of early origins of replication and tRNA were similar to those used in [13].

Additional file

Additional file 1: Additional-documentation. This document gives more information concerning the filtering of fragments and the normalization procedure.

Competing interests

The authors declare no conflicts of interests.

Author's contributions
AC, RK and JM designed the analysis. AC and HMN performed the analysis. AC, HMN, MM, RK and JM interpreted the data. AC, RK and JM wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements
The research that led to these results was funded by ANR PIRIBIO grant ANR-09-PIRI-0024. MM is the recipient of an Association pour la Recherche sur le Cancer fellowship (20100600373). This project receives funding from the European Research Council under the 7th Framework Program (FP7/2007-2013) / ERC grant agreement 260822 to RK.

Author details
[1] LPTMC, UMR 7600, Tour 12-13/13-23, Boîte 121, 4, Place Jussieu, 75252 Paris Cedex 05, France. [2] Institut Pasteur, Groupe Imagerie et Modélisation, Department of Cellular Biology and Infection, F-75015 Paris, France. [3] CNRS, URA2582, F-75015 Paris, France. [4] University Pierre et Marie Curie, Cellule Pasteur, 75252 Paris Cedex 05, France. [5] Institut Pasteur, Spatial regulation of genomes group, Department of Genomes and Genetics, F-75015 Paris, France. [6] CNRS, UMR3525, F-75015 Paris, France.

Received: 6 April 2012 Accepted: 21 August 2012
Published: 30 August 2012

References

1. Misteli T: **Beyond the sequence: cellular organization of genome function.** *Cell* 2007, **128**(4):787–800. [http://dx.doi.org/10.1016/j.cell.2007.01.028]
2. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, Zhang J, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei CL, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, *et al*: **Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.** *Cell* 2012, **148**(1-2):84–98. [http://dx.doi.org/10.1016/j.cell.2011.12.014]
3. Knott SRV, Peace JM, Ostrow AZ, Gan Y, Rex AE, Viggiani CJ, Tavaré S, Aparicio OM: **Forkhead transcription factors establish origin timing and long-range clustering in S. cerevisiae.** *Cell* 2012, **148**(1-2):99–111. [http://dx.doi.org/10.1016/j.cell.2011.12.012]
4. Dekker J, Rippe K, Dekker M, Kleckner N: **Capturing chromosome conformation.** *Science* 2002, **295:**1306–1311.
5. Simonis M, Klous P, Splinter E, Moshkin Y, Willemsen R, de Wit E, van Steensel B, de Laat W: **Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C).** *Nat Genet* 2006, **38**(11):1348–1354. [http://dx.doi.org/10.1038/ng1896]
6. Dostie J, Richmond TA, Arnaout RA, Selzer RR, Lee WL, Honan TA, Rubio ED, Krumm A, Lamb J, Nusbaum C, Green RD, Dekker J: **Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements.** *Genome Res* 2006, **16**(10):1299–1309. [http://dx.doi.org/10.1101/gr.5571506]
7. Hakim O, Misteli T: **SnapShot: Chromosome confirmation capture.** *Cell* 2012, **148**(5):1068.e1–1068.e2. [http://dx.doi.org/10.1016/j.cell.2012.02.019]
8. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO, Sandstrom R, Bernstein B, Bender MA, Groudine M, Gnirke A, Stamatoyannopoulos J, Mirny LA, Lander ES, Dekker J: **Comprehensive mapping of long-range interactions reveals folding principles of the human genome.** *Science* 2009, **326**(5950):289–293. [http://dx.doi.org/10.1126/science.1181369]
9. Kalhor R, Tjong H, Jayathilaka N, Alber F, Chen L: **Genome architectures revealed by tethered chromosome conformation capture and population-based modeling.** *Nat Biotechnol* 2012, **30:**90–98. [http://dx.doi.org/10.1038/nbt.2057]
10. Zhang Y, McCord RP, Ho YJ, Lajoie BR, Hildebrand DG, Simon AC, Becker MS, Alt FW, Dekker J: **Spatial organization of the mouse genome and its role in recurrent chromosomal translocations.** *Cell* 2012, **148**(5):908–921. [http://dx.doi.org/10.1016/j.cell.2012.02.002]
11. Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS, Ren B: **Topological domains in mammalian genomes identified by analysis of chromatin interactions.** *Nature* 2012, **485**(7398):376–380. [http://dx.doi.org/10.1038/nature11082]
12. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, ichi Noma K: **Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation.** *Nucleic Acids Res* 2010, **38**(22):8164–8177. [http://dx.doi.org/10.1093/nar/gkq955]
13. Duan Z, Andronescu M, Schutz K, McIlwain S, Kim YJ, Lee C, Shendure J, Fields S, Blau CA, Noble WS: **A three-dimensional model of the yeast genome.** *Nature* 2010, **465**(7296):363–367. [http://dx.doi.org/10.1038/nature08973]
14. Rodley CDM, Bertels F, Jones B, O'Sullivan JM: **Global identification of yeast chromosome interactions using Genome conformation capture.** *Fungal Genet Biol* 2009, **46**(11):879–886. [http://dx.doi.org/10.1016/j.fgb.2009.07.006]
15. Sexton T, Yaffe E, Kenigsberg E, Bantignies F, Leblanc B, Hoichman M, Parrinello H, Tanay A, Cavalli G: **Three-dimensional folding and functional organization principles of the Drosophila genome.** *Cell* 2012, **148**(3):458–472. [http://dx.doi.org/10.1016/j.cell.2012.01.010]
16. Dekker J: **The three 'C' s of chromosome conformation capture: controls, controls, controls.** *Nat Methods* 2006, **3:**17–21. [http://dx.doi.org/10.1038/nmeth823]
17. Yaffe E, Tanay A: **Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.** *Nat Genet* 2011, **43**(11):1059–1065. [http://dx.doi.org/10.1038/ng.947]
18. Gascoigne DK, Ryan MP, Taft J, Mattick JS: **Reassessment of the Hi-C analysis of human genome architecture.** 2011. [http://matticklab.com/index.php?title=File:HiCMain.pdf]
19. Dohm JC, Lottaz C, Borodina T, Himmelbauer H: **Substantial biases in ultra-short read data sets from high-throughput DNA sequencing.** *Nucleic Acids Res* 2008, **36**(16):e105. [http://dx.doi.org/10.1093/nar/gkn425]
20. Shore D, Langowski J, Baldwin RL: **DNA flexibility studied by covalent closure of short fragments into circles.** *Proc Natl Acad Sci U S A* 1981, **78**(8):4833–4837.
21. Du Q, Smith C, Shiffeldrim N, Vologodskaia M, Vologodskii A: **Cyclization of short DNA fragments and bending fluctuations of the double helix.** *Proc Natl Acad Sci U S A* 2005, **102**(15):5397–5402. [http://dx.doi.org/10.1073/pnas.0500983102]
22. Ruault M, Meyer AD, Loïodice I, Taddei A: **Clustering heterochromatin: Sir3 promotes telomere clustering independently of silencing in yeast.** *J Cell Biol* 2011, **192**(3):417–431. [http://dx.doi.org/10.1083/jcb.201008007]
23. Therizols P, Duong T, Dujon B, Zimmer C, Fabre E: **Chromosome arm length and nuclear constraints determine the dynamic relationship of yeast subtelomeres.** *Proc Natl Acad Sci U S A* 2010, **107**(5):2025–2030. [http://dx.doi.org/10.1073/pnas.0914187107]
24. Tjong H, Gong K, Chen L, Alber F: **Physical tethering and volume exclusion determine higher-order genome organization in budding yeast.** *Genome Res* 2012, **22**(7):1295–1305. [http://dx.doi.org/10.1101/gr.129437.111]
25. Rienzi SCD, Collingwood D, Raghuraman MK, Brewer BJ: **Fragile genomic sites are associated with origins of replication.** *Genome Biol Evol* 2009, **1:**350–363. [http://dx.doi.org/10.1093/gbe/evp034]
26. Haeusler RA, Pratt-Hyatt M, Good PD, Gipson TA, Engelke DR: **Clustering of yeast tRNA genes is mediated by specific association of condensin with tRNA gene transcription complexes.** *Genes Dev* 2008, **22**(16):2204–2214. [http://dx.doi.org/10.1101/gad.1675908]
27. Thompson M, Haeusler RA, Good PD, Engelke DR: **Nucleolar clustering of dispersed tRNA genes.** *Science* 2003, **302**(5649):1399–1401. [http://dx.doi.org/10.1126/science.1089814]
28. Phillips JE, Corces VG: **CTCF: master weaver of the genome.** *Cell* 2009, **137**(7):1194–1211. [http://dx.doi.org/10.1016/j.cell.2009.06.001]

29. Botta M, Haider S, Leung IXY, Lio P, Mozziconacci J: **Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide.** *Mol Syst Biol* 2010, **6:**426. [http://dx.doi.org/10.1038/msb.2010.79]
30. Langmead B, Trapnell C, Pop M, Salzberg SL: **Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.** *Genome Biol* 2009, **10**(3):R25. [http://dx.doi.org/10.1186/gb-2009-10-3-r25]

doi:10.1186/1471-2164-13-436
Cite this article as: Cournac *et al.*: Normalization of a chromosomal contact map. *BMC Genomics* 2012 **13**:436.

Chapter 4

Genome assembly from contact data

Contents

4.1 Introduction

Frederic Sanger is one of the few scientists who received two Nobel prices in his career. One of his greatest scientific achievement was the development of the so called Sanqer sequencing technique in 1977,Sanger, Frederick and Nicklen, Steven and Coulson, Alan R [123]. For the first time it was possible to read long DNA sequences of a thousand base pairs. The main drawback of this technique is its low throughput: it takes up to one day to sequence a million base pairs sample.

With the recent advances in sequencing techniques it is common to acquire more than twenty millions bp per hour in a single experiment. However, contrary to classic Sanger sequencing, only relatively short reads of a few hundreds bp can be obtained, Monya Baker, 2012, [6]). Therefore despite these great technical breakthrough, it is still impossible to read a whole chromosome directly. Instead the short sequences are pieced together by sophisticated computer programs, assemblers, to form long DNA sequences called contigs (figure 4.1). In an ideal world assemblers algorithms would produce as many contigs as there are chromosomes in the studied samples. However because of repeated sequences, genome complexity, and technical limitations, this situation is almost never reached. Thus, time consuming and expensive experiments are necessary in order to aggregate the contigs into scaffolds which will approximate the true genomic structure. At the time we write this thesis, there is no rigorous metric to estimate the quality of an assembly. Because of this strong limitation, it is very common to run many times, with different algorithm, the assembly process. The methodology we will introduce in this chapter, will allow, not only to correct and enhance existing genomic structure from contact data, but will also provide a rigorous probabilistic framework to assign a probabilistic score to a given scaffold.

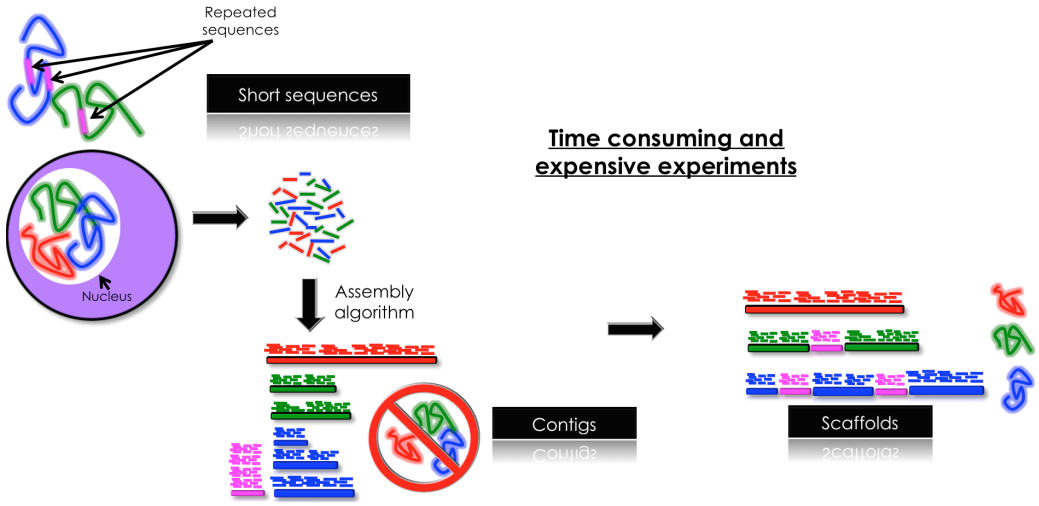


Figure 4.1: De novo genome assembly workflow. The nuclear DNA is cut randomly into small pieces which will be read by modern sequencers. Assembly algorithms piece together these short reads into largest contiguous sequences called contigs. Then, additional biochemical experiment are performed in order to create scaffolds of contigs which approximate the true genomic structure.

4.1.1 De novo genome assembler

We will give now a brief overview of some of the most popular assembly algorithm before introducing their limitations and some of the techniques used to construct scaffolds.

Assembly programs Basically, assembler algorithm fall into two categories[1]

- Greedy algorithms. These computer programs try to create contigs by computing some heuristic and meta heuristic about the sequence coverage of the short reads. Given two sequences $r_1 = (AGCTAT)$ and $r_2 = (AAGCTA)$, an overlapping score $s(r1, r2) = f(5)$ is computed, allowing to create the extended sequence $c = (AAGCTAT)$. Some of the most popular algorithm embedded in this category are PHRAP [57] and TIGR, [132].

- Graph based algorithms. These methods rely heavily on the work performed by Nicolaas de Bruijn in 1946 (Compeau, Phillip EC and Pevzner, Pavel A and Tesler, Glenn, 2011, [24]). First reads are split into smaller sequences of k elements called k-mers. Then the goal of the program is to find the smallest circular super string of nucleotide which contains all available k-mers. In a De Bruijn graph this task is achieved by finding an Eulerian cycle using the Euler algorithm (Compeau, Phillip EC and Pevzner, Pavel A and Tesler, Glenn, 2011, [24]). VELVET, (Zerbino, Daniel R and Birney, Ewan 2008,[150]), AbYss (Simpson, Jared T et al, 2009, citesimpson2009abyss) and SOAPDenovo (Li Ruiqiang et al, 2010 [77]) are among the most popular algorithm using graph representation.

Limitations Despite their empirical efficiency, there is still a huge lack of statistical tools to robustly assess the validity of the output of existing assemblers.(Howison, Mark and Zapata, Felipe and Dunn, Casey W2013,[62]). Because of repeated sequences both Graph Based and Greedy algorithm will encounter ambiguities in their process yielding in the best case, breaks of contigs, and in the worst case mis-assembled areas. The situation

[1]Please note that we do not perform a review of all existing methods. We give an insight about the main existing techniques.

get even more difficult in case of diploid or n ploid genomes where every allele of a gene has to be correctly located. This problem is refered as haplotype phasing.

4.1.2 Scaffold completion

PCR amplification PCR amplification is one of the easiest but also the more low throughput way to connect contigs to each other. Basically one will try to amplify the junction of two contig A and B. If there is amplification then the likelihood of having the contig A->B is high. Therefore, to verify every junction of a 100 contigs assembly, it is necessary to perform $200^2 = 40000$ controls, which can be quite time consuming.

Optical mapping A recent technique based on optical probes has been recently developed to address some of the main limitations introduced above (Lam, Ernest T et al , 2012, [74]). This method allows visually to correct and to solve the haplotype phasing problem. First DNA molecules of interest are labeled with some specific dies. Then they get stretched and elongated in nano-arrays where optical detection processes of the probes allow to determine their linear proximity.

4.2 Principle of Hi-C driven genome assembly

As described in chapter 3, a typical Hi-C experiment requires a fully assembled genome to infer the spatial conformations of the chromosomes. The method that we will introduce in this section is based on the reversed process. We will show that from the contact data it is possible to:

- retrieve the linear organization of the chromosomes,
- address the problem coming from repeated sequences,
- and provide a probabilistic score to any given genome.

4.2.1 Main concept

Because of the semi flexible nature of the chromatin fiber, loci which are very close linearly will interact much more than others. For instance two loci that are 15 kbp apart will interact more than two loci that are 1 mbp apart. This specificity is easily verified: in every genome wide contact matrix we can see a strong diagonal signal. This basic observation leads to the following statement:

Hypothesis 1. *High linear genomic proximity implies strong spatial proximity and therefore a high 3C signal* [2].

The main idea behind this project is to assume the reciprocal proposition of this implication statement:

Hypothesis 2. *High 3C signal implies high linear genomic proximity.*

[2] See a) in figure 4.2

Analogy with jigsaw puzzles The genome of the yeast Saccharomyces cerevisiae is fully assembled and 3C as well as HiC experimental data sets are available for this specie. In figure 4.2 a) the strong diagonal signal confirms hypothesis 1. Suppose now that the genome that is used to map the produced reads is wrong (figure 4.2 2). In this case, the matrix that is produced by the standard analysis procedure turns into a scrambled jigsaw puzzle. Based on the typical ill-patterns, reassembling the genome would be easy, at least visually. The spatial contact data provide a strong hint about the connectivity of the contigs and this simple example gives a strong indication of the relevance of this method.

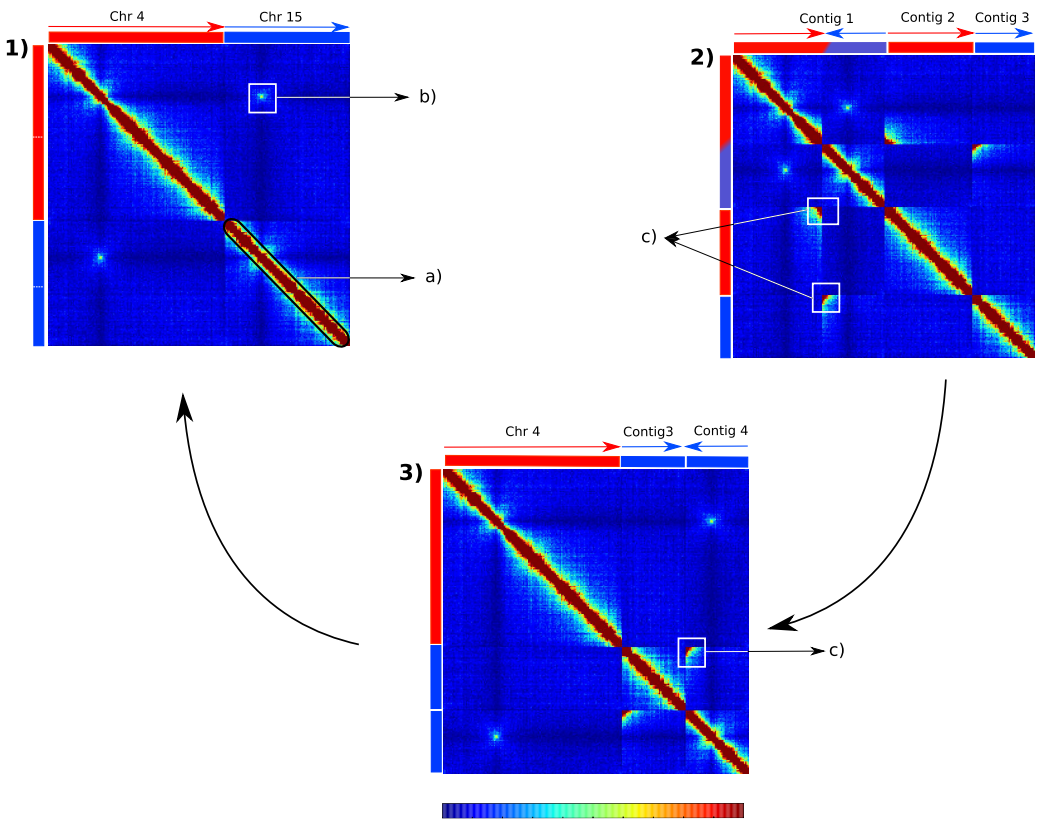


Figure 4.2: Re-assembly by contact data: a jigsaw puzzle.
1) displays the 3C contact data mapped on chromosome 4 and chromosome 15 of yeast saccharoymices cerevisiae. There is a ratio of two orders of magnitude between the diagonal signal (a) and a typical inter centromeric contact signal such as (b). 2) displays the same data aligned on a wrong genome. The upstream part of chromosome 4 at position 781795 bp has been translocated to the upstream part of chromosome 15 at position 533791, leaving the two remaining blocs of these chromosomes free. Therefore, the initial reads are no longer mapped on two chromosomes but on three contigs. From 2) one could easily re-order the contigs to produce matrix 3) and eventually retrieve the fully assembled genome in 1).

From now on we will assume that the genome of the organism that is used to perform the HiC or 3C experiment is either not finished or does not perfectly match their real linear genomic structure.

4.2.2 A naive approach

We have developed a very simple recursive algorithm and run it on simulated data. This algorithm works as follows:

The whole genome is split into restriction fragments. Therefore, each contig can be seen as an ordered string of oriented fragments. A restriction fragment has at most two[3] adjacent fragments: a left and a right partner. First

[3]A fragment located at the extremity of a contig will have only one partner.

we compute the two most frequent interacting fragments with a given third one. Then we connect recursively fragments to each other until an incompatibility arises. The entire procedure is illustrated in figure 4.3 and detailled in algorithm 1 and algorithm 2.

Algorithm 1 extend_chain

Require: L the list of available fragments
Require: C the list of the two most detected partners for each fragment
Require: f a restriction fragment
Require: d the direction of the extension
Ensure: An extended chain of fragments started at f in the direction d

```
update L : L = L - f
V_f = C(f) ∩ L
pick at random v ∈ V
check retro compatibility: r = f ∈ C(v)
if r then
  if {v} ≠ ∅ and r then
    if d = left then
      return [extend_chain(v, d), f]
    else
      return return [f, extend_chain(v, d) ]
    end if
  else
    return [f]
  end if
else
  return ∅
end if
```

Algorithm 2 build contigs

Require: L the list of all restriction fragments
Require: C the list of the two most detected partners for each fragment
Ensure: A set of ordered strings of fragments

```
while length(L) > 0 do
  (v_l, v_r) = C(f)
  C_f = [extend_chain(v_l, left), f, extend_chain(v_r, right]
end while
```

On simulated data with no Poisson noise the algorithm performs as expected and we obtain perfect recovery of the contiguity of the restriction fragments. However, obviously this situation is over simplified. We will now discuss the limitations of this algorithm and the requirements for a robust reconstruction method.

4.2.3 Limitations and requirements

The chromosome conformation capture experiments are fundamentally counting procedures. Therefore, a slightly more realistic artificial data set can be produced easily by considering the 3C experiment as the output of a Poisson process.

Limitations In this situation the algorithm fails to reconstruct the original contigs and many loops are created. The failure of the method exhibits the fact that even under perfect conditions, where no experimental artifacts are added, the raw contact counts cannot be used directly as a robust indication of the immediate

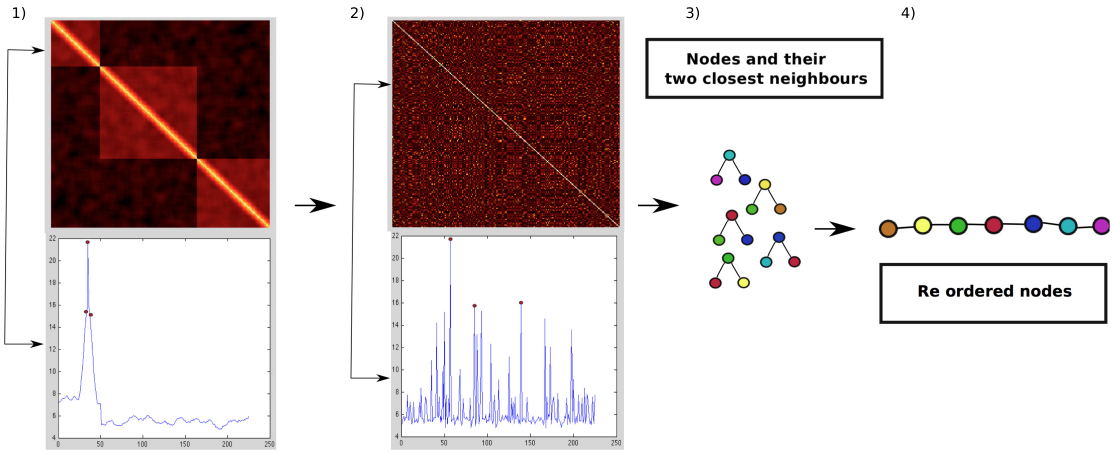


Figure 4.3: Recursive building of contigs.1)Simulated contact matrix over 3 contigs (top).The typical interaction profile (bottom) of a single fragment allows to characterize the two closest neighbors 3). After scrambling of the initial matrix 2) the algorithm succeed to recover the full connectivity of the fragments 4).

vicinity of a restriction fragment.

There is nothing surprising about this failure because the data that is produced by 3C experiments is the result of at least two complex stochastic processes:

- The first random input arises directly from the protocol itself. As described in section 3, many biases and artifacts must be taken into account before any further analysis can take place. Besides these experimental aspects it is crucial to take into account that the experiments are driven on dynamic objects.
- This is where the second stochastic process takes place. Chromosomes of a single cell nucleus are very dynamic polymer objects which physical properties are neither constant over time neither over their monomers. Knowing that 3C experiments are performed over billions of nuclei, a special effort has to be made to take into account the physical properties of the chromosomes.

Requirements A robust reconstruction algorithm should take into account both the biases of the experiments and especially, the physical properties of the chromatin fibers. In the following section we will introduce a rigorous probabilistic framework to infer genomic structures from 3C data.

4.3 A probabilistic framework for genome assembly

4.3.1 Bayesian inference

An over determined system A genome is defined as a set of chromosomes. Each chromosome consists of an ordered sequence of oriented restriction fragments. Lets consider a genome of p chromosomes and n restriction fragments.

Classic results from convex optimization and euclidian distance matrices [55] state that the rank of an euclidian distance matrix (EDM) of a three dimensional object of at least 5 points is bound by 5.

Therefore, in an ideal world with an ideal experimental setup, that is;

- if we were observing a single genomic spatial architecture,

- and if we had access to the full inter fragments euclidian distance matrix,

the system would be clearly over determined and classic optimization procedures would allow us to retrieve the unique spatial conformations of the chromosomes.
In our case we are not interested in the 3D organization of the genome but in the linear configuration. In the ideal world described above, the upper bound of the rank of the EDM would not be 5 but 4, meaning‘ that there is even more redundant information in the data. In fact the exact number of values needed for perfect recovery is even lower.
Each chromosome starts with a fragment origin f_0. Consequently, every fragment is fully characterized by its orientation, the id of the chromosome that it belongs to and its position related to the corresponding fragment origin. Therefore, the linear structure of the genome is completely specified by $3 \times n$ values that are below $4 \times n$.
In the 3C experimental framework:

- we are looking at a population of billions of highly dynamic genomic spatial structures,
- the experiments produce an inter-fragment frequency of contact matrix.

Probabilistic ranking Beyond the fact that 3C is a population based experiment, we can reasonably assume that the linear structure of all the organisms present in the samples is the same[4]. Even if we are not in the ideal situation of inter euclidian distance between fragments, polymer models give non negligible knowledge about the probability of contacts between two loci that are bound to the same chromatin fiber.
The overall goal of our project is to infer the complete genomic structure of the studied organisms in an objective way.What we mean by objective is that the ranking that we will assign to each of our estimations of the real structure will rely on the data and our knowledge of polymer physics. In Probability Theory, The logic of science [65], E.T Jaynes defines an objective inference method as a process who does not depend on the "personality of the user". Given the same data set and the same prior knowledge, two objective methods must lead to the same conclusions[5].
The conditional probability $Pr(\mathscr{G}|D, I)$ provides a quantification of how realistic is a genome $\mathscr{G}$ given the data D and a polymer contact model I. Therefore, the main scope of this section will consist of defining this quantity and introducing algorithmic methods to explore the space of highly likely structures.

4.3.2 Definitions

We define:

- f_i as the restriction fragment which index is i,
- len(f_i) as the base pair length of the fragment f_i,
- gc(f_i) as the mean GC content of the fragment f_i.
- $F = \{f_0, f_1, \ldots, f_n\}$ as the set of all initial restriction fragments
- $\varphi(k, i) = \varphi_k^i$ returns the id and the orientation of the fragment located at position i in contig k.

[4]We will explore the situation of multiple genomes further on in the manuscript
[5]Therefore any of the available assembly program available is objective...

- $\mathrm{id}\left(f_{\varphi_k^i}\right)$ returns the initial id of the restriction fragment $f_{\varphi_k^i}$
- $f_{\varphi_k^i}$ refers, for simplicity of notation to the oriented initial restriction fragment $\overrightarrow{f_{\mathrm{id}\left(f_{\varphi_k^i}\right)}}$, whose position and orientation are encapsulated in φ_k^i.
- $\mathrm{cont}\left(f_{\varphi_k^i}\right) = k$ returns the id of the contig where the restriction fragment is located.
- Φ_i is the set of all the occurrences of the initial restriction fragment f_i in the current genome.

$$\Phi_i = \left\{\varphi_x^y, \text{such that id}\left(f_{\varphi_x^y}\right) = i\right\}$$

- $\mathrm{m}_t(f_i)$ is the number of occurrences of the initial restriction fragment f_i in the current genome. $\mathrm{m}_t(f_i) = card(\Phi_i)$
- $\mathscr{C}_k = \left(f_{\varphi_k^0}, \ldots, f_{\varphi_k^{l_k-1}}\right)$ as a contig which contains l_k restriction fragments. The index k is based on the ascending length of the contig within a genome.
- $\mathrm{circ}\left(\mathscr{C}_k\right)$ returns the circularity of $\mathscr{C}_k$.

$$\mathrm{circ}\left(\mathscr{C}_k\right) = \begin{cases} 1, & \text{if } \mathscr{C}_k \text{is circular} \\ 0, & \text{otherwise} \end{cases} \tag{4.1}$$

- $\mathrm{nFrag}\left(\mathscr{C}_k\right)$ is the number of restriction embedded in $\mathscr{C}_k$.
- $\mathrm{len}\left(\mathscr{C}_k\right)$ is the bp length of the contig $\mathscr{C}_k$.

$$\mathrm{len}\left(\mathscr{C}_k\right) = \sum_{f_i in \mathscr{C}_k} \mathrm{len}\left(f_i\right)$$

- $\mathscr{G}_0 = \{\mathscr{C}_1^0, \ldots, \mathscr{C}_{N_0}^0\}$ as the initial genome used to map the reads from the 3C experiment.
- $\mathscr{G} = \{\mathscr{C}_1, \ldots, \mathscr{C}_N\}$ as a candidate genome made of N contigs.
- R as the set of all the raw paired reads which have been sequenced.
- R_m^0 as the set all all the raw paired reads mapped on $\mathscr{G}_0$
- R_u^0 as the set all all the raw paired reads which have not been mapped on $\mathscr{G}_0$
- D as the $n \times n$ observed contig wide contact matrix produced after the mapping of the raw reads R on the initial genome $\mathscr{G}_0$.
- I as a model of polymer contacts.
- $E_{\mathscr{G}}$ as the $n \times n$ expected contig wide contact matrix generated by I for a given genome $\mathscr{G}$.
- $d_{\mathscr{G}}(f_{\varphi_k^i}, f_{\varphi_l^j})$ as the genomic distance between two restriction fragments $f_{\varphi_k^i}, f_{\varphi_l^j}$located on the same chromosome $\mathscr{C}_k$. Therefore, $d_{\mathscr{G}}(f_{\varphi_k^i}, f_{\varphi_l^j})$ is defined if and only if $k = l$. The detail of the distance computing is illustrated in figure 4.4.

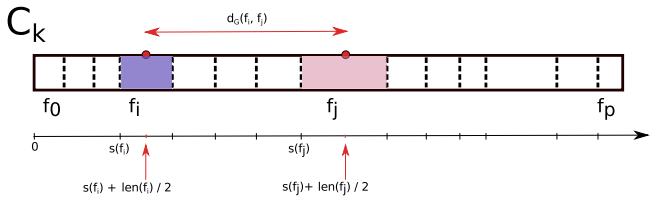


Figure 4.4: Distance between two fragments

- G is the set of all possible genomes $\mathscr{G}$ made of those n restriction fragments. The cardinal of G is infinite since we do allow more than one instance of a given fragment. If we would not allow fragments to occur more than once, $card(\mathsf{G})$ would still be an overwhelming number. $card(\mathsf{G}) > \text{Bell}(n)$[6].

- $\mathscr{P}_\lambda$, $\mathscr{B}_{N,p}$ respectively as the Poisson distribution which parameter is λ and the binomial distribution with N number of trials and a success probability equal to p.

- I embedded :

  –

  all the prior knowledge we have about the sample analysed. as a prior polymer model allowing us to predict the expected number of contacts between two restriction fragments separated by a given distance.

4.3.2.1 Bayes formula

The way to compute the conditional probability $Pr(\mathscr{G}|D, I)$ is given by the Bayes rule. Bayes' theorem states that:

$$Pr(\mathscr{G}|D, I) = \frac{Pr(D|\mathscr{G}, I) \times Pr(\mathscr{G}|I)}{Pr(D|I)}$$

Or,

$$\underbrace{Pr(\mathscr{G}|D, I)}_{\text{Posterior probability}} \propto \underbrace{Pr(D|\mathscr{G}, I)}_{\text{Likelihood of the data}} \times \underbrace{Pr(\mathscr{G}|I)}_{\text{Prior probability}} \tag{4.2}$$

So the initial problem has been translated into the search of the likelihood of the data given a set of chromosomes, $Pr(D|\mathscr{G}, I)$, and the prior probability $Pr(\mathscr{G}|I)$.

4.3.3 Modeling 3C data

In this section we will first give a brief overview of the structural and spatial organization of chromosomes. Then we will describe the most common polymer physics model and their limitations before introducing the simple analytic model that we will use in our probabilistic model.

4.3.3.1 Polymer physics

Chromosome packing The genetic data coding the development and functional organization of all known living organisms is encrypted in DNA. This molecule is a long polymer chain whose repeated monomers are nucleotids. In eukaryote organisms the majority of the DNA's genetic information is packed in discrete entities called chromosomes, while in procaryotes, this information is organized in circular conformations that are also known as plasmid. We will shortly describe the different levels of DNA string folding in eukaryotes organisms.

[6] Bell(n) is the number of partitions of a set of size n

Chromatin The double strand helix does not reside completely elongated inside the cell nucleus. Instead, it is warped around eight blocks of protein cores called histones. These spools of 150 bp DNA are called nucleosomes. Adjacent nucleosomes are linked to each other by 50 bp of DNA thus forming the "10nm" chromatin fiber (Kornberg, 1974 [71]) also known as the "beads on a string" structure.

Heterochromatin Chromatin can adopt an even more compact structure called heterochromatine. This super coiled DNA structure has been described as the "30nm" chromatin fiber. However, recent studies have given solid evidence to doubt the existence of this structure (Fussner et al., 2011 [48]; Eltsov et al. [40], 2008; Maeshima et Eltsov, 2008 [83]). Instead, chromatin might exist at different local compaction levels.

Spatial functional organization The study of the local and global structure of the chromatin fiber, especially during the interphase, is an active field of research. Remodeling chromatin plays a key role in many biological processes. For instance, chromatin methylation triggers heterochromatin formation which tends to silence genes embedded in these areas. Conversely, chromatin demethylation allows the decompaction of DNA which ensures a better access of the transcription machinery to the genes.
Another key aspect of chromatin high scale organization is the formation of intra chromosomal loops: close spatial proximity of the enhancer and promoter triggers the transcription for some genes (Kadauke et Blobel, 2009 [66]).

This very brief description of chromosomal structure and organization gives us insight into the non exhaustive list of parameters that can modify the probability of contact between two loci in a cell nucleus:

- the respective local levels of compaction of the two loci,
- the biological processes linked to their locations.
- the position of the cell in its cycle.

Computational models The modeling of polymers is a very prolific area of research, both in mathematics and physics. Hence, during the last thirty years many computational models have been developed in order to explain and predict the dynamic and static characteristics of nucleic acid polymers. Before going further we would like to define some concepts that are needed to understand the basics of polymer modeling:

- Kuhn length. A linear polymer can be seen as an idealized chain of N segments of length l. The physical quantity l is also referred to as Kuhn length (Rubinstein and Colby, 2003 [119]). The higher l, the higher the stiffness of the polymer.
- Monte Carlo simulation. This class of algorithms allows us to compute statistics of very complex systems for which the analytic equations are too difficult to solve. It relies on heavy usage of random number generation. In polymer modeling, people usually generate many random conformations of the chain and compute statistics on samples that satisfy some predefined constraints.
- Equilibrium. A polymer is said to be at equilibrium when the average polymer size, or the average polymer dynamics, do not change over time (Zimmer and Rosa, 2013).
- Average square end-to-end distance, $< R2(L) >$. As a single physical object, a polymer chain is characterized by its overall average size, which is often expressed as a function of its contour (curvilinear) length, L.

The average square end-to-end distance is defined as the square distance between the chain ends, averaged over all possible conformations the chain can assume in space due to random fluctuations.

- Freely Joint Chain model. The FJC model is the simplest polymer model. In this model, one assumes that Kuhn segments can randomly orient in any direction, independent of the orientation taken by the other bonds (Zimmer and Rosa, 2013).

Most of the computational models are based on complicated computer algorithms which simulate the dynamics of chromosomes. One of the main challenges they try to tackle is to predict and confirm experimental observations such as the formation of intra chromosomal loops ('random-walk/giant-loop' model, Sachs et al, 1995 [121]). Another controversial question they address is whether or not the chromosomes reach a state of equilibrium. It is well known that after mitosis, chromosomes are no longer compact but decondensate so that the transcription machinery can access the genes. Grosberg and his collegues (1988 [59], 1993 [58]) were the first to argue that chromosomes could exist in an "out of equilibrium" state they called "crumpled globule" (Zimmer and Rosa, 2013). Recently, this hypothesis was confirmed by experimental observations obtained by HiC (Lieberman-Aiden et al., 2009 [78]).
Generally, all these models require expert knowledge of the chromosomes' environment and characteristics (in order to setup the simulations), intense computing, followed by Monte Carlo simulations in order to extract both dynamic and static characteristics of the polymers. Since we do not know much about the studied genome we decided to adopt a very simple analytic model to describe the probability of contacts between two loci.

4.3.3.2 A simple analytic model

The probability of interaction between two proteins that are bound on the same chromatin fiber has been described analytically by Rippe [116]. The value corresponds to the local concentration j_M in moles per liter of one binding site in the proximity of another. Chromatin fiber is considered as a freely jointed chain, an idealized chain of N segments of length l. l Is also referred to as Kuhn length.
For a circular DNA fiber, j_M is given by:

$$j_M(n) = 0.53 \times \left(n - \frac{n^2}{N}\right)^{-\frac{3}{2}} \times \exp\left(\frac{d-2}{n - \frac{n^2}{N} + d}\right) \times l^{-3} \tag{4.3}$$

where n is the distance expressed in Kuhn segments along the chain between two sites. To turn n into a genomic distance (m) it is necessary to define L_M, the length per monomer unit(nm/kb) and λ_{Kuhn}, the Kuhn length. Thus:

$$n = \frac{m \times L_M}{l} \tag{4.4}$$

L_M and l define the stiffness of the fiber.

The first part of the equation describes the behavior expected for an ideal chain. The $-\frac{3}{2}$ value of the power corresponds to the theoretical decay computed for polymers at equilibrium state.

The second part of the equation lowers the influence of the FJC model at short distances. The parameter d regulates the contribution of this expression.

For linear polymers, the equation corresponds to the following expression by setting $N = \infty$:

$$j_M(n) = 0.53 \times n^{-\frac{3}{2}} \times \exp\left(\frac{d-2}{n+d}\right) \times l^{-3} \tag{4.5}$$

Experimental ranges of the values of the parameters L_m, l are summarized in table 4.1 (see Rippe, 2001 [116]).

Nucleic acid chain	Length L_m of monomer unit	Kuhn length l (nm)	Monomer per Kuhn length
Single chromatin fiber	8.6 nm/kb	60	7 kb
Chromatin fiber	9.6 np/kb	137-440	14-46 kb
Metaphase chromosome	34 nm/Mb	300-5400	9-60 Mb

Table 4.1: Length and flexibility of nucleic acid chains (Rippe, 2001, [116])

The first hypothesis we form is that contact frequencies are equal, up to a scale factor A, to the local concentration value, j_M, described above. Let $j_c(n)$ be the expected number of contacts between two loci separated by n Kuhn lengths. We have:

$$\begin{aligned} j_c(n, A) &= A * j_M(n) \\ &= A * 0.53 \times \left(n - \frac{n^2}{N}\right)^{-\frac{3}{2}} \times \exp\left(\frac{d-2}{n - \frac{n^2}{N} + d}\right) \times l^{-3} \end{aligned} \tag{4.6}$$

Now, if we inject equation 4.4 into equation 4.6 we can express the number of contacts as a function of the genomic distance s as follows:

$$j_c(s, A) = A * \left(s - \frac{s^2}{N_{bp}}\right)^{-\frac{3}{2}} \times \exp\left(\frac{d-2}{s - \frac{s^2}{N_{bp}} + d}\right) \tag{4.7}$$

where N_{bp} is the total length of the circular chromosome. Thanks to this formula the problem of evaluating the compaction and the stiffness has vanished.

However, the equation above is only correct in case of an ideal polymer at equilibrium. In a recent review (Mirny, 2011 [94]) as well as in the original HiC paper (Lieberman et al, 2009 [78]) the expected power law decay has been described both for the equilibrium globule state and the fractal state (figure 4.5). To take into account the full spectrum of states we set the power of the first term of the equation 4.7 as a free parameter. Thus:

$$j_c(s, A, \alpha) = A * \left(s - \frac{s^2}{N_{bp}}\right)^{-\alpha} \times \exp\left(\frac{d-2}{s - \frac{s^2}{N_{bp}} + d}\right) \tag{4.8}$$

The equivalent equation for a linear, non circular polymer becomes:

$$j_c(s, A, \alpha) = A * \left(s^{-\alpha} \times \exp\left(\frac{d-2}{s+d}\right)\right) \tag{4.9}$$

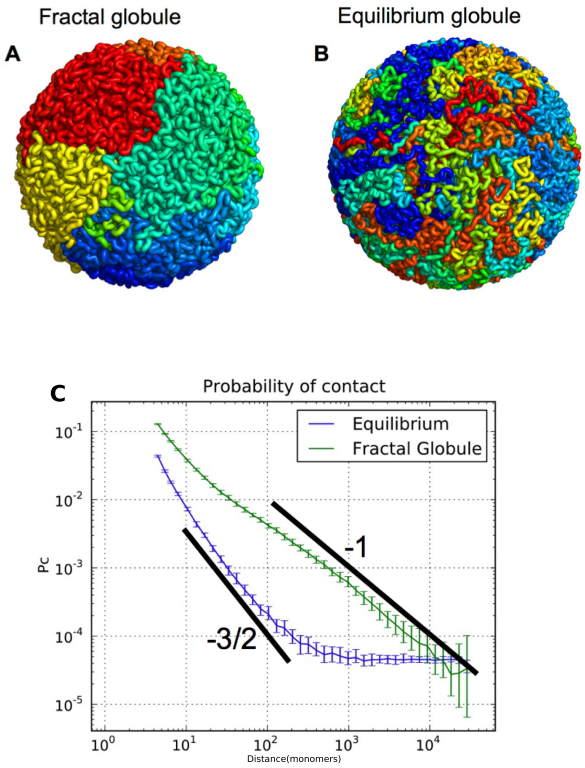


Figure 4.5: Fractal globule and equilibrium globule (Mirny, 2011 [94]). Panels A) and B) display the output of a polymer dynamic simulation based on a classic FJC model. In A) the polymer collapses in a hierarchy of globules. First, small local globules are formed. Then they start to fold on each other, creating the so-called fractal globules. As expected, chromosome territories are naturally created as shown by the spatial colored homogeneous region. Conversely, at equilibrium (panel B) the polymer is much more entangled. After Monte Carlo simulations, the distributions of contact probabilities of the two configuration are estimated: the slope of the distribution curve of the crumpled globule is estimated to be -1 when the equilibrium globule exhibits a power law of -1.5.

Before going further we would like to insist on the fact that the behavior of the parameter d is highly correlated to the living space of the scaling factor A. If $d >= 2$, instead of decreasing the value of j_c at small distances it starts to amplify the output of the first exponential term. Therefore d lives in the interval $[0, 2]$.

4.3.3.3 Experimental biases

In chapter 3 we have detailed the analysis of the raw reads produced by Hi-C or 3C experiments. Because of their inner characteristics, restriction fragments do not have a uniform probability of being captured. We will explain now, why we cannot use directly the normalization procedures discussed previously:

- The methods described in Cournac et al, 2012, [27], and Imakaev et al, 2012, [63] share many similarities in their strategies. As discussed in chapter 3, the produced corrected matrix is no longer a contact matrix but an "equal-visibility" matrix. The likelihood function that we will introduce further relies explicitly on the fact that Hi-C or 3C experiments are counting procedures. Therefore, these two approaches are by definition discarded.

- The normalization developped by Yaffe and Tanay (2010 [149]) relies on an optimization procedure that addresses three streams of b00iases that affect HiC or 3C signals: the local GC content of the restriction fragments, their length and their mappability. This probabilistic approach naturally fits in the bayesian

framework that we are building. However, a strong adaptation will be needed since computing the local GC content of every read requires a fully assembled genome.

For the reasons mentioned above we will adopt the following approach: Because we cannot assume that the initial genome, used to map the 3C or HiC read, is correct, the mappability bias cannot be correctly estimated and will be neglected. We will focus on two major biases: the fragments' length and their local GC content.
A study realized by Benjamini and Speed (2012 [8]) has shown that the GC enrichment signal present in many high throughput sequencing techniques is not reproducible but always follows a uni-modal distribution. Therefore, a very simple way to estimate the GC bias matrix (figure **??**) is to consider it as a two dimensional symmetric gaussian distribution whose mean value c and standard deviations σ_x and σ_y are free parameters:

$$\Gamma_{gc}(f_i, fj) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(\frac{(B_{gc}(f_i) - c)^2}{2\sigma_x^2} + \frac{(B_{gc}(f_j) - c)^2}{2\sigma_y^2}\right) \tag{4.10}$$

In Hi-C or 3C protocols loci that were close to each other are linked by formaldehyde bridges. Then, after the lysis of nuclei by sodium dodecyl sulfate (SDS), DNA is digested by restriction enzymes. To avoid the capture of false contacts (i.e. between random, non spatially close fragments) ligation is performed only on those complexes of DNA and formaldehyde that are diluted and in the solubilized phase of preparation. As explained in Yaffe and Tanay (2010 [149]), fragment ligation is strongly affected by the size of the fragments. Because of their reduced mobility, small fragments will have less opportunities to ligate with others than long fragments. Yaffe and Tanay (2010 [149]) tried to estimate this bias by an optimization procedure which produces an enrichment contact matrix with respect to a discrete grid of fragment lengths. This approach, which is unsupervised and mainly driven by the data, could be improved by using polymer physics: knowing the statistical volumes occupied by two fragments of different lengths, analytic formula $\Gamma_{length}(f_i, f_j)$ could characterize the shared space by two fragments depending on their size. This measure would be directly proportional to the probability of ligation with respect to the fragments length and could directly be injected in our model.
However, a recent study by Gavrilov et al. (2013 [49]) has brought some very interesting insights in the ligation step of the 3C protocol. Thanks to FISH and Electron microscopy experiments they provided the proof that only a small portion of the true 3C signals exists in the diluted and solubilized sample of chromatin and formaldehyde complexes used to form the ligation. In fact, the topological constraints of the chromatin organization remained unchanged within the cell thanks to the formation of meshes of chromatin and formaldehyde. These structures, which are found mainly in non lysed nuclei, have been called Active Chromatin Hub (ACH) by Gavrilov et al. (2013 [49]). In this situation the probability of capturing contacts with respect to the fragment length is not expected to be described as above. The ligation does not happen only within complexes but also between complexes. Because an analytic solution of this problem would be very challenging to obtain, it would be easier to use polymer simulation coupled with Monte Carlo sampling methods to robustly estimate the target distribution.
We define $\Gamma(f_i, f_j)$ as the function correcting the GC content bias and the fragment length bias. Since we consider the influence of these biases as being multiplicative we have:

$$\Gamma(f_i, f_j) = \Gamma_{gc}(f_i, f_j) \times \Gamma_{length}(f_i, f_j) \tag{4.11}$$

Therefore:

$$\begin{aligned} j_c(f_i, f_j, A, \alpha, \Gamma) &= j_c(d_{\mathscr{G}}(f_i, f_j), A, \alpha) * \Gamma(f_i, f_j) \\ &= A * \left(s^{-\alpha} \times \exp\left(\frac{d-2}{s+d}\right)\right) * \Gamma(f_i, f_j) \end{aligned} \tag{4.12}$$

Due to time constraints, so far none of these approaches have been included in our analysis. Therefore, in the rest of this report the function $\Gamma(f_i, f_j)$ will be treated as the Uniform distribution over the space of fragment lengths and GC content.

Now that we have defined the theoretical physical model needed to infer the expected number of contacts between two unique restriction fragments, and the way experimental biases should be handled, we are very close to give an explicit formula of the posterior probability of a genome.

4.3.3.4 Expected contacts

The equation 4.12 gives the expected 3C or HiC signal between two unique fragments that are located on the same linear contig.

Expected cis contacts The following expression summarizes the previous results for two fragments that are located on the same contig $\mathscr{C}_k$.

$$j_c(f_{\varphi_k^i}, f_{\varphi_k^j}, A, \alpha, \Gamma) = \begin{cases} A * s^{-\alpha} \times \exp\left(\frac{d-2}{s+d}\right) * \Gamma(f_{\varphi_k^i}, f_{\varphi_k^j}), & \text{if } \mathscr{C}_k \text{is linear} \\ A * \left(s - \frac{s^2}{\text{len}(\mathscr{C}_k)}\right)^{-\alpha} \times \exp\left(\frac{d-2}{\left(s - \frac{s^2}{\text{len}(\mathscr{C}_k)}\right) + d}\right) * \Gamma(f_{\varphi_k^i}, f_{\varphi_k^j}), & \text{otherwise} \end{cases} \tag{4.13}$$

Expected trans contacts The number of contacts between two fragments that are located on two different contigs is a constant denoted by δ_{nuc}. This parameter is perhaps the one for which we have the least prior knowledge. The spatial organization of a genome is specific to a given cell type and a given organism. For instance, in yeast cerevisiae it is known that centromeres cluster around the spindle pole body. Consequently, for every 3C experiment that is performed on this organism there is a high inter chromosomal contact signal detected for every centromere pair. A polymer model without this information could not predict this specific signal, and this is why HiC or 3C experiments are so powerful.

It is very important to notice that the curve described in equation 4.13 is monotonically decreasing. Since intra chromosomal signals cannot be weaker than inter chromosomal contacts, δ_{nuc} defines intrinsically the limit from which it is impossible to distinguish cis from trans contacts. Lets Δ_τ be this limit. We have:

$$\delta_{nuc} = A * \Delta_\tau^{-\alpha} \times \exp\left(\frac{d-2}{\Delta_\tau + d}\right) * \Gamma(f_i, fj)$$

Let J_c be the expected number of contacts between two fragments in a given genome $\mathscr{G}$. We have:

$$J_c(f_{\varphi_k^i}, f_{\varphi_k^j}, A, \alpha, \Gamma) = \begin{cases} j_c(f_{\varphi_k^i}, f_{\varphi_k^j}, A, \alpha, \Gamma), & \text{if } k = l \text{ and if } d_{\mathscr{G}}(f_{\varphi_k^i}, f_{\varphi_l^j}) < \Delta\tau \\ \delta_{nuc}, & \text{otherwise} \end{cases} \tag{4.14}$$

Expected contacts between repeated fragments Since restriction fragments can have multiple copies we still have to give an explicit formula that takes into account this parameter. When a fragment is duplicated it collects the sum of all the contacts captured by its aliases. The expected number of contacts between two initial fragments f_i and f_j is given by:

$$E(f_i, f_j, A, \alpha, \Gamma) = \sum_{\varphi_i \in \Phi_i} \sum_{\varphi_j \in \Phi_j} J_c(f_{\varphi_k^i}, f_{\varphi_k^j}, A, \alpha, \Gamma) \tag{4.15}$$

4.3.3.5 Relevance of the initial genome $\mathscr{G}_0$

A critical process in our method is the initial mapping of the raw reads on the initial genome $\mathscr{G}_0$. The restriction map of $\mathscr{G}$ defines explicitly F the set of all the available restriction fragments. Therefore any structural estimation of the genome relies on F. In order to illustrate this idea we will briefly describe to problematic situations:

- Lets consider the situation where the genome of the organism on which the 3C experiment is driven, is perfectly assembled. We define $\mathscr{G}_0$ as the set which contains all the chromosome except one. Therefore, a significant part of the reads will not be mapped and from the deficient restriction map created it will be impossible to infer the structure of the missing chromosome.
- We can imagine an extremely worse situation where $\mathscr{G}_0$ is the null genome. We have $\mathscr{G}_0 = \emptyset$ and $F = \emptyset$. The first consequence of this initialization is that any reads are mapped and therefore no contacts are detected. The second consequence is that no genome can be estimated since F is the empty set.

In order to reflect theses issues in the method we are developing we define the mappability score, $S_{map}(R, \mathscr{G}_0)$, as follow:

$$S_{map}(R) = \begin{cases} \dfrac{\text{card}\left(R_m^0\right)}{\text{card}(R)}, & \text{if card}(R) > 0 \\ 1, & \text{otherwise} \end{cases} \tag{4.16}$$

We are now able to explicit the computation of the likelihood of the data.

4.3.3.6 Full posterior probability

From the beginning our goal is to give a probabilistic score to a given genome with respect to our current knowledge and the observed 3C or HiC data. In the bayesian framework, this score is called posterior probability.

Nuisance parameters In order to compute the expected number of contacts we have defined many auxiliary variables that we are going to summarize here:

- The first parameter of every 3C experiment is the genome used to map the raw reads R, $\mathscr{G}_0$. From the initial genome are built the restriction fragments necessary to build our estimates.
- The polymer contact model J_c depends on many variables in order to generate estimated contacts between two fragments: the scale factor A, the slope of the power law α, the maximum discernible intra chromosomal distance Δ_τ, δ_{nuc} the expected number of trans contacts and the low distance parameter d.
- The bias correction model Γ relies also on auxiliary parameters, but for the reasons mentioned above we will not add them to this summary.

Let ξ be the set of all the auxiliary parameters:

$$\xi = \{\mathscr{G}_0, A, \alpha, \Delta_\tau, \delta\text{nuc}, d\}$$

In Bayesian inference these parameters, which are called nuisance parameters, are estimated in the same fashion as the structure itself (Rieping Wolfgang and Habeck Michael and Nilges Michael, 2005, [115]). Therefore the posterior probability relation (equation 4.2)becomes

$$\underbrace{Pr(\mathscr{G}, \xi|R, I)}_{\text{Posterior probability}} \propto \underbrace{Pr(R|\mathscr{G}, \xi, I)}_{\text{Likelihood of the data}} \times \underbrace{Pr(\mathscr{G}, \xi|I)}_{\text{Prior probability}} \tag{4.17}$$

The Bayes theorem us allows to compute this value by reversing the problem: up to a normalizing factor this value is equal to $Pr(R|\mathscr{G}, \xi, I)$, the likelihood of the data given a genome $\mathscr{G}$, nuisance parameters ξ and a model I.

A Poisson process Let us explain in plain English what the likelihood is. 3C or HiC experiments produce a set of reads R which are mapped on a referenced genome G_0. These mapped reads, R^0_m, are used to fill up a contig wide contact matrix D where each entry $D[i, j]$ corresponds to the number of paired reads captured by the procedure between the restriction fragment f_i and the restriction fragment f_j. The likelihood of $D[i, j]$ given a model I, nuisance parameters ξ, and a genome $\mathscr{G}$ is naturally the probability of counting $D[i, j]$ contacts between these fragments knowing that, thanks to our model and our estimated genome, we expect $E[i, j]$ contacts. This probability corresponds exactly to the definition of the Poisson random variable.
Therefore we have:

$$\begin{aligned} P(D[i,j]|\mathscr{G}, \xi, I) &= \mathscr{P}_{E[i,j]}\left(D[i,j]\right) \\ &= \exp\left(-E[i,j]\right) \times \frac{E[i,j]^{D[i,j]}}{D[i,j]!} \end{aligned} \tag{4.18}$$

The capture of contacts between every pair of fragments is a set of independent processes, thus the likelihood of the observed contact matrix is given by:

$$\begin{aligned} P(D|\mathscr{G}, \xi, I) &= \prod_{i>j} P(D[i,j]|\mathscr{G}, I) \\ &= \prod_{i>j} \exp\left(-E[i,j]\right) \times \frac{E[i,j]^{D[i,j]}}{D[i,j]!} \end{aligned} \tag{4.19}$$

The likelihood of the observed matrix data is performed on a partition of the raw reads. This split of the data set is performed implicitly by the initial genome. In order to retrieve the full likelihood of the raw data, it suffices to take into account the weight of the raw data which has served to fill D. The mappability score allows to quantify the ratio of reads aligned on $\mathscr{G}_0$. Therefore, we have:

$$P(R|\mathscr{G}, \xi, I) = P(D|\mathscr{G}, \xi, I) \times S_{map}(R) \tag{4.20}$$

4.3.4 Discussion

The Bayesian framework provides an elegant way to assign a probabilistic score to a genome. However some details must be taken into account:

- First, the score relies explicitly on the initial mapping of the raw 3C or HiC reads on $\mathscr{G}_0$. Therefore, even if more information are available in the raw materials, since our knowledge is limited to the contact matrix, we will no be able, without redefining our model, to infer what is happening at the sub fragment scale.
- Then at the fragment scale, the probabilistic score we have defined previously can not discriminate between two genome whose contigs contains the same fragments, at the same positions but with different orientations. However, we will see in the following section how it is possible to bypass this limitation.
- The last thing we have to take care of is the meaning of missing data. Typically, when no contacts at all are detected between two fragments, does that mean that the experiment failed to capture these information, or is it a true biological signal? Our probabilistic score handles this situation "out of the box" and it should be noticed that it gives us a strong hint on the relevance of the initial genome used to map the raw reads. As a matter of fact, many aligner software will discard reads which map at multiple locations on a given genome. Therefore if a restriction fragment falls in such a region, its global coverage will be close to zero. Both the mappability score and the expected contact matrix will penalize this initial setup.

Now that we have defined a proper way to compute the posterior probability of a genome and the auxiliary parameters we have just introduced it remains to explore this conditional distribution. This task is very difficult to achieve and very sophisticated methods have been developed in order to tackle this problem. In the next section we will present the strategies we have implemented before evaluating their performance and accuracy.

4.4 Markov Chain Monte Carlo algorithms

Our overall goal in this section will be to describe methods that allow to find the maximum a posteriori (MAP) of the posterior distribution defined previously. First we will introduce the basics of Monte Carlo integration. Then we will present and discuss algorithm we have developed to infer the correct genomic structure.

4.4.1 Principles

4.4.1.1 Monte Carlo

The original Monte Carlo approach is due to to Stanislaw Marcin Ulam who produced a great range of theoretical and applied results both in mathematics and physics. In order to compute complicated integrals for which analytic formulas are difficult or intractable to obtain, he proposed to use the increasing capability of computers to generate random numbers.

As explained in Andrieu et al , 2003 [3], let us consider the following situation, where we want to compute:

$$F = \int_a^b f(x)dx$$

Let $p(x)$ be a probability density function defined over the interval (a, b) and $\{x^i\}_{i=1}^N$ an i.i.d set of samples drawn from $p(x)$. We can approximate this distribution with the following empirical function:

$$p_N(x)\frac{1}{N}\sum_{i=1}^{N}\delta_{x^i}$$

Then if we can write $f(x)$ as the product of a function $h(x)$ and the density $p(x)$ it is possible to approximate F with the following expression:

$$\widehat{F_N} = \frac{1}{N}\sum_{i=1}^{N} h(x^i)$$

By the strong law of large numbers, it can be shown that $\widehat{F_N}$ converge almost surely (a.s) to F:

$$\widehat{F_N} = \frac{1}{N}\sum_{i=1}^{N} h(x^i) \xrightarrow[N\longrightarrow\infty]{a.s.} F = \int_a^b h(x)p(x)dx = \int_a^b f(x)dx$$

This strategy is the main core of the Monte Carlo principle. One of the main advantages of this method over deterministic approaches is that the points used for the integration concentrate in region of high probability. An extensive literature is available on the way to generate efficiently the samples used to perform the integration. In most of the situations it is impossible to generate random samples directly from $p(x)$. This is why a more sophisticated class of algorithm called Markov Chain Monte Carlo (MCMC) have been developed.

4.4.1.2 Markov Chain Monte Carlo

The overall goal of MCMC algorithms is to generate a Markov chain which will explore the integration space. The chain is designed such that it will spend more time in high probability areas. Before going further we will define roughly what a Markov Chain is.

Markov Chain Let $\mathscr{X} = \{s_1, s_2, \ldots, s_m\}$ denote a finite state space, and X_t the value of a random variable at time t over $\mathscr{X}$. X is said to be a Markov process if it satisfies the following Markov property:

$$Pr\left(X_{t+1} = s_j | X_0 = s_k, \ldots, X_t = s_i\right) = Pr\left(X_{t+1} = s_j | X_t = s_i\right)$$

This property guarantee that the next state of the process depends only on its current position. A Markov chain is, by definition, the sequence of random variable $(X_0, \ldots, X_n)$ produced by a Markov process.
Let T denote the transition probabilities matrix of the Markov chain X. T returns the probability that a process at state $X_t = s_i$, moves at state $X_{t+1} = s_j$. It is common in the literature to use the following notation to describe T:

$$T(s_i, s_j) = Pr(X_{t+1}|X_t) = Pr(X_t \longrightarrow X_{t+1})$$

Note that each row of T sums to one:

$$\sum_{s_j} T(s_i, s_j) = 1$$

If T remains constant for all t the chain is said to be homogeneous. The probability $\pi_i(t)$ of the chain to be at state i at time t is given by:

$$Pr(X_t = s_i) = \pi_i(t)$$

and by the Chapman-Kolmogorov theorem we have:

$$\begin{aligned}\pi_i(t) &= \sum_k Pr(X_t = s_i | X_{t-1} = s_k) \times Pr(X_{t-1} = s_k) \\ &= \sum_k Pr(X_{t-1} \longrightarrow X_t) \times \pi_k(t-1) \\ &= \sum_k T(k,i) \times \pi_k(t-1)\end{aligned}$$

Let $\pi(t)$ be the row vector of the state space probabilities at iteration t.

$$\begin{aligned}\pi(t) &= (\pi_0(t), \ldots, \pi_m(t)) \\ &= \pi(t-1)T\end{aligned}$$

In order to introduce intuitively how does the Markov process work we will describe some properties of the chain thanks to a simple example.

A simple example Let us consider the state space $\mathscr{X} = \{\text{On time (O)}, \text{Delayed (D)}, \text{Canceled (C)}\}$ and the presence of Professor Cuthbert Calculus[7] at his weekly student appointment as the realization of a Markov process. The transition probability matrix T is defined as follow:

$$T = \begin{array}{c} \\ P(.|O) \\ P(.|D) \\ P(.|C) \end{array} \begin{array}{c} \begin{array}{ccc} O & D & C \end{array} \\ \begin{pmatrix} 0.5 & 0.25 & 0.25 \\ 0.5 & 0 & 0.5 \\ 0.25 & 0.25 & 0.5 \end{pmatrix} \end{array}$$

where, for example, the first entry of the matrix ((O,O)) is the probability of the professor being "On time" at the next meeting given that he is "On time" at the current appointment.
If we assume that, today, Cuthbert delayed his meeting ($\pi(0) = (0,1,0)$) , what is the probability of his presence during the next weeks? We have:

$$\begin{aligned}\pi(1) &= \pi(0)T = (0.5, 0, 0.5) \\ \pi(2) &= \pi(0)T^2 = (0.375, 0.25, 0.375) \\ \pi(3) &= \pi(0)T^{10} = (0.40625, 0.1875, 0.40625) \\ \pi(10) &= \pi(0)T^{10} = (0.4, 0.2, 0.4)\end{aligned}$$

If the professeur was on time at hist current appointment ($\pi(0) = (1,0,0)$) we would have:

$$\begin{aligned}\pi(1) &= \pi(0)T = (0.5, 0.25, 0.25) \\ \pi(2) &= \pi(0)T^2 = (0.4375, 0.1875, 0.375) \\ \pi(3) &= \pi(0)T^3 = (0.40625, 0.2031, 0.390625) \\ \pi(10) &= \pi(0)T^{10} = (0.4, 0.2, 0.4)\end{aligned}$$

[7] also known as Professeur Tryphon Tournesol

The conclusion of this numerical tests, is that after a few iterations, the chain reaches a stationary distribution: the presence of the professor at the weekly meeting is independent of the starting value.
The stationary π^* distribution verifies:

$$\pi* = \pi^* T$$

Two necessary conditions for the chain to have a stationary distribution are:

- The chain is irreducible. This characteristic means that any state of the space can be reach by the chain with a positive probability.
- The chain is aperiodic: it cannot get trapped in cycles.

Detailed babance condition In order to guarantee the existence of a unique stationary distribution $\pi*$ a sufficient condition, also referred as the detailed balance condition, is that, for all i and j:

$$\pi_i^* T(s_i, s_j) = \pi_j^* T(s_j, s_i)$$

or,

$$\pi_i^* Pr(X_{t+1} = s_j | X_t = s_i) = \pi_j^* Pr(X_{t+1} = s_i | X_t = s_j)$$

It is possible to extend, the previous results to continuous state spaces. In such a situation the transition matrix becomes an integral kernel K and at equilibrium the stationary distribution satisfies:

$$\pi^*(y) = \int \pi^*(x) K(x, y) dy$$

4.4.1.3 The Metropolis-Hastings algorithm

As we explained before, the basic idea of the MCMC methods is to tackle the difficult task of generating the i.i.d samples from the distribution $p(x)$ by using an irreducible and aperiodic Markov chain.
Let $p(x)$ be a distribution known up to a normalizing constant. This is a very common situation in Bayesian statistic: the posterior probability is proportional to the likelihood of the data. Therefore even if the likelihood is analytically defined, the normalizing factor K, necessary to obtain a proper distribution, remains extremely difficult to compute. Let us write $p(x)$ as follow:

$$p(x) = \frac{1}{K} f(x)$$

The Metropolis-Hastings algorithm (figure 3) generates samples from the invariant distribution $p(x)$ by generating a Markov Chain as explained bellow:

1. Start with an initial value x_0 such that $f(x_0) > 0$
2. Given the current point x sample a candidate value x^* according to a jumping or proposal distribution, $q(x^*|x)$ (written also as $q(x \longrightarrow x^*)$). There is no restriction on the choice of the distribution q.
3. Compute an acceptance probability, $\mathscr{A}(x, x^*)$, defined as follow:

$$\mathscr{A}(x, x^*) = \min\left(1, \frac{p(x^*)q(x|x^*)}{p(x)q(x^*|x)}\right) = \min\left(1, \frac{f(x^*)q(x|x^*))}{f(x)q(x^*|x)}\right)$$

Note that the normalizing constant K naturally vanishes from the equation.

4. Accept the candidate point x^* with probability $\mathscr{A}(x, x^*)$ otherwise remain at x.

Algorithm 3 Metropolis Hasting algorithm

Require: $x_t = x_0, f(x_0) > 0$
Ensure: A markov chain x
  for $t = 0$ to N **do**
    sample u $\sim \mathscr{U}_{[0,1]}$
    sample $x^* \sim q(x^*|x_t)$
    compute $r = A(x_t, x^*)$
    if $r >= u$ **then**
      $x_{t+1} = x^*$
    else
      $x_{t+1} = x_t$
    end if
  end for

To prove that the MH algorithm truly samples from the stationary distribution $p(x)$ it suffices to verify that the detailed balance conditions holds. The demonstration of this property is detailed in Walsh, 2004 [142].

3C genome sampling Now let us give some details on how to use MCMC in the framework of 3C or HiC genome inference.

We have defined in equation 4.17, the posterior probability of an infered genome and some nuisance parameters given 3C data and some prior knowledge of polymer physics: $Pr(\mathscr{G})$. From now on we will consider the prior probability as a flat distribution. Therefore $p(\mathscr{G}, \xi)$, the posterior probability and $f(R, \mathscr{G}, \xi, I)$, the likelihood of the data, are equal up to a normalizing constant K. We have:

$$\underbrace{Pr(\mathscr{G}, \xi|R, I)}_{\text{Posterior probability}} \quad \propto \quad \underbrace{Pr(R|\mathscr{G}, \xi, I)}_{\text{Likelihood of the data}} \tag{4.21}$$

and,

$$\begin{aligned} p(\mathscr{G}, \xi) &= Pr(\mathscr{G}, \xi|R, I) \\ &= \frac{1}{K} \times Pr(R|\mathscr{G}, \xi, I) \\ &= \frac{1}{K} \times f(R, \mathscr{G}, \xi, I) \end{aligned} \tag{4.22}$$

As we said previously our goal is to generate samples from $p(\mathscr{G}, \xi)$. The MH algorithm provides a powerful way to explore this distribution. A very sensitive part of the MCMC algorithm is the design of the proposal distribution.

4.4.2 Implementation

Before going deeper into the description of the algorithmic methods we have developed, we will give some insight about the way the data are stored in memory and the computing framework we will use to implement our methods.

4.4.2.1 Data structure

A typical 3C or HiC experiments generates millions of reads which will first be mapped on a reference genome and used to fill up a contig wide contact matrix. The size of this matrix depends on the restriction enzyme used for the experiment. The higher the cutting frequency of the enzyme and the bigger will be the matrix. This is one of the reasons why we adopted a common approach in image processing to store these information.

Pyramids of contact matrices A pyramid of matrices is a multiscale representation of the contig wide contact data (figure 4.6). At each level of the pyramid, initial adjacent fragments are concatenated to create virtual fragments. In figure 4.6 a compression factor of three is used to create the pyramid. At the moment,

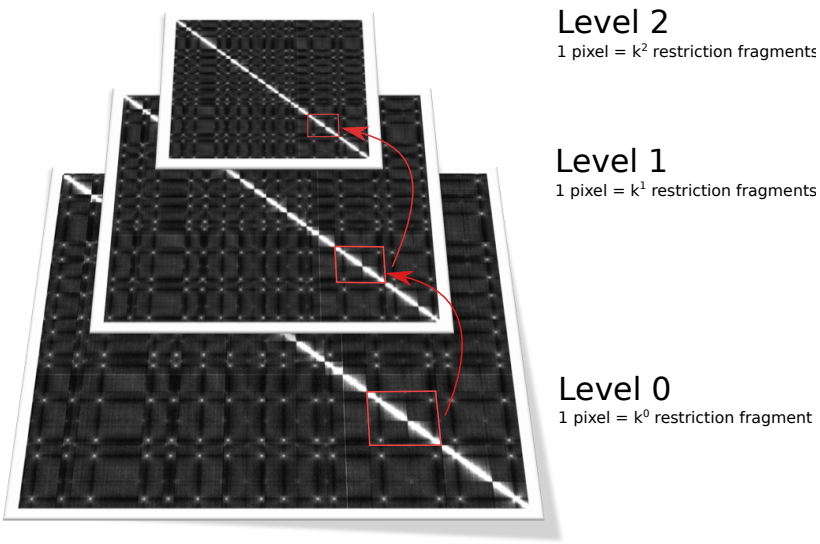


Figure 4.6: 3 levels HiC pyramid. The first level of the pyramid corresponds to the raw data. The pixel D_0[i,j] collects the number of paired reads sequenced mapped on the fragment f_i and the fragment f_j. At the upper level, a pixel $D_1[i,j]$ corresponds to number of contacts detected between two pools of k initial fragments $\{f_i^0, \ldots, f_i^k\}$ and $\{f_j^0, \ldots, f_j^k\}$. All the upper levels of the pyramid are constructed in the same fashion.

each level of the pyramid is stored in a dense[8]manner.

There are at least two reason which legitimate the use of this representation:

1. Memory limitation. The frequent cutter enzyme DpnII generates 35914 restriction fragments on the yeast s.cerevisiea genome. Therefore, the full matrix weight, 2.5 GB in memory. A ten times bigger genome with the same frequency of restriction site will be 250 GB. The pyramid representation allows to perform low resolution analysis at no expense.

2. Fragments orientation. As we mentioned earlier in this chapter, the posterior probability can not distinguish between two genomic structures which differ only by the orientation of the initial fragments. By concatenating initial fragments we are able to virtually polarize the contact data (figure 4.7) to the expense of a loss in resolution. We are aware that at the moment, this operation may bias our analysis. Since we

[8]Because of the high sparsity nature of the data it will be mandatory to adopt a sparse representation of the contact matrices. Typically a 500 000 fragments matrix weight almost 512 Gb in memory...

can not break the content of the virtual fragments, this data trick add a persistence of the initial assembly in all our estimations. For instance the exact position of breaks or translocation might be affected by the binning.

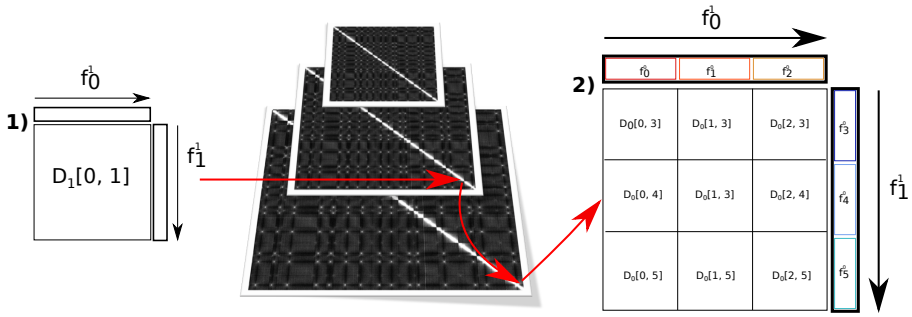


Figure 4.7: Orientation of the fragments. By concatenating restrictions fragments it is possible to orientate the newly created virtual fragments.

Notations As a matter of fact, the ratio k used to build the pyramid, defines at each level l:

- a new restriction map F_l, associated to a restriction enzyme whose cutting frequency is k times lower.
- an initial genome $\mathscr{G}_0^l$ whose atomic elements are the binned restriction fragments made of F_0
- an observed matrix D_l as displayed in figure 4.6.

For simplicity of notation, from now on, we will assume that the genome $\mathscr{G}_0$ refers to the genome related to a given level $l > 0$ of the pyramid. Moreover, in order to take into account the orientation of the fragments, the likelihood of the structure will always be computed on the level $l - 1$ of the pyramid.

4.4.2.2 Distributed computing

The algorithms we will introduce further in this manuscript rely on intense iterative computing. For instance the determination of the likelihood requires to estimate at each data point (i.e each pixel of a contact matrix) the expected number of contact before comparison with the observed data. As the number of restriction site increases, the number of data point to evaluate grows in a polynomial way. An increase by a factor 10 of the number of initial fragments implies a hundred times more computing operations to perform. If these evaluations were performed sequentially that would imply that the same algorithm would run 100 times slower for a 10 times bigger data set.

The current tendency in modern programming is to distribute the computing task as often as possible. For instance the ISD software package developed by Rieping W. Nilges M. and Habeck M., 2008 [114], to explore the conformation space of protein, makes great use of 50 nodes of a computer cluster. Nowadays it is very common to encounter modern laptop computers with up to 16 CPU cores. Despite this great improvement, these equipment cannot compete with computer clusters.

However with the rapid development of the video game industry a special effort have been done to increase the computing capabilities of Graphic Processing Unit (GPU). These devices are dedicated to compute simultaneously billions of trigonometric operation in order to render very complex 3D environments. Common graphic

cards available in the market possess more than 900 processors. Today it is possible to harness the overwhelming computing power of GPU to make scientific or general purpose computing. Therefore no need of a huge computer cluster to realize high performance computing.
Actually there are two languages available for GPGPU(General Purpose computing on GPU):

- OpenCL is an open framework which allows to write program which can be executed both on GPUs and CPUs.
- CUDA is a proprietary framework (Nvidia) whih allows to program only on GPU.

While CUDA and OpenCL are very similar in their syntax, the available libraries for scientific computing are much more developed in the former. Both CUDA and OpenCL are C based programming language, however it is possible to use them in the very popular framework Python thanks to the libraries developed by Andreas Klockner, 2009, [69]. Roughly, our programs are organized as follows:

- Genomic structures are generated on the GPU.
- Nuisance parameters are generated on the CPU.
- The likelihood is computed on the GPU.
- The main program which handles all the processing unit (CPU or GPU) runs on the CPU.

4.4.3 Proposal Distribution

The posterior distribution we have defined relies on two different kind of object:

- The nuisance parameters: $\xi = \{\mathscr{G}_0, A, \alpha, \Delta_\tau, \delta_{\text{nuc}}, d\}$. For now we will consider the initial genome used to map the raw reads as constant. Therefore the nuisance parameters will be $\xi = \{A, \alpha, \Delta_\tau, \delta_{\text{nuc}}, d\}$.
- The current genome : $\mathscr{G}_t$

This setup, obviously, forces us to define two different jumping schemes: one for the nuisance parameters and one for the estimated genome.

Gibbs Sampling In order to sample the joint distribution of the genome and the parameters of the model, we will use a Gibbs sampler, Geman, Stuart and Geman, Donald, 1984, [50]. In an iterative manner, we will alternatively perform the following operations:

- Start with initial parameters $\xi_0, \mathscr{G}_0$.
- $\xi_{t+1} \backsim Pr(\xi|R, I, \mathscr{G}_t)$. The genome stays constant and we update only the nuisance parameters.
- $\mathscr{G}_{t+1} \backsim Pr(\mathscr{G}|R, I, \xi_{t+1})$. With the updated parameters of the model we sample a new structure.

In the next sections we will describe the sampling strategies adopted for the model parameters, and the genomic structure.

4.4.3.1 Sampling the nuisance parameters

First we have to define the initial parameters ξ_0 .

Initialization ξ_0 is approximated by using simple optimization routines. Based on the initial genome $\mathscr{G}_0$, we compute the frequency of contact as a function of genomic distance averaged across $\mathscr{G}_0$. This process produced two vector of equal length:

- x_{bin}, a vector of genomic distance. The minimum value of x_{bin} is 0 while its maximum value $\widehat{\Delta_\tau}$ is equal to the genomic length of the longest contig in $\mathscr{G}_0$
- y_{obs}, a vector of contact frequency.$y_{obs}(s)$ is the mean number of contacts between fragments separated by s kb.

We also compute the mean number of contacts between two fragments which do not belong to the same contig δ^0_{nuc}. Then we perform a non linear least square optimization routine which minimizes:

$$E_{\widehat{\xi}} = \| y_{obs} - y_{exp} \|^2 \quad = \sum_{s \in x_{bin}} (y_{obs}(s) - j_c(s, \widehat{\xi}))^2$$

where y_{exp} is the expected number of contacts produced by j_c for each genomic distance in x_{bin} given the parameters $\widehat{\xi} = \{A, d, \alpha\}$. The parameters Δ_τ and δ_{nuc} are not estimated by this method. Therefore $\widehat{\xi_0}$ is defined as follow:

$$\widehat{\xi_0} = \{A_0, \alpha_0, d_0\} = \underset{\widehat{\xi}=\{A,\alpha,d\}}{\arg\min} \; E_{\widehat{\xi}}$$

We estimated Δ_τ by using a quasi-Newton method (Broyden algorithm) to find the root of the expression:

$$j_c(\Delta_\tau, \widehat{\xi_0}) - \delta^0_{nuc} = 0$$

Then, the initial nuisance parameters are:

$$\xi_0 = \left\{A_0, \alpha_0, \Delta^0_\tau, \delta^0_{nuc}, d_0\right\}$$

The figure 4.8 summarizes the output of the initialization process.

Proposal distribution Now that the nuisance parameters are initialized it remains to define how we will modify the current vector of nuisance parameters $\xi_t = \{A_t, \alpha_t, d_t, \Delta^t_\tau, \delta_{nuc}\}$. Our strategy cannot rely on independent moves apply to each of the parameters. A modification of any parameter $\theta \in \{A, \alpha, d\}$ has a direct consequence on the value of j_c for a genomic distance of Δ^t_τ. Therefore, for each move applied to a parameter different than Δ_τ, it will be necessary to update the latter. Moreover, since Δ_τ and δ_{nuc} are directly link one to the other it suffices to modify one of them.
The algorithm works as follow:

1. Pick at random $\theta \in \{A, \alpha, \Delta_\tau, d\}$, the parameter to modify.
2. Generate a new candidate θ^* as follow:

$$\theta^* = \theta_t + \epsilon_\theta \text{ , with } \epsilon_\theta \sim \mathscr{N}(0, \sigma_\theta)$$

   where θ_t is the current value of θ and $\mathscr{N}(0, \sigma_\theta)$ a centered normal distribution whose standard deviation is specific to each parameter.

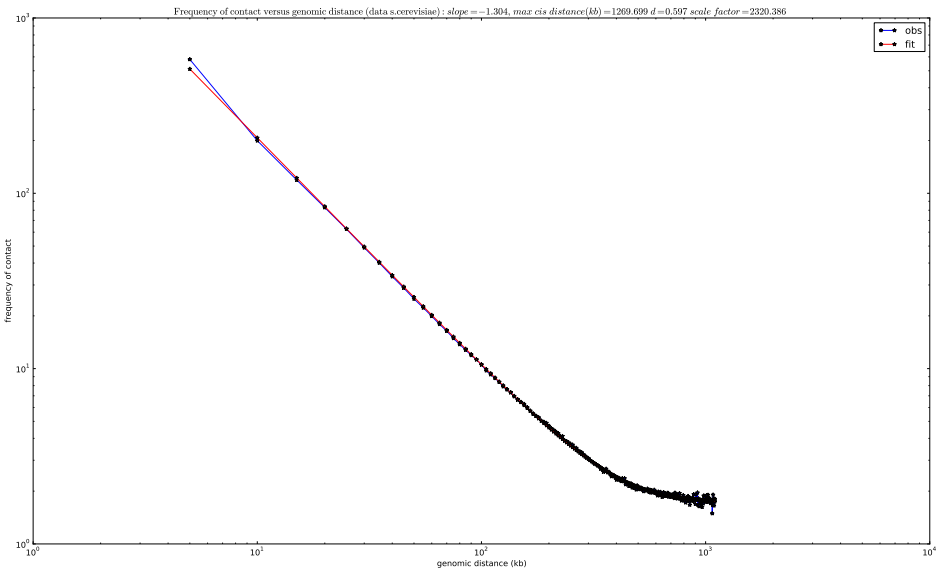


Figure 4.8: Initialization of the polymer model parameters. The figure displays the observed contact frequency (red curve) as a function of genomic distance. The blue curve represents the contact frequency generated by the polymer model previously fitted on the experimental data.

3. All parameters different than Δ_τ modify the expected number of contacts at Δ_τ. In order to keep the model coherent, it is necessary to correct the value of Δ_τ. Therefore:

   - If $\theta \in \{A, \alpha, d\}$, as described previously, we estimate the new value Δ_τ^* by solving the following equation:

     $$j_c(\Delta_\tau^*, \widehat{\xi}^*) - \delta_{nuc} = 0$$

     where $\widehat{\xi}^*$ embedded θ^* and all the other parameters except Δ_τ.

   - Otherwise we have $\Delta^* = \theta^*$ and we set $d_{nuc}^* = j_c(\Delta_\tau^*, \{A_t, \alpha_t, d_t\})$

4. Set ξ^* as the new candidate set of nuisance parameters whose parameters have been updated.

5. Accept ξ^* with the probability:

   $$r = min\left(1, \frac{p(\mathscr{G}_t, \xi^*)}{p(\mathscr{G}_t, \xi_t)}\right)$$

We recall that $p(\mathscr{G}_t, \xi^*) = Pr(\mathscr{G}, \xi | R, I)$ is the posterior probability we want to explore.
Note that since the probability of the moves we apply are symmetric, the acceptance ratio r corresponds to the probability defined in the original Metropolis algorithm (Metropolis, Nicholas and Ulam, Stanislaw, 1949 [92], Metropolis et al, 1953 [91]).
Now that we have described how the sampling of nuisance parameters will be performed we have to define how we will propose new genomes.

4.4.3.2 Genome proposal distribution

A genome is a multidimensional object and in order to build our proposal distribution we have to know how to jump from one genome to the other. We will introduce some very simple mechanisms which mimic common situations that can arise during the cell life.

4.4.3.3 A nature based algorithm

During the cell cycle many events can re organize the linear structure of the genome. As a matter of fact, the linear sequence of DNA is constantly altered and repaired by very sophisticated mechanisms (Ober, Raimund J and Ram, Sripad and Ward, E Sally, 2004 [101]). In yeast Saccharomyces cerevisiae, double strand break (DSB) play a key role in mating type switching (Haber, James E, 1998, [60]). Therefore, in the same way, we will generate new candidate genome $\mathscr{G}_{t+1}$ by performing mutations on a given genome $\mathscr{G}_t$.
The alterations we will compute will depend both on the restriction enzyme used to perform the experiment and the initial genome $\mathscr{G}_0$ on which the raw reads have been mapped. Concretely we will perform re ordering and duplication of the restriction fragments.

Basic operations As a matter of fact, with just five basic operations, it is possible to generate all possible combinations of orientated fragments and thereby all possible genomes given a restriction map, in a finite number of operations:

- $\mathsf{Split}(\mathscr{G}_t, \varphi_k^i, up)$: This operation performs a double strand break at position i on the contig $\mathscr{C}_k$. If $up = 0$, then the break is done on the upstream region of the fragment and otherwise on the downstream region of the fragment. This operation yields two new contigs.
- $\mathsf{Paste}(\mathscr{G}_t, \varphi_k^i, \varphi_l^j)$: This operation performs a ligation between $\mathscr{C}_k$ and $\mathscr{C}_l$ if and only if $i \in \{0, \text{len}(\mathscr{C}_k) - 1\}$ and $l \in \{0, \text{len}(\mathscr{C}_l) - 1\}$. The relative orientation of $f_{\varphi_k^i}$ and $f_{\varphi_l^j}$ in their respective contig is conserved.
- $\mathsf{Duplicate}(\mathscr{G}_t, \varphi_k^i)$: This operation virtually duplicate the initial restriction fragment corresponding to $f_{\varphi_k^i}$.
- $\mathsf{Delete}(\mathscr{G}_t, \varphi_k^i)$: This operation virtually delete the instance of the initial restriction fragment corresponding to $f_{\varphi_k^i}$.
- $\mathsf{Flip}(\mathscr{G}_t, \varphi_k^i)$: This operation switch the orientation of the fragment $f_{\varphi_k^i}$. Since the Paste operation cannot modify the current orientation of a fragment, Flip cannot be seen as a combination of two Split and two Paste.

Each of these mutations comes with it reciprocal operation:

- $\mathsf{Paste}^{-1} = \mathsf{Split}$
- $\mathsf{Duplicate}^{-1} = \mathsf{Delete}$
- $\mathsf{Flip}^{-1} = \mathsf{Flip}$

Any complex modification performed over a genome can be translated into a sequence of these five basic operations.
For simplicity we write the modification $\theta \in \{\mathsf{Paste}, \mathsf{Split}, \mathsf{Duplicate}, \mathsf{Delete}, \mathsf{Flip}\}$ applied to a genome $\mathscr{G}_t$ as follows:

$$\theta_t(x), \text{ where } x \text{ is the set of parameters needed to perform the operation.}$$

A new genome $\mathscr{G}^*$ is generated after the application of a sequence of the previously defined basic operations.

We use this notation to modify $\mathscr{G}_t$:

$$\mathscr{G}^* = \mathsf{Mod}_{\mathscr{G}_t,}(L), \text{ where } L = (\theta_t^0(x_0), \ldots, \theta_t^k(x_k))$$

4.4.3.4 Naive genome sampling

We will describe now how a very simple (too simple) Metropolis Hastings algorithm could be implemented using the genomic moves we have just defined.

Basic Metropolis Hastings sampling The algorithm works as follows:

1. Select at random a move $\theta \in \{\mathsf{Paste}, \mathsf{Split}, \mathsf{Duplicate}, \mathsf{Delete}, \mathsf{Flip}\}$
2. Depending on θ pick at random a set of parameters x_* and let $\theta_t(x_*)$ be the corresponding modification. The probability of choosing $\theta_t(x_*)$ is given by $T(\theta_t(x_*))$. Conversely the probability of choosing the reverse operation $(\theta_t(x_*))^{-1}$ is $T((\theta_t(x_*))^{-1})$
3. Generate a new genome $\mathscr{G}_* = \mathsf{Mod}_{\mathscr{G}_t,}(\theta_t(x_*))$
4. Accept $\mathscr{G}_*$ with the probability:

$$\begin{aligned} r &= min\left(1, \frac{p(\mathscr{G}_*, \xi_t)q(\mathscr{G}_* \longrightarrow \mathscr{G}_t)}{p(\mathscr{G}_t, \xi_t)q(\mathscr{G}_t \longrightarrow \mathscr{G}_*)}\right) \\ &= min\left(1, \frac{p(\mathscr{G}_*, \xi_t)T(\theta_t(x_*))}{p(\mathscr{G}_t, \xi_t)T((\theta_t(x_*))^{-1})}\right) \end{aligned}$$

Limitations Let $\Omega^{\mathscr{G}_t}_{\mathsf{Paste}}$ (resp $\Omega^{\mathscr{G}_0}_{\mathsf{Split}}$) be the set of all possible paste (resp split) operations over a genome $\mathscr{G}_t$. Assuming that $\mathscr{G}_t$ is made of n fragments we have:

- $\text{card}(\Omega^{\mathscr{G}_t}_{paste}) \leqslant \dfrac{n*(n-1)}{2}$
- $\text{card}(\Omega^{\mathscr{G}_t}_{split}) = n$

On a relatively small structures such as the genome of yeast cerevisiae, a frequent cutter restriction enzyme (DpnII for instance) generates tens of thousands of restriction fragments. Therefore with n $\geqslant$ 1000, the set of all possible combined paste operations, $\Omega^{\mathscr{G}_t}_{\mathsf{Paste}}$, is huge [9]. The odds of selecting relevant moves will be very low. Since some modifications require a specific order of basic mutations (for instance translocations) the situation gets even worse. This is why it will be mandatory to better sampling schemes.

4.4.3.5 Image processing sampler

As we explain at the beginning of this chapter, improving the assembly with contact data can be seen as solving a jigsaw puzzle. Visually, one can target relevant modifications and apply them in order to improve the estimated genomic linear structure.

Graphical signature of the mutations We will describe now the graphical signatures produced by simple linear incongruities of the genome. A correcting mutation is associated to each of these patterns.

[9] $\Omega^{\mathscr{G}_t}_P \subset (\Omega^{\mathscr{G}_t}_P)^{N-2}$ and $\text{card}(\Omega^{\mathscr{G}_t}_P) = \dfrac{N!(N-1)!}{2^{N-2}}$

Split If two regions, A and B, have to be split apart, their inter contacts matrix, D_{AB}, possesses relatively less hits than their respective intra contacts matrices. The junction area displays a specific local pattern on the diagonal of D. Typical split patterns are displayed in figure 4.9.

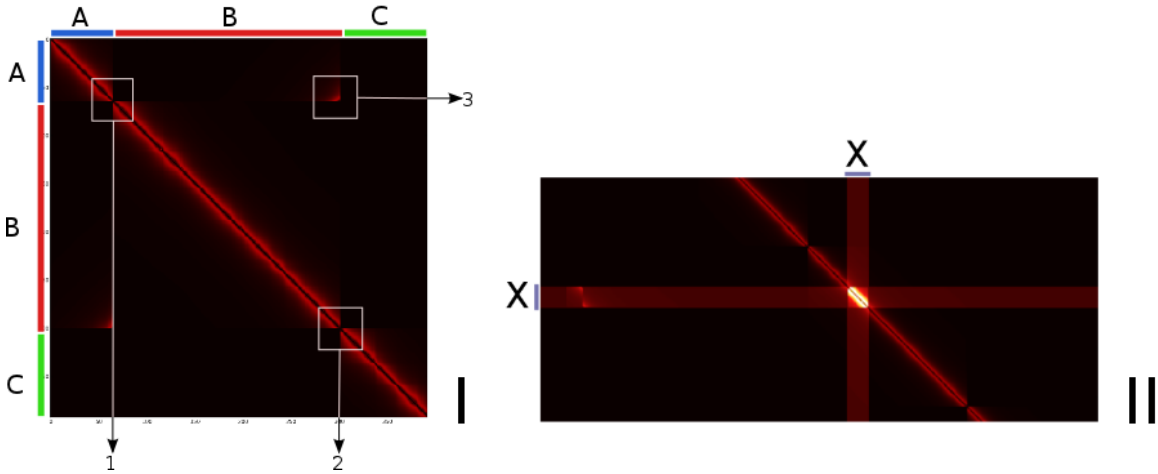


Figure 4.9: Graphical pattern for split and paste. I/ 1) displays a split signal between the right extremity of contig A and the left extremity of contig B. 2) displays a split signal between the right extremity of B and the left extremity of C. 3) displays a paste signal between the right extremity of A and the left extremity of C. II/ X is a duplicated bloc. The horizontal and vertical stripes indicate the enrichment of contacts throughout the whole genome of the fragments embedded in X.

Paste If two regions which have to be pieced together are split by the current linear genomic structure, two specific patterns can be detected in the observed contacts matrix D: a split and a paste signal(respectively 2 and 3 in figure 4.9).

Duplication A duplicated area of the genome will exhibit proportionally more hits than a non duplicated DNA bloc. The graphical signature of a duplication is diplayed in figure 4.9 II.

Probability of a move We will explain here how we will build custom distributions over $\Omega^{\mathscr{G}_t}_{paste}$ and $\Omega^{\mathscr{G}_t}_{split}$. Then we will describe how it is possible to sample duplications in a non uniform way.

Distribution of Paste and Split Both in the Paste and the Split cases, the goal is to detect corners. This is a common task in computer vision and plenty of algorithm have already been developed to perform this operation. The "matching pattern" algorithm allows to define a score and consequently a probability over $\Omega^{\mathscr{G}_t}_{paste}$ and $\Omega^{\mathscr{G}_t}_{split}$.

At each pixel $\mathsf{Match}(D, i, j, H, \mathscr{G}_t)$ evaluates a match criterion for all the patterns $h \in H$ and yield the score of the most likely pattern present. H_p is the set of the four rotated paste pattern. The bank of all these patterns is displayed in figure 4.10.

H_p is the bank of paste patterns and H_s the bank of split patterns. We define the probability of Paste and Split as follows:

$$T^{\mathscr{G}_t}_{\mathsf{Paste}}(i,j) = \frac{\mathsf{Match}(D, i, j, H^k_p, \mathscr{G}_t)}{\sum\limits_{i>j} \sum\limits_{H^k_p \in H_p} \mathsf{Match}(D, i, j, H^k_p, \mathscr{G}_t)}$$

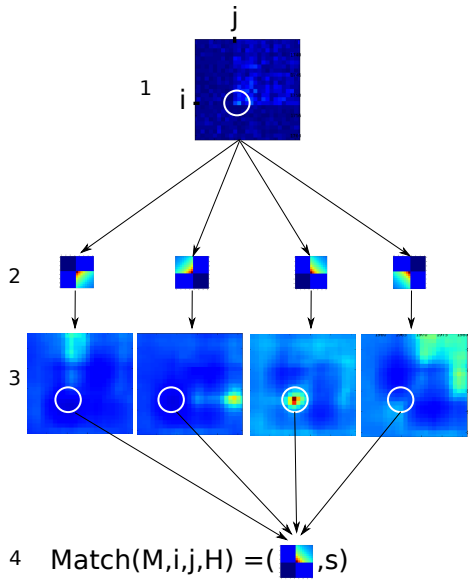


Figure 4.10: Matching feature process. 1) The observed pattern(sub matrix of the observed data). 2) The filter bank. 3)The filters responses.The stronger the signal (blue-weak $\longrightarrow$ red-strong) the more likely the pattern is present.4) Output of the Match at position (i,j)

$$T^{\mathscr{G}_t}_{\mathsf{Split}}(i) = \frac{\mathsf{Match}(D, i, H_s, \mathscr{G}_t)}{\sum_i \mathsf{Match}(D, i, i, H_s, \mathscr{G}_t)}$$

Therefore, thanks to the rejection method, it becomes possible to generate random Paste (resp Split) candidates from the distribution $P^{\mathscr{G}_t}_{\mathsf{Paste}}$ (resp $P^{\mathscr{G}_t}_{\mathsf{Split}}$).

Since the Flip operation may generate very similar pattern to the Split signal we set:

$$P^{\mathscr{G}_t}_{\mathsf{Flip}} = P^{\mathscr{G}_t}_{\mathsf{Split}}$$

Detection of Duplicate (and Delete) We define $Cg(f_i)$, the contact-coverage of an initial restriction fragment f_i, as the number of contacts it has captured:

$$Cg(f_i) = \sum_{j \neq i} D(i, j)$$

Note that our definition of coverage does not have much to see with its classic formulation. In shotgun sequencing the whole genome is cut in small pieces which are expected to be amplified uniformly by PCR. Therefore the coverage is the mean number of reads corresponding to a given nucleotides sequence of the genome.

In our case this measure is biased since we extract and sequence preferably the areas closed to a restriction site. Moreover since we capture ligation events, the spatial functional organization of the genome implies that some region will naturally capture more contacts than others. For instance the clustering of centromeres at the spindle pole body (Rable conformation), encountered in almost all yeast makes that centromeric region will capture more contacts than others.

To summarize, our definition of coverage is altered by the three dimensional organization of the genome and and the biases intrinsic of the method (GC content, size of the restriction fragments, etc...).

Hopefully, the situation is not completely despairing for two reasons:

1. The magnitude of the number of contacts captured between two fragments located at two different chro-

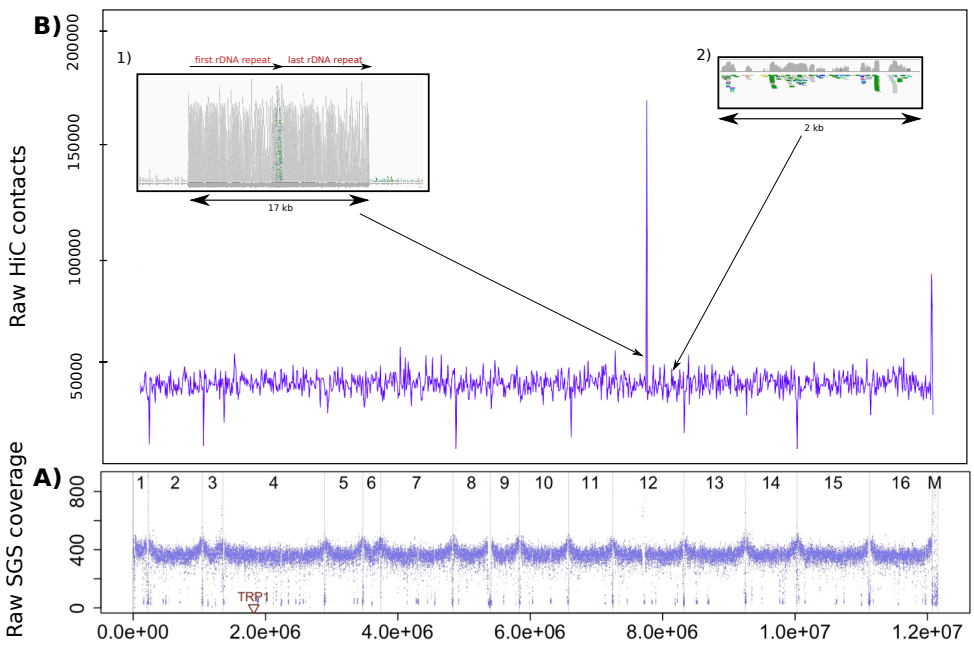


Figure 4.11: Comparison between shotgun sequencing coverage and Hi-C contacts counts. A) Genome wide contact frequency of an HiC experiment performed on yeast cerevisiae. 1) Enrichment of detected contacts and aligned sequences on the rDNA region of the chromosome 12 of yeast cerevisiae. 2) Reads are localized preferably close to restriction sites leaving desert coverage areas. B) Shotgun sequencing coverage performed also on yeast cerevisiae. Both HiC and SGS data display roughly the same tendency.

mosomes or at high genomic distance on the same chromosome, is much lower than the number of contacts captured between fragments at short distances. Therefore we can fairly expect that the impact of the specific spatial contact will not interfere much our estimation of the coverage.

2. All these biases are additive and this should be reflected in the value of $Cg(f_i)$.

A perfect experimental setup should provide at the same time both 3C and shotgun reads. While the contact data can provide estimation about the gap between two genomic regions, the shotgun reads yield a valuable prior knowledge about the sequence which can fill these gaps.

It is of great importance to realize how critical is the estimation of the number of copies of a single fragment. Let us consider the simplest situation where a 3C experiment is performed over a perfectly homogeneous population of genome. In this case, the structure we try to infer is unique. Therefore, if we do not put any limit on the number of copies of the fragments, our stochastic mechanism could produce millions of copies of the chromosomes by decreasing the scale factor of the polymer contact model.

In order to avoid this problem we decided to adopt a very simple strategy[10]. The number of contacts captured by a any fragment follows a Gaussian distribution which is completely defined by the experimental data (figure 4.12:

- μ_{C_g} is the mean of $\{C_g(f_i)|f_i \in F\}$,
- σ_{C_g} is the standard deviation of $\{C_g(f_i)|f_i \in F\}$.

[10]We are aware that it will be mandatory for meta-genomic experiment, to optimize this part of the process

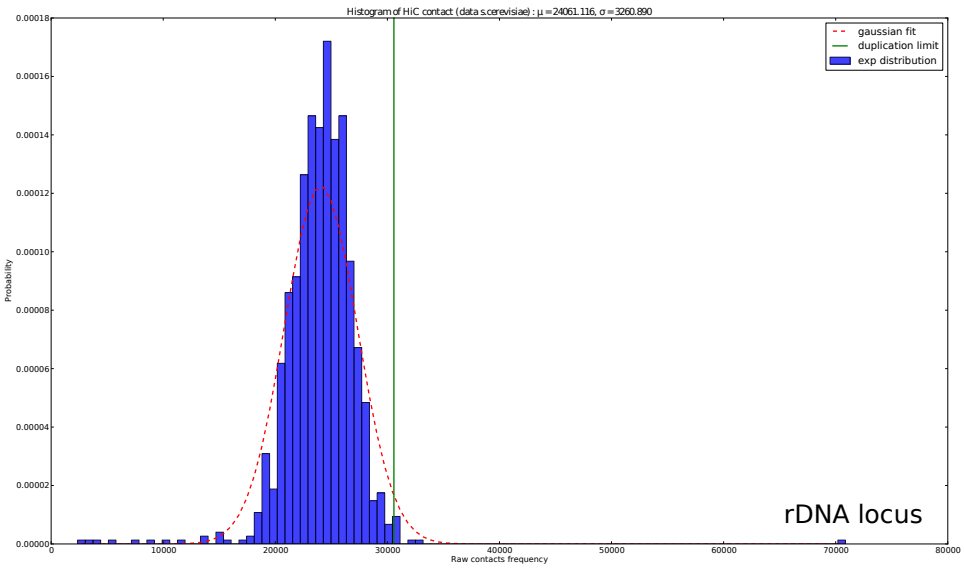


Figure 4.12: Distribution of raw contacts frequency. The figure displays both the experimental distribution of contacts and the fitted Gaussian distribution for an HiC experiment performed on yeast saccharomyces cerevisiae. It is known that around 100 copies of the rDNA genes are located on the chromosome XII. Because of technical limitations of the aligner software we are not able to see a 100 fold increase of contacts frequency.

We recall that $m_t(f_i)$ is the number of copies of the fragment f_i in the genome $\mathscr{G}_t$. Therefore we have:

$$P_{\mathsf{Duplicate}}(\mathscr{G}_t, \varphi_k^i) = \begin{cases} \frac{1}{2}\left(1 + erf\left(\frac{\left(\frac{C_g(f_i)}{m_t(f_i)} - \mu_{C_g}\right)^2}{2\sigma_{C_g}^2}\right)\right) \approx 0.83, & \text{if } \frac{C_g(f_i)}{m_t(f_i)} \geqslant 2\sigma_{C_g} \\ 0.17, & \text{otherwise} \end{cases} \tag{4.23}$$

$$P_{\mathsf{Delete}}(\mathscr{G}_t, \varphi_k^i) = \begin{cases} 0, & \text{if } m_t(f_i) = 0 \\ 1 - P_{\mathsf{Duplicate}}(\mathscr{G}_t, \varphi_k^i), & \text{otherwise} \end{cases} \tag{4.24}$$

We are now able to provide a full sampling strategy based on these distributions.

Sampling scheme Our image based genome sampler works as follows:

1. Start with an initial genome $\mathscr{G}_0$.
2. Given the current genome $\mathscr{G}_t$, compute the transition probability for Paste and Split operations.
3. Pick at random a modification $\theta \in \{\mathsf{Paste}, \mathsf{Split}, \mathsf{Duplicate}, \mathsf{Delete}, \mathsf{Flip}\}$.
4. Given θ, pick at random $\theta_t(x_*)$ with respect to the corresponding distribution $P_\theta^{\mathscr{G}_t}(x)$.
5. Compute $P_{\theta^{-1}}^{\mathscr{G}_*}(x_{*-1})$, the probability of the reversed transformation $(\theta_t(x_*))^{-1}$
6. Generate a new genome $\mathscr{G}_* = \mathsf{Mod}_{\mathscr{G}_t,}(\theta_t(x_*))$

7. Accept $\mathscr{G}_*$ with the probability:

$$r = min\left(1, \frac{p(\mathscr{G}_*, \xi_t)q(\mathscr{G}_* \longrightarrow \mathscr{G}_t)}{p(\mathscr{G}_t, \xi_t)q(\mathscr{G}_t \longrightarrow \mathscr{G}_*)}\right)$$
$$= min\left(1, \frac{p(\mathscr{G}_*, \xi_t)P(\theta_t(x_*))}{p(\mathscr{G}_t, \xi_t)P((\theta_t(x_t))^{-1})}\right)$$

Pre-evalutation The genome of the biomass-degrading fungus Trichoderma reesi is made of 77 contigs, Martinez et al, 2008 [88], on which a systematic analysis of repeated sequences have been performed. An assembly of 77 contigs is often considered as a very good scafold.
A HiC experiment have been performed on this specie and around 31 millions (3170955) paired reads were produced. After mapping on the reference genome and filtering 7 millions (7958927) contacts were generated.

Results The results of the simulation are summarized below in figure 4.13 and in figure 4.14. In this simulation duplication of the fragments were not authorized yet.

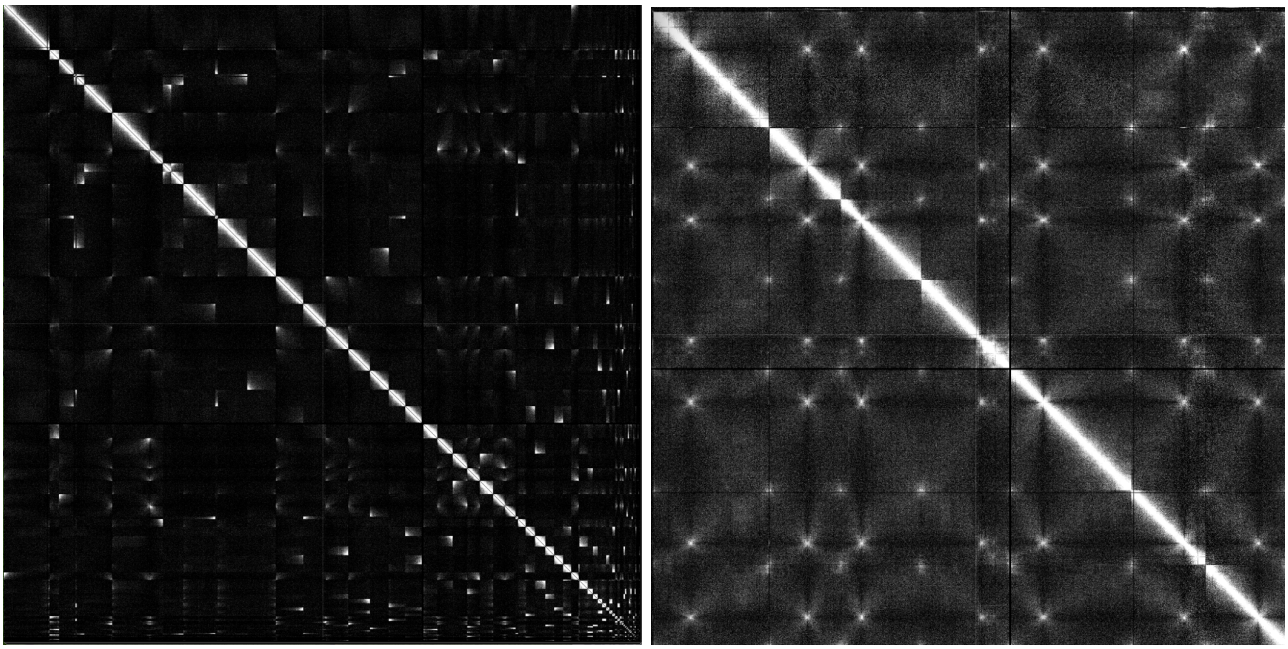


Figure 4.13: Snapshot of the final iteration of the MCMC procedure. On the left is displayed the initial matrix. The 77 contigs are ordered from the longest to the smallest. The last iteration structure is displayed on the right. Seven long contigs have been produced while 22 small fragments remain free.

Limitations This preliminary test gives us precious indication about the performance of the method. The fact that the algorithm could not sample relevant contigs candidate including the 22 small fragments is a direct consequence of the transition probability based on corner detection:

- First the filter bank is biased by the initialization of the method.
- Changes involving small fragments are much more difficult to detect.
- Intra chromosomal modifications because of the high signal displays along the diagonal.

Then another critical aspect of this approach arise from the fact that at every time point it is necessary to redraw the matrix in memory. On small genome of around ten thousands fragments the operation can easily

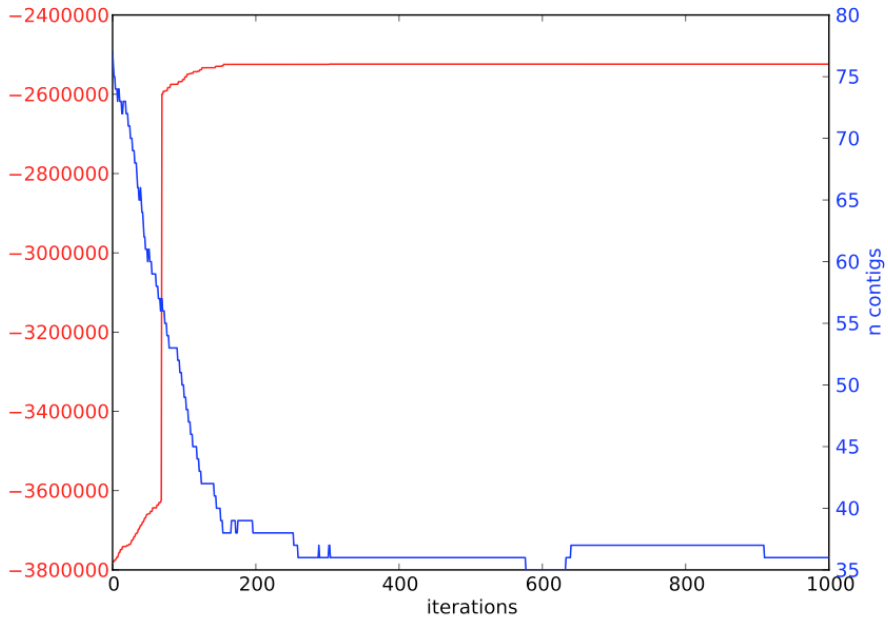


Figure 4.14: Trace of the simulation. The red curve shows the evolution of the likelihood throughout the 1000 steps simulation. The blue curve gives the number of contig available in the genomic structure at every time step.

be distributed on the GPU. However below a hundred thousands fragment, the dense representation of the matrices makes impossible the computing of local convolutions.

4.4.3.6 High vicinity sampling

In order to get rid of the limitations of the image based sampler we decided to adopt a completely different approach based on the hypothesis 2 we formulated at the beginning of this chapter:

Hypothesis. *High 3C signal implies high linear genomic proximity.*

First we will introduce a new class of sampler called Multiple-Try Metropolis developed by Jun S. Liu and Wing Hung Wong, 2000, [79].

Multiple-Try Metropolis (MTM) algorithm One of the main issues in common Metropolis-Hastings implementations is directly linked to the magnitude of the local moves generated at each iteration. The higher is the jump between candidates, the lower will be the acceptance ratio. Conversely small moves will be more likely accepted to the expense of a slow convergence of the process. The Multiple-Try algorithm proposes to break these limitations by realizing a local optimization of the moves before accepting or not a candidate. Let $T(X, Y)$ be a symmetric or not proposal jumping distribution which respects the condition that:

$$T(X, Y) > 0 \Longleftrightarrow T(Y, X) > 0$$

This condition guarantee that the Markov Chain will not be trapped in some subsets of the structures space. The authors of the method define:

$$w(X, Y) = \pi(X)T(X, Y)\lambda(X, Y)$$

where $\pi(X)$ is a complex distribution we try to explore.

Roughly, the algorithm works as follows (Jun S.Liu and Wing Hung Wong, 2009, [79]):

1. Draw k iid candidates, $y_1, \ldots, y_k$, from $T(X, .)$ and compute $w(y_j, X)$ for $j = 1, \ldots, k$
2. Pick $Y = y$ among the candidate set $\{y_1, \ldots, y_k\}$ with probability proportional to $w(y_j, X)$, $j = 1, \ldots, k$.
3. Then draw $x_1^*, \ldots, x_{k-1}^*$, from the proposal distribution $T(Y, .)$ and let $x_k^* = X$.
4. The candidate Y is accepted with probability:

$$r_g = min\left\{1, \frac{w(y_1, X) + \ldots + w(y_k, X)}{w(x_1^*, Y) + \ldots + w(x_k^*, Y)}\right\}$$

The MTM transition rule described above satisfied the detailed balance condition and therefore generate a Markov Chain whose invariant distribution is π (see Jun S.Liu and Wing Hung Wong, 2009, [79]). Before introducing the adaptation of this algorithm to genome sampling we will briefly describe how the likelihood of the data behaves with respect to local modifications applied to a given genome $\mathscr{G}_t$.

Local optimization Let $V_f(f_i, f_j)$ be the function which returns the normalized number of contact between the initial fragment f_i and all the other restriction fragments:

$$V_f(f_i, f_j) = \frac{D[f_i, f_j]}{\sum\limits_{f_k \neq f_i} D[f_i, f_k]}$$

Of course we have:

$$V_f(\varphi_k^i, \varphi_l^j) = V_f(f_{id(\varphi_k^i)}, f_{id(\varphi_l^j)})$$

Figure 4.15 displays the empirical contact frequency of the initial fragment f_{400} which is located on the chromosome 7 of s.cerevisiae.

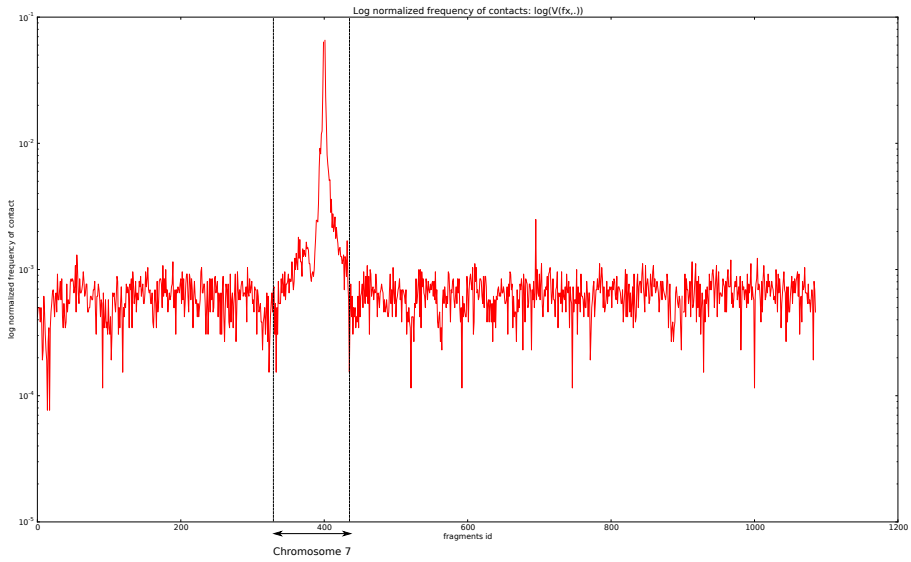


Figure 4.15: Log normalized contacts frequency $V_f(f_{400}, .)$. The fragment f_{400} is located at position 675143 bp on chromosome 7. The contact data comes from an HiC experiment performed on yeast s.cerevisiae.

We define two new composite mutations (figure 4.16):

- $\mathsf{PopOut}(\mathscr{G}_t, \varphi_k^i)$. This operation performs two double strand breaks at each extremity of the restriction fragment φ_k^i and the the repair is performed by joining the upstream strand to the downstream strand. The consequence of these two Split and this Paste is that the fragment φ_k^i get ejected from the contig k. In the following of the manuscript we might also refer to PopOut as Eject.

- $\mathsf{Insert}(\mathscr{G}_t, \varphi_k^i, \varphi_k^i, up)$. First, this operation pop out the fragment φ_k^i and insert it at the upstream (up = upstream) or downstream side of the fragment φ_l^j.

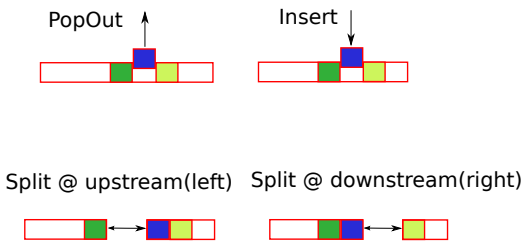


Figure 4.16: Graphical representation of PopOut, Insert and Split

Let $f_X = f_{\varphi_k^i}$. $V_X = \{f_{Y_1}, \ldots, f_{Y_k}\}$ is a set of k iid fragments sampled with respect to the empirical distribution $V_f(f_X, .)$. Mut is a set of simple, {Paste, Split, Duplicate, Delete, Flip} and composite mutations. To simplify the notation we will define an arbitrary index on Mut. Therefore $\mathsf{Mut} = \{\theta_1, \ldots, \theta_m\}$ where for instance θ_i can be Split.
Ψ_X is the set of all modified genomes produced as follows:

$$\Psi_{f_X}^t = \left\{\mathscr{G}_k^i = \mathsf{Mod}_{\mathscr{G}_t,}(L), L = \theta_i(f_X, f_{y_k}, \alpha), \text{with } \theta_i \in \mathsf{Mut}, f_{y_k} \in V_X, \alpha \in \mathsf{Aux}(\mathsf{Mut})\right\}$$

where Aux(Mut) is the set of mutations auxiliary parameters such as the upstream or downstream area to split. As explained in figure 4.17, given a fragment f_X, by applying small moves to a genome $\mathscr{G}_t$ it is possible to determine which of these changes will generate the most optimal modified genome $\mathscr{G}_*$:

$$\mathscr{G}_* = \arg\max_{\mathscr{G} \in \Psi_{f_X}^t} P(\mathscr{G}|D, \xi)$$

It is obvious that the fact of reducing the research space to the mutations performed with V_X, provides an easy way lower the computing charge with no loose of efficiency. Now that we have defined the concept of local optimization we are able to adapt it to the MTM algorithm previously described.

MTM genome sampler This algorithm adopts the same strategy as the published canonical version. However, since our transition function will differ from the one used in the original algorithm we will give the proof that the detailed balance equation is verified.
We call Mut_{rev} a set of self contained reversible mutations (simple or composite). We set $N_m = \text{card}(\mathsf{Mut}_{rev})$ and to simplify the notation we write:

$$\pi(\mathscr{G}) = P(\mathscr{G}|D, \xi)$$

Given a genome $\mathscr{G}_t$, the algorithm works as follows:

1. Pick at random a restriction fragment $f_{\mathbf{X}} = f_X$.

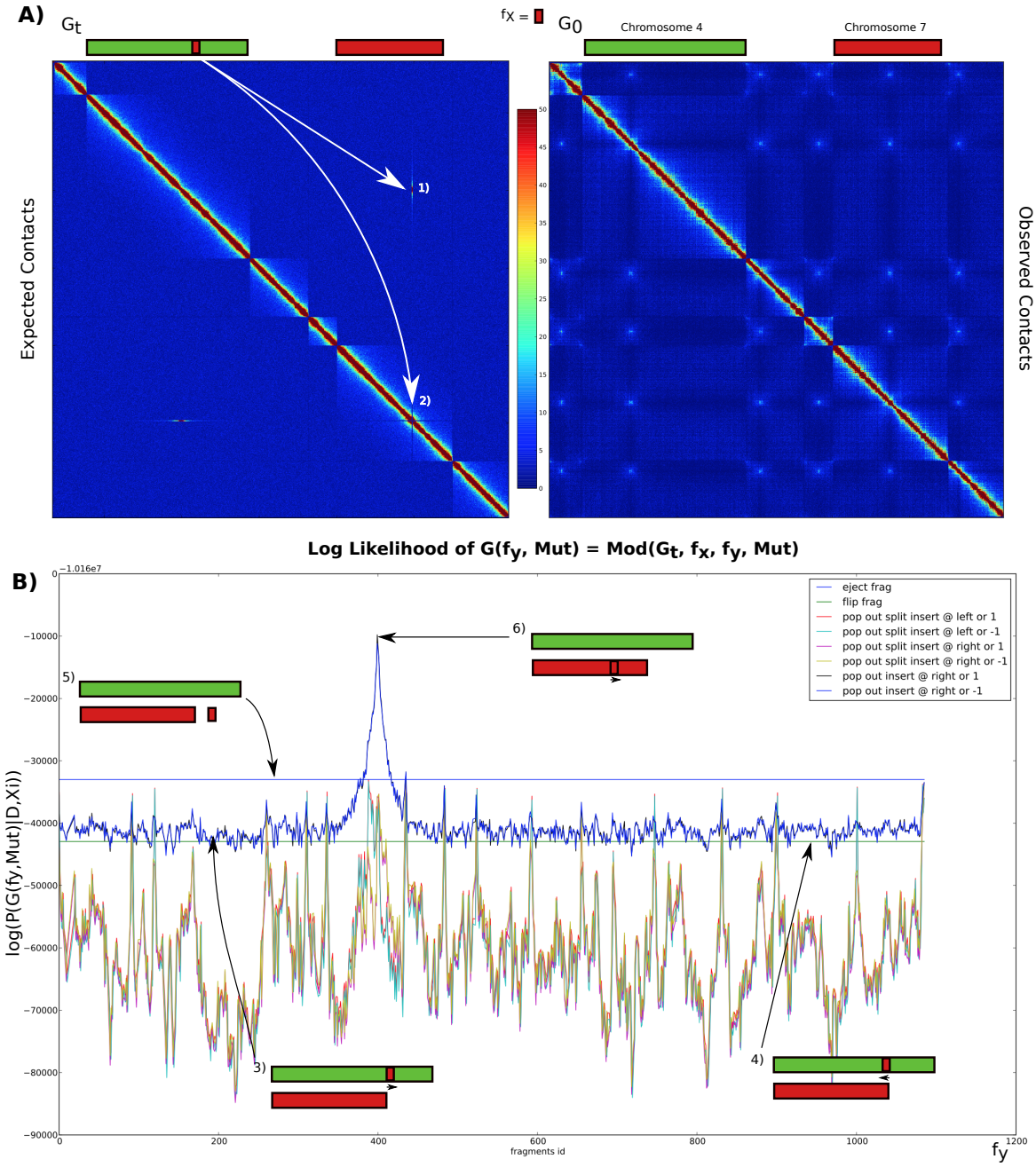


Figure 4.17: Local structure optimization. On panel A are display the expected and observed sub contact matrices (chromosomes 3, 4, 5, 6, 7 and 8) of an HiC experiment performed on yeast s.cerevisiae. The parameters of the model are the one computed at the initialization as displayed in figure 4.8. The raw data have been mapped on a perfectly assembled genome $\mathscr{G}_0$. Conversely in the genome $\mathscr{G}_t$, the restriction fragment $f_X = f_{400}$ has been inserted on the chromosome 4. This is why typical ill patterns such as 1) and 2), can be seen in the expected contact matrix.
On panel B, are displayed the log likelihood of modified genomes $\mathscr{G}_* = \mathsf{Mod}_{\mathscr{G}_t,}(\theta_t(x_*))$ where θ might be Flip, Eject, Insert or a combination of PopOut, Split and Insert. The argument $x_* = (f_X, f_Y, ., .)$ has a constant element which is the fragment f_X while f_Y is not fixed. Therefore each f_Y axis point of the graph gives the likelihood of a modified genome $\mathscr{G}_* = \mathsf{Mod}_{\mathscr{G}_t,}(\theta_t((f_X, f_Y, ., .))$. The score of the current genome $\mathscr{G}_t$ 3), is very close to the one of the genome whose f_X has been flipped 4). As expected the highest score 6) corresponds to the correct structure $\mathscr{G}_0$. Please note that the genome in which the fragment f_X is ejected, 5), exhibits a higher score than many structures where the fragments f_X is wrongly located on chromosome 7.

2. Draw k_0 iid contacts neighbors $V_X = \{f_{y_1}, \ldots, f_{y_{k_0}}\}$ with probability proportional to $V_f(f_X, .)$ and pick

$k \leq k_0 \times N_m$ genomes[11] at random in $\Psi^t_{f_X}$.

$$G_X = \{\mathscr{G}_{Y_1}, \ldots, \mathscr{G}_{Y_k}\}$$

For every $\mathscr{G}_{Y_j} \in G_X$ the corresponding mutation parameters are given by:

$$\sigma(\mathscr{G}_{Y_j}) = (\theta_a, f_{y_b}, \alpha), \text{such that } \mathscr{G}_{Y_j} = \mathsf{Mod}_{\mathscr{G}_t,}(\theta_a(f_X, f_{y_b}, \alpha)$$

3. Select $\mathscr{G}_{\mathbf{Y}} = \mathscr{G}_Y$ among G_X with probability proportional to:

$$w(\mathscr{G}_Y, \mathscr{G}_t) = \pi(\mathscr{G}_Y) \times T(\mathscr{G}_Y, \mathscr{G}_t, f_X)$$

where:

$$\begin{aligned} T(\mathscr{G}_Y, \mathscr{G}_t, f_X) &= P(\sigma(\mathscr{G}_Y)|f_X) \\ &= V_f(f_{y_b}, f_X) \times \frac{1}{N_m} \times P(\alpha|\theta_a) \end{aligned}$$

The probability $P(\alpha|\theta_a)$ is uniform over the set of auxiliary parameters specific to θ_a.

4. Draw k_0 iid contacts neighbors $V_Y = \{f_{x_1}, \ldots, f_{x_{k_0}}\}$ with probability proportional to $V_f(f_Y, .)$ and generate k_0 genomes:

$$\Psi^*_{f_Y} = \left\{\mathscr{G}^i_{x_k} = \mathsf{Mod}_{\mathscr{G}_Y,}(L), L = \theta_i(f_Y, f_x, \alpha), \text{with } \theta_i \in \mathsf{Mut}_{rev}, f_x \in V_Y, \alpha \in \mathsf{Aux}(\mathsf{Mut})\right\}$$

Then, pick $k-1$ candidates at random in $\Psi^*_{f_Y}$, $\left\{\mathscr{G}_{X_1}, \ldots, \mathscr{G}_{X_{k-1}}\right\}$ and set:

$$G_Y = \left\{\mathscr{G}_{X_1}, \ldots, \mathscr{G}_{X_{k-1}}\right\} \cup \{\mathscr{G}_t\}$$

5. Accept $\mathscr{G}_Y$ with probability:

$$r_g = \min\left\{1, \frac{\sum\limits_{\mathscr{G}_{y_k} \in G_X} w(\mathscr{G}_{y_k}, \mathscr{G}_t)}{\sum\limits_{\mathscr{G}_{x_k} \in G_Y} w(\mathscr{G}_{x_k}, \mathscr{G}_Y)}\right\}$$

Theorem 1. *Let N be the maximum number of fragments allowed by the experiment. N account for duplicated and non duplicated fragments. If N is constant, the MTM transition kernel described above satisfies the detailed balance condition and therefore generate a Markov Chain whose invariant distribution is π.*

Proof. We will follow the step of the proof given by Jun S.Liu and Wing Hung Wong, 2009, [79]. Let $A(\mathscr{G}_t, \mathscr{G}_Y)$ be the transition probability for moving from $\mathscr{G}_t$ to $\mathscr{G}_Y$ as described above. The two genomes are different and

[11] Please note that the mutation set Mut_{rev} has to be designed in a way to avoid duplication of structures.

let I indicate which of $\mathscr{G}_{y_j}$ has been selected. Since: $w(\mathscr{G}_Y, \mathscr{G}_t) = \pi(\mathscr{G}_Y)T(\mathscr{G}_Y, \mathscr{G}_t)$, we can write:

$$\begin{aligned}
\pi(\mathscr{G}_t)A(\mathscr{G}_t, \mathscr{G}_Y) &= \pi\left(\mathscr{G}_t\right) P\left[\bigcup_{j=1}^{k}\{\left(\mathscr{G}_{\mathbf{Y}_j} = \mathscr{G}_Y\right) \cap (I = j) \cap (f_{\mathbf{X}} = f_X)\}|\mathscr{G}_t\right] \\
&= k\pi(\mathscr{G}_t)P\left[(\mathscr{G}_{\mathbf{Y}_k} = \mathscr{G}_Y) \cap (I = k) \cap (f_{\mathbf{X}} = f_X)\,|\mathscr{G}_t\right] \\
&= k\pi(\mathscr{G}_t)\int \ldots \int P(f_{\mathbf{X}} = f_X)T(\mathscr{G}_t, \mathscr{G}_Y)T(\mathscr{G}_t, \mathscr{G}_{Y_1})\ldots T(\mathscr{G}_t, \mathscr{G}_{Y_{k-1}}) \\
&\times \frac{w(\mathscr{G}_Y, \mathscr{G}_t)}{\sum\limits_{j=1}^{k} w(\mathscr{G}_{Y_j}, \mathscr{G}_t)} \times \min\left\{1, \frac{\sum\limits_{j=1}^{k} w(\mathscr{G}_{Y_j}, \mathscr{G}_t)}{\sum\limits_{j=1}^{k} w(\mathscr{G}_t, \mathscr{G}_{Y_j})}\right\} \\
&\times T(\mathscr{G}_Y, \mathscr{G}_{X_1})\ldots T(\mathscr{G}_Y, \mathscr{G}_{X_{k-1}})df_X \times d\sigma_{Y_1} \times \ldots \times d\sigma_{Y_{k-1}} \times d\sigma_{X_1} \times \ldots \times d\sigma_{X_{k-1}} \\
&= \pi(\mathscr{G}_Y)A(\mathscr{G}_Y, \mathscr{G}_t)
\end{aligned}$$

with σ_{Y_i} the set of all the mutations parameters. Since $P(f_{\mathbf{X}} = f_X) = \dfrac{1}{N}$, with N the total number of restrictions fragments, we retrieve the same formula as derived in the original paper and therefore the detailed balance equation is proved. □

Because of time constraints, the implementation of this sampler is still under construction and we do not have yet any direct preliminary results to show the performance of this algorithm. However we have developed and implemented the main concept introduced in this sampler in a stochastic optimization routine which gives us great confidence in the efficiency of this approach.

4.5 Stochastic optimization algorithm

We will introduce in this section a procedure which belongs to the class of stochastic local search algorithm. First we will describe the method before evaluating its performance. Then we will present some results obtained on three organisms for which a correct assembly is still not available.

4.5.0.7 Description

In the previous section we have introduced the concept of local optimization based on some reversible mutations. The reversible condition is a strong constraints necessary to assure the convergence of the Markov Chain toward the target distribution. The Markov Chain we will build in this section will not have the same limitations. Therefore, we will not be able to provide a robust theoretical proof of the convergence of the whole procedure.

Unconstrained local optimization Let Mut_u be a set of reversible and non reversible mutations. In addition of the five basic operations described in the previous section, Mut_u posses the following operations:

- The five basics mutation: $\{\mathsf{Paste}, \mathsf{Split}, \mathsf{Duplicate}, \mathsf{Delete}, \mathsf{Flip}\}$
- $\mathsf{PopOut}(\mathscr{G}_t, \varphi_k^i) = \mathsf{Eject}(\mathscr{G}_t, \varphi_k^i)$
- $\mathsf{Insert}(\mathscr{G}_t, \varphi_k^i, \varphi_l^j, up)$

- $\mathsf{Translocate}(\mathscr{G}_t, \varphi_k^i, up_{k,i}, \varphi_l^j, up_{l,j})$: If $up_{i,k} = 1$ (resp $up_{k,i} = 0$), this operation pastes the upstream (resp downstream part) of the fragment φ_k^i to the upstream part (resp the downstream part) of the fragment φ_l^j if $up_{l,j} = 1$ (resp $up_{l,j} = 0$).
- $\mathsf{PSI}(\mathscr{G}_t, \varphi_k^i, \varphi_l^j, up)$. This composite operation performs first, $\mathsf{PopOut}(\mathscr{G}_t, \varphi_k^i)$, then $\mathsf{Split}(\mathscr{G}_t, \varphi_l^j, up)$ and $\mathsf{Insert}(\mathscr{G}_t, \varphi_k^i, \varphi_l^j, up)$.

With these operations, we allow the program to explore widely the genome conformations space at every iteration. The overall work-flow of the algorithm the same as the Gibbs sampler described previously:

1. Start with an initial guess of the genome $\mathscr{G}_0$ and initial values ξ_0 of the parameters of the polymer model.
2. A classic Metropolis Hastings algorithm updates the parameters of the model:$\xi_{t+1} \backsim P(\xi|D, I, \mathscr{G}_t)$
3. Then a stochastic optimization procedure updates the structure: $\mathscr{G}_{t+1} = \text{Stoc Optim}(P(\mathscr{G}_t|D, I, \xi_{t+1})$

Algorithm The genome optimization procedure, Stoc Optim, works as follow:

1. Start with an initial genome structure $\mathscr{G}_t$.
2. Draw at random a restriction fragmentf_X.
3. Pick k contact neighbors $V_X = \{f_{Y_1}, \ldots, f_{Y_k}\}$ with probability proportional to $V_f(f_X, .)$.
4. Generate $N_g = k \times \text{card}(\mathsf{Mut}_u$ candidates genomes:

$$\Psi_{f_X}^t = \left\{\mathscr{G}_k^i = \mathsf{Mod}_{\mathscr{G}_t,}(L), L = \theta_i(f_X, f_{y_k}, \alpha), \text{with } \theta_i \in \mathsf{Mut}_u, f_{y_k} \in V_X, \alpha \in \mathsf{Aux}(\mathsf{Mut}_u)\right\}$$

5. Set $\mathscr{G}_{t+1} \in \Psi_{f_X}^t$ with probability proportional to $P(\mathscr{G}|D, I, \xi_t)$

Before evaluating the method we would like to highlight some details of the implementation of the algorithm.

Implementation Please note that at every iteration we do not need to recompute the whole likelihood matrix. As a matter of fact it suffices to know which operation has been performed to update the likelihood only at the data point where the structure have changed. This optimization holds also for duplicated fragments. Moreover in order to check the behavior of the iterative process, we have implemented a real time matrix viewer which allows one to visualize the consequences of the mutations on the structure. We also included a three dimensional genome viewer which will be useful for further metagenomic analysis.

4.5.0.8 Evaluation

Yeast *S. cerevisiae* (BY4741) To evaluate our method we performed a preliminary test on a very simple data set[12]. A HiC experiment has been produced on yeast *S. cerevisiae* (BY4741). Around 52 millions (51 804 551) paired reads have been produced. 41 % (21 457 086 millions contacts) of the library was correctly mapped on the genome GCF_000146045.1.

We filled up a pyramid of contact matrix whose compression factor was arbitrary set to 3. Since the restriction

[12]Experiments on simulated data have also been performed but because of time constraints we are unable to display them in this manuscript.

enzyme used to perform the experiment was DpnII, the restriction map was made of 35913 fragments. We performed our analysis at the third level of the pyramid were the matrix D_3 represented the captured contacts over 1086 binned restriction fragments. The second level of the pyramid, necessary to orientate the fragments contained 3240 fragments.
We broke the initial 16 chromosomes into 1086 contigs. Then we launched our optimization routines. The results of the first 43440 iterations is displayed in figure 4.18. The study of the behavior of the embeded Metropolis Hastings algorithm has not been yet done. However we defined a very simple metric which provides valuable information about the position of a given genome with respect to the initial one.

Distance between genomes Let $Ne(f_i)_t$ be the set of the neighbors of f_i at time t. The distance between the initial genome $\mathscr{G}_0$ and $\mathscr{G}_t$, a given genome at time point t, $d(\mathscr{G}_0, \mathscr{G}_t)$, is computed by penalizing the differences between the $Ne(f_i)_0$ and $Ne(f_i)_t$.

Structural variant: analysis of duplications We performed the same analysis on a structural variant of yeast *S. cerevisiae*, YKF1246. The genome of this mutant is organized as displayed in figure **??**. An HiC experiment has been performed on this strand and around 36 millions (35958716) paired reads were produced. These reads were aligned in the same fashion as described below. We generated a 3 scales pyramid of contact matrices and perform the same analysis we did on the baker yeast. The results of the process is displayed in figure 4.19.

4.5.0.9 Discussion

The performance of the algorithm are very encouraging for these two data set. From more than a thousands fragments, the algorithm succeeded to retrieve the great majority of the fragments contiguity. Please note that the reason why the rDNA locus get correctly connected is still under investigation. A further analysis is desirable to investigate the effect of the nuisance parameters on the stability of the genome thoughout the experiment.

4.6 Results

These encouraging results gave us great confidence to try our method to correct the assembly of three organisms for which the genomic structure is poorly known.

4.6.1 Malaysian yeast

The yeast UWOPSO3-461.4 is a wild strand of s.cerevisiae which is commonly observed in Malaysia. We will refer to this strand as the malaysian yeast. Despite the fact that the lineage of this organism is well known, no correct assembly has been produced yet for this strand. The only draft assembly possesses more than 3000 contigs which maximum size is around 100 kb.
We performed an HiC experiment on this organism and mapped the obtained sequences on the reference genome of yeast s.cerevisiae. The results of the stochastic optimization procedure is displayed in figure 4.21

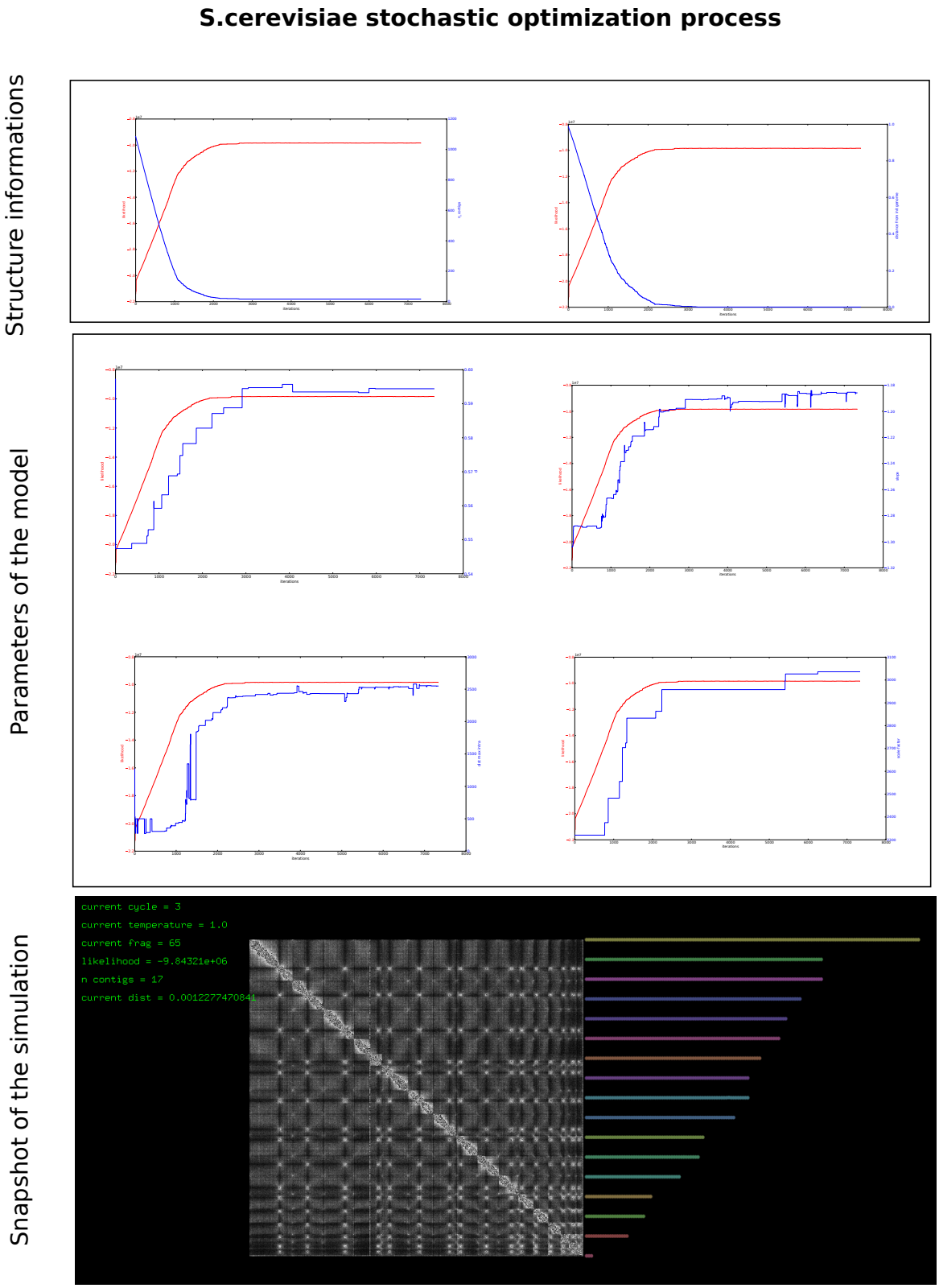


Figure 4.18: Results of the stochastic optimization procedure performed on yeast s.cerevisiae. The algorithm converges quickly to the right structure as displayed in the first panel.

Validations In order to check the validity of our method, classic PCR amplification have been performed at some chimeric junction proposed by the algorithm. The output of this procedure is displayed in figure 4.22.

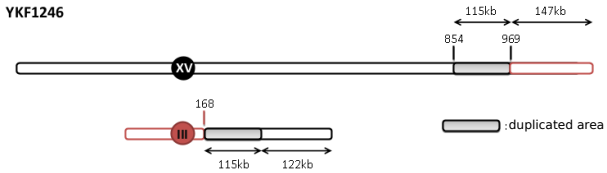


Figure 4.19: The yeast YKF1246 is a structural variant of yeast s.cerevisiae. A block of 115 kb is duplicated and inserted just after the centromeric region of the chromosome 3. Then two translocations between the chromosome 3 and the chromosome 15 are displayed.

4.6.2 Trichoderma reesei

Non evolved strain: QM6A As explained previously, the genome of the biomass-degrading fungus Trichoderma reesi is made of 77 contigs, Martinez et al, 2008 [88], on which a systematic analysis of repeated sequences have been performed. The non evolved strain, qm6a, used to perform this basic assembly has been sequenced in a HiC fashion.

The results of our process is displayed in figure 4.23 and figure 4.25. From 77 contigs we ended up with 11 contigs: 7 chromosomes and 4 small restriction fragments of 20 kb each **??**.

Industrial strain : rutc30 Such as the strand qm6a ,the genome of the industrial strain of trichoderma reeise, rutc30, is not known. We applied our method to this organism and get the following results : figure 4.26. From 77 contigs we ended up with 11 contigs: 7 chromosomes and 4 small restriction fragments of 20 kb each.

4.7 Conclusion

In this section we have introduced a promising framework for genome assembly. We are aware that the preliminary results we gave deserve to be analyzed further. However we have great confidence in the evolution of this program toward a full sampling procedure. Beside the characterization of the linear structure of the genome, HiC data allow also to characterize typical functional features. In the next chapter we will show how it is possible to detect centromeres and rDNA loci in organisms having a Rabl genome configuration.

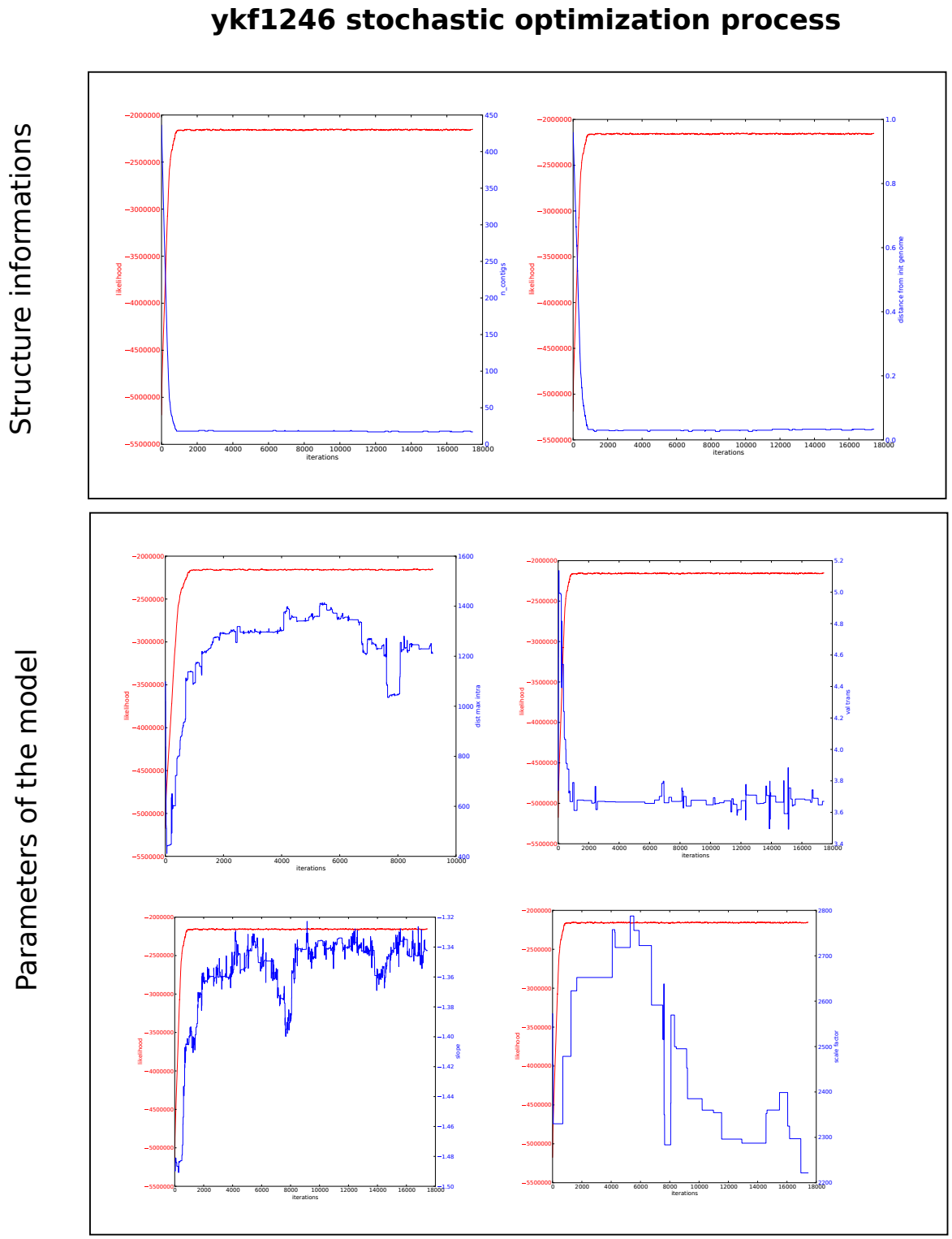


Figure 4.20: Results of the stochastic optimization procedure performed on yeast YKF1246. The algorithm converges quickly to the expected structure as displayed in the first panel.

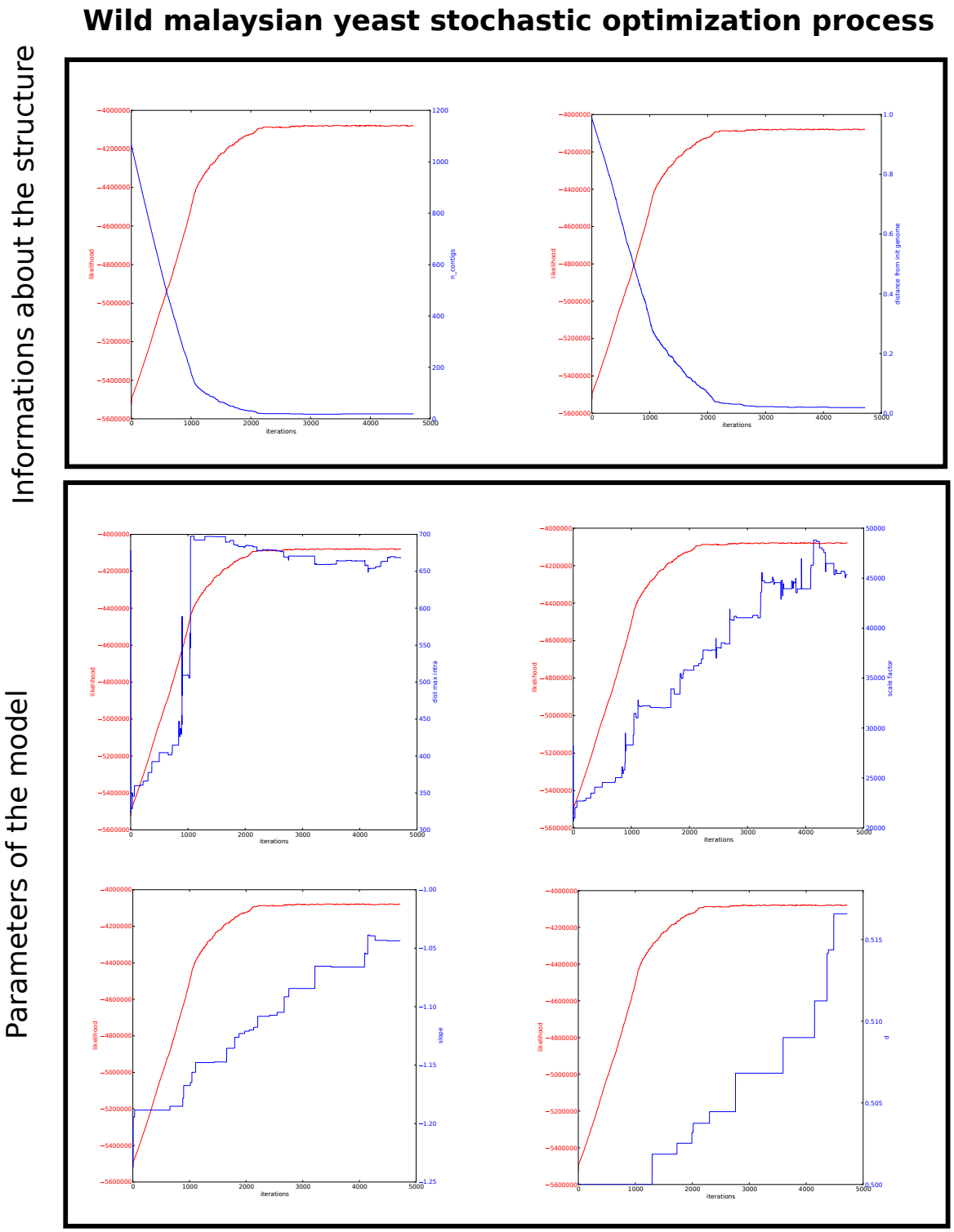


Figure 4.21: Results of the stochastic optimization procedure performed on the Malaysian yeast UWOPSO3-461.4. The algorithm converges quickly to a structure of 16 chromosomes. This number can easily be verified by looking at the inter centromeric region of the re ordered contact matrix.

A)

		chromosome	location	orientation	sequence
pair n°1	primer 1	VIII	466000	+	GCTGATCTGGAGCCAAGAC
	primer 2	X	203000	+	CCCATATCTCAGATTTCACATCAG
pair n°2	primer 1	VIII	466000	+	GCTGATCTGGAGCCAAGAC
	primer 2	X	203000	-	CTGATGTGAAATCTGAGATATGGG
pair n°3	primer 1	VIII	400000	-	CGGCATCAGCGAAGAAC
	primer 2	VII	425000	+	CCTTCAGTCCAGATGGAGC
pair n°4	primer 1	XI	459000	-	GGGAAACCGCCAAATGC
	primer 2	VII	321000	+	CGGGAGGGATAATGTCC
pair n°5	primer 1	IX	225000	+	CCACAGGTAACTCCA
	primer 2	IX	235000	-	GATGCACAACGTACAG

B)

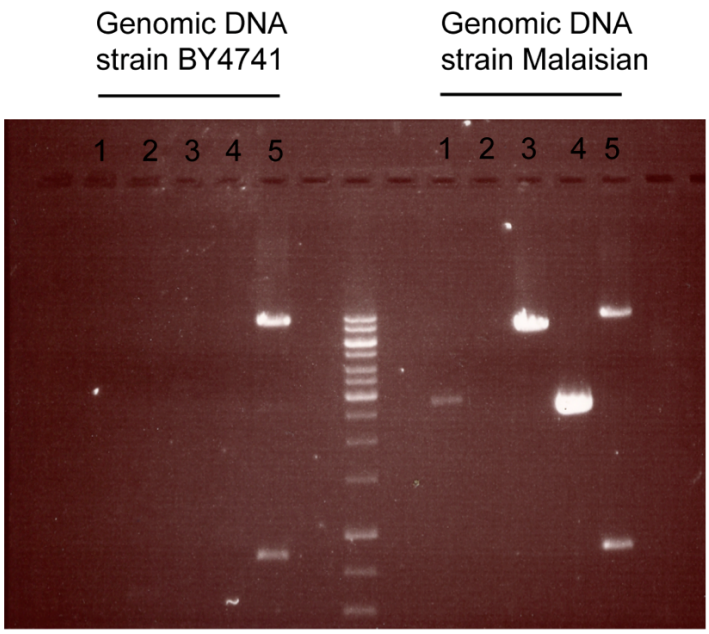


Figure 4.22: Validation of the HiC based reassembly. A) table of primers pairs used to test the assembly of S. cerevisiae Malaisian strain. Chromosome, location and orientation indicated in the table came from S288C reference genome. B)gel with the different PCR product obtained using the pairs of primers indicated in A using Genomic DNA from BY4741 or Malaisian strain as template.

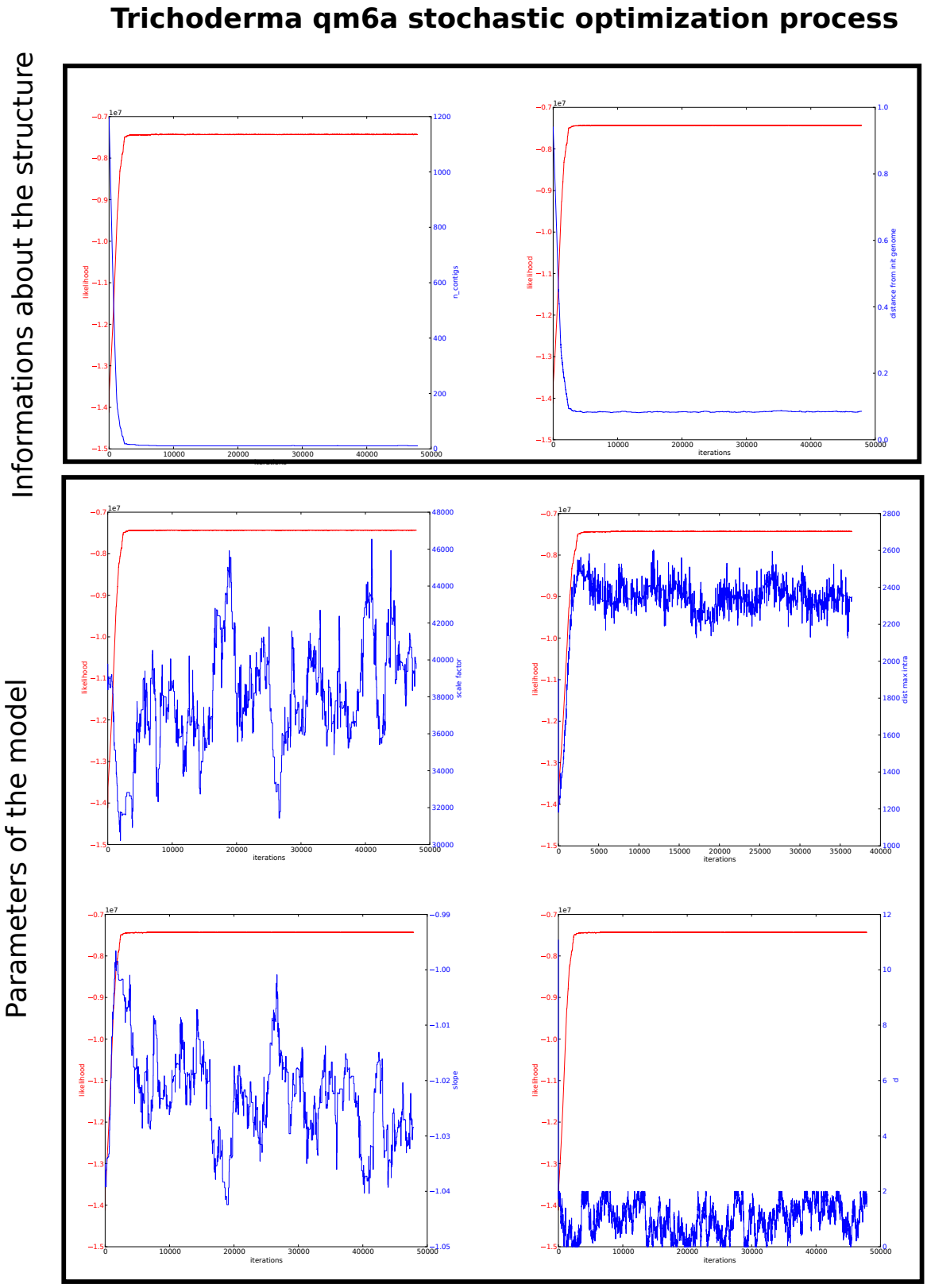


Figure 4.23: Results of the stochastic optimization procedure performed on the fungus trichoderma qm6a.

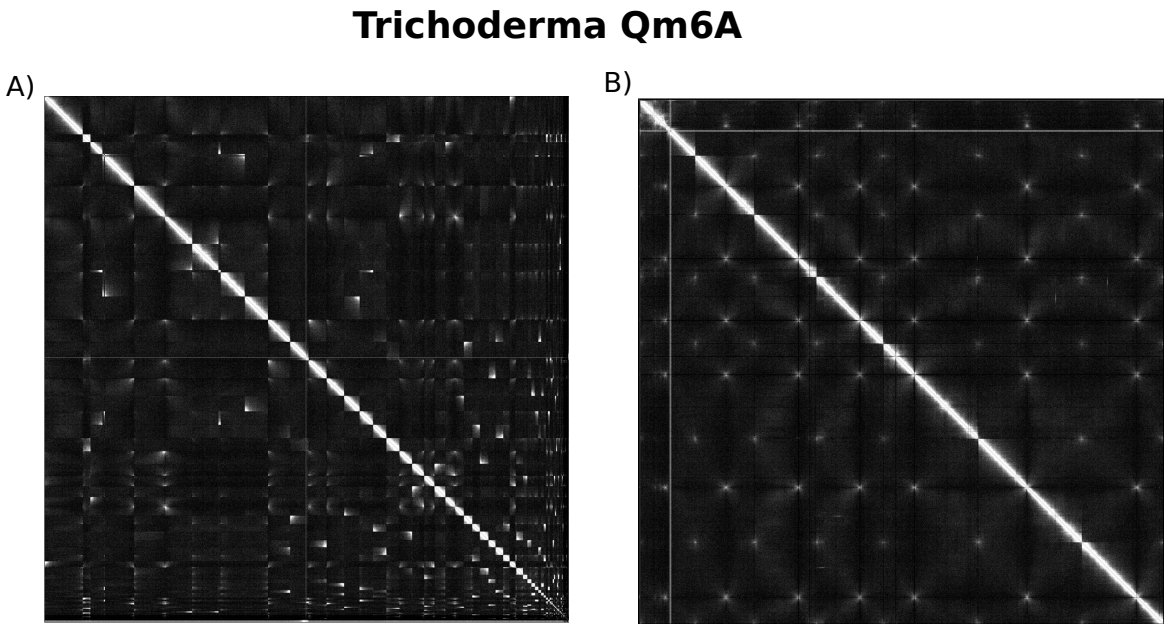
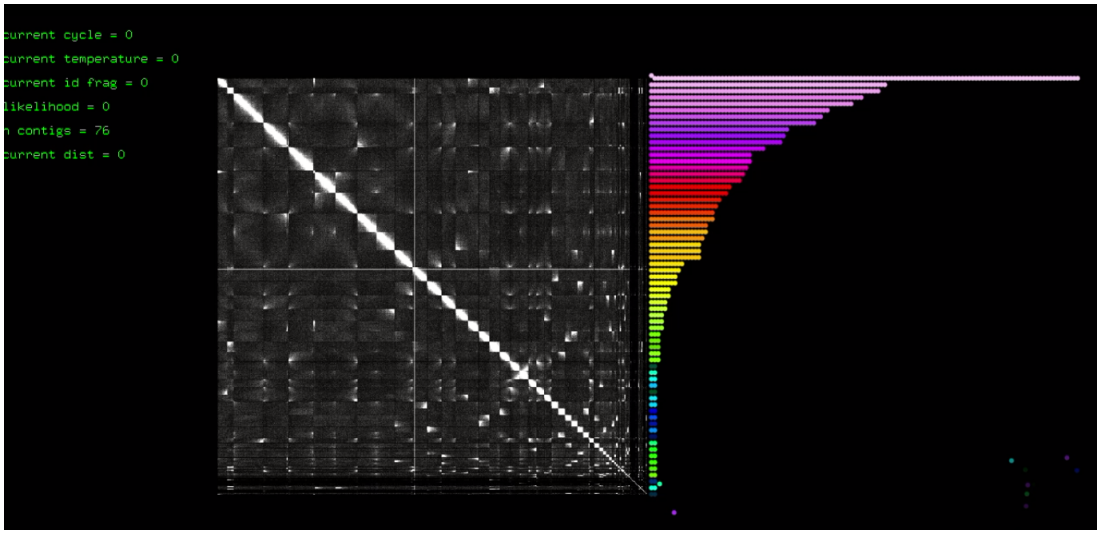


Figure 4.24: Trichoderma qm6a contact matrix. A) Matrix before the stochastic optimization. B) Matrix after optimization.

A)



B)

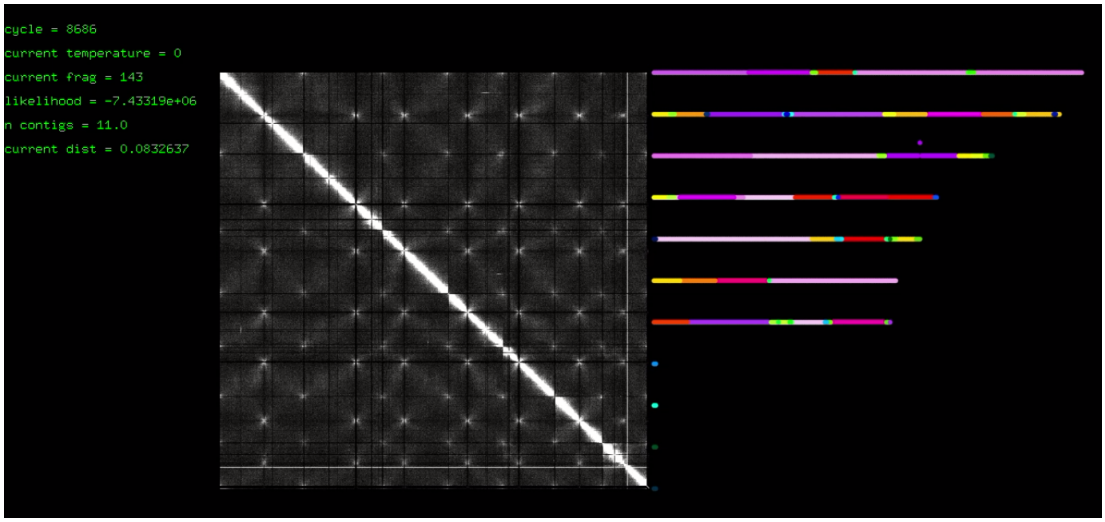


Figure 4.25: Snapshot of the viewer of the optimization procedure. A) Before optimization. B) after optimization

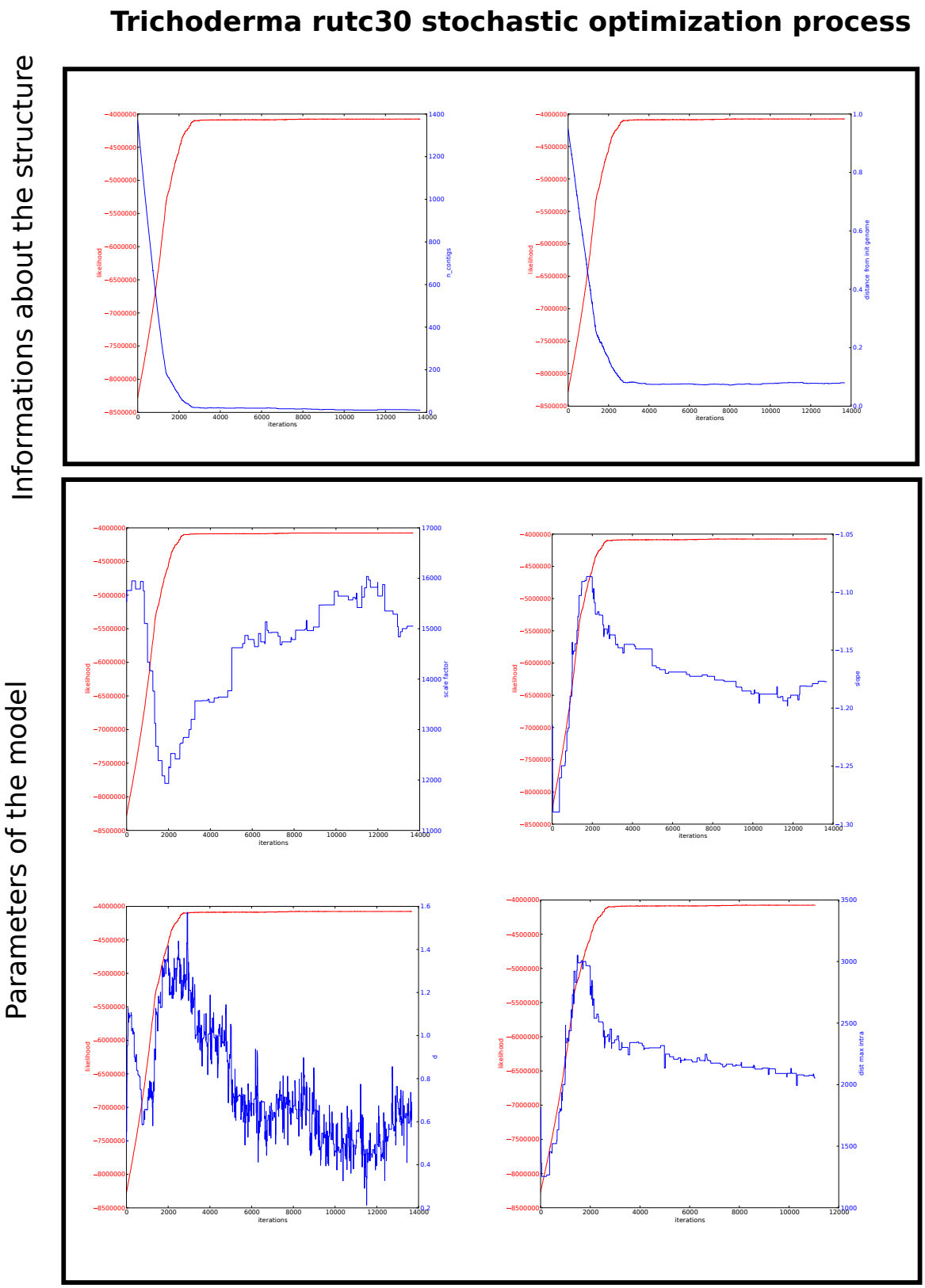


Figure 4.26: Results of the stochastic optimization procedure performed on the industrial fungus, trichoderma rutc30.

Chapter 5

Identification of centromeres from contact data

Contents

Genome-wide Chromosome Conformation Capture (3C) techniques are powerful means to decipher the three-dimensional organization of genomes. Combined with genomic and epigenetic annotations, these techniques provide important new information about the interplays between gene regulation, chromosome organization, and chromatin state and have led to important insights in a range of model organisms. In this chapter we demonstrate an unconventional application of these techniques by showing how functional information about the sequence *per se* can be extracted directly from the chromosome contact map. We use these approaches to complete genomic annotation of several yeast species. Specifically, we show that the genomic positions, or discrimination between several putative positions, of co-localizing centromeres can be readily identified, and the technique allows to conveniently identify the position or ribosomal DNA clusters, and that even in species where classical computational approaches fail. We first validate this technique on the budding yeast *Saccharomyces cerevisiae* genome, where centromere and rDNA positions are known. We then successfully apply it to the well-studied and intriguing RNAi-containing yeast *Naumovozyma castellii* where centromere positions cannot be determined with standard techniques, to *Kuraishia capsulata* in order to discriminate between multiple predicted centromeric positions, and to *Debaryomyces hansenii* where rDNA positions are uncharacterized. By completing these genomes, interesting observations can be drawn. For instance, in *N. castellii*, although most of centromeres are characterized in regions of conserved synteny with neighboring species, no consensus sequences can be identified suggesting that centromeric binding proteins and/or the mechanisms involved have significantly diverged.

5.1 Introduction

De novo sequencing of genomes is typically followed by analyses aiming to identify functional genomic features such as genes, non-coding RNAs or regulatory sequences. This important so-called annotation step raises non-trivial questions, and led to the development of complex bioinformatics approaches taking advantage of multiple datasets. For instance, transcriptome analysis is conveniently used to annotate expressed coding sequences (Grabherr et al., 2011 [56], Saha et al., 2002 [122]) and synteny conservation between related species can reveal or confirm the presence of regulatory elements (Gordon, Byrne and Wolfe, 2011 [54], Kellis, Birren and Lander, 2004 [67], Stark et al., 2007 [130]). Complementary to automated annotation through comparative approaches, experimental approaches such as ChIP-seq or MNase-seq have been conveniently used to map epigenetic marks, replication origins, or other functional elements of the genome (Roy et al., 2010 [118], Wang et al., 2012 [143]). However, such tools are sometimes unable to detect non-coding functional sequences: for example, origins of replication, centromeres and rDNA positions have sometimes proven difficult to annotate with a high degree of confidence in several genomes. A compelling example is represented by the inability to identify the centromeres of the hemiascomycetes species *Naumovozyma castellii* through comparative genomics (Gorden, Byrne and Wolfe, 2011 [54]). A more anecdotic example is the imprecision surrounding the number and positions of rDNA clusters in the genome of *Debaryomyces hansenii*, ranging from one to three and not indicated in the genomic sequence (Dujon et al., 2004 [39], Jacques et al., 2010 [64], Corredor et al., 2003[26]).
Genomic chromosome conformation capture (3C) assays measure the physical contact frequencies between DNA sequences (Dekker et al., 2002 [32], De Laat and Dekker, 2012 [31], Lieberman-Aiden, 2009 [78], Duan et al., 2010 [38]), providing important insights into both genomic organization and topological changes of chromatin domains that accompany cell differentiation or development. 3C data are typically analyzed in light of epigenetic marks and other genomic annotations. Here, we use genome-wide 3C data to unveil functional elements of eukaryotic genomes that escape comparative genomic analysis. Specifically, we take advantage of nuclear architecture features to precisely determine the positions of centromeres in the yeast species *Naumovozyma castellii* (Cliften et al., 2006 [22]). We show that this approach discriminates ambiguous results from bioinformatics analysis, such as in *K. capsulata*. Finally, it also allowed us to complete the annotation of *D. hansenii* rDNA locus, by revealing the presence of a unique intrachromosomal cluster. This method can probably be extended to discrete DNA sequences motifs interacting in the nuclear space.

5.2 Material and Methods

We took advantage of the peculiar behavior of centromeres and rDNA cluster in the yeast nucleus to develop a robust approach to characterize them experimentally.
First, yeast centromeres are tethered near a pole of the nucleus via microtubules attached to the microtubule organizing center (MTOC, or Spindle Pole Body in yeast), leading to centromere clustering. In the budding yeast *Saccharomyces cerevisiae*, this clustering causes distinct peaks of interchromosomal contact frequencies in the raw genome-wide contact matrix (Duan et al., 2010 [38]). We developed an algorithm that automatically recognizes these specific contact enrichments and estimates the genomic coordinates of centromeres. Centromeric positions are therefore experimentally characterized based on the biology of centromeres, and not on computational analysis based on motives recognition algorithms as it is usually the case. On the contrary, this approach

discriminates between multiple positions equally fitted from computational analysis.

Second, ribosomal DNA is organized as a cluster of repeats in the genome of all eukaryotes sequenced so far. In *S. cerevisiae*, and other species, these repeats are organized into the nucleolus, that occupies a discrete volume within the nuclear space opposite to the SPB. This organization combined with the large size of this cluster result in the rDNA creating what looks like an intra-chromosomal barrier in the contact matrix of the chromosome carrying it. The position of a rDNA cluster in a genome is therefore easily identifiable, even in the absence of any annotation or sequence in the reference sequence. We developed an algorithm that identifies pair-end reads where one read is well mapped on the genome assembly while the mate is not. We then retain the mates containing ribosomal sequences and look at the distribution of the associate read along the genome. The pics in the distribution indicate strong enrichments in ribosomal sequences, indicating the unique sequences at these positions were frequently captured with ribosomal DNA adjacent along the chromosome and revealing the presence of rDNA clusters.

The flowchart in Figure 5.1 provides an overview of the workflow, each of which will be described in a distinct subsection below.

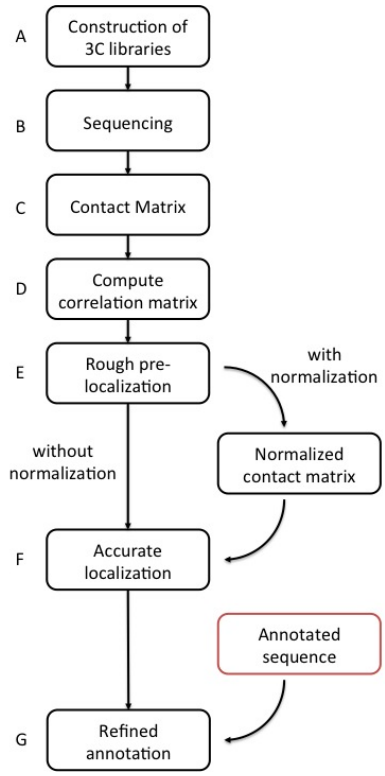


Figure 5.1: Experimental and computational workflow.

5.2.1 Generation of genome-wide chromosome contact frequency matrices

3C libraries of the yeast species *S. cerevisiae* (BY4741), *N. castellii* (CBS4309), *D. hansenii* (CBS767), and *K. capsulata* (CBS1993) were generated from log-phase cells growing in YPD medium and as previously described (Dekker et al., 2002 [32], Oza et al., 2009 [106]), but using a frequent cutter (DpnII) as in Sexton et al. (2012 [126]). Briefly, the cells were cross-linked with 1% formaldehyde, and resuspended in DpnII restriction buffer. They were then processed through the 3C procedure described in Oza et al. (2009 [106]) to generate 3C libraries subsequently processed into Illumina libraries. 3C libraries were sheared and resulting fragments

between 400 and 800bp were sequenced using 100bp pair-end sequencing on an Illumina HiSeq2000. All 3C-seq experiments raw data were then processed as follow: first, short reads were mapped on the genomes of *S. cerevisiae* (GCF_000146045.1), *N. castellii* (GCF_000237345.1), *D. hansenii* (GCF_000006445.1), and *K. capsulata* (Morales et al., in revision) using bowtie 2 in local and very sensitive modes (Langmead and Salzberg, 2012 [76]). Only pairs of reads with a Mapping Quality above 30 were retained, and unexpected contact reads were discarded (see Cournac et al., 2012 [27] for details). PCR duplicates were also removed. All reads (except PCR duplicates) that were not retained were included into a pool of "leftover reads". After alignment of individual reads on the reference genome we built a 2D histogram where the value of each 2D bin (pixel) indicates how many reads fall into the corresponding pair of genomic segments. The genomic partition defining these segments was based on the restriction enzyme cutting sites, rather than on constant genomic intervals. For the *S. cerevisiae*, the DpnII restriction enzyme leads to a contact matrix M_0 of size $m_0 \times m_0$, with $m_0 =$ 35914. At this resolution however, the contact matrix is very sparse, and hence noisy. The signal-to-noise ratio can be improved at the expense of genomic resolution by binning the reads into larger genomic intervals. We therefore considered three additional matrices, M_k ($k = 1,2,3$) obtained by summing non-overlapping blocks of $3^k \times 3^k$ pixels. For *S. cerevisiae*, these matrices have genomic bins of $R_1 = 1{,}233 \pm 1{,}095$ bp, $R_2 = 3{,}696 \pm 1{,}919$ bp and $R_3 = 11{,}034 \pm 3{,}455$ bp, (mean $\pm$ standard deviation) and size $m_1 = 9712$, $m_2 = 3{,}240$ and $m_3 = 1{,}086$, respectively (Figure 5.2A – C, respectively).

For small genomic bin sizes R, the limited signal-to-noise ratio of these matrices can complicate the identification of contact frequency enrichments. Computing a correlation matrix, as initially done in (20), allows to strongly increase the contrast of contact patterns. Following Lieberman-Aiden et al. (2009 [78]), we computed a new matrix C from M, where $C(i,j)$ is defined as the Pearson correlation coefficient of the rows i and j of M:

$$C(i,j) = \frac{\sum_{l=1}^{m} \left(M(i,l) - \overline{M}(i)\right)\left(M(j,l) - \overline{M}(j)\right)}{\sqrt{\sum_{l=1}^{m} \left(M(i,l) - \overline{M}(i)\right)^2}\sqrt{\sum_{l=1}^{m} \left(M(j,l) - \overline{M}(j)\right)^2}}$$

and $\overline{M}(i) = \frac{1}{m}\sum_{l=1}^{m} M(i,l)$ is the average value of row i of matrix M. Note that the correlation was computed separately for the interchromosomal and intrachromosomal parts of the matrix (Figure 5.2D).

5.2.2 Rough pre-localization of centromeric regions from *cis* contacts

In the correlation matrix C, the blocks corresponding to intrachromosomal (*cis*) contacts within pericentromeric regions exhibit a characteristic "butterfly pattern" (see Figure 5.3A). This pattern can be explained by the clustering of centromeres near the spindle pole body (SPB) and the polymer brush-like organization of chromosomes in this region, whereby the two chromosome arms are stretched out away from the SPB (Wong et al., 2012 [146]). As a result, the centromere is sequestered away from other loci along the chromosome, leading to a depletion of contacts along the yellow dotted lines in Figure 5.2E, while loci on opposite arms located at similar genomic distances from the centromeres tend to be in proximity, leading to contact enrichments along the "anti-diagonal" (pink dotted line in Figure 5.2E).

We took advantage of this pattern for the automated identification of centromeres by defining a "centromere

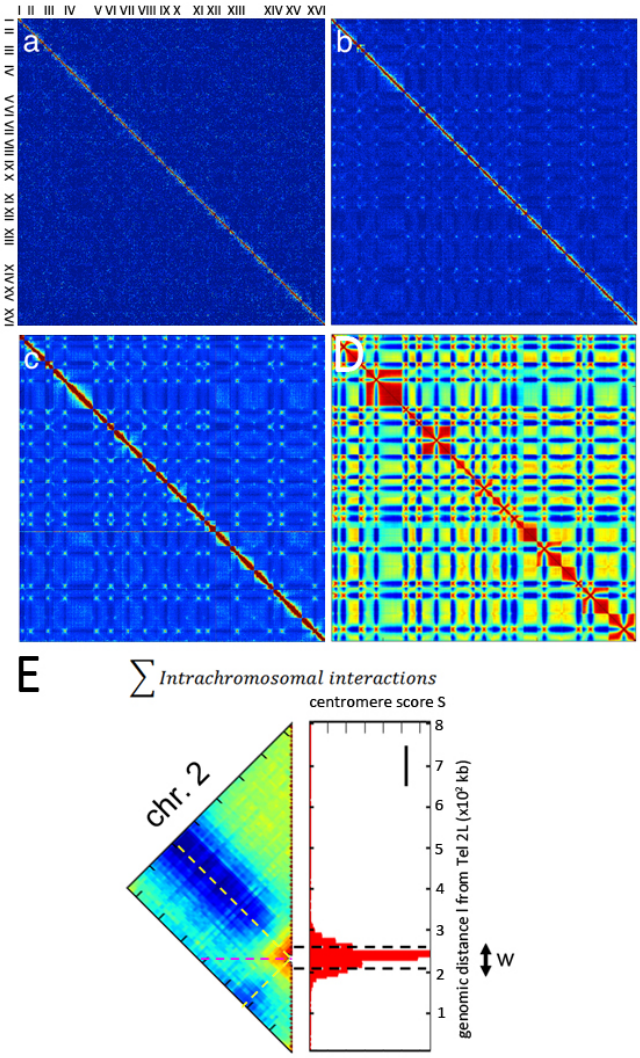


Figure 5.2: Contact frequency matrices M_1, M_2 and M_3, for *S. cerevisiae*, at three levels of genomic resolution: (**A**) R_1 = 1,233 bp, (**B**) R_2 = 3,696 bp, (**C**) R_3 = 11,034 bp. The 16 chromosomes of *S. cerevisiae* are labeled from I to XVI. The strong diagonal is due to intrachromosomal contacts. Note the peaks corresponding to contacts between centromeres from different chromosomes. (**D**) Correlation matrix for *S. cerevisiae*: each element of the matrix is the Pearson coefficient between the vectors i and j of the matrix of contacts (bin size of 3kb). (**E**) Zoom on intra-chromosomal contacts for chromosome 2. The centromere score $S(l)$ for each bin l is plotted along the sub-matrix (scale bar = 100kb). The peak of this distribution defines the center of a 40kb window w likely to contain the centromere.

score" as:

$$S(l) = \frac{\frac{1}{2l-1}\sum_{i=1}^{2l-1} C(l, 2l-i)}{\frac{1}{p}\sum_{j=1}^{p} C(l,j)} \text{ for } l = 1, 2, ...E[(p+1)/2] \text{ and}$$

$$S(l) = \frac{\frac{1}{2(p-l)+1}\sum_{i=2l-p}^{p} C(l, 2l-i)}{\frac{1}{p}\sum_{j=1}^{p} C(l,j)} \text{ for } l = E[(p+1)/2]+1, ..., p-1, p$$

where p is the number of rows of the submatrix and $E(x)$ denotes the largest integer $\leq x$. Thus, for each genomic bin l, $S(l)$ is the ratio of the average correlation along the anti-diagonal passing through $C(l,l)$ and the average correlation along the row l of C. The 'centromere score' $S(l)$ is expected to be largest for l near the actual position of the centromere (Figure 5.2E). Note that for acrocentric chromosomes, the peak of S can differ significantly from the true centromere position. Therefore, for each chromosome k, we used the location

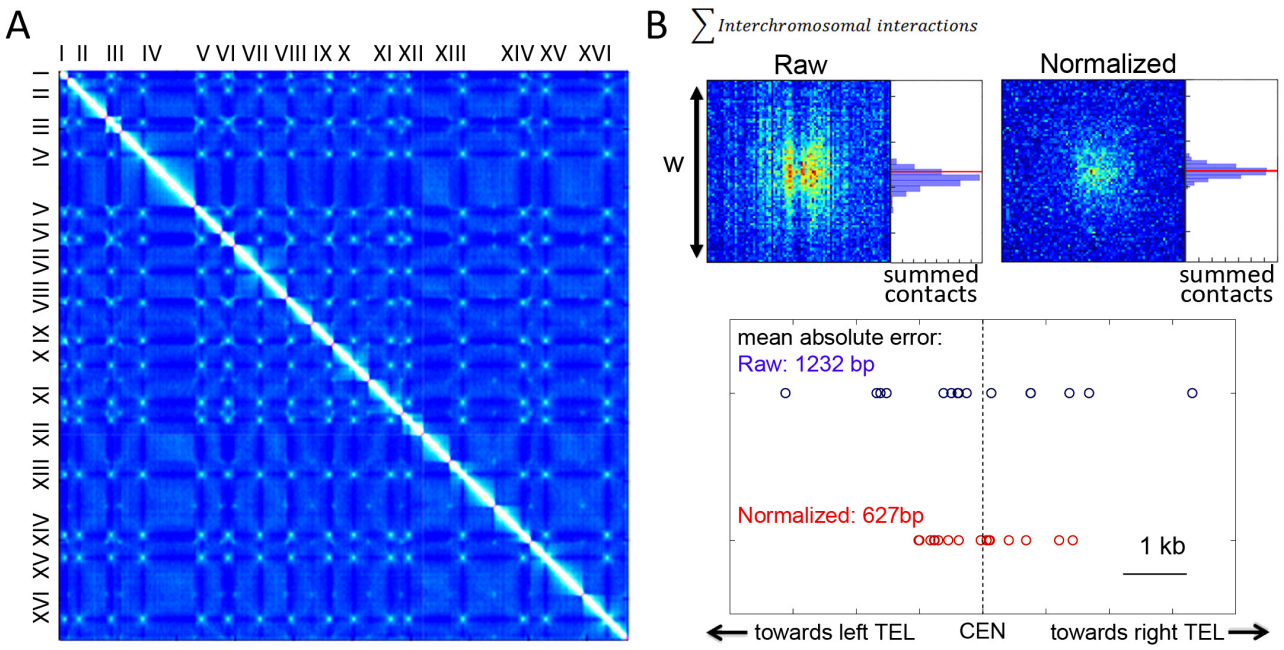


Figure 5.3: (**A**) Normalized contact frequency matrix N_1 for *S cerevisae*. (**B**) Summed trans-contact matrices corresponding to the submatrix of size w (Figure 5.2E) for chromosome 2. On the right side of each sub-matrix we plot the distribution of the centromere localizations obtained using bootstrapping. The true centromere position is indicated with a red line. The diagram below represents the distribution of the 16 centromere positions as estimated from raw and normalized data (blue and red circles, respectively). x-axis: distances along the chromosome, centered on the position of the centromere (scale bar = 1kb).

of this maximum, $l_0 = \arg\max S(l)$ to define a genomic interval $[i_L(k); i_R(k)] = [l_0 - 20kb; l_0 + 20kb]$ along the chromosome that we expect to contain the centromere. The size of the interval is arbitrary and depends on the size of the chromosome: it must be kept within its boundaries, and has to be large enough so that the Gaussian fit can be applied correctly (see below). For *S. cerevisiae* we used window sizes of 40kb. A more accurate localization of the centromere is performed in the next step, as described below.

5.2.3 Refined estimation of centromere position from trans contacts

In principle, the position of a given centromere could be obtained using only the *cis* contact submatrix for the corresponding chromosome, or alternatively using only the trans contact submatrix involving one other chromosome. However, since contact matrices are histograms obtained from a limited number of reads, they are subject to Poisson noise, which imposes a fundamental limit to the localization accuracy (much as in single molecule localization, see e.g. Langmead and Salzberg, 2012 [76]). For improved localization accuracy, we therefore took advantage of the redundancy provided by the $Nchr$ -1 distinct *trans* contact patterns available for each of the $Nchr$ centromeres ($Nchr = 16$ for *S cerevisiae*). This approach was applied both on the raw contact matrix M and on the normalized matrix N, obtained as described in Cournac et al. (2012 [27]). Briefly, this normalization step aims to correct for experimental biases affecting the transformation from ligation product counts into contact frequencies (a different approach as the procedure to correct for them described in Yaffe and Tanay, 2011 [149]). The procedure employs an iterative algorithm that enforces all rows and columns to have unit Euclidian norm, i.e. it ensures that $\sum_{i=1}^{m} N(i,j) = 1$ for all $j = 1...m$ and $\sum_{j=1}^{m} N(i,j) = 1$ for all $i = 1...m$ where m is the size of the matrix. For details, see Cournac et al. (2012 [27]) or Imakaev et al. (2012 [63]) for a related approach. The normalization has the overall effect to increase the contrast of the contact data, and to attenuate

noise in the raw data (Figure 5.3A). Specifically, for each chromosome, we carved out $Nchr$ -1 submatrices of size 40 kb × 40 kb corresponding to *trans* contacts and defined by the $[K - 20kb; k_0 + 20kb]$ intervals obtained above (if necessary, the size of this matrix was reduced to that of the smallest interval, such that all submatrices had the same size). Note that in computing the superposed matrix we did not use the intrachromosomal contact data because of the bias for acrocentric chromosomes mentioned above. These submatrices were then summed, yielding a single "superposed" contact matrix Ak (for the centromere of chromosome k; Figure 5.3B):

$$A_k = \sum_{l \in [1, N_{chr}] l \neq k} M\left(i_L(l)...i_R(l), i_L(k)...i_R(k)\right)$$

For normalized data, M is simply replaced by N (Figure 5.3B). The next step consists in projecting this summed contact matrix into a 1D profile:

$$F_k(i) = \sum_{j=1}^{p} A(i,j)$$

As apparent from Figure 5.3B, normalization typically produces a less noisy profile, allowing more accurate identification of the centromere-related peak.

Finally, in order to accurately estimate the centromere position, we implemented a Gaussian fitting procedure similar to that commonly used for single molecule localization (Ober, Ram and Ward, 2004 [100], Thompson, Larson and Webb, 2002 [137]). Specifically, we used an iterative algorithm that aims to minimize the mean squared difference: $H(a, b, i_c, \sigma) = \sum_i [F_k(i) - G(i; a, b, i_c, \sigma)]^2$ between F and the Gaussian function:

$$G(i; a, b, i_c, \sigma) = a \exp\left(-\frac{(i - i_c)^2}{2\sigma^2}\right) + b$$

where a, b, i_c and σ are the parameters to be fitted, i.e. we seek:

$$(\widehat{a}, \widehat{b}, \widehat{i}_c, \widehat{\sigma}) = \arg\min H(a, b, i_c, \sigma)$$

Thus the final estimated position of the centromere for chromosome k is given by $\widehat{i}_c$.

Application of this procedure to our normalized *S. cerevisiae* contact data and comparison with the genomic annotation revealed that the centromeres could be localized with a mean absolute error of only 627 bp (1232 bp without normalization) - demonstrating that this functionally important locus can be accurately located from the contact data alone (Figure 5.3B).

5.2.4 Confidence intervals and effect of coverage and normalization on localization accuracy

In order to provide a robust roadmap for future studies, we next quantified how centromere localization accuracy is affected by the influence of coverage (i.e. the sequencing depth), binning, and the normalization procedure.

First, we used a bootstrapping approach to estimate confidence intervals of the computed centromere localization and to examine the influence of coverage. Specifically, we simulated many contact frequency matrices with an expected total number of reads either equal to, or smaller than the experimentally obtained matrix M (which for *S. cerevisiae* totals $N_{reads,Sc} = 21,457,086$). To do this, we generated $N_{bs} = 500$ contact matrices $M_{bs,k}, k = 1..N_{bs}$ where $M_{bs,k}(i,j)$ is a random integer value drawn from a Poisson distribution of density

$\lambda(i,j) = fM(i,j)$, where $f \leq 1$ indicates the coverage relative to the original matrix. Thus the expected total number of contacts in $M_{bs,k}$ is fN_{reads}. We then used each of the random contact matrices $M_{bs,k}$ to compute an independent estimate of the centromere positions.

Centromere position confidence interval. For $f = 1$, the distribution of these estimates provides a measure of the uncertainty with which the centromere positions have been determined from the original contact data. We compared the distribution of localization errors for the 16 centromeres of *S. cerevisiae* to the normal distribution of mean 0 and variance given by the bootstrap samples. The two distributions cannot be distinguished by a Kolmogorov-Smirnov test (p = 0.12; Figure 5.4A). This suggests that the confidence intervals determined by the bootstrap estimates correctly reflect actual localization uncertainties.

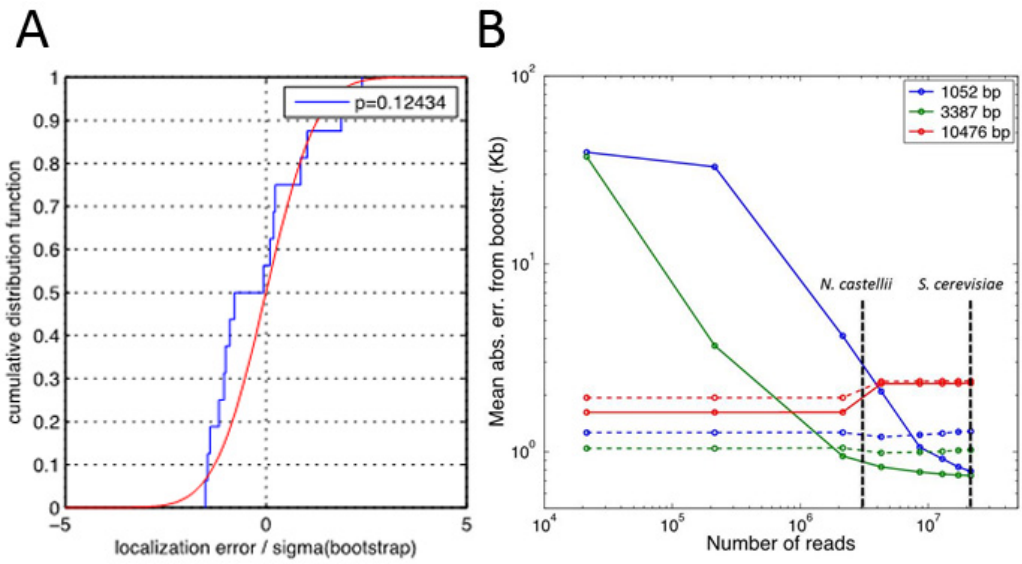


Figure 5.4: (**A**) The distribution of localization errors for the 16 *S. cerevisiae* chromosomes normalized by the standard deviation of corresponding bootstrap estimates (cumulative distribution shown as a blue curve) is consistent with a normal distribution (red curve). (**B**) Effect of coverage, normalization and binning on localization accuracy. The mean absolute localization error for the 16 *S. cerevisiae* centromeres is plotted as function of the number of reads for normalized (solid curves) and raw contact data (dashed curves) and for three resolutions (bin size median indicated by the color legend).

Effect of coverage, normalization and binning. To examine the effect of coverage (or sequencing depth), which determines the total number of reads, we applied the bootstrapping method to a range of f smaller than 1 (i.e. to matrices where the number of reads has been down-sampled), specifically: $f = 0.8, 0.6, 0.4, 0.2, 0.1, 0.01, 0.001$. For each value of f we computed the centromere position error from the N_{bs} samples (relative to the ground truth) and the mean over the 16 centromeres and the N_{bs} samples. This mean error was plotted as function of the mean number of reads in the bootstrapped samples (fN_{reads}) in Figure 5.4B. As expected, the localization accuracy generally improves with coverage, provided that the contact data are binned at adequate genomic resolution and that the qualities of the libraries are equivalent. Also, normalization improves localization accuracy for high coverage ($N_{reads} > 2.10^6$), but gives much poorer results for low coverage, where the raw data should be preferred and provide more consistent accuracy. This result underlines the complexity of contact matrix analysis studies that have to take into account the quality of the matrix, the sequencing depth, the binning, and the normalization procedure. The graph provides a means to determine the likely optimal choice of binning and normalization options for the DpnII enzyme applied on a budding yeast genome. Using bins

smaller than 3kb does not significantly affect localization accuracy of normalized data (Figure 5.4B).

5.2.5 Identification of rDNA loci in chromosome contact matrices

In order show that contact matrixes can allow the characterization of ribosomal gene clusters, we proceed as follow. First, a contact matrix of *S. cerevisiae* was generated where the bins containing the two rDNA repeats of the reference genome were removed (XII::451575-468931). The pair-end Illumina reads were remapped on this modified genome (including the mitochondrial DNA). We then selected in the pool of "leftover reads" all the pairs where one mate would map unambiguously on the genome (mapping quality above 30), and the other mate would not (i.e. the mapping quality field in the sam file is a start symbol) eliminating reads containing unknown bases (N). These unmapped sequences were blasted on a sequence dataset containing yeast ribosomal sequences (if available, preferentially ribosomal elements of the species of interest) retrieved from the NCBI server. The Blast parameters were:

$$blast2 - pblastn - e2e - 30$$

to keep only highly significant hits. The corresponding mates were then mapped along the genome divided into bins (Figure 5.5). The peak in distribution was clearly apparent on chromosome XII (∼10,000 hits compare to

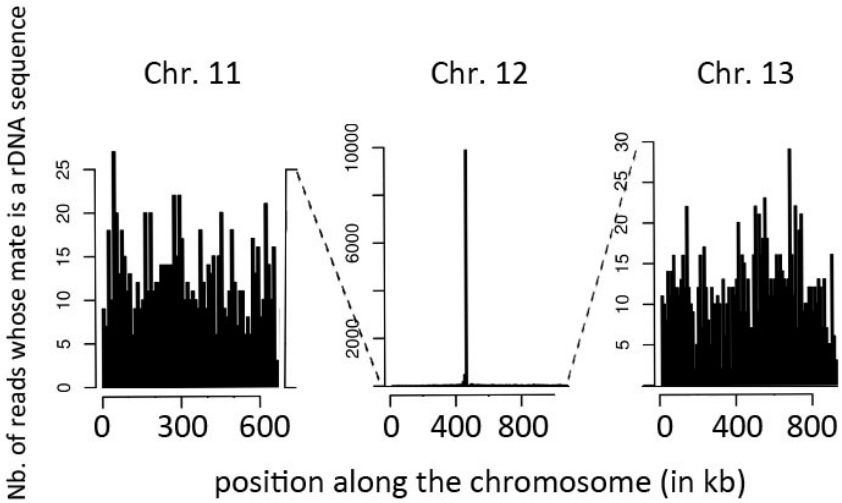


Figure 5.5: Correlation matrix of *N. castellii* (10 chromosomes; *A*). *B*) Zoom on intra-chromosomal contacts of chromosome 3. The peak of this distribution defines the center of a 20kb window w likely to contain the centromere (scale bar = 100kb). **C**) List of computed centromere positions for the ten chromosomes of *N. castellii*

an average of 20 along the rest of the genome), and by zooming in the distribution we were able to precisely estimate the position of the ribosomal gene cluster to lie within the XII:: xxx – xxx window.

Interestingly, if the mitochondrial ribosomal genes are retained in the analysis, the positions of nuclear mitochondrial DNA, i.e. pieces of mitochondrial DNA that has been inserted within the nuclear genome over time (NUMT; Imakaev et al., 2012 [63]) also appear clearly. Although the noise also increases (intrinsic to the experimental procedure), the peaks of ribosomal DNA NUMT appear clearly along the genome, including numerous positions not identified through classical computational analysis (Corredor et al., 2003 [26]).

5.3 Results and Discussion

We then proceed to test our different approaches on uncharacterized or ambiguous genomic features of several yeast species, e.g. centromeres and ribosomal DNA locus. Centromeres of the well-studied yeast *Saccharomyces cerevisiae* are very compact (125bp). The consensus sequence is composed of three centromere DNA elements (CDEI, II, III; Lieberman-Aiden et al., 2009 [78]). CDE I and III present a strong consensus core region and flank CDEII which is characterized by a strong AT rich content but a high sequence variability (>90%). Interestingly, such organization of a point centromere made of three consensus sequences is also found in most *Saccharomycetaceae* species studied so far, and this feature has been an efficient landmark to predict centromeres positions in sequenced species (Gordon, Byrne and Wolfe, 2011 [54]).

Disclosing true centromeres among computational predictions in Kuraishia capsulata The genome of the nitrate

assimilating yeast *K. capsulata* has been recently sequenced and assembled into seven chromosomes (Morales et al., submitted). A search for CDEI and III consensus sequences also failed to identify putative centromeres. However, an alternative computational analysis approach searching for composition-bias in GC content and motif recognition as described in Bailey and Elkan (1994, [5]) and Gordon, Byrne and Wolfe (2009 [53]) prove more successful and led to the characterization of nine putative centromeric regions (with chromosome four containing three; Morales et al., submitted; Table 1). In order to confer an experimental validation of these results and see if we could discriminate between ambiguous sequences we performed a genomic 3C experiment on *K. capsulata* and sequenced the resulting library. The quality of the matrix was relatively poor despite an important coverage ($N_{reads,Nc}$ = 16,446,227; Figure 5.6A), as seen by the "flatness" of the matrix and as quantified by the ratio between mitochondrial and genomic DNA interactions (AC and RK, personal communication). Despite the apparent noise, each chromosome still exhibits a discrete region presenting a strong enrichment in interactions with the corresponding other chromosomal regions, similar to centromeric behavior in *S. cerevisiae*. We followed the procedure described above and characterized for each of the seven chromosomes cis-contact matrixes the genomic intervals containing centromeric regions (Figure 5.6B). Given the low coverage in informative reads of the matrix, we opted for a binning of 2kb, and assessed from the analysis above that little if any improvement would result from SCN normalization. We proceeded to superpose the *trans*-submatrices containing the centromeric regions defined from the cis-contacts. A Gaussian fit was applied as described, and the coordinates of centromere positions along with the precision calculated (Figure 5.6B). Quite remarkably, the regions identified experimentally through this approach overlapped exactly with those obtained after computational analysis for the six chromosomes exhibiting a single, unambiguous putative centromere position. In addition, the region identified on chromosome four as the centromere overlapped with only one of the three putative positions identified from the composition bias analysis, allowing the annotation of this position as the true centromere (Table 5.1). This first analysis indicates that careful analysis of contact matrix can successfully and efficiently back up computational annotation, experimentally confirming and eventually disambiguating weak predictions.

Identification of centromeres in Naumovozyma castelliiv

We then turned to N. castellii, an organism in which centromeric regions remained elusive to date (Gordon, Byrne and Wolfe, 2011[54]). We built a genomic 3C library of *N. castellii* CBS 4309 strain and generated the

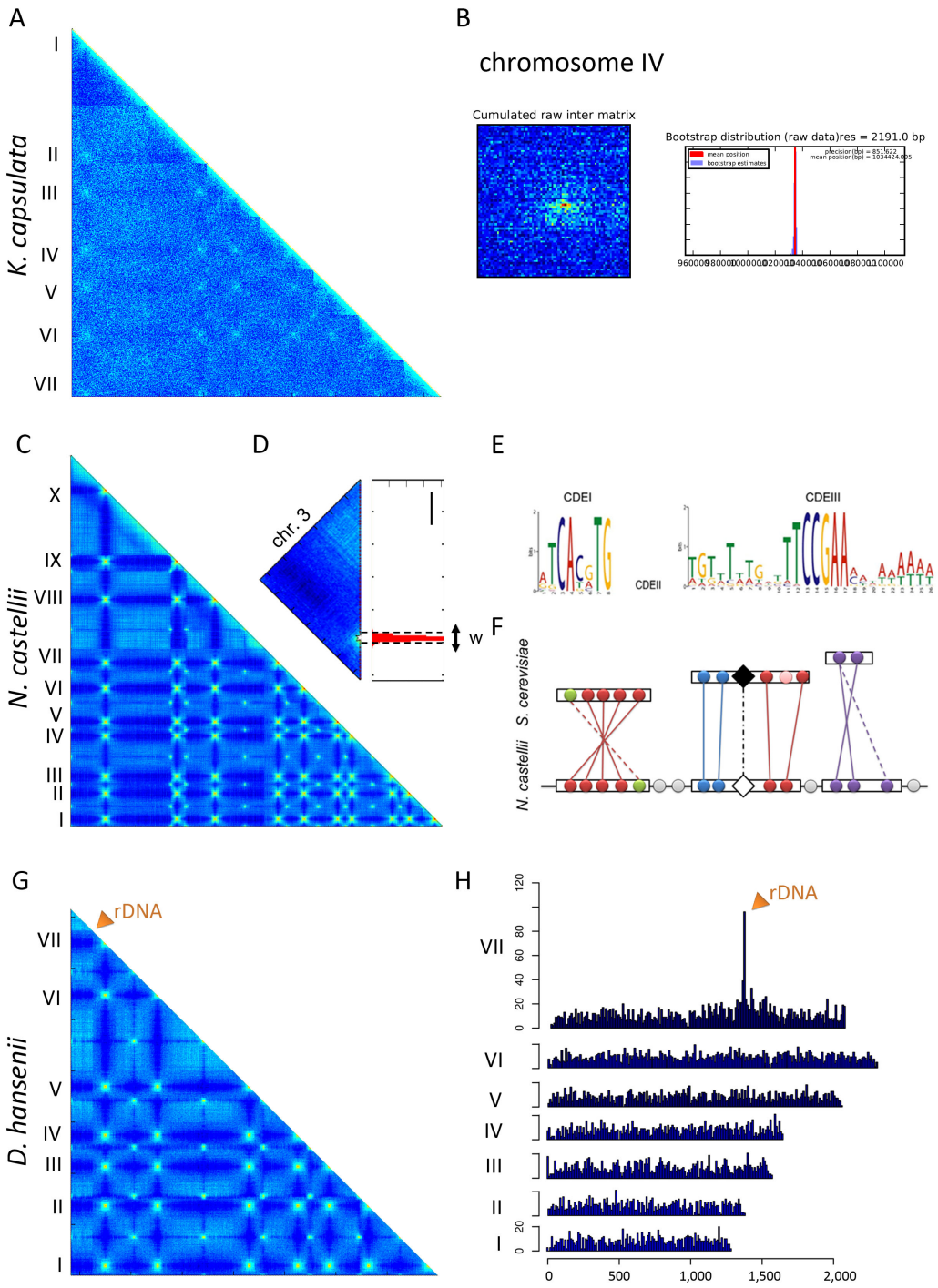


Figure 5.6: (**A**) Identification of CDEI and CDEIII consensus sequences in a mix of intergenic regions from Lachancea centromeric regions and *N. castellii* centromere sequences predicted from the 3C data. The signal identified corresponds only to *Lachancea* sequences. (**B**) Schematic representation of synteny conservation between a centromeric region of *N. castellii* (bottom line) and *S. cerevisiae* (three upper blocks). Grey circles: genes between syntenic blocks (in black rectangles). Full colored circles: conserved genes. Black diamonds: known centromere. Empty diamond: predicted centromere.

corresponding contact matrix (N_{chr} = 10; $N_{reads,Nc}$ = 3,265,947 contacts; Figure 6C). Following through the procedure described above, we characterized for each of the ten chromosomes cis-contact matrixes the genomic intervals containing centromeric regions (Figure 5.6D). From the *S. cerevisiae* analysis, we estimated that the optimal binning for a 3M reads raw contact matrix to 3kb bins and that normalization through SCN was likely to improve the results (Figure 5.4B). Therefore, we generated this matrix and proceeded to superpose the trans-

Centromeres of *Kuraishia capsulata*

#chr.	Predicted* Start	Predicted* End	3C mean position	Precision (bp)
1	466903	470996	470602	2426
2	1546884	1551323	1547201	6741
3	910200	913556	911375	3651
	469800	472200	X	
4	1033064	1035337	1034424	851
	476500	478800	X	
5	574146	576900	572890	2686
6	604123	607946	606172	3120
7*	1101261	1106974	1099909	3548

Centromeres of *Naumovozyma castellii*

#chr.	3C mean position	Precision (bp)	Supported by synteny
1	1047129	681	YES
2	864103	570	YES
3	973309	489	-
4	535959	720	YES
5	576626	1102	-
6	206931	527	YES
7	591720	1055	YES
8	293288	718	-
9	376666	746	-
10	183626	922	YES

Table 5.1

submatrices containing the centromeric regions defined from the cis-contacts. The Gaussian fit was applied as described, and the coordinates of centromere positions along with the precision calculated (Table 5.1).

N. castellii, although positioned within a clade encompassing species where CDE have been identified through computational analysis, is an intriguing exception in this regard (and the same is true for composition-bias searches; Cliften et al., 2006 [22], Gordon, Byrne and Wolfe, 2011 [54]). We hypothesized that CDE sequences may have escaped from former investigations because of important divergence of the consensus sequence, and performed a computational analysis focusing on the region identified through genomic 3C as the ones containing centromeres. First, from the computed coordinates of the centromere of *N. castellii*, intergenic sequences of the flanking coding DNA sequences (CDS) were recovered and submitted to the motif finder algorithm MEME (Lynch et al., 2010 [82]) under the zoops (zero or one motif per sequence) or the oops (only one motif per sequence) modes. No significant motif could be identified from these first analyses. In order to guide the motif finder program, intergenic sequences from *N. castellii* were included into a set of 63 intergenic regions known to contain centromeres from 8 other yeast species (from the *Lachancea* clade). These regions were used as a validation of the CDE I and III detection approach. We clearly identified CDEI and CDEIII consensus sequences (Figure 5.6E) but all of these motifs corresponded to centromere regions of *Lachancea* species, whereas no CDEI and only a very weak CDEIII signal was observed for *N. castellii* regions (and no signature of a CDEII region was found upstream of CDEIII).

For an independent verification, we analyzed the synteny conservation between these chromosomal regions and pericentromeric regions of neighboring species (*S. cerevisiae* and *Zygosaccharomyces rouxii*). To do so, we identified 1) conserved syntenic blocks or 2) synteny breakpoint within the *S. cerevisiae* and *Zygosaccharomyces*

rouxii genomes encompassing or flanking the computed centromeric position in *N. castellii*, respectively. We then determined if the conserved syntenic block also encompassed a centromeric region within these species, or if the synteny breakpoint was consistent with the presence of a centromere in *N. castellii* (Figure 5.6F). Four out of ten centromeres belonged to the first category. Of the remaining six, two felt within the second category, resulting in a computed position compatible with the Ancestral centromeric locations for six out of ten chromosomes. The remaining four centromeres we characterized do not lie in regions of conserved synteny with the two species studied, which can be explained easily by the accumulation of multiple rearrangements that will hide further the evolutionary relatedness of these regions. It is likely that extending this approach to more closely related species will unveil more links and increase the number of centromeres linked to ancestral positions.

Interestingly, Gordon, Byrne and Wolfe (2009, [53]) also sought without success for consensus centromere sequences at putative ancestral centromeric locations in *N. castellii*. Here, we show that, although CDS are not identifiable within these regions, the centromere function is still linked to these ancestral positions for at least six out of ten chromosomes. This suggests that the centromeric binding proteins and/or the mechanisms involved have evolved significantly in this lineage. Interestingly, and perhaps not coincidently, RNA interference is also conserved in this species.

Identification of ribosomal DNA locus in Debaryomyces hansenii

The genome of *D. hansenii*, a cryotolerant and osmotolerant marine yeast important in the agro-food industry, lacks annotation of the ribosomal DNA locus (Dujon et al., 2004 [39]). We generated a genomewide contact matrix of its seven chromosomes (N_reads, Nc = 7,020,925 contacts; Figure 5.6G), and proceeded as described above to identify the position of ribosomal DNA locus (or loci; Material and Methods). We found a peak on chromosome G in the distribution of reads along the genome for which the other mate corresponds to rDNA (Figure 5.6H). By zooming in the distribution, the position of the ribosomal DNA cluster of *D. hansenii* was identified at 1,354,000 pb (Figure 6H). This region corresponds to an intergenic region containing a pseudogene and a gap, according to the published reference genome (Deha2G::1,353,661-1,356,925, available on www.genolevure.org). This region was blasted on the NCBI database, revealing two small (75bp) regions matching with ribosomal DNA at positions 1,354,446 and 1,355,863. We therefore inferred the position of a large, unique ribosomal DNA cluster within this window on chromosome G, ruling out the hypothesis regarding the existence of three intrachromosomal clusters in this genome.

Overall, we showed that genome-wide chromosome conformation capture can be used to unveil important functional elements invisible to standard genomic analyses. It is likely that our standardized procedures will allow to identify other functional elements, such as ribosomal DNA loci, from contact data matrixes.

Chapter 6

Conclusion

In this work we have presented a probabilistic approach for genome assembly from high-throughput chromosome conformation capture data. For this we have first defined the theoretical framework with which we are dealing. We have considered the different levels of spatial genome architecture and current techniques to investigate this architecture. After this we have discussed different techniques of image analysis from both a technical as well as a biological approach. Subsequently, we have shown in an already published paper how we can normalize a chromosomal contact map. This multidisciplinary preparation in combination with a mathematical point of view, allowed us to to develop algorithms that eventually lead to a completely assembled genome. We have finished with a methodology of how we can identify centromeres from contact data.

Up to this point we were unable to read entire genomes directly. Instead, we only disposed of small genomic pieces in the form of contigs. The method we have introduced in this manuscript has the potential, not only to correct and finish assembly, but also to give a robust score to existing genomes. Moreover, we have great confidence that this method could be applied to even more complex situations like metagenomic studies.

6.1 Future work

In the future we would like to continue to improve our method. Even if we have delivered proof and results that our program works, there are some details that have not yet been completely integrated at this point. For example we hope to be able very soon to propose a full sampler algorithm. With this improvement, the program would be fully autonomous.

Further, we have explored some interesting ideas which are less directly related to our main thesis work and have to be left to future work. Below we will shortly elaborate on one of them.

6.1.1 Mating type switching in S.cerevisiae: Tracking of epigenetic signals throughout life cycle

6.1.1.1 Introduction

Yeast s. cerevisiae is a single celled eukaryote which can be either haploid or diploid. Haploid individuals come in two types: a-type and α-type. The sexual orientation is coded by a single locus, MAT-α/a, located on the chromosome III. Each haploid cell produces a specific mating pheromone (a-factor for a-cells and α-factor for

α-cells) which allows an opposite type cell to detect the presence of a mating partner. Mating is only possible between opposite types and results in a new diploid cell which possesses the two sexual types.

Besides this well known behavior, yeast cells have the striking possibility to switch mating type every generation by means of a highly sophisticated mechanism of site-specific recombination (Haber, 1998[60]). First, a site specific endonuclease HO creates a single double strand break within the MAT locus. Exonucleases cut the DNA ends resulting in the destruction of the MAT allele. Silenced alleles of MAT-a and MAT-α are located at the extremity of the chromosome III. The silencing of those two regions called HMR (Hidden Mat Right) and HML (Hidden Mat Left) is performed by a tightly regulated epigenetic mechanism. The repair of the DSS is almost always performed by using the opposite MAT allele of the cell (Coïc et al., 2011[23]). Therefore, MAT-α (resp MAT-a) will recombine with hidden MAT-a (resp MAT-α). It has been shown that donor preference is independent of the sequences carried by MAT or within the hidden MAT regions. This preferential selection is due to an element called recombination enhancer (RE), located 17kb away from the HML on the left arm of the chromosome III (figure 6.1). The epigenetic regulation of RE is specific to the cell type and to the transcription activity (Ercan et al., 2005[42]).

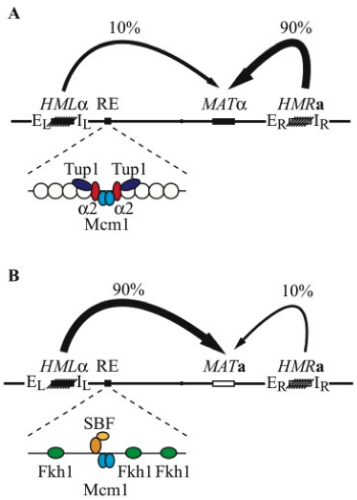


Figure 6.1: RE regulation and donor preference during mating type switching in yeast (Coïc et al., 2011[23]).

6.1.1.2 Tracking of epigenetic signals throughout life cycle

Imprinting is the mechanism by which mammals are able to distinguish the paternal or maternal origin of an allele. Therefore, vital messages are transmitted from the parents to their offspring without modification of the genetic information. Perturbation of this process can result in cancer or in dramatic diseases (i.e. Beckwith–Wiedemann syndrome).

The overall goal of this proposal is to answer the following questions:

- **Can the process of mating-type switching be a sophisticated mechanism to carry non genetic information throughout the life cycle of yeast?**
- Optionally, would it be possible to store data and start computing in a yeast network?

Since we want to track information over generations it will be necessary:

- to define explicitly observable epigenetics events;

- to define a set of events which could yield stress to the cells;
- to develop an automated live imaging system able to locate events at high spatial and time resolution;
- to develop the statistical methods and computing tools to detect relevant variations in the acquired signals.

6.1.1.3 Biological and imaging experiments

One of the most direct consequences of epigenetic modification is the change of flexibility of the chromatin. For instance, it has been shown that the mobility of the left arm of the chromosome III induces a shift in the donor choice during mating-type switch. Bressan, Vazquez and Haber (2004[13]) have constructed a new strain of yeast allowing the 3d positionning of the HML, HMR and MAT (figure 6.2).

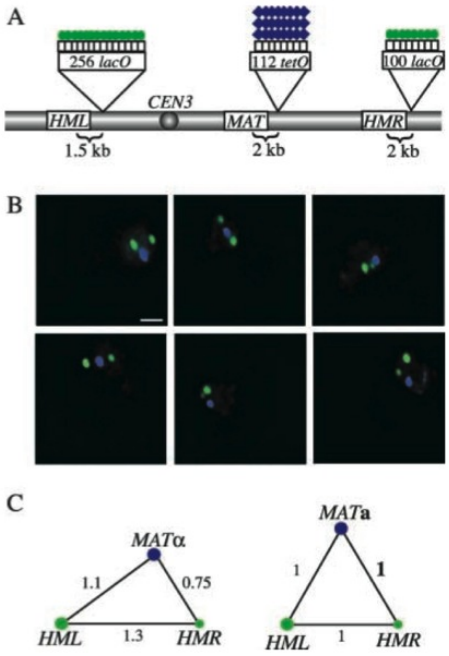


Figure 6.2: 3D positioning of tagged mating-type loci by fluorescence microscopy (Bressan, Vazquez and Haber, 2004[13]).

This construction would allow us to track the way DSB HO induced is repaired during mating-type switch events. However, it will still be necessary to construct a strain with one of the two hidden MAT labeled (TetR-gfp for instance) to distinguish α cells from a cells.

6.1.1.4 Statistics and computing tools

To verify if there are any information leakages at any step of the cell cycle we will have to compute all the transition probabilities and check the independences between mothers and daughters (figure 6.3). The best mathematical tool to study these complex relations would be the Bayesian network framework.

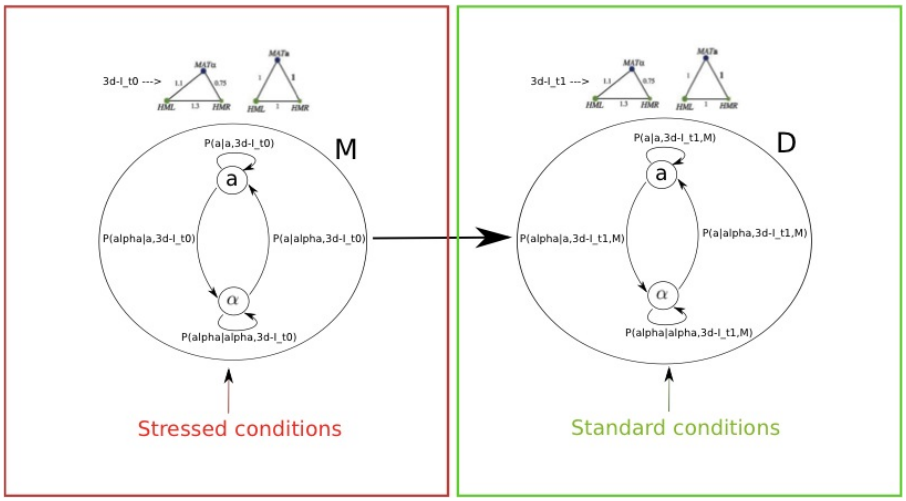


Figure 6.3: Mating-type switching process. The transition probabilities are modified by hidden epigenetic factors. The live imaging system allows us to infer some of those activities. M refers to "mother" and D to "daughter". Robust and reliable statistical tools will have to be implemented in order to track any relevant leakages.

Bibliography

[1] James R Abney, Bryan Cutler, Misty L Fillbach, Daniel Axelrod, and Bethe A Scalettar. Chromatin dynamics in interphase nuclei and its implications for nuclear structure. *The Journal of cell biology*, 137(7):1459–1468, 1997.

[2] Can Alkan, Saba Sajjadian, and Evan E Eichler. Limitations of next-generation genome sequence assembly. *Nature methods*, 8(1):61–65, 2010.

[3] Christophe Andrieu, Nando De Freitas, Arnaud Doucet, and Michael I Jordan. An introduction to mcmc for machine learning. *Machine learning*, 50(1-2):5–43, 2003.

[4] Nathalie J Arhel, Sylvie Souquere-Besse, Sandie Munier, Philippe Souque, Stéphanie Guadagnini, Sandra Rutherford, Marie-Christine Prévost, Terry D Allen, and Pierre Charneau. Hiv-1 dna flap formation promotes uncoating of the pre-integration complex at the nuclear pore. *The EMBO journal*, 26(12):3025–3037, 2007.

[5] Timothy L Bailey and Charles Elkan. Fitting a mixture model by expectation maximization to discover motifs in bipolymers, 1994.

[6] Monya Baker. De novo genome assembly: what every biologist should know. *Nature Methods*, 9(4):333–337, 2012.

[7] Marisa S Bartolomei, Sharon Zemel, and Shirley M Tilghman. Parental imprinting of the mouse h19 gene. *Nature*, 351(6322):153–155, 1991.

[8] Yuval Benjamini and Terence P Speed. Summarizing and correcting the gc content bias in high-throughput sequencing. *Nucleic acids research*, 40(10):e72–e72, 2012.

[9] Rosa Bernardi and Pier Paolo Pandolfi. Structure, dynamics and functions of promyelocytic leukaemia nuclear bodies. *Nature Reviews Molecular Cell Biology*, 8(12):1006–1016, 2007.

[10] Andreas Bolzer, Gregor Kreth, Irina Solovei, Daniela Koehler, Kaan Saracoglu, Christine Fauth, Stefan Müller, Roland Eils, Christoph Cremer, Michael R Speicher, et al. Three-dimensional maps of all chromosomes in human male fibroblast nuclei and prometaphase rosettes. *PLoS biology*, 3(5):e157, 2005.

[11] Giuseppe Borsani, Rossana Tonlorenzi, M Christine Simmler, Luisa Dandolo, Danielle Arnaud, Valeria Capra, Markus Grompe, Antonio Pizzuti, Donna Muzny, Charles Lawrence, et al. Characterization of a murine gene expressed from the inactive x chromosome. *Nature*, 351(6324):325–329, 1991.

[12] Miguel R Branco and Ana Pombo. Intermingling of chromosome territories in interphase suggests role in translocations and transcription-dependent associations. *PLoS biology*, 4(5):e138, 2006.

[13] Debra A Bressan, Julio Vazquez, and James E Haber. Mating type–dependent constraints on the mobility of the left arm of yeast chromosome iii. *The Journal of cell biology*, 164(3):361–371, 2004.

[14] Neil Brockdorff, Alan Ashworth, Graham F Kay, Penny Cooper, Sandy Smith, Veronica M McCabe, Dominic P Norris, Graeme D Penny, Dipika Patel, and Sohaila Rastan. Conservation of position and exclusive expression of mouse xist from the inactive x chromosome. *Nature*, 351(6324):329–331, 1991.

[15] Carolyn J Brown, Andrea Ballabio, James L Rupert, Ronald G Lafreniere, Markus Grompe, Rossana Tonlorenzi, and Huntington F Willard. A gene from the region of the human x inactivation centre is expressed exclusively from the inactive x chromosome. *Nature*, 349(6304):38–44, 1991.

[16] Ghislain G Cabal, Auguste Genovesio, Susana Rodriguez-Navarro, Christophe Zimmer, Olivier Gadal, Annick Lesne, Henri Buc, Frank Feuerbach-Fournier, Jean-Christophe Olivo-Marin, Eduard C Hurt, et al. Saga interacting factors confine sub-diffusion of transcribed genes to the nuclear envelope. *Nature*, 441(7094):770–773, 2006.

[17] Séverine Chambeyron and Wendy A Bickmore. Chromatin decondensation and nuclear reorganization of the hoxb locus upon induction of transcription. *Genes & development*, 18(10):1119–1130, 2004.

[18] Julie Chaumeil, Patricia Le Baccon, Anton Wutz, and Edith Heard. A novel role for xist rna in the formation of a repressive nuclear compartment into which genes are recruited when silenced. *Genes & development*, 20(16):2223–2237, 2006.

[19] Chien-Hui Chuang, Anne E Carpenter, Beata Fuchsova, Terezina Johnson, Primal de Lanerolle, and Andrew S Belmont. Long-range directional movement of an interphase chromosome site. *Current Biology*, 16(8):825–831, 2006.

[20] Jonathan R Chubb, Shelagh Boyle, Paul Perry, and Wendy A Bickmore. Chromatin motion is constrained by association with nuclear compartments in human cells. *Current Biology*, 12(6):439–445, 2002.

[21] Christine Moulton Clemson, John A McNeil, Huntington F Willard, and Jeanne Bentley Lawrence. Xist rna paints the inactive x chromosome at interphase: evidence for a novel rna involved in nuclear/chromosome structure. *Journal of Cell Biology*, 132(3):259–276, 1996.

[22] Paul F Cliften, Robert S Fulton, Richard K Wilson, and Mark Johnston. After the duplication: gene loss and adaptation in saccharomyces genomes. *Genetics*, 172(2):863–872, 2006.

[23] Eric Coïc, Joshua Martin, Taehyun Ryu, Sue Yen Tay, Jané Kondev, and James E Haber. Dynamics of homology searching during gene conversion in saccharomyces cerevisiae revealed by donor competition. *Genetics*, 189(4):1225–1233, 2011.

[24] Phillip EC Compeau, Pavel A Pevzner, and Glenn Tesler. How to apply de bruijn graphs to genome assembly. *Nature biotechnology*, 29(11):987–991, 2011.

[25] Le Cong, Ruhong Zhou, Yu-chi Kuo, Margaret Cunniff, and Feng Zhang. Comprehensive interrogation of natural tale dna-binding modules and transcriptional repressor domains. *Nature communications*, 3:968, 2012.

[26] Mauricio Corredor, Anne-Marie Davila, Serge Casaregola, and Claude Gaillardin. Chromosomal polymorphism in the yeast species debaryomyces hansenii. *Antonie van Leeuwenhoek*, 84(2):81–88, 2003.

[27] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. Normalization of a chromosomal contact map. *BMC genomics*, 13(1):436, 2012.

[28] Christoph Cremer, Thomas Cremer, and JW Gray. Induction of chromosome damage by ultraviolet light and caffeine: correlation of cytogenetic evaluation and flow karyotype. *Cytometry*, 2(5):287–290, 1982.

[29] T Cremer and C Cremer. Rise, fall and resurrection of chromosome territories: a historical perspective part ii. fall and resurrection of chromosome territories during the 1950s to 1980s. part iii. chromosome territories and the functional nuclear architecture: experiments and m. *European Journal of Histochemistry*, 50(4):223–272, 2009.

[30] Thomas Cremer and Marion Cremer. Chromosome territories. *Cold Spring Harbor perspectives in biology*, 2(3), 2010.

[31] Wouter de Laat and Job Dekker. 3c-based technologies to study the shape of the genome. *Methods*, 58(3):189–191, 2012.

[32] Job Dekker, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. Capturing chromosome conformation. *science*, 295(5558):1306–1311, 2002.

[33] Niall Dillon. Heterochromatin structure and function. *Biology of the Cell*, 96(8):631–637, 2004.

[34] Jesse R Dixon, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, 2012.

[35] Josée Dostie, Todd A Richmond, Ramy A Arnaout, Rebecca R Selzer, William L Lee, Tracey A Honan, Eric D Rubio, Anton Krumm, Justin Lamb, Chad Nusbaum, et al. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome research*, 16(10):1299–1309, 2006.

[36] David A Drubin, Arman M Garakani, and Pamela A Silver. Motion as a phenotype: the use of live-cell imaging and machine visual screening to characterize transcription-dependent chromosome dynamics. *BMC cell biology*, 7(1):19, 2006.

[37] Sheona P Drummond, Sandra A Rutherford, Helen S Sanderson, and Terry D Allen. High resolution analysis of mammalian nuclear structure throughout the cell cycle: implications for nuclear pore complex assembly during interphase and mitosis this paper is one of a selection of papers published in this special issue, entitled the nucleus: A cell within a cell. *Canadian journal of physiology and pharmacology*, 84(3-4):423–430, 2006.

[38] Zhijun Duan, Mirela Andronescu, Kevin Schutz, Sean McIlwain, Yoo Jung Kim, Choli Lee, Jay Shendure, Stanley Fields, C Anthony Blau, and William S Noble. A three-dimensional model of the yeast genome. *Nature*, 465(7296):363–367, 2010.

[39] Bernard Dujon, David Sherman, Gilles Fischer, Pascal Durrens, Serge Casaregola, Ingrid Lafontaine, Jacky De Montigny, Christian Marck, Cécile Neuvéglise, Emmanuel Talla, et al. Genome evolution in yeasts. *Nature*, 430(6995):35–44, 2004.

[40] Mikhail Eltsov, Kirsty M MacLellan, Kazuhiro Maeshima, Achilleas S Frangakis, and Jacques Dubochet. Analysis of cryo-electron microscopy images does not support the existence of 30-nm chromatin fibers in mitotic chromosomes in situ. *Proceedings of the National Academy of Sciences*, 105(50):19732–19737, 2008.

[41] Jesse M Engreitz, Amy Pandya-Jones, Patrick McDonel, Alexander Shishkin, Klara Sirokman, Christine Surka, Sabah Kadri, Jeffrey Xing, Alon Goren, Eric S Lander, et al. The xist lncrna exploits three-dimensional genome architecture to spread across the x chromosome. *Science*, 341(6147), 2013.

[42] Sevinc Ercan, Joseph C Reese, Jerry L Workman, and Robert T Simpson. Yeast recombination enhancer is stimulated by transcription activation. *Molecular and cellular biology*, 25(18):7976–7987, 2005.

[43] Andrew P Feinberg. Phenotypic plasticity and the epigenetics of human disease. *Nature*, 447(7143):433–440, 2007.

[44] Carmelo Ferrai, Sheila Q Xie, Paolo Luraghi, Davide Munari, Francisco Ramirez, Miguel R Branco, Ana Pombo, and Massimo P Crippa. Poised transcription factories prime silent upa gene prior to activation. *PLoS biology*, 8(1):e1000270, 2010.

[45] JT Finch, LC Lutter, D Rhodes, RS Brown, B Rushton, M Levitt, and A Klug. Structure of nucleosome core particles of chromatin. 1977.

[46] Archa H Fox, Yun Wah Lam, Anthony KL Leung, Carol E Lyon, Jens Andersen, Matthias Mann, and Angus I Lamond. Paraspeckles: a novel nuclear domain. *Current Biology*, 12(1):13–25, 2002.

[47] Alejandro F Frangi, Wiro J Niessen, Koen L Vincken, and Max A Viergever. Multiscale vessel enhancement filtering. In *Medical Image Computing and Computer-Assisted Interventation—MICCAI'98*, pages 130–137. Springer, 1998.

[48] Eden Fussner, Reagan W Ching, and David P Bazett-Jones. Living without 30nm chromatin fibers. *Trends in biochemical sciences*, 36(1):1–6, 2011.

[49] Alexey A Gavrilov, Ekaterina S Gushchanskaya, Olga Strelkova, Oksana Zhironkina, Igor I Kireev, Olga V Iarovaia, and Sergey V Razin. Disclosure of a structural milieu for the proximity ligation reveals the elusive nature of an active chromatin hub. *Nucleic acids research*, 41(6):3563–3575, 2013.

[50] Stuart Geman and Donald Geman. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

[51] Emmanuelle Gérard, François Guyot, Pascal Philippot, and Purificación López-García. Fluorescence< i> in situ</i> hybridisation coupled to ultra small immunogold detection to identify prokaryotic cells using transmission and scanning electron microscopy. *Journal of microbiological methods*, 63(1):20–28, 2005.

[52] SE Gerchman and V Ramakrishnan. Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. *Proceedings of the National Academy of Sciences*, 84(22):7802–7806, 1987.

[53] Jonathan L Gordon, Kevin P Byrne, and Kenneth H Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern saccharomyces cerevisiae genome. *PLoS Genetics*, 5(5):e1000485, 2009.

[54] Jonathan L Gordon, Kevin P Byrne, and Kenneth H Wolfe. Mechanisms of chromosome number evolution in yeast. *PLoS genetics*, 7(7):e1002190, 2011.

[55] J. C. Gower. Properties of euclidian and non euclidian distance matrices. 61:81–97, 1985.

[56] Manfred G Grabherr, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, et al. Full-length transcriptome assembly from rna-seq data without a reference genome. *Nature biotechnology*, 29(7):644–652, 2011.

[57] Phil Green. Phrap documentation, 1994.

[58] A Grosberg, Y Rabin, S Havlin, and A Neer. Crumpled globule model of the three-dimensional structure of dna. *EPL (Europhysics Letters)*, 23(5):373, 1993.

[59] A Yu Grosberg, Sergei K Nechaev, and Eugene I Shakhnovich. The role of topological constraints in the kinetics of collapse of macromolecules. *Journal de physique*, 49(12):2095–2100, 1988.

[60] James E Haber. Mating-type gene switching in saccharomyces cerevisiae. *Annual review of genetics*, 32(1):561–599, 1998.

[61] Ichiro Hiratani, Tyrone Ryba, Mari Itoh, Tomoki Yokochi, Michaela Schwaiger, Chia-Wei Chang, Yung Lyou, Tim M Townes, Dirk Schübeler, and David M Gilbert. Global reorganization of replication domains during embryonic stem cell differentiation. *PLoS biology*, 6(10):e245, 2008.

[62] Mark Howison, Felipe Zapata, and Casey W Dunn. Toward a statistically explicit understanding of de novo sequence assembly. *Bioinformatics*, 29(23):2959–2963, 2013.

[63] Maxim Imakaev, Geoffrey Fudenberg, Rachel Patton McCord, Natalia Naumova, Anton Goloborodko, Bryan R Lajoie, Job Dekker, and Leonid A Mirny. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nature Methods*, 9(10):999–1003, 2012.

[64] Noémie Jacques, Christine Sacerdot, Meriem Derkaoui, Bernard Dujon, Odile Ozier-Kalogeropoulos, and Serge Casaregola. Population polymorphism of nuclear mitochondrial dna insertions reveals widespread diploidy associated with loss of heterozygosity in debaryomyces hansenii. *Eukaryotic cell*, 9(3):449–459, 2010.

[65] Edwin T Jaynes. *Probability theory: the logic of science*. Cambridge university press, 2003.

[66] Stephan Kadauke and Gerd A Blobel. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms*, 1789(1):17–25, 2009.

[67] Manolis Kellis, Bruce W Birren, and Eric S Lander. Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, 428(6983):617–624, 2004.

[68] Elena Kiseleva, Sheona P Drummond, Martin W Goldberg, Sandra A Rutherford, Terence D Allen, and Katherine L Wilson. Actin-and protein-4.1-containing filaments link nuclear pore complexes to subnuclear organelles in xenopus oocyte nuclei. *Journal of cell science*, 117(12):2481–2490, 2004.

[69] Andreas Klöckner, Nicolas Pinto, Yunsup Lee, Bryan Catanzaro, Paul Ivanov, Ahmed Fasih, AD Sarma, D Nanongkai, G Pandurangan, P Tetali, et al. Pycuda: Gpu run-time code generation for high-performance computing. *Arxiv preprint arXiv*, 911, 2009.

[70] Claudia Köhler, Philip Wolff, and Charles Spillane. Epigenetic mechanisms underlying genomic imprinting in plants. *Annual review of plant biology*, 63:331–352, 2012.

[71] Roger D Kornberg. Chromatin structure: a repeating unit of histones and dna. *Science*, 184(4139):868–871, 1974.

[72] Steven T Kosak, Jane A Skok, Kay L Medina, Roy Riblet, Michelle M Le Beau, Amanda G Fisher, and Harinder Singh. Subnuclear compartmentalization of immunoglobulin loci during lymphocyte development. *Science*, 296(5565):158–162, 2002.

[73] Sreenivasulu Kurukuti, Vijay Kumar Tiwari, Gholamreza Tavoosidana, Elena Pugacheva, Adele Murrell, Zhihu Zhao, Victor Lobanenkov, Wolf Reik, and Rolf Ohlsson. Ctcf binding at the h19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to igf2. *Proceedings of the National Academy of Sciences*, 103(28):10684–10689, 2006.

[74] Ernest T Lam, Alex Hastie, Chin Lin, Dean Ehrlich, Somes K Das, Michael D Austin, Paru Deshpande, Han Cao, Niranjan Nagarajan, Ming Xiao, et al. Genome mapping on nanochannel arrays for structural variation analysis and sequence assembly. *Nature biotechnology*, 2012.

[75] Angus I Lamond and David L Spector. Nuclear speckles: a model for nuclear organelles. *Nature Reviews Molecular Cell Biology*, 4(8):605–612, 2003.

[76] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nature methods*, 9(4):357–359, 2012.

[77] Ruiqiang Li, Hongmei Zhu, Jue Ruan, Wubin Qian, Xiaodong Fang, Zhongbin Shi, Yingrui Li, Shengting Li, Gao Shan, Karsten Kristiansen, et al. De novo assembly of human genomes with massively parallel short read sequencing. *Genome research*, 20(2):265–272, 2010.

[78] 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, et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science*, 326(5950):289, 2009.

[79] Jun S Liu, Faming Liang, and Wing Hung Wong. The multiple-try method and local optimization in metropolis sampling. *Journal of the American Statistical Association*, 95(449):121–134, 2000.

[80] Vett Lloyd. Parental imprinting in drosophila. *Genetica*, 109(1-2):35–44, 2000.

[81] Stavros Lomvardas, Gilad Barnea, David J Pisapia, Monica Mendelsohn, Jennifer Kirkland, and Richard Axel. Interchromosomal interactions and olfactory receptor choice. *Cell*, 126(2):403–413, 2006.

[82] Denise B Lynch, Mary E Logue, Geraldine Butler, and Kenneth H Wolfe. Chromosomal g+ c content evolution in yeasts: systematic interspecies differences, and gc-poor troughs at centromeres. *Genome biology and evolution*, 2:572, 2010.

[83] Kazuhiro Maeshima and Mikhail Eltsov. Packaging the genome: the structure of mitotic chromosomes. *Journal of biochemistry*, 143(2):145–153, 2008.

[84] York Marahrens, Barbara Panning, Jessica Dausman, William Strauss, and Rudolf Jaenisch. Xist-deficient mice are defective in dosage compensation but not spermatogenesis. *Genes & Development*, 11(2):156–166, 1997.

[85] Lynn Margulis. *Origin of eukaryotic cells: Evidence and research implications for a theory of the origin and evolution of microbial, plant, and animal cells on the Precambrian earth.* Yale University Press New Haven, 1970.

[86] Clement L Markert. Parthenogenesis, homozygosity, and cloning in mammals. *Journal of Heredity*, 73(6):390–397, 1982.

[87] WF Marshall, A Straight, JF Marko, J Swedlow, A Dernburg, A Belmont, AW Murray, DA Agard, and JW Sedat. Interphase chromosomes undergo constrained diffusional motion in living cells. *Current Biology*, 7(12):930–939, 1997.

[88] Diego Martinez, Randy M Berka, Bernard Henrissat, Markku Saloheimo, Mikko Arvas, Scott E Baker, Jarod Chapman, Olga Chertkov, Pedro M Coutinho, Dan Cullen, et al. Genome sequencing and analysis of the biomass-degrading fungus trichoderma reesei (syn. hypocrea jecorina). *Nature biotechnology*, 26(5):553–560, 2008.

[89] James McGrath and Davor Solter. Completion of mouse embryogenesis requires both the maternal and paternal genomes. *Cell*, 37(1):179–183, 1984.

[90] Peter Meister, Benjamin D Towbin, Brietta L Pike, Aaron Ponti, and Susan M Gasser. The spatial dynamics of tissue-specific promoters during c. elegans development. *Genes & development*, 24(8):766–782, 2010.

[91] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21:1087, 1953.

[92] Nicholas Metropolis and Stanislaw Ulam. The monte carlo method. *Journal of the American statistical association*, 44(247):335–341, 1949.

[93] OL Miller and Barbara R Beatty. Visualization of nucleolar genes. *Science*, 164(3882):955–957, 1969.

[94] Leonid A Mirny. The fractal globule as a model of chromatin architecture in the cell. *Chromosome research*, 19(1):37–51, 2011.

[95] Yusuke Miyanari, Céline Ziegler-Birling, and Maria-Elena Torres-Padilla. Live visualization of chromatin dynamics with fluorescent tales. *Nature structural & molecular biology*, 20(11):1321–1324, 2013.

[96] Ian M Morison, Joshua P Ramsay, and Hamish G Spencer. A census of mammalian imprinting. *TRENDS in Genetics*, 21(8):457–465, 2005.

[97] Takashi Nagano, Yaniv Lubling, Tim J Stevens, Stefan Schoenfelder, Eitan Yaffe, Wendy Dean, Ernest D Laue, Amos Tanay, and Peter Fraser. Single-cell hi-c reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469):59–64, 2013.

[98] Attila Németh, Ana Conesa, Javier Santoyo-Lopez, Ignacio Medina, David Montaner, Bálint Péterfia, Irina Solovei, Thomas Cremer, Joaquin Dopazo, and Gernot Längst. Initial genomics of the human nucleolus. *PLoS genetics*, 6(3):e1000889, 2010.

[99] Elphège P Nora, Bryan R Lajoie, Edda G Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, Nynke L van Berkum, Johannes Meisig, John Sedat, et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature*, 485(7398):381–385, 2012.

[100] Raimund J Ober, Sripad Ram, and E Sally Ward. Localization accuracy in single-molecule microscopy. *Biophysical journal*, 86(2):1185–1200, 2004.

[101] Mark O'Driscoll and Penny A Jeggo. The role of double-strand break repair—insights from human genetics. *Nature Reviews Genetics*, 7(1):45–54, 2006.

[102] Ada L Olins and Donald E Olins. Spheroid chromatin units (ν bodies). *Science*, 183(4122):330–332, 1974.

[103] Cameron S Osborne, Lyubomira Chakalova, Karen E Brown, David Carter, Alice Horton, Emmanuel Debrand, Beatriz Goyenechea, Jennifer A Mitchell, Susana Lopes, Wolf Reik, et al. Active genes dynamically colocalize to shared sites of ongoing transcription. *Nature genetics*, 36(10):1065–1071, 2004.

[104] Cameron S Osborne, Philip A Ewels, and Alice NC Young. Meet the neighbours: tools to dissect nuclear structure and function. *Briefings in Functional Genomics*, 10(1):11–17, 2011.

[105] P Oudet, M Gross-Bellard, and P Chambon. Electron microscopic and biochemical evidence that chromatin structure is a repeating unit. *Cell*, 4(4):281–300, 1975.

[106] Pranav Oza, Sue L Jaspersen, Adriana Miele, Job Dekker, and Craig L Peterson. Mechanisms that regulate localization of a dna double-strand break to the nuclear periphery. *Genes & development*, 23(8):912–927, 2009.

[107] Eileen T O'Toole, Mark Winey, and J Richard McIntosh. High-voltage electron tomography of spindle pole bodies and early mitotic spindles in the yeast saccharomyces cerevisiae. *Molecular biology of the cell*, 10(6):2017–2031, 1999.

[108] Graeme D Penny, Graham F Kay, Steven A Sheardown, Sohaila Rastan, and Neil Brockdorff. Requirement for xist in x chromosome inactivation. *Nature*, 379(6561):131–137, 1996.

[109] Robert D Phair and Tom Misteli. High mobility of proteins in the mammalian cell nucleus. *Nature*, 404(6778):604–609, 2000.

[110] Miron Prokocimer, Maya Davidovich, Malka Nissim-Rafinia, Naama Wiesel-Motiuk, Daniel Z Bar, Rachel Barkan, Eran Meshorer, and Yosef Gruenbaum. Nuclear lamins: key regulators of nuclear structure and activities. *Journal of cellular and molecular medicine*, 13(6):1059–1085, 2009.

[111] Tobias Ragoczy, MA Bender, Agnes Telling, Rachel Byron, and Mark Groudine. The locus control region is required for association of the murine β-globin locus with engaged transcription factories during erythroid maturation. *Genes & development*, 20(11):1447–1457, 2006.

[112] Wolf Reik and Jörn Walter. Genomic imprinting: parental influence on the genome. *Nature Reviews Genetics*, 2(1):21–32, 2001.

[113] Marilyn B Renfree, Timothy A Hore, Geoffrey Shaw, Jennifer A Marshall Graves, and Andrew J Pask. Evolution of genomic imprinting: insights from marsupials and monotremes. *Annual review of genomics and human genetics*, 10:241–262, 2009.

[114] W. Rieping, M. Nilges, and M. Habeck. Isd: a software package for bayesian nmr structure calculation. *Bioinformatics*, 24(8):1104–1105, 2008.

[115] Wolfgang Rieping, Michael Habeck, and Michael Nilges. Inferential structure determination. *Science*, 309(5732):303–306, 2005.

[116] K. Rippe. Making contacts on a nucleic acid polymer. *Trends in biochemical sciences*, 26(12):733–740, 2001.

[117] CDM Rodley, F Bertels, B Jones, and JM O'Sullivan. Global identification of yeast chromosome interactions using genome conformation capture. *Fungal Genetics and Biology*, 46(11):879–886, 2009.

[118] Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher A Bristow, Lijia Ma, Michael F Lin, et al. Identification of functional elements and regulatory circuits by drosophila modencode. *Science*, 330(6012):1787–1797, 2010.

[119] Michael Rubinstein and Ralph H Colby. *Polymer physics*. OUP Oxford, 2003.

[120] Tyrone Ryba, Ichiro Hiratani, Junjie Lu, Mari Itoh, Michael Kulik, Jinfeng Zhang, Thomas C Schulz, Allan J Robins, Stephen Dalton, and David M Gilbert. Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. *Genome research*, 20(6):761–770, 2010.

[121] RK Sachs, G Van Den Engh, B Trask, H Yokota, and JE Hearst. A random-walk/giant-loop model for interphase chromosomes. *Proceedings of the National Academy of Sciences*, 92(7):2710–2714, 1995.

[122] Saurabh Saha, Andrew B Sparks, Carlo Rago, Viatcheslav Akmaev, Clarence J Wang, Bert Vogelstein, Kenneth W Kinzler, and Victor E Velculescu. Using the transcriptome to annotate the genome. *Nature biotechnology*, 20(5):508–512, 2002.

[123] Frederick Sanger, Steven Nicklen, and Alan R Coulson. Dna sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences*, 74(12):5463–5467, 1977.

[124] Michaela Schwaiger, Hubertus Kohler, Edward J Oakeley, Michael B Stadler, and Dirk Schübeler. Heterochromatin protein 1 (hp1) modulates replication timing of the drosophila genome. *Genome research*, 20(6):771–780, 2010.

[125] Yuri B Schwartz, Tatyana G Kahn, Per Stenberg, Katsuhito Ohno, Richard Bourgon, and Vincenzo Pirrotta. Alternative epigenetic chromatin states of polycomb target genes. *PLoS genetics*, 6(1):e1000805, 2010.

[126] Tom Sexton, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. Three-dimensional folding and functional organization principles of the< i> drosophila</i> genome. *Cell*, 148(3):458–472, 2012.

[127] Lindsay S Shopland, Christopher R Lynch, Kevin A Peterson, Kathleen Thornton, Nick Kepper, Johann von Hase, Stefan Stein, Sarah Vincent, Kelly R Molloy, Gregor Kreth, et al. Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. *The Journal of cell biology*, 174(1):27–38, 2006.

[128] Marieke Simonis, Petra Klous, Erik Splinter, Yuri Moshkin, Rob Willemsen, Elzo de Wit, Bas van Steensel, and Wouter de Laat. Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture–on-chip (4c). *Nature genetics*, 38(11):1348–1354, 2006.

[129] Charalampos G Spilianakis, Maria D Lalioti, Terrence Town, Gap Ryol Lee, and Richard A Flavell. Interchromosomal associations between alternatively expressed loci. *Nature*, 435(7042):637–645, 2005.

[130] Alexander Stark, Michael F Lin, Pouya Kheradpour, Jakob S Pedersen, Leopold Parts, Joseph W Carlson, Madeline A Crosby, Matthew D Rasmussen, Sushmita Roy, Ameya N Deoras, et al. Discovery of functional elements in 12 drosophila genomes using evolutionary signatures. *Nature*, 450(7167):219–232, 2007.

[131] MAH Surani, SC Barton, and ML Norris. Development of reconstituted mouse eggs suggests imprinting of the genome during gametogenesis. 1984.

[132] Granger G Sutton, Owen White, Mark D Adams, and Anthony R Kerlavage. Tigr assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology*, 1(1):9–19, 1995.

[133] Angela Taddei, Griet Van Houwe, Florence Hediger, Veronique Kalck, Fabien Cubizolles, Heiko Schober, and Susan M Gasser. Nuclear pore association confers optimal expression levels for an inducible yeast gene. *Nature*, 441(7094):774–778, 2006.

[134] Nobuo Takagi and Motomichi Sasaki. Preferential inactivation of the paternally derived x chromosome in the extraembryonic membranes of the mouse. 1975.

[135] Rajika Thakar, Geoff Gordon, and Amy K Csink. Dynamics and anchoring of heterochromatic loci during development. *Journal of cell science*, 119(20):4165–4175, 2006.

[136] D Thomann, DR Rines, PK Sorger, and G Danuser. Automatic fluorescent tag detection in 3d with super-resolution: application to the analysis of chromosome movement. *Journal of Microscopy*, 208(1):49–64, 2002.

[137] Russell E Thompson, Daniel R Larson, and Watt W Webb. Precise nanometer localization analysis for individual fluorescent probes. *Biophysical journal*, 82(5):2775–2783, 2002.

[138] Bas Tolhuis, Inhua Muijrers, Elzo de Wit, Hans Teunissen, Wendy Talhout, Bas van Steensel, and Maarten van Lohuizen. Genome-wide profiling of prc1 and prc2 polycomb chromatin binding in drosophila melanogaster. *Nature genetics*, 38(6):694–699, 2006.

[139] Bas Tolhuis, Robert-Jan Palstra, Erik Splinter, Frank Grosveld, and Wouter de Laat. Looping and interaction between hypersensitive sites in the active β-globin locus. *Molecular cell*, 10(6):1453–1465, 2002.

[140] Silvana van Koningsbruggen, Marek Gierliński, Pietá Schofield, David Martin, Geoffey J Barton, Yavuz Ariyurek, Johan T den Dunnen, and Angus I Lamond. High-resolution whole-genome sequencing reveals that specific chromatin domains from most human chromosomes associate with nucleoli. *Molecular biology of the cell*, 21(21):3735–3748, 2010.

[141] Julio Vazquez, Andrew S Belmont, and John W Sedat. Multiple regimes of constrained chromosome motion are regulated in the interphase< i> drosophila</i> nucleus. *Current Biology*, 11(16):1227–1239, 2001.

[142] B. Walsh. Markov chain monte carlo and gibbs sampling. 2004.

[143] Jie Wang, Jiali Zhuang, Sowmya Iyer, XinYing Lin, Troy W Whitfield, Melissa C Greven, Brian G Pierce, Xianjun Dong, Anshul Kundaje, Yong Cheng, et al. Sequence features and chromatin structure around the genomic regions bound by 119 human transcription factors. *Genome research*, 22(9):1798–1812, 2012.

[144] Bo Wen, Hao Wu, Yoichi Shinkai, Rafael A Irizarry, and Andrew P Feinberg. Large histone h3 lysine 9 dimethylated chromatin blocks distinguish differentiated from embryonic stem cells. *Nature genetics*, 41(2):246–250, 2009.

[145] Carl R Woese, Otto Kandler, and Mark L Wheelis. Towards a natural system of organisms: proposal for the domains archaea, bacteria, and eucarya. *Proceedings of the National Academy of Sciences*, 87(12):4576–4579, 1990.

[146] Hua Wong, Hervé Marie-Nelly, Sébastien Herbert, Pascal Carrivain, Hervé Blanc, Romain Koszul, Emmanuelle Fabre, and Christophe Zimmer. A predictive computational model of the dynamic 3d interphase yeast nucleus. *Current Biology*, 2012.

[147] Hugo Würtele and Pierre Chartrand. Genome-wide scanning of hoxb1-associated loci in mouse es cells using an open-ended chromosome conformation capture methodology. *Chromosome Research*, 14(5):477–495, 2006.

[148] Na Xu, Mary E Donohoe, Susana S Silva, and Jeannie T Lee. Evidence that homologous x-chromosome pairing requires transcription and ctcf protein. *Nature genetics*, 39(11):1390–1396, 2007.

[149] E. Yaffe and A. Tanay. Probabilistic modeling of hi-c contact maps eliminates systematic biases to characterize global chromosomal architecture. *Nature Genetics*, 2011.

[150] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome research*, 18(5):821–829, 2008.

[151] Bo Zhang, Josiane Zerubia, and Jean-Christophe Olivo-Marin. Gaussian approximations of fluorescence microscope point-spread function models. *Applied Optics*, 46(10):1819–1829, 2007.

[152] Zhihu Zhao, Gholamreza Tavoosidana, Mikael Sjölinder, Anita Göndör, Piero Mariano, Sha Wang, Chandrasekhar Kanduri, Magda Lezcano, Kuljeet Singh Sandhu, Umashankar Singh, et al. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra-and interchromosomal interactions. *Nature genetics*, 38(11):1341–1347, 2006.

Appendices

Appendix A

Image

SUPPLEMENT

FISH-quant: automatic counting of transcripts in 3D FISH images

Florian Mueller[1,2], Adrien Senecal[2], Katjana Tantale[3], Hervé Marie-Nelly[1], Nathalie Ly[2], Olivier Collin[2], Eugenia Basyuk[3], Edouard Bertrand[3], Xavier Darzacq[2], Christophe Zimmer[1]

[1] Institut Pasteur, Unité Imagerie et Modélisation, Centre National de la Recherche Scientifique, Unité de Recherche Associée 2582, 25-28 rue du Docteur Roux, 75015 Paris, France

[2] Institut de Biologie de l'Ecole Normale Supérieure, Functional Imaging of Transcription, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 8197, 46 rue d'Ulm, 75005 Paris, France

[3] Institut de Génétique Moléculaire de Montpellier, Centre National de la Recherche Scientifique, Unité Mixte de Recherche 5535, 34293 Montpellier Cedex 5, France

Correspondence to: Edouard Bertrand (edouard.bertrand@igmm.cnrs.fr), Xavier Darzacq (darzacq@ens.fr), Christophe Zimmer (czimmer@pasteur.fr)

Nature Methods: doi:10.1038/nmeth.2406

1. Workflow of FISH-quant

The general workflow for the counting of mature and nascent mRNA in FISH-quant is summarized in the following two schematics. The entire functionality can be controlled via graphical user-interfaces (GUI's). These interfaces are indicated by boxes with white headings on black background. A detailed description of how to use FISH-quant can be found in the documentation distributed together with source-code (http://code.google.com/p/fish-quant/). In this document also screen shots of the different GUI's can be found together with already processed example data. A more detailed description of the respective algorithms will be presented in the two following sections.

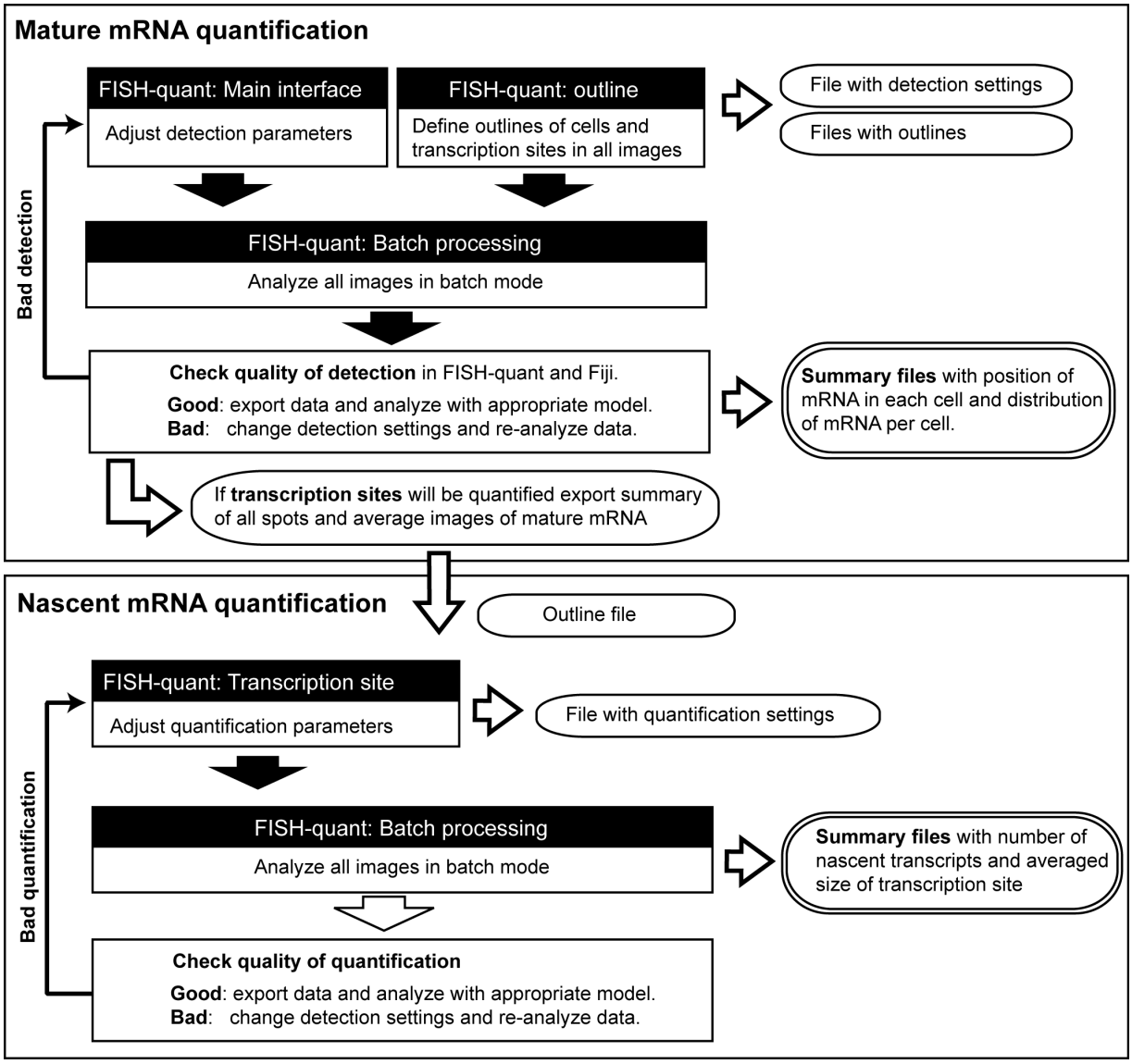


Figure S1. *Schematic of workflow for counting of mature and nascent mRNA in FISH-quant. Full arrows indicate processing steps, empty arrows point to generated data. Boxes with white headings on black background indicate GUI's. Output files are indicated with boxes with rounded edges. Final output files that can be analyzed with mathematical models are indicated with a doubled frame.*

Nature Methods: doi:10.1038/nmeth.2406

2. Algorithm for mature RNA detection and counting

The detection and counting of mature mRNA is based on established methods for single molecule detection in 3D[1]. In short, pre-detection of spots is performed on a filtered imaged followed by a fit with a 3D Gaussian function. Remaining spots are then counted in each cell. Figure S2 summarizes the different steps involved. Each step will be explained in more detail below.

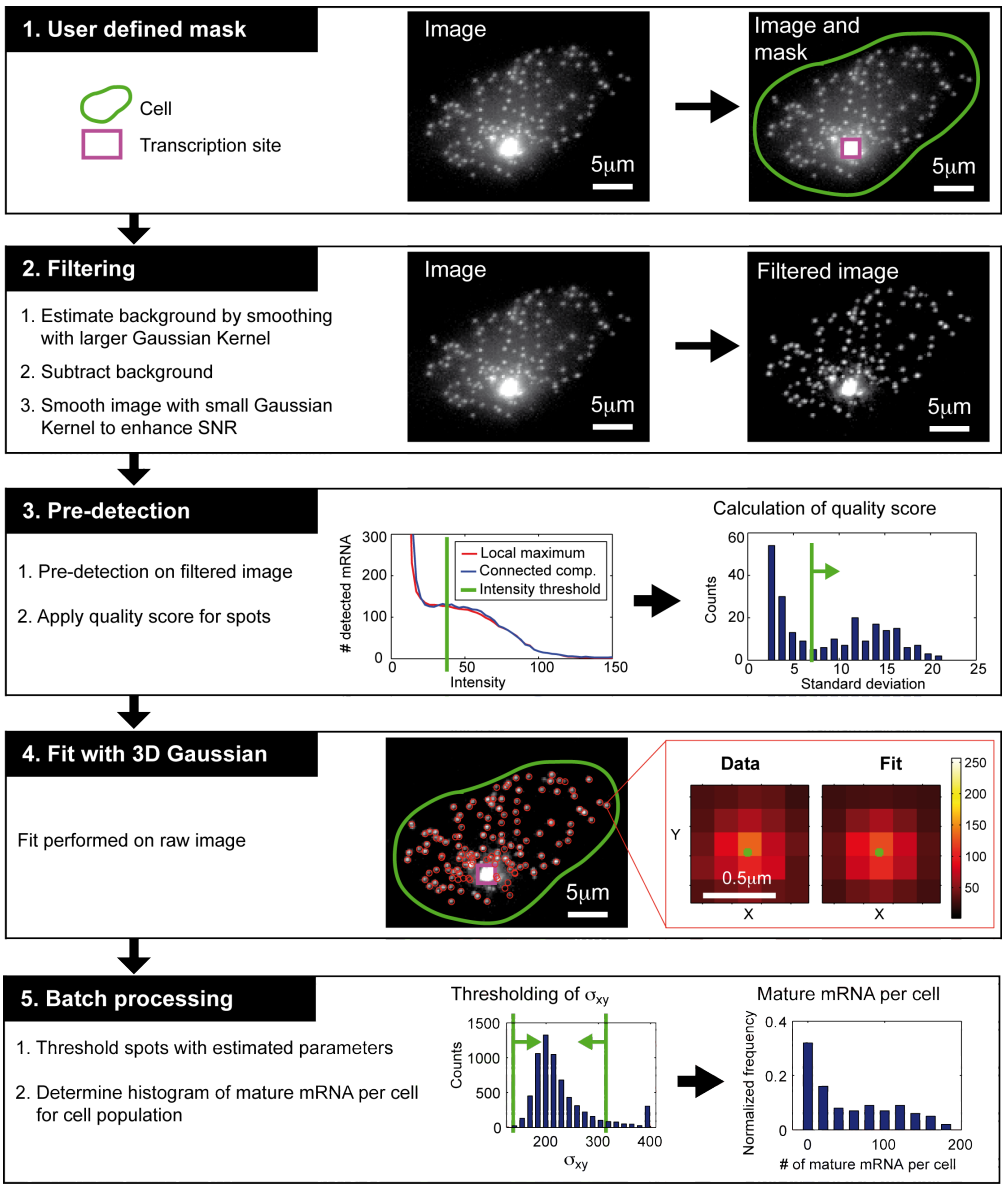


Figure S2. Algorithm for mature mRNA detection*. For illustration purposes images are shown as maximum intensity projections (MIP) in XY but analysis is performed in 3D.* ***(1)*** *User defines masks for the outline of cells, transcription sites, and optionally nuclei.* ***(2)*** *Images are filtered for improved pre-detection.* ***(3)*** *Pre-detection by local maximum detection or connected components. User has to define a minimum intensity threshold. Plot on the left shows number of detected spots as a function of this detection threshold. Similar results are obtained for either method (compare blue and red curve). A characteristic plateau is observed for the correct thresholds. Chosen threshold (green vertical line) yields a slight over-detection but gives a safety margin for batch detection. For each spot candidate a quality score is calculated (here the standard deviation of the intensity in the neighborhood of the*

Nature Methods: doi:10.1038/nmeth.2406

spot). User sets a minimum score (green line), ***(4)*** *Each candidate spot (red circles) is fit with a 3D Gaussian. Plots on the right show the MIP of the spot (left) and the best fit (right). The green circle indicates the identified center of the spot with sub-pixel localization accuracy.* ***(5)*** *Analysis can be performed in batch mode and results of many cells are pooled together. Spots can then be thresholded based on the different fitting parameters, e.g.* σ_{xy}*, as shown in the left plot. Final result is the number of mature mRNA per cell (right plot).*

2.1. Define mask with outline of cell and transcription site

The user draws a mask to outline the individual cells and the transcription sites (and optionally nuclei). The subsequent analysis is only performed within the cells; transcription sites are excluded from the analysis of mature mRNA (but will be processed separately, see Supplementary Note 4). FISH-quant also provides different methods to automatically detect transcription sites (See Supplementary Note 4.4).

2.2. Filtering of image for better pre-detection

We implemented a two-step filtering process to remove inhomogeneous background and increase the SNR[2]. This is achieved by a 2-step convolution of the image with a Gaussian Kernel using the function `gaussSmooth`[3]. First, the raw image I_{raw} is convolved with a large Gaussian Kernel to blur it and obtain a good approximation of the background. By default the standard deviation of this Kernel is set to 5 times the standard deviation of a Gaussian that best matches the theoretical PSF for the optical setup used[4]. Then this image is subtracted from the raw image. The resulting image is then filtered with a small Gaussian Kernel to enhance the SNR. By default the standard deviation of this Kernel is set as the standard deviation of the Gaussian best describing the theoretical PSF[4]. The filtered images I_{filt} is therefore obtained by

$$I_{filt} = G_{\sigma_{th}} * \left(I_{raw} - G_{5\sigma_{th}} * I_{raw}\right), \qquad [1]$$

where $*$ indicates convolution with the indicated Gaussian Kernel $G_{\sigma_{th}}$(the integrated intensity of this Kernel is 1).

2.3. Pre-detection of spots

Next, candidate spots are identified that will be subsequently fit with a 3D Gaussian function in the next step. We implemented two different methods to identify these candidates. Both methods are applied to the filtered image I_{filt} obtained in step B.

a. **3D local maximum detection**[1]. Identifies the local maxima with values greater then or equal to all voxels in the surrounding area with the function `nonMaxSupr`[3]. FISH-quant sets the radius of this area by default to twice the standard deviation of the Gaussian best describing the theoretical PSF[4].

b. **Connected components**[5]**.** Spot candidates are identified as connected components in 3D after thresholding the image with the Matlab function `bwconncomp`.

In either method the minimum intensity of a spot candidate must first be specified. When plotting the number of detected spots as a function of this intensity threshold a characteristic plateau can be found for a range of intensity values that yield an optimal detection (also described by Raj et al.[6]). We found identical curves for both pre-detection methods (Fig. S2). We manually place the intensity threshold towards the left part of the plateau, which leads to a slight over-detection, but the subsequent steps will remove false-positive detections.

Given the signal-to-noise ratio of typical FISH experiments (Supplementary Note 3.1), the determined spot candidates will encompass only a few false positives. Nevertheless we implemented an **additional quality check** to discriminate true spots from background noise. For this purpose we consider the

Nature Methods: doi:10.1038/nmeth.2406

intensity distribution surrounding the spot candidates[1]. In FISH-quant either the 3D curvature based on the Hessian matrix[1] or the standard deviation of the surrounding voxels can be estimated and serve as quality scores. We found that the standard deviation works more robustly for lower quality image. For both methods larger values of the quality score are obtained for good spots, and lower value for background and so a second threshold can be set to separate noise from actual spots. The remaining spots will then be fit with a 3D Gaussian function.

Note: Signal from individual, non-specifically bound probes is detected for some FISH experiments, especially if only a limited number of probes can be used to target the mRNA. Here the quality score alone might not be sufficient to differentiate background noise from real spots. A careful combination of intensity and quality score thresholding has to be applied. The estimated amplitude from the fit with the 3D Gaussian can be also used as an additional thresholding parameter (see below).

2.4. Spot fitting with 3D Gaussian

The remaining spot candidates are then fit with the following function, a 3D Gaussian integrated over the voxel. The fitting is performed in the raw image since filtering affects the localization accuracy and the intensity estimates[1].

$$I_{ijk} = B + A\frac{1}{x_{i,u}-x_{i,l}}\frac{1}{y_{j,u}-y_{j,l}}\frac{1}{z_{k,u}-z_{k,l}}\int_{x_{i,l}}^{x_{i,u}}\int_{y_{j,l}}^{y_{j,u}}\int_{z_{k,l}}^{z_{k,u}} G(x,y,z)\,dx\,dy\,dz, \qquad [2]$$

where I_{ijk} is the modeled intensity of voxel i , $x_{i,l}$, $y_{j,l}$, and $z_{k,l}$ denote the lower border of the voxel, $x_{i,u}$, $y_{j,u}$, and $z_{k,u}$ denote the upper border of the voxel, $G(x,y,z)$ is the Gaussian function give by Eq. [3], B is background of the image, and A is the amplitude of the Gaussian.

$$\mathrm{G}(x,y,z) = e^{-\frac{(x-\mu_x)^2+(y-\mu_y)^2}{2\sigma_{xy}^2}}\, e^{-\frac{(x-\mu_z)^2}{2\sigma_z^2}}, \qquad [3]$$

where σ_{xy} and σ_z are the width of the Gaussian in xy and z. μ_x , μ_y, and μ_z are the coordinates of its center in x, y, and z. The solution of this integral is provided in Matlab with the function `erf` and can be used after a simple renormalization. Images of individual spots I_{spot} are then fit with the Matlab function `lsqcurvefit` to minimize the squared sum of residuals R

$$R = R = \sum_i \sum_j \sum_k \left(I_{spot,ijk} - I_{ijk}\right)^2 \qquad [4]$$

thus yielding estimates of σ_{xy}, σ_z, μ_x, μ_y, μ_z, A, and B.

In a last step, spots can be selected by thresholding σ_{xy}, σ_z, as well as A. False positives, resulting from noise, usually have large σ_{xy} and σ_z compared to real spots and can therefore be easily removed. The remaining spots are then counted in each cell, providing the estimated number of mature mRNA.

2.5. Batch mode

We found that the various detection parameters can be defined robustly for images taken on the same day under identical imaging conditions. Therefore such a set of images can be analyzed with the same parameters. Accordingly FISH-quant offers a batch-processing module. The user can first define all outlines for all cells in the images and then process them fully automatically. The final thresholding based on the fitting parameters can then be adjusted based on the results of all fitted spots in all images.

Nature Methods: doi:10.1038/nmeth.2406

3. Validation of mature RNA detection in simulations and experiments

In this section we report quantitative validations of mature mRNA detection in simulations and in experimental data.

3.1. Localization accuracy of mature mRNA on simulated images

We validated the mRNA localization accuracy of FISH-quant in simulated 2D images. We compared the FISH-quant estimates to two other localization methods: a Maximum-Likelihood Estimation (MLE) and a recently developed algorithm based on radial symmetry[7]. Both of these methods reach accuracies that are near theoretical limits and were implemented in Matlab, facilitating their implementation and comparison with FISH-quant.

Signal-to-noise (SNR) in FISH images

It is well known that the localization accuracy of these methods depends strongly on the signal-to-noise ratio (SNR) of the images[8]. We therefore first quantified the signal-to-noise-ratio (SNR) in typical FISH images. We defined SNR as the ratio of estimated amplitude of the Gaussian A over the standard deviation of the background σ: $SNR = A/\sigma$. We obtained A from the actual fits with the Gaussian function, and σ by computing the mean and standard deviation of regions in cells containing only background (see equation [2] on page 5). The obtained SNR (Table S1) are high because in FISH individual mRNA molecules are labeled with several tens of fluorophores. Higher SNR was obtained for RPB1 probes labeled with Cy3 compared to Alexa 488, as expected given the difference of autofluorescence in the two colors and brightness of the dyes.

FISH experiment	SNR
RPB1: labeled with Cy3	33
RPB1: labeled with Alexa 488	7
Hygro-MS2x96-bGH	27

Table S1. *SNR of FISH images for different experiments presented in this study.*

Validation of localization accuracy in noisy images

To estimate the pointing accuracy of the three methods, we simulated pixelated images of a diffraction-limited point source with different, known sub-pixel localizations and different noise levels (Fig. S3). For each SNR, we simulated 800 images and fit them with the three methods.

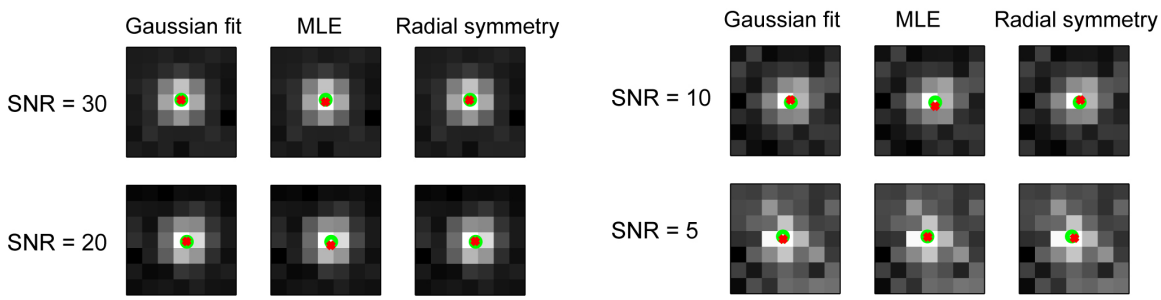


Figure S3. *Fit of simulated pixelated images in 2D. PSF was obtained from PSF ImageJ plugin PSF-Generator[9] on a fine pixel grid of 5nm (Emission wavelength = 568 nm, numerical aperture = 1.25, refractive index = 1.46). PSF was then placed at random sub-pixel locations and an image on larger pixel grid (100nm) was generated. We included noisy background by additive Gaussian noise and varied its standard deviation to obtain different SNR levels. Open green circles indicate the true locations of the PSF center, red spots indicate the position estimated by each localization method. Fig. S10 summarizes localization accuracy for different SNR.*

Nature Methods: doi:10.1038/nmeth.2406

We then computed for each SNR the median of the individual absolute localization errors $e = \sqrt{(x_c - x_m)^2 + (y_c - y_m)^2}$, where x_c and y_c are the known center coordinates of the point source, and x_m and y_m are the measured coordinates (Fig. S4, left). All three methods achieved similar detection accuracy for the simulated range of SNR.

We also analyzed simulated images with Poisson noise[7] for the same range of SNR and found again that the methods yielded comparable results (Fig. S4, right).

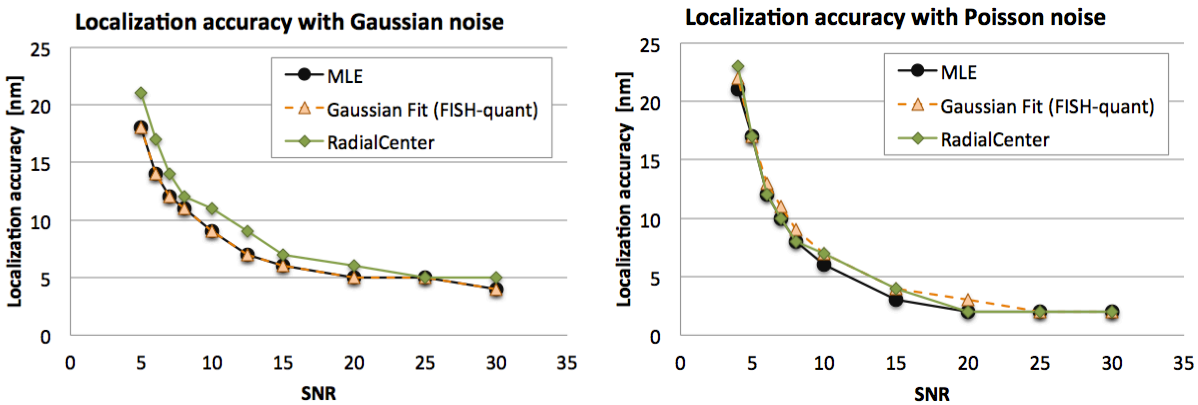


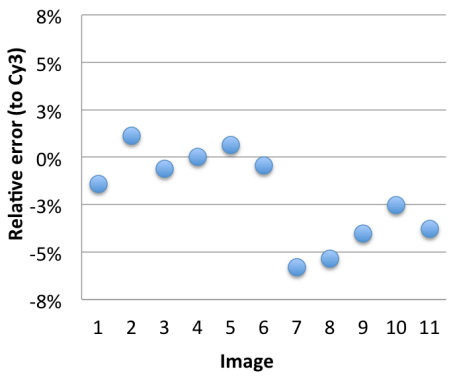
Figure S4**. Localization accuracy of Gaussian fit compared to MLE and radial center method. All methods provide accurate estimates for simulated images.* ***(Left plot) *Images with simulated additive Gaussian noise with different SNR.* ***(Right plot)*** *Images with simulated Poisson noise[7] with different SNR (SNR is defined relative to peak signal intensity).*

In summary, these simulations indicate that the localization accuracy of the Gaussian fit used in FISH-quant is high for realistic SNR levels of FISH images, and is very similar to accuracies achieved by state-of-the art localization methods. This stems from the high SNR of FISH images where each mRNA molecule is labeled by tens of fluorophores.

3.2. Experimental validation with dual-color FISH

We further tested the reliability of the mature mRNA detection experimentally by labeling the same target mRNA simultaneously in two colors. We defined a total of 30 probes for RPB1 and labeled them interleaved with Alexa 488 and Cy3 (see Supplementary Methods for details). We collected 11 images in both colors and used FISH-quant for the mature mRNA detection (Fig. 1b). We obtained excellent agreement for the number of mature mRNA and the estimates were within +/- 5% (Fig. S5 and Fig. 1c).

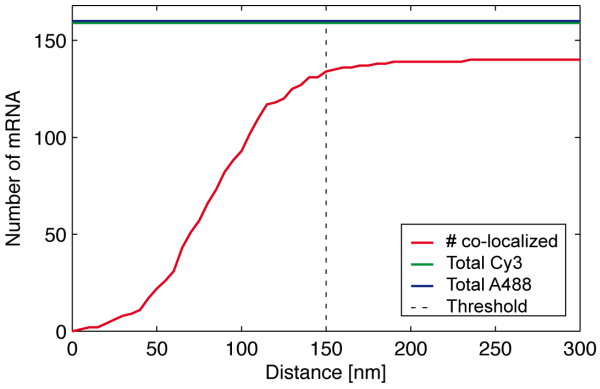
***Figure S5**. Relative error between estimated amount of mature mRNA molecules in dual color FISH against Alexa 488 and Cy3. Error is shown with respect to Cy3.*



Nature Methods: doi:10.1038/nmeth.2406

Next we investigated if the detected spots **co-localize**, i.e. if two spots detected in Alexa 488 and Cy3 correspond to the same mRNA molecule. For this purpose we first needed a quantitative definition of co-localization. To do this we calculated the 3D distance between each spots detected in Cy3 to all spots detected in Alexa 488 and vice-versa. We then determined for each spot the distance to the closest detected spot in the other color. Then we subtracted the average shift as a first order correction for chromatic aberration effects. We then counted for one color the number of spots that have at least one detected spot in the other color within a given distance (Fig. S6). These numbers increase with distance and reach a plateau after 125 nm indicating that this is the maximum distance between two co-localized spots corresponding to the same mRNA molecule. We then used a value of 150 nm to define co-localization of spots detected in two colors. We estimated the percentage of all spots that have a co-localized spot in the other color and determined the amount of co-localization for all images (Fig. 1d). We found that in each image 85%-90% of all detected spots co-localize with a spot in the other color.

Figure S6. *Number of detected spots having a neighboring detected spot in the other color within a given distance. Green line shows the total number of spots detected in Cy3, blue in Alexa 488. The red line shows the number of co-localized spots in Alexa 488 with respect to Cy3. The dashed black line shows the distance chosen to define co-localization. The co-localization for Alexa 488 is around 86% and for Cy3 85%.*



In summary, the dual-color FISH experiment demonstrates the high reliability of mature mRNA detection in FISH-quant.

Nature Methods: doi:10.1038/nmeth.2406

4. Algorithm for transcription site quantification and detection

The following sections describe in detail how FISH-quant quantifies the amount of nascent mRNA– or more precisely **the equivalent amount of full-length transcripts**. The quantification of the FISH signal at the transcription site yields the corresponding number of full length transcripts that would give rise to this signal. However, the signal could also stem from a larger number of partially transcribed transcripts. It is, however, not possible to differentiate between those two scenarios since the resulting signal will be the same. However, appropriate experimental design can minimize this problem and even be used to infer important properties of transcription[10–12]. When placing FISH probes towards the 3' end of the transcript only almost completed transcripts are visible. Alternatively, probes can be placed towards the 5' region to detect also incomplete nascent transcripts. Comparing the results of these two placement strategies for the same gene can be used to study polymerase clustering and transcriptional bursting[10], or to estimate the relative time taken by elongation versus 3′-end processing and release[12]. Further, by designing probe sets in different colors against different parts of the transcript the position of polymerase on the gene can be investigated[11].

Lastly, we note that FISH-quant is not limited to transcription sites, but can be **readily applied to other structures** with a dense accumulation of mRNA, e.g. P-bodies or stress granules[13].

First a brief motivation is given, followed by a detailed explanation of the implementation. The last section describes two different approaches implemented in FISH-quant to automatically detect transcription sites.

4.1. Motivation: spatially extended transcription sites

Transcription sites can have complex topologies such as elongated structures or V-shapes as observed for viruses, genes transcribing large repeated non-coding RNAs, and gene arrays[12,14–20]. We routinely observed transcription site that are larger than the diffraction limit (Fig. S7 and S19). Figure S7 shows the image of a typical transcription site and to the averaged image of individual mRNA molecules. This comparison shows that the transcription site is substantially larger than individual mRNA molecule (and the PSF).

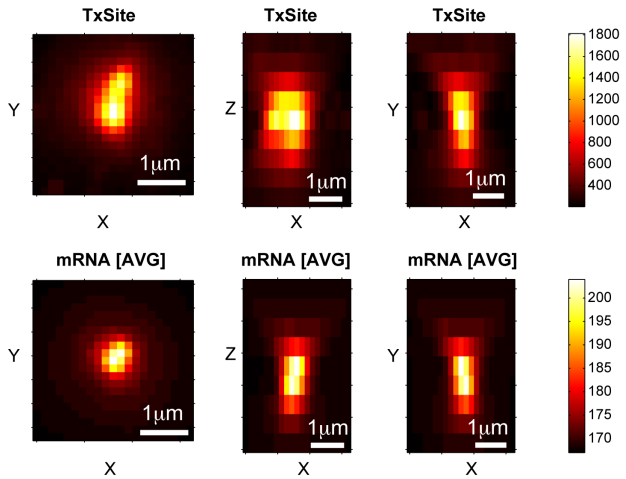


Figure S7. *Comparison of a typical transcription site (upper row) to the averaged image of 7500 individual mature mRNA molecules (lower row) for Hygro-MS2x96-bGH. Images are maximum intensity projections along the major axis as indicated in the axis label. The transcription site is larger than the individual mRNA molecule. Both images show defocusing pattern.*

To our knowledge no method is available to accurately quantify the number of nascent transcripts for such large and spatially extended transcription sites in 3D. In previous studies, the maximum intensity of the transcription site was divided by the averaged maximum intensity of the brightest voxel of individual mature mRNA in the cell[12]. Alternatively, the number of nascent mRNA can be inferred by calculating

Nature Methods: doi:10.1038/nmeth.2406

the ratio of the estimated amplitudes of the transcription site to that of the individual mRNA molecules. Either method neglects the spatial extent of transcriptions sites and therefore implicitly assumes that all transcripts are within a sub-resolution region. The example in Figure S7 illustrates, however, that's transcription sites can be larger and extended in 3D. We therefore implemented two new methods for the quantification of such sites as detailed below.

4.2. Transcription site quantification method 1: integrated intensity in 3D

We consider the spatial extent by comparing the total integrated intensity of the transcription site to the total integrated intensity of the individual mRNA molecules. FISH-quant directly considers the 3D image (quantifications based on integrated intensity have previously been applied to 2D maximum intensity projections[10]). We first average the images of the individual mRNA molecules detected as detailed in Supplementary Note 2. We then fit this image with Eq. [2] and calculate the integrated intensity under the fitted curve. Each transcription site is then fit with Eq. [2] and the integrated intensity under the Gaussian is calculated. The number of nascent transcripts is then inferred by dividing the integrated intensity of the transcription site by the integrated intensity of the individual mRNA molecules.

4.3. Transcription site quantification method 2: superposition of PSFs

In the second approach we considered the spatial extent by using the average image of the individual mRNA molecules to construct an image that best describes the recorded image of the transcription site. This approach is inspired by Gaussian mixture models (GMM) where a weighted sum of Gaussian functions is used to describe complex signals resulting from a superposition of overlapping Gaussians[1]. There are, however, two limitations to GMM that impede its application to FISH. First, a 3D Gaussian function can be satisfyingly used to fit and localize diffraction limited spot[1] but it fails to describe the observed complex diffraction patterns and other aberration effects due to misalignments of the microscope (Fig. S8). We therefore use directly the averaged image of all detected mature mRNA as described above rather than a Gaussian function to describe the signal of individual mRNA molecules. Second, in GMM the weight (amplitude) of the individual Gaussian functions are not restricted[1]. This can lead to an overestimation of the number of mRNA at the transcription site when Gaussian functions with increasingly small amplitudes are used to further improve the fit. In FISH-quant, we therefore restrict the range of the allowed amplitude to the range measured on the individual mRNA molecules.

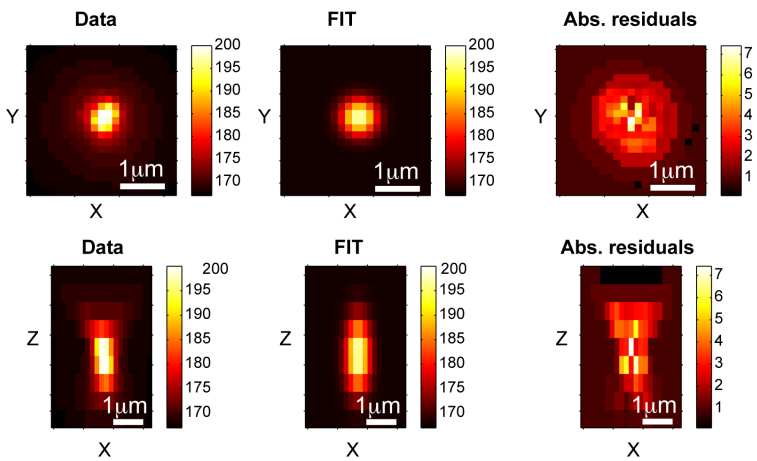


Figure S8. *Fit of averaged image of mRNA with 3D Gaussian. First row shows maximum intensity projections in XY, second row in XZ. First column shows the image, second column the best fit, and third column the absolute residuals. The fit describes the signal well but small systematic deviations can be seen for the diffraction patterns. This poses no problem for localization but it can results in an overestimation of the number of Gaussians used in the GMM to describe a bright transcription site. For such bright sites the diffraction patterns can become prominent and these additional Gaussians would be necessary to model them.*

We first **compared the averaged image of individual mRNA molecules** to images of 100 nm fluorescent beads (TetraSpeck, Invitrogen). We analyzed the images of the beads with the same workflow as the FISH data. While the beads were brighter than the individual mRNA molecules their estimated size

Nature Methods: doi:10.1038/nmeth.2406

was similar (Table S2). This argues that for our experimental system individual mRNA molecules are diffraction limited in size and we can safely average them without loosing spatial information.

Parameter	Beads (N=1,500)	**Hygro-MS2x96-bGH** (N=13,000)
$\boldsymbol{\sigma_{xy}}$	151 +/- 5 nm	175 +/- 35 nm
$\boldsymbol{\sigma_z}$	561 +/- 53 nm	577 +/- 128 nm
$\boldsymbol{A}$	428 +/- 101	84 +/- 24
$\boldsymbol{B}$	205 +/- 7	177 +/- 15

Table S2. *Comparison of fitting results for beads and mature mRNA in FISH-quant. Listed values are mean +/- standard deviation. Numbers in parenthesis indicate how many spots were considered.*

Next we analyzed the **distribution of the estimated amplitudes** of the fit with the 3D Gaussian for the individual mRNA. We found that this distribution is well described by a skewed normal distribution (Fig. S9). We used the Matlab command `normfit` to determine the mean value μ and standard deviation σ. Skewness s, and kurtosis k were determined with the Matlab commands `skewness` and `kurtosis` and are defined as follows:

$$s = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^3}{\left(\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2}\right)^3}, \quad k = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right)^2}. \qquad [5]$$

This distribution is caused by a number of different factors including detection noise, variable labeling efficiency of the FISH probes, variable number of hybridized probes per mRNA, and stochasticity of fluorescence.

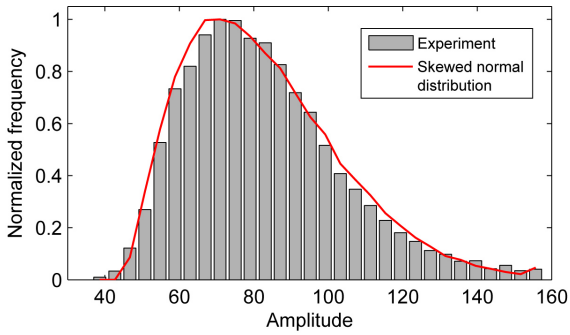


Figure S9. *Fit of distribution of estimated amplitudes with skewed normal distribution yields* $\mu = 83, \sigma = 21$, $s = 0.8$, *and* $k = 3.5$. *Red curve shows normalized histogram of 10,000 random numbers simulated with the Matlab function* `pearsrnd` *with the specified values.*

4.3.1. Algorithm for superposition of PSFs

The quantification method is summarized below and in Fig. S10. In short, the algorithm attempts to find the most probable 3D positions of mRNA's that give rise to the recorded image of the transcription site. This is achieved by an iterative process where individual mRNAs are placed in a model image until the best description of the actual image is obtained. Each step will be explained next in more detail.

a) Analyze transcription site and background of cell

The algorithm starts with a homogenous background image in which the individual mRNAs are placed. Different cells have different background values so we implemented an automated method to determine the best background. First, the background of the cell is analyzed by extracting the voxel intensities within the same z-planes as the transcription site. Then their mean μ_{cell} and standard deviation σ_{cell} are calculated and used to determine a possible range of background values B that will be tested (By default 10 values in the range [μ_{cell}-σ_{cell}, μ_{cell}+σ_{cell}]). The image is further cropped around the transcription site to restrict the area of analysis (I_{TS}).

Nature Methods: doi:10.1038/nmeth.2406

b) Properties of mature mRNA molecules

As an input for the algorithm the averaged image of the individual mature mRNA molecules (Fig. S7) and the distribution of the estimated amplitudes (Fig. S9) are imported (Supplementary Note 2).

c) Determine background

The algorithm (see below) is then applied 50 times for each background value B in the range [μ_{cell}-σ_{cell}, μ_{cell}+σ_{cell}]. These repetitions are needed because the algorithm uses random numbers.

d) Detailed analysis

Then the background value with the lowest residuals is chosen for a subsequent analysis and the algorithm is performed 100 times. Additionally, the averaged size of the transcription site is determined by calculating the average distance of all individual placed mRNA molecules to their center of mass.

Algorithm

1. Generate homogenous background image B with $B \in [\mu_{cell} - \sigma_{cell}, \mu_{cell} + \sigma_{cell}]$
2. Calculate sum of absolute residuals R_0 between B and the cropped image of the transcription site I_{TS}:

$$R_0 = \sum_{x,y,z} |B - I_{TS}| . \quad [6]$$

3. Iteratively add one image of the mature mRNAs I_M to the background B to obtain the model image $I_{G,N}$ (see below):

$$I_{G,N} = B + \sum_{i=1}^{N} I_M\,(x_{c,i}, y_{c,i}, z_{c,i}, A_i) , \quad [7]$$

where $I_{G,N}$ is the image obtained after placing N mRNA images. Each placed individual mRNA I_M has a different center specified by $x_{c,i}, y_{c,i}, z_{c,i}$ and amplitude A_i.

Iterative placement of mRNA

3.1. Subtract the model image from the preceding iteration ($I_{G,N-1}$) from the image of the transcription site I_{TS}. Note that $I_{G,0} = B$.

$$I_D = I_{G,N-1} - I_{TS} \quad [8]$$

3.2. Find voxel with maximum intensity in image I_D from Eq. [8].

3.3. Pick an amplitude A by random sampling of the skewed Gaussian distribution of the estimated amplitudes (Fig. S9).

3.4. Renormalize average image of mRNA to match amplitude from step 3.3.

3.5. Add this image to $I_{G,N-1}$ at location from step 3.2 to obtain the new model image $I_{G,N}$ as described in Eq. [7].

3.6. Calculate sum of absolute residuals R_N between model image and image of transcription site:

$$R_N = \sum_{x,y,z} |I_{G,N} - I_{TS}| . \quad [9]$$

3.7. Back to 3.1. until residuals R_N are larger than residuals R_0 estimated in step 2.

Analysis of results

4. The residuals R_N as a function of N follow a characteristic U-form (Fig. S10). For each run the number of mRNA's with the minimum residuals is determined and serves as an estimate of the number of nascent transcripts. Runs are repeated several times and the averaged number of nascent mRNA and the standard deviation is calculated.

Nature Methods: doi:10.1038/nmeth.2406

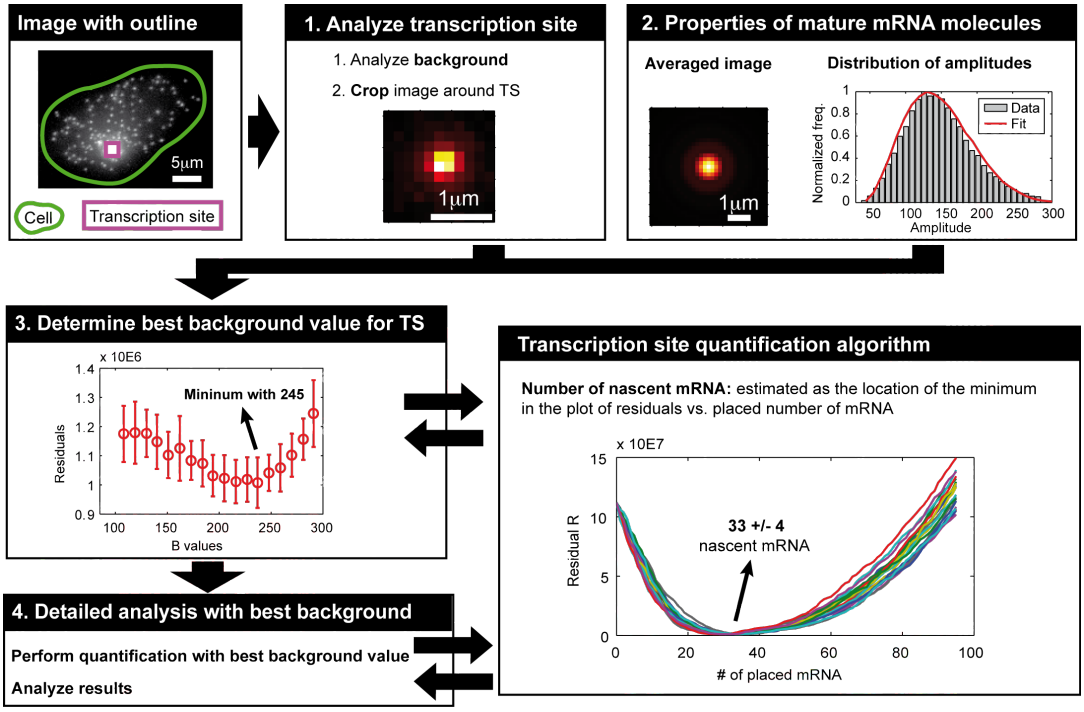


Figure *S10.* *Schematic of PSF superposition approach to quantify the amount of nascent transcripts.*

4.4. Automated detection of transcription sites

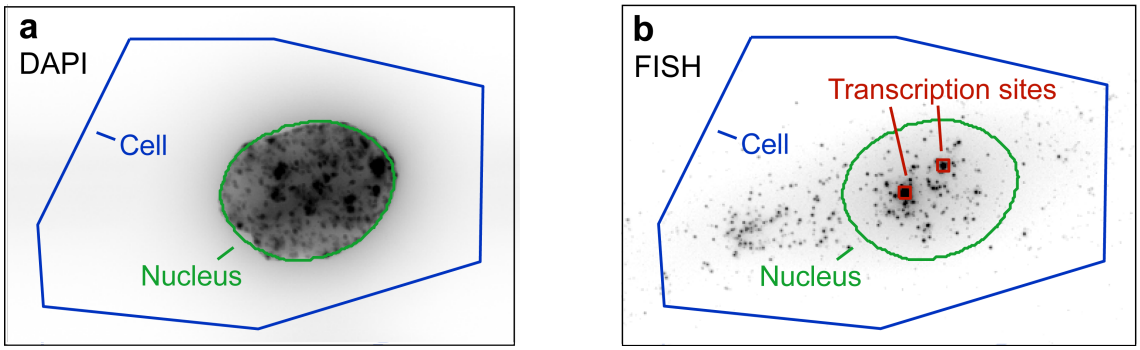
FISH-quant provides two different methods to automatically detect transcription sites. First, transcription sites can be identified as spots in the nucleus with higher intensities than the mature counterparts[10] (Supplementary Note 4.4.1.). Second, images of an independent label of the transcription site, like LacI, can be used to locate transcription sites in the FISH image (Supplementary Note 4.4.2.).

4.4.1. Automated detection based on intensity alone

Transcription sites are identified based on a user defined intensity threshold that separates them from mature mRNA. To reduce the number of false-positives the detection can be further restricted to the nucleus of each cell. This can be done either by loading a DAPI image, or an image of any other nuclear stain, and defining an additional intensity threshold for the DAPI signal or by defining the outline of the nucleus (Fig. S11a).

We applied this method on RNA-FISH images against the c-Fos gene in human fibroblasts 20 min after serum induction (Fig S11b). This gene is less expressed as the other genes used in this study (β-actin Hygro-MS2x96-bGH reporter), thus making transcription site identification more challenging. After a first round of FISH-quant analysis we found that the intensity of individual transcripts did not exceed 4000 units (Fig. S11c). We therefore set the intensity threshold for transcription site detection to 5000. We then restricted the automated detection to the outlined nuclei. Detected transcription sites were substantially brighter than mature transcripts (Fig. S11c). We visually verified more than 100 cells and found excellent agreement between the automatically detected and manually outlined sites.

Nature Methods: doi:10.1038/nmeth.2406



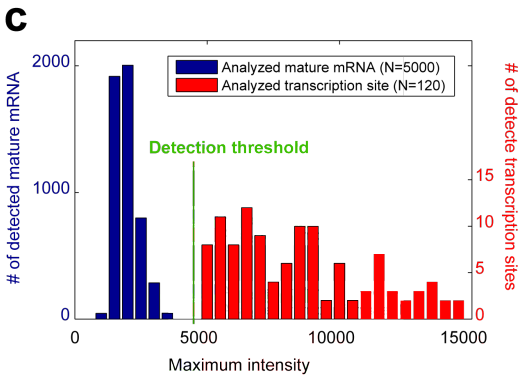


Figure S11. *Automated detection of transcription sites in FISH-quant for c-Fos (see Supplementary Methods for details).* ***(a)*** *DAPI stain is shown together with outline of cell (blue) and nucleus (green).* ***(b)*** *FISH signal is shown together with outline of cell and nucleus. Automatically detected transcription sites are shown in red.* ***(c)*** *Histograms of the maximum intensities of analyzed individual transcripts (blue) and the automatically detected transcription sites (red).*

4.4.2. Automated detection based on second marker

The detection method described in Supplementary Note 4.4.1. only works for sites that are sufficiently bright compared to mature mRNA. This poses, however, a problem at only weakly transcribing genes where in the lowest limit only one transcript is attached to the transcription site. The image of such a transcription site will result in the same diffraction limited spot as the image of its mature counterpart. It is therefore impossible to distinguish the two based on their intensity alone. Experimental approaches have been developed to circumvent this limitation by independently labeling the transcription site with a second marker. The most frequently used method is the LacI tagging approach[21]. Here the lac repressor (LacI) is fluorescently tagged and binds to arrays of lac operator sequences inserted close to the transcription sites on the chromosome. Alternatively, DNA FISH can be performed against the target gene to obtain independent labeling[11]. Lastly, mRNA FISH can be performed with special probes designed against the intron of the studied genes[11]. Most transcripts are spiced co-transcriptionally[22,23], so transcripts will only be visible at the transcription sites, while mature mRNA will generally not be detected. Ultimately, each of these methods produces a second image stack where the transcription sites are marked independently. In FISH-quant, these additional images can be used to automatically detect transcription sites also in the absence of FISH signal (Fig. S12).

We demonstrate the different approaches with an artificial reporter (β-globin-Luc-CFP-24MS2, Supplementary Methods). We constructed this reporter such that FISH can be performed against exons and introns. The reporter has also binding sites for LacI, so an independent visualization of the gene locus is possible with LacI-YFP (which was transfected as a plasmid). We acquired 4 image stacks for each field of view: DAPI, FISH against exons, FISH against introns, and LacI-YFP. This allows a direct comparison of these techniques. The exon FISH image shows a large number of mature mRNA, but also allows to detect a transcription site in cell 1 (Fig. S12b). In the intron FISH image, mature mRNA molecules are not visible, as expected, but the active transcription site can be clearly detected (Fig. S12c).

Nature Methods: doi:10.1038/nmeth.2406

Finally, the image of LacI-YFP also shows a silent transcription site in cell 2 that was not visible in the other two images (Fig. S12d). No transcription site could be detected in cell 3, because this cell did not express LacI-YFP.

In summary, FISH-quant provides different options for the automated detection of transcription sites. For strongly transcribing genes a detection based on the intensity of the transcription site alone can be sufficient. For weakly transcribing genes an independent label of the site can be used to reliably detect its location.

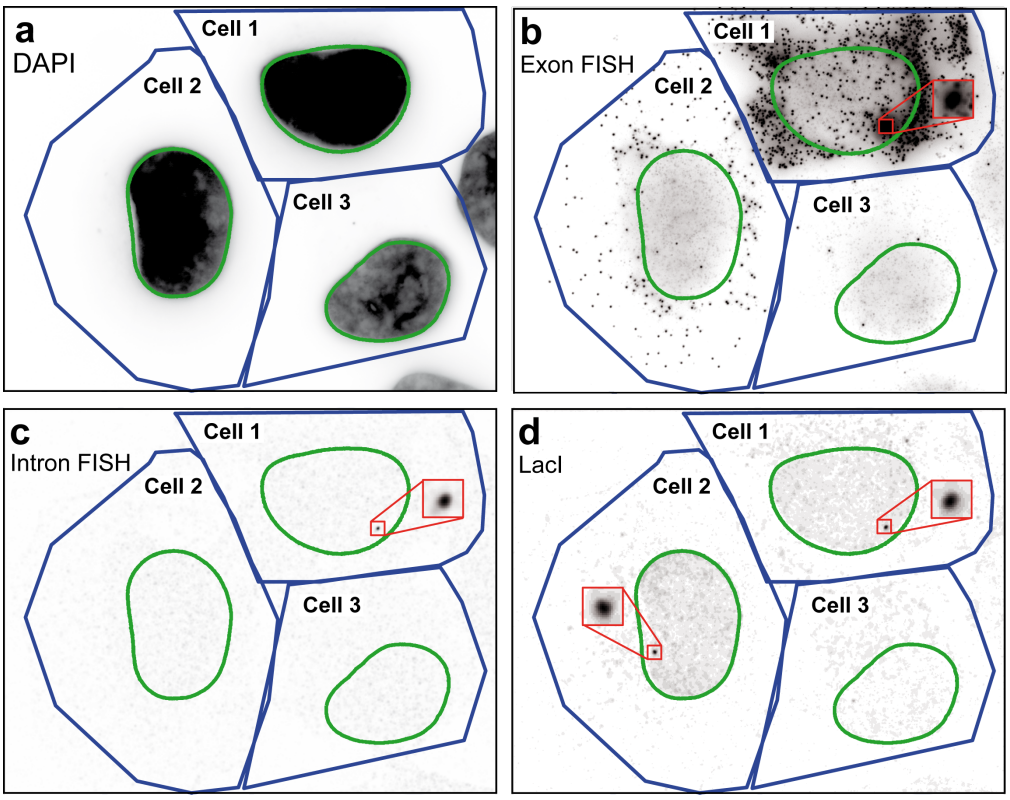


Figure S12. *Automated detection of transcription sites for β-globin-Luc-CFP-24MS2 in FISH-quant.* ***(a)*** *DAPI stain was used to outline the nuclei in the cells.* ***(b)*** *Detection based on intensity of FISH (against exon) signal. Only cell 1 contains a detected transcription site.* ***(c)*** *Detection with FISH against introns yields the same site as in b, since only transcriptionally active sites can be detected.* ***(d)*** *Detection with LacI (transiently expressed) yields two transcription sites: the same site in cell 1 detected in b and c, and a site detected in cell 2 containing only one transcript; in cell 3, the transfected LacI-YFP was not expressed and therefore did not allow to visualize the transcription site (this could be avoided by stably expressing LacI).*

Nature Methods: doi:10.1038/nmeth.2406

5. Validation of transcription site quantification on simulated images

We first evaluated the transcription site quantification methods on simulated data. In the following section we will refer to the different quantification methods with the following abbreviations:

- FISH-quant method based on integrated intensity: FQ-IntInt
- FISH-quant method based on superposing PSFs: FQ-PSFsup
- Method based on estimated amplitude: AMP
- Method based on maximum intensity: MaxInt

Further, we utilize the term FISH-quant methods to refer to FQ-IntInt and FQ-PSFsup, and simpler methods to refer to AMP and MaxInt, which both ignore the three-dimensional extent of the transcription sites.

5.1. Generation of artificial images of transcription sites

We showed that individual mature mRNA molecules are diffraction limited, i.e. their image can be described by the point-spread function (PSF) of the microscopes (Table S2). For the following simulations we therefore used a realistic 3D PSF obtained by the ImageJ plugin PSF-Generator[9] (Fig. S13, lower row).

We generated an image of a transcription site by superposing a pre-defined number of individual PSF's in a certain area. The resulting image of the transcription site depends a number of parameters: number of nascent mRNAs, amplitude of each placed mRNA, size of the transcription site, and noisy background (Fig. S13 shows an example for such a simulated site).

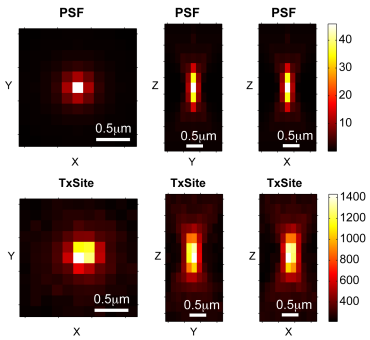


Figure S13. *Images of theoretical PSF and simulated transcription site. Images are shown as maximum intensity projections along the major axis.* ***(Upper row)*** *Theoretical PSF was generated with ImageJ PSF-Generator. Emission wavelength = 568 nm, numerical aperture = 1.25, refractive index = 1.46.* ***(Lower row)*** *Simulated transcription sites with 50 nascent transcripts and a radius of. No noise was added.*

We then simulated transcription sites with different spatial extent and varying amounts of nascent mRNA. We repeated the simulations for each condition 5 times and averaged the obtained estimates for each of the different quantification methods.

5.2. Transcription sites without spatial extent and no noise

First, we simulated transcription sites without any spatial extent and in absence of noise. All mRNA were placed at the same location with the same amplitude. No noisy background was added. Under these idealized conditions all quantification methods worked well for the entire tested range of nascent mRNA abundance (5-100) (Fig. S14a).

5.3. Transcription site with noise

For the next simulations, we placed the mRNA again at the exact same location but now considered the experimentally observed variability of their brightness and the effect of noise. We generated the

Nature Methods: doi:10.1038/nmeth.2406

amplitudes of the placed mRNA molecules randomly following a Gaussian distribution with a mean of 100 and standard deviation of 30, roughly similar to the empirically measured distribution of Figure S7. Furthermore, we simulated a noisy background by adding Gaussian noise of mean 500 and variable standard deviation to achieve a range of signal-to-noise ratios (Supplementary Note 3.1.). For the lowest experimentally observed SNR of 5, all methods again yielded accurate results (Fig. S14b).

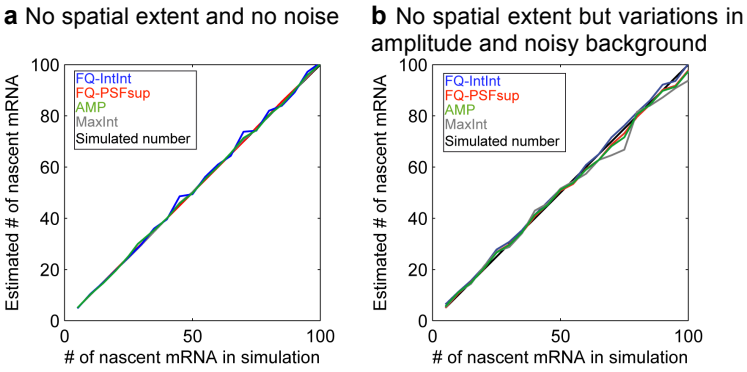


Figure S14. *Nascent mRNA counting in simulated images of sub-diffraction transcription sites. Plots show estimated number of nascent transcripts as a function of the simulated number of transcripts. Each data-point is the average of 5 individual simulations. FQ-PSFsup: red, FQ-IntInt: blue, MaxInt: green, AMP: gray.* ***(a)*** *Without fluctuations of the intensity of the placed mRNA's and no noisy background.* ***(b)*** *With variable amplitudes and noisy background (SNR =5). In either scenario all methods yield accurate estimates.*

We then investigated lower SNR than experimentally observed (Fig. S15). We still obtained accurate quantification for SNR as low as 1, but the methods started to fail for SNR below 1. As expected, the quantifications fail first for sites with fewer transcripts (Fig S15), whereas sites more enriched in transcripts still have sufficiently high signal to be quantified. However, such low SNR will not typically occur in typical FISH images (Supplementary Note 3.1.), otherwise individual mRNA could no longer be detected.

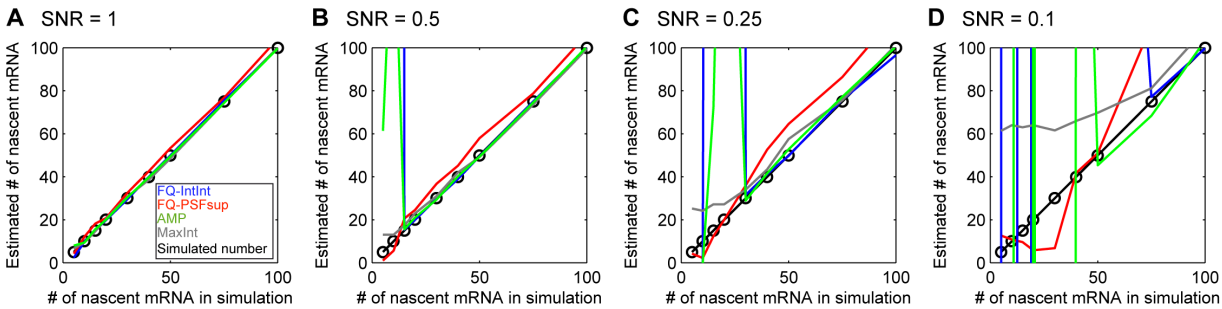


Figure S15*. Accuracy of transcription site quantification for very low SNR for individual mRNA molecules. At these noise levels individual mRNA molecules cannot be detected, so noise-free images of the individual mRNA molecules were used for the quantification.*

5.4. Spherical transcription sites

Next we investigated the impact of spatially extended transcription sites on the quantification results. We therefore simulated spherical transcription sites with increasing radius in which the mRNAs were randomly placed. We considered variations in the amplitude and added a noisy background with SNR=5 as described above. We obtained excellent agreement with the FQ-PSFsup for all radii and the estimates stayed within 3% of the true number (Compare red to black lines in Fig. S16 and Fig. S17). FQ-IntInt yielded good agreement as well and the estimates were within 15% error (Compare blue to black lines in Fig. S16, Fig. S17). The method based on comparisons of amplitude or maximum intensity, however, significantly underestimated the number of nascent transcripts by up to 80% (compare green and gray lines to black lines in Fig. S16 and Fig. S17).

Nature Methods: doi:10.1038/nmeth.2406

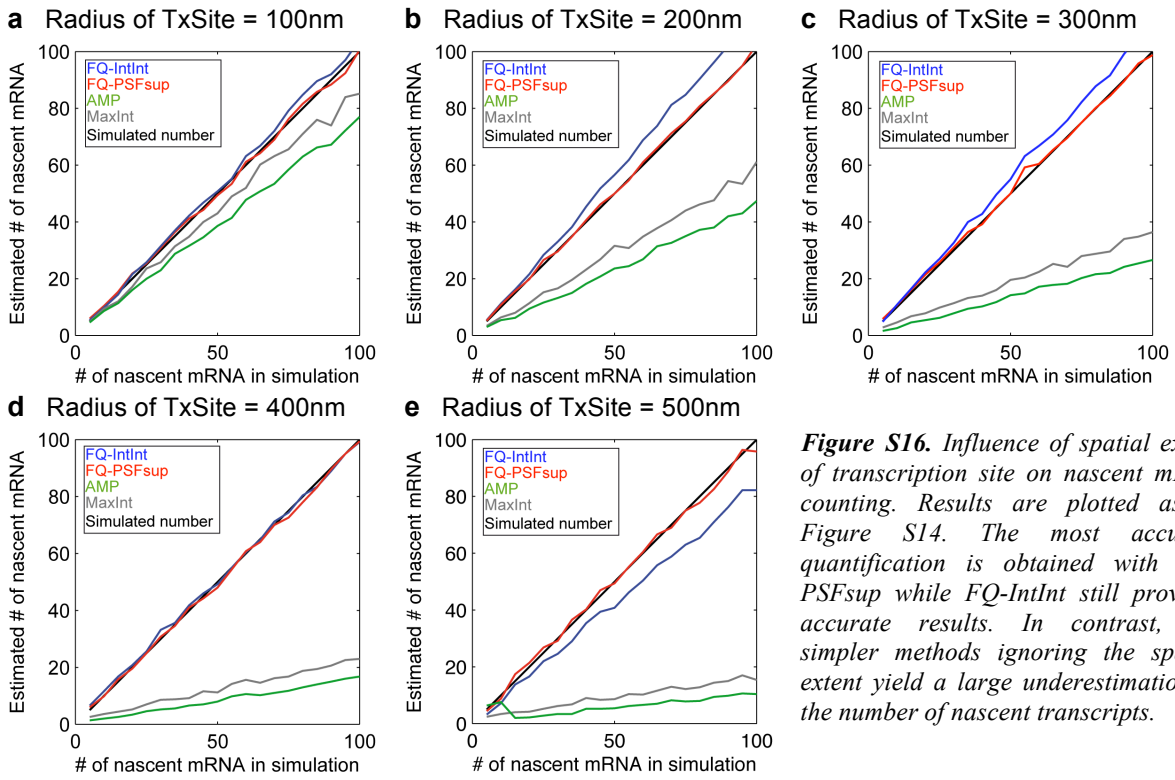


Figure S16. *Influence of spatial extent of transcription site on nascent mRNA counting. Results are plotted as in Figure S14. The most accurate quantification is obtained with FQ-PSFsup while FQ-IntInt still provides accurate results. In contrast, the simpler methods ignoring the spatial extent yield a large underestimation of the number of nascent transcripts.*

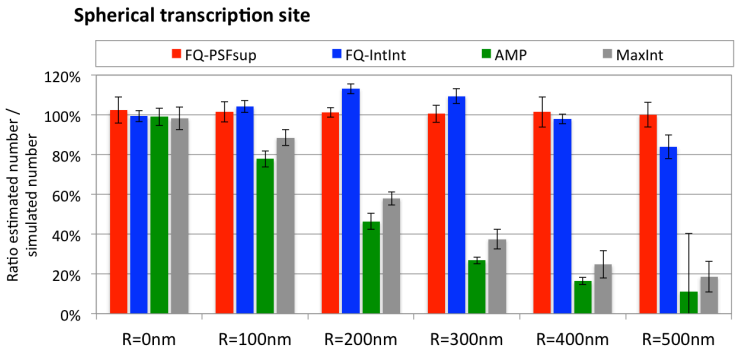


Figure S17. *Accuracy of transcription site quantification for spherical sites. For each individual simulation the ratio of the estimated number of nascent transcripts and the actually simulated number of nascent transcripts was calculated. Then median value and standard deviation of these ratios are shown as bar plots.*

5.5. Spherical transcription sites and noisy background

We then simulated the impact of different SNR on spherical transcription sites with a radius of 100nm (Fig. S16a and S17). As before, we found that varying SNR did not affect the quantification results. However, the simpler methods always underestimate the number of transcripts, while the FISH-quant methods provide accurate estimates (Compare Fig. S17 and Fig. S18). Similar results were found for larger transcriptions sites (data not shown).

Nature Methods: doi:10.1038/nmeth.2406

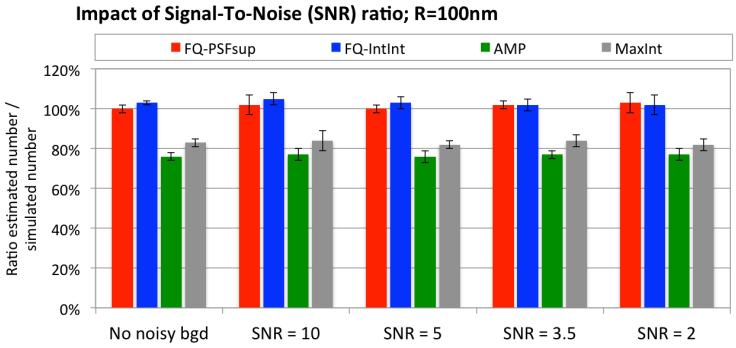


Figure S18. *Influence of SNR on accuracy of transcription site quantification for spherical site of radius 100nm. Results of quantification are presented as in Fig. S17.*

5.6. Ellipsoidal transcription sites

In the above simulations we assumed spherical transcription sites. As described above, many biological samples will, however, not show such perfect symmetry and more complex topologies such as elongated structures or V-shaped transcriptions sites can be observed, e.g. for viruses, genes transcribing large repeated non-coding RNAs, and gene arrays[12,14–19] (Fig. S19a). We therefore simulated ellipsoidal transcription sites to investigate the effect of less symmetrical sites. Ellipsoids were simulated with different ratios of the three semi-axes and rotated randomly in 3D to consider different spatial orientations (Fig 19b, c).

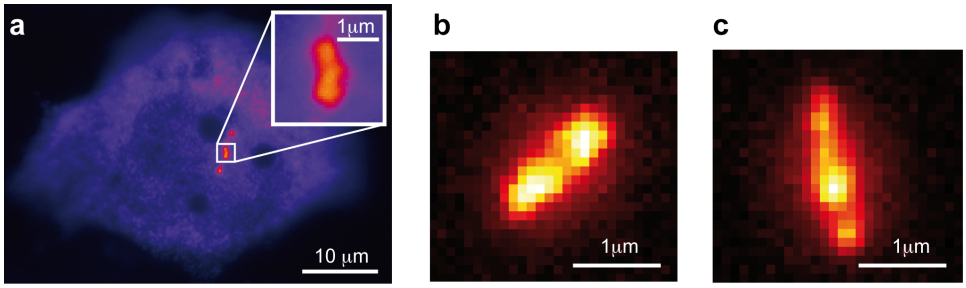


Figure S19. *Elongated transcriptions sites are simulated as ellipsoids.* ***(a)*** *U2OS cells expressing an HIV-1 reporter gene expressed from a gene array (Exo1 cells[12]). Cells were hybridized in situ with a probe against the reporter RNA, with single molecule sensitivity.* ***(b, c)*** *Simulated ellipsoidal transcription site. Semi-axes are 900 nm, 300 nm, and 300 nm (b) and 1000 nm, 250 nm and 250 nm (c). Ellipsoid are rotated counter-clockwise by 45°(b) and 100°(c). Rotation is 2D for illustration purposes only, in simulations 3D rotations were applied.*

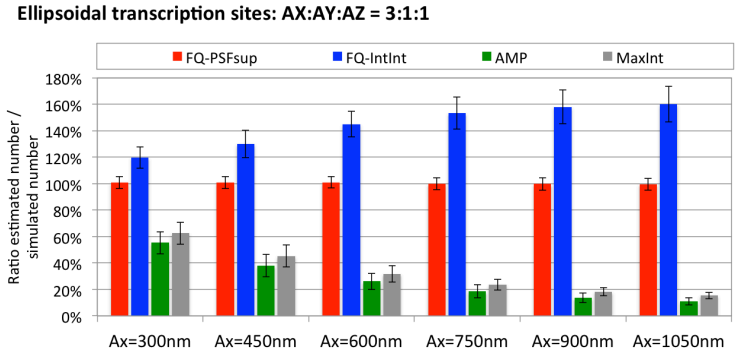
We simulated different mRNA abundances (5-100) and repeated each simulation 50 times. Validation results are shown in Fig. S20. The quantification with FQ-PSFsup stayed within 4% of the true number, while the quantification with FQ-IntInt led to an over-estimation of up to 60%. The simpler methods underestimated mRNA counts by up to 85%.

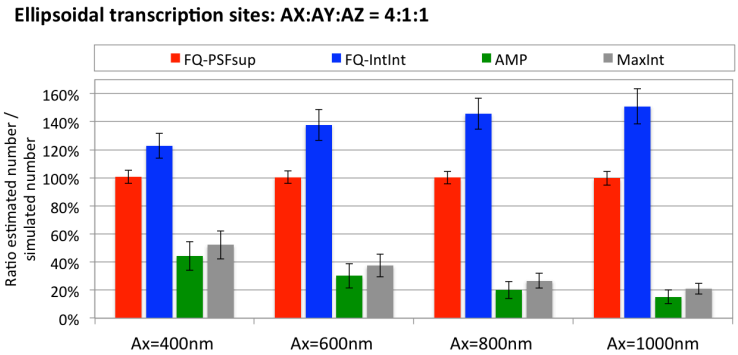
Nature Methods: doi:10.1038/nmeth.2406

Figure S20. *Summary of quantification for ellipsoidal transcription sites. Sites were simulated with different ratios and lengths of the semi-axes. In addition, sites were rotated randomly in 3D to consider different spatial orientations. Results of quantification are presented as in Fig. S17.*

(Upper plot) *Ellipsoidal transcription site with a ratio of the semi-axes AX:AY:AZ = 3:1:1. Length of the longest axis is indicated below each group of bars.*

(Lower plot) *Ellipsoidal transcription site with a ratio of the semi-axes AX:AY:AZ = 4:1:1.*





5.7. Analysis of the spatial extent of the transcription site

The PSF superposition approach does not only yield the number of nascent transcripts but also information about their spatial positioning. While the precise locations of individual mRNA molecules cannot be determined, we can still calculate ensemble quantities such as the averaged distance from their center of mass to measure the spatial extent of the transcription site. Figure S21 summarizes the results of the size measurement for the transcription sites simulated in the preceding sections. For each simulated size we averaged the estimated size by FQ-PSFsup (for sites with more than 20 transcripts). The estimated size was in good agreement to the simulated size for all considered transcription site geometries.

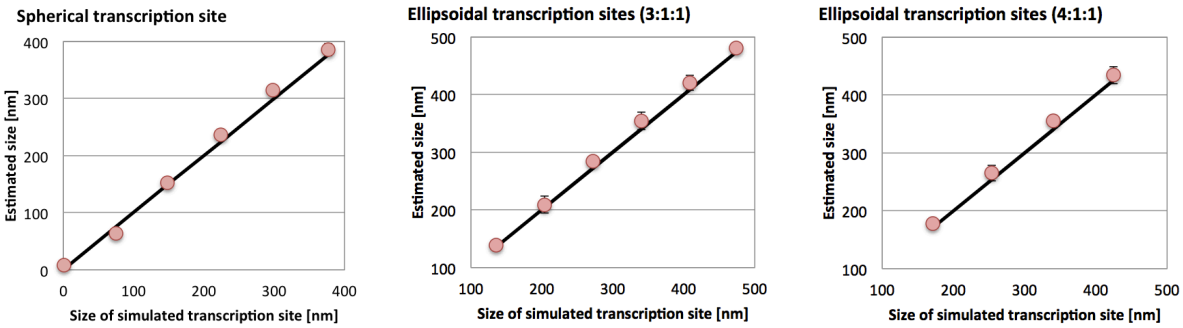


Figure S21. *Size of the transcription sites from Fig S17, S20. Size is measured as the averaged distance of each individual mRNA to the center of mass of the transcription site. Plots show the estimated size (y) vs. the actual size of the simulated size (x). The estimated size is in good agreement with the true size.*

Nature Methods: doi:10.1038/nmeth.2406

5.8. Summary of validation with simulations

In this section we investigated the impact of noise and spatial extent on the different transcription site quantification methods. Our results indicate that the quantification accuracy is not strongly affected by the typical noise observed in FISH, but instead depends on a proper consideration of the spatial extent.

We found that all methods are robust to experimental noise. This robustness can be explained by two factors. First, the signal-to-noise ratio (SNR) of individual mRNA molecules is typically very high in FISH images (Supplementary Note 3.1.), and transcription sites will have an even higher SNR. Second, by considering the averaged image of all detected mRNA molecules (frequently several thousands) we minimize the impact of noise in the analysis of individual mRNAs.

We found, however, that considering the spatial extent is important for a reliable quantification. We found that **FQ-PSsup** performed reliably for all simulated sites, independently of their spatial extent and simulated geometry, whereas the simpler methods ignoring this extent can grossly underestimate the number of transcripts. **FQ-IntInt** yielded accurate estimates for symmetrical, spherical sites but overestimated the number of transcripts for elongated, ellipsoidal sites.

The choice of the quantification methods therefore depends on the typical shape of the observed transcription site. For rather compact, symmetrical sites, both FISH-quant methods yield accurate results. We would therefore recommend using both methods and verify if the results obtained are comparable as an internal quality-check for the quantification. FQ-IntInt has the advantage of being computationally faster than FQ-PSFsup. So if computational time becomes an issue FQ-IntInt can be used alone. For spatially elongated transcription sites, e.g. as can be found for viruses, genes transcribing large repeated non-coding RNAs, and gene arrays, we recommend using FQ-PSsup since only this method accurately quantifies the number of nascent transcripts for these more complex structures. The methods based on a comparison of amplitude or maximum intensity underestimated the number of nascent transcripts for larger, spatially extended site. However, theses methods still provided accurate results for diffraction limited sites.

We quantified the spatial extent of the transcription site by calculating the average distance of each transcript to the center of the site. The distance estimated by FQ-PSFsup was in good agreement with the actual size of the transcription site. These estimates could therefore be used to quantify the spatial extent of the sites and relate this to biological properties such as the decondensation state of a transcriptionally active locus.

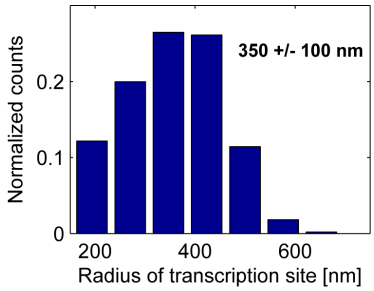
Nature Methods: doi:10.1038/nmeth.2406

6. Validation of transcription site quantification with experimental data

6.1. Hygro-MS2x96-bGH: transcription site quantification

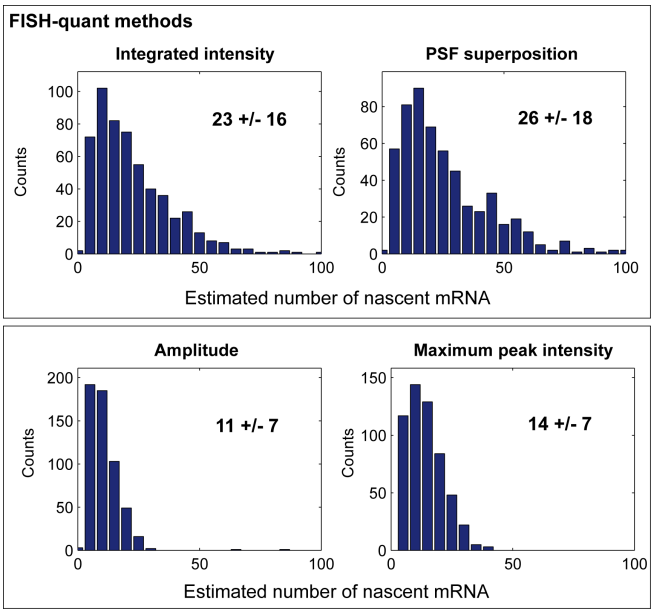
For experimental validation, we analyzed transcription sites in Hygro-MS2x96-bGH cells. The estimated size with FQ-PSFsup reveals that the majority of the sites are larger than the diffraction limit (Fig. S22).

***Figure S22**. Estimated radius of the transcription sites forHygro-MS2x96-bGH. In addition the mean value +/- standard deviation is reported.*



In our simulations we found that for sites that are larger than the diffraction limit the FISH-quant methods estimated larger numbers of nascent transcripts than the simpler methods (Supplementary Note 5). The quantification results of the amount of the nascent mRNA in Hygro-MS2x96-bGH cells with the four methods revealed identical trends (Fig. S23). The FISH-quant methods estimated twice as many nascent transcripts than the simpler methods.

***Figure S23.** Transcription site quantification with the different quantification methods. A total of 552 transcription sites were analyzed. Each plot shows the histogram of the amount of nascent mRNA per site as estimated with the method indicated in the title. In addition the mean value +/- standard deviation is reported.*



6.2. Hygro-MS2x96-bGH: FISH-quant vs. RNAse protection assay

We attempted to further validate the transcription site quantification by comparing the ratio of mature vs. nascent mRNA estimated by FISH to the values estimated by RNAse protection assay (RPA, Supplementary Methods). Using a probe that spans the 3'-end cleavage and polyadenylation site, RPA allows to detect 3'-end cleaved and uncleaved mRNA. The ratio of cleaved vs. uncleaved mRNA can then be used as an approximation of the ratio of mature vs. nascent mRNA and thus compared to the ratio estimated by FISH-quant. We used samples from the same day and experiments were performed in

Nature Methods: doi:10.1038/nmeth.2406

triplicates. Nevertheless, the accuracy of RPA is limited. RPA measurements are based on the quantification of bands from a gel (Fig. S24a). Because nascent mRNA is not very abundant, it appears as a dim band and therefore its quantification is prone to uncertainties. Despite these limitations, RPA can be used to estimate the order of magnitude of nascent vs. mature mRNA.

The cleaved/uncleaved ratio estimated by RPA fell between the values obtained by the FISH-quant methods and the simpler methods (Fig. S24b). Because of its limited accuracy, RPA cannot be used to favor one method over the other. Nevertheless, it confirms the general validity of using imaging based methods to measure mRNA content.

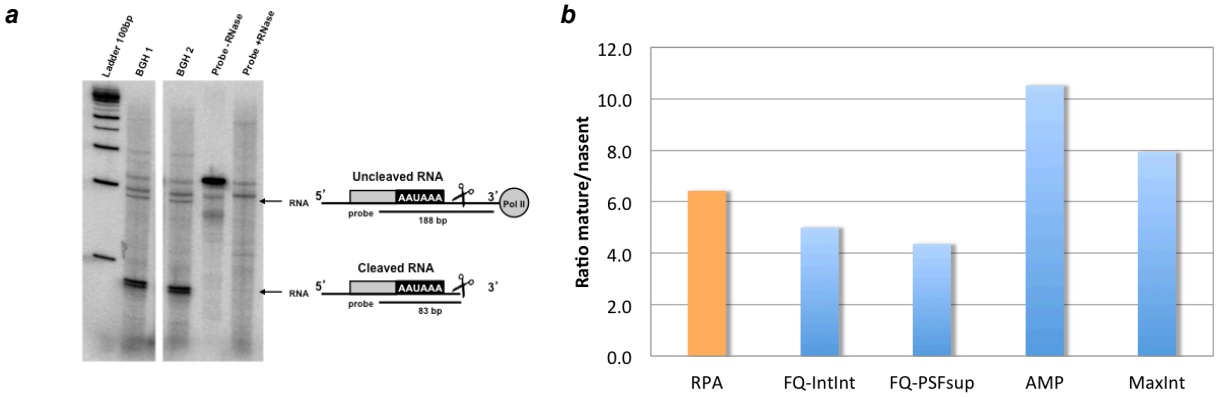


Figure S24. *Cleaved (mature) vs Uncleaved (nascent) mRNA ratio determination by Ribonuclease Protection Assay (RPA) and FISH methods.* ***(a)*** *The protected radiolabeled probe hybridized to complementary RNA is separated on polyacrylamide gel. Schematization of Cleaved and Uncleaved mRNAs with the position and the length of the hybridized probe indicated on the right.* ***(b)*** *Ratio of mature vs. nascent Hydro-MS2x96-bGH mRNA as estimated by RPA and FISH methods.*

6.3. Validation of transcription site quantification on β-actin

We then analyzed the transcription activity of β-actin. This gene has been studied by mRNA FISH in the pioneering study of Femino et al[11] and showed strong activation after serum induction. We repeated the experiment in U2OS cells and obtained similar results. We see practically no active transcription sites before induction and a strong activity 20 min after serum induction. As before the FISH-quant methods estimated larger numbers of nascent transcripts than the simpler methods (Fig. S25).

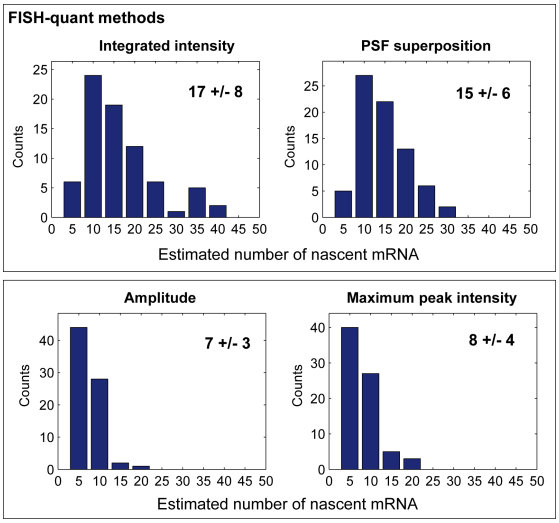
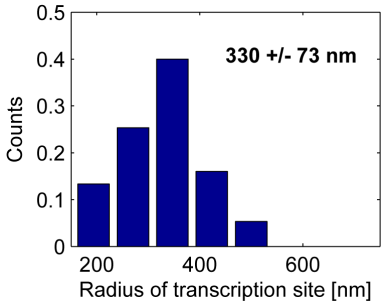


Figure S25. *Transcription site quantification for β-actin 20 min after serum induction with the different quantification methods. Each plot shows the histogram of the amount of nascent mRNA per site as estimated with the method indicated in the title. In addition the mean value +/- standard deviation is reported.*

Nature Methods: doi:10.1038/nmeth.2406

The estimated size of the transcription is larger than the diffraction limit (Fig. S26) providing evidence why the FISH-quant methods estimated larger number of nascent transcripts.

Figure S26. *Estimated radious of the transcription sites for β-actin 20 min after serum induction. In addition the mean value +/- standard deviation is reported.*



Nature Methods: doi:10.1038/nmeth.2406

Supplementary Methods

Plasmids and cell lines

Hygro-MS2x96-bGH reporter

Frt-hygro fragment from pCDNA5/FRT was inserted in BamHI site of BAC2+bS86 (XXV)[24]. Bovine growth hormone (bGH) polyadenylation signal was inserted in MluI site of BAC2+bs96-frthygro to generate the Hygro-MS2x96-bGH reporter. The reporter was stably integrated in frt site of flp-in-293 cell line (Life Technologies) as recommended by manufacturer. The resulting cell lines flp-in-293-Hygro-MS2x96-bGH were cultured at 37°C in DMEM with 10% FBS and 50 µg/ml of hygromycin.

β-globin-Luc-CFP-24MS2

The plasmids pFRT/LacZeo, pOG44 and pcDNA5/FRT were supplied by Invitrogen, pTet-On by Clontech, pBslacO containing 40 LacI binding sites is a gift from M. Ackermann, Institute of Virology, Zurich, Switzerland[25] and pSV2-EYFP/ lac repressor is a gift from DL. Spector, Cold Spring Harbor Laboratory, New York, USA[26]. A pTet-globin-Luc-CFP-24MS2 was generated by inserting in the pTet-globin-CFP-18MS2-2 construct[27] the PCR amplified Luc gene at the BstXI site and 24MS2 repeats from pSL-24X[28] replacing the 18MS2. The FRT-TOLCM vector was generated by inserting pTet-globin-Luc-CFP-24MS2 in the pFRT/LacZeo backbone with restriction enzymes ApaI and NruI. The cell line U2OS was cultured at 37°C in low glucose DMEM with 10% FBS (optionally supplemented with 150µg/ml Zeocin or 100 mg/ml Hygromycin). Following calcium phosphate transfection of a mixture of 1:10 of pFRT/LacZeo and 9:10 of pBslacO, U2OS cell clones having integrated pFRT/LacZeo were selected on Zeocin. The clones containing a single tandem array of lac operator sites next to a single FRT site were selected. One of these clones (A33-8) was then co-transfected with pOG44 and FRT-TOLCM by FuGENE (Roche). One clone (A33-8-T1) that had integrated the single copy of FRT-TOLCM plasmid at the FRT site via Flp recombinase mediated DNA recombination was selected with Hygromycin.

c-Fos: Normal Human Dermal Fibroblast Cell

FISH against c-FOS was performed in Normal Human Dermal Fibroblast Cell (NHDF). Cells were isolated from the dermis of adult skin (Promocell, C-12302). Primary fibroblast cultures were maintained at 37°C in DMEM-F12 (Invitrogen), supplemented with 10% FBS. Primary cell cultures were transferred into collagen (Gibco) coated culture plates and maintained in DMEM-F12, supplemented with 10% FBS. The culture medium was removed 1 day after passage to start serum starvation, the cells were washed with PBS and fresh medium containing no FBS was added for 24h. The culture medium was removed and fresh medium containing 10% FBS was added for 20min before fixation.

RBP1 and β-actin: U2OS

FISH against RBP1 and β-actin was performed in U2OS cells. Cells were cultured at 37°C in DMEM with 10% FBS.

In situ hybridization and imaging

In situ hybridization was performed as described previously[28]. The formamide concentration for FISH in hybridization and washing mixture was 30% against MS2, and 40% against RPB1, β-actin, β-globin-Luc-CFP-24MS2, and c-Fos. 10ng (MS2, RPB1, β-globin-Luc-CFP-24MS2, and c-Fos) or 5ng (β-actin) of probes were used per 50 microliters of the hybridization mixture.

For **β-globin-Luc-CFP-24MS2** cells were co-transfected with pTet-On and pSV2-EYFP/ lac repressor by FuGENE (Roche) 18h before fixation.

Nature Methods: doi:10.1038/nmeth.2406

Sequence of MS2 probe (X stands for amino-allyl T). This probe binds 48 times to the Hygro-MS2x96-bGH reporter. It was labeled with Cy3 (GE Healthcare).

AXACATGGGTGATCCTCATGTTACCCAXGCTCTAGCACACATGGGTGATCCTCATGTXA

Sequence of probes against RBP1 (X stands for amino-allyl T). Even-numbered probes were labeled with Alexa 488 (Life technology) and odd-numbered probes with Cy3 (GE Healthcare) following the protocols of suppliers. 0.5ng of each probe was used for hybridization.

Probe 01	CCCXCAGTCGTCXCTGGGTATXTGATGCCACCCXCCGTCACAGACATXCG
Probe 02	TXTCTTTGGTCAGAXCCTCGTCACCCXCAGGTTGTXCCACACCGAACTXG
Probe 03	GGAXCTTCTTCTCCXGAGAGTCCTCAXTAACGTGCTXCCATTCCGCAXAC
Probe 04	AAACACXCCTCATCXGAGATGCGTTXGAAGATCTCAXGCACTCGCTCXG
Probe 05	TXGATCTTCACGAXGTCAGCCAGTTXGTGAGTCAGGXCATCCTGGTXACG
Probe 06	XGGAGGAGCTTCACAXCCTCTGCAAXGACATGGGCCGCXGCGCCGTTCXG
Probe 07	CGAXCACCATTGTCXCGGATGATGTACXTGGCGCCTGGGXACTGGCTGTXC
Probe 08	ACAXGTGCCGTTCCACCXTATAGCCGGXCTGCAGGXGAAGGTCACXGGG
Probe 09	XCGGAGTTGXCACACTAAGATXCAAGCGAAAGGXAGACCATGGGAGAAXGC
Probe 10	CXCTGCTCGCGXCTCCAGAGACXGTGGCAGGXGCAAGTTCATCTCAXCC
Probe 11	XGGGACCTXCCCATCCCACGXCGACAGGAACAXCAGGAGGTTCAXCAC
Probe 12	XCGGGATGGGTACXGTGGGTACGGAXACAATTGATGXGACCAGGTATGAXG
Probe 13	XCGATGAGGAGCCAGXTGTTAATGACAGXCTGAATGTXGGAGTAGAAGAGXA
Probe 14	XGCTTGGCCTTCTXAATAGTGTTCXGAATGTCCTGGXAAGTCTTAGAAXC
Probe 15	XGGAACCTTTAGCXCCGGACACGACCAXAGACTTGAAGXTATTGTATXCAG
Probe 16	XCACATCAAAGXCAGGCATTTCAXAGTAGACATXCACCCATTCCTGAXCC
Probe 17	GXGAGCTTCCGGXCAGTCATGTGCXTCCGATCCAGCXCCACCCGCAACAGXA
Probe 18	XCTTCCTCCTCXTGCATCTTGTXCTCATCGCTGTXCATGATGCGAAXACG
Probe 19	XCAGCATGXTGGACTCGAXGCAGCGCAGGAAGACAXCATCATCCATCTXG
Probe 20	CCAXCCTCCGTGAXGATGATCTTCTTCTXGTTGTCTGTCXGTGGCAAGXG
Probe 21	XTTCTCACXCAGCACCCGCAXCAAGCTCACGCCGXCCGTCTCCAGGAXCCAC
Probe 22	AAGXGTCGGTAATXGACATAGGAGCCAXCAAAGGAGAXGACGTGGTACAGXA
Probe 23	XAAGGAACACTXCATGAGTGGTCCXGTGTCCTGGCGGXTGACTCCGTGXCG
Probe 24	ACGXTGGCGAGTAGCXGGGAGACAXGGCACCACCTGGXGAAGGGATGXAG
Probe 25	GGAGAGGXCGGTGAGTAGCXGGGTGACGTXGGCGAATAGCXGGGTGATGXG
Probe 26	AAXTGGGACTGGTXGGAGAATAGTXCGGGCTGGXGGGTGAGTAACTXGGG
Probe 27	AXAGGTGGGACXGGTAGGCGAGXACTTGGGAGAGGXGGGTGAATATTXGG
Probe 28	TCXCCTCGTCACTGXCATCCGGGCTGAXAGCCGGGCTXGTGAGACTGXAG
Probe 29	TCXGCATCAGAAACGGGAXCCAGAAGTXCACCGGGAGCXCTGCCACAAGGXT
Probe 30	XCTTTGTTCTXCCCGAGGATCAGCXGTAACCACXCACAGCAGGAACXACCC

Sequence of probes against β-actin (X stands for amino-allyl T). Probes were labeled with Cy3 (GE Healthcare).

Probe 01	AXTGTAGAAGGXGTGGTGCCAGAXTTTCTCCATGXCGTCCCAGTTGGXGA
Probe 02	GCCXGGATAGCAACGXACATGGCTGGGGXGTTGAAGGXCTCAAACAXGAT
Probe 03	GAAGXCCAGGGCGACGXAGCACAGCTXCTCCTTAATGXCACGCACGATXT
Probe 04	AXGTCCACGTCACACXTCATGATGGAGXTGAAGGTAGXTTCGTGGAXGCC
Probe 05	XAACGCAACTAAGTCAXAGTCCGCCXAGAAGCATTXGCGGTGGACGAXGGA

Sequence of β -globin-Luc-CFP-24MS2 probes (X stands for amino-allyl T). Exonic probes were labeled with Cy3 (GE Healthcare) and intronic probes with Cy5 (GE Healthcare) following the protocols of suppliers. 0.5ng of each probe was used for hybridization.

Nature Methods: doi:10.1038/nmeth.2406

β-globine Exon 1	AGGAGXCAGGTGCACCAXGGTGTCTGTTXGAGGTTGCTAGXGAACACAGTA
β-globine Exon 2	GCCCAXAACAGCAXCAGGAGTGGACAGAXCCCCAAAGGACXCAAAGAACC
β-globine Exon 3-CFP	XGAACAGCTCCXCGCCCTTGCXCACCATGAATTCXTTGCCAAAGTGAXGG
Luciferase 1	GCGGXTCCATCCTCXAGAGGATAGAAXGGCGCCGGGCCXTTCTTTATGXT
Luciferase 2	TGTXCCAGGAACCAGGGCGXATCTCTTCAXAGCCTTATGCAGXTGCTCTCXA
Luciferase 3	XCCAACCGAACGGACAXTTCGAAGTAXTCCGCGTACGXGATGTTCACCXCG
Luciferase 4	XAACCGGGAGGXAGATGAGATGXGACGAACGTGTACAXCGACTGAAAXCCC
Luciferase 5	XAAAATAGGAXCTCTGGCAXGCGAGAATCXGACGCAGGCAGTTCTAXGCGG
MS2 NBX (12 repeat)	CXAGGCAATXAGGTACCTXAGGATCTAAXGAACCCGGGAATACXGCAGAC
β-globine Exon1-Intron 1	GXCTTGTAACCTXGATACCAACCXGCCCAGGGCCXCACCACCAACTTCATA
β-globine Intron 1	XCAGTGCCTAXCAGAAACCCAAGAGXCTTCTCTGTCXCCACATGCCCAGXA
β-globine Intron 2	XAGCAAAAGGGCCXAGCTTGGACXCAGAATAAXCCAGCCTTAXCCCAACCA

Sequence of probes against c-Fos (X stands for amino-allyl T). Probes were labeled with Cy3 (GE Healthcare).

c-Fos 188E	CXCGTAGTCTGCGTXGAAGCCCGAGAACAXCATCGTGGCGGXTAGGCAAAXA
c-Fos 288	XGACAGGCGAGCCCAXGCTGGAGAAGGAGXCTGCGGGTGAGTGGXAGTAAGXA
c-Fos 1123	XCCGGACTGGXCGAGATGGCAGXGACCGTGGGAAXGAAGTTGGCACXGGAG
c-Fos 1806E	XTGCGGCATTXGGCTGCAGCCAXCTTATTCCTTXCCCTTCGGATTCXCCT
c-Fos 2007E	XGGCAATCTCGGXCTGCAAAGCAGACXTCTCATCTTCXAGTTGGTCTGXC
c-Fos 2083	AGGXCATCAGGGATCTXGCAGGCAGGXCGGTGAGCXGCCAGGATGAACTA
c-Fos 2270E	GAAGXCATCAAAGGGCXCGGTCTTCAGCXCCATGCTGCXGATGCTCTXGA
c-Fos 2382	XAGCCACTGXGCAGAGGCTCCCAGXCTGCTGCAXAGAAGGACCCAGAXAGG
c-Fos 2485	XGAAGACGAAGGAAGACGXGTAAGCAGXGCAGCTGGGAGXACAGGTGACXT
c-Fos 2676	AXGTGTTTCTCCXCTCTGTAAXGCACCAGCXCGGGCAGTGGCACTTGXGG
c-Fos 2727	TXCACGCACAGAXAAGGTCCXCCCTAGGTCXACAGGAACCCXCTAGGGAA
c-Fos 2781	CXTGAGTCCACACAXGGATGCTTXCAAGTCCTXGAGGCCCACAGCCXGGT
c-Fos 2832	XGGAACAATACACACXCCATGCGTTTXGCTACATCXCCGGAAGAGGXAAGG
c-Fos 2883	CCAGGCCXGGCTCAACAXGCTACTAACXACCAGCTCTCXGAAGTGTCACXG

The modified oligonucleotide probes for MS2, RPB1, and c-Fos were synthesized by J-M. Escudier (Plateforme de synthèse d'Oligonucléotides modifiés de l'Interface Chimie Biologie de l'ITAV, Toulose, France). The modified oligonucleotide probes for β-actin and β-globin-Luc-CFP-24MS2 were synthesized by Eurogentec (Seraing, Belgium).

Imaging of MS2. 3D image stacks of cells after in situ hybridization were captured on a 100x NA 1,4 wide-field microscope (DMRA; Leica) equipped with a camera (CoolSNAP HQ; Roper Scientific and controlled by MetaMorph software (Universal Imaging Corp.). Pixel-size of 160 nm. Z-stacks of 61 images with a 300-nm Z-step were used.

Imaging of RPB1, β-actin, and β-globin-Luc-CFP-24MS2. 3D image stacks of cells after in situ hybridization were captured on a 100x NA 1.4 wide-field microscope (ECLIPSE Ti; Nikon) equipped with a camera (CoolSNAP HQ; Roper Scientific) and controlled by MetaMorph software (Universal Imaging Corp.). Pixel-size of 160 nm. Z-stacks of 51 images with a 200-nm Z-step were used.

RNAse protection assay

188 nucleotides fragment of bGH poly-adenylation signal spanning RNA cleavage site was amplified by PCR and cloned in pCRII-TOPO vector (Invitrogen). 32P-UTP labeled antisense RNA probe was synthesized with T7 RNA polymerase using Riboprobe in vitro transcription kit (Promega). RNAse protection assay was performed using Ambion kit RPAIII, according to manufacturer protocol. The protected fragments were run on 6% denaturing acrylamide gel, which was dried and exposed in Phosphoimager. 188 nucleotides band corresponded to noncleaved mRNA and 83 nucleotides band corresponded to cleaved mRNA. The intensities of the bands were quantified by ImageJ.

Nature Methods: doi:10.1038/nmeth.2406

References

1. Thomann, D., Rines, D. R., Sorger, P. K. & Danuser, G. Automatic fluorescent tag detection in 3D with super-resolution: application to the analysis of chromosome movement. *J Microsc* **208,** 49–64 (2002).

2. Henriques, R. *et al.* QuickPALM: 3D real-time photoactivation nanoscopy image processing in ImageJ. *Nat. Methods* **7,** 339–340 (2010).

3. Piotr's Matlab Toolbox. at <http://vision.ucsd.edu/~pdollar/toolbox/doc/>

4. Zhang, B., Zerubia, J. & Olivo-Marin, J.-C. Gaussian approximations of fluorescence microscope point-spread function models. *Appl. Opt.* **46,** 1819–1829 (2007).

5. Raj, A., Peskin, C. S., Tranchina, D., Vargas, D. Y. & Tyagi, S. Stochastic mRNA synthesis in mammalian cells. *PLoS Biol* **4,** e309 (2006).

6. Raj, A., Van den Bogaard, P., Rifkin, S. A., Van Oudenaarden, A. & Tyagi, S. Imaging individual mRNA molecules using multiple singly labeled probes. *Nat Meth* **5,** 877–879 (2008).

7. Parthasarathy, R. Rapid, accurate particle tracking by calculation of radial symmetry centers. *Nat. Methods* **9,** 724–726 (2012).

8. Ober, R. J., Ram, S. & Ward, E. S. Localization accuracy in single-molecule microscopy. *Biophys. J* **86,** 1185–1200 (2004).

9. Kirshner, H., Sager, D. & Unser, M. in *Proceedings of the Twelfth International Conference on Methods and Applications of Fluorescence Spectroscopy, Imaging and Probes* 154 (2011).

10. Zenklusen, D., Larson, D. R. & Singer, R. H. Single-RNA counting reveals alternative modes of gene expression in yeast. *Nat. Struct. Mol. Biol* **15,** 1263–1271 (2008).

11. Femino, A. M., Fay, F. S., Fogarty, K. & Singer, R. H. Visualization of single RNA transcripts in situ. *Science* **280,** 585–590 (1998).

12. Boireau, S. *et al.* The transcriptional cycle of HIV-1 in real-time and live cells. *J. Cell Biol.* **179,** 291–304 (2007).

13. Decker, C. J. & Parker, R. P-bodies and stress granules: possible roles in the control of translation and mRNA degradation. *Cold Spring Harb Perspect Biol* **4,** a012286 (2012).

14. Müller, W. G., Walker, D., Hager, G. L. & McNally, J. G. Large-scale chromatin decondensation and recondensation regulated by transcription from a natural promoter. *J. Cell Biol* **154,** 33–48 (2001).

15. Darzacq, X. *et al.* In vivo dynamics of RNA polymerase II transcription. *Nat. Struct. Mol. Biol.* **14,** 796–806 (2007).

16. Lawrence, J. B., Singer, R. H. & Marselle, L. M. Highly localized tracks of specific transcripts within interphase nuclei visualized by in situ hybridization. *Cell* **57,** 493–502 (1989).

17. Bellemer, C. *et al.* Microprocessor dynamics and interactions at endogenous imprinted C19MC microRNA genes. *J. Cell. Sci.* **125,** 2709–2720 (2012).

Nature Methods: doi:10.1038/nmeth.2406

18. Vitali, P., Royo, H., Marty, V., Bortolin-Cavaillé, M.-L. & Cavaillé, J. Long nuclear-retained non-coding RNAs and allele-specific higher-order chromatin organization at imprinted snoRNA gene arrays. *J. Cell. Sci.* **123,** 70–83 (2010).

19. Royo, H. *et al.* Bsr, a nuclear-retained RNA with monoallelic expression. *Mol. Biol. Cell* **18,** 2817–2827 (2007).

20. Wegel, E. & Shaw, P. Gene activation and deactivation related changes in the three-dimensional structure of chromatin. *Chromosoma* **114,** 331–337 (2005).

21. Belmont, A. S. & Straight, A. F. In vivo visualization of chromosomes using lac operator-repressor binding. *Trends Cell Biol.* **8,** 121–124 (1998).

22. Tilgner, H. *et al.* Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. *Genome Res.* **22,** 1616–1625 (2012).

23. Schmidt, U. *et al.* Real-time imaging of cotranscriptional splicing reveals a kinetic model that reduces noise: implications for alternative splicing regulation. *J. Cell Biol.* **193,** 819–829 (2011).

24. Golding, I., Paulsson, J., Zawilski, S. M. & Cox, E. C. Real-time kinetics of gene activity in individual bacteria. *Cell* **123,** 1025–1036 (2005).

25. Fraefel, C. *et al.* Spatial and temporal organization of adeno-associated virus DNA replication in live cells. *J. Virol.* **78,** 389–398 (2004).

26. Tsukamoto, T. *et al.* Visualization of gene activity in living cells. *Nat. Cell Biol* **2,** 871–878 (2000).

27. Darzacq, X. *et al.* Stepwise RNP assembly at the site of H/ACA RNA transcription in human cells. *J. Cell Biol.* **173,** 207–218 (2006).

28. Fusco, D. *et al.* Single mRNA molecules demonstrate probabilistic movement in living mammalian cells. *Curr. Biol.* **13,** 161–167 (2003).

Nature Methods: doi:10.1038/nmeth.2406