The University of Manchester

Multi-scale Analysis of Chromosome and Nuclear Architecture

A thesis
submitted to The University of Manchester
for the degree of
doctor of philosophy (PhD)
in the faculty of Faculty of Life Sciences

Pedro Humberto Olivares-Chauvet

Faculty of Life Sciences
2012

Contents

Final word count: 36457

List of Figures

List of Tables

ABSTRACT

Mammalian nuclear function depends on the complex interaction of genetic and epigenetic elements coordinated in space and time. Structure and function overlap to such a degree that they are usually considered as being inextricably linked. In this work I combine a experimental approach with a computational one in order to answer two main questions in the field of mammalian chromosome organization.

In the first section of this thesis, I attempted to answer the question to what extent does chromatin from different chromosome territories share the same space inside the nucleus? This is a relatively open question in the field of chromosome territories. It is well-known and accepted that interphase chromosomes are spatially constrained inside the nucleus and that they occupy their own territory, however, the degree of spatial interaction between neighbouring chromosomes is still under debate. Using labelling methods that directly incorporate halogenated DNA precursors into newly replicated DNA without the need for immuno-detection or in situ hybridization, we show that neighbouring chromosome territories colocalise at very low levels. We also found that the native structure of DNA foci is partially responsible for constraining the interaction of chromosome territories as disruption of the innate architecture of DNA foci by treatment with TSA resulted in increased colocalisation signal between adjacent chromosomes territories.

The second major question I attempted to answer concerned the correlation between nuclear function and the banding pattern observed in human mitotic chromosomes. Human mitotic chromosomes display characteristic patterns of light and dark bands when visualized under the light microscope using specific chemical dyes such as Giemsa. Despite the long standing use of the Giemsa banding pattern in human genetics for identifying chromosome abnormalities and mapping genes, little is known about the molecular mechanisms that generate the Giemsa banding pattern or its biological relevance. The recent availability of many genetic and epigenetic features mapped to the human genome permit a high-resolution investigation of the molecular correlates of Giemsa banding. Here I investigate the relationship of more than 50 genomic and epigenomic features with light (R) and dark (G) bands. My results confirm many classical results, such as the low gene density of the most darkly staining G bands and their late replication time, using genome-wide data. Surprisingly, I found that for virtually all features investigated, R bands show intermediate properties between the lightest and darkest G bands, suggesting that many R bands contain G-like sequences within them. To identify R bands that show properties of G bands, I employed an unsupervised learning approach to classify R bands on their genomic and epigenomic properties and show that the smallest R bands show a tendency to have characteristics typical of G bands. I revisit the evidence supporting the boundaries of G and R bands in the current cytogenomic map and conclude that inaccurate placement of weakly supported band boundaries can explain the intermediate pattern of R bands. Finally, I propose an approach based on aggregating data from multiple genomic and epigenomic features to improve the positioning of band boundaries in the human cytogenomic map. My results suggest that contiguous domains showing a high degree of uniformity in the ratio of heterochromatin and euchromatin sub-domains define the Giemsa banding pattern in human chromosomes.

DECLARATION

The University of Manchester

Candidate Name: Pedro Humberto Olivares-Chauvet

Faculty: Faculty of Life Sciences

Thesis Title: Multi-scale Analysis of Chromosome and Nuclear Architecture

Declaration to be completed by the candidate:

I declare that no portion of this work referred to in this thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Signed: Date: March 4, 2013

COPYRIGHT

[1]This excludes material already printed in academic journals, for which the copyright belongs to said journal and publisher. Pages for which the author does not own the copyright are numbered differently from the rest of the thesis.

ACKNOWLEDGEMENTS

Finally, after all this time, it is a pleasure for me to be writing these lines. Along this way I have been always thinking about all the people that had supported me through this adventure. But before I get to names, there is someone else I would like to thank.

Is not an exaggeration to say that 99% of this work has been done with open source and free software. From the most simple command-line applications to the super computer cluster of the faculty (thanks also to people like Nick Gresham), passing through the operative system I use day to day and the graphics design software. Without all the volunteers and people who develops open source and free software, I would not have been able to achieve what I did on this work. Before expressing my gratitude to anyone else, I want to thank all these faceless people out there.

Also in the faceless side of acknowledgements, I give thanks the Mexican tax payers who sponsored my studies through CONACyT.

The second group of people to thank is all those that I forget to mention in the next lines.

The second-and-a-half group of people I want to acknowledge is not the people who where here with me, but the ones that I missed during these four years: Pablo Rodriguez, Ana Gutierrez, Jose Carlos Ramon alias “El Charli”, Carlos Espinosa, Federico Sanchez Quinto, the famous Hermes “Chafo” Oropeza, Mary and Juanita too. Also to Flavia Zorrilla who made the effort to visit the neighborhood where Morrisey lived, and taking advantage of that, also visited me. I want to thank and wish the best to Alejandro de Coss who will join us soon in the UK. Friends as Mayra Furlan, Edgar Gonzales and Erandi Ayala from the academic side, with special mention to Felix Recillas for the discipline I learned from him. And lastly in this category to the ones that are not around anymore, Lilia Gutierrez.

But life takes from you as much as it gives back. I am so glad that during these years I met all these simians and I thank for: Matias Vidal an his inspiring curiosity about reality, Cristobal Espinoza for all the adventures and advices, Alicia O’Donald for all her unique

warmth and happiness, and also for her excellent company as a disciplined swimmer. Ricardo Rodriguez for his “eeeeeeeeehhhh” and great sense of humor. I thank for all the good times to my “Mancunian parents”: Max Haeussler and Helene Auger. Knowing Silvia Pineapple and Cesar Camacho helped me immensely against homesickness. Wambui Mburathi for all the laughs, drinks and food we shared. Thanks too to the Mexican cycling society.

I specially thank Apolinar Maya for all his unconditional help at any moment, for all the adventures together and for treating me as a member of his family. Obvious greetings to Asia del Tigre and Matyldita.

I thank also to all the members of the Jackson and Bergman lab, specially to Michael Barton who represented the mayor source of tricks.

I am very grateful with David Romero. Thanks to him I came here and I am going there. Federico Sanchez ’papa’ for his support back in the day and during my PhD, my thanks as well to his family.

I want to thank to my supervisor Dean, who first invited me to Manchester, for all the patience he had with me and my rubbish English, but specially for being always funny. The simplicity and friendliness of his person is what I take. I also thank Casey for promoting always the most rigorous and serious perspective in science, making me aware of the big science and small science. I thank him as well for the lessons we shared at the most difficult times of life (hi Martin!) as much as in the good ones. I thank you bosses, you are good friends!

Thank you Wendy and Matt for the great experience that my Viva voce was. Thank you for taking the time to read my work and for the great discussions we had.

And finally I want to dedicate this work to the closest persons I have. Ana Kozomara, whom with I learned many things and will be part of me always. And at last my parents and sister, Romel Olivares, Michelle Chauvet and Irene Olivares who I have always been proud of.

Nomenclature

3C	Chromosome conformation capture
3D	3-dimensional
Alu	Arthrobacter luteus restriction endonuclease site
BED	Browser Extensible Data
BP	Band-pass
bp	Basepairs
CDS	Protein coding sequence
CGIs	CpG Islands
ChIP	Chromatin Immunoprecipitation
CsCl	Caesium chloride
CT-IC	Chromosome Territory-Interchromatin Compartment Model
DAM	DNA adenine methyltransferase
DHFR	Dihydrofolate Reductase Locus
dTTP	2'-Deoxythymidine Triphosphate
dUTP	2'-Deoxyuridine, 5'-Triphosphate
EF	Emission Filter
ESC	Embryonic Stem Cells
FACS	Fluorescence-activated Cell Sorting
FRAP	Fluorescence Recovery after Photobleaching
FRAP	fluorescence recovery after photobleaching
Gb	Giga base pairs
GFP	Green Fluorescent Protein
GO	Gene Ontology
HDAC	Histone Deacetylase
HeLa	Henrietta Lacks cells
HiLo	Hi Low indicator lookup table
hiPSC	Human Induced Pluripotent Stem Cells
HP1	Heterochromatin protein 1
HSs	DNase I hypersensitive sites
ICCB	Intensity correlation coefficient-based
ICD	Interchromosome Domain
ICN	Interchromosomal Network Model
IR	Infra-red
LINE	Long Interspersed Nuclear Elements
LP	Low-pass
LTR	Long Terminal Repeat
LUT	Lookup table
NA	Numerical Aperture
nm	Nanometer
PC	Pearson's coefficient
PML	Promyelocytic leukemia bodies
Pol	RNA polymerase II
PR	Perichromatin region
Rif1	Rap1-interacting-factor-1
rRNA	Ribosomal ribonucleic acid
SARS	Scaffold Attachment Region Se-

quences

scRNA Small Cytoplasmic Ribonucleic Acid

SINE Short Interspersed Nuclear Element

SMC Structural Maintenance of Chromosomes protein family

snRNA Small nuclear ribonucleic acid

TE Transposable Element

TSA Trichostatin A

TSS Transcription start site

VRML Virtual Reality Modelling Language

WIG Wiggle File Format

Chapter 1

AN OVERVIEW OF CHROMOSOME AND NUCLEAR ARCHITECTURE

1.1 General Perspective

This introductory chapter provides a general overview of the field of chromosome and nuclear organization. Each results chapter provides a more detailed introduction that contextualizes the key questions relevant to that chapter.

The general introduction attempts to follow a similar structure and hierarchy to that seen inside the mammalian nucleus. Beginning from the smallest scale, we discuss how chromatin organises at the level of nucleosomes and how epigenetics represent the platform for complexity to develop. We describe the different scales at which chromatin is structured and compartmentalised in order to control different nuclear processes. Finally we finish by exploring the largest scales of chromosome organization represented by condensed mitotic chromosomes.

1.2 Chromatin

1.2.0.1 The nucleosome and epigenetics

One of the most remarkable features of the eukaryotic nucleus is its ability of organize the extremely large molecule that the genome represents in space. Since the discovery of the structure of DNA in 1953 [1], major advances in the understanding of how the nucleus is able to organize itself in space and function had been achieved. It is now well-known that DNA in its naked form is almost never found in the living cell. It is arranged in a protein-nucleic acid complex that is able to self-organize at different hierarchical levels [2]. This complex is known as chromatin, which literally means 'stainable material', a term coined by the German anatomist Walther Flemming [3] on seminal work of mi-

croscopy in the late 1870's.

The most basic level of chromatin organization is defined by the local, stable interaction of DNA with a protein octamer called the 'nucleosome', first isolated in 1974 by Roger Kornberg [4]. The term nucleosome was first used specifically for this nuclear element in 1975 by Oudet [5]. The nucleosome is composed of a pair of each of 4 core proteins called histones (H2A, H2B, H3 and H4) to which DNA wraps around 2 times in approximately 165 basepair (bp). It took more than two decades to resolve the crystal structure of the nucleosome [6].

Histones are not only responsible for an efficient compaction of DNA, which ranges in the order of 5- to 10-fold [4], but also for playing an important role in the regulation of the biological function of the DNA they organize.

Histones contain N-terminal regions that protrude out of the body of the nucleosomes and the amino acids located on these 'tails' are target for numerous covalent modifications [7]. These modification are the foundation of an additional layer of complexity that orchestrates intricate cellular functions, such as regulation of gene expression which in turn manifests at the level of cellular communication and development of multi-cellular organisms. The study of these posttranslational histone modifications, together with DNA methylation, is the focus of the field of epigenetics.

Epigenetics

The term 'epigenetics' defines the inheritance of cellular traits (phenotype), through non-genetic mechanisms, from one cell generation the next one. They can be summarized in two main categories, modifications that directly mark the DNA (methylation) or, as mentioned above, modifications that lead to covalent addition of different chemical moieties to amino acids residing in histones, usually on their protruding tails.

The large number of possible modifications that histone tails can undergo, plus the interdependence and relationship between these modification gave birth to the model of the 'histone code' [8] which states that the biological function of a region of the genome can be decoded by understanding the combination of its histone modifications.

In general, there are two types of histone modifications; ones where the residues added

to the histones are able to change the electric charge (e.g. acetylation) of the nucleosome, and ones where signalling 'flags' can be recognized by other proteins that are able to perform specific functions [7] (Table 1.1). Acetylation is a classical example of chromatin modification that changes the electric charge of nucleosomes as it removes positive charges from the histones, decreasing the strength of the interaction between the nucleosome and the negative charge of the DNA, thus rendering a more open conformation (e.g. acetylation lysine 16 of histone H4) [9, 10, 7]. An example of a signaling mark is the methylation of histone H3, which promotes binding of HP1 protein [11]. The synergistic activity of different chromatin marks show different biological functions, for instance, the combination of HP1, methylation of histone H3 lysine 9 (H3K9me) and H2A.Z promote a more compact chromatin configuration [12]. A summary of the most common chromatin modifications is shown in Table 1.1.

Chromatin structure is directly linked to its function, therefore changes in its spatial conformation will translate to different biological properties. Nucleosomes are understood as the unit of chromatin organization but the cell functions at more complex levels of chromatin organization.

1.3 Chromatin structure

1.3.1 The interphase chromatin

The unit of chromatin organization that the nucleosome represent is periodically repeated giving form to a chromatin fiber which resembles a fiber of 'beads on a string' of 10 nanometer (nm) diameter (Fig. 1.1).

In mammalian cells, the average distance between a nucleosome and the next is $\sim$35 bp. This piece of DNA is generally referred to as 'linker DNA' for obvious reasons. Depending on the biological function chromatin presents, the length of the linker DNA can vary and nucleosome can be re-positioned in order to fit a particular function. For instance nucleosome positioning has an influence in gene regulation, as the regulatory sequences can be exposed/hidden from the regulating trans-acting elements [14, 15]. Additionally, particular patterns of nucleosome arrangements have been observed for the insulator protein CTCF which is able to re-arrange $\sim$20 nucleosomes around its binding sites [16, 17].

Histone modification or variant	Biological functions
H2A.Z	Histone 2 variant (H2A.Z) associated with regulatory elements of dynamic chromatin
H3K4me1	Associated with enhancers and other distal elements. Enriched downstream of TSS
H3K4me2	Associated with promoters and enhancers
H3K4me3	Associated with promoters and TSS
H3K9ac	Active mark commonly associated with promoters
H3K27ac	Mark of active regulatory elements; may distinguish active enhancers and promoters from their inactive counterparts
H3K79me2	Transcription-associated mark, with preference for 5' end of genes
H3K9me1	Preference for the 5' end of genes
H4K20me1	Preference for the 5' end of genes
H3K36me3	Elongation mark associated with transcribed portions of genes, with preference for 3' regions after intron 1
H3K9me3	Associated with inactive chromatin. Found in constitutive heterochromatin and repetitive elements
H3K27me3	Repressive mark established by polycomb complex. Found in repressive domains and silent developmental genes

Table 1.1: Summary of posttranslational modification of histone proteins and their biological functions, from [13]. TSS = Transcription Start Site

Occupation of the linker DNA by the histone H1, or 'linker histone' is know to mediate and stabilize the interaction between adjacent nucleosomes on the polynucleosome fiber and allow the organization of chromatin at higher-order levels [18]. H1-rich chromatin shows a specific structural motif supporting its role as a key higher-order element [19], with depletion of H1 leading to improper folding of chromosomes in mitosis [20, 21].

Aaron Klug proposed the solenoid 30 nm fiber model as a supercoiled configuration of the 10 nm chromatin fiber based on EM imaging of chromatin fibers reconstituted *in vitro* [22] (Fig. 1.1). This model has been challenged recently [23], arguing against it by evidence based on more modern studies of cryo-EM of chromatin that could not detect such level of organization [24, 25].

It has been proposed that, in analogy to protein secondary structure, chromatin *in vivo* can self-organize into different configurations, as the arrangement of specific periodic sequences of amino acids in proteins give rise to particular structural domains [18], however, it is still not clear the mechanisms that organize chromatin above the 30 nm fiber.

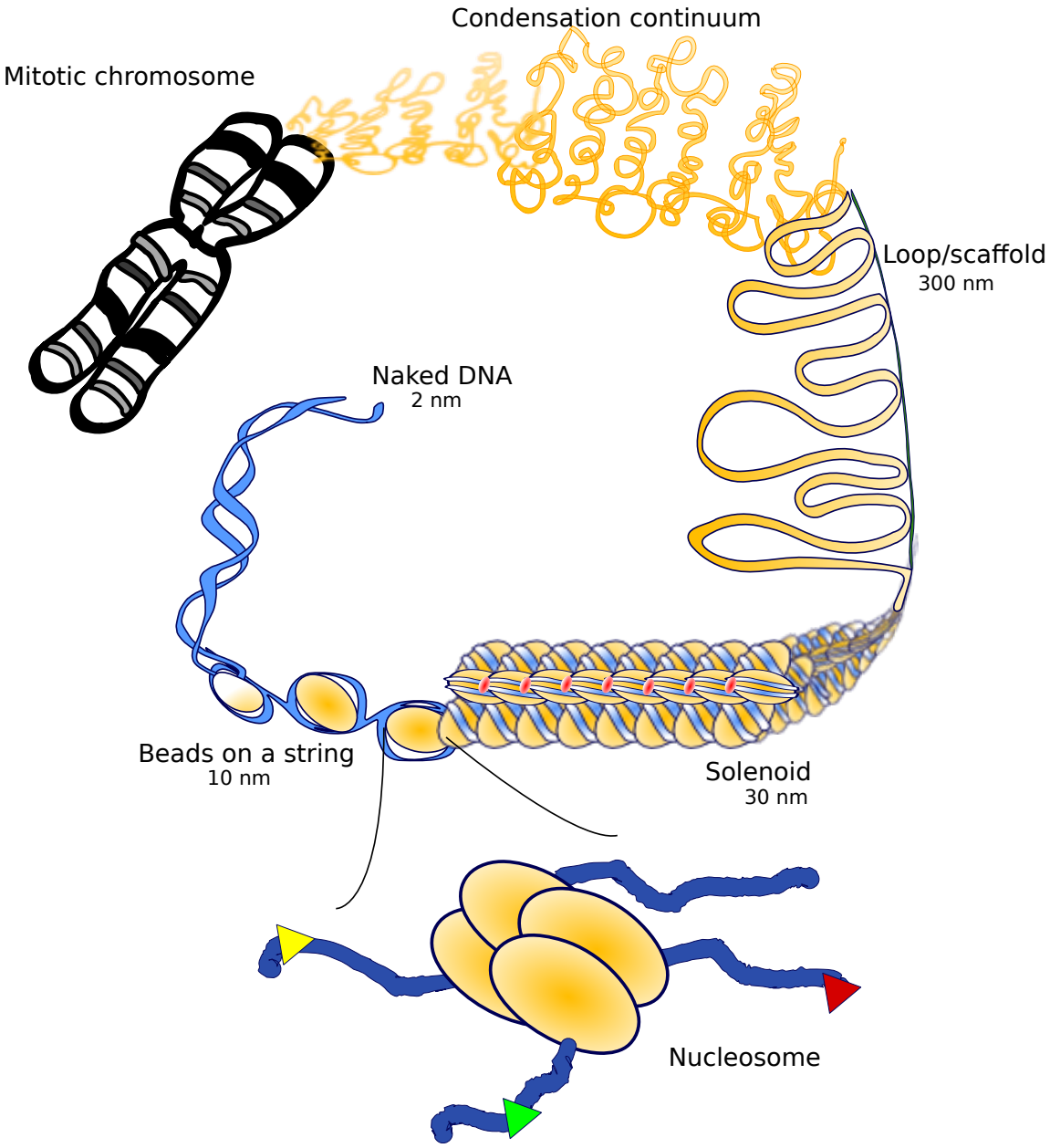


Figure 1.1: Hierarchical compaction of DNA and chromatin. Naked DNA is not found inside the mammalian nucleus, instead it wraps around histone proteins and give structure to the basic unit of chromatin organization, the nucleosome. Nucleosomes are the target of different posttranslational modifications with different biological functions. Nucleosomes connect to each other through DNA, resembling a "beads-on-a-string" fiber (10 nm). Under certain conditions, chromatin has been observed as a fiber of 30 nm in diameter. Chromatin is able to structure into higher-order conformations which result in the condensed mitotic chromosomes.

1.3.1.1 Local Chromatin Domains

Despite the gap in understanding of the molecular mechanisms that give structure to chromatin above the 10 or 30 nm scale, there is evidence that chromatin organizes itself into constrained spatial domains.

Discrete chromatin domains can be observed in light microscopy by staining DNA of synchronized HeLa cells with short pulses of thymidine analogues during S phase [26, 27] (Fig. 1.3) and confirmed recently observed in EM [28] as 100 nm structures (Fig. 1.3a). These replication foci are persistent after many cell generations and range from the order of hundreds of kb to several Mb [29], and interestingly, the same distribution of DNA foci can be seen as mirror images between daughter cells [30, 31], suggesting the inheritance of local architecture from one generation to the next.

These domains contain on average 1Mb of DNA [26, 27], and based on this feature, are usually referred in literature as the '1Mb chromatin domains'. Modern models such as the 'fractal globule' model [32] (Fig. 1.2), incorporate this level of organization of chromatin and have recently confirmed the presence of DNA foci by analysis of genome-wide chromatin interaction maps [32, 33].

1.3.2 Higher-order chromatin structure

1.3.2.1 Heterochromatin and Euchromatin as functional compartments

As mentioned above, the ways in which chromatin organizes above the proposed solenoid fiber of 30 nm is very complex and there is no accepted model in the field (Fig. 1.1). However some interesting models have been proposed and are discussed at the end of this section.

Many years before the description of the structure of DNA, the main nuclear compartments had already been studied. Heterochromatin was first described in the early 20th century due to its remarkable staining intensity using basic dyes [34, 35]. This chromatin form is found specifically in the periphery of the nucleus, covering the nuclear envelope from the inside and surrounding the nucleolus. The rest of the nuclear volume is occupied by euchromatin where transcription takes place. Heterochromatin depends on the association of certain non-histone proteins that, as an ensemble, provide it with its biological

properties. A classical instance of such proteins is the heterochromatin protein 1 (HP1) family.

It was generally thought that heterochromatin was a static compartment of the nucleus where even molecular access was restricted, thus, any gene that happened to be 'buried' in heterochromatic regions would be doomed to silencing. This view has changed as fluorescence recovery after photobleaching (FRAP) experiments and green fluorescent protein (GRP)-HP1 [36, 37]. These experiments showed a remarkably quick recovery of the fluorescent signal after laser-bleaching of heterochromatin regions. These findings suggest that HP1 turnover is happening constantly and that repression and compaction of heterochromatic domains is an active process rather than a static, determined state. In addition to this, not all the cells share the same heterochromatin regions. The concept of facultative heterochromatin defines such chromatin that can be present as euchromatin in some cases and heterochromatin in others [38]. These changes in function are usually related to cell differentiation during development. A very popular example that shows the degree of plasticity of the nuclear compartments is provided by an inverted occupation of euchromatin and heterochromatin in the nuclear space. When comparing the nuclei in retinal pigment epithelial cells of nocturnal and diurnal mammals, Boris Joffe's group [39] showed that heterochromatin is not localized at the periphery of the nucleus but is rather found at the center, and vice versa, euchromatin occupies the nuclear regions typical of heterochromatin in the normal mammalian nucleus. This observation suggest a model in which this configuration of chromatin serves as 'collecting lenses' to improve the night vision of these animals [39].

Chromatin as a fractaloid structure

Molecular dynamic approaches have been implemented to simulate the different scenarios in which chromatin can be structured. In these studies chromatin is modeled as a polymer and the parameters used in the model are based on biological properties of chromatin such as the beads-on-string structure and the length of the DNA around each nucleosome, the length of the linker DNA. More complex models incorporate additional information such as topology-regulating proteins (CTCF and cohesins), etc.

A popular model called the "equilibrium globule" has been proposed [40, 41, 42], where the modelled polymer shows a very compact configuration of highly knotted nature. This model follows a random walk inside the nuclear volume. When the random-walk reaches the limit of the nuclear volume, it turns into a different direction and follows a new random walk-path.

Inspired in the model of the "crumpled globule" proposed by Grosbeg et al. [43, 44], Job Dekker's group has recently proposed the scale-independent organization of chromatin, referred as the fractal globule model, based on genome-wide interaction maps derived from the chromosome conformation capture (3C) (Fig. 1.2). This model was developed by comparing the decay of chromatin contacts, as a function of the linear genomic distance, with computer simulations of polymer following space-filling curves, such as the Peano or Hilbert curves [32]. More recent Hi-C experiments with a higher resolution have confirmed the emergence of topologically associated domains of 1Mb of size [33].

The advantages of this idea is that this kind of spatial configuration is less prone to entanglements, re-modelling of chromatin requires less energetic costs and is very space-efficient as it allows access to a large number of sequences in a small volume. The notion that chromatin organizes as a three-dimensional fractal has been re-enforced by the behaviour of fluorescent marks and how they diffuse in living cells [45]. The fractal globule model has been compared the equilibrium model and shown to be closer to how chromatin folds inside the nuclear volume.

1.4 Chromatin in the mitotic chromosome

1.4.1 Chromatin Condensation

One of the most obvious structures that can be observed under the light microscope in the mammalian nucleus are chromosomes in their condensed configuration during mitosis. How chromosomes condense during mitosis still remains as a big question in the field as the degree of compaction of the DNA in the nucleus falls in the range of 10,000- to 20,000-fold [35]. One of the first steps towards a better understanding of the mitotic chromosome was achieved by Taylor and colleagues in 1957, they demonstrated that one

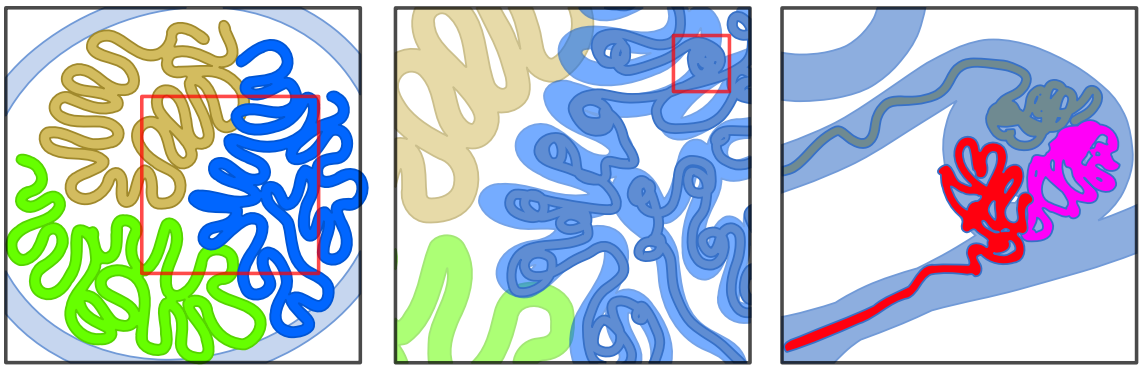


Figure 1.2: The fractal globule model. This model proposes a scale-independent structure of chromatin. A fractal globule is conformed by an unentangled fiber that spontaneously folds in such a way that globular structures are made from smaller globular structures, which in turn are made of even smaller globular structures. One advantage of this structure as a model for chromatin organization, is that this kind of object lacks knots and can be easily unfolded and refolded [32].

single molecule of DNA represented a single chromatid of each chromosome from telomere to telomere (*uninemy*) [46].

Different structural arrangements of chromatin in the mitotic chromosome preparations could be observed depending on the buffer conditions used. High salt and detergent containing buffers were able to extract all the histones and most of the nonhistone proteins and by using this method DNA 'halos' were observed, showing a core proteinaceous structure that resembled the shape of the condensed intact chromosome, with morphological features such as parallel sister chromatids joined together at the centromere [47]. These observations led to the proposal of the loop/scaffold model. This model stated that loops of DNA emanate from a mitotic core axis/scaffold made of nonhistone proteins [47, 48, 49].

In the early 80's, when isolating chromosomal scaffolds and extracting the residual proteins, two metalloproteins were found: Sc1 and Sc2, of 170,00 and 135,000 daltons respectively [50]. Sc1 protein was found to be Topoisomerase II [51, 52] and Sc2, a member of the Structural Maintenance of Chromosomes (SMC) protein family [53, 54] that includes condensins and cohesins. Both Sc proteins are essential for proper chromosome condensation. A few years later, based on the isolation of nonhistone protein scaffolding structures from mitotic chromosomes preparations and electron microscopy, the core structure of mitotic chromosomes model was confirmed [55]. This fibrous structure showed identical sizes and similar landmarks as the ones observed in intact chromo-

somes, e.g. kinetochore attachment regions and the chromatid axis. It was also observed that these chromosome preparations were able to fold and unfold, reversibly, depending on the buffer conditions used, suggesting a that the same dynamic chromosome condensation mechanism present in mitosis was somehow encoded in the spreads [55].

The radial loop model implied that there could be specific DNA sequences that bind to the chromosome scaffold. These sequences were then found by digesting with nuclease and analyzing the sequences that were left behind in the proteins scaffold and were named Scaffold Attachment Regions (SARs). SARs showed no specific sequence motif but AT-rich sequences with high-affinity to Topoisomerase II. It was proposed that the banding pattern of chromosomes observed after special staining procedures was due to different densities of SARs that resulted into transitions in the size of the loops [56].

Mitotic chromosomes show a constant diameter but the mechanisms that define such feature are still not very well understood [35]. If SARs are the genetic elements that define the looping frequency, a constant diameter would then require equally spaced SARs along the genome sequence so the length of each emanating loop from the scaffold have the same length. This was shown to be wrong as different lines of evidence suggested that chromosome constant width is independent of DNA sequence. Naturally occurring tandem repeats as the dihydrofolate reductase (DHFR), reported by Bickmore's group [57], show no difference in width in the condensed chromosomes. Furthermore, engineered chromosomes did not present variation of their diameter despite variations on the density of SARs [58]. It was later confirmed that there are non genetic elements orchestrating this complex process. For instance, in histone H1 depleted systems, chromosomes appeared longer and with a reduced chromosome diameter [20, 21], supporting the role of the linker histone H1 as a higher-order structure determinant factor.

A hierarchical folding model ('axial glue' model [59]), that challenges the radial loop model, was proposed based on the observed progressive thickening of the chromosomes and a condensed axial distribution of condensin [60]. Additionally, there is evidence based on engineered, Lac repressor stainable chromosome segments, that suggests the existence of a structure of around 200-300 nm that turns around, producing a 400-600 nm chromatid [58].

To date there is no generally accepted model for mitotic chromosome structure that satisfies all the observations. The process of chromatin condensation after genome duplication and prior to cell division may involve complex interdependent events between genetic and non genetic factors.

1.5 Chromatin structure and genomic function

1.5.1 Relationship between the S phase programme and higher-order chromatin organization

Many different nuclear processes are known to depend on chromosome architecture. Proper gene expression needs a chromatin milieu that allows the transcriptional machinery to be loaded and the regulatory elements in *cis* to be located by their respective transcription factors. Initiation of DNA replication is also known to require particular chromatin conditions to fire. In this section we will explore in more detail the link between replication and chromatin architecture as it represents the best instance of the structure:function equivalence of the eukaryotic nucleus.

As the dimension of eukaryotic genomes is several orders of magnitude larger than bacterial genomes, replication is parallelized in order to complete this task in a time efficient manner. It is known that the genome is replicated at different rates and in a differential firing sequence of origins throughout S phase [61]. This temporal pattern is known as the S phase programme.

An important element for the regulation of replication timing is the activation of origins of replication at different times of S phase. In fission yeast (*Schizosaccharomyces pombe*) , levels of different proteins regulated by cell cycle proteins associate at specific times in order to fire the set of origins under regulation. For instance, Sld3,Sld7 and Cdc45 proteins are essential for the temporal order of origin firing [62], in conjunction with the telomere-binding protein Taz1 [63] that controls the timing regulation of replication of almost one half of the late replicating origins in the genome. By repressing initiation of replication by DNA elements located close to late replicating origins during early S phase. TRF1 and TRF2 are the human counterpart of Taz1 [63]. In addition, Rap1-interacting-factor-1 (Rif1) also regulates the determination of late-replicating domains in fission yeast

[64] and in human [65]. Replication of DNA in higher-eukaryotes shows a more complex regulation and dynamics.

Constitutively active genes, house-keeping genes for instance, tend to replicate early during S phase [66]. On the other hand, tissue-specific genes tend to replicate during the later stage of S phase in most tissues, and replicate early in the tissue of expression [67]. It has been observed that budding yeast does not show any correlation between transcriptional activity and time of replication [68] whereas human lymphoblastoid cells [69, 70] and fruit fly Kc cells [71] show the opposite trend, with only a few cases of chromosomal domains replicating independently of their transcriptional activity. Even though gene expression certainly has some degree of influence on the time of replication, the relationship between transcription and early replication cannot be generalized. Furthermore, our understanding of the molecular mechanisms and how the genomic landscape and chromatin modifications influence replication is still very poor.

The first genome-wide approach made to understand the link between genome replication and transcriptional activity in higher eukaryotes came from a time of replication genomic profile performed in *Drosophila melanogaster* [71]. Kc cells were sorted using fluorescence-activated cell sorting (FACS) by DNA content in early-S and late-S populations then pulse labeled with BrdU. BrdU-labeled DNA from each sample was then isolated and co-hybridized on a microarray representing 6,500 genomic regions in which 5,543 belonged to genes (over 40% of all *D. melanogaster* genes). Results from the microarray were then confirmed by semi-quantitative PCR and compared with expression data from the same array platform including 5,077 genes, where $\sim$32% of the genes showed no expression. A general trend linking transcription with replication was found; the earlier in S phase when a gene is replicated, the more probable it is to be expressed.

Different large-scale microarray hybridization analyses addressing the S phase temporal programme in human cells have been performed [72, 69, 70, 73, 74]. Of this list, the most comprehensive of these studies was based on genome-tiling microarrays of the ENCODE pilot regions [74]. By comparing array data of total RNA with time of replication, the general trend linking transcriptional activity with early-replicating regions was confirmed. Early replicating regions are $\sim$5.34 times more transcribed than late replicat-

ing regions. The most significant result from this study was the confirmation of previous observations suggesting that some proportion of the genome replicates asynchronously (pan-S phase; pan-S) [73] and adjusted it from 60% to a 20% estimation of pan-S segments. Asynchronous replication is defined when a chromosome segment yields both timing patterns, early and late. In other words, it does not show a time-specific pattern of replication. By performing interphase FISH, it was demonstrated that the asynchrony in replication was between alleles of the same gene in the same cell and not in dissimilarities in replication patterns between chromosomes from distinct cells. Genomic regions showing a pan-S behaviour were enriched in H3K4 methylation and H3K9 di-methylation, histone modifications characteristic of active and silenced chromatin respectively.

Not all silent chromatin modifications correlate with replication timing. Polycomb regulated domains, characterized by the H3K27me3 post-translational chromatin modification do not show any correlation with late replicating regions [75, 76], mutant cells lacking a crucial element of the polycomb group showed no difference in their pattern of replication timing [77].

Most genome-wide analysis addressing the temporal replication patterns of the genome segment S phase in early and late replicating fractions. Sub-chromosomal segments left between these two extremes are classified as middle replicating. A recent study in lymphocyte mouse cells (L1210 cell line) using genomic arrays representing 80% of the sequenced mouse genome (1.9 Giga base pairs (Gb)) segmented S phase in 7 time zones [66]. Cells were collected by using a 'retroactive' synchronization method based on membrane eluted mitotic cells called the 'baby machine' [78, 79]. This method relies on separation of mitotic cells by letting a stream of medium drop from a membrane covered by growing cells. Unsynchronized cells were pulse labeled with BrdU and then collected from the membrane at different time points. BrdU-labeled DNA was immuno-precipitated and samples validated by PCR using genes characteristic of Early, Mid and Late phases looking for enrichment on their respective temporal replication signature. Samples were then hybridized onto genomic microarrays. It was found that 9% of the genome shows a pan-S replicating pattern; this study also confirmed by FISH that the pan-S patterns were due to allelic variation in time of replication rather than variation of time of replication in

the cell population. As previously observed in human cells [73], mouse cells also show large replicons over the range of 400 kb occupying early to late transition regions. Once each region of the genome had a temporal zone assigned, different correlation analysis were performed.

David Gilbert and colleagues [76] found that replication domain changes are correlated with activation of gene transcription in both species and even more interesting, that cells from the same cell type but different organisms (hESC and mEpiSCs) showed more similarities in their replication timing profiles than to other cell types from the same organism.

Farkash-Amar et al. [66] compared the murine time of replication pattern with human lymphocytes replication timing profiles from [69] and showed that time of replication is conserved, raising the question whether time of replication is evolutionary constrained. The evolutionary conserved replication domains were confirmed later using more data sets [80, 76]. A very interesting discovery from [81] is that large-scale chromatin domains correspond to synteny blocks, suggesting that these physical-units correlate with recombination events.

In addition to evolutionary conservation of replication timing, it is remarkable that using a tissue specificity index [81] as a comparison parameter, Farkash-Amar and colleagues found that the more tissue specific a gene is, the later it replicates. Furthermore, genes were classified depending on their gene ontology (GO) category, surprisingly lymphocyte inducible genes that are inactive show early replicating patterns independently of their transcriptional activity. Table 1.2 summarizes GO classification of genes given their replication time zones.

It has also been proposed that epigenetic status is determined, and thus inherited from cell to cell, in strong relation with the S phase programme. DNA micro-injected into nuclei of rat cells during early and late S phase was associated with transcriptionally competent chromatin marks and inactive chromatin marks respectively [82]. This observation poses S phase not just as a duplication of genetic information but as a key step in inheritance of epigenetic directly linking S phase with development and cellular differentiation.

Gene type	Replicating Time	GO Category
Housekeeping	Early	Metabolism Transport Transcription Cell cycle
Not expressed but inducible in lymphocytes	Early	Stress response Apoptosis
Tissue specific expressed in lymphocytes	Early	Immune response Lymphocyte activation
Tissue specific not expressed in lymphocytes	Mid and Late	Mid brain development Sensory perception of smell Sensory perception of taste Keratinization
Not expressed but inducible in lymphocytes	Early	Stress response Apoptosis

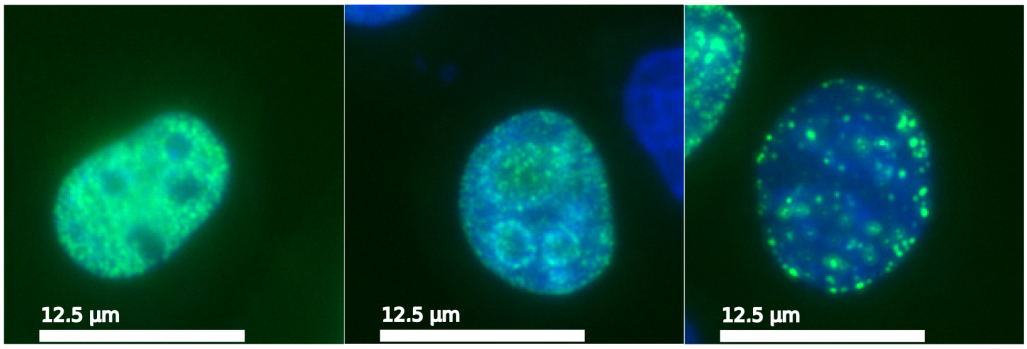
GO - Gene Ontology

Table 1.2: GO classification of replication time zones in mouse lymphocytes [66]

When placing the S phase programme in the context of multi-cellular organism somatic development, there is evidence supporting the hypothesis that there exists a developmental plasticity on the time of replication dependent on which lineage the cell is committed to [83, 84, 85]. Comparison of replication-timing profiles before and after differentiation between mouse embryonic stem cells has shown that $\sim$20% of the genome changes the replication-timing patterns upon differentiation [75]. Most of the replication timing rearrangement of replication domains involves the consolidation of chromosomal regions showing mixed timing profiles to the same replication timing. The rearrangement of replication domains in differentiated cells tends to match replication timing zones with isochores, suggesting that time of replication aligns better to sequence composition in differentiated cells and suggests that the mis-alignment of replication timing to sequence composition is a trait of pluripotency [75].

When pulse labeling DNA of proliferating cells and immuno-detection is performed with fluorescent-conjugated antibodies, different patterns can be observed depending on the stage of S phase where labeling occurred [86] (Fig. 1.3). Early S phase is characterized by a homogeneous distribution of replication foci with sizes in the range of 100-150 nm in diameter [29]. Mid S phase shows larger foci in the range of 150-250 nm and a less uniform distribution of foci, with foci clustering towards the nuclear periphery and perin-

ucleolar regions. Finally, late S phase patterns, which are characterized by the aggregation of replication foci into larger spots, generally localized in the nuclear periphery. These patterns show the equivalence between the S phase programme and the large-scale chromosome architecture inside the nuclear volume [87]. This equivalence has been reported using analytical tools based on derivatives of 3C experiments [80, 76]. The correlation of replication timing profiles and the resulting eigenvectors of PCA applied to the multi-dimensional chromatin interaction matrices is in the order of R=0.80 genome wide [76].



(a) Early S phase **(b)** Mid S phase **(c)** Late S phase

Figure 1.3: Replication timing and nuclear architecture. DNA pulse labelling during DNA replication at different times reveals the spatio-temporal progression of the replication machinery in different regions of the nucleus. **(a)** Early S phase is characterized by a homogeneous replication foci distribution with sizes in the range of 100-150 nm in diameter. **(b)** Mid S phase shows larger foci in the range of 150-250 nm. Foci distribution is less uniform clustering to the nuclear periphery and perinucleolar regions. **(c)** Finally late S phase pattern main feature is the agglutination of replication foci into larger spots. Panels show HeLa cells after pulse labelling with BrdU.

Chapter 2

METHODS

2.1 Innate Structure of DNA Foci Restricts the Mixing of DNA from Different Chromosome Territories

2.1.1 Cell Culture

HeLa cells were grown in DMEM (Dulbecco's Modified Eagle's Medium; Sigma-Aldrich) supplemented with penicillin and streptomycin, L-glutamine, and 5% fetal bovine serum (DMEM 5% FBS) at 37° C . Cells were cultured in 25cm^2 polystyrene angled neck cell culture flasks (Corning Inc.). For cell imaging, culture dishes containing coverslips were used and mounted into slides after labelling and treatment, details below.

2.1.2 Visualizing replication foci in human cells

Carrier mediated transfection of the thymidine triphosphate (dTTP) analogues, Cy3-dUTP or AlexaFluor488-dUTP (AF488-dUTP), was performed with FuGENE 6 Kit (Roche Applied Science) [88]. 12 μl PBS was mixed with 3 μl FuGENE 6 and kept on ice for 5 min. Subsequently 1 μl of AF488-dUTP was added and incubated on ice for 10 min. 8 μl drops of the transfection mix were pipetted on a piece of parafilm and then the coverslip, cells facing the drop, rested for 10 min over the parafilm strip on ice. The coverslips were then rinsed with cold PBS and placed in warm medium. After a chase period of 24 hours, cells divided and a second pulse of dTTP analogue has performed following the same procedure explained above; this time with Cy3-dUTP. Cells were left to grow in warm complete medium for 1 or 2 days for proper random segregation of chromosomes in daughter cells. After labeling, cells were rinsed twice with PBS at room temperature and then fixed for 10 minutes at room temperature with 4% paraformaldehyde, rinsed again three times with PBS and mounted into glass slides (Superfrost, Menzel-Glaser) using

Vectashield (Vector Laboratories) as mounting medium to maintain the fluorescence of the samples.

2.1.3 TSA treatment

After labelling of DNA foci by transfection of fluorescent DNA precursors and random segregation of labelled chromosomes after mitosis, cells were incubated for further 24 hours in the presence of the histone deacetylase inhibitor TSA (Sigma) at concentrations of 50 and 100 ng/ml. After incubation with the drug, cells were rinsed twice with PBS at room temperature in order to wash the medium. Cell were fixed for 10 minutes with 4% paraformaldehyde at room temperature, then processed for microscopy as mentioned above.

2.1.4 Confocal microscopy

Confocal imaging for the preliminary colocalisation analysis was performed using a Zeiss LSM510META confocal microscope. Sections were collected using a 100× (1.45 NA) lens. Channel settings were set as follows:

- Green Channel: 488 nm laser line at 2% intensity; BP 500-530 IR filter.
- Red Channel: 543 nm laser line at 32% intensity; LP 545 filter.

For the second, more detailed, colocalisation analysis confocal imaging was performed using a Zeiss LSM710 microscope using an 100× (1.46 NA) objective. Voxel dimensions were 0.8 × 0.8 × 0.34 microns and images of a XY resolution of 988 × 988 pixels; pinhole settings of 1.0 Airy unit. Amplifier and detector gain and offset were optimally chosen by the instrument for each field acquired. Channel settings were set as follows:

- Green Channel: EF1 filter used with a SPI wavelength range of 493-543 nm.
- Red Channel: EF2 filter used with a SPI wavelength range of 566-681 nm.

2.1.5 Image analysis and model building

Volumetric projections, such as the one showed in Fig. 3.9 were generated using Z stacks and processed for median filter in Imaris® (Bitplane) software.

2.1.5.1 High-throughput image analysis

Custom Jython scripts were written using the Fiji software suite [89] with the aid of the of 3-D filters suite of plugins [90]. Fiji is a free and open-source project based in ImageJ [91] for image processing. For colocalisation analysis we made use of the JACoP plugin which offers most of the different colocalisation methods [92].

Figure 2.1 shows a diagram of the image processing pipeline. Image stacks were duplicated in order to compare raw images and processed images. The duplicated stack was first split into the corresponding channels. Each channel was filtered individually using a $2 \times 2 \times 2$ 3D median filter [89, 90] to remove noise. Thresholding was then determined by calculating the standard deviation of the mean intensity values for each channel. Volumetric estimation of colocalisation was performed based on the results of this step (see next section). Voxels below the threshold value were set to zero. In parallel, Mander's coefficient was calculated using the same thresholding values. Colocalisation analysis was performed to the different sets of images and results processed with R programming language [93].

Estimation of volumetric colocalisation

Volumes were estimated by multiplying the total number of colocalised voxels by the volume covered by a single voxel ($0.8 \times 0.8 \times 0.34$ microns). The criteria to consider a voxel colocalised required that both channels indicated signal values above a threshold value. The threshold value was empirically determined to be equal to the standard deviation of the distribution of pixel intensities in the corresponding channel across the whole Z stack.

2.1.5.2 3D modelling

To visualize 3-D interactions between chromosome territories, as in Fig. 3.10, coordinates of each of the fluorescent tags were exported individually into Virtual Reality Modelling Language (VRML) format using Imaris® software. VRML files were exported to *3ds* format using an open-source, platform-free 3D-design suite (http://www.blender.org/). These files were imported into Autodesk® 3ds Max® (www.autodesk.com/3dsmax) and

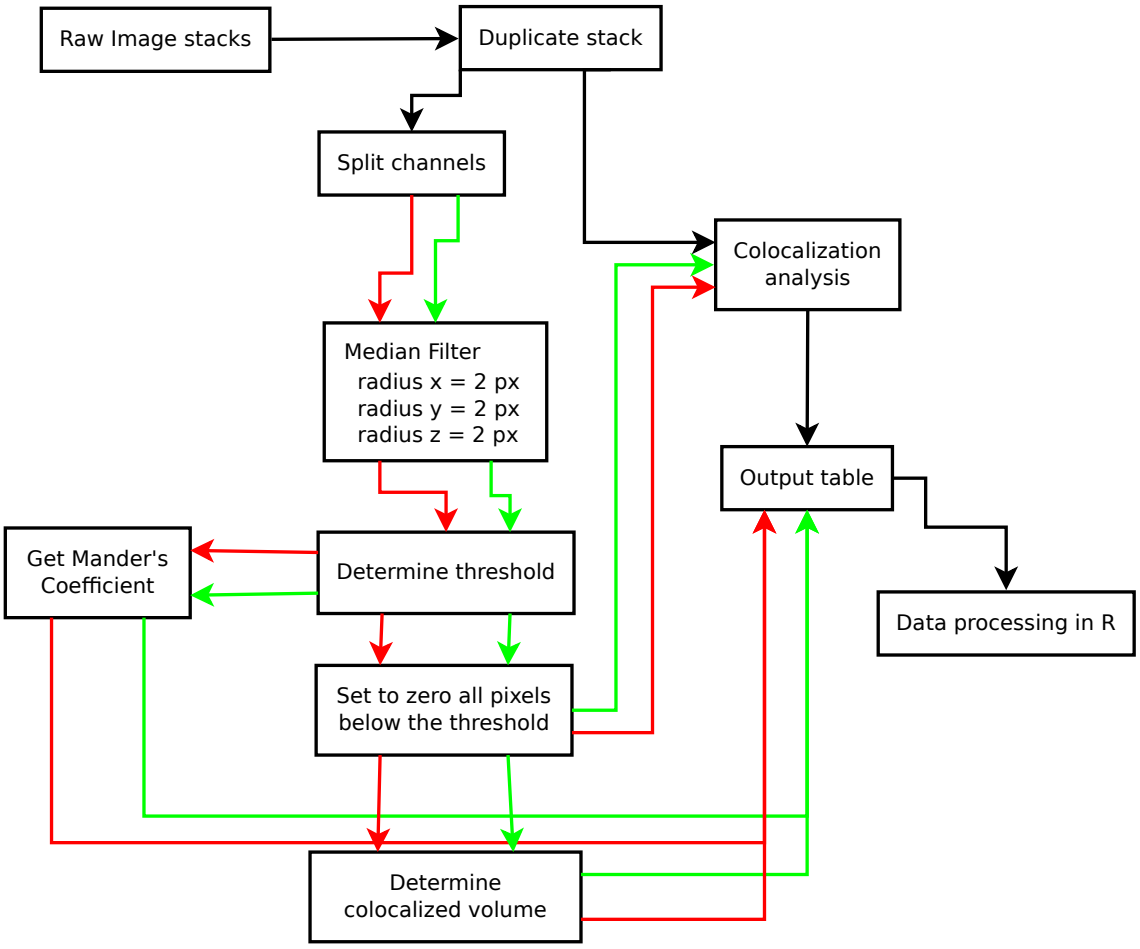


Figure 2.1: Image processing pipeline

imported files merged in a single MAX file to facilitate image rendering, 3-D modelling and animation. This procedure using 3ds Max's built-in compound modifiers models the 3-D shape of the chromatin compartment using the continuity of labeled DNA foci to define the chromatin space.

2.2 Post-genomic analysis of the banding pattern of human mitotic chromosomes

For the measurement of the different genomic features studied in Chapter 4, each data set was downloaded from its respective source and processed afterwards for statistical analysis. The method varied depending on the file format of each source but all of them went through the same processing pipeline, which required the data to be transformed into Browser Extensible Data (BED) format (*http://genome.ucsc.edu /FAQ/ FAQformat.html format1*). This format requires three basic fields which, specify for each block of data the chromosome where it belongs and its position on it with a start and end coordinate.

Once each data track was transformed into BED format, we made use of the utility *overlapSelect* from J. Kent's source tree [94], which filters BED blocks based on overlapping ranges of genomic feature blocks. By running *overlapSelect* with the BED file for a given feature and BED file with the coordinates of the chromosomal bands, we could assign to which band each individual block of the different features belonged relative to the banding pattern. Statistical analyses and all the plots were done in the R programming language [93].

Depending on the nature of the feature under evaluation, different data processing steps were performed. The next sections provide the sources for each data set used in this work in addition to appropriate technical details when required. Features that overlapped with band boundaries were excluded from the analysis. All analyses were performed with coordinates based on the human genome assembly Mar.2006 (NCBI36/hg18)

2.2.1 Properties of bands

The general properties of bands were obtained from the *cytoband* track from the UCSC genome browser [95, 96]. This genomic track is based on the prediction of borders through the implementation of a Hidden Markov Model which called each band in the cytogenetic map and assigned its corresponding coordinates at the level of nucleotide sequence in the first draft of the human genome [97]. This table contains the start and end coordinates for each band, the chromosome to which it belongs, the Giemsa staining intensity and the name of the band based on its position on each arm of the chromosome.

All the different genome tracks in this work were matched to this set of coordinates using *overlapSelect*.

2.2.2 Sequence Features of bands

Differences in the structure of genes and their distribution among bands

The file table containing the information for genes (*ensGene*) file was downloaded from UCSC Table browser corresponding to the assembly Mar.2006 (NCBI36/hg18) on the 9th of Feb 2011. Genes mapping to chrN_random and haplotypes were eliminated.

For the gene-based analysis, the table was filtered through a custom Python script, which only selected the protein coding genes. For each of the genes selected, only the longest transcript registered for that gene was kept and the genes that overlapped with the coordinates of a band where excluded from the analysis. Additionally, nested genes were ignored.

CpG Islands

The CpG island coordinates table (*cpgIslandExt*) was obtained from the Table Browser from UCSC genome browser. This table is based on [98].

GC content of bands

To calculate the GC content for each band we used the utility *hgGcPercent* from J. Kent's source tree [94] which is able to calculate the GC content for the coordinates provided in a BED file. We used the coordinates obtained from the *cytoband* file mentioned above and measured the exact GC content of each individual chromosomal band.

Analysis of GC content as isochores

The table of coordinates used to determine the position of each isochore family was obtained from the consensus table from Isobase [99]. This table is able to capture the advantages of each method used for the detection of isochores and presents a unified source for the study of the distribution of isochore families in the human genome.

Repetitive elements

The repetitive elements table (RepeatMasker) was obtained from the Table Browser of the UCSC Genome Browser. It is synchronized with the most current versions of the RepeatMasker software and repeat libraries (RepBase) [100, 101].

2.2.3 Histone Modifications

The chromatin data was derived from ChIP-Seq experiments. Data was downloaded from *http:// dir.nhlbi.nih.gov /papers/lmi/ epigenomes/ hgtcell.aspx* in the vstep format, transformed to BED format with an in-house developed python script and then overlapped to the cyto-bands file. All coordinates were from the Human Mar. 2006 (hg18) assembly.

2.2.4 Replication timing features

Replication timing profiles were downloaded from [102] and [103], then followed the standard overlapping procedure.

2.2.5 Higher-order chromatin organization of Giemsa bands

Differential compaction of bands in mitotic chromosomes

In order to measure the degree of compaction that chromosomes undergo during the mitotic condensation of chromatin we compared the proportional lengths of chromosomal bands from the ideograms in [104] to the linear length of the bands in the *cytobands* set of coordinates from [97]. We defined the index of compaction based on the log ratio of ideogram length/genomic length.

DNaseI hypersensitivity sites

The data files for the DNaseI Hypersensitivity by DNase-seq experiments from the ENCODE project and the University of Washington [105, 106] were downloaded from *http:// hgdownload-test.cse.ucsc.edu /goldenPath/ hg18/ encodeDCC/ wgEncodeUwDnaseSeq/* in the WIG file format. Files were converted to the BED format and overlapped with the chromosomal band coordinates.

Chromatin interaction maps

Hi-C data was downloaded from *http:// www.ncbi.nlm.nih.gov /geo/query /acc.cgi?acc = GSE18199* based on the chromatin interaction maps from [32]. The eigenvectors (details on Section 4.6.2.1) provided by the authors, were used to measure the preference towards compartment A and B and averaged from each individual. Values from the eigenvectors were first transformed to a BED format file and overlapped to the band coordinates table.

Nucleolus-associated domains (NADs) and Lamin-associated domains (LADs)

Data was obtained from [107] and [108] as BED files. Given the large block size of the tables, the data was processed so data blocks were split into smaller 1000 bp blocks prior to the overlap step. Afterwards, files were processed in the standard way used for other genomic features.

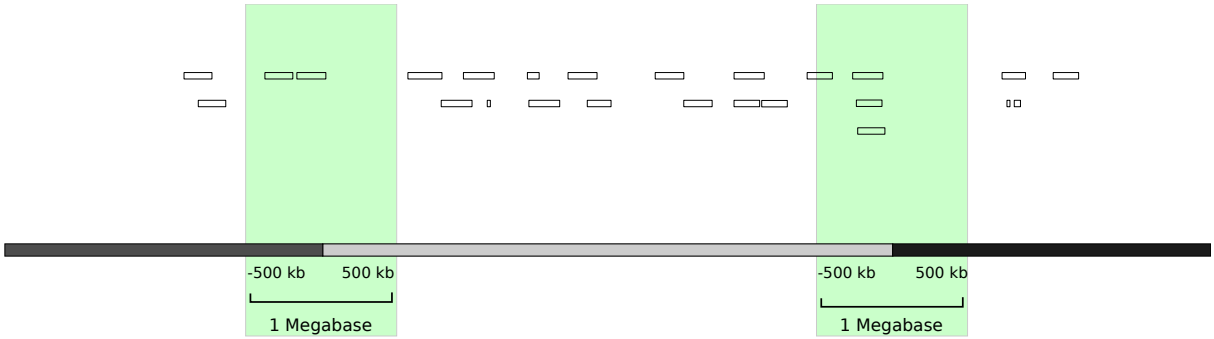


Figure 2.2: Description of scoring strategy based on BAC data. Density of BACs inside of a 1 Mb window around both borders of each band is measured.

2.3 Inaccuracies on the cytogenomic map

2.3.1 Border scoring based on density of border surrounding regions

In Section 5.1.2 we estimate a score for the assignment of each chromosomal band from [97] based on the BAC data they used for bridging the human genome sequence with the cytogenetical map [109].

As the strategy we used is able to 'see' beyond the range separating the most proximal BAC of a border to the border itself, we can detect regions of low BAC-density surrounding the border. Shown in Figure 2.2, the band borders have small distances to the most proximal BAC, however, by measuring the density around the border, we can detect the overall BAC support landscape and score the borders in a more reliable and robust manner.

To estimate the reliability of each band, we summed the total number of BAC probes surrounding the up- and down-stream borders of the band.

2.3.2 Unsupervised Machine Learning for the identification of G-like R bands

There are many different methods for the automatic classification of datasets. One of the most common is the K-means clustering algorithm. Once defining a K, this algorithm is able to find K domains within the dataset. This algorithm works in 4 steps in general.

The first step consists on the random generation of K different "means". These means represent the center of each of the K groups and will be adjusted afterwards in order to

be aligned to the place in the data space that best represents that group in particular. After the assignment of K means, the distance between each data point and each mean is measured in all the possible dimensions (genomic features). Each data point is then associated with its closest mean and is grouped with the cluster represented by that particular mean. After this preliminary clustering step, the algorithm finds a centroid in the space of values defined by the data points associated with each mean. The centroid of each cluster becomes the new mean for the next iteration of the algorithm. The next iteration of the algorithm goes back to the measurement of distances between data points and means and re-associates each data point to its closest mean. This process is repeated until there is no change of a data point of being associated from one of the K means to any other.

The implementation we used for this algorithm is the Sparse K-means clustering algorithm [110], available as an R library package [93]. This implementation is able to automatically select which dimensions in the data are the ones that define the sub-structures within it most effectively. When it performs the clustering steps, it pays more attention to the key features and ignores the redundant ones.

We applied this algorithm in a range of K values from 2 to 9 in order to re-classify the R bands (gneg band category) depending on the whole collection of genomic features measured for each band type.

2.3.3 Segmentation based on active vs inactive blocks

The table of chromatin state blocks reported on [111] was downloaded from *http://compbio.mit.edu/ ChromatinStates/ map_allstates.bed* . States from the 'promoter', 'transcribed' and 'active intergenic' categories were labelled as *'active'* and aggregated as a single state. The states from the 'repetitive' and 'suppressed' categories were all labelled as *'inactive'* and aggregated into a single state. State regions were split into non-overlapping 200 bp segments. We defined the active/inactive (AI) ratio of a region by dividing the total number of active 200bp-blocks between the total number of 200bp-blocks of the inactive state.

By binning the genome into N bps long windows and measuring the AI ratio per bin, we were able to reconstruct the Giemsa pattern *in silico* at different scales. For each value of N explored, bins of genome was first split into N-long windows and the AI ratio was

computed. The AI ratio was plotted using a color code that ranged from a light gray that represented the pure active (values close or equal to 1) from and a dark grey, the pure inactive state (values close or equal to 0).

We explored 3 main scale-ranges of N. The first range covered window sizes from 10 kb to 40 kb by increments of 6.6 kb. The second range covered windows sizes from 40 kb to 1 Mb by increments of 25 kb. Finally, the third range covered the larger scales, from window sizes of 1 Mb to 1.75 Mb by window size increments of 150 kb. To visualize these results we translated the AI ratio into a gray scale and plotted it as a heatmap, where dark gray represented values close to 0 and light gray to 1.

2.3.3.1 Inflection profiles

To create inflection profiles based on the aggregation of chromatin states in [111] we transformed the categorical value 'active' or 'inactive' to 1 and -1, respectively. This binary representation of the data allowed us to process it as numerical signal of numeric values. We applied the cumulative sum function to the vector of transformed numerical values and plotted it as an inflection profile.

2.3.3.2 Identification of relevant inflection points

To identify inflection points that represented transitions of large scale domains and not only small deviations, we performed a multi-scale analysis based in the annotation of local minima and maxima in sliding windows of different sizes. For each data point in the sequence we annotated the local maximum and minimum covered by the window. We repeated this step on each of the data points. Each time a minimum or maximum was determined, an entry in a table was created for that coordinate and the table registered a score of how many times each entry was called. The score of each entry incremented by one each time the entry was repeated.

We were able to pin point the relevant inflection points by ranking them based on their occurrence score. For each chromosome, only the first n places in rank were used. We calculated n as two times the number of band borders seen on chromosome.

Chapter 3

INNATE STRUCTURE OF DNA FOCI RESTRICTS THE MIXING OF DNA FROM DIFFERENT CHROMOSOME TERRITORIES

3.1 Introduction

During mitosis, mammalian cell chromosomes appear as discrete and well-defined bodies that allow the proper segregation of genetic information to the two daughter cells. However, the compact conformation of chromosomes that grants their easy visualization and resolution is not present during the rest of the cell cycle. During interphase, chromosomes decondense and diffuse inside the nuclear envelope appearing indistinguishable from each other. Chromosomes adopt a structural configuration permitting proper gene regulation, DNA repair and DNA replication. It is now well established that in the interphase nucleus, rather than being fully dispersed, chromosomes present a nonrandom spatial distribution throughout the nuclear volume where each chromosome is constrained to its own spatial domain known as a chromosome territory (CT) [31]. The shape of CTs is highly variable and does not show any regular pattern. Despite the general agreement about the common properties that characterize CTs, there is no consensus view of the functional and structural relationship between different CTs nor about the functional interactions of CTs with the rest of the nucleoplasm. The difference in opinion is due to the different lines of evidence that support each model. As all of them rely on different preparations and visualization methods, each model is constrained to what its own evidence suggests. This situation has resulted in the development of several models of CTs and chromatin organization, the main ones being: the giant-loop model, lattice model, chromosome territory-interchromatin compartment model (CT-IC) and the interchromo-

somal network model (ICN).

3.1.1 Giant-loop model

Based on the systematic comparison of the interphase distance between pairs of FISH probes against their respective genomic linear separation, Yokota and colleagues proposed a simple model of interphase chromatin structure named the giant-loop model [112, 113]. The experiment consisted in the hybridization of two selected fluorescent probes and the repeated measurement of the 2-dimensional distance between them observed under the microscope. The mean-square of the distance between a given pair of probes was then compared to their linear genomic distance. A second pair of probes was designed by increasing the linear genomic distance of the previous set of probes followed by a systematic repetition of the hybridization, measurement and comparison steps. This process was iterated throughout a range of 0.15 to 190 megabases separating the two probes. The 2-dimensional distance as a function of genomic distance showed two different linear phases with a transition around the 2 Mbp point. This suggests that there are two levels of organization of the chromatin at scales higher than 100 kbp. The giant-loop model proposes that each of the two levels of organization fold in a *random walk* fashion. The first level consists of large chromatin loops in the range of 1 to 3 Mbp in size while the second level of organization serves as a scaffold of loop-attachment regions that connects the loops formed at the first level of organization (Fig. 3.1). This model is compatible with experimental evidence showing that distal active genes migrate to previously assembled transcription factories [114]. Furthermore, co-regulated genes can be preferentially recruited to the same transcription factory regardless of the chromosome in which they are located [115, 116, 117, 118, 119]. Recent studies derivated from 3C techniques have shown that active regions of the genome, located at different chromosomes can interact [120] more frequently as they are found closer to the periphery of their chromosome territory.

3.1.2 Lattice model

This model proposed by Dehghani and colleagues [122] was developed from electron spectroscopic imaging (ESI) of the eukaryotic nucleus. ESI is also known as energy

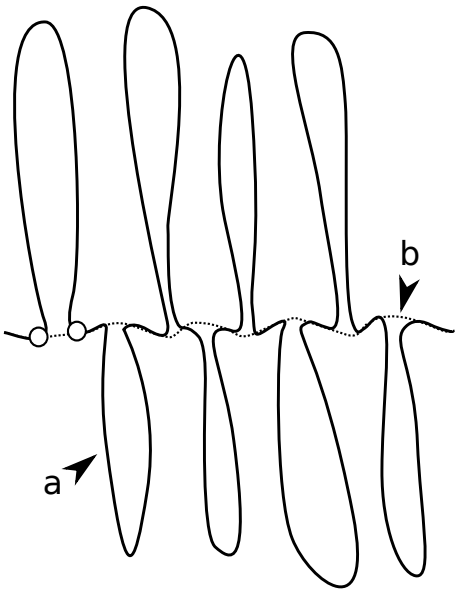


Figure 3.1: Giant-loop model. Two levels of loop organization of the chromatin are proposed by this model. The first level consists of flexible, loose loops covering several Mbp (a), responsible for bringing distant loci into proximity at regions denominated "loop attachment points" depicted as white circles in the diagram. Loops follow a random path along the nuclear space and form a secondary level of organization when consecutive loop attachment points lie together, forming a continuous flexible path termed the "backbone"; dotted line in the diagram. The scaffold provided by the backbone also follows a random walk through the nuclear space. (b) Putative protein complexes would mediate and stabilize loop formations. Insulator proteins such as CTCF are good candidates for carrying out this role [121]. (Diagram modified from [113])

filtered transmission electron microscopy. This imaging technique relies on the image projected by the loss of energy that the electrons suffer when interacting with the sample. The detection devices are able to filter the electrons by their level of energy thereby providing the ability to filter the signal in ranges of energy specific for different atoms in the sample. In other words, the composition of the sample will interfere with electrons in such a way that the relative enrichment of different atoms in the sample will "cast a shadow" unique to the particular element examined (Fig. 3.2 A). The classical example is the comparison of the energy-loss pattern of electrons with the sample due to phosphorous atoms versus nitrogen atoms, representing nucleic acids and proteins respectively. Relative to the power of fluorescent microscopy, ESI is also able to delineate the borders of condensed and decondensed chromatin regions in the nucleus based in the structures of individual chromatin fibers.

The lattice model proposes a porous, lattice-like network of chromatin fibers with a diameter of 10 and 30 nm. Chromosomes would organize as a nucleoprotein array that fills almost contiguously the nucleoplasmic space. The mesh-like properties of the chromatin

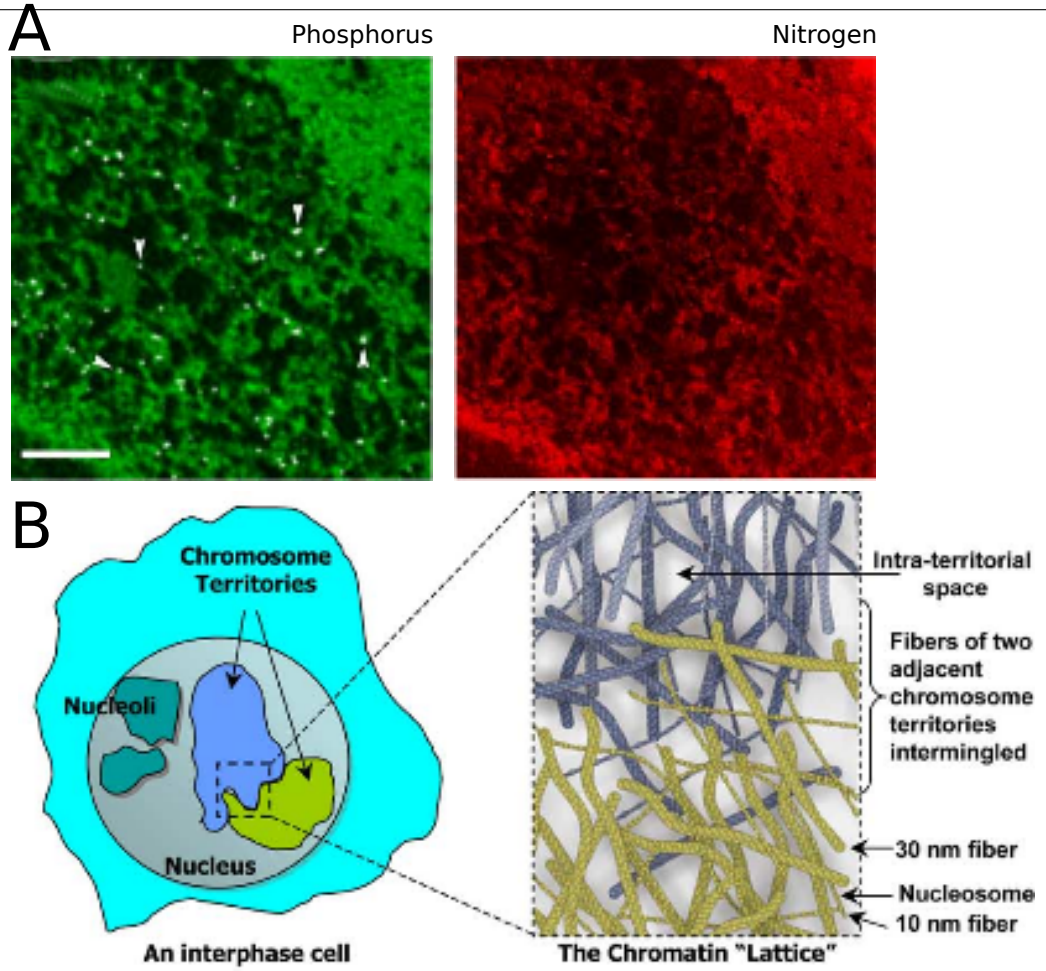


Figure 3.2: Lattice model of chromatin organization. (A) ESI derived phosphorus map in green represents the location of DNA inside the nucleus. It also shows that gold-tagged RNA-polymerase is uniquely detected in euchromatic regions. The red panel represents proteins by ESI tuned for detecting electrons that had interacted with nitrogen (pictures taken from [122]). (B) Schematic comparison of the appearence of chromatin under the light microscope showing a dense chromatin body in contrast with ESI that presents a lattice-like network of 10 and 30 nm chromatin fibers (modified from [122]).

fiber network would allow the free diffusion of nonchromatin nucleoplasmic elements and machinery. This model supports the idea of intermingling of neighbouring CTs, partially rejecting the notion of the interchromatin compartment (see section 3.1.3) based on the observations under the ESI that large spaces or channels between CTs are not apparent. Dehghani and colleagues [122] also take into account complementary lines of evidence to validate their model. The list of evidence includes (1) fluorescence recovery after photobleaching (FRAP) experiments demonstrating that relatively large protein complexes are able to "roam" freely within the nuclear space, including the interior of CTs [123], (2) that transcription can occur in the core of CTs and it is not confined to the periphery of CTs [124, 125] and (3) the observation that chromatin exists mostly as 10 and 30 nm fibers

as revealed by conventional transmission electron microscopy [126]. However, recent work has challenged the model of basic organization of the chromatin into 30 nm fibers based on the development of new molecular techniques such as genome-wide interacting maps and cryo-electron microscopy which is able to image sample slices as thin as 70 nm [23, 25].

3.1.3 The chromosome territory-interchromatin compartment model

Nearly two decades ago, Peter Lichter and colleagues [127] introduced the model of the functional compartmentalization of the cell nucleus into two main components: chromosome territories and the space between them named the interchromosome domain (ICD). The ICD was supposed to provide the cell with a space for the interaction of DNA sequences and the rest of the nuclear machinery in order to accomplish their biological function. Their model was based on the spatial exclusion of the splicing and transcriptional machinery from the space occupied by chromosomes [127]. A natural assumption of this model was that biologically functional elements, such as genes or regulatory sequences, would be exposed to the ICD, therefore they should be found at the periphery of CTs. Despite several studies supporting the idea of active elements being recruited to the surface of the CT [128, 129, 130, 131, 132], there is evidence that this is not necessarily true and genes can occupy any region of the CTs independently of their transcriptional status [133, 124, 125, 134, 135].

The CT-IC model is based on the appearance of CTs by the technique of 'chromosome painting' under the fluorescence microscope. Individual painted chromosomes appear as dense nuclear bodies without any internal cavities or room for the occupancy of other nucleoplasmic elements. However, parallel lines of evidence pointed that chromatin was organized into smaller chromatin domains as a fundamental unit of chromatin organization [136, 137, 138] that could be visualized as replication foci as stable structures across different cell generations after cell division [29, 28]. As a consequence, the most recent version of the CT-IC model updated the concept of the ICD from interchromosome domain to interchromatin domain (IC) portraying CTs as the largest unit of chromatin organization composed of smaller higher-order units of chromatin domains (Fig. 3.3 top panel), most of the time arranged into a configuration 10 times above the compaction level of the 30 nm

fiber [137, 138]. Chromatin would now be represented as a dynamic, 3-dimensional (3D) network [137] of channels and lacunas, or cavities, originating from the nuclear pores resembling the texture of a sponge. Nonchromatin machinery and factors could freely travel through the network inside the volume of the CT, opening the possibility for interaction of biological functional units hosted in the interior of their respective CT. Transcriptionally inactive regions of the chromosome would be characterized by a narrowing of the IC channels without reaching a point of total collapse [138, 139]. The IC has been observed by electron microscopy [124] and there is strong evidence to support its existence as demonstrated by the reversible manipulation of the compaction status of chromatin by increasing the concentration of divalent cations (osmolarity) of the growing medium [138]. By changing the osmolarity of the medium, chromatin can shrink making the IC noticeable [138]. The CT-IT model is complemented by the perichromatin region (PR) model. This model proposes that a ribonucleoprotein domain of a width of around 100-200 nm surrounds chromatin domains and interconnects the different sub-compartments of the 3D chromatin network [137, 140]. Different nuclear processes appear to be happening in the PR region as a variety of nuclear sub-compartments or macromolecular complexes such as speckles, promyelocytic leukemia(PML)-bodies, replisome assemblies, and RNA pol II foci show specific topological distribution and relationships towards the chromatin network depending on the function that they are performing [138].

A key feature of the CT-IT model states that neighbouring CTs can touch, nevertheless the level of chromatin intermingling between CTs is insignificant or non-existent [137, 138].

3.1.4 The interchromosomal network (ICN)

This recent model suggested by Branco et al. [141] is based in the visualization and analysis of ultrathin cryosections of 150 nm. These authors argue that this sample preparation is able to preserve the sample without disrupting the chromatin nanostructure. In this model they challenge two of the main foundations of the CT-IC model, arguing against the low-levels of chromatin fiber intermingling between neighbouring CTs and the existence of the interchromosomal domain. They introduced a new method named cryo-FISH and measured colocalisation of neighbouring CTs using light microscopy, then confirmed

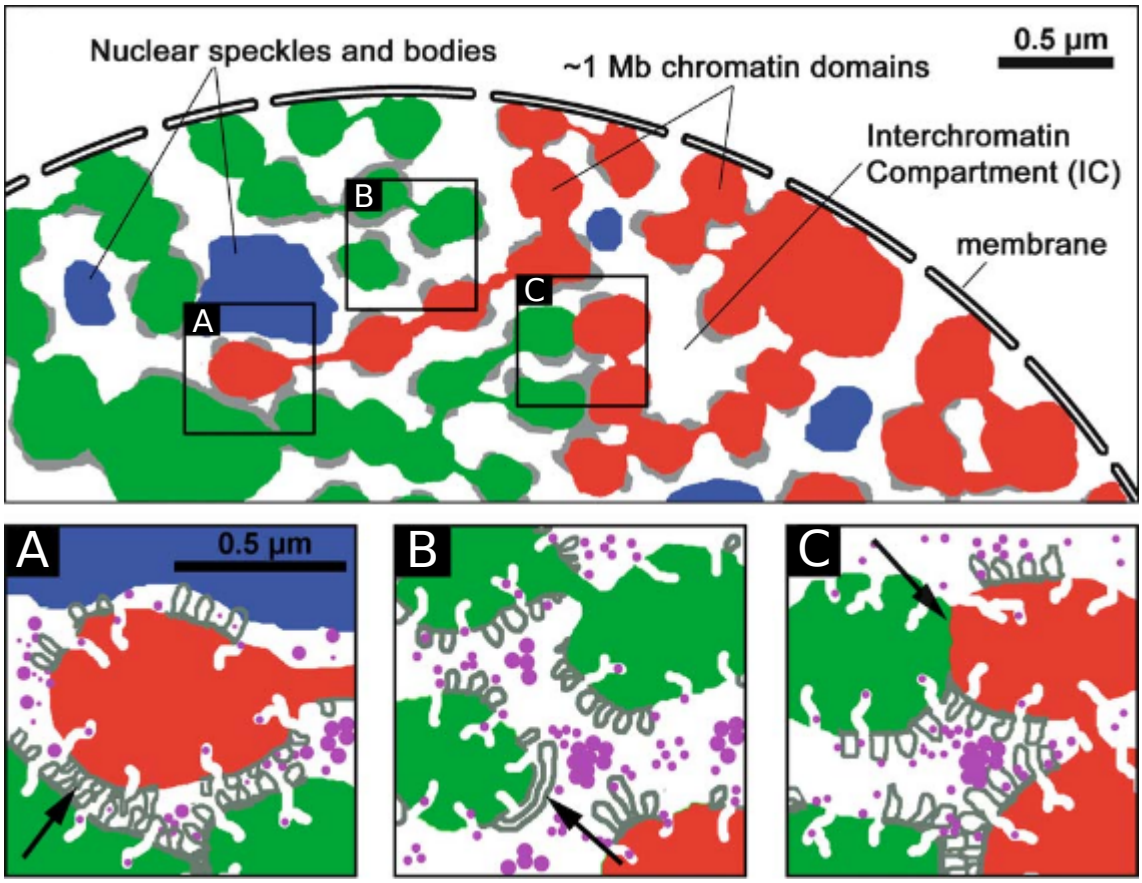


Figure 3.3: Chromosome Territories-Interchromatin Compartment model. In this diagram CTs are shown in red and green and are separated by the IC in white (top panel). Nuclear speckles and bodies as blue. Perichromatic region is shown in grey . (A) The IC can infiltrate in between CTs and also inside chromatin domains. Very narrow tunnels can access the most internal sectors the different domains, allowing access to the transcription and splicing machinery (purple). CTs can invade the space of adjacent territories but only as higher-order stuctures. PR is thought to guide this processes (black arrow). (B) In wider IC channels, transcription factories can recruit functional elements from different CTs and big chromatin loops, emerging from the PR can appear (black arrow). (C) Small areas of contact between adjacent CTs can occur without significant levels of intermingling of chromatin fibers (adapted from [138]).

their observations by electron microscopy using immunogold particles. They found that 41% of the FISH signal from chromosome 3 CT contained FISH signal from at least other CT in human lymphocytes. Then they supported their observations by stereological analyses of intermingling volumes of 24 pairs. Their observations are in agreement with chromosome translocation data in the same cell type [142]. Furthermore, Branco and colleagues [141] showed that by restricting transcriptional activity they were able to re-shape the interactions between chromosomes and concluded that nuclear processes such as transcription can drive changes in chromosome organization.

The ICN model is in direct conflict with some of the postulates of other models [143]. In particular, this model challenges the existence of the IC by affirming that there is exten-

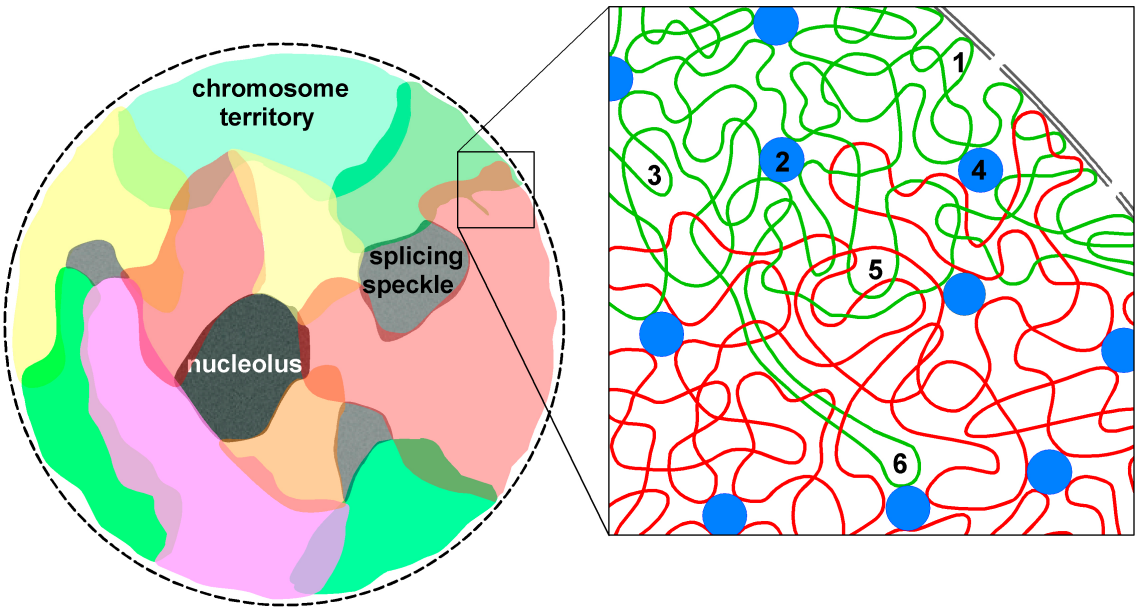


Figure 3.4: The interchromosomal network model. This model rejects the existence of the IC and allows chromosome to expand in their vicinity, rendering a continuous chromatin network. (1) The spatial constraints imposed by different structures (nucleoli or the nuclear envelope) set the limits for the levels of intermingling between chromosomes (5). (2 & 4) Active transcription factories can be located inside the CT as well as in the borders of adjacent CTs where they are able to recruit loci from different chromosomes. Some interchromosome contacts are also constrained by the CT and are kept in the interior (3). Some rare large loops can extend deep inside the territory of another chromosome (6). Image modified from [141].

sive intermingling of chromatin fibers between neighbouring CTs yielding a continuous body of chromatin. This opens the question of whether CTs are really self-contained structures inside the nuclear space and begs for a clarification of the real degree of intermingling between neighbouring CTs and its functional contribution to interchromosomal interactions and gene regulation.

To attempt to answer this question, we have quantified the degree of spatial interaction of neighbouring CTs in HeLa cells. By labelling chromosomes with fluorescent DNA precursors at the moment of DNA synthesis [88], after fixation, we were able to perform optical serial sectioning of cell nuclei without perturbing the native chromatin structure. Subsequently we measured the degree of 3-dimensional colocalisation of the spatial signal of neighbouring CTs. Our studies confirm the postulate of the CT-IC model about interchromosomal contacts, showing that neighbouring CTs do not intermingle and physically interact only at very low levels.

3.2 Results

3.2.1 Labelling strategy for identification of individual chromosome territories

The labelling strategy that we implemented takes advantage of two natural properties of the cell. First, when the cell goes through mitosis and divides, the chromosomes are randomly segregated to the two daughter cells. Second, DNA replication is semi-conservative, this means that newly synthesized DNA carries the template strand from the original DNA molecule together with the complementary newly synthesized strand [144, 46, 88].

By alternating rounds of DNA-labelling and cell-division cycles, individual CTs with unique labels can be obtained within the same cell Fig. 3.5. First, we pulse-labeled with AF488-dUTP (green) and let the cells grow for one cycle (top panel in Fig. 3.5). At this stage, daughter cells showed all their chromosomes with the green fluorescent label on them. We pulse-labeled one more time but this time using Cy3-dUTP (red) as the marker (mid-panel in Fig. 3.5). Following the second labelling step, daughter cells showed a mixture of green/red and red chromosomes which can be resolved into uniquely labeled chromosomes by letting cells grow and divide for 1-2 more days after the second pulse (bottom panel in Fig. 3.5). At the level of the cell population, we observed different labelling scenarios as seen on Fig. 3.5 bottom panel, where some cells will show only the red staining, and some others with both green and red label. At the level of individual cells, in the cases of cells with both colours incorporated, the same proportion of chromosomes with each colour were observed. These particular cases where hand selected for microscopy.

The use of antibodies or FISH as detection methods, typically require a step of DNA denaturation that allows exposure of the epitope or complementary sequence for proper association with its respective fluorochrome-associated element. Alternative methods are desired as the denaturing step locally disrupts the native structure of the chromatin. Our labelling method circumvents this problem given that the DNA precursors used for labelling are already conjugated with fluorochromes therefore there is no need for denatu-

ration of DNA.

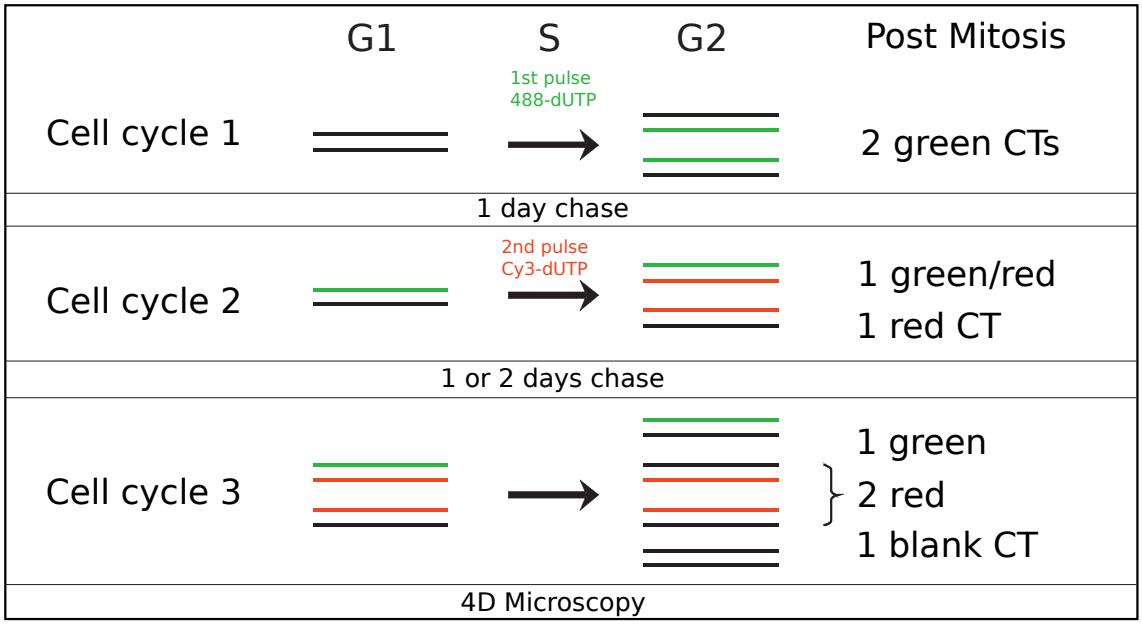


Figure 3.5: Labelling of individual chromosome territories. In order to resolve DNA foci into uniquely labeled domains, we labelled HeLa cells using AF488-dUTP in the first cell cycle (top panel) and Cy3-dUTP in the second one (mid panel). Cells were grown for 2 or 3 more generations in order to split the signal of the two labels used.

3.2.1.1 Image processing and filter selection

After labelling of CTs, cells were fixed and prepared for laser confocal microscopy. As we are working with 3-dimensional data, image processing has to be adjusted to a third dimension where pixels are extruded to fit the extra dimension. A 3D pixel is called a voxel; Fig. 3.6. As we are observing continuous objects in space, each voxel is somehow influenced by the surrounding intensity-landscape of the neighbouring voxels. For this reason, image analysis tools, such as image filter algorithms, also have to be adjusted to a third dimension in order to take into account the relationship between voxels in space.

After image acquisition, image processing is critical for the proper biological interpretation of the data. In this work in particular the importance of this point is accentuated as we are trying to properly measure the degree to which two objects may share the same space. Inappropriate manipulation of images may lead to incorrect interpretations of the data. This makes of critical importance the setting of threshold values that truly define real signal from background signal and noise.

To better understand how image filtering processes work, we need to introduce the

concept of the filter kernel. Image filters process images in a voxel-by-voxel fashion and modify the intensity value of each voxel taking into account the values of the neighbouring voxels. The kernel of a filter defines the region around the voxel under evaluation. The kernel can be seen as the spectrum and range in which the filter will be applied in order to define the value of a given voxel. It establishes the sampling parameters to define the set of voxels in which the mathematical functions will be applied in order to improve the signal. Depending on the desired function of the filter, a filter kernel can vary in shape and length. For this analysis we used an isotropic kernel (symmetric along all axes) with a radius of 2 voxels.

Proper image filtering requires an understanding of the sources and the nature of the different kinds of noise inherent in the system used. Confocal microscopy incorporates electronic noise at the moment of image acquisition. Electronic noise is characterized by occasional high-intensity pixels surrounded by pixels showing very low levels or no signal at all. This kind of noise is commonly known in the field of image analysis as 'salt and pepper noise' (Fig. 3.7 top left and 3.8 top row).

Using a filter that smoothes or spreads the intensity values of voxels across the image would be a mistake in this kind of study as the high-intensity levels added by the noise would degrade the signal and compromise the quality of the data along the edges of the objects imaged. The typical filter that performs this kind of signal diffusion is the Gaussian or mean filter (Fig. 3.7 mid-column and 3.8 mid-row). Depending on the parameters set for the kernel, the filter takes the intensity values of the voxels covered by the kernel and assigns the mean value of the set to the voxel in turn. One disadvantage of this filtering technique is that diffusion of the noise will add higher levels of overall background signal as shown in Fig. 3.7.

To avoid the addition of unreal voxel intensity values into the dataset we used an alternative filter that instead of averaging the intensity values for the voxels included in the kernel, returns the median value. Voxels with high intensity values surrounded by low intensity voxels will not spread its signal, as it is the case with a mean filter. The median value of a distribution of number is less sensitive to outliers (in this case represented by high intensity voxels), thus 'salt and pepper' noise can be efficiently removed from the

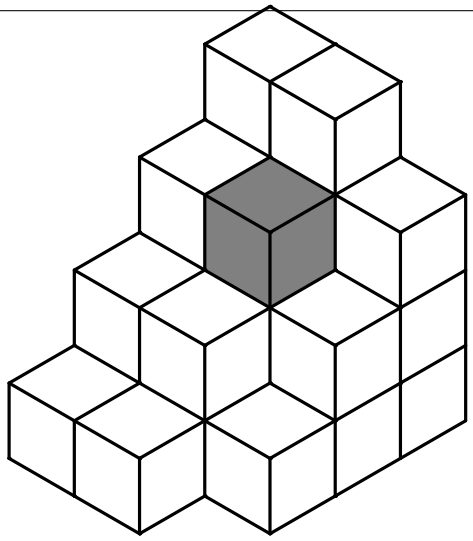


Figure 3.6: Voxels are 3D pixels. From (http://upload.wikimedia.org/wikipedia/commons/b/bc/Voxels.svg)

dataset with just a minor erosion of the real signal (Fig. 3.8 bottom row).

Figure 3.7 shows the green channel with an alternative (HiLo; Hi and Low indicator) lookup table (LUT) that instead of being a linear gradient from black to green, represents pixels with zero intensity values as blue, saturated pixels as red and a gradient in the gray-scale for the values in between. This display of the images is very effective for accurate visualization of extreme values in the pixel-intensity distribution of an image. From this example, the advantages of using a median filter compared to a mean filter are very clear.

3.2.2 Chromosome territories are discrete structures

Light optical serial sectioning of nuclei and image filtering was performed and volumetric projection of Z stacks are shown in Fig. 3.9 A. Typical patterns of DNA replication-based labelling were observed as discrete chromatin domains. The structure of individual DNA foci could be resolved which formed groups that gave shape to discrete CTs as shown in Figure 3.9 B.

Objects labeled in both, green and red, channels showed well demarcated boundaries and regions of apparent colocalisation (seen in yellow on Fig. 3.9 C) were only present in the borders of the CTs. The apparent colocalisation between neighbouring domains was only seen with volumetric Z-stack projections. Further examination of the phenomenon showed that at the level of individual Z-sections the colocalisation signal was almost unnoticeable. Figure 3.9 panel D shows an individual slice from the Z-stack showing that

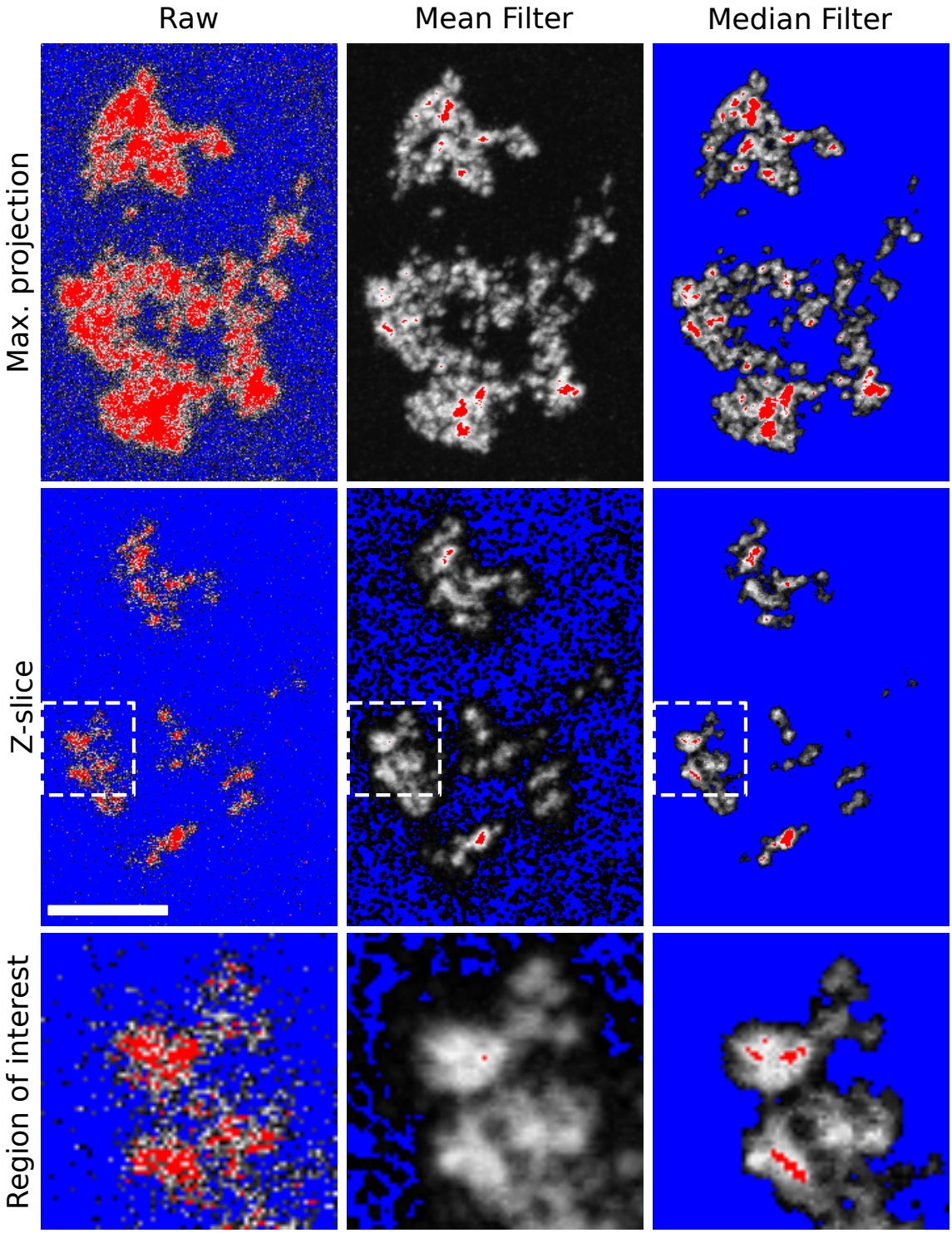


Figure 3.7: Median filtering is more effective than mean filtering to remove noise and background signal. Scale bar represents 5 micrometers. The HiLo LUT represents zero-valued pixels as blue, saturated pixels as red and values in between both extremes, as a gray scale.

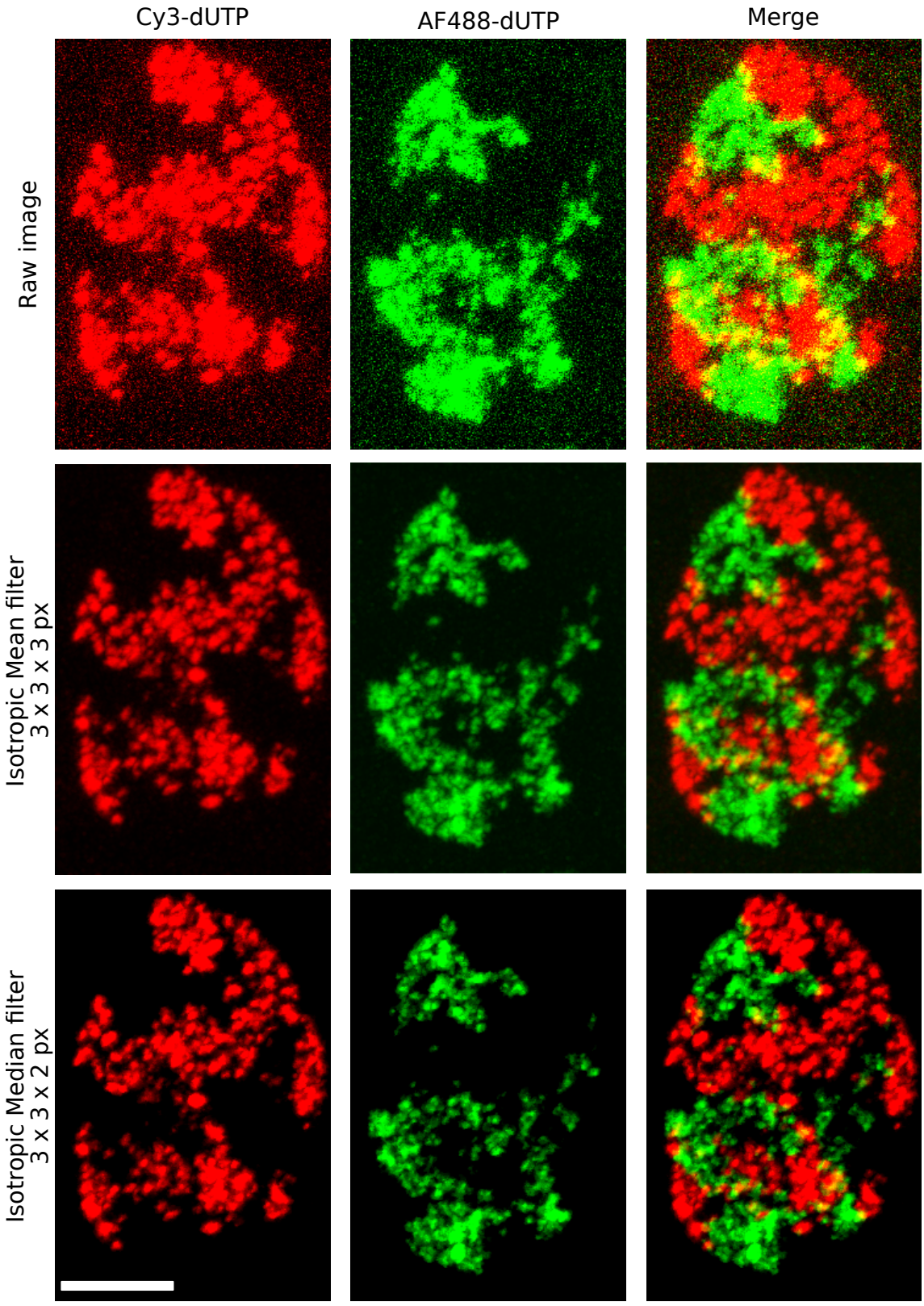


Figure 3.8: Comparison of filtering methods using both channels. Using median filtering (bottom panel), signal from chromatin domains and CTs is conserved better at the edges and structures appear sharper relative to mean filtered images (mid panel). Regions of apparent colocalisation can be seen as yellow signal. Maximum Z-projections of image stacks. Scale bar represents 5 micrometers.

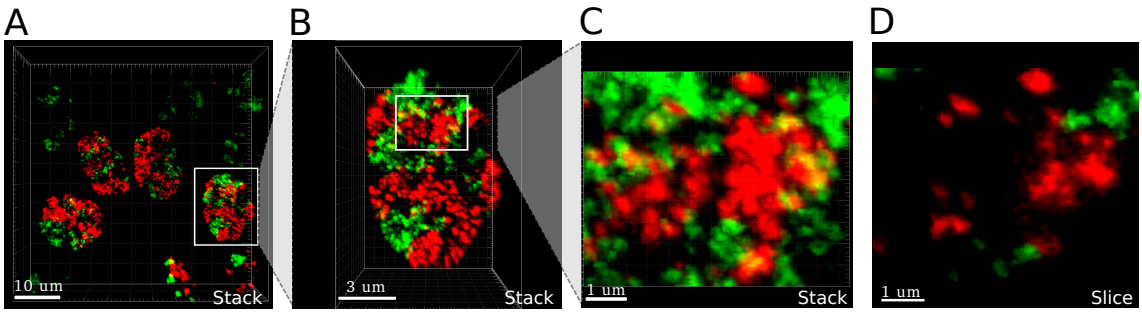


Figure 3.9: Visualization of chromatin domains and chromosome territories. (A) Cells showing homogeneous labelling in both channels were manually selected. Regions of apparent colocalisation were observed when images were transformed into volumetric objects the Imaris® suite (B and C, show the whole nucleus and cropped regions respectively). (D) After closer examination of individual Z-slices, regions of colocalisation were nos observed.

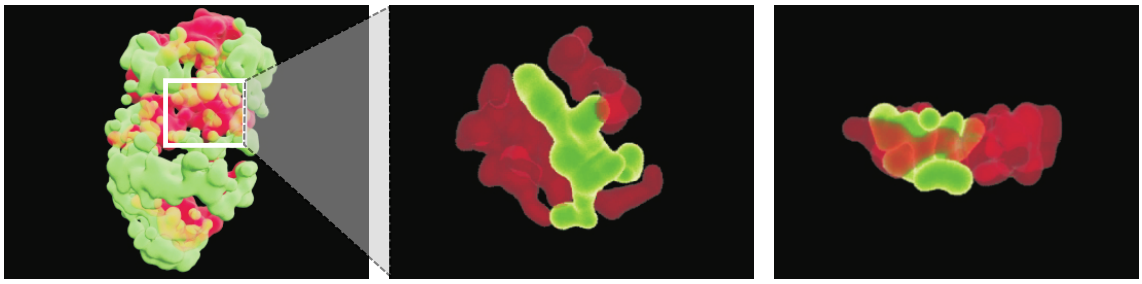


Figure 3.10: 3D reconstruction of labelled nuclei uncovers mis-colocalisation. Apparent colocalisation results from the juxtaposition of green and red channel signal and the angle of observation. 3D reconstruction of chromatin domains based on imaging data allows the rotation of objects and unmasks the mis-colocalisation effect. For details on the transformation see section 2.1.5.2 in the Materials and Methods chapter.

the colocalisation signal may be an illusion from the volumetric projection.

To confirm that the colocalisation signal observed was due to juxtaposition of images and not true spatial colocalisation of neighbouring CTs, we generated a 3D model of the CTs using the processed image data. By rotation of the 3D model we confirmed that regions of apparent colocalisation appeared when higher-order structures of one CTs protruded inside the spatial territory of another chromosome. Figure 3.10 shows the reconstructed 3D model of a whole nucleus and volumetric cropped region where colocalisation appears to happen, when in reality it does not. Rotation of the model shows that CTs are spatially distinct.

Together with individual optical sectioning of nuclei and 3D reconstruction of the labeled nuclei, we confirmed that CTs occupy their unique space inside the nucleus. However, proper colocalisation analysis required a systematic quantification of correlation of the signal of both channels for each voxel in the image to establish this result more rigor-

ously.

3.2.3 Quantitative measurement of inter-chromosomal mixing

Instead of relying on the visual estimation of colocalisation by overlaying the green and red channel to test inter-chromosomal chromatin intermingling, we quantified the level of colocalisation in our samples by measuring the total volume of the nucleus with both signals present at a significant level and the amount of DNA within the colocalised volume.

Many different analytical tools have been developed to perform a systematic, objective analysis of colocalisation. The most common ones are based on the intensity levels of each voxel in both channels and their statistical relationships [92, 145]. Alternative methods that are not based on the intensity of signal but on definition of objects and the spatial overlap of them were not considered in this study, as they are most useful for isotropic objects such as speckles or localised foci and not for amorphous masses of signal such as the ones that characterize CTs.

3.2.3.1 Evaluation of colocalisation by intensity correlation coefficient-based methods

Intensity correlation coefficient-based (ICCB) methods attempt to quantify levels of colocalisation in a global statistical manner. The relationship between the voxel intensities of two images can be described by a linear regression. The slope of the linear regression will describe the rate of association of one channel relative to the other but this does not give any information regarding how good the data fitting to the plotted line. The Pearson's Coefficient (PC) is an estimate of the extent of fit of the linear regression model.

Pearson's Coefficient

Pearson's coefficient values range from -1 to +1, inclusive. Negative values represent exclusion of the signal in one channel against the other, zero values represent no correlation and +1 represents complete colocalisation of both channels. For instance, if a linear model with a slope (rate) where the green voxels are twice the intensity of the red voxels, the meaning of a +1 PC is that **all** the green voxels are **exactly** two times more intense

than the red voxels.

The relationship between two channels can be visualized in a scatter plot, or fluorogram, that represents one channel on each axis. Perfect colocalisation will plot all the points along the diagonal, noise-corrupted signal will spread points along the diagonal whereas noise or background will be scattered all around the plot. Differences in intensity of the two channels, together with bleedthrough (or crosstalk) between them, will also change the distribution of the points and the slope of the fitted line (Fig. 3.11 A).

One disadvantage of the PC as a metric for spatial overlap is that it cannot distinguish the signal that is not colocalising in one channel against the other. For instance, there can be cases where all green voxels colocalise with the red ones but not all the red voxels colocalise with green voxels. Based on the PC, alternative methods for the estimation of global colocalisation have been developed, which measure the proportion of an individual channel that contributes to the colocalisation signal.

Mander's Coefficient

Mander's coefficient are channel specific metrics (e.g. M1 and M2 for green and red channels, respectively) with values that fall in the range of 0 to 1, where a value of 1 means that all voxels of that channel colocalise with the other channel; 0 values represent no overlap between the channels.

The Mander's coefficient is computed as follows. For the green channel, it is defined as the ratio of the sum of intensities of the green signal where the red signal is above a defined threshold between the total intensity of green channel; and vice versa for the red channel. This operation implies the estimation or definition of a threshold value as the background signal has to be eliminated from the image, otherwise the proportion of voxels above it would be significant and would severely distort the calculations. In our analysis we empirically found that the best way to systematically determine the threshold point for background removal in the images was the value of the standard deviation of the mean of the intensity values across the image for each channel.

Given the susceptibility of the colocalisation coefficient to noise and background it is important to emphasize the advantages of our labelling strategy, which is not based on

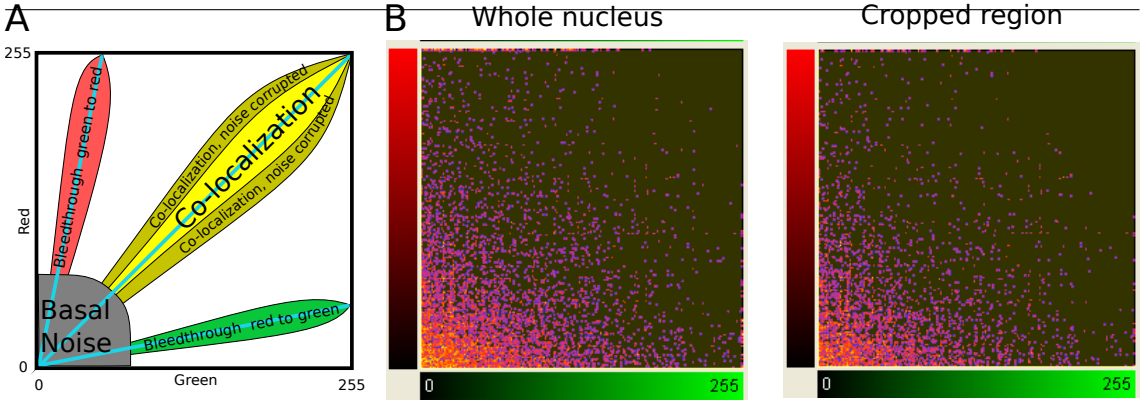


Figure 3.11: 2-Dimensional histogram for colocalisation visualization. (A) Schematic representation of colocalisation regions in the 2D histogram and the effect of noise together with bleedthrough between channels. Adapted from [92]. (B) 2D histogram of raw images of a complete nucleus with a PC of -0.0194 compared to the corresponding 2D histogram of a cropped region from the same cell and a PC of -0.0437. The corresponding nucleus and cropped regions are shown in Fig. 3.9 B and C.

immuno-detection systems that generally increases the levels of background signal. Examples of 2D histograms from the nucleus shown in Fig. 3.9 are displayed in Fig. 3.11 B for the whole nuclear volume and for a cropped region where levels of apparent colocalisation were higher. The spread of the data points confirms that there is no significant colocalisation either in the whole nuclear volume or as in the cropped regions where the highest apparent colocalisation signal was detected.

Optically sectioned nuclei (n=10) like the one shown in Fig. 3.9 were analyzed to monitor levels of channel colocalisation. Table 3.1 shows a comparison of how the colocalisation coefficients vary depending on how images were processed (Table 1). Unprocessed images show only a small decrease in colocalisation after thresholding, especially in the Mander's coefficient values as the low-intensity voxels are removed from the image. Independent of how images are treated, low levels of colocalisation are present in all cases, with negative average PC and low Mander's coefficients. Median filtering is able to remove background signal effectively and represent correctly the low levels of colocalisation in contrast with mean Gaussian filtering that by spreading the signal alongside the labeled structures increases the level of apparent colocalisation.

Due to unlabeled chromatin and chromatin-free regions a high proportion of the voxels in the data lack any signal in either channel. Nevertheless these black voxels influence the numerical analysis. To get rid of this influence, we repeated the same colocalisation measurement in cropped volumetric regions of interest (ROI) (as represented by Fig. 3.9

Summary 10 nuclei	No threshold			Thresholded		
Coloc. Metric	PC	Mander's		PC	Mander's	
		Green	Red		Green	Red
Raw Files (no filtering)	-0.03646	0.18368	0.05448	-0.05216	0.09689	0.0401
Median Filter 3x3x3	-0.07823	0.11228	0.01549	-0.10285	0.04953	0.01616
Gaussian Filter 0.08 μm	-0.04365	0.5957	0.42345	-0.11055	0.1041	0.07763

Table 3.1: Analysis of different approaches for signal colocalisation. A variety of automatic and manual protocols were tested to monitor levels of colocalisation in samples generated throughout this study. Pearson's and Mander's coefficients were used as indicators of colocalisation between different channels representing different CTs. Numbers belong to confocal series for 10 different nuclei and cropped regions as the ones shown in Fig. 3.9 B and C. Despite the different conditions used, the impact of colocalisation levels remained low. Using median filtering increases colocalisation dramatically by spreading the edges of the labeled structures. Simple median filtering improves the quality of the images without this distortion. Thresholding images for removal of low intensity voxels reduces level of colocalisation slightly.

Category	Measure	Value $\pm SD$
Colocalisation coefficients	PC	-0.07 $\pm$0.04
Colocalisation coefficients	M Green	0.05 $\pm$0.04
Colocalisation coefficients	M Red	0.08 $\pm$0.06
Volumetric statistics	Colocalised volume μm^3	0.28$\pm$0.24
Volumetric statistics	Colocalised % of cropped box	0.96%$\pm$0.85
Volumetric statistics	Green volume % of cropped box	17.3%$\pm$7
Volumetric statistics	Red volume % of cropped box	11.4% $\pm$3
Volumetric statistics	Total labeled % of cropped box	$\sim$28.7%

Table 3.2: Comparison of image filters and Colocalisation analysis of whole nuclei and cropped regions.

panel B-C) of the nuclei. We were particularly interested in measuring the extent of colocalisation in regions of the labeled nuclei where signal in both channels was most intense and where apparent colocalisation was found (Table 3.2).

Cropped regions included 28.5 μm^3 of the nuclear space and also showed very low levels of colocalisation, with less than 1% of the space inside the cropped box having significant signal levels in both channels in the same voxels. To test if chromatin structure could account for the colocalisation observed we extended the analysis to larger numbers of cells and perturbed the integrity of foci by modifying histone acetylation.

3.2.4 Local environment defines the integrity of DNA foci

In the previous analysis we manually chose cells that had similar intensity distributions between the two channels. Selection bias may exist in this data set, for that reason we randomly chose cells for this analysis. For a more detailed colocalisation analysis we improved the microscopy by using a Zeiss LSM 710 microscope. As chromosome structure is tightly linked to post-translational modification of histones we asked whether alteration of the chromatin acetylation levels could influence the degree of crosstalk be-

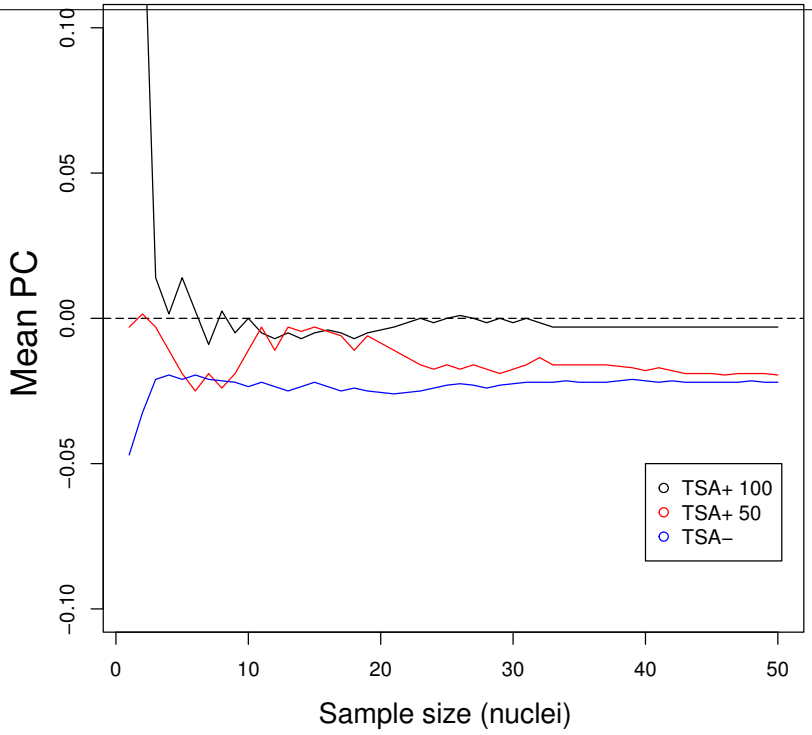


Figure 3.12: Little or no variation is seen above sample sizes of 30.

tween neighbouring CTs.

Double-labeling of CTs was performed as before by alternative rounds of modified DNA precursor pulses and cell division cycles. After mitotic segregation of individual CTs we exposed the cells to the histone deacetylase (HDAC) inhibitor TSA [146] at two different concentrations (50 and 10 ng/ml; TSA+ 50 and TSA+ 100 respectively) for further 24 hours. A control, untreated group (TSA-) was labeled in parallel. Confocal optical sectioning was performed on randomly selected nuclei that were labeled in euchromatic regions. Euchromatin regions are easily identified as they show the typical early S phase pattern of homogeneous abundant signal across the whole cell nucleus in contrast to mid-late S phase that clusters signal at the nuclear periphery and constitutive heterochromatin clumps.

A sample size (n) of ∼30 nuclei is usually used for this kind of experiment [138]. Figure 3.12 shows that after a sample size above 30 the variation of the data is minimum. For each condition assayed we used a sample size of at least 50 nuclei. Z-sections were also subjected to a 3D median filtering step and colocalisation analysis as before (see Fig. 2.1 in the Methods section for details of the high-throughput image processing).

In accordance with the results from the previous analysis, we observed that regions of

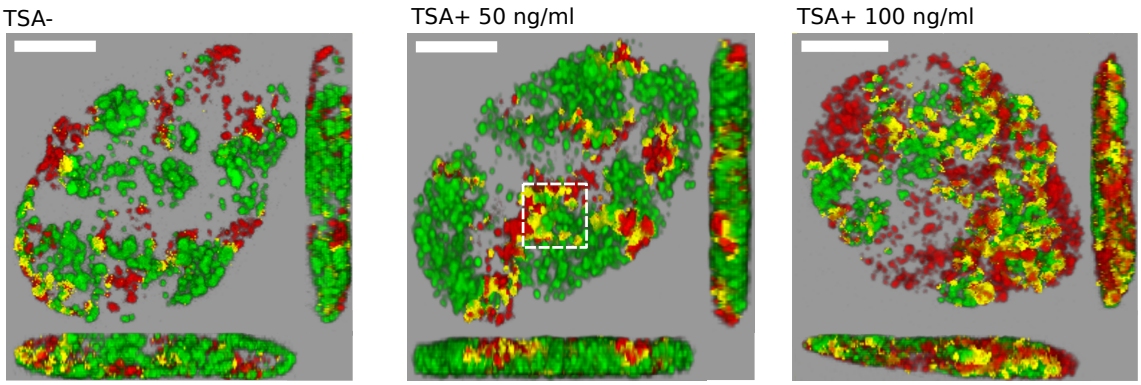


Figure 3.13: Chromatin epigenetic status contributes to territory confinement. TSA-induced hyperacetylation of histones disrupts the native architecture of DNA foci. These changes are reflected on the levels of colocalisation observed between neighbouring CTs. White box in mid panel shows a typical example of the cropped regions used for a more detailed analysis in Table 3.3. Scale bars of 5 μm are shown on individual panels.

colocalisation were detected only along the boundaries of adjacent domains, as maximum Z-projection of representative cells for each treatment show in Fig. 3.13. Cropped ROIs analysis was also performed for these experiments; data is summarized in Table 3.3.

Low levels of colocalisation were detected in this analysis. For instance, the cropped ROIs that showed highest colocalisation index contained only 0.55$\pm$0.6% of their voxels with colocalisation signal. Consistent with the results from the preliminary colocalisation analysis in the previous section, the average percentage of voxels of the ROI that were labeled by either channel was 27.8% of the volume of the ROI.

In the case of the cells under TSA treatment we observed a slight, but significant, increase in the levels of the colocalisation signal detected by colocalisation coefficients as shown in Figure 3.13. The effect was dose dependent as the samples exposed to a concentration of 100 ng/ml showed an increase of colocalising signal relative to the ones exposed to only 50 ng/ml and to the control group. This result suggests that by disrupting chromatin native structure by hyperacetylation, regions from neighbouring CTs will partially overlap in space.

	Vol. Coloc (μm^3)	%Coloc.	%Green	%Red	PC	M Green	M Red	%occupied
Nuclei n=50								
TSA-	6.03 $\pm$4.49	0.30 $\pm$0.22	8.48 $\pm$2	4.73 $\pm$1.7	-0.0209 $\pm$0.01	0.03 $\pm$0.02	0.05 $\pm$0.02	13.21
TSA+ 50	12.23 $\pm$10	1.01 $\pm$0.8	12.06 $\pm$4.2	7.71 $\pm$2.5	-0.0156 $\pm$0.02	0.07 $\pm$0.04	0.11 $\pm$0.07	19.77
TSA+ 100	23.90 $\pm$19	0.98 $\pm$0.8	6.86 $\pm$2.33	9.75 $\pm$2.7	0.0060 $\pm$0.05	0.12 $\pm$0.08	0.09 $\pm$0.07	16.61
ROIs n=50								
TSA-	0.36 $\pm$0.4	0.55 $\pm$0.6	13.87 $\pm$4.3	13.93 $\pm$5.6	-0.0994 $\pm$0.03	0.03 $\pm$0.03	0.04 $\pm$0.03	27.81
TSA+ 50	1 $\pm$0.6	1.8 $\pm$1.6	19$\pm$8.3	19$\pm$5.7	-0.13$\pm$0.07	0.08 $\pm$0.06	0.09 $\pm$0.06	59
TSA+ 100	1.55 $\pm$2.7	2.15 $\pm$2.07	17.50 $\pm$4.5	21.97 $\pm$5.6	-0.1087 $\pm$0.08	0.12 $\pm$0.1	0.1 $\pm$0.1	40

Table 3.3: Colocalisation analysis in cropped regions After treatment with TSA.

Similar results were observed independently of how images were handled. Moreover, as a complementary control, we performed colocalisation analysis for raw images without a median filtering step or thresholding. Figure 3.14 A shows a comparison of the distribution between each experimental condition, before and after image processing. The trends were also present even in raw images, which nevertheless showed higher variability in the distribution of PCs (Table 3.4 and 3.5).

Using Pearson's correlation coefficient is useful as a general estimation of the degree of colocalisation, however, as an abstract numerical index it does not provide a direct biological interpretation of what it represents. To have a more meaningful understanding of the real colocalised signal observed we also performed a volumetric assessment of colocalisation. We performed a simple count of the number of voxels containing significant signal from both channels and translated it into volumetric units (Fig. 3.14 B).

In the untreated control group, cells showed on average 6 μm^3 of colocalising signal, representing 0.3% of the nuclear volume. In the case of TSA-treated cells the colocalising volume increased as a function of TSA concentration. In these cells the colocalised volume increased up to 24 μm^3. Image processing did not significantly change the colocalised volume as can be visually confirmed by the overlap of notches in the box plots of Fig. 3.14 B.

To complement the volumetric analysis we transformed the colocalised data to proportions of the total volume in each nucleus and ROIs. As before, ROIs were selected as before by manually selecting the regions with the highest intensity of labelling, a typical example is shown in as a white box in Figure 3.13.

Figure 3.15 shows box plots comparing the distribution of the three different treatments. Proportions of colocalisation increase as the total volume measured decreases when analyzing whole nuclei versus cropped ROIs. However, in all cases, colocalisation signal occupies less than 1% of the volume analysed and is influenced positively by treatment with TSA.

Pair-wise Mann-Whitney		
Nuclei		
	TSA+ 50 ng/ml	**TSA+ 100 ng/ml**
Pearson's Coefficient		
TSA-	6.2E-01	5.5E-06
TSA+ 50ng/ml	-	1.83E-03
Colocalised Volume		
TSA-	2.48E-04	9.457E-12
TSA+ 50ng/ml	-	8.254-05
Colocalised Percentage		
TSA-	7.616E-10	4.923E-11
TSA+ 50ng/ml	-	8.931E-01
ROIs		
	TSA+ 50 ng/ml	**TSA+ 100 ng/ml**
Pearson's Coefficient		
TSA-	5.88E-03	3.519E-02
TSA+ 50ng/ml	-	4.645E-01
Colocalised Volume		
TSA-	3.705E-08	5.922E-11
TSA+ 50ng/ml	-	1.564E-01
Colocalised Percentage		
TSA-	8.593E-09	1.544E-10
TSA+ 50ng/ml	-	6.951E-01

Table 3.4: Pair-wise Mann-Whitney statistical analysis of colocalisation results. As the distribution of the values from the three different treatments do not follow a normal distribution, non-parametric statistical methods were used.

Kruskal-Wallis test	
Nuclei	**ROIs**
Pearson's Coefficient	
3.88E-05	1.48E-02
Colocalised Volume	
1.166E-11	1.13E-11
Colocalised Percentage	
1.68E-12	1.513E-11

Table 3.5: Kruskal-Wallis test of colocalisation results.

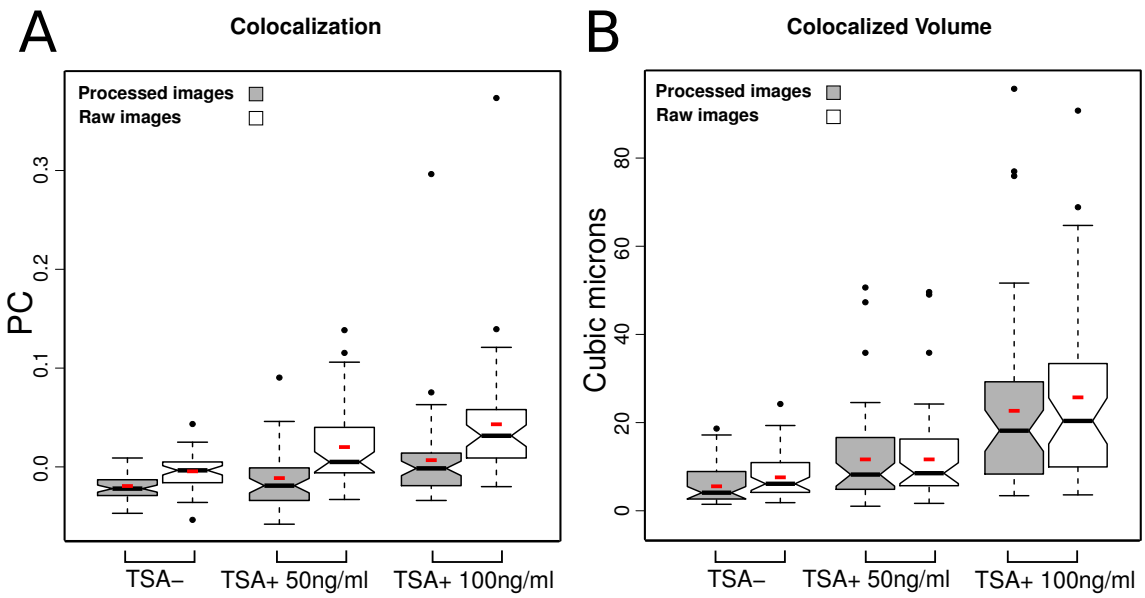


Figure 3.14: High-throughput image analysis. (A) Untreated cells showed a negative Pearson's coefficient consistent with the low levels of colocalisation in the sample. TSA treatment showed a significant increase in Pearson's coefficient, demonstrating increased colocalisation. In order to develop a quantitative estimate of channel colocalisation, voxel-level channel intensities were extracted and the volume (measured in μm^3) of colocalised voxels calculated. The trends are conserved regardless of image processing. Small red boxes on the box plots represent the mean value for each distribution.

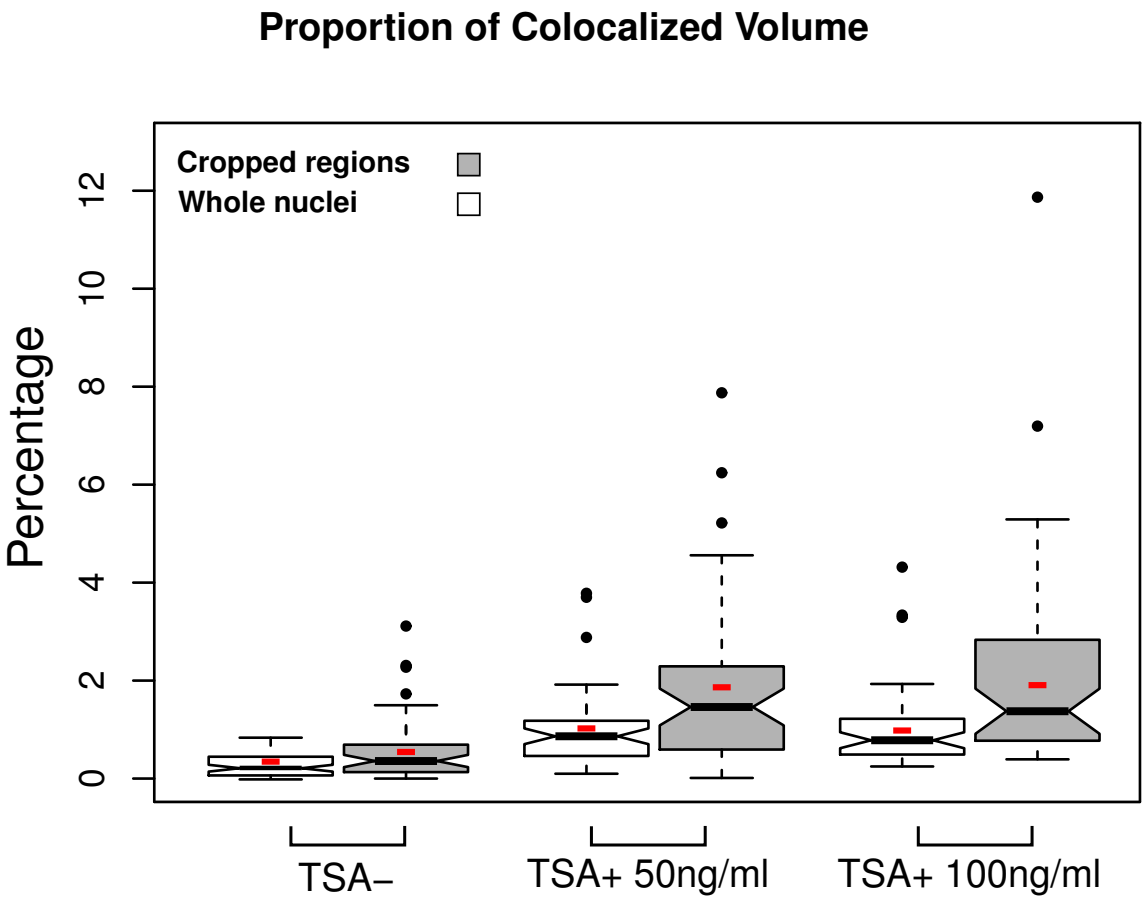


Figure 3.15: Proportion of colocalised volume.The colocalised volumes were calculated as a percentage of the total nuclear volume and in cropped regions. The effect of TSA is more evident after analysis of regions of interest defined by homogeneous intensity and the effect of unlabeled voxels removed. Small red boxes on the box plots represent the mean value for each distribution.

3.3 Discussion

The models proposed for the understanding of chromosome territories such as the giant-loop, lattice, ICN and CT-IC are not necessarily exclusive but complementary. The conflicting views in the field had been compared to the parable of the blind men and the elephant: "... present models reflect rather the nuclear sites and scales of resolution where their studies were undertaken than a deep understanding of the global architecture and functional implications of the elephant (the nucleus)..." [138]. The reason this analogy is used is that experimental strategies of different nature had been applied for understanding chromosome architecture and each of them focuses on the strengths of their method and extrapolate their observations as a general feature of chromosome organization.

The pitfalls in the interpretation of each model is inherent of their respective system. For instance, live cell imaging approaches are very informative about the dynamics of CTs and their interactions, however, the low levels of illumination required for live cell imaging limit the resolution at which experiments can be performed. In the same way, molecular analysis, such as chromosome conformation capture (3C) and its derivatives, provide the power of detecting long range interactions with high specificity, however, the information provided is based on cell populations. These molecular methods lose completely the information of single cells. Given the complex arrangement of pros and cons of the methods available, the need for an integrative model, able to reconcile the different efforts in the field is becoming more obvious.

We measured the extent of chromatin mixing between adjacent CTs with fluorescent DNA precursors incorporated at the time of DNA replication and therefore labeled newly replicated DNA directly without the need for additional steps that might compromise native chromosome architecture [147]. Individual DNA foci were observed and clustered into higher-order domains and CTs. The levels of mixing between adjacent CTs were detected at very low levels (less than 1%) or not at all, in agreement with previous studies [148, 149]. Our method should be able to detect high-order chromatin structures with the same sensitivity at any part of CTs and capture the intermingling regions, as density of DNA has been observed to be constant across the surface and the interior of CTs as well as in intermingling sectors [141]. In addition, as our method does not rely on hybridization of

fluorescent probes, it is not susceptible to staining differences observed with commercial chromosome-specific 'paints' as the one reported in [150].

Parallel to the advantages of the method we implemented, it is important to point out its limitations. Even though out labelling method avoids the possible biases from traditional staining methods, events of thin chromatin threads, looping out from their CT could potentially show such a small signal intensity that makes impossible their detection given the microscopy used. A conjugation of information gathered from different 3C based approaches [32, 120] and custom FISH probe pools [150] in order to find good candiates of events of long range interactions as previously shown for the gene-dense regions of 11p15.5 on human fibroblasts [138].

Our results of structurally unperturbed chromatin based on light microscopy with little background signal contradict FISH experiments that presumably do not alter chromatin structure and electron microscopy [141]. It is important to emphasise that FISH shows some disadvantages compared to our method. First, in order to expose DNA to the FISH probe and allow hybridization, DNA has to be denatured. Denaturation can compromise the native local structure of chromatin. Secondly, FISH yields higher levels of background signal and the manual thresholding strategy used in these experiments can be misleading.

Nuclear organization at the scale of the mega-base domain has been observed to be stable across cell cycles [29, 148, 28] suggesting that this level of organization defines and constrains the dynamics of chromatin. Moreover genome-wide interaction maps based in chromosome conformation capture (3C) have shown that chromatin domain interactions occur preferentially within themselves [32, 33], however, gene activity can influence their dynamics and enhance the spectrum of interactions [120].

Genome function, such as gene activity, is known to depend on the epigenetic landscape it presents and the connection between function and chromatin epigenetic status extends as well into structural changes. Euchromatic regions are known to be more fluid and characterized by a more accessible chromatin configuration. These characteristics are accompanied by its respective post-translational modifications of histones, such as acetylation of histones.

We altered the chromatin epigenetic status of DNA foci in order to measure the ex-

tent to which DNA foci innate structure limits the interaction dynamics. Global hyperacetylation of chromatin due to inhibition of histone H3 and H4 deacetylases by TSA caused an increase of the colocalising volume of adjacent CTs of about 4 times relative to the control group. This observations confirmed the idea that chromatin environment influences the mechanisms that maintain DNA foci as a structural integrity.

This suggests that regions of apparent intermingling observed by Pombo [141] could just represent a specific cases of local enrichment of particular combination of chromatin marks and not a general property of chromosome organization as they also show how transcriptional activation could rearrange chromatin structure [141].

The observed increase of colocalisation signal due to hyperacetylation of histones is also in accordance with a relatively new line of evidence that has shown long-range genomic associations that are linked to gene expression [151, 152]. Several instances of long-range chromatin physical contact have been reported [118, 153, 154, 117, 116, 155, 156, 157, 158], even from regions located in different chromosomes [114, 119]. Nevertheless, it is important to point out that, single-cell analysis has demonstrated that these associations are not frequent and can be found only in a small fraction of the cell population studied ($\sim$10%) [114, 159, 115]. It remains an open question how and in which circumstances long-range interactions happen, what is the real frequency of these associations and how they reconcile with alternative models of nuclear architecture

3.3.1 Conclusions and open questions

From a functional point of view it is important to understand how the interaction between CTs and the dynamic set of interaction networks provided by them establishes during development and tissue differentiation and even if these rare interactions could have an influence in genome evolution.

In this chapter we made used of fluorescent DNA precursors to label individual chromosome territories and evaluated the degree of spatial overlap between them. We confirmed the generally accepted notion that chromatin from different chromosome territories does not mix extensively. In addition, we observed that colocalisation signal was positively affected when disturbing native chromatin structure by promoting a more open conformation by treating cells with histone deacetylase inhibitor TSA.

An integrative model should arise after the different pieces of evidence of all the efforts are considered. New genome-wide single-cell analysis of interactomes promise to shed some light into the field as of cell-population based studies may mask the real biological phenomenon, the overlay of single cell data should reconstruct the patterns observed in cell population based studies. Nevertheless, the problem of how stable long-range interactions are and the relationship they show with their genomic landscape is still an open question that live cell imaging, in addition to high-throughput image analysis, can address.

Chapter 4

POST-GENOMIC ANALYSIS OF THE BANDING PATTERN OF HUMAN METAPHASE CHROMOSOMES

4.1 Introduction

In the early years of genetics research, after the total number of chromosomes in healthy human adults was determined [160], and the link between Down syndrome and other congenital disease were linked to chromosomal aberrations [161], the field of human cytogenetics made great advances in a period of time beginning from the early 60's to late 90's.

The development of fluorescent dyes in 1968 represented a technological breakthrough that accelerated this field [162]. Exposure of chromosome spreads to these dyes produced reproducible banding patterns that provided a great frame of references for the study of human genetics. Variations and alternatives to this technique followed.

The conference of Paris in 1971 resulted in the formation of the International System for Human Cytogenetic Nomenclature (ISCN) [163, 164]. This commitee defined the basic rules for chromosome classification and nomenclature and developed a standard diagrammatic representation of chromosome landmarks. This 'ideogram' has undergone many rounds of improvements and updates since its proposal, with the latest version released in 2009 [165].

Additional techniques allowed more rapid development of the cytogenetics field as custom probes representing specific sequences could be mapped to the chromosomes based on radioactively labeled probes hybridized in situ [166], bridging the gap between sequence data and cytogenetic maps. This technique evolved to incorporate fluorescent nucleotides instead of radioactive ones, namely Fluorescent In Situ Hybridization (FISH)

[167].

Many techniques were used in order to reveal different, complementary, banding patterns in the cytogenetic map. Each one was able to reflect different properties of the chromosomes, such as C-banding for staining centromeres, Quinacrine (Q)-banding for highlighting AT-rich DNA, Giemsa (G)-banding for GC-rich DNA, Reverse (R)-banding as a inverted version of G-banding, Telomere (T)-banding for very GC-rich G-bands and telomeres, and replication-based banding that was able to monitor different stages of S phase depending on the time of pulse labeling in proliferating cell cultures [168]. The technique that showed the sharpest and most distinguishable patterns was the G-banding, which remains as the standard method for cytogeneticists. In seminal studies by Yunis et al. [169] they systematically screened the largest 5 human chromosomes at 4 different stages of mitosis and set the basic cytogenetic maps. Chromosome condensation, from late prophase to metaphase, consists of a continuum of condensed states from long, thin and flexible, to short, thick and rigid bodies (Fig. 4.1). Each of these stages showed similar intermediate patterns. Patterns from early mitotic states (late prophase) showed larger amounts of bands, in contrast to the later stages (metaphase) which used to show fewer. Identifying bands in high-resolution maps was more subjective and imprecise but given the similarity between maps at different stages of mitosis, the low-resolution banding patterns of metaphase chromosomes served as a reference for calling bands in the higher-resolution patterns of late prophase. This phenomenon suggested a nested structure of bands, were dark bands were usually composed of smaller sub-bands [169]. This idea has been confirmed recently [170, 171]. Before the G-banding method developed by Yunis et al. [169] was introduced, ideograms were only portrayed as black and white bands. Given that their technique proved to be very efficient for the detection of G-bands, 4 different shades of grey that reflected the Giemsa staining intensity, were incorporated to the maps (Fig. 4.1). In addition, ideograms that represented different levels of resolution were created at the 300, 850 and 1,250 level

The Giemsa dye is composed of a mixture of eosine and five compounds that cover a range of methylated forms thiazine. These variants range from the un-methylated thiazine molecule to the to tetra-methylated variant; thionine, Azure C, Azure A, Azure B and

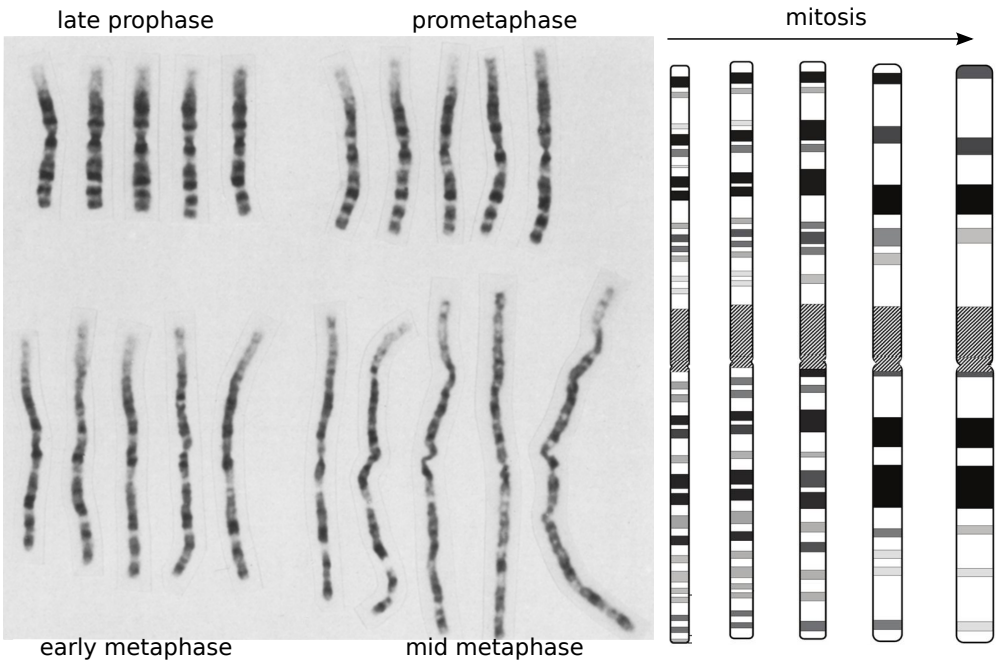


Figure 4.1: Giemsa staining of chromosome 1. The G-banding pattern changes depending on the stage in mitosis in which chromosome spreads are prepared. Early spreads produce higher resolution patterns (850 bands) and later spreads produce thicker, low-resolution patterns (300 bands). Photo of spreads modified from [169] and ideograms modified from [104].

metheylene blue, respectively [172]. These molecules interact with the phosphate groups in DNA and "side stack" along it [172].

At the beginning, the general properties of the DNA found on R- and G-bands were determined by different methods directly on chromosome spreads on glass slide preparations. Given the harsh conditions used for the staining to work, DNA was not in its native form. The first time DNA for each type of band was isolated and studied was achieved by separation of DNA by density gradients of bromodeoxyuridine (BrdU) pulse-labeled DNA during replication. Synchronized Chinese hamster cells were pulse-labeled at different stages of S phase and then DNA separated into early- and late-replicating DNA. The late-replicating DNA showed higher AT levels than the early fraction. In addition, depending on the time of labeling, R- or G-banding patterns were observed [173] and interestingly, intermediate R/G patterns did not appear. This reflects the interruption of DNA synthesis in the middle of S phase, called the 3C pause where the transition of synthesis from R- to G-bands synthesis happens. This result suggested that R bands are replicated strictly before G-bands.

G. Bernardi suggested the isochore theory of organization of the human genome [174]

based on the observation of five fractions after centrifugation of genomic DNA in in Caesium chloride (CsCl) gradients. These fractions represented five different classes of isochores. Each of these families were characterized by different GC levels that conferred them specific buoyant densities. Bernardi speculated that the banding pattern matched the isochore distribution. He later confirmed his ideas in chromosome 21 [175] and then the rest of the human genome [170]: the GC-poor and GC-rich isochore families occupy preferentially G bands and R bands respectively, in accordance to previous observations of differential sequence composition of bands [173].

In a rather intricate review that compiled most of the information available at the time, Holmquist re-classified R-bands into 4 different "flavours" based on the levels of GC content and *Alu* repeat density [176] at the 400 band level. This observation pointed out the relationship between band-type and sequence composition and suggested that repetitive sequences represented a major driving factor [177].

Additional lines of evidence support the idea that sequence composition contributes to the determination of the banding pattern through chromatin structure. By observing the differential fluorescence quenching of sequence-specific dyes in mitotic chromosomes, it was suggested that there are differences in the tightness and length of chromatin loops attached to the AT-rich scaffold of mitotic chromosomes [56]. This phenomenon is not only observed in metaphase but appears also in the interphase nucleus. The distance between FISH probes of R- and G-bands was measured in the interphase nucleus and revealed that chromatin folding is different in both regions [178]. At a higher-order scale the different types of bands have been shown to occupy different compartments of the eukaryote nucleus [179].

A more direct relationship between sequence composition and biological function was found as clusters of broadly expressed genes matched the differential GC content blocks [180, 181]. Many of the features that differentiate G- from R-bands, such as gene-expression levels, differential enrichment levels of SINES and LINES, were found to cluster and called "RIDGES" and "anti-RIDGES", where RIDGES are equivalent to T-bands [182], surprisingly the authors made no association with the rest of the banding pattern.

After the sequencing of the human genome, the potential to understand the different factors that determined the banding pattern increased substantially. Using the first drafts of the human genome, Niimura and colleagues attempted to recreate *in silico* the chromosomal banding pattern based on GC content data at different window sizes [183]. This attempt partially resembled the banding pattern according to [104].

A parallel effort that relied in the first drafts of the human genome bridged the gap between the genome sequence and the cytogenetic map. Based on the BAC clones used for the Human Genome Project i[184, 109], Furey and colleagues integrated the cytogenetic map from [104] with locations in the genomic sequence by using a dynamic programming algorithm [97]. The set of coordinates produced by this work are the standard reference in the genomic data resources, such as the UCSC Genome Browser [95, 96].

Alternative high-resolution maps for the banding pattern that did not take into account the clone hybridization data but take into account the isochore distribution, have been proposed [185]. Isochore maps were compared to Furey's maps visually and show good correspondence, however, a more robust and automatic approach is not possible as the genomic coordinates from the isochore map in [185] are not available. Furthermore, the identification of isochores can vary from how the computational approach is implemented [99]. For example the data from the Constantini method from Bernardi's group has been excluded from approaches to unify the available isochore data sets because "no hard threshold was used, and in many cases subjective decisions were made as to whether or not to merge windows, making the Constantini method as described in the original publication hardly fully automatable..." [99].

With improvements in the human genome assembly currently available (hg18) and the plethora of genome-wide data sets available, we performed a comparative analysis of different features at different scales in an attempt to understand the functional genomic elements that define the banding pattern. All the comparisons we performed were based on the coordinates defined by the Furey method.

4.2 Properties of bands

In human, chromosomal bands are classified as Giemsa (G)-negative (gneg or R bands; 414 in total) and G-positive bands (G bands). The latter class comprises four

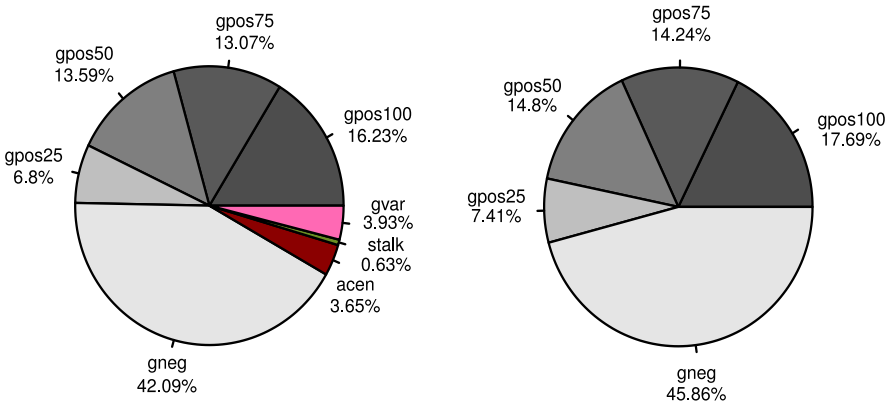


Figure 4.2: Proportion of the genome covered by each class of band. The genome is essentially split in halves by the G and R bands, while special band types as acrocentromeric regions, stalks and variable heterochromatin regions represent only a small proportion of the human genome.

different levels of increasing staining intensity: gpos25, gpos50, gpos75 and gpos100 with a total of 87, 121, 89 and 81 bands respectively. In addition, there is a subset of regions of the genome that are not classified as Giemsa bands, such are: variable length heterochromatic regions (gvar), pericentromeric regions (acen) and stalks belonging to acrocentric chromosomes (stalk). Figure 4.2 .

Using the band coordinates reported in the UCSC Genome Browser [97] we measured the amount of base pairs (Bps) that each class of band covered. We observed that almost half of the genome (42%) is not stained by the Giemsa dye. The largest proportion of the G-positive section was classified as gpos100. Approximately $\sim$8% of the genome did not fall into the Giemsa band categories (gvar, acen and stalk). For further analysis we ignored the portion of the genome covered by these classes.

We also found that Giemsa bands show a differential distribution of their size depending on their staining intensity. The darkest G bands showed the largest sizes whereas G-light bands showed the smallest sizes. Figure 4.3 A plots the density distribution of the sizes of bands per class. R bands (gneg) presented a broad distribution in size whereas the different G band classes showed less variation in size. As mentioned before, G bands showed a positive correlation between staining intensity and band size. G-light bands and gneg had mean values below the genomic average of 3.5 Mb $\pm$1.9 (SD). The smallest group was represented by gpos25 bands with a mean band size of 2.4 Mb $\pm$1.2, followed by gneg with 3.1 Mb $\pm$1.6 and gpos50 bands with 3.4 Mb $\pm$1.6. Only gpos75 and

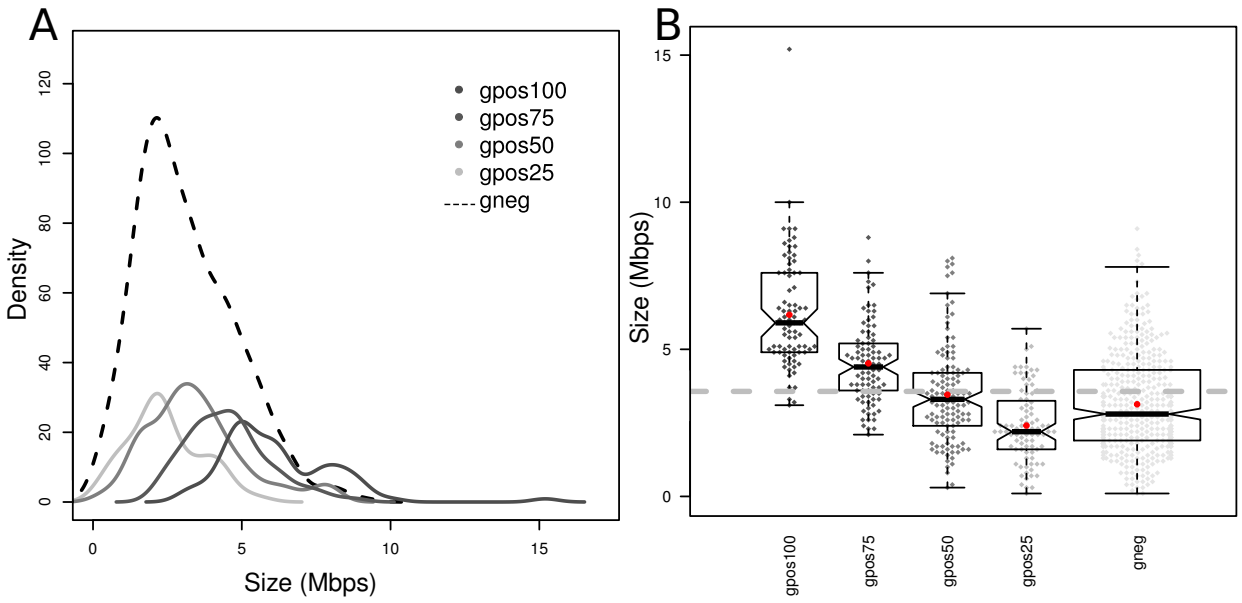


Figure 4.3: Size distribution of bands. (A) Density distributions of the sizes of each band type show that R bands tend to be smaller on average than G bands. Within G bands there is also a differential distribution of sizes as bands tend to be larger as they get darker. (B) R bands and G-light bands are smaller in average relative to the genomic mean band size (dashed line). Box plots are graphical representations of the distribution of a sample, they show the minimum and maximum values as horizontal lines, lower and upper quartiles as the extremes of the box and the central line of the box represents the median of the distribution. Each dot in the plot represents an individual band. The differences of the median values of the distributions are statistically significant (as the notches drawn in each distribution box plot do not overlap; Notches give reference for a 95% confidence interval for the difference in two medians [186]). Mean values of the distributions are represented by red circles.

gpos100 classes showed a mean band size above the genomic average with 4.5 Mb $\pm$1.3 and 6.1 Mb $\pm$1.8, respectively (Figure 4.3 B).

We next deepened our analysis in order to evaluate the different correlates available at genome-wide levels and high-resolution for the human genome (hg18) at three different scales: (1) Analysis of the banding pattern at the sequence level, (2) epigenetic status of chromatin level, (3) S phase programme differences across classes of bands and (4) the higher-order organization of bands.

4.3 Sequence Features of bands

4.3.1 Differences in the structure of genes and their distribution on bands

Previous work has characterized the distribution and structure of genes [177, 97] within each band type. The most comprehensive of these attempts is the work done by

Furey and colleagues, with the first draft of the human genome sequence in 2003. Our analysis reproduces this work with improved data in terms of genome assembly. We analysed different structural properties of genes and how they vary depending on the class of band in which they are found. Figure 4.4 shows box plots comparing each of the features measured on a per band basis. We focused only on protein coding genes and excluded genes that overlapped with band borders to simplify the analysis; 20,315 genes were analysed in total (see Methods 2.2.2).

In accordance with previous studies [177, 97]], R bands contain more genes than G bands. From the set of genes used in this analysis 62% were found in R bands and on average R bands showed 2× more genes than gpos100 (Fig. 4.4 Upper left). As there is a differential distribution of band size across classes, we measured the density of genes (genes per Mb) instead of just the number of genes found per class. We found the same trend of light bands being gene-rich and each class showed gene densities statistically significantly different from the rest. This time, gpos25 showed the highest median density of genes 1.4× more than the genome median and 3.63× more than the gene-poor regions in gpos100. In addition, when measuring the proportion of the band that comprised a gene and coding sequences we found that gneg (R bands) and gpos25 showed median values above the genomic median (Fig. 4.4 Bottom panels).

We extended the analysis to a per-gene based analysis and addressed more specific features of genes as a function of the class of band in which they reside 4.5. Interestingly, the length of transcripts is positively correlated with the darkness of the band. Darkest bands showed the longest transcripts almost 2× larger than the genomic mean, 37,103 bp and 19,637 bp respectively. This difference could not be explained by the length of exons, nor the length of the coding sequences (CDS) as the gpos100, gpos75 and gpos50 were only slightly larger than gpos25 and gneg bands. In contrast, the main structural difference between genes accounted for the length of intronic regions. Total length of introns followed the same trend as the total length of the transcript: the darkest bands (gpos100, gpos75 and gpos50) showed the highest median values relative to the global genomic median, consistent with previous results showing that genes inside RIDGES are characterised by a smaller intron length relative to genes located in ANTIRIDGES [182].

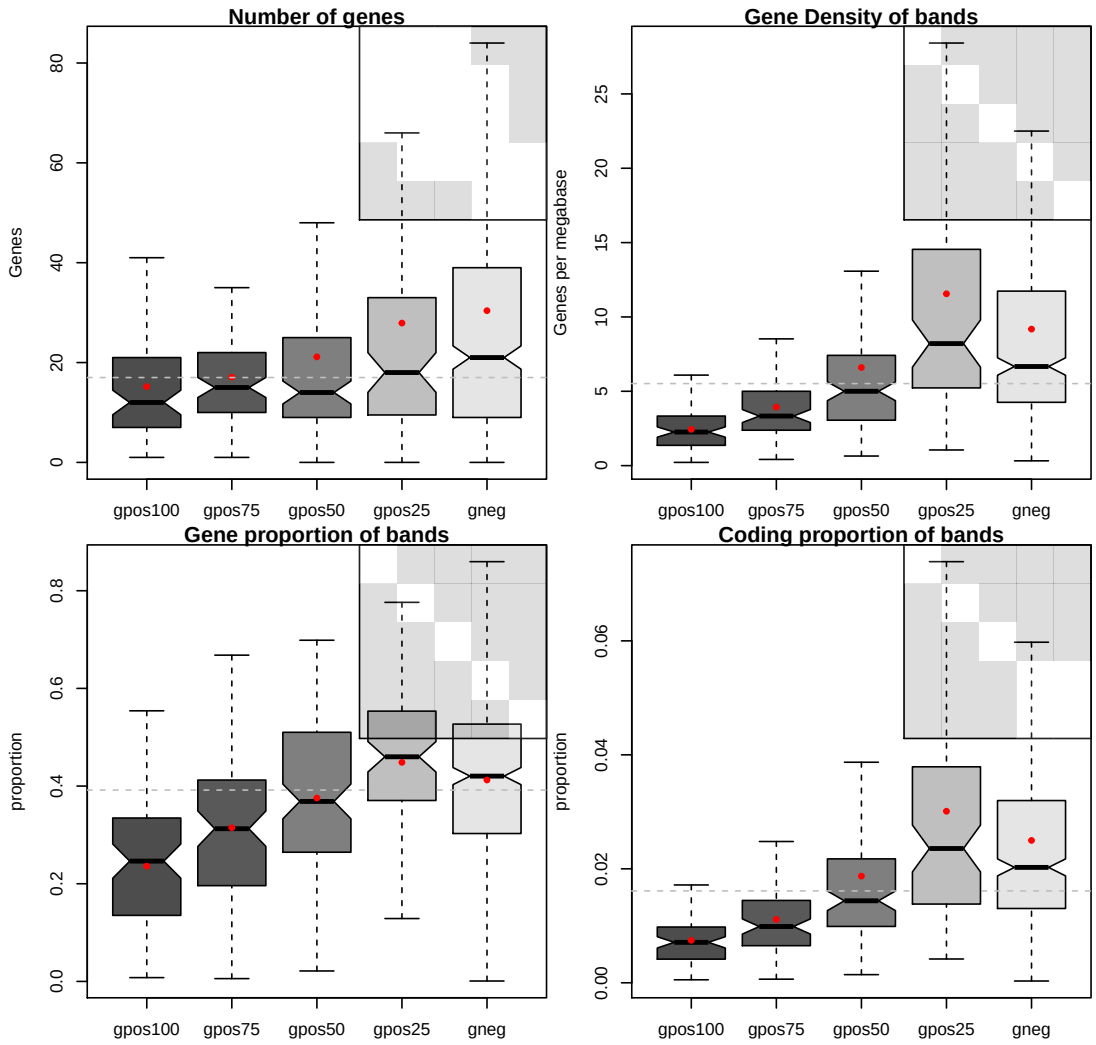


Figure 4.4: Characteristics of genes at the band level. The data units in this analysis are bands, so each data point represents the average value of the metric studied for each single band. Gene density is significantly higher in R bands. Inner panels show a colour code for the p-value of the Wilcoxon rank sum tests for a pair-wise comparison the distributions of all classes of bands against the rest. The matrix is in the same order as in the respective box plot x-axis. Gray cells represent statistical difference between the corresponding pair compared; p-values smaller than 0.05. White value represent no statistical difference between the two groups. Dotted lines represent the genomic median value. Red dots represent mean values of the distributions.

There was a 2-fold difference between the largest median value (gpos100) relative to the smallest median intron length (gpos25): 3,705 Bps and 1,813 respectively. We found no difference in the number of exons in genes with a median of 7 exons per gene for all the five classes of band. Finally, according to the differential gene-density of bands, the length of intergenic regions was positively correlated with G-staining intensity. R bands and gpos25 showed narrow distributions and low median values regarding the size of intergenic regions (18,447.5 and 16,367.5 respectively). Previous work has reported similar results where gene density shows the highest scores for T-bands, followed by R

bands and G bands the lowest gene-dense regions [187].

4.3.2 CpG Islands

It is known that CpG islands (CGIs) are closely related to transcription start sites (TSS) and promoter regions of genes. In accordance to gene distribution trends of the previous section, the density of CpG islands also drops as G bands get darker (Figure 4.6 Left panel). The general properties of CGIs do not vary substantially with regards to the kind of band they occupy. There is only a slight increase in the median length of CGIs in gpos50, gpos75 and gpos100 relative to gpos25 and gneg.

4.3.3 Chromosomal bands show characteristic differences in sequence composition

4.3.3.1 GC content of bands

It has been previously reported that the staining intensity of Giemsa bands is correlated to the percentage of the nucleotides guanine and cytosine (GC content) [168, 173, 174, 56, 97]. It is generally believed that R bands are more GC rich than G bands but when the proportion of GC of each band was measured, the bands containing the highest GC content were gpos25 class with an average of 43.6% $\pm$3.6 (SD). In addition R bands showed more heterogeneity in GC content than G bands as seen in the broad distribution of data points in Figure 4.7.

As GC content at the band level is not a very informative trait we further analysed GC content using a more robust metric by analyzing the distribution of isochore families proposed by Bernardi [174].

4.3.3.2 Analysis of isochores

Isochores are blocks, larger than 300 kb, of homogeneous GC content. They have been classified in five different categories, or families, defined by their degree of GC content. Namely H3, H2, H1, L1 and L2 ranging from more than 53% to less than 37% of GC content; H stands for high whereas L for low GC content. Isochores where first identified by ultra-centrifugation of bovine DNA on density gradients, yielding two different bands [188]. However, even though there is experimental evidence supporting the concept of

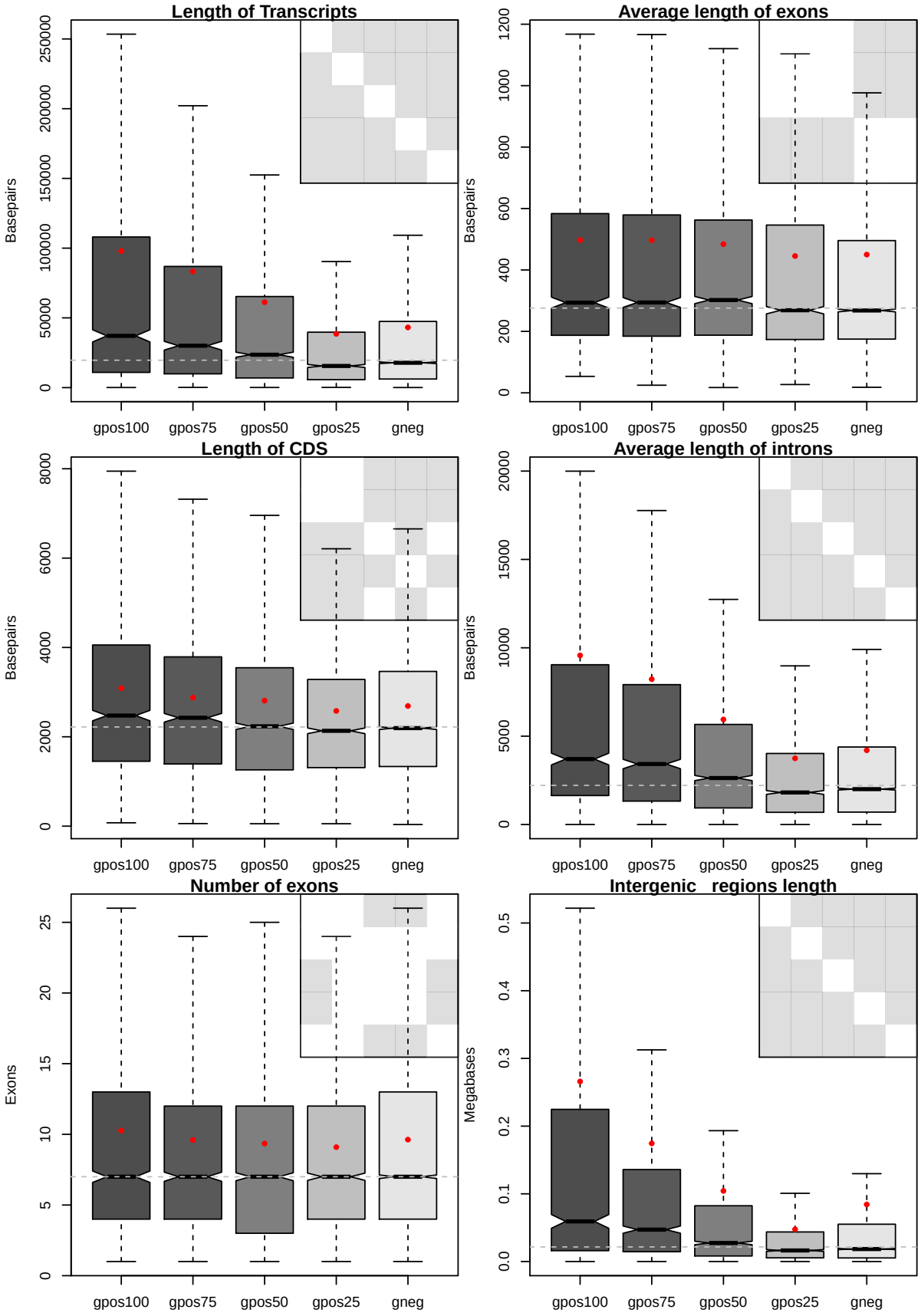


Figure 4.5: Analysis at the gene level. The data points measured in this figure are individual genes, the layout is the same as in Fig. 4.4. Inner panels show a colour code for the p-value of the Wilcoxon rank sum tests for a pair-wise comparison of the mean ranks of the distributions of all classes of bands against the rest. The symmetrical matrix represents each kind of band in the same order as in the respective box plot panel. Gray cells represent statistical difference between the corresponding pair compared; p-values smaller than 0.05. White values represent no statistical difference between the two groups. Red dots represent mean values of the distributions.

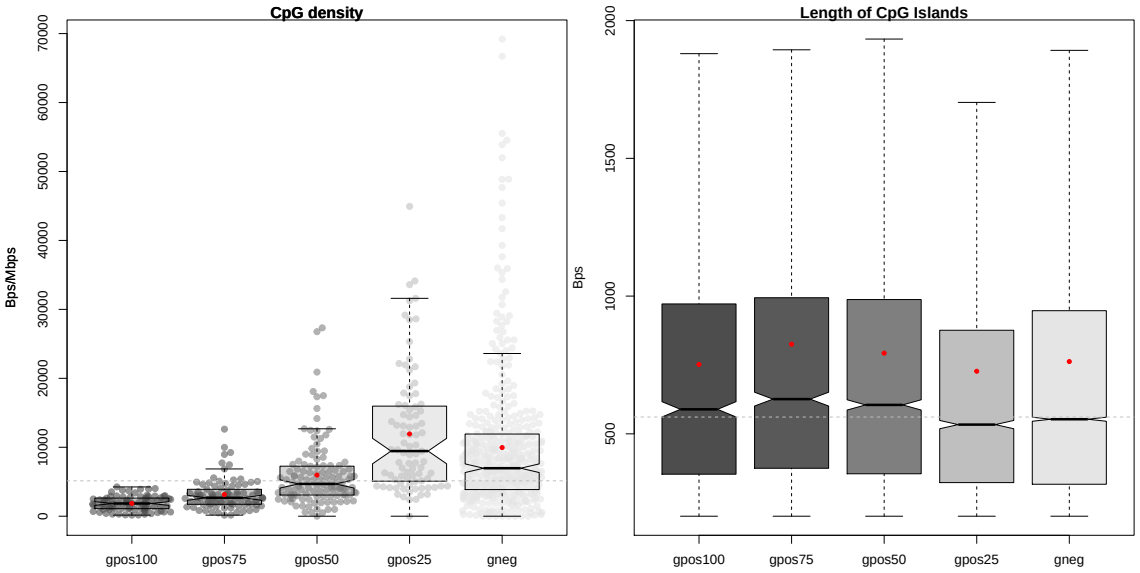


Figure 4.6: CpG islands. Left panel shows the density of CpG island signal at the band level, estimated as the amount of base pairs that correspond to CGI in a 1 Mb window. Each dot in the plot represents one of the 792 bands. Right panel shows the difference in length of CGIs, at the CGI level. As there are more than 27,000 data points these plots do not show each of the data points individually. Dotted lines represent genomic median values.

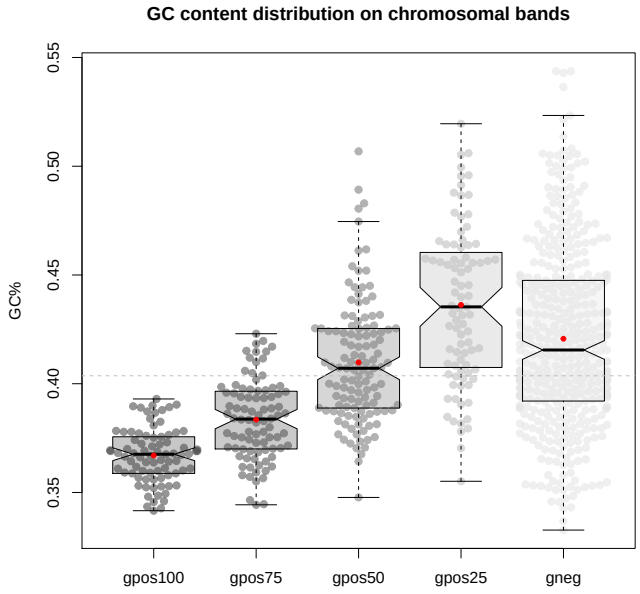


Figure 4.7: GC content of bands. R bands show more variable GC levels than G bands. Same layout as left panel in Fig. 4.6.

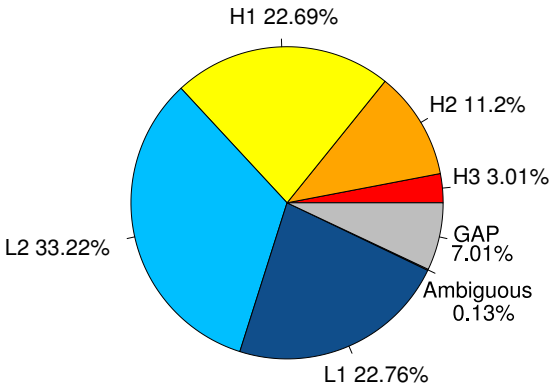


Figure 4.8: Segmentation of the genome as isochore families showed a stricking similarity to the proportional profile of the genome covered by G and R bands. Most of the genome belongs to low GC families of Isochores. The proportion of the genome covered by high GC content families matches the proportion of the genome covered by genes.

isochores, the actual determination and identification of them at the sequence level is not a trivial task. Analysing band sequence composition using isochore distribution was more informative than just measuring the ratio of GC versus AT in the genome.

Different computational approaches have been implemented towards the characterization of the isochore distribution in the human genome [189, 190, 191, 192, 185, 193], which led to different definitions of borders between adjacent isochores. Frishman's group compared all the available approaches and generated a consensus isochore dataset that matched features of the majority of the different methods without losing relevant contributions inherent to each particular method [99]. We used this set of coordinates in order to simplify our analysis (see Methods 2.2.2).

According to the consensus from Frishman, more than half ($\sim$56%) of the genome is covered by low GC content isochore families, congruent with the observation that $\sim$50% of the total genome is classified as G-positive bands. Furthermore, the consensus isochore data classified $\sim$7% as not belonging to any of the isochore classes in parallel with $\sim$8% of the genome not categorized as a G nor R band (Figures 4.8 and 4.2). The highest GC content family, H3, only occupied $\sim$3%, a small proportion of the genome, a number similar to the portion of the genome occupied by genes.

The density of the different isochores was calculated for each type of band by measuring the occurrences of each isochore family per megabase (Figure 4.9). As expected from the GC content distribution, the darkest bands were almost devoid of H isochore families. The density of isochores is highest for gpos25, in agreement with gpos25 being the

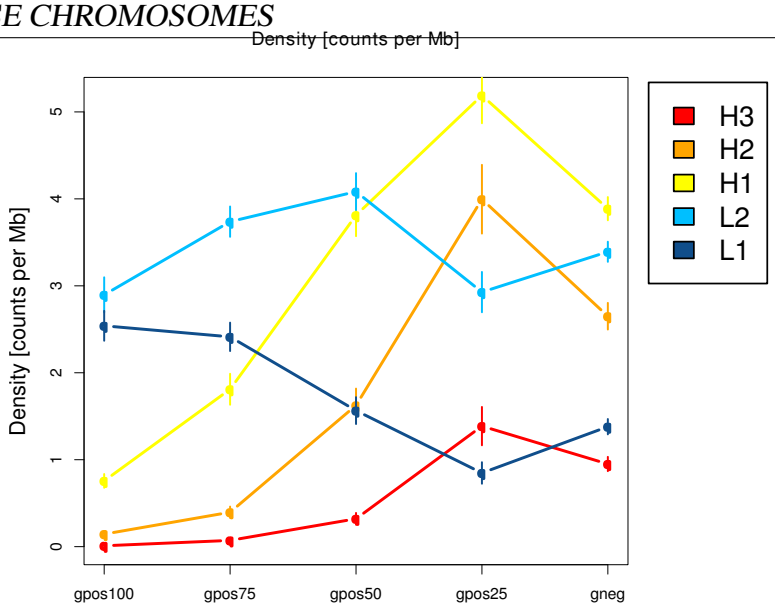


Figure 4.9: Isochores. The GC richest isochore families are more abundant in R bands and the lightest G bands.

class of bands with the highest density of genes shown in Figure 4.4 (Upper left panel). Interestingly, gpos50 showed a good balance of H and L families. Another interesting observation is the fact that, contrary to what should be expected from a decrease of L families towards light bands, gneg bands showed a higher density of L families relative to the slightly darker bands gpos25.

In order to appreciate better the compositional structure of bands as long, homogeneous GC content blocks, we measured the proportion of each class of band belonging to the different isochore families (Fig. 4.10). In this view it was easier to appreciate the differences in isochore composition of bands. Based on this measure, we found that $\sim$50% of gneg bands are divided into halves of each isochore family, similar to gpos25.

4.3.3.3 Repetitive elements

Taking into account that 50-60% of the human genome is shaped by repetitive sequences [100, 194] we extended our sequence-based analysis to understand the differential occupancy of repeats along the five Giemsa band classes. Based on the table of coordinates of repetitive elements from RepeatMasker [100] (see Methods 2.2.2) we found that $\sim$50% of the sequence covered by Giemsa bands are of a repetitive origin. Generally speaking, repetitive elements can be found in the form of transposable elements (TE) and non-TE sequences. Non-TE repeats include simple tandem and satellite repeats oc-

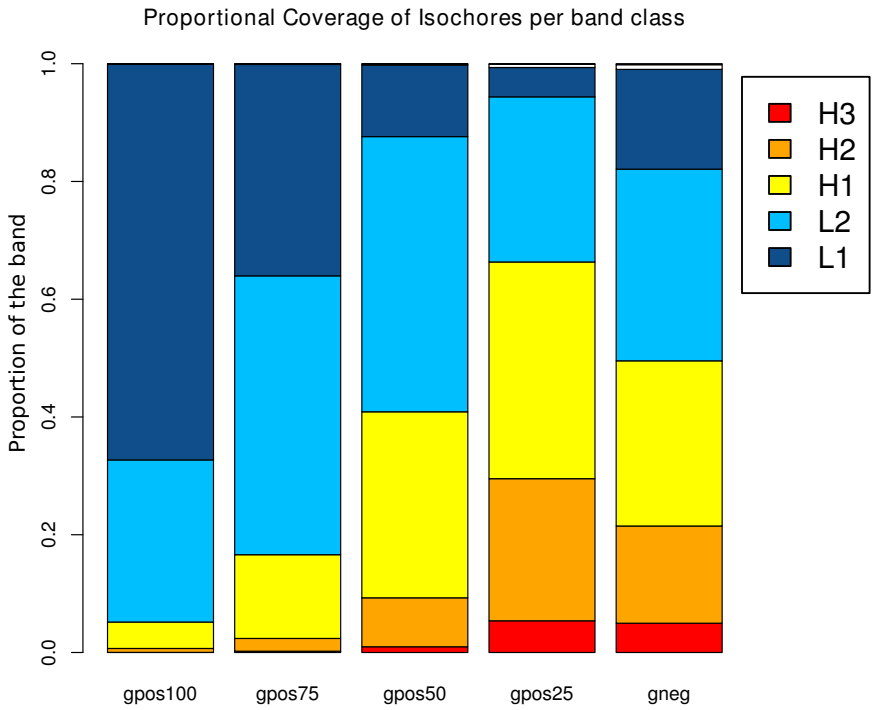


Figure 4.10: Proportional coverage of isochores per band class. There is a shift in the proportion of which families occupy the majority of each band as we progress from dark to light bands.

cupying a small proportion of the sequenced human genome, but satellite repeats occupy a large part of the unsequenced genome. TEs are more abundant than satellite repeats and show a much more complex structure in their population and behaviour. TEs can be classified as DNA transposons (3%) and retrotransposons. Retrotransposons can be found as Long Terminal Repeat (LTR) and Non-LTR elements, covering respectively about one tenth and one third of the human genome (Fig. 4.11).

To further dissect the distribution of repeat classes by band classes we measured individually the density of repetitive sequences for each repeat class in each band class. To measure the density of signal, instead of just counting the number of instances of each repeat, avoiding the problem of over counting due to fragmentation of TEs, we measured the average signal density at the megabase level. In other words, how many base pairs belong to a repetitive element per megabase. We aggregated the results in three sets: as elements that do not show any preference of occupancy either G or R bands, repetitive elements preferentially found in R bands and repetitive elements preferentially found in G bands.

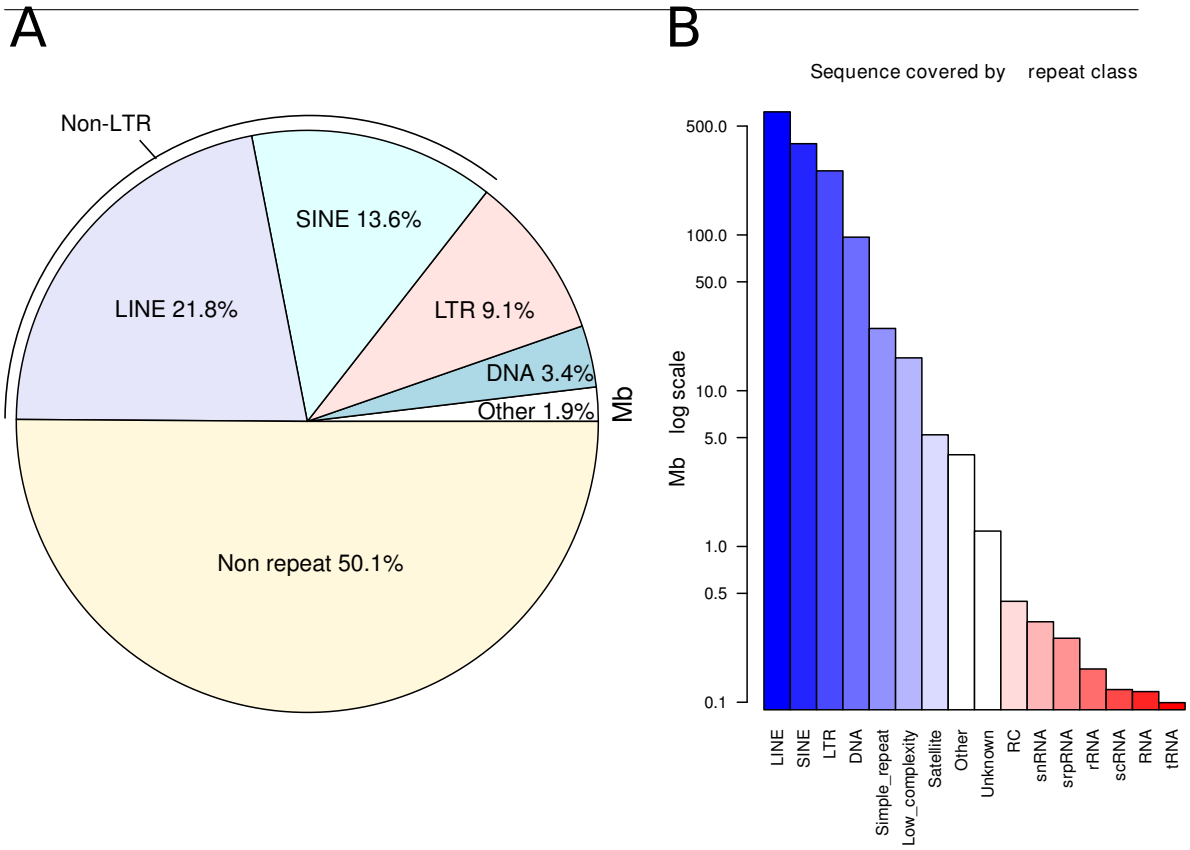


Figure 4.11: Genome coverage by repeat class. Half of the human genome is of a repetitive nature or origin. (A) Proportional coverage of each repetitive element class. (B) Total sum of megabases covered by each of the repetitive elements included in this study. Same layout as in previous box plots.

Repetitive elements preferentially occupying G bands.

The classes of repetitive elements that are highly enriched in G bands were the long interspersed nuclear elements (LINEs), LTR elements, low complexity repeats and satellite repeats, which together cover around 31.74% (896.4 Mb) of the human genome. Figure 4.12 summarizes this data. In all cases the enrichment of these repeats decreased as bands got lighter. Unexpectedly, R bands showed higher enrichment levels than the gpos25 class.

Further dissection of LINE families showed that the strongest correlation between band staining intensity and density of repetitive elements was found in the L1 family. Opposite to L2 families and CR1 that show no real preference of occupancy. Dong-R4 family was found to be similarly enriched in gpos100 and gpos75 bands, relative to the rest, which showed almost total absence of this family.

All families of LTRs, excluding the ERVK family, showed the same trend of enrich-

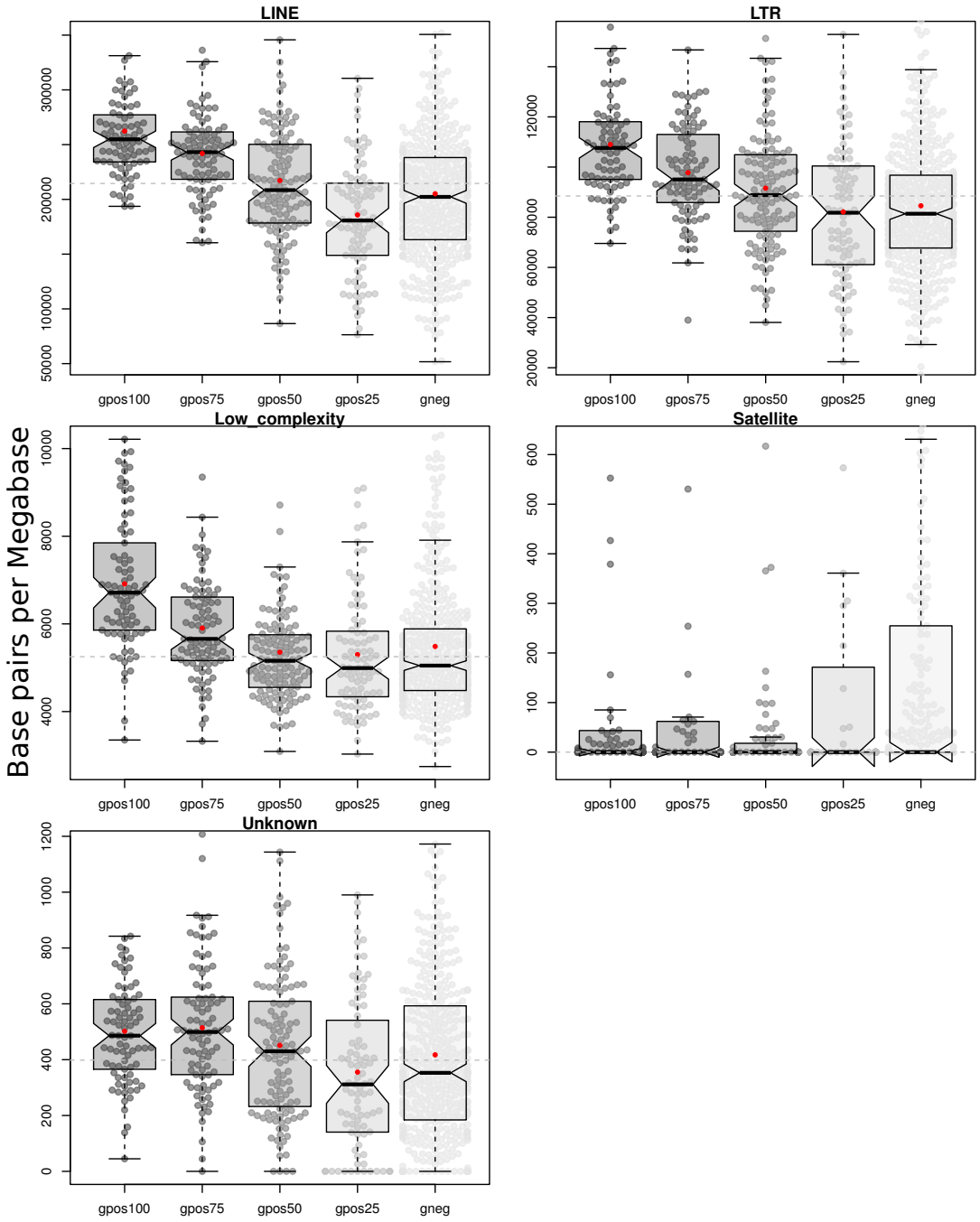


Figure 4.12: Repetitive elements enriched in G bands. Density of repetitive elements measured as base pairs of repetitive sequence per megabase. Each data point represents one band. Dashed line represent genomic median score and red dots represent mean values. Outliers were removed for clarity. Same layout as in previous boxplots.

ment in dark bands. Even though ERVK repeats were more densely present in the darkest bands, some of the gpos25 bands showed a slight enrichment.

In addition, a minor group of repetitive sequences classified as *unknown* also showed a preference to occupy dark bands instead of light bands. In total, this group covers only 1.2 Mb.

Repetitive elements enriched in R bands.

R bands were found to be highly enriched in TEs, in particular short interspersed nuclear elements (SINEs). With a weaker enrichment signal, R bands were also found to be preferentially occupied by satellite repeats and several repetitive sequences associated with different kinds of RNA molecules such as small nuclear RNAs (snRNA), ribosomal (rRNA) and small cytoplasmic RNA (scRNA). Grouping SINES and the RNA repeat classes they covered ∼13.97% (385.9 Mbs) of the human genome with a total of 1,765,107 repetitive sequences, where ∼99% (1,749,764) were copies only of Alu elements. Alu elements are 2.4× more dense in light G bands than the darkest ones.

SINEs were 1.3× more dense in gpos25 relative to the genomic overall, followed by gneg. There was more than a 2-fold difference between the SINE poorest regions in gpos100 and the richest in gpos25. When analyzing SINE families we found that DeuSINES do not follow the general trend and were found to be almost absent in R bands and light G bands. Once again, we found the unexpected result that R bands showed more a more G band-like behaviour than gpos25.

Repetitive elements found without preferential occupancy.

A very small fraction of the genome (4.43% or 122.8 Mb) is covered by repetitive elements that were not found to occupy preferentially any kind of band. This minority cohort comprises DNA transposons, signal recognition particle RNAs (srpRNA), rolling circle (RC) elements, simple repeats, RNA repeats and tRNA repeats. The distribution of each individual repeat class is shown in Figure 4.14.

4.3.4 Summary of sequence features

Figure 4.15 shows a summarizing heatmap representing all the relative signal changes across the five Giemsa band categories for all the sequence-based genomic features tested. Table 4.3.4 shows the absolute values for the different features measured in this section. The gene and CpG data was averaged from a gene-based analysis to a per-band analysis in order to fit this table and match the rest of the data. After the analysis of the genomic features that characterize each Giemsa band class at the sequence level, we observed that

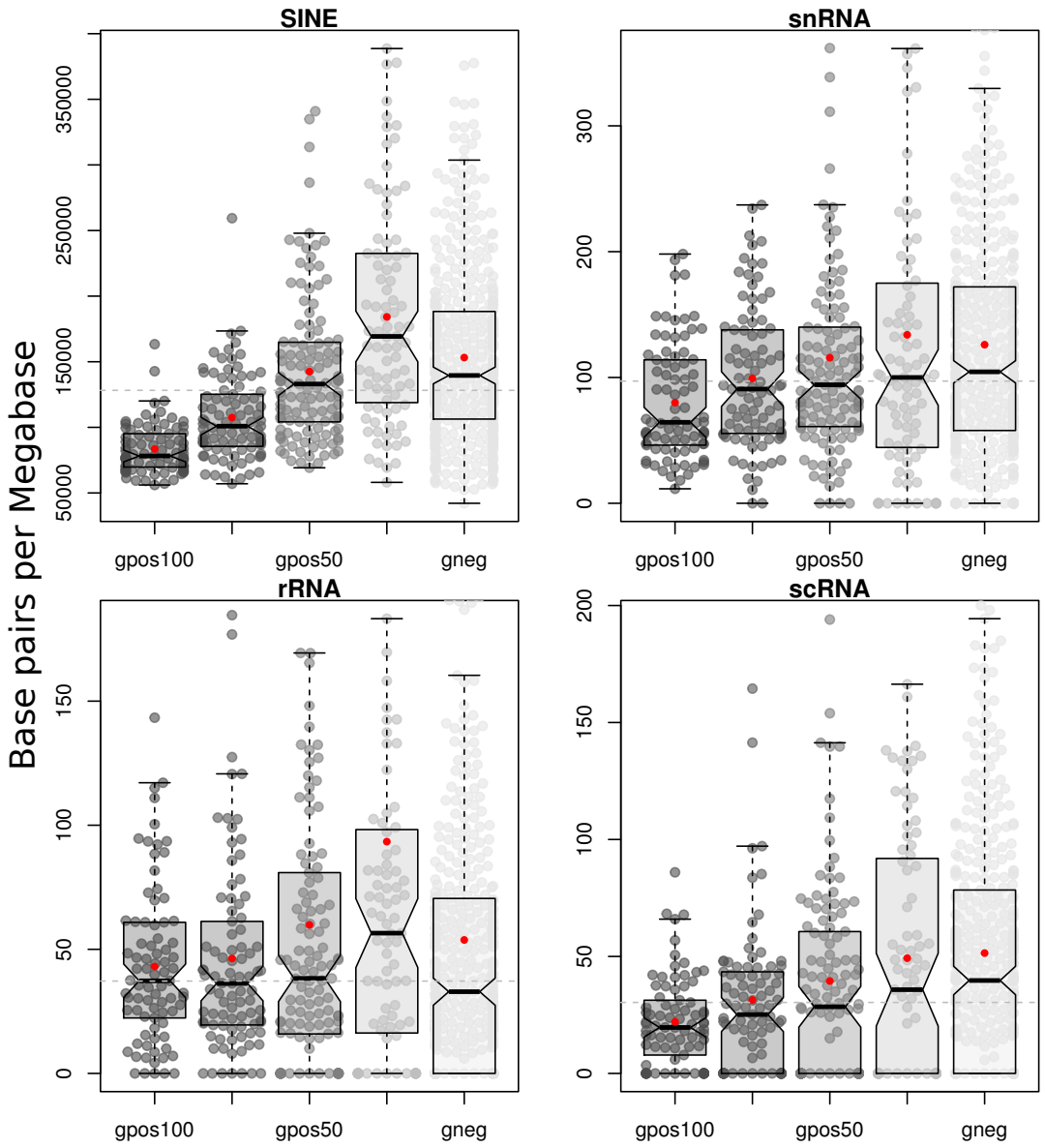


Figure 4.13: Repetitive elements enriched in R bands. Same layout as in previous box plots.

at different scales there are the banding pattern through differences between band classes. As a broad explanation, we can say that Giemsa dark bands showed low levels of G and C nucleotides and as bands get lighter the GC levels increased. This trend, manifested as well in terms of isochore families, high GC content isochore families were almost absent in Giemsa dark bands and preferentially populated light G and R bands.

As repetitive elements are usually derived from transposable elements, with well understood molecular mechanisms, they represent a cornerstone in the understanding of not only genome evolution but nuclear function and organization, specially as they represent at least half of the human genome. Interestingly, there is a clear preference for

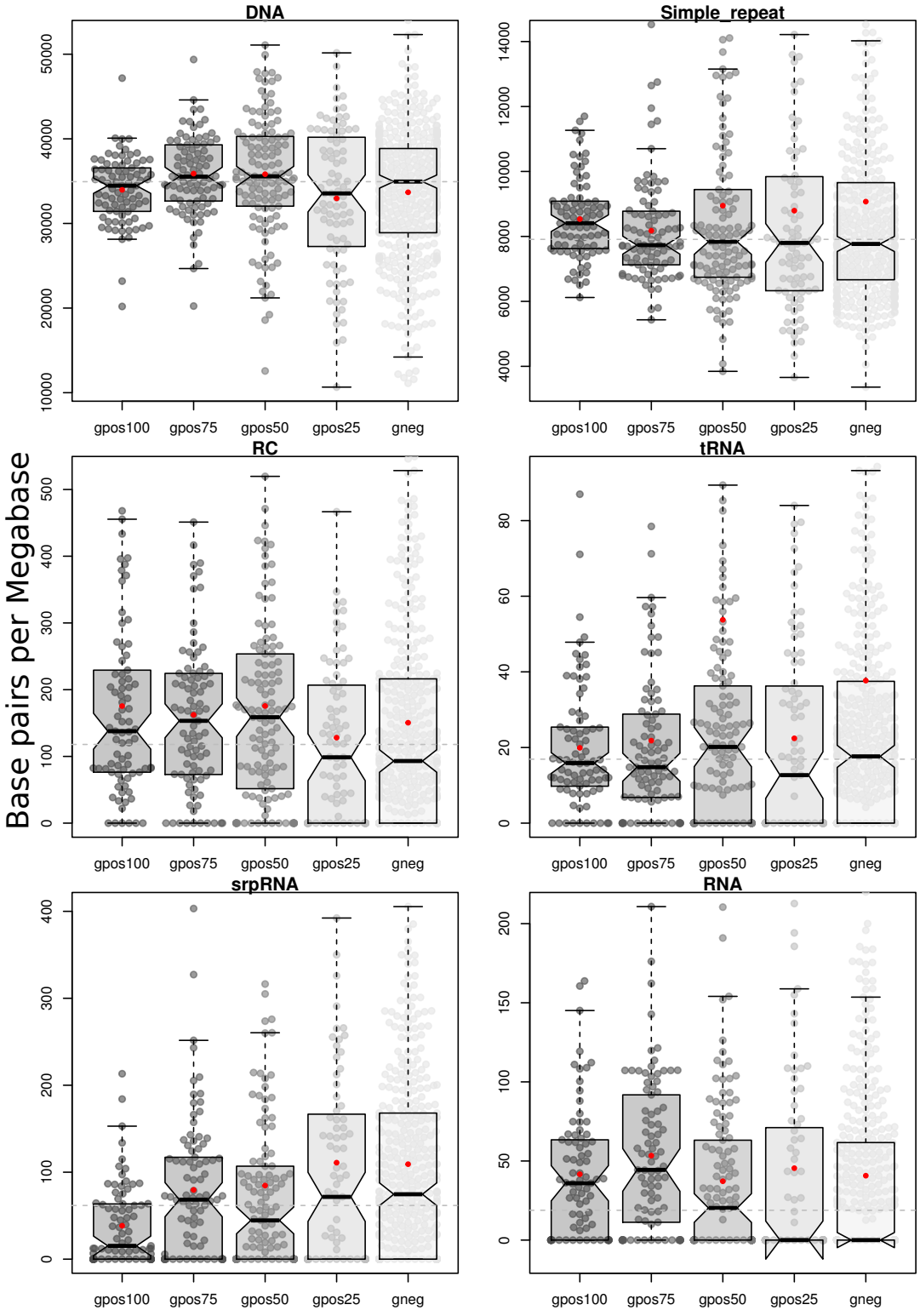


Figure 4.14: Repetitive elements that do not show any preferential occupancy in the genome. Same lay out of previous box plots.

autonomous transposable elements such as LINEs, and specially the L1 family, to occupy dark G bands in contrast to the not autonomous transposable elements, such as Alu elements, which showed a bias for inserting in R bands and light G bands. The genome is usually thought as protein coding sequence or non-coding sequence, and not coding elements usually thought as regulatory elements. A deeper understanding of the function of the genome will require the definition of 'structure-coding' regions of the genome where repetitive sequences will certainly play an important role.

Even though the genome sequence represents the platform on which biological processes are built, it does not completely determine the function of the eukaryotic nucleus or the phenotype of a cell. Epigenetic factors provide the next layer of complexity in the regulation of genomic function as they do not rely completely, and even in some times can override, the information encoded genetically. To pursue a better and more complex understanding of the underlying epigenetic characteristics that could define the banding pattern, we extended our data mining analysis to incorporate genome-wide data sets of posttranslational modifications of histone tails.

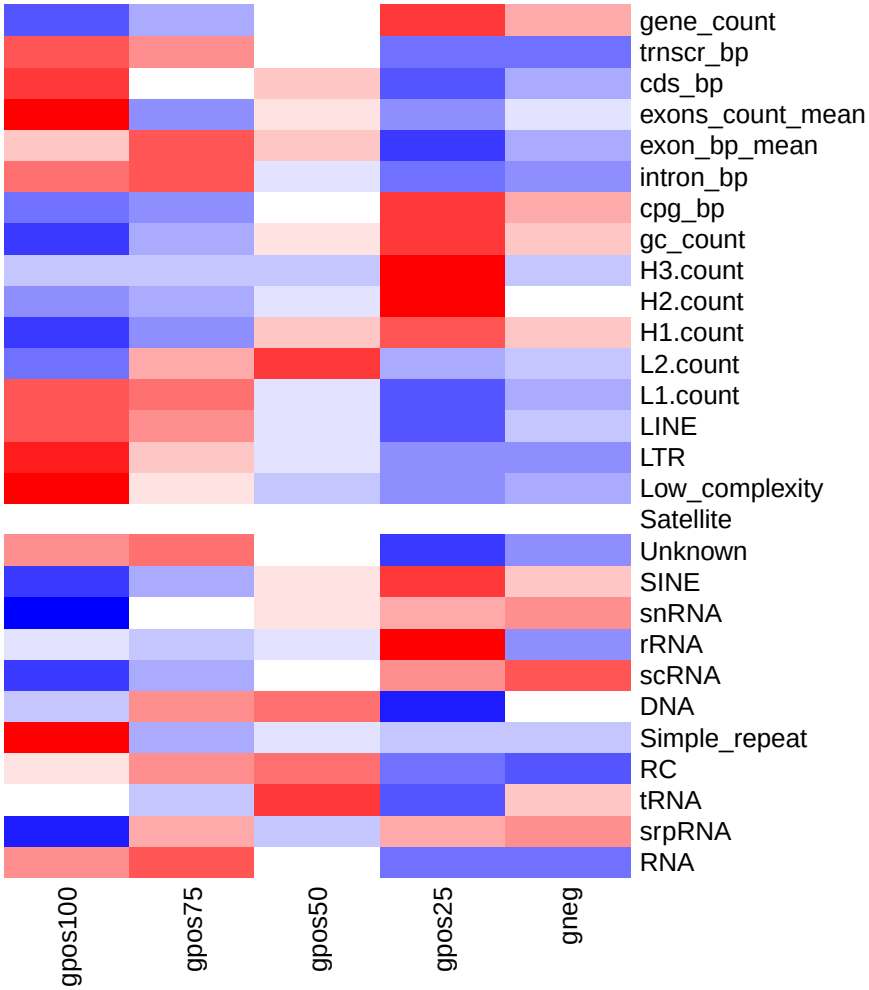


Figure 4.15: Summary of enrichment of sequence-based features. The colour code represents a transition from low (blue) to high values (red) relative to the range of values of each row; middle point in the range of values in white.

		Genomic Summary						Median per band class				
Genomic Feature	Units	Min	1st.Qu.	Median	Mean	3rd.Qu.	Max	gpos100	gpos75	gpos50	gpos25	gneg
Genes												
No. of Genes	Genes/Mbps	0	3.14	5.4	7.66	9.42	44.09	2.26	3.33	4.81	8.18	6.67
Transcript LG	Bps/Mbps	0	36960	60980	79220	96450	925700	98989.11	93111.12	71635.75	49733.04	52527.94
CDS LG	Bps/Mbps	0	2468	2864	2948	3341	11520	3075	2911.8	2959.53	2752.33	2813.19
Mean No. Exons	Exons	0	8	9.74	9.97	11.59	39.5	10.28	9.55	9.85	9.55	9.67
Mean Exon LG	Mean Bps	0	359.3	434.2	487.9	543.1	2341	449.75	462.53	447.43	410.44	427.25
Mean Intron LG	Mean Bps	0	3614	5770	7676	8383	230000	8208.67	8449.8	6189.16	4755.06	4991.4
Mean CGI LG	Mean Bps	0	2753	5132	7972	9872	69210	1843.92	2671.95	4684.38	9450	6961.02
GC content												
GC	Percent	33.27	38.13	40.36	41.11	43.59	54.37	36.76	38.39	40.71	43.53	41.55
H3 Isochores	Counts/Mbps	0	0	0	0.71	0.62	10.71	0	0	0	0.53	0
H2 Isochores	Counts/Mbps	0	0	0.62	2.13	3.64	25	0	0.16	0.67	3.5	1.35
H1 Isochores	Counts/Mbps	0	1.11	2.86	3.46	5.46	13.85	0.53	1.43	3.44	4.6	3.61
L2 Isochores	Counts/Mbps	0	1.76	3.15	3.44	4.86	13.08	2.67	3.61	4.05	2.92	3.04
L1 Isochores	Counts/Mbps	0	0.37	1.11	1.58	2.29	12.86	2.46	2.22	1.15	0.45	0.79
Repetitive elements												
LINE	Bps/Mbps	51720	175400	214500	214800	249400	658800	254846.4	243122.3	208448.8	180702.3	202269.2
LTR	Bps/Mbps	17220	71210	88470	89350	104200	315100	107623.05	95056.92	88821.71	81786.25	81385.11
Low Complexity	Bps/Mbps	740	4610	5252	5638	6239	27140	6713.51	5657.92	5158.85	4992.31	5048.49
Satellite Rep.	Bps/Mbps	0	0	0	1900	120.9	265900	0	0	0	0	0
Unknown Rep.	Bps/Mbps	0	220.5	398.3	435.2	608.1	1817	485.93	499.23	430	311.43	352.87
SINE	Bps/Mbps	42130	92980	128200	142700	173800	388700	78164.88	100905.43	133016.36	169290	139541.19
snRNA	Bps/Mbps	0	54.46	97.11	117.6	157	994.6	64.38	90.81	94.19	100	104.49
rRNA	Bps/Mbps	0	12.67	37.21	57.12	71.91	1893	37.3	36.25	38.29	56.52	32.93
scRNA	Bps/Mbps	0	0	30.34	44.11	62.95	390	19.64	25.15	28.48	35.79	39.72
DNA	Bps/Mbps	7714	30220	34940	34210	38930	60190	34458.59	35529.81	35567.06	33550	34947.98
Simple Rep.	Bps/Mbps	3360	6822	7909	8866	9445	71250	8406.78	7730.96	7835.96	7799.38	7768.07
RC	Bps/Mbps	0	29.43	117.9	155.7	223	1862	137.76	153.33	158.75	98.64	93.08
tRNA	Bps/Mbps	0	0	16.91	34.87	34.21	3770	15.88	14.79	20.14	12.69	17.65
srpRNA	Bps/Mbps	0	0	61.72	95.08	143.9	968	15.29	68.41	44.57	71.63	74.63
RNA	Bps/Mbps	0	0	18.85	42.21	66.78	470	35.93	44.37	20.28	0	0

Table 4.1: Summary of enrichment of sequence-based features. LG = length.

4.4 Chromatin features

Epigenetic changes are at the core of nuclear processes like gene regulation and gene expression. They are also involved in the higher-order organization of chromatin structure. These changes manifest at two main levels: methylation of DNA and posttranslational modification of histones. Our data mining analysis focuses on the second level.

For some decades, epigenetic analysis of the cell could only be performed at a single locus of interest because the molecular biology techniques used could only offer information of a limited number of pre-designed PCR targets. With the advent of high-throughput screening technologies, such as microarrays and second generation sequencing, the whole genome can now be surveyed and quantification of the spread of a large collection of post-translational histone modifications can be performed.

When analyzing the human genome at the sequence level, as in the previous section, the differences that determine one cell-type from the rest cannot be detected as, it can be assumed, all cells have essentially the same genome. In this sense, the cell-type from which the genome sequence was determined will not introduce any bias into our analysis. This cannot be said for epigenetic data since the differences at the epigenetic level can occur between two or more different cell-types. Taking this into account, we made sure that when comparing the cytogenetic map to chromatin data, the epigenetic data matched the same cell-type or, at least, both belonged to the same cell lineage. Cytogeneticists use peripheral blood lymphocytes as a reference for the standard cytogenetic map because of the ease acquiring of large numbers of cells, their ability to proliferate, their cell synchronization efficiency and the quality of chromosome spreads [169]. For this reason we used the data from Barski et al. [195] who performed genome-wide chromatin immunoprecipitation (ChIP) in human $CD4^+$ T lymphocytes of 20 histone lysine and arginine methylations in addition to different chromatin related factors: the histone variant H2A.Z, RNA polymerase II and the insulator binding protein CTCF.

To simplify the visualization of the data and given that the units for measuring ChIP-seq data are more arbitrary, we switched from the box plot representation to a colour coded heat maps and tables. Figure 4.16 visually summarises the distribution and relative enrichment of the different chromatin marks evaluated.

4.4.1 Chromatin modifications enriched in R bands

Generally speaking, open chromatin marks were found to be enriched in R bands, but in most cases, gpos25 showed the highest values of open chromatin marks. Unexpectedly, some marks associated with gene silencing were also found to be enriched in these two classes of bands. The reason for this phenomenon is that some gene-silencing marks (such as H3K79me3) operate at the level of gene promoters, therefore gene desert regions cannot incorporate promoter-specific chromatin marks.

Open chromatin marks are enriched in regions where RNA synthesis is occurring while genes are being transcribed. As expected from the distribution of genes and their high-density of occurrence in gpos25 and gneg bands, these band classes showed the highest levels of transcriptional activity as measured by RNA Pol II occupancy. Relative enrichment of RNA polymerase II (PolII) binding was respectively 1.31× and 1.22× higher in gpos25 and gneg than the overall genomic score and 1.89× and 1.75× higher than the gene-poorest regions.

Several other chromatin modifications and elements, such as insulators, are associated with gene transcription, they are thought to provide a molecular milieu that facilitates access to the different regulating elements, present a more flexible, open chromatin structure, or even mediate and stabilize long-range interaction and the general topological distribution of chromatin domains. CTCF is the best known insulator protein and it is known to be enriched in euchromatic, gene-active regions. Our analysis confirmed this distribution, as gene-rich and transcriptionally active regions showed the strongest enrichment signal in gpos25, followed by gneg and then decreasing linearly as bands got darker, with gpos100 being the CTCF-poorest class.

The following section will describe the distribution of individual active chromatin marks relative to Giemsa bands. The order in which the marks appear in the text follows a descending order of the strength in the signal found for each mark. The first marks showed the strongest preference for the lightest bands and the last ones showed the strongest signal towards darkest bands. The order of the rows in the heatmap in Fig. 4.16 follows the same order.

Table 4.2 at the end of this section shows the fold changes of each band class relative

to the genomic median. The reason for reporting data in this format is that ChIP-seq experiments do not have any specific unit of measurement and the raw signal of each track is not necessarily equivalent between different ChIP-seq experiments.

Methylation of Histone 3, lysine 4 (H3K4)

H3K4 is associated with activated genes. The three methylation states of H3K4 (H3K4me1,H3K4me2 and H3K4me3) are enriched in the regions surrounding TSS. Levels of H3K4me3 are positively correlated with gene expression [195]. H3K4me1 is enriched outside of promoter regions that interact with functional enhancers in different types of cells [196, 197, 198]. H3K4me2 is known to lead chromatin decondensation and to drive the formation of loops that separate active chromatin from heterochromatic bulks [199]. Interestingly, even though the bulk of evidence points methylations in lysine 4 of histone 3 as an active mark, H3K4me3 has been found to colocalise with, the apparently antagonistic mark H3K27me3, into domains known as "bivalent" in embryonic stem cells that provide the cell with plasticity upon differentiation [200]. In our analysis, all these three marks were found to be highly enriched in gneg and, specially, gpos25 bands.

Monomethylation of Histone 3, lysine 9 (H3K9me1)

This chromatin modification is linked to gene activation and has been found to correlate with H4K20me1 and correlates with high levels of gene expression, showing a preference towards the 5' end of genes [195]. It has been found in insulators and hypersensitive sites (HSs). The band class that showed the highest enrichment of this mark was gpos25, followed by gneg.

Monomethylation of Histone 3, arginine 2 (H3R2me1)

H3R2me1 even though this mark was not found to be in active, or inactive promoters Barski et al. [195], in yeast correlates with gene transcription as seen in [201]. It is catalyzed by CARM1, an enzyme involved in activation of nuclear hormone receptor-mediated transcription [202]. Our analysis showed that this mark is preferentially found in gpos25 and gneg bands.

Methylation of Histone 3, lysine 79 (H3K79)

Histones with this covalent posttranslational modification in its monomethylated form (H3K79me1) are localized in active promoters. On the contrary, the trimethylated form of this modification (H3K79me3), is enriched in silent gene promoters. The dimethylated form of this mark (H3K79me2) does not show any preference for occupying gene promoters, in terms of gene transcription [195]. The enrichment signal distribution of all these three forms of H3K79 followed the same trend of occupancy along bands that genes showed. As mentioned before, even though H3K79me3 is not an active chromatin mark in human, as it is in yeast [203], it is a promoter-specific silencing modification therefore it cannot be found independently of genes. In our analysis these marks showed a strong preference for occupying gpos25 and R bands, with gpos25 the most enriched. H3K79me2 has been shown to be associate with gene transcription activity [204, 203, 205].

Monomethylation of Histone 2B, lysine 5 (H2BK5)

This poorly known chromatin modification was analysed for the first time in the Barski data set [195]. It was found to be associated downstream of active promoters. In our analysis it showed a strong preference for occupying gpos25 and R bands, with gpos25 the most enriched in this chromatin mark.

Monomethylation of Histone 4, lysine 20 (H4K20me1)

Correlated with H3K9me1, this chromatin modification is found downstream of active promoters and follows the same trend of all the previous gene-associated chromatin modifications, with gpos25 being the most enriched class.

Mono and trimethylation of Histone 3, lysine 36 (H3K36me1, H3K36me3)

These chromatin marks are associated with regions downstream of TSSs, nevertheless H3K36me1 shows only a slight preference for active promoters [195] while H3K36me3 positively correlates with active genes [206]. We found that these two marks were also enriched in gpos25 and gneg bands, nevertheless the trend was slightly less strong than

the other chromatin marks reported above.

Histone H2A variant H2A.Z

This highly conserved histone variant shows associations with regulatory elements at DNase I hypersensitive sites and insulators [195]. H2A.Z is known to be associated with promotes of active genes [207] and genes that are repressed by the polycomb complex [208]. We found that the distribution of the histone variant H2A.Z showed a negative correlation with staining intensity, the darkest bands showed low levels of signal that increased linearly as bands got lighter. The bands showing the highest enrichment of the histone variant H2A.Z were gpos25 and gneg, both showing the same levels of enrichment.

Monomethylation of Histone 3, lysine 27 (H3K27me1)

This mark is one that typically defines euchromatin domains. It is related to active promoters downstream of TSSs. We found that in the darkest bands (gpos100, gpos75, and gpos50) this chromatin modification showed the same levels of enrichment as each other. The three of them below the genomic median. Gpos25 and gneg bands were the most enriched and both showed the similar enrichment values.

4.4.2 Chromatin modifications without specific distribution along bands.

Dimethylation of Histone 4, arginine 2 (H4R3me2)

We detected a slight increase in the occupancy of this chromatin mark in gpos25 but besides this, all band classes showed similar levels of H4R3me2.

Bi and trimethylation of Histone 3, lysine 27 (H3K27me3, H3K27me2)

In general, H3K27 methylation is related to gene repression. Similar to the related modification H3K27me1, the H3K27me3 showed a preference for light bands, with very low levels in gpos100 bands. The other bands showed similar trends, nevertheless this time the most highly occupied band classes for this chromatin modification were gpos50

and gpos25. Even though this mark is not particularly enriched in light bands, it is clearly underrepresented in the darkest class of bands.

H3K27me2 showed the same levels of occupation in all band classes, with only a minor increase in the signal in gpos50.

4.4.3 Chromatin modifications preferentially found in G bands

Dimethylation of Histone 3, arginine 2 (H3R2me2)

The original screening of H3R2me2 did not report any particular localization of this chromatin mark, nevertheless, and in agreement with other studies [209, 210], we found that H3R2me2 constitutes a repressive mark of heterochromatin. Most band classes showed the same levels of H3R2me2 without including the gpos100 class, which showed clear enrichment of this mark relative to the genomic median score.

Trimethylation of Histone 4, lysine 20 (H4K20me3)

In contrast to the related posttranslational modification H4K20me1, H4K20me3 is a chromatin mark previously reported as associated with heterochromatin. Out results confirmed that this chromatin modification follows the opposite trend of the euchromatic marks and the band class most enriched with it was gpos100, followed by the less dark bands gpos75 and gpos50. Only gpos100 and gpos75 median scores are above the overall genomic median. The class of band that showed the weakest signal for H4K20me3 was gpos25.

Di and Trimethylation of Histone 3, lysine 9 (H3K9me2, H3K9me3)

This pair of chromatin modifications, known to be characteristic of transcriptional repression [211, 212], showed strong signal in the darkest bands and became increasingly weak as bands got lighter. For both marks gpos100 showed the strongest signal and for gpos25 the weakest.

4.4.4 Summary of chromatin modifications

Our analysis of genome-wide data sets of the location and enrichment levels of post-translational modification of histone tails and other chromatin factors, showed that euchromatic chromatin marks were enriched in the lighter band classes, in particular gpos25 bands. Transcriptional activity also showed strong preference for light bands, as the presence of RNA PolII was highly enriched in these classes. As efficient gene expression requires an appropriate topological environment it is reasonable to have found the insulator factor CTCF and the histone variant H2A.Z, which are known to be associated with regulatory regions, also enriched in the light classes of bands.

On the other hand, Giemsa dye intensity was positively correlated to constitutive heterochromatin marks such as H3K9me2, H3K9me3 and H4K20me3, plus the lesser known histone modification H3R2me2. Heterochromatic marks signal dropped as chromosomal bands got lighter. Nevertheless, some gene-silencing marks, such as H3K79me3, were found to be enriched in gpos25 and R bands.

At this point we were interested in making sense of the combination of chromatin modifications and their biological functions as a whole. As cells undergo drastic global changes in their epigenetic makeup in order to adjust to cell-lineage commitments. Chromosome organization and structure are also affected during the lineage commitment process. For reasons that are not very well understood, epigenetic changes coupled with architectural changes have been shown to heavily influence the temporal dynamics with which the cell replicates its DNA during S phase. As replication timing reflects the aggregate epigenetic and structural changes of the cell, it represents the interface where epigenetics and higher-order chromatin dynamics meet. Using replication timing as a surrogate of the overall epigenetic status of chromatin, we extended our analysis to assess the differences in replication timing across band classes. Our results are consistent with the distribution of chromatin states in *D. melanogaster* [213] where 'black' chromatin covers half of the genome and is characterized by close chromatin marks and tissue specific genes, red and yellow states would correspond to gneg and gpos25 with open chromatin marks and early replication timing.

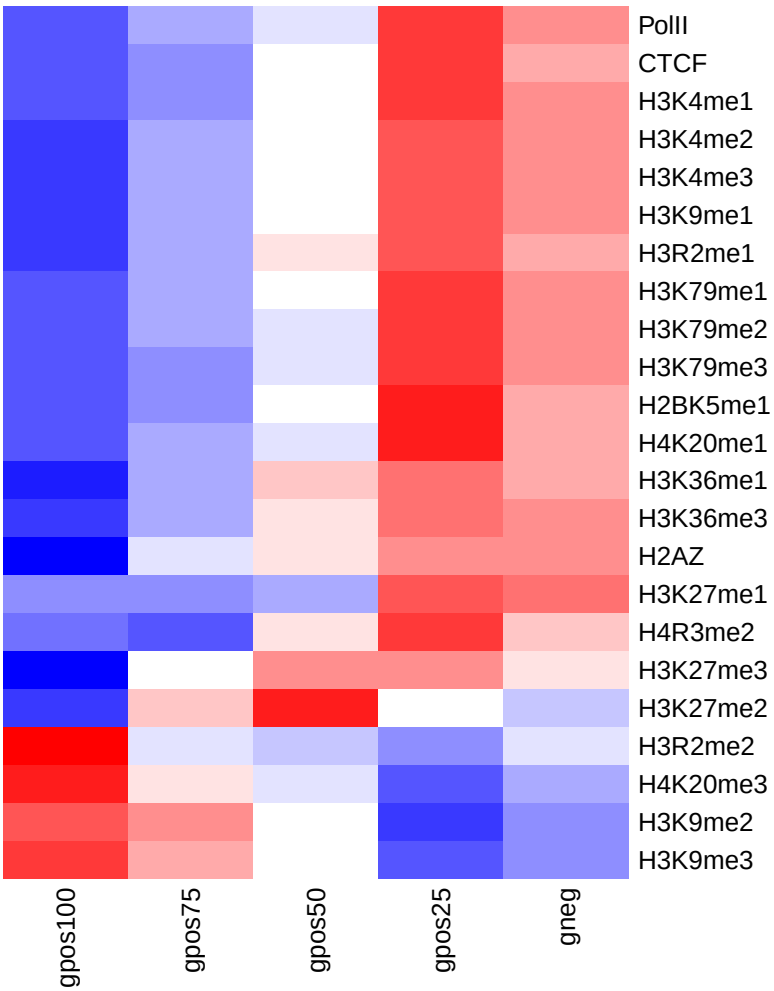


Figure 4.16: Distribution of post-translational modification of histones and other chromatin factors across bands. The colour code is the same as in Fig. 4.15.

4.5 Replication timing features

DNA replication in mammalian cells is a very complex nuclear process regulated tightly in space and time. Given the huge size of the genome, the cell has to replicate its DNA in parallel from numerous origins, so that, different regions of the genome undergo replication at the same time [61]. Nevertheless, some sectors of the genome are replicated earlier than others. This order in which the synthesis of DNA happens along the genome at different times is known as the replication timing programme or S phase programme [69, 70].

Similar to the acquisition of chromatin data, replication timing data now benefits from high-throughput technologies. There are different approaches to generate genome-wide profiles to express time of replication as a function of chromosomal position in different

Mark	Genomic Summary						Fold change relative to the Genomic Median				
	Min	1st.Qu.	Median	Mean	3rd.Qu.	Max	gpos100	gpos75	gpos50	gpos25	gneg
PolII	0	0.73	1	1.17	1.47	4.08	0.69	0.82	0.9	1.31	1.21
CTCF	0	0.83	1	1.06	1.23	3.18	0.83	0.89	0.98	1.17	1.08
H3K4me1	0	0.55	1	1.22	1.66	7.55	0.4	0.59	0.87	1.57	1.29
H3K4me2	0	0.61	1	1.08	1.48	4.47	0.45	0.68	0.89	1.38	1.26
H3K4me3	0	0.76	1	1.03	1.25	2.97	0.66	0.8	0.93	1.18	1.12
H3K9me1	0	0.63	1	1.07	1.45	4.41	0.48	0.69	0.93	1.35	1.24
H3R2me1	0	0.85	1	1.01	1.18	1.9	0.79	0.88	1.01	1.13	1.07
H3K79me1	0	0.52	1	1.24	1.74	5.4	0.39	0.66	0.91	1.62	1.34
H3K79me2	0	0.47	1	1.27	1.8	7.71	0.43	0.64	0.83	1.51	1.34
H3K79me3	0	0.54	1	1.27	1.76	8.16	0.49	0.65	0.89	1.59	1.34
H2BK5me1	0	0.61	1	1.26	1.59	8.57	0.49	0.66	0.93	1.62	1.23
H4K20me1	0	0.53	1	1.52	1.85	16.98	0.37	0.61	0.9	1.82	1.34
H3K36me1	0	0.85	1	0.97	1.13	2.16	0.81	0.92	1.03	1.08	1.04
H3K36me3	0	0.73	1	1.02	1.26	4.74	0.62	0.79	0.99	1.2	1.14
H2AZ	0	0.83	1	0.98	1.17	1.76	0.77	0.93	1.01	1.08	1.07
H3K27me1	0	0.88	1	1.04	1.21	2.75	0.92	0.92	0.94	1.1	1.09
H4R3me2	0	0.94	1	0.99	1.08	1.79	0.98	0.98	1	1.03	1.01
H3K27me3	0	0.79	1	0.98	1.19	3.73	0.79	0.98	1.1	1.09	1.02
H3K27me2	0	0.9	1	1	1.14	2.13	0.97	1.02	1.05	1.01	0.99
H3R2me2	0	0.95	1	0.98	1.05	1.66	1.02	1	1	0.99	1
H4K20me3	0	0.86	1	1.25	1.22	33.54	1.22	1.07	1.01	0.9	0.96
H3K9me2	0	0.71	1	0.97	1.23	2.02	1.27	1.22	1.08	0.79	0.91
H3K9me3	0	0.72	1	1.04	1.31	4.01	1.5	1.28	1.08	0.78	0.89

Table 4.2: Summary of chromatin modifications across band classes. As ChIP-seq experiments do not have an absolute scale for measurement, it is just raw signal of reads, each chromatin modification has been normalized to the genomic median value for each specific track. This way each band class shows the fold change in signal relative to the general genomic read signal strength.

cell types. These approaches differ by the quantification strategy of the time of replication and the high-throughput platform. In general, all these strategies show strikingly similar results. The next sub-section will briefly describe the most important datasets, the cell types they cover and what is the added value of each against the others.

Early/Late S phase ratios [76]

This study relies on unsynchronized cell populations sorted by DNA content and FACS, separating S phase in two halves; early S and late S fractions. Each sample is labelled and co-hybridized with high-resolution tilling arrays. The log2 (early/late) ratio is interpreted as time of replication, where the highest and lowest values are the earliest and latest replication times, respectively.

G1/S phase ratios and Gaussian convolution namely TimeEX-seq [103]

This was the first attempt to generate a time of replication profile by using second generation sequencing platforms (Illumina's GA and SOLiD). The rationale behind this approach is based on the premise that the sequences that are replicated at the onset of S

Mark	Genomic Summary						Median per band class				
Cell type	Min	1st.Qu.	Median	Mean	3rd.Qu.	Max	gpos100	gpos75	gpos50	gpos25	gneg
Erythroids *	0	0.39	0.61	0.57	0.78	1	0.29	0.4	0.58	0.75	0.69
ES *	0	0.26	0.42	0.44	0.61	1	0.18	0.28	0.41	0.62	0.5
BG01 ESC	0	0.26	0.42	0.44	0.59	0.98	0.2	0.3	0.41	0.59	0.49
BG01 NPC	0.02	0.31	0.53	0.52	0.73	0.98	0.23	0.35	0.5	0.71	0.6
Lymphoblastoid	0.03	0.24	0.43	0.45	0.64	0.98	0.21	0.28	0.37	0.63	0.53
H7 ESC	0	0.29	0.45	0.47	0.64	1	0.24	0.32	0.44	0.62	0.52
H9 ESC	0	0.21	0.39	0.41	0.58	1	0.18	0.25	0.39	0.58	0.46
iPS4 hiPSC	0.02	0.25	0.41	0.44	0.61	0.99	0.2	0.31	0.4	0.61	0.48
iPS5 hiPSC	0.02	0.23	0.37	0.42	0.59	0.96	0.19	0.27	0.37	0.61	0.47

Table 4.3: Summary of replication timing changes across band types. The scale with which time of replication is measured goes from 0 to 1. The earliest replication times correspond to values of 1 and close to 1. Late replication timing is represented by values close or equal to 0. * = Data corresponding to the TimEx method from [103]. The rest of the data sets correspond to [76] and [102].

phase will be overrepresented on a cell population. Thus by comparing the signal from G1 sorted cells and S phase cells, copy number variation will reflect time of replication of the sequences, where higher copy number means early replication and vice versa.

Analysis

Using publicly available datasets from the two different methods mentioned above, we compared the average replication timing of the five band classes across eight different cell types. The cell types analysed included three different types of embryonic stem cells (ESC), two types of human induced pluripotent stem cells (hiPSC) and two cell types from the erythroid lineage.

The heat map in Figure 4.17 confirms that for all the cell types and methods analysed, dark G bands replicate latest in S phase. However, gpos25, and not gneg bands, showed the earliest replication timing, consistent with the results of the previous sections regarding gene density and histone modifications. It is notable that regardless of the cell type and method used, each band class showed almost the same average time of replication the summarized in Table 4.3.

To visually compare the similarities between the replication timing profiles and the Giemsa banding pattern we transformed the replication timing data and plotted it as a heatmap. The heatmap represents the replication timing of chromosome 1 into different intensities along a gray scale. Early replication timing values were assigned a light gray colour and late replication timing values a dark one.

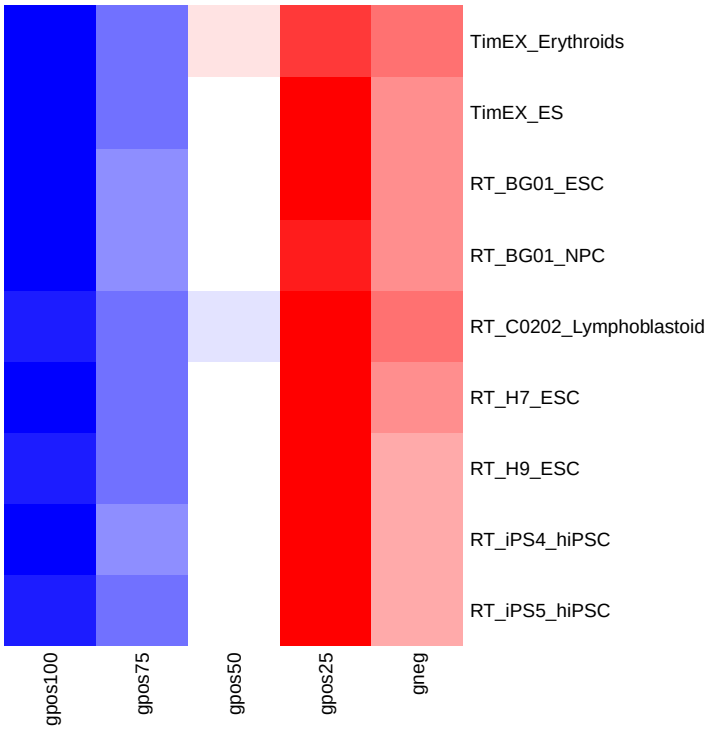


Figure 4.17: Replication timing profiles across bands for different cell lines. The temporal order in which the genome undergoes replication shows a striking correlation with the banding pattern. Early replication regions are found in R bands and gpos25 bands. This result is in contradiction with the general belief that R bands replicate strictly before G bands.

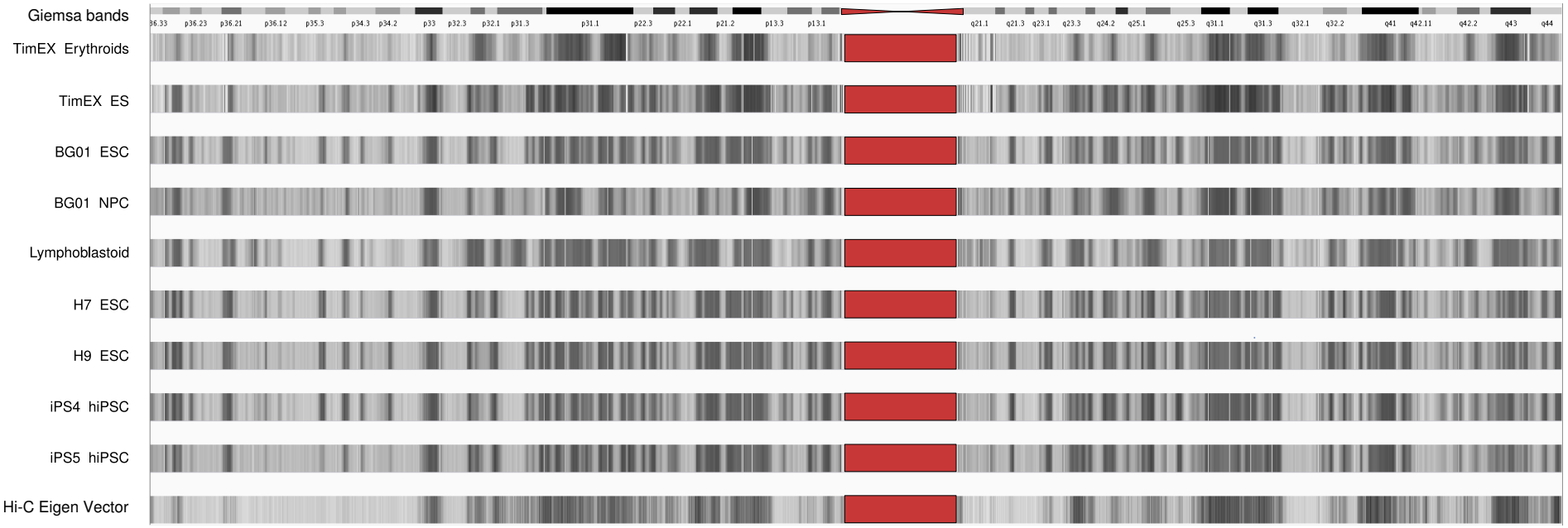


Figure 4.18: *In silico* chromosome banding. By representing the replication timing range of values as a gradient of gray tones along chromosome one, the Giemsa banding pattern of chromosomes was partially reconstructed. Cell types closer to the erythroid lineage showed more resemblance to the original Giemsa patter. The bottom track represents data from Hi-C experiments showing a strong similarity against replication timing profiles.

It has been reported previously that replication timing profiles show a striking correlation with genome-wide chromatin interaction maps based on high-throughput analysis of chromosome conformation capture libraries [80, 76] (see bottom track in 4.18). The equivalence of replication timing with chromatin interactions would imply that the same trends regarding replication timing would be conserved when analyzing higher-order nuclear organization data. For this reason, in addition of extending the band analysis to chromatin interaction maps, we studied several nuclear organization-related features such as genome-wide datasets of lamin-associated [108] and nucleolus-associated domains [107], chromatin accessibility [106] and chromatin compaction estimated by comparing the cytogenetic map to the cytogenomic map.

4.6 Higher-order organization of Giemsa bands

4.6.1 Chromatin compaction

4.6.1.1 Differential compaction of bands in metaphase chromosomes

Before the human genome sequence was available, the length of each chromosomal band was estimated based on the actual physical size observed under the microscope and compared, as a proportion, to the sum of band lengths for the corresponding chromosome arm. Given that the cytogenetic map was the only reference to the human genome, this strategy to measure bands sizes could not be verified independently. In addition, it did not allow a direct comparison of the length of two bands if they were located in different chromosome arms or allow estimation of the total amount of DNA contained in each band.

With the human genome sequence now available and the band border coordinates estimated by Furey et al. [97], a more precise estimation of the real length of bands in genomic space is possible. We transformed the genomic linear-length of each band (in base pairs) into proportional lengths based on the length of the arm in which they are located (Fig. 4.19 A). This allows a direct comparison with the relative sizes of bands from the cytogenetic map.

Using the hg18 assembly of the human genome, we compared the proportional sizes of the *cytogenetic* map [104] to the proportional sizes of the *cytogenomic* map and determined a compaction index based on their ratio (ideogram size/genomic size), as seen in panel B of Figure 4.19 for chromosome 1.

We observed that there was a different level of DNA compaction within each band class, with the darkest bands containing relatively more DNA on average when compared to R bands (gneg) and light G bands (gpos25), as previously reported for the analysis based on the draft sequence of the human genome [97].

R bands and gpos25 bands in the genome were shown to be, on average, 2.5% and 2.4% smaller than in the ideogram. The opposite trend was found for gpos50, gpos75 and gpos100 which where shown to be, respectively, 7.5%, 11% and 19.6% longer in the linear genomic sequence than seen under the microscope (Figure 4.19; panel C). In other words, the total amount of DNA that the darkest bands contain is larger than the total

DNA light bands hold. These results suggest two models, one suggesting that each class of band undergoes different levels of chromatin compaction when chromosomes condense for cell division. The second model would state that the difference in condensation of the chromatin covered by each band class pre-exists in the interphase nucleus. FISH experiments measuring the distance between one pair of probes for each band type [178] and micro-array data of open chromatin fibers [214] support the latter model. The question of whether R bands and G bands are condensed in different ways during mitosis remains open.

The diagram of chromosome 1 shown in Figure 4.19 panel B allows the direct visualization of the different degrees of compaction measured for individual bands. Values above the zero-line indicate bands that look smaller in the ideogram than in the genomic map, and vice versa for bands below the line. This view allows us to show that while at the genome-wide level (Figure 4.19; panel C) there is a clear specific compaction index for each band class, not all the bands followed this trend. Notably, it is seems that the smallest bands tend to be influenced by their neighbouring band, as seen in the small cluster of light G bands downstream of the centromere in Figure 4.19 panel B.

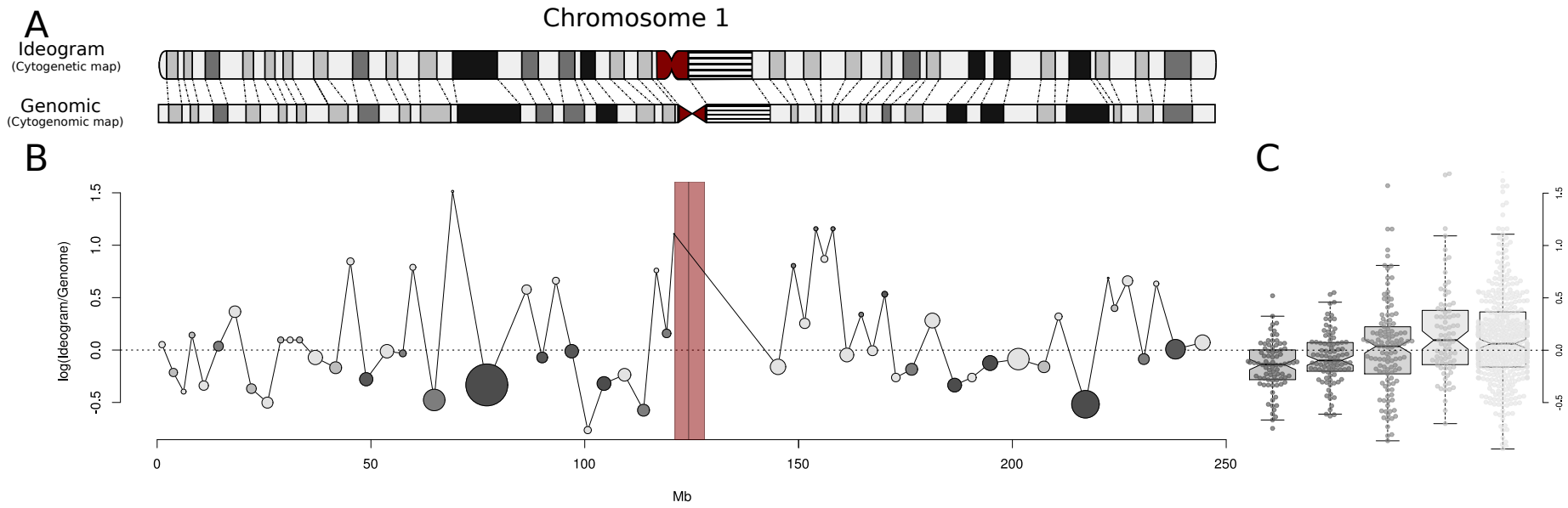


Figure 4.19: Comparison of cytogenetic and cytogenomic map. Panel A, Relative sizes of bands observed in cytological preparations of chromosomes (cytogenetic map; top) and in the human genome (cytogenomic map; bottom). Panel B, Compaction index for chromosome 1. The different kinds of chromatin present on each class of band show specific compaction indices. Each ban is represented as a circle with a diameter proportional to the length of the band. Values with a log compaction index smaller than zero look smaller in the cytogenetic map than in the cytogenomic; vice versa for bands above zero. Panel C, R bands and light G bands tend to look proportionally larger in chromosome ideograms than their respective ones in the linear coordinates of the human genome. Dark bands seem to contain a more compact conformation as their size in the cytogenomic map tends to be significantly larger than in the cytogenetic map. The trend is typically not followed by small bands, suggesting that small bands are more sensitive to their surrounding context. For instance, the small light G bands downstream of the centromere around the 150 Mb mark.

4.6.1.2 DNaseI hypersensitivity sites

As the different compaction indices found for each class of band could represent a phenomenon specific to chromosome condensation prior to mitosis, in order to measure the degree of chromatin accessibility in the interphase nucleus, we analysed genome-wide data from DNaseI hypersensitivity site assays based on the digital DNaseI method [106] and deep sequencing. The screening of DNaseI hypersensitivity was performed in 48 different cell types summarized in Table 4.4.

The presence of PolII is strong evidence supporting gene transcription and hence chromatin accessibility. Based on our previous analysis of gene density and PolII occupation, we expected to find gpos25 and gneg bands with the highest levels of DNaseI digestion signal. This is exactly what we found, for all the 48 different cell types measured, gpos25 and gneg bands showed the highest levels of DNaseI digestion signal. Even though we were most interested in the cell types closer to the lymphocyte lineage (such as Gm06990, Gm12865, Gm12878), all the cell types analysed showed the same general trend regarding DNA accessibility at the scale of chromosomal bands.

One of the most important characteristics of the eukaryote nucleus is the compartmentalization of its functional domains. Chromatin is not just a thread floating freely in the nucleoplasm, it organizes as self-contained compartments (see Chapter 3). Different strategies to map and measure the properties of nuclear compartments had been developed in the past years. Following the strategy described previously for analysis of epigenetic marks, we analysed the relationship of the banding pattern and a variety of nuclear architecture data sets.

4.6.2 Compartmentalization of the genome and its relationship with the Giemsa banding pattern

4.6.2.1 Chromatin interaction maps

The recent fusion of the 3C methodology with massive parallel high-throughput sequencing, stands as a promising opportunity for a more precise understanding of the structural organization of chromatin in the interphase nucleus. From the several methods developed so far, we focused only on the method developed in Lieberman et al. [32] called Hi-C.

Name	Tissue of origin
Gm06990	Human B-Lymphocyte
Gm12865	B-Lymphocyte
Gm12878	Lymphoblastoid
Jurkat	T Lymphoblastoid
JurkatG1	T Lymphoblastoid
Ag04449	Fetal thigh fibroblast
Ag04450	Fetal lung fibroblast
Ag09309	Adult human toe fibroblast
Ag09319	Adult human gum tissue fibroblast
Ag10803	Adult human abdominal skin fibroblasts
Bjtert	Skin fibroblasts
Hcf	Human cardiac fibroblasts
Hgf	Human gingival fibroblasts
Nhdfneo	Neonatal human dermal fibroblasts
Nhlf	Normal human long fibroblasts
Hcm	Human cardieletal muscle
SkmcV2	Human skeletal muscle
Sknshra	Neuroblastoma cell line differentiated with retinoic acid
Caco2	Colorectal carcinoma
Helas3	Human epithelial carcinoma
Mcf7	Breast cancer
Panc1	Pancreatic carcinoma
Saec	Small airway epithelial
Cmk	Human acute megakaryocytic leukemia
Hl60	Human promyelotic leukemia
Hl60V2	Human promyelotic leukemia
K562	Human mylogenous leukaemia
Nb4	Acute promyelocytic leukemia
Nb4V2	Acute promyelocytic leukemia
Hmec	Human mammary epithelial
Hepg2	Human hepatocellular liver carcinoma
H1es	Human embryonic stem cells
H7es	Undifferentiated human embryonic stem cells
Hae	Human amniotic epithelial
Hcpe	Human choroid plexus epithelial cells
Hee	Human esophageal epithelial cells
Hipe	Human iris pigment epithelial cells
Nhek	Normal human epidermal keratinocytes
Hnpce	Human non-pigment ciliary epithelial
Hrcepic	Human renal epithelial cells
Hre	Human renal epithelial cells
Hrpe	Human retinal pigment epithelial cells
Huvec	Human umbilical vein endothelial cells
Skmc	Human sklastoma cell line differentiated with retinoic acid
SknshraV2	Neuroblastoma cell line differentiated with retinoic acid
Th1	Primary human Th1 T Cells
Th2	Primary human Th2 T Cells

Table 4.4: List of cell types analysed for DNaseI hypersensitivity

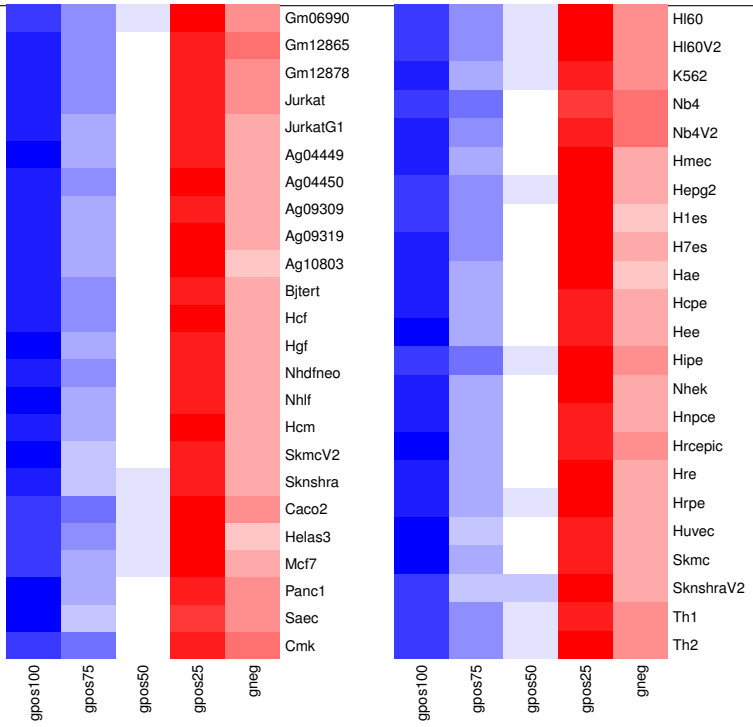


Figure 4.20: DNase I hypersensitivity. DNaseI sensitivity is directly reflected by the raw tag density of the DNase-seq protocol. In agreement with the preference for open chromatin marks, density of genes and PolII enrichment, gneg bands and gpos25 bands showed the highest signal of DNaseI hypersensitivity. The trend was maintained regardless of the cellular lineage.

Hi-C experiments quantify the frequency/stability of physical interactions between all possible pairs of loci in the genome. It is an "all against all" view of the nucleus, and therefore, given its multi-dimensional nature, Hi-C data has to be processed depending on the questions in hand. Dimension reduction by principal component analysis (PCA) showed that the nucleus organizes into self-contained multi-scale physical domains, confirming the presence of euchromatin and heterochromatin as two mayor mutually exclusive nuclear compartments, named A and B in the original work in Dekker's lab [32]. We used the original Lieberman data set as it was the only available at the time of this analysis. Improved data sets are available to date from Tanay group [215]. As the original work was relatively low in resolution (100 kb) and the scale of bands is in the range of several megabases, the improved binning suggested by [215] would reflect a minimal benefit.

The output of PCAs are special vectors called 'eigenvectors', which can be interpreted as a genomic track, similar to any other genome-wide data set where each element in the vector represents a window covering a fraction of the genome. PCA-reduced data allowed us to perform a measurement of the preferential association of the nuclear com-

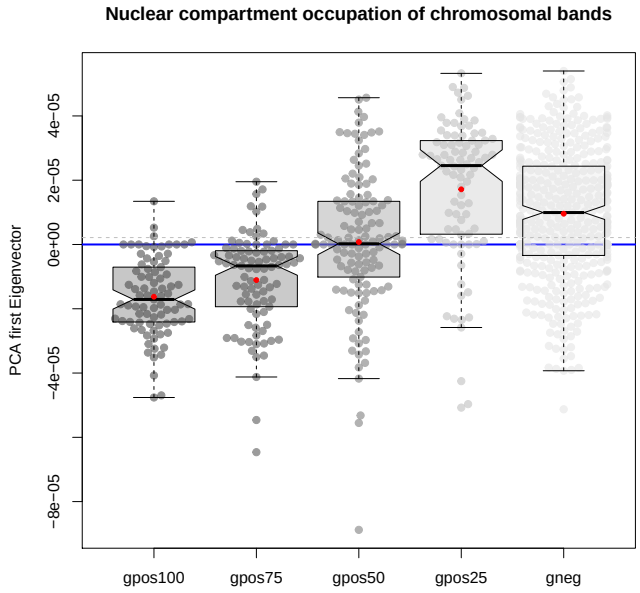


Figure 4.21: Nuclear compartment occupation of chromosomal bands. Eigenvector values distribution across bands. Values above zero (blue line) occupy the euchromatic compartment (compartment A) of the nucleus while values below zero occupy the heterochromatic compartment (compartment B). Each data point represents a single chromosomal band, red dots represent mean values for each category and dashed line represents the genomic median value.

partments with each band class. In the range of values of the eigenvector, values above zero represent the euchromatic compartment of the nucleus and values below zero the heterochromatin compartment. The original experiment was performed in karyotypically normal lymphoblastoid cell line (GM06990), therefore compatible with the cytogenetic map.

Given that the G darkest bands showed all the chromatin modifications classically associated with heterochromatin, we expected to find the darkest bands to show the lowest scores of the eigenvector. In contrast to R bands and gpos25 bands that should be present the highest scores of all the band classes. The results fitted our predictions (Fig. 4.21). Effectively, the darkest G bands are the ones to be found with scores representing the heterochromatin compartment followed by intermediate scores of the less dark bands until finally reaching the R bands and gpos25 bands occupying the euchromatic compartment of the nucleus. Once again we observed a G-like behaviour from R bands when compared to gpos25 which showed the highest values in the eigenvectors.

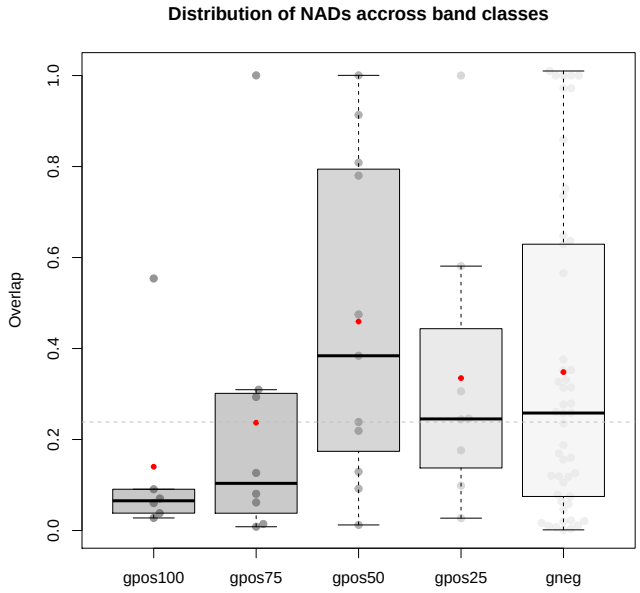


Figure 4.22: Distribution of NADs across band classes. Only bands with NAD signal were used for this analysis represented by points in the box plots. Notches are not plotted as in some cases they are bigger than the hinges of the box plots. Y-axis represent the proportion of a band that overlaps with a NAD.

4.6.2.2 Nucleolus-associated domains (NADs)

The nucleolus is most prominent sub-nuclear compartment. Is is structured by a complex spacial arrangement of repetitive sequences and protein complexes where ribosomal RNA (rRNA) is transcribed. Recent studies identified the regions of the genome that co-purify with the nucleolus [107, 216] in HeLa cells. These regions are termed the nucleolus-associated domains (NADs). In our sequence-based analysis (see Fig. 4.15 for reference), gpos25 were the band class most enriched with the repetitive elements belonging to rRNA sequences. Based on this result, we would expect gpos25 bands to show the richest NADs signal. When measuring the overlap of NADs with the coordinates of chromosomal bands, gpos50 showed the highest score. NADs are known to replicate in the second half of S phase, thus this result is in agreement with the replication timing of gpos50 bands (Fig. 4.17).

4.6.3 Lamin-associated domains (LADs)

The inner-face of the nuclear envelope associates with a fibrillar protein network called the nuclear lamina [217], which controls important regulatory and structural pro-

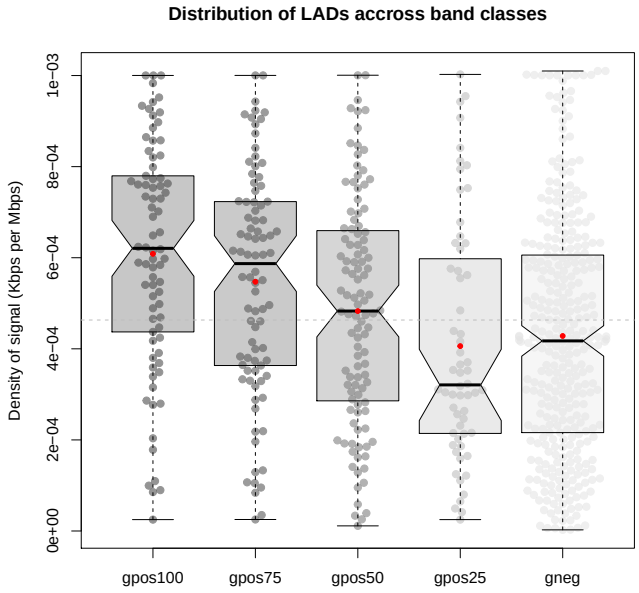


Figure 4.23: Distribution of LADs across band classes. Consistent with distribution of heterochromatin and replication timing data, lamin associated domains are present in dark G bands. Each point represents an individual band and dashed line represents the genomic median value. Red dots represent mean values.

cesses that allow a regulated physical connection between the nucleoplasm and the cytoplasm. It is believed that the interaction of chromosomes with the nuclear lamina allows structural polarization of interphase chromosome, which is a structural regulator of genomic function [218].

Based on (DNA adenine methyltransferase) DAM-id experiments in combination with microarray technologies, the regions of the human genome interacting with the nuclear lamina were identified [108]. As heterochromatin is essentially found in the peripheral regions of the nucleus we anticipated that gpos100 should be the most enriched in the LAD microarray data. The overlap of the band coordinates with the lamin domains confirmed our expectations and gpos100 showed the strongest signal. As a consequence, gpos25, followed by gneg bands showed almost complete absence of LAD signal.

		Genomic Summary						Median per band class				
Genomic Feature	Units	Min	1st.Qu.	Median	Mean	3rd.Qu.	Max	gpos100	gpos75	gpos50	gpos25	gneg
Compaction Ratio	log(Ideogram LG/Genomic LG)	-1.2	-0.2	0.01	0.08	0.25	3.29	-0.14	-0.1	0.04	0.1	0.06
Gm06990 HSs	Fold change *	169.62	2.85	0.95	-0.47	-1.9	-243.51	2.06	2.06	1.58	-2.37	-0.32
Hi-C	Eigenvector	-399.1	-32.76	4.5	5.85	52.29	285.2	-104.08	-32.59	0.36	43.99	24.03
NADS	Fold change **	0.01	0.44	1	1.78	2.61	8.01	0.73	0.93	2.34	0.95	1.08
LADS	Kbps/Mbps	0	387	1168	1512	2254	10800	3611	2462	1405	523	864.5

Table 4.5: Nuclear architecture features of chromosomal bands. Summary statistics for each of the genomic tracks evaluated. Fold changes relative to: *= Eigenvalue equal to zero; **= genomic median

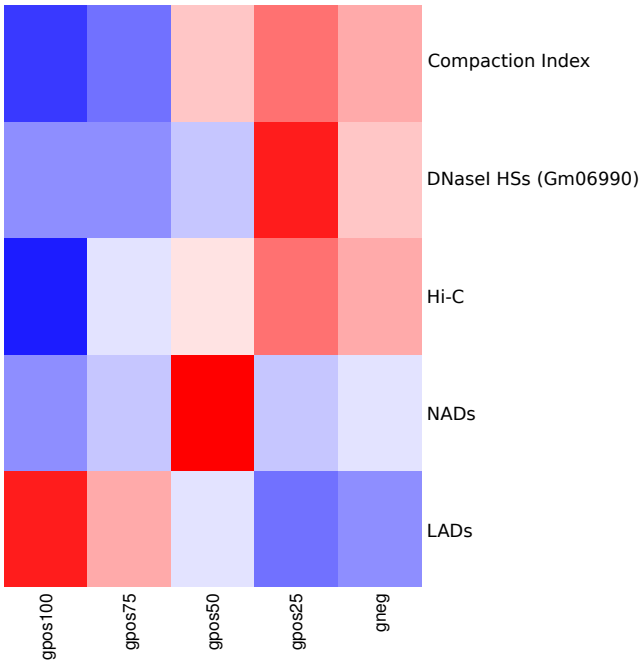


Figure 4.24: Higher-order chromatin architecture features

4.7 Discussion

Giemsa bands shown strong correlations with different genomic features at different scales. The darkest G bands (gpos100) contain gene-poor regions, low GC content and absence of H isochore families, high levels of L1 transposable elements, constitutive heterochromatin epigenetic marks and the latest time of replication. At the other extreme, gpos25 and R bands showed the densest gene regions, the highest levels of GC content and exclusive presence of the H3 isochore families, transposable elements of the Alu type, euchromatin marks associated with gene activation and active transcription and the earliest replication timing. Intermediate scores between the two extremes of the banding pattern (gpos100 and gneg) showed a increasing/decreasing gradient of the levels of presence of the different features that determine the band phenotypes. Additionally, the blocks of the genome that represent R bands show a broader range in sizes with a median smaller than the genomic median. The most striking result is the fact that, contrary to what was expected, R bands were overtaken by the gpos25 class in all the features measured.

The differential distribution of gene sizes can be explained by the levels of gene density observed in R bands and in particular, the kind of gene that they host. Housekeeping genes are being transcribed all the time and therefore do not need sophisticated means

of regulation. For this reason they show a more generic architecture, which allows their expression in a routine manner. In contrast to developmentally regulated genes, which require more genomic space to host the different regulatory elements that allow their proper fine-tuned expression and regulation in time, housekeeping genes need less space. In addition, as seen from the measurement of the distribution of L1 transposons to preferentially "jump" to heterochromatin and their transposition mechanisms, which leaves copies of them when transposing, it is reasonable to believe that the regions of the genome that host these kind of element will expand across evolutionary times-scales.

Since the early years of cytogenetics, one the first differences observed between the two classes of bands in the Giemsa banding pattern was the characteristic levels of GC content for each band type. A more robust and informative way of measuring GC content was through the long homogeneous levels found in what Bernardi defined as isochores. Even though interpreting the genome as repeating units of different GC content showed a more ordered partition of the genome and better frame of references for the definition of bands, it lacks a functional interpretation and mechanism to explain such a distribution and the biological forces driving it, in terms of genome function and genome evolution.

Repetitive elements provide a better perspective to understand why the genome shows the distribution of variable GC content blocks. Each family of TEs has its own levels of GC and they represent functional units with its own dynamics and well-understood molecular mechanisms of spreading. With this view, we are measuring the actual biological processes that drive genome evolution and not just the traces of their activity by abstracts metrics as the GC content.

An interesting question is if the repetitive elements typically found in G bands, such as LINEs, are actually molding and driving heterochromatin regions of the genome or they need heterochromatin in order to insert and propagate. As mentioned above, as the dark regions of the genome are significantly larger in size and the property of LINE elements to increase in copy number when transposing, it may be the case that these repetitive elements are actually defining the regions that will be seen as dark shades when the Giemsa staining procedure is performed.

Surprisingly, even though our results confirmed the previous observations that LINEs

are highly enriched in G bands, there were some cases of gneg bands and gpos25 (seen as outliers) that showed even higher values than the darkest G bands. In addition to this case, there were many features that showed such completely opposite results and are discussed below.

Unexpectedly, given that SINEs are characteristic of R bands, we found that not all the sub-families of SINEs are found in R bands. The DeuSINE family is actually absent from the R type. This family is one of the oldest found in the genome. A possible explanation is that after long evolutionary time, the genome found a way of silencing these elements, which are just a transposable element relic found in heterochromatin regions.

In addition to genetic data, the different band classes show specific epigenetic landscapes. Chromatin modifications associated with gene regulation were preferentially found in R bands, regardless of whether they were associated for activation or silencing of gene promoters. As mentioned before, the reason for the presence of gene silencing modifications in regions of open chromatin is explained by the promoter-specific nature of these marks and will be naturally found in regions hosting genes, which obey a more finer scale of genome function regulation, in contrast to classic heterochromatin marks even if they are repressive. Conversely, gene-poor regions therefore will not host gene-specific marks. For instance, the levels of methylation of H3K9 correlated with the staining intensity of the G bands. The more methylated this modification was, the darker bands it preferentially occupied. This is a good example on how, even though epigenetics influence genome function, there is an underlying hierarchical organization where the genome sequence acts as a platform, which will allow more layers of complexity to flourish, depending on the biological function of that particular region of the genome.

As previously known from the replication-based banding techniques that linked the S phase programme to the banding pattern [168], we observed that DNA from R bands is replicated from the onset of S phase to the first half of it. As suggested before [87], this strong correlation suggests that time of replication can be seen as a surrogate that represents or summarizes the epigenetic landscape of the genome and represents a link between higher-order chromatin structure and the combination of epigenetic factors driving genome function. Unexpectedly, we observed that the established dogma stating that

R bands replicate rigorously before G bands did not apply for the gpos25 band class. In all the cell-types analysed we observed this behaviour. The unexpected behaviour of these band categories is discussed below.

Bands also showed differences at the larger scales of nuclear architecture configuration. G bands hold more DNA per space unit, suggesting higher levels of chromatin compaction, in accordance to previous models [56] and observations [178]. This is consistent with the spatial distribution of chromatin observed under the microscope that shows more tightly organized, late-replicating and lamin-associated heterochromatin compartment at the peri-nuclear region of the nucleus and the open, fluid euchromatin compartment occupying the central nuclear volume [219, 29]. Our analysis of the Hi-C interaction maps also fit these observations as G bands occupy one of the two compartments determined by chromatin contacts and R bands the opposite one in addition to the distribution of nucleolus associated domains.

To summarize, our detailed, high-resolution analysis of the distribution of genomic features of different natures and at different scales, confirmed the observations of the field in the last decades, but also yielded unexpected findings regarding the features associated with gpos25 and R bands. The explanation of why there is such a discrepancy is explored and discussed in the next chapter.

Chapter 5

INACCURACIES ON THE CYTOGENOMIC MAP

General genomic properties of R and G bands have been known for many years. For instance, R bands were known to be associated with euchromatin such as high gene density, high GC content, enrichment of SINE elements, and replicate during the early stages of S phase and strictly before G bands. In the same way, G bands were known to be characterized by features typically associated with heterochromatin. In chapter 4 we explored a large variety of genomic features at a level of resolution that has not been reported previously and partially confirmed some well-known properties of Giemsa bands.

Contrary to expectations, the data presented so far showed that the gpos25 band class, and not gneg bands, was always the most enriched class for factors commonly associated with euchromatin or, conversely, the most impoverished for heterochromatin features. The summary heatmap in Fig. 5.1 clearly shows that for features that switch signal from one extreme to the other, gpos25 bands represented the corresponding highest (or lowest) extreme of the range of values. It seems that the real R bands are represented by the gpos25 class. This is a contradiction because gpos25 bands are determined by their staining intensity under the microscope but their data corresponds to R-like bands. This observation opens two alternative explanatory models.

The first alternative suggests that the established dogma in the chromosome research field is wrong and in reality R bands are not representatives of the pure euchromatic features (i.e. they do not necessarily replicate before G bands, can have less genes than G bands, etc.). The second, more simple explanation, suggests that there are major inaccuracies in the cytogenomic map. Specific band boundaries are perhaps unrealistic, given the structural nature of mitotic chromosomes where in reality, bands fade from one to the next without a sharp border. However estimated band borders should demarcate well enough

the cytological boundaries between bands, and if they do not, genomic and epigenomic features may be associate with the wrong band category.

Band borders in the current cytogenomic map were estimated by a computational method developed by T. Furey [97] that relied on data from mapping BAC clones to chromosome spreads and detecting them by FISH. BAC probes were sequenced and aligned to the genomic sequence map to generate a link between the physical cytogenetic map and the first assemblies of the human genome.

The logical design of this method is simple and effective, however, it is not perfect and there are two main factors that decrease its precision. First, resolving the position a BAC FISH probe hybridized in the chromosome spreads can be subjective as it is just a visual observation. This problem is exacerbated in cases where the band in turn shows a small size or when the BAC probe is in the transition zone between two bands and the signal of the FISH probe is in the border of both of them. The second source of noise comes from the fact that the BAC mapping procedure was performed by different laboratories belonging to the BAC Resource Consortium [109].

In this chapter we explore the weak points of the Furey method based on the quality of the BAC-mapping data used for the integration of the cytogenetic map to the genomic map. We provide evidence that band boundaries are not accurately determined in many cases. We then suggest a method to improve location of band boundaries.

5.1 Identification of misassigned bands

5.1.1 Revisiting Furey's method

As briefly mentioned above, the method implemented by Furey is based on fluorescence in situ hybridization experiments of more than 9000 BAC clones. As each BAC clone covers a unique segment of the human genome it is possible to link that particular segment of the genome to its corresponding band by FISH. Using this method, the high-resolution cytogenetic map comprising 850 bands could be recreated at the sequence level.

As a first step, the most prominent chromosomal landmarks were used to anchor the map. These features are the centromeres, the variable heterochromatin regions and the

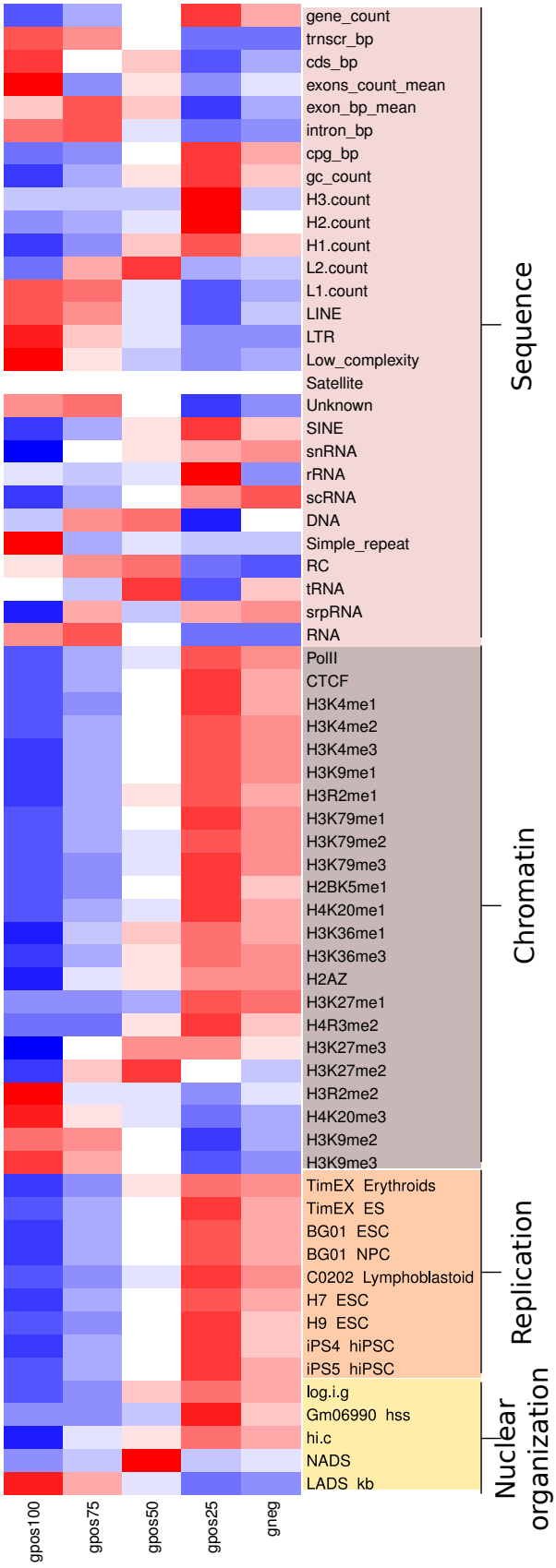


Figure 5.1: Distribution of genomic features for the different classes of bands. Measuring different genomic features across the five classes of Giemsa bands usually showed linear transitions from one extreme to the other, either from depletion to enrichment or vice versa. In other words, the feature that was enriched (or depleted) in a particular band, showed the opposite trend on its antagonistic class. This linear trend was always broken between the gpos25 and gneg classes.

stalks from acrocentric chromosomes, such as 13, 14, 15, 21 and 22. The second step for the map scaffolding made use of the estimated sizes of the bands under the microscope [104]. These two steps defined a temporary map which was then improved by the incorporation of the BAC clone hybridization.

The FISH data was first filtered as they observed that 10% of the total probes were directly contradicting one another and not all the laboratories used the same technique. Therefore each FISH result had to be scored depending on its laboratory of origin. Scores over six orders of magnitude were given to BAC clones mapped by the National Cancer Institute (NCI) as their mapping technique provided highly accurate results [184, 109].

For the proper estimation of the boundaries between bands, a Hidden Markov Model (HMM) was used. HMMs are commonly used in the field of computational biology as they provide a precise method for the identification of unknown states in a given data set. We consider a state in this case to a portion of the genome under analysis to which we want to segment and label. In the particular case of the bands, the states assigned by the HMM were equivalent to the name of the corresponding band. They named this algorithm Bander [97].

One of the weakest points of this method is that there are about 9000 BACs hybridized. If we consider the human genome to be approximately 3 billion bps and the average length of BACs of 200 kbps, we would need at least $\sim$15,000 BACs to cover the human genome at $1\times$. What these numbers tell us is that there may be portions of the human genome which are less well covered by BACs and these same BAC-poor regions can alter the precision with which Bander estimates the boundaries of bands and bias the results.

As a first attempt, we screened the density of BACs along the genome, and our prediction of regions of low BAC density was confirmed. Figure 5.2 shows the distribution of BAC midpoint as black asterisks at the bottom of the plot. We were concerned only with the middle point of the BAC as the HMM used in Bander only takes into account the position of the middle point of the probe regardless of the total length of the sequence. Under these circumstances, even though a chromosome may seem to be well covered by BACs in a coarse scale, upon closer examination there were regions of several megabases without a BAC. An example of these cases is shown inside the red box in Fig. 5.2 where

a border was called in the absence of supporting data.

In more detail, if there are large gaps without any BAC within, the HMM will tend to call a border next to the closest FISH probe that supports a change of band regardless of the position of the FISH probe. After large stretches without FISH data, the HMM could try to 'hang' to any data point supporting transition between one band type and the next. This 'hanging' phenomenon can only happen when the FISH probes on each extreme of a gap support two different bands. Two clear examples can be appreciated for each case after long gaps lacking BAC probes in the red box of Fig. 5.2. In the middle of the large dark band, we can see that even though there are relatively large gaps without a BAC probe, as there is no transition between band types, the HMM works fine. This is not the case for the immediate downstream thin R band which shows how the HMM 'hangs' to the coordinate of the BAC probe at the beginning of the gap. In addition, in order to respect the original cytogenetic map from the ISCN [104], the HMM was designed in such a way that it called all the bands in the map. Therefore some bands were called arbitrarily, or without strong BAC support, by the HMM in order to call another band with more evidence (see small red box in Fig. 5.2).

To quantify the potential degree of error by the border estimation algorithm, we measured the length of gaps between the BAC clones surrounding each of the band borders. The range of the gap sizes went from a minimum of 6,008 bp to a maximum of 32,200,000 bp, with a median size of 435,500 bp (top panels in Fig. 5.3). The great majority of the bands showed gaps smaller than 1 Mb (75.7%) but only 12.1% showed gaps smaller than 100 kb. The 100 kb window is relevant to this analysis as the genome was split in 100 kb windows in the original work for the estimation of coordinates for chromosomal bands [97]. Therefore the majority of band boundaries were supported by BACs further away than the minimum resolution of Furey's method (100 kb).

As not all the borders fell in the middle of the gap between the flanking BAC clones, we extended this analysis for the distances up- and downstream of each border, independently of each other. This measurement would allow us to spot any directional bias systematically introduced by the HMM.

The measurement of the up- and downstream distances from the border to the clos-

est BAC is shown in the middle panels of Fig. 5.3. It confirmed that the HMM did not introduce any bias, as the distribution of distances towards each side was symmetric. Statistical comparison of the up-stream distances against the down-stream distances at different scales did not show significant variation in a wide range of cut-off values. Bottom panel of Fig. 5.3 shows the p-value changes returned by the Wilcoxon's test at different cut-off values. The cut-off values represent the sizes for the distances from the closest BAC up- and downstream of the borders, similar to the cut-offs represented by the gray and red box in the upper panels.

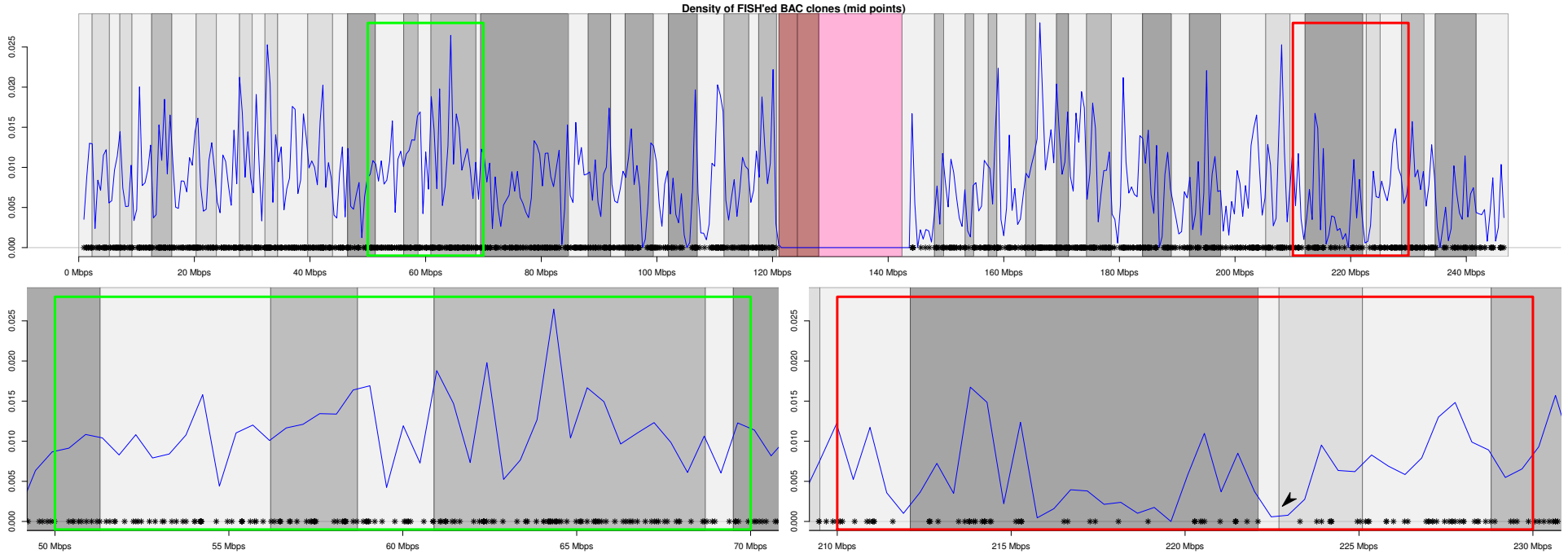


Figure 5.2: FISH probe density along chromosome 1. Top panel shows the whole chromosome, bottom panels show a zoomed view for the corresponding, green or red, square windows. The green region shows a portion of the chromosome with a relatively homogeneous BAC density. The red region shows the opposite, a region of the chromosome with large gaps between BAC probes (black asterisks in the bottom of each plot). The blue line represents the average density of BAC probes along the chromosome, shown as an aid for the proper visualization of BAC clone density used for the determination of Giemsa band borders. It is important to understand that the HMM does not necessarily call a border in highly dense BAC regions, so peaks of the blue line do not necessarily match band borders. Nevertheless, the accuracy of the HMM implemented in Furey's method depends on a homogeneous density of BAC clones. As there are regions of the genome with a low density of BACS it is expected that these regions of the genome will be more susceptible to misassignment of band borders by the computational method as observed in the thin R band around the 222 Mbp coordinate in the right bottom panel (black arrow).

5.1.2 Scoring chromosomal bands by BAC density support

The initial screening of the BAC coverage of the genome showed that a significant proportion (∼30%) of the borders appeared relatively far from the closest BAC, thus, potentially, be poorly supported. This basic measurement is not useful to identify the individual bands that had been misassigned due to a low density of BACs around the transition zones between two neighbouring bands.

For this reason we estimated a score for each border based on the density of BACs within a region with a diameter of 1 Mb around each border and then labeled each band depending on the quality of its borders. The justification for using this measure rather than the distance to the closest BAC is that the HMM used for the border prediction depends on the BAC data to call one band or another and this does not necessarily mean that bad supported borders will be far away from a BAC. It may even be the opposite case. In cases where the BAC data shows gaps of information and the next BAC supports the transition from one type of band to another, the HMM will naturally call a change of band just where the BAC is. Therefore a bad supported band will show a short distance between its position and the coordinates of the most proximal BAC. See Methods.

Figure 5.4 shows the distribution of BAC densitys in the upstream extreme of the band in the y-axis and the downstream density score on the x-axis. The left panel in Figure 5.4 A shows the distribution of scores for all kinds of bands. The rest of the panels show the distribution of scores for each of the band sub-classes. There is no significant bias in any of the band categories towards any particular extreme of the band.

The final score each band received was calculated as the sum of the BAC density scores of its up- and downstream borders. Figure 5.4 B shows distribution of scores based on the BAC density of the borders. The median number of BACs in a ±500 kb window around all the borders is seven. This value was selected as the cutoff for determining how well supported each band was. Values below the median are less well supported than values above the median. The inner panel of Figure 5.4 B shows the distribution of scores for each band class. The only classes that seem to show higher quality scores relative to the genomic median are gpos50 and gpos25 bands. Wilcoxon Rank Sum test confirmed for both cases to be statistically significantly greater than the genomic median

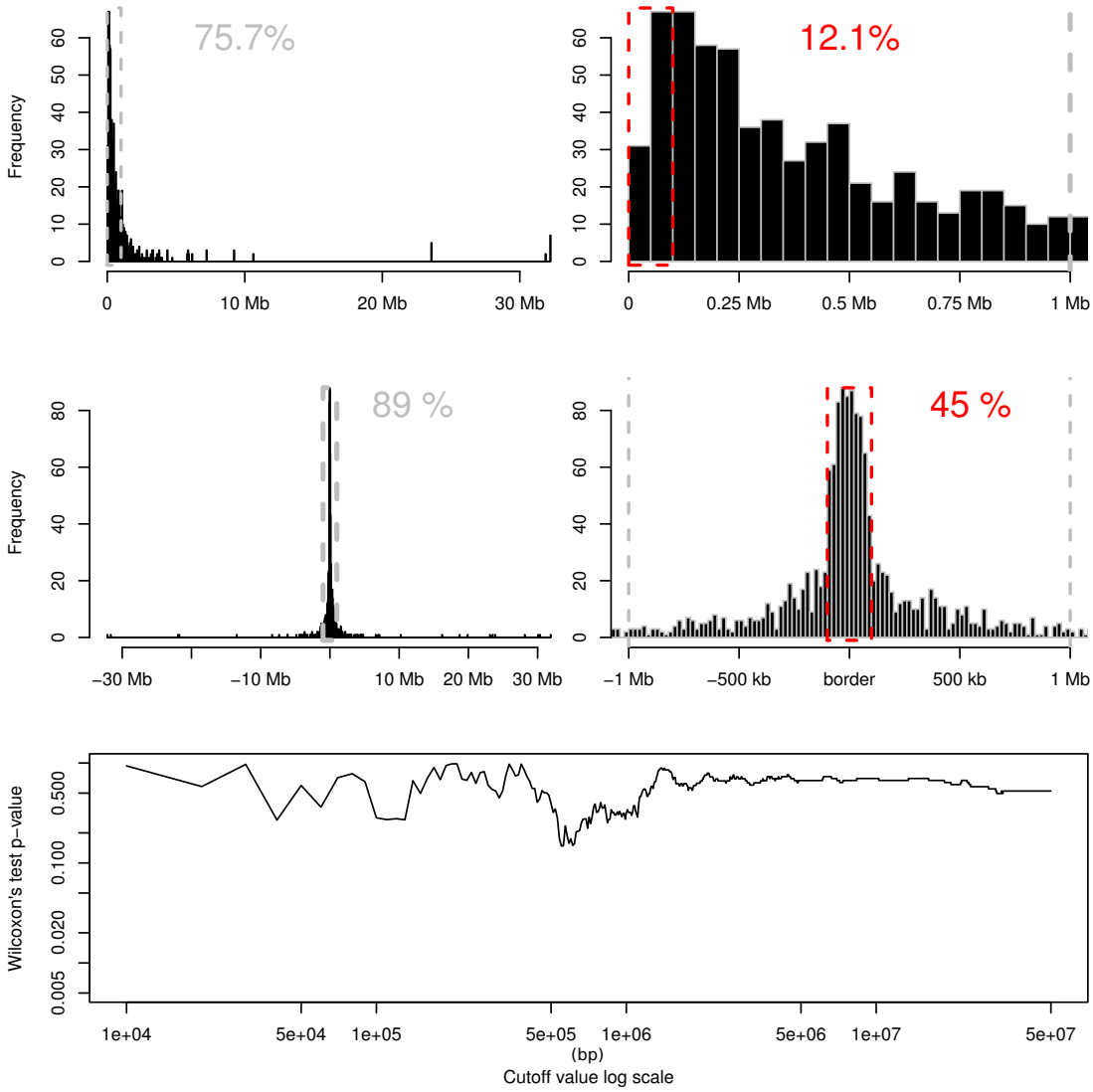


Figure 5.3: Screening of quality of predicted band borders. The distance of the flanking BAC probes (namely gaps) for each of the band borders was measured. The size distribution of gaps is shown in the upper panel. The great majority of the borders (75.7%) had a BAC clone supporting its assignment in a window smaller than 1 Mb and only 12.1% within a 100 kb window. Middle panels show the same analysis performed for each region, up- and downstream of the border. Under this view, approximately half of the borders are 100 kb away from a BAC clone. In an attempt to measure any directional bias from the HMM we compared the distribution of sizes of the upstream and downstream cohorts at different cutoff values, from 10 kb to 50 Mb. We found no significant difference at any scale. Bottom panel shows the p-values returned by the Wilcoxon's test for all the cutoff values assayed.

value (p-values of 0.001845 and 0.0007382 for gpos50 and gpos25, respectively).

5.1.3 Unsupervised Machine Learning for the identification of G-like R bands

In addition to providing evidence for badly supported bands based on the BAC mapping data, we wanted to see if we could use the different genomic features studied in Chapter 4 to classify the bands based on the combinatorial enrichment of these features. It is important to note that we did not attempt to re-classify R bands in the same way as they have been sub-classified previously as T-bands, "mundane", Alu rich or Alu poor [220, 176]. Given that R bands showed intermediate ambiguous characteristics, we attempted to decompose R bands into different sub-classes and ask whether any of those classes shared the same data profiles as G bands. This is one of the typical problems of machine learning: how, based on a set of features, can you define groups where elements within a group are more similar between them than to the rest of the elements in the dataset.

To complete this task, we used a special variant of the K-means algorithm which is able to adaptively select the set of variables to be used for the classification process. This method works under the premise that the variables that define the real clusters in the data are only a small fraction of the total used and performing the clustering analysis using all the features available can mask the real structure in the dataset. The selected group of relevant features will then be given higher weights during the clustering process. This variant of the K-means algorithm is called sparse K-mean clustering [110].

65 genomic features were selected, the same as the ones displayed above in Fig. 5.1, for the clustering analysis and a range from 2 to 9 Ks was evaluated. Figure 5.5 shows the output of each run of the clustering algorithm for all the levels of K assayed. The first row shows the distribution of normalized values for each of the genomic features without performing clustering analysis. Similar to the first heatmap in this chapter (Fig. 5.1), it is very clear how the extreme (either top or bottom) is always occupied by the gpos25 bands and not the gneg bands (drawn in red in this first row). After the re-classification of the gneg band class using the sparse clustering algorithm (K2) we observed that the new subclasses followed the opposite trends and the expected behaviour of R bands was shown

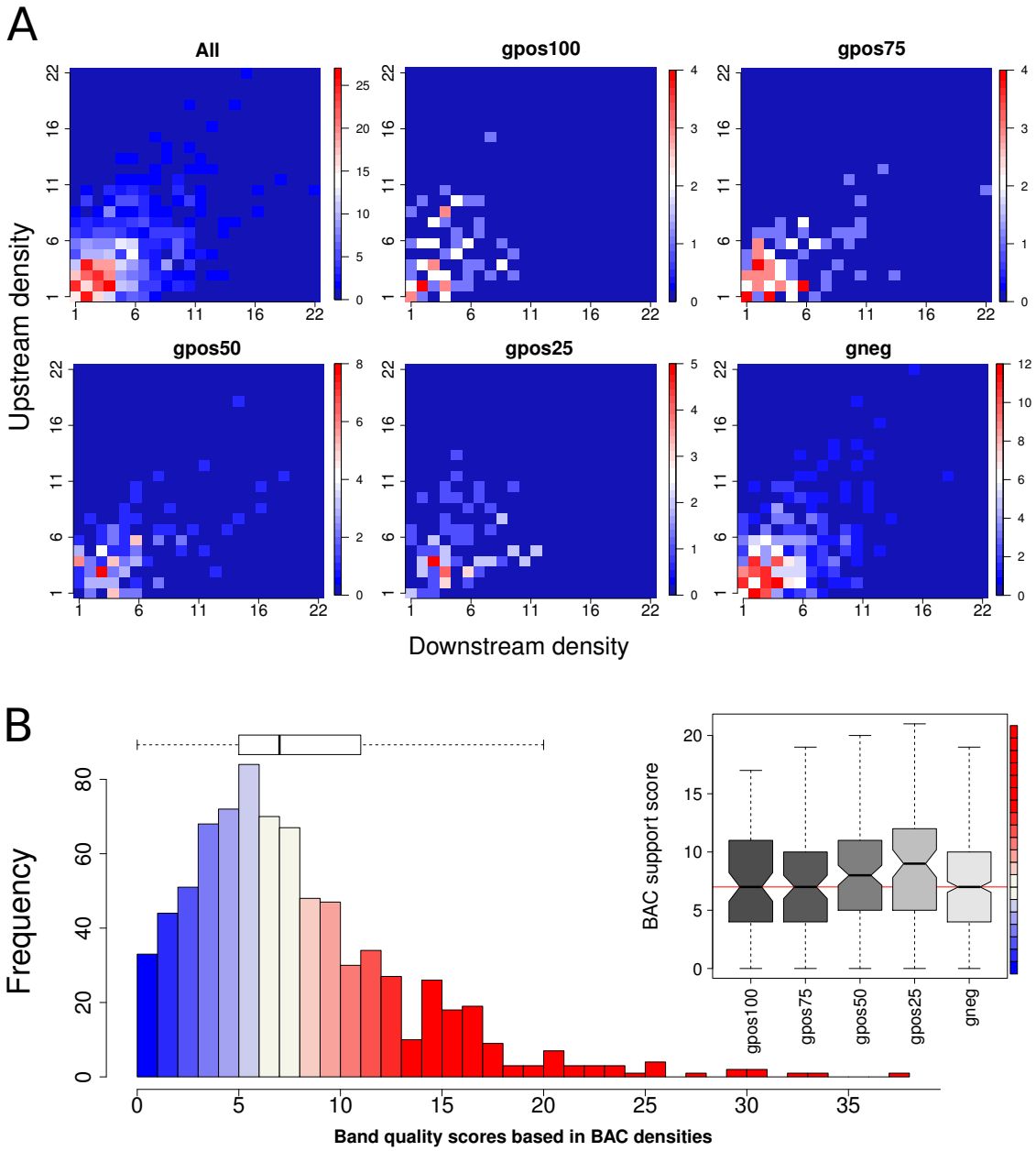


Figure 5.4: Spectrum of quality of bands based on the number of BACs supporting each band. (A) Distribution of each band border in a 2-D scoring space. The strength of BAC clones supporting each border was estimated by the amount of BAC clones within a 1 Mb region around each border. After each border quality was estimated, the distribution of each band based on the scores of its corresponding up- and downstream borders was plotted in a scoring space represented by the upstream scores in the y-axis and the downstream scores on the x-axis. The score spectrum was plotted for all and each of the different band classes individually. The colour code of each panel is normalized to the maximum value for each panel and is shown on the right axis. This representation of the data shows that there is no significant bias towards any extreme of the borders. In addition it shows that there is an important proportion of the bands that show poorly supported band borders. (B) The final score for each band was estimated by the addition of the quality scores of each of the borders, up- and downstream, of each band. The histogram shows the distribution of the final scores for all the bands with a median value of seven (white) as shown by the horizontal box plot on the top. Bands with a poor score were labeled with blue colours whereas bands with high scores were labeled with red. When comparing the distribution of each band class against the genomic median, gpos50 and gpos25 showed a statistically significant difference against the genomic mean. The colour code complementing the box plot comparison of scores represent the same colour distribution than the scores histogram.

by one the newly defined clusters (cyan in the K2 row). In addition, from the two new categories, the group with the smallest band size was represented by the G-like R bands (red in the K2 row). By successive increments of K we found that the new categories thrown by each run of the clustering algorithm showed intermediate behaviour from the distribution of values from one extreme to the other. Most importantly, when we varied the value of K we always found the same observation: there was always a new class of R bands that looked like G bands and, most importantly, it was always comprised of the smallest R bands.

As the algorithm will always find as many Ks as it is asked for, caution has to be exercised for the proper selection of this parameter. The uniqueness of the groups returned will tend to decrease as K increases as the algorithm will find less significant differences as the criteria to set a new group and fit the parameter K. To avoid estimating more groups than necessary, we chose 4 as the optimum value for K as it was able to fetch both extremes, either the canonical R bands and the wrongly delimited G-like R bands without creating too many intermediate groups (Fig. 5.6 top). In this classification the canonical R bands, showing the highest density of genes, correspond to T-bands [173] and clusters of highly-expressed genes [221].

With the re-classification of R bands based on the combinatorial enrichment of functional genomic features on one hand, and the individual scoring of bands based on the density of BACs around its borders on the other, we had enough evidence to confirm the initial speculation that the cytogenomic map presented major inconsistencies.

To formally test this idea, we compared the distribution of BAC-support scores against the newly defined categories by the sparse clustering algorithm and K=4 and found that the G-like R bands (cluster 3; C.3) showed the lowest BAC support scores, opposite to the canonical R bands (cluster 2; C.2) which showed BAC support scores higher than the genomic median, as seen in Fig. 5.6. The differences were statistically significant (p-value = 0.006 for C.3 and 6.348e-05 for C.2; Wilcoxon test).

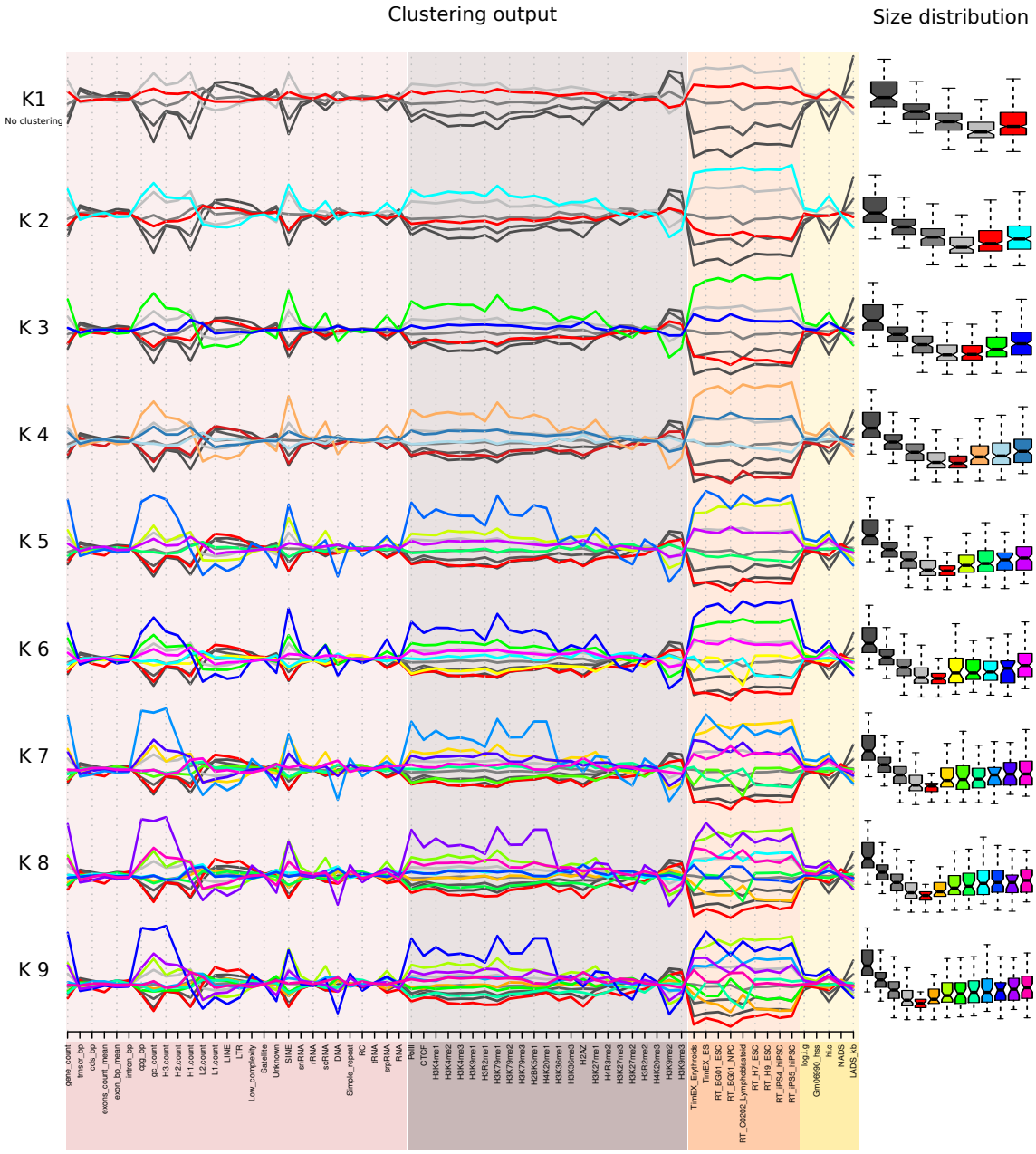


Figure 5.5: Parameter exploration of the machine learning methods applied. Using a total number of 65 genomic features, sparse clustering analysis was performed. Each of the runs was able to differentiate between wrongly delimited R bands from the canonical R bands. Each row of this figure shows the normalized value for each feature used in the analysis for all the 4 G band classes (in grey tones) in addition to the sub-classes found after clustering R bands. For the same cohorts shown in each plot, a complementing box plot showing the size distribution of the cohorts of that row. A striking result of these analyses is the fact that G-like R bands (red) are always the sub-class of R bands with the smallest band sizes.

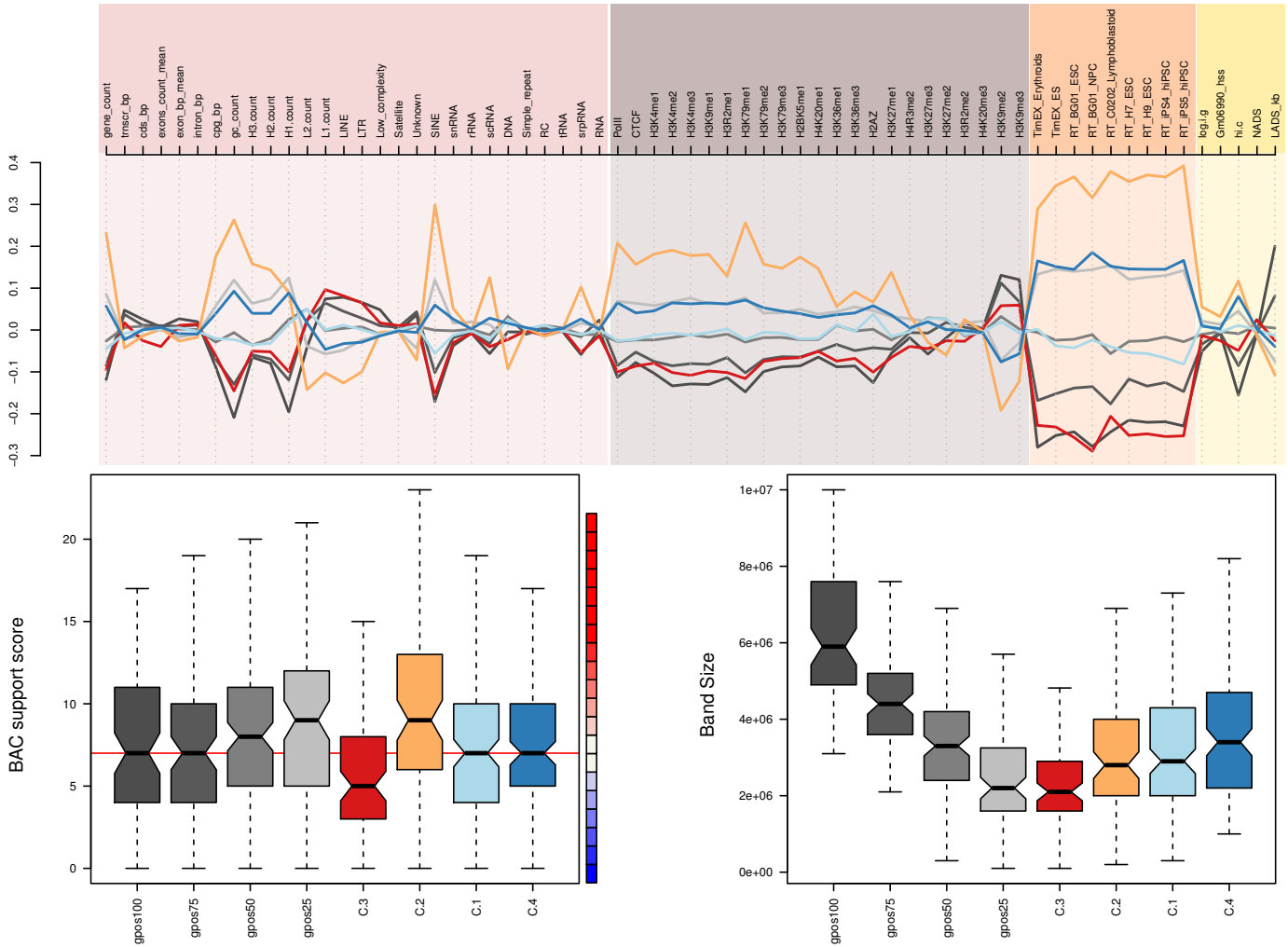


Figure 5.6: Reclassification of R bands by the K-means clustering algorithm. By performing clustering analysis based on the genomic data at different levels, we were able to differentiate the R bands that did not follow the expected behaviour. Poorly BAC-supported bands coincide with the misbehaving bands at the level of genomic data, whereas the best supported ones show the characteristics expected from typical R bands. The smallest bands were the most sensitive to inaccuracies at the border definition stage. Obtained clusters are shown as C.1-C.4.

This is a key result as it tells us that the R bands defined with robust BAC support are the ones showing the classical behaviour from what is expected from the cytogenetics literature. The C.2 group showed the highest scores for features typically associated with R bands such as the highest gene density, GC levels, H3 isochores, SINE TEs, presence of PolII and all the chromatin marks characteristic of euchromatin, earliest replication timing, etc.

Likewise, this result shows that the smallest bands are the most sensitive to inaccuracies in band borders. If a small band coordinates are shifted by a couple of megabases and its original size is relatively small, the vast majority of its body will present features in contradiction to its real nature and therefore will present the wrong profile when assessed for the genomic correlates that are covered in such a range. This explains why we had an unexpected behaviour in all the analysis performed in Chapter 4, R bands are the smallest in average and represent one extreme in the spectrum. This effect, while still present in the rest of the bands is less obvious as the rest of the bands are usually larger and their nature is not defined by the pure presence of a particular kind of features rather than a mixture of them.

A genome-wide comparison of the two metrics to assess band quality is shown in the summarizing Figure 5.7.

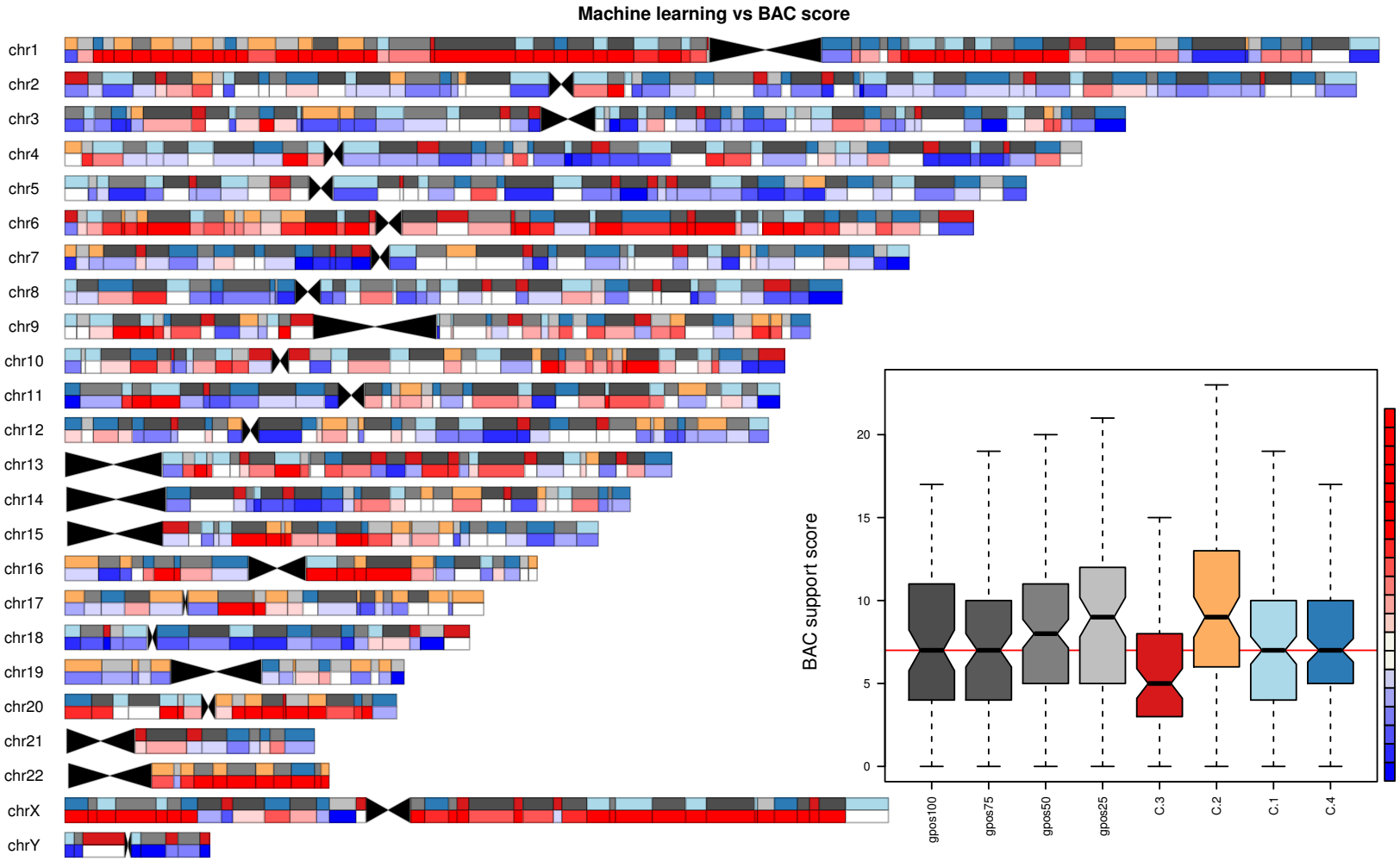


Figure 5.7: Comparison band score by two methods: Clustering analysis vs BAC support. Each band is represented twice, one for each scoring method. The top represents the results from clustering analysis and bottom shows a colour representing the position in the distribution of BAC support scoring. Black triangles represent the short arms of acrocentric chromosomes and centromeres. The G-like R bands (C.3) are usually paired with low BAC score bands (in blue and light blue colours). The colour code next to the box plot represents the BAC score range of values.

5.2 Suggested improvement of cytogenomic map

The evidence thus far shows that the chromosomal bands exhibit co-occurrence of different sets of features that are predictive of the different band states. This co-association of features is a good candidate for a surrogate of what, at the scale of the optical microscope, may appear as intercalated light and dark bands. Based om this premise, we attempted to improve band boundaries.

Using the machine learning approach we could detect groups of segments (in this case chromosomal bands) that shared specific combinations of different genomic variables. To extend this analysis we could have segmented the genome into a set of blocks that were distinct from the chromosomal bands, for instance 100 bp windows, and then cluster these genome segments based on the genomic features they show. The output of this analysis would then provide a set of different functional "flavours" of the genome. As a very similar approach has been previously performed by the Kellis group [111] in order to segment the genome based on the combinatorial pattern of chromatin epigenetic data. We decided not to repeat this analysis and used instead the data from this work.

The aim of Kellis' analysis was similar to ours, however, they used a different approach. Instead of using a clustering algorithm to define blocks of the genome with shared chromatin profiles, they implemented a multivariate HMM, which was able to assign states based on "recurrent and spatially coherent combinations of chromatin marks" [111].

Using this approach they reported 51 different chromatin states that fell in 5 broad categories: promoter-associated (1%), transcribed (15%), active intergenic (18%), repressive (64%) and repetitive (2%) states. Each of these states represented a specific combination of chromatin marks obtained in 200 bp windows (see heatmap in Fig. 5.8 B). The source of the chromatin data used to feed the multivariate HMM belonged to the ChIP-seq work from Barski et al. [195], the same that we used in our previous analysis.

Figure 5.8 shows a region of chromosome 1 with all of the 51 states labeled with the same colour as the broad class to which they belong. In the original work from Ernst et al. [111], the distribution of the 51 states along the 5 different band classes was measured using the same coordinates estimated by Furey [97]. In the heatmap of Fig. 5.8 B the

same swapped signal between gpos25 and gneg classes we found in our analysis can be observed for most of the states. However this pattern was not reported by Ernst et al.

In order to synthesize the information that each state holds, we fused the states that functionally matched the characteristics we observed from the canonical R bands and created two mutually exclusive states named 'active' and 'inactive'. The active state inherited 3 from the 5 broad categories: promoter, transcribed and active intergenic states. The inactive state inherited the remaining 2 states, the repetitive and the suppressed state. Fusing states in this manner allowed us to segment the genome in only two categories of antagonist biological properties at the high resolution of 200 bp blocks. The final result of the state fusing step can be seen at the bottom of Figure 5.8 A.

As an attempt to evaluate the segmentation of the genome in these two states as a predictor of Giemsa bands, we measured the ratio of active versus inactive blocks for each class of band and normalized it to the length of the band, in order to get an average signal across each band class (Fig. 5.8 C). The correlation between the staining intensity and the proportion of the active state against the inactive state was surprisingly strong. Dark bands showed almost complete absence of the active state, which increased gradually as bands got lighter. The classes of bands whose proportion of active states seemed to be the highest, belonged to gpos25 and gneg, as expected. However, once again we found that gpos25 and gneg bands were not clearly discriminated, with the gpos25 category showing the highest median score.

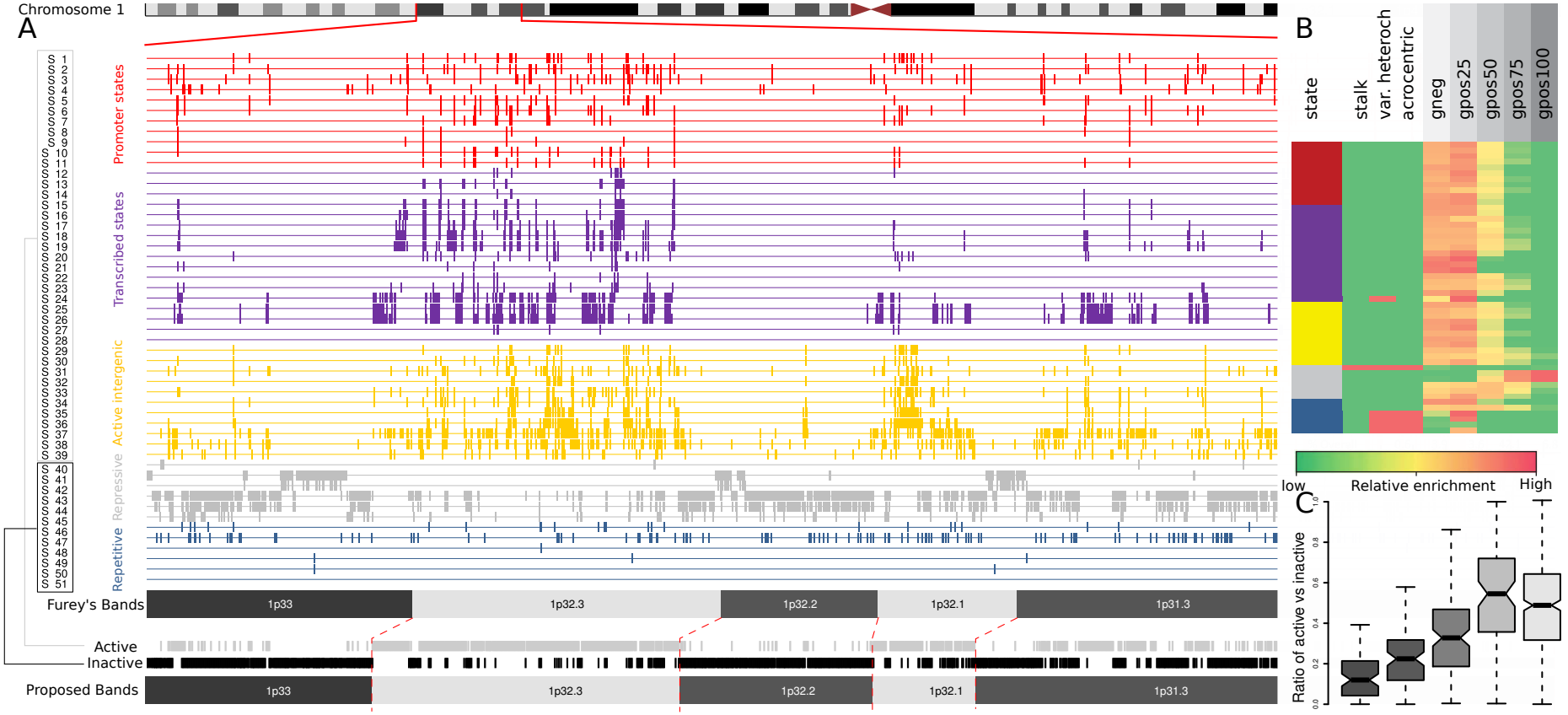


Figure 5.8: Fusion of functional chromatin states into active and inactive blocks. (A) Region of chromosome 1 showing the occupation of functional chromatin states defined by Kellis [111]. Individual states belong to 5 different broader states shown in colour. Promoter, transcribed and active intergenic state, in red, purple and yellow respectively, are fused into the larger category of active blocks. Repressive and repetitive states, gray and dark blue respectively, are aggregated into the inactive block category. The result of the fusion is shown at the bottom where light gray represents the active blocks and black, the inactive. There is a striking similarity between the distribution of active and inactive blocks and the cytogenomic map created by Furey, however, some inaccuracies can be observed. The proposed correction is shown by red dotted lines at the bottom if panel A. (B) Heatmap representing the enrichment of each individual state across Furey's cytogenomic. The same contamination issue we detected can be seen here where gpos25 shows the highest enrichment values for active chromatin states. Heatmap adapted from [111]. (C) After partitioning the genome into only two main states, active and inactive, we performed an analysis of the distribution of the active/inactive ratio across the five classes of bands. The staining intensity is directly correlated to the proportion of the active state against the inactive, where Giemsa darkest bands show a majority of inactive states and R bands and G-light bands for the active state.

5.2.1 Segmentation based on active vs inactive blocks

To further explore the notion that the banding pattern observed in the microscope is the result of large regions of the genome with homogeneous proportions of active and inactive states, which at different levels determine the staining intensity of a band, we aggregated the signal of the ratio of active/inactive states (referred to henceforth as the AI ratio) at different scales. Technical details on how this analysis was performed can be found in the section 2.3.3 from the Materials and Methods chapter.

Figure 5.9 shows the *in silico* reconstruction of the chromosome banding pattern based on the AI ratio at different scales. The pattern was reconstructed at different scales, nevertheless, given the binning method used, the dark signal showed shifts from scale to scale as the underlying genomic data was not centered at all times. This is one disadvantage of the method used. Nevertheless, for the purposes of visual comparison of the maps valuable conclusions could be drawn. The scale that showed the optimal reconstruction of the banding pattern was at the 40 kb scale. In this scale the "pixelation" effect due to large bin sizes was not so strong. Binning the AI ratio at larger scales (> 1 Mb), could partially reconstruct the banding pattern observed at lower resolutions of chromosome spreads, which showed around 10 large bands in chromosome 1.

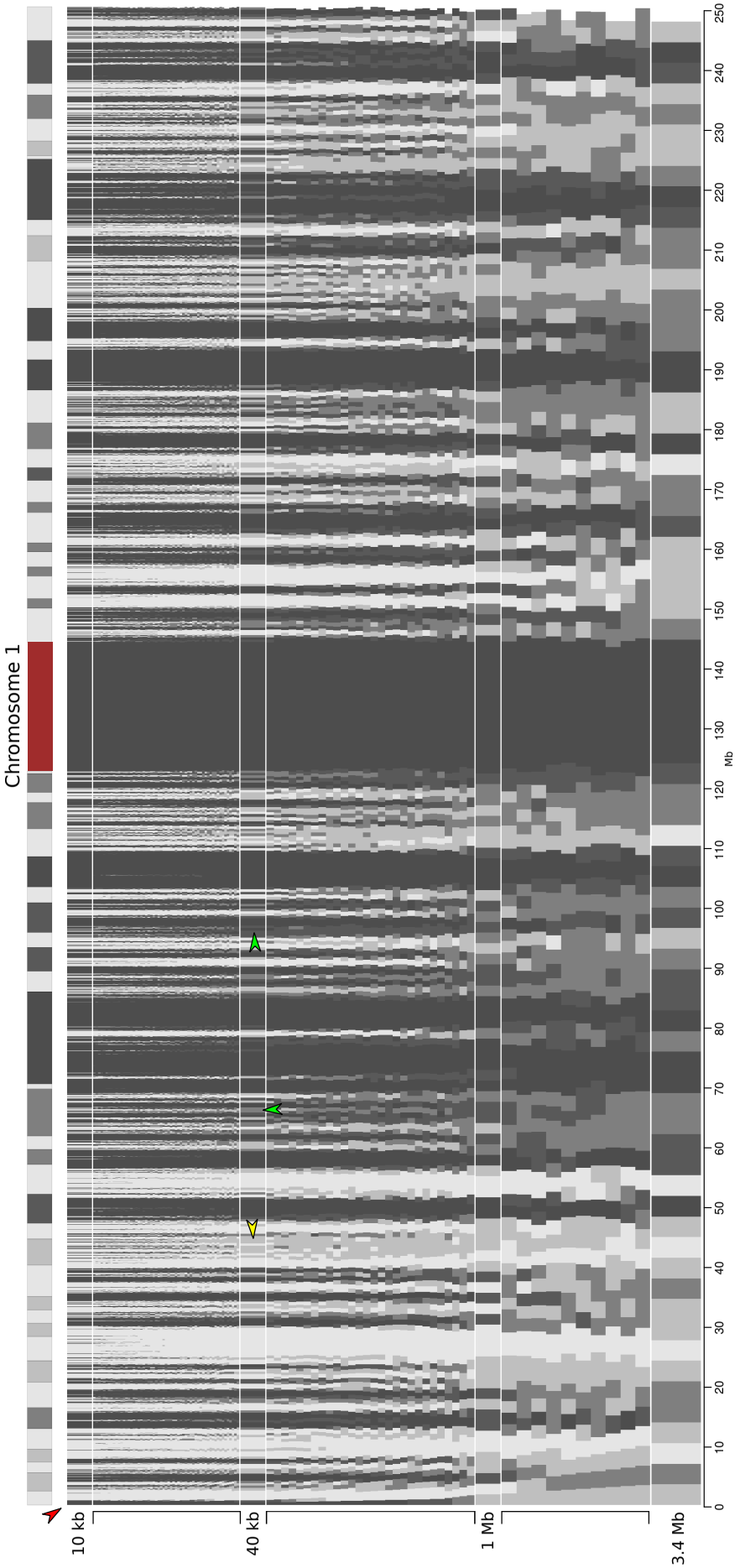


Figure 5.9: Multi-scale aggregation of AI ratio signal. The Giemsa banding pattern can be reconstructed at different scales based on the AI ratio. We calculated the AI ratio at different window-sizes on 3 different ranges. The first range covered window sizes from 10 kb to 40 kb by increments of 6.6 kb. The second range covered windows sizes from 40 kb to 1 Mb by increments of 25 kb. Finally, the third range covered the larger scales, from window sizes of 1 Mb to 1.75 Mb by window size increments of 150 kb. AI ratios were plotted as gray tones. One disadvantage of this visualization method is that for large bin sizes there is an offset between the data and the bin, which produces a "pixelation" effect.

This visualization approach demonstrated that light G bands, such as the gpos25 class, could be reconstructed in regions where the balance between states was skewed towards the active states. In other words, low-density clusters of thin inactive blocks interspersed by active blocks, represent the lightest gray bands (yellow arrow in Fig. 5.9). The same could be observed with intermediate dark bands, such as the gpos50 class, where the proportion of inactive was higher than the active states, therefore looking like a dark band under the microscope (green arrow). With this method we also observed bands that did not appear in the original ideogram, such as the one pointed by the red arrow. As these are close to telomeric regions, our hypothesis is that this extreme band does not appear in all the chromosome preparations or is more difficult to spot and was not included in the standard cytogenetic map or be cell-type specific. See Fig. 4.1 for visual reference of the chromosome spreads.

Based on these observations, in addition to the quantitative analysis of the proportional AI ratios (box plot in Fig. 5.8 C), we concluded that remodelling the Giemsa chromosomal banding pattern based on the fusion of data from [111] was a reasonable strategy to follow.

An alternative way of interpreting blocks of active or inactive states is by translating their categorical label to numerical values. We transformed the contiguous categorical data into sequences of 1 and -1 values. This transformation is the key operation for our further analyses, as by plotting the cumulative sum of these sequences produced signal profiles where transitions from one state to the other could be seen precisely as inflection points. 1 for active and -1 for inactive.

In agreement with the visual comparison of AI ratios, by plotting the cumulative sum of sequences of 1s and -1s representing each state and comparing the coordinates of the major inflections in the data we found that the majority of the band borders were located in the vicinity of inflection points that represent either peaks or troughs in the plot. The distances between band borders and inflections varied from border to border but a good general resemblance with the banding pattern was seen. Going from right to left, uphill regions in the cumulative sum plot represent light bands (as they are blocks of 1s) and downhill regions represent dark bands (as they are blocks of -1s). A comparison of the

cumulative sum profile from chromosome 1 and Furey's bands is shown in Fig. 5.10. The rough equivalence between the inflection profile and the banding pattern is the basis for the correction we propose of the cytogenomic map coordinates.

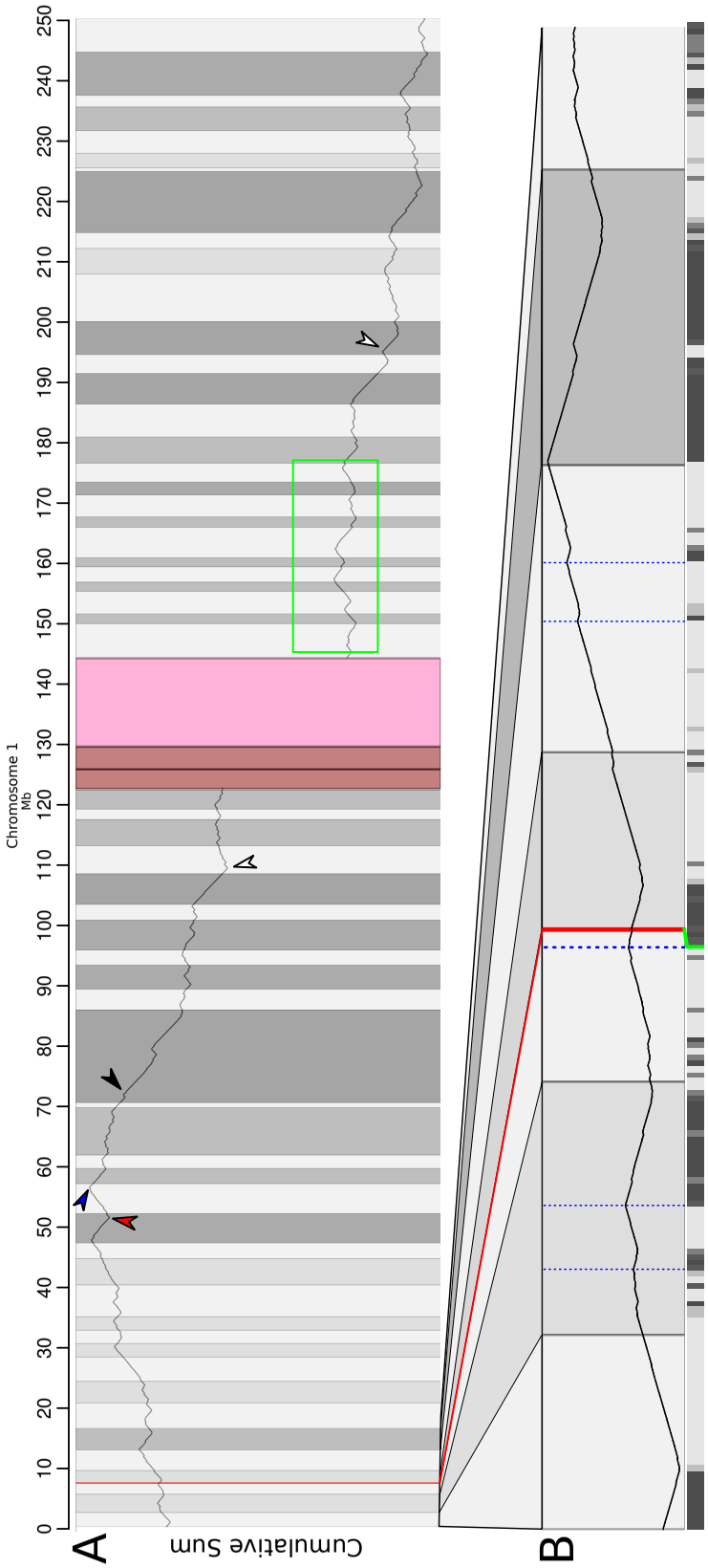


Figure 5.10: Representation of transitions between states by inflection profiles. The categorical variable from the 200 bp blocks of 'active' or 'inactive' states was transformed to 1 and -1 numerical values, respectively. (A) The cumulative sum of this sequence is plotted and appears as an inflection profile where peaks (blue arrow) coincide with transitions from R bands to G bands and vice versa for troughs (red arrow). Furey's chromosomal bands are superimposed for a comparison. The banding pattern follows the inflection profile at the largest scale with some inaccuracies (white arrows) and some offset regions (green box). Closer examination of the inflection profile reveals inflection points at low scales. (B) One iteration of the algorithm that adjusts Furey's coordinates to the transitions between Kellis' states is shown for the band in red that turns green after correction. Blue lines represent relevant inflections (see text). The bottom track depicts the AI ratio as a heatmap, equivalent to the inflection profile.

As Furey's coordinates are based on BAC mapping and thus encode information that Kellis' states lack, we decided to rely on them as a reference and then fine-tune the location of the Furey's borders depending on the type of inflection that each particular border needed. For instance, the inflection separating a gneg in the left side from a gpos100 band to the right side would have to be a peak in the inflection profile (blue arrow in Fig. 5.10). In addition, as Furey based his analysis on the set of bands reported in [104], our corrected map would represent a corrected version of the same map established in the ISCN.

The assignment of band borders to the closest inflection point is not a trivial task. Upon closer examination, the inflection profile did not show monotonous transitions between peaks and troughs, it presents texture at very small scales (Fig. 5.10 B). Using a method that measures local minima and maxima at different scales, we scored each inflection point (see section 2.3.3.2 in Methods) and then, for each band border in Furey's band table, we reset its position to match the coordinates of the closest scored inflection point.

One iteration of the algorithm that corrects the border coordinates is shown in Fig. 5.10 B, for the red border. Blue dotted lines represent the possible inflection that this border can be assigned to; smaller inflections are ignored as a result of the scoring process. Given that the border in turn (red) represents the transition between an R band and a G band, the inflection sought has to be a peak and the closest is chosen (thicker blue dotted line). This method was applied to each border.

Preliminary results based on this strategy are shown in Fig. 5.11 the set of corrected coordinates (bottom) follow the pattern of active/inactive states more closer than the original coordinates (top).

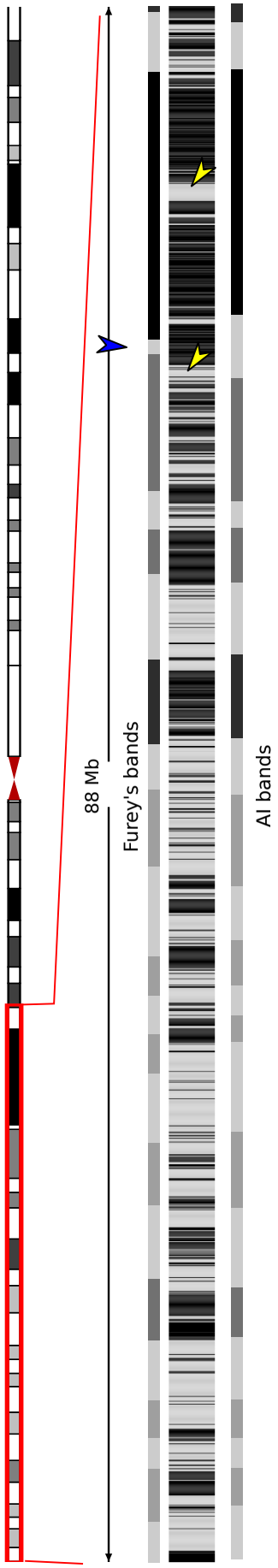


Figure 5.11: Correction based on segmentation of the genome on active and inactive states. Band boundaries are now in better agreement with the chromatin data. The implemented algorithm still shows some inconsistencies as shown with the blue arrow as there are two possible places where the small R band could fit, shown in yellow arrows.

5.3 Discussion

In this chapter we revisited the data on which Furey based his cytogenetic map and explored potential sources of error. We found that, even though Furey's method is effective in the determination of chromosomal band coordinates on a large scale, its robustness is compromised over regions of the genome that are not well supported by BACs. Using two independent methods, we were able to identify the most inaccurately determined bands.

First, we scored bands based on their BAC support and then identified the less well BAC-supported boundaries and bands. The second method made use of an unsupervised machine learning technique. We clustered R bands based on the co-occurrence of genomic traits. By this method, we identified the R bands that showed the expected set of genomic features, and those R bands with an unexpected behaviour, that was more similar to G bands (G-like R bands). When comparing results from the BAC scores and the clustering method, we found that the bands with a poor BAC support were in the same cluster as the G-like R bands. The opposite was seen for well BAC-supported bands, which were found in the cluster with the expected genomic features known to be associated with R bands. Furthermore, as expected, we found that the bands that were most affected by the inaccurate assignment of borders were the smallest in size, since small shifts in location of boundaries can have a bigger influence on the properties of small bands..

After establishing that the current cytogenomic map has significant inaccuracies in some cases, we decided to improve the location of the band borders. Our analysis in Chapter 4 showed that there is a subset of genomic features unique to each extreme of the Giemsa spectrum (gpos100 and gneg), such as gene density, the presence/absence of LINES, SINES, specific replication timing, specific chromatin modifications, etc. The intermediate band classes, such as gpos25-75 were shown to be only a gradient in the balance these defining genomic features. Previous work from Kellis lab [111] segmented the genome based on the combinatorial presence of the same genomic features just mentioned above. We used this segmentation of the genome as the basis for the improved determination of band boundaries by updating Furey's coordinates, which incorporate BAC data.

Current work is in progress to optimize this method and a validation method is re-

quired. Recent work from Ernst et al. has systematically segmented 9 different human cell types [222] using the same strategy as before [111]. By performing the transformation of chromatin states to inactive and inactive blocks that we propose here, it is possible to determine the cell-type specific, high-resolution Giemsa pattern of these 9 samples. If our method for predicting band boundaries has value, the preparation of Giemsa-stained chromosome spreads using these 9 cell types should match the *in silico* predicted ones.

Chapter 6

GENERAL DISCUSSION AND PERSPECTIVES

The mammalian interphase nucleus hosts innumerable biological processes in a complex interdependent network that operates at multiple scales. For instance, gene transcription and initiation of DNA replication require specific chromatin environments that are the result of both, local [223] and large-scale factors [149]. Locally, transcribed genes adopt a particular spatial configuration [224] through the interaction of genetically determined elements (enhancer, promotes and transcription factor binding sites) [225, 152] with the transcription machinery. These processes are dependent on a local epigenetic milieu which also contributes to the local topology determined by loops of chromatin fibers [226].

The establishment of transcriptionally active hubs is a dynamic process that happens under different regimes. For instance, transcription factories assemble constitutively with house-keeping genes as seen in PolII-interaction maps [224]. Factories can also be rapidly induced upon certain stimulus [227, 228] or during development, where the conformational re-arrangement of chromatin domains take place in order to fit the topological requirements for transcription, like the locus control region of the alpha globin locus in human [229]. These local rearrangements follow a hierarchy and can happen only if their local domain is located in the correct large-scale landscape. For instance, the higher-order reconfiguration of chromatin domains upon differentiation from embryonic stem cells to neural precursor cells, measured by time of replication, reveals the order in which nuclear structural changes correlate with the activation of neuron-specific genes [75, 76]. In Chapter 4 we saw that replication timing can be seen as a surrogate for chromatin epistate and higher-order configuration and its correlation with the banding pattern would suggest two ideas. First, that the cell regulates genome function by the organization of

higher-order domains by allowing large regions to occupy the active compartment of the nucleus [230]. Second, these changes would be reflected at the banding pattern level, which would be more or less obvious depending on the size of the rearranged domains. This idea is in agreement with the different type of function that genes show depending on their staining intensity (equivalent to replication timing [66]). In other words, there is a positive relationship between tissue-specificity of a gene and the type of large-scale chromatin domain (chromosomal band) that genes occupy. This is also reflected in the internal structure of genes, as the size of their introns increases positively with staining intensity. This suggests that tissue-specific genes require more genomic space to "fill up" with regulatory elements which will provide the finer level of regulation these genes need, opening the question as to how these mechanisms have evolved.

Given the vast proportion of the human genome that is covered by repetitive elements, they could represent a key link between genomic function and genome evolution. The exact number is still not clear but recent studies report that the proportion of the human genome with a repetitive origin is close to two thirds of it [194]. Moreover, recent studies report SINE elements to be involved in the definition of DNA foci [231, 33]. Another interesting feature of transposable elements is their ability to self-replicate and shape the genome not only by expanding it or truncating coding sequences but also by transporting architectural elements, such as CTCF binding sites when transposing [232]. Another interesting point arises from the evolutionary conserved synteny blocks between human and mouse that do not only share sequence, but also higher-order chromatin structures [66, 80, 76]. This data suggest that structural modules of defined biological function are conserved and are shuffled during evolution. Repetitive sequences could represent 'wild card' regions that allow the recombination of the conserved functional structural domains.

An alternative view of the relevance of repetitive sequences is that they can encode structural motifs by the repetition of the same sequence blocks, in addition to protein-coding DNA and non-coding regulatory elements, the repetitive landscape of the genome could be seen as structure-coding DNA. Such 'scaffolding domains' could give rise to the different patterns of gene-rich and gene desert regions observed in mouse [218]. Large repetitive blocks of LINEs, as in some R bands, could potentially work as anchors that

tether and stabilize large domains of chromatin as reported with centromeric DNA [179]. In the same line of thought, we would expect different chromosome "personalities" derived from the differences in their Giemsa banding pattern and chromosomal environment. Chromosomes varying in size, density of genes and repetitive blocks tend to occupy different locations inside the nuclear volume [233]. A very good example is the comparison of chromosome 18 and 19. These human chromosomes have a similar length but very different gene content, chromosome 18 is very gene-poor whereas chromosome 19 very gene-rich. This is reflected in the preference for the nuclear periphery from chromosome 18 and a less compact morphology of chromosome 19 [234]. Furthermore, chromosome 19 tends to interact more frequently with other gene-rich chromosomes towards the centre of the nucleus [230]. The transcription of SINE repeats [231]- highly present in these chromosomes - can contribute to the centric 'buoyancy' of the chromosomes with large R bands.

But how much do chromosomes interact with each other? Even though we did not measure the levels of interaction between specific chromosomes we showed the importance of local chromatin organization to constrain the interaction of neighbouring chromosome territories (CT). When disrupting DNA foci integrity by hyperacetylation of histone tails in HeLa cells, we observed an increase in the colocalisation signal between two neighbour CTs, relative to the very low levels of signal co-occurrence in normal conditions. One of the technical highlights of our study is the automation of the imaging analysis, which avoided possible selection bias due to human intervention of samples, in addition to allowing the analysis of larger sample sizes. High-resolution tethered Hi-C experiments have recently shown that active regions of the genome do interact more frequently among them despite the chromosome territory they occupy [120]. These active regions are preferentially found on the borders of the CTs when FISH probes for representative regions of different chromosomes were measured [120] but these interactions occur in only a small fraction of the cells a do not interact exclusively. As the labelling method used for quantification of colocalised CTs could potentially miss thin chromatin fibers further experiments using specific FISH probes can shed light on this problem.

In this work, we have explored the elements driving chromosome composition and

architecture, with particular emphasis in the structure:function duality of chromosome biology. We made this exploration from an experimental and computational angle, making use of advanced tools on both fronts. Base on this we proposed a correction of the cytogenomic band coordinates based on types of chromatin, rather than visual estimation of FISH probes. Future improvements of the method that we propose herein will allow a more robust definition of the coordinates and comparison with other cell types, but most importantly, will help in the design of experiments which will validate of our model and help is gain deeper insights into genomic and epigenomic determinants of chromosome and nuclear architecture.

Chapter 7

Bibliography

[1] J. D. WATSON and F. H. CRICK, "The structure of DNA.," *Cold Spring Harbor symposia on quantitative biology*, vol. 18, pp. 123–131, 1953.

[2] G. Felsenfeld and M. Groudine, "Controlling the double helix," *Nature*, vol. 421, pp. 448–453, Jan. 2003.

[3] N. Paweletz, "Walther Flemming: pioneer of mitosis research," *Nat Rev Mol Cell Biol*, vol. 2, pp. 72–75, Jan. 2001.

[4] R. D. Kornberg, "Chromatin structure: a repeating unit of histones and DNA.," *Science (New York, N.Y.)*, vol. 184, pp. 868–871, May 1974.

[5] P. Oudet, M. Gross-Bellard, and P. Chambon, "Electron microscopic and biochemical evidence that chromatin structure is a repeating unit.," *Cell*, vol. 4, pp. 281–300, Apr. 1975.

[6] K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent, and T. J. Richmond, "Crystal structure of the nucleosome core particle at 2.8 Å resolution," *Nature*, vol. 389, pp. 251–260, Sept. 1997.

[7] T. Kouzarides, "Chromatin modifications and their function.," *Cell*, vol. 128, pp. 693–705, Feb. 2007.

[8] T. Jenuwein and C. D. Allis, "Translating the Histone Code," *Science*, vol. 293, pp. 1074–1080, Aug. 2001.

[9] A. G. Ladurner, C. Inouye, R. Jain, and R. Tjian, "Bromodomains mediate an acetyl-histone encoded antisilencing function at heterochromatin boundaries.," *Molecular cell*, vol. 11, pp. 365–376, Feb. 2003.

[10] M. Shogren-Knaak, H. Ishii, J.-M. M. Sun, M. J. Pazin, J. R. Davie, and C. L. Peterson, "Histone H4-K16 acetylation controls chromatin structure and protein interactions.," *Science (New York, N.Y.)*, vol. 311, pp. 844–847, Feb. 2006.

[11] S. A. Jacobs and S. Khorasanizadeh, "Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail.," *Science (New York, N.Y.)*, vol. 295, pp. 2080–2083, Mar. 2002.

[12] J. Y. Fan, D. Rangasamy, K. Luger, and D. J. Tremethick, "H2A.Z alters the nucleosome surface to promote HP1alpha-mediated chromatin fiber folding.," *Molecular cell*, vol. 16, pp. 655–661, Nov. 2004.

[13] "An integrated encyclopedia of DNA elements in the human genome," *Nature*, vol. 489, pp. 57–74, Sept. 2012.

[14] R. T. Simpson, "Nucleosome positioning: occurrence, mechanisms, and functional consequences.," *Progress in nucleic acid research and molecular biology*, vol. 40, pp. 143–184, 1991.

[15] C. Jiang and B. F. Pugh, "Nucleosome positioning and gene regulation: advances through genomics," *Nature Reviews Genetics*, vol. 10, pp. 161–172, Mar. 2009.

[16] Y. Fu, M. Sinha, C. L. Peterson, and Z. Weng, "The Insulator Binding Protein CTCF Positions 20 Nucleosomes around Its Binding Sites across the Human Genome," *PLoS Genet*, vol. 4, pp. e1000138+, July 2008.

[17] A. Shaw, P. Olivares-Chauvet, A. Maya-Mendoza, and D. A. Jackson, "S-phase progression in mammalian cells: modelling the influence of nuclear organization.," *Chromosome Research*, vol. 18, pp. 163–178, Jan. 2010.

[18] C. L. Woodcock and S. Dimitrov, "Higher-order structure of chromatin and chromosomes," *Current Opinion in Genetics & Development*, vol. 11, pp. 130–135, Apr. 2001.

[19] J. Bednar, R. A. Horowitz, S. A. Grigoryev, L. M. Carruthers, J. C. Hansen, A. J. Koster, and C. L. Woodcock, "Nucleosomes, linker DNA, and linker histone form

a unique structural motif that directs the higher-order folding and compaction of chromatin," *Proceedings of the National Academy of Sciences*, vol. 95, pp. 14173–14178, Nov. 1998.

[20] T. J. Maresca, B. S. Freedman, and R. Heald, "Histone H1 is essential for mitotic chromosome architecture and segregation in Xenopus laevis egg extracts.," *The Journal of cell biology*, vol. 169, pp. 859–869, June 2005.

[21] T. J. Maresca and R. Heald, "The long and the short of it: linker histone H1 is required for metaphase chromosome compaction.," *Cell cycle (Georgetown, Tex.)*, vol. 5, pp. 589–591, Mar. 2006.

[22] J. T. Finch and A. Klug, "Solenoidal model for superstructure in chromatin.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 73, pp. 1897–1901, June 1976.

[23] E. Fussner, R. W. Ching, and D. P. Bazett-Jones, "Living without 30nm chromatin fibers," *Trends in Biochemical Sciences*, vol. 36, pp. 1–6, Jan. 2011.

[24] K. Maeshima, S. Hihara, and M. Eltsov, "Chromatin structure: does the 30-nm fibre exist in vivo?," *Current opinion in cell biology*, vol. 22, pp. 291–297, June 2010.

[25] M. Eltsov, K. M. Maclellan, K. Maeshima, A. S. Frangakis, and J. 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 of the United States of America*, vol. 105, pp. 19732–19737, Dec. 2008.

[26] H. Nakamura, T. Morita, and C. Sato, "Structural organizations of replicon domains during DNA synthetic phase in the mammalian nucleus.," *Experimental cell research*, vol. 165, pp. 291–297, Aug. 1986.

[27] H. Ma, J. Samarabandu, R. S. Devdhar, R. Acharya, P. C. Cheng, C. Meng, and R. Berezney, "Spatial and temporal dynamics of DNA replication sites in mammalian cells.," *The Journal of cell biology*, vol. 143, pp. 1415–1425, Dec. 1998.

[28] K. Koberna, A. Ligasová, J. Malínský, A. Pliss, A. J. Siegel, Z. Cvacková, H. Fidlerová, M. Masata, M. Fialová, I. Raska, and R. Berezney, "Electron microscopy of DNA replication in 3-D: evidence for similar-sized replication foci throughout S-phase.," *Journal of cellular biochemistry*, vol. 94, pp. 126–138, Jan. 2005.

[29] D. A. Jackson and A. Pombo, "Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells.," *The Journal of cell biology*, vol. 140, pp. 1285–1295, Mar. 1998.

[30] T. Cremer, C. Cremer, H. Baumann, E. K. Luedtke, K. Sperling, V. Teuber, and C. Zorn, "Rabl's model of the interphase chromosome arrangement tested in Chinese hamster cells by premature chromosome condensation and laser-UV-microbeam experiments.," *Human genetics*, vol. 60, no. 1, pp. 46–56, 1982.

[31] T. Cremer and M. Cremer, "Chromosome Territories," *Cold Spring Harbor Perspectives in Biology*, vol. 2, Mar. 2010.

[32] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome," *Science*, vol. 326, pp. 289–293, Oct. 2009.

[33] J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren, "Topological domains in mammalian genomes identified by analysis of chromatin interactions.," *Nature*, vol. 485, pp. 376–380, May 2012.

[34] E. Heitz, "Das Heterochromatin der Moose," vol. 69, pp. 762 – 818+, 1928.

[35] C. L. Woodcock and R. P. Ghosh, "Chromatin Higher-order Structure and Dynamics," *Cold Spring Harbor Perspectives in Biology*, vol. 2, May 2010.

[36] R. Festenstein, S. N. Pagakis, K. Hiragami, D. Lyon, A. Verreault, B. Sekkali, and D. Kioussis, "Modulation of Heterochromatin Protein 1 Dynamics in Primary Mammalian Cells," *Science*, vol. 299, pp. 719–721, Jan. 2003.

[37] T. Cheutin, A. J. McNairn, T. Jenuwein, D. M. Gilbert, P. B. Singh, and T. Misteli, "Maintenance of Stable Heterochromatin Domains by Dynamic HP1 Binding," *Science*, vol. 299, pp. 721–725, Jan. 2003.

[38] P. Trojer and D. Reinberg, "Facultative heterochromatin: is there a distinctive molecular signature?," *Molecular cell*, vol. 28, pp. 1–13, Oct. 2007.

[39] I. Solovei, M. Kreysing, C. Lanctôt, S. Kösem, L. Peichl, T. Cremer, J. Guck, and B. Joffe, "Nuclear architecture of rod photoreceptor cells adapts to vision in mammalian evolution.," *Cell*, vol. 137, pp. 356–368, Apr. 2009.

[40] C. Münkel and J. Langowski, "Chromosome structure predicted by a polymer model," *Physical Review E*, vol. 57, pp. 5888–5896, May 1998.

[41] J. Mateos-Langerak, M. Bohn, W. de Leeuw, O. Giromus, E. M. M. Manders, P. J. Verschure, M. H. G. Indemans, H. J. Gierman, D. W. Heermann, R. van Driel, and S. Goetze, "Spatially confined folding of chromatin in the interphase nucleus," *Proceedings of the National Academy of Sciences*, vol. 106, pp. 3812–3817, Mar. 2009.

[42] L. A. Mirny, "The fractal globule as a model of chromatin architecture in the cell.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 19, pp. 37–51, Jan. 2011.

[43] Grosberg, S. K. Nechaev, and E. I. Shakhnovich, "The role of topological constraints in the kinetics of collapse of macromolecules," *Journal de Physique*, vol. 49, no. 12, pp. 2095–2100, 1988.

[44] A. Grosberg, Y. Rabin, S. Havlin, and A. Neer, "Crumpled Globule Model of the Three-Dimensional Structure of DNA," *EPL (Europhysics Letters)*, vol. 23, pp. 373+, July 2007.

[45] A. Bancaud, S. Huet, N. Daigle, J. Mozziconacci, J. Beaudouin, and J. Ellenberg, "Molecular crowding affects diffusion and binding of nuclear proteins in heterochromatin and reveals the fractal organization of chromatin.," *The EMBO journal*, vol. 28, pp. 3785–3798, Dec. 2009.

[46] J. H. Taylor, P. S. Woods, and W. L. Hughes, "The organization and duplication of chromosomes as revealed by autoradiographic studies using tritium-labeled thymidinee.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 43, pp. 122–128, Jan. 1957.

[47] J. R. Paulson and U. K. Laemmli, "The structure of histone-depleted metaphase chromosomes.," *Cell*, vol. 12, pp. 817–828, Nov. 1977.

[48] K. W. Adolph, S. M. Cheng, and U. K. Laemmli, "Role of nonhistone proteins in metaphase chromosome structure.," *Cell*, vol. 12, pp. 805–816, Nov. 1977.

[49] M. P. Marsden and U. K. Laemmli, "Metaphase chromosome structure: evidence for a radial loop model.," *Cell*, vol. 17, pp. 849–858, Aug. 1979.

[50] C. D. Lewis and U. K. Laemmli, "Higher order metaphase chromosome structure: Evidence for metalloprotein interactions," *Cell*, vol. 29, pp. 171–181, May 1982.

[51] W. C. Earnshaw, B. Halligan, C. A. Cooke, M. M. Heck, and L. F. Liu, "Topoisomerase II is a structural component of mitotic chromosome scaffolds.," *The Journal of cell biology*, vol. 100, pp. 1706–1715, May 1985.

[52] S. M. Gasser, T. Laroche, J. Falquet, E. Boy de la Tour, and U. K. Laemmli, "Metaphase chromosome structure. Involvement of topoisomerase II.," *Journal of molecular biology*, vol. 188, pp. 613–629, Apr. 1986.

[53] N. Saitoh, I. G. Goldberg, E. R. Wood, and W. C. Earnshaw, "ScII: an abundant chromosome scaffold protein is a member of a family of putative ATPases with an unusual predicted tertiary structure.," *The Journal of cell biology*, vol. 127, pp. 303–318, Oct. 1994.

[54] T. Hirano and T. J. Mitchison, "Topoisomerase II does not play a scaffolding role in the organization of mitotic chromosomes assembled in Xenopus egg extracts.," *The Journal of cell biology*, vol. 120, pp. 601–612, Feb. 1993.

[55] W. C. Earnshaw and U. K. Laemmli, "Architecture of metaphase chromosomes and chromosome scaffolds.," *The Journal of cell biology*, vol. 96, pp. 84–93, Jan. 1983.

[56] Y. Saitoh and U. K. Laemmli, "Metaphase chromosome structure: bands arise from a differential folding path of the highly AT-rich scaffold.," *Cell*, vol. 76, pp. 609–622, Feb. 1994.

[57] B. A. Sullivan and W. A. Bickmore, "Unusual chromosome architecture and behaviour at an HSR.," *Chromosoma*, vol. 109, pp. 181–189, June 2000.

[58] Y. G. Strukov, Y. Wang, and A. S. Belmont, "Engineered chromosome regions with altered sequence composition demonstrate hierarchical large-scale folding within metaphase chromosomes.," *The Journal of cell biology*, vol. 162, pp. 23–35, July 2003.

[59] N. Kireeva, M. Lakonishok, I. Kireev, T. Hirano, and A. S. Belmont, "Visualization of early chromosome condensation: a hierarchical folding, axial glue model of chromosome structure.," *The Journal of cell biology*, vol. 166, pp. 775–785, Sept. 2004.

[60] K. A. Hagstrom and B. J. Meyer, "Condensin and cohesin: more than chromosome compactor and glue.," *Nature reviews. Genetics*, vol. 4, pp. 520–534, July 2003.

[61] R. Hand, "Eucaryotic DNA: organization of the genome for replication.," *Cell*, vol. 15, pp. 317–325, Oct. 1978.

[62] S. Tanaka, R. Nakato, Y. Katou, K. Shirahige, and H. Araki, "Origin Association of Sld3, Sld7, and Cdc45 Proteins Is a Key Step for Determination of Origin-Firing Timing," *Curr Biol*, vol. 21, pp. 2055–2063, Dec. 2011.

[63] A. Tazumi, M. Fukuura, R. Nakato, A. Kishimoto, T. Takenaka, S. Ogawa, J.-h. Song, T. S. Takahashi, T. Nakagawa, K. Shirahige, and H. Masukata, "Telomere-binding protein Taz1 controls global replication timing through its localization near late replication origins in fission yeast," *Genes & Development*, vol. 26, pp. 2050–2062, Sept. 2012.

[64] M. Hayano, Y. Kanoh, S. Matsumoto, C. Renard-Guillet, K. Shirahige, and H. Masai, "Rif1 is a global regulator of timing of replication origin firing in fission yeast.," *Genes & development*, vol. 26, pp. 137–150, Jan. 2012.

[65] S. Yamazaki, A. Ishii, Y. Kanoh, M. Oda, Y. Nishito, and H. Masai, "Rif1 regulates the replication timing domains on the human genome.," *The EMBO journal*, vol. 31, pp. 3667–3677, Sept. 2012.

[66] S. Farkash-Amar, D. Lipson, A. Polten, A. Goren, C. Helmstetter, Z. Yakhini, and I. Simon, "Global organization of replication time zones of the mouse genome.," *Genome research*, vol. 18, pp. 1562–1570, Oct. 2008.

[67] G. P. Holmquist, "Role of replication time in the control of tissue-specific gene expression.," *American journal of human genetics*, vol. 40, pp. 151–173, Feb. 1987.

[68] M. K. Raghuraman, E. A. Winzeler, D. Collingwood, S. Hunt, L. Wodicka, A. Conway, D. J. Lockhart, R. W. Davis, B. J. Brewer, and W. L. Fangman, "Replication dynamics of the yeast genome.," *Science (New York, N.Y.)*, vol. 294, pp. 115–121, Oct. 2001.

[69] K. Woodfine, H. Fiegler, D. M. Beare, J. E. Collins, O. T. McCann, B. D. Young, S. Debernardi, R. Mott, I. Dunham, and N. P. Carter, "Replication timing of the human genome," *Human Molecular Genetics*, vol. 13, pp. 191–202, Jan. 2004.

[70] K. Woodfine, D. M. Beare, K. Ichimura, S. Debernardi, A. J. Mungall, H. Fiegler, P. P. Collins, N. P. Carter, and I. Dunham, "Replication timing of human chromosome 6.," *Cell cycle (Georgetown, Tex.)*, vol. 4, pp. 172–176, Jan. 2005.

[71] D. Schubeler, D. Scalzo, C. Kooperberg, B. van Steensel, J. Delrow, and M. Groudine, "Genome-wide DNA replication profile for Drosophila melanogaster: a link

between transcription and replication timing," *Nat Genet*, vol. 32, pp. 438–442, Nov. 2002.

[72] Y. Watanabe, A. Fujiyama, Y. Ichiba, M. Hattori, T. Yada, Y. Sakaki, and T. Ikemura, "Chromosome-wide assessment of replication timing for human chromosomes 11q and 21q: disease-related genes in timing-switch regions.," *Human molecular genetics*, vol. 11, pp. 13–21, Jan. 2002.

[73] Y. Jeon, S. Bekiranov, N. Karnani, P. Kapranov, S. Ghosh, D. MacAlpine, C. Lee, D. S. Hwang, T. R. Gingeras, and A. Dutta, "Temporal profile of replication of human chromosomes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, pp. 6419–6424, May 2005.

[74] N. Karnani, C. Taylor, A. Malhotra, and A. Dutta, "Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas," *Genome Research*, vol. 17, pp. 865–876, June 2007.

[75] I. Hiratani, T. Ryba, M. Itoh, T. Yokochi, M. Schwaiger, C.-W. W. Chang, Y. Lyou, T. M. Townes, D. Schübeler, and D. M. Gilbert, "Global reorganization of replication domains during embryonic stem cell differentiation.," *PLoS biology*, vol. 6, pp. e245+, Oct. 2008.

[76] T. Ryba, I. Hiratani, J. Lu, M. Itoh, M. Kulik, J. Zhang, T. C. Schulz, A. J. Robins, S. Dalton, and D. M. Gilbert, "Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types.," *Genome research*, vol. 20, pp. 761–770, June 2010.

[77] H. F. Jørgensen, V. Azuara, S. Amoils, M. Spivakov, A. Terry, T. Nesterova, B. S. Cobb, B. Ramsahoye, M. Merkenschlager, and A. G. Fisher, "The impact of chromatin modifiers on the timing of locus replication in mouse embryonic stem cells.," *Genome biology*, vol. 8, pp. R169+, Aug. 2007.

[78] M. Thornton, K. L. Eward, and C. E. Helmstetter, "Production of minimally disturbed synchronous cultures of hematopoietic cells.," *BioTechniques*, vol. 32, May 2002.

[79] C. E. Helmstetter, M. Thornton, A. Romero, and K. L. Eward, "Synchrony in human, mouse and bacterial cell cultures–a comparison.," *Cell cycle (Georgetown, Tex.)*, vol. 2, no. 1, pp. 42–45, 2003.

[80] E. Yaffe, S. Farkash-Amar, A. Polten, Z. Yakhini, A. Tanay, and I. Simon, "Comparative analysis of DNA replication timing reveals conserved large-scale chromosomal architecture.," *PLoS genetics*, vol. 6, pp. e1001011+, July 2010.

[81] I. Yanai, H. Benjamin, M. Shmoish, V. Chalifa-Caspi, M. Shklar, R. Ophir, A. Bar-Even, S. Horn-Saban, M. Safran, E. Domany, D. Lancet, and O. Shmueli, "Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification.," *Bioinformatics (Oxford, England)*, vol. 21, pp. 650–659, Mar. 2005.

[82] J. Zhang, F. Xu, T. Hashimshony, I. Keshet, and H. Cedar, "Establishment of transcriptional competence in early and late S phase.," *Nature*, vol. 420, pp. 198–202, Nov. 2002.

[83] S. Selig, K. Okumura, D. C. Ward, and H. Cedar, "Delineation of DNA replication time zones by fluorescence in situ hybridization.," *The EMBO journal*, vol. 11, pp. 1217–1225, Mar. 1992.

[84] I. Simon, T. Tenzen, R. Mostoslavsky, E. Fibach, L. Lande, E. Milot, J. Gribnau, F. Grosveld, P. Fraser, and H. Cedar, "Developmental regulation of DNA replication timing at the human beta globin locus.," *The EMBO journal*, vol. 20, pp. 6150–6157, Nov. 2001.

[85] P. Perry, S. Sauer, N. Billon, W. D. Richardson, M. Spivakov, G. Warnes, F. J. Livesey, M. Merkenschlager, A. G. Fisher, and V. Azuara, "A dynamic switch in the replication timing of key regulator genes in embryonic stem cells upon neural induction.," *Cell cycle (Georgetown, Tex.)*, vol. 3, pp. 1645–1650, Dec. 2004.

[86] D. A. Jackson, "S-phase progression in synchronized human cells.," *Experimental cell research*, vol. 220, pp. 62–70, Sept. 1995.

[87] D. A. Jackson, "Nuclear organization: Uniting replication foci, chromatin domains and chromosome structure," *BioEssays*, vol. 17, no. 7, pp. 587–591, 1995.

[88] A. Maya-Mendoza, P. Olivares-Chauvet, A. Shaw, and D. A. Jackson, "S phase progression in human cells is dictated by the genetic continuity of DNA foci.," *PLoS Genetics*, vol. 6, pp. e1000900+, Apr. 2010.

[89] J. Schindelin, I. Arganda-Carreras, E. Frise, V. Kaynig, M. Longair, T. Pietzsch, S. Preibisch, C. Rueden, S. Saalfeld, B. Schmid, J.-Y. Y. Tinevez, D. J. J. White, V. Hartenstein, K. Eliceiri, P. Tomancak, and A. Cardona, "Fiji: an open-source platform for biological-image analysis.," *Nature methods*, vol. 9, pp. 676–682, July 2012.

[90] E. Iannuccelli, F. Mompart, J. Gellin, Y. Lahbib-Mansais, M. Yerle, and T. Boudier, "NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-FISH experiments," *Bioinformatics*, vol. 26, pp. 696–697, Mar. 2010.

[91] C. A. Schneider, W. S. Rasband, and K. W. Eliceiri, "NIH Image to ImageJ: 25 years of image analysis," *Nature Methods*, vol. 9, pp. 671–675, June 2012.

[92] S. Bolte and F. P. Cordelières, "A guided tour into subcellular colocalization analysis in light microscopy," *Journal of Microscopy*, vol. 224, pp. 213–232, Dec. 2006.

[93] R Development Core Team, *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

[94] J. Kent, "The source tree - Genomewiki."

[95] D. Karolchik, R. Baertsch, M. Diekhans, T. S. Furey, A. Hinrichs, Y. T. Lu, K. M. Roskin, M. Schwartz, C. W. Sugnet, D. J. Thomas, R. J. Weber, D. Haussler, W. J. Kent, and University of California Santa Cruz, "The UCSC Genome Browser Database.," *Nucleic acids research*, vol. 31, pp. 51–54, Jan. 2003.

[96] T. R. Dreszer, D. Karolchik, A. S. Zweig, A. S. Hinrichs, B. J. Raney, R. M. Kuhn, L. R. Meyer, M. Wong, C. A. Sloan, K. R. Rosenbloom, G. Roe, B. Rhead,

A. Pohl, V. S. Malladi, C. H. Li, K. Learned, V. Kirkup, F. Hsu, R. A. Harte, L. Guruvadoo, M. Goldman, B. M. Giardine, P. A. Fujita, M. Diekhans, M. S. Cline, H. Clawson, G. P. Barber, D. Haussler, and W. James Kent, "The UCSC Genome Browser database: extensions and updates 2011.," *Nucleic acids research*, vol. 40, Jan. 2012.

[97] T. S. Furey and D. Haussler, "Integration of the cytogenetic map with the draft human genome sequence.," *Human molecular genetics*, vol. 12, pp. 1037–1044, May 2003.

[98] M. Gardiner-Garden and M. Frommer, "CpG Islands in vertebrate genomes," *Journal of Molecular Biology*, vol. 196, pp. 261–282, July 1987.

[99] T. Schmidt and D. Frishman, "Assignment of isochores for all completely sequenced vertebrate genomes using a consensus.," *Genome biology*, vol. 9, pp. R104+, June 2008.

[100] A. F. A. Smit, R. Hubley, and P. Green, "RepeatMasker Open-3.0," 1996-2004.

[101] J. Jurka, "Repbase Update: a database and an electronic journal of repetitive elements," *Trends in Genetics*, vol. 16, pp. 418–420, Sept. 2000.

[102] N. Weddington, A. Stuy, I. Hiratani, T. Ryba, T. Yokochi, and D. M. Gilbert, "ReplicationDomain: a visualization tool and comparative database for genome-wide replication timing data.," *BMC bioinformatics*, vol. 9, no. 1, 2008.

[103] R. Desprat, D. Thierry-Mieg, N. Lailler, J. Lajugie, C. Schildkraut, J. Thierry-Mieg, and E. E. Bouhassira, "Predictable dynamic program of timing of DNA replication in human cells," *Genome Research*, vol. 19, pp. 2288–2299, Dec. 2009.

[104] U. Francke, "Digitized and differentially shaded human chromosome ideograms for genomic applications.," *Cytogenetics and cell genetics*, vol. 65, no. 3, pp. 206–218, 1994.

[105] P. J. Sabo, M. Hawrylycz, J. C. Wallace, R. Humbert, M. Yu, A. Shafer, J. Kawamoto, R. Hall, J. Mack, M. O. Dorschner, M. McArthur, and J. A. Stam-

atoyannopoulos, “Discovery of functional noncoding elements by digital analysis of chromatin structure.,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, pp. 16837–16842, Nov. 2004.

[106] P. J. Sabo, M. S. Kuehn, R. Thurman, B. E. Johnson, E. M. Johnson, H. Cao, M. Yu, E. Rosenzweig, J. Goldy, A. Haydock, M. Weaver, A. Shafer, K. Lee, F. Neri, R. Humbert, M. A. Singer, T. A. Richmond, M. O. Dorschner, M. McArthur, M. Hawrylycz, R. D. Green, P. A. Navas, W. S. Noble, and J. A. Stamatoyannopoulos, “Genome-scale mapping of DNase I sensitivity in vivo using tiling DNA microarrays,” *Nature Methods*, vol. 3, pp. 511–518, June 2006.

[107] A. Németh, A. Conesa, J. Santoyo-Lopez, I. Medina, D. Montaner, B. Péterfia, I. Solovei, T. Cremer, J. Dopazo, and G. Längst, “Initial Genomics of the Human Nucleolus,” *PLoS Genet*, vol. 6, pp. e1000889+, Mar. 2010.

[108] L. Guelen, L. Pagie, E. Brasset, W. Meuleman, M. B. Faza, W. Talhout, B. H. Eussen, A. de Klein, L. Wessels, W. de Laat, and B. van Steensel, “Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions.,” *Nature*, vol. 453, pp. 948–951, June 2008.

[109] V. G. Cheung, N. Nowak, W. Jang, I. R. Kirsch, S. Zhao, X. N. Chen, T. S. Furey, U. J. Kim, W. L. Kuo, M. Olivier, J. Conroy, A. Kasprzyk, H. Massa, R. Yonescu, S. Sait, C. Thoreen, A. Snijders, E. Lemyre, J. A. Bailey, A. Bruzel, W. D. Burrill, S. M. Clegg, S. Collins, P. Dhami, C. Friedman, C. S. Han, S. Herrick, J. Lee, A. H. Ligon, S. Lowry, M. Morley, S. Narasimhan, K. Osoegawa, Z. Peng, I. Plajzer-Frick, B. J. Quade, D. Scott, K. Sirotkin, A. A. Thorpe, J. W. Gray, J. Hudson, D. Pinkel, T. Ried, L. Rowen, G. L. Shen-Ong, R. L. Strausberg, E. Birney, D. F. Callen, J. F. Cheng, D. R. Cox, N. A. Doggett, N. P. Carter, E. E. Eichler, D. Haussler, J. R. Korenberg, C. C. Morton, D. Albertson, G. Schuler, P. J. de Jong, B. J. Trask, and BAC Resource Consortium, “Integration of cytogenetic landmarks into the draft sequence of the human genome.,” *Nature*, vol. 409, pp. 953–958, Feb. 2001.

[110] D. M. Witten and R. Tibshirani, "A framework for feature selection in clustering.," *Journal of the American Statistical Association*, vol. 105, pp. 713–726, June 2010.

[111] J. Ernst and M. Kellis, "Discovery and characterization of chromatin states for systematic annotation of the human genome.," *Nature biotechnology*, vol. 28, pp. 817–825, Aug. 2010.

[112] R. K. Sachs, G. van den Engh, B. Trask, H. Yokota, and J. E. Hearst, "A random-walk/giant-loop model for interphase chromosomes.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, pp. 2710–2714, Mar. 1995.

[113] H. Yokota, G. van den Engh, J. E. Hearst, R. K. Sachs, and B. J. Trask, "Evidence for the organization of chromatin in megabase pair-sized loops arranged along a random walk path in the human G0/G1 interphase nucleus.," *The Journal of Cell Biology*, vol. 130, pp. 1239–1249, Sept. 1995.

[114] C. S. Osborne, L. Chakalova, K. E. Brown, D. Carter, A. Horton, E. Debrand, B. Goyenechea, J. A. Mitchell, S. Lopes, W. Reik, and P. Fraser, "Active genes dynamically colocalize to shared sites of ongoing transcription," *Nature Genetics*, vol. 36, pp. 1065–1071, Sept. 2004.

[115] S. Schoenfelder, T. Sexton, L. Chakalova, N. F. Cope, A. Horton, S. Andrews, S. Kurukuti, J. A. Mitchell, D. Umlauf, D. S. Dimitrova, C. H. Eskiw, Y. Luo, C.-L. L. Wei, Y. Ruan, J. J. Bieker, and P. Fraser, "Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells.," *Nature genetics*, vol. 42, pp. 53–61, Jan. 2010.

[116] S. Lomvardas, G. Barnea, D. J. Pisapia, M. Mendelsohn, J. Kirkland, and R. Axel, "Interchromosomal interactions and olfactory receptor choice.," *Cell*, vol. 126, pp. 403–413, July 2006.

[117] J. Q. Q. Ling, T. Li, J. F. F. Hu, T. H. Vu, H. L. L. Chen, X. W. W. Qiu, A. M. Cherry, and A. R. Hoffman, "CTCF mediates interchromosomal colocalization between

Igf2/H19 and Wsb1/Nf1.," *Science (New York, N.Y.)*, vol. 312, pp. 269–272, Apr. 2006.

[118] B. Tolhuis, R. J. J. Palstra, E. Splinter, F. Grosveld, and W. de Laat, "Looping and interaction between hypersensitive sites in the active beta-globin locus.," *Molecular cell*, vol. 10, pp. 1453–1465, Dec. 2002.

[119] C. G. Spilianakis, M. D. Lalioti, T. Town, G. R. Lee, and R. A. Flavell, "Interchromosomal associations between alternatively expressed loci," *Nature*, vol. 435, pp. 637–645, May 2005.

[120] R. Kalhor, H. Tjong, N. Jayathilaka, F. Alber, and L. Chen, "Genome architectures revealed by tethered chromosome conformation capture and population-based modeling," *Nat Biotech*, vol. 30, pp. 90–98, Jan. 2012.

[121] J. E. Phillips and V. G. Corces, "CTCF: master weaver of the genome.," *Cell*, vol. 137, pp. 1194–1211, June 2009.

[122] H. Dehghani, G. Dellaire, and D. P. Bazett-Jones, "Organization of chromatin in the interphase mammalian cell.," *Micron (Oxford, England : 1993)*, vol. 36, pp. 95–108, Feb. 2005.

[123] R. D. Phair and T. Misteli, "High mobility of proteins in the mammalian cell nucleus.," *Nature*, vol. 404, pp. 604–609, Apr. 2000.

[124] P. J. Verschure, I. van Der Kraan, E. M. Manders, and R. van Driel, "Spatial relationship between transcription sites and chromosome territories.," *The Journal of cell biology*, vol. 147, pp. 13–24, Oct. 1999.

[125] N. L. Mahy, P. E. Perry, S. Gilchrist, R. A. Baldock, and W. A. Bickmore, "Spatial organization of active and inactive genes and noncoding DNA within chromosome territories.," *The Journal of cell biology*, vol. 157, pp. 579–589, May 2002.

[126] A. S. Belmont and K. Bruce, "Visualization of G1 chromosomes: a folded, twisted, supercoiled chromonema model of interphase chromatid structure.," *The Journal of cell biology*, vol. 127, pp. 287–302, Oct. 1994.

[127] R. M. Zirbel, U. R. Mathieu, A. Kurz, T. Cremer, and P. Lichter, "Evidence for a nuclear compartment of transcription and splicing located at chromosome domain boundaries.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 1, pp. 93–106, July 1993.

[128] A. Kurz, S. Lampel, J. E. Nickolenko, J. Bradl, A. Benner, R. M. Zirbel, T. Cremer, and P. Lichter, "Active and inactive genes localize preferentially in the periphery of chromosome territories.," *The Journal of cell biology*, vol. 135, pp. 1195–1205, Dec. 1996.

[129] J. M. Bridger, H. Herrmann, C. Münkel, and P. Lichter, "Identification of an interchromosomal compartment by polymerization of nuclear-targeted vimentin.," *Journal of cell science*, vol. 111 (Pt 9), pp. 1241–1253, May 1998.

[130] M. Reichenzeller, A. Burzlaff, P. Lichter, and H. Herrmann, "In vivo observation of a nuclear channel-like system: evidence for a distinct interchromosomal domain compartment in interphase cells.," *Journal of structural biology*, vol. 129, pp. 175–185, Apr. 2000.

[131] J. M. Bridger, C. Kalla, H. Wodrich, S. Weitz, J. A. King, K. Khazaie, H.-G. G. Kräusslich, and P. Lichter, "Nuclear RNAs confined to a reticular compartment between chromosome territories.," *Experimental cell research*, vol. 302, pp. 180–193, Jan. 2005.

[132] K. Richter, M. Reichenzeller, S. M. Görisch, U. Schmidt, M. O. Scheuermann, H. Herrmann, and P. Lichter, "Characterization of a nuclear compartment shared by nuclear bodies applying ectopic protein expression and correlative light and electron microscopy.," *Experimental cell research*, vol. 303, pp. 128–137, Feb. 2005.

[133] D. Cmarko, P. J. Verschure, T. E. Martin, M. E. Dahmus, S. Krause, X. D. Fu, R. van Driel, and S. Fakan, "Ultrastructural analysis of transcription and splicing in the cell nucleus after bromo-UTP microinjection.," *Molecular biology of the cell*, vol. 10, pp. 211–223, Jan. 1999.

[134] M. O. Scheuermann, J. Tajbakhsh, A. Kurz, K. Saracoglu, R. Eils, and P. Lichter, "Topology of genes and nontranscribed sequences in human interphase nuclei.," *Experimental cell research*, vol. 301, pp. 266–279, Dec. 2004.

[135] K. Küpper, A. Kölbl, D. Biener, S. Dittrich, J. von Hase, T. Thormeyer, H. Fiegler, N. P. Carter, M. R. Speicher, T. Cremer, and M. Cremer, "Radial chromatin positioning is shaped by local gene density, not by gene expression.," *Chromosoma*, vol. 116, pp. 285–306, June 2007.

[136] R. Berezney, D. D. Dubey, and J. A. Huberman, "Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci.," *Chromosoma*, vol. 108, pp. 471–484, Mar. 2000.

[137] A. E. Visser, F. Jaunin, S. Fakan, and J. A. Aten, "High resolution analysis of interphase chromosome domains," *Journal of Cell Science*, vol. 113, pp. 2585–2593, July 2000.

[138] H. Albiez, M. Cremer, C. Tiberi, L. Vecchio, L. Schermelleh, S. Dittrich, K. Küpper, B. Joffe, T. Thormeyer, J. von Hase, S. Yang, K. Rohr, H. Leonhardt, I. Solovei, C. Cremer, S. Fakan, and T. Cremer, "Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 14, pp. 707–733, Nov. 2006.

[139] J. Rouquette, C. Genoud, G. H. Vazquez-Nin, B. Kraus, T. Cremer, and S. Fakan, "Revealing the high-resolution three-dimensional network of chromatin and interchromatin space: a novel electron-microscopic approach to reconstructing nuclear architecture.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 17, pp. 801–810, Aug. 2009.

[140] S. Fakan and R. van Driel, "The perichromatin region: a functional compartment in the nucleus that determines large-scale chromatin folding.," *Seminars in cell & developmental biology*, vol. 18, pp. 676–681, Oct. 2007.

[141] M. R. Branco and A. Pombo, "Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations," *PLoS Biol*, vol. 4, pp. e138+, Apr. 2006.

[142] J. Arsuaga, K. M. Greulich-Bode, M. Vazquez, M. Bruckner, P. Hahnfeldt, D. J. Brenner, R. Sachs, and L. Hlatky, "Chromosome spatial clustering inferred from radiogenic aberrations.," *International journal of radiation biology*, vol. 80, pp. 507–515, July 2004.

[143] M. R. Branco and A. Pombo, "Chromosome organization: new facts, new models.," *Trends in cell biology*, vol. 17, pp. 127–134, Mar. 2007.

[144] D. Zink, T. Cremer, R. Saffrich, R. Fischer, M. F. Trendelenburg, W. Ansorge, and E. H. K. Stelzer, "Structure and dynamics of human interphase chromosome territories in vivo," *Human Genetics*, vol. 102, pp. 241–251, Feb. 1998.

[145] O. Ronneberger, D. Baddeley, F. Scheipl, P. J. Verveer, H. Burkhardt, C. Cremer, L. Fahrmeir, T. Cremer, and B. Joffe, "Spatial quantitative analysis of fluorescently labeled nuclear structures: problems, methods, pitfalls.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 16, pp. 523–562, May 2008.

[146] M. Yoshida, S. Horinouchi, and T. Beppu, "Trichostatin A and trapoxin: novel chemical probes for the role of histone acetylation in chromatin structure and function.," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 17, pp. 423–430, May 1995.

[147] A. Maya-Mendoza, E. Petermann, D. A. Gillespie, K. W. Caldecott, and D. A. Jackson, "Chk1 regulates the density of active replication origins during the vertebrate S phase.," *The EMBO journal*, vol. 26, pp. 2719–2731, June 2007.

[148] A. E. Visser and J. A. Aten, "Chromosomes as well as chromosomal subdomains constitute distinct units in interphase nuclei.," *Journal of cell science*, vol. 112 (Pt 19), pp. 3353–3360, Oct. 1999.

[149] S. Goetze, J. Mateos-Langerak, H. J. Gierman, W. de Leeuw, O. Giromus, M. H. Indemans, J. Koster, V. Ondrej, R. Versteeg, and R. van Driel, "The three-dimensional structure of human interphase chromosomes is related to the transcriptome map.," *Molecular and cellular biology*, vol. 27, pp. 4475–4487, June 2007.

[150] S. Boyle, M. J. Rodesch, H. A. Halvensleben, J. A. Jeddeloh, and W. A. Bickmore, "Fluorescence in situ hybridization with high-complexity repeat-free oligonucleotide probes generated by massively parallel synthesis.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 19, pp. 901–909, Oct. 2011.

[151] P. Fraser and W. Bickmore, "Nuclear organization of the genome and the potential for gene regulation.," *Nature*, vol. 447, pp. 413–417, May 2007.

[152] S. Schoenfelder, I. Clay, and P. Fraser, "The transcriptional interactome: gene expression in 3D," *Current Opinion in Genetics & Development*, vol. 20, pp. 127–133, Apr. 2010.

[153] C. G. Spilianakis and R. A. Flavell, "Long-range intrachromosomal interactions in the T helper type 2 cytokine locus.," *Nature immunology*, vol. 5, pp. 1017–1027, Oct. 2004.

[154] C. P. Bacher, M. Guggiari, B. Brors, S. Augui, P. Clerc, P. Avner, R. Eils, and E. Heard, "Transient colocalization of X-inactivation centres accompanies the initiation of X inactivation.," *Nature cell biology*, vol. 8, pp. 293–299, Mar. 2006.

[155] N. Xu, C.-L. Tsai, and J. T. Lee, "Transient Homologous Chromosome Pairing Marks the Onset of X Inactivation," *Science*, vol. 311, pp. 1149–1152, Feb. 2006.

[156] H. Würtele and P. Chartrand, "Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture

methodology.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 14, pp. 477–495, July 2006.

[157] M. Simonis, P. Klous, E. Splinter, Y. Moshkin, R. Willemsen, E. de Wit, B. van Steensel, and W. de Laat, "Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation captureâĂŞon-chip (4C)," *Nature Genetics*, vol. 38, pp. 1348–1354, Oct. 2006.

[158] Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson, "Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions.," *Nature genetics*, vol. 38, pp. 1341–1347, Nov. 2006.

[159] C. S. Osborne, L. Chakalova, J. A. Mitchell, A. Horton, A. L. Wood, D. J. Bolland, A. E. Corcoran, and P. Fraser, "Myc dynamically and preferentially relocates to a transcription factory occupied by Igh.," *PLoS biology*, vol. 5, pp. e192+, Aug. 2007.

[160] J. H. Tjio and A. Levan, "The chromosome number of man," *Hereditas*, vol. 42, no. 1-2, pp. 1–6, 1956.

[161] J. Lejeune, M. Gautier, and R. Turpin, "Etude des chromosomes somatiques de neuf enfants mongoliens," *Comptes rendus hebdomadaires des séances de l'Académie des sciences*, 1959.

[162] T. Caspersson, S. Farber, G. E. Foley, J. Kudynowski, E. J. Modest, E. Simonsson, U. Wagh, and L. Zech, "Chemical differentiation along metaphase chromosomes.," *Experimental cell research*, vol. 49, pp. 219–222, Jan. 1968.

[163] "Paris Conference (1971): Standardization in human cytogenetics.," *Cytogenetics*, vol. 11, no. 5, pp. 317–362, 1972.

[164] "Paris Conference (1971), supplement (1975) Standardization in human cytogenetics.," *Cytogenetics and cell genetics*, vol. 15, no. 4, pp. 203–238, 1975.

[165] L. Willatt and S. M. Morgan, "Shaffer LG, Slovak ML, Campbell LJ (2009): ISCN 2009 an international system for human cytogenetic nomenclature," *Human Genetics*, vol. 126, no. 4, pp. 603–604.

[166] J. G. Gall and M. L. Pardue, "Formation and detection of RNA-DNA hybrid molecules in cytological preparations.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 63, pp. 378–383, June 1969.

[167] G. T. Rudkin and B. D. Stollar, "High resolution detection of DNA-RNA hybrids in situ by indirect immunofluorescence.," *Nature*, vol. 265, pp. 472–473, Feb. 1977.

[168] D. E. Comings, "Mechanisms of chromosome banding and implications for chromosome structure.," *Annual review of genetics*, vol. 12, pp. 25–46, 1978.

[169] J. J. Yunis, J. R. Sawyer, and D. W. Ball, "The characterization of high-resolution G-banded chromosomes of man," *Chromosoma*, vol. 67, pp. 293–307, Dec. 1978.

[170] M. Costantini, O. Clay, C. Federico, S. Saccone, F. Auletta, and G. Bernardi, "Human chromosomal bands: nested structure, high-definition map and molecular basis," *Chromosoma*, vol. 116, pp. 29–40, Feb. 2007.

[171] N. Kosyakova, A. Weise, K. Mrasek, U. Claussen, T. Liehr, and H. Nelle, "The hierarchically organized splitting of chromosomal bands for all human chromosomes," *Molecular Cytogenetics*, vol. 2, no. 1, pp. 4+, 2009.

[172] D. E. Comings and E. Avelino, "Mechanisms of chromosome banding," *Chromosoma*, vol. 51, pp. 365–379, Dec. 1975.

[173] G. Holmquist, M. Gray, T. Porter, and J. Jordan, "Characterization of Giemsa dark- and light-band DNA.," *Cell*, vol. 31, pp. 121–129, Nov. 1982.

[174] G. Bernardi, "The isochore organization of the human genome.," *Annual review of genetics*, vol. 23, pp. 637–661, 1989.

[175] K. Gardiner, B. Aissani, and G. Bernardi, "A compositional map of human chromosome 21.," *The EMBO journal*, vol. 9, pp. 1853–1858, June 1990.

[176] G. P. Holmquist, "Chromosome bands, their chromatin flavors, and their functional features.," *American journal of human genetics*, vol. 51, pp. 17–37, July 1992.

[177] J. M. Craig and W. A. Bickmore, "Chromosome bands–flavours to savour.," *BioEssays : news and reviews in molecular, cellular and developmental biology*, vol. 15, pp. 349–354, May 1993.

[178] H. Yokota, M. J. Singer, G. J. van den Engh, and B. J. Trask, "Regional differences in the compaction of chromatin in human G0/G1 interphase nuclei.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 5, pp. 157–166, May 1997.

[179] C. Carvalho, H. M. Pereira, J. Ferreira, C. Pina, D. Mendonça, A. C. Rosa, and M. Carmo-Fonseca, "Chromosomal G-dark bands determine the spatial organization of centromeric heterochromatin in the nucleus.," *Molecular biology of the cell*, vol. 12, pp. 3563–3572, Nov. 2001.

[180] M. J. Lercher, A. O. Urrutia, and L. D. Hurst, "Clustering of housekeeping genes provides a unified model of gene order in the human genome," *Nat Genet*, vol. 31, pp. 180–183, June 2002.

[181] M. J. Lercher, A. O. Urrutia, A. Pavlícek, and L. D. Hurst, "A unification of mosaic structures in the human genome.," *Human molecular genetics*, vol. 12, pp. 2411–2415, Oct. 2003.

[182] R. Versteeg, B. D. C. van Schaik, M. F. van Batenburg, M. Roos, R. Monajemi, H. Caron, H. J. Bussemaker, and A. H. C. van Kampen, "The human transcriptome map reveals extremes in gene density, intron length, GC content, and repeat pattern for domains of highly and weakly expressed genes.," *Genome research*, vol. 13, pp. 1998–2004, Sept. 2003.

[183] Y. Niimura and T. Gojobori, "In silico chromosome staining: reconstruction of Giemsa bands from the whole human genome sequence.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 797–802, Jan. 2002.

[184] I. R. Kirsch, E. D. Green, R. Yonescu, R. Strausberg, N. Carter, D. Bentley, M. A. Leversha, I. Dunham, V. V. Braden, E. Hilgenfeld, G. Schuler, A. E. Lash, G. L. Shen, M. Martelli, W. M. Kuehl, R. D. Klausner, and T. Ried, "A systematic, high-resolution linkage of the cytogenetic and physical maps of the human genome," *Nat Genet*, vol. 24, pp. 339–340, Apr. 2000.

[185] M. Costantini, O. Clay, F. Auletta, and G. Bernardi, "An isochore map of human chromosomes.," *Genome research*, vol. 16, pp. 536–541, Apr. 2006.

[186] R. McGill, J. W. Tukey, and W. A. Larsen, "Variations of Box Plots," *The American Statistician*, vol. 32, no. 1, pp. 12–16, 1978.

[187] J. M. Craig and W. A. Bickmore, "The distribution of CpG islands in mammalian chromosomes.," *Nature genetics*, vol. 7, pp. 376–382, July 1994.

[188] J. Filipski, J. P. Thiery, and G. Bernardi, "An analysis of the bovine genome by Cs2SO4-Ag density gradient centrifugation.," *Journal of molecular biology*, vol. 80, pp. 177–197, Oct. 1973.

[189] C. T. Zhang, J. Wang, and R. Zhang, "A novel method to calculate the G+C content of genomic DNA sequences.," *Journal of biomolecular structure & dynamics*, vol. 19, pp. 333–341, Oct. 2001.

[190] V. E. Ramensky, V. J. Makeev, M. A. Roytberg, and V. G. Tumanyan, "Segmentation of long genomic sequences into domains with homogeneous composition with BASIO software.," *Bioinformatics (Oxford, England)*, vol. 17, pp. 1065–1066, Nov. 2001.

[191] J. L. Oliver, P. Carpena, M. Hackenberg, and P. Bernaola-Galván, "IsoFinder: computational prediction of isochores in genome sequences.," *Nucleic acids research*, vol. 32, July 2004.

[192] C.-T. T. Zhang, F. Gao, and R. Zhang, "Segmentation algorithm for DNA sequences.," *Physical review. E, Statistical, nonlinear, and soft matter physics*, vol. 72, Oct. 2005.

[193] N. Haiminen and H. Mannila, "Discovering isochores by least-squares optimal segmentation.," *Gene*, vol. 394, pp. 53–60, June 2007.

[194] A. P. J. de Koning, W. Gu, T. A. Castoe, M. A. Batzer, and D. D. Pollock, "Repetitive Elements May Comprise Over Two-Thirds of the Human Genome," *PLoS Genet*, vol. 7, pp. e1002384+, Dec. 2011.

[195] A. Barski, S. Cuddapah, K. Cui, T.-Y. Roh, D. E. Schones, Z. Wang, G. Wei, I. Chepelev, and K. Zhao, "High-Resolution Profiling of Histone Methylations in the Human Genome," *Cell*, vol. 129, pp. 823–837, May 2007.

[196] T.-Y. Roh, S. Cuddapah, and K. Zhao, "Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping," *Genes & Development*, vol. 19, pp. 542–552, Mar. 2005.

[197] T.-Y. Roh, S. Cuddapah, K. Cui, and K. Zhao, "The genomic landscape of histone modifications in human T cells," *Proceedings of the National Academy of Sciences*, vol. 103, pp. 15782–15787, Oct. 2006.

[198] N. D. Heintzman, R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford, and B. Ren, "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome," *Nat Genet*, vol. 39, pp. 311–318, Mar. 2007.

[199] S. Chambeyron and W. A. Bickmore, "Chromatin decondensation and nuclear reorganization of the HoxB locus upon induction of transcription.," *Genes & development*, vol. 18, pp. 1119–1130, May 2004.

[200] B. E. Bernstein, T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber, and E. S. Lander, "A bivalent chromatin structure marks key developmental genes in embryonic stem cells.," *Cell*, vol. 125, pp. 315–326, Apr. 2006.

[201] A. Kirmizis, H. Santos-Rosa, C. J. Penkett, M. A. Singer, R. D. Green, and T. Kouzarides, "Distinct transcriptional outputs associated with mono- and

dimethylated histone H3 arginine 2," *Nat Struct Mol Biol*, vol. 16, pp. 449–451, Apr. 2009.

[202] W. An, J. Kim, and R. G. Roeder, "Ordered cooperative functions of PRMT1, p300, and CARM1 in transcriptional activation by p53.," *Cell*, vol. 117, pp. 735–748, June 2004.

[203] D. K. Pokholok, C. T. Harbison, S. Levine, M. Cole, N. M. Hannett, T. I. I. Lee, G. W. Bell, K. Walker, P. A. Rolfe, E. Herbolsheimer, J. Zeitlinger, F. Lewitter, D. K. Gifford, and R. A. Young, "Genome-wide map of nucleosome acetylation and methylation in yeast.," *Cell*, vol. 122, pp. 517–527, Aug. 2005.

[204] D. Schübeler, D. M. MacAlpine, D. Scalzo, C. Wirbelauer, C. Kooperberg, F. van Leeuwen, D. E. Gottschling, L. P. O'Neill, B. M. Turner, J. Delrow, S. P. Bell, and M. Groudine, "The histone modification pattern of active genes revealed through genome-wide chromatin analysis of a higher eukaryote.," *Genes & development*, vol. 18, pp. 1263–1271, June 2004.

[205] A. Morillon, N. Karabetsou, A. Nair, and J. Mellor, "Dynamic lysine methylation on histone H3 defines the regulatory phase of gene transcription.," *Molecular cell*, vol. 18, pp. 723–734, June 2005.

[206] A. J. Bannister, R. Schneider, F. A. Myers, A. W. Thorne, C. Crane-Robinson, and T. Kouzarides, "Spatial distribution of di- and tri-methyl lysine 36 of histone H3 at active genes.," *The Journal of biological chemistry*, vol. 280, pp. 17732–17736, May 2005.

[207] R. S. Illingworth, C. H. Botting, G. R. Grimes, W. A. Bickmore, and R. Eskeland, "PRC1 and PRC2 Are Not Required for Targeting of H2A.Z to Developmental Genes in Embryonic Stem Cells," *PLoS ONE*, vol. 7, pp. e34848+, Apr. 2012.

[208] M. Ku, J. Jaffe, R. Koche, E. Rheinbay, M. Endoh, H. Koseki, S. Carr, and B. Bernstein, "H2A.Z landscapes and dual modifications in pluripotent and multipotent stem cells underlie complex genome regulatory functions," *Genome Biology*, vol. 13, pp. R85+, Oct. 2012.

[209] E. Guccione, C. Bassi, F. Casadio, F. Martinato, M. Cesaroni, H. Schuchlautz, B. Luscher, and B. Amati, "Methylation of histone H3R2 by PRMT6 and H3K4 by an MLL complex are mutually exclusive," *Nature*, vol. 449, pp. 933–937, Oct. 2007.

[210] D. Hyllus, C. Stein, K. Schnabel, E. Schiltz, A. Imhof, Y. Dou, J. Hsieh, and U.-M. Bauer, "PRMT6-mediated methylation of R2 in histone H3 antagonizes H3 K4 trimethylation," *Genes & Development*, vol. 21, pp. 3369–3380, Dec. 2007.

[211] F. Lienert, F. Mohn, V. K. Tiwari, T. Baubec, T. C. Roloff, D. Gaidatzis, M. B. Stadler, and D. Schübeler, "Genomic Prevalence of Heterochromatic H3K9me2 and Transcription Do Not Discriminate Pluripotent from Terminally Differentiated Cells," *PLoS Genet*, vol. 7, pp. e1002090+, June 2011.

[212] Z. Wang, C. Zang, J. A. Rosenfeld, D. E. Schones, A. Barski, S. Cuddapah, K. Cui, T.-Y. Y. Roh, W. Peng, M. Q. Zhang, and K. Zhao, "Combinatorial patterns of histone acetylations and methylations in the human genome.," *Nature genetics*, vol. 40, pp. 897–903, July 2008.

[213] G. J. Filion, J. G. van Bemmel, U. Braunschweig, W. Talhout, J. Kind, L. D. Ward, W. Brugman, I. J. de Castro, R. M. Kerkhoven, H. J. Bussemaker, and B. van Steensel, "Systematic protein location mapping reveals five principal chromatin types in Drosophila cells.," *Cell*, vol. 143, pp. 212–224, Oct. 2010.

[214] N. Gilbert, S. Boyle, H. Fiegler, K. Woodfine, N. P. Carter, and W. A. Bickmore, "Chromatin architecture of the human genome: gene-rich domains are enriched in open chromatin fibers.," *Cell*, vol. 118, pp. 555–566, Sept. 2004.

[215] E. Yaffe and A. Tanay, "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture.," *Nature genetics*, vol. 43, pp. 1059–1065, Nov. 2011.

[216] S. van Koningsbruggen, M. Gierliński, P. Schofield, D. Martin, G. J. Barton, Y. Ariyurek, J. T. den Dunnen, and A. I. Lamond, "High-Resolution Whole-Genome Sequencing Reveals That Specific Chromatin Domains from Most Human

Chromosomes Associate with Nucleoli," *Molecular Biology of the Cell*, vol. 21, pp. 3735–3748, Nov. 2010.

[217] T. Dechat, S. A. Adam, P. Taimen, T. Shimi, and R. D. Goldman, "Nuclear Lamins," *Cold Spring Harbor Perspectives in Biology*, vol. 2, Nov. 2010.

[218] L. S. Shopland, C. R. Lynch, K. A. Peterson, K. Thornton, N. Kepper, J. von Hase, S. Stein, S. Vincent, K. R. Molloy, G. Kreth, C. Cremer, C. J. Bult, and T. P. O'Brien, "Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence," *The Journal of Cell Biology*, vol. 174, pp. 27–38, July 2006.

[219] P. Hozák, D. A. Jackson, and P. R. Cook, "Replication factories and nuclear bodies: the ultrastructural characterization of replication sites during the cell cycle.," *Journal of cell science*, vol. 107 (Pt 8), pp. 2191–2202, Aug. 1994.

[220] B. Dutrillaux, "Nouveau système de marquage chromosomique: les bandes T," vol. 41, no. 4, pp. 395–402, 1973.

[221] H. J. Gierman, M. H. G. Indemans, J. Koster, S. Goetze, J. Seppen, D. Geerts, R. van Driel, and R. Versteeg, "Domain-wide regulation of gene expression in the human genome," *Genome Research*, vol. 17, pp. 1286–1295, Sept. 2007.

[222] J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein, "Mapping and analysis of chromatin state dynamics in nine human cell types," *Nature*, vol. 473, pp. 43–49, May 2011.

[223] P. R. Cook, "A model for all genomes: the role of transcription factories.," *Journal of molecular biology*, vol. 395, pp. 1–10, Jan. 2010.

[224] G. Li, X. Ruan, R. K. Auerbach, K. S. S. Sandhu, M. Zheng, P. Wang, H. M. M. Poh, Y. Goh, J. Lim, J. Zhang, H. S. S. Sim, S. Q. Q. Peh, F. H. H. Mulawadi, C. T. T. Ong, Y. L. Orlov, S. Hong, Z. Zhang, S. Landt, D. Raha, G. Euskirchen, C.-L. L. Wei, W. Ge, H. Wang, C. Davis, K. I. Fisher-Aylor, A. Mortazavi, M. Gerstein, T. Gingeras, B. Wold, Y. Sun, M. J. Fullwood, E. Cheung, E. Liu, W.-K. K.

Sung, M. Snyder, and Y. Ruan, "Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation.," *Cell*, vol. 148, pp. 84–98, Jan. 2012.

[225] G. P. Patrinos, M. de Krom, E. de Boer, A. Langeveld, A. M. Imam, J. Strouboulis, W. de Laat, and F. G. Grosveld, "Multiple interactions between regulatory regions are required to stabilize an active chromatin hub.," *Genes & development*, vol. 18, pp. 1495–1509, June 2004.

[226] W. de Laat and F. Grosveld, "Spatial organization of gene expression: the active chromatin hub.," *Chromosome research : an international journal on the molecular, supramolecular and evolutionary aspects of chromosome biology*, vol. 11, no. 5, pp. 447–459, 2003.

[227] E. V. Volpi, E. Chevret, T. Jones, R. Vatcheva, J. Williamson, S. Beck, R. D. Campbell, M. Goldsworthy, S. H. Powis, J. Ragoussis, J. Trowsdale, and D. Sheer, "Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei.," *Journal of cell science*, vol. 113 (Pt 9), pp. 1565–1576, May 2000.

[228] C. Ferrai, S. Q. Xie, P. Luraghi, D. Munari, F. Ramirez, M. R. Branco, A. Pombo, and M. P. Crippa, "Poised transcription factories prime silent uPA gene prior to activation.," *PLoS biology*, vol. 8, pp. e1000270+, Jan. 2010.

[229] D. R. Higgs, M. A. Vickers, A. O. Wilkie, I. M. Pretorius, A. P. Jarman, and D. J. Weatherall, "A review of the molecular genetics of the human alpha-globin gene cluster," *Blood*, vol. 73, pp. 1081–1104, Apr. 1989.

[230] S. Boyle, S. Gilchrist, J. M. Bridger, N. L. Mahy, J. A. Ellis, and W. A. Bickmore, "The spatial organization of human chromosomes within the nuclei of normal and emerin-mutant cells," *Human Molecular Genetics*, vol. 10, pp. 211–219, Feb. 2001.

[231] V. V. Lunyak, G. G. Prefontaine, E. Núñez, T. Cramer, B.-G. Ju, K. A. Ohgi, K. Hutt, R. Roy, A. García-Díaz, X. Zhu, Y. Yung, L. Montoliu, C. K. Glass, and

M. G. Rosenfeld, “Developmentally Regulated Activation of a SINE B2 Repeat as a Domain Boundary in Organogenesis,” *Science*, vol. 317, pp. 248–251, July 2007.

[232] D. Schmidt, P. C. Schwalie, M. D. Wilson, B. Ballester, A. Gonçalves, C. Kutter, G. D. Brown, A. Marshall, P. Flicek, and D. T. Odom, “Waves of Retrotransposon Expansion Remodel Genome Organization and CTCF Binding in Multiple Mammalian Lineages,” *Cell*, vol. 148, pp. 335–348, Jan. 2012.

[233] H. Tanabe, F. A. Habermann, I. Solovei, M. Cremer, and T. Cremer, “Non-random radial arrangements of interphase chromosome territories: evolutionary considerations and functional implications.,” *Mutation research*, vol. 504, pp. 37–45, July 2002.

[234] J. A. Croft, J. M. Bridger, S. Boyle, P. Perry, P. Teague, and W. A. Bickmore, “Differences in the Localization and Morphology of Chromosomes in the Human Nucleus,” *The Journal of Cell Biology*, vol. 145, pp. 1119–1131, June 1999.

Appendix A

RELATED PUBLICATIONS

A.1 S-phase progression in mammalian cells: modelling the influence of nuclear organization

Shaw A, **Olivares-Chauvet P**, Maya-Mendoza A, Jackson DA. S phase progression in mammalian cells: modelling the influence of nuclear organization. Chromosome Research. 2010;18:163-178.

Summary

In this review we explored the different considerations for properly modelling S phase in mammalian cells. Base on the nature of DNA replication in mammalian cells and the data sets available we propose a basic modelling framework.

Contribution

General discussion of the manuscript, in particular regarding the putative role of the synergistic association of CTCF and cohesins for the formation and stabilisation of chromatin loops. CTCF nucleosome repositioning properties, in addition to spontaneous entropy-driven higher-order can lead to chromatin looping where cohesins can stabilize it.

Data analysis concerning the replication timing of G and R bands.

Chromosome Research (2010) 18:163–178
DOI 10.1007/s10577-010-9114-2

S-phase progression in mammalian cells: modelling the influence of nuclear organization

Alex Shaw · Pedro Olivares-Chauvet · Apolinar Maya-Mendoza · Dean A. Jackson

Published online: 13 February 2010

Abstract The control of DNA replication is of fundamental importance as cell proliferation demands that identical copies of the genetic material are passed to the two daughter cells that form during mitosis. These genetic copies are generated in the preceding S phase, where the entire DNA complement of the mother cell must be copied exactly once. As part of this process, it is known that different regions of mammalian genomes are replicated at specific times of a temporally defined replication programme. The key feature of this programme is that active genes in euchromatin are replicated before inactive ones in heterochromatin. This separation of S phase into periods where different classes of chromatin are duplicated is important in maintaining changes in gene expression that define individual cell types. Recent attempts to understand the structure of the S-phase timing programme have focused on the use of genome-wide strategies that inevitably use DNA isolated from large cell populations for analysis. However, this approach provides a composite view of events that occur within a population without knowledge of the cell-to-cell variability across the population. In this review, we attempt to combine information generated using genome-wide and single cell strategies in order to develop a coherent molecular understanding of S-phase progression. During this integration, we have explored how available information can be introduced into a modelling environment that best describes S-phase progression in mammalian cells.

Keywords DNA replication · S-phase programme · replication origins · DNA foci · replicon clusters · nuclear organization · modelling S phase

Abbreviations

BAC	Bacterial artificial chromosome
BrdU	5-bromo-2'-deoxyruridine
CDK	cyclin dependent kinase
CTCF	CCCTC-binding factor
FISH	Fluorescence in situ hybridization
ES	Embryonic stem cell

Introduction

The sheer complexity of the replication process is evident from the size of the genome in human cells—proliferating human cells have a diploid ($2n$=46) genome of roughly 6×10^9bp DNA. Inevitably, this demands that synthesis proceeds from numerous points—about 40,000 are used in each cell cycle—

Responsible Editors: Marie-Noelle Prioleau and Dean Jackson

A. Shaw · P. Olivares-Chauvet · A. Maya-Mendoza · D. A. Jackson (✉)
Faculty of Life Sciences, University of Manchester,
MIB, 131 Princess Street,
Manchester M1 7DN, UK
e-mail: dean.jackson@manchester.ac.uk

that are scattered throughout the genome (Berezney et al. 2000; DePamphilis et al. 2006). Such synthetic initiation points, termed origins of DNA synthesis, are of fundamental importance in defining the efficacy of the replication process as they provide targets for binding of the replication machinery and facilitate replication licensing, which ensures that DNA is replicated once and only once during each cell division cycle (Blow and Dutta 2005).

Controlling the activation of DNA synthesis is a key decision point in the eukaryotic cell cycle and understanding how origins of DNA synthesis are first established on chromatin and then selected for activation is of fundamental importance in replication control (Mendez and Stillman 2003; Gilbert 2004). Across eukaryotes the synthetic machinery is highly conserved. However, with the evolution of organismal complexity and associated increases in genome size it is clear that higher eukaryotes face a substantial challenge in order to ensure that their genomes are replicated intact. With this in mind, it is not surprising that the replication of mammalian genomes takes many times longer than primitive eukaryotes, which provide our most tractable model systems; under optimal growth condition budding yeast replicate their genomes in ~1 h whereas human cells require ~10 h. This ~10-fold difference in the duration of S phase is not directly linked to the size of replication units, which on average only vary by ~3-fold. Indeed, the key difference is that in mammalian cells only about 10–15% of the genome is associated with replicons that are engaged in synthesis at any point during S phase. Like unicellular yeasts (Raghuraman et al. 2001), mammalian cells replicate specific regions of the genome at defined times of S phase (Goen and Cedar 2003; Aladjem 2007) so that active genes within open or dynamic euchromatin tend to be replicated early in S phase while more condensed heterochromatin replicates later.

It is reasonable to assume that a temporally structured S phase is likely to be of biological importance. Notably, cell differentiation in higher eukaryotes correlates with cell type specific patterns of replication timing (Hiratani et al. 2008), which in broad terms reflect changes in gene expression. This link between replication timing and gene expression may be of fundamental importance in maintaining patterns of expression as there is some evidence that histones with different post-translational modifications are deposited on DNA that is replicated during early or late S phase (Zhang et al. 2002; Lande-Diner et al. 2009). Moreover, the fact that early and late firing replication origins can be regulated by a molecular switch that involves the differential activation of potential origins based on their interaction with specific cyclin-CDK complexes (Donaldson 2005; Katsuno et al. 2009) implies that there is a biological imperative to maintain the timing programme.

The S-phase programme

S phase in mammalian cells is structured in time so that euchromatin is replicated before heterochromatin. At admittedly low resolution, this separation is clear from the discrete labelling of chromosomal bands visualized on metaphase chromosomes after prior labelling of DNA—using precursor analogues such as bromo-deoxyuridine (BrdU)—during different intervals of the preceding S phase (Holmquist 1987; Drouin et al. 1994). This type of cytological labelling suggests that the replication of R-band chromatin during early S phase is essentially completed before the synthesis in the chromosomal G-bands can begin. Moreover, the idea that R- and G-band synthesis occurs at discrete times of S phase is reinforced by the observation that in some cell types perturbation of the precursor pools reveals a distinct ‘3C-pause’ which appears to represent the time of switching from early to late synthesis (Drouin et al. 1994; Strehl et al. 1997). This broad-scale timing of chromatin domains of the size of chromosomal bands (i.e. ~10 Mbp) was confirmed in the seminal experiments of Carter et al. (Woodfine et al. 2004), who used hybridization of DNA isolated from G1 and S phase human lymphoblast cells to map the timing of replication based on DNA content. This early study provided a low resolution (~1 Mbp) genome-wide map of the human timing programme.

Recent advances in microarray design and analysis (TimEX) and deep sequencing (TimEX-seq) approaches have confirmed the basic conclusions of this seminal study (Desprat et al. 2009). These high resolution studies provide precise estimates of copy number variation based on the use of Gaussian convolution (noise filtering) to integrate massive numbers of highly redundant measurements. The

following points summarise the key findings of this high resolution analysis:

1. Replication proceeds with a clear temporal programme, with regions of the genome of ~1 Mbp being assigned to replication timing domains with a time resolution of 1-2 h. Importantly, replication domains of this size are unlikely to represent single replicons as forks of ~500 kbp would take at least 5 h to complete synthesis.
2. Replication domains in early and mid/late S phase are distinct and in both cases synthesis initiates within large zones that contain a high density of potential initiation sites. Many long replicons (>250 kbp) link the early and mid/late replicating regions. These 'temporal transition regions' couple the early and mid/late replicating domains and represent ~5% of the genome where replication origins are highly dispersed.
3. Replication timing of a gene locus correlates with the level of gene activity within the locus. Regions that are replicated very early in S phase tend to contain genes with very high levels of transcription whereas regions with genes that have only low levels of transcription tend to be replicated during mid-S phase. Regions of the genome that are expressed late in S phase are remote form transcribed genes.

These general conclusions have been confirmed and extended in many recent studies that have explored both the general timing programme (Farkash-Amar et al. 2008; Hiratani et al. 2008; Desprat et al. 2009) and specific locations where synthesis can begin (Cadoret et al. 2008; Sequeira-Mendes et al. 2009). For a much better appreciation of this information the reader is referred to excellent articles elsewhere in this volume (Farkash-Amar and Simon 2010; Pope et al. 2010; Cadoret and Prioleau 2010).

While the segmentation of mammalian genomes into early and late replicating domains appears robust the time resolution of these experiments is often poor and in some cases large regions of the genome appear to engage synthesis over many hours (Jeon et al. 2005; Karnani et al. 2007). The mechanisms of origin selection clearly influence the efficiency with which different regions of the genome are replicated at precise times of S phase. In this respect, it is important to recognise that the activation of synthesis at individual potential origins is stochastic. Potential origins will each have different probabilities of firing during each cell cycle, presumably as a consequence of the local chromatin environment. As a result, in individual cells, the majority of potential origins are not used during a particular S phase and most potential origins are replicated passively by forks that emanate from adjacent replication units. Hence, individual potential origins will only provide initiation sites for synthesis in a minority of cells and any specific locus will be replicated at a time that relates to its position relative to the nearest active origin (Hamlin et al. 2010). The analysis of replication intermediates using both 2D-gels (Mesner et al. 2006) and DNA fibres (Lebofsky et al. 2006; Desprat et al. 2009) confirms this stochastic view of origin firing.

Analysis of DNA replication using single cell approaches

Genome-wide approaches undoubtedly provide valuable insight into the distribution of replication origins and the replication timing programme (Farkash-Amar and Simon 2010; Cadoret and Prioleau 2010 in this volume). However, the variable efficiency of origin activation raises obvious questions about mechanisms of origin firing that cannot be addressed using genome-wide strategies, which use large numbers of cells and so are unable to detect any cell-to-cell variability. Hence, it is of clear value to interpret the genome-wide data in the context of complementary studies performed on individual cells.

It has been recognised since the seminal studies of Nakamura et al. (1986) that replication in mammalian cells takes place within specialized nuclear domains where many active replisomes are clustered together. Many subsequent studies (see Jackson 1995 and Zink 2006 for reviews) have used a wide range of modified replication precursor analogues (BrdU, biotin-dUTP and Cy3-dUTP are frequently used examples) to confirm that mammalian cells perform replication at discrete replication sites, which contain groups of polymerase complexes within synthetic factories (Hozak et al. 1994; Leonhardt et al. 2000). The replication machinery within individual factories performs synthesis of small groups of contiguous active replicons, which are replicated together and activated at similar times (Jackson and Pombo 1998; Ma et al. 1998). These replicon clusters can be

visualized as 'DNA foci' that contain ~1 Mbp of DNA (Cremer and Cremer 2001). Importantly, these functional targets for DNA synthesis have been shown to represent stable structural units of sub-chromosomal organization. As S phase proceeds, the structure of active centres of DNA synthesis changes according to a predictable programme (Fig. 1), which reflects the disposition of different chromatin classes within the nucleus (Cremer and Cremer 2001; Goetze et al. 2007).

Mapping S-phase progression at the level of DNA foci

A highly structured replication programme (Fig. 1) implies that specific regions of the genome are selected for synthesis at predictable times. This could of course reflect the stochastic activation of different classes of potential replication origins, based perhaps on their interaction within different cyclin/CDK complexes (Katsuno et al. 2009). However, studies in single cells have suggested that the organization of DNA foci contributes to S-phase progression. Analysis of the time of replication of DNA foci in different cell cycles has shown that the same DNA foci are activated with high efficiency (>90%) at the onset of S phase (Jackson and Pombo 1998; Ma et al. 1998). This implies that a robust mechanism regulates the selection of replicon clusters that are targets for synthesis as cells enter S phase. In addition, as S phase proceeds any newly activated replication sites appear to lie next to sites that were engaged in synthesis during the previous period of S phase (Manders et al. 1992). This suggests that the spatial architecture of chromatin foci might be a key determinant of S-phase progression with the sequential activation of foci occurring following a 'next-in-line' principle (Manders et al. 1992). This has been confirmed using an analysis of mid/late replication factories in living cells (Sporbert et al. 2002), where analysis of the simplified patterns of active sites allowed the spatial relationship of foci to be mapped at high resolution.

Are structure-function links defined by DNA foci?

If the structure of DNA foci plays a significant role in defining the architecture of the replication programme it is important to understand how individual foci are defined. In fact, very little is known about the structure of foci and the molecular principles that might allow stable structures to be established and maintained. There is evidence that chromatin foci are maintained by epigenetic chromatin states. For example, the analysis of sub-chromosomal regions with interspersed gene islands (gene-rich regions) and gene deserts (regions with very few active genes) shows that the two chromatin classes are separated into discrete foci with chromatin that does not mix (Fig. 2; Shopland et al. 2006; Goetze et al. 2007). If such specific examples define a general feature of chromatin organization, it is not unreasonable to suggest that chromatin status might dictate the replication timing of foci with different chromatin epi-states. Supporting this model, replication timing in yeast and notably the transition from euchromatin to heterochromatin replication is defined by the acetylation status of histone in the two chromatin compartments (Vogelauer et al. 2002).

While mechanisms that link the structure of DNA foci and their replication timing are a matter for speculation, our understanding of potential links is clearly hindered by deficiencies in our knowledge of foci structure. It is known, for example, that global chromatin loops in mammalian cells correlate with replicon size (Buongiorno-Nardelli et al. 1982; Courbet et al. 2008), perhaps to provide a memory of replicon structure that is transmitted for one cell generation to the next. But how such loops relate to function and the structure of the template within DNA foci is unclear. Historically, numerous studies have described the behaviour of genomic elements such as nuclear scaffold and matrix attachment regions, locus control regions and domains insulators (reviewed in West and Fraser 2005) that together define the architecture of chromatin domains in mammalian cells. More recently, the insulator protein CTCF has emerged as a good candidate to define boundary elements that punctuate the genome to form higher-order chromatin domains (Phillips and Corces 2009). Intriguingly, sites of CTCF binding have also been shown to be sites of cohesin accumulation, suggesting that they might assume special structural properties that contribute to architecture of chromatin loops (Parelho et al. 2008; Hadjur et al. 2009). In addition, hotspots of CTCF binding have been shown to establish unique features in the local chromatin environment (Fu et al. 2008; Zhang et al. 2008), which might contribute to the formation of

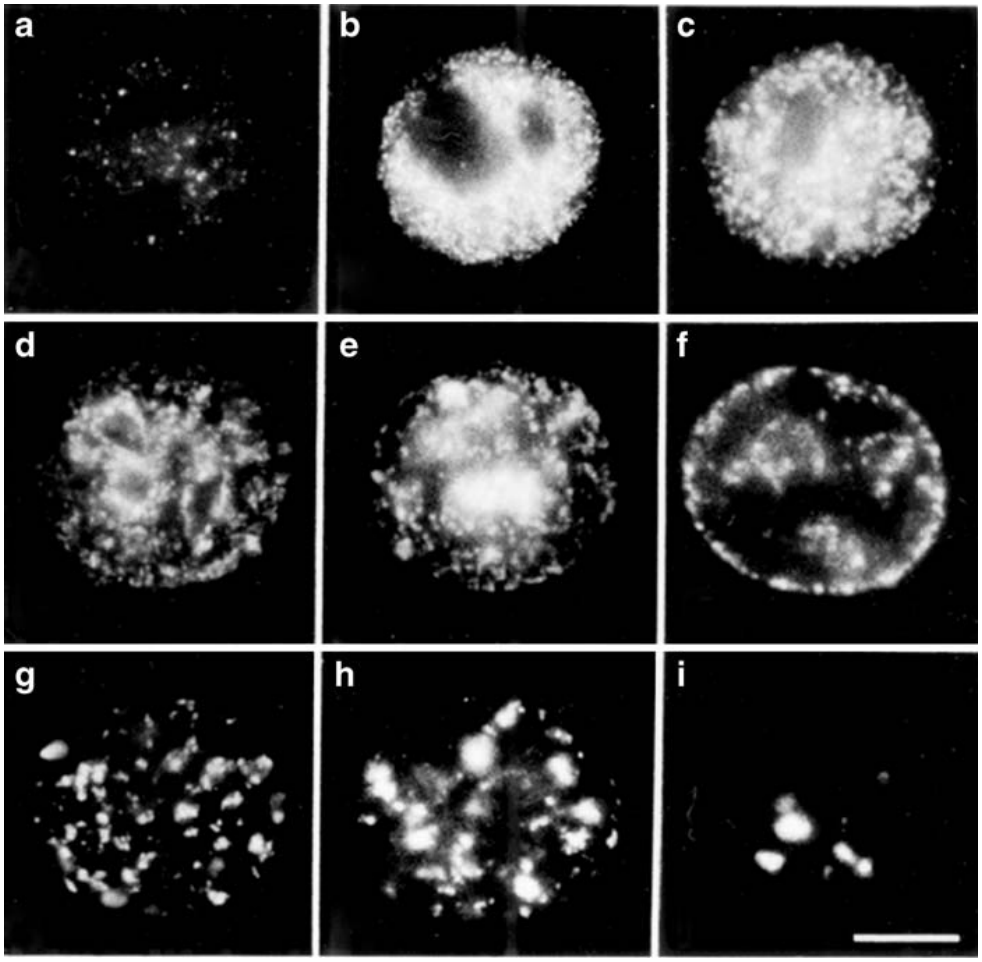


Fig. 1 The spatial distribution of active replication sites during S phase. During S phase, different classes of chromatin are replicated at different times. Chromatin that contains the majority of transcribed genes, within chromosomal R bands, is replicated over the first ~4 h of S phase. During this period, active sites of DNA synthesis are in discrete foci dispersed throughout the nuclear interior (**a–c**). At mid-S phase, replication begins to switch to more inert chromatin and patterns of replication foci that reflect the peripheral location of heterochromatin are seen (**d–f**). Finally, heterochromatic blocks of late replicating chromatin are duplicated within the nuclear interior (**g–i**). Images shown are replication sites labelled in permeabilized HeLa cells—using biotin-dUTP—that were fixed and indirectly immuno-labelled under conditions that preserve nuclear organization. The bar is 5 microns. For more details see Hozak et al (1994). Reproduced with permission from the Company of Biologists

entropy-driven higher-order chromatin conformations (St-Jean et al. 2008).

Single molecule analysis on DNA Fibres

While the analysis of replication foci in situ provides some molecular insight to support the genome-wide studies, the analysis of DNA foci within nuclei is also limited in scope by resolution. The low resolution information within foci in situ can, however, reveal additional high resolution information if the DNA that they contain is analyzed after preparation of spread DNA fibres. Labelled DNA fibres can then be used to develop detailed information about fork rates and the distribution of active replicons and how individual replicons are activated in different cell cycles (Jackson and Pombo 1998; Takebayashi et al. 2001). Most importantly, DNA fibres prepared from cells that were labelled with different replication precursor analogues during consecutive cell cycles provided compelling evidence that structurally stable replicon clusters generate DNA foci that represent both structural and functional sub-chromosomal units (reviewed in Maya-Mendoza et al. 2009).

In recent years, the analysis of DNA fibres prepared from cells that have been labelled with a range of replication precursors has revealed funda-

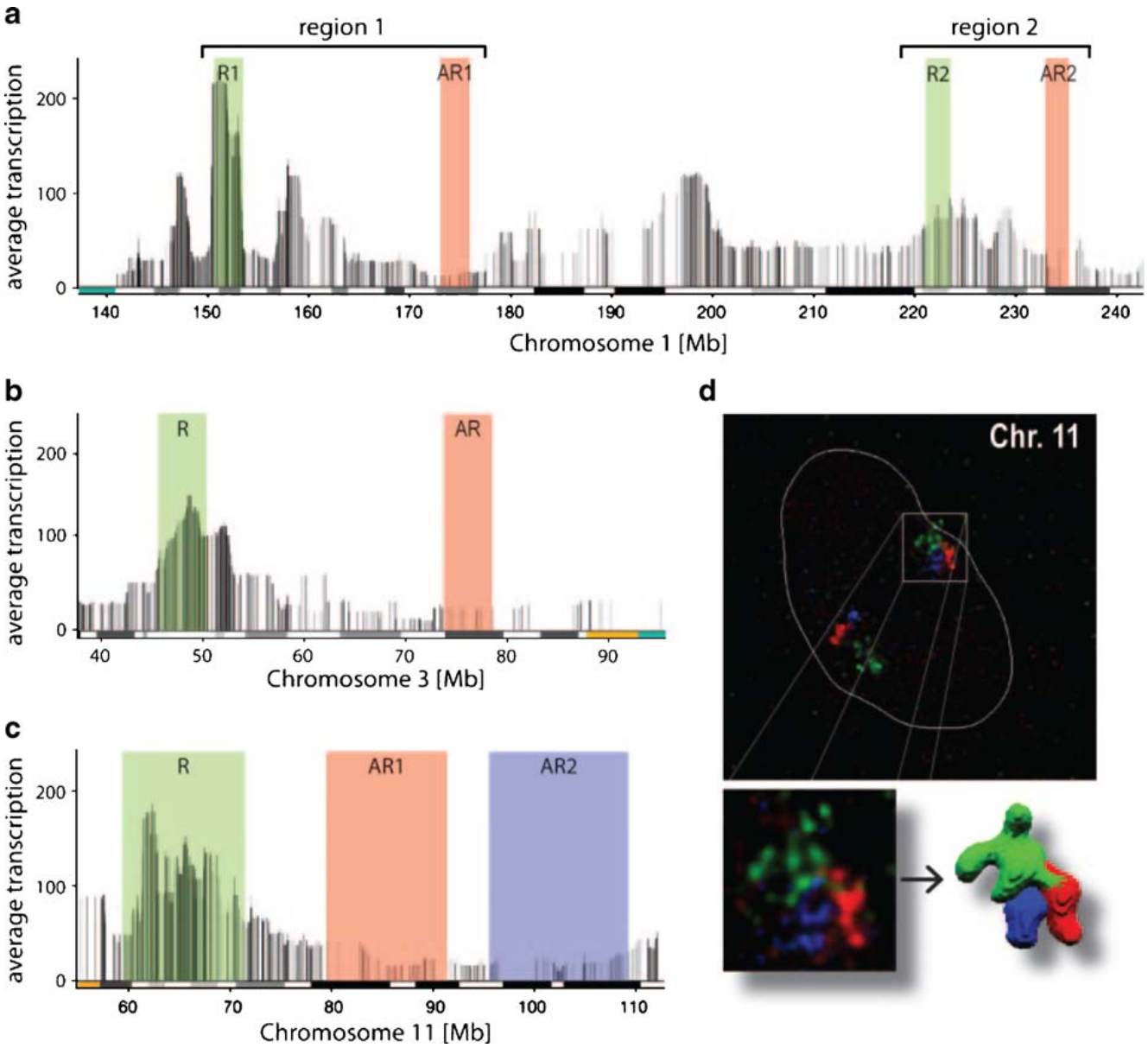


Fig. 2 DNA foci are structural units of higher-order chromatin folding. Mammalian genomes are folded into chromatin domains that assume a variety of chromatin environments as a result of local patterns of gene expression. Genomic regions that are rich in active genes—known as gene ridges (*R*) or gene islands—are separated by gene-poor domains—known as anti-ridges (*AR*) or gene deserts. The local architecture of three chromosomal loci with interspersed ridges and anti-ridges are shown above (**a**–**c**). To test the nuclear distribution of the different chromatin compartments, the three regions (one ridge and two anti-ridges) highlighted in (**c**) were visualized in situ using FISH (fluorescent in situ hybridization) to probe the target loci (**d**). The FISH probes for the three target loci were prepared from contiguous arrays of BAC clones, which spanned the regions shown. The three pools of BAC clones were differentially labelled prior to hybridization. Visualization of the labelled probes shows that three regions under analysis are constrained within discrete local domains. Notably, while the BAC pools cover ~10 Mbp of DNA for each region, in all cases the fluorescent signal was concentrated locally in domains of ~500 nm. Based on their number, each of these domains contains roughly 1 Mbp of DNA. The gene-rich and gene-poor compartments are self-contained (i.e. discrete) and the chromatin environment within the compartments defines the volume occupied (gene-rich compartments are more open) and the spatial architecture of the domains within the each chromosome territory. Images taken from Goetze et al. (2007) and published with permission of the American Society of Microbiology

mental information about the structure of eukaryotic replicons and the replication programme (reviewed by Tuduri et al. 2010). Studies evaluating the activation of potential origins across specific chromosomal regions have been especially informative (Lebofsky et al. 2006; Conti et al. 2007). A key feature of these studies has been the recognition that potential sites of initiation of DNA synthesis are typically distributed throughout ~10 kbp chromatin domains. However, pre-initiation complexes that are selected from these zones to activate synthesis are recognised inefficiently, so that in individual cells only about 1/3rd support initiation during a particular cell cycle. Activated origins appear to be selected at random so that different combinations of active origins are seen in different cells (Lebofsky et al. 2006). Moreover, adjacent active origins, which are typically separated by roughly 100 kbp of DNA in mammalian cells, are often seen to be activated at similar times and in most cases synthesis proceeds with forks that grow at very similar rates (Lebofsky et al. 2006; Conti et al. 2007).

Simulating S-phase progression in mammalian cells

As so little is understood about the molecular principles that regulate S-phase progression we wanted to assess if in silico simulations could be developed to model features of nuclear organization that contribute to the chromatin environment and drive the S phase programme. To do this, we have attempted to incorporate information described above that is derived from both genome-wide and single cell studies. In considering possible mechanisms, one might begin by suggesting two extreme scenarios. In the first, the activation of potential origins might be fundamentally stochastic, so that initiation is driven by random choice with the proviso that the chromatin environment modulates choice so that different regions of the genome will be replicated preferentially at different times. Euchromatin is known to engage synthesis before heterochromatin and it is possible to argue that subtle differences in chromatin structure might contribute to replication timing within these chromatin compartments. In the second, the chromatin environment defines the sites that are selected for initiation of synthesis at the onset of S phase but thereafter replication spreads from these primary initiation sites so that the downstream replication programme is defined by the activation of genetically adjacent chromatin domains along chromosomes. This scenario represents origin activation driven by a next-in-line model of S-phase progression. Of course, as these extremes are not mutually exclusive the molecular mechanism of progression in vivo might involve a mixture of stochastic and genetically coupled activation events.

Modelling the chromatin environment

Published models to describe eukaryotic DNA replication have focussed predominantly on stochastic models of origin activation. Most attention has focussed on organisms with simple replication programmes (Lygeros et al. 2008; Herrick et al. 2002; Rhind 2006) and only recently have the models been used to explore aspects of replication in the S phase of somatic mammalian cells (Goldar et al. 2009; Ge and Blow 2009). A comprehensive analysis of these published models is presented elsewhere in this volume (Rhind et al. 2010; Hyrien and Goldar 2010).

In mammalian cells, local chromatin environments play a major role in S-phase progression. Hence, any viable model of S phase must incorporate parameters related to the orderly synthesis of the major chromatin compartments and evaluate established features of organization related to the mechanisms involved. In particular, any model of the mammalian S phase must incorporate replicon clusters (within DNA foci) as the basic targets for DNA synthesis and evaluate how replication spreads between these structures. Here, in order to simulate the activation of replicon clusters, we have taken data for the distribution of replicons within replicon clusters from Jackson and Pombo (1998); primary data sets were used to model the profile of inter-origin separations within clusters and used in combination with the published distribution of active replicons/cluster. The distribution of replicons within replicon clusters that are replicated at different times of S phase have a similar average structure (Maya-Mendoza et al. 2007), despite differences in their spatial organization and nuclear distribution (Shopland et al. 2006; Goetze et al. 2007; see Fig. 2).

Hence, in considering the different features that define the chromatin environment we propose that a biologically informative simulation of the mammalian

S phase should incorporate the following conditions during modelling:

1. DNA in chromosomal R- and G-bands is replicated preferentially at defined times of S phase, with synthesis of R-band chromatin in early S phase and G-band chromatin in mid/late S phase. The differential probability of origin activation will be determined by expression of appropriate cyclin-CDK complexes (Katsuno et al. 2009).
2. Throughout S phase, replicons are activated in small groups within functional replicon clusters (Jackson and Pombo 1998).
3. Clusters that are active during consecutive intervals of S phase are defined predominantly by chromosome structure.
4. Mammalian S phase is regulated by a mechanism that restricts the absolute level of synthesis, so that only 10–15% of the genome is engaged in synthesis at any time. The mechanism that drives this 'replication rheostat' is unknown.
5. Replication of different chromatin classes occurs at different rates (Takebayashi et al. 2001).

A basic modelling framework

Using these experimentally defined conditions, model development has obvious potential to inform our understanding of mechanisms that drive the spread of replication throughout mammalian genomes. In the analysis that follows, models were implemented in Matlab and tested by fitting to the natural duration of S phase in order to assess biological efficacy; mammalian S phase takes ~10 h to complete and for the purpose of simulation we have restricted initiation to the first 8 h of this period. Using human chromosome 6 to build our model (Goldar et al. 2009), we first estimated the number of replicon clusters required to complete S phase (Fig. 3) using cluster architectures from Jackson and Pombo (1998) and variable fork rates from Takebayashi et al. (2001). In this simple form of simulation, all clusters have an equal probability of being activated. Hence, the simulation defines the absolute number of activation events required to complete synthesis and estimates the probability of cluster firing during defined intervals in order to perform synthesis in the desired time. As expected, as potential origins are consumed the probability that remaining origins will be activated increases so that replication completes on schedule (Fig. 3). The pattern of activation seen in this profile reflects the structure of replicon clusters—the lengths of replicons within clusters dictates the timing when their synthesis can complete and this is linked to activation of new clusters. The decline in initiation events towards the middle of S phase is a consequence of the reduced rate of fork elongation at that time.

The profile of cluster activation seen in Fig. 3 can also be represented to show the absolute levels of DNA synthesis as S phase proceeds. This readout is used to map the success of S phase in the simulations shown in Fig. 4. In these simulations we incorporate key features of the chromatin environment into the model using R- and G-band coordinates taken from the UCSC Table Browser with the March 2006 genome assembly (Karolchik et al. 2003). These chromosome banding patterns were applied to the simulation and the probabilities of activation within euchromatin and heterochromatin adjusted to mimic the effect of chromatin environment on the activation of potential origins. We also used expression data from Katsuno et al (2009) to simulate the effect on differential activation of R- and G-band replicon clusters given that the G-band clusters are activated by increasing Cyclin A-CDK1 expression towards late S phase—the availability of Cyclin A-CDK1 was modelled to rise starting at 2 h after the onset of S phase and reach a peak 4 h later (Fig. 4b). During this compound simulation, R band replication was activated at the onset of S phase and proceeded as before until the increasing expression of Cyclin A-CDK1 allowed origin activation within G-band clusters.

Simulations were developed to identify optimal probabilities of origin activation as defined by the amount of deviation from the average synthetic quota (defined by the replication rheostat) required to complete synthesis within 10 hours. The simulation incorporates variable activation probabilities of G-band clusters as Cyclin A-CDK1 expression increases between 2–6 h of S phase. With a sigmoidal expression profile, the minimum variation from the DNA quota per minute was found to occur with a maximum G band cluster firing efficiency of 0.0045/cluster/min (Fig. 4b). A slightly higher maximal probability of 0.0054/cluster/min was seen when the increase in expression was linear.

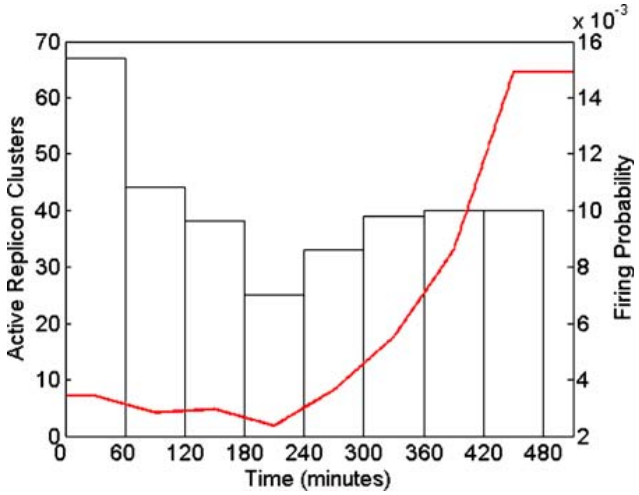


Fig. 3 Calculation of replicon cluster firing efficiencies. This simulation describes the architecture of replicon clusters and the probability of cluster firing during replication of human chromosome 6. This chromosome contains 171 Mbp of DNA, so simple calculations allow us to determine the fraction of the chromosome that must be replicated within each hour window of S phase when initiation can occur for 8 h. Using cluster architectures from Jackson and Pombo (1998), this simulation calculates the number of replicon clusters that must be activated to ensure the required amount of DNA synthesis within each hour of S phase (*bars*). We used the published cluster architectures and replicon lengths, which approximated to a normal distribution (μ=140.6238 kbp, σ=58.8192). Cluster architectures were generated independently for each simulation by random sampling of the experimentally derived data sets (average values of 5,000 independent simulations are shown). Replication of individual clusters was programmed to proceed at constant rate and variable fork rates across S phase were smoothed to prevent discontinuities in the simulation. Then, as the distribution of replicon clusters defines the number of active clusters that will be needed to complete synthesis in the required time the necessary firing probabilities can be calculated (*red line*). In this example, the average profile of cluster architectures requires that 326 foci are activated to replicate the 171 Mbp chromosome (most foci contain 250–1,000 kbp of DNA). It is assumed that all unreplicated clusters have an equal probability of activation. Hence, for each time point of S phase the simulation uses the absolute number of synthetic units to estimate the probability/cluster/minute that is required to complete synthesis on schedule. At the onset of S phase, 67 clusters are activated to engage the required level of synthesis, with a probability of $67/326 \times 60 = 0.034$ clusters/min During the 7th hour of the simulation 40 of the remaining 80 clusters are activated so the probability of activation increases to $40/80 \times 60 = 0.083$ clusters/min

The firing efficiency profiles generated using the optimal conditions (with adjustments after 6 h to compensate for the dwindling pool size) are shown in Fig. 4c. These outputs include the effects of cluster banding on the existing model framework and approach a realistic biological representation of chromosome structure in vivo. As shown in Fig. 4c, the different patterns of increasing cyclin expression had only a slight effect on S-phase progression; the sigmoidal profile was used in later models.

Spatial architectures of replication foci

So far our analysis has simulated the effects of replicon clustering within DNA foci, variable fork rates throughout S phase and the differential activation of potential origins during early and mid/late S phase based on their chromatin environment. To add molecular complexity to the simulations, we next evaluated how models might be affected by different mechanisms of S-phase progression (see Fig. 5). This aspect of the modelling is designed to assess how next-in-line and stochastic models of cluster activation influence S-phase progression. Simulations were performed using the conditions developed in Fig. 4 to test which parameters give the best fit to the established S phase duration (Fig. 6). In this analysis, different modelling environments were compared using an end-time where 95% of DNA was replicated; this limits the effect of rare events that can lead to very long end-times. To simulate the effect of a next-in-line mechanism of origin activation different multiplier values (between 1 and 5000) were incorporated into the model. This feature alters the probability with which replicon clusters are selected for activation based on changes in the chromatin environment that arise during replication of neighbouring clusters. A low resolution scan of the parameter space, comparing a range of maximum firing efficiencies for the sigmoidal curve (between 0.0001 and 0.0083/cluster/min), highlights a number of regions of biological interest (Fig. 6). In this phase plot, each of these areas of interest indicates the impact of different parameter sets and thus different mechanisms that are driving the progress of S phase (Fig. 7).

The following conclusions were drawn from simulations that test three alternative models of S-phase progression:

1. Origin selection is stochastic

A null hypothesis that ignores any relationship between DNA foci would simply alter the probabilities of origin activation towards late S phase, based

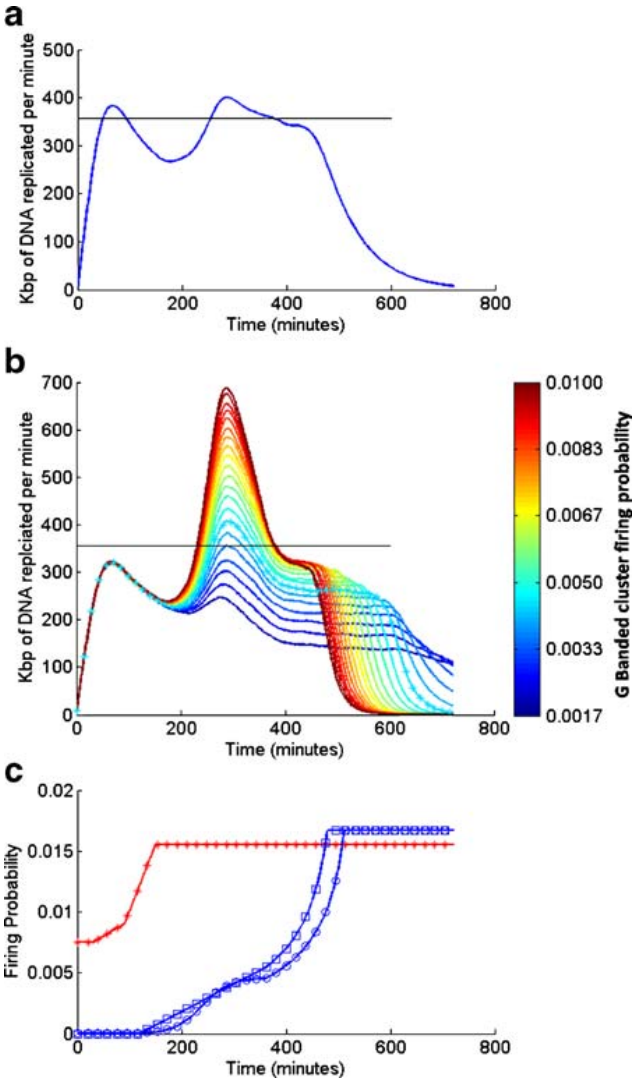


Fig. 4 Modelling DNA replication across S phase. The distribution of cluster firing probabilities generated in Fig. 3 was used to simulate an averaged profile of DNA replication throughout S phase (**a**). A linear representation of human chromosome 6 was created and divided into replicon clusters using distribution data from Jackson and Pombo (1998) and distribution of firing probabilities applied. Each cluster has the potential to be activated during time steps of 1 min. Once activated, DNA within each replicon of a cluster replicates according to the specific fork speed relevant to the particular period of S phase and terminates on meeting a neighbouring fork. The *blue line* shows the progress of synthesis (DNA replicated in kbp/min averaged over 5,000 simulations) and black line the quota of DNA synthesis required to complete S phase on schedule. A modified version of the simulation shown in (**a**) was generated to accommodate the effect of different chromatin environments in chromosomal R- and G-bands (**b**). Using the R/G band configuration across human chromosome 6, probabilities of cluster activation were calculated first in R-bands and then in G-bands using a range of different potential maximum values as Cyclin A-CDK1 concentrations increased. Data shown were generated by modelling a sigmoidal increase in Cyclin A-CDK1 expression between 2–6 h of S phase. A linear increase in expression was also tested (not shown). The range of maximum firing probabilities is shown in the accompanying colour-bar. Once the maximum value is reached at 6 h, the probability is adjusted to account for the decreasing pool of unreplicated clusters. Each coloured plot of DNA output therefore refers to the DNA replicated (kbp/min) under different maximum G Band firing probabilities. Output is averaged over 1,000 simulations for each parameter and firing probabilities are measured per cluster per minute. The optimal probability of 0.0045, giving the closest adherence to the DNA quota, is highlighted (*cyan stars*). Rates of synthesis within different clusters throughout S phase can be transformed to monitor overall levels of synthesis as S phase proceeds (**c**). Firing probabilities were generated as before with R-bands firing (*red line*) during early S phase followed by the optimized firing of G-band clusters (blue/cyan lines). The optimal G-band cluster firing probabilities were the maximum values giving the closet fit to the DNA replication quota. With linear increase in Cyclin-CDK expression (*blue line with squares*), the probability of cluster firing within G-bands peaked with optimal probability of 0.0054/cluster/min at 6 h. With sigmoidal increase (*blue line with circles*), an optimal probability of 0.0045/cluster/min at 6 h was seen

on expression of activating cyclin-CDK complexes (Fig. 4). In this case, the maximum probability of G-band firing defines the behaviour of the model. A maximum probability of 0.004/cluster/min was therefore tested as a case study (Fig. 6, position a). This parameter set gives an average variation from quota of 74.14 kbp/min and completes 95% of DNA replication within 8.4 h, with absolute completion by 10.8 h. The standard deviation at absolute completion was 67.0 min. This mechanism therefore provides a stable and timely completion of S phase. However, this model does generates a high level of ab initio cluster activation of 41.5%. Additionally, whilst the ratio of single sided firing events to dual sided is 2.2, this is a consequence of the high levels of ab initio firing. Importantly, the distribution of origin firing in this case is skewed very late into S phase and predicts a level of very late synthesis that is not seen experimentally. Predictably, increasing the maximum probability of activation results in a shift in activation but also leads synthesis to complete at unrealistically early times. Allowing a small effect of fork elongation on cluster firing probabilities (with a maximum probability of 0.0033/cluster/min and a ×2 increase in cluster firing if forks are encroaching—Fig. 6, position b) reduces the length of S phase even though the maximum probability of activation is reduced. The variation of completion times is seen to rise slightly however, showing that this limited spatial effect has little beneficial consequence on the behaviour of the system.

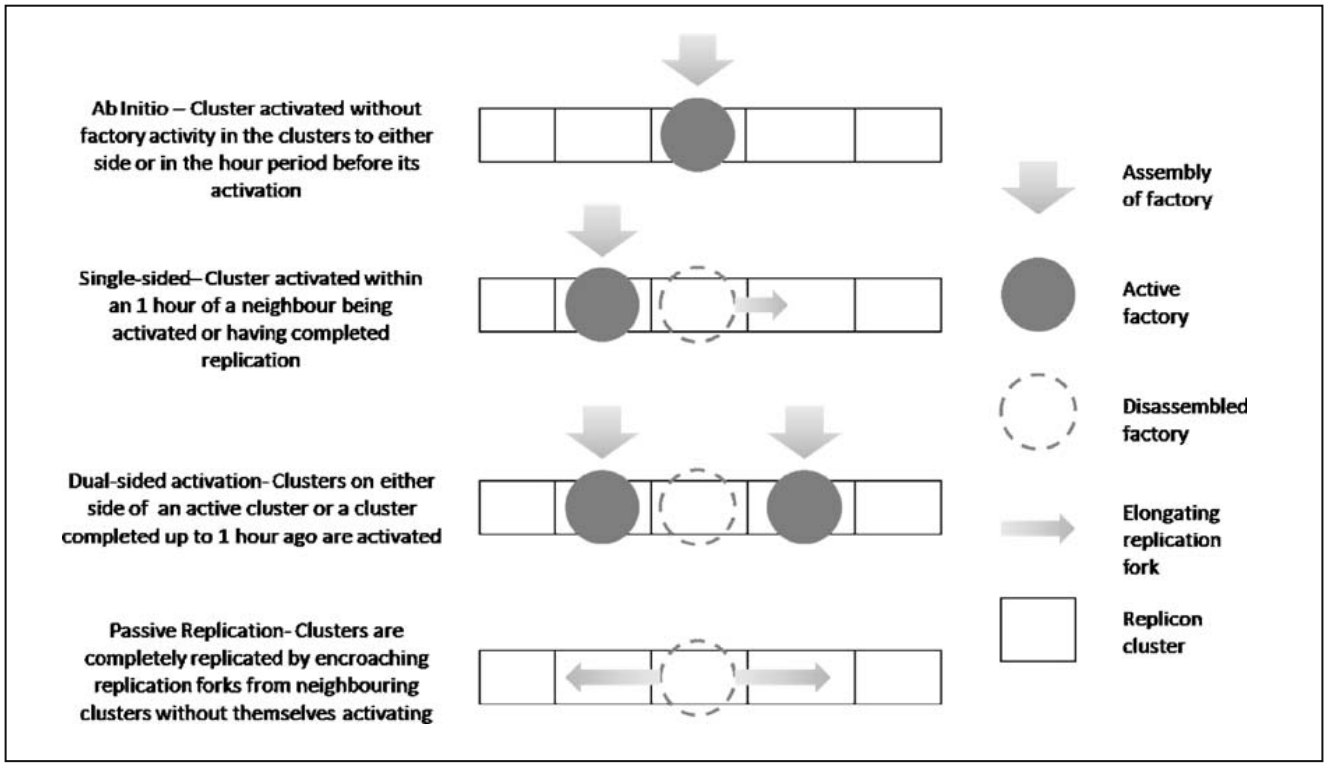


Fig. 5 Mechanisms of S phase propagation

2. Encroaching forks drive cluster firing

Next-in-line models of S-phase progression predict that the spread of encroaching forks is the driving factor that increases local firing probabilities. To simulate this, the model was set with a low maximum value of 0.0008/cluster/min for origin firing efficiency and a high multiplier value of 1,500, so that clusters with encroaching forks have a high probability of engaging synthesis (Fig. 6, position c). With these settings, 95% of DNA replication is completed in 8.3 h and total completion occurs within 11.2 h on average. The distribution of these completion times is more varied than in model (1), with a standard deviation of 73.7 min for the absolute completion times and 21.6 min for the 95% completion times. However, while the next-in-line conditions produce more variable end-times the dynamics of cluster firing give a better fit to biological profiles of origin activation (Goldar et al 2009) and yield a reduced rate (17%) of ab initio activation events.

3. Hybrid-driven cluster firing

A final possibility is that alterations in G band cluster firing efficiencies are driven by a mixture of the mechanisms explored in (1) and (2). This was simulated in the model through a multiplier value for fork encroachment of ten and a maximum firing probability of 0.0022/cluster/min (Fig. 6, position d). The combination of factors still gives a 95% completion time of 8.4 h with absolute completion in 11.0 h. The variation of the completion times lies between that of the two alternative models, as does the rate of ab initio activations at 31.0%. With a ratio of single activation events to dual activation of 1.86, these conditions allow a significant increase in activation by encroaching forks relative to the stochastic model. However, the spatial effects are not strong enough to drive a high ratio of dual cluster activation events, as is seen at higher levels of spatial activation by fork encroachment.

To explore how changes in the chromatin environment might influence the switching of synthesis between neighbouring replicon clusters, we performed simulation that incorporated sub-optimal fork elongation rates in order to mimic possible fork stalling, which might occur as synthesis switches from one replication cluster to the next. Variable probability settings in the range 1–50% were used to simulate different extents of fork failure. From these simulations, it is evident that the ‘fork elongation’

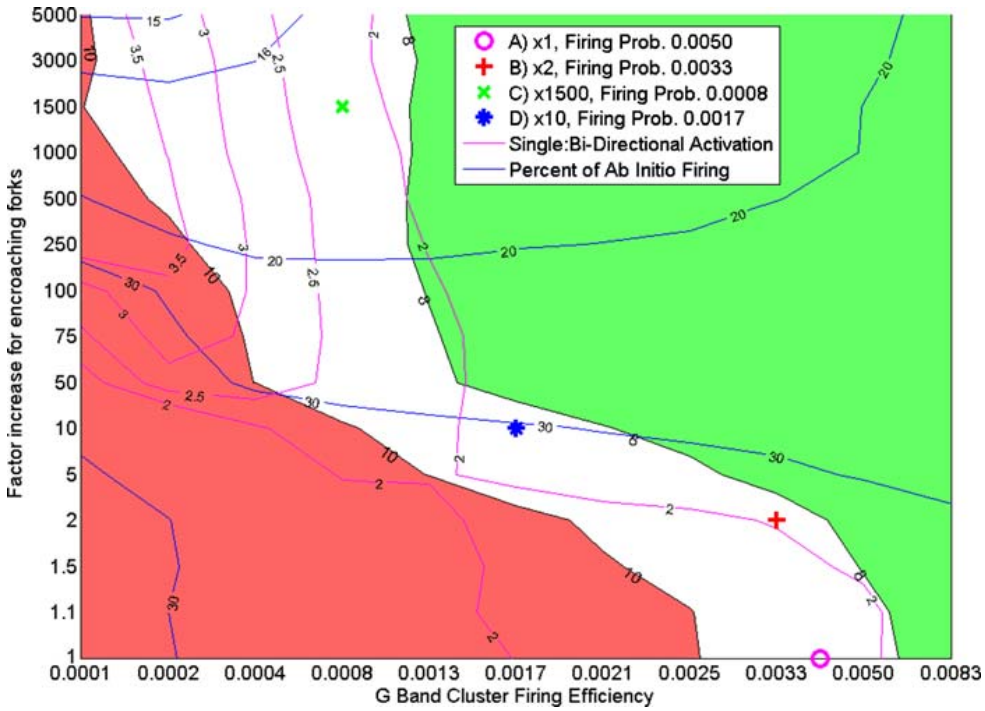


Fig. 6 A phase diagram of the explored parameter space. Using the method demonstrated in Fig. 4b, different firing probabilities were tested against a range of values to model spatial activation of DNA foci, using models described in Fig. 5. As synthesis within active clusters completes the extending forks growing out from the flanking replicons begin to interact with chromatin of neighbouring clusters. Here, we test how this influences the probability of activation within the adjacent cluster—the extent of this increase was modelled over a range of probabilities from ×1 (no change) to ×5,000 (highly probable). Given these parameter sets, an approximated phase space is created, which displays a number of key results: *Black contours* indicate completion times for replicating 95% of DNA. The red area indicates parameters giving a 95% completion time over 10 h and the green area indicates parameter settings giving a 95% completion time of less than 8 h. The *white area* therefore represents a set of biologically relevant parameters within which S phase would complete on schedule. To assist interpretation, additional features of interest have been imposed over the analysis: (1) *magenta contours* indicate the ratio of single/dual activation events as described in Fig. 5—contours are labelled intermittently along their length (numbered 2, 2.5, 3, 3.5); (2) *blue contours* indicate percentage of ab initio firing events—contours are labelled intermittently along their length (numbered 15, 16, 20, 30). Biologically interesting positions (**a**–**d**) are indicated by coloured icons on the figure and discussed as case studies in the text

model is most susceptible to fork failure. Even so, a 6% probability is required to drive completion of 95% of DNA replication beyond 9 h and a 16% probability of failure is required to prolong S phase beyond 10 h. The 'hybrid model' is less sensitive to fork failure and completes 95% of DNA replication within 9 h even with a 15% chance of fork failure. In this case, S phase completes within 10 h as long as the probability of fork failure does not exceed 34%. Interestingly, increased levels of fork stalling also drives the hybrid model to generate a higher ratio of single/dual coupled activation events, while the spatially driven model maintains a constant ratio.

Predictions of replication timing profile generated by the final model were tested for biological efficacy by comparison with timing profiles generated using TimEX-seq protocols from human ES cells (Desprat et al. 2009). The replication timing data for chromosome 6 was segmented into 100 time windows and a cumulative frequency profile showing the progress of DNA synthesis was generated (Fig. 7c). The whole chromosome profile was then segmented into R- and G-band regions using the recognised coordinates (see above) to generate separate timing profiles for the two major chromatin compartments. Comparison of the our S phase simulation with the TimEX-seq profiles (Fig. 7) shows that the replication timing data generated from human ES cells map closely to the data generated by our in silico simulations. Similarities were most evident at the level of total synthesis, where in both cases the accumulation of replicated DNA was essentially linear. However, the individual

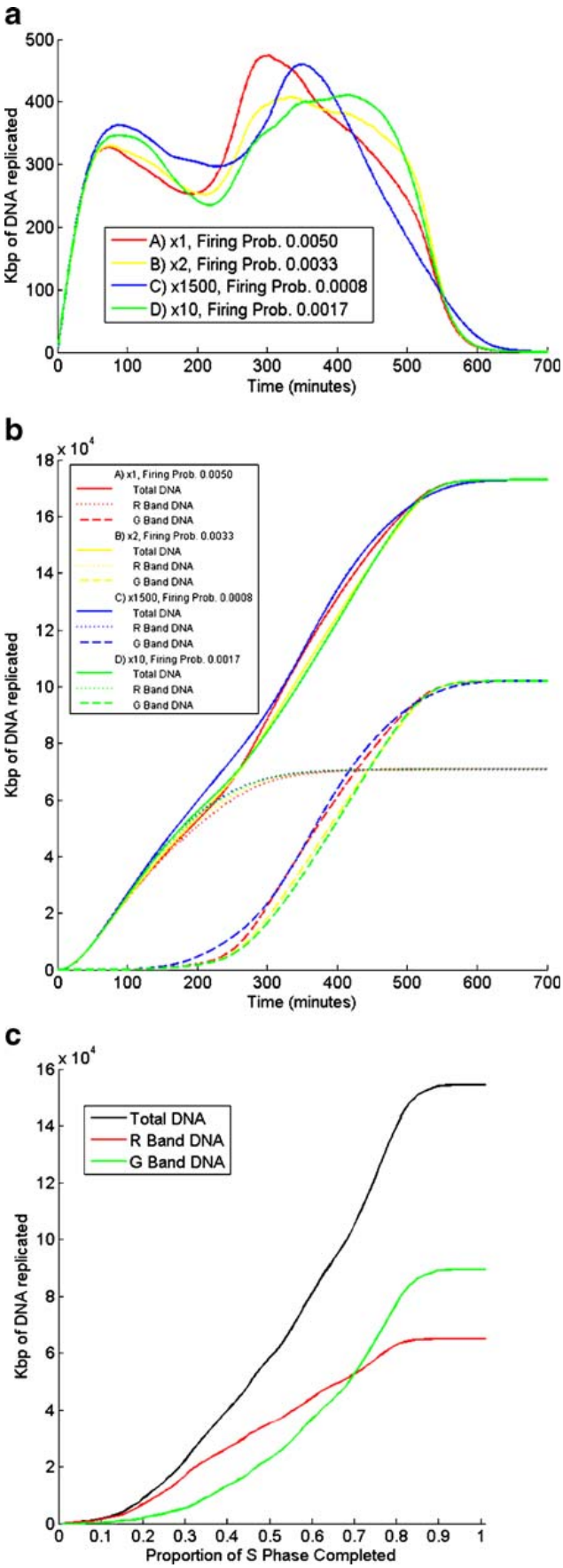


Fig. 7 Testing the models—comparison with genome-wide replication timing data. Four sets of simulations (**a**, **b**) were performed using the parameter sets highlighted in Fig. 6. For each, a model was created as described in Fig. 4, using one of a range of firing probabilities for G band clusters. For each firing probability, spatial effects were then tested based on the activation of clusters by encroaching forks (Fig. 5). Different plots (*coloured lines*) indicate parameters used in each set of simulations (see keys). Amplification factors (xn) define the adjusted firing probability that was applied when a cluster is activated by encroaching replication forks. Maximum firing probability refers to the probability of firing of a G-band cluster at the 6-h time point, based on the optimal concentration of activating cyclin-CDK complexes at that time. For each set of simulations (averages of 1,000 independent simulations are shown), the amount of DNA replicated (kbp/min) at each time point is determined (**a**) and converted into a cumulative replication profile (**b**), which shows the progress of synthesis. For each case study, solid lines indicate the total DNA replicated and *broken lines* display DNA synthesis within chromosomal R-bands (*dotted lines*) and G-bands (*dashed lines*). **c** The in silico simulations shown (**b**) were tested against experimentally derived profiles using the TimEX-seq data set from human ES cells (Desprat et al. 2009). The replication profile for chromosome 6 was generated by segmenting the published TimEX-seq data into 100 time intervals. This data set showing the amount of synthesis at different points throughout S phase was converted into a cumulative frequency plot of genome duplication across the sample. Plots showing S-phase progression were generated for the entire chromosome (Total DNA) and individually for chromosomal R- and G-bands, as shown

profiles for replication of R- and G-band DNA show significant discordance. This was particularly evident during mid-S phase, when the TimEX-seq data showed a higher level of G-band replication and prolonged R-band synthesis. Based on these profiles, the basic assumption that synthesis of R- and G-band DNA occurs during mutually distinct periods of S phase appears to be flawed. Hence, while the preference to engage synthesis in R-bands before G-bands is clear, the data do not suggest that an obligatory mechanism ensures that the cytologically defined chromosomal bands are replicated in a strict temporal order.

Conclusions and perspective

It has been known for many years that sites of initiation of DNA synthesis in mammalian cells are closely linked to local levels of RNA synthesis and that in general terms synthesis in gene-rich chromosomal R-bands occurs early in S phase and G-band

synthesis occurs later. Hence, the synthesis of mammalian genomes is thought to follow a temporal programme, which could be of fundamental biological importance if distinct chromatin states are specifically reproduced at defined times of S phase.

In this review, we set out to assess how different experimental approaches have been used to inform our understanding of DNA synthesis in mammalian cells and then assimilate ideas from different sources into a model of S-phase progression. Simulations were then used as an in silico approach to test alternative models of S phase. We tested a number of basic features related to genome architecture and local chromatin environments and then focussed on alternative mechanisms that might allow synthesis to propagate throughout the mammalian genome. In particular, we assessed how replication might spread between replication domains that contain ~1 Mbp of DNA. Specifically, we evaluated the behaviour of models of S phase that were based on both the stochastic activation of replication domains and the sequential activation of genetically linked DNA foci, according to the 'next-in-line' hypothesis of S-phase progression (Manders et al. 1992; Sporbert et al. 2002). As an alternative to these extremes, we considered a hybrid model, which incorporates a combination of S phase propagation using the next-in-line principle together with a level of external or ab initio activation events that are not influenced by encroaching forks from neighbouring replicons. Such initiation events might arise with different probabilities at different times of S phase, for example in response to changes in expression of specific cyclin-CDK complexes as S phase proceeds (Katsuno et al. 2009). The hybrid model incorporates a spatial component and temporal features related to changes in the chromatin environment. This model also accommodates a variable probability of origin activation so that the probability of clusters firing within G-bands remains low, but is enhanced by the presence of encroaching replication forks. Interestingly, we find that while this hybrid model is less reliant on fork elongation than the basic fork encroachment model, it shares some of the spatial dynamic benefits whilst being less susceptible to fork stalling. The fitness of this model is thus at least partially reliant on the probability that forks progress from one cluster to the next and appears to provide the best representation of the system in vivo.

In testing a range of alternative models, we have defined a parameter space that is likely to describe the biologically relevant mechanisms of S-phase progression in mammalian cells (region of interest highlighted in Fig. 6). Under the optimal parameter settings, comparison with experimental data shows that the model provides an excellent representation of replication for human chromosome 6 during the mammalian S phase (Fig. 7). However, we note a significant discrepancy between experimental data (Desprat et al. 2009) and our simulations of replication timing for designated R- and G-bands. This failure of the model implies that the chromosome-wide timing and order of R- and G-band replication is not defined with high precision. In particular, it is notable that while early cytological studies described a clear temporal separation in R- and G-band replication (Drouin et al. 1994) genome-wide analysis of the timing programme has shown that R-bands replicate before G-bands but that replication of the cytologically defined DNA compartments occurs throughout S phase (Desprat et al. 2009). Many features of the replication process might contribute to this observation. In particular, while genome-wide studies give a composite view of synthesis within huge cell populations it is clear that potential origins are used inefficiently so that the time of replication of specific chromosomal regions must reflect their location relative to the nearest active origin. While regions of the genome that have a high-density of active genes provide hot-spots for initiation of DNA replication—these will likely correlate with active regions at the onset of S phase – regions with lower levels of transcriptional activity provide weak targets for initiation and appear to replicate inefficiently, so that many potential origins are not used in most cells.

Based on our analysis, it is clear that the temporal restriction of R- and G-band replication to specific periods of S phase is an over-simplification that must be re-evaluated if we are to develop biologically robust models of S-phase progression. Specifically, it will become necessary to move away from the low resolution cytological chromosomal banding patterns, which generally incorporate chromosomal sub-domains of 5–20 Mbp, and towards high-resolution patterns of chromatin epi-states that better reflect local patterns of gene expression. Such improvements in resolution should provide a better insight into the molecular mechanisms that drive the mammalian S phase so that synthesis is performed with the efficacy

required to ensure the preservation of genome integrity.

Acknowledgements The authors are indebted to BBSRC (AS, DJ), the Wellcome Trust (AM-M) and CONACyT (National Council for Science and Technology, Mexico; PO-C) for support.

References

Aladjem MI (2007) Replication in context: dynamic regulation of DNA replication patterns in metazoans. Nat Rev Genet 8:588–600

Berezney R, Dubey DD, Huberman JA (2000) Heterogeneity of eukaryotic replicons, replicon clusters and replication foci. Chromosoma 108:471–484

Blow JJ, Dutta A (2005) Preventing re-replication of chromosomal DNA. Nat Rev Mol Cell Biol 6:476–486

Buongiorno-Nardelli M, Micheli G, Carri MT, Marilley M (1982) A relationship between replicon size and supercoiled loop domains in the eukaryotic genome. Nature 298:100–102

Cadoret J-C, Prioleau M-N (2010) Genome-wide approaches to determining origin distribution. Chromosome Res (in press)

Cadoret J-C, Meisch F, Hassan-Zadeh V et al (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci USA 105:15837–15842

Conti C, Sacca B, Herrick J, Lalou C, Pommier Y, Bensimon A (2007) Replication fork velocities at adjacent replication origins are co-ordinately modified during replication in human cells. Mol Biol Cell 18:3059–3067

Courbet S, Gay S, Arnoult N et al (2008) Replication fork movement sets chromatin loop size and origin choice in mammalian cells. Nature 455:557–560

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

DePamphilis ML, Blow JJ, Ghosh S et al (2006) Regulating the licensing of DNA replication origins in metazoan. Curr Opin Cell Biol 18:231–239

Desprat R, Thierry-Mieg D, Lailler N et al (2009) Predictable dynamic programme of timing of DNA replication in human cells. Genome Res 19:2288–2299

Donaldson AD (2005) Shaping time: chromatin structure and the DNA replication programme. Trends Genet 21:444–449

Drouin R, Holmquist GP, Richer CL (1994) High-resolution replication bands compared with morphological G-bands and R-bands. Advances Hum Genet 22:47–115

Farkash-Amar S, Simon I (2010) Genome-wide analysis of the replication programme in mammals. Chromosome Res (in press)

Farkash-Amar S, Lipson D, Polten A et al (2008) Global organization of replication time zones of the mouse genome. Genome Res 18:1562–1570

Fu Y, Sinha M, Peterson CL, Weng Z (2008) The insulator binding protein CTCF positions 20 nucleosomes around its binding sites across the human genome. PLoS Genet 4: e1000138+

Ge XG, Blow JJ (2009) A model for DNA replication showing how dormant origins safeguard against replication fork failure. EMBO Rep 10:406–412

Gilbert DM (2004) In search of the holy replicator. Nat Rev Mol Cell Biol 5:848–854

Goen A, Cedar H (2003) Replicating by the clock. Nat Rev Mol Cell Biol 4:25–32

Goetze S, Mateos-Langerak J, Gierman HJ et al (2007) The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. Mol Cell Biol 27:4475–4487

Goldar A, Marsolier-Kergoat M, Hyrien O (2009) Universal temporal.profile of replication origin activation in eukaryotes. PLoS ONE 4:1–7

Hadjur S, Williams LM, Ryan NK et al (2009) Cohesins form chromosomal cis-interactions at the developmentally regulated IFNG locus. Nature 460:410–413

Hamlin JL, Mesner LD, Dijkwel PA (2010) A winding road to origin discovery. Chromosome Res (in press)

Herrick J, Jun S, Bechhoefer J, Bensimon A (2002) Kinetic model of DNA replication in eukaryotic organisms. J Mol Biol 320:741–750

Hiratani I, Ryba T, Itoh M et al (2008) Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol 6:2220–2236

Holmquist GP (1987) Role of replication time in the control of tissue-specific gene expression. Am J Hum Genet 40:151–173

Hozak P, Jackson DA, Cook PR (1994) Replication factories and nuclear bodies: the ultrastructural characterization of replication sites during the cell cycle. J Cell Sci 107:2191–2202

Hyrien O, Goldar A (2010) Mathematical modelling of eukaryotic DNA replication. Chromosome Res (in press)

Jackson DA (1995) Nuclear organization: uniting replication foci, chromatin domains and chromosome structure. BioEssays 17:587–591

Jackson DA, Pombo A (1998) Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J Cell Biol 140:1285–1295

Jeon Y, Bekiranov S, Karnani N et al (2005) Temporal profiles of replication of human chromosomes. Proc Natl Acad Sci USA 202:6419–6424

Karnani N, Taylor C, Malhotra A, Dutta A (2007) Pan-S replication patterns and chromosomal domains defined by genome-tiling arrays of ENCODE genomic areas. Genome Res 17:865–876

Karolchik D, Baertsch R, Diekhans M et al (2003) UCSC genome browser database. Nucl Acids Res 31:51–54

Katsuno Y, Suzuki A, Sugimura K et al (2009) Cyclin A-CDK1 regulates the origin firing programme in mammalian cells. Proc Natl Acad Sci USA 106:3184–3189

Lande-Diner L, Zhang JM, Cedar H (2009) Shifts in replication timing actively affect histone acetylation during nucleosome reassembly. Mol Cell 34:767–774

Lebofsky R, Heilig R, Sonnleitner M, Weissenbach J, Bensimon A (2006) DNA replication origin interference increases the

spacing between initiation events in human cells. Mol Biol Cell 17:5337–5345

Leonhardt H, Rahn HP, Weinzierl P et al (2000) Dynamics of DNA replication factories in living cells. J Cell Biol 149:271–279

Lygeros J, Koutroumpas K, Dimopoulos S et al (2008) Stochastic hybrid modeling of DNA replication across a complete genome. Proc Natl Acad Sci USA 105:12295–12300

Ma H, Samarabandu J, Devdhar RS et al (1998) Temporal dynamics of DNA replication sites in mammalian cells. J Cell Biol 143:1415–1425

Manders EM, Stap J, Brakenhoff GP, van Driel R, Aten JA (1992) Dynamics of 3-dimensional replication patterns during the S-phase, analyzed by double labelling of DNA and confocal microscopy. J Cell Sci 103:857–862

Maya-Mendoza A, Petermann E, Gillespie DA, Caldecott KW, Jackson DA (2007) Chk1 regulates the density of active replication origins during the vertebrate S phase. EMBO J 26:2719–2731

Maya-Mendoza A, Tang CW, Pombo A, Jackson DA (2009) Mechanisms regulating S phase progression in mammalian cells. Front Biosci 14:4199–4213

Mendez J, Stillman B (2003) Perpetuating the double helix: molecular machines at eukaryotic DNA replication origins. BioEssays 25:1158–1167

Mesner LD, Crawford EL, Hamlin JL (2006) Isolating apparently pure libraries of replication origins from complex genomes. Mol Cell 21:719–726

Nakamura H, Morita T, Sato C (1986) Structural organizations of replication domains during DNA synthetic phase in the mammalian nucleus. Exp Cell Res 165:291–297

Parelho V, Hadjur S, Spivakov M et al (2008) Cohesins functionally associate with CTCF on mammalian chromosome arms. Cell 132:422–433

Phillips JE, Corces VG (2009) CTCF: master weaver of the genome. Cell 137:1194–1211

Pope BD, Hiratani I, Gilbert DM (2010) Domain-wide regulation of DNA replication timing during mammalian development. Chromosome Res (in press)

Raghuraman MK, Winzeler EA, Collingwood D et al (2001) Replication dynamics of the yeast genome. Science 294:115–121

Rhind N (2006) DNA replication timing: random thoughts about origin firing. Nat Cell Biol 12:1313–1316

Rhind N, Yang SC-H, Bechhoefer J (2010) Reconciling stochastic origin firing with defined replication timing. Chromosome Res (in press)

Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A et al (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet 5:e1000446

Shopland LS, Lynch CR, Peterson KA et al (2006) Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. J Cell Biol 174:27–38

Sporbert A, Gahl A, Ankerhold R, Leonhardt H, Cardoso MC (2002) DNA polymerase clamp shows little turnover at established replication sites but sequential de novo assembly at adjacent origin clusters. Mol Cell 10:1355–1365

St-Jean P, Vaillant C, Audit B, Arneodo A (2008) Spontaneous emergence of sequence-dependent rosette like folding of chromatin fiber. Phys Rev E 77:e061923

Strehl S, Lasalle J, Lalande M (1997) High-resolution analysis of DNA replication domain organization across an R/G-band boundary. Mol Cell Biol 17:6157–6166

Takebayashi SI, Manders EM, Kimura H, Taguchi H, Okumura K (2001) Mapping sites where replication initiates in mammalian cells using DNA fibres. Exp Cell Res 271:263–268

Tuduri S, Tourrière H, Pasero P (2010) Defining replication origin efficiency using DNA fiber assays. Chromosome Res (in press)

Vogelauer M, Rubbi L, Lucas I, Brewer BJ, Grunstein M (2002) Histone acetylation regulates the time of replication origin firing. Mol Cell 10:1223–1233

West AG, Fraser P (2005) Remote control of gene transcription. Hum Mol Genet 14:R101–R111

Woodfine K, Fiegler H, Beare DM et al (2004) Replication timing of the human genome. Hum Mol Genet 13:191–202

Zhang JM, Xu F, Hashimshony T, Keshet N, Cedar H (2002) Establishment of transcriptional competence in early and late S phase. Nature 420:198–202

Zhang N, Kuznetsov SG, Sharan SK, Li K, Rao PH, Pati D (2008) A handcuff model for the cohesin complex. J Cell Biol 183:1019–1031

Zink D (2006) The temporal programme of DNA replication: new insights into old questions. Chromosoma 115:273–287

A.2 S phase progression in human cells is dictated by the genetic continuity of DNA foci.

Maya-Mendoza A, **Olivares-Chauvet P**, Shaw A, Jackson DA. S phase progression in human cells is dictated by the genetic continuity of DNA foci. PLoS Genetics. 2010;6:e1000900.

Summary

By using modified DNA precursors as a label, after consecutive labeling pulses, we show that the S phase programme follows a genetic order instead of activating dormant origins by spatial proximity. We compared the distribution if foci sizes observed in the microscope against replication timing profiles and show that replication domains in the replication timing profiles share the same size distribution.

Contribution

General discussions for the manuscript, data analysis and bioinformatic analysis.

OPEN ACCESS Freely available online

PLoS GENETICS

S Phase Progression in Human Cells Is Dictated by the Genetic Continuity of DNA Foci

Apolinar Maya-Mendoza, Pedro Olivares-Chauvet, Alex Shaw, Dean A. Jackson*

Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom

Abstract

DNA synthesis must be performed with extreme precision to maintain genomic integrity. In mammalian cells, different genomic regions are replicated at defined times, perhaps to preserve epigenetic information and cell differentiation status. However, the molecular principles that define this S phase program are unknown. By analyzing replication foci within discrete chromosome territories during interphase, we show that foci which are active during consecutive intervals of S phase are maintained as spatially adjacent neighbors throughout the cell cycle. Using extended DNA fibers, we demonstrate that this spatial continuity of replication foci correlates with the genetic continuity of adjacent replicon clusters along chromosomes. Finally, we used bioinformatic tools to compare the structure of DNA foci with DNA domains that are seen to replicate during discrete time intervals of S phase using genome-wide strategies. Data presented show that a major mechanism of S phase progression involves the sequential synthesis of regions of the genome because of their genetic continuity along the chromosomal fiber.

Citation: Maya-Mendoza A, Olivares-Chauvet P, Shaw A, Jackson DA (2010) S Phase Progression in Human Cells Is Dictated by the Genetic Continuity of DNA Foci. PLoS Genet 6(4): e1000900. doi:10.1371/journal.pgen.1000900

Editor: Jeannie T. Lee, Massachusetts General Hospital, Howard Hughes Medical Institute, United States of America

Received April 16, 2009; **Accepted** March 8, 2010; **Published** April 8, 2010

Funding: This research was supported by The Biotechnology and Biological Sciences Research Council (grant BBS/B/06091 AM-M, DAJ), The Wellcome Trust (grant 080172/Z AM-M, DAJ), and CONACyT (National Council for Science and Technology, Mexico; PO-C). The funders had no role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dean.jackson@manchester.ac.uk

Introduction

DNA synthesis in eukaryotes must be performed with absolute precision as any defects compromise genetic integrity. In all eukaryotes, DNA is duplicated during S phase of the cell cycle, which is regulated to ensure that DNA synthesis is completed before mitosis can begin [1–3]. During synthesis, different regions of the genome are replicated at specific times [4–6], perhaps as a part of a fundamental mechanism that ensures the preservation of epigenetic information [7]. Within this timing program, chromatin within gene-rich chromosomal R-bands is known to begin early in S phase, before synthesis of heterochromatic G-bands takes place. This general structure can be revealed at low resolution, using cytological chromosome banding [8], and at higher resolution using genome-wide strategies [9–15].

Recent developments in genome-wide analysis have revolutionized our ability to define the structure of S phase in higher eukaryotes. However, detailed analysis of the replication program has been limited by our understanding of the molecular mechanisms that control how specific origins are used at different times. In mammalian cells, recent studies have shown that local chromatin environments define a general preference for origins that are activated during early S-phase [10–15]. Regions that engage synthesis at the onset of S phase frequently have a locally high gene density and correspondingly high levels of RNA synthesis. In addition, more detailed analysis is beginning to explore how local chromatin features such as the distribution of CpG islands [14] and local chromatin accessibility [15] contribute to patterns of origin selection.

Single cell studies provide an alternative strategy for understanding S phase progression. Active sites of DNA synthesis can be revealed as replication foci [16,17], which contain groups of replicons that are replicated together within dedicated replication factories [18]; such replicon clusters typically contain 3–5 replicons within ~1 Mbp of DNA [19,20]. DNA foci are thought to represent fundamental unit of chromosome structure [19–23] that are defined by local chromatin environments [23–25] and replicated during defined intervals of S phase [26,27]. Perhaps importantly, foci that are replicated during consecutive intervals of S phase appear to lie side-by-side in nuclei [28,29], suggesting that their organization contributes to replication timing.

During S phase, the organization of replicons within replicon clusters defines how long individual DNA foci are engaged in synthesis. In HeLa cells, during early S phase, the average speed of fork elongation is ~1.5 kbp/min/fork [19,30]. As the average distance between adjacent origins in replicon clusters is ~150 kb (90% of adjacent origins are typically ~50–250 kb apart) most will be engaged in synthesis for 1–2 h before the internal forks from adjacent replicons meet and terminate by fork fusion. When this occurs, the rate of synthesis can only be maintained if new origins are activated. Hence the progressive activation and completion of synthesis within the ~1 Mbp DNA foci defines a replication timing program within which different cohorts of foci are replicated within time zones that occupy ~1–2 hours of S phase.

Mechanisms of origin selection that define S phase timing are known to show remarkable plasticity during cell differentiation [10,12,15]. However, within a particular cell type, the extent to which DNA replication is deterministic – and hardwired by

Author Summary

Eukaryotic DNA synthesis is regulated with exquisite precision so that genomes are replicated exactly once before cell division occurs. In simple eukaryotes, chromosomal loci are preferentially replicated at specific times of S phase, in part because of their differential sensitivity to cell cycle regulators and in part as a result of random choice. Mammals, with ~250-fold larger genomes, have more complex replication programs, within which different classes of chromatin replicate at defined times. While the basic regulatory mechanisms in higher eukaryotes are conserved, it is unclear how their much more complex timing program is maintained. We use replication precursor analogues, which can be visualized in living or fixed cells, to monitor the spatial relationship of DNA domains that are replicated at different times of S phase. Analyzing individual chromosome, we show that a major mechanism regulating transitions in the S phase timing program involves the sequential activation of replication domains based on their genetic continuity. Our analysis of the mechanism of S phase progression in single cells provides an alternative to genome-wide strategies, which define patterns of replication using cell populations. In combination, these complimentary strategies provide fundamental insight into the mechanisms of S phase timing in mammalian cells.

chromosome structure – or stochastic – and so varies for cell to cell – remains a matter of debate. To address this question, we designed an experimental approach that would allow us to analyze the spread of DNA synthesis throughout nuclei of individual cells (Figure 1). Sites of DNA synthesis within DNA foci were labeled with thymidine analogues using pulse and pulse-chase-pulse strategies and analyzed over many days. Initially, labeled foci are distributed throughout all chromosomes but as cells proliferate random mitotic segregation reduces the number of labeled chromosomes within individual cells so that chromosome territories (CT) and their DNA foci are resolved. Immediately after labeling it is impossible to establish the extent to which adjacent foci are related by their spatial and/or genetic continuity, as the alternative models are indistinguishable. However, following chromosome segregation, the plasticity of CT structure [24] allows the spatially and genetically determined models to be distinguished (Figure 1B). Hence, over many cell division cycles, the analysis of individual CTs provides a high-resolution memory of cis- and trans-activation events that define the replication timing program.

We used 3D and 4D light microscopy to analyze the organization of DNA foci within individual CTs of nuclei and mitotic chromosomes. We show that the sequential replication of DNA foci is defined by their genetic association along individual chromosomes. To visualize the genetic association directly, we analyzed individual DNA fibers from cells that were labeled during sequential 1 h intervals of S phase. We conclude that the sequential activation of adjacent replicon clusters represents a major mechanism of S phase progression. Indeed, once early synthesis has begun, only a minority – about 10% - of *de novo* initiation events are genetically uncoupled from sites that were engaged in synthesis earlier during S phase. Finally, in order to integrate this conclusion with the analysis of replication using genome-wide strategies, we used bioinformatic tools to show that the structure of replicon clusters within DNA foci and lengths of replication timing domains correlate with extremely high significance. This is consistent with DNA foci being the stable higher-order units of chromatin packaging that define the replication timing program in mammalian cells.

Results

S phase progression is defined by the spatial organization of DNA foci

In HeLa cells in early S phase, the template for DNA synthesis is folded into DNA foci that can be labeled with a variety of modified thymidine analogues and visualized in both living and fixed cells (Figure S1). Different pulse and pulse-chase-pulse strategies can then be used to evaluate the relationship of foci that are engaged in DNA synthesis during different intervals of S phase (Figure S2). In mid/late S phase, the spatial relationship of foci that were labeled during consecutive intervals of S phase is evident because distinct patterns of active sites are seen at this time (Figure S1A and S1C). In early S phase (Figure S1B), in contrast, spatial analysis at the time of labeling is much less informative because of the high density of active sites.

To evaluate the alternative models of S phase progression described in Figure 1, cells were labeled with two consecutive pulse-labels and grown for many days to leave ~3 labeled CTs/cell (Figure 1C, Figure 2, Figure 3). As a control, we first monitored the co-association of labeled foci in metaphase, as this defines their distribution within individual chromosomes (Figure 1). Metaphase images, from cells that were labeled during early S phase, showed that all labeled chromosomes within double-labeled cells contained early S phase foci that incorporated both the 1st and 2nd precursors. However, as chromosome condensation during metaphase limits the resolution of the spatial analysis, we next monitored the level of co-association within interphase CTs [23]. Analysis of CTs showed that foci labeled with the 1st replication precursor were within 500 nm of a focus labeled with the 2nd. In addition, time-lapse imaging of foci in living cells showed this co-association to be maintained when cells were monitored for up to 3 hours. Throughout the imaging time course (Figure 2 and Video S1, S2, S3), individual CTs showed dramatic plasticity [31], with shape transformations during cell movement resulting in early S phase foci displaying frequent relative positional shifts of 0.2–0.6 μm over 30 min. Notably, during these shifts, the association of adjacent foci labeled during the 1st and 2nd pulses was always maintained (25 CTs were analyzed by live imaging and labeled foci showed the same behavior in all cases).

The sequential activation of DNA foci is defined by their spatial continuity within individual chromosomes

To reinforce the interpretation of time-lapse imaging, we evaluated the spatial relationship of interphase foci labeled during consecutive intervals of early S phase using 3D confocal microscopy (Figure 3). To test the models described in Figure 1, we measured the spatial separation of nearest foci containing the 1st and 2nd precursors using both consecutive pulses and pulses separated by intervening unlabeled periods of 1 or 2 hours. Experiments were performed using both fixed (Figure 3B and 3C) and living (Figure 3D and 3E) cells. Living cells were analyzed directly and fixed cells were processed for 3D confocal imaging by indirect immuno-labeling.

Following image capture, image analysis software was used to define the center of mass of labeled sites (Figure 3B6) and then measure the 3D separation of the nearest sites labeled during the 1st and 2nd pulses (Figure 3G and 3H). Under the experimental conditions used, DNA foci in HeLa cells have an average diameter of ~350 nm (Figure 3F). Moreover, as living and fixed cells show the same size distribution, our experimental strategies do not

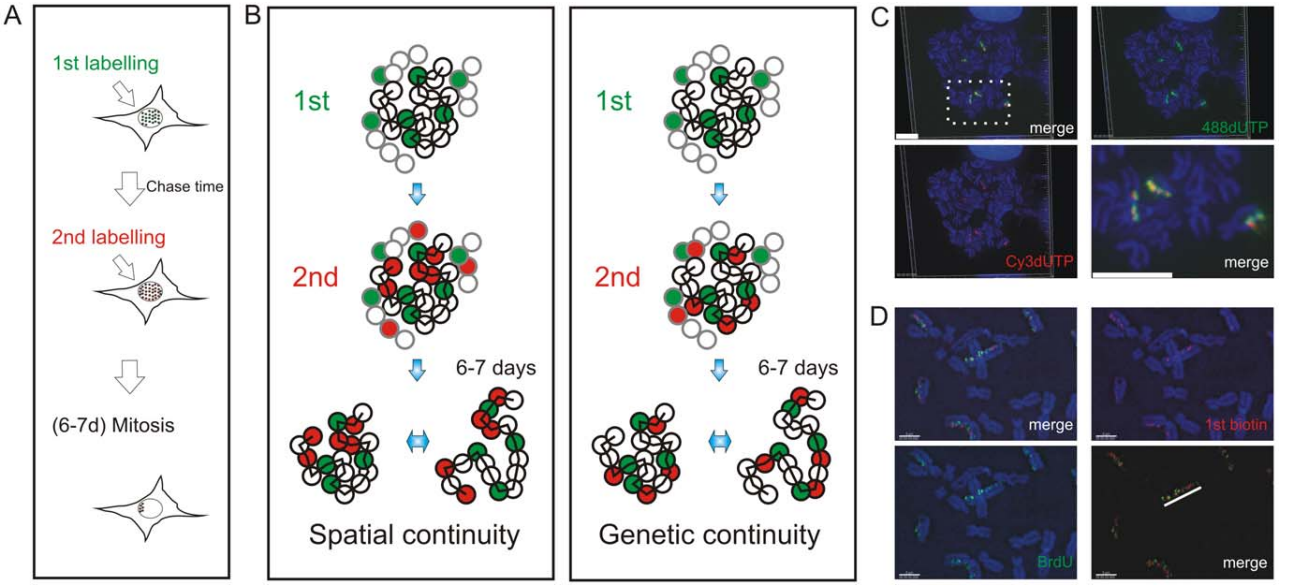


Figure 1. Double-labeled replication foci are segregated in specific regions of mitotic chromosomes. Different dUTP analogues were incorporated into newly replicated DNA and individual chromosomes resolved by random mitotic segregation over 6–7 days (A). Different models (B) show possible relationships between individual DNA foci that are replicated at different times of S phase. In each panel, the replication foci of a single CT (spheres with black rims) and parts of three adjacent CTs (spheres with grey rims) are shown. Foci within the central CT are genetically linked along the chromosome fiber (black zig-zag line). During pulse labeling, some foci are labeled during the 1st pulse (green) and others during the 2nd (red). At this time, the alternative models are indistinguishable, with all green foci lying adjacent to neighboring red foci. 6–7 days later, the foci of individual CTs can be visualized as the surrounding CTs are no longer labeled. The innate plasticity of CTs (2 inter-changeable forms are shown) supports distinct predictions about S phase progression: i) if progression is based on spatial continuity of foci at the time of labeling subsequent changes in CT structure will degrade the side-by-side relationship of foci whereas ii) if progression is based on genetic continuity the side-by-side relationship will be preserved. HeLa cells (C) were labeled with AF448-dUTP (green) and Cy3-dUTP (red), grown for 6 days and DAPI-stained chromosome spreads prepared. Deconvolution microscopy shows that 100% (n = 65 chromosomes from 25 metaphase plates) of the labeled chromosomes incorporated both dUTP analogues and that all labeled regions (note that labeling appears in chromosomal bands at this level of resolution) contained both analogues. A merge of the individual channels and a high-resolution merge of the highlighted region (rectangle) are shown to emphasise co-association of the 1st and 2nd labels. Diploid human fibroblasts (D) were labeled with biotin-dUTP and BrdU with an intervening unlabeled period of 1h. Labeled chromosomes were resolved by random mitotic segregation (6 days) and confocal imaging performed following indirect immuno-fluorescence using specific antibodies to biotin (red) and BrdU (green). Individual red and green channels and a channel merge were overlaid on the DAPI-stained chromosomes as shown. Merged images with the DAPI removed (D, bottom right panel) were used to demonstrate the co-association of foci along individual chromosomes - the white line highlights the labeled foci along one chromatid of a single chromosome. Scale bars: 10 μm in (C) and 5 μm in (D).
doi:10.1371/journal.pgen.1000900.g001

appear to disrupt local chromatin architecture during processing and imaging (Figure 3F; t test association probability $p<0.07$ $n = 60$). While analysis of both fixed and living cells demonstrates the stability of foci with sizes of 300–500 nm, we note that recent advances in light microscopy (3D-SIM and SMI microscopy) reveal that individual foci can be resolved into sub-domains with an average size of ~125 nm [Cristina Cardoso and Vadim Chagin, Technische Universität Darmstadt, personal communication].

When unsynchronized cells were labeled with consecutive pulses, most foci were labeled with both precursors (Figure 3B and 3D); as synthesis within individual foci is not synchronized, a minority of foci might be labeled with only one precursor because they began or completed synthesis during the 1st or 2nd labeling periods. However, when the pulses were separated by 1 hour (Figure 3C and 3E) ~50% of foci were labeled with only one analogue (43.5% of foci in living cells ($n = 200$) and 52% in fixed cells ($n = 146$)). Nearest neighbor analysis was used to explore this spatial relationship quantitatively (Figure 3G and 3H). With consecutive pulses, the average center-center separation of the nearest red and green labeled sites was ~150 nm (Figure 3B and 3D) – as most foci are double labeled this center-to-center separation is less than the average diameter of individual foci. With an intervening chase of 1 h, the separation between adjacent foci labeled during the 1st and 2nd pulses increased to ~350 nm (Figure 3G and 3H). As this center-to-center separation is similar to the average diameter of foci in early S phase (Figure 3F) foci labeled during the consecutive intervals of S phase must lie close to or touching their nearest neighbor.

Two important controls emphasize the significance of this nearest neighbor analysis. First, we analyzed individual foci that were labeled simultaneously with 2 replication precursor analogues (Figure S3). This defines the reliability of distance measurements and the effect of experimental noise on the precision of data generated by the analysis. To demonstrate a worst-case-scenario, red and green foci with >2-fold average intensities were seen to give an average separation of no more than 75 nm (Figure S3). Second, we also measured the separation of foci labeled during either 1st or 2nd pulse to define the distribution of foci that were labeled with each precursor. Under all labeling conditions used, the average separation of nearest early S phase foci was ~500 nm (Figure S4), which is highly significantly different to the separation of neighboring foci labeled by consecutive pulses with an intervening chase (t test = 2.955E-12 comparing separation of BrdU foci in Figure S4D with separation of biotin and BrdU foci in Figure 3H).

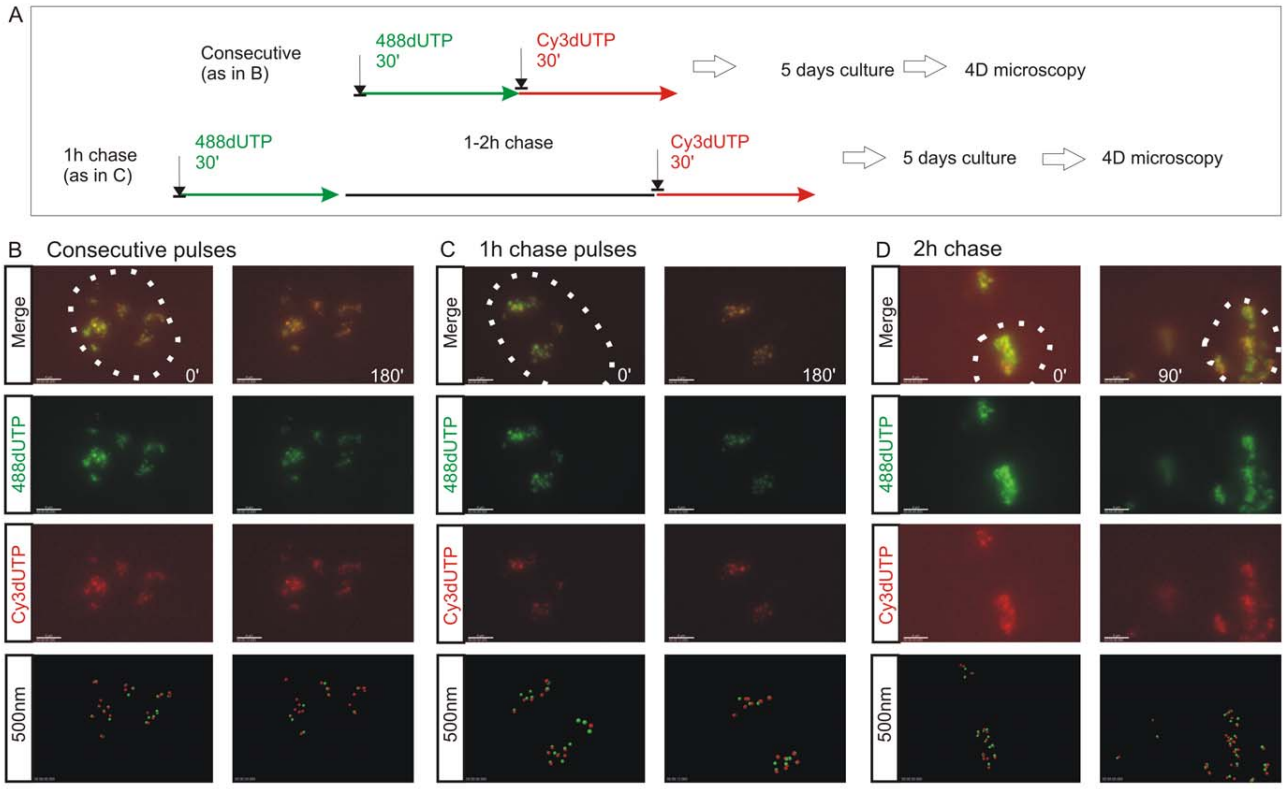


Figure 2. The spatial architecture of DNA foci is maintained in living cells. 4D time-lapse imaging was used to monitor the dynamic behavior of DNA foci (A–D). HeLa cells were labeled with consecutive pulses of AF488-dUTP (green) and Cy3-dUTP (red) with different times of intervening chase (A) and individual CTs resolved by mitotic segregation (B–D). Using consecutive pulses with no intervening unlabeled period (B), all CTs were labeled with both precursors, which were also co-associated within sub-regions of individual CTs (B) (Video S1, S2). CTs are seen to be highly dynamic, yet despite changes resulting from cell movement the spatial co-association of 1st and 2nd pulse-labels was always maintained throughout the imaging time course. Clear spatial co-association of the 1st and 2nd pulses was also seen when pulses were separated by unlabeled chase periods of 1h (C) (Video S3) and 2h (D), with adjacent foci labeled during the 1st and 2nd pulses maintaining separations of ~500 nm (B, 1 h chase: 390+/−148 nm n = 53; D, 2 h chase: 438+/−141 nm n = 57). For each labeling program (B–D), typical examples show isolated CTs within individual cells (nuclei are marked by dotted lines) that were imaged at 15 min intervals using time-lapse 3D microscopy for 3 h or more (data not shown). Individual green and red channels together with a two channel overlay (merge) and centers of mass of foci labeled in red and green channels (500 nm: labeled sites are depicted by foci of 500 nm diameter) are shown (B–D). For each experiment (B–D), 2 time points (0 and 90 or 180 min) are shown to emphasise changes in the structure of foci within individual CTs over time. Scale bars: 4 μm.
doi:10.1371/journal.pgen.1000900.g002

In a parallel study, we also performed a nearest neighbor analysis using normal diploid human fibroblasts (MRC5; Figure S5). While these diploid cells appear to have slightly larger early foci (513+/−116 nm; n = 200) than HeLa cells, perhaps as a consequence of their flattened shape, foci labeled with a separation of 1 h nevertheless maintained a strict side-by-side relationship (separation was 556+/−114nm; n = 155). These experiments show that DNA foci labeled during consecutive intervals of S phase retain a nearest neighbor relationship independently of changes in CT structure, consistent with the spatial relationship at the time of labeling being defined by the genetic connectivity of DNA foci along chromosomes. The significance of this strict side-by-side relationship was reiterated using *in silico* simulations to model the activation of DNA foci (Figure S6).

The replication timing program correlates with the spatial context of DNA foci

We next attempted to reinforce the links between S phase progression and the genetic continuity of DNA foci by monitoring the distribution of foci labeled during widely separated intervals of S phase. First, we analyzed cells in early S phase after labeling replication foci with 3 sequential replication precursors each separated by 1 hour (Figure S7). As expected, the separation of both consecutive labels – the separation between the 1st–2nd and 2nd–3rd precursors - was ~350 nm (Figure S7). However, a significantly larger separation of ~500 nm was seen when the separations of sites labeled with the 1st and 3rd precursors was measured (Figure S7C). This shows that even though the folding of DNA foci within individual CTs is complex and dynamic (Figure 2) the foci labeled at different times of early S phase show a progressive separation over time.

This progressive synthesis of early S phase replication foci is consistent with synthesis spreading along chromosomes at a rate of ~200 nm/h. Over longer periods - with separations of >4 hours - the linear continuity of labeled sites is difficult to define because nearest neighbor relationship are degraded by chromosome folding (Figure S8) and the distribution of euchromatin and heterochromatin in CTs [17,24,25]. Based on this observation, we would not rule out the possibility that early and mid/late S phase have distinct characteristics. Towards the end of early S phase, as the replication of euchromatin completes, many forks appear to pass from the early to mid/late replication domains [10–12]. At

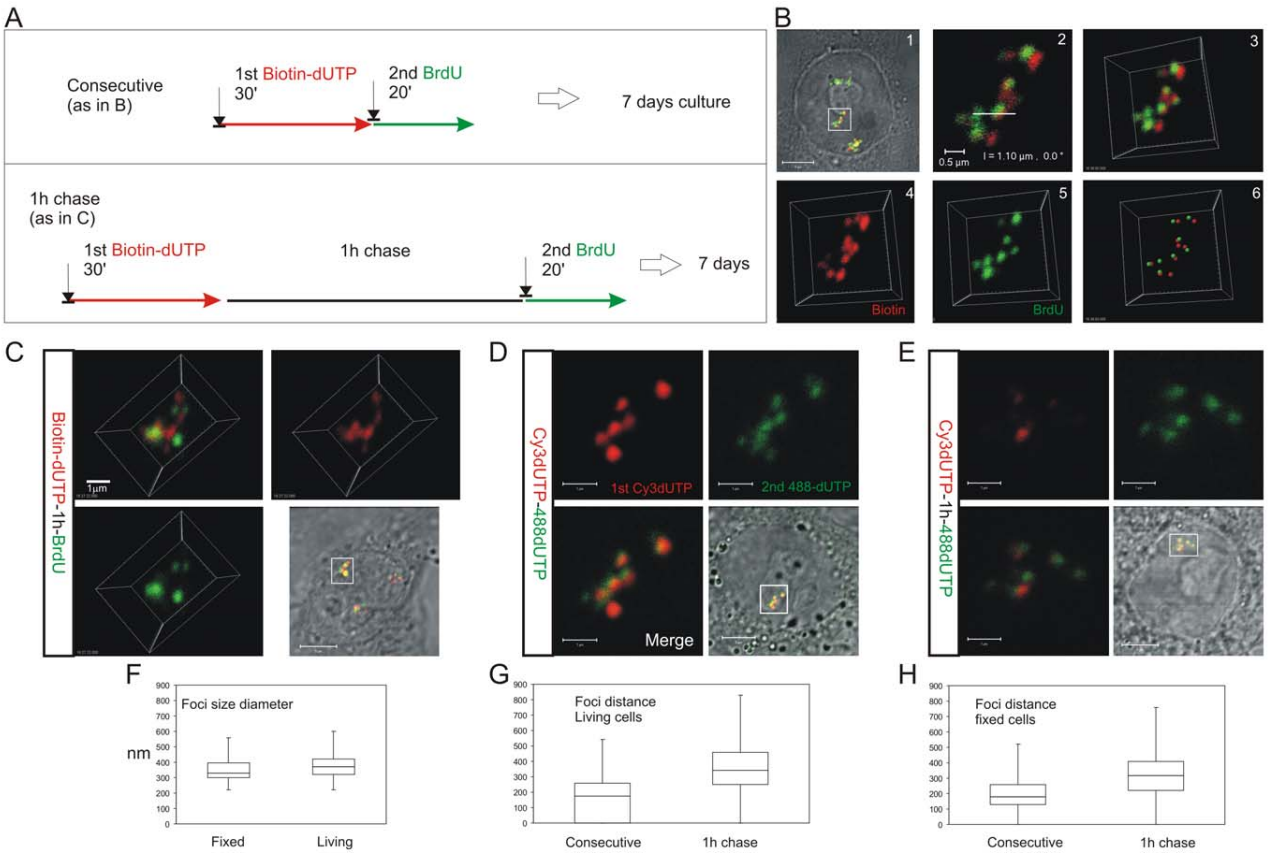


Figure 3. The S phase program is defined by the temporal activation of DNA foci at adjacent positions within CTs. HeLa cells were labeled with consecutive pulses of biotin-dUTP and BrdU either without or with an intervening unlabeled chase and grown for 6–7 days to resolve labeled CTs (A). Cells with individual labeled CTs were analyzed by confocal microscopy (B). Following consecutive pulses of biotin-dUTP (red) and BrdU (green) a cell with 3 CTs was selected (B1) and confocal sections of an individual CT (boxed area) taken (B2 shows a single confocal section) to produce a 3D projection of the entire CT (B3). Individual channels from the 3D projection were separated (B4,5) and mass centers of the labeled foci defined and combined (B6). Within this CT most foci are double-labeled though some are only labeled with the 1st or 2nd precursor. Double labeled CTs were analyzed following consecutive pulses (B,D) or pulses with an intervening chase (C,E) to monitor spatial continuity during S phase progression. Equivalent cells were analyzed by indirect immuno-fluorescence in fixed cells (B,C), after replication with biotin-dUTP (red) and BrdU (green), or under live imaging conditions (D,E), after replication with AF488-dUTP (green) and Cy3-dUTP (red). (D–E) show 2D confocal sections and (C) a 3D maximum projection of CTs highlighted (square) in the corresponding phase contrast images. Individual foci were measured to define their size distribution (F; diameter of foci; n = 60). Nearest neighbor analysis of labeled foci was performed (G,H) to define the separation of adjacent foci labeled with the 1st and 2nd analogues during consecutive pulses and pulses with an intervening chase. In both living (G; n = 200; t test: $p<2.05E-31$) and fixed (H; n = 167; t test: $p<4.7E-18$) cells the labeling patterns differed with a high degree of statistical significance. Scale bars: 5 and 0.5 μm in (B) and 5 and 1 μm in (C,D), respectively.
doi:10.1371/journal.pgen.1000900.g003

this time of S phase, a significant fraction (~5%) of chromatin is replicated by forks that extend for at least 500 kbp. Such temporal transition regions in the replication program [10] apparently engage synthesis for many hours without encountering and activating potential origins in heterochromatin.

Visualizing replication domains on single DNA molecules defines the genetic contribution to S phase progression

Nearest neighbor analysis is consistent with a genetically defined next-in-line model, which operates in cis within individual CTs (Figure 1). We next wanted to evaluate the extent to which this cis activation defines S phase progression. In nuclei, however, analysis is compromised by the dynamic properties of DNA foci within individual CTs. To avoid this limitation, we analyzed the genetic relationships of replication pulses along individual DNA fibers (Figure 4). DNA fibers were prepared by direct spreading of cells labeled with biotin-dUTP and BrdU with an intervening 1 h chase. Spreads were prepared directly from cells without prior DNA isolation in order to image isolated ~1–2 Mbp DNA fibers. As careful spreading, with only ~5 labeled cells per spread, prevents mixing of fibers from individual labeled cells [32], this approach allows us to capture biotin-labeled fibers from cells that were engaged in DNA synthesis during the 1st labeling period. Regions of spreads with dispersed biotin-labeled fibers were located and randomly selected fields recorded; low magnification was used so that each imaging field contained fibers with at least 0.8 Mbp of DNA. In 144 fields, from 4 equivalent experiments, the fibers analyzed contained 450 Mbp of DNA in total.

Double-labeled fibers were analyzed, as any forks growing throughout the labeling period will incorporate both 1st and 2nd precursors, which will be separated by a predictable distance that reflects the rate of fork elongation (Figure 4A and Figure S9). As seen before [19], the active replicons are often clustered into small groups that typically contain ~0.5–1 Mbp of DNA. This

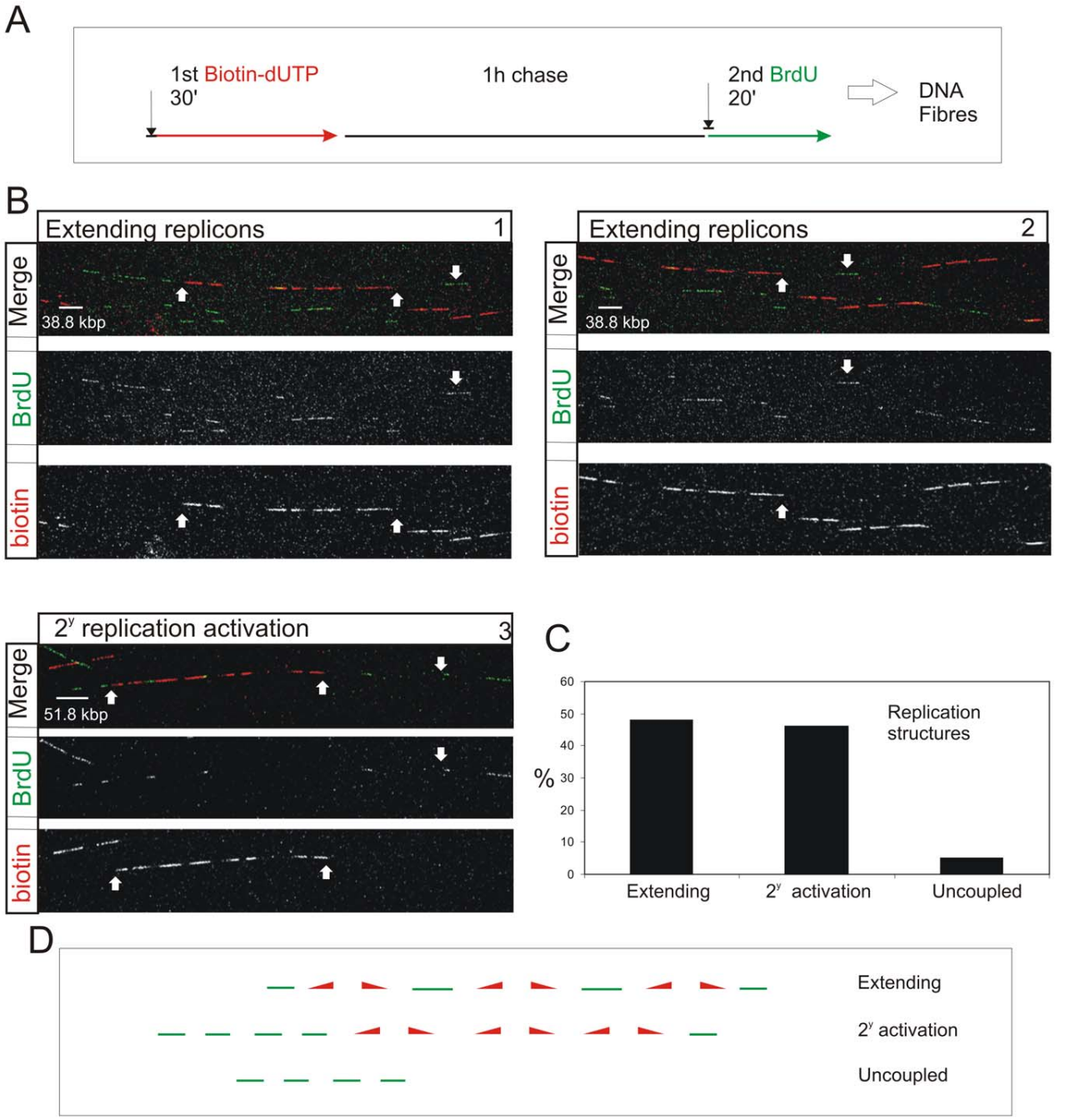


Figure 4. S phase progression correlates with the sequential activation of replicon clusters as defined by their genetic continuity along individual chromosome. Cells were pulse-labeled with biotin-dUTP and BrdU separated by 1 h without label (A) and double-labeled DNA fibers of >0.8 Mbp in length collected (B). Typical examples (B) show two major classes, where the 1st and 2nd pulse labels were incorporated into genetically adjacent replicon clusters. (B) panels 1–2 show a single fiber that extends over two adjacent imaging fields; the up pointing arrows show part of the replicon cluster labeled with biotin-dUTP during the 1st pulse; down pointing arrows show BrdU incorporation between two growing replication forks. Panel 3 shows a typical cluster with four active replicons, which were labeled during the 1st pulse, and two adjacent replicon clusters (defined my multiple Br-labeled tracks) activated during the 2nd pulse. In other clusters the labeling was confined within a single active cluster that was labeled during both periods of incorporation (Figure S9). To analyze genetic continuity, BrdU incorporation was monitored in the vicinity of stretches of biotin labeled DNA of >0.8 Mbp DNA with labeling properties expected for early S phase replicon clusters (B; n = 50). Double labeled fibers were scored in two classes (C,D): 1) Extending replicons - contained biotin-labeled replicons with internal forks labeled with BrdU during the 2nd pulse. 2) Clusters with secondary activation - contained multiple BrdU patches in the DNA fiber adjacent to the biotin-labeled cluster. In the same spread fields, fibers containing tracks labeled uniquely with BrdU (ie >250kbp from biotin-labeled tracks; D) were also recorded (C). The sizes of scale bars are shown on individual panels.
doi:10.1371/journal.pgen.1000900.g004

clustering is exemplified by the DNA fibers shown in Figure 4B. The first example (Panels 1 and 2) shows two adjacent imaging fields that contain a single fiber of >1.5 Mbp. This fiber has 3 replicons in the center and 2 on the right that were active during the 1st pulse (biotin in red). These replicons are linked genetically as replication in the DNA between them is completed during the 2nd pulse (BrdU in green). On the left of the same fiber, three patches are labeled during the 2nd pulse, showing that replicons in the adjacent DNA are activated during the 2nd labeling period. The short cluster shown in panel 3 contains 4 active replicons with an average size of 90 kbp. In this particular example, secondary origins are activated in replicons on both sides of the central cluster during the 2nd labeling period.

Using fibers like those shown (Figure 4B), two distinct classes of double-labeled fiber were scored, based on labeling within the proximal flanking DNA (Figure 4C and 4D). Replicon clusters with 'extending' forks were scored when replicons labeled during the 1st pulse were flanked by single DNA tracks labeled during the 2nd pulse, consistent with continued elongation of the out-growing forks from the flanking replicons of the primary cluster. Replicon clusters with 'secondary activation' were scored when DNA flanking the primary cluster also contained multiple tracks labeled during the 2nd pulse, which is only possible if additional forks are activated within the flanking DNA. The structure of replicons within the primary (biotin-labeled) clusters defines the frequency of these two populations (Figure S9). Notably, clusters with extending forks had widely dispersed origins (~200 kbp apart on average) whereas clusters with secondary initiations within the flanking DNA had shorter inter-origin distances (~125 kbp apart on average). This difference presumably reflects the temporal relationship between the completion and activation of synthesis in adjacent replicon clusters.

Preparation and staining of DNA fibres that contain >1 Mbp of DNA is technically challenging. However, the use of quality controls to monitor spreading and measurement of the labeled tracks (Figure S9) ensure reliability of the data generated. In all of the scored fibres, the separation of the biotin- and BrdU-labeled tracks was consistent with fork elongation rates within the normally accepted range for early S phase of 1–2 kbp/min (Figure S9). In these fibers, the continuity of the labeled tracks demonstrates that the underlying DNA strand must be intact throughout the labeled region.

To complete this analysis, we recorded single-labeled regions in order to define *de novo* initiation events that were remote from previously active replicons and thus 'uncoupled' (Figure 4C and 4D) from synthesis during the 1st labeling period. In the random fields used in this analysis, only 5% of labeled tracks were seen to be uniquely BrdU-labeled (Figure 4C). These observations suggest that genetically adjacent DNA foci are replicated during consecutive intervals of S phase. This genetic spread of synthesis appears to be a major mechanisms, as while the stochastic activation of potential origins is not precluded, remote initiation events, which are uncoupled from previously active replication foci, account for no more than 10% of initiation events once S phase has begun.

Individual DNA foci correlate with genome-wide replication timing domains

During our analysis of replication foci within individual cells we deliberately used a holistic approach in order to avoid any bias that might arise if specific genomic regions were targeted for analysis. To validate our conclusions, we next attempted to integrate the single cell data (Figure 1, Figure 2, Figure 3, Figure 4) with genome-wide data sets [10–15], which define the average pattern of synthesis across cell populations. To compare the structure of genome-wide timing domains with replication foci, we first defined the distribution profile of replication timing domains on selected regions of a specific human chromosome (Figure 5A) using genome-wide data sets taken from Desprat et al. [10]. Randomly selected regions of human chromosome 6 with ~10 Mbp of DNA (1 region is shown in Figure 5B) were sampled and points of inflection in the data readout used to define peaks in the timing profile. Individual peaks represent domains of discrete replication timing and peak heights (Figure 5B) define the average time of replication of the domain across the cell population analyzed – the highest peaks are replicated predominantly at the onset of S phase. When replication domains from different regions of chromosome 6 were combined the resulting distribution profile (Figure 5A) showed the average domain to contain 529.5+/−208.0 kbp of DNA.

For comparison with the timing data, we generated a series of distribution profiles that simulate the DNA content of populations of DNA foci in human cells. Profiles were generated using published data [19] that describes the distribution of replicon sizes and the number of replicons/cluster in human HeLa cells. In the two distribution profiles shown (Figure 5C) the first describes a typical profile for a population of 112 DNA foci – for direct comparison with the data set in Figure 5A – and the second shows the profile for a much larger sample. With average DNA contents of 527.9+/−312.2 kbp and 549.0+/−306.2 kbp of DNA, respectively, these simulations show that the DNA contents of replication timing domains and DNA foci have a high degree of similarity, with correlation coefficients in excess of 0.9 (Figure 5).

Figure 5 also shows the timing relationship between adjacent replication domains using genome-wide analysis of cell populations. The early replicating band p12.3 shows an example of how replication proceeds across a chromosomal domain, which in this typical example contains ~5 Mbp of DNA. At the left side of this region, 6 timing domains (seen as peaks on the timing profile) are clearly structured so that the central region (Figure 5B, region a) is replicated at the onset of S phase and the adjacent flanking regions (Figure 5B, regions b–d) are replicated sequentially as S phase proceeds. While the structure of peaks and valleys in the timing profile shows that individual cells in the population activate replication of the respective domains at slightly different times, the general trend is clearly consistent with the sequential activation of genetically adjacent timing domains across this region of chromosome 6 in human ES cells.

This comparison highlights a number of fundamental features of chromatin organization that define the efficacy of DNA replication. Most importantly, it is clear that the amount of DNA within both DNA foci and replication timing domains is dramatically different from the average size of individual replicons, which typically contain 100–150 kbp of DNA in human cells [19,20]. This implies that the replication timing domains must contain groups of replicons that are replicated together. In addition, if individual timing domains were single replicons it would only be possible to duplicate 1.5×10^9 bp or ~25% of their DNA in an S phase of 10 hours, given that synthesis during S phase of a diploid mammalian cells involves ~750 replication sites at any time [16–23]. Hence, the co-replication of replicons clusters within replication timing domains is necessary to complete synthesis on schedule.

While the evidence for replication timing domains that contain multiple replicons is overwhelming, it is notable that individual replicons are not evident at the resolution provided by genome-wide analysis (Figure 5B). This is likely to reflect the redundancy of potential origins, which in human cells are present in ~10-fold

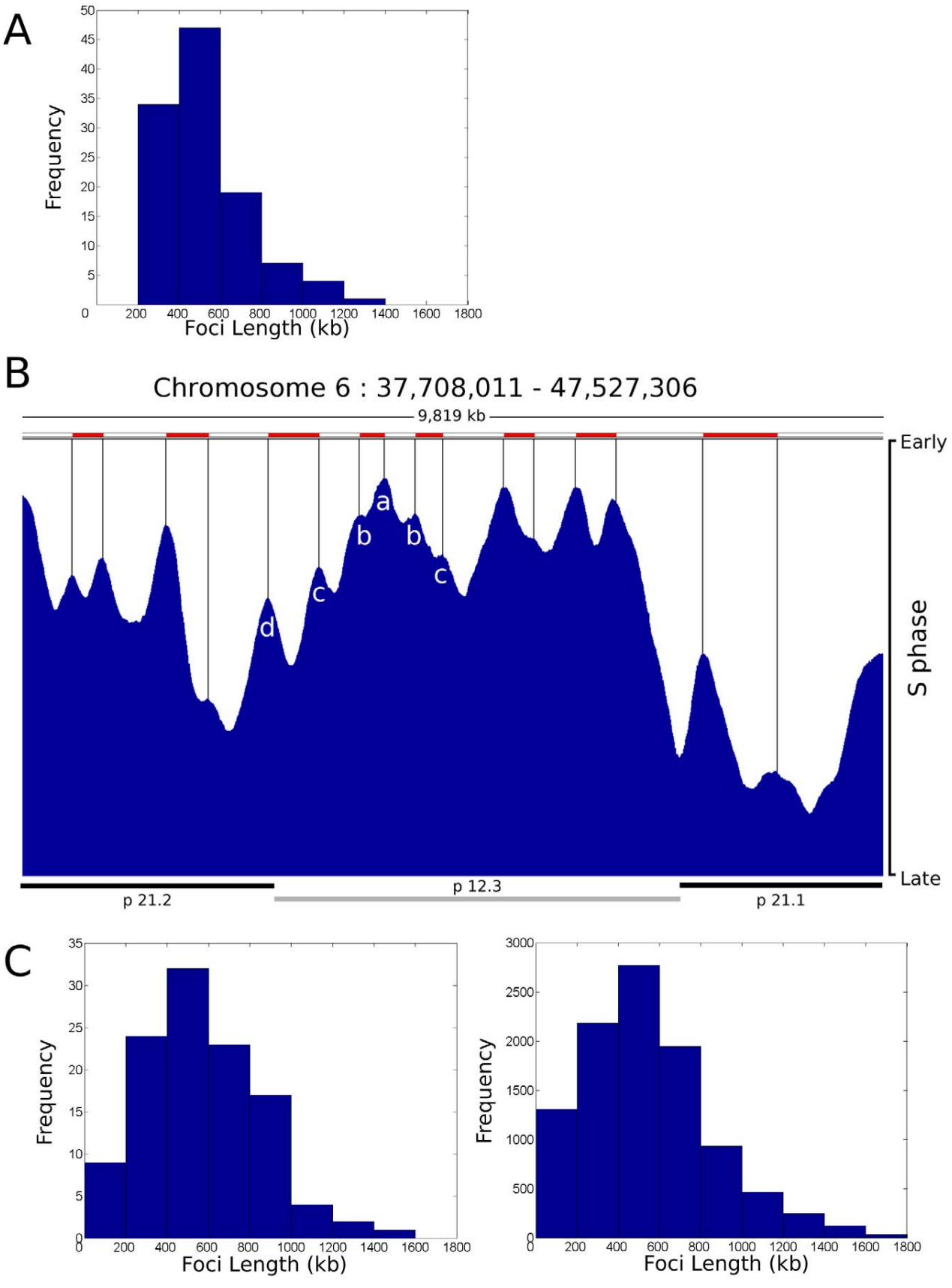

A
Frequency
Foci Length (kb)
B
Chromosome 6 : 37,708,011 - 47,527,306
9,819 kb
Early
S phase
Late
a
b
b
c
c
d
p 21.2
p 12.3
p 21.1
C
Frequency
Foci Length (kb)
Frequency
Foci Length (kb)

Figure 5. Replication timing domains correlate with DNA foci. A distribution profile for the length of replication timing domains was generated (A) using randomly selected regions of human chromosome 6 (n = 112, representing 59 Mbp (35%) of ch6), using data from [10]. Points of inflection in the timing profile were used to define replication timing domains – peaks corresponding to 6 such domains are identified in the center of the region shown (peaks a–d in B). The typical region shown (B) contains 1 central chromosomal R-band (light grey bar below) flanked by two G-bands. The G-band on the left is cytologically light staining and replicates during early S phase whereas as the one on the right is dark staining and replicates late in S phase. Domains in R- and G-bands were analysed separately, but as no significant difference was seen a composite genome-wide profile was generated. Distribution profiles for the length of DNA in individual DNA foci were also generated using data from [19]. Data derived from the profiles was as follows: (A) Mean length, 529.5+/−208.0 kbp, 90% data within 274.7–934.6 kbp; (C) right, simulation for 112 clusters – Mean length, 527.9+/−312.2 kbp, 90% data within 125.7–1,055.2 kbp; (C) left, simulation for 10,000 clusters – Mean length, 549.0+/−306.2 kbp, 90% data within 140.4–1,144.0 kbp. Correlation Coefficients for each pair of profiles were as follows: A:C_{112} = 0.9193; A:C_{10000} = 0.9100; C_{112}:C_{10000} = 0.9820.
doi:10.1371/journal.pgen.1000900.g005

excess relative to actual sites where DNA synthesis initiates [1–3,33]. Features of the local chromatin environment are thought to contribute to origin selection and define the relative efficiency with which different potential origins are used. Even so, origin activation clearly has a strong stochastic component so that different sites are used in different cells (Figure S10). As a result, the timing domains seen in population studies must generate a composite activation profile, which reflects how potential origins are used. The use of different potential origins in different cells will effectively smooth synthesis across chromatin domains so that the distribution of individual replicons is not seen. This means that replicon structure is defined by initiation events within individual cells and that the functional domains that are defined by DNA foci, and not the individual replicons themselves, are the regulatory targets for DNA synthesis.

The organization of DNA within chromosome territories defines the location of replication factories within the inter-chromatin compartment

The efficacy of a timing program that propagates using the genetic continuity of DNA foci will require that initiation sites that are used at the onset of S phase have an appropriate distribution throughout the genome. Notably, replication foci visualized in metaphase are uniformly spread along chromosomes (Figure 1). While it is not known how this is achieved, genome-wide approaches show that replication will often begin in regions of the genome that are rich in features linked to gene expression [10–15]. Interestingly, this conclusion was drawn from single cell studies 15 years ago [34], based on the co-localization of replication factories and active transcription sites at the onset of S phase.

Potential origins are thought to be equivalent when they are established well before the onset of S phase [1–3]. Hence, origin selection at the beginning of S phase must reflect the local chromatin environment within nuclear domains where replication factories are assembled. In this regard, it is notable that early replication factories are associated with nuclear domains that contain open chromatin whereas replication during mid/late S phase spreads to the chromatin-dense nuclear domains (Figure 6). This is confirmed by the structure of sites that contain nascent DNA (Figure 6A), which are located within the chromatin compartment at the interface between the chromatin and inter-chromatin nuclear domains [22,23]. During synthesis, the organization of active sites means that DNA foci, which contains the unreplicated template, and the associated factories and nascent product occupy discrete nuclear compartments (Figure 6C). This spatial separation means that during replication of a DNA focus that was labeled with BrdU in an earlier cell cycle the nascent product shows very little immediate co-localization with Br-DNA within the template containing focus. Subsequently, as the nascent chromatin matures, a period of 1–2 h is required before almost complete co-localization is seen (Figure 6D). This arrangement shows how the spatial architecture of the template-containing DNA foci and synthetic factories (Figure 6C) contribute to the dynamic behavior of chromatin during S phase.

Discussion

Eukaryotic cells have such complex genomes that DNA synthesis must be highly regulated in order to ensure the preservation of genome integrity and epigenetic modifications that define cell type. Surprisingly little is known, however, about the molecular principles by which this is achieved in higher eukaryotes. One key feature of the process, which has been appreciated for many years, is that replication of euchromatin and heterochromatin is structured temporally to occur preferentially during early and mid/late S phase, respectively [8]. This temporal restriction correlates with the differential activity of specific cyclin-CDK complexes [35] and the replication of different classes of chromatin, as defined by post-translational histone modification [36,37], during early and mid/late S phase.

While the spatial architecture of DNA foci appears to contribute to the structure of the mammalian S phase, the molecular mechanisms involved are not known. To address this question, we designed a single cell strategy to identify molecular links between chromosome organization and the timing of DNA synthesis (Figure 1). Analysis at the level of single cells is based on the structure of DNA foci, which are both functional units of DNA replication and structural units of chromosome organization [17,22,23]. The architecture of structural foci within chromosomal sub-domains has been analyzed in numerous recent studies. High-resolution analysis of the distribution of chromatin in domains of 2–10 Mbp has clearly demonstrated that foci typically contain 0.5–1 Mbp of DNA [24,25,38]. The most comprehensive study has shown that foci with ~1 Mbp of DNA are a common feature of genome organization [25] and that foci within transcriptionally active and inactive chromatin domains have distinct properties and nuclear distributions [25]. The spatial architecture of the 1 Mbp DNA domains has been analyzed in detail over length scales ranging from 0.5 to 75 Mbp [39]. Notably, the domains in nuclei are separated in relation to their genetic co-ordinates in the range 0 to 3–5 Mbp but little further separation is seen when sequences are further apart, because of the 3D folding of chromosomes within CTs [39].

S phase timing is defined by the connectivity of DNA foci

Here, we wanted to assess how higher-order chromatin organization contributes to the S phase timing program in mammalian cells. To do this, we evaluated the relative importance of direct (genetic) and indirect (spatial) chromatin interactions during S phase progression (Figure 1). DNA foci were labeled at different times of S phase and their spatial organization analyzed within individual CTs. Using a nearest neighbor analysis of DNA foci (Figure 1, Figure 2, Figure 3, and Video S1, S2, S3), together with an analysis of labeling continuity on stretched DNA fibers (Figure 4), we show that DNA foci that were labeled during

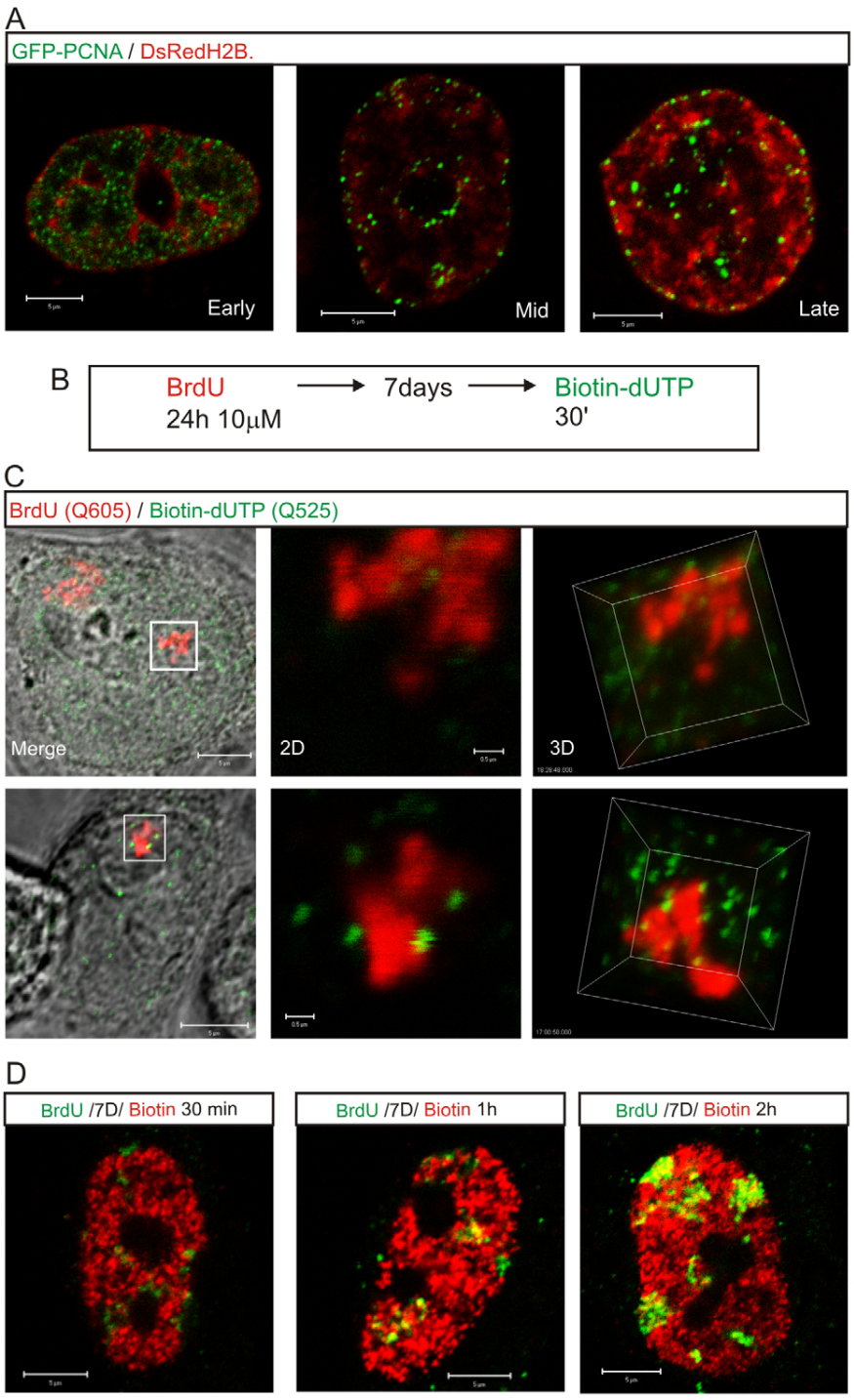


Figure 6. The proximity of DNA foci and the inter-chromatin domain defines the location of sites that are permissive for replication factory assembly. Active sites of DNA synthesis are shown by 3D imaging to be spatially separated from the substrate containing DNA foci (A,C). The distribution of replication factories was monitored using live cell imaging in cells transiently expressing GFP-PCNA (green) and histone H2B-DsRed (red) 24–48 h post-transfection (A). For a high-resolution view (C), entire CTs were labeled with BrdU (red), resolved by mitotic segregation and sites of nascent replication pulse-labeled with biotin-dUTP (green) as shown (B). Labeled sites were visualized using Q-dots and high-resolution images (60 slices with 100 nm Z steps) collected to assess the relative distribution of nascent sites and associated CTs during early (C, top) and mid/late (C, bottom) S phase. Highlighted regions (white boxes) are shown at high magnification in 2D and 3D, as indicated. Using the same labeling program (B) and different chase periods (for biotin labeling: biotin-dUTP is consumed in 10–15 min so longer incubations incorporate the initial labeling pulse followed by an unlabeled chase) co-localization of the BrdU (green) and biotin (red) labels was evaluated in confocal sections of fixed cells by indirect immuno-fluorescence (D) to monitor the location of newly replicated DNA. Following the 2nd pulse, the typical early S phase cell shown had only 11% of voxels in biotin-labeled foci that also contained BrdU. Following 1 and 2 h chase periods the level of co-localization increased to 31% and 59%, respectively, again in the typical early S phase cells shown. Scale bars: 5 and 0.5 μm in panels with individual nuclei and high-magnification, respectively.
doi:10.1371/journal.pgen.1000900.g006

consecutive intervals of S phase maintain a strict spatial co-association over many cell cycles. This demonstrates that foci labeled during consecutive intervals of S phase are genetic neighbors along chromosomes and provides strong evidence that this relationship underlies a 'next-in-line' mechanism of S phase progression [28,29]. Importantly, our experimental design is not directed to specific chromosomal loci or specific times of the cell cycle but instead uses an unbiased and holistic analysis of DNA foci, which are replicated during early S phase; as the labeled foci are not constrained by synthesis at the time of analysis their distribution must reflect a preferred organizational steady state within CTs.

As S phase proceeds, the majority of foci engage synthesis for 1–2 h (Figure S2) before the termination of synthesis by fusion of internal forks is coupled to activation of origins within adjacent DNA foci. The invasion of outgrowing forks into the genetically adjacent foci is one mechanism that in principle could cause structural alterations that allow or stimulate *de novo* origin activation. However, our analysis shows that this is not an inevitable outcome, as some forks grow without encountering conditions where *de novo* origin activation will occur; such regions might have a low density of potential origins [10–12]. Forks with these characteristics have been described using both DNA fibers [reviewed in 22] and in recent genome-wide studies [10–12], where extended forks of >250 kbp (representing ~5% of the genome) correlate with the 'temporal transition regions' that link replication during early and mid/late S phase. This transition from early to mid/late S phase correlates with a timing transition that can be revealed as a '3C-pause' in DNA synthesis under some conditions of replicative stress [40].

Genome-wide approaches to map replication timing

Single cell studies and genome-wide analysis of replication in cell populations provide complimentary strategies to explore DNA synthesis. Hence, it is important to understand the strengths and limitations of these strategies and evaluate how key information can be combined to develop a general model of S phase progression. A specific advantage of the genome-wide approach is that replication timing is anchored directly to DNA sequence and annotated features such as chromatin architecture and transcriptional activity. In doing this, genome-wide strategies also provide a composite view of DNA synthesis, which can be interpreted to define the average behavior of cells in the population. Such population approaches have shown that large regions of mammalian genomes are replicated during predictable intervals of S phase and that this generally correlates with features of the chromatin environment, so that highly expressed regions of the genome are replicated early during S phase [10–15]. The fact that syntenic regions of the human [9] and mouse [11] genomes replicate at equivalent times implies that this general principle is conserved.

During DNA synthesis, cells must also preserve the epigenetic information in chromatin that defines cell type specific patterns of gene expression. In exploring this aspect of mammalian S phase, genome-wide studies have shown that large genomic regions alter their replication timing when cells are induced to differentiate [10,12,15] and that distinct changes in replication timing arise as cells become epigenetically committed to differentiation [41]. Such changes raise obvious questions about mechanisms that link chromatin domains that are selected for synthesis during different periods of S phase and how these might relate to the next-in-line model of S phase progression [28,29]. As described above, such changes are presumably linked to changes in the local chromatin environment, which modulates the efficiency with which potential origins are established and used.

While the ability to relate replication timing to DNA sequence and chromatin features, such as histone modifications, is compelling [10–15], one limitation of studies based on cell populations is that any cell-to-cell variability is lost. This is inevitable as population-based approaches will smooth any biological complexity that we might expect to see as experimental noise. In contrast, analysis of DNA synthesis within individual nuclei and on isolated DNA fibers [5,20], is able to reveal detail related to the specific events that occur within individual cells. However, despite obvious experimental differences, our attempt to integrate data from genome-wide and single cell studies has shown that replicon clusters within domains that contain ~500 kbp of DNA provide the functional targets during replication of mammalian genomes (Figure 5). Moreover, evidence discussed above shows how data derived from single cells and cell populations support a general model for S phase progression that is in part based on the stochastic activation of potential replication origins and in part on the sequential activation of replication domains, based on their genetic continuity along chromosomes.

A model of S phase progression

The preferential accessibility of potential origins within open chromatin and the differential sensitivity of early and late origins to different cyclin/CDK complexes are major regulators of origin selection. These properties then dictate the efficiency with which different loci – such as potential replication origins (pre-RCs; Figure 7B) – interact with the inter-chromatin compartment where active replication factories are formed (RF; Figure 7B2). Origin selection is never-the-less stochastic, as most potential origins are replicated passively throughout S phase [6]. However, once S phase has begun, our data suggest that a next-in-line principle [28,29] defines the efficiency with which origins can be activated in the downstream replication program, so that only a minority (at most 10%) of *de novo* initiation events are uncoupled from synthesis within previously active replicon clusters (Figure 4). As replication within engaged replicon clusters approaches completion, the external forks might drive structural perturbations in neighboring foci that alter the exposure of potential origins to the replication machinery and so increases the probability of their activation (Figure 7B3). In this way, the genetic continuity of DNA foci along the chromosomal fiber provides a fundamental determinant of S phase progression in mammalian cells.

In the absence of genetically defined initiation sites, it is interesting to speculate how the mammalian cells have evolved to ensure that their genetic information is preserved during cell proliferation. Given the demand for precision, it is perhaps surprising that a key regulatory principle involves the random activation of potential initiation sites that are significantly more numerous than necessary to perform synthesis on schedule [33]. This stochastic feature of initiation and the redundancy of potential origins ensures that the system has sufficient tolerance to complete synthesis on schedule if the synthetic environment happens to change; any condition that result in slowing or stalling of the engaged forks are counteracted by local increases in origin density [reviewed in 42]. This regulatory mechanism operates at the level of DNA foci, and recent studies have suggested that a replication-dependent memory mechanism, based on the structure of DNA loops, ensures that appropriate levels of synthesis are maintained from one cell cycle to the next [43].

During S phase, the co-ordinated activation of groups of replicons within DNA foci will reduce the number of active synthetic sites that are required to complete synthesis. In addition,

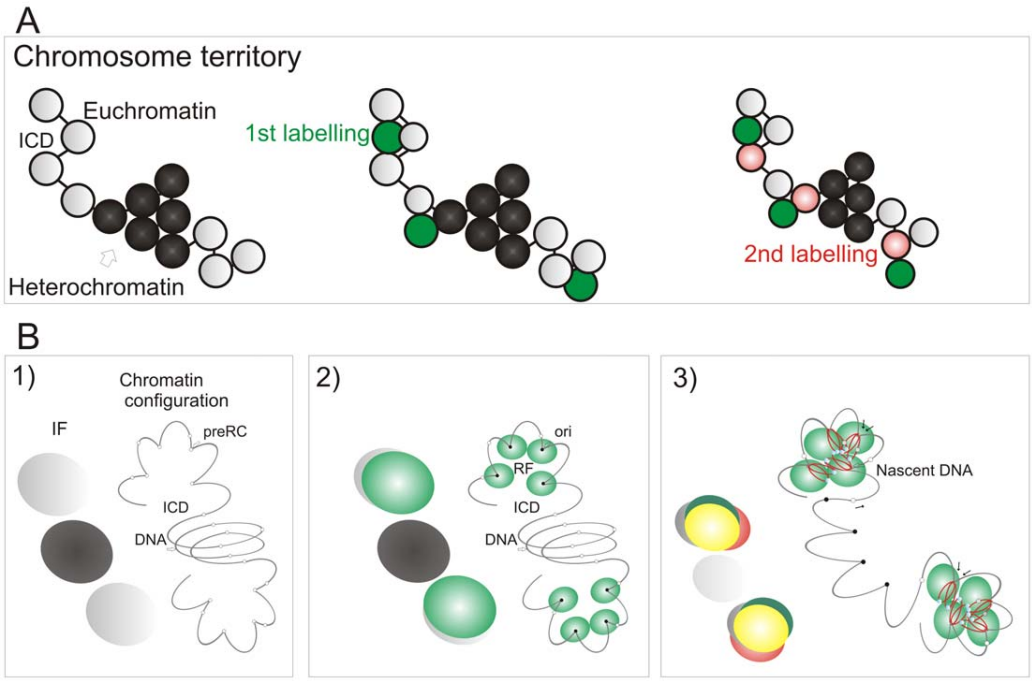


Figure 7. A model linking the organization of replicon clusters to S phase progression. A model (A,B) for S phase progression shows how the spatial and genetic continuity of DNA foci together with the organization of DNA foci and CTs relative to the interchromatin domain regulate the selection of active foci as S phase proceeds. CTs (A, one is shown) are composed of discrete DNA foci (coloured spheres), with structural characteristics that are defined by the epigenetic status of DNA to yield open and accessible euchromatic foci (grey) or more condensed and relatively inaccessible heterochromatic foci (black). The structure, accessibility—relative to the inter-chromatin domain (ICD)—and sequential labeling of adjacent foci provide 3 key determinants that define the course of S phase (B). Potential initiation sites (pre-RC complexes—small open circles) scattered throughout the chromatin fiber (line) interact by chance with the replication machinery (small green circles; B2) to initiate synthesis at a fraction of pre-RCs (now functional origins—small filled circles) within a local replication factory (RF—of clustered replisomes). As synthesis continues, chromatin fibers are reeled into the active synthetic factory and nascent strands displaced from the factory surface (B3). Eventually, the internal forks from adjacent replicons fuse and terminate. The outgrowing forks continue to grow and at some point structural changes in genetically linked chromatin (B3) increase the probability of activating origins within the adjacent foci. Three large spheres on the left of each panel in (B) depict the structures that would be visualized using fluorescent microscopy (IF): Grey—the structure of DNA foci that would be seen by prior labeling in vivo (for example with Cy3-dUTP); Green—location of active replication complexes and factories; Red—the nascent DNA; Yellow—overlap of red and green structures.
doi:10.1371/journal.pgen.1000900.g007

as replicon clusters engage synthesis together, within dedicated replication factories [18], this organization minimizes the time that adjacent replicons are replicating before their growing forks meet and fuse to terminate synthesis. Growing forks are complex structures that are inevitably more prone to damage and recombination than DNA packaged into normal chromatin, hence limiting the number of exposed forks will minimize the risk of damaging the genome. In addition, the sequential activation of replicon clusters based on their genetic continuity along chromosomes will also limit the number of isolated forks. Hence, we propose that the orderly synthesis of replicon clusters within DNA foci has evolved as a mechanism to ensure that higher eukaryotes can duplicate their genomes with the required efficiency while ensuring the preservation of both genetic and epigenetic information.

Materials and Methods

Labeling replication foci in situ

HeLa cells were grown in DMEM (Sigma) with 5% FBS and antibiotics. MRC5 cells were grown in MEM with 10% FBS and antibiotics. Replication foci were pulse-labeled in culture medium containing 10 µM bromo-deoxyuridine (BrdU) or labeled with modified replication precursor analogues: Cy3-; AlexaFluor488-(AF488-); biotin-; or digoxigenin-dUTP as described by Maya-Mendoza et al. [44]. Active replication factories were defined by transient expression of GFP-PCNA [29] or indirect immuno-fluorescence with a PCNA specific antibody (Immuno Concepts; Auto I.D. serum No 6006; 1/1000; 15 h; 4°C). Chromatin was visualized by transient expression of DsRed-histone-H2B. Unsynchronized cells were used throughout this study; this was a deliberate choice to avoid synchrony-dependent artefacts and preserve the natural structure of the S phase program.

DNA fiber experiments

DNA fiber spreads were prepared as previously described [19,44] using very low densities of labeled cells – of 10^3 cell/spread only 5–10 were labeled in these experiments. This low density minimizes DNA bundles and tangles within labeled fibers and allows visualization of Mbp fibers. In addition, the low density of labeled cells allows analysis of fibers from individual labeled cells. BrdU labeled tracks were detected with BrdU anti-sheep antibody (Biodesign; M20105S; 1:1000 dilution; 1 h at 20°C) and biotin-11-dUTP tracks using a mouse monoclonal antibody (Clone BN-34, Sigma; 1:1000 dilution; 1 h; 20°C). Primary antibodies were detected using Cy3- or AF488-conjugated donkey anti-sheep and anti-mouse secondary antibodies. The slides with DNA fibers were mounted with 50:50 PBS-glycerol.

Fibers were examined using a Zeiss LSM 510META confocal microscope using a 40× lens, labeled tracks measured using the LSM software and converted to kbp using a conversion factor of 1 μm = 2.59 kbp [19]; under these imaging conditions a single imaging field contains ~0.8 Mbp DNA. Double-labeled fibers were imaged only in dispersed, untangled areas of the DNA spread, to ensure the continuity of adjacent replicon clusters on individual DNA fibers. Routine quality control for spreading was performed using direct DNA labeling with YOYO-1 (Figure S9F) or cells labeled for >24 h with 10 μM BrdU, to give fully Br-labeled fibers (Figure S9G).

Immuno-fluorescence and direct labeling of DNA foci

DNA foci labeled with BrdU, biotin-dUTP or digoxigenin-dUTP were visualized by indirect immuno-fluorescence as described [19,44]. Cells were grown on coverslips, pulse labeled (directly or by transfection) and fixed in 4% paraformaldehyde. Fixed cells were acid treated (for BrdU labeling) and washed 3× in PBS, treated with 0.5 Triton ×100 in PBS, rinsed 3× in PBS, 3× PBS+ (PBS plus 1% BSA and 0.1% Tween 20), blocked for 1 h and incubated for 1 h with the appropriate antibody. Secondary antibodies were conjugated with Cy3, AF488, AF647 and Qdot reagents (Invitrogen). For 2nd or 3rd pulse detection, cells incubated after first detection including secondary antibody, were washed 3× in PBS and 3× in PBS+ and incubated with the appropriate first and second antibodies. In some experiments we used BrdU anti-rat (Immunologicals Direct Clone BU 1/75; 1:1000 dilution; 1 h; 20°C) and a secondary anti-rat antibody conjugated with Qdot-605. Streptavidin-Qdot-525 was used to identify sites containing biotin-dUTP. Finally, slides were washed 3× in PBS+, 3× in PBS, incubated with 5 μg/ml Hoechst 33258 (Sigma) for 10 min, rinsed 3× in PBS and mounted with either Vectashield or Prolong mounting media. Mitotic chromosomes were prepared as described [44].

For confocal imaging, samples were examined using a Zeiss LSM 510META confocal microscope and 100× (1.45 NA) lens. 3D images were generated using Z stacks and processed in Imaris software. In order to ensure optimal imaging performance, instrument alignment was performed at regular intervals by Zeiss. Chromatic shift was corrected using multi-coloured TetraSpeck florescent beads; the maximum tolerated shifts were 50 nm in X–Y and 100 nm in Z (Figure S3B). To minimize chromatic shift, for all experimental conditions extreme care was taken to balance labeling intensities in different imaging channels. In addition, for each indirect labeling experiment multiple samples were prepared so that each replication pulse could be labeled with the different secondary reagents used. 4D time-lapse imaging was performed using a Deltavision microscope with a CoolSNAP-HQ2 camera and Olympus objective (100x; 1.4 NA). The intensity of light during imaging was kept to 32% using an acquisition speed of 100–200 ms. Chromosome spreads were captured using a Deltavision microscope and images deconvolved using 5–10 iterations and pre-filter cut-off values (microns) of 0.05.

The 3D and 4D images were analyzed using Imaris software. For LSM images of individual CTs a 0.02 μm Gaussian filter was applied. For nearest neighbor analysis, 3D projections were generated in Imaris software from confocal Z series and software used to identify 3D labeled sites and the mass centers of individual sites (foci). Individual channels were processed separately. The co-ordinates of the mass centers were then used to define the spatial relationship between adjacent foci, either within or between channels. For presentation, the imaging software represents the mass centers of DNA foci as computer generated spheres that correspond in size to average foci. Images generated in doing this are clearly artificial and while providing an accurate representation of the positions of foci are not intended to provide a realistic representation of the foci themselves.

Bioinformatic analysis of replication timing domains

Replication timing data from human ES cells [10] was taken from the Integrative Genomics Viewer website at: http://www.broadinstitute.org/igv. For analysis, we choose to use human chromosome 6, as we have used this chromosome recently to model S phase [45]. To map the replication timing domains, ~10 Mbp regions were selected at random and points of inflection defined to identify peaks in the timing profile. Distances between adjacent peaks were then taken from the browser to develop a profile of distributions.

Profiles of distributions for replication foci were generated using parameters for the distribution of replicons per cluster and the length of replicons [19]. For simulation, the primary data for replicon length was approximated to a normal distribution ($\mu = 140.6238$kbp, $\sigma = 58.8192$), which was then sampled to determine the length of each individual replicon and assimilated into replicon clusters using the published frequencies of replicons/cluster. Simulations were implemented in Matlab.

Supporting Information

Figure S1 Three colour labeling to assess the spatial continuity of replication foci at different times of S phase. Nascent DNA synthesis in unsynchronized HeLa cells was labeled by indirect immuno-fluorescence after consecutive incorporation pulses using combinations of biotin-dUTP (blue), digoxigenin-dUTP (green) and BrdU (red). In some experiments the active factories were labeled using antibodies to PCNA (red). High-resolution 3D confocal images (1 μm sections are shown) of typical examples demonstrate how the 3 channel labeling can be utilized to define the structure of individual sites and the spatial continuity that links the separate pulses. Mid/late S phase patterns (A,C) provide discrete foci with clear structure and spatial connectivity. In early S phase, in contrast (B), while differentially labeled domains within individual foci can be identified with ease the complexity of the foci means that foci labeled during consecutive time zones of S phase will inevitable lie in close proximity. For (A–C), boxed areas in panel 1 are shown at high magnification in panels 2 and 3 and the intensity plots in panel 4 are scans along the line indicated in panel 2. The labeling protocol is shown on the left of the figure. Because cells were fixed immediately after incorporation, any labeling asymmetry presumably reflects the synthetic polarity that arises when DNA foci are replicated by a dedicated synthetic factory. Scale bars: 5 and 0.5 μm.
Found at: doi:10.1371/journal.pgen.1000900.s001 (9.44 MB TIF)

Figure S2 Spatio-temporal relationship of active replication factories and DNA foci. To establish the temporal separation between replication foci labeled during different replication time zones (A) HeLa cells were pulse labeled with biotin-dUTP (red), chased for 30, 60, and 120 min in medium and pulse labeled with BrdU (green). Separation of individual foci was seen following an intervening chase period of ~60 min in early S phase and ~120 min during mid and late S phase (A and insets at high magnification). (B) shows the percentage of imaging voxels in which the two precursors co-localized during early S phase following different chase intervals using 3D imaging (n = 25 nuclei/sample). (C) shows the size of replication foci during early, mid and late S phase (n = 200 for each pattern). Scale bars: 5 and 0.5 μm.
Found at: doi:10.1371/journal.pgen.1000900.s002 (9.22 MB TIF)

Figure S3 Chromatic shift influences the precision of co-localization during spatial analysis of DNA foci. HeLa cells were transfected at the same time using 488-dUTP and Cy3-dUTP, cultured for 7 days and chromatic shift evaluated (A). Confocal sections of individual imaging channels were recorded and mass centers (maximal intensities) of labeled foci defined by Imaris imaging software. Distances between the identified centers of labeled sites were then measured (78.37+/−53.48 nm shift, n = 68) to define the extent of chromatic shift. Chromatic shift due to instrument alignment was corrected using multi-coloured TetraSpeck florescent beads (B) — the maximum tolerated shifts were 50 nm in X–Y and 100 nm in Z; alignment was performed at regular intervals by Zeiss engineers. Scale bars: 1 and 2 μm in (A) and (B), respectively.
Found at: doi:10.1371/journal.pgen.1000900.s003 (6.78 MB TIF)

Figure S4 Structural analysis of DNA foci in individual CTs. Replication foci of unsynchronized HeLa cells were pulse-labeled to incorporate selected replication precursor analogues into nascent DNA. Cells were labeled with consecutive pulses of biotin-dUTP and BrdU both without (A) and with (B) an intervening 1h chase. Cells were then grown for 6–7 days to resolve the labeled CTs. After this time, cells with discrete labeled territories were analyzed using confocal microscopy. Pseudo-shapes were generated by image processing software to define the boundaries of labeled foci. In this example, shapes defined by the biotin labeling are transposed onto the other images to demonstrate the separation of labels in the different channels. In some experiments, CTs were also labeled with Qdot-conjugated secondary antibodies (C) to allow increased section density and Z resolution. (D) shows single channel (eg biotin to biotin or BrdU to BrdU) nearest neighbor analyzes for the labeled DNA foci within individual CTs. Scale bars: 5 and 0.5 μm.
Found at: doi:10.1371/journal.pgen.1000900.s004 (7.39 MB TIF)

Figure S5 Chromosome territories in human fibroblasts. CTs of MRC5 cells were analyzed after 6–7 days in culture. Cells were pulse labeled with biotin-dUTP (30 min; red) and subsequently with BrdU (20 min; green) following growth in fresh medium for 0, 1, or 2 h. Cells were fixed and sites of incorporation detected using indirect immuno-fluorescence and confocal microscopy; projections of confocal Z-stacks are shown. Using the pulse-chase (1 h)-pulse strategy, labeled early S phase foci of MRC5 cells were 513+/−116 nm (n = 200) in diameter and foci labeled during the 1st and 2nd pulses were 556+/−114 nm (n = 155) apart. Scale bars: 5 μm.
Found at: doi:10.1371/journal.pgen.1000900.s005 (9.12 MB TIF)

Figure S6 Different models of S phase progression. During S phase, the distribution of active sites that is defined by incorporation of labeled nucleotides into DNA foci allows identification of early, mid and late S phase cells. Multiple pulses with different timing separations can be used to monitor transitions between these different periods (A). However, DNA foci within the nuclear space are so highly crowded that defining the molecular principles that underlie the timing program is technically challenging. Three obvious models might account for the structure of the timing program. (A,1) - the genetic continuity between foci might provide an innate mechanism that allows foci to be replicated in a particular pattern once a specific set of foci is activated at the onset of S phase. (A,2) – a mechanism of spatial continuity might operate if once active factories are assembled the subsequent completion of synthesis allows factories to interact with the nearest unreplicated DNA foci. If factories disassemble when synthesis is complete, decay of active sites might provide a local high concentration of synthetic components that stimulates the assembly of new factories within the same nuclear domain. (A,3) – random activation of DNA foci within distinct chromatin compartments – eg euchromatin and heterochromatin – might explain the timing program if, for example, different CDK/cyclin complexes are required to activate origins within different chromatin compartments. (A) shows how these different models can be analyzed using the distribution of labeled foci within individual CTs during interphase and single chromosomes during metaphase. Random S phase progression can be modeled using statistical tools and MathLab software (B). Two examples are shown (B), which mimic the appearance of confocal sections. To simulate foci within diploid mammalian nuclei we generated random distributions of 350 spheres with 500 nm diameter – the foci - within a single large sphere of 10 μm diameter – the nucleus (Figure S6B). We assumed that S phase contained 10 time zones of 1 hour each so that 10% of foci were active at any particular time. With these assumptions, nuclei contain a total of 3500 foci that would occupy 44% of the total nuclear volume, as expected in proliferating diploid mammalian cells. Notably, the randomly generated patterns displayed similar structural features to foci seen during early S phase, yet when two randomly generated channels (single colour images) were overlaid (double colour images) the 1:1 co-association of nearest red and green neighbors that was seen experimentally in cells was never reproduced. The importance of spatial continuity is clearly evident in labeled cells, even immediately following labeling when the density of labeled foci is too high to allow detailed analysis in early S phase (C), though analysis in mid S phase (D) is possible. The same conclusion is reached if labeled cells are grown prior to analysis to resolve the labeled CTs by random chromosome segregation (E,F). Using precursors that can be imaged without processing, during both interphase (E) and metaphase (F), chromosomes labeled using a pulse-chase (2h)-pulse strategy always retain a high degree of co-association between sites labeled with the 1st and 2nd pulse labels. Using this live cell imaging approach, all CTs analyzed during interphase correspond with individual labeled chromosomes during metaphase. Scale bars: 5 and 0.5 μm in (C,D), and 10 μm in (E,F).
Found at: doi:10.1371/journal.pgen.1000900.s006 (8.74 MB TIF)

Figure S7 Three colour labeling to assess the genetic continuity of replication foci in chromosome territories. HeLa cells were labeled with sequential pulses of AF488-dUTP, Cy3-dUTP and BrdU each separated by unlabeled periods of 1 h (A). After 7 days, cells were fixed and BrdU detected using indirect immuno-labeling with rat anti-BrdU and anti-rat IgG conjugated with AF647 (B). Individual image channels were recorded for each precursor and the mass centers for individual foci defined by Imaris imaging software. Nearest neighbor analysis was then performed using all possible pair-wise combination (C): 1st–2nd pulses = 414.88+/−111.36 nm; 2nd–3rd = 376.96+/−109.64 nm; 1st–3rd = 487.17+/−137.66 nm; n = 150. Scale bars: 5 and 1 μm, as indicated on individual panels.
Found at: doi:10.1371/journal.pgen.1000900.s007 (9.67 MB TIF)

Figure S8 Extended pulse separations preclude nearest neighbor analysis. HeLa cells were pulse-labeled with AF488-dUTP, chased for 4 or 5 h and pulse-labeled with Cy3-dUTP. After 7 days, cells were fixed and images collected. As before, individual CTs contain distinct labeled sites of ~400 nm, which correspond to DNA foci that are labeled with the different precursors. Under these conditions, all sites are labeled uniquely with only one precursor. Moreover, patterns of foci labeled in the two channels are clearly unrelated, with foci labeled during the 1st and 2nd pulses populating distinct regions of individual CTs. CTs within 2

typical cells are shown. The magnified image (below) is a 2.5× view of the region highlighted (boxed area, above). Separate imaging channels and a channel merge are shown. Scale bars: 10 and 5 μm, as indicated on individual panels.
Found at: doi:10.1371/journal.pgen.1000900.s008 (6.65 MB TIF)

Figure S9 Structure analysis of DNA fibers defines genetic continuity during the S phase progression. HeLa cells were pulse-labeled (30 min) with biotin-dUTP grown for 1 h in medium and then pulse-labeled (20 min) with BrdU. DNA fibers from the labeled cells were spread on to glass slides and active replicons visualized by confocal microscopy after indirect immuno-labeling. Double labeled fibers of ~1–2 Mbp in length were recorded and analyzed. Typical examples of stalled replication forks (A) and long extending replicons (B) are shown. The analysis of the distance between replication forks (C; distances measurements using Zeiss software are superimposed on the images) correlates well with the labeling and chase times used, given rates of synthesis in the range 1–2 kb/min/fork. Using 5–10 cells/spread, almost all biotin-labeled fibers contain associated forks that are labeled with BrdU (see typical examples shown in C). A minority – 5% in each of 4 experiments (144 image fields like those shown) – of fibers in the double labeled regions of a spread were labeled only with BrdU (D shows typical image fields; n = 144). This suggests that de novo initiation events that occur as S phase proceeds are almost always coupled to existing active sites. The average separation of origins in clusters with extending forks and *de novo* (secondary) activation of adjacent clusters was 181.2+/−87.5 kbp and 119.6+/−47.0 kbp, respectively (E). DNA fiber integrity and distribution was assessed routinely by YOYO-1 staining—typical staining of a biotin-labeled sample is shown (F). DNA fiber integrity during BrdU labeling is also evident from the integrity of the labeled fibers—staining of biotin labeled forks on a fully labeled DNA fibre are shown (G). *In situ* labeling, using the same labeling program (H), shows how the complex patterns of incorporation into replication foci (foci 1–3) can be attributed to the distribution of replication structures on nascent DNA fibers (replicons shown in cartoon form below). Scale bars: 50 μm in (D), 5 and 0.5 μm in (F).
Found at: doi:10.1371/journal.pgen.1000900.s009 (9.18 MB TIF)

Figure S10 Using genome-wide and single cell approaches to analyze replication timing. (A–C) show the structure of 3 well-characterised examples of initiation sites for mammalian DNA synthesis. At some sites, local gene structure determines that replication might initiate at a specific site (A)—the human lamin B2 locus represents a paradigm for this class of origin. Some replicons have dispersed potential sites of initiation, which contain preferred initiation sites within them (B)—the mammalian DHFR locus is a good example of this class of initiation domain. Finally, some loci contain regions (C) with hotspots of replication initiation that contain many possible sites within clusters of potential origins that cover about 10 kbp. The example shown contains 4 potential initiation zones, which may be treated as individual replicons (C1–4), but in the cells can be activated unpredictably—selection is stochastic—so that different cells initiate synthesis from different sites across the locus [see 20 for details]. The cartoon in (D) depicts an imaginary DNA locus of ~1 Mbp, which contains each of these three classes of initiation domain. In the cell, this locus would fold to occupy a single DNA focus. Analysis of replication across the locus using DNA fibres isolated from individual cells would reveal a range of patterns, such as the two depicted in (D1–2). However, a genome-wide analysis designed to define replication timing across the locus (D3) would give a more complex picture that incorporates all possible initiation events across the cell population used.
Found at: doi:10.1371/journal.pgen.1000900.s010 (0.42 MB TIF)

Video S1 Time-lapse analysis of DNA foci dynamics—consecutive pulse labels. The time-lapse series from the experiment in Figure 2B shows how individual foci labeled with consecutive pulses are dynamic within CTs so that adjacent sites labeled with the 1st and 2nd precursor always maintain complete co-association. Using a live cell imaging protocol that maintains cell viability for at least 24 h, images shown were taken at 15 min intervals for 3 h. Video S1 shows the mobility of foci directly (1 frame/second), without further processing.
Found at: doi:10.1371/journal.pgen.1000900.s011 (0.69 MB MOV)

Video S2 Time-lapse analysis of DNA foci dynamics—consecutive pulse labels. A representation of Video S1 in which image processing software was used to replace each labeled site in the green (1st) and red (2nd) channels with a sphere of 500 nm; the spheres and original labeled sites have coincident centers of mass. Individual images in the video are presented at a rate of 1 frame/second.
Found at: doi:10.1371/journal.pgen.1000900.s012 (0.34 MB MOV)

Video S3 Time-lapse analysis of foci dynamics—consecutive pulse labels with an intervening 1 h unlabeled period. The time-lapse series from the experiment in Figure 2C was prepared as described in the legend to Video S1. Even with 1 h and 2 h (not shown) unlabeled periods between the two pulses, foci containing the 1st and 2nd precursors maintain complete spatial co-association over an imaging time course of 3 h. As CT shape changes significantly over the imaging time course, the persistent co-association of neighboring foci is clearly consistent with them being genetically linked along chromosomes.
Found at: doi:10.1371/journal.pgen.1000900.s013 (0.18 MB MOV)

Author Contributions

Conceived and designed the experiments: AMM DAJ. Performed the experiments: AMM. Analyzed the data: AMM POC AS. Wrote the paper: AMM DAJ. Conceived and designed key experimental strategies: DAJ AMM. Performed cell biology and imaging: AMM. Performed bioinformatic analysis and in silico experiments: POC AS.

References

1. Blow JJ, Dutta A (2005) Preventing re-replication of chromosomal DNA. Nat Rev Mol Cell Biol 6: 476–486.
2. DePamphilis ML, Blow JJ, Ghosh S, Saha T, Noguchi K, et al. (2006) Regulating the licensing of DNA replication origins in metazoa. Curr Opin Cell Biol 18: 231–239.
3. Sclafani RA, Holzen TM (2007) Cell cycle regulation of DNA replication. Annu Rev Genet 41: 237–280.
4. Machida YJ, Hamlin JL, Dutta A (2005) Right place, right time, and only once: replication initiation in metazoans. Cell 123: 13–24.
5. Norio P, Kosiyatrakul S, Yang Q, Guan Z, Brown NM, et al. (2005) Progressive activation of DNA replication initiation in large domains of the immunoglobulin heavy chain locus during B cell development. Mol Cell 20: 575–87.
6. Mesner LD, Crawford EL, Hamlin JL (2006) Isolating apparently pure libraries of replication origins from complex genomes. Mol Cell 21: 719–726.
7. Groth A, Rocha W, Verreault A, Almouzni G (2007) Chromatin challenges during DNA replication and repair. Cell 128: 721–733.

8. Drouin R, Lemieux N, Richer CL (1990) Analysis of DNA replication during S-phase by means of dynamic chromosome banding at high resolution. Chromosoma 99: 273–280.
9. Woodfine K, Fiegler H, Beare DM, Collins JE, McCann OT, et al. (2004) Replication timing of the human genome. Hum Mol Genet 13: 191–202.
10. Desprat R, Thierry-Mieg D, Lailler N, Lajugie J, Schildkraut C, et al. (2009) Predictable dynamic program of timing of DNA replication in human cells. Genome Res 19: 2288–2299.
11. Farkash-Amar S, Lipson D, Polten A, Goren A, Helmstetter C, et al. (2008) Global organization of replication time zones of the mouse genome. Genome Res 18: 1562–1570.
12. Hiratani I, Ryba T, Itoh M, Yokochi T, Schwaiger M, et al. (2008) Global reorganization of replication domains during embryonic stem cell differentiation. PLoS Biol 6: e245. doi:10.1371/journal.pbio.0060245.
13. Cadoret J-C, Meisch F, Hassan-Zadeh V, Luyten I, Guillet C, et al. (2008) Genome-wide studies highlight indirect links between human replication origins and gene regulation. Proc Natl Acad Sci USA 105: 15837–15842.
14. Sequeira-Mendes J, Diaz-Uriarte R, Apedaile A, Huntley D, Brockdorff N, et al. (2009) Transcription initiation activity sets replication origin efficiency in mammalian cells. PLoS Genet 5: e1000446. doi:10.1371/journal.pgen.1000446.
15. Hansen RS, Thomas S, Sandstrom R, Canfield TK, Thurman RE, et al. (2010) Sequencing newly replicated DNA reveals widespread plasticity in human replication timing. Proc Natl Acad Sci USA 107: 139–144.
16. Jackson DA (1995) Nuclear organization: uniting replication foci, chromatin domains and chromosome structure. BioEssays 17: 587–591.
17. Zink D (2006) The temporal program of DNA replication: new insights into old questions. Chromosoma 115: 273–287.
18. Hozak P, Hassan AB, Jackson DA, Cook PR (1993) Visualization of replication factories attached to a nucleoskeleton. Cell 73: 361–373.
19. Jackson DA, Pombo A (1998) Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J Cell Biol 140: 1285–1295.
20. Lebofsky R, Heilig R, Sonnleitner M, Weissenbach J, Bensimon A (2006) DNA replication origin interference increases the spacing between initiation events in human cells. Mol Biol Cell 17: 5337–5345.
21. Ma H, Samarabandu J, Devdhar RS, Acharya R, Cheng PC, et al. (1998) Spatial and temporal dynamics of DNA replication sites in mammalian cells. J Cell Biol 143: 1415–1425.
22. Berezney R, Dubey DD, Huberman JA (2000) Heterogeneity of eukaryotic replicons, replicon clusters, and replication foci. Chromosoma 108: 471–484.
23. Cremer T, Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet 2: 292–301.
24. Shopland LS, Lynch CR, Peterson KA, Thornton K, Kepper N, et al. (2006) Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. J Cell Biol 174: 27–38.
25. Goetze S, Mateos-Langerak J, Gierman HJ, de Leeuw W, Giromus O, et al. (2007) The three-dimensional structure of human interphase chromosomes in related to the transcriptome map. Mol Cell Biol 27: 4475–4487.
26. Goen A, Cedar H (2003) Replicating by the clock. Nat Rev Mol Cell Biol 4: 25–32.
27. Aladjem MI (2007) Replication in context: dynamic regulation of DNA replication patterns in metazoans. Nat Rev Genet 8: 588–600.
28. Manders EMM, Stap J, Brakenhoff GJ, van Driel R, Aten JA (1992) Dynamics of three-dimensional replication patterns during the S-phase, analyzed by double labeling of DNA and confocal microscopy. J Cell Sci 103: 857–862.
29. Sporbert A, Gahl A, Ankerhold R, Leonhardt H, Cardoso MC (2002) DNA polymerase clamp shows little turnover at established replication sites but sequential de novo assembly at adjacent origin clusters. Mol Cell 10: 1355–1365.
30. Takebayashi S, Sugimura K, Saito T, Sato C, Fukushima Y, et al. (2004) Regulation of replication at the R/G chromosomal band boundary and pericentromeric heterochromatin of mammalian cells. Exp Cell Res 304: 162–174.
31. Bornfleth H, Edelmann P, Zink D, Cremer T, Cremer C (1999) Quantitative motion analysis of subchromosomal foci in living cells using four-dimensional microscopy. Biophys J 77: 2871–2886.
32. Blow JJ, Gillespie PJ, Francis D, Jackson DA (2001) Replication origins in Xenopus egg extracts are 5–15 kb apart and are activated in clusters that fire at different times. J Cell Biol 152: 15–26.
33. Hyrien O, Marheineke K, Goldar A (2003) Paradoxes of eukaryotic DNA replication: MCM proteins and the random completion problem. BioEssays 25: 116–125.
34. Hassan AB, Errington RJ, White NS, Jackson DA, Cook PR (1994) Replication and transcription sites are colocalized in human cells. J Cell Sci 107: 425–434.
35. Katsuno Y, Suzuki A, Sugimura K, Okumura K, Zineldeen DH, et al. (2009) Cyclin A-CDK1 regulates the origin firing program in mammalian cells. Proc Natl Acad Sci USA 106: 3184–3189.
36. Zhang JM, Xu F, Hashimshony T, Keshet N, Cedar H (2002) Establishment of transcriptional competence in early and late S phase. Nature 420: 198–202.
37. Lande-Diner L, Zhang JM, Cedar H (2009) Shifts in replication timing actively affect histone acetylation during nucleosome reassembly. Mol Cell 34: 767–774.
38. Albiez H, Cremer M, Tiberi C, Vecchio L, Schermelleh L, et al. (2006) Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. Chromosome Res 14: 707–733.
39. Mateos-Langerak J, Bohn M, de Leeuw W, Giromus O, Manders EMM, et al. (2009) Spatially confined folding of chromatin in the interphase nucleus. Proc Natl Acad Sci USA 106: 3812–3817.
40. Fetni R, Drouin R, Richer CL, Lemieux N (1996) Complementary replication R- and G-band patterns induced by cell blocking at the R-band/G-band transition, a possible regulatory checkpoint within the S phase of the cell cycle. Cytogenet Cell Genet 75: 172–179.
41. Hiratani I, Ryba T, Itoh M, Rathjen J, Kulik M, et al. (2010) Genome-wide dynamics of replication timing revealed by in vitro models of mouse embryogenesis. Genome Res 20: 155–169.
42. Maya-Mendoza A, Tang CW, Pombo A, Jackson DA (2009) Mechanisms regulating S phase progression in mammalian cells. Front Biosci 14: 4199–4213.
43. Courbet S, Gay S, Arnoult N, Wronka G, Anglana M, et al. (2008) Replication fork movement sets chromatin loop size and origin choice in mammalian cells. Nature 455: 557–560.
44. Maya-Mendoza A, Petermann E, Gillespie DA, Caldecott KW, Jackson DA (2007) Chk1 regulates the density of active replication origins during the vertebrate S phase. EMBO J 26: 2719–2731.
45. Shaw A, Olivares-Chauvet P, Maya-Mendoza A, Jackson DA (2010) S phase progression in mammalian cells: modelling the influence of nuclear organization. Chromosome Res 18: 163–178.

A.3 Innate structure of DNA foci restricts the mixing of DNA from different chromosome territories.

Olivares-Chauvet P, Fennessy D, Jackson DA, Maya-Mendoza A. 2011. Innate Structure of DNA Foci Restricts the Mixing of DNA from Different Chromosome Territories. Plos One 6(12): 1-13

Summary

Publication relevant to the work shown in Chapter 3. In order to clarify a relatively open question in the field of chromosome territories (CT) we performed confocal imaging of differentially labeled individual CTs and colocalisation analysis. We found colocalisation signal at only at very low levels. We repeated this analysis with TSA cells and found that disruption of native chromosome architecture increased colocalisation signal in only small levels.

Contribution

General discussion and writing of the manuscript in addition to DAJ and AM-M. Experiments regarding CT colocalisation, confocal microscopy, data analysis, image processing and bioinformatic analysis.

OPEN ACCESS Freely available online

PLoS one

Innate Structure of DNA Foci Restricts the Mixing of DNA from Different Chromosome Territories

Pedro Olivares-Chauvet[•], **Dorota Fennessy, Dean A. Jackson***, **Apolinar Maya-Mendoza***[•]

Faculty of Life Sciences, University of Manchester, Manchester, United Kingdom

Abstract

The distribution of chromatin within the mammalian nucleus is constrained by its organization into chromosome territories (CTs). However, recent studies have suggested that promiscuous intra- and inter-chromosomal interactions play fundamental roles in regulating chromatin function and so might define the spatial integrity of CTs. In order to test the extent of DNA mixing between CTs, DNA foci of individual CTs were labeled in living cells following incorporation of Alexa-488 and Cy-3 conjugated replication precursor analogues during consecutive cell cycles. Uniquely labeled chromatin domains, resolved following random mitotic segregation, were visualized as discrete structures with defined borders. At the level of resolution analysed, evidence for mixing of chromatin from adjacent domains was only apparent within the surface volumes where neighboring CTs touched. However, while less than 1% of the nuclear volume represented domains of inter-chromosomal mixing, the dynamic plasticity of DNA foci within individual CTs allows continual transformation of CT structure so that different domains of chromatin mixing evolve over time. Notably, chromatin mixing at the boundaries of adjacent CTs had little impact on the innate structural properties of DNA foci. However, when TSA was used to alter the extent of histone acetylation changes in chromatin correlated with increased chromatin mixing. We propose that DNA foci maintain a structural integrity that restricts widespread mixing of DNA and discuss how the potential to dynamically remodel genome organization might alter during cell differentiation.

Citation: Olivares-Chauvet P, Fennessy D, Jackson DA, Maya-Mendoza A (2011) Innate Structure of DNA Foci Restricts the Mixing of DNA from Different Chromosome Territories. PLoS ONE 6(12): e27527. doi:10.1371/journal.pone.0027527

Editor: Joanna Mary Bridger, Brunel University, United Kingdom

Received December 22, 2010; **Accepted** October 19, 2011; **Published** December 21, 2011

Funding: Funding was from: BBSRC (grant BBS/B/06091), The Wellcome Trust (grant 080172/Z/06) and EU (Project LSBH-CT511965). The funding source for PO-C is credited to: CONACyT (National Council for Science and Technology, Mexico). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: dean.jackson@manchester.ac.uk (DAJ); Apolinar.Maya-Mendoza@manchester.ac.uk (AMM)

• These authors contributed equally to this work.

Introduction

Within the nucleus of higher eukaryotic cells [1–3] individual chromosomes are folded to occupy spatially discrete chromosome territories (CTs) (reviewed in [4–6]). DNA foci, which typically contain 250–1,000 kbp of DNA, provide the fundamental subunits of higher order chromatin folding within CTs. Though the molecular mechanisms that define the structure of foci are unclear, it has been known for many years that discrete foci are stable entities over many cell generations and that they contain multiple units of DNA synthesis, which are replicated together at specific times of S phase [7,8]. This temporal regulation of replication, within defined cohorts of DNA foci, emphasises the importance of links between chromosome structure and function, while preserving epigenetic information during cell proliferation [9,10].

As stable structures of higher-order chromatin folding, DNA foci might be expected to suppress DNA mixing [11,12]. In fact, the dynamic mobility of chromatin within mammalian CTs is generally constrained at less that 1 μm and once nuclei are formed, following mitosis, the relative spatial distribution of CTs is largely preserved [4,5]. The structure of individual CTs is however plastic [13,14], so that chromatin within individual territories might assume a variety of alternative configurations [15]. Extreme examples of alternative patterns of chromatin folding are most evident in gene-rich chromosomal domains - such as the human MHC locus - which are able to form extended chromatin loops that spread away from the linked CT when gene expression is induced [16]. However, dynamic analysis of defined endogenous loci has not been possible and, as a result, large artificially-tagged ectopic repeats have been used to analyze chromatin mobility in mammalian cells [17].

Over the past few years an alternative view of chromosome structure has emerged, which challenges the idea that CTs are self-contained and proposes that significant mixing of DNA can occur [2,18]. Clear evidence for long-range chromatin looping evolved from the analysis of intra-chromosomal interactions during gene expression, using chromosome conformation capture (3C) technologies. More surprisingly, while evaluating the extent of the regulatory interaction it became clear that genes from different CTs were also able to co-associate at common sites of gene expression [19,20]. However, validation of specific inter-chromosomal interactions within individual cells typically demonstrated that only ~10% of the loci in question were co-associated when transcribed [19,21,22]. Nevertheless, recent innovations in analysis of genome-wide interaction networks or functional 'interactomes', have placed unprecedented emphasis on understanding how chromatin dynamics facilitate the formation of gene interactions networks, which in turn might contribute to the regulation of gene expression in mammalian cells [18,23].

If long-range chromosomal interactions make a significant contribution to the regulation of gene expression in higher eukaryotes, it is important to understand the range and extent of interactions that this involves. To address this issue, we have used single cell imaging techniques to monitor chromatin mixing in human HeLa cells. DNA foci were pulse-labeled using fluorescent dNTP analogues that incorporate during replication and remain stably associated with labeled CTs for at least 14 days. After labeling, mitotic segregation reveals discrete chromatin domains with clearly defined DNA foci, so that the dynamic properties of foci and interactions between foci of neighboring CTs can be assessed. We show that while individual foci are spatially dynamic their DNA is generally locally constrained and so limits mixing between neighboring CTs.

Results

Chromosome territories are discrete structures

We tested the extent of DNA mixing between CTs using established protocols that allow CTs and individual DNA foci to be visualized in living cells [24,25]. Cells were pulse-labeled with AF488-dUTP, grown for 24 h and then pulse-labeled with Cy3-dUTP and grown for a further 1–2 days (Fig. 1). Because replication is semi-conservative and mitotic chromosome segregation is random, this protocol yields cells that contain uniquely red or green labeled CTs together with a minority of CTs that are unlabeled (Fig. 1A).

Live cell analysis (Fig. 1B and video S1) showed that the identity of CTs is preserved for many hours with little or no interaction between neighboring CTs. However, as resolution is limited by the low levels of illumination used during live cell imaging we also performed imaging on fixed cells (Fig. 1C). Post-fixation analysis, in the absence of processing that might perturb chromosome structure at the resolution analyzed by light microscopy, allows the structure of the differentially labeled chromatin domains and distribution of their foci to be visualized (Fig. 1Ci; projections of complete Z stacks are shown). With this type of analysis, the structure of DNA foci is clearly preserved and foci are clustered into local domains that represent individual or small groups of CTs. Notably, the boundaries between adjacent green and red domains are clearly defined (see isolated channels in Fig. S1) and regions of apparent co-localization between the differentially labeled regions (yellow regions; high magnification views in Fig. 1Civ–v) were restricted to these boundary domains. However, rotation of the 3D image suggested that many sites of apparent localization resulted from the spatial overlap of adjacent foci in projections of optical sections and not true co-localization within individual voxels of the 3D image (Fig. 1D and videos S2 and S3). To address this issue, we next attempted to place numerical limits on the low-level co-localization seen by measuring both the nuclear volume occupied by the co-localized regions and the amount of labeled DNA within these domains of chromatin mixing.

Quantitative measurement of inter-chromosomal mixing

A number of strategies have been described for monitoring levels of co-localization in confocal images (reviewed in refs [27,28]). Pearson's and Mander's coefficients provide a qualitative insight into degrees of co-localization on double stained confocal images. The Mander's coefficient is scaled from 0 to 1, where a value of 1 represents complete co-localization and a value of 0 no overlap between the imaging channels. For the Pearson's coefficient (PC) the scale is 1 to −1. On this scale, 1 represents complete co-localization and negative numbers represent exclusion, with −1 representing samples with no overlap at all.

While Pearson's and Mander's coefficients are used routinely for analysis of signal co-localization it is important to recognise that these values are heavily influenced by the way in which noise and background labeling in the sample is treated. Quantitative image processing is notoriously challenging, principally because of uncertainties in setting threshold values that reliably define true signal from various sources of noise [26]. However, the labeling protocol applied here avoids traditional sources of background staining (such as those that arise during immuno-labeling), as the fluorescent replication precursor analogues are incorporated directly into DNA. By the time imaging is performed, essentially all fluorescent molecules added to cells are covalently bound to DNA and make no contribution to background.

Another challenge of image analysis is that raw images often contain electronic noise, which is characterized by isolated voxels with high signal intensities. Imaging software provides different strategies to remove such noise. Gaussian filtering averages the signal in neighboring voxels and so smoothes noise and allows extraneous signal to be subsequently removed by thresholding. For quantitative analysis, however, one limitation of this approach is the incorporation of unreal voxel intensity values into the data set, which tend to degrade the signal and compromise the integrity of structures by spreading of their edges. Median filtering provides an alternative strategy in which each voxel in the image is assigned the median value of all of the immediately adjacent voxels. Hence, isolated high-intensity voxels will be eliminated, while voxels storing real signal are essential unaltered by the filtering step.

After labeling (Fig. 1A), maximum projections of complete Z series were collected (Fig. 1Ci; taken using a 100× lens). Individual nuclei with classical patterns seen during early S phase were then selected for detailed analysis based on the balance of labeling in the red and green channels (Fig. 1Cii shows an electronic zoom of the selected cell from Fig. 1Ci). Prior to image analysis, low level electronic noise in the zoomed image was extracted using a median filter (3×3×3 voxels). We then applied an empirical approach for manual background adjustment and found that within our samples the best estimate of background was represented by a signal corresponding to the standard deviation of the average signal intensity across all labeled voxels in the image (Fig. S1). Ten nuclei like the typical sample shown (Fig. 1Cii) were then analyzed to monitor levels of channel co-localization (Table 1). Across this sample, all approaches yielded a negative average Pearson's coefficient and low Mander's coefficient, consistent with very low levels of co-localization between the two imaging channels.

We also performed co-localization analysis after selecting regions of interest to exclude the contribution of black voxels that lie outside the nucleus (Fig. 1Civ–vi). Importantly, for this analysis we selected nuclear regions with the highest levels of apparent co-localization with adjacent red and green chromatin domains (Fig. 1Civ). Within these regions, the average Pearson's coefficient within such cropped regions (n = 10; vol = 28.5 μm^3) was −0.07 ±0.04 and the Mander's coefficient was 0.05 ±0.04 and 0.08 ±0.06 in the green and red channels, respectively. Within the selected regions of interest, the average co-localized volume covered 0.28±0.24 μm^3 occupying 0.96%±0.85 of the cropped box. The green signal occupied on average 17.3%±7 of the imaging voxels and the red signal 11.4%±3, so the labeled space represents ~30% of the volume in these selected regions. By analyzing the most highly intermingled chromatin domains within individual nuclei, this analysis provides an upper limit for the

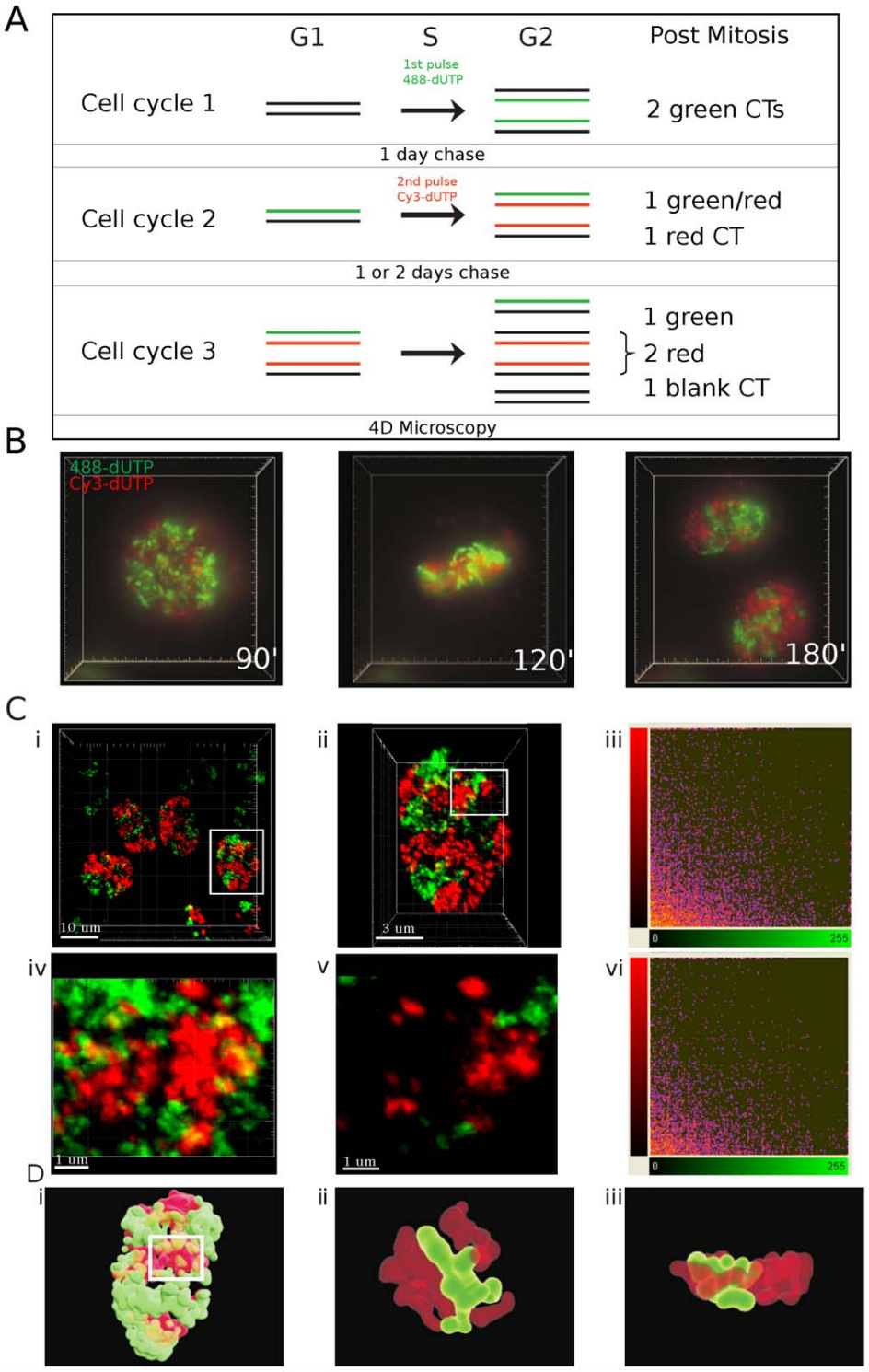


Figure 1. DNA foci are discrete higher-order chromatin structures. DNA foci within HeLa cells were labelled in two consecutive cell cycles - using AF488-dUTP in the first cycle and Cy3-dUTP in the second – and grown for a further 1–2 days to resolve the labeled DNA foci into uniquely labeled nuclear domains (A). In this context, the domains represent clusters of CTs as individual CTs cannot be resolved with confidence. Labeled cells were analyzed using 3-D time-lapse microscopy (DeltaVision) for up to 24 h (B; time-lapse frames are taken from Supplementary video S1), to confirm that chromosome in mitosis are labeled with either the red or green fluorescent precursor. Cells like those shown (B) were also fixed and imaged (Zeiss LSM510META) without further processing (C). For image analysis, cells with similar intensities in the two imaging channels were manually selected (Ci - white box), confocal Z stacks collected and 3-D projections generated (Cii). Nuclei like that shown in Cii were used for co-localization analysis, using all voxels within the image (Ciii). For this example, co-localization analysis using Imaris software (Ciii), gave a Pearson's coefficient (−0.0194) consistent with very weak co-localization. The region highlighted in Cii (white box) contained the majority of voxels containing signal from

the red and green channels and was selected for further analysis. A high magnification view of the cropped region shows the local structure of chromatin domains, both in the entire Z series (Civ) or individual sections (Cv), with discrete patches of red and green labeling and little mixing (yellow) of signal from the two channels. Co-localization analysis using Imaris software without background adjustment (Cvi) showed this typical sample to have a Pearson's coefficient of −0.0437. A surface rendered video simulation of chromatin domains in the cell from Cii was also generated to show the distribution of interaction interfaces within the sample (D; see Supplementary video S2 to section though the 3-D reconstruction). High-power views of the 3-D region highlighted (white box in Di) are shown (Dii–iii). This modeling demonstrates the complex structure of interaction surfaces, with many surface protrusions from one chromatin domain interdigitating with folds or channels within the neighboring domain. Two snapshots from the 3-D reconstruction (Supplementary video S3) are shown. Scale bars of 10, 3 or 1 μm are shown on individual panels.
doi:10.1371/journal.pone.0027527.g001

proportion of the total nuclear volume in which inter-chromosomal mixing of the labeled chromatin is seen.

Co-localization analysis is most reliable when the two imaging channels are labeled with similar intensities and signal fills the full dynamic range of the detectors used. Hence, for the preliminary analysis shown (Fig. 1 and Table 1) co-localization analysis was performed on manually selected images with similar intensities and label distribution in the two imaging channels. However, as image selection might bias analysis, we next analyzed larger data sets without prior sample selection (Fig. 2A).

The local chromatin environment defines the integrity of DNA foci

As part of a detailed analysis of the structure of DNA foci in untreated cells, we also evaluated if the integrity of foci was influenced by the local chromatin environment. Molecular mechanisms that define the higher order structure of chromatin domains are unknown. However, as foci within the euchromatin and heterochromatin compartments – which are labeled at defined times of S phase – persist over many cell generations it is reasonable to suggest that the chromatin environment contributes to the preservation of these structures. To evaluate if the epigenetic status of chromatin influences the structure of DNA foci, we analyzed foci in cells treated with the histone deacetylase (HDAC) inhibitor TSA [29]. As before, replicating DNA was labeled using double-pulse strategies and individual CTs resolved through random mitotic segregation during cell proliferation. Cells with discrete foci were then treated with TSA and imaging performed 24 h later (Fig. 2).

As discussed above, for detailed quantitative analysis, double-labeled cells were randomly selected and 3D images stack generated (Fig. 2A–B); as before, only nuclei with labeled early S phase foci (i.e. euchromatic) were used for subsequent analysis. Cell populations were processed either as raw images or after filtering and thresholding as described (Fig. S1) and statistical tests performed (not shown) to establish that the analysis of 50 cells/sample was sufficient to ensure reliability of the data. In parallel samples, cells were treated using 2 concentrations of TSA (Fig. 2), which were selected based on the extent of changes in acetylation of histones H3 and H4 (Fig. 2C). Untreated (control) samples contained discrete DNA foci that were distributed in distinct domains with regions of co-localization restricted to the boundaries of adjacent domains. In this data set, cropped regions with the highest levels of co-localization contained only 0.55+/−0.6% of co-localized voxels when, on average, 27.8% of voxels in the selected regions were labeled (Table 2).

When cells were treated with TSA a clear increase in channel co-localization was seen (Fig. 2 and Table 2). When Pearson's correlation coefficient was used as an indicator of co-localization, differences were statistically significant when cells were treated with TSA at 100 ng/ml and an intermediate level of co-localization was seen when 50 ng/ml was used (Fig. 2D and Table 3). Importantly, the same trends were seen when analysis was performed on raw images, without processing, or after median filtering and thresholding (Fig. 2Di). However, as Pearson's correlation coefficient provides an abstract indicator of channel cross-talk or co-localization, we also deconstructed images and used a volex level co-localization analysis to calculate the volume of voxels that contained both labels (Fig. 2Dii and E). This analysis confirmed that the co-localized volume in the nuclei of untreated control cells was restricted to ~6 μm^3, representing 0.3% of the nuclear volume, whereas following treatment with TSA at 100 ng/ml the co-localized volume increased to ~24 μm^3 (Table 2).

Many experiments support the idea that euchromatic and heterochromatic DNA foci have distinct characteristics that contribute to the spatial organization of CTs [4–6]. To assess

Table 1. Analysis of different approaches for signal co-localization.

Summary 10 nuclei	No threshold			Thresholded		
Coloc. Metric	PC	Mander's		PC	Mander's	
		Green	Red		Green	Red
Raw files (no filtering)	−0.03646	0.18368	0.05448	−0.05216	0.09689	0.0401
Median Filter 3×3×3	−0.07823	0.11228	0.01549	−0.10285	0.04953	0.01616
Gaussian Filter 0.08 μm	−0.04365	0.5957	0.42345	−0.11055	0.1041	0.07763

A variety of automatic and manual protocols were tested to monitor levels of co-localization in samples generated throughout this study. Confocal series were collected (with sequential imaging in the labeling channels) and data files imported into image analysis software (Imaris suite). Pearson's and Mander's coefficients were used as indicators of the extent of co-localization between different channels (see text). Entire confocal series for 10 different nuclei (like those shown in Fig. 1C) were used to analyze apparent co-localization between the imaging channels using the different conditions identified in the Table, as discussed in the text. It is notably that the different conditions used have only a superficial impact of levels of co-localization, with very weak co-localization seen in all cases. Simple median filtering improves the quality of the images and decreases apparent co-localization relative to the unprocessed images. However, using Gaussian filtering as an alternative dramatically increases the apparent co-localization, by spreading the edges of the labeled structures (this is evident from differences in the respective Mander's coefficients). Thresholding after preliminary processing (filtering) eliminates low-intensity volexs but reduces levels of co-localization only slightly.
doi:10.1371/journal.pone.0027527.t001

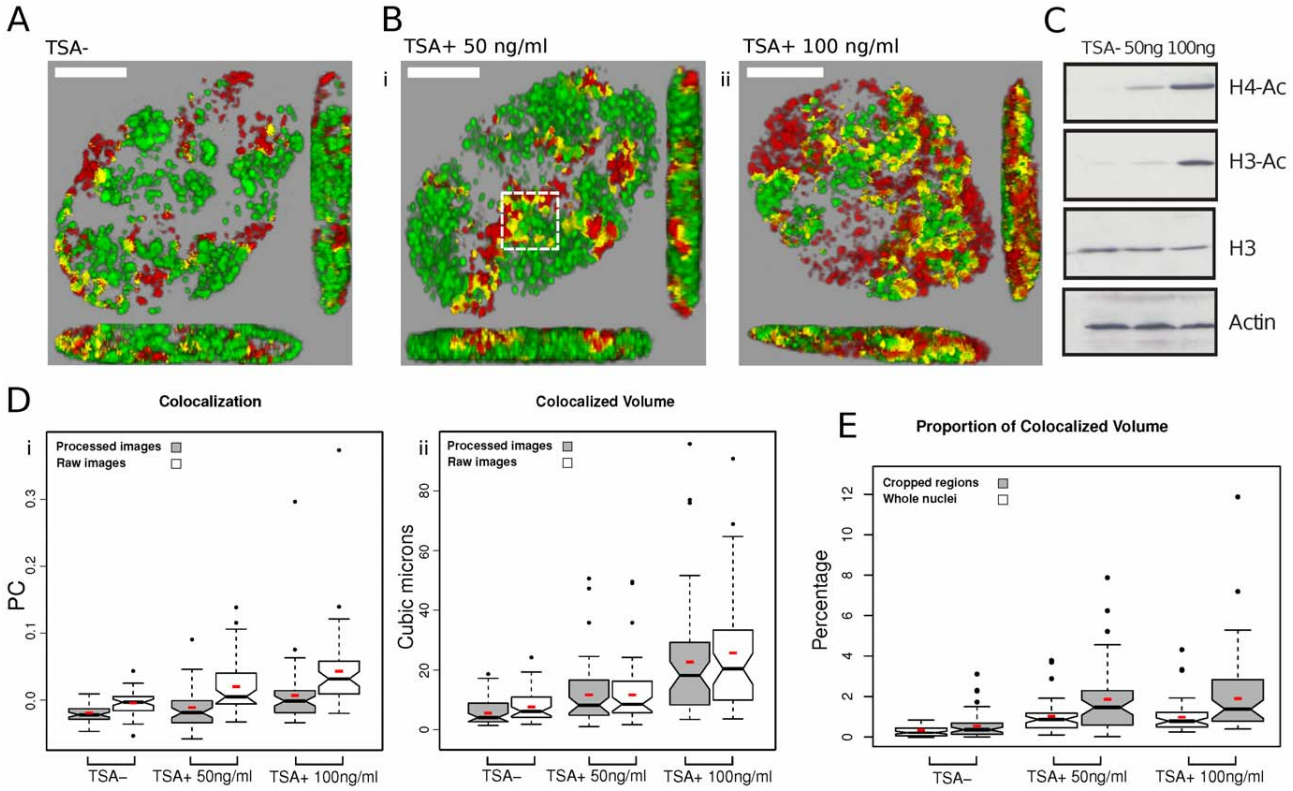


Figure 2. Chromatin epi-states define foci structure. AF488- and Cy3-dUTP were incorporated into DNA foci are described in the legend to Figure 1. Prior to imaging, cells were treated with TSA (24 h) at the concentrations shown (A–E). For imaging, samples were fixed and confocal Z series collected (Zeiss LSM710) and processed (A–B); imaging was performed on double-labeled nuclei but without selection for labeling intensity. Changes seen in the structure of DNA foci following treatment with TSA (Bi–ii) correlated with changes in global histone acetylation by Western Blot analysis using specific antibodies to pan-Ac+ histones H3 and H4 (C). Image processing of the confocal projections (n = 50 per sample) was performed using Fiji and jacop software. Analysis was perform on raw images, without processing, and on the same images after processing as describe in the legend to Figure 1 (D). As seen in Figure 1, untreated cells (A) gave a negative Pearson's coefficent (Di) consistent with low levels of colocalization in the sample. Following TSA treatment (B), a significant increase in Pearson's coefficient was recorded, demonstrating increased co-localization (Di). In order to develop quantitative estimate of channel co-localization, voxel-level channel intensities were extracted and the volume (μm^3) of co-localized voxels calculated (Dii). Finally, the co-localized volumes were calculated as a proportion (%) of the total nuclear volume (whole nuclei) and the volume of the most highly labeled regions (cropped regions – the boxed region in Bi shows a typical example) of individual nuclei (E; Table 2). Small red boxes on the box plots represent the mean value for each distribution. Scale bars of 5 μm are shown on individual panels.
doi:10.1371/journal.pone.0027527.g002

Table 2. Comparison between co-localization results for entire nuclei and cropped regions.

	Vol Coloc. (μm^3)	% Coloc.	% Green	% Red	PC	M Green	M Red	%occupied
Nuceli n = 50								
TSA−	6.03±4.49	0.30±0.22	8.48±2	4.73±1.7	−0.020±0.01	0.03±0.02	0.05±0.02	13.21
TSA+ 50	12.23±10	1.01±0.8	12.06±4.2	7.71±2.5	−0.015±0.02	0.07±0.04	0.11±0.07	19.77
TSA+ 100	23.90±19	0.98±0.8	6.86±2.33	9.75±2.7	0.006±0.05	0.12±0.08	0.09±0.07	16.61
ROIs n = 50								
TSA−	0.36±0.4	0.55±0.6	13.87±4.3	13.93±5.6	−0.099±0.03	0.03±0.03	0.04±0.03	27.81
TSA+ 50	1 ±0.6	1.8±1.6	19±8.3	19±5.7	−0.13±0.07	0.08±0.06	0.09±0.06	39
TSA+ 100	1.55±2.7	2.15±2.07	17.50±4.5	21.97±5.6	−0.1087±0.08	0.12±0.1	0.1±0.1	40

Data from experiments described in Figure 2 are shown. Cells were labeled and processed as described and 50 whole nuclei of selected cropped regions (ROIs) from each nucleus were analyzed. Cropped region of the same surface area were selected to contain the most highly co-localized nuclear region. A typical example is highlighted in Figure 2Bi (boxed area) in which a regions of substantial chromatin mixing (yellow) lies at the junction of discrete red and green chromatin domains. The total volumes of nuclei and selected cropped regions were 1953±946 μm^3 (n = 150) and 65±7 μm^3 (n = 150), respectively. No statistically significant differences in the nuclear volume of TSA treated cells were seen.
doi:10.1371/journal.pone.0027527.t002

Table 3. Statistical Analysis of co-localization results.

Pair-wise Mann-Whitney		
Nuclei		
	TSA+50 ng/ml	**TSA+100 ng/ml**
Pearson's Coefficient		
TSA−	6.2E-01	5.5E-06
TSA+ 50 ng/ml	x	1.83E-03
Co-localized Volume		
TSA−	2.482E-04	9.457E-12
TSA+ 50 ng/ml	x	8.154-05
Co-localized Percentage		
TSA−	7.616E-10	4.923E-11
TSA+ 50 ng/ml	x	8.931E-01
ROIs		
	TSA+50 ng/ml	**TSA+100 ng/ml**
Pearson's Coefficient		
TSA−	5.88E-03	3.519E-02
TSA+ 50 ng/ml	x	4.645E-01
Co-localized Volume		
TSA−	3.705E-08	5.922E-11
TSA+ 50 ng/ml	x	1.564E-01
Co-localized Percentage		
TSA−	8.593E-09	1.544E-10
TSA+ 50 ng/ml	x	6.951E-01
Kruskal-Wallis test		
Nuclei		**ROIs**
Pearson's Coefficient		
3.88E-05		1.481E-02
Co-localized Volume		
1.166E-11		1.13E-11
Co-localized percentage		
1.68E-12		1.513E-11

As the distributions of the values from the three different treatments are not normal, non-parametric methods were used. The Mann-Whitney test was used for pair-wise tests and the Kruskal-Wallis test for multiple comparisons.
doi:10.1371/journal.pone.0027527.t003

how these specialized chromatin states contribute to CT structure, we analyzed DNA foci within isolated CTs that were labeled with biotin-dUTP (early S phase) and BrdU (mid/late-S phase) using a pulse-chase-pulse strategy (Fig. 3). After labeling, cells were grown for 5 days to reveal isolated CTs, treated with TSA for 24 h and the structure of DNA foci and CTs analyzed. In comparison with untreated control cells from the same labeled population (Fig. 3A), the DNA foci of cells treated with TSA were clearly swollen and dispersed (Fig. 3B–C), consistent with the local mixing of adjacent foci seen along the boundaries of neighboring CT (Fig. 2B). However, despite the clear structural deterioration and associated >2-fold increase in CT volume (Fig. 3D) widespread mixing of the early and late chromatin domains was not seen (Fig. 3), suggesting that even following TSA treatment some residual higher-order structure is preserved. Based on these observations, we propose that the chromatin environment has a significant influence on the structure of DNA foci and that patterns of interaction between foci contribute to the spatial architecture of CTs.

Analysis of completely labeled CTs

Fluorescently labeled replication precursor analogues have distinct technical advantages for image analysis (Figs. 1 and 2) but are limited by the extent of incorporation achieved; under conditions used here, which do not compromise the rate of DNA synthesis, the modified precursors are only incorporated into growing replication forks for ~15 min before the labeled precursor dNTPs are consumed. This limitation means that alternative strategies must be used to provide global estimates of DNA mixing within mammalian nuclei.

We considered two strategies for estimating genome-wide levels of inter-chromosomal DNA mixing. First, because euchromatin and heterochromatin occupy discrete nuclear domains the distribution of these chromatin compartments is non-uniform [4–6]. Hence, even in nuclei with partially labeled genomes it should be possible to estimate the maximum extent of DNA mixing locally using volumes of the nucleus in which the majority of DNA foci are labeled. This approach was tested above (Figs. 1 and 2) using crops of the most highly labeled nuclear regions. Within these highly labeled domains, typically ~30% of the cropped nuclear volume was occupied by labeled DNA. Moreover, as 50% or less of the nuclear volume is occupied by chromatin [30], we can estimate that >60% of the chromatin space present in the selected regions will be labeled. Given the short duration of labeling with modified dNTPs this might seem surprising. However, when the structure of DNA foci that had been pulse-labeled with biotin-dUTP were compared with foci labeled with BrdU throughout S phase no significant differences were seen (Fig. 4). This shows that when the DNA within individual foci is only partially labeled the plasticity of chromatin folding means that the labeled and adjacent unlabeled regions cannot be resolved. In fact, as typical foci have 3–5 replicons and so 6–10 extending replication forks, it is not surprising that the labeled and unlabeled regions cannot be resolved by confocal microscopy.

For the second approach we analyzed cells with DNA that had been labeled with BrdU throughout S phase. Prior to this analysis we performed an extensive analysis of labeling specificity [31], and found that DNA foci labeled with BrdU or biotin-dUTP could be distinguished with good sensitivity and without cross-talk between the imaging channels. A pulse-chase-pulse-chase labeling strategy (Fig. 4D) was used to monitor the interaction between DNA foci of neighboring Br and biotin-containing nuclear domains (Fig. 4E). As before (Figs. 1 and 2) regions of co-localization were clearly restricted to boundaries between these neighboring domains. In the cropped region highlighted (Fig. 4Ei), the Pearson's coefficient is −0.3036, consistent with exclusion of signal in the two labeling channels (Fig. 4Ev). As in this typical example, even when individual CTs are completely labeled with BrdU, we see that only ~1% of the nuclear volume contains both Br and biotin-labeled DNA.

Chromatin dynamics in living cells

In seminal studies on CT dynamics, Cy3-labeled DNA foci in HeLa and SH-EP N14 (a neuroblastoma cell line) cells were shown to undergo constrained random diffusion with rare examples of directional motion correlating with changes in cell shape [13]. Later, a more sophisticated analysis of temporal dynamics was performed using a ~10 Mbp artificial repeat that was able to bind lacI-GFP [17,32]. However, largely because of technical limitations, we have only limited understanding of dynamic changes that

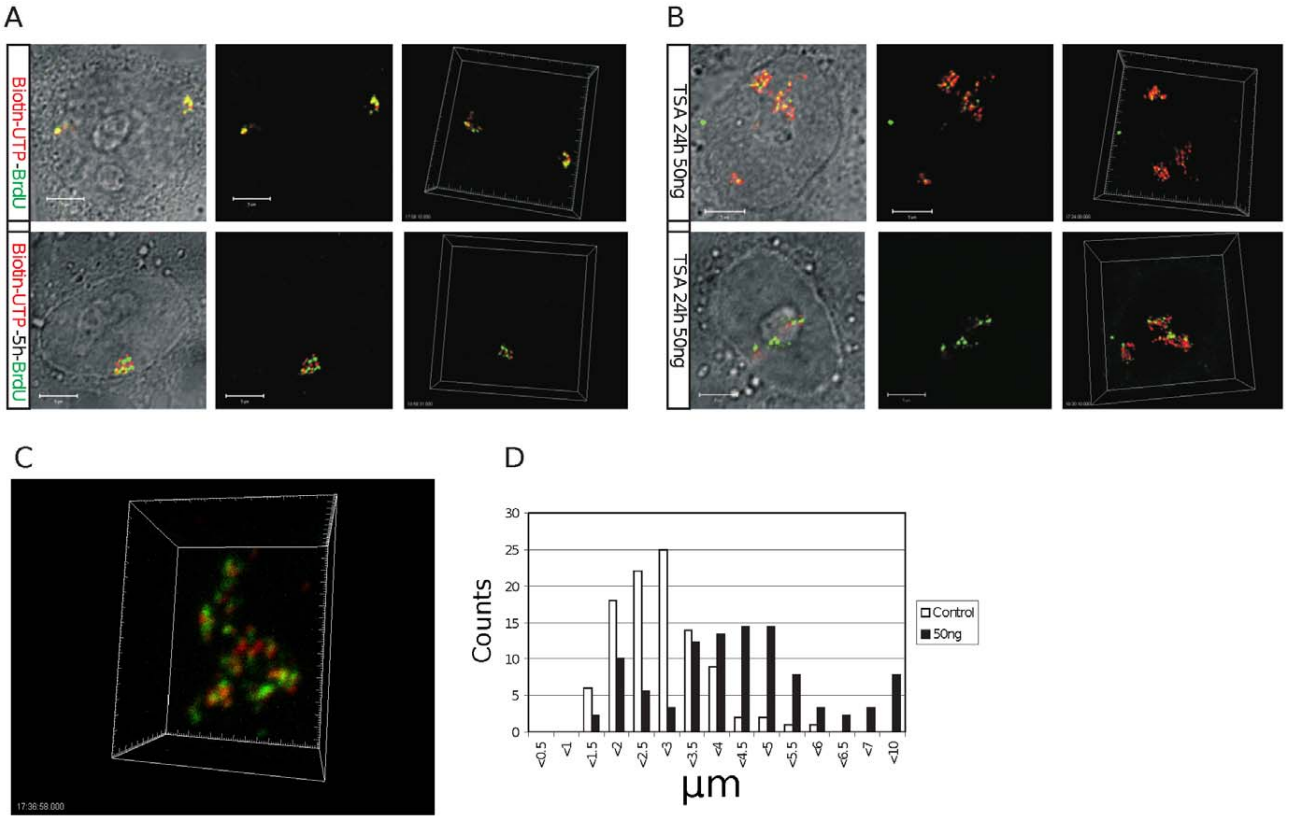


Figure 3. The chromatin environment contributes to the local and long-range architecture of DNA foci and CTs. HeLa cells were pulse-labeled with biotin-dUTP and BrdU either sequentially or with an unlabeled intervening chase of 5 h and then grown for 6–7 days to resolve individual CTs by random mitotic segregation. Cultures were then divided into 2 and treated without (A) or with (B–C) 50 ng/ml of TSA for 24 h. Simple visual inspection showed CTs to be visibly disorganized and expanded (C – shows an isolated CT from the sample in B) with notable deterioration in the structure of DNA foci, which appeared irregular and diffuse relative to untreated controls. Expansion of CTs was confirmed by measuring the diameter (long axis) of individual territories (D). Though the structural changes were obvious in pulse-labeled samples (compare A and B) for this analysis we used cells that were labeled with BrdU throughout S phase so that the boundaries of individual CTs could be identified with confidence. The diameter of CTs (D) within control cells (open bars; dia = 2.65 μm +/−1.25; n = 100) was seen to increase by 1.59-fold in TSA treated cells (closed bars; dia = 4.21 μm +/−1.68; n = 90; t test $-p<1.3\times10-12$). Under these conditions, there was no significant change in the average nuclear volume of the two samples. Scale bars of 5 μm are shown on individual panels.
doi:10.1371/journal.pone.0027527.g003

occur when DNA foci engage DNA or RNA polymerases to function as a synthetic template [13,14,32–34].

As structural transitions related to chromatin function must increase the probability of DNA mixing within the inter-chromatin domain, we next wanted to evaluate if live cell imaging could be used to define the structural stability of DNA foci. Because DNA foci within euchromatin and heterochromatin have well-characterized nuclear organization [15] and dynamic properties [14], it is possible to use foci labeled at different times of S phase as metastable landmarks to map the relative movement of individual foci. A typical example of this approach is shown in Figure 5. Replication foci were labeled with AF488-dUTP during early S phase and Cy3-dUTP 5 hours later (Fig. 5A). Using this labeling program, the relative spatial stability of heterochromatic foci labeled during mid-S phase (red) can be used as anchor points to align CTs at different times during the imaging series and so increase the confidence with which the location of individual foci can be assigned. In the example shown, the overall shape of the CTs and local architecture of individual foci is maintained throughout the imaging time-course even though it is not unusual to see local transformations in the shape of individual CTs; the example shown here is seen to rotate around its vertical axis (Fig. 5B).

Even though the quality of foci is limited by the imaging set-up (low laser power) used during live cell imaging, the type of data shown in Figure 5 allows unambiguous identification of discrete foci using image processing software (Fig. 5B). In this typical example, individual foci are most obvious along the periphery of CTs (Fig. 5B - regions highlighted in white ovoids). In such areas of the sample, the ability to track and assign co-ordinates for individual foci during time-lapse imaging allows the location and movement of individual foci to be monitored with confidence. As noted before [13,14], we found euchromatic foci to be locally dynamic, typically moving ~0.5 μm over periods of 15 minutes (Fig. 5D). However, dramatic directional movements were never sustained for long periods. Instead, foci appeared to oscillate within CTs so that individual territories maintained their relative position and general shape for many hours. Relative to euchromatic foci, heterochromatic foci were frequently clustered and showed significantly reduced mobility (Fig. 5D). This correlates with heterochromatic foci being preserved as temporally stable clusters of structurally inert chromatin. The architecture of mid/late replicating DNA foci correlates with the structural polarization of CTs and corresponding programme of DNA synthesis in mammalian cells [35].

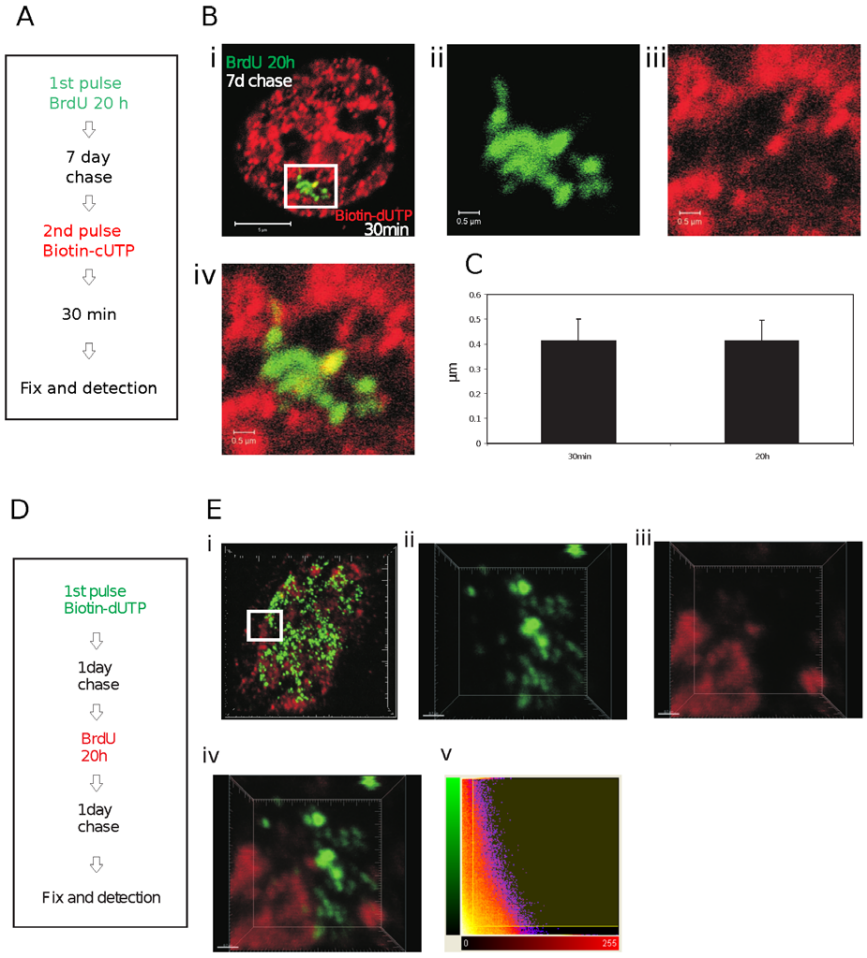


Figure 4. Estimating global levels of chromatin mixing. Pulse-labeling with conjugated replication precursors such as Cy3- or biotin-dUTP yields labeled DNA foci in which only about 15% of DNA contains the modified precursor. As we are able to measure DNA mixing at the boundaries of such foci in highly labeled nuclear volumes, it is important to know if apparent volumes of foci are influenced by the extent of modified precursor incorporation. HeLa cells were labeled as shown (A) and processed by indirect immuno-labeling. Confocal projection of double-labeled cells like that shown (B) were collected and the diameters of foci measured (C) using Imaris software in selected regions as shown (a zoom of the boxed region in Bi is shown as individual channels in Bii and Biii and Biv shown the channel merge). Foci that were pulse-labeled with biotin-dUTP had an average diameter of 0.413+/−0.087 µm (mean+/−SD; n = 100). Foci that were labeled for the entire S phase with BrdU had an average diameter of 0.414+/−0.081 µm (mean+/−SD; n = 100). Bars are 5 and 0.5 µm in low and high power images, respectively. To increase the extent of labeling in one imaging channel, foci within individual CTs were labeled with either BrdU or biotin-dUTP using the labeling scheme shown (D). Samples were fixed and processed to visualize site of incorporation by indirect immuno-labeling using secondary antibodies conjugated with Qdots™; Qdots are very stable during illumination and allowed sampling using 50 nm Z steps (92 slices in the examples shown) and multiple scans without bleaching. Individual cells were selected and confocal projections generated (E). Regions from selected cells were analyzed to identify the extent of co-localization between the two labeling channels. A 3-D reconstruction of the region highlighted in (Ei) was used for further analysis (Eii–Ev: Eii (green) and Eiii (red) show signal in the separate labeling channels; Eiv an overlay of the red and green channels and Ev a co-localization analysis using Imaris software). Note the discrete nature of the labeled sites in both labeling channels and almost complete lack of sites of overlap (yellow) in the channel merge (Eiv) – in this typical example the co-localized volume was 0.96%. Scale bars are 5 and 0.7 µm in low and high power images, respectively.
doi:10.1371/journal.pone.0027527.g004

Discussion

Competing models of nuclear organization have addressed how prevailing views of CT structure and chromatin dynamics might be resolved [36–38]. Traditionally, fluorescent in situ hybridization (FISH) has been used to define the distribution of DNA from individual chromosomes. This 'chromosome painting' of intact nuclei showed CTs to be discrete structures [4]. However, quantitative analysis of low-level surface mixing is technically challenging within the 3D volume of an entire CT. To address this point, Branco and Pombo [39] applied routine FISH techniques to ~200 nm cryosections. With this approach, the borders of neighboring CTs were seen to contain extensive domains of inter-chromosomal mixing. For example, when PHA stimulated human lymphocytes were analyzed, ~40% of the chromatin-rich compartment – corresponding to ~20% of the nuclear volume - was estimated to contain DNA from more than one chromosome.

In higher eukaryotes, we have limited information about the range and scale of chromatin dynamics and the potential for inter-chromosomal mixing in living cells. Individual CTs within mammalian nuclei are known to be locally plastic [13,15] and many have structures that change both locally [14] and at longer range [16,40] in response to changes in gene expression. However,

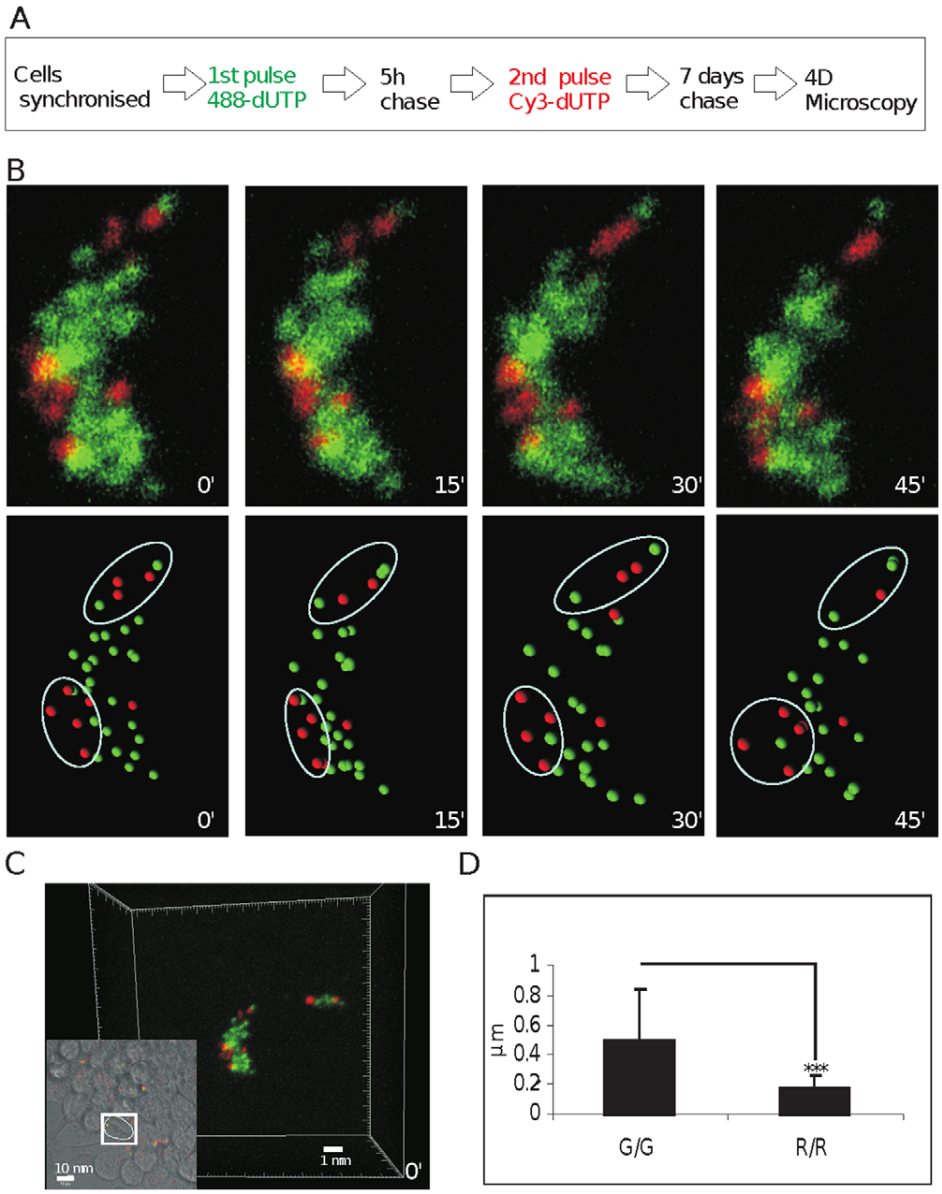


Figure 5. Differential dynamic behavior of DNA foci labeled during early and mid/late S phase. Early replicating euchromatic foci were pulse-labeled with AF488-dUTP (green) and mid/late replicating foci with Cy3-dUTP (red) - an optimal pulse separation of 5 h was established experimentally (A). Individual CTs were resolved by mitotic segregation for 7 days (A–C) and confocal time-lapse microscopy (Zeiss LSM510META) performed over 1–2 h with sampling every 15 min. Raw images (B; upper panels) show maximum projections of Z stacks for a typical isolated CT (from the cell highlighted in C). Raw images were imported into Imaris software in order to determine the mass centers of individual labeled sites (DNA foci). Software-generated spheres (250 nm; B; lower panels), which represent the mass centers of discrete foci, were used to develop 3-D coordinates to define changes in separation of paired neighboring foci (n = 42). For each time-lapse series used (B), 4 3-D projections were generated at 15 min intervals and specific regions identified (white ovoids) where foci could be tracked without ambiguity. In each image, the separation between all neighboring pairs of assigned foci (within each ovoid) was determined. Finally, the 4 data sets were used to calculate 3 relative separations that define the change in separation of assigned foci in µm/15 min (µm in D). The relative dynamic properties shown relate to early foci within early replicating chromosomal domains (G/G) and mid/late foci within mid/late replicating domains (R/R). Scale bars of 1 and 10 µm are shown on individual panels.
doi:10.1371/journal.pone.0027527.g005

detailed dynamic studies on specific endogenous loci have not been reported. Moreover, with live imaging, it is extremely difficult to reliably measure subtle changes in shape and intensity of 3D structures based on fluorescent time-lapse imaging. Hence, it is unclear how the structure of DNA foci changes when functions such as DNA or RNA synthesis are performed (e.g. [13,14]).

The molecular mechanisms that define the structural properties of DNA foci have not been explored in detail. It is known that individual foci within euchromatin and heterochromatin are discrete entities, implying that the local chromatin environment contributes to their structure [12,15]. It is not clear if individual foci have strictly defined boundaries or how possible boundaries

might be formed. Even so, a recent study using an unbiased genome-wide 3C approach – termed Hi-C - has demonstrated that the analysis of cell populations shows DNA to be clustered into ~1,000 kbp chromatin domains [41]. The DNA domains that were predicted during analysis of the Hi-C data show a strong size correlation with DNA foci [8] and replication timing domains [42]. Interestingly, bioinformatic analysis has predicted that at least part of the structural organization of the higher-order chromatin domains or globules correlates with the distribution of the insulator protein CTCF on DNA [43,44] and the association of active genes within common transcription factories [43–45].

Structural plasticity of CTs and DNA foci

In this study, we used a single cell approach to evaluate if the structure and dynamic behaviour of CTs could be rationalized with the formation of wide-scale genomic interaction networks, which can be crudely defined by the extent of inter-chromosomal mixing within nuclei. We used fluorescent thymidine analogues to label DNA foci in living cells and then used light microscopy to monitor both the structure and dynamic behaviour of individual foci and CTs. Using this approach, we saw little evidence for extensive zones of DNA mixing between the foci in adjacent chromatin domains (Figs. 1, 2). Indeed, analysis of the co-localization of DNA from neighboring CTs suggests that nuclear domains in which chromatin from different CTs is freely mixed represents a small fraction – probably no more than 1% - of the chromatin space. The very limited mixing that we see agrees with other reports [11,12] but appears at odds with the existence of widespread chromatin domains within which DNA from different chromosomes is mixed [39].

Whether this apparent discrepancy results from experimental differences or innate differences in the cell systems used is presently unclear. Our analysis is dependent on metabolic labeling during DNA replication and transformed human HeLa cells are ideally suited for this approach. Pombo and colleagues used freshly isolated peripheral human lymphocytes, which in some cases were activated by PHA treatment [39,46]. One possibility might be that the observed differences reflect changes in higher-order chromatin architecture in transformed and specialized cells. Various technical limitations might also contribute to the differences seen using different experimental strategies. Notably, our preferred analytical strategy simply involves fixation and analysis of higher-order DNA structures that were fluorescently labeled during DNA replication (Fig. 2). Pombo et al. used technically elegant in situ hybridization to visualize the distribution of CTs. Unlike our non-destructive labeling, hybridization demands that the target DNA is denatured, which even in fixed samples might involve some loss of local structure. In addition, our metabolic labeling approach allows visualization of DNA foci [8] in samples without background (i.e. unincorporated) label. In contrast, when analysis is based on DNA hybridization, samples often contain low-level background staining, which makes true signal difficult to define [36].

Dynamic DNA foci and chromatin looping

Our observations suggest that the chromatin in human HeLa cells does not undergo wide-scale inter-chromosomal mixing (Figs. 1, 2, 3, 4). From our analysis, we estimate that within individual cells only ~1% of DNA is found to occupy nuclear sites where DNA from different chromosomes is likely to be freely mixed. This level of potential interaction does however reflect a snap-shot in time and it is also important to emphasize that CTs [13–16] and their constituent DNA foci [14,15] are dynamic and able to engage in structural transformations (Fig. 5), so that different loci might interact with numerous other loci at different times. It is reasonable then to assume that such changes will respond to the functional state of chromatin, and not difficult to imagine how post-translational histone epi-states define a chromatin landscape, which also contributes to patterns of DNA interaction. In addition, while specific patterns of inter-chromosomal interactions might form preferred steady-state structures in differentiated cells it is important to consider how such interactions might be influenced by the formation of chromosomes and their CTs during cell division. Chromosome condensation will inevitably disrupt inter-chromosomal DNA interactions that exist during interphase and so reset the interaction networks to a structural ground-state that will be based on local structure.

While DNA foci with ~1 Mbp of DNA are widely accepted as fundamental higher-order features of chromosome structure surprisingly little is known about the molecular principles that regulate chromatin function within these structures. Though the formation of foci is unlikely to reflect a single mechanism, it is notable that the foci which form within the euchomatin and heterochromatin compartments are distinct. This is consistent with the local chromatin environment contributing to the structure and stability of individual foci. To test this possibility, we perturbed the local chromatin environment within DNA foci by manipulating the acetylation status of histones using the histone deacetylase inhibitor TSA. After treatment with TSA, under conditions that increased global histone acetylation ~5-fold, clear changes in the structure of DNA foci were seen (Figs. 2, 3). Notably, foci became more open or dispersed and this correlated with a 4-fold increase (Table 2) in the volume of nuclear domains where DNA from adjacent chromosomes was intermingled. TSA-induced changes in the structure of DNA foci also correlated with a more general disorganization of CTs, which showed widely variable structures and increased size (Fig. 3). These experiments show that the chromatin environment contributes to the structure of DNA foci so that when the chromatin environment is perturbed a corresponding deterioration in the structure of DNA foci and CTs is seen.

Inside the nucleus, DNA and RNA synthesis are performed within the inter-chromatin compartment, and not within the chromatin-rich DNA foci themselves [36–38]. Because of this spatial separation, it is self-evident that chromatin loops must be extruded from the foci towards the active sites during synthesis. This requirement for movement of the chromatin fibre raises the possibility that chromatin loops continually escape from the surface of structural foci in order to probe the inter-chromatin space where favourable synthetic environments might be encountered. During this process, extended chromatin loops from neighboring territories might occupy the same nuclear space and so have a high probability of interacting, for example by binding to a common transcription factory. The analysis presented here suggests that at any time the extended loops represent a very small amount – ~1% or less - of the mammalian genome. Even so, it is important to recognise that the single cell analysis used is unable to explore the range (spread) of isolated chromatin fibres and it remains an open question if extended chromatin fibres are able to persist as a result of stable interactions within the inter-chromatin space [1–3].

Our experiments imply that extended loops that spread from the surface of CTs are generally short-range and probably short-lived. If long-range (>μm) open loops are able to form in our experimental system, these must be rare in individual cells or below the level of detection of our analysis. In fact, chromatin loops that extend well outside the normal boundaries of CTs are not uncommon and provide an obvious means of increasing the range of inter-chromosomal contacts while maintaining the

normal higher-order packaging density of DNA [36]. While such extruded loops provide one class of CT remodelling, we believe that our observations support a model in which the majority of inter-chromosomal contacts form locally at the boundaries of CTs, in domains where chromatin architecture might be open and dynamic. In this regard, it is notable that recent studies using electron spectroscopic imaging [47] have suggested that the majority of chromatin in mammalian cells is in the form of 10 nm chromatin fibers, which in differentiated cells fold locally to form higher-order DNA foci [30]. Interestingly, the chromatin fibers within embryonic stems cells appear to be much more chaotic, perhaps implying that ES and differentiated somatic cells have quite different principles of higher-order chromatin organization.

Analysis of widespread chromatin dynamics in different cell types supports this possibility. It is well-known that in differentiated cells chromatin structure is spatially stable over long periods of time [48] whereas similar experiments performed in ES cells shows their genome organization to be extremely plastic [Thomas Cremer, personal communication]. This implies that stable higher-order structures seen following differentiation are not a major chromatin feature in developmentally primitive cells. Even so, indirect functional evidence does show that some level of higher-order structure is present in stem cells. Notably, replication timing domains that are seen following cell commitment correlate with ~0.5–1 Mbp chromatin domains – the DNA foci [8] – and similar replication structures are seen both in mouse ES cells and in their committed and differentiated descendants [42].

Conclusions and perspectives

Genome-wide studies offer a promiscuous view of inter-chromosomal interactions, which suggest a significant degree of intermingling between DNA from different CTs (e.g. [41]). However, to date such experiments have been performed on large cell populations and provide a view of potential interactions without having the power to predict the frequency of these interactions within individual living cells. Moreover, unbiased analysis of potential genome interactions using Hi-C clearly shows that intra-chromosomal interactions within CTs are at least 2 orders of magnitude more frequent than inter-chromosomal interactions (see Figure 2 in [41]); this level is consistent with the potential for chromatin mixing described herein. Hence, while the formation of extensive interaction networks within mammalian cells appears to conflict with the idea that individual CTs are spatially self-contained [4], dynamic changes at the interaction interfaces of neighboring CTs (Fig. 5) can be sufficient to allow the formation of widespread gene interactions while preserving CTs as higher-order chromatin structures. As a growing body of evidence supports the formation of cell type specific 'interactomes' during cell differentiation [18–23], it is important to understand how different patterns of gene expression correlate with the formation of interaction networks and how these interactions define spatial and temporal changes in genome structure and function within individual cells.

Materials and Methods

Visualizing replication foci in human cells

HeLa cells were grown in the presence of different dTTP analogues to label sites of DNA synthesis, as described in detail by Maya-Mendoza et al. [49]. The following precursors were used: AlexaFluor488-dUTP (AF488-dUTP); Cy3-dUTP; biotin-dUTP and bromo-deoxyuridine (BrdU). AF488-dUTP and Cy3-dUTP were visualized either in living cells using time-lapse light microscopy or by confocal microscopy after fixation using routine procedures. For fixation, cells growing on glass coverslips were rinsed briefly in PBS (1 sec; 20°C to remove medium and fixed in 4% paraformaldehyde (15 min; 0°C). These fixation conditions preserved the structure of chromatin domains present in living cells and no changes in structure of the chromatin foci was seen under the imaging conditions used. Fixed cells were washed 3× in PBS, treated with 0.5% Triton ×100 in PBS, rinsed 3× in PBS, incubated with 5 μg/ml Hoechst 33258 (Sigma) for 10 min, rinsed 3× in PBS and mounted with either Vectashield or Prolong mounting media. Alternatively, DNA foci were labeled by indirect immuno-fluorescence [49]. Where secondary fluorescent antibodies were replaced by Qdots the following changes were applied: 1) permeabilization was altered to 1% Triton ×100 for 10 min; 2) Qdots (1/500 dilution) were applied to coverslips in 24 well plates and incubation performed for 15 h at 4°C with shaking (orbital rocker); fixation, primary antibody incubation and washes were as for routine immuno-labeling. We note that in our hands the performance of Qdots was very variable from batch to batch, with some batches giving high background staining. Qdots were from Invitrogen: streptavidin conjugated Qdot-525 was used to detect biotin labeled CTs and secondary anti-rat antibody conjugated with Qdot-605 to detect BrdU. TSA was purchased from Sigma. Western blotting was performed as described [50] using appropriate antibodies (Abcam), as shown.

For confocal imaging, samples were examined using a Zeiss LSM510META confocal microscope following well-established imaging protocols [27,28]. Labeling conditions were selected to minimize background noise and the microscope configuration was selected to reduce bleed-through between imaging channels to negligible levels. In order to ensure optimal imaging performance, instrument alignment was performed at regular intervals by Zeiss. Multi-coloured TetraSpeck florescent beads were used to monitor point spread functions and correct chromatic shift; maximum tolerated shifts were 50 nm in X-Y and 100 nm in Z. To minimise chromatic aberrations, great care was also taken to balance labeling intensities in different imaging channels. Confocal sections were collected through a 100× (1.45 NA) lens and 3-D images generated using Z stacks and processed in Imaris® software. For LSM510 image acquisition the following channel settings were used: green −488 nm laser line at 2% intensity with a BP 500–530 IR filter; red – 543 nm laser line at 32% of intensity and LP 545 filter. 4-D time-lapse imaging was performed using either a DeltaVision microscope with a CoolSNAP-HQ2 camera and Olympus objective (100×; 1.4 NA) or Zeiss LSM510META confocal microscope using the settings detailed above. The Deltavision system was used for long-term imaging experiments (e.g. Fig. 1), with the intensity of light during imaging kept to 32% using an acquisition speed of 100–200 ms. The conditions used allow imaging for at least 2 days without influencing cell viability or cell cycle parameters. Because of the zoom facilities, the Zeiss system was used when foci-level resolution was required (Fig. 5). As above, the light intensity was reduced to the minimum required to resolve individual foci and the imaging conditions used were shown not to prevent subsequent cell division.

For detailed co-localization analysis (Fig. 2), confocal imaging was performed using a Zeiss LSM710 microscope using instrument setting equivalent to those detailed above to minimize bleed-through between channels and background levels. Z-stacks were acquired for each sample with voxel dimensions of 0.8×0.8×0.34 microns, for X, Y and Z respectively with an XY resolution of 988×988 pixels and a pinhole setting of 1.0 Airy unit. Amplifier and detector gain and offset were optimally chosen by the instrument for each field acquired. For the Alexa-488 channel an EF1 filter set was used with a SPI wavelength range from 493–

543 nm. For the Cy3 channel an EF2 filter set was used with a SPI wavelength range from 566–681 nm.

Image analysis and model building

3-D and 4-D images were analyzed using Imaris® software (Bitplane). For confocal images, Z stacks were processed using Imaris® software after applying a Gaussian or Median filter. Imaris® software was used to process 3-D projections, identify individual foci and assign coordinates for mass centers of each focus. Individual channels were processed separately. Co-ordinates of mass centers were used to define the spatial relationship between adjacent foci, either within or between channels. The mass centers can be represented by computer generated spheres that correspond in size to average foci. Such images are artificial and while providing an accurate representation of the positions of foci are not intended to provide a realistic representation of the foci themselves. Imaris® imaging software was used to isolate cropped regions, with the same crop volumes used for equivalent samples.

For high-throughput image analysis, in-house scripts were developed using Fiji software [51] with the aid of the suite of 3-D filters [52]. Co-localization analysis was performed with JACoP [28] and co-localized volumes estimated by multiplying the number of co-localized voxels by the volume covered by a single voxel. Co-localized voxels where defined as voxels for which both channels indicated values above a threshold point, equal to the standard deviation of the distribution of pixel intensities in the corresponding channel.

To visualize 3-D interactions between CTs (e.g. Fig. 2D), coordinates of each of the fluorescent tags were exported individually into Virtual Reality Modelling Language (VRML) format using Imaris® software. VRML files were exported to 3ds format using an open-source, platform-free 3d-design suite (http://www.blender.org/). These files were imported into Autodesk® 3ds Max® (www.autodesk.com/3dsmax) and imported files merged in a single MAX file to facilitate image rendering, 3-D modelling and animation. This procedure using 3ds built-in compound modifiers models the 3-D shape of the chromatin compartment using the continuity of labeled DNA foci to define the chromatin space.

Supporting Information

Figure S1 Analysis of co-localization in cells labeled with 488-dUTP and Cy3-dUTP. This figure demonstrates how manual thresholding was used for quantitative channel co-localization. The sample described in Figure 1 was used for analysis and the specific nucleus shown in (Cii) use in this example. As in Figure 1, LSM (Zeiss) data files were uploaded into Imaris software and individual channels (3-D) isolated (A: three images on top show the unprocessed green (left) and red (center) channels and the channel merge (right; co-localized sites are shown yellow). Data from the imaging files is extracted as a screen shot on the right. Using data like this we performed a detailed empirical analysis of the behavior of sites of co-localization, using co-localization intensity plots (B). Using these plots, manual thresholding was applied to eliminate background noise. Threshold settings (shown by yellow lines in the intensity co-loclization plot (B)) were adjusted sequentially in order to establish the minimum level that eliminated noise that was clearly unrelated to the real signal (define by nuclear location). Using this approach, the minimum value for thresholding correlated with the standard deviation of the data intensity in the separate imaging channels. When voxels below this intensity were subtracted from the images noise was essentially eliminated without degrading the structure of the true signal (in A, compare raw images (top panel) and equivalent images after noise reduction (lower panel)). Following nosie reduction, the filtered images were analyzed to identify levels of co-localization (data panel below). Finally, the voxels showing co-localization after background subtraction were extracted (C), for comparison with levels of apparent co-localization in the primary image (A: yellow voxels, top right).
(TIF)

Video S1 Preservation of relative spatial architecture of CTs in response to cell movement. Video showing the time-lapse series that includes the individual images shown in Figure 1B. Video rate - 1 frame/sec. 0 to 360 mins.
(MOV)

Video S2 Regions of apparent co-localization between neighboring CTs result from foci that lie in close juxtaposition in nuclear space. This video shows how co-localization alters during Z sectioning of the space-filling model presented in Figure 1Di. Note that while zones of apparent co-localization often appear along the borders where neighboring CTs meet (these appear yellow while panning through the image) high-resolution analysis shows that these rarely represent true co-localization. In fact, sectioning through the nucleus shows almost complete separation of the green- and red-labeled DNA.
(AVI)

Video S3 CT architecture generates frequent regions of interdigitation along the boundaries where neighboring CTs meet. This video shows a high-magnification 3D rotational view of the region shown in Figure 1Dii–iii(from the region highlighted (white box) in Fig. 1Di). Note that domains protruding from the surface of both CTs are able to pass into the neighboring territory. However, while foci from the individual CTs interact within the same nuclear space the structural integrity of the foci appears to be preserved so that DNA interactions are restricted to the surfaces where adjacent foci touch. Such experiments do not support the existence of extensive nuclear domains where DNA from two or more CTs is freely mixed, although it is important to note that DNA within individual foci will also be dynamic so that DNA at the surface of individual DNA foci will also change with time.
(AVI)

Acknowledgments

We thank Casey Bergman and Chi Tang for help and advice.

Author Contributions

Conceived and designed the experiments: PO-C DF DAJ AM-M. Performed the experiments: PO-C DF AM-M. Analyzed the data: PO-C DF AM-M. Contributed reagents/materials/analysis tools: PO-C AM-M. Wrote the paper: PO-C DAJ AM-M.

References

1. Misteli T (2007) Beyond the sequence: cellular organization of genome function. Cell 128: 787–800.
2. Fraser P, Bickmore W (2007) Nuclear organization of the genome and the potential for gene regulation. Nature 447: 413–417.

3. Kumaran FI, Thakar R, Spector DL (2008) Chromatin dynamics and gene positioning. Cell 132: 929–934.
4. Cremer T, Cremer C (2001) Chromosome territories, nuclear architecture and gene regulation in mammalian cells. Nat Rev Genet 2: 292–301.
5. Lanctot C, Cheutin T, Cremer M, Cavalli G, Cremer T (2007) Dynamic genome architecture in the nuclear space: regulation of gene expression in 3 dimensions. Nat Rev Genet 8: 104–115.
6. Cremer T, Cremer M (2010) Chromosome Territories. Cold Spring Harb Perspect Biol 2: a003889.
7. Jackson DA, Pombo A (1998) Replicon clusters are stable units of chromosome structure: evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells. J Cell Biol 140: 1285–1295.
8. Maya-Mendoza A, Olivares-Chauvet P, Shaw A, Jackson DA (2010) S phase progression in human cells is dictated by the genetic continuity of DNA foci. PLoS Genet 6: e1000900.
9. Zhang JM, Xu F, Hashimshony T, Keshet N, Cedar H (2002) Establishment of transcriptional competence in early and late S phase. Nature 420: 198–202.
10. Lande-Diner L, Zhang JM, Cedar H (2009) Shifts in replication timing actively affect histone acetylation during nucleosome reassembly. Mol Cell 34: 767–774.
11. Visser AE, Aten JA (1999) Chromosomes as well as chromosomal subdomains constitute distinct units in interphase nuclei. J Cell Sci 112: 3353–3360.
12. Goetze S, Mateos-Langerak J, Gierman HJ, de Leeuw W, Giromus O, et al. (2007) The three-dimensional structure of human interphase chromosomes is related to the transcriptome map. Mol Cell Biol 27: 4475–4487.
13. Bornfleth H, Edelmann P, Zink D, Cremer T, Cremer C (1999) Quantitative motion analysis of subchromosomal foci in living cells using four-dimensional microscopy. Biophys J 77: 2871–2886.
14. Pliss A, Malyavantham K, Bhattacharya S, Zeitz M, Berezney R (2009) Chromatin dynamics is correlated with replication timing. Chromosoma 118: 459–470.
15. Shopland LS, Lynch CR, Peterson KA, Thornton K, Kepper N, et al. (2006) Folding and organization of a contiguous chromosome region according to the gene distribution pattern in primary genomic sequence. J Cell Biol 174: 27–38.
16. Volpi EV, Chevret E, Jones T, Vatcheva R, Williamson J, et al. (2000) Large-scale chromatin organization of the major histocompatibility complex and other regions of chromosome 6 and its response to interferon in interphase nuclei. J Cell Sci 113: 1565–1576.
17. Levi V, Ruan Q, Plutz M, Belmont AS, Gratton E (2005) Chromatin dynamics in interphase cells revealed by tracking in a two-photon excitation microscope. Biophys J 89: 4275–4285.
18. Schoenfelder S, Clay I, Fraser P (2010) The transcriptional interactome: gene expression in 3D. Curr Opin Genet Dev 20: 127–133.
19. Osborne CS, Chakalova L, Brown KE, Carter D, Horton A, et al. (2004) Active genes dynamically colocalize to shared sites of ongoing transcription. Nat Genet 36: 1065–1071.
20. Spilianakis CG, Lalioti MD, Town T, Lee GR, Flavell RA (2005) Interchromosomal associations between alternatively expressed loci. Nature 435: 637–645.
21. Osborne CS, Chakalova L, Mitchell JA, Horton A, Wood AL, et al. (2007) Myc dynamically and preferentially relocates to a transcription factory occupied by Igh. PLoS Biol 5: 1763–1772.
22. Schoenfelder S, Sexton T, Chakalova L, Cope NF, Horton A, et al. (2010b) Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat Genet 42: 53–61.
23. Ragoczy T, Groudine M (2010) Getting connected in the globin interactome. Nat Genet 42: 16–17.
24. Zink D, Cremer T, Saffrich R, Fischer R, Trendelenburg MF, et al. (1998) Structure and dynamics of human chromosome territories in vivo. Hum Genet 102: 241–251.
25. Manders EMM, Kimura H, Cook PR (1999) Direct imaging of DNA in living cells reveals the dynamics of chromosome formation. J Cell Biol 144: 813–821.
26. Murray JM, Appleton PL, Swedlow JR, Waters JC (2007) Evaluating performance in three-dimensional fluorescence microscopy. J Microsc 228: 390–405.
27. Ronneberger O, Baddeley D, Scheipl F, Verveer PJ, Burkhardt H, et al. (2008) Spatial quantitative analysis of fluorescently labeled nuclear structures: Problems, methods, pitfalls. Chromosome Res 16: 523–562.
28. Bolte S, Cordelieres FP (2006) A guided tour into subcellular colocalization analysis in light microscopy. J Microsc 224: 213–232.
29. Yoshida M, Horinouchi S, Beppu T (1995) Trichostatin A and trapoxin: novel chemical probes for the role of histone acetylation in chromatin structure and function. Bioessays 17: 423–430.
30. Bazett-Jones DP, Li R, Fussner E, Nisman R, Dehghani H (2008) Elucidating chromatin and nuclear domain architecture with electron spectroscopic imaging. Chromosome Res 16: 397–412.
31. Manders EMM, Stap J, Brakenhoff GJ, van Driel R, Aten JA (1992) Dynamics of three-dimensional replication patterns during the S-phase, analysed by double labeling of DNA and confocal microscopy. J Cell Sci 103: 857–862.
32. Chuang CH, Carpenter AE, Fuchsova B, Johnson T, de Lanerolle P, et al. (2006) Long-range directional movement of an interphase chromosome site. Curr Biol 16: 825–831.
33. Edelmann P, Bornfleth H, Zink D, Cremer T, Cremer C (2001) Morphology and dynamics of chromosome territories in living cells. Biochim Biophys Acta 1551: M29–M40.
34. Chubb JR, Boyle S, Perry P, Bickmore WA (2002) Chromatin motion is constrained by association with nuclear compartments in human cells. Curr Biol 12: 439–445.
35. Zink D (2006) The temporal programme of DNA replication: new insights into old questions. Chromosoma 115: 273–287.
36. Albiez H, Cremer M, Tiberi C, Vecchio L, Schermelleh L, et al. (2006) Chromatin domains and the interchromatin compartment form structurally defined and functionally interacting nuclear networks. Chromosome Res 14: 707–733.
37. Heard E, Bickmore W (2007) The ins and outs of gene regulation and chromosome territory organisation. Curr Opin Cell Biol 19: 311–316.
38. Branco MR, Pombo A (2007) Chromosome organization: new facts, new models. Trends Cell Biol 17: 127–134.
39. Branco MR, Pombo A (2006) Intermingling of chromosome territories in interphase suggests role in translocation and transcription-dependent associations. PLoS Biol 4: 780–788.
40. Mahy NL, Perry PE, Bickmore WA (2002) Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. J Cell Biol 159: 753–763.
41. Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, et al. (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326: 289–293.
42. Ryba T, Hiratani I, Lu J, Itoh M, Kulik M, et al. (2010) Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res 2010 Jun;20(6): 761–70.
43. Botta M, Haider S, Leung IX, Lio P, Mozziconacci J (2010) Intra- and inter-chromosomal interactions correlate with CTCF binding genome wide. Mol Syst Biol 6: 426.
44. Baù D, Sanyal A, Lajoie BR, Capriotti E, Byron M, et al. (2011) The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat Struct Mol Biol 18: 107–114.
45. Jackson DA, Hassan AB, Errington RJ, Cook PR (1993) Visualization of focal sites of transcription within human nuclei. EMBO J 12: 1059–1065.
46. Branco MR, Branco T, Ramirez F, Pombo A (2008) Changes in chromosome organization during PHA-activation of resting human lymphocytes measured by cryo-FISH. Chromosome Res 16: 413–426.
47. Fussner E, Ching RW, Bazett-Jones DP (2011) Living without 30 nm chromatin fibers. Trends Biochem Sci 36: 1–6.
48. Strickfaden H, Zunhammer A, van Koningsbruggen S, Köhler D, Cremer T (2010) 4D chromatin dynamics in cycling cells: Theodor Boveri's hypotheses revisited. Nucleus 1: 284–297.
49. Maya-Mendoza A, Petermann E, Gillespie DA, Caldecott KW, Jackson DA (2007) Chk1 regulates the density of active replication origins during the vertebrate S phase. EMBO J 26: 2719–2731.
50. Chen S, Maya-Mendoza A, Zeng K, Tang CW, Sims PF, et al. (2009) Interaction with checkpoint kinase 1 modulates the recruitment of nucleophosmin to chromatin. J Proteome Res 8: 4693–4704.
51. Schindelin J (2008) "Fiji Is Just ImageJ". Proceedings of the 2nd ImageJ User & Developer Conference.
52. Iannuccelli E, Mompart F, Gellin J, Lahbib-Mansais Y, Yerle M, et al. (2010) NEMO: a tool for analyzing gene and chromosome territory distributions from 3D-FISH experiments. Bioinformatics 26: 696–697.