Sorbonne Université
École doctorale Complexité du Vivant
Institut Pasteur
Unité de Régulation Spatiale des Génomes
Institut Pasteur

Thèse de doctorat en Bioinformatique

Exploring the genomic complexity of bacterial infection in 3D

Cyril Matthey-Doret

Directeur de thèse: **Pr. Romain Koszul**

Présentée et soutenue publiquement le 1er Octobre 2021

Jury

Pr. X (PU)	Président
Dr. A (DR)	Rapporteur
Dr. B (MCU)	Rapporteur
Dr. C (CR)	Examinateur
Pr. Romain Koszul (PR)	Examinateur
Dr. D (CR)	Examinateur

Exploring the genomic complexity of bacterial infection in 3D

Thèse de doctorat en Bioinformatique

Rapporteurs: Dr. A et Dr. B

Examinatrice: Dr. C

Directeur de thèse: Pr. Romain Koszul

Sorbonne Université

École doctorale Complexité du Vivant

Unité de Régulation Spatiale des Génomes

Institut Pasteur

25-28 Rue du Docteur Roux

75724 Paris Cedex 15

Abstract

Numerous bacteria and viruses use cells from different species to ensure their proliferation. This mode of reproduction implies the pathogen must escape the host immune system and reprogram its metabolism to sustain its own needs. These changes are often detrimental to the host cell and cause pathologies or death. The intracellular bacteria which use this mode of operation have been the focus on many studies aiming to understand their "hijacking" mechanisms. Recent advannces in genomics have largely stimulated research in this field by offering the possibility to decipher the sequence of genes expressed during infection. Several intracellular bacteria secrete "effectors" into the host cytoplasm which interact with its protein and affect its genetic expression program. Recently, studies in *Legionella pneumophila*, an experimental model for intracellular bacteria, have showed it was able to change the epigenetic state of its host. This kind of modifications allow rapid physiological changes and are intimately linked to the spatial organisation of the gneome. This tridimensional organisation allows to regulate the activity of large regions by modifying their compaction, or to activate genes by forming long range interactions in the sequence.Here, we study the relationships between intracellular pathogens and their hosts through spatial regulation of the genome. We use the model species *Legionella pneumophila* and *Salmonella enterica* to explore structural changes taking place in the host chromosomes and their link with genetic expression.

Résumé

De nombreuses bactéries et virus utilisent les cellules d'autres espèces pour assurer leur prolifération. Ce mode de reproduction implique que le pathogène doit échapper au système immunitaire son hôte et reprogrammer son métabolisme pour subvenir à ses propres besoins. Ces changements s'opèrent souvent au détriment de la cellule hôte et causent des pathologies ou la mort. Les bactéries intracellulaires utilisent ce mode de fonctionnement et font l'objet de nombreuses études dans le but de comprendre leurs mécanismes de "piratages". Les récents progrès en génomique ont largement stimulés la recherche dans ce domaine en offrant la possibilité de déchiffrer la séquence des gènes exprimés durant l'infection. De nombreuses bactéries intracellulaires sécrètent des "effecteurs" dans le cytoplasme de leur hôte qui vont interagir avec ses protéines et affecter son programme d'expression génétique. Récemment des études dans la légionelle (*Legionella, pneumophila*), un modèle expérimental pour les bactéries intracellulaires, ont démontré qu'elle était capable de changer la régulation épigénétique de son hôte. Ce genre de modifications permettent des changement physiologiques rapides et sont intimement liés a l'organisation spatiale du génome. Cette organisation tridimensionnelle permet de réguler l'activité de large régions en modifiant leur compaction, ou d'activer des gènes en formant des interactions à longues distance. Ici, nous étudions les relations entre les pathogènes intracellulaires et leurs hôtes à travers la régulation spatiale du génome. Nous utilisons en particulier les modèles *Salmonella enterica* et *Legionella pneumophila* pour explorer les changement de structure qui surviennent dans les chromosomes de leurs hôtes de leurs lien avec l'expression génétique.

Acknowledgements

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Contents

List of Figures

List of Tables

I

Introduction

> *Flattery isn't the highest compliment – parasitism is.*
>
> — **Gregory Benford**
> Shipstar

In this first part, we introduce the complex relationships between hosts and their parasites. We also discuss the evolutionary implications of these associations. We then focus on two model systems for infection biology: *Salmonella enterica* and *Legionella pneumophila*. We then provide an overview of how recent advances in genomics have pushed the knowledge of these systems and their current limitations.

1 Host parasite interactions

A large number of organisms throughout the tree of life establish stable interactions with different species. These interactions are often classified according to their perceived impact on the fitness of their members. We traditionally talk about parasitism for interactions with one-way benefits, and mutualism when the interaction has a positive impact on all parties involved. Rather than a dichotomous classification, the difference between parasitism and mutualism is better viewed as a spectrum, depending on the Fitness cost and benefit of the relationship to the host (Fig. I.A).

Biological interactions are observed at different scales, from nanometer-scale virophages infecting giant viruses to fungi forming mycorrhyzal networks spanning several meters [1, 2] allowing exchange of nutrients with plants root systems. These interactions shape the evolutionary trajectories of the species involved and their genomic landscapes. These changes can sometimes result in drastic transitions in the organisms' lifestyle.

This can be the case for example with intracellular bacteria forming symbiosis with their host cells, known as endosymbionts. The *Wolbachia* genus is a famous example of endosymbiotic bacteria infecting arthropod species. These bacteria are reproductive parasites which can be transmitted vertically through infection of the host female's eggs [3]. Some *Wolbachia* have altered the reproductive capabilities of their sexual host species to reproduce asexually by *Parthenogenesis* [4]. This effectively removes all males from the host population, benefitting the bacterium which can only be transmitted through females. In some species, infection by *Wolbachia* has even become necessary for reproduction. While the bacterium takes advantage of its host reproduction, it also provides numerous advantages such as

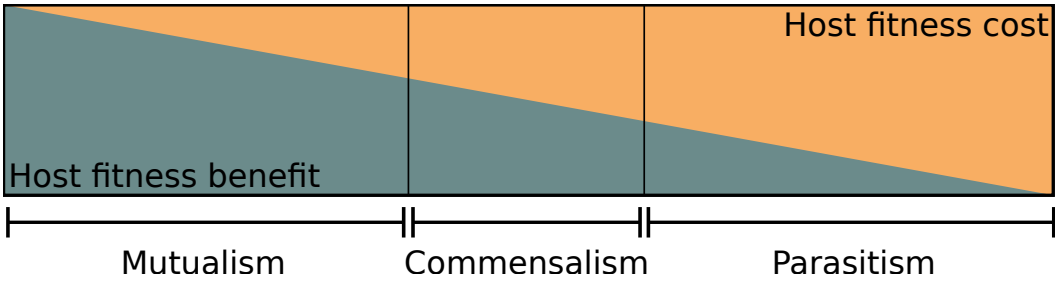


Fig. I.A: Parasitism - Mutualism spectrum: A spectrum of host Fitness cost underlies common terms used to described a biological interactions.

resistance to viruses in flies and mosquitoes [5, 6] and help with vitamin synthesis in bed bugs [7], illustrating the blurry line between parasitism and mutualism.

In this work, we focus on bacterial endosymbionts. Living directly inside of their host's cytoplasm, their genomic fate is most tightly linked to their host.

1.1 The evolutionary context of intracellular parasitism

Intracellular bacteria can either be facultative or obligatory endosymbionts. Obligatory endosymbionts can only replicate inside of their host cells. This is the case of several genera of obligate intracellular bacteria, such as *Rickettsia* or *Chlamydia*. These parasites are unable to reproduce outside of their host and become reliant on it for most metabolic pathways. The host cytoplasm being an isolated environment, obligate intracellulars have limited opportunity for recombination with other strains. Small populations of asexual organisms unable to recombine are at the mercy of *Muller's ratchet*, the progressive accumulation of mutations and loss of genetic material. They undergo a process known as genome reduction: Pathways provided by the host need no longer be encoded by the parasite and are therefore lost [8]. This process eventually leads to the parasite becoming completely reliant on its host for survival.

Facultative parasites bacteria opt for a different strategy, often with larger host ranges. These bacteria can complete their life cycle without the need for a host. They can reproduce in the extracellular space and be transmitted between different species. An analogy often used to describe the evolutionary dynamics of intracellular parasites with their hosts is the "arms race". Each organism evolves novel strategies (i.e. weapons) to improve its own fitness at the expense of the other. This is the case for intracellular bacteria such as *Legionella* or *Salmonella*, which can secrete a large arsenal of effector proteins into their host's cytoplasm. These proteins will manipulate host pathways to sustain the parasite's reproduction and protect it against host defenses. Many of these proteins are redundant in the sense that they interact with the same host proteins or pathways and can complement each other if one is defective [9].

Perfectly redundant genes should be subject low selective pressures, making them susceptible to genetic drift and therefore unstable [10]. It is therefore thought that the functions of redundant genes in intracellular bacteria have partial overlap, such as different affinity for certain substrates or the ability to function in different conditions or infection stages [9]. Selective pressure would therefore be applied

on these specificities. This is likely an important phenomenon for parasites with a broad host range or environments susceptible to changes.

Most intracellular bacteria incorporate genes from their hosts into their genome. Such genetic transfers are known as Horizontal Gene Transfer (HGT) and are a major contributor to bacterial genomes, with an estimated 80% of genes being ionvolved in HGT at some point in their history [11]. More recently, HGT from bacteria to eukaryotes have also been detected in eukaryotic genomes. Although they are much less frequent 0.04-6.49% of genes in microbial eukaryotes [12]), gene transfers from intracellular microorganisms to eukaryotic hosts are thought to have catalyzed major shifts in environmental niche. Examples are the terrestrial colonization of plants and extremophile eukaryotes such as sea ice diatoms which acquired ice binding proteins from prokaryotes [12].

All these exchanges illustrate the complex evolutionary dynamics of intracellular life; genetic material can be passed not only from the host to the parasite, but also between different endosymbionts and to the host.

1.2 Amoeba as a host model

Free living amoeba are ubiquitous unicellular organisms found in soil and various bodies of water, such as rivers, lakes [13] or even puddles [14]. They graze on bacterial biofilms, feeding on microorganisms by phagocytosis. This lifestyle exposes them to a large number of bacteria and viruses and they are host to many endosymbionts.

Amoeba offer a great experimental model, as many species are easy to grow in laboratory conditions and can be used for infection experiments. Despite their extensive use as an infection model, only a few species have high quality genome assemblies available, and the genomics of free living amoeba are still largely unknown. For example the *Acanthamoeba castellanii* genome has evidence for highly variable ploidy levels [15] and horizontally acquired genes [16]. These peculiar genomic features are likely important in their interactions with endosymbionts. Indeed, high ploidy levels have been proposed as a mean for asexual amoeba to escape Muller's ratchet through homologous recombination between haplotypes [15].

Similarly, the amoeba *Paulinella chromatophora* has photosynthetic organelles whose genome benefits from HGT from endosymbionts, as they counteract Muller's ratchet. This interesting observation, provides an interesting track to explain the conservation of horizontal gene transfers in the genome of free living amoeba *A. castellanii* [16].

Their long term coevolution with endosymbionts make free living amoeba an interesting model for evolutionary biology and ecology, but they are also an important organism for public health as they are the reservoir of human pathogens such as *Legionella pneumophila*. Besides, many free living amoeba have a biphasic life cycle, living as trophozoite to feed and reproduce, and transforming into cysts in harsher conditions. This encystation process makes them even more important from the standpoint of public health; intracellular bacteria infecting amoeba are able to survive water chlorination or antibiotic treatments by using the encysted amoebae as shelters.

1.3 *Legionella pneumophila*

L. pneumophila is an important model for studying intracellular bacteria and infects a range of 15 species of amoebae and ciliated protozoa in the wild [17]. It can also infect lung macrophages of humans and other mammalians. In humans, this can cause a severe pneumonia known as *Legionnaire's disease* [18]. Human to human transmission of *L. pneumophila* is extremely rare [19], making infection of macrophages an evolutionary dead-end for the bacterium. *L. pneumophila* is a major public health concern as it can contaminate water distribution systems and cause major outbreaks. The outbreak which lead to the identification of this bacterium and after which the bacteria was named happened at a convention of the American legion, Philadelphia, in 1976. This outbreak resulted in 182 cases, 29 of whom died. Since then, outbreaks are associated to *Legionella* every year with over 32,000 cases reported between 1995 and 2005 [20].

Unlike other bacteria on which phagocytic cells prey (Fig. I.Ba), when it is engulfed by a predatory cell, *Legionella* evades the lysosomal degradation route and survives in a special vacuole, the *Legionella* Containing Vacuole (LCV) (Fig. I.Bb). It does so using its type IV secretion system to secrete around 300 effector proteins into the host cytoplasm and rewire the host metabolic and signalling pathways. Many of those effectors contain eukaryotic domains and likely originate from inter-domain HGT [21]. Through their secretion, the bacterium is able to create a niche inside of the host cell with stable conditions and ample nutrients nutrients where it can proliferate.

1.3.1 Life cycle

L. pneumophila follows a biphasic life cycle. It can survive in the extracellular environment and thrives in fresh water. It can eith spread planktonically as a free living organism using its flagella to reach new hosts, or by associating with biofilms

[22, 23]. This extracellular phase is called the "transmissive form", as bacteria will search for new host cells but will not replicate [24]. In contrast, when entering a host cell, the bacterium enters the "replicative form". In that phase, the bacterium takes advantage of the abundant resources and nutrient available in the host cell to replicate as much as possible.

Switching between replicative and transmissive phases requires consequent morphogenetic and metabolic changes, mobilizing expression changes in almost half of the known genes [23]. Low nutrient and high stress conditions, cause *L. pneumophila* to enter transmissive phase, activating genes related to motility and virulence, such as its type IV secretion system. When entering the replicative phase, genes related to sugar and gluconate uptake and amino-acid catabolism are upregulated instead. The bacteria become acid resistant and replicate in the LCV until the nutrient pool is depleted.

Comparison of gene expression profiles between *L. pneumophila* grown *in vitro* in the absence of host, and *in vivo* with the amoeba *A. castellanii* revealed that changes associated with progression from exponential growth to stationary phase are similar to those observed between replicative and transmissive phases [25]. In *in vitro* cultures, stationary phase refers to the time when bacteria stop replicating for lack of nutrients. This suggests that the biphasic life cycle of *L. pneumophila* is governed by the presence of nutrient in the environment [26].

The master regulator underlying this switch is thought to be the Carbon storage regulator protein A (CsrA). CsrA is an RNA binding protein with over 400 target transcripts identified, including 40 effector proteins and genes related to virulence and motility. In replicative phase, CsrA binds its target transcripts to repress their translation. When nutrients are running low, *L. pneumophila* produces the alarmone (p)ppGpp, which triggers the expression of noncoding transcripts with strong affinity for CsrA. This prevents CsrA from binding its targets and enables the translation of virulence genes [27].

1.3.2 Host interactions

While inside the host, *L. pneumophila* consumes products from the host cell for energy production. It relies mainly on serine, threonine and other amino acids but can also scavenge carbohydrates such as gluconate [25]. Those nutrients are transferred from the host cytoplasm to the LCV by transporters on the LCV membrane [28]. The bacterium can increase the availability of nutrients in the host cell using its effector proteins. One example is the AnkB effector which can poly-ubiquitinate host proteins, causing their degradation by the host proteasome, resulting in amino

acids which can then be imported into the LCV and consumed [29]. Other effectors block host protein translation to increase the pool of free amino acid available for consumption by *L. pneumophila* [30].

The host trafficking system is hijacked to recruit mitochondria and Endoplasmic Reticulum (ER) membrane vesicles to the LCV. This is likely achieved by modulating the activity of host GTPases, such as Arf1, Sar1 and Rab1 [31]. Some *Legionella* effectors directly affect the host actin cytoskeleton, which is important in many cellular processes including vesicle trafficking [32, 33].

L. pneumophila also ensures successful infection by promoting host cell survival. The effector SdhA interferes with host cell apoptosis by inhibiting caspases [34]. All these interferences with host cell signalling pathways are likely bound to affect its expression programm but it was recently found that one of the effectors secreted by *L. pneumophila* directly affects the host epigenetic state. This effector, named RomA, is a histone methyltransferase which can alter the histone methylation state throughout the host genome and affects the expression of a large number of genes [35].

There is still much to learn about the interaction between *Legionella* effectors and its host regulation, but direct modification of nucleosomes reveals a new level of intimacy between bacterial endocymbions and their host. Besides, epigenetics and gene expression are tightly connected with spatial genome organization in eukaryotes. This gives a new promising angle to approach the study of host-pathogen interactions.

1.4 *Salmonella enterica*

Unlike *L. pneumophila*, *S. enterica* species are known to infects various birds, reptiles and mammals [36]. It is also a model for intracellular bacterial infections and a major human pathogen. *Salmonella* is a facultative intracellular parasite which can infect macrophages, dendritic, epithelial and microfold (M) cells. It is usually transmitted by ingestion of contaminated food and colonizes the gastrointestinal tract. *Salmonella* isolates are classified into 2,500 serovars based on their lipopolysaccharides and flagellar antigens. While most serovars, referred to as "non-typhoidal" cause salmonellosis, a self-limiting enteritis, "typhoidal" serovars are human restricted and cause a systemic disease known as typhoid fever [37].

Every year, it is estimated that there are 16.6 million cases of typhoid fever causing 600,000 deaths in the world, and 1.3 billion cases of acute gastroenteritis associated

with *Salmonella*, responsible for 3 million deaths [38]. Most of the current knowledge on *Salmonella* infection biology was built on the non-typhoidal serovar *S. enterica* subsp. *enterica* serovar Typhimurium [37].

Much like *Legionella*, when *Salmonella* enters the host cell, it is engulfed into a *Salmonella* Containing Vacuole (SCV) and secretes effector proteins into the host cytoplasm. This is done via two independent type 3 secretion systems (T3SS) named SPI1 and SPI2. These two systems are encoded by and named after the *Salmonella pathogenicity island*, which *Salmonella* likely acquired through horizontal gene transfer.

The mechanisms employed by *Salmonella* to infect host cells are similar to *Legionella*. For example, they encode effectors that also activate the host gene Arf1 to promote bacterial uptake and actin polymerization [37].

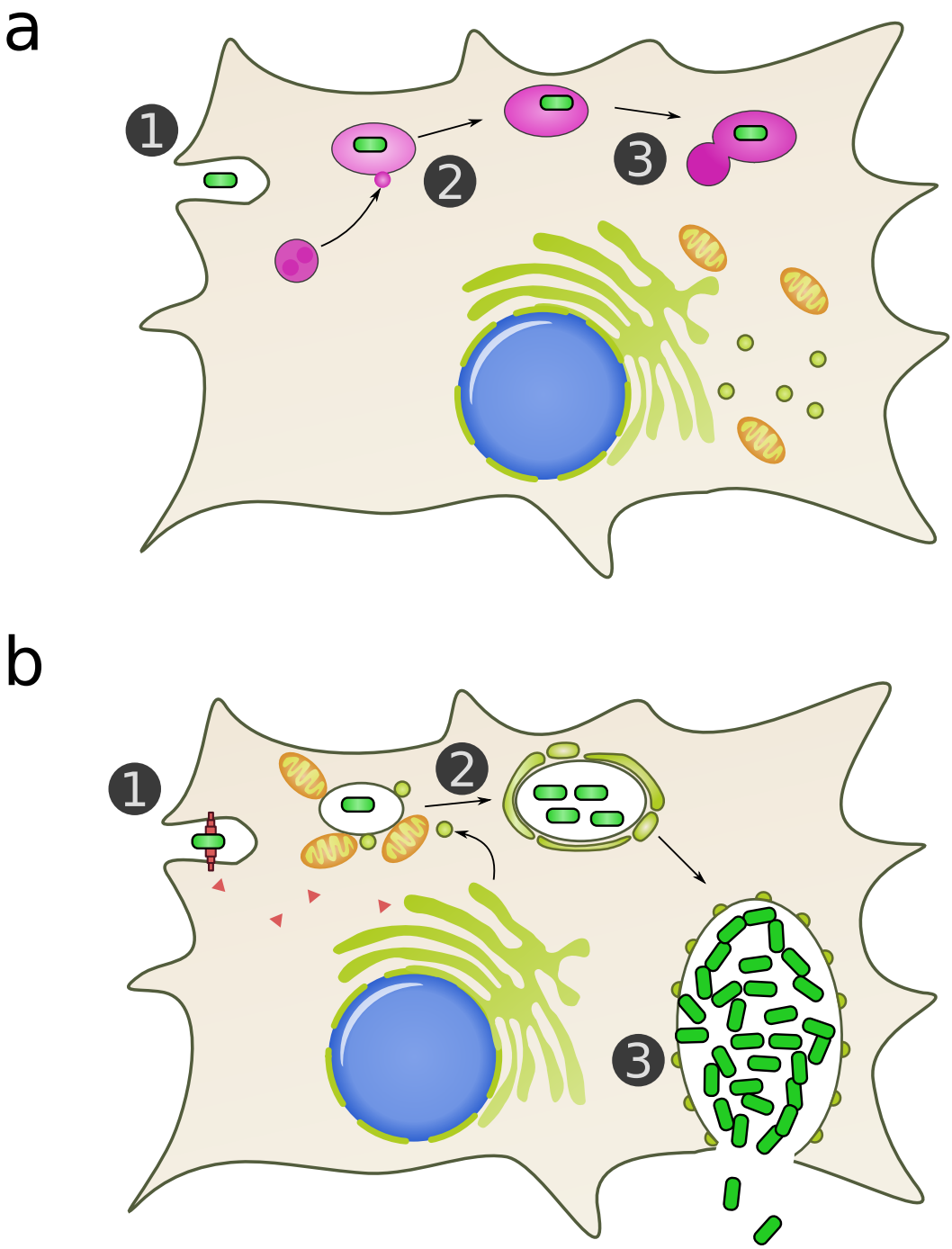


Fig. I.B: Infection by *Legionella*: **a**: Non infectious bacteria (green) are phagocytized by amoeba or macrophages (1), the early and late endosomes (pink) acidify the compartment (2), and it finally merges with the lysozome (3) where the bacteria is degraded. **b** Upon phagocytosis, *Legionella* uses its type IV secretion system to secrete effector proteins (red triangles) into the cytoplasm and evades the endosome route (1). Instead, it stays in a "*Legionella* containing vesicle" (LCV) and recruits mitochondria (orange) and endoplasmic reticulum-derived vescicles (yellow) (2). The bacteria keeps replicating in the LCV until it bursts out and infects other cells.

2 Infection through the lense of genomics

The toolset to detect and investigate bacterial infection traditionally included biochemical assays and microscopy. The recent technological advances in DNA sequencing have spurred a rapid extension of this toolset with NGS-derived methods. Here we introduce the different ways genomics can provide biological insights into the biology of bacterial pathogens.

2.1 Pathogen characterization

The most fundamental task related to infection in biomedical research is to detect the presence of infectious agents and characterize them. This allows to test patients presenting suspicious symptoms for the presence of known pathogens, or determine the pathogenicity of a particular strain.

Genotyping was traditionally achieved using molecular biology techniques, such as Restriction Fragment Length Polymorphism (RFLP) or Pulsed Field Electrophoresis Gel (PFGE) [39]. These techniques rely on the negative charge of the DNA molecules. When put in a gel submitted to an electromagnetic field, these molecules will migrate along the electrical current. The migration distance is proportional to the size of molecules. After migration is complete, the gel can be treated with chemical markers to highlight the location of DNA molecules. This will reveal discrete bands of similar-length DNA fragments. Together these bands form a bar-code which can be interpreted by the scientist to draw conclusion about the number and size of these fragments. In the case of RFLP, the genome is prealably digested by *Restriction enzymes*. The digestion will result in a series of fragments whose lengths can be seen on the gel. Bacterial genotypes wil have different mutations which will affect the digestion pattern and resulting barcode on the gel.

While these methods work well to determine differences between alleles, they do not inform us on the actual DNA sequence involved. The advent of DNA sequencing made it possible to directly link phenotype with associated sequences of nucleotides. Whole Genome Sequencing (WGS) provides accurate information on an organism's genotype, down to down to the Single Nucleotide Polymorphism (SNP), allowing to define genotypes at a finer scale. The main shortcoming of WGS is its higher cost than other genotyping techniques, but the recent plummeting of sequencing costs

have made it releatively affordable. These advantages have made WGS a popular approach in clinical settings.

2.2 Genomics to probe homeostasis

When host cells are exposed to or infected by a pathogen, their homeostatic state is disrupted. This disruption is a combination of alterations caused by the pathogen to colonize the host cell and host-triggered immune reactions to improve its survival. These two components can usually be unentangled in infection experiments by using a disabled pathogen. The pathogen will still harbour the antigens triggering host reactions, but will be unable to cause any harms. One can then deduce the pathogen-caused disruptions by comparing the infection results from a standard and disabled pathogen.

Multiple levels of regulation are affected upon infection, from signalling to epigenetic modifications [40], and over the years, a vast arsenal of NGS techniques have been developed to read these regulatory states.

The most frequently used feature is gene expression. The transcribed RNAs present in the sample can be reverse-transcribed into cDNA and sequenced and the relative abundance of each gene's transcript allows to quantify the expression of the whole genome, known as the transcriptome. Transcriptomes can then be compared between different conditions to find out which genes undergo perturbations during infection.

Many levels of regulation allow to fine tune gene expression in eukaryote (Fig. I.C). Regulatory elements can be directly encoded by the sequence and trigger the recruitment of protein complexes to fine tune gene expression. Epigenetic changes, in the form of chemical modification of histone proteins offer yet another way to regulate gene expression in eukaryotes. The amount of these epigenetic marks can be measured along the genome using another NGS-derived technique known as Chromatin Immuno-Precipitation Sequencing (ChIPseq). In ChIPseq, the chromatin sample is crosslinked with formaldehyde to generate covalent bonds between proteins and DNA. The sample is then sonicated to break the DNA into smaller fragments. Beads coated with specific antibodies against a protein of interest (e.g. an epigenetic mark) are then added and the beads are then precipitated to retrieve them. The crosslinked is then reversed and the DNA fragments purified. This allows to retrieve all genomic regions that were bound to the protein of interest.

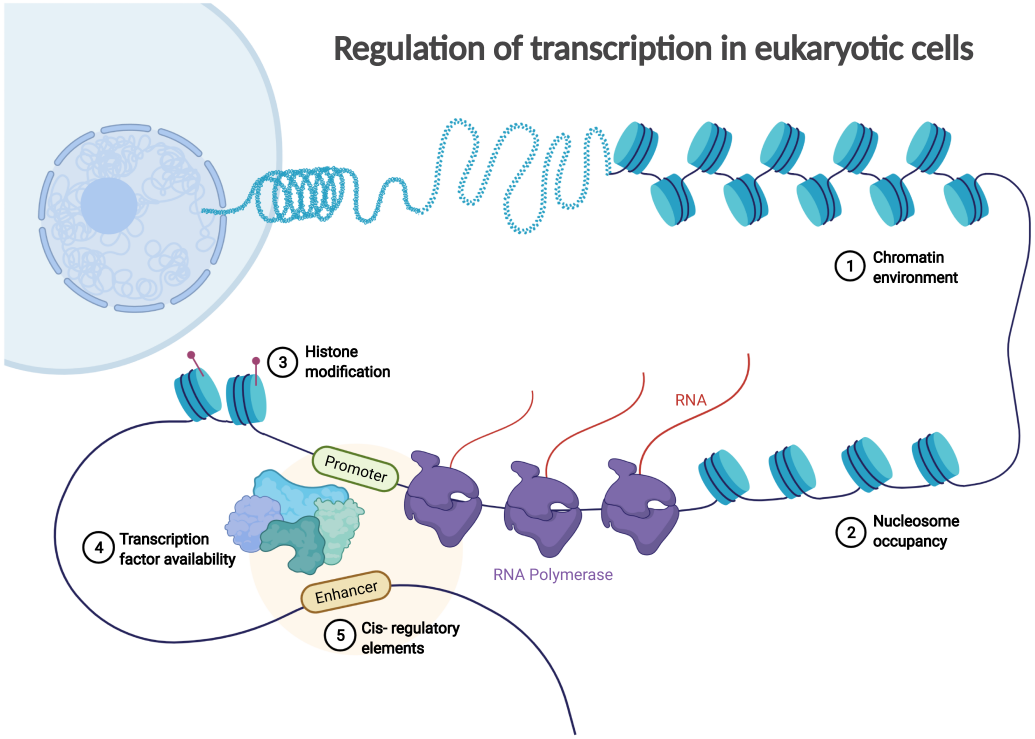


Fig. I.C: Regulation of transcription in eukaryotic cells. Visual summary of the different levels at which transcription can be regulated. At the largest scale (1), the chromatin environment can form structures affecting transcription. The open space between nucleosome can also affect accessibility of protein complexes to gene sequences (2). Chemical modifications on histone proteins form an epigenetic code defining the recruitment of transcriptional complexes on the genome (3). The availablity of those factors (4) and the proximity of regulatory sequences such as enhancers (5) provide another level of transciptional regulation. Reprinted from "Regulation of Transcription in Eukaryotic Cells", by BioRender.com (2020). Retrieved from https://app.biorender.com/biorender-templates

2.3 Capturing chromosome conformation

Although DNA is a linear (or circular) molecule, it can fold back on itself and form three-dimensional structures which have several benefits compared to a linear structure. These benefit include compactness: For example, the human chromosome 1 consists of 250 millions nucleotides each spaced by 0.34nm [41]. If straightened, the chromosome would be 85mm long, but the whole genome fits in a nucleus where the diameter is around $10\mu m$. Another key feature of genome folding is the regulation of gene expression through higher order structures. Compacting large regions of the genome by spreading of *Heterochromatin* allows to downregulate their activity. Smaller scale structures allow to fine tune gene regulation more locally. For example, *Chromatin* loops can bring enhancer and promoters in close spatial proximity even if they would otherwise be far apart on the linear sequence. Compact chromatin domain also form local neighbourhoods where different loci are in close spatial proximity, while loci in distinct domains are isolated from each other. All these levels of spatial regulation are important to understand the coordination of the gene expression programme with other celluar processes.

2.3.1 3C technologies

The use of genomics to investigate the three-dimensional organisation of the genome started with the invention of Chromosome Conformation Capture (3C) [42]. This technique allowed to measure the frequency of physical interactions a pair of loci (Fig. I.D). This is done by crosslinking the genome with formaldehyde, which forms stable bonds between DNA and proteins, and subsequently digesting the genome with a restriction enzyme. Genomic regions closer in space will be crosslinked together more frequently, resulting in pairs or complexes of chromatin fragments from different genomic regions that were spatially interacting. The digested genome is then religated. Loci which are closer in space are religated more often with each other in the population of cells. The crosslink is then reverted and primers from two regions of interest are added to perform qPCR. This allows to measure the quantity of religated products containing the two loci of interest. Thus, 3C allows to quantify the spatial proximity between two known genomic loci.

Since then, many derivatives of the 3C technique have been developed. The most significant improvement was brought by Hi-C (Fig. I.D, right). This method shares several steps with 3C, the main difference being that next generation sequencing is used instead of qPCR. This allows to quantify the interaction frequency of all versus all loci in the genome instead of using specific primers for a pair of locus. In Hi-C, fragment ends are filled with biotin prior to religation [43]. The religation products

are then pulled-down using streptavidin beads (which have high affinity for biotin), which allows to specifically retain digested and religated products for sequencing.

The information generated by Hi-C experiments is a list of contacts between all pairs of restriction fragments in the genome. These contacts are most commonly visualized using contact maps (Fig. I.Ea), which are two entry tables represented as color-coded matrices. The color of each value in the matrix is proportional to its value and reflects the interaction frequency between the associated pair of fragments. Those contact maps are an indirect representation of the tri-dimensional folding of chromosomes and are rich in information.

The various folding structures formed by chromatin are created by DNA binding proteins. The most common example is the CCCTC-binding binding factor (CTCF), a transcription factor that also acts as an "architectural protein" structuring chromatin. Molecular motors such as cohesin and other Structural Maintenance of Chromosomes (SMC) family proteins slide along DNA to operate various roles. When cohesin is loaded onto the chromosome, it can extrude two strands of DNA in opposite directions through it's ring-shaped structure, a process known as loop extrusion [44]. When cohesin encounters a roadblock protein such as CTCF the extrusion stops, forming a chromatin loop and maintaining contact between the two DNA strands. Depending on the location of those roadblocks, this can form stable interactions between distant genomic regions.

Each spatial structure formed by chromatin is reflected on the contact map as a distinct pattern (Fig. I.Eb). At the largest scale, chromosomes are isolated from each other in the nucleus, occupying distinct "chromosome territories". This is reflected as dark squares along the diagonal of the whole genome contact map as each chromosome interacts more with itself than any other chromosome (Fig. I.Ea). Within each chromosome chromatin forms "insulation domains", known as Topologically Asscociating Domain (TAD)s in mammals. Genes sharing the same domain are in close proximity, while being isolated from genes in neighbouring domains . Genes within the same domain also tend to be co-regulated [45]. Domains form dark squares along the diagonal of a chromosome, due to the enriched intra-domain interactions at the expense of inter-domain contacts (Fig. I.Eb, bottom). At a finer scale, chromatin loops are visible on contact maps as dots away from the diagonal (Fig. I.Eb, middle). The coordinates of those dots correspond to the genomic positions of roadblocks which stopped the extrusion process.

In many eukaryotes, chromatin is also segregated into active and inactive compartments, commonly known as "euchromatin" and "heterochromatin" or A/B compartments. The A (active) compartment has higher GC content, gene density and gene expression than its counterpart [43]. A and B compartments also occupy separate

spaces in the nucleus, whereas the A compartment is located towards the middle of the nucleus, the B compartment is relegated to the nuclear periphery and associated with lamina domains [46]. This spatial segregation results in preferential physical interactions within the same compartment, which are reflected on Hi-C contact maps by a plaid-like pattern.

2.3.2 Analysis of chromosome contact maps

The most visible element on any chromosome contact map is the diagonal gradient reflecting the power-law relationship between genomic distance and contact frequency. This is often called the distance-decay function or $P(s)$ where s means genomic distance and P probability of contacts. The slope of the $P(s)$ in itself holds information on the relative contribution of short range and long range contacts in the chromosome, which is linked to chromosome compaction.

Due to the high intensity of this gradient, other patterns of biological relevance are often obscured on the contact map. A common preprocessing step to account for it is to apply an observed over expected (o/e) normalization of the Hi-C map, where each pixel is divided by the average of its diagonal (Fig. I.Fa). Lower intensity patterns such as domains, compartments and chromatin loop are easier to perceive on the resulting map.

After o/e normalization, the compartment signal is generally the most salient feature on the contact maps. It can be extracted using Principal Component Analysis (PCA) on the normalized contact map [47] and retrieving the first few eigenvectors (i.e. principal component) explaining the most variance (Fig. I.Fb). In some cases where compartment signal is weak (e.g. noisy datasets), the compartment signal may not be contained in the first eigenvector. A robust approach is to select the eigenvector with the strongest absolute correlation to an external signal known to be associated with active chromatin, such as gene expression or GC content [48] (Fig. I.Fc). The sign of the eigenvector is arbitrary, and it must be "phased" with the feature by flipping its sign to ensure a positive correlation with the said feature. In the phased vector, regions in the A compartment will contain positive values and *vice versa*. The positions at which the sign changes are boundaries between different compartments (Fig. I.Fd).

Many genomes are segmented into TADs containing frequently interacting loci, and often co-regulated genes. Regions in separate TADs are insulated from each other, and the strength of this insulation can be quantified using an "insulation score". The insulation score of a given region is defined as the intensity of contacts across that region (upstream with downstream) [49] and can be represented as a numerical

track along the genome. Improved metrics such as the relative insulation score have since been developed to successfully detect TADs [50]. The relative insulation score (RI) at a locus *s* between bins *k* and *k+1* with a predetermined window size *w* is defined as:

$$U(w,s) = \sum_{i=-w}^{-1} \sum_{j=0}^{i+1} M_{k+i,k+j} \tag{2.1}$$

$$D(w,s) = \sum_{i=-w+1}^{0} \sum_{j=1}^{w+i} M_{k+i,k+j} \tag{2.2}$$

$$B(w,s) = \sum_{i=1}^{w} \sum_{j=i+1}^{w+1} M_{k+i,k+j} \tag{2.3}$$

$$RI(w,s) = \frac{U(w,s) + D(w,s) - B(w,s)}{U(w,s) + D(w,s) + B(w,s)} \tag{2.4}$$

Where U and D are contacts in the upstream and downstream regions, respectively and B are the contacts between U and D. This can be visually represented as a triangle sliding along the diagonal of the Hi-C matrix (Fig. I.G). By contrast the original insulation score consisted only in computing B (Eq. 2.3).

At a smaller scale than TADs, chromatin loops contain valuable information about interactions between regulatory elements. Some tools detect these loops by searching for local enrichment of contacts, appearing as dots away from the diagonal [51, 52]. However, current loop detection algorithms suffer from low detection rates (recall). As such, an alternative approach is to focus on a set of genomic intervals of interest (e.g. binding sites of a transcription factor) and compute a window average of all pairs of intervals. The resulting average, often called pile-up, can be used to visualize the presence of chromatin loops between regions, or be compared between mutants [53].

Quantifying contact changes is useful in the study of many biological processes, such as differentiation or cell cycle progression. In that regard, the most global comparison, is to compute a similarity metric between pairs of samples. This is also useful as quality control, to estimate technical (replicates) and biological (conditions) variability. Different comparison metrics have been used, such as the sum of differences between Hi-C matrices [54], correlation coefficients [55] or distance between the matrix eigenvectors [56].

Rather than computing a single metric for each sample, most applications of Hi-C require the identification of regions where the chromatin behaviour changes. Several

methods aiming to achieve this are adapted from existing count-based algorithms designed for RNA-seq [57–59]. In this analogy, they consider each bin of the genome as a "gene" and their contacts as an expression count. A discrete probability distribution is then fitted to the bin counts and used to identify bins with significant contact changes consistant across replicates. This approach relies on solid statistical grounds, but it often does not address the question at hand. Often times, when analysing Hi-C data, we are interested in finding specific structures appearing or disappearing rather than a simple contact change at a region. The development of methods to discover relevant changes in chromatin conformmation is still an active area of research.

2.4 Combining layers of biological informations

The central dogma of biology - "DNA makes RNA and RNA makes protein" - describes a linear set of reactions carrying the flow of information in living organisms. It is now known that these reactions by themselves are hardly sufficient to explain the complexity of biological processes. The fine tuning required for proper regulation is achieved through feedback loops and cross-talk between the different types of molecules (Fig. I.H). Common examples are methylation of DNA by proteins to reduce gene expression [60], noncoding RNAs recruiting proteins to repress transcription [61] or directly repressing translation by preventing ribosome binding [62, 63].

There is now a growing area of research focusing on the development of methods that combine these layers of information. They aim to gain an integrative view of biology to better model the behaviour of molecular networks. This is done by combining "omics" datasets measuring various biomolecules, such as genetic mutations, gene expression, protein binding, histone modifications or protein abundance.

One of the main challenges is to find efficient ways to combine these informations to extract meaningful biological information. More often than not, they are analysed separately to find regions of deregulation common to the different layers. But there have already been attempts at fully integrating these levels of informations

Another challenge is the difficulty to combine different datasets due to technical heterogeneities or biological differences such as different strains or experimental conditions.

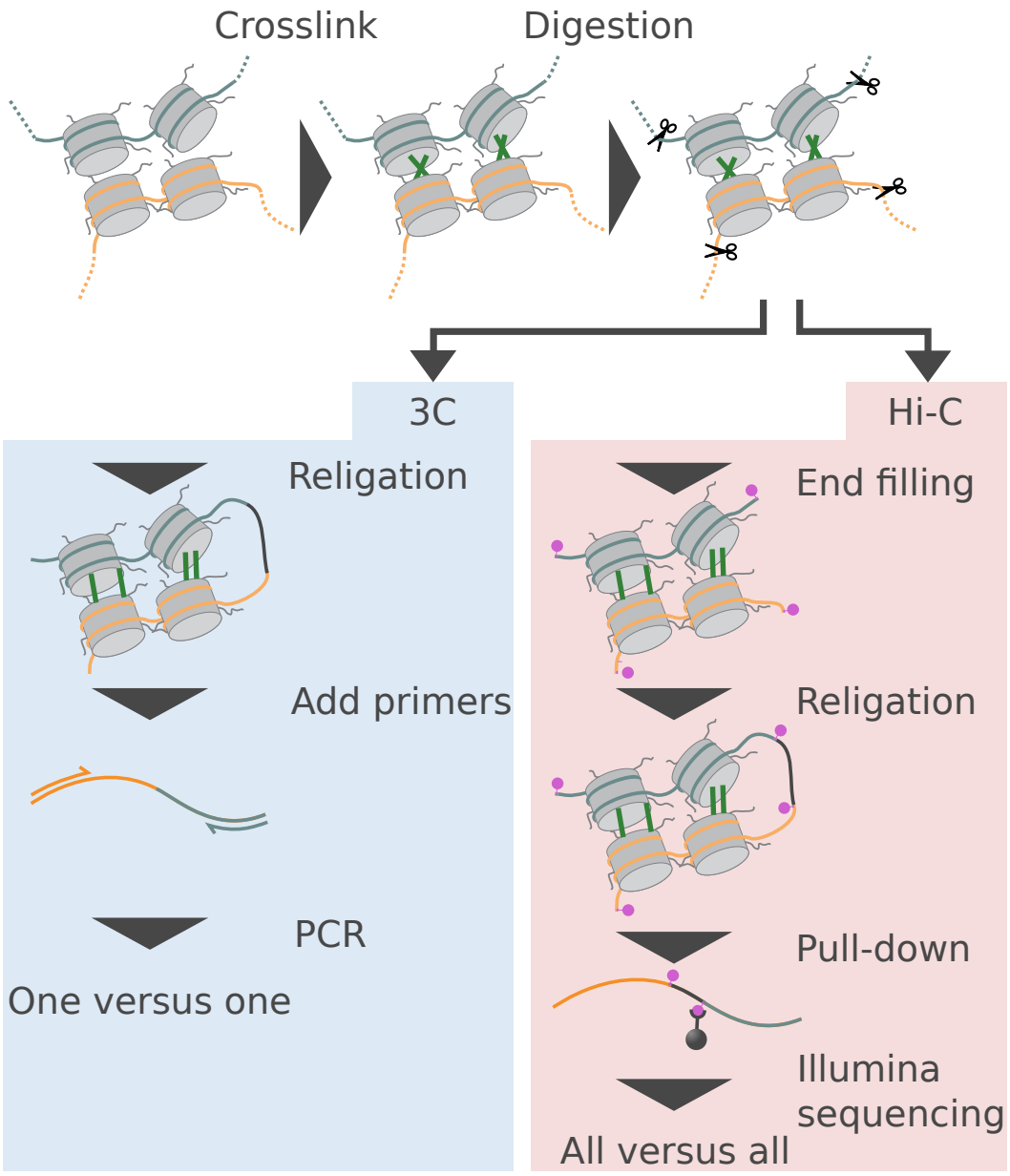


Fig. I.D: Chromosome conformation capture protocol: Chromosome conformation capture protocol share common steps (top): The chromatin is first crosslinked to form covalent DNA-protein bonds and then digested using a restriction enzyme. The Hi-C protocol subsequently differs from the original 3C protocol. In 3C (left), fragments are religated, the crosslinked is reversed and specific primers are added to amplify a pair of known loci. This allows to quantify interaction between 2 loci. In Hi-C (right), the fragments ends are filled with biotinylated nucleotides (pink), religated and the crosslinked is reversed. Streptavidin beads are then used to pull down religation products which are then sequenced.

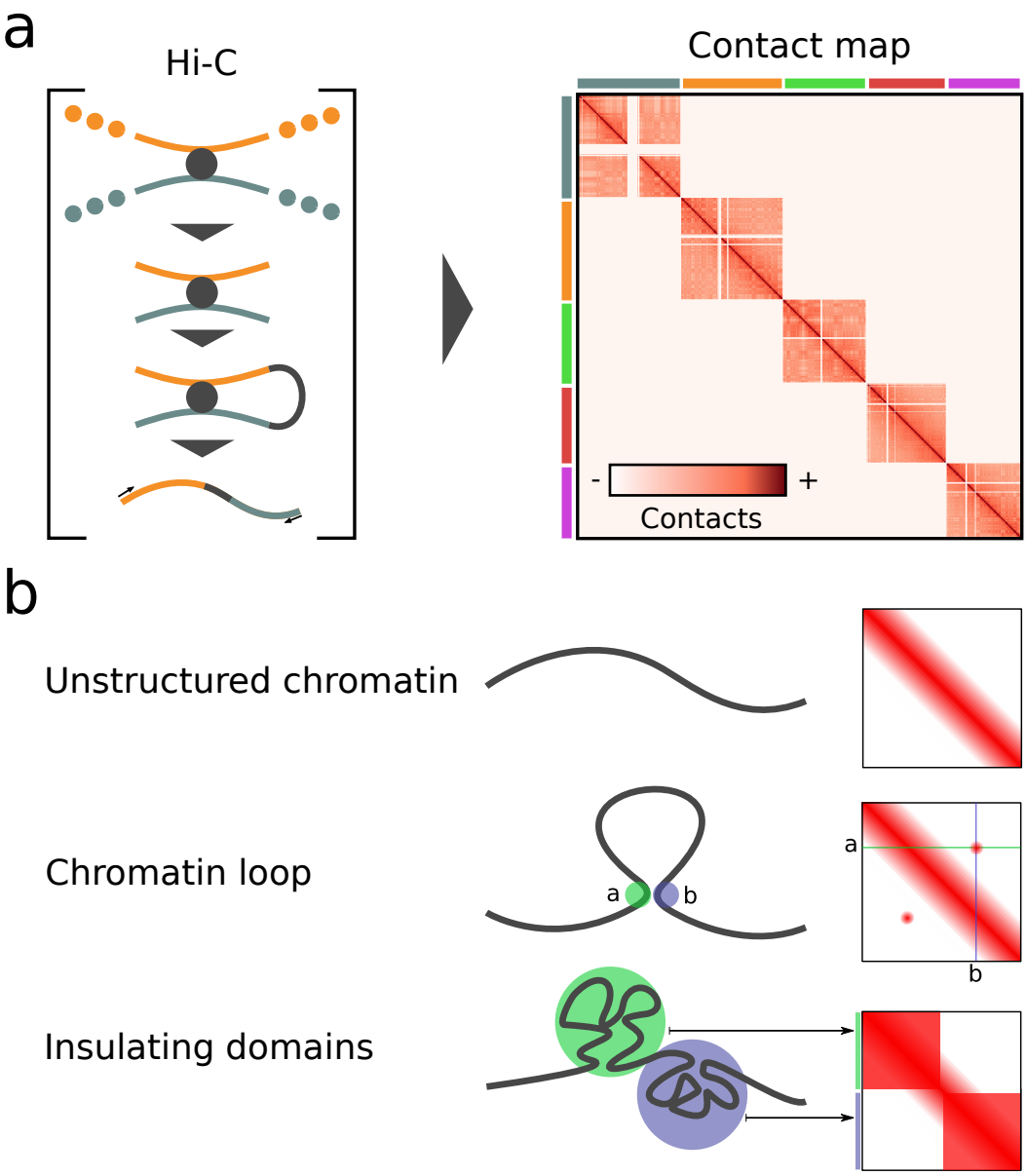


Fig. I.E: Interpretation of Hi-C contact maps: **a**: The Hi-C protocol (left) generates millions of read pairs representing contacts between genomic loci in a population of cells. Those contacts can be stored into an all-versus-all contact matrix (right) averaging all contacts in the population. Each chromosome in the matrix forms a square of strong self-interactions along the diagonal due to chromosomal territories. **b:** Within each chromosomal map, different contact patterns reflect specific conformations (right). The main feature of a contact map is the diagonal gradient (top) caused by the contact decay according to genomic distances. Chromatin loops between two anchor loci are visible as dots away from the diagonal (middle). Insulation domains form squares along the diagonal of a chromosome where loci within the same domain interact strongly, but interactions between domains are depleted

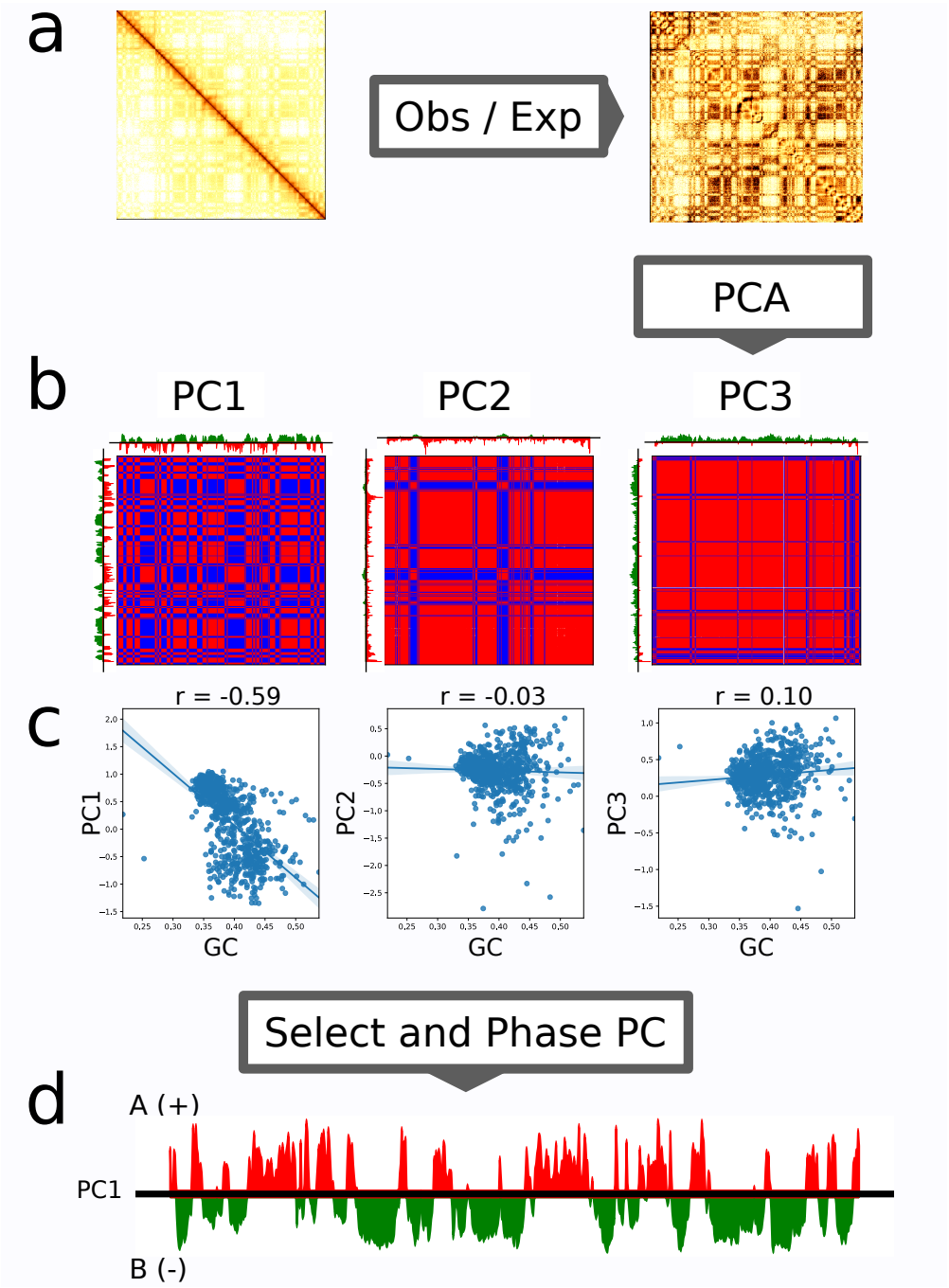


Fig. I.F: Representation and analysis of chromatin compartments in Hi-C. **a:** observed over expected (o/e) normalization is applied to the balanced contact map to remove the distance-decay gradient. Higher frequency of interactions within the same compartment result in a plaid-like pattern on chromosome contact maps. **b:** PCA is applied to the o/e normalized contact map and the first few principal coponents (PC) are retained. For visualization, each PC is shown alongside with its outer, yielding the rank-1 reconstruction of the contact map. The outer product matrix is binarized (negative=blue, positive=red) to show the compartmentalization. **c:** The correlation of each PC with GC content is computed, to select the PC with the highest absolute correlation. **d:** The sign of PCs being meaningless, the selected PC is phased (by changing its sign in case of negative correlation) to ensure positive values represent A compartment.

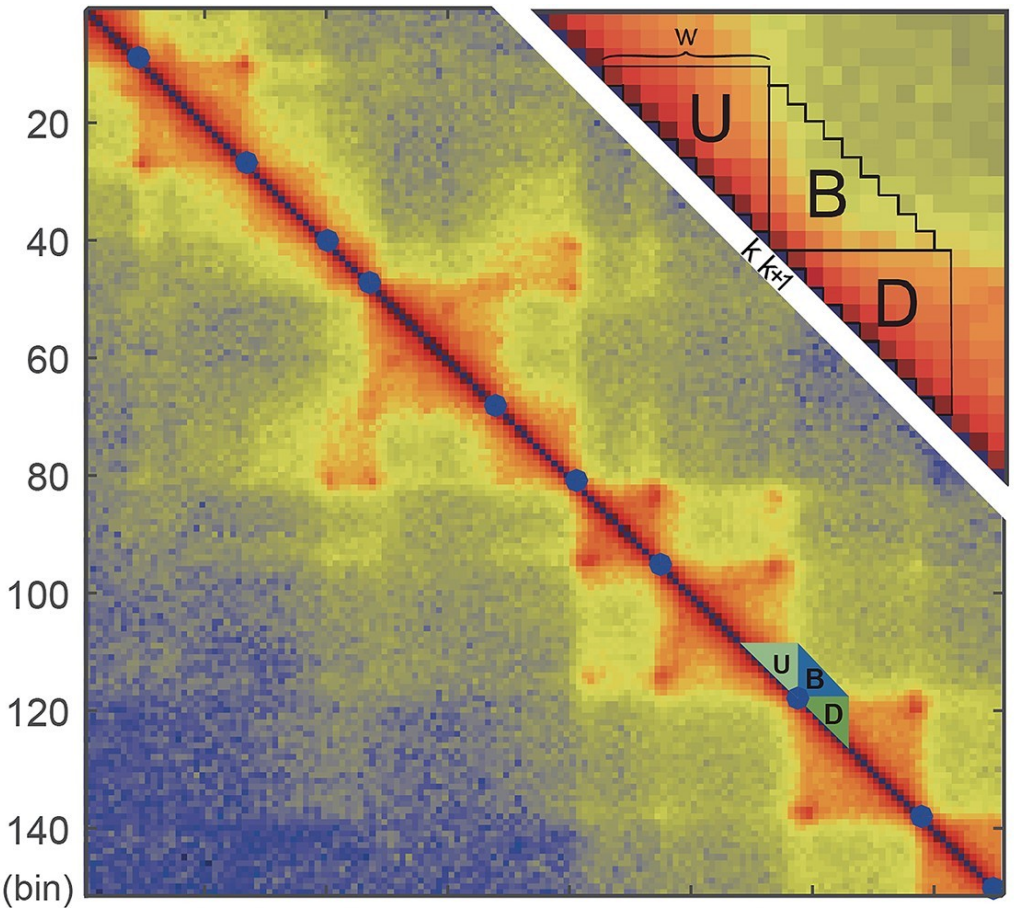


Fig. I.G: Visual illustration of the relative insulation score. Computing the relative insulation score at bin k involves computing the average interactions between upstream (U) and downstream (D) regions, denoted as B, as well as the average contacts within U and B. The key parameter when computing insulation is the window size (w), determining the size of the U, B and D. Figure adapted from [50]

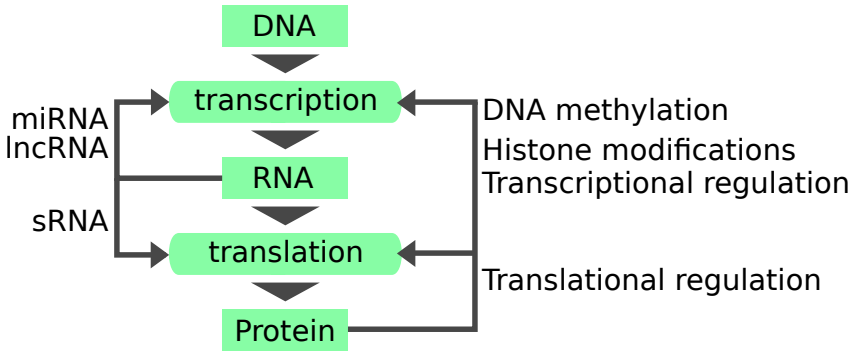


Fig. I.H: Central dogma of molecular biology. Products and reactions from the central dogma are shown in green, with grey arrows showing some of the regulatory interactions between the different biomolecules.

3 The importance of genome assembly

Most of the genomic techniques presented before require a complete reference genome as downstream analyses will rely on the relative position of different biological elements on the genome sequence to draw biological conclusions.

A good phylogenetic representation of sequenced genomes is also crucial for comparative analyses. This allows for example to identify recent HGT events. Here we describe in more detail the process of genome assembly and its relevance to infection genomics.

3.1 From contigs to chromosomes

Genome assembly consists in reconstructing the linear sequence of the genome from the readings of DNA sequencing technologies. Although the final assembly depends on the quality of these readings, the algorithms used to combine their information are also crucial.

In the early days of genome sequencing, the Sanger method was used to read DNA sequences [64]. Sanger is a low throughput, but highly accurate sequencing method. This technology allowed to unveil the complete genome sequences of viruses [65, 66] and the yeast genome [67]. A common practice at the time, was to clone a small genomic region of ~10kb into a plasmid, and fragment it [68]. The resulting fragments were sequenced and the sequencing readout, in the form of gels, had to be deciphered by scientists, one nucleotide at a time. The cloned region was then assembled manually by searching for overlaps between fragments.

Further technological improvements allowed to automate the sequencing process to tackle the sequencing of larger eukaryotic genomes. Early genome sequencing projects were performed using laborious and costly experimental methods, such as Bacterial Artificial Chromosome (BAC), which involved cloning long overlapping pieces of DNA of the genome into bacteria. These pieces were then experimentally amplified and sequenced in parallel. Ovelapping ends from each of those sequences had to be aligned to recover the entire chromosome sequence. The first genome sequencing projects were sizable undertakings requiring the collaboration of many research groups throughout the world [67, 69, 70], but technological advancements

progressively reduced the cost and time required. A decisive change was the development of shotgun sequencing [71], which involves randomly sequencing regions to cover the entire genome.

With the advent of Next Generation Sequencing (NGS), shotgun sequencing became the standard for whole genome sequencing. NGS has much higher throughput than Sanger sequencing, allowing to sequence megabases of DNA very quickly. However, it can only read short sequences at a time, referred to as *Read*s. Classic overlap-based genome assembly algorithms used in previous sequencing projects could not scale to such large numbers of short reads. This called for the development of more efficient genome assembly algorithms.

The goal of an assembler is to generate a highly contiguous genome sequence from a large number of short reads. Early algorithms computed pairwise alignments between all reads to build an overlap graph (Fig. I.Ia). The genome could then be assembled by finding th e Hamiltonian path of the graph, which passes once through every node. However, finding this approach is computationally expensive and cannot be used with high sequencing throughput [72]. This lead to the development of de Bruijn-based assembly algorithms, which many most modern still use [73, 74]. de Bruijn assemblers split reads into short *K-mer*s which they use to generate a de Bruijn graph. In these graphs, k-mer sequences represent edges, and the overlap between adjacent k-mers within reads are the nodes (Fig. I.Ib). To assemble a genome, assemblers need to find the Eulerian path, which passes through every edge once. However this is often not possible because of repeated sequences in the genome, sequencing errors and haplotypes [75]. Whenever a repeated sequence is longer than the read itself, the graph can not be solved and heuristics have to be used. Rather than a single fully resolved genome, the resulting assemblies usually have a relatively high number of independent pieces called *Contig*s.

Third generation sequencing partially alleviates this issue by generating long albeit less accurate reads. Read lengths up to hundreds of thousands of basepairs can be generated, which allows to span most repeated regions. Recently, these technologies were used to generate telomere-to-telomere assemblies of several human chromosomes [76, 77]. Third generation sequencing techniques still suffer from their lower base calling accuracy resulting in assemblies with high point error rates (>10%) and indels [78, 79]. To remove these errors, some methods have been developed to correct long reads before assembly, either by correcting long reads among themselves [80], or using a separate set of short accurate reads to erase sequencing errors in long reads [81]. Most long reads correction tools are also unable to differentiate between SNPs and sequencing errors, which result in the loss of haplotype information and prevents the generation of haplotype-resolved assemblies. Some long read correction methods have recently been developed to preserve haplotypes information

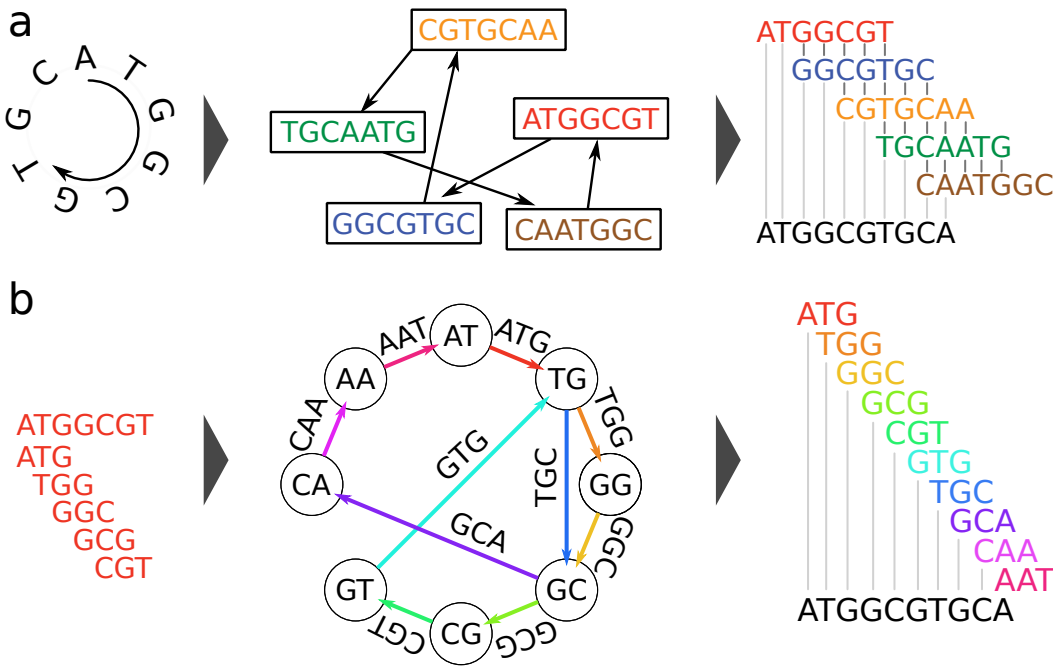


Fig. I.1: Graphs in genome assembly: A small circular genome is sequenced and the resulting reads are shown in color **a**: Early assembling techniques computed all pairwise alignments between reads to represent them as nodes in an overlap graph and their overlaps as edges. The genome sequence can be retrieved by finding a path going through each read exactly once. **b**: Modern assemblers first split reads into their constituent k-mers and represent the k-mers as edges in a de Bruijn graph where nodes are the k-1 overlap between two k-mers located in the same read. The path going through each edge once is computed to solve the graph. K-mers are extracted from the edges visited to retrieve the genome sequence. Adapted from [72]

[82]. One major drawback of read correction methods is their high computational cost, as they require to align high number of reads to each others. An alternative strategy is to use the uncorrected reads to assemble the genome and perform errror correction directly on the assembly, a process known as *Polishing*. Traditional short read polishers work by aligning short reads to the assembly and replacing each position of the assembly by the consensus of short reads [83]. Additionally, they can correct larger scale misassemblies such as indels by using the pair-end information and alignment discrepancies [84]. Some polishers have obtained better polishing accuracy by combining the information in short and long reads [85].

More recently the emergence of specialized technologies aimed at scaffolding have allowed to generate even more continuous and correct genomes at reduced costs. One example is the recent rebirth of optical mapping to introduce fluorescent probes into chromosomes at specific sites [86]. The order of these probes and their relative distance form barcodes which can then be used to scaffold genome assemblies, reorder and merge contigs. This is often combined with Hi-C to generate highly continuous assemblies even in the presence of repeated sequences.

A growing number of genome assemblies combine several of these different technologies to bring the number of scaffolds as close as possible to the real number of chromosomes (Fig. I.J).

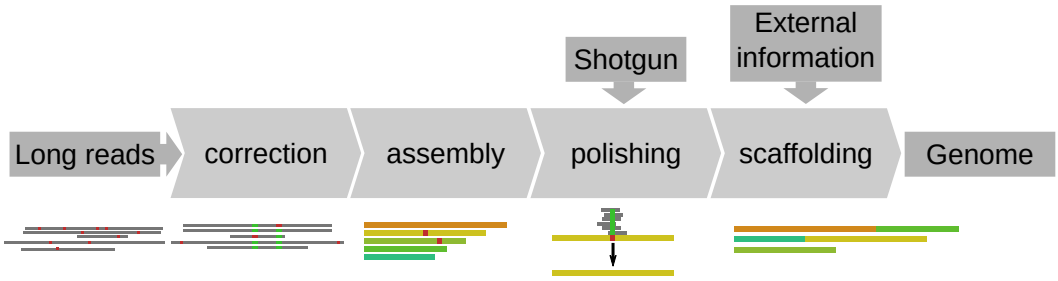


Fig. I.J: Example of a typical assembly pipeline using third generation sequencing. The error prone long reads are first corrected by pairwise comparisons. The corrected reads are assembled into contigs using their overlaps. The remaining sequencing errors in the assembly are removed by polishing with accurate short reads. Other sources of information can then be used to combine contigs into scaffolds.

3.2 Phylogenetic representation

A common way to analyze the genome of new microorganisms is to compare it to other species. To achieve this, one needs to have other closely related genomes available. A common case where dense species genome representation is required is when attempting to detect HGT.

HGT detection methods often rely on discordance between gene trees and species trees. A horizontally transferred gene between two distant species would show strong sequence similarity [87]. For this reason, detection of recent events requires genomes of closely related organisms as a comparison point.

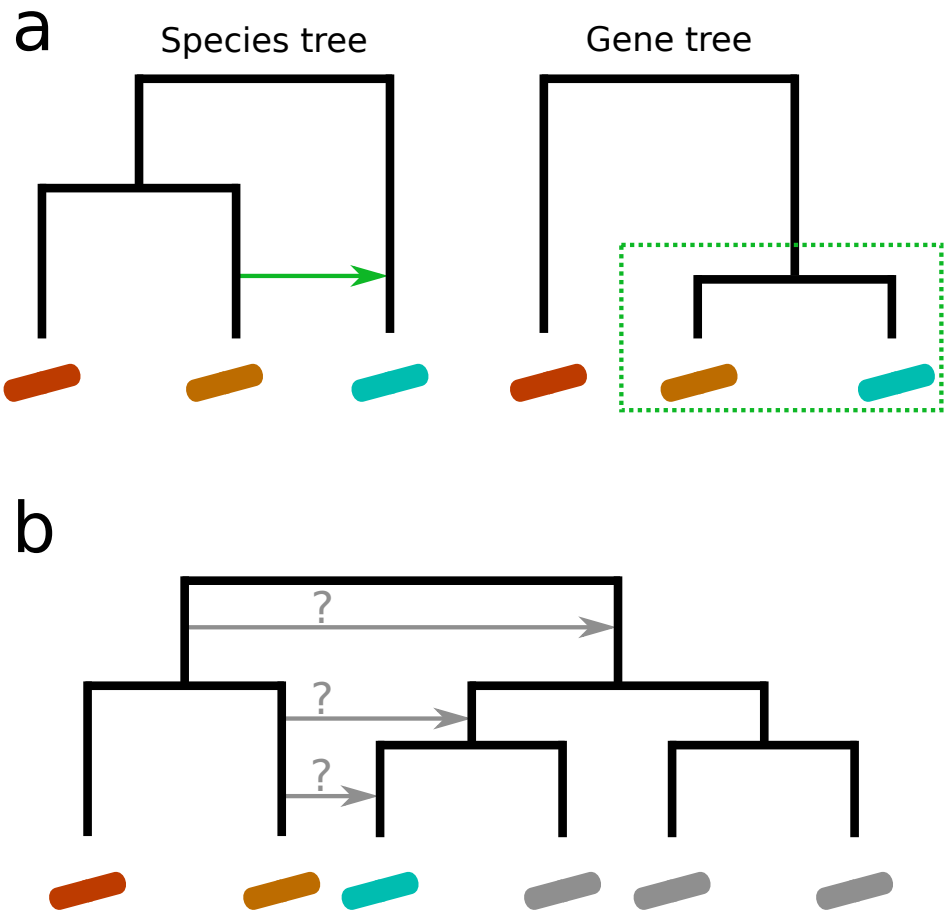


Fig. I.K: Phylogenetic representation of an HGT event. **a:** An HGT event between two species (shown with a green arrow) can be detected through discrepencies between the species (left) and gene (right) trees. **b:** In cases where genomes of closely related species are unavailable (greyed out organisms), the origin of the horizontal transfer cannot be accurately inferred (possible events shown with grey arrows).

Another frequent analysis when comparing a group of strains or species of microorganisms is to define the common set of genes they share, known as pangenome. This also allows the identification of genes specific to a particular genome, known as accessory genome. Such sets can be helpful to determine metabolic reactions associated with species or niches, however they heavily depend on the proportion of available species in the group.

Lately, several large consortia [88–90] undertook the daunting task of sequencing thousands of organisms throughout the tree of life. For the aforementioned reasons,

these large collaborations are likely to greatly improve the power of comparative genomic analyses results in the future.

3.3 The transition to genome graphs

Until recently, all reference genomes were exclusively stored as linear (or circular) sequences of DNA. This linear sequence is often obtained from a mix of multiple individuals, or alleles within an individual. It is effectively a semi-arbitrary combination of multiple haplotypes collapsed into an artificial consensus sequence. A more accurate alternative is to produce a reference sequence graph instead [91]. Given a collection of haplotypes, individuals, or strains of a species, one can generate a graph where identical regions are collapsed, while sample-specific variants form bubbles retaining the genetic variability. As this approach is relatively recent, few algorithms have been developed to operate on sequence graphs, making their applications very limited.

The shift to genome graphs is promising for the analysis of bacterial samples, where alignment can be performed on multiple strain references at the same time. Doing this with a collection of linear genome incurs mapping bias due to ambiguous alignments between redundant regions between references [92]. Similarly, genome graphs also allow systematic alignment to different alleles in polyploid organisms, solving the issue of allele-specific mapping bias in linear references [93].

4 Thesis objectives

Throughout this first part, we have laid out the scope of host-pathogen interactions and summarized the current state of genomics in relation to regulation and 3D genomes. Although genomics is a fast changing field, there is a need for computational tools to extract meaningful biological information from the wealth of data.

Throughout the next part, we will introduce our contributions to the field and main results. In the first chapter, we explain our methodological developments related to chromosome conformation capture technologies. In the second chapter, we will present our chromosome scale genome assembly of *A. castellanii*. We then use this resource for our main findings on the genomic changes happening during infection by *L. pneumophila*. Chapter 3 will focus on murine bone macrophages infection by *S. enterica* and the genomic alterations it entails. Finally, in chapter 4, we will discuss additional results related to the implications of viral integrations linked to hepatocellular carcinoma in the human genome. We will end with part 3 where we discuss various aspects of genomics in infection biology, including prospects and limitations.

In this work, we develop accessible and performant methods to extract information from 3C technologies and use them to identify changes happening during infections in various organisms. We then use external data such as gene expression to assess the genes involved in those alterations and discuss how they could be associated with the infection process.

II

Results

quote

In this second part, we present new results produced in the frame of this work. We start by describing tools and algorithms developed to address the questions at hand. In later parts, we dive into the biological results and discuss their significance.

1 Extracting biological signal from contact maps

Most genomics methods generate a large amount of information, most of which is not directly relevant for the problem at hand. One of the main challenges emanating from genomics data is to distill this information and extract only the relevant signal.

In the case of Hi-C and other 3D genomics techniques, the resulting signal is a collection of contacts recorded between pairs of genomic regions. These contacts reflect the average genome structure from a population of cells and are subject to various biases.

The spatial features and changes of interest are diluted in the population and can be obfuscated by noise. Detecting these changes requires a set of bias correction and signal detection methods which are still in their early developments.

In this section we review the recent methodlogical developments that allow to correct the Hi-C signal and present new methods to extract biological features from these datasets. These developments proved necessary to tackle the questions raised in further chapters.

1.1 Streamlined and reproducible Hi-C processing

Pre-processing of Hi-C data, to convert Next Generation Sequencing (NGS) reads into chromosomal contact matrices involves several steps that will impact the resulting signal. The sequencing reads themselves can be the result from religation of two distinct loci (Fig. II.A). These chimeric reads cannot be aligned reliably with generic methods and need to be cut for proper alignment. Chimeric reads become more problematic when increasing the read size relative to restriction fragment length.

Not all read pairs generated by Hi-C experiments represent valid spatial interactions. Some restriction fragments are sequenced without religation and other fragments religate on themselves (Fig. II.B). The various interaction types can be separated based on the strand of origin of their individual reads. In theory, and in practice at

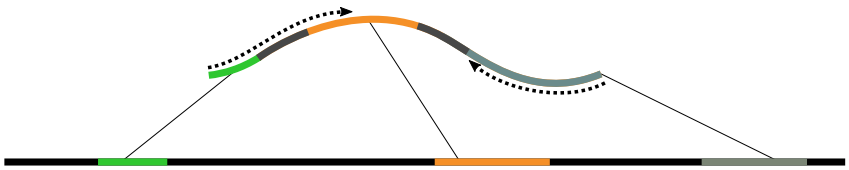


Fig. II.A: Chimeric reads in Hi-C: Example of a Hi-C fragment resulting in a chimeric read. The Hi-C fragment contains 3 different regions (green, orange and grey) which have been religated together. The paired-end sequencing reads are shown as dotted line. The sequencing read spanning the green and orange region will be chimeric and not map to a unique region.

long ranges, one would expect religations to be strand agnostic and to have an equal represenation of all four possible combinations (++, –, +-, -+). In reality, this is never the case at short range contacts, due to the enrichment of self-religation (-+) and dangling ends (or undigested fragments, +-) [94].

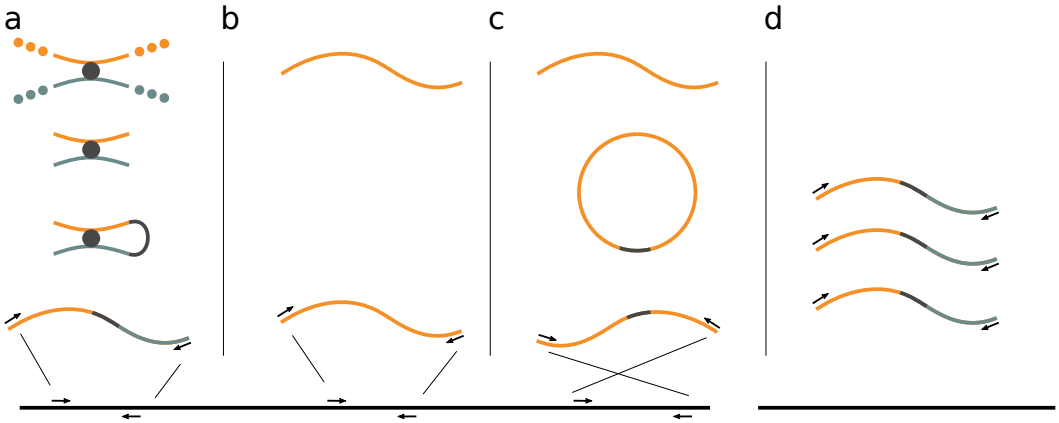


Fig. II.B: Type of interactions generated from Hi-C experiments: **a:** Valid interaction resulting from the religation of two distant loci in physical contact. **b:** Spurious interaction caused by the sequencing of a single restriction fragment, or undigested sequence. **c:** Spurious interaction resulting from the self-religation and breakage of a fragment. **d:** Interactions caused by PCR duplicates. Both reads have the exact same coordinates for all PCR duplicate pairs.

These biases must be accounted when processing Hi-C data. This can be achieved by identifying and filtering out faulty interactions based on their strands.

This preprocessing is often performed using custom scripts and prone to errors, bugs and lack of informations about parameters. In an effort to improve reproducibility and accessibility of Hi-C analysis, we developed hicstuff, an open source Hi-C pipeline that incorporate all the aforementioned steps, along with several downstream processing utilities (Fig. II.C).

Hicstuff can properly align chimeric reads, by digesting them *in-silico* at religation sites, or using iterative mapping (Fig. II.E) where reads are truncated and iteratively extended until they align unambiguously. The Hi-C pairs are then assigned a numerical index according to the restriction they originate from (Fig. II.D) and

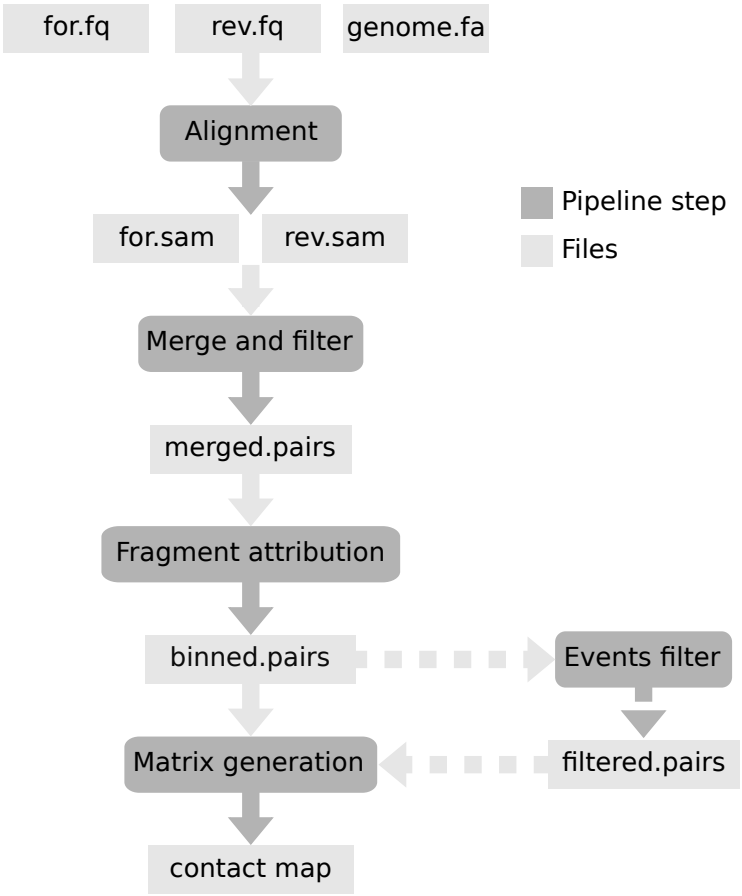


Fig. II.C: Overview of the hicstuff pipeline: Consecutive steps towards the generation of a contact map from sequencing reads, along with the intermediate files are shown as a directed acyclic graph.

artifactual contacts are filtered out using the strand information. Contacts in each bin combination are then summed into a "contact matrix", which is stored in sparse format to spare memory. To allow compatibility with various programs, it can generate sparse matrices in 3 possible formats (COO, bedgraph2d and cool).

Hicstuff is meant to be easily accessible [95], even to non-expert users. It has a comprehensive online documentation and tutorials and the program and its dependencies are installed with a single command. The code is written in python and is exposed both via a Command Line Interface (CLI) to use it as an executable, and an Application Programming Interface (API) to import it as a python library. It is covered by unit tests which are automatically executed on each new release, on the cloud by a continuous integration service to reduce the likelihood of bugs. Hicstuff runs well with default parameters, but has many options to fit most common use cases. It works regardless of genome size organism.

The pipeline also provides reproducibility through an automatic logging of every intermediate result in the pipeline as well as the input parameters used.

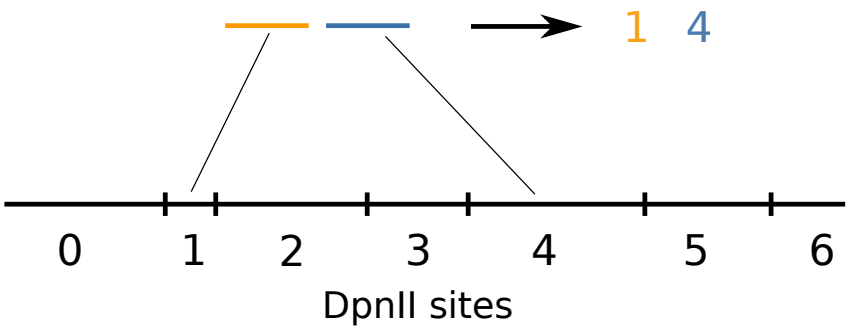


Fig. II.D: Fragment attribution of Hi-C contacts: The genome is segmented into discrete bins according to the positions of restriction sites. Hi-C reads are assigned an index according to the restriction fragment to which they aligned.

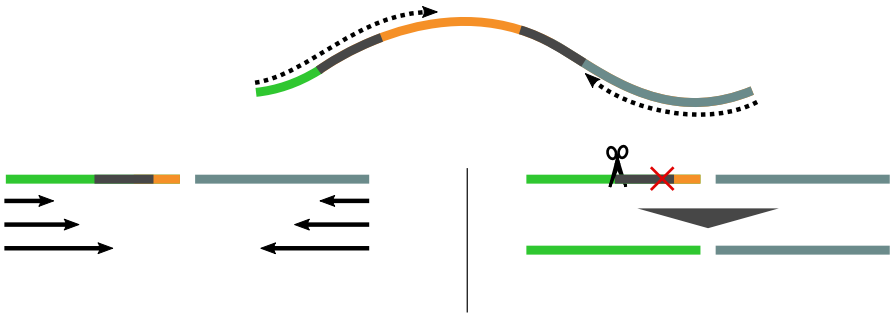


Fig. II.E: Iterative alignment of a Hi-C pair: The Hi-C fragment consists of 3 regions religated together (top). One sequencing read spans two regions (orange and green). Iterative alignment is used to uniquely align the resulting chimeric read (left). The read is truncated to a short length (e.g. 20bp) and iteratively extended until it aligns to a unique position in the genome. Alternatively, reads which do not map uniquely can be digested *in-silico* at known religation sites (right) to remove the chimeric part. The digested reads are then realigned.

The project has already fostered a modest community of users which are offering their contributions, suggest features or report issues they encounter.

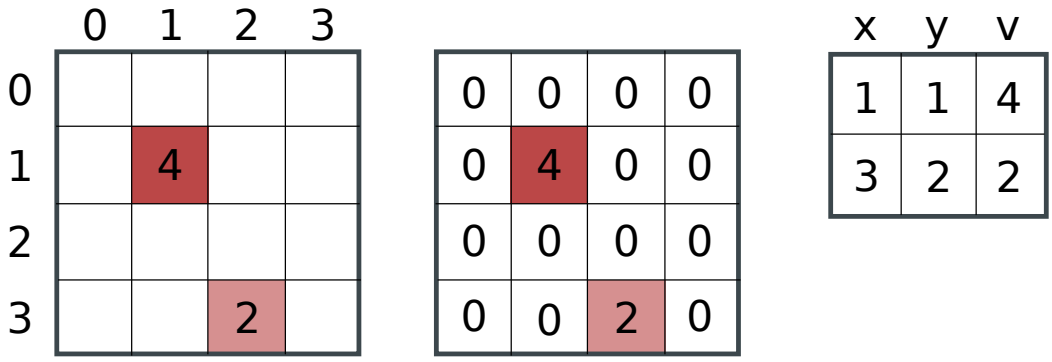


Fig. II.F: Dense and sparse matrix representation: . In Hi-C, matrices are very sparse (i.e. mostly contain 0s), (left). In dense matrix representation, we store all values explicitely. The information stored is highly redundant (middle). Such matrices can be stored efficiently using a sparse representation where only non-zero values are stored explicitely along with their coordinates (right).

1.2 Feature detection with Chromosight

The downstream analysis of chromosome contact maps often involves looking for signals reflecting biologically relevant spatial interactions. Several specialized approaches for pattern detection have been proposed in the past. Each of these methods use a set of specific rules to detect one particular type of pattern. For example, HICCUPS [96] detects chromatin loops by scanning each pixel of the contact map for contact enrichment compared to surrounding pixels.

These specialized methods present several drawbacks, including strong dependence on parameters, poor generalization to non-model species and poor detection rates. These shortcomings motivated us to work on a more generalized pattern detection method to identify arbitrary patterns in chromosome contact maps.

Chromosight uses template matching to identify features on a chromosome contact map. This technique consists in scanning the input Hi-C matrix with a smaller "kernel" image corresponding to the pattern of interest (e.g. a loop) to identify input regions bearing similarity to the kernel. This has the added benefit of allowing to swap the kernel to detect a different feature.

One of the main algorithmic challenges of applying a convolution-based method to Hi-C data is the size of matrices. Hi-C matrices are notoriously large, but they are also extremely sparse (loci do not interact with each other). As a consequence, sparse matrix representation is generally used to handle Hi-C data (Fig. II.F). In the case of large genomes, such as that of *Homo sapiens*, this is necessary to store an entire chromosome's contact map into a regular computer's memory. One of the main drawbacks of sparse representation is that most algorithms are slower and

harder to implement on such matrices. No implementation of convolution for sparse matrices was openly available, which prompted us to write an efficient method to scan the billion of pixels from Hi-C maps in reasonable time. Fortunately, the convolution problem can be reformulated as a matrix multiplication by transforming the input matrices (see A.1), and matrix multiplication is a standardized operation that has been highly optimized in low level libraries, including for sparse matrices.

Chromosight is a python package that takes cool files as input. During its development, we put special attention on good software practice mentioned in section 1.1 to make it easy to use and accessible. This was done by spending time documenting the python API and the CLI as well as putting tutorial examples. Furthermore, the program is covered by a suite of unit tests set up with continuous integration. On every new release, Chromosight is automatically distributed on PyPI, bioconda and dockerhub to accomodate the different use cases and pipelines.

Chromosights algorithm, results and benchmark against state of the art loop detection methods are presented in details in the following pages. The algorithmic details used to tackle the sparse convolution problem are presented in section A.1.

nature COMMUNICATIONS

ARTICLE

https://doi.org/10.1038/s41467-020-19562-7 OPEN

Computer vision for pattern detection in chromosome contact maps

Cyril Matthey-Doret[1,2], Lyam Baudry[1,2,5], Axel Breuer[3,5], Rémi Montagne[1], Nadège Guiglielmoni[1], Vittore Scolari[1], Etienne Jean[1], Arnaud Campeas[3], Philippe Henri Chanut[3], Edgar Oriol[3], Adrien Méot[3], Laurent Politis[3], Antoine Vigouroux[4], Pierrick Moreau[1], Romain Koszul[1✉] & Axel Cournac[1✉]

Chromosomes of all species studied so far display a variety of higher-order organisational features, such as self-interacting domains or loops. These structures, which are often associated to biological functions, form distinct, visible patterns on genome-wide contact maps generated by chromosome conformation capture approaches such as Hi-C. Here we present Chromosight, an algorithm inspired from computer vision that can detect patterns in contact maps. Chromosight has greater sensitivity than existing methods on synthetic simulated data, while being faster and applicable to any type of genomes, including bacteria, viruses, yeasts and mammals. Our method does not require any prior training dataset and works well with default parameters on data generated with various protocols.

[1] Institut Pasteur, Unité Régulation Spatiale des Génomes, CNRS, UMR 3525, C3BI USR 3756, Paris, France. [2] Sorbonne Université, Collège Doctoral, F-75005 Paris, France. [3] ENGIE, Global Energy Management, Paris, France. [4] Institut Pasteur, Synthetic Biology Group, Paris, France. [5] These authors contributed equally: Lyam Baudry, Axel Breuer. ✉email: romain.koszul@pasteur.fr; acournac@pasteur.fr

Proximity ligation derivatives of the chromosome conformation capture (3C) technique[1] such as Hi–C[2] or ChIA-PET[3] determine the average contact frequencies between DNA segments within a genome, computed over hundreds of thousands of cells.These approaches have unveiled a wide variety of chromatin 3D structures in a broad range of organisms. For instance, in all species studied so far, sub-division of chromosomes into self-interacting domains associated with various functions have been observed[4,5] (Fig. 1a). In addition, chromatin loops bridging distant loci within a chromosome (from a few kb to a Mb) are also commonly detected by Hi–C, such as during mammalian interphase[6] or yeast mitotic metaphase[7–9]. Other spatial structures are more peculiar, and sometimes specific to some organisms. For instance, the contact maps of most bacteria display a secondary diagonal perpendicular to the main one[10–12], reflecting the bridging of chromosome replichores (i.e. arms) by the structural maintenance of chromosome complex (SMC) condensin[10], a ring-shaped molecular motor able to entrap and travel along DNA molecules[13]. Smaller straight, or loosely bent, secondary diagonals, also perpendicular to the main diagonal, can also be observed in some maps, reflecting potentially long DNA hairpins or dynamic sliding asymmetrical contacts (Fig. 1a). Such "hairpin-like" configuration is for instance observed near the origin of replication of the *Bacillus subtilis* genome, were it was originally described as a "bow shaped" structure[10]. The formation of these different structures can vary depending on the stage of the cell cycle,[7,10,14], the state of cell differentiation[15] or viral infection[16]. Different molecular mechanisms have been proposed to explain the patterns visible on the contact maps, and for a similar pattern, these mechanisms or their regulation can differ. Although detailing these mechanisms is beyond the scope of the present work, one can note that in mammals the CCCTC-binding factor (CTCF) protein is enriched at loop anchors (i.e. the regions bridged together). It has been proposed that CTCF acts as a roadblock to the SMC molecular motor cohesin, which travels along chromatin. Cohesins promote the formation of chromatin loops, potentially through a loop extrusion mechanisms in which two chromatin filaments are extruded through the cohesin ring[17]). When cohesin encounters a roadblock along one of the filament, chromatin displacement stops in this direction. As a consequence, two roadblocks at two distant loci will stop cohesin progression along both filaments, resulting in a stabilised loop. Such stable loops are then visible in bulk genomics techniques such as Hi–C (for more insights on the putative mechanisms, see for instance[17,18]). Other patterns such as the perpendicular "hairpin" can be explained by alternative scenarios, for instance where cohesin is continuously loaded at a discrete position along the chromatin while being unloaded before hitting a roadblock. A single roadblock combined with continuous cohesin loading in an adjacent locus could result in a bent, bow-shaped pattern, as proposed in[10,19,20]. A large body of work, exploiting genetics and chromosome engineering approaches, aims at characterising the regulation and the functional relationships of these 3D features with DNA processes such as repair, gene expression or

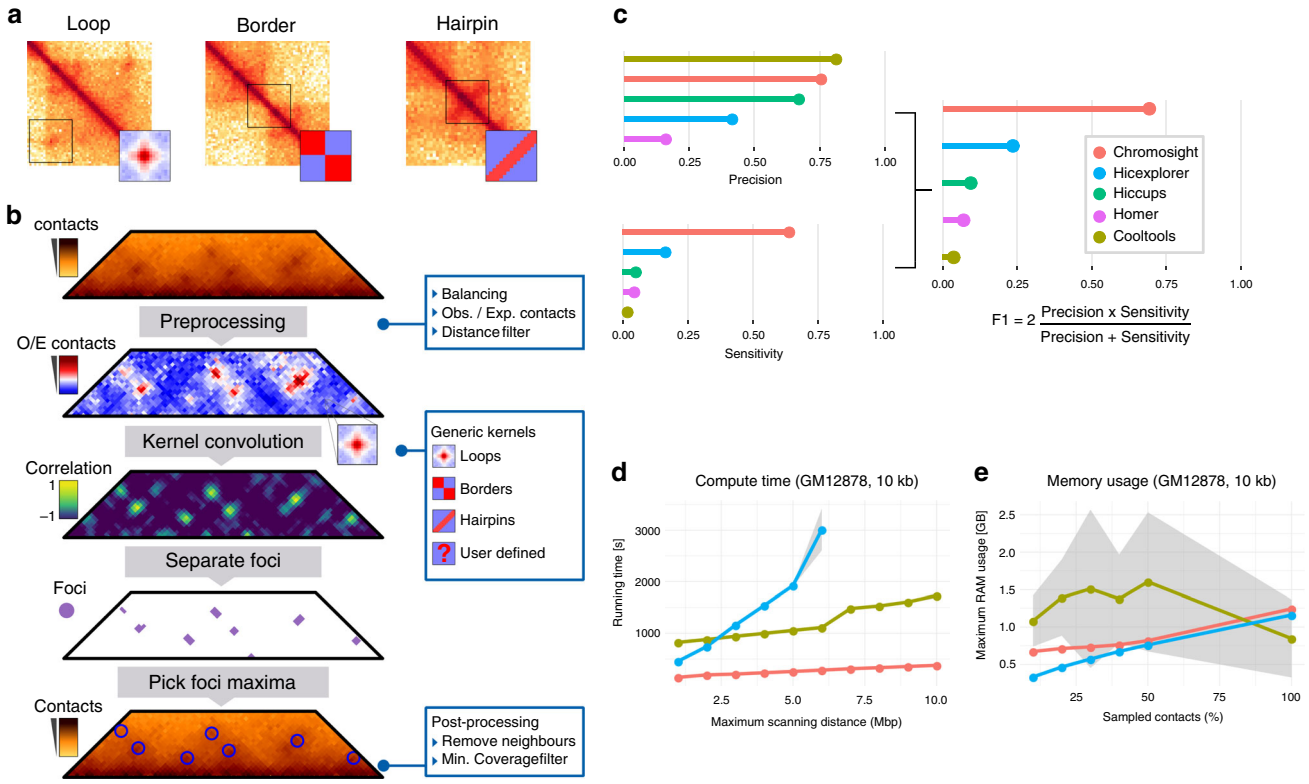


Fig. 1 Chromosight workflow and benchmark. a Examples of distinct patterns visible on contact maps (loop, border and hairpin) and the corresponding chromosight kernels. **b** Matrix preprocessing involves normalisation balancing followed by the computation of observed/expected contacts. Only contacts between bins separated by a user-defined maximum distance are considered. The preprocessed matrix is then convolved with a kernel representing the pattern of interest. For each pixel of the matrix, a Pearson correlation coefficient is computed between the kernel and the surrounding window. A threshold is applied on the coefficients and a connected component labelling algorithm is used to separate groups of pixels (i.e. foci) with high correlation values. For each focus, the coordinates with the highest correlation value are used as the pattern coordinates. Coordinates located in poorly covered regions are discarded. **c** Comparison of Chromosight with different loop callers. Top: F1 score, Precision and Sensitivity scores assessed on labelled synthetic Hi-C data. Higher is better. **d** Run-time. **e** Memory usage according to maximum scanning distance and the amount of subsampled contact events, respectively. Means and standard deviations (grey areas) are plotted.

segregation. Although most structural features can be identified by eye on the contact maps, automated detection is essential to quantify and facilitate the biological and physical interpretation of the data generated through these experiments. While border detection can be achieved quite efficiently using different methods (segmentation, break-point detection, etc; ref. [21]), the calling of loops, as well as other more peculiar features such as "hairpin-like" signals, remains challenging.

Most tools aiming at detecting DNA loops in contact maps rely on statistical approaches and search for pixel regions enriched in contact counts, such as Cloops[22], HiCCUPS[23], HiCExplorer[24], diffHic[25], FitHiC2[26], HOMER[27]. These programs can be computationally intensive and take several hours of computation for standard human Hi–C datasets (reviewed in ref. [22]), or require specialised hardware such as GPU (HiCCUPS). In addition, most if not all of them were developed from, and for, human data. As a consequence, they suffer from a lack of sensitivity and fail to detect biologically relevant structures not only in non-model organisms but also in popular species with compact genomes such as budding yeast (*Saccharomyces cerevisiae*) or bacteria where the scales of the structures are considerably smaller than in mammalian genomes. Here we present *Chromosight*, an algorithm that, when applied on mammalian, bacterial, viral and yeast genome-wide contact maps, quickly and efficiently detects and/or quantifies any type of pattern, with a specific focus on chromosomal loops. Different species were chosen to reflect the diversity of genome-wide contact maps observed in living organisms. For instance, loop contact patterns have been observed in these four clades, but with very different scales and visibility. In human (genome size: ~3 Gb), interphase chromosomes display loops bridging chromatin loci separated by ~20 kb to 20 Mb. The structures are reflected by well-defined, discrete dots in the contact maps, away from the main diagonal. In contrast, the mitotic chromosomes of *S. cerevisiae* and fission yeast *Schizosaccharomyces pombe* (genome sizes: ~12 Mb) organise into arrays of loops spanning ~5–50 kb, i.e. much smaller than the loops observed along mammalian interphase chromosomes[7–9]. Because of their proximity to the main diagonal in standard Hi–C experiments, the signal generated by those loops is more difficult to call. Loops have been observed in bacteria as well. For instance, in *B. subtilis* (genome size: 4.1 Mb), a few weak, discrete loop signals were observed but never directly quantified[10]. In addition to loops, self-interacting domains have also been described in these different species, that differ in size and nature. For instance, topologically associating domains[4,28] have a mean size of 1 Mb (from 200 kb to 6 Mb) in human and mice, compared to the small, chromosome interacting domains (CID) of bacteria that range in size between a few dozens to a couple hundreds kb[10,29,30]. Besides this limitation, most programs are limited to domain or loop calling and remain unable to call de novo different contact patterns such as DNA hairpins or the asymmetric patterns seen in species such as *B. subtilis*[10].

Results

Presentation and benchmark of Chromosight. Chromosight takes a single, whole-genome contact map in sparse and compressed format as an input. It applies a balancing normalization procedure[31] to attenuate experimental biases. A detrending procedure, to remove distance-dependent contact decay due to polymeric behaviour, is then applied, which consists in dividing each pixel by its expected value under the polymer behaviour (Fig. 1b). A template (kernel) representing a 3D structure of interest (e.g. a loop, a boundary,...) is fed to the program and sought for in the image of the contact map through two steps (Fig. 1b). First, the map is subdivided into sub-images correlated to the template; then, the sub-images with the highest correlation values are labelled as template representations (i.e. potential matches, see Methods). Correlation coefficients are computed by convolving the template over the contact map. To reduce computation time, the template can be approximated using truncated singular value decomposition (tSVD) (Supplementary Note 1[32]). To identify the regions with high correlation values (i.e. correlation foci), Chromosight uses Connected Component Labelling (CCL). Finally, the maximum within each correlation focus is extracted and its coordinates in the contact map determined.

We decided to benchmark Chromosight against 4 existing programs by running them in loop-calling mode on synthetic Hi–C data mimicking mitotic chromosomes of *S. cerevisiae* ("Methods" and Supplementary Fig. 1). Whereas Chromosight displays a precision (i.e. proportion of true positives among detected patterns) comparable to the other programs, its sensitivity (i.e. proportion of relevant patterns detected) is more than threefold higher (~70%) compared to the second-best program Hicexplorer (~20%) (Fig. 1c). As a result, Chromosight's F1 score, a metric that considers both precision and sensitivity, is also threefold higher, reflecting the effectiveness of the program at detecting more significant loops in this synthetic case study (Supplementary Fig. 2a). To further benchmark the program's performance, we ran the three best CPU-based programs (Cooltools, Hicexplorer, Chromosight) on high resolution (10 kb), human genome-wide experimental contact maps. Chromosight outperforms existing methods regarding computing time (Fig. 1d), without straining RAM (Fig. 1e). For instance, on a single CPU core, it detects loops at maximum distance of 5 Mb within ~5 min compared to ~17 and 30 min for Cooltools and Hicexplorer, respectively.

To get a sense of the differences between the softwares when applied to experimental human contact maps, we compared them with default parameters on Hi–C data generated from GM12878 cell lines[33]. Compared to Chromosight, we first noticed that other programs missed multiple loops which were clearly visible on the maps (e.g. Supplementary Fig. 3a). For instance, Chromosight found 85% of the loops detected by Cooltools, the software with the highest precision in our benchmark, while overall identifying a much larger number of loops (37,955 vs. 6264, respectively) (Supplementary Fig. 3c). We then measured the proportion of loops with both anchors overlapping CTCF peaks identified from ChIP-seq[34]. Almost all (~95%) loops detected by Hiccups and Cooltools, the most conservative programs, co-localize with CTCF enriched sites, compared to ~64% for the loops detected by Chromosight and Hicexplorer (Supplementary Fig. 3b). Chromosight (and Hicexplorer) indeed detects multiple weaker loops, visible on the maps and arranged in grid-like patterns, but often with only one anchor falling into a well-defined CTCF enriched site. Some of these weaker loops' anchors may be less enriched in CTCF, which would cause ChIP-seq peak calling algorithms to discard them because of parameters such as intensity thresholds, or minimum inter-peak distances. This means that more sensitive loop callers could result in lower CTCF peak overlap, not because of inaccurate detection, but rather because of the CTCF peaks cutoffs. On the other hand, less sensitive loop callers would call the strongest loops associated with the strongest CTCF peaks. We can also not exclude that a portion of the less intense loops called by Chromosight are linked to different protein complexes or mechanisms. More investigations will further dissect the nature of these loops.

Detection and quantification of loops in a compact genome. Hi–C contact maps of budding and fission yeast chromosomes

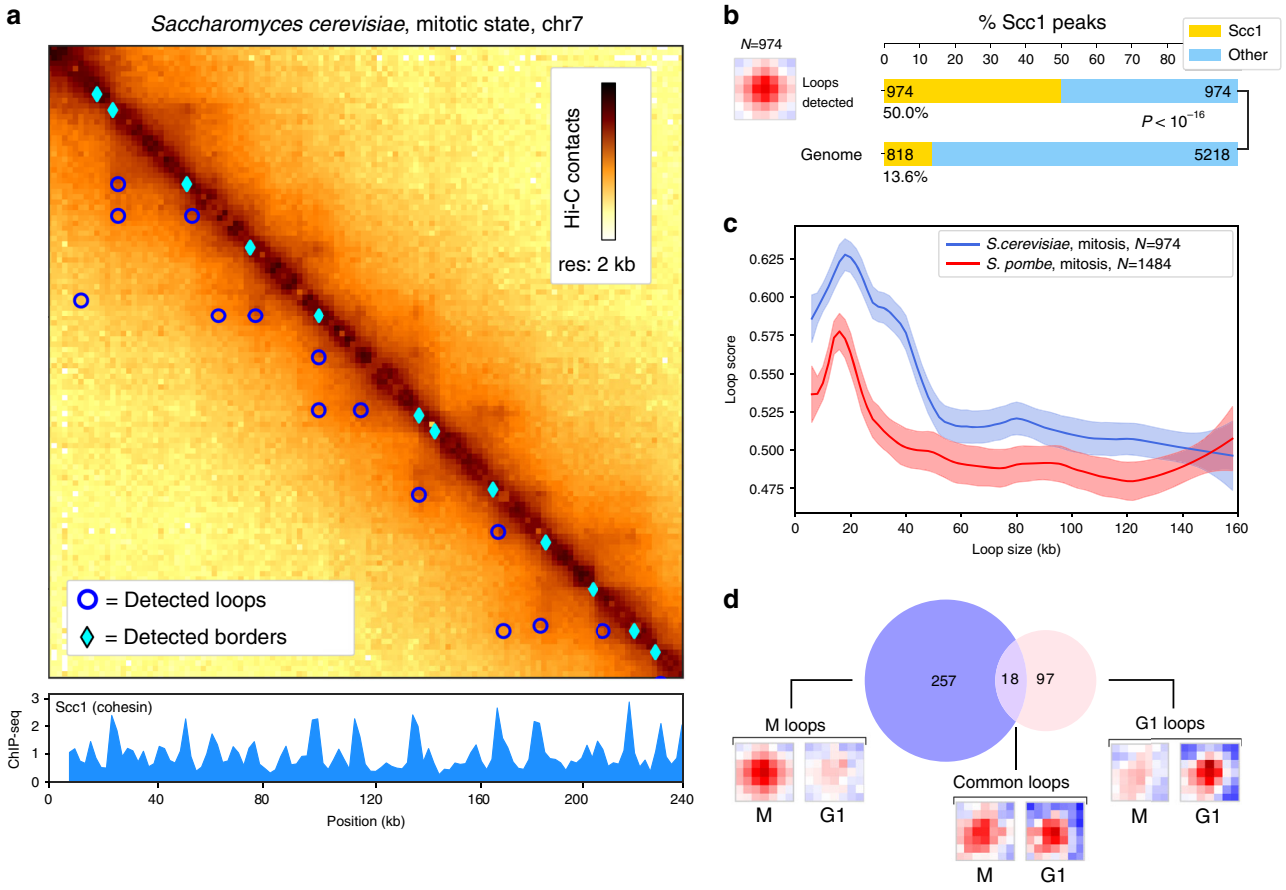


Fig. 2 Applications on yeast genomes. a Zoom-in of the contact map of chromosome 5 of *S. cerevisiae* with synchronised ChIP-Seq signal of Scc1 protein (cohesin) at 2 kb resolution with detected loops and border patterns,[8]. The darker, the more contacts. **b** Pileup plots of windows centered on detected loops with the number of detections. Barplots of the proportion of Scc1 peaks for anchors of detected loops and associated *p*-value (Fisher test, two-sided). **c** Loop spectrum showing scores in function of the loop size in *S. cerevisiae* (974 loops) and *S. pombe* (1484 loops). Curves represent lowess-smoothed data for easier interpretation with 95% confidence intervals. **d** Number of loops detected only in G1 phase, M phase, or in both. For each category, the pileup of each set of coordinates is shown for both G1 and M conditions (mitotic data[8] subsampled from 44M to 5.8M contacts for comparison with G1[7]).

generated from synchronised cells during meiosis[35] and mitosis[7–9] display arrays of chromatin loops. Recent work further showed that *S. cerevisiae* mitotic loops are mediated and regulated by the SMC complex cohesin[7,8]. Chromosight loop calling on data from ref. [8] identified 974 loops along *S. cerevisiae* mitotic chromosomes (Fig. 2a). An enrichment analysis shows that half (50%) of the anchors of those mitotic loops consist in loci enriched in the cohesin subunit Scc1 (Fig. 2b), ($P < 10^{-16}$). The loop signal spectrum in mitosis shows the most stable loops are ~20 kb long (Fig. 2c). This size is also found in the *S. pombe* yeast, which has longer chromosomes.

On the other hand, loop calling on contact maps generated from cells in G1, where cohesin does not stably binds to chromosomes, yielded only 115 loops (Fig. 2d and Supplementary Fig. 4a). Interestingly, this pool of loops appears different from the group of loops detected during mitosis suggesting that cohesin independent processes act on chromosomal loop formation in yeast (Fig. 2d and Supplementary Fig. 4a). Notably, loop anchors were enriched in highly expressed genes (HEG) (Supplementary Fig. 4a).

To validate the biological relevancy of the loops detected by Chromosight during mitosis, we further analysed their dependency and association to cohesin using the quantification mode implemented in the program (Methods and Supplementary Fig. 5a). This mode allows to precisely compute the correlation scores on a set of input coordinates with a generic kernel. We computed the "loop spectrum" (Loop score versus size) for pairs of cohesin ChIP-seq peaks separated by increasing genomic distances. A characteristic size of 20 kb was clearly visible on the spectrum during mitosis, whereas the spectrum in G1 appeared flat (Supplementary Fig. 5b). This analysis highlights the role of cohesin in mediating regular loop structures during mitosis and shows how Chromosight can be used to precisely quantify spatial patterns like chromosome loops.

To test the ability of Chromosight to detect loops in a genetically disturbed context, they were called on contact data of a mutant depleted for the SMC holocomplex member Pds5 (Precocious Dissociation of Sisters)[7]. This protein regulates cohesin loop formation through two independent pathways[7], and its depletion leads to the formation of loops over longer distances than in wild-type yeast. One anchor of loops in Pds5 depleted cells appeared to be the centromeres, as suggested by visual inspection of the maps[7]. However, loop patterns are shadowed by a strong boundary signal appearing at the centromeres, which makes their visual identification challenging. Loop calling using Chromosight confirmed this observation, as the anchors of the loops called were strongly enriched at centromeric regions (Supplementary Fig. 4b, $P < 10^{-16}$)). This analysis shows that Chromosight is able to robustly quantify global reorganisation of genome architecture.

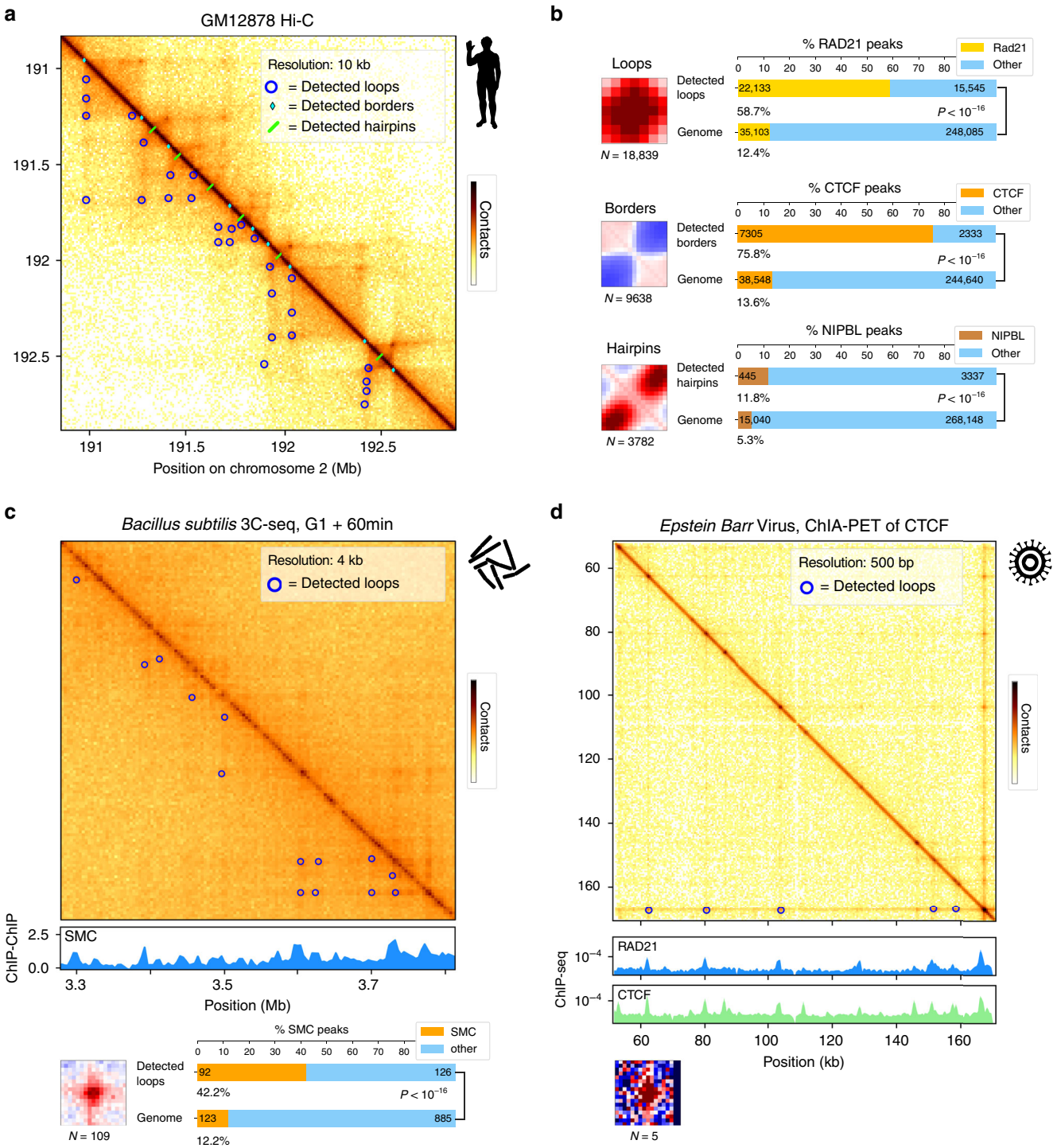


Fig. 3 Applications to various genomes. a Zoom-in of contact map for chromosome 2 of *Homo sapiens* at 10 kb resolution[36] with Chromosight detection of loop, border and hairpin patterns. The darker, the more contacts. **b** Left: pileup plots of windows centered on detected loops, borders and hairpins with the number of detections. Right: bar plots showing proportion in Rad21 peaks for detected loops, proportion in CTCF peaks for detected borders and proportion of NIPBL peaks for detected hairpins and associated *p*-value (Fisher test, two-sided). **c** Detection of loops in the *B. subtilis* genome. Subset of the *B. subtilis* genome-wide contact map near the replication origin. The darker, the more contacts. Loops are called with Chromosight and annotated with blue circles. Under the contact map the ChIP-chip signal deposition of *B. subtilis* SMC is plotted[10]. The pileup plot of the detected loops, and a bar plot showing enrichment of SMC in the anchors of the detected loops (Fisher test, two-sided), are indicated underneath. **d** Contact map of the Epstein Barr virus genome[38]. Called loops using Chromosight are indicated with blue circles. The ChIP-seq deposition signal of Rad21 and CTCF is plotted under the map. Associated pileup plot of the detections is indicated underneath.

Finally, we called domain boundaries (Fig. 1a, border kernel) on the G1 maps, identifying 473 instances of boundaries mostly associated with HEG as well (Supplementary Fig. 4b).

Exploration of various genomes and patterns. To further test the versatility of Chromosight, we called all three kernels described in Fig. 1a, i.e. loops, borders and hairpins (Supplementary Fig. 6) in Hi–C contact maps of human lymphoblastoids (GM12878)[36] (Fig. 3a).

With default parameters, Chromosight identified 18,839 loops (compared to ≃10,000 detected in ref. [6]) whose anchors fall mostly (~ 58%, $P < 10^{-16}$) into loci enriched in cohesin subunit Rad21 (Fig. 3b). Decreasing the detection threshold (Pearson coefficient parameter) allows to detect lower intensity but relevant patterns (Supplementary Fig. 7a). The program also identified 9638 borders, ~75% of which coincide with CTCF binding sites, compared to ~14% expected ($P < 10^{-16}$). In human, TADs are known to be delimited by CTCF-enriched sites, suggesting that Chromosight does indeed correctly identify boundaries involved in TADs delimitation. Finally, Chromosight detected 3,782 hairpin-like structures (Fig. 3b), a pattern not systematically sought for in Hi–C maps. The chromosome coordinates for this pattern appeared enriched in cohesin loading factor NIPBL (2 fold effect, $P < 10^{-16}$), suggesting that these hairpin-like structures could be interpreted as cohesin loading points (Supplementary Fig. 6). To test for a role of cohesin and NIPBL in generating these patterns, we quantified loops and hairpins on contact maps generated from cells depleted either in cohesin or NIPBL. Both conditions were associated with a disappearance of the detected patterns (Supplementary Fig. 8), further supporting their formation hypothesis. Finally, we called loops de novo along the genomes of various animals from the DNA Zoo project[37], showing that stable loops of ≃100–150 kb are a conserved feature of animal genomes (Supplementary Fig. 9).

The loop detection efficiency was also tested using noisier, compact genomic contact maps. We applied it on the 3C-seq data generated from bacterium *B. subtilis*[10]. Chromosight identified 109 loops distributed throughout the chromosome (Fig. 3c). Annotation of loop anchor positions showed a strong enrichment with the bacteria Smc-ScpAB condensin complexes (Fig. 3c). Some of these loops were surprisingly large, bridging loci separated by more than 100 kb (Supplementary Fig. 10) (for a genome size of 4.1 Mb). Several of these large loops may correspond to the bridging of replichores at positions symmetric with respect to the origin of replication (Supplementary Fig. 10). This is in agreement with[10] which showed how SMC condensin SMC-ScpAB complexes loaded at sites adjacent to the origin of replication of the chromosome tether the left and right chromosome arms together while traveling from the origin to the terminus.

Finally, we used Chromosight to detect loops on contact data generated using pair-end tag sequencing (ChIA-PET)[38], which captures contacts between DNA segments associated to a protein of interest. We used ChIA-PET data for CTCF from human lymphoblastoids[38] binned at a very high resolution (500 bp). Lymphoblastoids are immortalised B lymphocytes, they contain episomes of the Epstein Barr Virus (EBV), a DNA virus that is approximately 172 kb in size and is involved in the development of certain tumours[39]. Surprisingly, Chromosight detected several loops (5) inside the genome of the Epstein Barr virus[38]. These loops, of a few dozen kb in size, coincide with the position of the cohesin (Rad21) and CTCF binding sites present along the viral genome (Fig. 3d). Such interactions have been suggested from 3C qPCR data[40]. Automatic detection now unambiguously supports a specific viral chromosome structure that could impact the transcriptional regulation and metabolism of the virus[40].

Application to different proximity ligation protocols. Besides Hi–C, Chromosight can be applied on contact data generated with alternative protocols developed to explore various aspect of chromosomal organisation (Fig. 4a). We retrieved publicly available datasets from asynchronous human cells spanning a range of techniques (i.e. ChIA-PET, DNA SPRITE, HiChIP and Micro-C) from the 4D Nucleome Data Portal[41], and applied loops detection in the resulting contact maps. In situ ChIA-PET[42] quantifies the contact network mediated by a specific protein of interest thanks to the addition of an immunoprecipitation step. Chromosight required adjustment of a single parameter to produce visually satisfying loop calling in in situ ChIA-PET data. We then performed loop detection on DNA Split-Pool Recognition of Interactions by Tag Extension (SPRITE) data[43]. This approach requires cross-linking and fragmentation of chromatin but does not use ligation. Instead, it splits the content into 96-well plates with barcode molecules in each well. The barcode signature allows clustering of complexes that were originally part of a higher-order chromatin structure in the nucleus. Chromosight was able to detect patterns that visually correspond to loops, although the noise present in this original proof-of-principle dataset made detection challenging. We then analysed HiChIP data[44], a protocol similar to ChIA-PET but with a better signal-to-noise ratio, and that requires a lower amount of input DNA. The results of loop calling on HiChIP matrices were very close to those from Hi–C (Fig. 4a). Finally, loops were called on the Micro-C data recently generated from human embryonic stem cells (hESC)[45]. Micro-C uses MNase digestion and a dual cross-link procedure, which allows a contact resolution down to the nucleosome scale. This approach resulted in the highest number of loops (~45,000 Fig. 4b); a visual inspection confirmed that most of them appeared relevant. The number of detected loops in each protocol is directly dependent on the coverage, but these analyses show that Chromosight can conveniently be used for the analysis of data generated through various proximity ligation protocols with minimal, if any, tuning.

In parallel to the loop calling mode, we also used Chromosight in its quantify mode to measure the loop signal between pairs of cohesin peaks as a function of their genomic distance for the different protocols in asynchronous human cells (Fig. 4c). The resulting spectra were quite similar, with loop scores peaking around 120 kb for each protocol. Surprisingly, a secondary peak was also clearly visible at 250 kb, corresponding to about twice the fundamental frequency. This peak was clearest with the Micro-C data. These peaks were absent from dataset generated directly on mitotic condensed chromosomes ($T = 0$ from ref. [46]), but using the same ChIP-seq dataset (Supplementary Fig. 8c). The median distance between cohesin peaks called from ChIP-seq was 468 kb, suggesting that this parameter didn't introduce a bias accounting in the 120 kb. This double peak in the distribution of cohesin contacts as a function of their genomic distance in interphase cells remains to be validated independently, and its signification characterised.

Point and click mode. In addition to the kernels presented here (loops, borders, hairpins), visual inspection of the contact maps may inspire scientists to seek for new patterns of interest for quantitative analysis. We have therefore included a "point and click" mode that allows easy manual inspection of Hi–C contact maps to select patterns identified by users. The user clicks on positions corresponding to patterns of interests. For each position, a window will be drawn by the program. A new kernel is

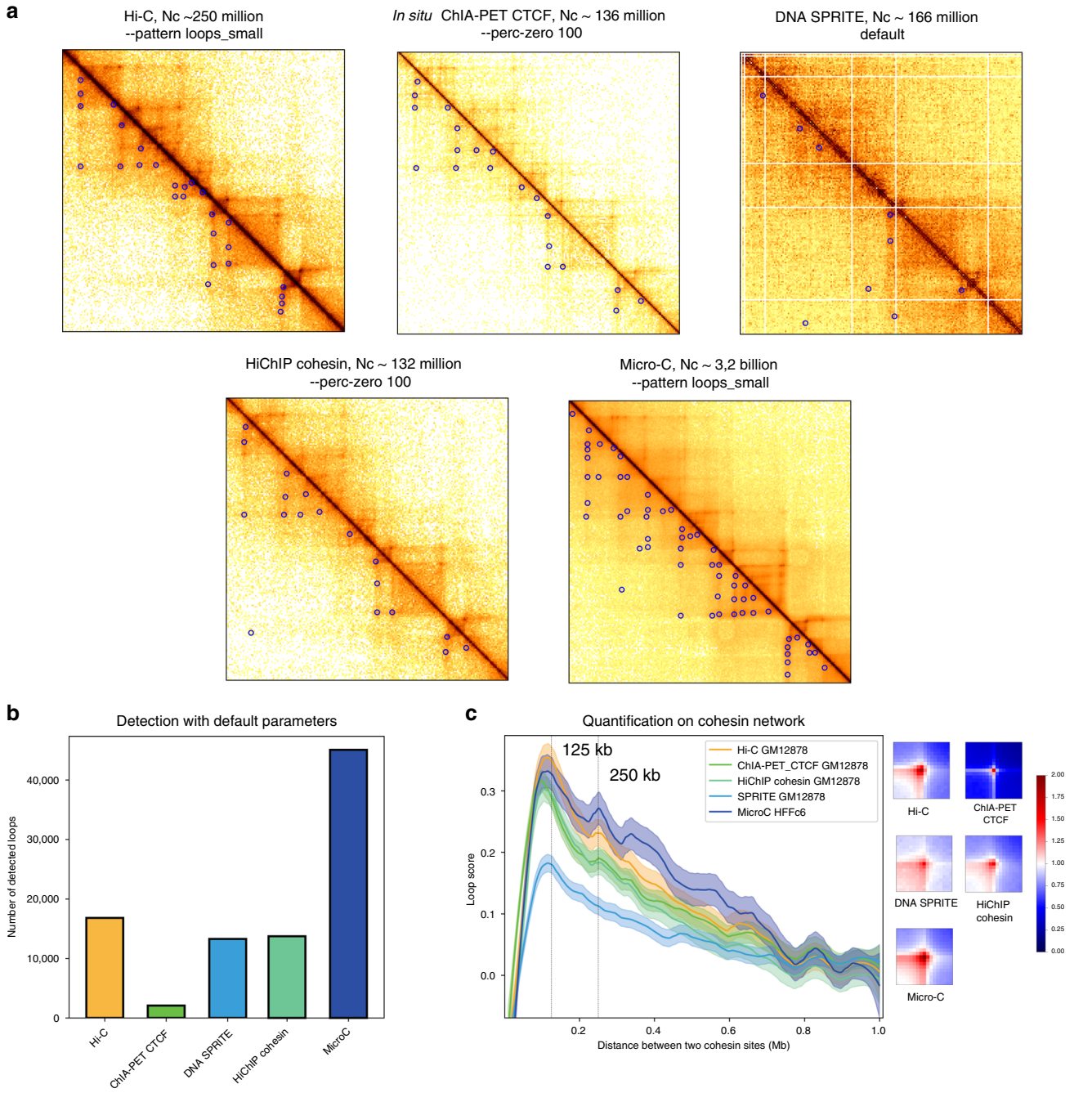


Fig. 4 Analyses with data from alternative contact technologies. a Magnification of *Homo sapiens* chromosome 2 contact maps generated with five different experimental methods (around STAT1 gene; bin:10 kb): Hi-C[36], In situ ChIA-PET of CTCF[42], DNA SPRITE[43], HiChIP of cohesin[44], Micro-C[45]. All cells are cycling GM12878 cell types except for Micro-C (hESC). Blue circles: loops detected using Chromosight. The corresponding number of reads in each of the genome-wide map is indicated above the panels. The parameter (if any) notified to Chromosight is also indicated above each map. **b** Number of loops detected using Chromosight with default parameters for the five datasets. **c** Left: loop spectrum computed using Chromosight in quantify mode on pairs of cohesin peaks for the five datasets (Methods). Curves represent lowess-smoothed data with 95% confidence intervals. Right: associated pileup plots of the quantified positions for the five different experimental methods.

then automatically generated by summing all windows and applying a Gaussian filter to attenuate the fluctuations resulting from the small number of selected positions. This kernel can then be used in the other modes of Chromosight (detection, quantification) for further analyses.

We illustrate this functionality to investigate the pattern of centromere-centromere interactions in yeast. Yeasts contact maps are scattered with cross-shaped dots corresponding to inter-chromosomal contacts between peri-centromeric positions. This cross-shaped pattern is characteristic of the Rabl configuration of those genomes, where all centromeres are maintained in the vicinity of each other at the level of the microtubule organising center[47,48]. As a result, peri-centromeric regions collide with each other more frequently than with the rest of the genome, resulting in a distinct trans pattern. In budding yeast, the 16 centromeres result in 120 discrete, inter-chromosomal cross-shaped dots. We selected (by double-clicking) 15 patterns of these *S. cerevisiae* centromere contacts. The resulting kernel was then used to

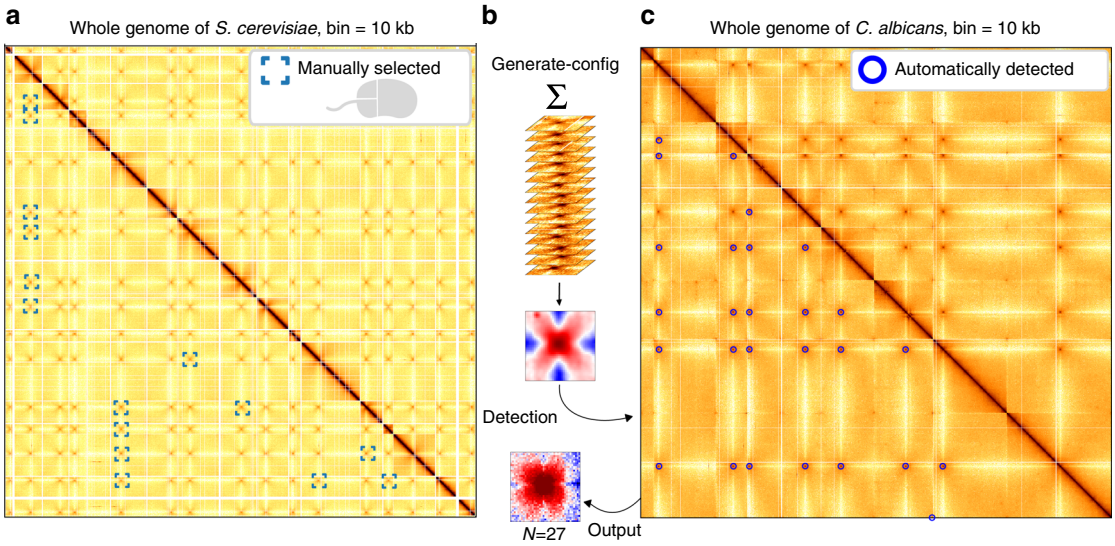


Fig. 5 Point and click mode. a Whole-genome contact map of *S. cerevisiae*[8] with 15 inter-centromere patterns that were selected by hand. Darker means more contacts. **b** Chromosight generates a new kernel by summing all the selected patterns and applying a Gaussian filter. **c** Chromosight detection of the inter-centromeres patterns in the whole-genome contact map of *C. albicans*[49] with the resulting pileup plot of the 27 detections.

perform the detection of similar structures in the genome contact map of another yeast species, *Candida albicans*, a diploid opportunistic pathogen which contains 8 pairs of chromosomes (resolution: 5 kb, ref. [49]).

Using the kernel generated de novo from the *S. cerevisiae* contact map, Chromosight automatically detected 26 out of the 28 inter-centromeric patterns of *C. albicans*, along with one false positive (most likely a genome misassembly, located at the edge of the map) (Fig. 5). These positions are nevertheless sufficient to point at centromere positions, and can for instance then be used to characterise their genomic coordinates[47].

Note that, although subtelomeric regions in yeast tend to cluster in yeast nuclei and therefore display discrete contacts reminiscent of those of peri-centromeric contacts, Chromosight was able to discriminate between those two patterns, detecting specifically inter-centromeric interactions. The program was therefore able to correctly assess the subtle geometrical differences between these two patterns. Overall, this analysis shows the ability of Chromosight to quickly detect any type of user-defined pattern. We anticipate that many more patterns will be added to the catalogue of visual patterns linked to different molecular mechanisms of chromosome architecture.

Discussion

In this work, we present Chromosight, a computer vision program to detect 3D structures in chromosome contact maps. We show that Chromosight outmatches other programs designed to detect chromosome loops, and that it can be used to extract other biologically relevant patterns generated through different chromosome capture derivatives.

Chromosight is versatile and we expect that additional pattern configurations will be added by the community, such as stripes, bow-shaped patterns, patterns associated to misassemblies or structural variations (e.g. inversions, translocations...) or any pattern of interest that the user can propose. The approach could therefore be used to investigate structural rearrangements in cancer cells, for instance, although the sensitivity of the program to detect rearrangements taking place in only a fraction of a population of cells remains to be tested. Similarly, the potential of the approach to develop new Hi–C based genome scaffolding algorithms could also be explored in the future[50,51]. The program has a great flexibility that allows to work with diverse biological data and address different questions, either using the de novo calling mode or the quantification mode. For instance, the possibility of varying the size of the loop kernel allows to optimise it for different conditions: larger kernels are more tolerant to noisy data (Fig 3c) as they dampen the fluctuations whereas smaller kernels allow to detect loops very close to the main diagonal (Supplementary Fig. 7).

A possible extension of the present approach is the addition of an iterative feedback step to the general flowchart of the current algorithm. Indeed, the output pileup after the first run of detection can be reused in another iteration of detection on the same data. This step could allow a finer adaptation to the data and to detect patterns a little further away from the initial kernel while keeping the basic characteristics.

With decreasing sequencing costs, new experimental protocols and optimised methods for amplifying specific genomic regions, we expect that the folding of the genomes of many species will be investigated in the near future using chromosome contact techniques. The algorithmic approach we present here provides a computational and statistical framework for the discovery of new principles governing chromosome architecture.

Methods

Simulation of Hi–C matrices. Simulated matrices were generated using a bootstrap strategy based on Hi–C data from chromosome 5 of mitotic *S. cerevisiae*[7] at 2 kb resolution. Three main features were extracted from the yeast contact data (Supplementary Fig. 1): the probability of contact as a function of the genomic distance ($P(s)$), the positions of borders detected by HicSeg v1.1[52] and positions of loops detected manually on chromosome 5. Positions from loops and borders were then aggregated into pileups of 17×17 pixels. We generated 2000 simulated matrices of 289×289 pixels. A first probability map of the same dimension is generated by making a diagonal gradient from $P(s)$ representing the polymer behaviour. For each of the 2000 generated matrices, two additional probability maps are generated. The first by placing several occurrences of the border pileup on the diagonal, where the distance between borders follows a normal distribution fitted on the experimental coordinates. The second probability map is generated by adding the loop kernel 2–100 pixels away from the diagonal with the constraint that it must be aligned vertically and horizontally with border coordinates. For each generated matrix, the product of the $P(s)$, borders and loops probability maps is then computed and used as a probability law to sample contact positions while keeping the same number of reads as the experimental map. This simulation method is implemented in the script chromo_simul.py, which can be found on the github repository: https://github.com/koszullab/chromosight_analyses_scripts.

Benchmarking. To benchmark precision, sensitivity and F1 score, the simulated Hi–C data set with known loop coordinates were used. Each algorithm was run with a range of 60-180 parameter combinations (Supplementary Fig. 2) on 2000 simulated matrices and F1 score was calculated on the ensemble of results for each parameter combination separately (Supplementary Table 1). For each software, scores used in the final benchmark (Fig. 1) are those from the parameter combination that yielded the highest F1 score.

For the performance benchmark, HiCCUPS and HOMER were excluded. The former because it runs on GPU, and the latter because it uses genomic alignments as input and is much slower. The dataset used is a published high coverage Hi–C library[36] from human lymphoblastoid cell lines (GM12878). To compare RAM usage across programs, this dataset was subsampled at 10%, 20%, 30%, 40% and 50% contacts and the maximum scanning distance was set to 2 Mbp. To compare CPU time, all programs were run on the full dataset, at different maximum scanning distances, with a minimum scanning distance of 0 and all other parameters left to default. All programs were run on a single thread, on a Intel(R) Core(TM) i7-8700K CPU at 3.70 GHz with 32 GB of available RAM.

Software versions used in the benchmark are Chromosight v0.9.0, hicexplorer v3.3.1, cooltools v0.2.0, homer 4.10 and hiccups 1.6.2. Input data, scripts and results of both benchmarks are available on Zenodo (https://doi.org/10.5281/zenodo.3742095)

Preprocessing of Hi–C matrices. Chromosight accepts input Hi–C data in cool format[53]. Prior to detection, Chromosight balances the whole-genome matrix using the ICE algorithm[31] to account for Hi–C associated biases. For each intrachromosomal matrix, the observed/expected contact ratios are then computed by dividing each pixel by the mean of its diagonal. This erases the diagonal gradient due to the power-law relationship between genomic distance and contact probability, thus emphasising local variations in the signal (Fig. 1b). Intra-chromosomal contacts above a user-defined distance are discarded to constrain the analysis to relevant scales and improve performances.

Calculation of Pearson coefficients. Correlation coefficients are computed by convolving the template over the contact map. Convolution algorithms are often used in computer vision where images are typically dense. Hi–C contact maps, on the other hand, can be very sparse. Chromosight's convolution algorithm is therefore designed to be fast and memory efficient on sparse matrices. It can also exclude missing bins when computing correlation coefficients. Those bins appear as white lines on Hi–C matrices and can be caused by repeated sequences or low coverage regions.

The contact map can be considered an image $\mathrm{IMG}_{\mathrm{CONT}}$ where the intensity of each pixel $\mathrm{IMG}_{\mathrm{CONT}}[i, j]$ represents the contact probability between loci i and j of the chromosome. In that context, each pattern of interest can be considered a template image $\mathrm{IMG}_{\mathrm{TMP}}$ with M_{TMP} rows and N_{TMP} columns.

The correlation operation consists in sliding the template ($\mathrm{IMG}_{\mathrm{TMP}}$) over the image ($\mathrm{IMG}_{\mathrm{CONT}}$) and measuring, for each template position, the similarity between the template and its overlap in the image. We used the Pearson correlation coefficient as a the measure of similarity between the two images. The output of this matching procedure is an image of correlation coefficients $\mathrm{IMG}_{\mathrm{CORR}}$ such that

$$\mathrm{IMG}_{\mathrm{CORR}}[i,j] = \mathrm{Corr}\left(\mathrm{IMG}_{\mathrm{CONT}}\left[i-\frac{M_{\mathrm{TMP}}}{2} : i+\frac{M_{\mathrm{TMP}}}{2}, j-\frac{N_{\mathrm{TMP}}}{2} : j+\frac{N_{\mathrm{TMP}}}{2}\right],\ \mathrm{IMG}_{\mathrm{TMP}}\right) \quad (1)$$

where the correlation operator $Corr(\,\cdot\,,\,\cdot\,)$ is defined as

$$\begin{aligned}\mathrm{Corr}\,(\mathrm{IMG}_X, \mathrm{IMG}_Y) &= \frac{\mathrm{cov}\,(\mathrm{IMG}_X, \mathrm{IMG}_Y)}{\mathrm{std}\,(\mathrm{IMG}_X)\cdot \mathrm{std}\,(\mathrm{IMG}_Y)} \\ &= \frac{\sum_{(m,n)\in X\cap Y}(\mathrm{IMG}_X[m,n]-\overline{\mathrm{IMG}_X})\cdot(\mathrm{IMG}_Y[m,n]-\overline{\mathrm{IMG}_Y})}{\sqrt{\sum_{(m,n)\in X\cap Y}(\mathrm{IMG}_X[m,n]-\overline{\mathrm{IMG}_X})^2}\cdot\sqrt{\sum_{(m,n)\in X\cap Y}(\mathrm{IMG}_Y[m,n]-\overline{\mathrm{IMG}_Y})^2}}\end{aligned} \quad (2)$$

where $\overline{\mathrm{IMG}} = \frac{1}{|X\cap Y|}\sum_{(m,n)\in X\cap Y}\mathrm{IMG}[m,n]$, $X \cap Y$ is the set of pixel coordinates that are valid in image IMG_X and in image IMG_Y, and $|X \cap Y|$ is the number of valid pixels in IMG_X and IMG_Y. A pixel in $\mathrm{IMG}_{\mathrm{CONT}}$ is defined as valid when it is outside a region with missing bins.

Separation of high-correlation foci. Selection is done by localising specific local maxima within $\mathrm{IMG}_{\mathrm{CORR}}$. We proceeded as follows: first, we discard all points (i, j) where $\mathrm{IMG}_{\mathrm{CORR}}[i, j] < \tau_{\mathrm{CORR}}$. An adjacency graph A_{dxd} is then generated from the d remaining points. The value of $A[i, j]$ is a boolean indicating the (four-way) adjacency status between the ith and jth nonzero pixels. The scipy implementation of the CCL algorithm for sparse graphs[54] is then used on A to label the different contiguous foci of nonzero pixels. Foci with less than two pixels are discarded. For each focus, the pixel with the highest coefficient is determined as the pattern coordinate.

Patterns are then filtered out if they overlap too many empty pixels or are too close from another detected pattern. The remaining candidates in $\mathrm{IMG}_{\mathrm{CORR}}$ are scanned by decreasing order of magnitude: every time a candidate is appended to the list of selected local maxima, all its neighbouring candidates are discarded. The proportion of empty pixels allowed and the minimum separation between two patterns are also user defined parameters.

Biological analyses. Pairs of reads were aligned independently using Bowtie2 (v2.3.4.1) with --very-sensitive-local against the *S. cerevisiae* SC288 reference genome (GCF000146045.2). Uncuts, loops and religation events were filtered as described in ref. [55]. Contact data were binned at 2 kb and normalised using the ICE balancing method[31]. Hi–C matrices were generated from fastq files using hicstuff v2.3.0[56]. Detection for biological analyses of yeast and human data was performed with default parameters using a 7×7 loop kernel available in Chromosight using --pattern loops_small unless mentioned otherwise. For enrichment analysis, cohesin peaks were defined using ChIP-seq data from[57]. Raw reads were aligned with bowtie2 and only mapped positions with Mapping Quality superior to 30 were kept and signals were also binned at 2 kb to synchronise with Hi–C data. Peaks of cohesins were considered with ChIP/input > 1.5 and peaks closer than 10 kb to centromeres or rDNA were removed.

Annotation of highly expressed genes was done using RNA-seq data from[8]. Alignment was done as above. The distribution of the number of reads for each 2 kb bin was computed and the top 20% of the distribution were considered bins with high transcription. For border annotation, a set of plus or minus 1 bin on the detected positions is used. For human data, hg19 genome assembly was used with same strategy for alignment, construction and normalisation of contact data. ChIPseq peaks were retrieved from UCSC database (Supplementary Table 2). *B. subtilis* data were aligned with the PY79 genome version and the SMC signal was extracted using ChIP-chip data from[58] and processed as described previously[10,59]. Peaks were annotated with the find_peaks function from scipy (v1.4.1), with parameters threshold = 0.1, width = 50. ChIA-PET data were processed as Hi–C data except that the contact maps were binned at a 500bp resolution. Epstein-Barr virus (EBV) genome, strain B95-8 (V01555.2) sequence was used to align the reads from EBV. For the detection in the different proximity ligation protocols, we retrieved publicly available data sets from the 4D Nucleome Data Portal[41], and applied loops detection in the resulting contact maps of the mcool files at 10 kb resolution with the default settings by possibly changing one option that is indicated in (Fig. 4a).

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

All data associated with this study are publicly available and their reference numbers are listed in Supplementary Tables 2 and 3. Intermediate results, benchmark code and data are available on Zenodo (https://doi.org/10.5281/zenodo.3742095).

Code availability

Software and documentation available at https://github.com/koszullab/chromosight. All scripts required to reproduce figures and analyses are available at https://github.com/koszullab/chromosight_analyses_scripts.

Received: 12 June 2020; Accepted: 16 October 2020;
Published online: 16 November 2020

References

1. Dekker, J., Rippe, K., Dekker, M. & Kleckner, N. Capturing chromosome conformation. *Science* **295**, 1306–1311 (2002).
2. Lieberman-Aiden, E. et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* **326**, 289–293 (2009).
3. Fullwood, M. J. et al. An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature* **462**, 58–64 (2009).
4. Dixon, J. R. et al. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* **485**, 376–380 (2012).
5. Nora, E. P. et al. Spatial partitioning of the regulatory landscape of the x-inactivation centre. *Nature* **485**, 381–5 (2012).
6. Rao, S. S. P. et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–80 (2014).
7. Dauban, L. et al. Regulation of cohesin-mediated chromosome folding by eco1 and other partners. *Mol. Cell* **77**, 1279–1293 (2020).
8. Garcia-Luis, J. et al. Fact mediates cohesin function on chromatin. *Nat. Struct. Mol. Biol.* **26**, 970–979 (2019).
9. Tanizawa, H., Kim, K.-D., Iwasaki, O. & Noma, K.-I. Architectural alterations of the fission yeast genome during the cell cycle. *Nat. Struct. Mol. Biol.* **24**, 965–976 (2017).

10. Marbouty, M. et al. Condensin-and replication-mediated bacterial chromosome folding and origin condensation revealed by hi-c and super-resolution imaging. *Mol. cell* **59**, 588–602 (2015).
11. Umbarger, M. A. et al. The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. *Mol. Cell* **44**, 252–264 (2011).
12. Marbouty, M., Baudry, L., Cournac, A. & Koszul, R. Scaffolding bacterial genomes and probing host-virus interactions in gut microbiome by proximity ligation (chromosome capture) assay. *Sc. Adv.* **3**, e1602105 (2017).
13. Nasmyth, K. & Haering, C. H. Cohesin: Its roles and mechanisms. *Ann. Rev. Gen.* **43**, 525–558 (2009).
14. Naumova, N. et al. Organization of the mitotic chromosome. *Science* **342**, 948–953 (2013).
15. Bonev, B. et al. Multiscale 3d genome rewiring during mouse neural development. *Cell* **171**, 557–572 (2017).
16. Heinz, S. et al. Transcription elongation can affect genome 3d structure. *Cell* **174**, 1522–1536 (2018).
17. Fudenberg, G. et al. Formation of chromosomal domains by loop extrusion. *Cell Rep.* **15**, 2038–2049 (2016).
18. Banigan, E. J. & Mirny, L. A. Loop extrusion: theory meets single-molecule experiments. *Curr. Opin. Cell Biol.* **64**, 124–138 (2020).
19. Wang, X., Brandão, H. B., Le, T. B. K., Laub, M. T. & Rudner, D. Z. *Bacillus subtilis* smc complexes juxtapose chromosome arms as they travel from origin to terminus. *Science* **355**, 524–527 (2017).
20. Brandão, H. B. et al. Rna polymerases as moving barriers to condensin loop extrusion. *Proc. Natl Acad. Sci. USA* **116**, 20489–20499 (2019).
21. Forcato, M. et al. Comparison of computational methods for hi-c data analysis. *Nat. Methods* **14**, 679 (2017).
22. Cao, Y. et al. Accurate loop calling for 3d genomic data with cloops. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz651 (2019).
23. Durand, N. C. et al. Juicer provides a one-click system for analyzing loop-resolution hi-c experiments. *Cell Systems* **3**, 95–98 (2016).
24. Ramírez, F. et al. High-resolution tads reveal dna sequences underlying genome organization in flies. *Nat. Commun.* **9**, 189 (2018).
25. Lun, A. T. L. & Smyth, G. K. diffhic: a bioconductor package to detect differential genomic interactions in hi-c data. *BMC Bioinform.* **16**, 258 (2015).
26. Kaul, A., Bhattacharyya, S. & Ay, F. Identifying statistically significant chromatin contacts from hi-c data with fithic2. *Nat. Protoc.* https://doi.org/10.1038/s41596-019-0273-0 (2020).
27. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and b cell identities. *Mol.Cell* **38**, 576–589 (2010).
28. Dali, R. & Blanchette, M. A critical assessment of topologically associating domain prediction tools. *Nucleic Acids Res.* **45**, 2994–3005 (2017).
29. Le, T. B. K., Imakaev, M. V., Mirny, L. A. & Laub, M. T. High-resolution mapping of the spatial organization of a bacterial chromosome. *Science* **342**, 731–734 (2013).
30. Lioy, V. S. et al. Multiscale structuring of the e. coli chromosome by nucleoid-associated and condensin proteins. *Cell.* **172**, 771–783 (2018).
31. Imakaev, M. et al. Iterative correction of hi-c data reveals hallmarks of chromosome organization. *Nat. Methods* **9**, 999–1003 (2012).
32. Haralick, R. M. & Shapiro, L. G. *Computer and Robot Vision* 1st edn (Addison-Wesley Longman Publishing Co., Inc., USA, 1992).
33. Rao, S. S. P. et al. A 3d map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell* **159**, 1665–1680 (2014).
34. Karolchik, D. The UCSC table browser data retrieval tool. *Nucleic Acids Res.* **32**, 493D–496 (2004).
35. Muller, H. et al. Characterizing meiotic chromosomes' structure and pairing using a designer sequence optimized for hi-c. *Mol. Syst. Biol.* **14**, e8293 (2018).
36. Ghurye, J. et al. Integrating hi-c links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15**, e1007273 (2019).
37. Dudchenko, O. et al. De novo assembly of the *Aedes aegypti* genome using hi-c yields chromosome-length scaffolds. *Science* **356**, 92–95 (2017).
38. Tang, Z. et al. Ctcf-mediated human 3d genome architecture reveals chromatin topology for transcription. *Cell* **163**, 1611–27 (2015).
39. Küppers, R. B cells under influence: transformation of b cells by epstein-barr virus. *Nat. Rev. Immunol.* **3**, 801–12 (2003).
40. Arvey, A. et al. An atlas of the epstein-barr virus transcriptome and epigenome reveals host-virus regulatory interactions. *Cell Host Microbe* **12**, 233–45 (2012).
41. Dekker, J. et al. The 4d nucleome project. *Nature* **549**, 219–226 (2017).
42. Li, X. et al. Long-read chia-pet for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nat. Protoc.* **12**, 899–915 (2017).
43. Quinodoz, S. A. et al. Higher-order inter-chromosomal hubs shape 3d genome organization in the nucleus. *Cell* **174**, 744–757 (2018).
44. Mumbach, M. R. et al. Hichip: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods* **13**, 919–922 (2016).
45. Krietenstein, N. et al. Ultrastructural details of mammalian chromosome architecture. *Mol. Cell* **78**, 554–565 (2020).
46. Abramo, K. et al. A chromosome folding intermediate at the condensin-to-cohesin transition during telophase. *Nat. Cell Biol.* **21**, 1393–1402 (2019).
47. Marie-Nelly, H. et al. Filling annotation gaps in yeast genomes using genome-wide contact maps. *Bioinformatics* **30**, 2105–2113 (2014).
48. Mizuguchi, T., Barrowman, J. & Grewal, S. I. Chromosome domain architecture and dynamic organization of the fission yeast genome. *FEBS Lett.* **589**, 2975–2986 (2015).
49. Burrack, L. S. et al. Neocentromeres provide chromosome segregation accuracy and centromere clustering to multiple loci along a candida albicans chromosome. *PLOS Genet.* **12**, e1006317 (2016).
50. Flot, J.-F., Marie-Nelly, H. & Koszul, R. Contact genomics: scaffolding and phasing (meta) genomes using chromosome 3d physical signatures. *FEBS Lett.* **589**, 2966–2974 (2015).
51. Baudry, L. et al. instagraal: chromosome-level quality scaffolding of genomes using a proximity ligation-based scaffolder. *Genom. Biol.* https://doi.org/10.1186/s13059-020-02041-z (2020).
52. Lévy-Leduc, C., Delattre, M., Mary-Huard, T. & Robin, S. Two-dimensional segmentation for analyzing hi-c data. *Bioinformatics* **30**, i386–i392 (2014).
53. Abdennur, N. & Mirny, L. A. Cooler: scalable storage for hi-c data and other genomically labeled arrays. *Bioinformatics* https://doi.org/10.1093/bioinformatics/btz540 (2019).
54. Pearce, D. J. *An Improved Algorithm for Finding the Strongly Connected Components of a Directed Graph* (Victoria University, Wellington, 2005).
55. Cournac, A., Marie-Nelly, H., Marbouty, M., Koszul, R. & Mozziconacci, J. Normalization of a chromosomal contact map. *BMC Genom.* **13**, 436 (2012).
56. Matthey-Doret, C. et al. hicstuff: Simple library/pipeline to generate and handle hi-c data. *Zenodo*, https://doi.org/10.5281/zenodo.4066351 (2020).
57. Hu, B. et al. Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq. *Nucleic Acids Res.* https://doi.org/10.1093/nar/gkv670 (2015).
58. Gruber, S. & Errington, J. Recruitment of condensin to replication origin regions by parb/spooj promotes chromosome segregation in *B. subtilis*. *Cell* **137**, 685–696 (2009).
59. Marbouty, M. et al. Metagenomic chromosome conformation capture (meta3c) unveils the diversity of chromosome organization in microorganisms. *eLife* **3**, e03318 (2014).

Acknowledgements

This work was initiated during a Hackathon between Institut Pasteur scientists and ENGIE engineers. We would like to thank all the people that allow the organisation of this event especially Anne-Gaelle Coutris, Romain Tchertchian and Olivier Gascuel. Julien Mozziconacci, Frédéric Beckouët and all the members of Spatial Regulation of Genomes unit are thanked for stimulating discussions and feedback. This work used the computational and storage services (TARS cluster) provided by the IT department at Institut Pasteur, Paris. C.M.-D. was supported by the Pasteur—Paris University (PPU) International PhD Program. A.B. works within the framework of a "Mécénat Compétence" contract of the company ENGIE. V.S. is the recipient of a Roux-Cantarini Pasteur fellowship. This research was supported by funding to R.K. from the European Research Council under the Horizon 2020 Program (ERC grant agreement 771813) and by ANR JCJC 2019, "Apollo" allocated to A.C.

Author contributions

All authors contributed to the design of the algorithm. C.M.-D., A.B., L.B., A.C. implemented it. C.M.-D., R.M., L.B. compared to other algorithms. L.B. and A.C. designed strategy for simulations of data. C.M.-D., P.M., R.K. and A.C. analysed biological data and interpreted results. C.M.-D., A.B., L.B., R.K. and A.C. wrote the paper. All authors read and approved the final paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41467-020-19562-7.

Correspondence and requests for materials should be addressed to R.K. or A.C.

Peer review information *Nature Communications* thanks Vera Pancaldi, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary information: Computer vision for pattern detection in chromosome contact maps

Cyril Matthey-Doret[1,2], Lyam Baudry[1,3,♮], Axel Breuer[4,♮], Rémi Montagne[1,3], Nadège Guiglielmoni[1,3], Vittore Scolari[1,3], Etienne Jean[1,3], Arnaud Campeas[4], Philippe Henri Chanut[4], Edgar Oriol[4], Adrien Meot[4], Laurent Politis[4], Antoine Vigouroux[5], Pierrick Moreau[1,3], Romain Koszul[1,3,*], Axel Cournac[1,3,*]

Supplementary Note 1

Fast 2D-convolution via SVD

Optionally, Chromosight's convolution algorithm can be accelerated further by approximating the template. This is done using truncated singular value decomposition (tSVD) to decompose the template into two sets of vectors whose product contain most of the information in the template, while reducing the number of operations needed in the convolution. This note explains the acceleration of 2D-convolution by using the SVD decomposition of the kernel. It was inspired from section 6.4.2 in *Computer and Robot Vision Vol. I* by Haralick and Shapiro (1992) [1].

The general case

Suppose that the contact map $\mathrm{IMG}_{\mathrm{CONT}}$ and the template $\mathrm{IMG}_{\mathrm{TMP}}$ have respectively size $(M_{\mathrm{CONT}}, N_{\mathrm{CONT}})$ and $(M_{\mathrm{TMP}}, N_{\mathrm{TMP}})$.

The convolution of $\mathrm{IMG}_{\mathrm{CONT}}$ by $\mathrm{IMG}_{\mathrm{TMP}}$, noted $\mathrm{IMG}_{\mathrm{CONT}} * \mathrm{IMG}_{\mathrm{TMP}}$, is an array such that

$$\left(\mathrm{IMG}_{\mathrm{CONT}} * \mathrm{IMG}_{\mathrm{TMP}}\right)[i,j] := \sum_{m=0}^{M_{\mathrm{TMP}}-1} \sum_{n=0}^{N_{\mathrm{TMP}}-1} \mathrm{IMG}_{\mathrm{CONT}}\left[i+m, j+n\right] \times \mathrm{IMG}_{\mathrm{TMP}}[m,n] \tag{1}$$

for $i = 1, \ldots, M_{\mathrm{CONT}} - M_{\mathrm{TMP}} + 1$ and $j = 1, \ldots, N_{\mathrm{CONT}} - N_{\mathrm{TMP}} + 1$. Otherwise stated $\mathrm{IMG}_{\mathrm{CONT}*\mathrm{TMP}} := \mathrm{IMG}_{\mathrm{CONT}} * \mathrm{IMG}_{\mathrm{TMP}}$ is an array of size $(M_{\mathrm{CONT}} - M_{\mathrm{TMP}} + 1, N_{\mathrm{CONT}} - N_{\mathrm{TMP}} + 1)$.

The computation of $\left(\mathrm{IMG}_{\mathrm{CONT}} * \mathrm{IMG}_{\mathrm{TMP}}\right)[i,j]$ requires $2\,M_{\mathrm{TMP}}\,N_{\mathrm{TMP}}$ operations, composed of $M_{\mathrm{TMP}}\,N_{\mathrm{TMP}}$ additions and $M_{\mathrm{TMP}}\,N_{\mathrm{TMP}}$ multiplications.

The separable case

Suppose that the template is *separable* i.e. there exists two vectors $\mathrm{U}_{\mathrm{TMP}}$ and $\mathrm{V}_{\mathrm{TMP}}$, with respective size (M_{TMP}) and (N_{TMP}), such that

$$\mathrm{IMG}_{\mathrm{TMP}}[m,n] = \mathrm{U}_{\mathrm{TMP}}[m]\,\mathrm{V}_{\mathrm{TMP}}[n]. \tag{2}$$

The operations in equation (1) can then be re-arranged more efficiently:

$$\left(\mathrm{IMG}_{\mathrm{CONT}} * \mathrm{IMG}_{\mathrm{TMP}}\right)[i,j] = \sum_{m=0}^{M_{\mathrm{TMP}}-1} \left(\sum_{n=0}^{N_{\mathrm{TMP}}-1} \mathrm{IMG}_{\mathrm{CONT}}\left[i+m, j+n\right] \times \mathrm{V}_{\mathrm{TMP}}[n] \right) \times \mathrm{U}_{\mathrm{TMP}}[m] \tag{3}$$

$$= \sum_{m=0}^{M_{\mathrm{TMP}}-1} \mathrm{IMG}_{\mathrm{CONT}*\mathrm{V}_{\mathrm{TMP}}}[i+m, j] \times \mathrm{U}_{\mathrm{TMP}}[m] \tag{4}$$

where

$$\mathrm{IMG}_{\mathrm{CONT}*\mathrm{V}_{\mathrm{TMP}}}[i,j] := \sum_{n=0}^{N_{\mathrm{TMP}}-1} \mathrm{IMG}_{\mathrm{CONT}}\left[i+m, j+n\right] \times \mathrm{V}_{\mathrm{TMP}}[n] \tag{5}$$

The computation of an element $\mathrm{IMG}_{\mathrm{CONT}*\mathrm{TMP}}[i,j]$ costs N_{TMP} multiplications and N_{TMP} additions. According to equation (4), the computation of $\mathrm{IMG}_{\mathrm{CONT}*\mathrm{V}_{\mathrm{TMP}}}[i,j]$ requires $M_{\mathrm{TMP}} + N_{\mathrm{TMP}}$ multiplications and $M_{\mathrm{TMP}} + N_{\mathrm{TMP}}$ additions.

Consequently, the evaluation of $\mathrm{IMG}_{\mathrm{CONT}*\mathrm{TMP}}[i,j]$ costs $2\,(M_{\mathrm{TMP}} + N_{\mathrm{TMP}})$ operations in the separable case, which compares favorably to the $2\,M_{\mathrm{TMP}}\,N_{\mathrm{TMP}}$ operations required in the general case.

The SVD case

Next, suppose that the template has a representation as the sum of K separable kernels:

$$\mathrm{IMG}_{\mathrm{TMP}}[m,n] = \sum_{k=1}^{K} \mathrm{U}_{\mathrm{TMP}}[m,k]\,\mathrm{V}_{\mathrm{TMP}}[k,n]. \tag{6}$$

The number of operations involved in evaluating $\mathrm{IMG}_{\mathrm{CONT}*\mathrm{TMP}}[i,j]$ is $2\,(M_{\mathrm{TMP}} + N_{\mathrm{TMP}})$ for each kernel plus $K-1$ additions necessary to sum up the contribution of each kernel. In total, there are hence $2\,K\,(M_{\mathrm{TMP}} + N_{\mathrm{TMP}}) + K - 1$ operations.

The template $\mathrm{IMG}_{\mathrm{TMP}}$ is not necessarily equal to the superposition of K separable kernels, but it can always be approximated by such a superposition. The (truncated) SVD algorithm discussed below allows to construct such an approximation.

The Singular Value Decomposition (SVD) factorizes any rectangular matrix A of size (M,N) as

$$A = U\,D\,V \tag{7}$$

where U is a (M,M) orthogonal matrix, V is a (N,N) orthogonal matrix and D is a (M,N) matrix all of whose nonzero entries are on the diagonal and are positive.

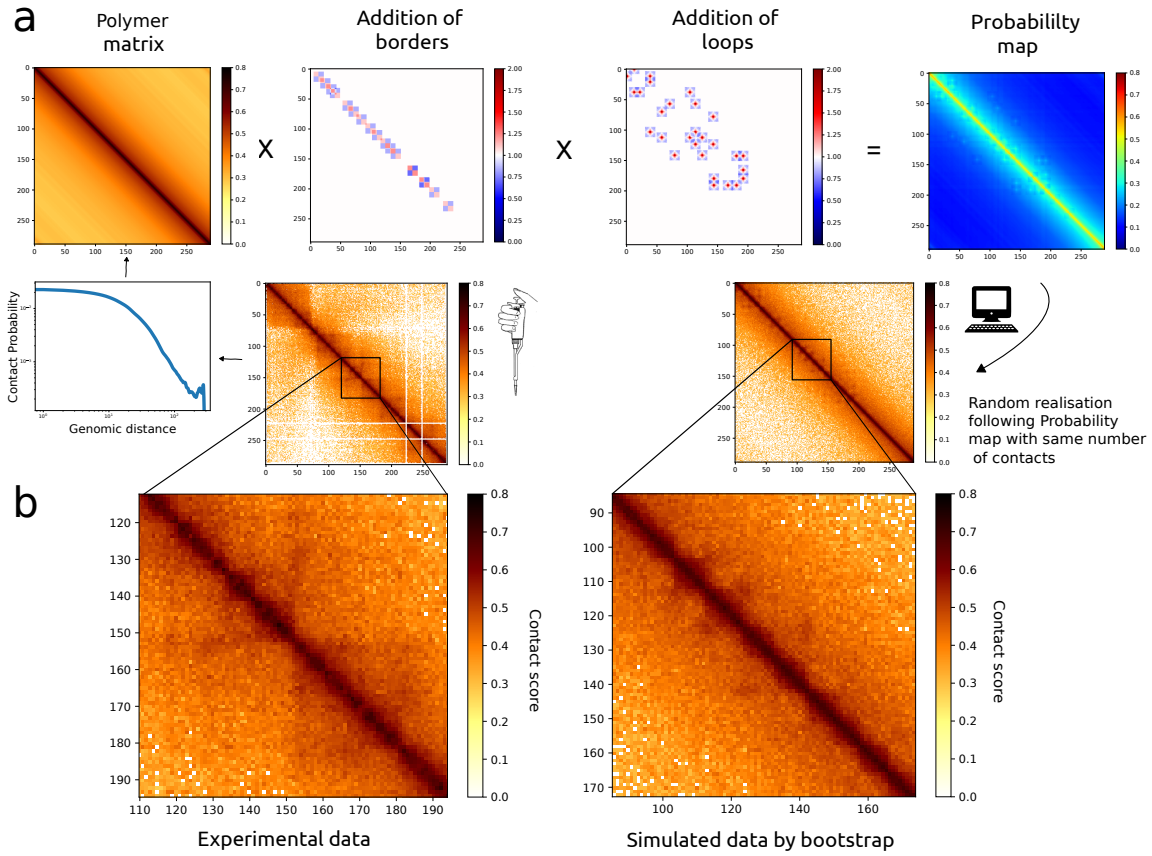
Given any template $\mathrm{IMG}_{\mathrm{TMP}}$, it is possible to approximate it by retaining only the K largest singular values in the SVD of $A = \mathrm{IMG}_{\mathrm{TMP}}$, such that:

$$\mathrm{U}_{\mathrm{TMP}}[:,k] = \sqrt{D[k,k]}\,U[:,k] \tag{8}$$

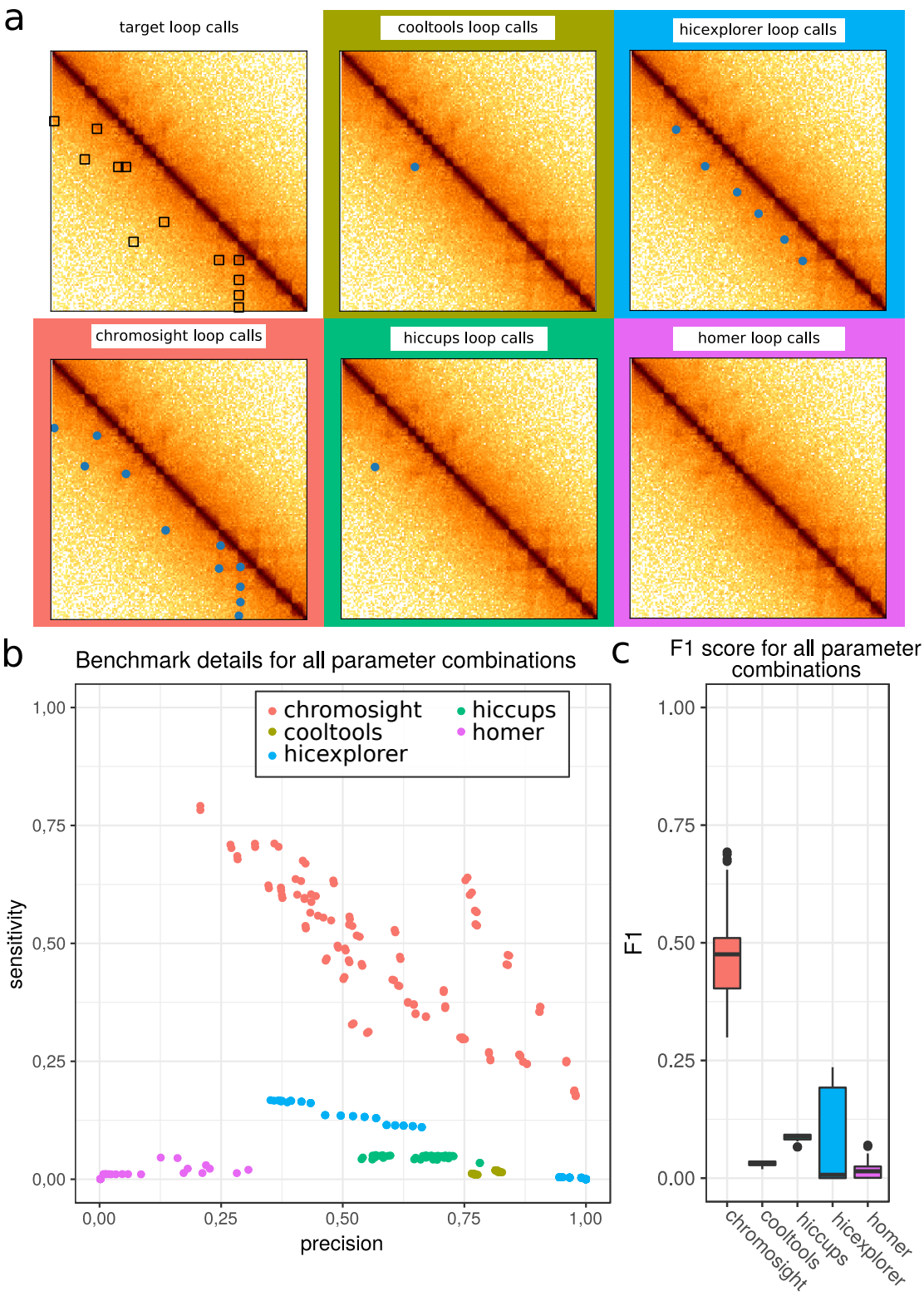
$$\mathrm{V}_{\mathrm{TMP}}[k,:] = \sqrt{D[k,k]}\,V[k,:] \tag{9}$$

$$\tag{10}$$

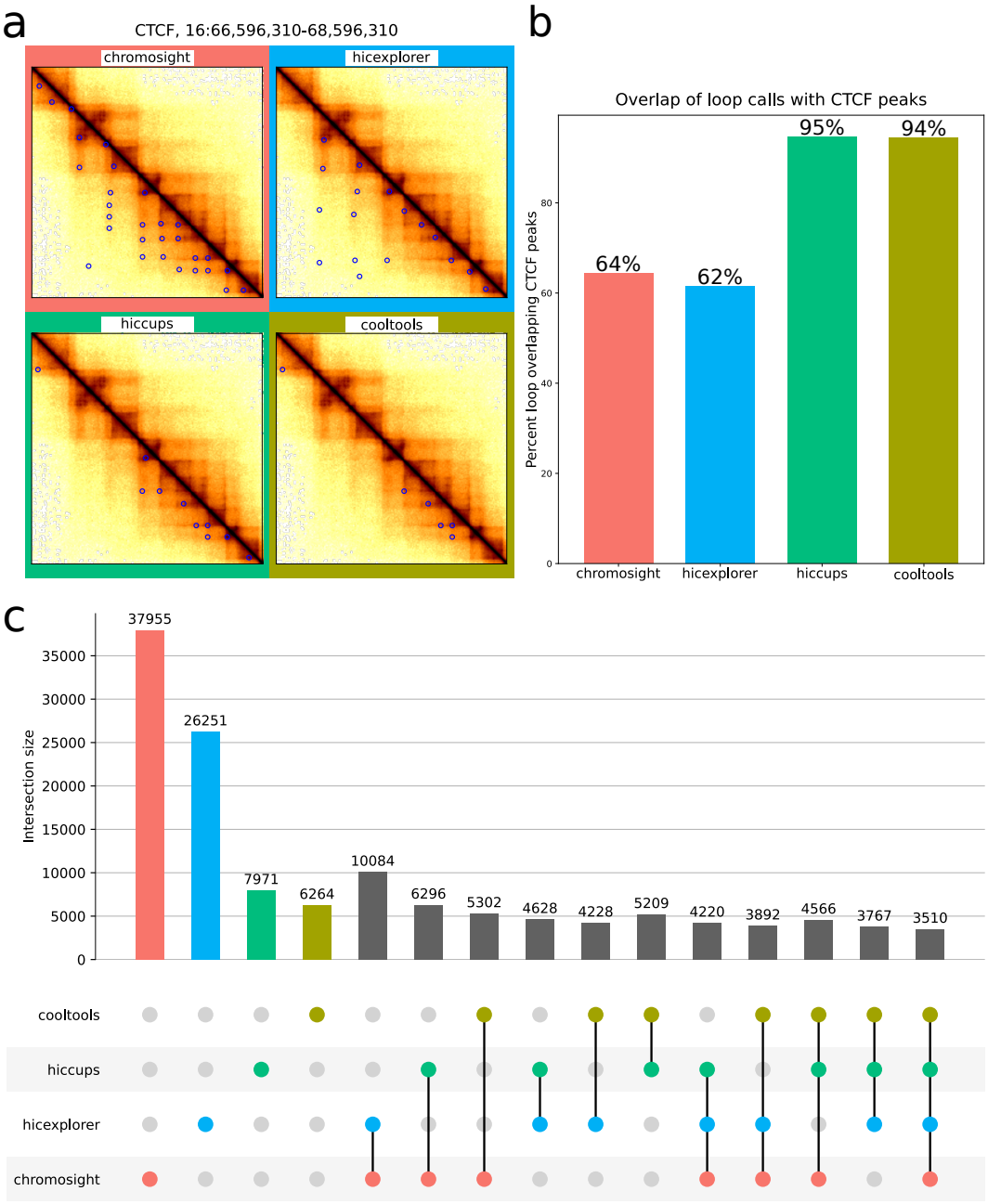
Let's give a toy example of the operations spared by using a SVD approach. Suppose that $M_{\mathrm{TMP}} = N_{\mathrm{TMP}} = 17$, the standard convolution would require $2 \times 17 \times 17 = 578$ operations per point. In contrast, if we use a SVD convolution with $K = 1$, the number of operations reduces to 68, which represents only 12% of the brute force approach. Even with $K = 8$, we are below 50% of the brute force approach.



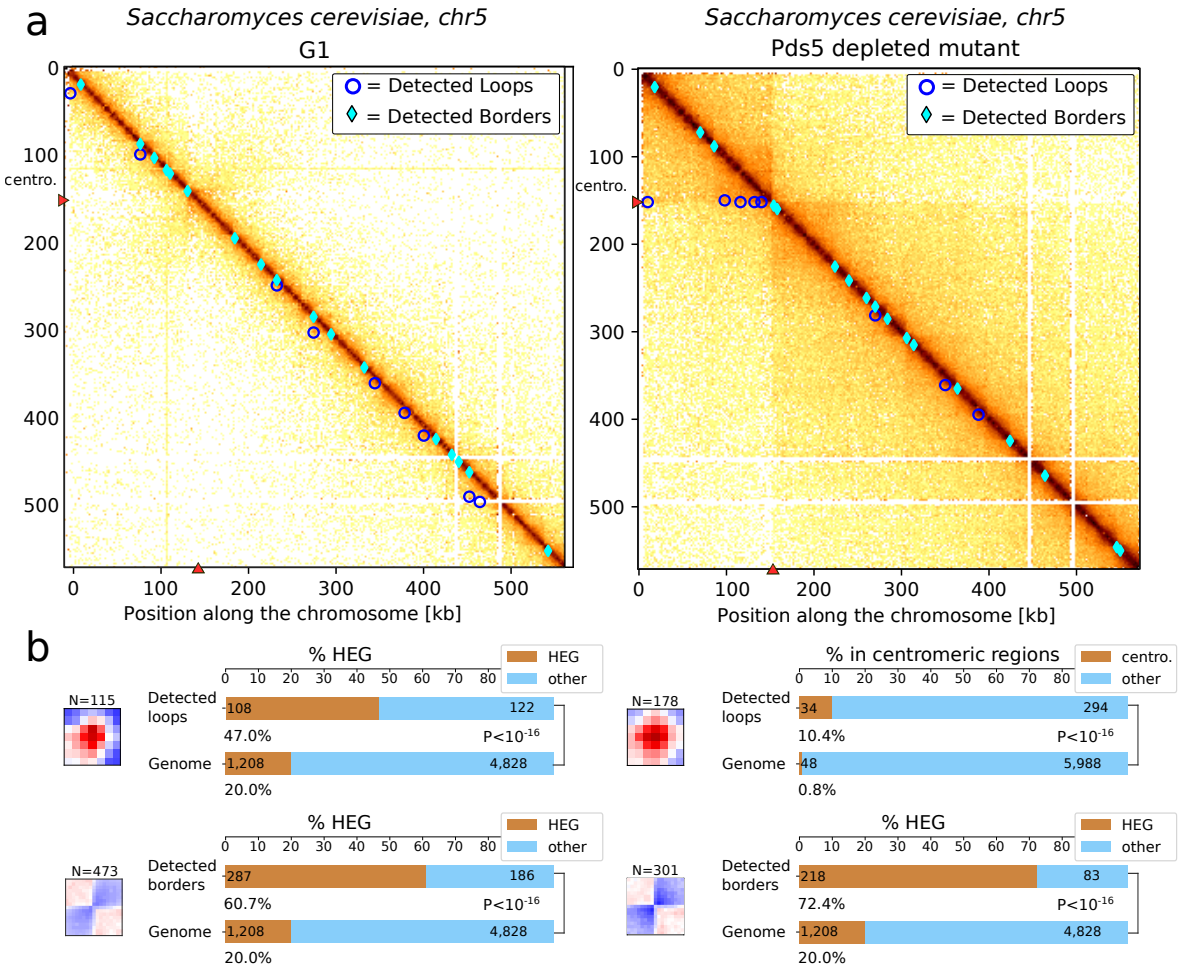
Supplementary Figure 1: Strategy for the generation of simulated contact data for benchmark tests of different loop calling algorithms a, The simulated data were generated with a bootstrap approach based on contact data generated for yeast *S. cerevisiae* in mitotic phase [2]. Three main features of the contact data were extracted: the probability of contact as a function of the genomic distance (P(s), Polymer matrix), presence of borders and presence of loops. The positions and intensity of border and loop patterns were defined thanks to pile-up signals from patterns detected by eye on the contact maps. Their positions were chosen according to a law of probabilities based on experimental data (see Methods). The product of the 3 feature matrices results in a probability matrix (**a**, right). This matrix is used as a probability law to sample contact positions while keeping the same number of reads as the experimental map. **b**, Zoom of contact maps for experimental and simulated data showing patterns of loops and borders. (Icons: [3], Perhelion / Wikimedia Commons, CC-BY-SA-3.0.)



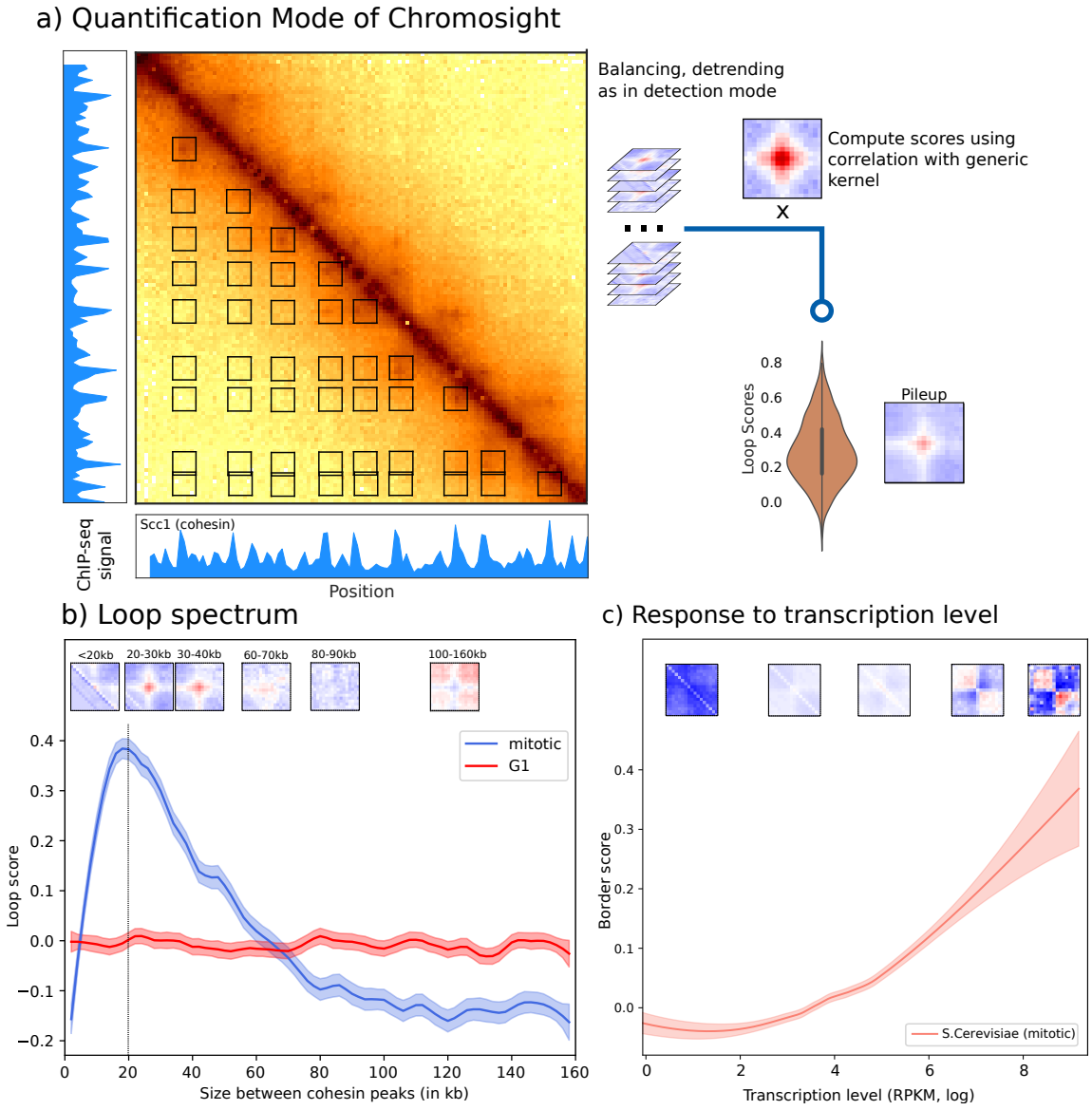
Supplementary Figure 2: Comparison of different loop callers on simulated data. **a**, Example region from a synthetic matrix with real loop calls (top left) and loops detected by all algorithms used in the benchmark using the combinations of parameters which yielded the highest F1 score. **b**, Precision and sensitivity for all algorithms on synthetic matrices, on the whole range of parameters tested. **c**, Distribution of F1 Scores for each algorithm for the range of parameters. Medians are shown as a black band inside boxplots. Hinges show the first and third quartiles and whiskers extend from the hinge to the furthest value within 1.5 times the inter-quartile distance (between first and third quartiles).



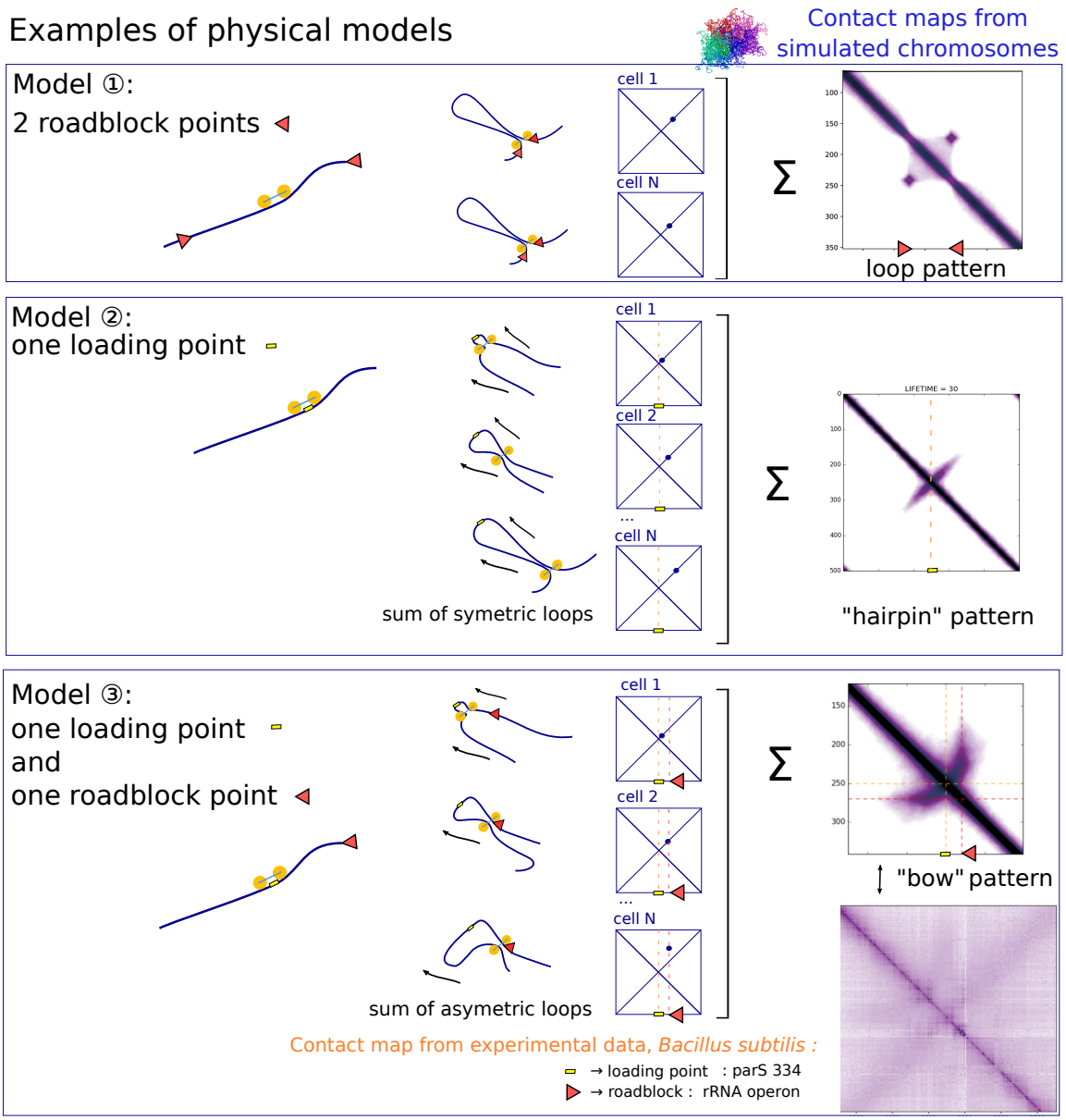
Supplementary Figure 3: Comparison of different loop callers on experimental data. **a**, Contact maps representating a region of +/- 1Mb around the CCCTC-binding factor (CTCF) gene of GM12878 (GSE63525, [4]), at 10kb resolution with coordinates of detected loops for different loop calling softwares, with default parameters. **b**, Proportion of loops with both anchors overlapping CTCF peaks [5]. An overlap is considered if loop anchors and CTCF peaks are within 10kb distance. **c**, Upsetplot showing the number of loops detected in GM12878 by each combination of softwares. Loops are considered identical if they are within 10kb of each other. For each combination of softwares, the intersection ($\cap$) of detected coordinates is shown.



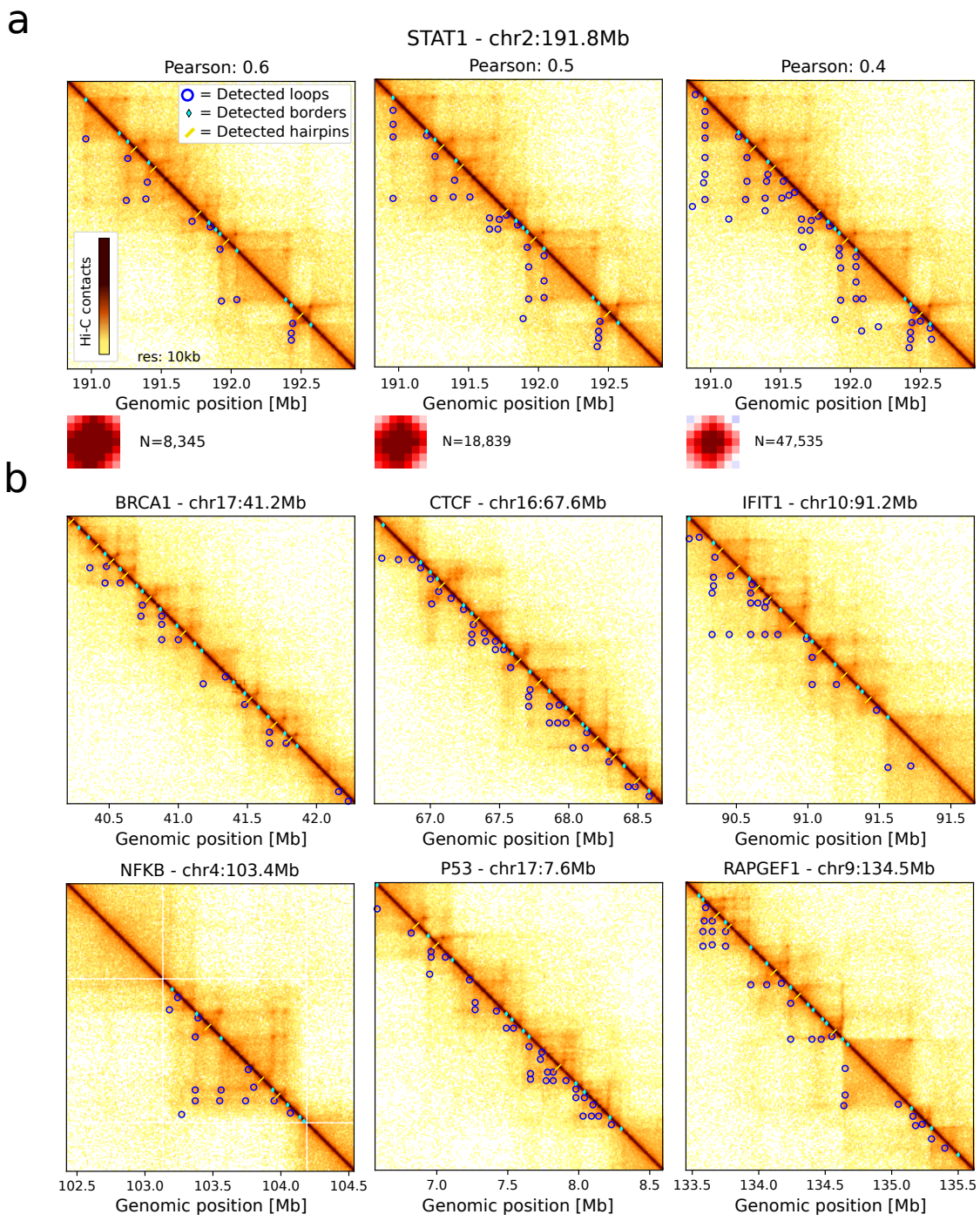
Supplementary Figure 4: Detection of loop and border patterns in yeast contact data a, Detection of loops and borders in Hi-C data of *S. cerevisiae* synchronised in G1 and for a mutant depleted in the protein Pds5 [2]. **b**, Bar plots showing enrichment in highly expressed genes (HEG) for detected loops in G1 and an enrichment in centromeric regions for the Pds5 mutant (Precocious Dissociation of Sisters gene). Bar plots showing enrichment in highly expressed genes for detected borders in G1 and Pds5 mutant. (Fisher test, two-sided)



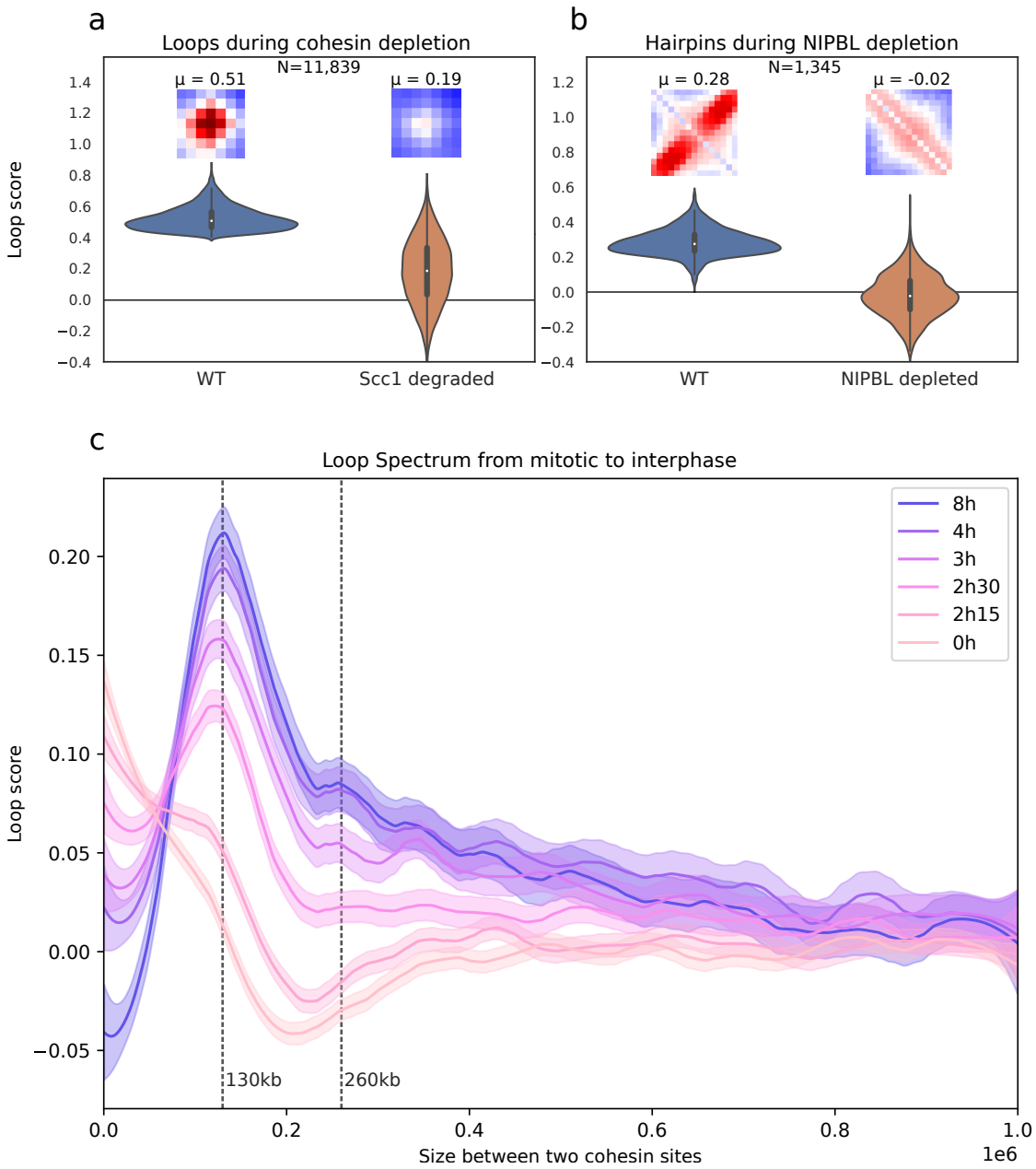
Supplementary Figure 5: Applications of quantification mode on yeast contact data a, Chromosight quantification mode workflow: sub-matrices from certain 2D genomic positions are extracted from balanced and detrended matrices (as in detection mode). Correlation with the kernel is then computed for each sub-matrix and the mean of all the sub-matrices is giving a pileup visualisation. Such 2D coordinates can be, for instance, pairs of protein enrichment peaks called from ChIP-seq data. **b**, Loop spectrum computed for the cohesin peaks network. The loop score is given as a function of distance between cohesin peaks for cells in mitotic state (data from [2]). Curves represent lowess-smoothed data with 95% confidence intervals. **c**, Plot showing the border score as a funtion of transcription levels in *S. cerevisiae*, (contact data and transcriptome data from [6]). The curve represents lowess-smoothed data with 95% confidence intervals.



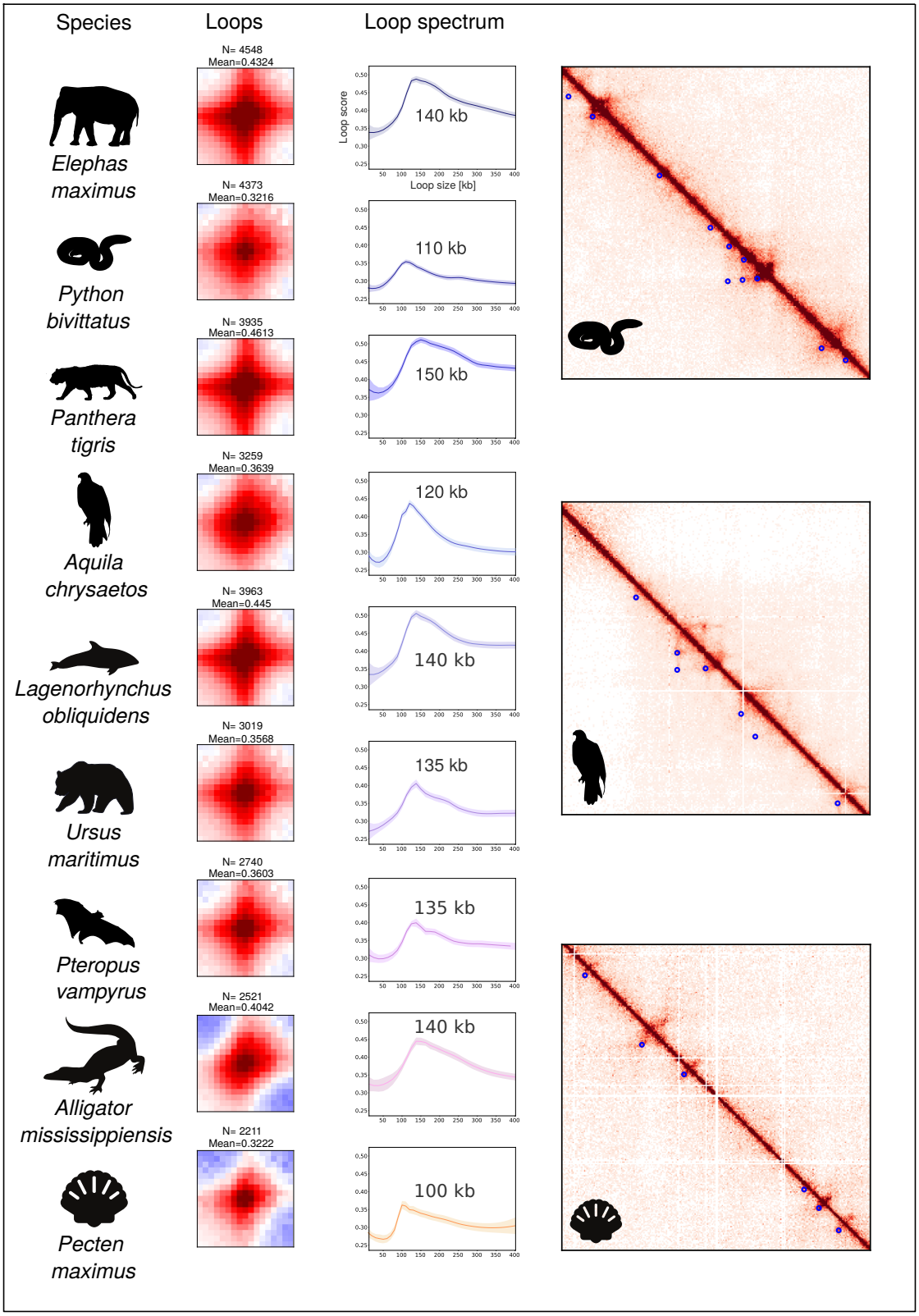
Supplementary Figure 6: Toy models that can link visual patterns and physical models. **Model 1**: loop extruding motors with two roadblock points leading to a loop pattern. **Model 2**: loop extruding motors with a specific loading point leading to a hairpin pattern. **Model 3**: loop extruding motors with a specific loading point and a single roadblock leading to a bow pattern. The bow pattern has been observed in contact data from *Bacillus subtilis* bacteria [7, 8]. By connecting the simulation and experimental contact data, the identified roadblock is a highly transcribed gene, (rDNA operon) and the loading site corresponds to the ParS 334 site. Molecular dynamics simulations were performed using OpenMM [9] and libraries with default parameters of [10].



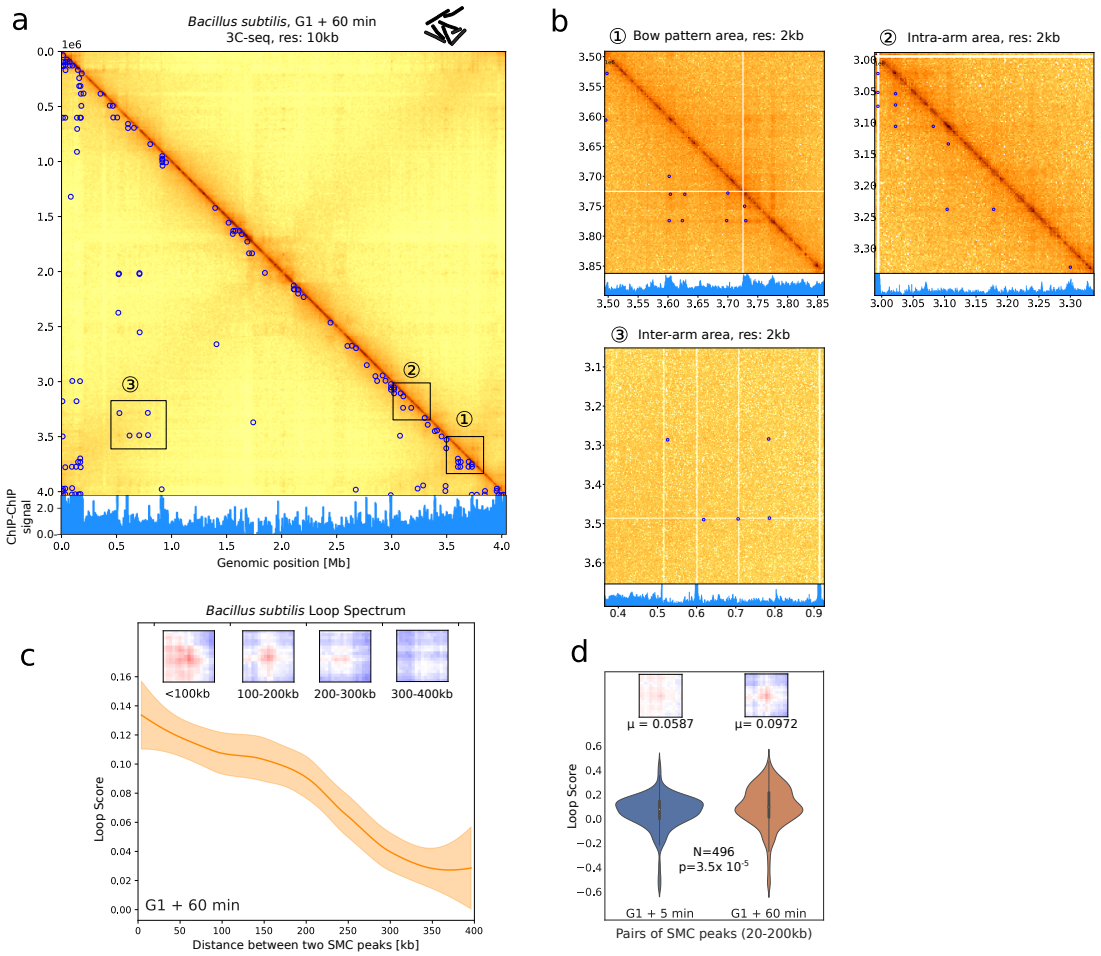
Supplementary Figure 7: Detection of loops, borders, hairpins in human Hi-C data. a, Effect of decreasing Chromosight's Pearson coefficient detection threshold on loop detection. Contact map in the vicinity of STAT1 gene in Hi-C data of lymphoblastoids [11] and total number of detected loops are shown for 3 Pearson threshold values: 0.6, 0.5, 0.4. Decreasing the Pearson threshold allows the detection of weak patterns. **b**, Zoom of contact maps 2 Mb around different genes of interest: BRCA1, CTCF, IFIT1, NFκB, P53, RAPGEF1 in Hi-C data of [11]. Detection done with Pearson coefficient parameter set to 0.5.



Supplementary Figure 8: Applications of quantification of loops and hairpins on human contact data. **a**, Comparison of loop score distributions in WT (*Homo sapiens*, HeLa cells) and in mutant cells depleted in Scc1 [12] for loops detected in WT condition. Associated pileup plots of windows centered on detected loops in WT condition. μ: median of loop scores. **b**, Comparison of hairpin score distributions in WT (*Mus musculus*, liver cells) and in mutant cells depleted in NIPBL [13] for hairpins detected in the WT condition. Associated pileup plots of windows centered on detected hairpins in WT condition. **c**, Loop spectrum showing correlation scores with the loop kernel for pairs of Rad21 ChIP-seq peaks separated by increasing distances, at different time points during release from mitosis into G1 (*Homo sapiens*, HeLa S3 cells) [14]. Curves represent lowess-smoothed data with 95% confidence intervals.



Supplementary Figure 9: Detection of loops, borders, hairpins in various animals from the DNA Zoo project [15]. From left to right, name of the species, associated pileup plots for the called loops and loop spectra computed on the positions of detected loops. The loop spectrum gives the size at which the detected loops have the highest scores. Curves represent lowess-smoothed data with 95% confidence intervals. Zooms on the right show examples of detected loops on Python bivittatus, Aquila chrysaetos and Pecten maximus, respectively. Detection has been performed on a standard laptop with a calculation time of less than 5 min for each pattern per organism. Credits for vectorized images: T. Michael Keesey (*Elephas maximus*), Steven Traver (*Panthera tigris*), Anthony Caravaggi (*Aquila chrysaetos*), Chris Huh (*Lagenorhynchus obliquidens*). Others are in the public domain and all are available on phylopic.org.



Supplementary Figure 10: Detection and quantification of loops in 3Cseq data of Bacillus subtilis. **a**, Detection of loops in 3Cseq data of *Bacillus subtilis* [16]. Genome contact map is shown at 10 kb resolution annotated with detected loops (carried on 2 kb data, 17x17 loop kernel). ChIP-chip signal of Structural Maintenance of Chromosomes proteins (SMC) is plotted under the map. Note that the origin of replication is at the end of the reference genome (bottom right of the contact map). **b**, Zooms of 3 genomic regions highlighted in panel a: in the bow pattern, in intra-arm region or in inter-arms area. **c**, Quantification of loop signal for pairs of SMC peaks for different sizes. Associated pileups of patterns for 4 size ranges are shown above. The curve represents lowess-smoothed data with 95% confidence intervals. **d**, Quantification of loop signals for pairs of SMC peaks between 20 and 200 kb in 2 conditions: G1 + 5 min and G1 + 60 min. Mean of loop scores and associated p-value (Paired Mann Withney U test, two-sided).

software	parameter	values	best F1
chromosight	--window-size	10,15,20	15
chromosight	--min-dist	0,40000	40000
chromosight	--pearson	0.30,0.35,0.40,0.45,0.50	0.30
chromosight	--min-separation	0,50000	0
hicexplorer	--windowSize	10,15,20	10
hicexplorer	--peakWidth	4,5,6,7,8	5
hicexplorer	--peakInteractionsThreshold	10,20,30	10
hicexplorer	--pValuePreselection	0.01,0.02,0.05,0.1	0.05
cooltools	--max-loci-separation	100000,200000,1000000,2000000	2000000
cooltools	--max-nans-tolerated	5,10,15,20	10
cooltools	--dots-clustering-radius	14000,19000,34000,39000	14000
hiccups	-p	1,2,4,6	1
hiccups	-i	6,10,14	14
hiccups	-f	0.05,0.1,0.2	0.1
homer	-poissonLoopGlobalBg	0.0001,0.001	0.001
homer	-poissonLoopLocalBg	0.01,0.05,0.1	0.05
homer	-window	2000,5000,10000	2000

Supplementary Table 1: Parameters used in the benchmark. Name and values of all parameters tested in the benchmark for each software. The best F1 column indicates which value yielded the best F1 score on the simulated dataset.

Organism	Experiment type	Figure	Ref	Identifier
S. cerevisiae	Hi-C, mitotic (nocodazole synchr.)	**Fig** 2	[6]	SRR7706226, SRR7706227
S. cerevisiae	Hi-C, G1 (alpha factor synchr.)	**Fig** 2	[2]	SRR8769554
S. pombe	Hi-C, Mitotic phase, 40 min	**Fig** 2	[17]	SRR5149256
S. cerevisiae	Hi-C, Pds5 depleted, mitotic (cdc20 synchr.)	**Sup Fig** 4	[2]	SRR8769553
H. sapiens	Hi-C, GM12878, asynchronous	**Fig** 3	[11]	SRR6675327
H. sapiens	Hi-C, HeLa cells, WT	**Sup Fig** 8	[12]	GSM2747745
H. sapiens	Hi-C, HeLa cells, depleted in Scc1	**Sup Fig** 8	[12]	GSM2747747
M. musculus	Hi-C, liver cells	**Sup Fig** 8	[13]	GSE93431
M. musculus	Hi-C, liver cells, depleted in NIPBL	**Sup Fig** 8	[13]	GSE93431
H. sapiens	Hi-C, HeLa cells during cell cycle (R2, T0, T2h15, T2h30, T3h, T4h, T8h)	**Sup Fig** 8	[14]	GSM3909703 GSM3909697 GSM3909696 GSM3909694 GSM3909691 GSM3909686
B. subtilis	3Cseq in G1 + 60 min	**Fig** 3	[7]	SRR2214080
B. subtilis	3Cseq in G1 + 5 min	**Sup Fig** 10	[7]	SRR2214069
Epstein Barr Virus	ChIA-PET of CTCF in GM12878 cells	**Fig** 3	[18]	SRR2312566
H. sapiens	In situ ChiA-PET, GM12878, asynchronous	**Fig** 4	[19]	4DNFIMH3J7RW
H. sapiens	DNA SPRITE, GM12878, asynchronous	**Fig** 4	[20]	4DNFIUOOYQC3
H. sapiens	HiChIP , GM12878, asynchronous	**Fig** 4	[21]	GSE80820_HiChIP_GM_cohesin.hic
H. sapiens	Micro-C , hESC, asynchronous	**Fig** 4	[22]	4DNFI9FVHJZQ
C. albicans	Hi-C, asynchronous	**Fig** 5	[23]	SRR3381672
E. maximus	Hi-C, asynchronous	**Sup Fig** 9	[15]	Elephas_maximus rawchrom.hic
P. bivittatus	Hi-C, asynchronous	**Sup Fig** 9	[24]	Python_bivittatus rawchrom.hic
P. tigris	Hi-C, asynchronous	**Sup Fig** 9	[25]	Panthera_tigris rawchrom.hic
A. chrysaetos	Hi-C, asynchronous	**Sup Fig** 9	[26]	Aquila_chrysaetos rawchrom.hic
L. obliquidens	Hi-C, asynchronous	**Sup Fig** 9	[15]	Lagenorhynchus_obliquidens rawchrom.hic
U. maritimus	Hi-C, asynchronous	**Sup Fig** 9	[27]	Ursus_maritimus rawchrom.hic
P. vampyrus	Hi-C, asynchronous	**Sup Fig** 9	[28]	Pectorus_vampyrus rawchrom.hic
A. mississippiensis	Hi-C, asynchronous	**Sup Fig** 9	[29][30]	Alligator_mississi-ppiensis rawchrom.hic
P. maximus	Hi-C, asynchronous	**Sup Fig** 9	[31]	Pecten_maximus rawchrom.hic

Supplementary Table 2: Different contact datasets analysed in the present study. The last column indicates either the identifier for the raw reads available on the Short Read Archive server (SRA) (https://www.ncbi.nlm.nih.gov/sra), the identifier of the .cool files accessible on the Gene Expression Omnibus server (GEO) https://www.ncbi.nlm.nih.gov/geo or the name of hic files from DNA zoo project available on https://www.dnazoo.org/assemblies [15] from which the analysis were made. mcool files coming from 4DN portal were downloaded from the server https://data.4dnucleome.org [32].

Organism	Experiment type	Figure	Ref	Identifier
S. cerevisiae	RNA-seq, mitotic (nocodazole synchr.)	**Fig** 2	[6]	SRR7692240
S. cerevisiae	ChIP-seq, Scc1PK9 IP G1 releasing 60min	**Fig** 2	[33]	SRR2065097, SRR2065092
H. sapiens	ChIP-seq CTCF	**Fig** 3	[5]	wgEncodeAwgTfbsBroad Gm12878CtcfUniPk.narrowPeak
H. sapiens	ChIP-seq RAD21	**Fig** 3	[5]	wgEncodeAwgTfbsHaib Gm12878Rad21V0416101UniPk
H. sapiens	ChIP-seq NIPBL	**Fig** 3	[5]	GSM2443453_GM12878_NIPBL Rep1_ 2WCE_Narrow_Peaks_peaks.narrowPeak
B. subtilis	ChIP-chip of SMC	**Fig** 3	[34]	GSE14693
Epstein Barr Virus	ChIP-seq CTCF	**Fig** 3	[35]	SRR036682
Epstein Barr Virus	ChIP-seq RAD21	**Fig** 3	[18]	SRR2312570

Supplementary Table 3: Other genomic datasets used in the present study. The last column indicates either the identifier for the raw reads available on the Short Read Archive server (SRA) (https://www.ncbi.nlm.nih.gov/sra), the identifier of the ChIP-chip files accessible on the Gene Expression Omnibus server (GEO) https://www.ncbi.nlm.nih.gov/geo or the identifier of ChIP-seq peak files available on http://genome.ucsc.edu.

References

1. Haralick, R. M. & Shapiro, L. G. *Computer and Robot Vision* 1st. ISBN: 0201569434 (Addison-Wesley Longman Publishing Co., Inc., USA, 1992).
2. Dauban, L. *et al.* Regulation of Cohesin-Mediated Chromosome Folding by Eco1 and Other Partners. *Molecular Cell* **77,** 1279–1293.e4. https://doi.org/10.1016/j.molcel.2020.01.019 (Mar. 2020).
3. OpenWetWare. *BISC209/S13: Use and Care of Micropipets — OpenWetWare,* [Online; accessed 5-October-2020]. 2013. https://openwetware.org/mediawiki/index.php?title=BISC209/S13:_Use_and_Care_of_Micropipets&oldid=666811.
4. Rao, S. S. P. *et al.* A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. eng. *Cell* **159,** 1665–1680. ISSN: 1097-4172 (Dec. 2014).
5. Karolchik, D. The UCSC Table Browser data retrieval tool. *Nucleic Acids Research* **32,** 493D–496. https://doi.org/10.1093/nar/gkh103 (Jan. 2004).
6. Garcia-Luis, J. *et al.* FACT mediates cohesin function on chromatin. *Nat. Struct. Mol. Biol.* **26,** 970–979 (Oct. 2019).
7. Marbouty, M. *et al.* Condensin-and replication-mediated bacterial chromosome folding and origin condensation revealed by Hi-C and super-resolution imaging. *Molecular cell* **59,** 588–602 (2015).
8. Banigan, E. J., van den Berg, A. A., Brandão, H. B., Marko, J. F. & Mirny, L. A. Chromosome organization by one-sided and two-sided loop extrusion. *eLife* **9.** ISSN: 2050-084X. http://dx.doi.org/10.7554/eLife.53558 (Apr. 2020).
9. Eastman, P. *et al.* OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. *PLOS Comp. Biol.13(7): e1005659.* (2017).
10. Goloborodko, A., Imakaev, M. V., Marko, J. F. & Mirny, L. Compaction and segregation of sister chromatids via active loop extrusion. *Elife* **5,** e14864 (2016).
11. Ghurye, J. *et al.* Integrating Hi-C links with assembly graphs for chromosome-scale assembly. *PLoS Comput. Biol.* **15,** e1007273 (Aug. 2019).
12. Wutz, G. *et al.* Topologically associating domains and chromatin loops depend on cohesin and are regulated by CTCF, WAPL, and PDS5 proteins. *EMBO J.* **36,** 3573–3599 (Dec. 2017).
13. Schwarzer, W. *et al.* Two independent modes of chromatin organization revealed by cohesin removal. *Nature* **551,** 51–56 (Nov. 2017).
14. Abramo, K. *et al.* A chromosome folding intermediate at the condensin-to-cohesin transition during telophase. *Nat. Cell Biol.* **21,** 1393–1402 (Nov. 2019).
15. Dudchenko, O. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science* **356,** 92–95. https://doi.org/10.1126/science.aal3327 (Mar. 2017).
16. Marbouty, M. *et al.* Metagenomic chromosome conformation capture (meta3C) unveils the diversity of chromosome organization in microorganisms. eng. *Elife* **3,** e03318. http://dx.doi.org/10.7554/eLife.03318 (2014).

17. Tanizawa, H., Kim, K.-D., Iwasaki, O. & Noma, K.-I. Architectural alterations of the fission yeast genome during the cell cycle. *Nat. Struct. Mol. Biol.* **24,** 965–976 (Nov. 2017).

18. Tang, Z. *et al.* CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell* **163,** 1611–27 (Dec. 2015).

19. Li, X. *et al.* Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nature Protocols* **12,** 899–915. ISSN: 1750-2799. http://dx.doi.org/10.1038/nprot.2017.012 (Mar. 2017).

20. Quinodoz, S. A. *et al.* Higher-Order Inter-chromosomal Hubs Shape 3D Genome Organization in the Nucleus. *Cell* **174,** 744–757.e24. ISSN: 0092-8674. http://dx.doi.org/10.1016/j.cell.2018.05.024 (July 2018).

21. Mumbach, M. R. *et al.* HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods* **13,** 919–922. ISSN: 1548-7105. http://dx.doi.org/10.1038/nmeth.3999 (Sept. 2016).

22. Krietenstein, N. *et al.* Ultrastructural Details of Mammalian Chromosome Architecture. *Molecular Cell* **78,** 554–565.e7. ISSN: 1097-2765. http://dx.doi.org/10.1016/j.molcel.2020.03.003 (May 2020).

23. Burrack, L. S. *et al.* Neocentromeres Provide Chromosome Segregation Accuracy and Centromere Clustering to Multiple Loci along a Candida albicans Chromosome. *PLOS Genetics* **12** (ed Mellone, B. G.) e1006317. ISSN: 1553-7404. http://dx.doi.org/10.1371/journal.pgen.1006317 (Sept. 2016).

24. Castoe, T. A. *et al.* The Burmese python genome reveals the molecular basis for extreme adaptation in snakes. *Proceedings of the National Academy of Sciences* **110,** 20645–20650. https://doi.org/10.1073/pnas.1314475110 (Dec. 2013).

25. Cho, Y. S. *et al.* The tiger genome and comparative analysis with lion and snow leopard genomes. *Nature Communications* **4.** https://doi.org/10.1038/ncomms3433 (Sept. 2013).

26. Bussche, R. A. V. D., Judkins, M. E., Montague, M. J. & Warren, W. C. A Resource of Genome-Wide Single Nucleotide Polymorphisms (Snps) for the Conservation and Management of Golden Eagles. *Journal of Raptor Research* **51,** 368–377. https://doi.org/10.3356/jrr-16-47.1 (Sept. 2017).

27. Liu, S. *et al.* Population Genomics Reveal Recent Speciation and Rapid Evolutionary Adaptation in Polar Bears. *Cell* **157,** 785–794. https://doi.org/10.1016/j.cell.2014.03.054 (May 2014).

28. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478,** 476–482. https://doi.org/10.1038/nature10530 (Oct. 2011).

29. John, J. A. S. *et al.* Sequencing three crocodilian genomes to illuminate the evolution of archosaurs and amniotes. *Genome Biology* **13.** https://doi.org/10.1186/gb-2012-13-1-415 (Jan. 2012).

30. Rice, E. S. *et al.* Improved genome assembly of American alligator genome reveals conserved architecture of estrogen signaling. *Genome Research* **27,** 686–696. https://doi.org/10.1101/gr.213595.116 (Jan. 2017).

31. Kenny, N. J. *et al.* The Gene-Rich Genome of the Scallop Pecten maximus. https://doi.org/10.1101/2020.01.08.887828 (Jan. 2020).

32. Dekker, J. *et al.* The 4D nucleome project. *Nature* **549,** 219–226. https://doi.org/10.1038/nature23884 (Sept. 2017).

33. Hu, B. *et al.* Biological chromodynamics: a general method for measuring protein occupancy across the genome by calibrating ChIP-seq. *Nucleic Acids Research,* gkv670. https://doi.org/10.1093/nar/gkv670 (June 2015).

34. Gruber, S. & Errington, J. Recruitment of condensin to replication origin regions by ParB/SpoOJ promotes chromosome segregation in B. subtilis. eng. *Cell* **137,** 685–696. http://dx.doi.org/10.1016/j.cell.2009.02.035 (May 2009).

35. McDaniell, R. *et al.* Heritable Individual-Specific and Allele-Specific Chromatin Signatures in Humans. *Science* **328,** 235–239. https://doi.org/10.1126/science.1184655 (Mar. 2010).

1.3 Change detection across biological conditions

Change detection is a common issue in the field of signal processing and remote sensing. Given two or more input signals such as images, we want to find portions that differ between the two inputs. This principle can also be applied to Hi-C contacts, where we can detect regions of contact maps that differ between biological conditions.

Change detection in Hi-C contact maps is required whenever we want to identify genomic regions whose spatial organization is altered between two conditions.

Many approaches can be taken to detect these changes. Some of them, such as diffhic, formulate the problem similarly to a differential expression RNAseq analysis using contact counts instead of read counts. This approach has the benefit of being straightforward, but it only finds contact increase. Local increase in contacts can represent a specific spatial interactions, but also differential accessibility or insulation. which could be caused by a number of different phenomena.

We developed pareidolia, a software package for change detection with an apriori on the type of signal to detect. The method is "supervised" in the sense that it requires a kernel representing the feature of interest. Pareidolia relies on Chromosight's backend to convert the contact map of each condition into a map of correlation coefficients representing similarity with the feature of interest. Change detection is then performed on these coefficients. As a consequence, rather than looking for contacts increase, pareidolia looks for changes in feature intensity, such as border sharpness or looping intensity.

1.3.1 Pareidolia algorithm

Pareidolia works by comparing one or several samples issued from two conditions such as treatments or timepoints.

Assuming two conditions $t = t_0, t_1$, where multiple samples (replicates) $(r_1, r_2, ..., r_R)$ can share the same condition. The contact matrix from each sample $H_{r,t}$ is first convoluted with a kernel K representing the pattern of interest, using Chromosight's API. In the resulting matrix M, each value $M_{r,t}[i, j]$ is a Pearson correlation coefficient with the kernel. M is computed as described in equation

Change detection can then be performed on the correlation maps using two different methods.

The first method is inspired by median filtering-based background formation (Fig. II.G). We generate a background matrix for each condition (timepoint), whose values are defined as the median of all replicates' correlation matrices from that condition (Eq. 1.1)

$$B_t[i,j] = median(M_{1,t}[i,j], M_{2,t}[i,j], ..., M_{R,t}[i,j]) \tag{1.1}$$

We then compute the matrix of squared errors V between each replicate and their condition's median background (Eq. 1.2). This technical variability (among replicates) is then used to filter out noisy regions. This is done by generating a signal-to-noise ratio (SNR) map S (Eq. 1.3) and applying a threshold to it.

$$V_t = \sum_{i=1}^{r} (M_{i,t} - B_t)^2 \tag{1.2}$$

$$S = \frac{1}{t} \sum_{i=0}^{t} \frac{B_i}{V_i} \tag{1.3}$$

Pixels in noisy regions (below the SNR threshold) are then removed. To further reduce false detections, a density filter can be applied to discard detections located in low coverage regions. If a kernel of size K was used for detection, the proportion of nonzero values within a window of size K surrounding each position must be above a threshold to be considered.

Pareidolia can either return the change intensity on a set of predetermined loop positions provided as input, or perform *de-novo* detection of differential loops on the Hi-C map. For detection, pareidolia uses Chromosight's implementation of the connected component labelling algorithm for sparse matrices.

1.3.2 Results on experimental data

To showcase the performances of Pareidolia, we used it to measure loop changes upon depletion of CTCF in murine cells using published data [97]. CTCF acts as a roadblock for the motor protein cohesin which travels along the chromatin fiber. Cohesin accumulates at CTCF binding sites, forming stable chromatin loops between pairs of CTCF binding sites.

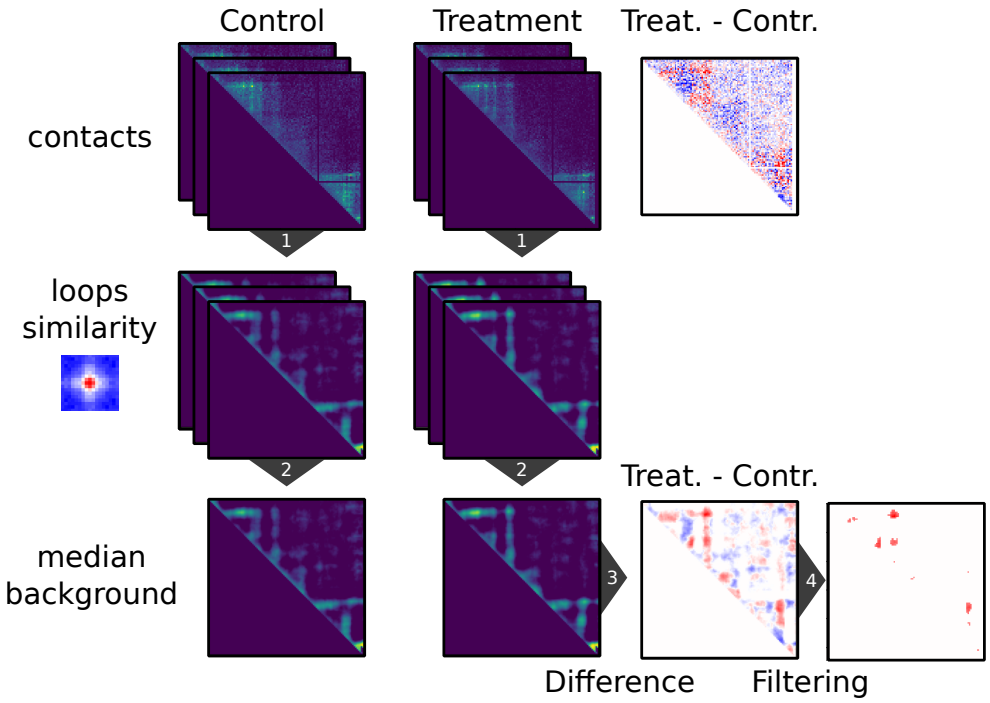


Fig. II.G: Pareidolia algorigthm. From top to bottom: The Hi-C contact maps of several replicates in two conditions are shown, as well as the difference between conditions. Chromosight's convolution algorithm is used on each sample (1) to generate a map of correlation coefficient with the kernel of interest (loops in this case). For each condition, a median background is computed among replicates (2). The difference between these two background is then extracted (3) and filtered (4) using a SNR threshold and a percentile threshold.

These looping interactions have been shown to disappear in the absence of cohesin [98] or CTCF [97]. Here we show an example use of Pareidolia to quantify these 3D changes.

The dataset used here consists of CTCF-AID mutant mouse embryonic embryonic stem cells (ES-E14TG2a). When auxin treatment is applied, the CTCF-AID recombinant protein is degraded. We use pareidolia to compare chromatin loops before and after auxin treatments, using all 4 samples (2 replicates per condition).

With default parameters, pareidolia identifies a total of 2997 disappearing differential loops and 845 appearing loops (Fig. II.H).

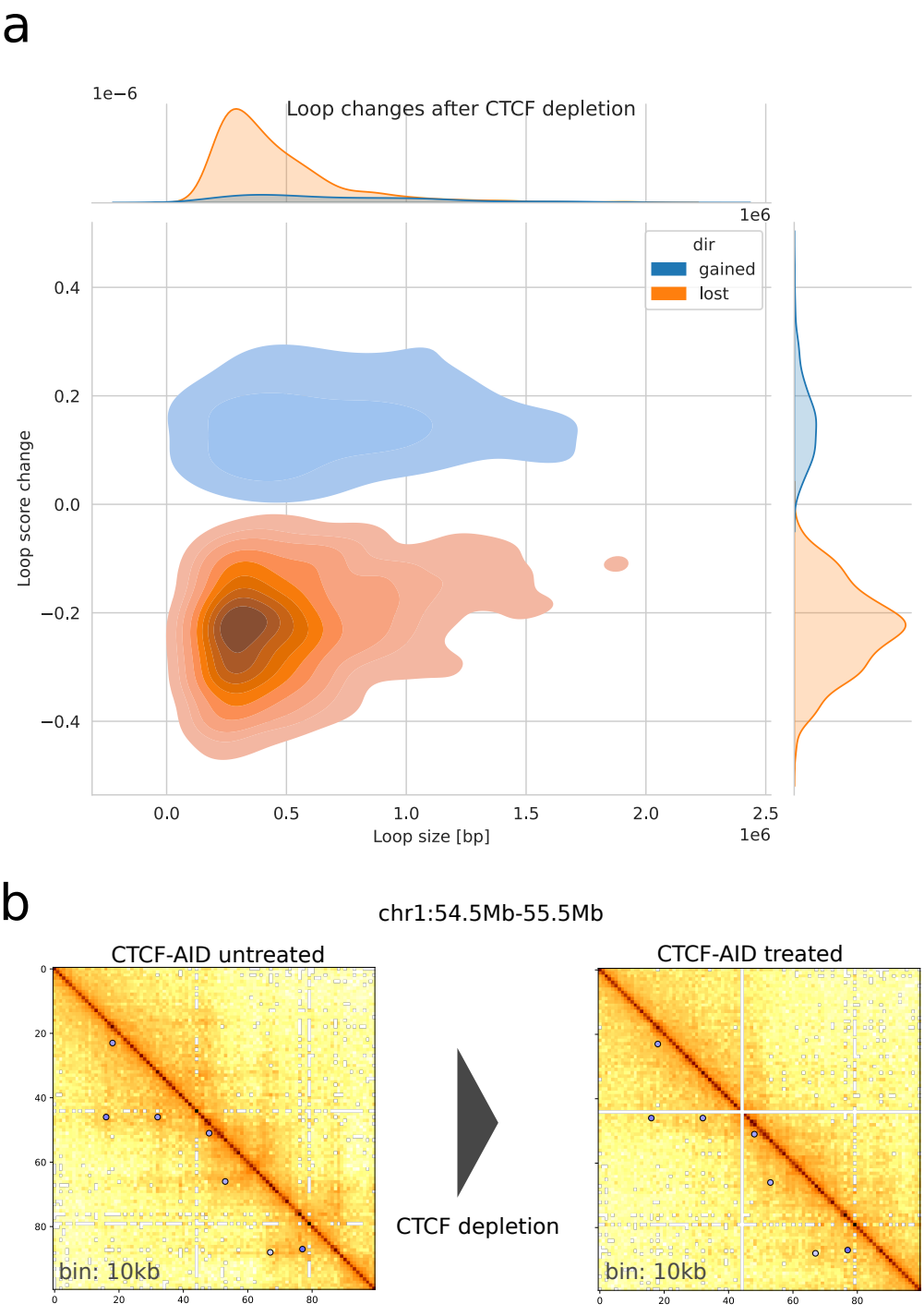


Fig. II.H: Pareidolia results on CTCF degradation experiments from [97]: **a:** Distribution of chromatin loop change and size upon CTCF depletion as detected by pareidolia. **b:** Zoom on a region of the Hi-C map from mouse chromosome 1. Disappearing loops detected by pareidolia are highlighted in blue. For visualization purpose, replicates were merged in Hi-C matrices shown. Processed data retrieved from `https://data.4dnucleome.org`, accessions 4DNES87HWQAX and 4DNES7UKQHOX.

2 Infection of *Acanthamoeba castellanii* by *Legionella pneumophila*

Legionella pneumophila alters its host signal transduction, metabolism and gene regulation upon infection. In addition to all these changes, it also affects host histone marks, which are known to be related to gene regulation and genome architecture. In this chapter, we investigate the genome structure of *A. castellanii* and how it is affected during infection by *L. pneumophila*.

2.1 Genome assembly

As with most other genomics techniques, a prerequisite of Hi-C analyses is to have a high quality reference genome with clearly delimited chromosomes. At the time of writing, the A. castellanii reference genome is split into 384 scaffolds which do not represent chromosomes. This prompted us to generate a chromosome-level genome assembly.

2.2 Genome architecture of *A. castellanii*

The genome of *A. castellanii* is 45Mbp long and pulsed field gel electrophoresis experiments suggested that it has around 20 chromosomes [99]. The 5S ribosomal DNAs are dispered throughout all chromosomes, unlike in most species, where they are clustered in tandem repeats.

2.3 Strains comparison

Several strains of *A. castellanii* have been isolated throughout history.These strains may originate from different ecological niche of geographical location and have been cultivated in labs for long preriods. As a result, they can differ in various phenotypes, including susceptibility to infection. Comparing such divergent strains can help us understand what genomic features are important for pathogen susceptibility.

2.4 Changes during infection

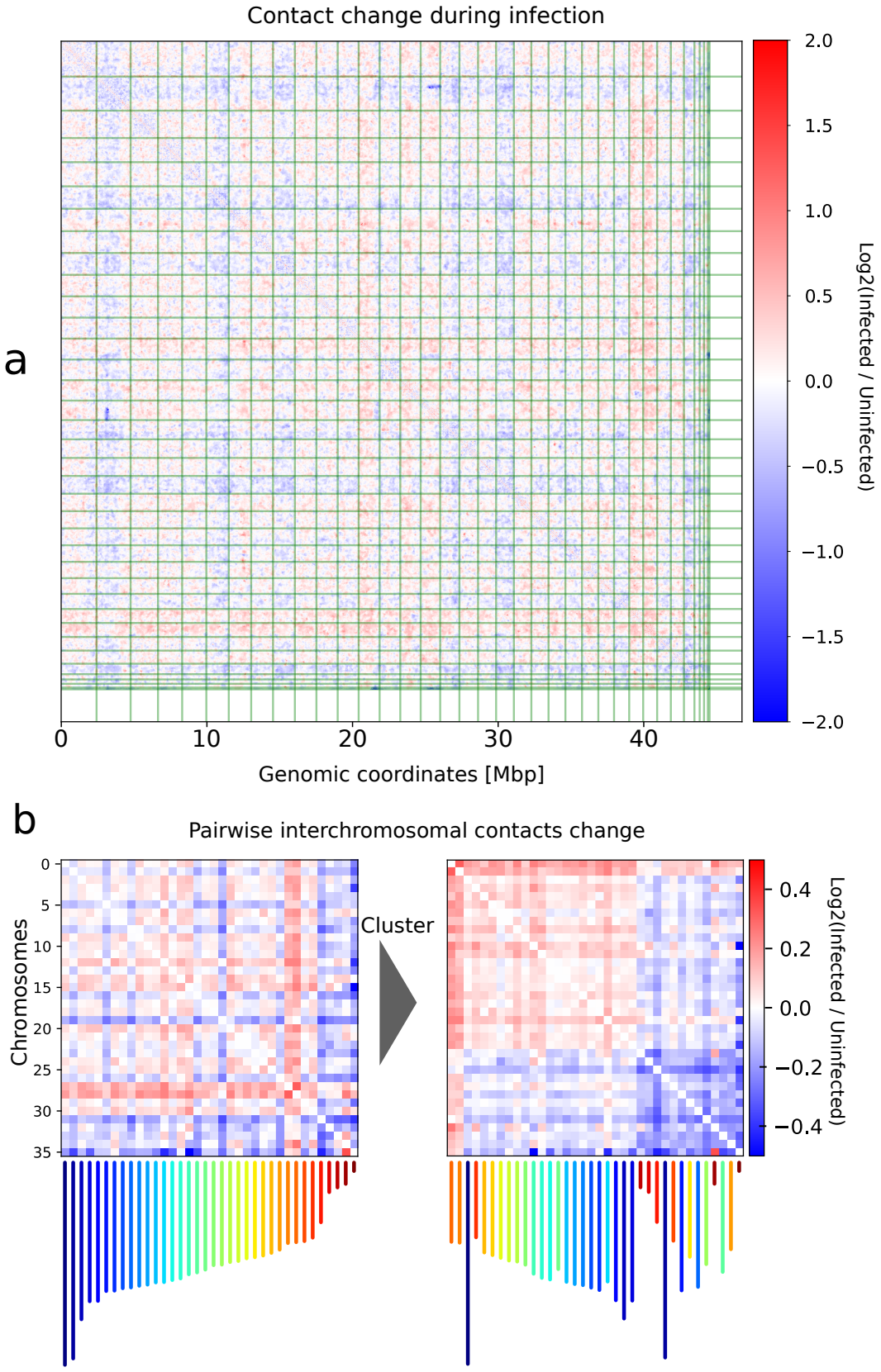


Fig. II.I: Interchromosomal contacts changes during infection by *Legionella*: **a:** Whole genome contact map of *A. castellanii str. C3* and log ratio map of *L. pneumophila*-infected over control. Green lines delimit scaffolds. **b:** Mean contact changes during infection between all pairs of chromosomes before (left) and after(right) clustering chromosomes.

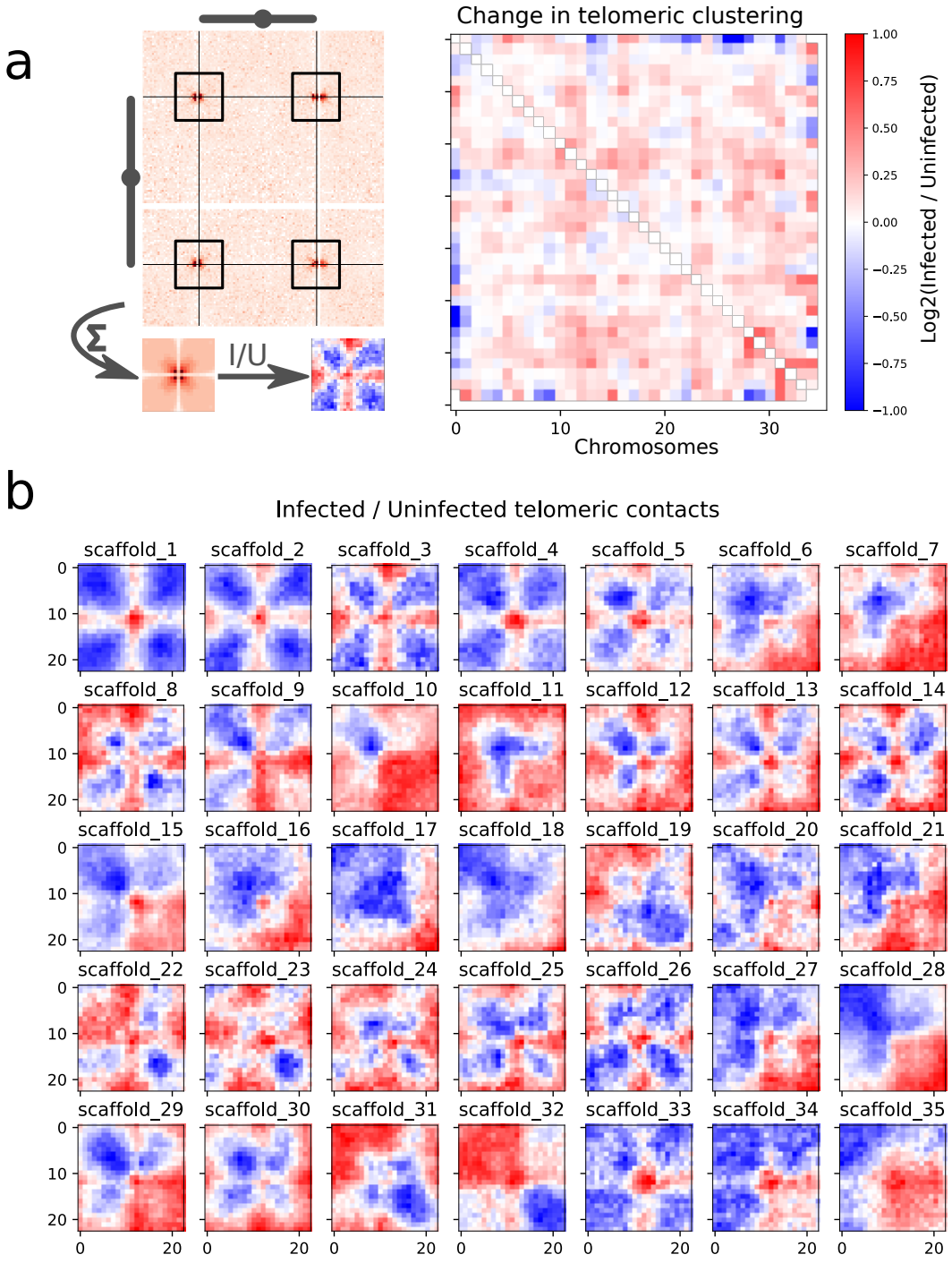


Fig. II.J: Telomeric interactions changes during infection by *Legionella*: **a:** Illustration of the extraction process for inter-telomeric contact from two chromosomes. All telomeric windows are into a pileup for a given chromosome and the ratio between infected (I) and uninfected (U) pileups can be visualised. Change in telomeric pattern intensity during infection for all pairs of chromosomes. The intensity is the Pearson correlation coefficient between the telomeric pileup and each telomeric window. Each value in the matrix represents a pair of chromosome, consisting of the average of 4 telomeric windows. Intrachromosomal windows are excluded. **b:** Telomeric contacts ratio for all chromosomes. For each chromosome, the pileup contains the windows of its telomeric interactions with all other chromosomes.

3 Infection of murine bone macrophages by *Salmonella enterica*

3.1 Large scale changes

3.2 Picking up local differences

3.3 Integrating gene expression with 3D changes

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

4 Viral insertions in the human genome

Viruses represent another type of host pathogen interactions. Some of those viruses, including the hepatitis B virus (HBV), have the ability to insert their genetic material into the genome of their host. Depending on the genomic landscape around those insertions, it can have various effects on the host cells.

It is known that hepatocellular carcinoma (HCC), a type of liver cancer, are frequently associated with HBV insertions.

Here we investigate the location of those viral insertions in multiple HCC cell lines, the epigenetic marks present at those regions and how they affect the local 3D structure.

4.1 Genome structure of HBV-infected cancer cell lines

4.2 Detection of viral insertions

4.3 Epigenetic states at inserted regions

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

III

Discussions and conclusion

quote

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

1 Limitations of genomics in infection biology

1.1 Correlation is not causality

1.2 Reproducibility and reliability challenges

Results from genomics analyses are especially sensitive to the parameters and methods used. This makes reproducibility in bioinformatics of utmost importance. Just like RNAseq, Hi-C has important technical variability which needs to be accounted for using multiple replicates.

It was proposed that RNAseq experiments for differential expression analysis should comprise at least 6 replicates and ideally 12 [100]. While this is probably true for most omics experiments, this entails a high cost which is often the limiting factor when designing experiments in genomics.

The core issue with low replicate numbers is the impossibility to distinguish between technical variability due to the assay and biological variability of interest. As a consequence, when fewer replicates are used, lower effect size (fold changes in the case of gene expression) become undetectable. This is especially problematic when studying gene regulation, where small changes in expression can be important.

Unlike RNAseq, where the standard for analyses is well established and most softwares are able to account for replicates and experimental design, most softwares available for Hi-C analysis cannot use replicate information. This limits the power of analysis to detect only very strong changes. The lack of standard also causes a general fragmentation of bioinformatic tools, with many redundant tools of variable quality. One recurrent issue is the absence or low quality of unit tests and documentation, which are unfortunately still not regarded as standard in the computational biology community. Unit tests could be viewed as an equivalent to control experiments in molecular biology, as they validate each logic piece of the software using inputs with known truths. Software lacking these controls is more likely to have undetected bugs that could impact results and lead to false conclusion. Another issue with newer techniques like Hi-C is the lack of standardisation for file formats and practices. This can result in incompatibilities between programs and introduce errors during conversions that alter the data.

Some general practices can of course be adopted to address these issues, such as writing comprehensive documentation, solid tests and maintain software, but as it stands, there is no incentive to do so in academia. Ultimately, these directives need to be enforced globally in the peer review process by journals, so that tools need to meet quality standards to reach publication.

Although this will increase the effort and time required to develop methods, this also means the resulting tools will be more reliable, easier to use and more widely adopted. Fortunately, recent years have seen an increasing adoption of good practices in bioinformatic software and the cool format is now supported by the majority of Hi-C analysis tools. This could mean that the quality of academic software for bioinformatics will undergo major improvements in the forseeable future.

2 Perspectives of genomics for infection biology

2.1 The 3D genome and the advent of deep learning

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

IV

Appendices

A Supplementary information

A.1 Sparse convolution in Chromosight

The explanation below describes how Chromosight reformulates convolution into a matrix multiplication problem to better handle large sparse matrices. The algorithm is inspired from [101]. For brevity, we call the operation "convolution" throughout the section, although cross-correlation could be considered more accurate as we do not transpose the kernel. Let S be the signal (Hi-C) matrix and K the kernel matrix.

$$S = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 4 & 1 \\ 1 & 1 & 3 \end{bmatrix} K = \begin{bmatrix} 10 & 12 \\ 11 & 13 \end{bmatrix} \tag{A.1}$$

The dimensions of the desired convolution output are defined by:

$$(m_S - m_K + 1) \times (n_S - n_K + 1) \tag{A.2}$$

Note this corresponds to a convolution in "valid" mode, where edge values are truncated.

We transform each column of the kernel into a Toeplitz matrix with the same number of columns as the input signal. In this matriix, each value along the diagonals is constant.

$$T_0 = \begin{bmatrix} 10 & 11 & 0 \\ 0 & 10 & 11 \end{bmatrix} \qquad T_1 = \begin{bmatrix} 12 & 13 & 0 \\ 0 & 12 & 13 \end{bmatrix} \tag{A.3}$$

The convolution of the signal and kernel can now be replaced by a sum of dot products between the signal and Toeplitz matrices built from the column filters. For each dot product, the signal is shifted according to the order of filters to respect operations performed during convolution.

$$C = S * K \tag{A.4}$$

$$= S[:, 0 : sn - kn + 1] \cdot T_0 + S[:, 1 : sn - kn + 2] \cdot T_1 \tag{A.5}$$

Where $\cdot$ is the matrix dot product operator and $*$ is the convolution operator. The complete convolution algorithm used in chromosight is given as pseudocode in algorithm 1.

Algorithm 1 Calculate $C = S * K$ using matrix products

Require: S, $m_S \times n_S$ matrix
Require: K, $m_K \times n_K$ matrix
Ensure: $m_S >= m_K, n_S >= n_K$
  Let $\{T_0, ..., T_{n_K}\}$ be $m_K \times n_S$ matrices
  $y \leftarrow 0$
  while $y \neq n_K$ **do**
    $t \leftarrow 0$
    while $t \neq n_S$ **do**
      $T_y[t, :] \leftarrow K[:, t]^T$
    end while
  end while
  Let C be a $(m_S - m_K + 1) \times (n_S - n_K + 1)$ matrix
  $C \leftarrow \sum_{i=0}^{n_K} T_{n_K} \cdot S[:, i : sn - kn + 1 + kj]$ {Signal shifted according to each filter}

Bibliography

[1] N. C. Johnson, J. H. Graham, and F. A. Smith. „Functioning of Mycorrhizal Associations along the Mutualism-Parasitism Continuum“. en. In: *New Phytologist* 135.4 (Apr. 1997), pp. 575–585 (cit. on p. 2).

[2] Marc-André Selosse, Franck Richard, Xinhua He, and Suzanne W. Simard. „Mycorrhizal Networks: Des Liaisons Dangereuses?“ en. In: *Trends in Ecology & Evolution* 21.11 (Nov. 2006), pp. 621–628 (cit. on p. 2).

[3] Jonathan Knight. „Meet the Herod Bug“. en. In: *Nature* 412.6842 (July 2001), pp. 12–14 (cit. on p. 2).

[4] R. Stouthamer, J. A. J. Breeuwer, R. F. Luck, and J. H. Werren. „Molecular Identification of Microorganisms Associated with Parthenogenesis“. en. In: *Nature* 361.6407 (Jan. 1993), pp. 66–68 (cit. on p. 2).

[5] L. M. Hedges, J. C. Brownlie, S. L. O'Neill, and K. N. Johnson. „Wolbachia and Virus Protection in Insects“. en. In: *Science* 322.5902 (Oct. 2008), pp. 702–702 (cit. on p. 3).

[6] Luís Teixeira, Álvaro Ferreira, and Michael Ashburner. „The Bacterial Symbiont Wolbachia Induces Resistance to RNA Viral Infections in Drosophila Melanogaster“. en. In: *PLoS Biology* 6.12 (Dec. 2008). Ed. by Laurent Keller, e1000002 (cit. on p. 3).

[7] N. Nikoh, T. Hosokawa, M. Moriyama, et al. „Evolutionary Origin of Insect-Wolbachia Nutritional Mutualism“. en. In: *Proceedings of the National Academy of Sciences* 111.28 (July 2014), pp. 10257–10262 (cit. on p. 3).

[8] John P. McCutcheon and Nancy A. Moran. „Extreme Genome Reduction in Symbiotic Bacteria“. en. In: *Nature Reviews Microbiology* 10.1 (Jan. 2012), pp. 13–26 (cit. on p. 3).

[9] Soma Ghosh and Tamara J. O'Connor. „Beyond Paralogs: The Multiple Layers of Redundancy in Bacterial Pathogenesis“. en. In: *Frontiers in Cellular and Infection Microbiology* 7 (Nov. 2017), p. 467 (cit. on p. 3).

[10] U. Bergthorsson, D. I. Andersson, and J. R. Roth. „Ohno's Dilemma: Evolution of New Genes under Continuous Selection“. en. In: *Proceedings of the National Academy of Sciences* 104.43 (Oct. 2007), pp. 17004–17009 (cit. on p. 3).

[11] T. Dagan, Y. Artzy-Randrup, and W. Martin. „Modular Networks and Cumulative Impact of Lateral Transfer in Prokaryote Genome Evolution“. en. In: *Proceedings of the National Academy of Sciences* 105.29 (July 2008), pp. 10039–10044 (cit. on p. 4).

[12] Julia Van Etten and Debashish Bhattacharya. „Horizontal Gene Transfer in Eukaryotes: Not If, but How Much?“ en. In: *Trends in Genetics* 36.12 (Dec. 2020), pp. 915–925 (cit. on p. 4).

[13] David T John and Marsha J. Howard. „Seasonal Distribution of Pathogenic Free-Living Amebae in Oklahoma Waters“. In: *Parasitology Research* 80 (1995), pp. 193–201 (cit. on p. 4).

[14] Ryota Sakamoto, Akira Ohno, Toshitaka Nakahara, et al. „Legionella Pneumophila in Rainwater on Roads“. en. In: *Emerging Infectious Diseases* 15.8 (Aug. 2009), pp. 1295–1297 (cit. on p. 4).

[15] Sutherland K. Maciver. „Asexual Amoebae Escape Muller's Ratchet through Polyploidy“. en. In: *Trends in Parasitology* 32.11 (Nov. 2016), pp. 855–862 (cit. on p. 4).

[16] Michael Clarke, Amanda J Lohan, Bernard Liu, et al. „Genome of Acanthamoeba Castellanii Highlights Extensive Lateral Gene Transfer and Early Evolution of Tyrosine Kinase Signaling“. en. In: *Genome Biology* 14.2 (2013), R11 (cit. on p. 4).

[17] T J Rowbotham. „Preliminary Report on the Pathogenicity of Legionella Pneumophila for Freshwater and Soil Amoebae.“ en. In: *Journal of Clinical Pathology* 33.12 (Dec. 1980), pp. 1179–1183 (cit. on p. 5).

[18] Paul H. Edelstein and Craig R. Roy. „Legionnaires' Disease and Pontiac Fever“. In: *Mandell, Douglas, and Bennett's Principles and Practice of Infectious Diseases*. Vol. 2. 2014, pp. 2633–2644 (cit. on p. 5).

[19] Ana M. Correia, Joana S. Ferreira, Vítor Borges, et al. „Probable Person-to-Person Transmission of Legionnaires' Disease“. en. In: *New England Journal of Medicine* 374.5 (Feb. 2016), pp. 497–498 (cit. on p. 5).

[20] Joseph E. McDade. „Legionella and the Prevention of Legionellosis“. In: *Emerging Infectious Diseases*. Vol. 14. 2008, 1006a–1006 (cit. on p. 5).

[21] Karim Suwwan de Felipe, Sergey Pampou, Oliver S. Jovanovic, et al. „Evidence for Acquisition of Legionella Type IV Secretion Substrates via Interdomain Horizontal Gene Transfer“. en. In: *Journal of Bacteriology* 187.22 (Nov. 2005), pp. 7716–7726 (cit. on p. 5).

[22] Hubert Hilbi, Christine Hoffmann, and Christopher F. Harrison. „Legionella Spp. Outdoors: Colonization, Communication and Persistence“. en. In: *Environmental Microbiology Reports* 3.3 (2011), pp. 286–296 (cit. on p. 6).

[23] Michael Steinert, Ute Hentschel, and Jörg Hacker. „Legionella Pneumophila: An Aquatic Microbe Goes Astray“. In: *FEMS Microbiology Reviews* 26.2 (June 2002), pp. 149–162 (cit. on p. 6).

[24] Brenda Byrne and Michele S. Swanson. „Expression of Legionella pneumophilaVirulence Traits in Response to Growth Conditions“. en. In: *Infection and Immunity* 66.7 (July 1998), pp. 3029–3034 (cit. on p. 6).

[25] Holger Brüggemann, Arne Hagman, Matthieu Jules, et al. „Virulence Strategies for Infecting Phagocytes Deduced from the in Vivo Transcriptional Program of Legionella Pneumophila“. en. In: *Cellular Microbiology* 8.8 (2006), pp. 1228–1240 (cit. on p. 6).

[26] Giulia Oliva, Tobias Sahr, and Carmen Buchrieser. „The Life Cycle of L. Pneumophila: Cellular Differentiation Is Linked to Virulence and Metabolism“. en. In: *Frontiers in Cellular and Infection Microbiology* 8 (Jan. 2018), p. 3 (cit. on p. 6).

[27] Tobias Sahr, Christophe Rusniok, Francis Impens, et al. „The Legionella Pneumophila Genome Evolved to Accommodate Multiple Regulatory Mechanisms Controlled by the CsrA-System“. en. In: *PLOS Genetics* 13.2 (Feb. 2017). Ed. by Jörg Vogel, e1006629 (cit. on p. 6).

[28] Hagen Wieland, Susanne Ullrich, Florian Lang, and Birgid Neumeister. „Intracellular Multiplication of Legionella Pneumophila Depends on Host Cell Amino Acid Transporter SLC1A5“. en. In: *Molecular Microbiology* 55.5 (2005), pp. 1528–1537 (cit. on p. 6).

[29] Christopher T. Price, Souhaila Al-Khodor, Tasneem Al-Quadan, et al. „Molecular Mimicry by an F-Box Effector of Legionella Pneumophila Hijacks a Conserved Polyubiquitination Machinery within Macrophages and Protozoa“. en. In: *PLOS Pathogens* 5.12 (Dec. 2009), e1000704 (cit. on p. 7).

[30] Justin A. De Leon, Jiazhang Qiu, Christopher J. Nicolai, et al. „Positive and Negative Regulation of the Master Metabolic Regulator mTORC1 by Two Families of Legionella Pneumophila Effectors“. en. In: *Cell Reports* 21.8 (Nov. 2017), pp. 2031–2038 (cit. on p. 7).

[31] Ralph R. Isberg, Tamara J. O'Connor, and Matthew Heidtman. „The Legionella Pneumophila Replication Vacuole: Making a Cosy Niche inside Host Cells“. en. In: *Nature Reviews Microbiology* 7.1 (Jan. 2009), pp. 13–24 (cit. on p. 7).

[32] Yao Liu, Wenhan Zhu, Yunhao Tan, et al. „A Legionella Effector Disrupts Host Cytoskeletal Structure by Cleaving Actin“. en. In: *PLOS Pathogens* 13.1 (Jan. 2017), e1006186 (cit. on p. 7).

[33] Irina Saraiva Franco, Nadim Shohdy, and Howard A. Shuman. „The Legionella Pneumophila Effector VipA Is an Actin Nucleator That Alters Host Cell Organelle Trafficking“. en. In: *PLOS Pathogens* 8.2 (Feb. 2012), e1002546 (cit. on p. 7).

[34] Rita K. Laguna, Elizabeth A. Creasey, Zhiru Li, Nicole Valtz, and Ralph R. Isberg. „A Legionella Pneumophila-Translocated Substrate That Is Required for Growth within Macrophages and Protection from Host Cell Death“. en. In: *Proceedings of the National Academy of Sciences* 103.49 (Dec. 2006), pp. 18745–18750 (cit. on p. 7).

[35] Monica Rolando, Serena Sanulli, Christophe Rusniok, et al. „Legionella Pneumophila Effector RomA Uniquely Modifies Host Chromatin to Repress Gene Expression and Promote Intracellular Bacterial Replication“. en. In: *Cell Host & Microbe* 13.4 (Apr. 2013), pp. 395–405 (cit. on p. 7).

[36] S. Uzzau, D. J. Brown, T. Wallis, et al. „Host Adapted Serotypes of *Salmonella Enterica*“. en. In: *Epidemiology and Infection* 125.2 (Oct. 2000), pp. 229–255 (cit. on p. 7).

[37] Doris L. LaRock, Anu Chaudhary, and Samuel I. Miller. „Salmonellae Interactions with Host Processes“. en. In: *Nature Reviews Microbiology* 13.4 (Apr. 2015), pp. 191–205 (cit. on pp. 7, 8).

[38] Tikki Pang, Zulfiqar A. Bhutta, B.Brett Finlay, and Martin Altwegg. „Typhoid Fever and Other Salmonellosis: A Continuing Challenge“. en. In: *Trends in Microbiology* 3.7 (July 1995), pp. 253–255 (cit. on p. 8).

[39] Margarita M Ochoa-Díaz, Silvana Daza-Giovannetty, and Doris Gómez-Camargo. „Bacterial Genotyping Methods: From the Basics to Modern“. In: *Host-Pathogen Interactions: Methods and Protocols*. Ed. by Carlos Medina and Francisco Javier López-Baena. New York, NY: Springer New York, 2018, pp. 13–20 (cit. on p. 10).

[40] Monica Rolando and Carmen Buchrieser. „Legionella Pneumophila Type IV Effectors Hijack the Transcription and Translation Machinery of the Host Cell“. en. In: *Trends in Cell Biology* 24.12 (Dec. 2014), pp. 771–778 (cit. on p. 11).

[41] R. Langridge, H.R. Wilson, C.W. Hooper, M.H.F. Wilkins, and L.D. Hamilton. „The Molecular Configuration of Deoxyribonucleic Acid“. en. In: *Journal of Molecular Biology* 2.1 (Apr. 1960), 19–IN11 (cit. on p. 13).

[42] J. Dekker. „Capturing Chromosome Conformation“. en. In: *Science* 295.5558 (Feb. 2002), pp. 1306–1311 (cit. on p. 13).

[43] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, et al. „Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome“. en. In: *Science* 326.5950 (Oct. 2009), pp. 289–293 (cit. on pp. 13, 14).

[44] Geoffrey Fudenberg, Maxim Imakaev, Carolyn Lu, et al. „Formation of Chromosomal Domains by Loop Extrusion“. en. In: *Cell Reports* 15.9 (May 2016), pp. 2038–2049 (cit. on p. 14).

[45] Elphège P. Nora, Bryan R. Lajoie, Edda G. Schulz, et al. „Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre“. en. In: *Nature* 485.7398 (May 2012), pp. 381–385 (cit. on p. 14).

[46] Bas van Steensel and Andrew S. Belmont. „Lamina-Associated Domains: Links with Chromosome Architecture, Heterochromatin, and Gene Repression“. en. In: *Cell* 169.5 (May 2017), pp. 780–791 (cit. on p. 15).

[47] Bryan R. Lajoie, Job Dekker, and Noam Kaplan. „The Hitchhiker's Guide to Hi-C Analysis: Practical Guidelines“. en. In: *Methods* 72 (Jan. 2015), pp. 65–75 (cit. on p. 15).

[48] Sergey Venev, Nezar Abdennur, Anton Goloborodko, et al. *Open2c/Cooltools: V0.4.0*. Zenodo. Apr. 2021 (cit. on p. 15).

[49] Emily Crane, Qian Bian, Rachel Patton McCord, et al. „Condensin-Driven Remodelling of X Chromosome Topology during Dosage Compensation“. en. In: *Nature* 523.7559 (July 2015), pp. 240–244 (cit. on p. 15).

[50] Fengling Chen, Guipeng Li, Michael Q Zhang, and Yang Chen. „HiCDB: A Sensitive and Robust Method for Detecting Contact Domain Boundaries“. In: *Nucleic Acids Research* 46.21 (Nov. 2018), pp. 11239–11250 (cit. on pp. 16, 21).

[51] Neva C. Durand, James T. Robinson, Muhammad S. Shamim, et al. „Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom“. en. In: *Cell Systems* 3.1 (July 2016), pp. 99–101 (cit. on p. 16).

[52] Fidel Ramírez, Vivek Bhardwaj, Laura Arrigoni, et al. „High-Resolution TADs Reveal DNA Sequences Underlying Genome Organization in Flies“. en. In: *Nature Communications* 9.1 (Jan. 2018), p. 189 (cit. on p. 16).

[53] Ilya M Flyamer, Robert S Illingworth, and Wendy A Bickmore. „Coolpup.Py: Versatile Pile-up Analysis of Hi-C Data“. In: *Bioinformatics* 36.10 (May 2020), pp. 2980–2985 (cit. on p. 16).

[54] Oana Ursu, Nathan Boley, Maryna Taranova, et al. „GenomeDISCO: A Concordance Score for Chromosome Conformation Capture Experiments Using Random Walks on Contact Map Graphs“. In: *Bioinformatics* 34.16 (Aug. 2018), pp. 2701–2707 (cit. on p. 16).

[55] Tao Yang, Feipeng Zhang, Galip Gürkan Yardımcı, et al. „HiCRep: Assessing the Reproducibility of Hi-C Data Using a Stratum-Adjusted Correlation Coefficient“. en. In: (), p. 37 (cit. on p. 16).

[56] Koon-Kiu Yan, Galip Gürkan Yardımcı, Chengfei Yan, William S Noble, and Mark Gerstein. „HiC-Spector: A Matrix Library for Spectral and Reproducibility Analysis of Hi-C Contact Maps“. In: *Bioinformatics* 33.14 (July 2017), pp. 2199–2201 (cit. on p. 16).

[57] Aaron T.L. Lun and Gordon K. Smyth. „diffHic: A Bioconductor Package to Detect Differential Genomic Interactions in Hi-C Data“. In: *BMC Bioinformatics* 16.1 (Aug. 2015), p. 258 (cit. on p. 17).

[58] John C Stansfield, Kellen G Cresswell, and Mikhail G Dozmorov. „multiHiCcompare: Joint Normalization and Comparative Analysis of Complex Hi-C Experiments“. In: *Bioinformatics* 35.17 (Sept. 2019), pp. 2916–2923 (cit. on p. 17).

[59] Sven Heinz, Christopher Benner, Nathanael Spann, et al. „Simple Combinations of Lineage-Determining Transcription Factors Prime Cis-Regulatory Elements Required for Macrophage and B Cell Identities“. en. In: *Molecular Cell* 38.4 (May 2010), pp. 576–589 (cit. on p. 17).

[60] A. Zemach, I. E. McDaniel, P. Silva, and D. Zilberman. „Genome-Wide Evolutionary Analysis of Eukaryotic DNA Methylation“. en. In: *Science* 328.5980 (May 2010), pp. 916–919 (cit. on p. 17).

[61] Miao Wang, Chen Guo, Liang Wang, et al. „Long Noncoding RNA GAS5 Promotes Bladder Cancer Cells Apoptosis through Inhibiting EZH2 Transcription“. en. In: *Cell Death & Disease* 9.2 (Feb. 2018), p. 238 (cit. on p. 17).

[62] Cynthia M. Sharma, Fabien Darfeuille, Titia H. Plantinga, and Jörg Vogel. „A Small RNA Regulates Multiple ABC Transporter mRNAs by Targeting C/A-Rich Elements inside and Upstream of Ribosome-Binding Sites“. In: *Genes & Development* 21.21 (Nov. 2007), pp. 2804–2817 (cit. on p. 17).

[63] Branislav Večerek, Isabella Moll, and Udo Bläsi. „Control of Fur Synthesis by the Non-Coding RNA RyhB and Iron-Responsive Decoding“. In: *The EMBO Journal* 26.4 (Feb. 2007), pp. 965–975 (cit. on p. 17).

[64] F. Sanger, S. Nicklen, and A. R. Coulson. „DNA Sequencing with Chain-Terminating Inhibitors“. en. In: *Proceedings of the National Academy of Sciences* 74.12 (Dec. 1977), pp. 5463–5467 (cit. on p. 22).

[65] F. Sanger, A.R. Coulson, G.F. Hong, D.F. Hill, and G.B. Petersen. „Nucleotide Sequence of Bacteriophage λ DNA“. en. In: *Journal of Molecular Biology* 162.4 (Dec. 1982), pp. 729–773 (cit. on p. 22).

[66] R. Baer, A. T. Bankier, M. D. Biggin, et al. „DNA Sequence and Expression of the B95-8 Epstein—Barr Virus Genome“. en. In: *Nature* 310.5974 (July 1984), pp. 207–211 (cit. on p. 22).

[67] S. G. Oliver, Q. J. M. van der Aart, M. L. Agostoni-Carbone, et al. „The Complete DNA Sequence of Yeast Chromosome III“. en. In: *Nature* 357.6373 (May 1992), pp. 38–46 (cit. on p. 22).

[68] Agnès Thierry, Cécile Fairhead, and Bernard Dujon. „The Complete Sequence of the 8 · 2 Kb Segment Left of MAT on Chromosome III Reveals Five ORFs, Including a Gene for a Yeast Ribokinase“. en. In: *Yeast* 6.6 (1990), pp. 521–534 (cit. on p. 22).

[69] F Collins and D Galas. „A New Five-Year Plan for the U.S. Human Genome Project“. en. In: *Science* 262.5130 (Oct. 1993), pp. 43–46 (cit. on p. 22).

[70] M. D. Adams. „The Genome Sequence of Drosophila Melanogaster“. In: *Science* 287.5461 (Mar. 2000), pp. 2185–2195 (cit. on p. 22).

[71] J. Craig Venter, Mark D. Adams, Eugene W. Myers, et al. „The Sequence of the Human Genome“. en. In: *Science* 291.5507 (Feb. 2001), pp. 1304–1351 (cit. on p. 23).

[72] Phillip Compeau, P. Pevzner, and G. Tesler. „How to Apply de Bruijn Graphs to Genome Assembly.“ In: *Nature biotechnology* (2011) (cit. on pp. 23, 24).

[73] J. T. Simpson, K. Wong, S. D. Jackman, et al. „ABySS: A Parallel Assembler for Short Read Sequence Data“. en. In: *Genome Research* 19.6 (June 2009), pp. 1117–1123 (cit. on p. 23).

[74] D. R. Zerbino and E. Birney. „Velvet: Algorithms for de Novo Short Read Assembly Using de Bruijn Graphs“. en. In: *Genome Research* 18.5 (Feb. 2008), pp. 821–829 (cit. on p. 23).

[75] Jared T. Simpson and Mihai Pop. „The Theory and Practice of Genome Sequence Assembly“. en. In: *Annual Review of Genomics and Human Genetics* 16.1 (Aug. 2015), pp. 153–172 (cit. on p. 23).

[76] Karen H. Miga, Sergey Koren, Arang Rhie, et al. „Telomere-to-Telomere Assembly of a Complete Human X Chromosome“. en. In: *Nature* 585.7823 (Sept. 2020), pp. 79–84 (cit. on p. 23).

[77] Glennis A. Logsdon, Mitchell R. Vollger, PingHsun Hsieh, et al. „The Structure, Function and Evolution of a Complete Human Chromosome 8“. en. In: *Nature* (Apr. 2021), pp. 1–7 (cit. on p. 23).

[78] Jason L Weirather, Mariateresa de Cesare, Yunhao Wang, et al. „Comprehensive Comparison of Pacific Biosciences and Oxford Nanopore Technologies and Their Applications to Transcriptome Analysis“. In: *F1000Research* 6 (June 2017) (cit. on p. 23).

[79] Miten Jain, Sergey Koren, Karen H. Miga, et al. „Nanopore Sequencing and Assembly of a Human Genome with Ultra-Long Reads“. en. In: *Nature Biotechnology* 36.4 (Apr. 2018), pp. 338–345 (cit. on p. 23).

[80] Pierre Morisse, Camille Marchet, Antoine Limasset, Thierry Lecroq, and Arnaud Lefebvre. „Scalable Long Read Self-Correction and Assembly Polishing with Multiple Sequence Alignment“. en. In: *Scientific Reports* 11.1 (Jan. 2021), p. 761 (cit. on p. 23).

[81] Jeremy R. Wang, James Holt, Leonard McMillan, and Corbin D. Jones. „FMLRC: Hybrid Long Read Error Correction Using an FM-Index“. In: *BMC Bioinformatics* 19.1 (Feb. 2018), p. 50 (cit. on p. 23).

[82] Guillaume Holley, Doruk Beyter, Helga Ingimundardottir, et al. „Ratatosk: Hybrid Error Correction of Long Reads Enables Accurate Variant Calling and Assembly“. In: *Genome Biology* 22.1 (Jan. 2021), p. 28 (cit. on p. 25).

[83] Robert Vaser, Ivan Sović, Niranjan Nagarajan, and Mile Šikić. „Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads“. en. In: *Genome Research* 27.5 (Jan. 2017), pp. 737–746 (cit. on p. 25).

[84] Bruce J. Walker, Thomas Abeel, Terrance Shea, et al. „Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement“. en. In: *PLOS ONE* 9.11 (Nov. 2014), e112963 (cit. on p. 25).

[85] Ritu Kundu, Joshua Casey, and Wing-Kin Sung. *HyPo: Super Fast & Accurate Polisher for Long Read Genome Assemblies*. en. Preprint. Bioinformatics, Dec. 2019 (cit. on p. 25).

[86] Ernest T Lam, Alex Hastie, Chin Lin, et al. „Genome Mapping on Nanochannel Arrays for Structural Variation Analysis and Sequence Assembly“. en. In: *Nature Biotechnology* 30.8 (Aug. 2012), pp. 771–776 (cit. on p. 25).

[87] Matt Ravenhall, Nives Škunca, Florent Lassalle, and Christophe Dessimoz. „Inferring Horizontal Gene Transfer“. en. In: *PLOS Computational Biology* 11.5 (May 2015). Ed. by Shoshana Wodak, e1004095 (cit. on p. 26).

[88] Genome 10K Community of Scientists. „Genome 10K: A Proposal to Obtain Whole-Genome Sequence for 10 000 Vertebrate Species“. en. In: *Journal of Heredity* 100.6 (Nov. 2009), pp. 659–674 (cit. on p. 26).

[89] Monica Poelchau, Christopher Childers, Gary Moore, et al. „The I5k Workspace@NAL—Enabling Genomic Data Access, Visualization and Curation of Arthropod Genomes“. en. In: *Nucleic Acids Research* 43.D1 (Jan. 2015), pp. D714–D719 (cit. on p. 26).

[90] *Darwin Tree Of Life*. https://www.darwintreeoflife.org/ (cit. on p. 26).

[91] Deanna M Church, Valerie A Schneider, Karyn Meltz Steinberg, et al. „Extending Reference Assembly Models“. en. In: *Genome Biology* 16.1 (Dec. 2015), p. 13 (cit. on p. 27).

[92] Heng Li, Xiaowen Feng, and Chong Chu. „The Design and Construction of Reference Pangenome Graphs with Minigraph“. en. In: *Genome Biology* 21.1 (Dec. 2020), p. 265 (cit. on p. 27).

[93] Bryce van de Geijn, Graham McVicker, Yoav Gilad, and Jonathan K Pritchard. „WASP: Allele-Specific Software for Robust Molecular Quantitative Trait Locus Discovery“. en. In: *Nature Methods* 12.11 (Nov. 2015), pp. 1061–1063 (cit. on p. 27).

[94] Axel Cournac, Hervé Marie-Nelly, Martial Marbouty, Romain Koszul, and Julien Mozziconacci. „Normalization of a Chromosomal Contact Map“. en. In: *BMC Genomics* 13.1 (2012), p. 436 (cit. on p. 31).

[95] Cyril Matthey-Doret, Lyam Baudry, Amaury Bignaud, et al. *Simple Library/Pipeline to Generate and Handle Hi-C Data*. https://github.com/koszullab/hicstuff. Mar. 2021 (cit. on p. 32).

[96] Suhas S.P. Rao, Miriam H. Huntley, Neva C. Durand, et al. „A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping“. en. In: *Cell* 159.7 (Dec. 2014), pp. 1665–1680 (cit. on p. 34).

[97] Elphège P. Nora, Anton Goloborodko, Anne-Laure Valton, et al. „Targeted Degradation of CTCF Decouples Local Insulation of Chromosome Domains from Genomic Compartmentalization“. en. In: *Cell* 169.5 (May 2017), 930–944.e22 (cit. on pp. 67–69).

[98] Suhas S.P. Rao, Su-Chen Huang, Brian Glenn St Hilaire, et al. „Cohesin Loss Eliminates All Loop Domains“. en. In: *Cell* 171.2 (Oct. 2017), 305–320.e24 (cit. on p. 68).

[99] David L. Rimm, Thomas D. Pollard, and Philip Hieter. „Resolution of Acanthamoeba Castellanii Chromosomes by Pulsed Field Gel Electrophoresis and Construction of the Initial Linkage Map“. en. In: *Chromosoma* 97.3 (Nov. 1988), pp. 219–223 (cit. on p. 70).

[100] Nicholas J. Schurch, Pietá Schofield, Marek Gierliński, et al. „How Many Biological Replicates Are Needed in an RNA-Seq Experiment and Which Differential Expression Tool Should You Use?“ en. In: *RNA* 22.6 (Jan. 2016), pp. 839–851 (cit. on p. 78).

[101] *Neural Network - 2-D Convolution as a Matrix-Matrix Multiplication*. https://stackoverflow.com/question d-convolution-as-a-matrix-matrix-multiplication (cit. on p. 82).

Colophon

This thesis was typeset with LaTeX 2_ε. It uses the *Clean Thesis* style developed by Ricardo Langner. The design of the *Clean Thesis* style is inspired by user guide documents from Apple Inc.

Download the *Clean Thesis* style at http://cleanthesis.der-ric.de/.

Declaration

Je soussigné Cyril Matthey-Doret certifie que le manuscrit présenté en vue de la soutenance est le fruit d'un travail original et que toutes les sources utilisées ont été clairement indiquées.

Je certifie, de surcroît, que je n'ai ni copié ni utilisé des idées ou des formulations tirées d'un ouvrage, article ou mémoire, en version imprimée ou électronique, sans mentionner précisément leur origine et que les citations sont expressément signalées entre guillemets (ou par une autre disposition graphique sans ambiguïté).

Conformément à la loi, le non-respect de ces dispositions me rend passible de poursuites devant la commission disciplinaire et les tribunaux de la République française pour plagiat universitaire.

Fait à Paris le 1er Octobre 2021

Cyril Matthey-Doret